

# Bootstrapping and Empirical Bayes Methods Improve Rhythm Detection in Sparsely Sampled Data

Alan L. Hutchison,<sup>\*,†,‡,1</sup>  Ravi Allada,<sup>§</sup> and Aaron R. Dinner<sup>‡,||,¶,1</sup>

<sup>\*</sup>Medical Scientist Training Program, <sup>†</sup>Graduate Program in the Biophysical Sciences, <sup>‡</sup>Institute for Biophysical Dynamics, University of Chicago, Chicago, Illinois, <sup>§</sup>Department of Neurobiology, Northwestern University, Evanston, Illinois, <sup>||</sup>Department of Chemistry, and <sup>¶</sup>James Franck Institute, University of Chicago, Chicago, Illinois

**Abstract** There is much interest in using genome-wide expression time series to identify circadian genes. However, the cost and effort of such measurements often limit data collection. Consequently, it is difficult to assess the experimental uncertainty in the measurements and, in turn, to detect periodic patterns with statistical confidence. We show that parametric bootstrapping and empirical Bayes methods for variance shrinkage can improve rhythm detection in genome-wide expression time series. We demonstrate these approaches by building on the empirical JTK\_CYCLE method (eJTK) to formulate a method that we term *BooteJTK*. Our procedure rapidly and accurately detects cycling time series by combining information about measurement uncertainty with information about the rank order of the time series values. We exploit a publicly available genome-wide data set with high time resolution to show that *BooteJTK* provides more consistent rhythm detection than existing methods at typical sampling frequencies. Then, we apply *BooteJTK* to genome-wide expression time series from multiple tissues and show that it reveals biologically sensible tissue relationships that eJTK misses. *BooteJTK* is implemented in Python and is freely available on GitHub at <https://github.com/alanlhutchison/BooteJTK>.

**Keywords** bioinformatics, rhythm detection, gene expression analysis, circadian, empirical Bayes

Periodic patterns (rhythms) are pervasive in biology at the molecular, cellular, organismal, and ecological scales. It can be challenging to detect these patterns with confidence of their significance, however, because biological dynamics are intrinsically noisy, and often it is feasible to obtain only a few samples of a potentially periodic process. To address this issue, several statistical methods have recently been developed to identify cycling time series despite

sparse sampling (Hutchison et al., 2015; Thaben and Westermark, 2014; Deckard et al., 2013; Yang and Su, 2010; Hughes et al., 2010; Keegan et al., 2007).

Empirical JTK\_CYCLE (eJTK; Hughes et al., 2010; Hutchison et al., 2015) and RAIN (Thaben and Westermark, 2014) are nonparametric methods that analyze the rank order of measurements. While this approach makes them sensitive to waveforms of arbitrary shape, it does not incorporate information about

1. To whom all correspondence should be addressed: Alan L. Hutchison, Medical Scientist Training Program, Graduate Program in the Biophysical Sciences, Institute for Biophysical Dynamics; e-mail: [alanlhutchison@uchicago.edu](mailto:alanlhutchison@uchicago.edu). Aaron R. Dinner, Institute for Biophysical Dynamics, Department of Chemistry, James Franck Institute, University of Chicago, Chicago, Illinois, USA; e-mail: [dinner@uchicago.edu](mailto:dinner@uchicago.edu).

JOURNAL OF BIOLOGICAL RHYTHMS, Vol. 33 No. 4, August 2018 339–349

DOI: 10.1177/0748730418789536

© 2018 The Author(s)

Article reuse guidelines: [sagepub.com/journals-permissions](http://sagepub.com/journals-permissions)

the measurement uncertainty. Analysis of variance (ANOVA) is a parametric approach that does incorporate the variance of intra-time point measurements when identifying differences in mean values, but it is less sensitive than eJTK because it does not use information about the time order of the measurements (Hutchison et al., 2015). ARSER (Yang and Su, 2010) is likely the most successful parametric method at present, but the fact that it fits a time series to a sinusoidal curve by autoregressive spectral estimation makes it less sensitive to nonsinusoidal time series, as we discuss below.

Either explicit or implicit in these methods is comparison of the variance in measurements at each time point (i.e., across replicates/periods) to the variance from one time point to another over the period of interest. The cost and effort of sample preparation and measurement limit the number of replicates/periods obtained. As a result, the observed variance at each time point may poorly represent the true variance. In particular, if the data yield an estimate of the variance that is too small, a time series is more likely to be falsely identified as cycling because the apparent signal is large compared with the apparent noise. Properly accounting for small replicate numbers in estimating the variance has the potential to provide substantial gains in accuracy of rhythm detection and, in turn, to aid in understanding periodic biological processes.

To this end, we introduce an empirical Bayes (eBayes) procedure (Smyth, 2004; Lu and Stephens, 2016). In this approach, which is commonly employed in differential expression analysis (Trapnell et al., 2012; Ritchie et al., 2015), information from across a data set is combined to estimate a prior distribution for the standard deviation, and this prior is then used together with the individual measurements to estimate the variance at each time point. This “shrinks” the spread in variances (Suppl. Fig. S1; Lu and Stephens, 2016). Here, we use the empirical Bayes variance estimates to generate parametric bootstrap time series samples and then apply a rhythm detection algorithm to them. The parametric bootstrap (Efron and Tibshirani, 1994) is also established in bioinformatics and is applied in packages for RNA-Seq quantification (Bray et al., 2016) and differential expression (Pimentel et al., 2017). To the best of our knowledge, this is its first application in rhythm detection.

While the strategy that we present is general, we focus on its implementation with the empirical JTK\_CYCLE (eJTK) method, which we have demonstrated to outperform other algorithms (Hutchison et al., 2015). eJTK compares time series to a set of reference waveforms varying in phase (peak expression) and distance from peak to trough using a nonparametric pairwise rank order correlation, Kendall's  $\tau$ . Selecting

the best waveform presents a multiple-hypothesis testing problem, which eJTK solves by explicitly calculating the null distribution of the selection procedure to assign (one-sided)  $p$  values to resulting rhythmicity scores. This approach is accurate but relatively computationally costly because the null distribution must be reevaluated for each set of measurements (in order to distinguish time series that are missing observations for different sets of time points). In the present work, we reduce this expense significantly by fitting a Gamma distribution to test statistics for a small number of time series. This approximation makes eJTK, even in the context of the bootstrap, computationally economical.

