SUPPLEMENTARY INFORMATION
Design and Analysis of Large-Scale Biological Rhythm Studies: A Comparison of Algorithms for Detecting Periodic Signals in Biological Data

Anastasia Deckard[1,3*], Ron C. Anafi[4], John B. Hogenesch[5],
Steven B. Haase[2], John Harer[3]

[1]Program in Computational Biology and Bioinformatics, [2]Department of Biology, and [3]Department of Mathematics, Duke University, Durham, NC, USA; [4]Department of Medicine and [5]Department of Pharmacology, Institute for Translational Medicine and Therapeutics, University of Pennsylvania School of Medicine, Philadelphia, PA, USA.

Analysis	Time	# Samples	Period	Amp (peak-to-trough)	Phase Shift	Shapes	Gauss. Noise Levels	Profiles per Case	# Periods	Samples per Period
Noise, Resolution, and Shape	0-200	**50, 25, 17**	100	50 (100)	0 - per	cos cos2: per= 0.3333 * per, amp= 0.5 * amp, pshift= (pshift + 0.083) % per damp: 0.01 peak = 20 trend lin: 0.5 trend exp: 0.027 flat linear: (-0.5, 0.5) (see functions table for details)	***SD* = 0, 25, 50**	100	2	25, 12.5, 9
Recover Period	0-200	50	**50-150**	50 (100)	0 - per		***SD* = 0, 25, 50**	50	4 - 1.5	12.5 - 33
Recover Phase Shift	0-200	50	100	50 (100)	**0 - per**		***SD* = 0, 25, 50**	50	2	25
Recover Amplitude	0-200	50	100	**25-75 (50-150)**	0 - per		***SD* = 0, 25, 50**	50	2	25

Table S1: Synthetic data sets generated for each analysis. Each analysis used a set of synthetic data designed to capture behaviors of interest. Each set of synthetic data was made of several cases (shown in bold in the table).

Shapes	**Function**
cos	amp * cos(2*pi/per * t - pshift*(2*pi/per))
cos 2	per2 = per * 0.3333 amp2 = amp * 0.50 pshift2 = (pshift + (per2 * 0.25)) % per amp * cos(2*pi/per * (t - pshift)) + amp2 * cos(2*pi/per2 * (t - pshift2))
damp	amp * cos(2*pi/per * t - pshift*(2*pi/per)) * exp(-damp*t)
peak	amp * (-1 + 2 * fabs(cos(pi/per * t - pshift*(pi/per)))**peak)
trend exp	amp * cos(2*pi/per * (t - pshift)) + exp(trendexp * t)
trend	amp * cos(2*pi/per * t - pshift*(2*pi/per)) + (trend * t)
flat	0
linear	(slope * t)

Table S2: Functions of time (t) used to generate profiles. The types of periodic profiles are: cosine (cos), cosine two periods (cos 2), cosine damped (damp), cosine peaked (peak), cosine exponential trend (trend exp), and cosine linear trend (trend). The values for amplitude (amp), period (per), and phase shift (pshift) are selected from a uniform distribution within the defined minimum and maximum. For phase shift, the range is 0 to the period length. The values for the level of transformation for damp, peak, and trend are defined for a given set.

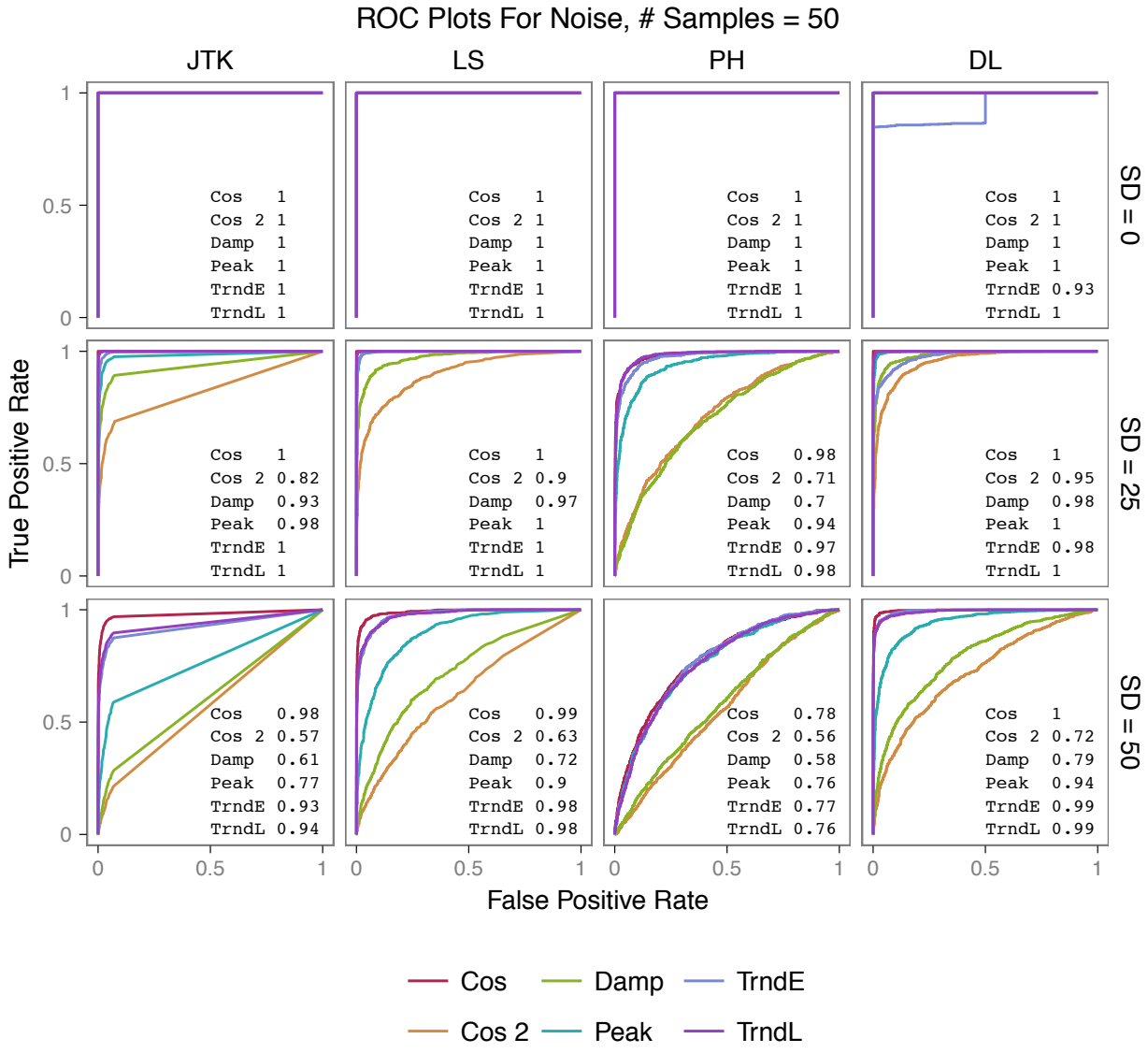


Figure S1: Algorithm performance on identifying periodic versus non-periodic profiles for different profile shapes and noise levels for 50 samples per profile, 1000 profiles per case. Receiver Operator Characteristic (ROC) plots shown with Area Under Curve (AUC). Performance degradation under increasing Gaussian noise with standard deviation = {0, 25, 50}. Used -ln (p-value or score).

