


# TimeTrial: An Interactive Application for Optimizing the Design and Analysis of Transcriptomic Time-Series Data in Circadian Biology Research

Elan Ness-Cohn,<sup>\*,†</sup>  Marta Iwanaszko,<sup>\*,†,‡</sup> William L. Kath,<sup>†,§,||</sup> Ravi Allada,<sup>†,||</sup>  
and Rosemary Braun<sup>\*,†,§,#,1</sup>

<sup>\*</sup>Biostatistics Division, Department of Preventive Medicine, Northwestern University, Chicago, Illinois, USA, <sup>†</sup>NSF-Simons Center for Quantitative Biology, Northwestern University, Evanston, Illinois, USA, <sup>‡</sup>Silesian University of Technology, Department of Systems Biology and Engineering, Gliwice, Poland, <sup>§</sup>Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL, USA, <sup>||</sup>Department of Neurobiology, Northwestern University, Evanston, IL, USA, and <sup>#</sup>Department of Physics and Astronomy, Northwestern University, Evanston, Illinois, USA

**Abstract** The circadian rhythm drives the oscillatory expression of thousands of genes across all tissues, coordinating physiological processes. The effect of this rhythm on health has generated increasing interest in discovering genes under circadian control by searching for periodic patterns in transcriptomic time-series experiments. While algorithms for detecting cycling transcripts have advanced, there remains little guidance quantifying the effect of experimental design and analysis choices on cycling detection accuracy. We present TimeTrial, a user-friendly benchmarking framework using both real and synthetic data to investigate cycle detection algorithms' performance and improve circadian experimental design. Results show that the optimal choice of analysis method depends on the sampling scheme, noise level, and shape of the waveform of interest and provides guidance on the impact of sampling frequency and duration on cycling detection accuracy. The TimeTrial software is freely available for download and may also be accessed through a web interface. By supplying a tool to vary and optimize experimental design considerations, TimeTrial will enhance circadian transcriptomics studies.

**Keywords** circadian biology, computational biology, biostatistics, experimental design guidelines, rhythm detection, benchmarking, gene expression analysis

The circadian rhythm, observed as a periodic 24-h behavioral and physiological cycle at the organismal level, is governed by an evolutionarily conserved set of core clock genes operating at the transcriptional and protein level. Consisting of only a few genes, the

circadian clock coordinates a vast array of cellular processes, including the cyclic expression of nearly half the genes across all tissues (Zhang et al., 2014). This rhythm can be entrained to environmental cues (zeitgebers) such as light, temperature, and food,

1 To whom all correspondence should be addressed: Rosemary Braun, Biostatistics Division, Department of Preventive Medicine, Northwestern University, 750 N Lake Shore Dr, Rubloff 11-166, Chicago, IL 60611-3008, USA; e-mail: rbraun@northwestern.edu.



allowing external stimuli to modulate time-of-day-specific functions. While numerous epidemiological studies have demonstrated significant links between circadian rhythms and human health (Levi and Schibler, 2007; Roenneberg et al., 2007; Chang et al., 2009; Puttonen et al., 2010; Kathale and Liu, 2014; Videnovic et al., 2014; Zhang et al., 2014; Patke et al., 2017; Braun et al., 2018), the underlying mechanisms linking the circadian clock to health outcomes remain largely unknown.

The advent of high-throughput omic technology now enables researchers to investigate these mechanisms in molecular detail by tracking the expression of thousands of transcripts over the course of the day, with the goal of identifying specific genes under circadian control. However, there are a number of analytical challenges in extracting rhythmic signals from noisy transcriptomic data. First, experimental limitations constrain the frequency and length of sampling, requiring inferences to be made from sparse or short time-series measurements. Second, cycling genes' expression profiles often do not exhibit sinusoidal trajectories; sharp peaks, damped oscillations, and additive linear trends have all been observed.

To address the complexities of circadian rhythm detection, a range of nonparametric methods have been developed (Hughes et al., 2010; Yang and Su, 2010; Thaben and Westermarck, 2014; Perea et al., 2015; Wu et al., 2016; Hutchison et al., 2018). These methods search for evidence of periodicity (e.g., by testing for correlations with template waveforms; Hughes et al., 2010; Hutchison et al., 2018). However, subsequent benchmarking studies demonstrate that different methods can yield conflicting results when run on the same data set (Serpedin et al., 2008; Deckard et al., 2013; Wu et al., 2014). Moreover, performance depends on the shape of the signal being detected, noise levels, and sampling schemes (Hughes et al., 2009; Deckard et al., 2013).

In addition to performance differences, there are also considerations of various methods' abilities to handle replicates, uneven sampling, missing data, and computational efficiency. In practice, the ability for methods to adequately handle these features directly affects a researcher's flexibility in experimental design. For instance, a method that can accommodate uneven sampling can allow for dense sampling at times of interest, with sparser sampling at other times. Because missing data often occur as a result of sequencing errors with greater likelihood as sample size increases (Gierliński et al., 2015), researchers benefit from algorithms that can handle missingness without the need to impute data. Finally, computational efficiency allows for data set sizes to grow while still processing the data in a reasonable amount of time. Taken together, methodological constraints

imply that the choice of cycling detection method will necessarily affect the optimal experimental design and vice versa.

These considerations, coupled with the need to limit costs, imply that designing an optimal circadian time-series experiment is a nontrivial task. While recommendations for experimental designs have been made (Hughes et al., 2007; Wu et al., 2014; Hughes et al., 2017), quantitative tools to flexibly and comprehensively weigh these considerations in the context of real data remain lacking. Moreover, while researchers have attempted to define criteria for method usage (Deckard et al., 2013; Wu et al., 2014; Hutchison et al., 2018), no guidance exists for custom sampling schemes, as previous studies used fixed sampling schemes and a limited number of waveform shapes.

To address these challenges, this article focuses on optimizing circadian rhythm detection by introducing a framework to evaluate the reliability of cycling detection as sampling schemes, waveform shapes, and cycling detection algorithms are varied. The results provide valuable evidence-based guidance for experimental design and analysis choices. As part of this work, we developed TimeTrial: an interactive, user-friendly, open-source software suite that enables circadian researchers to perform head-to-head comparisons of four leading cycle detection methods (JTK\_CYCLE [Hughes et al., 2010], ARSER [Yang and Su, 2010], RAIN [Thaben and Westermarck, 2014], and BooteJTK [Hutchison et al., 2018]; Suppl. Table S1) using both synthetic and real data. With TimeTrial, researchers can further explore these methods' performance under different noise levels, number of replicates, length of sampling, sampling resolution, and waveform shapes. An innovative feature of TimeTrial is the ability for the researcher to specify an arbitrary custom sampling scheme and obtain comparison of the cycling detection results, allowing them to gauge how their design choices may affect the experimental findings. Together, these results will enhance rigor and reproducibility in future circadian time-series experiments and improve tools to analyze cycling genes in increasingly large and complex datasets.

## RESULTS

TimeTrial was developed as a tool for the design and optimization of omic time-series experiments in circadian biology research. Consisting of two interactive applications using both synthetic and real data, TimeTrial allows researchers to explore the effects of experimental design on cycling detection, examine the reproducibility of cycling detection methods across biological data sets, and optimize experimental design for cycle detection. Applied to four cycle

detection methods (JTK\_CYCLE [Hughes et al., 2010], ARSER [Yang and Su, 2010], RAIN [Thaben and Westermarck, 2014], and BooteJTK [Hutchison et al., 2018]; Suppl. Table S1), our results reveal that no method consistently outperforms all others in all circumstances but rather that the performance depends on the sampling schemes and waveforms of interest. An interactive interface allows researchers to explore performance under different sampling schemes (including varying lengths, resolutions, and irregular sampling), providing valuable guidance for the optimization of circadian transcriptomic experiments given practical constraints (e.g., number of samples) and the signals of interest.