Our approach, which we term *BooteJTK*, combines freedom from restrictive assumptions regarding the shape of the waveform with incorporation of information about the uncertainty in each measurement. We demonstrate the method on simulated data and 2 circadian genome-wide expression data sets. The first data set is densely sampled with measurements of gene expression in mouse liver samples at 1-h intervals for 48 h (Hughes et al., 2009). This data set allows us to examine the performance (self-consistency) of the method as fewer time points are included. The second data set comprises gene expression measurements every 2 h for 48 h for 12 mouse tissues in continuous darkness (Zhang et al., 2014). This data set allows us to look at the consistency of rhythm detection across tissues. We find that fewer genes are rhythmic than previously believed, because of the more stringent requirement that the uncertainty in measurements be small relative to the amplitude of expression, in addition to the rank order of the values of the time series matching those of a reference waveform. Corroborating our more stringent results with core clock transcription factor targets (CCTs; Koike et al., 2012), we find no decrease in CCT enrichment between BooteJTK and eJTK. At the same time, we find increases in the conditional probabilities of rhythmic genes across tissues, suggesting circadian programs in different tissues may not be as distinct as previously thought. Put together, the results indicate that BooteJTK provides robust rhythm detection with improved consistency. The general principles and methods that we present here, the empirical Bayes and bootstrapping procedures, can be applied to other rhythm detection algorithms.

## METHODS

To test our method against other methods, we generated 1100 time series with a 24-h period sampled every 2 h in duplicate with Gaussian noise added to

each point. The time series were generated to include varying peak-to-trough time separations,  $w$ , with values ranging from 2 h to 22 h in steps of 2 h. Defining  $w' = 2\pi w / (24\text{h})$  and  $t' = 2\pi t / (24\text{h})$ ,

$$y = \begin{cases} \cos(\pi t' / w') & \text{for } t' \leq w' \\ \cos(\pi \frac{t' + 2(\pi - w')}{2\pi - w'}) & \text{for } t' > w' \end{cases} \quad (1)$$

Equation (1) simply defines a waveform that falls along a cosine curve from  $t = 0$  to  $t = w$ , then rises along a different cosine curve up until  $t = 24$  h. This waveform can be described as having peaks at  $t = 0$  and  $t = 24$  h and a trough at  $t = w$ . One hundred time series were generated for each  $w$  value. We varied the Gaussian standard deviation relative to the amplitude of the underlying cosine, a ratio we refer to as the noise level. To model the null hypothesis, we generated 1000 time series with the same number of time points but with values drawn only from the Gaussian distribution.

### BooteJTK Algorithm

Given a set of time series, our approach, BooteJTK, consists of the following steps:

1. Average replicate values and estimate variances via eBayes
2. Generate bootstrap time series
3. Run eJTK on the bootstrap time series
4. Compute an average test statistic
5. Estimate the  $p$  value

We discuss these steps in detail below.

### Empirical Bayes Variance Estimation

Empirical Bayes methods are an established part of many workflows for differential expression analysis (Smyth, 2004; Trapnell et al., 2012; Ritchie et al., 2015). These methods combine information from all the time points in the data to shrink the spread of standard deviation estimates. As a result, low standard deviation estimates are increased and high standard deviation estimates are decreased (Suppl. Fig. S1).

Empirical Bayes methods are useful in the present context because the number of time point replicates tends to be low (e.g., 2), so traditional statistical estimators for the standard deviation tend to be unreliable, and pooling data from multiple time points can moderate the errors from these estimators. In this article, replicates, or time point replicates, refer to measurements at mod 24-h time points (e.g., time 0 h and 24 h are replicates, times 2 h and 26 h are replicates,

etc.). The phrase *bootstrap replicates* refers specifically to parametric bootstrap resamplings of the original time series.

We use *voom* (specifically, *vooma* for microarrays; Ritchie et al., 2015) to obtain initial standard deviation estimates that account for mean-variance relationships. These estimates are then adjusted using the empirical Bayes procedure implemented in *vash* (Lu and Stephens, 2016). We note that we originally used *limma* (Ritchie et al., 2015) instead of *vash*, but that resulted in overdispersion and overestimates of rhythmicity  $p$  values, partially because of adjustment of small standard deviations away from zero.

### Calculating a Test Statistic by Bootstrapping eJTK

In nonparametric bootstrapping, data are resampled with replacement to create a distribution of simulated measurements that can in turn be used to compute statistics. In parametric bootstrapping, data are resampled from a distribution modeling the original data (Efron and Tibshirani, 1994). Parametric bootstrapping more readily incorporates variance estimates from empirical Bayes and outperformed nonparametric bootstrapping in our tests of the methods (Suppl. Fig. S2), likely owing to the limited number of time points. Thus, we use parametric bootstrapping. Specifically, we log-transform the expression measurements, a standard method for positively skewed data (Tukey, 1977), and model the resulting data at each time point as normally distributed with the mean directly calculated from the time point replicates and the variance modeled by the empirical Bayes procedure as described above.

Using this model of the original time series data, we generate  $n$  time series for this model and analyzed them with eJTK to determine their circadian characteristics: rhythmicity score ( $\tau$ ), phase (peak), and best-matching waveform. We averaged each of these  $n$  statistics across the model time series. While eJTK generally outputs integer multiples of the measurement interval for the peak and trough times (i.e., extrema), the means of these statistics can be noninteger, which allows for better representation of the times of the extrema when they do not coincide with the measurement times. Regardless, for the phase and trough, the mean values are close to the values output by eJTK. This is not necessarily the case for the rhythmicity score, as we now discuss.

In the context of eJTK, the Kendall's  $\tau$  statistic measures the correlation in rank order of the values of the time series of interest and the values of a discretized reference waveform; the rhythmicity score is the highest  $\tau$  across all tested reference series. A perfect match in rank order has  $\tau = 1$ . Adding noise to the values of a reference time series and comparing the

resulting rank order with the original one often results in  $\tau < 1$ , with  $\tau$  tending to decrease as the noise grows in comparison with the amplitude of the oscillation. As a result, the mean of the distribution of  $\tau$  values for the bootstrap resamples depends on both the rank order of time series values and the measurement uncertainty.

An additional issue is that the  $\tau$  distribution is skewed when  $\tau$  is close to the limits of its range ( $-1$  and  $1$ ). To stabilize the variance across the full range of possible rhythmicity scores, we average the Fisher transform of  $\tau$ :  $\tilde{\tau} = \text{arctanh}(\tau)$ , truncating the values to  $\pm \text{arctanh}(0.99)$  for  $|\tau| > 0.99$  to ensure that the  $\tilde{\tau}$  values are finite.

We also examined the accuracy of the phase estimation of BooteJTK on the simulated data (Suppl. Fig. S3). For the data described above with a variety of peak-to-trough distances, the standard deviation of the phase error for BooteJTK is 2.8 h, while for ARSER it is 1.6 h. The phase standard deviation calculated by BooteJTK provides a sense of the phase error. We found that the calculated phase standard deviation and the phase error of the method were negatively correlated with the strength of rhythmicity detected. While ARSER provides a lower standard deviation of the phase error, BooteJTK provides information on the phase standard deviation calculated from the bootstrap replicates, which provides information on the trustworthiness of the mean phase estimate.

