THE UNIVERSITY OF CHICAGO

IMPROVED STATISTICAL METHODS FOR ANALYZING CIRCADIAN RHYTHMS IN HIGH-THROUGHPUT DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

INTERDISCIPLINARY SCIENTIST TRAINING PROGRAM: BIOPHYSICAL SCIENCES

BY
ALAN L. HUTCHISON

CHICAGO, ILLINOIS
DECEMBER 2016

Copyright © 2016 by Alan L. Hutchison

All Rights Reserved

To my parents Barb and Rick, who prepared me intellectually, socially, emotionally, financially, and educationally for the rigor of thought and work needed to earn a PhD.

Make my funk the P-Funk

I want my funk uncut

Make my funk the P-Funk

I wants to get funked up.

-George Clinton

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES x

ACKNOWLEDGMENTS xi

ABSTRACT xii

1 INTRODUCTION 1
1.1 General Circadian Biology 1
1.2 History of Rhythm Detection and Observation 1
1.3 Three Rhythm Detection Problems 3
1.4 Types of Recent Experiments 4
1.5 Types of Recent Rhythm Detection Methods 5
1.6 Differential Rhythmicity 7

2 RHYTHM DETECTION: EMPIRICAL JTK_CYCLE 8
2.1 Abstract 8
2.2 Author Summary 9
2.3 Introduction 9
2.4 Methods 11
2.4.1 Overview 11
2.4.2 Improvements to the JTK_CYCLE Method 16
2.5 Results 21
2.5.1 Simulated Data Benchmarks 21
2.5.2 Microarray Metadataset 31
2.6 Discussion 43
2.7 Supplementary Figures 46

3 CORRECTING FOR DEPENDENT P-VALUES IMPROVES ACCURACY OF LEADING RHYTHM DETECTION METHODS 65
3.1 Abstract 65
3.2 Introduction 65
3.3 Methods 66
3.3.1 Empirically calculating p-values from simulated data 66
3.4 Results 68
3.4.1 Dependence of p-values from comparing reference waveforms with experimental time series 68
3.4.2 Dependence of p-values from comparing different rhythm detection results for the same time series 69
3.5 Discussion 75

4 BOOTSTRAPPING AND EMPIRICAL BAYES METHODS IMPROVE RHYTHM DETECTION IN SPARSELY SAMPLED DATA . . . 78
4.1 Abstract . . . 78
4.2 Introduction . . . 78
4.3 Methods . . . 81
4.3.1 Empirical Bayes variance estimation . . . 81
4.3.2 Bootstrapping eJTK . . . 82
4.3.3 Obtaining Accurate and Computationally Inexpensive P-values . . . 83
4.3.4 BooteJTK Outperforms Alternative Rhythm Detection Methods . . . 83
4.3.5 Computational Expense . . . 86
4.4 Results . . . 88
4.4.1 Effect of sampling frequency on rhythm detection . . . 88
4.4.2 BooteJTK Reveals More Biologically Consistent Circadian Rhythmic Gene Expression Across 12 Mouse Tissues . . . 90
4.5 Discussion . . . 96
4.6 Supplementary Figures . . . 100

5 DIFFERENTIAL RHYTHMICITY . . . 111
5.1 Abstract . . . 111
5.2 Introduction . . . 112
5.3 Methods . . . 115
5.3.1 Naïve Comparison Approach . . . 115
5.3.2 Differential Rhythmicity Method . . . 116
5.3.3 Comparison of Naïve and Differential Rhythmicity Methods . . . 117
5.4 Results . . . 120
5.4.1 Comparison of Light-Dark and Dark-Dark Circadian RNA Expression with BDR and DODR . . . 120
5.4.2 Comparison of Circadian Protein and RNA Time Series Rhythmicity 124
5.5 Discussion . . . 129
5.6 Supplementary Figures . . . 134