TimeTrial provides insights into cycling detection performance using both synthetic data and real data. The synthetic data provide precise control over the input data dynamics and noise, allowing the accuracy of cycling detection to be directly assessed when the ground truth (cycling/non-cycling) is known. The real data, which come from multiple studies, allow cycling detection methods to be assessed in terms of the reproducibility of the findings in biologically representative datasets.

### Synthetic Data: Simulating Gene Expression Dynamics

To comprehensively evaluate the performance of cycling detection methods for different patterns of temporal gene expression, we systematically created synthetic data sets consisting of varying number of replicates, sampling intervals, sampling lengths, and noise levels for a variety of waveform shapes; in total, these represent 240 combinations of conditions (Fig. 1A). One thousand “genes” were simulated with varying amplitudes, phases, and shape parameters (e.g., the envelope for damped/amplified waves) for each of the 11 base waveforms (Fig. 1B), yielding in total 11,000 simulated genes for each of the 240 conditions. The choice of waveform shapes was inspired by patterns observed in experimental circadian data sets (Suppl. Fig. 1) and is designed to give the user an avenue to explore the types of patterns that would be classified as cycling or noncycling for various sampling and analysis choices. Further details of the synthetic data sets can be found in the Methods section and the supplement.

### Biological Data: Reproducibility Analysis

While synthetic data have the advantage of a known ground truth, they have the drawback of not necessarily being representative of real biological data sets. On the other hand, measuring a method’s

accuracy using real data is limited, as the ground truth is not generally known. Instead, one may test the reproducibility of the results, under the assumption that a true biological signal should be consistently detected across multiple studies of the same condition.

To this end, we took a “cross-study concordance” approach in which we tested methods’ ability to consistently characterize a set of 12,868 genes measured in three independent studies as cycling or noncycling. By evaluating the rank correlation  $\rho$  of the cycling detection  $p$ -values obtained from the various studies, our analysis directly quantifies reproducibility. We analyzed three distinct mouse liver time-series expression sets (Hughes et al., 2009, 2012; Zhang et al., 2014) to evaluate the concordance. In addition, we down-sampled each data set to mimic the effect of sparser sampling. Additional details can be found in the Methods section.

### Experimental Design Recommendations

Using TimeTrial to analyze common sampling schemes in both synthetic and biological data sets, the following sections present recommendations for the experimental design framework with regard to selection of sampling scheme and concatenation. For the development of customized sampling schemes that may better fit the individual researcher’s needs, users are encouraged to explore TimeTrial’s custom sampling feature as described in the TimeTrial Capabilities and Usage section.

*Sampling resolution.* While long, frequently sampled time series provide the clearest picture of circadian dynamics, this must be balanced with practical considerations such as experimental cost. It is thus of interest to identify an optimal sampling scheme by varying sampling length, resolution, and replicates. To determine the limits of cycling detection as the frequency and length of sampling is reduced, we down-sampled three data sets and compared the results across all genes (Fig. 1C).

To determine whether a method provides consistent results at lower sampling, we compute the rank correlation  $\rho$  of the cycling detection  $p$ -values across all genes for different data sets and sampling schemes to quantitatively determine whether two methods yield the same ranking of cycling genes, without setting arbitrary  $p$ -value thresholds. Results revealed that 1-h and 2-h sampling resolutions show high correlation across all data sets and methods, with  $\rho$  values between 0.79 and 0.94 (Suppl. Fig. S2). This implies that at 1- and 2-h sampling, all methods reproducibly rank the same genes (from most to least

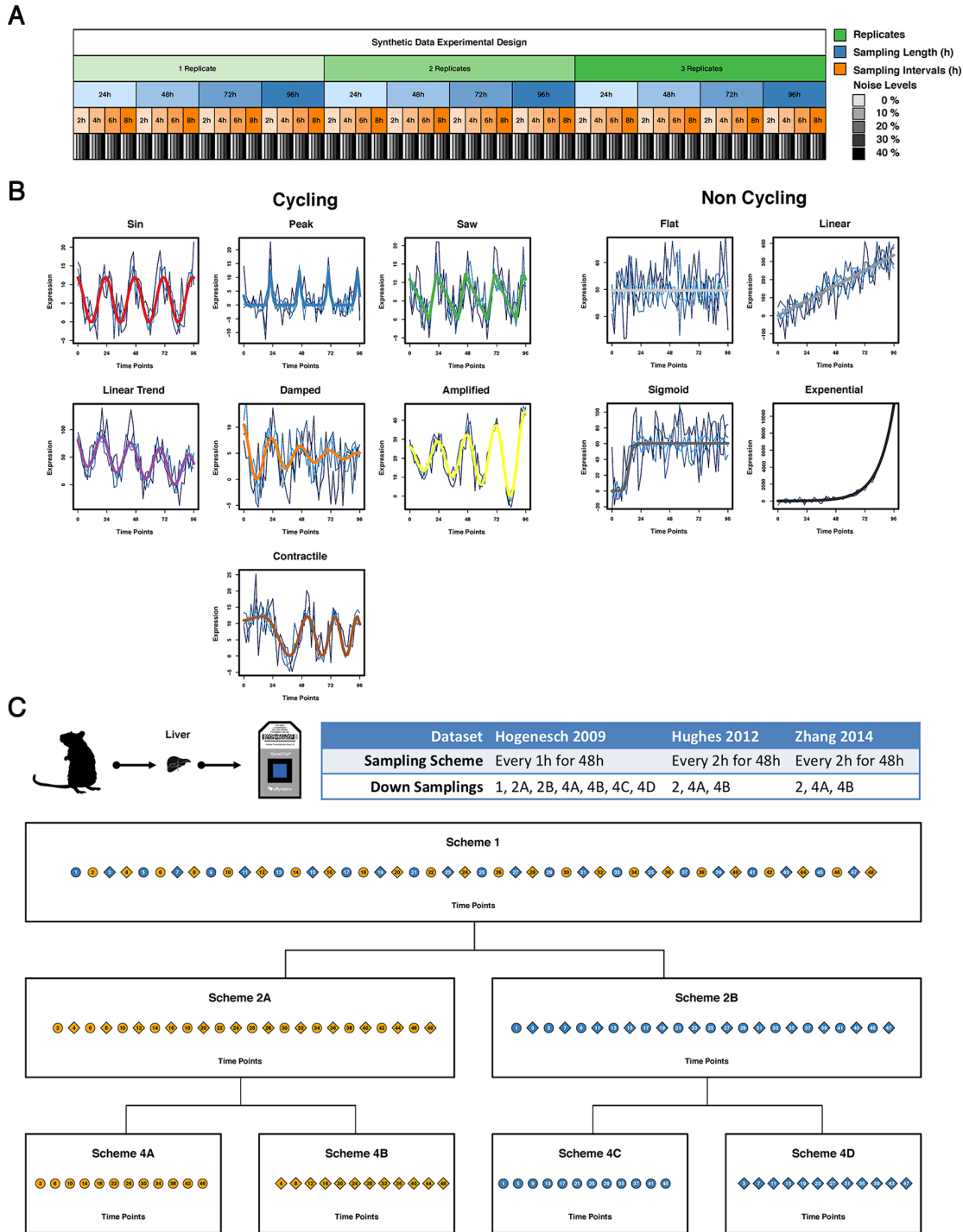
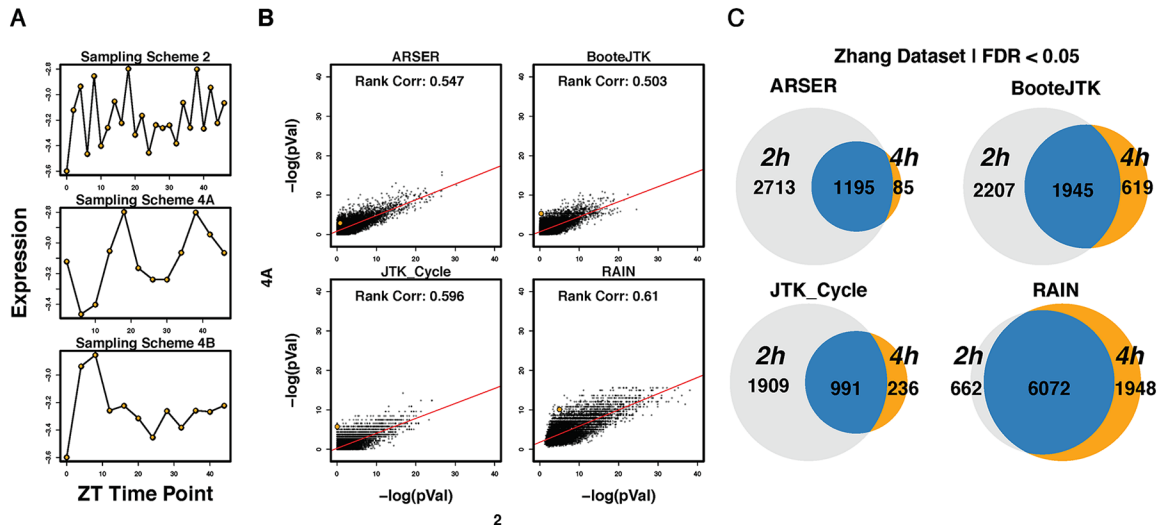


