Supplementary Materials for Likelihood-based Tests for Detecting Circadian Rhythmicity and Differential Circadian Patterns in Transcriptomic Applications

Haocheng Ding[1], Lingsong Meng[1], Andrew C. Liu[2], Michelle L. Gumz[3], Andrew J. Bryant[4], Colleen A. Mcclung[5], George C. Tseng[6], Karyn A. Esser[2], and Zhiguang Huo*[1]

[1]Department of Biostatistics, University of Florida
[2]Department of Physiology and Functional Genomics, University of Florida
[3]Division of Nephrology, Hypertension, and Renal Transplantation, Department of Medicine, University of Florida
[4]Division of Pulmonary, Critical Care, and Sleep Medicine, Department of Medicine, University of Florida
[5]Translational Neuroscience Program, Department of Psychiatry, Center for Neuroscience, University of Pittsburgh
[6]Department of Biostatistics, University of Pittsburgh

*Corresponding author

S1 Using the Box-Cox transformation to improve normality

S1.1 Simulation setting

In this section, we used simulated data to show when the normality assumption on the residuals (i.e., y_i - $\hat{y}_i$) is violated, we could use the Box-Cox transformation to rescue the normality. The simulation is based on Equation 1 of the main text, and parameters are set as below:

$$y_i = A\sin(\omega(t_i + \phi)) + C + \varepsilon_i, \tag{1}$$

- $n = 50$
- $A = 0$
- $\phi = 0$
- $\omega = 2\pi/24$
- $C = 20$
- $t_i \sim \text{UNIF}(0, 24)$
- $\varepsilon_i \sim \text{t}(3)$

where the error term follows a t-distribution with degree of freedom 3, which represents a severe violation of the normality assumption.

S1.2 Before data transformation, residuals violate the normality assumption

We estimated the parameters in Equation 1, obtained the fitted value $\hat{y}_i$, and calculated the residuals $y_i - \hat{y}_i$. Then we plotted the Q-Q plot of the residuals against the theoretical standard normal distribution (See Figure S1a). A Q-Q plot is a scatter plot comparing the distribution of two datasets. The quantiles of two datasets at the same level were plotted against each other. If the two datasets came from the same underlying distribution, we should see the scatter points roughly forming a straight line. Figure S1a shows that the scatter points deviated from the blue line, indicating a severe violation of the normality assumption. We further used Shapiro-Wilk normality test, which is a statistical test to assess the normality assumption. The resulting p-value is 0.000146, corroborating the violation of the normality assumption.

(a) Before data transformation

(b) After the Box-Cox transformation

Figure S1: Q-Q plot of the residuals against the theoretical standard normal distribution before data transformation. x-axis represents the theoretical quantiles; y-axis represents the sample quantiles. If the scatter points form a a straight line, it would indicate the residual distribution is from a normal distribution.

S1.3 The Box-Cox transformation

To improve normality, we employed a Box-Cox transformation. The Box-Cox transformation is an important technique in statistics such that after the transformation, the non-normally distributed residuals will become normally distributed. To facilitate the Box-Cox transformation, we employed Equation 2, which is under a linear model framework. To be specific, we used the *boxcox* function in R *MASS* package to achieve this transformation.

$$y_i = E\sin(\omega t_i) + F\cos(\omega t_i) + C + \varepsilon_i, \tag{2}$$

S1.4 After data transformation, residuals satisfy the normality assumption

Denote the Box-Cox tranformation data for sample i is z_i. We estimated the parameters in Equation 1, obtained the fitted value $\hat{z}_i$, and calculated the residuals $z_i - \hat{z}_i$. Then we plotted the Q-Q plot of the residuals against the theoretical standard normal distribution (See Figure S1b). It shows that the scatter points fit the blue line very well, indicating no violation of the normality assumption. We further used Shapiro-Wilk normality test. The resulting p-value is 0.6485, corroborating no violation of the normality assumption.

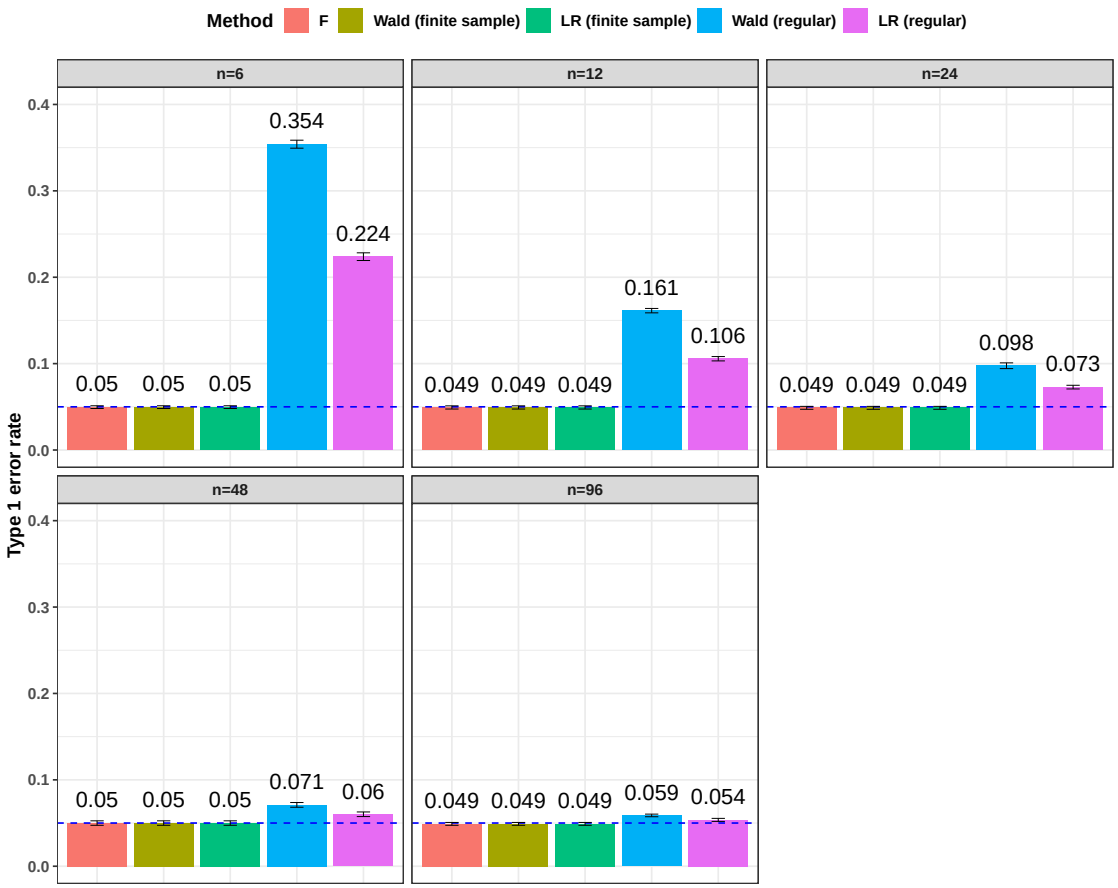


Figure S2: Type I error rate at nominal α level 5% for F test and the four likelihood-based methods in detecting circadian rhythmicity. The sample sizes were varied at n=6, 12, 24, 48, and 96. The blue dashed line is the 5% nominal level. A higher than 5% blue dashed line bar indicates an inflated type I error rate; a lower than 5% blue dashed line bar indicates a smaller than expected type I error rate; and a bar at the blue dashed line indicates an accurate type I error rate (i.e., p-value = 0.05). The standard deviation of the mean type I error rate was also marked on the bar plot.

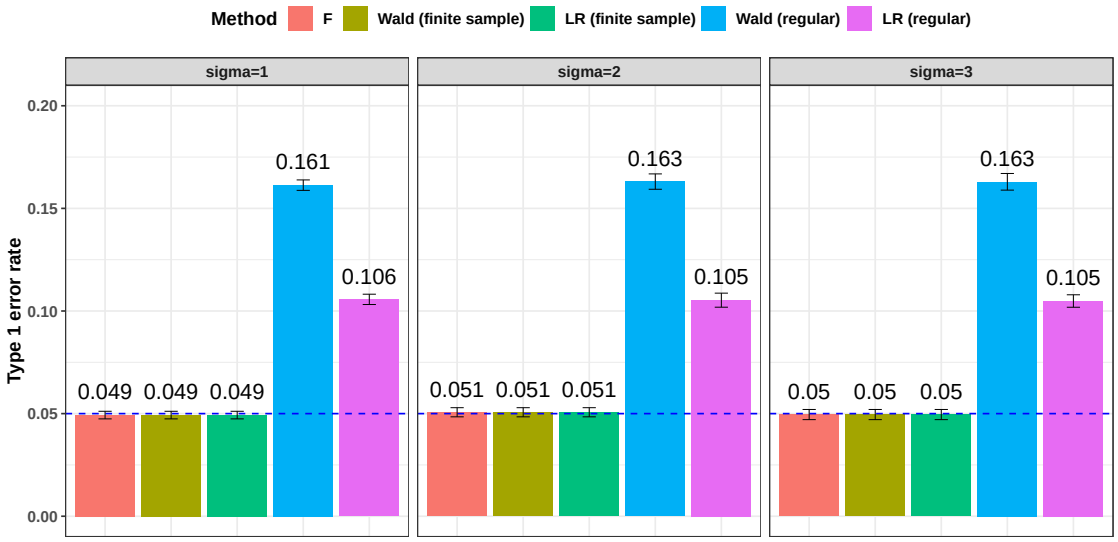


Figure S3: Type I error rate at nominal α level 5% for F test and the four likelihood-based methods in detecting circadian rhythmicity. The noise level were varied at $\sigma = 1, 2, 3$. The blue dashed line is the 5% nominal level. A higher than 5% blue dashed line bar indicates an inflated type I error rate; a lower than 5% blue dashed line bar indicates a smaller than expected type I error rate; and a bar at the blue dashed line indicates an accurate type I error rate (i.e., p-value = 0.05). The standard deviation of the mean type I error rate was also marked on the bar plot.