### Obtaining Accurate and Computationally Inexpensive $p$ Values

A  $p$  value is the likelihood under the null hypothesis of observing a value of a test statistic or a more extreme one. We previously generated the null distribution for eJTK by applying the method to  $10^6$  time series generated by selecting values from a Gaussian distribution with a constant mean (Hutchison et al., 2015). We applied the same approach for calculating the null distribution for our bootstrap method, generating  $n$  bootstrap replicates for each of the  $10^6$  time series, averaging the resulting  $\tau$  values to generate  $10^6$   $\tilde{\tau}$  values. We generate this null distribution to match the measurement times in the experimental time series. Because this numerical procedure to generate the null distribution represents most of the computational expense of eJTK and, in turn, BooteJTK, we sought an approximate analytical form for the null distribution. We found the Gamma distribution, which we previously used to model the null distribution of the rhythm detection method F24 (Wijnen et al., 2005; Hutchison et al., 2015), to be a reasonable choice (Suppl. Fig. S4). To assess this approach quantitatively, we computed  $p$  values with this model and

our earlier method (i.e., empirically from the histogram of  $10^6$   $\tilde{\tau}$  values) for a range of  $\tilde{\tau}$  values. The ratio of the analytical  $p$  values to the empirical  $p$  values is larger than 1 at very low  $p$  values and close to 1 for moderate  $p$  values close to typical significance thresholds (Suppl. Figs. S4B and D). This means that the  $p$  values obtained from the Gamma-distribution approximation are sufficiently accurate for use, but they favor the null hypothesis relative to the empirical values, leading to slightly fewer time series being considered rhythmic for a given significance threshold.

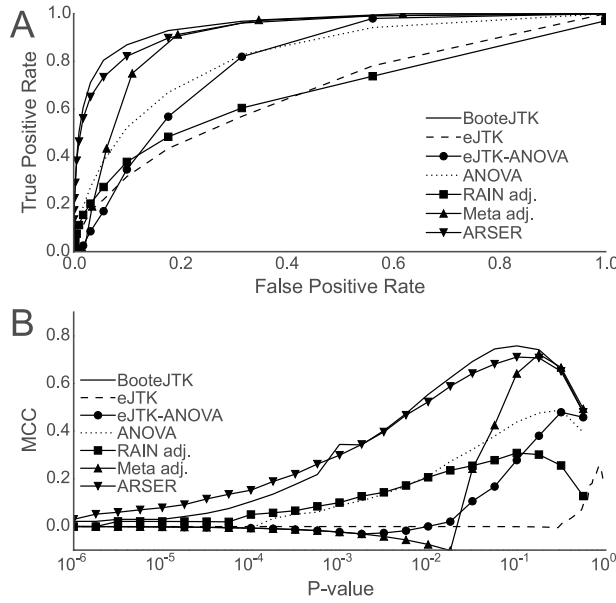
Having established the abilities of BooteJTK, we sought to minimize the computational cost of our method. For the simulated data described above, we found no discernible difference in  $\tilde{\tau}$  scores for  $n = 100, 50, 25$ , and 10 bootstrap samples, where  $n$  is the number of bootstrap samples used to calculate  $\tilde{\tau}$  (Suppl. Fig. S5). We use 25 bootstrap samples throughout the rest of this study. With 25 bootstrap samples, we obtain run times on a late-2013 iMac with a 3.5-GHz Intel Core i7-4771 processor and 16 GB of 1600 MHz DDR3 memory. For 1000 time series, the analysis by BooteJTK took 180 s and the analysis by eJTK took 8 s. Of this, less than 2 s is the integration of the Gamma distribution to translate  $\tilde{\tau}$  statistics to  $p$  values. With the improvements in the present article, the computational cost of eJTK is several orders of magnitude less than in Hutchison et al. (2015).

### BooteJTK Outperforms Alternative Rhythm Detection Methods

*BooteJTK Outperforms eJTK.* First, we compared BooteJTK to eJTK to understand the effect of the bootstrapping. In each case, we compared the 1100 simulated time series against 24-h reference waveforms with peak-to-trough separations varying as above (Eq. (1)) and phases varying every 2 h from 0 to 22 h (132 total reference waveforms). Figure 1 shows that BooteJTK significantly outperforms eJTK: the true-positive rate is higher for a given false-positive rate (FPR; Fig. 1A; Suppl. Fig. S6A), and the Matthews correlation coefficient (MCC) is higher for all  $p$  value cutoffs (Fig. 1B; Suppl. Fig. S6B). The MCC (Eq. (2)) combines the number of true-positives ( $T_p$ ), false-positives ( $F_p$ ), true-negatives ( $T_N$ ), and false-negatives ( $F_N$ ) to define a classifier that is 1 if it is perfect and 0 if it performs no better than random guessing.

$$MCC = \frac{T_p T_N - F_p F_N}{\sqrt{(T_p + F_p)(T_p + F_N)(T_N + F_p)(T_N + F_N)}} \quad (2)$$

To illustrate the differences between BooteJTK and eJTK applied to time series, we show 2 time series with the same  $\tau$  value for eJTK ( $\tau = 0.57$ ,  $p = 2 \times 10^{-3}$ ) but different values for BooteJTK ( $\tilde{\tau} = 0.66$  and  $0.97$ ,

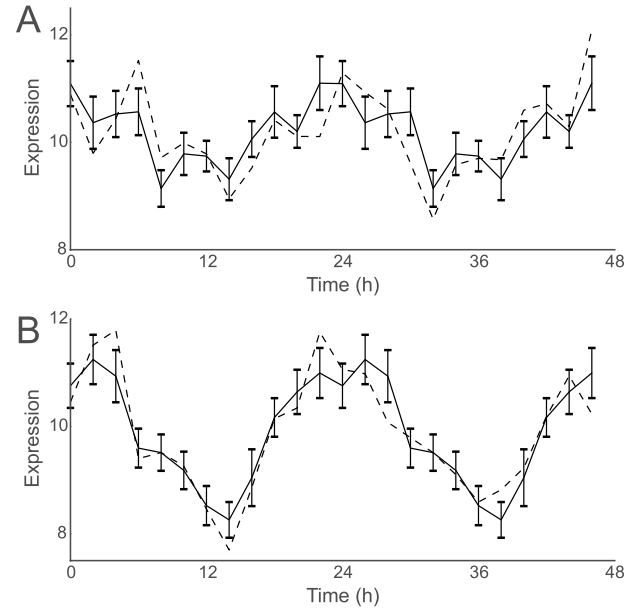


**Figure 1.** BooteJTK outperforms competing rhythm detection methods. (A) Receiver-operating characteristic curve and (B) Matthews correlation coefficient. BooteJTK is compared to RAIN (Thaben and Westermark, 2014; Hutchison and Dinner, 2017) with empirical  $p$  values, MetaCycle (Wu et al., 2016; Hutchison and Dinner, 2017) with Brown-corrected  $p$  values, empirical JTK (eJTK; Hutchison et al., 2015), ARSER (Yang and Su, 2010), analysis of variance (ANOVA), and Fisher-Brown integration (Fisher, 1925; Brown, 1975) of eJTK and ANOVA (eJTK-ANOVA). Simulated data are generated with noise-to-amplitude ratio of 1.00.