Figure 1. Experimental design, synthetic and biological data. (A) A total of 240 unique synthetic time-course data sets were generated in R with known ground truth. Each data set consisted of a different number of replicates, sampling intervals, sampling lengths, and noise levels as a percentage of the wave form amplitude. Each lowest-level rectangle represents an independent experiment. (B) Each data set consisted of 11 different classes of waveforms commonly seen in real biological data. Seven were classified as cycling and four as noncycling. Signals are shown with additive Gaussian noise as a percentage of the wave form amplitude (i.e., 10%, 20%, 30%, and 40%). All cycling waveforms had a set period of 24-h, except contractile, which varied over time. (C) Three mouse liver time-series expression sets were analyzed from the Gene Expression Omnibus database: Hogenesch 2009 (GSE11923), Hughes 2012 (GSE30411), and Zhang 2014 (GSE54650). The Hughes and Zhang studies, sampled every 2-h for 48-h, were down-sampled to two data sets sampled every 4-h (Hughes\_4A and Hughes\_4B; Zhang\_4A and Zhang\_4B). The Hogenesch study, sampled every 1-h for 48-h, was down-sampled to two data sets sampled every 2-h (Hogenesch\_2A and Hogenesch\_2B) and also into four data sets sampled every 4-h (Hogenesch\_4A, Hogenesch\_4B, Hogenesch\_4C, and Hogenesch\_4D).





**Figure 2.** Effect of experimental resolution. (A) Expression time courses of the P2ry10b gene at 2-h and the two down-sampled 4-h (4A and 4B) resolution from the Zhang data set. (B) Scatterplots of  $-\log(p\text{Vals})$  of 2-h sampling versus 4-h sampling across all methods.  $p$ -values are not corrected for false-discovery rate (FDR) for comparison purposes. The orange point denotes the P2ry10b gene highlighted in the right panel. (C) Venn diagram of genes detected as cycling (FDR-adjusted  $p$ -value  $< 0.05$ ) in the 2-h versus 4-h down-sampled Zhang data set across all methods. Genes detected as cycling in the 4A and 4B conditions were grouped together.

cycling) in the three independent data sets. It also suggests that different methods identify the same genes at the top of their respective lists, implying that the choice of methods does not significantly affect the results at these resolutions. In contrast, the results from 4-h sampling resolutions show poorer correlations across data sets and methods, with  $\rho$  values between 0.67 and 0.89 (Suppl. Fig. S2), implying that results become less reliably reproducible with 4-h sampling. Moreover, when working in the low-density regimes, there is less concordance between methods, suggesting that the choice of method has a large impact on the cycle detection results (Suppl. Fig. S3). Taken together, these results suggest that cycle detection is robust at 1- and 2-h sampling intervals but becomes significantly less reliable at 4-h sampling intervals.

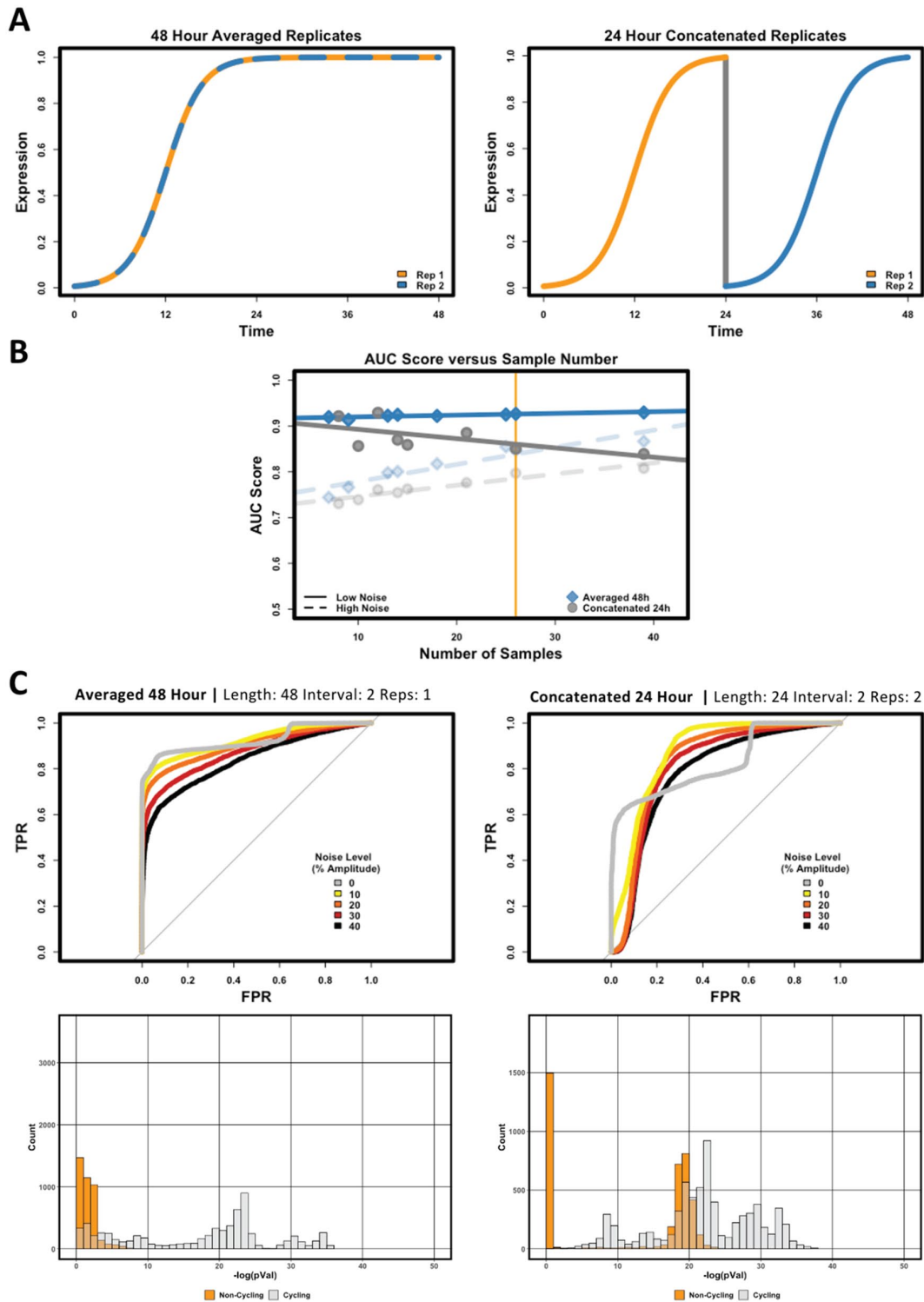
Importantly, we find that 4-h sampling not only misses cycling genes that are detected at 1- and 2-h sampling (false-negatives) but can also lead to erroneously calling noncycling genes as cycling (false-positives). Figure 2A illustrates a gene that is not rhythmic in the 2-h sampling but appears periodic when down-sampled to 4-h. A cluster of such genes can be seen in the lower right of the scatter plots in (Fig. 2B), where genes have a nonsignificant  $-\log(p)$  in the 2-h sampling scheme but a significant  $-\log(p)$  in the 4-h sampling scheme. Figure 2C shows the overlap of genes detected as cycling at FDR  $< 0.05$  under 2- and 4-h sampling. In addition, we investigated whether known core clock genes were consistently detected as cycling; although they are robustly detected in the 1-h and all 2-h data sets, they are less