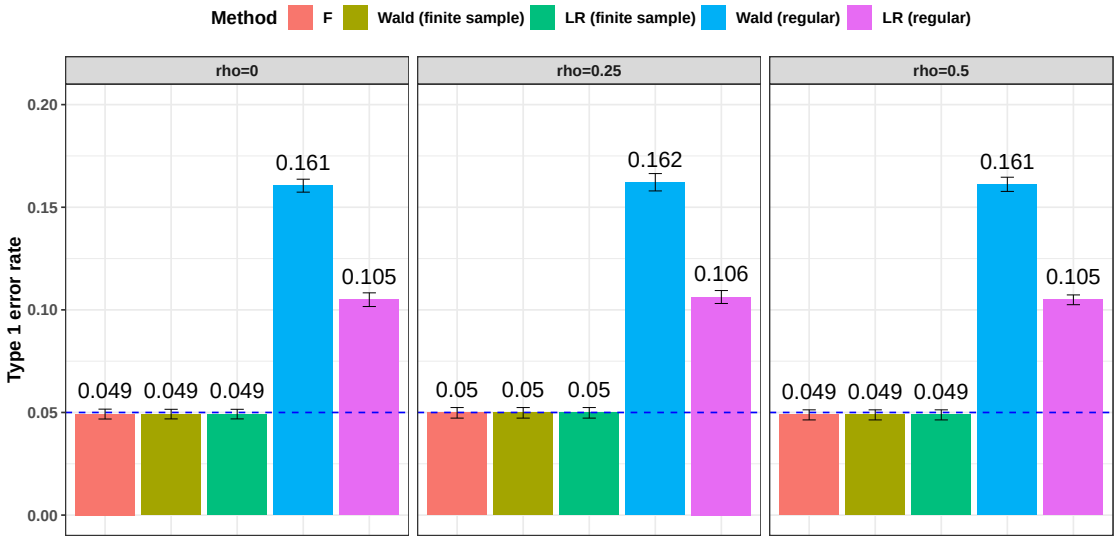


Figure S4: Type I error rate at nominal α level 5% for F test and the four likelihood-based methods in detecting circadian rhythmicity. The correlation strength between genes were varied at $\rho = 0, 0.25, 0.5$. The blue dashed line is the 5% nominal level. A higher than 5% blue dashed line bar indicates an inflated type I error rate; a lower than 5% blue dashed line bar indicates a smaller than expected type I error rate; and a bar at the blue dashed line indicates an accurate type I error rate (i.e., p-value = 0.05). The standard deviation of the mean type I error rate was also marked on the bar plot.

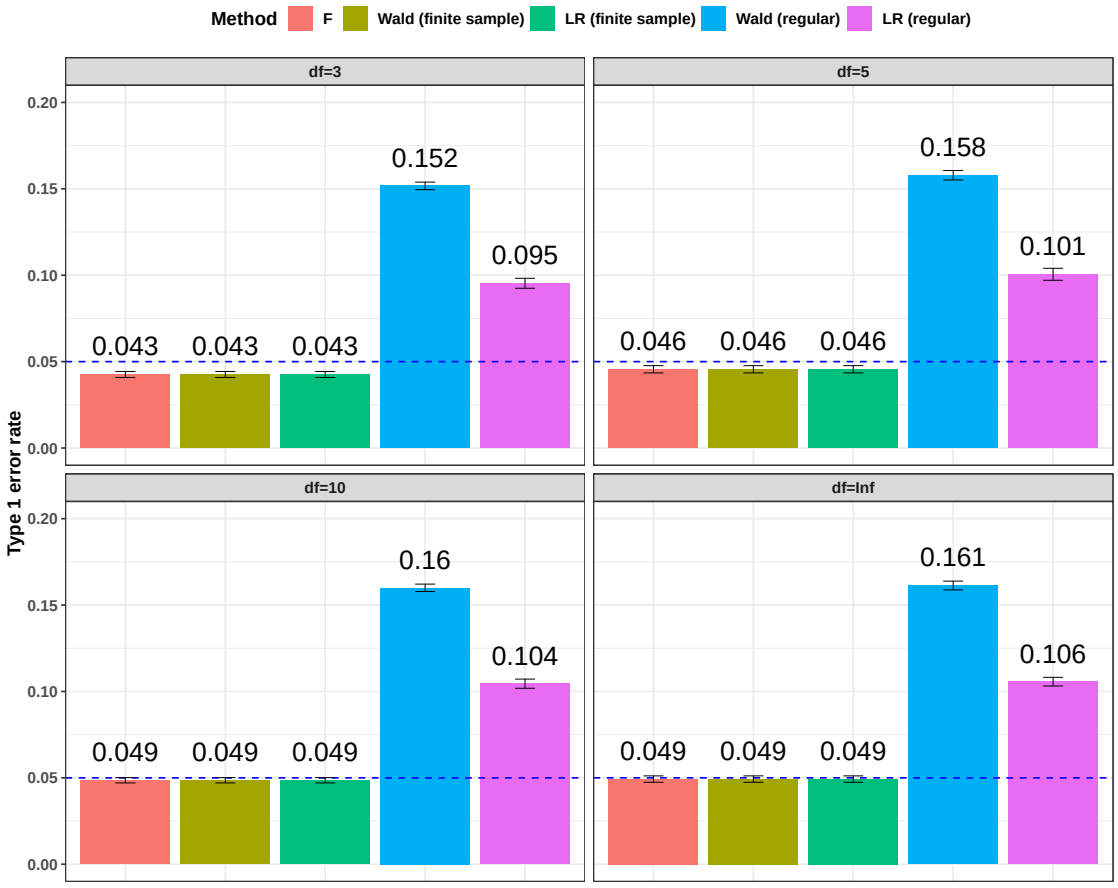


Figure S5: Type I error rate at nominal α level 5% for F test and the four likelihood-based methods in detecting circadian rhythmicity. The violation of the normality assumption *df* were varied at $df = 3, 5, 10, \infty$, where *df* is the degree of freedom of a t-distribution. When $df = \infty$, $t(\infty)$ is equivalent to a standard normal distribution (i.e., $N(0, 1)$). The blue dashed line is the 5% nominal level. A higher than 5% blue dashed line bar indicates an inflated type I error rate; a lower than 5% blue dashed line bar indicates a smaller than expected type I error rate; and a bar at the blue dashed line indicates an accurate type I error rate (i.e., p-value = 0.05). The standard deviation of the mean type I error rate was also marked on the bar plot.