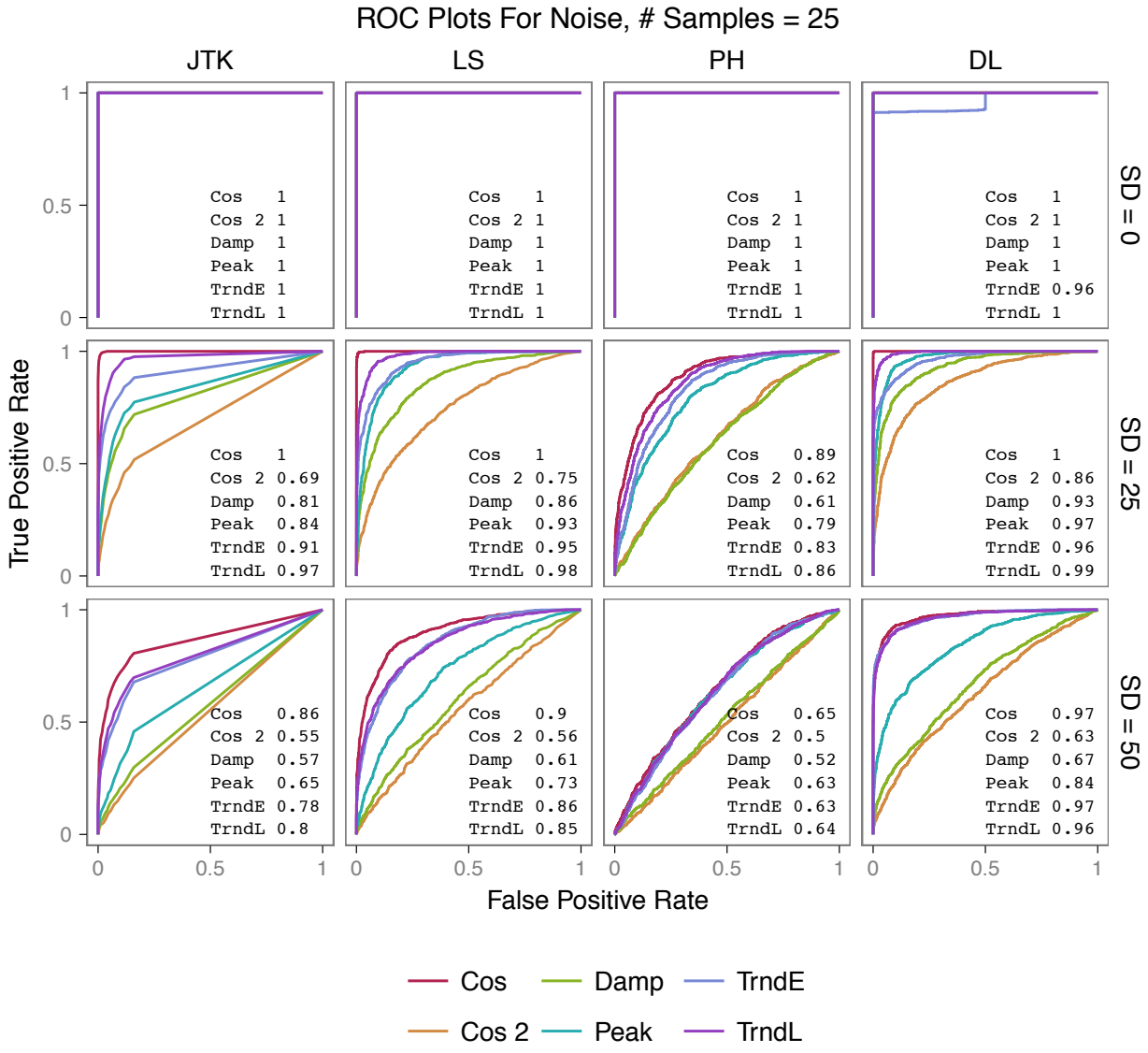


Figure S2: Algorithm performance on identifying periodic versus non-periodic profiles for different profile shapes and noise levels for 25 samples, 1000 profiles per case. Receiver Operator Characteristic (ROC) plots shown with Area Under Curve (AUC). Performance degradation under increasing Gaussian noise with standard deviation = {0, 25, 50}. Used -ln (p-value or score).

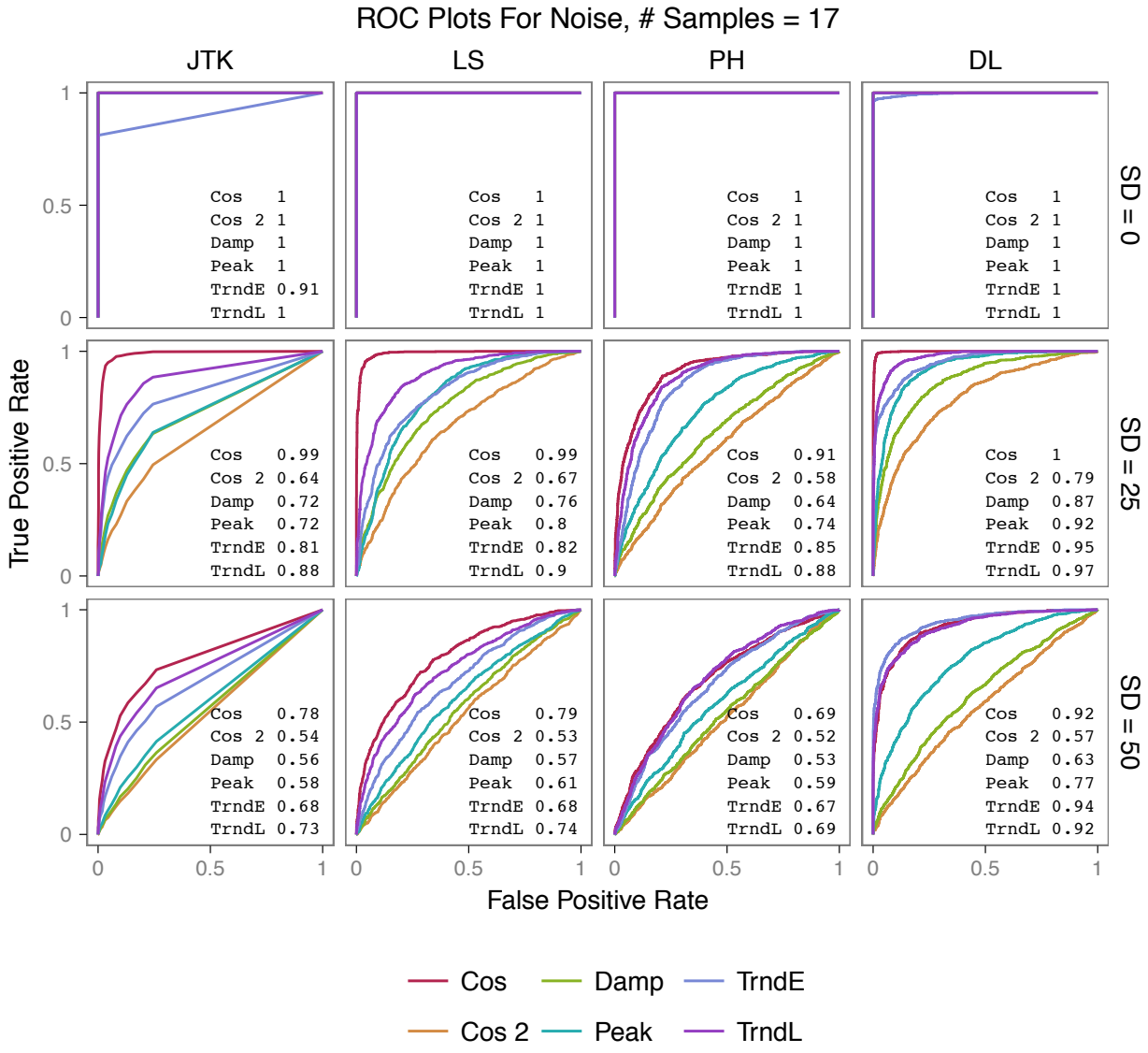


Figure S3: Algorithm performance on identifying periodic versus non-periodic profiles for different profile shapes and noise levels for 17 samples, 1000 profiles per case. Receiver Operator Characteristic (ROC) plots shown with Area Under Curve (AUC). Performance degradation under increasing Gaussian noise with standard deviation = {0, 25, 50}. Used -ln (p-value or score).

	# Samples for Two Periods, No Noise		
	50	**25**	**17**
LS p-value	1.26E-09	1.48E-04	4.74E-03
JTK p-value	2.44E-49	1.47E-17	2.31E-09
PH score	1.00E+00	1.00E+00	1.00E+00

Table S3: The effect of different sampling rates on p-values or scores. Scores are for the identical synthetic cosine curve with two full periods and no noise, but with number of samples = {50, 25, 17}. Both Lomb-Scargle and JTK CYCLE return p- values, and these methods and their statistical tests are affected by number of samples: a profile will receive lower p-values as the number of samples increases. Persistent Homology, however, does not use the number of samples when it computes scores; therefore, the score will not vary in relation to the number of time points.