6 CONCLUSION . . . 143

REFERENCES . . . 149

LIST OF FIGURES

2.1 JTK_CYCLE description 15
2.2 Calculating empirical p-values generates uniformly-distributed p-values when tested against randomly-generated data. 17
2.3 Simulated data line shapes 22
2.4 AUROC results for simulated data with 50% noise. 24
2.5 Higher numbers of replicates provide greater sensitivity compared to increased density of time points for the same number of samples. 27
2.6 Empirical JTK_CYCLE outperforms the other methods in the presence and absence of asymmetric time series. 30
2.7 Pdp1 gene expression from metadata 32
2.8 Empirical JTK_CYCLE with asymmetry search of 4 h (eJTK_aby4) identifies more genes than ANOVA, F24, and the other JTK_CYCLE methods. 34
2.9 Ontology groups identified by different methods 41
2.10 Phase distribution to eJTK_aby4 ontologies 42
2.S1 Gamma distribution accurately models the F24 null distribution. 46
2.S2 Trough and Cosine are highly correlated waveforms 47
2.S3 AUROCs for simulated data with 25% noise. 48
2.S4 Full set of comparisons used to evaluate the trade-off between increased numbers of replicates and increased densities of time points per period. 49
2.S5 Interpolation scheme for doubling replicate counts 50
2.S6 Interpolating the data points to generate pseudo-replicates improves AUROCs when the number of actual replicates is low. 51
2.S7 MCC results for simulated data show that JTK_CYCLE methods outperform ANOVA and F24 in the presence and absence of asymmetric time series. . . . 52
2.S8 Searching for asymmetric waveforms is detrimental if none are present, but is otherwise advantageous. 53
2.S9 Metadata results for known positive and negative examples 54
2.S10 Comparison of the p-value distributions of the original JTK_CYCLE method (with Bonferroni correction) with the empirical JTK_CYCLE method without (A) and with (B) asymmetry search. 55
2.S11 Comparison of the intersection and union of genes identified as rhythmic with Benjamini-Hochberg adjusted p-values less than 0.05 (blue bars) and 0.20 (red bars) for empirical JTK_CYCLE with different asymmetry searches. 55
2.S12 Metadata asymmetry aby4 vs. a08-16 comparison 56
2.S13 Using a cosine as a reference waveform instead of a triangle does not produce substantially different results in genes identified as cycling. 57
2.S14 Using a cosine as a reference waveform instead of a triangle does not produce substantially different results in annotation terms enriched for in genes identified as cycling. 58
2.S15 eJTK vs. Keegan *et al.* 59
2.S16 Comparison of genes identified as cycling by Keegan *et al.*, Wijnen *et al.*, and empirical JTK_CYCLE with asymmetry search by 4 h (eJTK_by4). 60

2.S17 Z-score expression time series of *cbt*, *tws*, *Est-Q*, and *ABGE* averaged across replicate time points. . . . 61
2.S18 KEGG pathway "dme00480: Glutathione metabolism" is enriched in genes identified as rhythmic by eJTK_aby4. . . . 62
2.S19 Gene ontology "GO:0055114 oxidation reduction" is enriched in genes identified as rhythmic by eJTK_aby4. . . . 63
2.S20 PIR keyword "alternative splicing" is enriched in genes identified as rhythmic by eJTK_aby4. . . . 64

3.1 RAIN does not produce p-values that have a uniform distribution under the null hypothesis. . . . 70
3.2 The p-values from the Brown correction of Fisher integration recapitulate the uniform distribution under the null regardless of the dependency of the p-values. 72
3.3 While uncorrected MetaCycle p-values are not uniform under the null hypothesis, the corrected method does produce uniformly distributed p-values under the null hypothesis. . . . 74
3.4 The Brown method for p-value integration provides more accurate results for rhythm detection method combination than the Fisher method. . . . 76