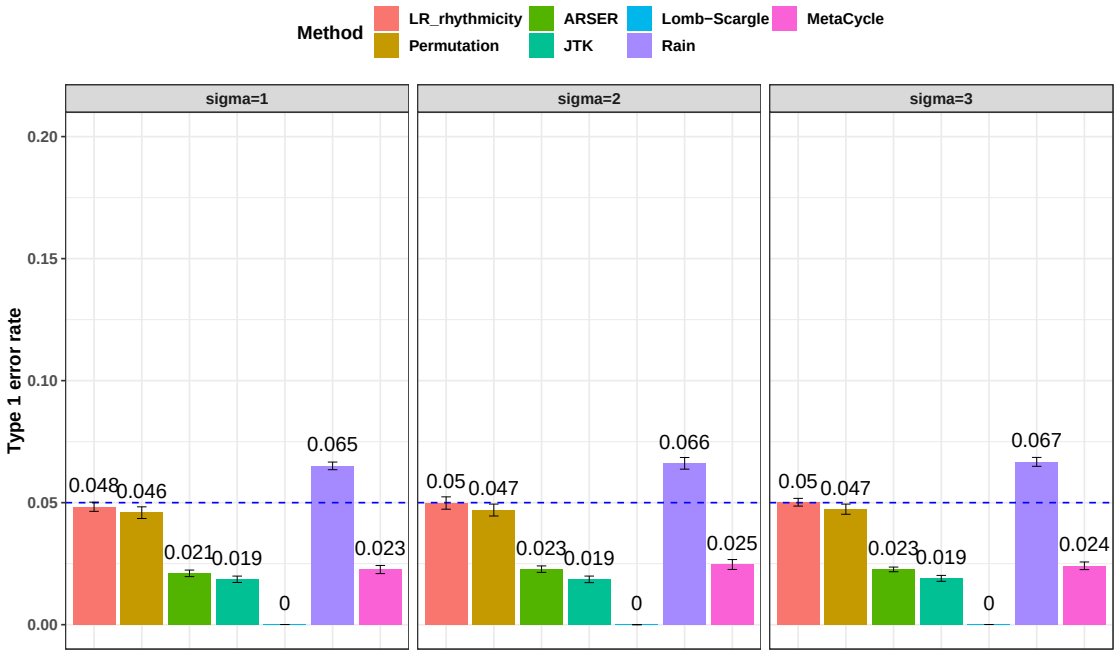


Figure S6: Type I error rate at nominal α level 5% for 7 different methods in detecting circadian rhythmicity. The noise level were varied at $\sigma = 1, 2, 3$. The blue dashed line is the 5% nominal level. A higher than 5% blue dashed line bar indicates an inflated type I error rate; a lower than 5% blue dashed line bar indicates a smaller than expected type I error rate; and a bar at the blue dashed line indicates an accurate type I error rate (i.e., p-value = 0.05). The standard deviation of the mean type I error rate was also marked on the bar plot.

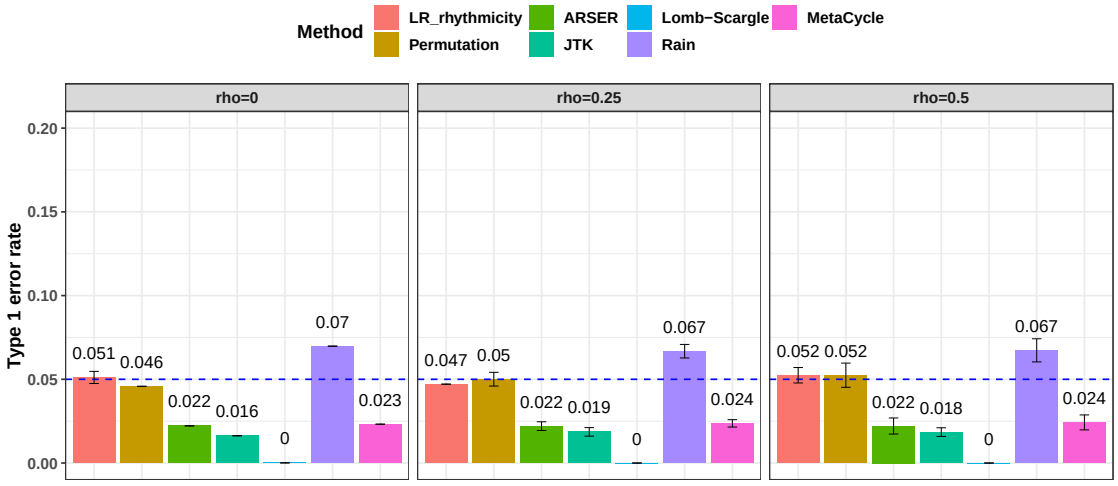


Figure S7: Type I error rate at nominal α level 5% for 7 different methods in detecting circadian rhythmicity. A higher than 5% blue dashed line bar indicates an inflated type I error rate; a lower than 5% blue dashed line bar indicates a smaller than expected type I error rate; and a bar at the blue dashed line indicates an accurate type I error rate (i.e., p-value = 0.05). The correlation strength between genes were varied at $\rho = 0, 0.25, 0.5$. The standard deviation of the mean type I error rate was also marked on the bar plot.

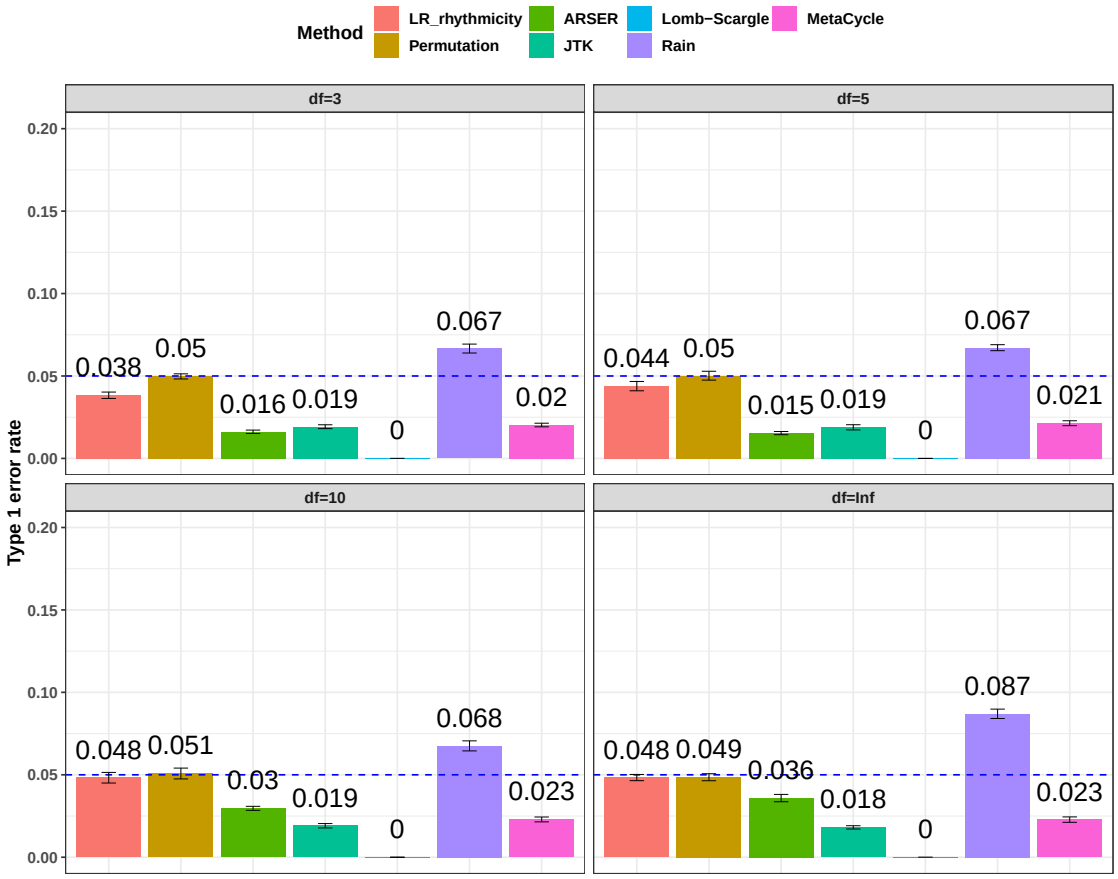


Figure S8: Type I error rate at nominal α level 5% for 7 different methods in detecting circadian rhythmicity. The violation of the normality assumption df were varied at $df = 3, 5, 10, \infty$, where df is the degree of freedom of a t-distribution. When $df = \infty$, $t(\infty)$ is equivalent to a standard normal distribution (i.e., $N(0, 1)$). The blue dashed line is the 5% nominal level. A higher than 5% blue dashed line bar indicates an inflated type I error rate; a lower than 5% blue dashed line bar indicates a smaller than expected type I error rate; and a bar at the blue dashed line indicates an accurate type I error rate (i.e., p-value = 0.05). The standard deviation of the mean type I error rate was also marked on the bar plot.

Figure S9: Power evaluation for LR_rhythmicity and the permutation test. The sample sizes were varied at n=6, 12, 24, 48, and 96. The standard deviation of the mean type I error rate was also marked on the bar plot.

Table S1: Sensitivity analysis by perturbing amplitude A, phase ϕ, basal level C, and noise level σ^2. First column indicates which parameter was varied while other parameters remained fixed (default setting). The default parameter setting was inherited from the main simulation setting for the circadian rhythmicity analysis.