reliably detected in the 4-h data, further corroborating the findings above.

One may then ask whether it is better to devote resources to more frequent sampling or to a greater number of replicates at lower sampling rates. From the standpoint of sequencing cost, sampling every 4-h for 48-h with 2 replicates is the same as sampling every 2-h for 48-h with 1 replicate. Tests with synthetic data show these 2 schemes are similarly powered in their ability to detect true cycling genes; however, the 2-h single-replicate scheme has the benefit of fewer false-positives compared with 4-h duplicate sampling (Fig. 2A). These findings suggest that sampling every 2-h for 48-h with a single replicate is advantageous over sampling every 4-h in duplicate.

**Concatenation bias.** Concatenation of replicate time series is common practice in the field of circadian biology (Duong et al., 2011), wherein researchers will concatenate two replicate 24-h series into a single 48-h series. This is done under the assumption that if a signal has a true 24-h period, concatenation of the signal will maintain its cyclic nature. However, concatenation of replicates can induce apparent 24-h periodicity for nonperiodic signals (Fig. 3A).

As we increase the sampling resolution in the low-noise regime, accuracy of the concatenated data decreases, as seen by the negative slope in Figure 3B (solid gray line). Paradoxically, this means our ability to correctly classify genes becomes worse with more samples with less noise. A closer look at the samples highlighted by the orange line in Figure 3B illustrates



**Figure 3. Concatenation bias.** (A) Cartoon representation of concatenated versus averaged replicates of sigmoidal waveform. Top: Nonoscillatory when sampled for 48-h with averaged replicates. Bottom: Artificially oscillatory when sampled for 24-h with concatenated replicates. (B) Area under the curve (AUC) scores of the 24-h and 48-h time courses from the synthetic data sets were plotted as a function of the number of samples using BooteJTK. Time courses varied in sampling interval and number of replicates. Twenty-four-hour time courses were concatenated, whereas the 48-h time courses were averaged across replicates. Solid lines represent AUC scores in the low-noise regimes (0% and 10%), and dashed lines represent AUC scores in the high-noise regimes (30% and 40%). The vertical line represents the samples highlighted in panel B of the figure. (C) Top: ROC curves for all noise levels at 48-h every 2-h with 1 replicate and 24-h every 2-h with 2 replicates sampling scheme. Both schemes have the same number of time points. Bottom: Histograms of  $-\log(p)$  values of cycling (light gray) and non-cycling (orange) waveforms in each condition. Larger  $-\log(p)$  values denote waveforms detected as more significantly cycling by BooteJTK.

the reason for this effect (24-h at 2-h with 2 replicates vs. 48-h at 2-h with 1 replicate; Fig. 3C). The receiver-operating characteristic (ROC) curves show that concatenation of two 24-h replicates increases the false-positive rate over the 48-h case due to periodic replication of any transient dynamics in the first 24-h period (Fig. 3C). As we increase our sampling resolution, we can better discern signal shape; coupled with concatenating expression patterns at the 24-h mark, this increased clarity causes methods searching for periodicity with a period of 24-h to erroneously classify sigmoidal, linear, and exponential signals as rhythmic. This finding corroborates other published studies (Hughes et al., 2009, 2017) that have likewise strongly recommended against concatenation. Taken together, these results imply time-series data should not be concatenated prior to statistical testing.

### TimeTrial Capabilities and Usage

The optimal choice of cycling detection method depends on sampling scheme, noise level, number of missing data points, number of replicates, and shape of the waveform of interest. Thus, in addition to the above recommendations, we provide TimeTrial as a freely available tool for users to explore cycling detection performance and to optimize their sampling schemes.

As described above, TimeTrial consists of 2 components: one that analyzes cycling detection accuracy using synthetic data and another that analyzes cycling detection reproducibility using real data. In the first TimeTrial component, using synthetic data, users can experiment with different number of replicates, sampling lengths, sampling resolutions, and noise levels to assess their effects on the detection of specific signal shapes (Fig. 4). By comparing the reported ROC curves, area under the curve (AUC) scores, and *p*-value distributions for each different sampling scheme and method, users can determine the optimal sampling scheme and method for cycle detection of various waveforms (i.e., cosine, peak, saw-tooth, etc.). Furthermore, users can assess the robustness of each method when noise is introduced by comparing the standard deviation AUC scores. Given a specified sampling scheme, methods with lower standard deviations in AUCs imply the cycling detection results are stable across increasing noise values, whereas larger standard deviations indicate results are more strongly influenced by the noise level.

In addition, users can determine how their choices affect the *p*-value distribution of signal shapes to determine how robust a sampling scheme and method are at separating specific types of cycling signals from noncyclers. Finally, users can adjust *p*-value thresholds and inspect output on a per-signal basis to

help determine appropriate threshold cutoffs for defining separation between signal shapes for downstream analysis.

In the second TimeTrial component, using real data, users can compare sets of genes detected as cycling by different methods and under different sampling schemes (Fig. 5). By comparing results across each data set and down sampling, users can perform a concordance analysis to determine which genes are picked up by each method and each sampling scheme. The TimeTrial interface allows the user to explore cycling detection results at various levels of significance and minimal fold-change (a criterion commonly used to reduce false-positives). Given the knowledge of how sampling scheme effects the detection of waveform shape from use of the synthetic data sets, users can test the ability to pick up these shapes in the biological data set and judge whether the waveform shapes being detected as cycling are representative of the patterns they wish to detect.