4.1 The BooteJTK $\tilde{\tau}$ null distribution can be modeled by a Gamma distribution. . 84
4.2 BooteJTK outperforms alternative methods on symmetric and asymmetric simulated data. . . . 87
4.3 BooteJTK identifies rhythmic genes more consistently than eJTK as data are downsampled. . . . 90
4.4 BooteJTK provides more consistent rhythm detection than eJTK between downsampled datasets. We quantified the overlap between results with different levels of downsampling by the probability that a probe is rhythmic in one dataset (a row) if it is rhythmic in another (a column). As no probes are found to be rhythmic when using eJTK on data downsampled to every 4 h, rows and columns for that conditions are not shown. . . . 91
4.5 BooteJTK reveals fewer genes with circadian rhythmic expression across 12 mouse tissues than eJTK for the Zhang *et al.* dataset. . . . 94
4.6 BooteJTK reveals fewer genes with circadian rhythmic expression across 12 mouse tissues than eJTK for the Zhang *et al.* dataset. . . . 95
4.S1 Empirical Bayes variance shrinkage of standard deviations . . . 101
4.S2 The eJTK $\tilde{\tau}$ null distribution can be modeled by a Gamma distribution. . . . 102
4.S3 BooteJTK outperforms alternative methods across noise levels and different peak-to-trough times for simulated data. . . . 103
4.S4 Examples of time series that have the same eJTK score but different BooteJTK scores . . . 104
4.S5 The BooteJTK $\tilde{\tau}$ scores for the same time series dataset show no substantial difference for 10, 25, 50, or 100 bootstrap samples. . . . 105
4.S6 BooteJTK identifies rhythmic genes more consistently than eJTK and RAIN as data are downsampled. . . . 106

4.S7 For the Hughes *et al.* dataset, downsampled data analyzed with BooteJTK have greater overlap in genes identified as rhythmic between samples as when analyzed with eJTK or RAIN. . . . 107
4.S8 No net difference exists between fraction of core clock genes rhythmic across tissues analyzed with BooteJTK as compared to eJTK. . . . 108
4.S9 Number of genes rhythmic in each tissue out of the 78 probes rhythmic in 9 or more tissues . . . 109
4.S10 The standard deviation of arrhythmic time series provides an approximation of the standard deviation of time points. . . . 110

5.1 Comparison of BDR methods against naïve methods on simulated data . . . 119
5.2 Comparison of BDR methods against naïve methods on comparison of Hughes *et al.* and Jouffe *et al.* data . . . 121
5.3 Rhythmicity measures from the Jouffe and Hughes datasets plotted against each other with the colors indicating differential rhythmicity measures . . . 125
5.4 Comparison of protein and RNA rhythmicity in Mauvoisin *et al.*, Robles *et al.*, and Guerreiro *et al.* datasets . . . 131
5.5 378 genes had similar relationships between RNA and protein behavior in Mauvoisin *et al.* and Robles *et al.* . . . 132
5.S1 Variance of time series rhythmicity can be predicted from simulated data by comparing ANOVA p-values . . . 135
5.S2 Rhythmicity and phase Z-scores are normally distributed and p-value distributions are uniformly distributed. . . . 136
5.S3 Bootstrap Differential Rhythmicity differential rhythmicity and differential phase p-values outperforms BooteJTK p-values, Phase Difference measurements, and T-statistic p-values in classification ability . . . 137
5.S4 Comparison of RNA expression time series noisiness from Jouffe *et al.* and Hughes *et al.* datasets using ANOVA finds equivalent levels of noisiness in gene-by-gene graph as well as overall distribution . . . 138
5.S5 Examples of genes with RNA expression time series from Jouffe *et al.* and Hughes *et al.* which are found to have converging or diverging differential rhythmicity results by BDR and DODR . . . 139
5.S6 Phase shift and standard deviation distributions for subsets of probes from Jouffe *et al.* and Hughes *et al.* comparison . . . 140
5.S7 Comparison of RNA expression noisiness to protein level noisiness as measured by ANOVA at a gene-by-gene comparison and a distribution level comparison 141
5.S8 Comparison of phases for genes whose RNA expression and protein levels are found not to be differentially rhythmic by BDR for Robles *et al.* and Mauvoisin *et al.* . . . 142

LIST OF TABLES

4.1 Select functional annotations for the 119 genes identified as rhythmic in 9 or more tissues by BooteJTK with a Benjamini-Hochberg adjusted p-value threshold of 0.05 from the Zhang *et al.* dataset analyzed the with DAVID webtool. . 97

5.1 Comparison of Jouffe *et al.* dataset and Hughes *et al.* dataset with BDR and PVT. 122

5.2 Comparison of BDR and DODR on Jouffe *et al.* and Hughes *et al.* data 123

5.3 Results for differential rhythmicity and phase analysis for Robles *et al.*, Mauvoisin *et al.*, and Guerreiro *et al.* 129

5.4 Genes overlapping in differential behavior between Mauvoisin *et al.* and Robles *et al.* 130

ACKNOWLEDGMENTS

I would like, first and foremost, to thank my wonderful wife, Kate Carter, whose unconditional support and understanding as I worked during evenings, weekends, road trips, and vacations was vital to my well-being and success.