	Type I error	Power
A = 1	0.051	0.804
A = 2	0.051	1.000
A = 3	0.051	1.000
$\phi = 1$	0.051	0.806
$\phi = 2$	0.051	0.806
$\phi = 3$	0.051	0.796
C = 1	0.050	0.807
C = 2	0.050	0.803
C = 3	0.051	0.804
$\sigma^2 = 1$	0.051	0.804
$\sigma^2 = 2$	0.051	0.494
$\sigma^2 = 3$	0.050	0.348

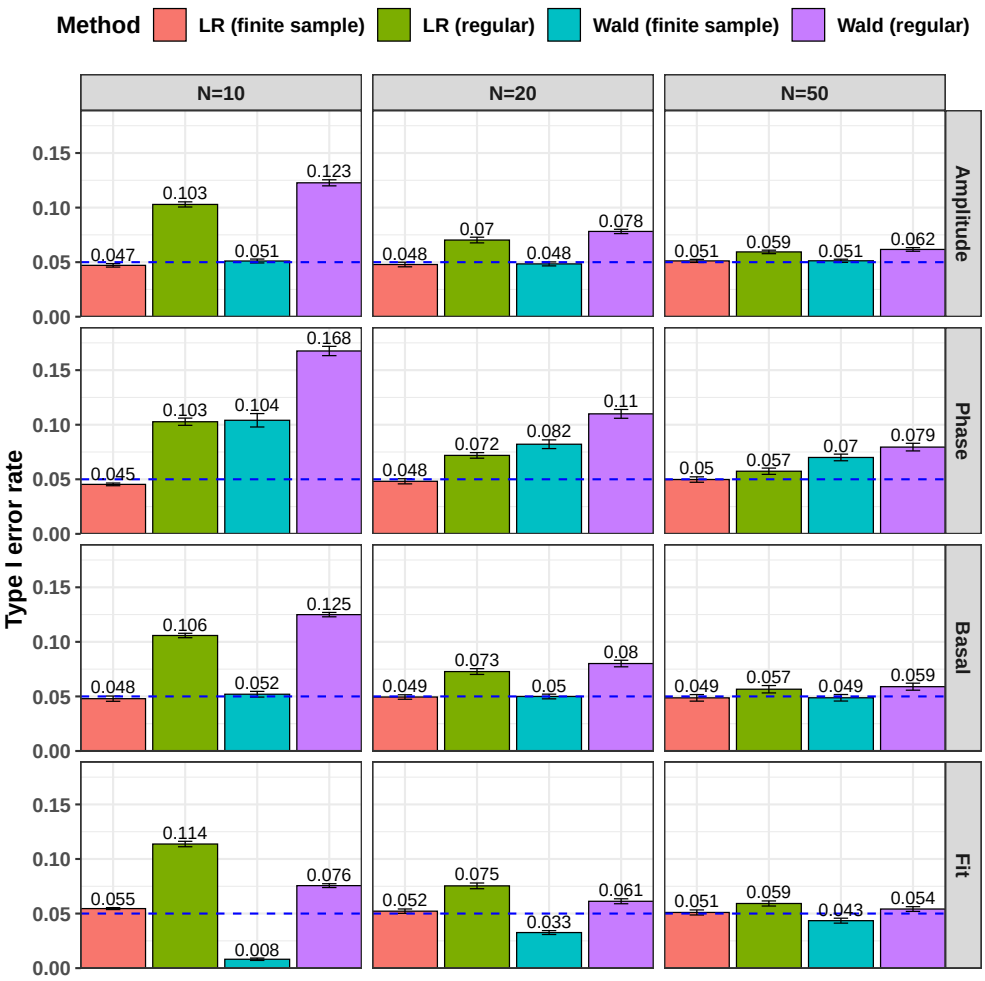


Figure S10: Type I error rate at nominal α level 5% for the 4 likelihood-based methods in detecting differential circadian patterns. The differential circadian patterns include differential amplitude (Amplitude), differential phase (Phase), differential basal level (Basal), and differential fit (Fit). The sample sizes were varied at N=10, 20, and 50. The blue dashed line is the 5% nominal level. A higher than 5% blue dashed line bar indicates an inflated type I error rate; a lower than 5% blue dashed line bar indicates a smaller than expected type I error rate; and a bar at the blue dashed line indicates an accurate type I error rate (i.e., p-value = 0.05). The standard deviation of the mean type I error rate was also marked on the bar plot.

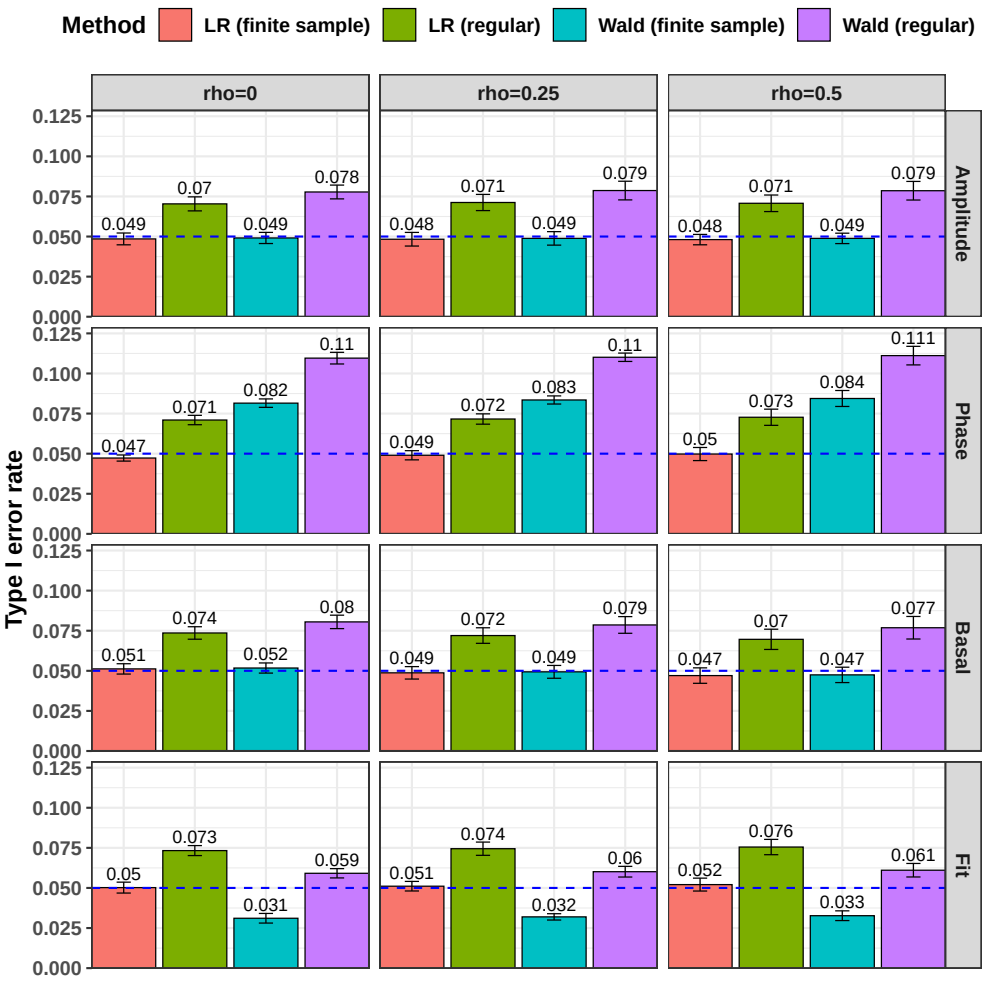


Figure S11: Type I error rate at nominal α level 5% for 6 different methods in detecting differential circadian patterns. The differential circadian patterns include differential amplitude (Amplitude), differential phase (Phase), differential basal level (Basal), and differential fit (Fit). The correlation strength between genes were varied at $\rho = 0, 0.25, 0.5$. The blue dashed line is the 5% nominal level. A higher than 5% blue dashed line bar indicates an inflated type I error rate; a lower than 5% blue dashed line bar indicates a smaller than expected type I error rate; and a bar at the blue dashed line indicates an accurate type I error rate (i.e., p-value = 0.05). The standard deviation of the mean type I error rate was also marked on the bar plot.