### Designing Circadian Experiments with TimeTrial

Most importantly, TimeTrial enables users to develop their own custom sampling scheme for cycle detection (Fig. 6). While sampling every hour for 48-h would be ideal for cycling detection, it is also expensive. TimeTrial allows the user to explore how scheduling fewer samples will affect the results and explore whether enhanced sampling at specific times of day can improve detection. Irregular sampling schemes may be beneficial for scientific or practical considerations. For instance, one may be interested in monitoring not only cycling genes but how their dynamics change immediately following an exposure; in this case, one might wish to bias samples toward the time immediately following the stimulus, at the expense of fewer samples later in the time course. The impact of these choices can be explored using TimeTrial's down-sampling tool to test how this alternative sampling scheme compares to the ideal sampling scheme. (Note that the custom analysis is performed using only JTK\_CYCLE and RAIN, since ARSER and BooteJTK require regularly spaced samples; see Suppl. Table S1.) Finally, users can further explore how adjusting the times and spacing of sampling might improve detection and query genes of interest to determine if a specific gene and/or core clock genes are detected.

The TimeTrial application is freely available for download at <https://github.com/nesscoder/TimeTrial> and may also be accessed through a web interface hosted on shinyapps.io: [https://nesscoder.shinyapps.io/TimeTrial\\_Synthetic/](https://nesscoder.shinyapps.io/TimeTrial_Synthetic/), [https://nesscoder.shinyapps.io/TimeTrial\\_Real/](https://nesscoder.shinyapps.io/TimeTrial_Real/). (Note that we

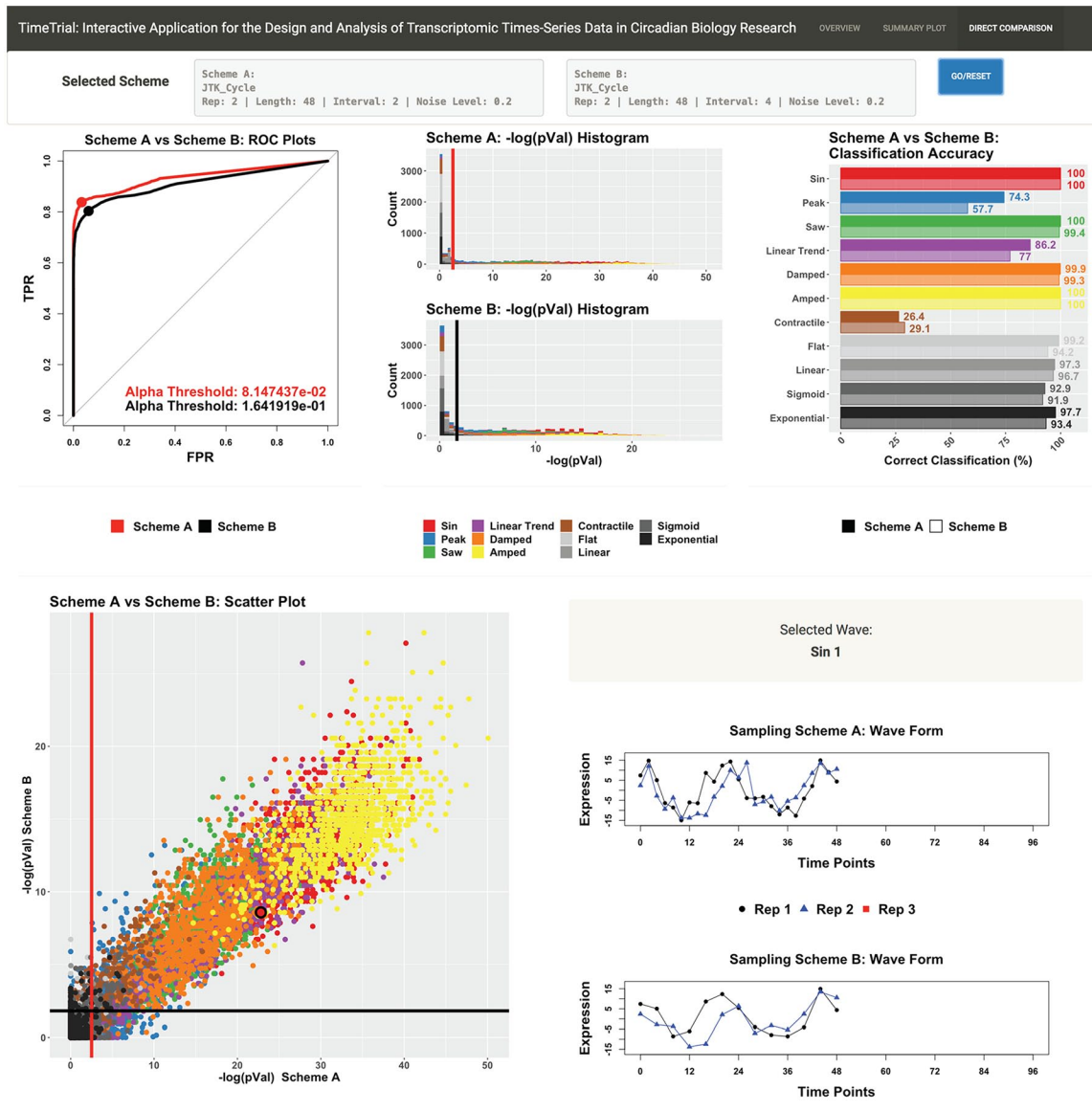


Figure 4. Synthetic data. TimeTrial: interactive application for circadian rhythm study design. The synthetic data set version of TimeTrial allows users to directly compare methods across different sampling lengths, sampling intervals, number of replicates, and noise levels. Users can further set different significance thresholds and take closer looks into the ability of methods to detect different waveform patterns. See <https://github.com/nesscoder/TimeTrial> and/or [https://nesscoder.shinyapps.io/TimeTrial\\_Synthetic/](https://nesscoder.shinyapps.io/TimeTrial_Synthetic/) for interactive plots and a complete tutorial.

recommend local installation from GitHub for greater speed and reliability, as it does not depend on the user's internet connection.) Full documentation and walk-through tutorials are provided to guide users in the optimal use of these tools.

## DISCUSSION

We developed TimeTrial, a tool that uses synthetic and real data to assist researchers in the design and analysis of circadian time-series experiments that

optimize cycling detection. We applied TimeTrial to explore the effects of experimental design on signal shape, examine cycling detection reproducibility across biological data sets, and optimize experimental design for cycle detection. By comparing the performance of different cycling detection algorithms under different sampling schemes, TimeTrial provides valuable guidance for the design of rigorous, reproducible circadian transcriptomics studies. We expect that these results will be of interest to experimentalists and computational researchers alike.