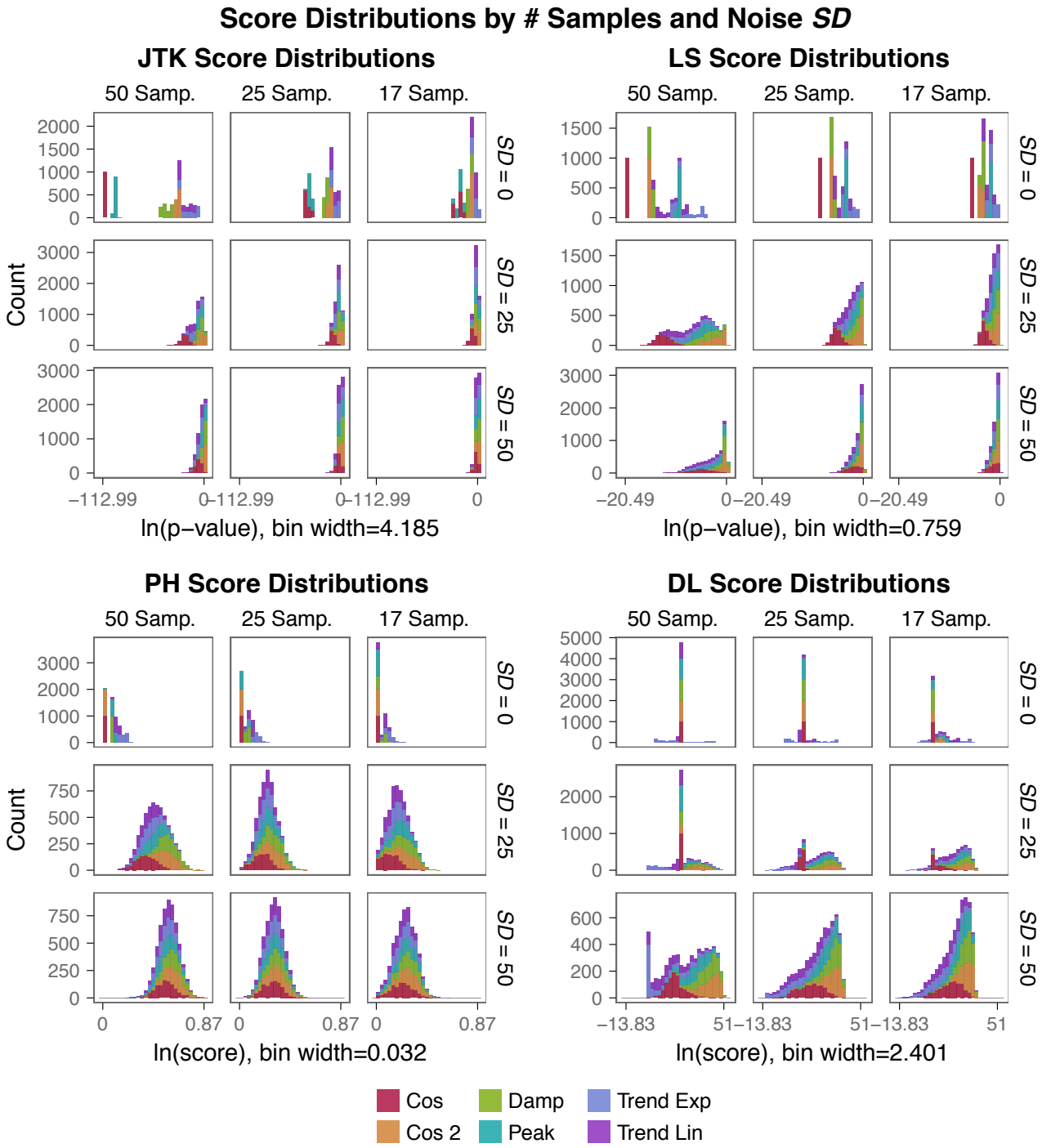


Figure S4: Algorithm biases for profile shapes. Histograms display scores returned for each different shape. Distributions of scores are by shape, with plots for data sets with difference number of samples = {50, 25, 17} and noise levels (Gaussian Noise SD = {0, 25, 50}. The same data set used in the ROC analysis was used. The x-axis shows the scores, log transformed, ranging from the lowest (best score) to the highest (worst score) returned by the algorithm. The y-axis shows the number of profiles receiving the score.

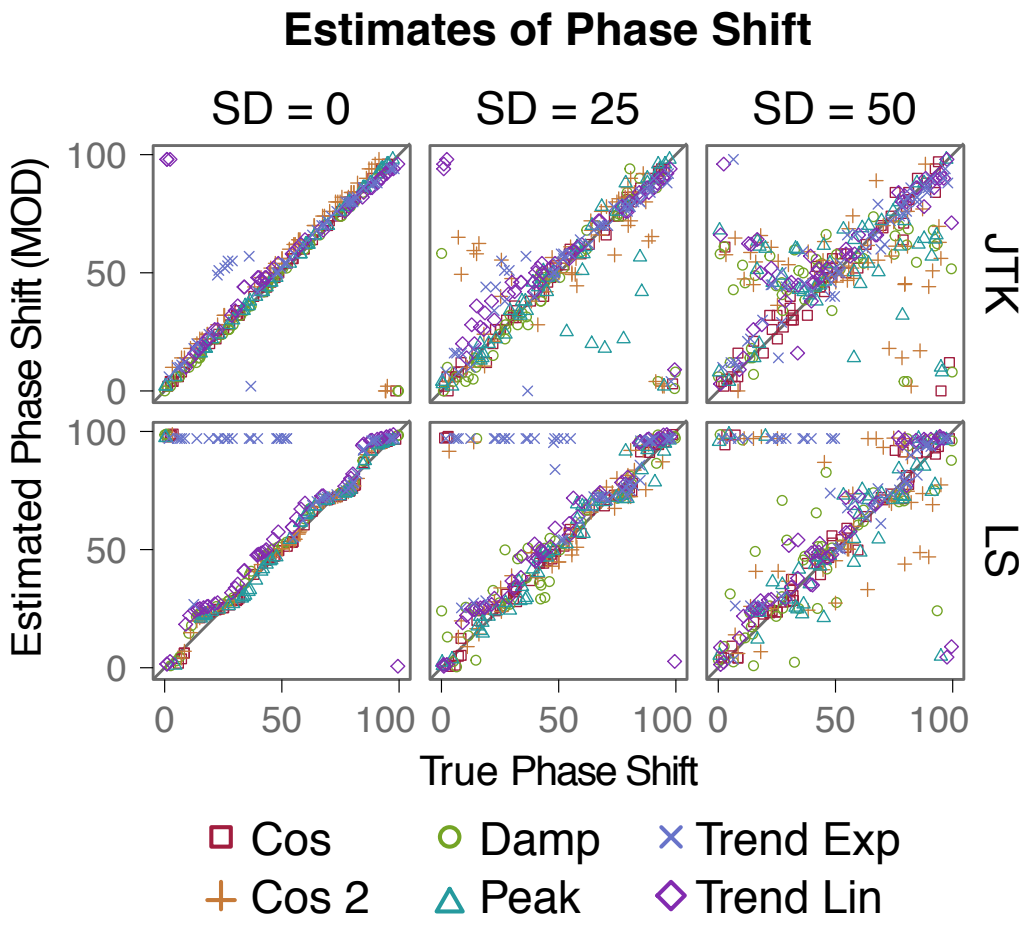


Figure S5: Phase estimates for different profile shapes and noise levels. Estimates for all profiles are shown. The \# samples = 50 for times 0-200 and Gaussian noise with *SD* = {0, 25, 50}. The black line indicates estimate = true. Plots of true phase shift versus estimated phase shift. The period was 100 and the phase shifts were 0-100, which covered every possible phase shift. The phase shift MODULO true period was used. The modulo operator forces numbers above a certain value to wrap around back to zero (e.g. 120 MOD 100 = 20).

Algorithm	Data Set	Parameters
LS	Yeast Cell Cycle	per_min: 64 per_max: 112 test_freq: 4
JTK	Yeast Cell Cycle	per_min: 64 per_max: 112 interval: 16
DL	Yeast Cell Cycle	num_permutations: 10000 period: 97.8
PH	Yeast Cell Cycle	per_min: 64 per_max: 112 degree: 2 combine: 0 geom_factor: 1 amp_factor: 0
LS	Yeast Metabolic Cycle	per_min: 96 per_max: 504 test_freq: 4
JTK	Yeast Metabolic Cycle	per_min: 96 per_max: 504 interval: 24
DL	Yeast Metabolic Cycle	period: 300 num_permutations: 10000
PH	Yeast Metabolic Cycle	per_min: 96 per_max: 504 degree: 2 combine: 0 geom_factor: 1 amp_factor: 0
LS	Plant Root Clock	per_min: 1.28 per_max: 12.16 test_freq: 4
JTK	Plant Root Clock	per_min: 1.28 per_max: 12.16 interval: 0.32
DL	Plant Root Clock	period: 6 num_permutations: 10000
PH	Plant Root Clock	per_min: 1.28 per_max: 12.16 degree: 2 combine: 0 geom_factor: 1 amp_factor: 0
LS	Mammal Circadian	min_per: 20 max_per: 28 test_freq: 4
JTK	Mammal Circadian	per_min: 20 per_max: 28 interval: 1
DL	Mammal Circadian	num_permutations: 10000 period: 24
PH	Mammal Circadian	per_min: 20 per_max: 28 degree: 2 combine: 0 geom_factor: 1 amp_factor: 0

Table S4: Running the Algorithms on Biological Data. For each algorith and data set, the parameters used to run the algorithm are listed.