I would like to thank my lab-mates in the Dinner Group, whose tolerance for my ideas, questions, difficulties in understanding, desire to decorate Aaron's door, and thirst for the Pub made graduate school much more enjoyable than it might have otherwise been.

I would like to thank my fellow students in the Biophysical Sciences and Medical Scientist Training Program, especially within my own cohort, for their support through classes, qualifying exams, practice talks, and parties.

I would like to thank the faculty and administration of the Biophysical Sciences and Medical Scientist Training Program for always having their doors open to me to discuss my plans, concerns, ideas and views of the program, and for being as flexible as they were in allowing me to craft my own program of study and research.

Finally, I would like to thank my PI, Aaron Dinner, whom I came to appreciate more and more as my time in graduate school progressed. I am incredibly thankful that he allowed me to pursue as many research ideas as I did, giving me time to explore them and assess their worth for further development. I am thankful that he involved himself as deeply as he did in guiding and challenging my ideas on my research, equipping me with set of lenses with which to view my own and others' research that I hope to be able to use for years to come. I am thankful that Aaron was interested not only in the quality of our publications and talks, but interested in our development as scientists, researchers, and thinkers, and valued our thoughts and opinions and frankly discussing them. I am also thankful that now it will never escape my attention when I use 'less' when I should use 'fewer'.

ABSTRACT

Biological rhythms are recognized as critical aspects of cellular, physiological, organismal, and ecological function. Circadian rhythms are 24 hour (h) physiological rhythms driven by photoperiod and temperature yet maintained in their absence, and occur throughout all kingdoms of life. Disruptions in circadian rhythms are both causes and consequences of many neurological and metabolic diseases, making their study relevant from both a basic science and a biomedical perspective. In the past two decades, high-throughput molecular methods have been applied to probe circadian regulation and behavior of RNA and protein expression levels throughout various organisms, tissues, and conditions. The high-dimensional data that result are sparse, noisy, and contain high levels of null results. Adequately and efficiently detecting rhythms in these data and identifying differences across datasets are ongoing challenges for which many statistical methods have been developed. I introduce three new methods for analyzing these data. In Chapter 2, I improve upon a popular rhythm detection method by more accurately calculating its p-values, which allows its further improvement to increase its sensitivity to asymmetric rhythmic time series; my new method is called empirical JTK_CYCLE (eJTK). In Chapter 3, I discuss incorrect assumptions regarding the independence of p-values in two leading rhythm detection methods, RAIN and MetaCycle, and suggest approaches to correct and improve those methods. In Chapter 4, I combine eJTK with empirical Bayes-based bootstrap replicates of experimental time series, adding sensitivity for time series with low relative uncertainty in their time point measurements; my new method is called Bootstrap eJTK (BooteJTK). I also enhance the efficiency of calculating accurate p-values for the method and identify two areas in p-value calculation in other methods in the field. In Chapter 5, I build upon BooteJTK to create a method that rigorously compares two time series from different conditions to determine if they have significantly different rhythmicity and phase. My methods allow for greater sensitivity of rhythmic time series detection while simultaneously providing improved rigor with regard to

noisy time series and identification of differences across datasets.

CHAPTER 1

INTRODUCTION

1.1 General Circadian Biology

Due to the rotation of the Earth and the 24 hour (h) photoperiod that results, many species have evolved physiologies to help anticipate this rhythmicity. Dis-regulation of rhythmicity, due to jet lag and shift work, for example, has been implicated in several types of neurologic and metabolic diseases [98], making these rhythms of interest from both a basic science and a biomedical perspective. There are four characteristics considered necessary for these *circadian* rhythms: 1) endogenous internal rhythms with a period 2) near 24 h that are 3) entrainable by stimuli such as light and temperature but whose period is 4) only weakly sensitive to temperature (temperature compensated) [6].