$p = 10^{-3}$  and  $p = 10^{-6}$ ; Figs. 2A, B). We plot the time point replicates sequentially. In both panels, the average time series is visually similar to the contributing time point replicates. However, the error bars (standard deviations) are much smaller in Figure 2B than in Figure 2A, resulting in the difference in  $p$  values in BooteJTK (but not eJTK).

**BooteJTK Outperforms RAIN.** We also compared BooteJTK to RAIN (Thaben and Westermark, 2014), another nonparametric method that uses reference waveforms. RAIN does not take into account the size of the noise relative to the amplitude of the time series, which we expect to be more important when testing experimental data. We previously found that RAIN underestimates  $p$  values (Hutchison and Dinner, 2017), so we computed empirical  $p$  values for RAIN using  $10^6$  simulations of the null distribution as described in Hutchison and Dinner (2017). We ran RAIN with the settings as follows: period = 24, deltat = 2, peak.border = c(0.1, 0.9), adjp.method = "ABH." We found that BooteJTK outperforms RAIN (Fig. 1).

**BooteJTK Outperforms ARSER on Asymmetric Time Series.** As mentioned in the introduction section, parametric methods do account for the size of the



**Figure 2.** Examples of 2 time series with the same eJTK  $\tau$  but different BooteJTK  $\tau$  values. Dashed lines show the original data, plotted sequentially. Black lines show averages over the 2 periods of data, mod 24 (e.g., time 0 h and 24 h averaged, 2 h and 26 h averaged, etc.), double plotted for comparison with the original data. Error bars indicate the standard deviation of the averaged time point replicates. The BooteJTK  $\tau$  values for (A) and (B) are 0.66 and 0.97, which correspond to  $p$  values of  $10^{-3}$  and  $10^{-6}$ . The eJTK  $\tau$  score is  $\tau = 0.57$ , which corresponds to a  $p$  value of 0.002. The noise-to-amplitude ratio used to generate both time series was 0.60.

noise relative to the amplitude of the time series, and ARSER (Yang and Su, 2010) is likely the best such method presently. Because ARSER fits a time series to a sinusoidal curve, it should outperform nonparametric methods when detecting time series that are approximated well by that waveform. However, we expect many biological time series to be nonsinusoidal (Hutchison et al., 2015; Thaben and Westermark, 2014) and show further evidence supporting this below. For this reason, we compared the MCC scores of BooteJTK and ARSER with given  $p$  value thresholds with our simulated time series in aggregate (Fig. 1) and separated by time from peak to trough (Suppl. Fig. S7). ARSER was run on default settings. While ARSER has higher MCC than BooteJTK for near-symmetric time series, BooteJTK outperforms ARSER for asymmetric time series with peak-to-trough times less than 6 h or greater than 18 h (Suppl. Fig. S7).

**BooteJTK Outperforms MetaCycle and Combined eJTK-ANOVA.** MetaCycle combines ARSER, the original JTK\_CYCLE, and Lomb-Scargle by Fisher integration (Fisher, 1925; Wu et al., 2016). Because the contributing methods are applied to the same experimental time series, their  $p$  values are dependent on each other, and it is necessary to employ the Brown

correction (Brown, 1975) to obtain accurate  $p$  values (Hutchison and Dinner, 2017). We find that BooteJTK outperforms MetaCycle with this correction (Fig. 1). Relatedly, we observed that BooteJTK combines eJTK's nonparametric test with an ANOVA-like treatment of measurement uncertainties. We thus wondered if Fisher-Brown integration of ANOVA and eJTK  $p$  values would perform similarly to BooteJTK. The fact that it did not (Fig. 1) shows that the bootstrapping procedure is a better way of combining these features.

## RESULTS

### Effect of Sampling Frequency on Rhythm Detection

While BooteJTK outperforms leading methods for simulated data, such time series can lack features of experimental data. Assessing the behavior of algorithms for experimental data can be challenging, however, because rhythmic expression has been independently verified for only a small fraction of the genome. Nevertheless, we expect rhythm detection to be accurate when a waveform is extensively sampled (with high frequency, over many periods), and we can study the consistency of each method as data are downsampled.

To this end, we applied BooteJTK to microarray data collected every 1 h for 48 h from mouse liver tissue under constant conditions (Hughes et al., 2009). As the original analysis of the data set was performed with JTK\_CYCLE (Hughes et al., 2010), we analyzed the data set with eJTK as well for comparison. In addition, we analyzed the data with adjusted RAIN and ARSER, as those methods performed well on our simulated data.

We treated the modulo 24 time points as replicates, providing 2 replicates every 1 h over 24 h. Since most transcriptomic circadian experiments have data collected every 2 h (Zhang et al., 2014) or 4 h (Flourakis et al., 2015; Perelis et al., 2015), we parsed the data (collected from CT18-CT65) into 2 data sets with measurements every 2 h (denoted 2a: CT18, CT20, etc. and 2b: CT19, CT21, etc.) and into 4 data sets with measurements every 4 h (denoted 4a, 4b, 4c, and 4d, starting at CT18, CT19, CT20, and CT21, respectively). We used the R package *gcrma* (Wu et al., 2018) to normalize the data (GEO GSE11923) and removed probes with constant expression. In all cases, we controlled the false-discovery rate with the Benjamini-Hochberg (BH) correction and took time series to be rhythmic if their adjusted  $p$  values were less than 0.05 (Fig. 3).