Lomb-Scargle

Gene	Period	p-value	Norm Plot
ACF4	104.33	0.042	
SKI7	96.41	0.045	
TPO3	98.91	0.046	
YIL001W	89.60	0.047	
YDR239C	71.85	0.050	
RRN10	110.38	0.051	
GCD6	79.33	0.052	
TCB2	105.78	0.052	
MGM1	105.78	0.054	
SMT3	112.00	0.055	
SRO7	101.55	0.056	
YLL032C	105.78	0.056	
COS6	88.56	0.056	
RIB5	85.57	0.056	
VPS60	97.64	0.057	
SCP1	90.67	0.057	
RAD50	87.54	0.057	
YLR297W	79.33	0.058	
TPM2	112.00	0.059	
YSP2	105.78	0.059	

JTK_CYCLE

Gene	Period	p-value	Norm Plot
SRP1	96	1.1E-04	
RRN10	112	1.1E-04	
SKI7	96	1.1E-04	
ECM32	96	1.1E-04	
PLB1	112	3.0E-04	
TCB2	112	3.0E-04	
ILV1	96	3.0E-04	
VBA4	112	3.0E-04	
PPH21	96	3.0E-04	
PSK2	96	3.0E-04	
AGP3	112	3.0E-04	
ALR2	96	3.0E-04	
MIC14	96	3.0E-04	
ARO3	112	7.0E-04	
PNS1	112	7.0E-04	
PRX1	96	7.0E-04	
GCD6	80	7.0E-04	
FUS3	112	7.0E-04	
CNE1	112	7.0E-04	
NPR1	96	7.0E-04	

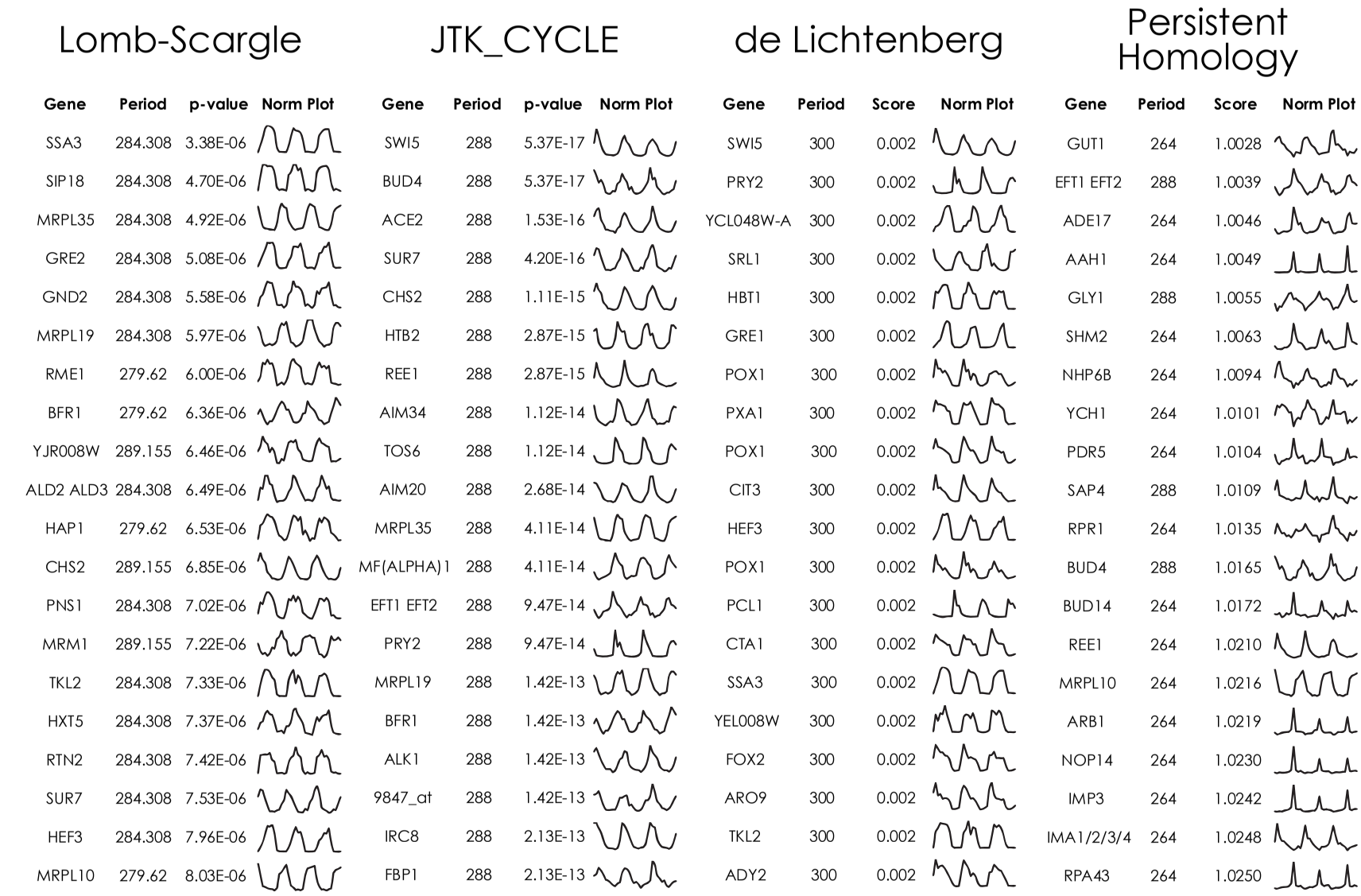
de Lichtenberg

Gene	Period	Score	Norm Plot
HTA2	97.8	0.0001	
NRM1	97.8	0.0005	
CLB1	97.8	0.0011	
SIM1	97.8	0.0023	
SRL1	97.8	0.0049	
YGP1	97.8	0.0104	
ALK1	97.8	0.0197	
CLB2	97.8	0.0293	
SWI5	97.8	0.0408	
BUD4	97.8	0.0415	
AIM20	97.8	0.0452	
GAS3	97.8	0.0826	
YJL118W	97.8	0.1222	
HST3	97.8	0.1242	
WSC2	97.8	0.1344	
SUR7	97.8	0.1458	
YNL058C	97.8	0.1532	
YMR001C-A	97.8	0.1635	
MCM5	97.8	0.2104	
TOS6	97.8	0.2311	

Persistent Homology

Gene	Period	Score	Norm Plot
MET18	80	1	
RAD51	80	1	
NIP100	80	1	
RAD50	80	1	
POP2	80	1	
STE6	80	1	
HIR3	80	1	
CTR9	80	1	
GYL1	80	1	
SSE1	80	1	
GYP6	80	1	
SRS2	80	1	
VCX1	80	1	
SAP190	80	1	
FYV10	80	1	
RAD16	80	1	
YJR098C	80	1	
GPI16	80	1	
NIT3	80	1	
PDR5	80	1	

Figure S6: Top 20 probes for Yeast Cell Cycle data for each algorithm. PH returned 253 probes with the same top score.



Gene	Period	p-value	Norm Plot	Gene	Period	p-value	Norm Plot	Gene	Period	Score	Norm Plot	Gene	Period	Score	Norm Plot
SSA3	284.308	3.38E-06		SWI5	288	5.37E-17		SWI5	300	0.002		GUT1	264	1.0028	
SIP18	284.308	4.70E-06		BUD4	288	5.37E-17		PRY2	300	0.002		EFT1 EFT2	288	1.0039	
MRPL35	284.308	4.92E-06		ACE2	288	1.53E-16		YCL048W-A	300	0.002		ADE17	264	1.0046	
GRE2	284.308	5.08E-06		SUR7	288	4.20E-16		SRL1	300	0.002		AAH1	264	1.0049	
GND2	284.308	5.58E-06		CHS2	288	1.11E-15		HBT1	300	0.002		GLY1	288	1.0055	
MRPL19	284.308	5.97E-06		HTB2	288	2.87E-15		GRE1	300	0.002		SHM2	264	1.0063	
RME1	279.62	6.00E-06		REE1	288	2.87E-15		POX1	300	0.002		NHP6B	264	1.0094	
BFR1	279.62	6.36E-06		AIM34	288	1.12E-14		PXA1	300	0.002		YCH1	264	1.0101	
YJR008W	289.155	6.46E-06		TOS6	288	1.12E-14		POX1	300	0.002		PDR5	264	1.0104	
ALD2 ALD3	284.308	6.49E-06		AIM20	288	2.68E-14		CIT3	300	0.002		SAP4	288	1.0109	
HAP1	279.62	6.53E-06		MRPL35	288	4.11E-14		HEF3	300	0.002		RPR1	264	1.0135	
CHS2	289.155	6.85E-06		MF(ALPHA)1	288	4.11E-14		POX1	300	0.002		BUD4	288	1.0165	
PNS1	284.308	7.02E-06		EFT1 EFT2	288	9.47E-14		PCL1	300	0.002		BUD14	264	1.0172	
MRM1	289.155	7.22E-06		PRY2	288	9.47E-14		CTA1	300	0.002		REE1	264	1.0210	
TKL2	284.308	7.33E-06		MRPL19	288	1.42E-13		SSA3	300	0.002		MRPL10	264	1.0216	
HXT5	284.308	7.37E-06		BFR1	288	1.42E-13		YEL008W	300	0.002		ARB1	264	1.0219	
RTN2	284.308	7.42E-06		ALK1	288	1.42E-13		FOX2	300	0.002		NOP14	264	1.0230	
SUR7	284.308	7.53E-06		9847_at	288	1.42E-13		ARO9	300	0.002		IMP3	264	1.0242	
HEF3	284.308	7.96E-06		IRC8	288	2.13E-13		TKL2	300	0.002		IMA1/2/3/4	264	1.0248	
MRPL10	279.62	8.03E-06		FBP1	288	2.13E-13		ADY2	300	0.002		RPA43	264	1.0250	

Figure S7: Top 20 probes for Yeast Metabolic Cycle data for each algorithm. DL returned 42 probes with the same top score.