From an experimental perspective, our results suggest several guidelines for designing circadian



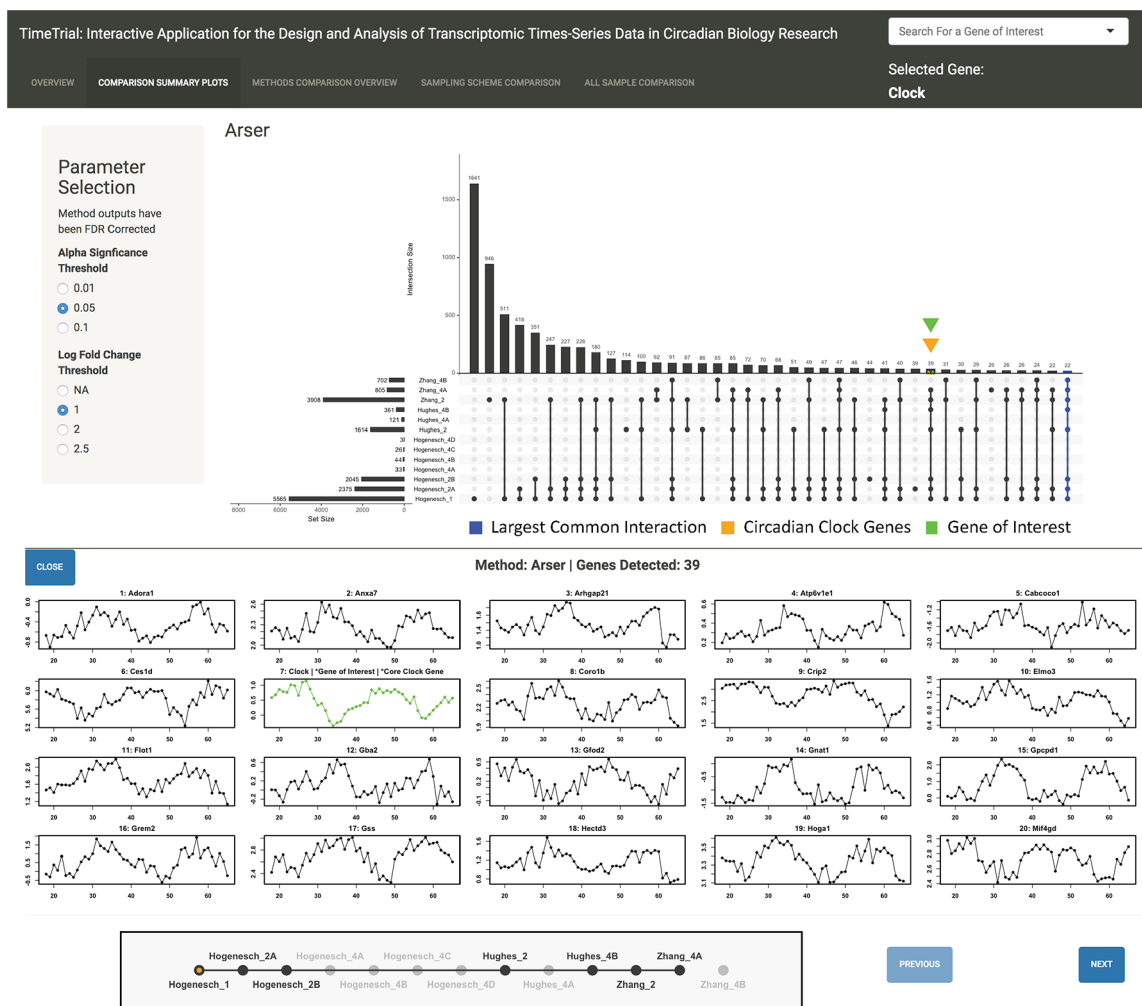


Figure 5. Real data. TimeTrial: exploring processed data. The real data set version of TimeTrial allows users to directly compare methods and sampling schemes. Users can set different significance and log-fold change thresholds to explore the ability of different methods to pick up circadian clock genes (orange triangle) and genes of interests (green triangle) across data sets. See <https://github.com/nesscoder/TimeTrial> and/or [https://nesscoder.shinyapps.io/TimeTrial\\_Real/](https://nesscoder.shinyapps.io/TimeTrial_Real/) for interactive plots and a complete tutorial.

time-series studies. First, we demonstrated that 24-h sampling with concatenation introduces biases that increase the number of false-positives, and therefore, this practice should be avoided (Hughes et al., 2017). Moreover, our findings suggest that 2-h resolution is required at a minimum to pick up the dynamical transcriptional changes that occur on the 24-h circadian scale (corroborating earlier findings; Hughes et al., 2009, 2017) and that a 2-h resolution with a single replicate is advantageous over a 4-h resolution in duplicate. We note that the errors produced with the 4-h schemes include both false negatives (missed cyclers) and false positives (noncyclers erroneously classified as cycling), compromising the reliability of results obtained from 4-h sampling schemes. Finally, we observe that different detection methods exhibit different performance depending on the underlying waveform shape, suggesting that the researcher

should consider the patterns of interest when selecting an analysis method. For instance, a researcher may decide that classifying a waveform with a strong linear drift as cyclic may be (un)desirable, in which case, a method that calls these patterns as (non) cycling should be selected.

From a computational perspective, our results provide a means to benchmark new methods based on their accuracy in synthetic data and reproducibility in real data. They also indicate methodological gaps. Notably, our findings did not define a clear overall “winner” among the methods tested, suggesting that there is still a need for methods that perform consistently well in multiple conditions for a variety of waveform shapes. Among these findings, we note that no method detects as cycling “contractile” waveforms in which the period changes (as might be the case when an environmental change is introduced),