1.2 History of Rhythm Detection and Observation

The first observation of endogenous circadian behavior was made in 1729 by de Mairan, who saw a plant of the *Mimosa* genus changing its leaf position in constant darkness after having been exposed to light-dark cycles [39]. Molecular biology entered the field of circadian biology with the discovery of the *period* gene in the fruit fly *Drosophila melanogaster* by Konopka and Benzer in 1971 [62] and the successful cloning of that gene in 1984 by Rosbash and Hall [119]. The following decades led to the discovery of roughly twenty more genes in *D. melanogaster*, *Mus musculus*, and other model organisms, that comprise the core clock genetic circuit. Many of these genes were discovered to be transcription factors, positively and negatively regulating their own transcription and the transcription of the other clock genes. With the advent of high-throughput technologies that could quantify the expression levels of thousands of RNAs at a time, exploration began into the downstream effects of the core clock involved in other physiological processes [98]. The development of next-generation

sequencing led to a large number of technologies using high-throughput sequencing of DNA and RNA to enable the investigation of circadian rhythms on DNA binding by transcription factors [61], methylation levels [64], and on the gut microbiome [63]. Differences were soon observed in the mechanisms and processes involved in circadian rhythms across tissues and conditions [4, 83, 95, 121, 84].

As this experimental progression from behavioral organismal observation to molecular assays has occurred, methods for observing, detecting, and classifying these rhythms have concurrently evolved to match the new statistical needs and realities generated by these experiments. Early measurements monitoring the locomotion of flies or mice, which follows a circadian pattern, could be continuously conducted and were typically performed over 10 days, with results being binned every 6 minutes, providing 240 measurements per circadian period and 2400 time points total [88]. These studies were generally performed on tens to hundreds of individuals [119, 107]. For these data, which have high numbers of accurate measurements per period over multiple periods and where rhythmicity is being detected across tens or hundreds of animals, methods such as Enright's periodogram, Fourier transforms, and cosine-fitting methods are common and perform well [88]. In the first microarrays examining gene expression of multiple genes, measurements were taken every 4 h over 2 periods, with each period treated as a replicate of the other, for 6 measurements per circadian period and 12 measurements overall, though thousands of genes were now able to be assayed [45, 75, 4, 83]. The introduction of these high-throughput time series to circadian biology has increased the rate of discovery and our understanding of the role of circadian rhythms in different conditions and tissues at the gene-specific level. The data generated by these gene-specific methods, however, present new and unique problems from the data gathered on the behavior of individual organisms.

1.3 Three Rhythm Detection Problems

High-throughput circadian experiments, such as microarray or RNA-Seq time series [7, 108], have a set of complicating factors that necessitate the use of careful statistical methods to identify rhythms.

The first limitation of circadian experimental time series is the low sampling rate. Data for circadian rhythms are usually collected across 24 h, and the sampling rates vary from experiment to experiment. High-throughput measurements are labor-intensive and financially expensive, which limits sampling rates. The most common experimental design is to collect data every 4 h [45, 75, 4, 83], with 6 h a less common sampling rate [63], and sampling every 2 h and 3 h more recently possible as costs of microarrays and RNA-Seq have decreased [57, 73, 91, 121]. For experiments involving living human subjects, often using blood or plasma samples, ethical considerations and standards also limit sampling [46, 76, 8]. Overall this means a circadian experiment can have anywhere from 4 to 12 time points in a period, with few exceptions [51, 46].

In addition to having a low sampling rate, circadian studies often only obtain two or three replicates of a given time point due to cost limitations. These replicates are either collected at the same time points or collected over several periods and then treated as replicates modulo 24 h [83, 121]. In model organism studies, data collection is destructive in nature as well; it is necessary to sacrifice one or several animals per time point to collect data, and therefore replicates are always from different animals. The process of quantifying transcript or gene levels from microarrays or RNA-Seq is also non-trivial; many types of methods have been developed to accurately obtain these values [115, 101, 90]. The biological variance in the replicates and the low number of replicates together result in noisy time series, necessitating advanced statistical methods to extract signal from noise.