Lomb-Scargle				JTK_CYCLE				de Lichtenberg				Persistent Homology			
Gene	**Period**	**p-value**	**Norm Plot**	**Gene**	**Period**	**p-value**	**Norm Plot**	**Gene**	**Period**	**Score**	**Norm Plot**	**Gene**	**Period**	**Score**	**Norm Plot**
AT3G52930	6.293	1.11E-05		AT1G03820	6.08	8.47E-12		RNS2	6	101.4		PBC1	6.08	1.0495	
FAR7	6.293	1.27E-05		CRR28	6.08	1.56E-11		AT5G12470	6	101.4		AT1G28400	10.56	1.0528	
AT1G03820	6.119	1.36E-05		PAP21	6.4	2.84E-11		PKT3	6	101.4		DER1	5.76	1.0532	
AT4G15940	6.293	1.40E-05		263507_s_a	6.4	3.82E-11		UNE5	6	101.4		AT1G63310	5.44	1.0589	
AT5G11310	6.293	1.43E-05		244917_at	6.4	5.12E-11		KCR1	6	101.4		AT3G03960	6.08	1.0594	
255159_at	6.293	1.46E-05		AVP2	6.08	6.83E-11		ACYB-2	6	101.4		TOM40	6.08	1.0624	
CYP702A3	6.293	1.58E-05		AT4G33390	6.08	6.83E-11		TUA6	6	101.4		AT5G11280	5.44	1.0637	
PAP21	6.119	1.73E-05		AT3G13590	6.4	9.10E-11		PBD2	6	101.4		AT5G18970	5.44	1.0658	
AT4G05160	6.119	1.84E-05		257329_at	6.4	1.60E-10		RALFL33	6	101.4		AT4G36660	6.08	1.0693	
AAC1	6.119	1.84E-05		AT5G11310	6.4	1.60E-10		MEE58	6	101.4		AT1G26640	5.76	1.0708	
AT3G04700	6.119	1.89E-05		GLU2	6.08	2.11E-10		AT4G16450	6	101.4		AT2G16460	6.72	1.0753	
AT1G62070	6.119	1.89E-05		ARF21	6.4	2.11E-10		AT4G17100	6	101.4		AT3G10780	6.08	1.0767	
244911_at	6.293	1.91E-05		FAR7	6.4	2.78E-10		AT4G15160	6	101.4		AT5G10070	12.16	1.0790	
AT1G49380	6.119	1.92E-05		AT5G48140	6.4	2.78E-10		AT1G25275	6	101.4		AT2G21390	6.08	1.0825	
AT5G28230	6.119	1.99E-05		AT3G14480	6.08	2.78E-10		AT5G08670	6	101.4		AT1G55160	6.72	1.0827	
PMS1	5.955	1.99E-05		PMS1	6.08	3.65E-10		HTA10	6	101.4		AT5G22270	5.44	1.0856	
AtGolS5	6.293	2.05E-05		AT5G39020	6.08	3.65E-10		ATDAD1	6	101.4		AT2G01220	6.72	1.0870	
AT2G06040	6.119	2.05E-05		AT2G39510	6.4	4.78E-10		RD21	6	101.4		252436_x_a	6.08	1.0886	
AT3G08610	6.119	2.06E-05		CYP702A3	6.4	4.78E-10		EMB1144	6	101.4		AT5G65440	1.6	1.0886	
FLP	6.293	2.07E-05		AT1G18260	6.08	6.24E-10		TUA3	6	101.4		AT2G32580	5.44	1.0887	

Figure S8: Top 20 probes for Plant Root Clock data for each algorithm. DL returned 530 probes with the same top score.

Lomb-Scargle

Gene	Period	p-value	Norm Plot
Elovl3	24.42	1.1E-08	
Tars	24.691	1.4E-08	
Gys2	24.465	1.4E-08	
Clpx	23.939	1.7E-08	
Sco2	24.691	2.3E-08	
Insig2	24.42	2.4E-08	
Arntl	24.51	2.6E-08	
Ubxn1	24.376	2.6E-08	
Slc37a4	23.939	3.1E-08	
Nr1d2	23.436	3.2E-08	
1110067D22Ri	23.939	3.3E-08	
Cml5	23.982	3.3E-08	
Cry1	24.6	3.4E-08	
Hsd3b7	24.645	3.4E-08	
Gda	23.727	3.6E-08	
Smagp	24.42	3.8E-08	
Thrsp	23.559	4.0E-08	
Insig2	23.896	4.0E-08	
Hacl1	25.936	4.1E-08	
Gda	23.232	4.2E-08	

JTK_CYCLE

Gene	Period	p-value	Norm Plot
Elovl3	25	1.0E-24	
Bnip3	24	1.6E-24	
St3gal5	24	3.9E-24	
St3gal5	24	9.0E-24	
Nampt	24	1.4E-23	
Gys2	25	4.6E-23	
Ubxn1	24	1.5E-22	
Smagp	25	1.5E-22	
Pppde1	24	1.5E-22	
Rorc	24	4.8E-22	
Clpx	24	6.9E-22	
Nampt	24	1.8E-20	
Nrg4	24	6.8E-20	
Cry1	25	9.5E-20	
Cml5	24	1.3E-19	
Adck3	25	2.5E-19	
Sco2	25	3.5E-19	
Phf17	24	6.6E-19	
Insig2	24	9.0E-19	
Tars	25	9.0E-19	

de Lichtenberg

Gene	Period	Score	Norm Plot
Tubb2a	24	1.4E-05	
Avpr1a	24	1.4E-05	
Elovl3	24	1.4E-05	
Arntl	24	1.4E-05	
Nr1d2	24	1.4E-05	
Slc30a10	24	1.4E-05	
Slc45a3	24	1.4E-05	
Per2	24	1.4E-05	
Nr1d2	24	1.4E-05	
Chka	24	1.4E-05	
Mthfr	24	1.4E-05	
Hist2h3c1	24	1.4E-05	
Chka	24	1.4E-05	
Lrfn3	24	1.4E-05	
Cldn1	24	1.4E-05	
Chka	24	1.4E-05	
Dbp	24	1.4E-05	
Trib3	24	1.4E-05	
Dbp	24	1.4E-05	
Per3	24	1.4E-05	

Persistent Homology

Gene	Period	Score	Norm Plot
St3gal5	23	1.004	
Nampt	23	1.005	
St3gal5	25	1.018	
Adam1b	27	1.020	
Bnip3	23	1.027	
Dlgap4	25	1.029	
Ubxn1	25	1.035	
Sco2	23	1.039	
Rgs16	23	1.043	
Nle1	28	1.049	
9130015A21Ri	28	1.051	
Gmeb1	28	1.058	
Kcnk10	27	1.059	
Runx3	28	1.059	
Tspan4	25	1.060	
Elovl3	23	1.061	
Zfand3	28	1.061	
Nampt	25	1.062	
Dnaja4	28	1.063	
Lpin1	23	1.065	

Figure S9: Top 20 probes for Mammal Circadian data for each algorithm. DL returned 52 probes with the same top score.