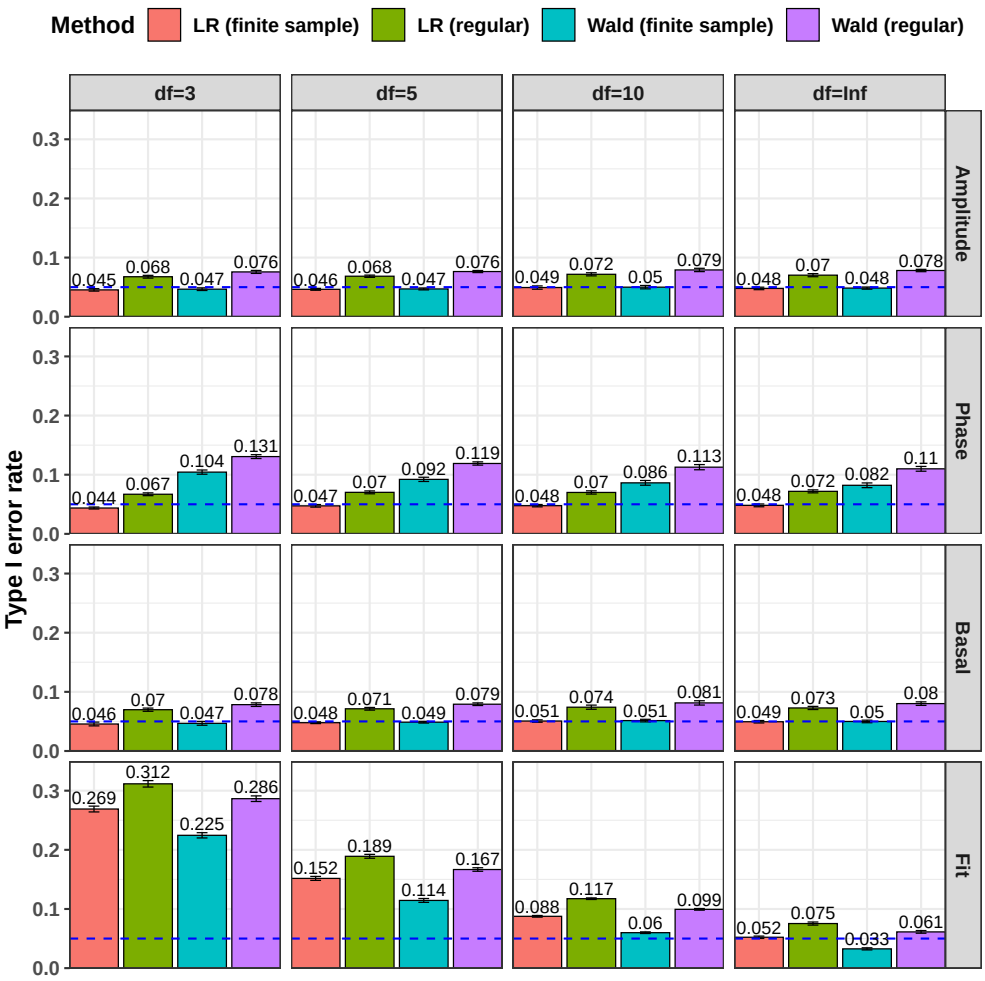


Figure S12: Type I error rate at nominal α level 5% for 6 different methods in detecting differential circadian patterns. The differential circadian patterns include differential amplitude (Amplitude), differential phase (Phase), differential basal level (Basal), and differential fit (Fit). The violation of the normality assumption df were varied at $df = 3, 5, 10, \infty$, where df is the degree of freedom of a t-distribution. When $df = \infty$, $t(\infty)$ is equivalent to a standard normal distribution (i.e., $N(0, 1)$). The blue dashed line is the 5% nominal level. A higher than 5% blue dashed line bar indicates an inflated type I error rate; a lower than 5% blue dashed line bar indicates a smaller than expected type I error rate; and a bar at the blue dashed line indicates an accurate type I error rate (i.e., p-value = 0.05). The standard deviation of the mean type I error rate was also marked on the bar plot.

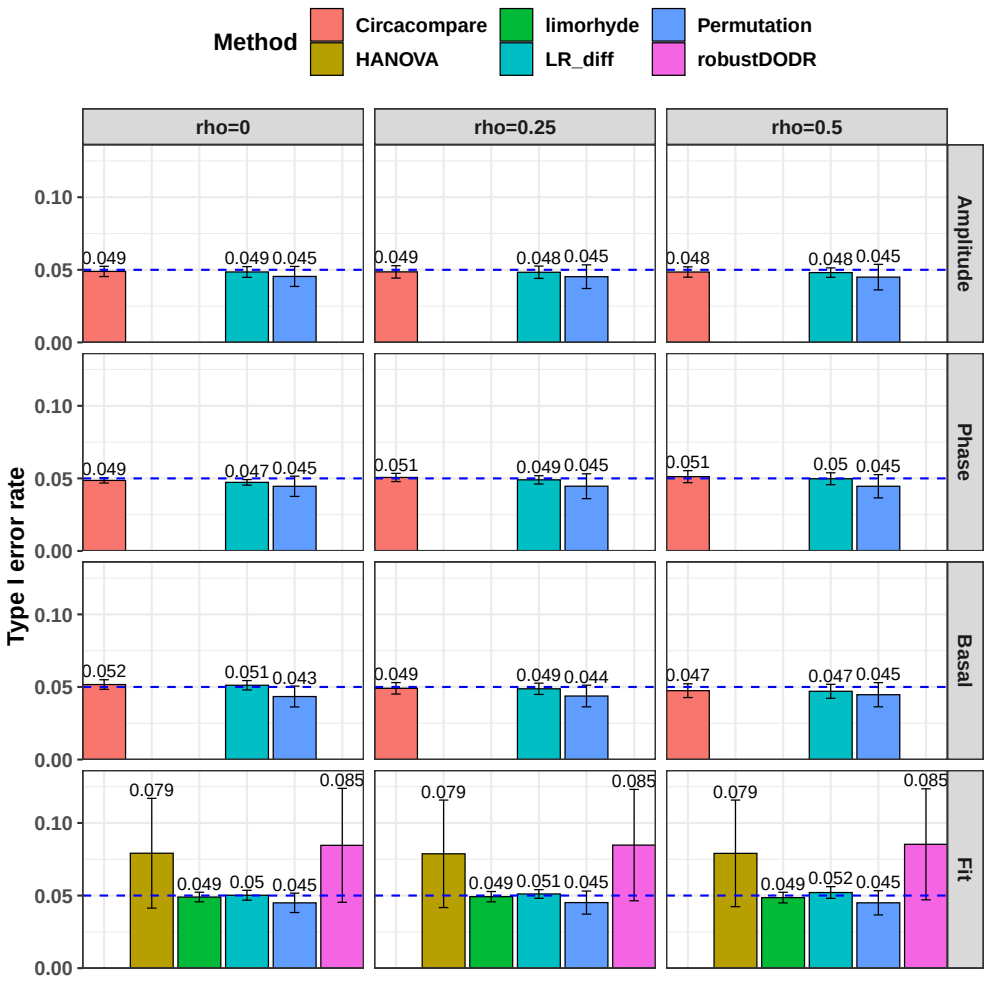


Figure S13: Type I error rate at nominal α level 5% for 6 different methods in detecting differential circadian patterns. The differential circadian patterns include differential amplitude (Amplitude), differential phase (Phase), differential basal level (Basal), and differential fit (Fit). The correlation strength between genes were varied at $\rho = 0, 0.25, 0.5$. The blue dashed line is the 5% nominal level. A higher than 5% blue dashed line bar indicates an inflated type I error rate; a lower than 5% blue dashed line bar indicates a smaller than expected type I error rate; and a bar at the blue dashed line indicates an accurate type I error rate (i.e., p-value = 0.05). The standard deviation of the mean type I error rate was also marked on the bar plot.

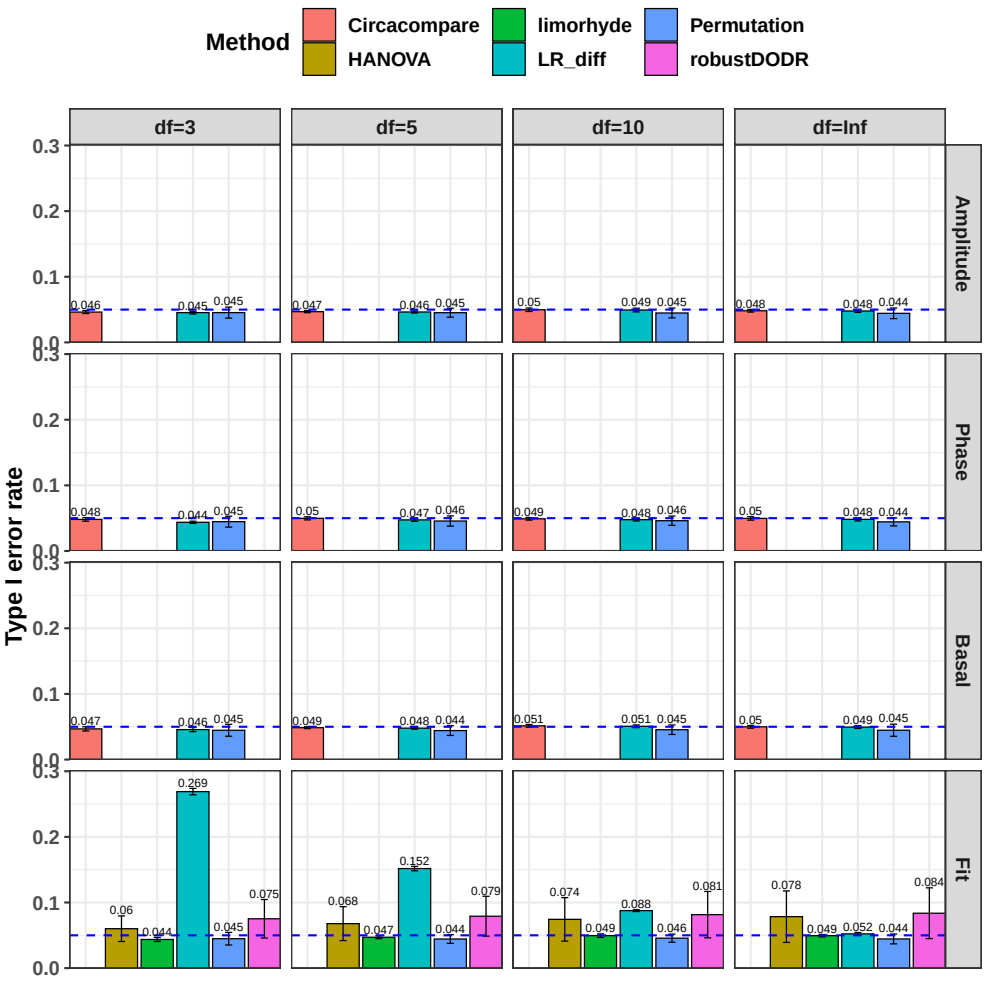


Figure S14: Type I error rate at nominal α level 5% for 6 different methods in detecting differential circadian patterns. The differential circadian patterns include differential amplitude (Amplitude), differential phase (Phase), differential basal level (Basal), and differential fit (Fit). The violation of the normality assumption df were varied at $df = 3, 5, 10, \infty$, where df is the degree of freedom of a t-distribution. When $df = \infty$, $t(\infty)$ is equivalent to a standard normal distribution (i.e., $N(0, 1)$). The blue dashed line is the 5% nominal level. A higher than 5% blue dashed line bar indicates an inflated type I error rate; a lower than 5% blue dashed line bar indicates a smaller than expected type I error rate; and a bar at the blue dashed line indicates an accurate type I error rate (i.e., p-value = 0.05). The standard deviation of the mean type I error rate was also marked on the bar plot.