The third hurdle to detecting rhythmicity in high-throughput data is the problem of multiple hypothesis testing. In the field of circadian biology, it is common to use the Benjamini-

Hochberg method to adjust p-values to control the False Discovery Rate (FDR) [13]. This adjustment increases the stringency of the adjustment dependent on the number of hypotheses being tested. Though p-values should be uniformly distributed between 0 and 1 if they are accurate, most rhythm detection methods do not produce p-values that are uniform. Instead, the p-values tend to be conservative, even if sometimes only slightly. This means that methods such as Storey's q-value procedure, for which increasing the hypotheses does not increase the stringency of the adjustment, are not appropriate [96]. Given the low power of some rhythm detection methods and the tens of thousands of genes or probes that are tested, the Benjamini-Hochberg adjustment can result in few or no genes being identified as rhythmic for a moderate threshold (such as FDR of 0.05). As a result, some studies report the number of genes that pass a p-value threshold, altogether ignoring the step of correcting for multiple hypothesis testing [25, 84, 63].

1.4 Types of Recent Experiments

Most circadian studies pursue an experimental progression. First, high-throughput time series experiments implicate a large number of potential genes as having rhythmic expression. This leads to the hypothesis that they are potentially regulated by the circadian clock. From this list, several genes are selected for further experiments, either more targeted time series measurement of RNA or protein expression, or perturbative experiments, such as RNAi, where the gene is mutated or molecularly disrupted to understand its physiological role. The criteria for selecting genes varies. One could choose the genes that are the most rhythmic, involved in a biological pathway or annotation of interest, or most connected in network-based models built from the data [120]. There is often a push to have as many genes as possible to choose from, which underlies the drive to use high FDR thresholds or to dispense with FDR correction altogether when identifying rhythmic expression. However, having a rhythm detection method that is not stringent, searches for the wrong features, or accepts

too many genes can bias the results against enriched biological annotations or high-yield targets worthy of follow-up experimental study as much if not more than being too stringent can.

1.5 Types of Recent Rhythm Detection Methods

In the past decade, several types of approaches have been taken to detect rhythms in this sparse, noisy, and high-throughput data. One approach attempts to fit data to a cosine, such as COSOPT [97], and ARSER [117]. Another tests rhythmicity by assessing power of a Fourier transform, such as F24 [111, 112] and Lomb-Scargle [67, 93]. A third approach correlates or matches to a reference waveform, such as JTK_CYCLE [52] and RAIN [99]. A final approach is to make no *a priori* assumptions about the reference waveform, such as Cyclohedron Test [78], Address Reduction [3], Stable Persistence [26, 20], and ANOVA [59]. Each of these methods has its benefits and limitations. The cosine-fitting and Fourier-based methods are parametric, making the assumption that the rhythmic time series they attempt to identify matches a sinusoid in terms of time from peak to trough and difference between the points. When these assumptions are met (for example, by a time series closely matching a sinusoid), these methods perform extremely well and generate extremely small p-values. When these assumptions are not met, however (for example, when the time from trough to peak is much shorter than the time from peak to trough), then these methods do not perform as well. The reference waveform-matching methods are non-parametric in nature: they don't use the time series measurements themselves, but instead use the ranks of those time series measurements relative to one another. This allows for the use of reference waveforms that can match different aspects of a rhythmic time series, such as phase or time from peak to through (width or asymmetry), which can be measured discretely due to finite time sampling, without having to consider the continuous possibilities of amplitude and distance between points. These non-parametric tests have greater sensitivity for waveforms that do

not match the pattern of a sinusoid (either by slope or width of peak) or have differences in asymmetry from a sinusoid. However, their non-parametric nature means that very slight but well-ordered changes in a time series can be detected as rhythmic. The reference waveform-free methods require fewer assumptions. ANOVA compares distance between means to the variance of all the points, while the Cyclohedron Test, Address Reduction, and Stable Persistence methods test for monotonicity in the time series. These methods do not generally perform as well as the others for rhythm detection. Some of these methods will be discussed and compared in the following chapters, where further description of them will occur. For additional details and content, I refer the reader to review articles containing comparisons of some of these methods [22, 122].