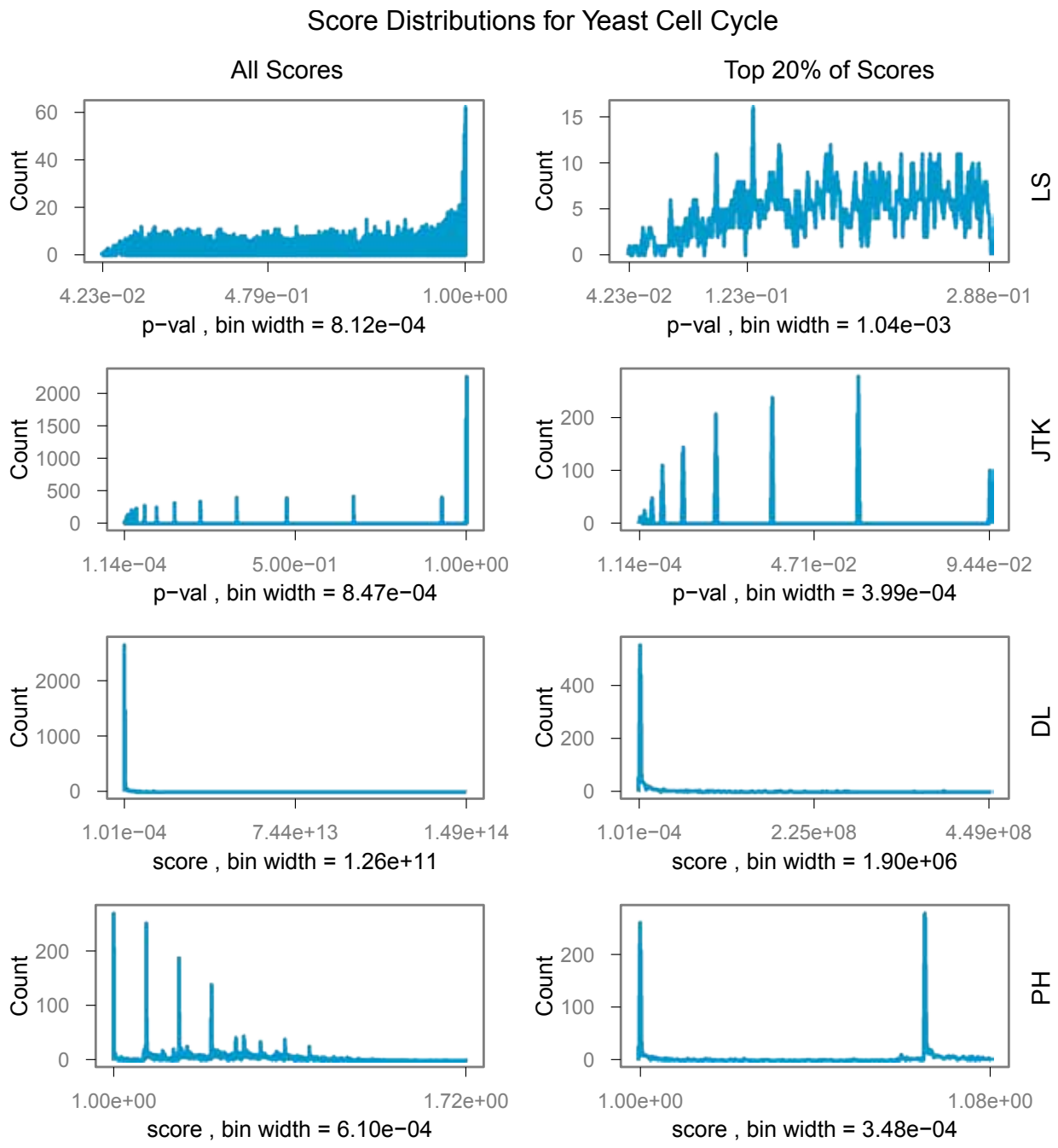


Figure S10: Score distributions for the Yeast Cell Cycle Data. Plots for each algorithm are in rows. Plots for all the scores are on the left, and plots only of the top 20% of scores are on the right. The frequency polygon function in ggplot2 was used to produce the plots in R, and the bin widths are shown in the x-axis title.

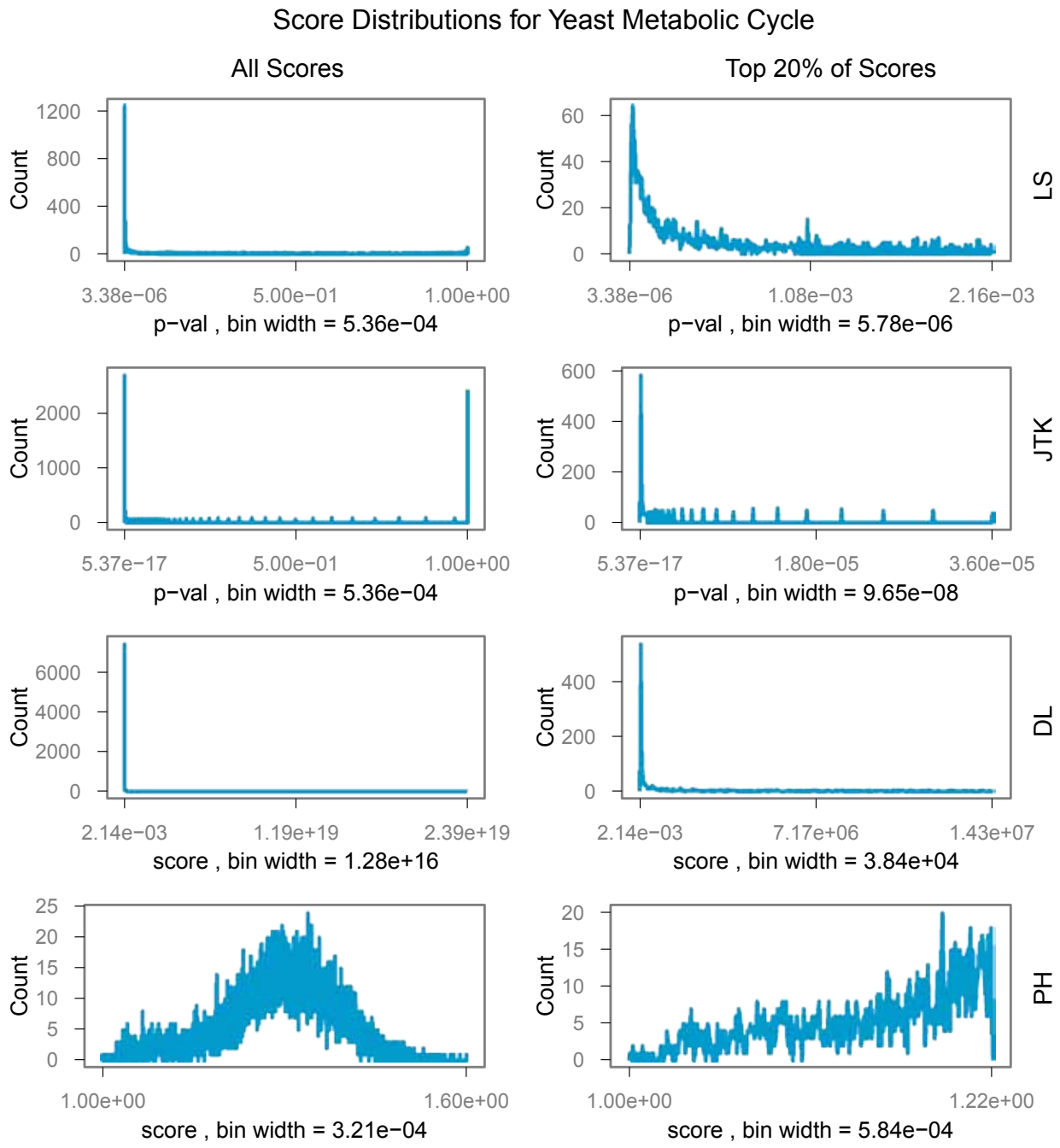


Figure S11: Score distributions for the Yeast Metabolic Cycle Data. Plots for each algorithm are in rows. Plots for all the scores are on the left, and plots only of the top 20% of scores are on the right. The frequency polygon function in ggplot2 was used to produce the plots in R, and the bin widths are shown in the x-axis title.

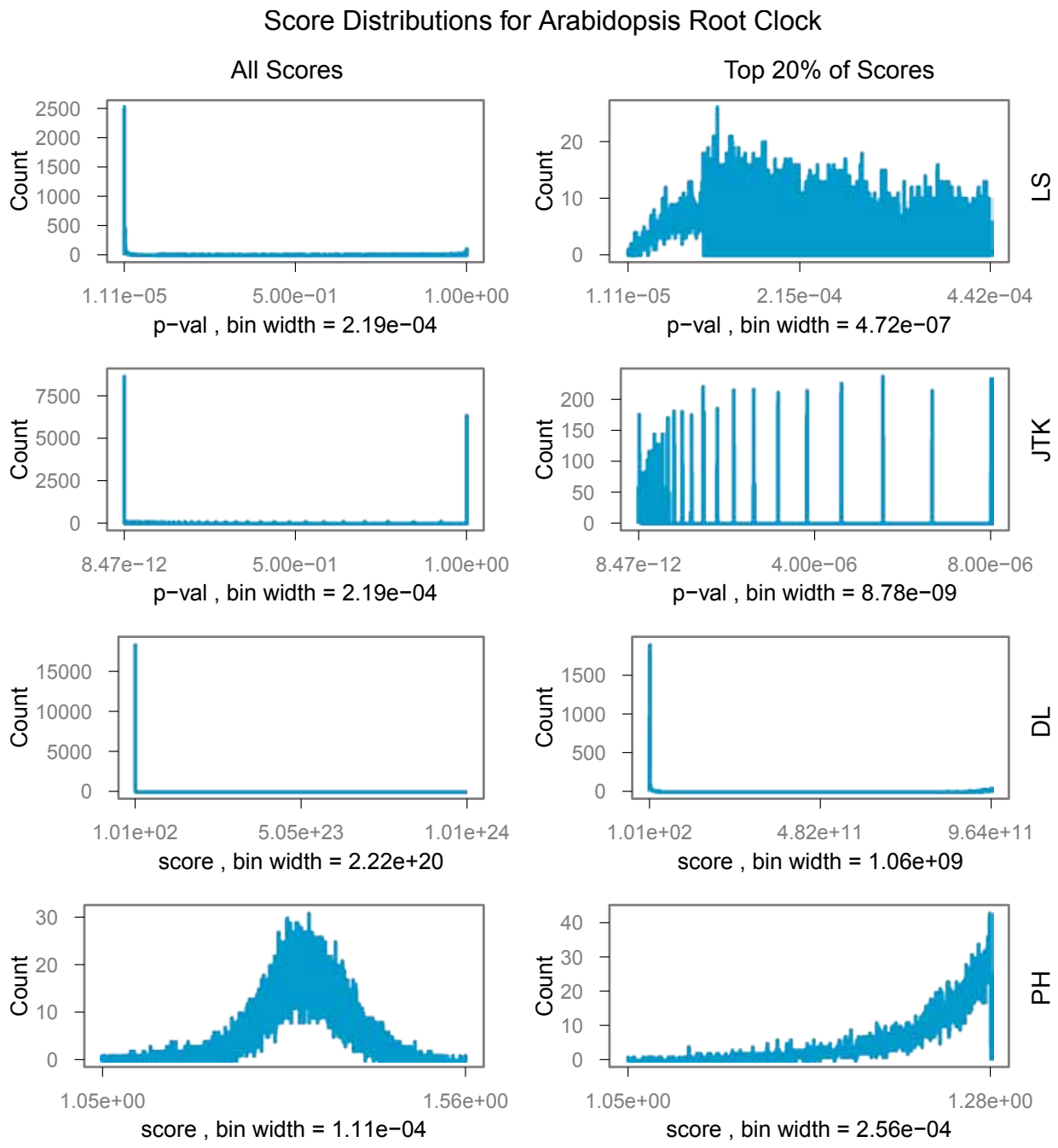


Figure S12: Score distributions for the Plant Root Clock Data. Plots for each algorithm are in rows. Plots for all the scores are on the left, and plots only of the top 20% of scores are on the right. The frequency polygon function in ggplot2 was used to produce the plots in R, and the bin widths are shown in the x-axis title.