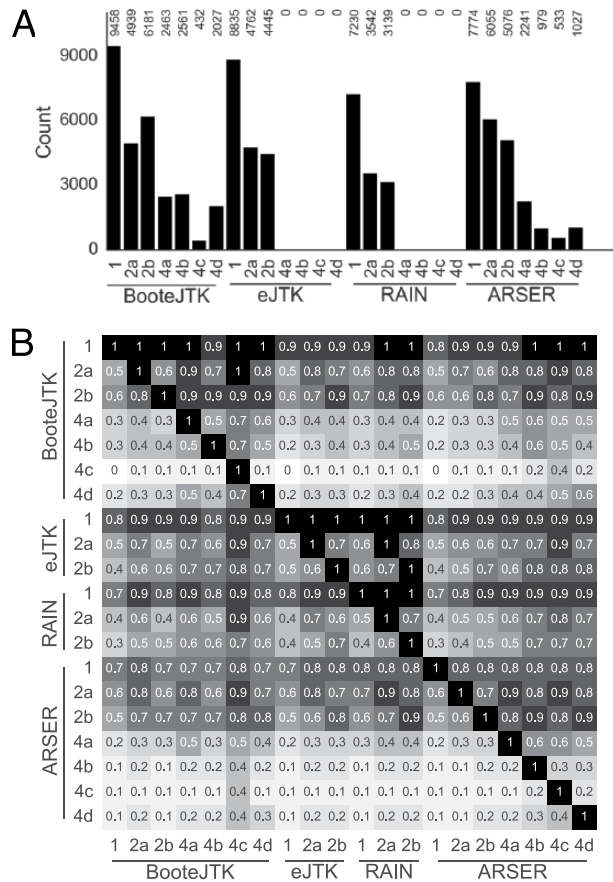
For the full data set, BooteJTK identifies more probes as rhythmic than eJTK, RAIN, and ARSER do. Downsampling the data to measurements every 2 h reduces the number of rhythmic probes in all cases (Fig. 3A). Downsampling the data to measurements every 4 h prevents eJTK and adjusted RAIN from finding any rhythmic probes at this significance threshold, while BooteJTK identifies thousands of the probes that it originally considered rhythmic. ARSER likewise detects more than eJTK or RAIN, but as we show in Figure 3B and discuss at the end of this section, many of the rhythmic probes ( $\approx 80\%$ ) in the downsampled data sets are not rhythmic in the full data set.

The overlap between results obtained from 2 data sets can be quantified by the conditional probability that a probe is rhythmic in one data set (a row in Fig. 3B) if it is rhythmic in another (a column in Fig. 3B). For example, 78% of probes rhythmic under BooteJTK when the time series are downsampled to even time points are also found to be rhythmic when downsampling to odd time points; 62% are rhythmic in the opposite case ([row 3, column 2] vs. [row 2, column 3], respectively; below, we use square brackets to denote positions in Fig. 3B and consistently write the row first and the column second). Because no biological differences should exist between equivalently downsampled data sets, the conditional probabilities provide a scale for evaluating differences across biologically distinct data sets. For eJTK, these values are 63% and 68% (Fig. 3B, [10,9], and [9,10], respectively); for RAIN, they are 61% and 68% (Fig. 3B, [13,12] and [12,13], respectively); and for ARSER, they are 62% and 73% (Fig. 3B, [16,15] and [15,16], respectively). These results indicate greater consistency for the BooteJTK results.

We also analyzed the overlap of the inverse case: the conditional probability that if a probe is found to be arrhythmic in one data set it is arrhythmic in another (Suppl. Fig. S8). We found that across methods, as the sampling rate decreased, the number of probes identified as arrhythmic increased. Examining consistency within a method with data sets downsampled to 2 h as above, we found that RAIN had the most consistency in assigning probes as arrhythmic (92% and 89%), followed by BooteJTK (89% and 78%), eJTK (87% and 85%), and ARSER (86% and 78%). Given that RAIN finds fewer probes as rhythmic than the other methods, it would be expected that it would more consistently find probes to be arrhythmic while downsampling. These results reinforce that BooteJTK provides strong consistency between results.

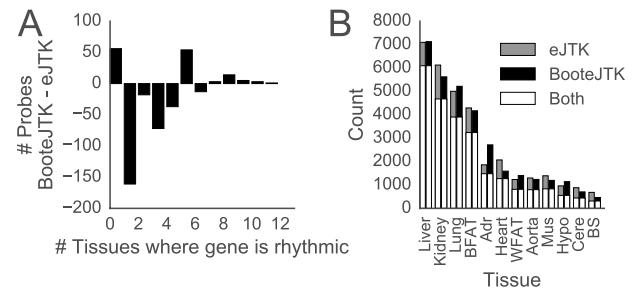
We suggested above that ARSER, although highly sensitive to sinusoids, was not as sensitive as BooteJTK to nonsinusoidal waveforms. We explored the impact of this in the Hughes data in Supplementary





**Figure 3. BooteJTK identifies rhythmic probes more consistently than eJTK, RAIN, and ARSER as data are downsampled.** The full data set (Hughes et al., 2009) is sampled every 1 h for 48 h (denoted 1); downsampled data sets are generated by keeping measurements at intervals of 2 h (denoted 2a and 2b, starting at CT18 and CT19, respectively) and 4 h (denoted 4a, 4b, 4c, and 4d, starting at CT18, CT19, CT20, and CT21, respectively). (A) Number of rhythmic probes at Benjamini-Hochberg  $<0.05$  for the indicated methods and data sets. (B) We quantified the conditional probabilities between results with different data sets by the probability that a probe is rhythmic in the row data set if it is rhythmic in the column data set. Algorithm-data set combinations without any rhythmic time series are not shown.

Figure S9. We show that many of the probes identified as rhythmic by BooteJTK have nonsinusoidal waveforms, which helps explain why BooteJTK detects more probes as rhythmic than ARSER for the 1-h data. Another potential explanation is that isolated randomly noisy time point measurements can lead to harsher penalties in parametric methods such as ARSER as opposed to nonparametric methods such as BooteJTK, eJTK, or RAIN. This is supported by detection of rhythmic time series in the downsampled data that ARSER does not detect in the full data set. Whereas BooteJTK, eJTK, and RAIN on the full data set identify nearly 100% of the rhythmic probes identified after downsampling (Fig. 3B [1,2-7], [8,9-10], and [11,12-13], respectively), ARSER detects



**Figure 4. BooteJTK reveals fewer probes with circadian rhythmic expression across 12 mouse tissues than eJTK for the Zhang et al. (2014) data set.** (A) Fewer probes are rhythmic in multiple tissues under BooteJTK than under eJTK (with Benjamini-Hochberg  $<0.05$  for both methods). (B) Fewer probes per tissue are identified as rhythmic with BooteJTK than with eJTK.