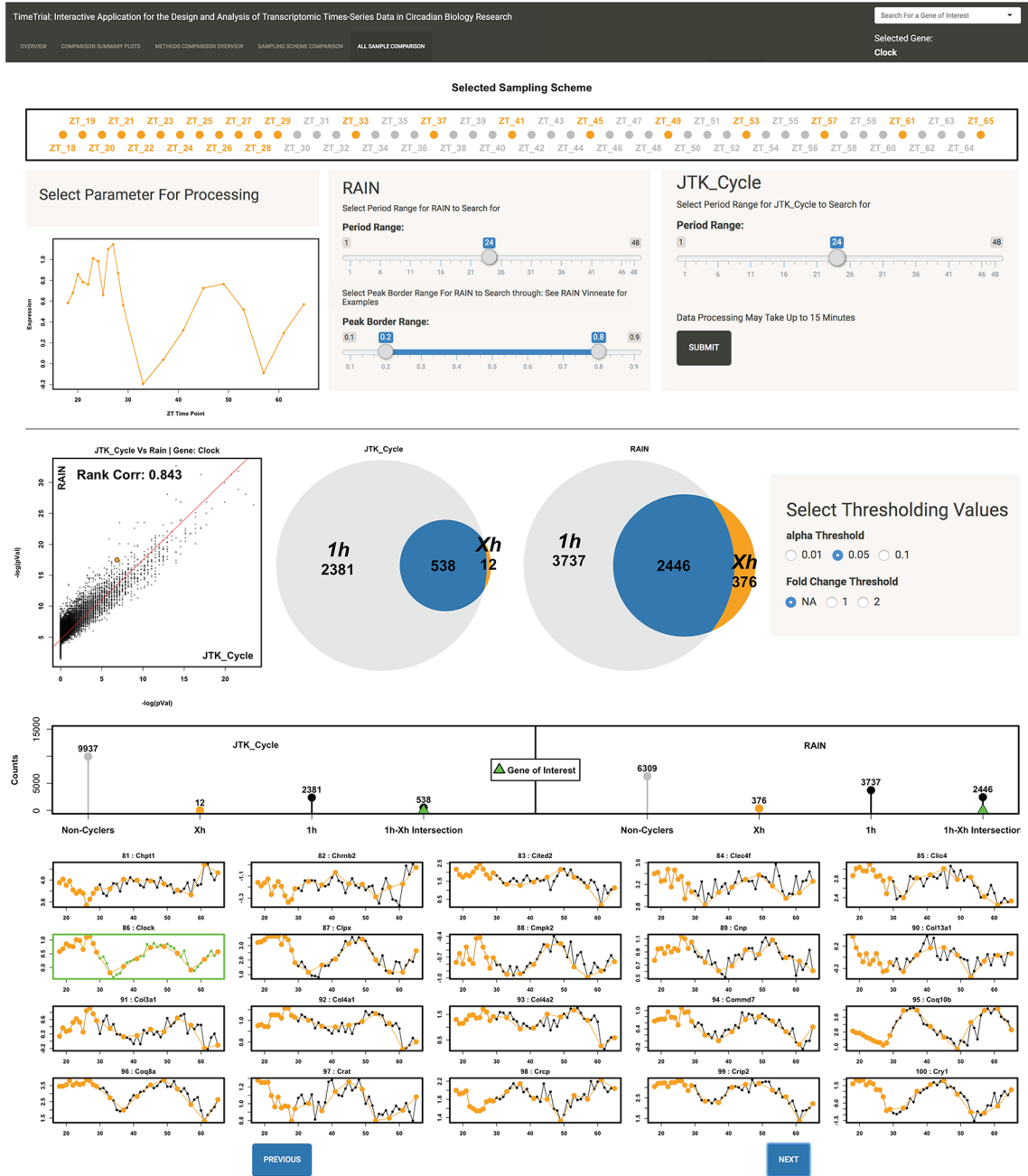


Figure 6. Real data. TimeTrial: testing arbitrary sampling schemes. The real data set version of TimeTrial allows users to define their own custom down-sampled sampling scheme and compare the results to that of sampling every 1-h for 48-h. The custom sampling scheme analysis is performed using only JTK\_CYCLE and RAIN, since ARSER and BooteJTK cannot handle uneven sampling (Suppl. Table S1). Users can further set different significance and log-fold change thresholds and query for genes of interest. See <https://github.com/nesscoder/TimeTrial> and/or [https://nesscoder.shinyapps.io/TimeTrial\\_Real/](https://nesscoder.shinyapps.io/TimeTrial_Real/) for interactive plots and a complete tutorial.

indicating the need for new methods in which such patterns are of interest. Our results also highlight the need for methods that have more robust performance with 4-h sampling designs; the development of such a method would make future experiments more feasible and enable reanalysis of existing data. Finally, we note that any new method should be designed to handle biological and technical replicates in a

justifiable way (without requiring concatenation and ideally without averaging so that the full information about the variance in the data is retained), permit missing data and/or uneven sampling, and be computationally efficient (Suppl. Table S1).

Researchers should also be aware of read-depth considerations when performing cycling detection using next-generation sequencing. Our analysis was

performed on publicly available microarray data, and thus read depth was not considered a factor in the present analysis. However, previous work has recommended optimal read depths for cycling detection: ~10 million reads per sample to detect >75% of cycling transcripts in fly RNA-seq studies and ~40 million reads per sample for studying mammals (Li et al., 2015). In the context of TimeTrial, read depth will effect the cost of sampling and thus acts as a constraint on the number of samples a researcher has at his or her disposal. Once the sample number is determined, TimeTrial can be used to help determine the optimal sampling scheme given this constraint. As more next-generation circadian time-series sequencing data become publicly available, future versions of TimeTrial will include the effects of read depth by allowing users to vary this parameter.

Our benchmarking approach is unique in that it provides an assessment of cycling detection for arbitrary sampling schemes. Previous studies were done using fixed sampling frequencies, number of replicates, and sampling lengths. Our analysis used a mixture of different sampling resolutions, number of replicates, sampling lengths, and noise levels. Ultimately, we attempted to model biological experimental practice, in which a fixed sampling scheme is not always possible as a result of monetary constraints. Our findings suggest that the sampling schemes and signal shape, rather than the cycling detection method, will have the largest impact on cycle detection. Thus, in designing a time-series experiment, researchers should contemplate the total number of samples at their disposal; how those samples should be used across replicates, length, and intervals; and how the chosen scheme allows cycling detection methods to pick up different signal shapes (i.e., symmetric, nonsymmetric, peak, trends, etc.).

In addition, our benchmarking approach is unique in that it explicitly considers the reproducibility of cycling detection results. By considering the concordance of genes detected as cycling across multiple independent data sets, we directly assess whether genes detected in one study would be validated in another. We propose that this assessment of reproducibility, rather than the number of cycling genes detected, should be the standard against which new methods are judged.

## METHODS

### Generating Synthetic Data Sets

A total of 240 unique synthetic time-course data sets, each comprising 11,000 expression profiles, were generated in R. Each data set consisted of a different

number of replicates (1, 2, 3), sampling intervals (2-h, 4-h, 6-h, 8-h), sampling lengths (24-h, 48-h, 72-h, 96-h), and noise levels as a percentage of the wave form amplitude (0%, 10%, 20%, 30%, 40%; Fig. 1A).

Within each condition, 11 base waveforms were simulated to mimic expression patterns observed in nature: periodic patterns, nonperiodic patterns, and dynamics that have a cyclic component but do not meet the strict definition of periodicity. Seven of these 11 shapes were considered cyclic (sine, peak, sawtooth, linear trend, damped, amplified, contractile), and 4 were considered noncyclic (flat, linear, sigmoid, and exponential); examples are given in Figure 1B. For each waveform in each condition, 1000 “genes” were simulated with varying amplitudes, phases, and shape parameters (e.g., the envelope for damped/amplified waves), yielding in total 11,000 simulated genes for each of the 240 conditions. Variation in amplitude of the underlying functions were drawn from a log-normal uniform distribution with a mean 1.302 and standard deviation 0.30, as modeled from real data amplitude distributions to simulate differences in amplitude between genes. Additional variation in the phase of underlying functions were drawn from a uniform distribution between 0 and  $2\pi$ , to simulate differences in phase between genes. The data were mean centered, as is common in preprocessing for cycle detection. A complete list of the waveform function definitions and source code for generating the data can be found in the supplementary material.

### Preprocessing Microarray Datasets

The CEL files from three mouse liver Affymetrix microarray time-series expression sets (Hogenesch 2009 - GSE11923 [Hughes et al., 2009], Hughes 2012 - GSE30411 [Hughes et al., 2012], Zhang 2014 - GSE54650 [Zhang et al., 2014]) were downloaded from the Gene Expression Omnibus database (GEO). In each experiment, wild-type C57BL/6J mice were entrained to a 12-h light, 12-h dark environment before being released into constant darkness. Mouse age, length of entrainment, time of sampling, and sampling resolution vary by experiment. The data were subsequently normalized by robust multi-array average (rma) using the Affy R Package (Gautier et al., 2004) and checked for quality control using the Oligo R Package (Carvalho and Irizarry, 2010), following each package’s vignette, respectively. Since each GEO data set used a different microarray platform (affy\_mouse430\_2, affy\_moex\_1\_0\_st\_v1, affy\_mogene\_1\_0\_st\_v1), each had a different set of probes. A common set of features needed to be identified to compare across microarrays. Probes for each data set were mapped to genes based on prealigned

databases specific to each microarray (mouse4302.db, moex10sttranscriptcluster.db, mogene10sttranscriptcluster.db). Multiple probes corresponding to a single gene were aggregated by taking the mean expression. A final 12,868 common set of genes across all three microarray platforms were used for subsequent analysis. See the supplement for code.

### Processing Microarray Data Sets

To characterize the effects of sampling schemes using real data, the three data sets were down sampled to simulate the effects of sampling at 2-h and 4-h intervals. The Hughes 2012 and Zhang 2014 data sets were sampled every 2-h for 48-h. Each of these experimental time-series were down-sampled to every 4-h to generate four additional time-series data sets. The Hogenesch 2009 data set, sampled every hour for 48-h, was down-sampled to 2 data sets every 2-h and 4 data sets sampled every 4-h to generate six additional time-series data sets. Ultimately, 13 data sets (3 original and 10 down-sampled) were processed by all 4 cycling detection methods (ARSER, BooteJTK, JTK\_CYCLE, and RAIN), using each method's recommended parameter settings as defined by the sampling length and interval (Fig. 1C). We thus aimed to assess a method's robustness by the ability to consistently score genes as cycling versus noncycling across experimental data sets and down-samplings. A complete list of the experimental parameters and source code can be found in the supplement.