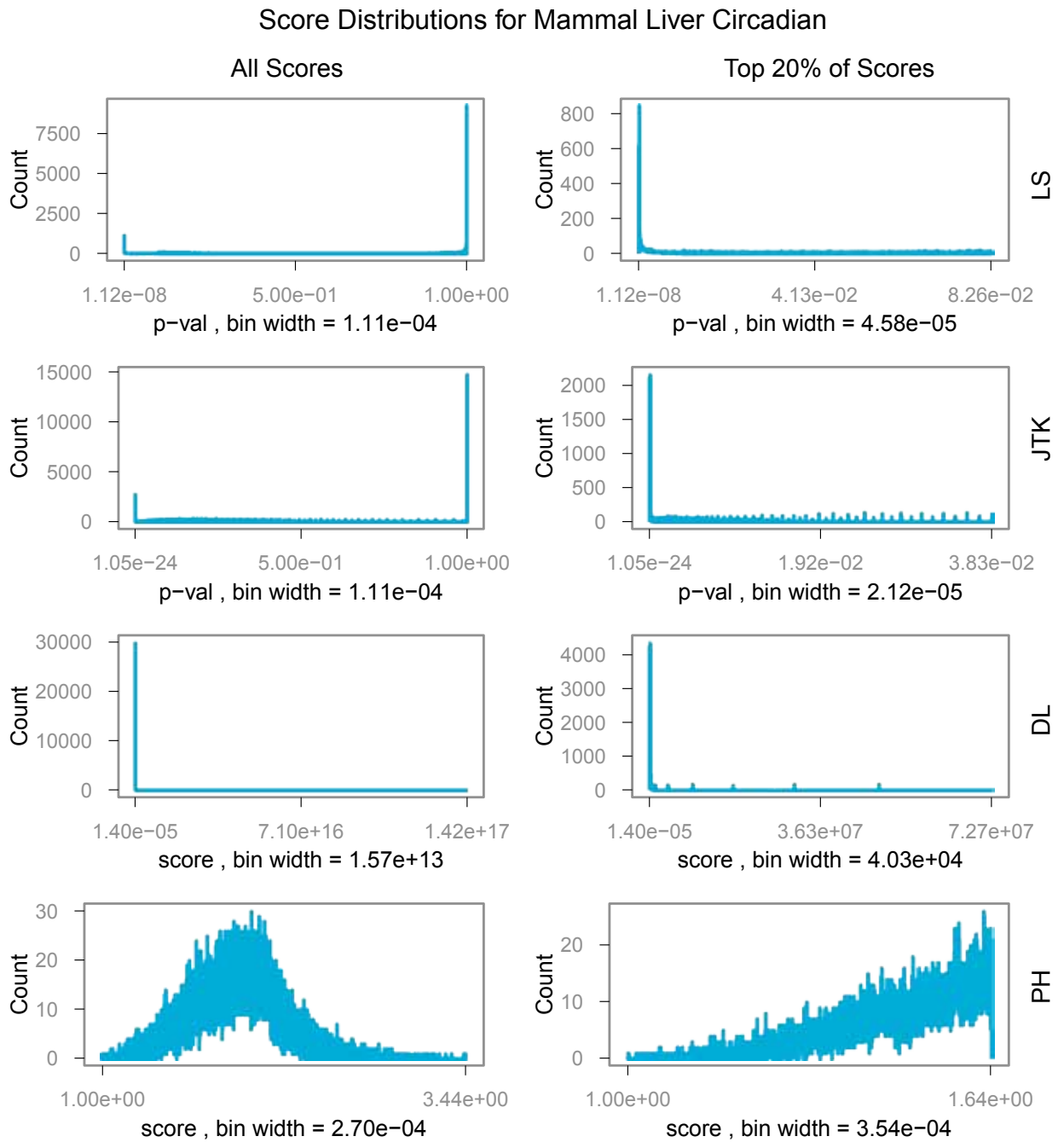


Figure S13: Score distributions for the Mammal Circadian Data. Plots for each algorithm are in rows. Plots for all the scores are on the left, and plots only of the top 20% of scores are on the right. The frequency polygon function in ggplot2 was used to produce the plots in R, and the bin widths are shown in the x-axis title.

1 Features of Algorithms

Here we cover some of the features of the algorithms, many of which are specified by the current implementation and not the algorithm. LS, JTK, and PH return estimates of the period from a range of periods searched, but DL only looks at one period. It is possible to run DL for each period of interest, but DL is slower than the other algorithms. LS and JTK also estimate the amplitudes and phase shifts, while DL and PH do not. The implementation of PH could be modified to detect phase shift by finding the persistent global maximum. Some algorithms can handle time series with missing time points and/or unevenly spaced time points. The implementation of PH will not process data sets with missing or uneven time points, but LS can as it was designed for this situation. JTK can handle missing time points, but its implementation currently only allows for specifying an even spacing between time points and then indicating which time points are missing. The algorithms' performance on handling missing or uneven time points was not evaluated. The features are summarized in Table S5.

Features	JTK	LS	DL	PH
Estimates significance (p-values)	y	y	n	n
Estimates period	y	y	n	y
Estimates amplitude	y	y	n	n
Estimates phase shift	y	y	n	n
Handles missing time points	y	y	y	n
Handles uneven time points	n*	y	y	n

Table S5: A summary of the features provided by each algorithm. Yes (y) and No (n). *JTK can handle missing time points, but does not directly handle unevenly spaced time points. The algorithms' performance on handling missing or unevenly spaced time points was not evaluated.

2 Run Time of Algorithms

When working with larger data sets, such as the circadian genome-wide RNA-Seq data (>200,000 features), the speed of the algorithms becomes important. To test their speed, we created synthetic data sets containing 100, 1k, 10k, or 100k profiles and having 10, 20, 40, or 80 samples. The profiles had two periods, with peak-to-trough amplitudes of 100, and Gaussian noise with standard deviation = 25. The algorithms were modified to suppress graphical output, but still write all resuls to files. Each data set was run twice through each algorithm and their times were averaged. The average run times (Table S6, Figure S14) were used to explore the growth of the run time.

A run time with a set input (for a given number of genes and number of samples) reflects the performance of the selected language and implementation in addition to the efficiency of an algorithm. PH was written in C++ while LS and JTK were written in R; an algorithm implemented in C++ is expected to run faster than the same algorithm implemented in R. As we deal with increasingly larger data sets, another concern is how

the run time scales as the size of the input increases. This is dependent on the number of steps an algorithm must perform for each input, or how efficient the algorithm is.

To explore how the algorithm scales as the size of the input increases, we show the running time as the number of samples or number of genes increases (Figure S14). For increasing numbers of genes, the execution times of the algorithms were approximately linear. For increasing numbers of samples, LS had sub-linear increase in execution time for up to 80 samples; e.g. twice as many samples took less than twice as long to run. However, JTK and PH both exhibited above linear growth for 20 to 40 and 40 to 80 samples.

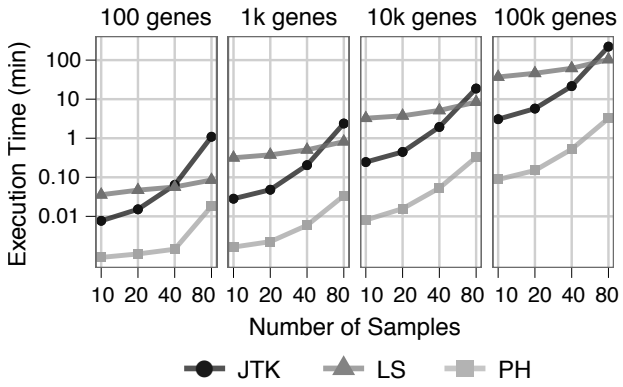


Figure S14: Relationship between algorithm run time and sample density, for several numbers of gene expression profiles. Run times are shown for JTK, LS, and PH.