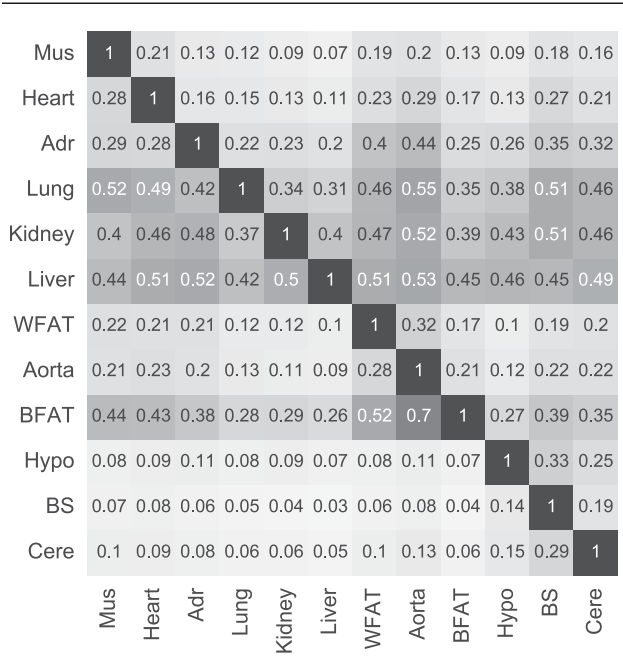
only about 80% of the rhythmic probes after down-sampling, as it does with the full data set (Fig. 3B, [14, 15-20]).

### BooteJTK Reveals More Biologically Consistent Circadian Rhythmic Gene Expression across 12 Mouse Tissues

Having established the improved rhythm detection of BooteJTK, we analyzed the 12-tissue mouse microarray time series in Zhang et al. (2014). The 12 tissues are adrenal, aorta, brown fat, brainstem, cerebellum, heart, hypothalamus, kidney, liver, lung, muscle, and white fat. Expression was sampled every 2 h for 48 h, which we again treated as duplicate measurements over 24 h. As we wish to compare the analysis of the data with BooteJTK to the original manner in which it was analyzed, we also applied eJTK, mirroring the application of the original JTK\_CYCLE method (Hughes et al., 2010) by Zhang et al. (2014).

#### BooteJTK Is More Stringent than eJTK but Exhibits Comparable Enrichment for Core Clock Targets (CCTs).

When applied to the 12-tissue microarray data set from Zhang et al. (2014), BooteJTK identifies 14,598 out of 25,268 probes (11,731/20,038 genes) as rhythmic in at least one tissue with BH  $<0.05$ , whereas eJTK finds 14,763 such probes (12,426 genes; Fig. 4A, B). This difference was to be expected for the reasons discussed above: BooteJTK compares the noise to the amplitude of a time series in addition to evaluating the rank order of the values. This added condition to the definition of rhythmicity should result in fewer time series being identified as rhythmic. However, we were concerned that the greater stringency might exclude actual rhythmic time series. To evaluate this possibility, we corroborated our rhythmic genes with CCTs identified by ChIP-Seq in mouse liver (Koike et al., 2012). Across the 12 tissues and between the 2



**Figure 5. Conditional probabilities of rhythmic probe identification between tissues.** The values shown are the conditional probabilities that a probe is rhythmic in a row tissue if it is rhythmic in a column tissue for BooteJTK.

methods, we found no meaningful difference in the fraction of CCT genes relative to the number of genes identified as rhythmic (mean difference  $-0.016$ , standard deviation  $0.022$ ). This result persisted as we increased the number of core-clock transcription factors necessary to consider a gene a CCT: the mean difference and standard deviation only became smaller as the requirements were made more stringent.

**Relationships between Tissues.** We examined the relationships between different tissues through the probability that a probe is rhythmic in one tissue conditioned on it being rhythmic in another. In Figure 5, row tissues are conditioned on column tissues as in Figure 3B, and the tissues are ordered by hierarchically clustering on the columns. Anatomically related tissues appear together in the plot—for example, the hypothalamus, brainstem, and cerebellum are on the right. It is important to note that the relationships between tissues are asymmetric: being rhythmic in the hypothalamus leads to a probability of 46% that a probe is rhythmic in the liver, but being rhythmic in the liver leads to a probability of 7% that a probe is rhythmic in the hypothalamus. To put these numbers in context, we can compare them to the data of Hughes et al. (2009): the conditional probabilities for 2 data sets from identical conditions downsampled to measurements every 2 h were 62% and 78%. The highest off-diagonal value in Figure 5A is for the probability that a probe is rhythmic in brown fat

conditioned on being rhythmic in the aorta (70%). This result is interesting, but we feel further study is warranted as this point is an outlier in Supplementary Figure S11A, and the presence of brown fat around the aorta can easily lead to contamination of aorta samples (Fitzgibbons et al., 2011).

In Supplementary Figure S10, we show the differences between the conditional probabilities from BooteJTK and eJTK. The columns corresponding to the brain tissues show marked differences. Zhang et al. (2014) discuss the technical difficulty of dissecting the brain regions separately, so using a robust method to analyze these data should be of particular importance. The consistent increase in rhythmic predictive value of other tissues for the hypothalamus and adrenals is due to the increase in probes identified as rhythmic by BooteJTK relative to eJTK, whereas the increase in rhythmic predictive value of the brain stem and cerebellum for other tissues is due to a decrease in probes identified as rhythmic by BooteJTK relative to eJTK (Suppl. Figs. S11B and S10).

**Genes That Are Rhythmic in Most Tissues.** Thirteen genes are identified as rhythmic by BooteJTK in all 12 tissues: *Arntl* (Bmal), *Nr1d1* (Rev-erbA), *Nr1d2* (Rev-erbB), *Dbp*, *Per1*, *Per2*, *Per3*, *Ciart* (Chrono), *Bhlhe41* (Dec2), *Tns2*, *Tsc22d3*, *Usp2*, and *Tspan4*. The first 8 listed are involved in the core clock machinery (Ukai and Ueda, 2010; Goriki et al., 2014). However, known core clock genes such as *Npas2*, *Tef*, and *Hlf* are identified as rhythmic in only 11 of the 12 tissues; they do not meet the significance threshold in the hypothalamus. Given our prior knowledge regarding these genes (Ukai and Ueda, 2010) and the evidence of their rhythmicity in other tissues, it is possible that they are in fact rhythmic in all 12 tissues, and experimental issues are responsible for the inability to detect them in all 12 tissues. As noted above, Zhang et al. (2014) suggest that the technical difficulty of dissecting the brain regions separately may negatively affect circadian rhythm identification in these tissues. We thus examined the 119 genes that BooteJTK identified as rhythmic in 9 or more of the tissues. Most of these genes were not rhythmic in the brainstem, hypothalamus, or cerebellum (Suppl. Fig. S12).