Table S2: Number of significant genes (p<0.01) on 3 real datasets among 6 methods. – indicates the method is not applicable under the setting.

	Brain Aging Data	Time Restricted Data		Mouse Exercise Data	
		Restricted	Unrestricted	Sedentary	Exercise
LR_rhythmicity	528	1407	935	621	752
ARSER	–	–	–	–	–
JTK	–	–	–	93	108
Lomb-Scargle	26	308	108	491	575
MetaCycle	–	–	–	398	493
Rain	–	421	678	1209	112

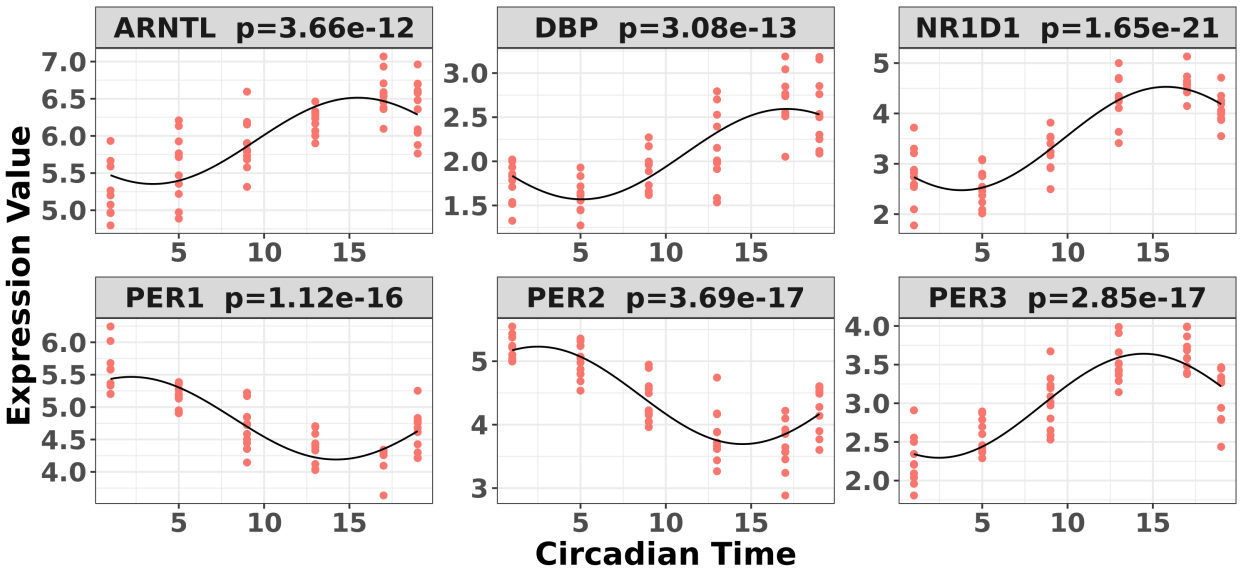


Figure S15: Circadian rhythmicity detected by LR_rhythmicity for 6 core circadian genes in the restricted group of the time-restricted feeding data, including PER1, PER2, PER3, ARNTL, NR1D1, and DBP.

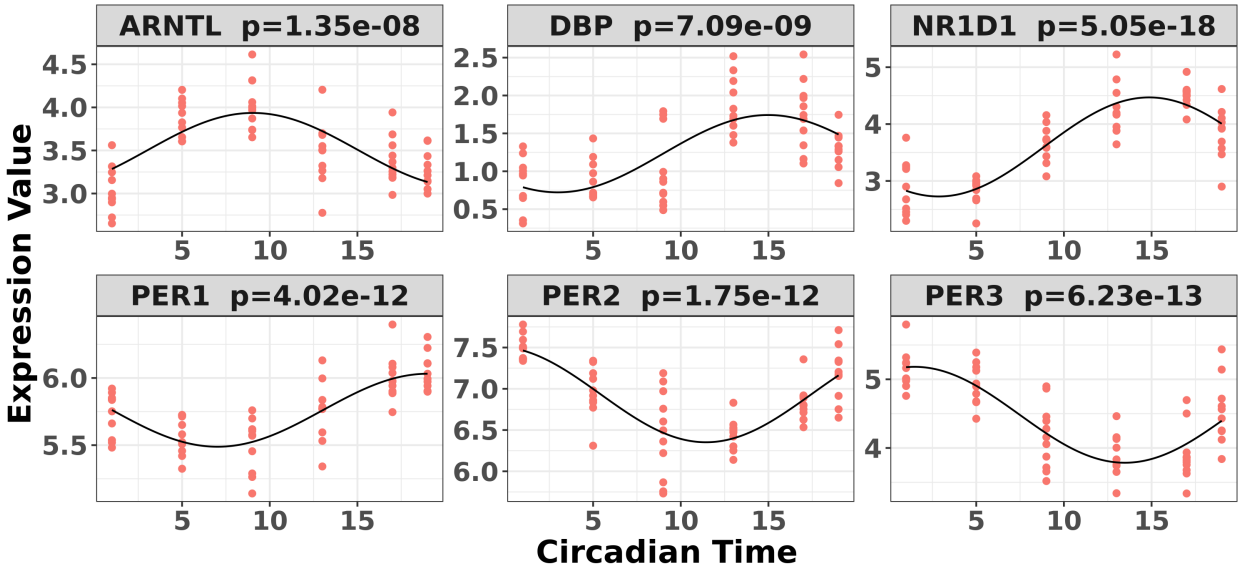
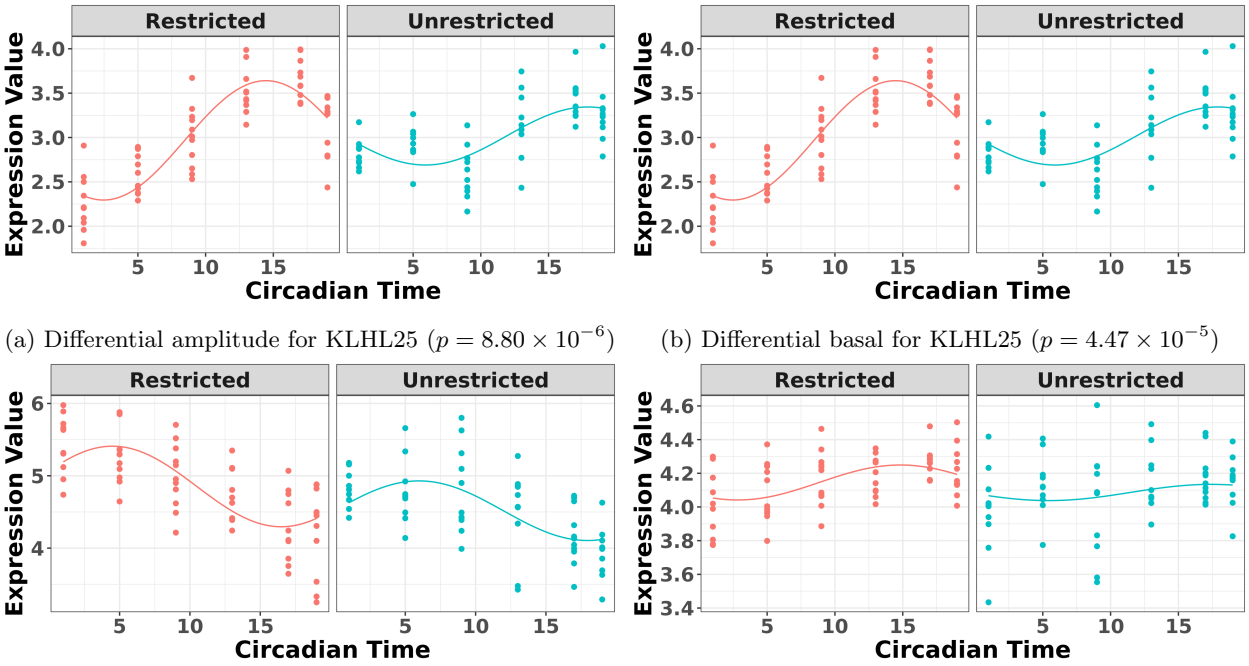


Figure S16: Circadian rhythmicity detected by LR_rhythmicity for 6 core circadian genes in the unrestricted group of the time-restricted feeding data, including PER1, PER2, PER3, ARNTL, NR1D1, and DBP.