### Application of Cycling Detection Algorithms

All data sets were processed by all four cycling detection methods (JTK\_CYCLE [Hughes et al., 2010], ARSER [Yang and Su, 2010], RAIN [Thaben and Westermarck, 2014], and BooteJTK [Hutchison et al., 2018]; Suppl. Table S1), using each method's recommended parameter settings as defined by the sampling length and interval. Since ARSER and BooteJTK do not have a built-in function for dealing with replicates, replicates were either averaged together or concatenated, following the two common practices in the field. JTK\_CYCLE and RAIN used the replicate procedures recommended in their documentation. A complete list of the experimental parameters and source code can be found in the supplement.

### ACKNOWLEDGMENTS

Research reported in this publication was supported by the NSF-Simons Center for Quantitative Biology at Northwestern University, an NSF-Simons MathBioSys

Research Center. This work was supported by a grant from the Simons Foundation/SFARI (597491-RWC) and the National Science Foundation (1764421). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation and Simons Foundation.

### AUTHOR CONTRIBUTIONS

E.N.C. and R.B. designed the research; E.N.C., M.I., W.L.K., R.A., and R.B. contributed to the design requirements of the TimeTrial; E.N.C developed the TimeTrial software, analyzed the data, and produced the tutorials; E.N.C. and R.B. wrote the article; all authors read and approved the final manuscript.

### CONFLICT OF INTEREST STATEMENT

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### ORCID iD

Elan Ness-Cohn  <https://orcid.org/0000-0002-3935-6667>

### REFERENCES

- Braun R, Kath WL, Iwanaszko M, Kula-Eversole E, Abbott SM, Reid KJ, Zee PC, and Allada R (2018) Universal method for robust detection of circadian state from gene expression. *Proc Natl Acad Sci U S A* 115:E9247-E9256. doi:10.1073/pnas.1800314115
- Carvalho BS and Irizarry RA (2010) A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 26:2363-2367. doi:10.1093/bioinformatics/btq431
- Chang AM, Reid KJ, Gourineni R, and Zee PC (2009) Sleep timing and circadian phase in delayed sleep phase syndrome. *J Biol Rhythms* 24:313-321. doi:10.1177/0748730409339611
- Deckard A, Anafi RC, Hogenesch JB, Haase SB, Harer J, and Valencia A (2013) Design and analysis of large-scale biological rhythm studies: a comparison of algorithms for detecting periodic signals in biological data. *Bioinformatics* 29:3174-3180. doi:10.1093/bioinformatics/btt541
- Duong HA, Robles MS, Knutti D, and Weitz CJ (2011) A molecular mechanism for circadian clock negative feedback. *Science* 332:1436-1439. doi:10.1126/science.1196766



- Gautier L, Cope L, Bolstad BM, and Irizarry RA (2004) Affy: analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20:307-315. doi:10.1093/bioinformatics/btg405
- Gierliński M, Cole C, Schofield P, Schurch NJ, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson G, Owen-Hughes T, et al. (2015) Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics* 31:3625-3630. doi:10.1093/bioinformatics/btv425
- Hughes M, Deharo L, Pulivarthy SR, Gu J, Hayes K, Panda S, and Hogenesch JB (2007) High-resolution time course analysis of gene expression from pituitary. *Cold Spring Harbor Symp Quant Biol* 72:381-386. doi:10.1101/sqb.2007.72.011
- Hughes ME, Abruzzi KC, Allada R, Anafi R, Arpat AB, Asher G, Baldi P, de Bekker C, Bell-Pedersen D, Blau J, et al. (2017) Guidelines for genome-scale analysis of biological rhythms. *J Biol Rhythms* 32:380-393. doi:10.1177/0748730417728663
- 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:e1000442. doi:10.1371/journal.pgen.1000442
- 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* 25:372-380. doi:10.1177/0748730410379711
- Hughes ME, Hong HK, Chong JL, Indacochea AA, Lee SS, Han M, Takahashi JS, and Hogenesch JB (2012) Brain-specific rescue of Clock reveals system-driven transcriptional rhythms in peripheral tissue. *PLoS Genet* 8:e1002835. doi:10.1371/journal.pgen.1002835
- Hutchison AL, Allada R, and Dinner AR (2018) Bootstrapping and empirical Bayes methods improve rhythm detection in sparsely sampled data. *J Biol Rhythms* 33:339-349. doi:10.1177/0748730418789536
- Kathale ND and Liu AC (2014) Prevalence of cycling genes and drug targets calls for prospective chronotherapeutics. *Proc Natl Acad Sci USA* 111:15869-15870. doi:10.1182/blood-2014-05-577825
- Levi F and Schibler U (2007) Circadian rhythms: mechanisms and therapeutic implications. *Ann Rev Pharmacol Toxicol* 47:593-628. doi:10.1146/annurev.pharmtox.47.120505.105208
- Li J, Grant GR, Hogenesch JB, and Hughes ME (2015) Considerations for RNA-seq analysis of circadian rhythms. *Methods Enzymol* 551:349-367. doi:10.1016/bs.mie.2014.10.020
- Patke A, Murphy PJ, Onat OE, Krieger AC, Özçelik T, Campbell SS, and Young MW (2017) Mutation of the human circadian clock gene CRY1 in familial delayed sleep phase disorder. *Cell* 169:203-215.e13. doi:10.1016/j.cell.2017.03.027
- Perea JA, Deckard A, Haase SB, and Harer J (2015) SW1PerS: sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. *BMC Bioinform* 16:257. doi:10.1186/s12859-015-0645-6
- Puttonen S, Härmä M, and Hublin C (2010) Shift work and cardiovascular disease: pathways from circadian stress to morbidity. *Scand J Work Environ Health* 36:96-108. doi:10.5271/sjweh.2894
- Roenneberg T, Kuehnle T, Juda M, Kantermann T, Allebrandt K, Gordijn M, and Merrow M (2007) Epidemiology of the human circadian clock. *Sleep Med Rev* 11:429-438. doi:10.1016/j.smrv.2007.07.005
- Serpedin E, Zhao W, Agyepong K, and Dougherty ER (2008) Detecting periodic genes from irregularly sampled gene expressions: a comparison study. *Eurasip J Bioinform Syst Biol* 2008:1-8. doi:10.1155/2008/769293
- Thaben PF and Westermark PO (2014) Detecting rhythms in time series with RAIN. *J Biol Rhythms* 29:391-400. doi:10.1177/0748730414553029
- Videnovic A, Noble C, Reid KJ, Peng J, Turek FW, Marconi A, Rademaker AW, Simuni T, Zadikoff C, and Zee PC (2014) Circadian melatonin rhythm and excessive daytime sleepiness in Parkinson disease. *JAMA Neurol* 71:463-469. doi:10.1001/jamaneurol.2013.6239
- 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* 32:3351-3353. doi:10.1093/bioinformatics/btw405
- Wu G, Zhu J, Yu J, Zhou L, Huang JZ, and Zhang Z (2014) Evaluation of five methods for genome-wide circadian gene identification. *J Biol Rhythms* 29:231-242. doi:10.1177/0748730414537788
- Yang R and Su Z (2010) Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics* 26:i168-i174. doi:10.1093/bioinformatics/btq189
- 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 USA* 111:16219-16224. doi:10.1073/pnas.1408886111