Examining the functional annotation of these genes revealed many ontologies to be expected of consistently rhythmic genes, such as rhythmic processes and transcription regulation (Table 1). However, additional functional annotations were identified, such as genes involved in the stress response, endoplasmic reticulum, pigment granules, and heat shock. A few other genes stand out as well. *Wee1* is rhythmic in 10 tissues (absent from hypothalamus and brainstem). *Wee1* regulates cellular division by inhibiting entry into mitosis (Kellogg, 2003)



**Table 1.** Select functional annotations from the DAVID webtool (Huang et al., 2009a) for the 119 genes identified as rhythmic in 9 or more tissues by applying BooteJTK with a BH adjusted  $p$  value threshold of 0.05 to the Zhang et al. (2014) data set.

Functional Annotation	Fold Enrichment <sup>a</sup>	BH	Genes
GO:0048511: rhythmic process	22.57	2.11e-11	ARNTL, CLOCK, CRY1, DBP, HLF, KDR, MMP14, NFIL3, NPAS2, NR1D1, PER1, PER2, PER3, TEF
GO:0048770: pigment granule	16.34	2.06e-04	HSP90AA1, HSPA5, HSPA8, MMP14, PDIA3, PDIA4, SLC1A5
SP-PIR stress response	24.54	3.56e-04	CIRBP, HSP90AA1, HSPA5, HSPA8, HSPB1, HSPH1
SP-PIR endoplasmic reticulum	4.07	8.68e-04	CALR, DGAT2, EPHX1, FMO1, FMO2, HERPUD1, HSPA5, LPIN1, P4HA1, PDIA3, PDIA4, POR, SCD2, SDF2L1, SERP1
SP-PIR transcription regulation	2.62	2.52e-03	ARNTL, BHLHE40, BHLHE41, CLOCK, CRY1, DBP, ELK3, HLF, KLF9, KLF15, LEO1, LITAF, LPIN1, NFIL3, NPAS2, NR1D1, NR1D2, PER1, PER2, PER3, TEF, THRA
IPR013126: heat shock protein 70	48.93	4.61e-02	HSPA5, HSPA8, HSPH1

BH = Benjamini-Hochberg adjusted  $p$  value; SP-PIR = Swiss-Prot Protein Information Resource keywords; GO = Gene Ontology keywords; IPR = InterPro.

a. Fold enrichment is relative to the expectation of observing the indicated functional annotation in a set of randomly selected genes.

and is known to be regulated by the core clock (Matsuo et al., 2003); more generally, it has been suggested that the cell cycle is under circadian control (Sandler et al., 2015). Two other genes involved in the cell cycle, *Cdkn1a* and *Calr*, are rhythmic in 9 or more tissues. Given that many of these tissues have little cell proliferation, these genes may be functioning in other processes, in which case we expect those processes to be influenced by the circadian clock as well. *Fmo1* and *Gst2* are identified as rhythmic in 10 tissues, while *Fmo2* is identified as rhythmic in 11 tissues. These genes are identified as being involved in drug metabolism by the DAVID webtool (Huang et al. 2009a, 2009b). Given the increasing interest in chronotherapeutics (Zhang et al., 2014), further research into these genes is warranted to better understand their involvement in circadian processes.

## DISCUSSION

We have shown that rhythm detection from genome-wide expression time series can be considerably improved by using an empirical Bayes approach to improve variance estimates from limited replicates and propagating the resulting estimates into the test statistic for eJTK (Hutchison et al., 2015) by a parametric bootstrap. Because eJTK itself is nonparametric, BooteJTK maintains sensitivity for arbitrarily shaped and scaled waveforms but accounts for experimental uncertainty when comparing measurements. We demonstrated that the method provides improved accuracy in identifying simulated rhythmic time series and improved consistency across related experimental data sets. More generally, we expect the framework that we have built around JTK\_CYCLE (Hughes et al., 2010; Hutchison et al., 2015)—empirical estimation of  $p$  values to account properly

for multiple-hypothesis testing, analytical approximation of the null distribution, variance shrinkage and stabilization, and bootstrapping—can be applied to other rhythm detection algorithms to obtain inexpensive and accurate  $p$  values, as we have here.

Our method uses time point replicates to estimate the variance in expression, which is then propagated to the rhythmicity estimate. For time series data without replicates, a different approach is needed. We suggest using the standard deviation of arrhythmic time series as a proxy for the standard deviation of the time points of rhythmic time series. In support of this idea, we show in Supplementary Figure S13 that the mean of the standard deviation of the arrhythmic ( $p > 0.8$ ) time series overestimates the standard deviation for the data from Hughes et al. (2009) by a factor of only about 1.5.

Our analysis of the mouse liver microarray data from Hughes et al. (2009) emphasizes the importance of sampling time series frequently and understanding the expected consistency of results. While BooteJTK and ARSER, in contrast to eJTK or RAIN, were still able to detect a small fraction of rhythmic genes, even downsampling the data to measurements every 2 h resulted in some differences in genes identified as rhythmic from odd-hour measurements and even-hour measurements. That said, sampling every 2 h is much better than the commonly used 4-h sampling rate, especially when comparisons are made across tissues or conditions.

Future studies can also use the values in Figure 3B and their analogs for other rhythm detection methods to better understand the overlap and consistency that should be expected when comparing data sets. Here, for example, comparing the tissue conditional probability results to the 2-h downsampled benchmark suggested that the aorta tissue samples of Zhang et al. (2014) were potentially contaminated by brown fat.

Multiple studies have discussed the tissue specificity of circadian rhythms (Panda et al., 2002; Storch

et al., 2002; Zhang et al., 2014). Our analysis with BooteJTK identifies a larger number of genes that are generally rhythmic than previously identified (positive bars at tissue numbers 8 to 12 in Fig. 4B). We expect the number of genes that are generally rhythmic to be even larger, however, given that current rhythm detection methods do not incorporate prior information about a given gene. We discuss above how including prior information provides us with a stronger belief that genes such as *Npas2*, *Hlf*, and *Tef* may be rhythmic in all 12 tissues. Conversely, one might doubt that a gene that is rhythmic in only one tissue is genuinely cycling, given the evidence from the other 11 tissues. We were able to address this concern in part by corroborating our results with ChIP-Seq data for core clock transcription factors (Koike et al., 2012), with the caveat that the ChIP-Seq data were performed only in the mouse liver. Given the increasing amounts of data now available, future methods should systematically integrate multiple sources and types of evidence for rhythm detection.