(a) Differential amplitude for KLHL25 ($p = 8.80 \times 10^{-6}$)

(b) Differential basal for KLHL25 ($p = 4.47 \times 10^{-5}$)

(c) Differential phase for TSPAN8 ($p = 1.37 \times 10^{-4}$)

(d) Differential fit for RUFY1 ($p = 7.70 \times 10^{-5}$)

Figure S17: The most significant genes showing four types of differential circadian patterns from the time-restricted feeding data.

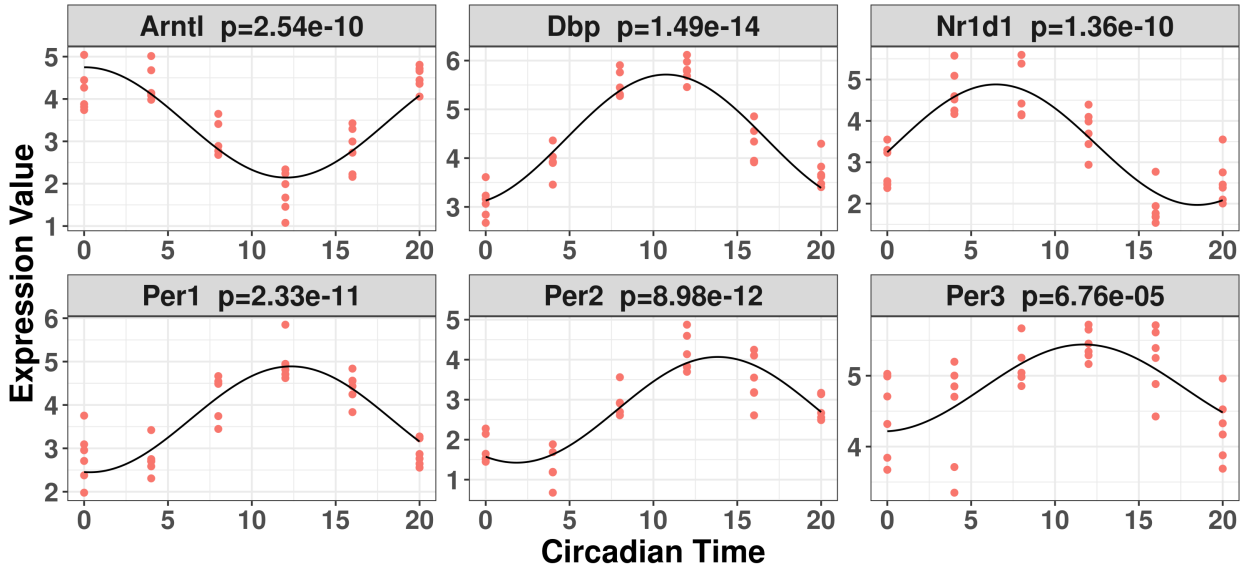


Figure S18: Circadian rhythmicity for 6 core circadian genes in the skeletal muscle data of sedentary mice, including *Per1*, *Per2*, *Per3*, *Arntl*, *Nr1d1*, and *Dbp*, using LR_rhythmicity.

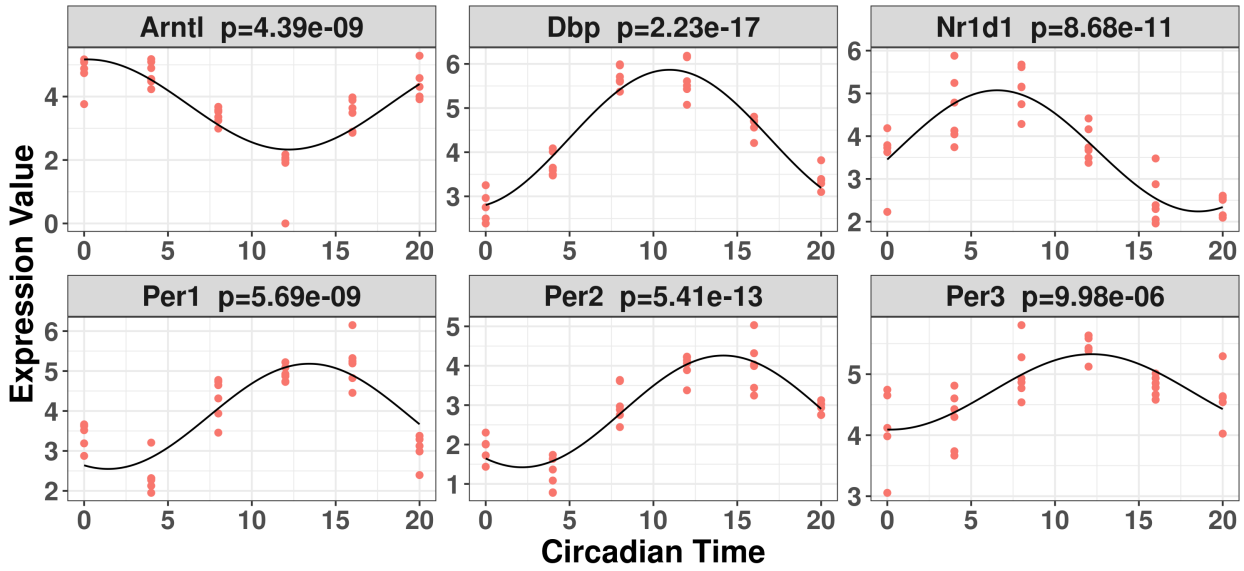
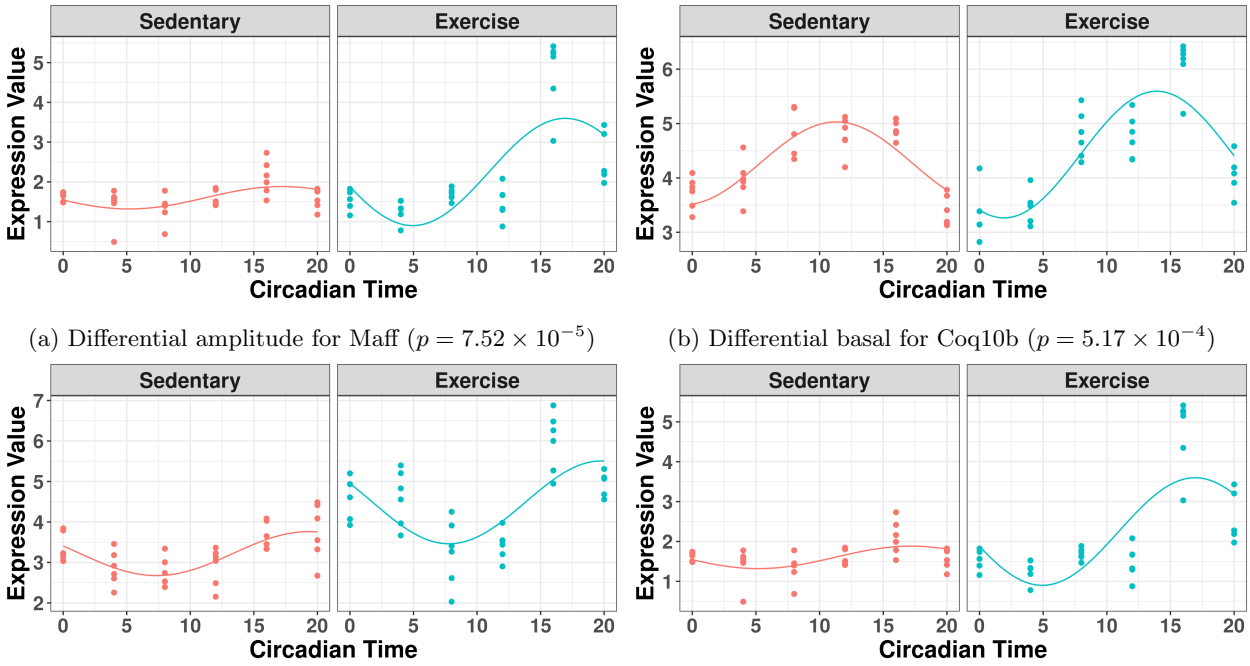


Figure S19: Circadian rhythmicity for 6 core circadian genes in the skeletal muscle data of exercise mice, including *Per1*, *Per2*, *Per3*, *Arntl*, *Nr1d1*, and *Dbp*, using LR_rhythmicity.