JTK Execution Times (minutes)

		# Genes			
		100	1000	10000	100000
# Samples	10	0.0077	0.0283	0.2451	3.0718
	20	0.0151	0.0480	0.4459	5.7297
	40	0.0637	0.2045	1.9401	21.5390
	80	1.0897	2.3810	18.7047	221.3169

LS Execution Times (minutes)

		# Genes			
		100	1000	10000	100000
# Samples	10	0.0357	0.3159	3.2490	36.9739
	20	0.0469	0.3752	3.7451	45.9281
	40	0.0564	0.5041	5.1495	61.6099
	80	0.0851	0.8154	8.4007	103.5927

PH Execution Times (minutes)

		# Genes			
		100	1000	10000	100000
# Samples	10	0.0009	0.0016	0.0081	0.0878
	20	0.0011	0.0023	0.0157	0.1496
	40	0.0014	0.0059	0.0531	0.5241
	80	0.0182	0.0338	0.3258	3.2279

Table S6: Run times on data sets with different numbers of samples and probes. Times in minutes for the algorithms to run. Rows are number of samples, and columns are number of genes that were run. Each time is an average from two runs on the same computer.

3 Algorithms

Lomb-Scargle (Lomb, 1976; Scargle, 1982): A set of sinusoidal signals that cover a range of periods are compared to the time series to generate a measure of correspondence. The significance of each of these is calculated, and the period of the most significant fit is returned. The explanation of this method in Scargle (1982) and Glynn *et al.* (2006) is recommended. The R-implementation was from (Glynn *et al.*, 2006). This implementation uses the Lomb-Scargle normalized periodogram as defined in Press and Rybicki (1989).

JTK_CYCLE (Hughes *et al.*, 2010): A set of profiles (user-defined, the default is sinusoidal) is generated to cover a range of periods and phase shifts. A pair-wise comparison of all points in a profile calculates whether they are increasing or decreasing in relation to one another. The increasing/decreasing pattern of the time series is then compared to the increasing/decreasing pattern of each reference profile to determine the statistical significance of the correlation. It uses the Jonckheere-Terpstra test and Kendall's tau to compute the significance. The period and phase shift for the reference profile with the most significant correlation (or an average if there are more than one) is Bonferroni-adjusted for multiple testing and returned. The implementation was in R from the author of the paper.

de Lichtenberg (de Lichtenberg *et al.*, 2005): To measure the significance of periodicity, a background distribution is generated by creating a set of random profiles by permuting a given profile's expression values. The p-value is the proportion of permuted profiles with Fourier score at least as large as the original profile's observed Fourier score. For the significance of regulation, the gene expression profile is compared to a set of random profiles generated by selecting a value from a randomly selected gene profile at each time point. The p-value for regulation (amplitude) is measured as the proportion of permuted profiles with standard deviation at least as large as a time series' observed standard deviation. The implementation in R from (Orlando *et al.*, 2008) was used (see Acknowledgements).

Persistent Homology (Cohen-Steiner *et al.*, 2010): PH normalizes the data from 0 to 1, and then pairs (in a subtle way) minima and maxima of a time series, treated as a function on the circle. A measure is obtained by summing the differences (persistence) between the maximum and the minimum of each pair. If there is only one minimum and maximum pair, the measure is one and is considered to be a perfect oscillation; thus the method is insensitive both to amplitudes and sinusoidal shape. Additional oscillations in the time series will create more minimum-maximum pairs, which will increase the score, indicating a less perfect profile. To determine period, sliding windows with widths equal to the range of periods are used; the period with the lowest score is returned. The last author of (Cohen-Steiner *et al.*, 2010) provided an implementation of the algorithm written in C++ (see Acknowledgements).

4 Data Sets

Yeast Cell Cycle Data (Orlando *et al.*, 2008): Wild-type strains of *S. cerevisiae* (derivatives of BF264-15Dau) were synchronized by elutriation. Samples were taken at 16 minute intervals starting at 30 minutes and ending at 254 minutes. There were two replicates in this experiment, for our analysis we used only the first replicate. The period for the cell cycle in this experiment is estimated to be 77.1 minutes for mother cells and 118.5 minutes for daughter cells (length of normal cell cycle of 77.1 plus daughter specific phase

of 41.4). The samples cover a recovery period and roughly two cell cycles. However, there is a stress shock response during the recovery period; we therefore ignored the first 2 time points and looked only at the last 13 time points. This microarray data was from the Affymetrix Yeast Genome 2.0 Array and was processed using dChip. This data set contains 15 time points for 5,900 probes. The data was provided by the authors (GEO accession GSE8799).

Yeast Metabolic Data (Tu *et al.*, 2005): Diploids of *S. cerevisiae* strain CEN.PK were grown to a high density, briefly starved and then given low concentrations of glucose. Samples were taken approximately every 23-25 minutes (sampling was not even at all time points) starting at 3973 minutes and ending at 4837 minutes. We evened the sample times in the data by making the sampling at every 24 minutes. Any blanks in the data were filled with zeros. The period of the yeast metabolic cycle is estimated to be ~300 minutes, and this data set covers approximately three cycles. This microarray data was from the Affymetrix Yeast Genome S98 Array. This data set contains 36 time points and 9,335 probes. The data was downloaded from GEO (GEO accession GSE3431).

Plant Root Clock Data (Moreno-Risueno *et al.*, 2010): The roots could not be synchronized, so instead the chronological order of different roots was inferred by analyzing the reporter DR5:GUS expression by RT-PCR. We applied evenly spaced time points to approximate the inferred timing. The period of the root clock is estimated to be ~6 hours and this data set covers roughly two cycles. This microarray data was from the Affymetrix Arabidopsis ATH1 Genome Array. This data set contains 39 time points and 22,801 probes. . The data was provided by the authors (GEO accession GSE21611).

Mammalian Circadian Rhythm Data Hughes *et al.* (2009): Wild-type C57BL/6J mice were synchronized by entraining them to an environment with 12 h light and 12 h dark for one week. They were then placed into total darkness. Samples were taken from the liver every hour starting at 18 hours after the first subjective day and ending at 65 hours. The period of the circadian rhythm is ~24 hours and this data set covers two circadian cycles. This microarray data was from Affymetrix Mouse Genome 430 2.0 array and was processed using GCRMA. This data set contains 48 time points and 45,101 probes. The data was provided by the authors (GEO accession GSE11923).

5 Running the Algorithms on the Biological Data Sets

The period of the cell cycle in the wild-type yeast was estimated to be 77.1 minutes for mother cells and 118.5 minutes for daughter cells (Orlando *et al.*, 2008). As a simplification, we assumed the period length would be the average of the mother and daughter = 97.8 minutes. For LS, JTK, and PH a period range of 64-112 minutes was used. For DL, the period was 97.8 minutes.

The yeast metabolic cycle data was evaluated by LS, JTK, and PH with a period range of 96 to 504 minutes and by DL with a period of 300 minutes.

The plant root clock data set was evaluated by LS, JTK, and PH with a period range of 1.28 to 12.16 hours and by DL with a period of 6 hours.

The mammalian circadian data was evaluated by LS, JTK, and PH with a period range of 20 to 28 hours. For DL, the period was set to 24 hours.

6 Data Analysis & Plotting

The R package ROCR was used to compute ROC and AUC (Sing *et al.*, 2005). The results from synthetic data were plotted in R using the ggplot2 package (Wickham, 2009).

References

Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Mileyko, Y. (2010). Lipschitz Functions Have L p -Stable Persistence. *Foundations of Computational Mathematics*, **10**(2), 127–139.

de Lichtenberg, U., Jensen, L. J., Fausbøll, A., Jensen, T. S., Bork, P., and Brunak, S. (2005). Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics*, **21**(7), 1164–1171.

Glynn, E. F., Chen, J., and Mushegian, A. (2006). Detecting periodic patterns in unevenly spaced gene expression time series using Lomb–Scargle periodograms. *Bioinformatics*, **22**(3), 310–316.

Hughes, M., Hogenesch, J. B., and Kornacker, K. (2010). JTK_CYCLE: An Efficient Nonparametric Algorithm for Detecting Rhythmic Components in Genome-Scale Data Sets. *Journal of Biological Rhythms*, **25**(372), 372–380.

Hughes, M. E., DiTacchio, L., Hayes, K. R., Vollmers, C., Pulivarthy, S., Baggs, J. E., Panda, S., and Hogenesch, J. B. (2009). Harmonics of circadian gene transcription in mammals. *PLoS genetics*, **5**(4), e1000442.

Lomb, N. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, **39**, 447–462.

Moreno-Risueno, M. A., Van Norman, J. M., Moreno, A., Zhang, J., Ahnert, S. E., and Benfey, P. N. (2010). Oscillating gene expression determines competence for periodic Arabidopsis root branching. *Science*, **329**(5997), 1306–1311.

Orlando, D., Lin, C., Bernard, A., Wang, J., Socolar, J., Iversen, E., Hartemink, A., and Haase, S. (2008). Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature*, **453**(7197), 944–947.

Press, W. H. and Rybicki, G. B. (1989). Fast algorithm for spectral analysis of unevenly sampled data. *Astrophysical Journal*, **338**, 277–280.

Scargle, J. (1982). Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data. *Astrophysical Journal*, **263**, 835–853.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). Rocr: visualizing classifier performance in r. *Bioinformatics*, **21**(20), 7881.

Tu, B., Kudlicki, A., Rowicka, M., and McKnight, S. (2005). Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, **310**(5751), 1152–1158.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.