## ACKNOWLEDGMENTS

We would like to thank Matthew Stephens for many discussions regarding bootstrapping and John Hogenesch for sharing data from Zhang et al. (2014). This work was completed in part with resources provided by the University of Chicago Research Computing Center. This work was supported by the Defense Advanced Research Projects Agency (D12AP00023) [www.darpa.mil/](http://www.darpa.mil/). ALH is a trainee of the National Institutes of Health Medical Scientist Training program at the University of Chicago (grant NIGMS T32GM07281; [www.nigms.nih.gov/](http://www.nigms.nih.gov/)) and was supported in part by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (award T32EB009412; [www.nibib.nih.gov/](http://www.nibib.nih.gov/)). The content is solely the responsibility of the authors and does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.


## CONFLICT OF INTEREST STATEMENT

The author(s) have no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## NOTE

Supplemental material is available for this article online.

## ORCID ID

Alan L. Hutchison  <https://orcid.org/0000-0003-4161-0772>

## REFERENCES

- Bray NL, Pimentel H, Melsted P, and Pachter L (2016). Near-optimal probabilistic RNA-seq quantification. *Nat Biotech* 34(5):525-527.
- Brown MB (1975) A method for combining non-independent, one-sided tests of significance. *Biometrics* 31(4):987-992.
- Deckard A, Anafi RC, Hogenesch JB, Haase SB, and Harer J (2013) Design and analysis of large-scale biological rhythm studies: a comparison of algorithms for detecting periodic signals in biological data. *Bioinformatics* 29(24):3174-3180.
- Efron B and Tibshirani RJ (1994) *An Introduction to the Bootstrap*. Boca Raton (FL): CRC Press.
- Fisher RA (1925) *Statistical Methods for Research Workers*. Edinburgh (UK): Oliver and Boyd.
- Fitzgibbons TP, Kogan S, Aouadi M, Hendricks GM, Straubhaar J, and Czech MP (2011) Similarity of mouse perivascular and brown adipose tissues and their resistance to diet-induced inflammation. *Am J Physiol Heart Circ Physiol* 301(4):H1425-H1437.
- Flourakis M, Kula-Eversole E, Hutchison AL, Han TH, Aranda K, Moose DL, White KP, Dinner AR, Lear BC, Ren D, et al. (2015) A conserved bicycle model for circadian clock control of membrane excitability. *Cell* 162(4):836-848.
- Goriki A, Hatanaka F, Myung J, Kim JK, Yoritaka T, Matsubara A, Forger D, and Takumi T (2014). A novel protein, CHRONO, functions as a core component of the mammalian circadian clock. *PLoS Biol* 12(4):e1001839.
- Huang DW, Sherman BT, and Lempicki RA (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37(1):1-13.
- Huang DW, Sherman BT, and Lempicki RA (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4(1):44-57.
- Hughes ME, DiTacchio L, Hayes KR, Vollmers C, Pulivarthy S, Baggs JE, Panda S, and Hogenesch JB (2009) Harmonics of circadian gene transcription in mammals. *PLoS Genet* 5(4):e1000442.
- Hughes ME, Hogenesch JB, and Kornacker K (2010) JTK\_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J Biol Rhythms* 25(5):372-380.
- Hutchison AL and Dinner AR (2017) Correcting for dependent p-values in rhythm detection. *bioRxiv* 118547. doi: 10.1101/118547.

- Hutchison AL, Maienschein-Cline M, Chiang AH, Tabei SMA, Gudjonson H, Bahroos N, Allada R, and Dinner AR (2015) Improved statistical methods enable greater sensitivity in rhythm detection for genome-wide data. *PLoS Comput Biol* 11(3):e1004094.
- Keegan KP, Pradhan S, Wang JP, and Allada R (2007) Meta-analysis of *Drosophila* circadian microarray studies identifies a novel set of rhythmically expressed genes. *PLoS Comput Biol* 3(11):2087-2110.
- Kellogg DR (2003) Wee1-dependent mechanisms required for coordination of cell growth and cell division. *J Cell Sci* 116(24):4883-4890.
- Koike N, Kim T-k, and Takahashi JS (2012) Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science* 338(August):1-10.
- Lu M, and Stephens M (2016) Variance adaptive shrinkage (vash): flexible empirical bayes estimation of variances. *Bioinformatics* 32(22): 3428-3434
- Matsuo T, Yamaguchi S, and Mitsui S (2003) Control mechanism of the circadian clock for timing of cell division in vivo. *Science* 302:255-260.
- Panda S, Antoch MP, Miller BH, Su AI, Schook AB, Straume M, Schultz PG, Kay SA, Takahashi JS, and Hogenesch JB (2002). Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell* 109(3):307-320.
- Perelis M, Marcheua B, Ramsey KM, Schipma MJ, Hutchison AL, Taguchi A, Peek CB, Hong H, Huang W, Omura C, et al. (2015) Pancreatic  $\beta$  cell enhancers regulate rhythmic transcription of genes controlling insulin secretion. *Science* 350(6261):aac4250.
- Pimentel H, Bray NL, Puente S, Melsted P, and Pachter L (2017) Differential analysis of RNA-Seq incorporating quantification uncertainty. *Nature Methods* 14:687-690.
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43(7):e47.
- Sandler O, Mizrahi SP, Weiss N, Agam O, Simon I, and Balaban NQ (2015) Lineage correlations of single cell division time as a probe of cell-cycle dynamics. *Nature* 519(7544):468-471.
- Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3(1):1-26.
- Storch K-F, Lipan O, Leykin I, Viswanathan N, Davis FC, Wong WH, and Weitz CJLB (2002) Extensive and divergent circadian gene expression in liver and heart. *Nature* 417(6884):78-83.
- Thaben PF and Westermark PO (2014) Detecting rhythms in time series with RAIN. *J Biol Rhythms* 29(6):391-400.
- Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, and Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562-578.
- Tukey JW (1977) *Exploratory Data Analysis*. Reading (MA): Addison-Wesley.
- Ukai H and Ueda HR (2010) Systems biology of mammalian circadian clocks. *Annu Rev Physiol* 72:579-603.
- Wijnen H, Naef F, and Young MW (2005) Molecular and statistical tools for circadian transcript profiling. *Methods Enzymol.* 393:341-365.
- Wu G, Anafi RC, Hughes ME, Kornacker K, and Hogenesch JB (2016) MetaCycle: an integrated R package to evaluate periodicity in large scale data. *Bioinformatics* 1-3(July):040345.
- Wu J, MacDonald J, Gentry J, and Irizarry R (2018) gcrma: Background Adjustment Using Sequence Information. R package version 2.52.0.
- Yang R and Su Z (2010) Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics* 26(12):i168-i174.
- Zhang R, Lahens NF, Ballance HL, Hughes ME, and Hogenesch JB (2014) A circadian gene expression atlas in mammals: implications for biology and medicine. *Proc Natl Acad Sci U S A* 111(45):16219-16224.