(a) Differential amplitude for Maff ($p = 7.52 \times 10^{-5}$)

(b) Differential basal for Coq10b ($p = 5.17 \times 10^{-4}$)

(c) Differential phase for Mt1 ($p = 2.99 \times 10^{-9}$)

(d) Differential fit for Maff ($p = 2.92 \times 10^{-7}$)

Figure S20: The most significant genes showing four types of differential circadian patterns from the mouse skeletal muscle data.

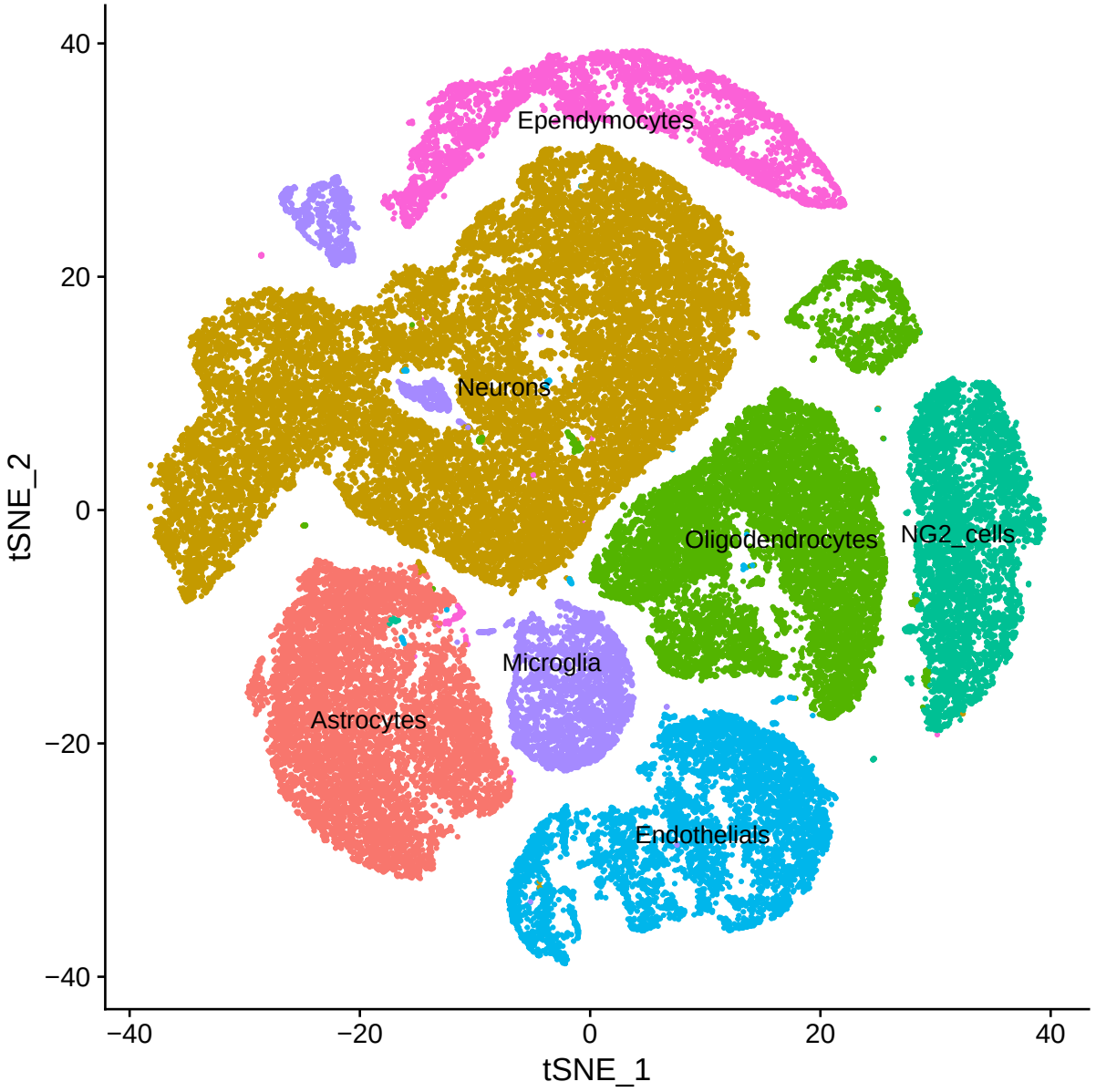
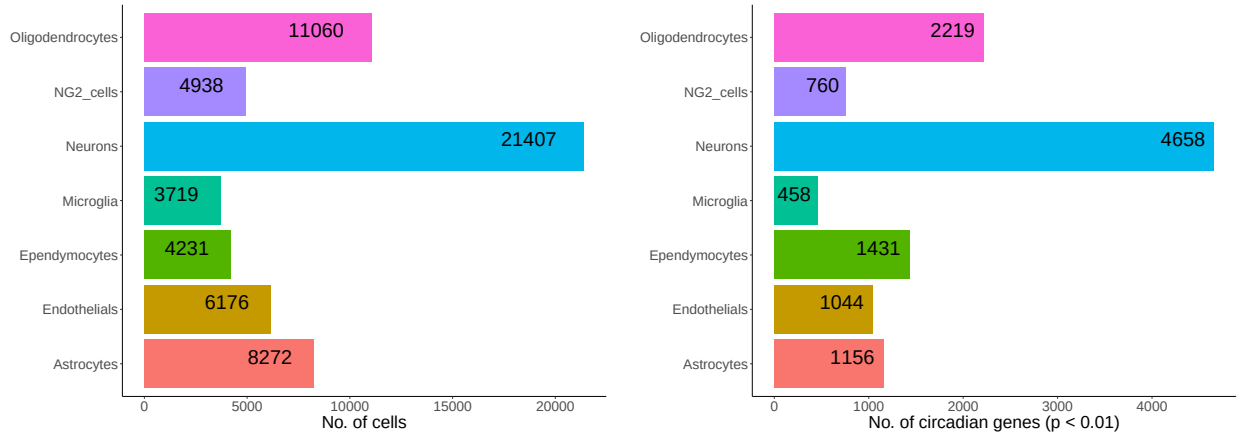


Figure S21: TSNE plot showing 59,803 cells from 7 unique cell types. These 59,803 cells were pooled from all 12 mouse single-cell RNA Sequencing samples. The 7 unique cell types were obtained by merging 18 original identified clusters.



(a) Number of cells for each cell type.

(b) Number of circadian genes ($p < 0.01$) for each cell type.

Figure S22: Number of cells and number of circadian genes of for each cell type.

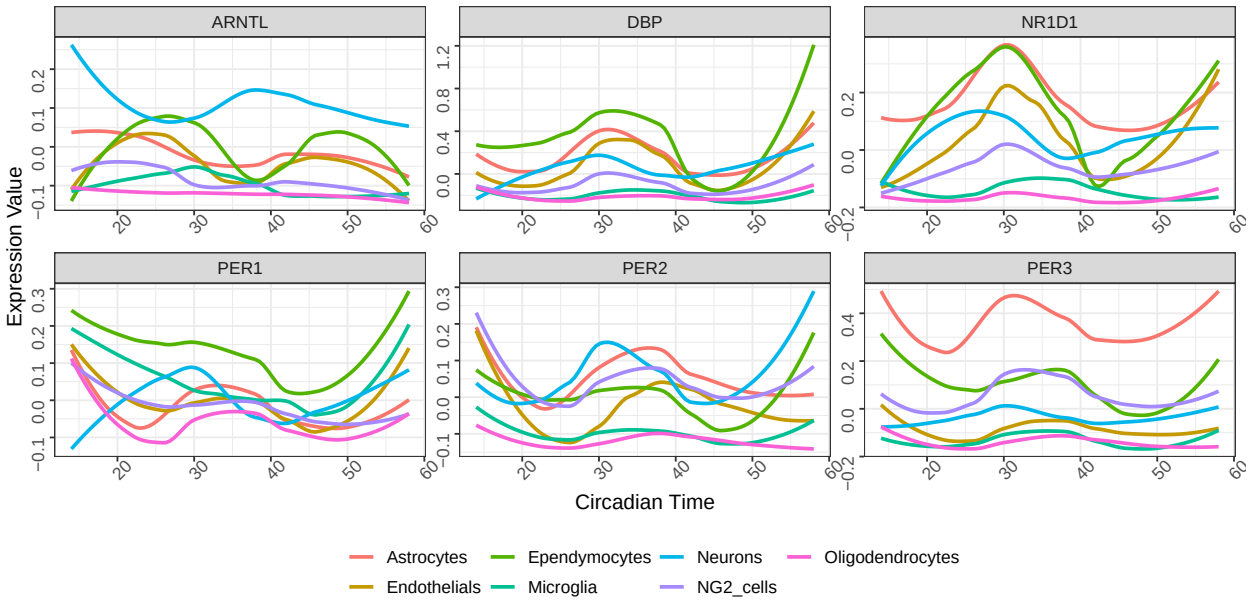


Figure S23: Cell-type specific circadian rhythmicity for 6 core circadian genes in the scRNAseq data, including *Arntl*, *Dbp*, *Nr1d1*, *Per1*, *Per2* and *Per3*.