Chapters 2 and 4 of this dissertation present two rhythm detection methods based on the non-parametric JTK_CYCLE algorithm [52]. Empirical JTK_CYCLE (eJTK, Chapter 2) improves on the p-value estimation of JTK_CYCLE via permutation to generate an empirical null distribution, which JTK_CYCLE p-values lack, and then allows for asymmetric waveforms to be compared to the experimental time series. Bootstrap empirical JTK_CYCLE (BooteJTK, Chapter 4) uses an empirical Bayes method to better estimate time point variances and then performs eJTK on bootstrap replicates of the original time series. Chapter 4 also introduces a procedure to reduce the computational cost of computing the null distribution and identifies improvements to two other rhythm detection approaches.

Chapter 3 of the dissertation examines ways to improve two of the more recent methods, RAIN [99] and [113], which make incorrect assumptions about the independence of p-values which result in underestimates, increasing the likelihood of false positives. These methods are corrected and improved using techniques discussed in Chapter 2, and then are compared against eJTK and BooteJTK in Chapter 4.

1.6 Differential Rhythmicity

As high-throughput experiments have become more common, studies have begun to make comparisons between time series collected under different conditions or in different tissues. An obvious question is whether a change in rhythmicity exists between the time series. The common way to perform this comparison has been to observe whether the rhythmicity p-values both occur below a significance threshold, signifying no change in rhythmicity, or if one p-value occurs above the threshold, signifying a change in rhythmicity. This approach does not account for the uncertainty in measurement or p-value calculation, however, and relies on the arbitrary choice of a rhythmicity p-value threshold. In May 2016, a method called Determination of Differential Rhythmicity was published. It fits sine waves to each time series and uses an ANOVA method to test whether the parameters of the fitted sine waves are significantly different [100]. In Chapter 5, we discuss a different approach to identify differential rhythmicity. Our approach, Bootstrap Differential Rhythmicity (BDR) uses BooteJTK phase estimates and rhythmicity test statistics. By using simulated data to predict the expected variance in difference between test statistics when the time series have the same rhythmicity, null hypothesis significance testing can be performed to determine if a gene's expression has changed rhythmicity between tissues or conditions.

The main chapters of this dissertation, Chapters 2, 3, 4, and 5, are written as stand-alone publishable units, and as such may contain introductions or discussions that are repetitive relative to one another. I hope this repetitiveness is helpful to the reader in keeping track of the concepts and issues discussed.

CHAPTER 2

RHYTHM DETECTION: EMPIRICAL JTK_CYCLE

2.1 Abstract

Robust methods for identifying patterns of expression in genome-wide data are important for generating hypotheses regarding gene function. To this end, several methods have been developed for detecting periodic patterns. We improve one such method, JTK_CYCLE, by explicitly calculating the null distribution such that it accounts precisely for multiple hypothesis testing and, in turn, by including non-sinusoidal reference waveforms. We term this method empirical JTK_CYCLE with asymmetry search, and we compare its performance to JTK_CYCLE with Bonferroni and Benjamini-Hochberg corrections for multiple hypothesis testing, as well as to five other rhythm detection methods: Cyclohedron Test, Address Reduction, Stable Persistence, ANOVA, and F24. We find that ANOVA, F24, and JTK_CYCLE consistently outperform the other three methods when data are limited and noisy; empirical JTK_CYCLE with asymmetry search gives the greatest sensitivity while controlling for the false discovery rate. Our analysis also provides insight into experimental design: for a fixed number of samples, better sensitivity and specificity are achieved with higher numbers of replicates than with higher sampling density. Application of the method to detecting circadian rhythms in a metadataset of microarrays that quantify time-dependent gene expression in whole heads of *Drosophila melanogaster* reveals annotations that are enriched among genes with highly asymmetric waveforms. These include a wide range of oxidation reduction and metabolic genes, as well as genes with transcripts that have multiple splice forms. The code for the empirical JTK_CYCLE with asymmetry method can be found at *https://github.com/alanlhutchison/empirical-JTK_CYCLE-with-asymmetry.*

2.2 Author Summary

Much biomedical research focuses on how the expression of genes changes over time. Many genes' activities vary periodically. For example, circadian rhythms repeat daily with the light-dark cycle. Understanding how such rhythms couple to biological processes requires statistical methods that can identify cycling time series in typical genome-wide data. In this paper, we improve on a method used to identify cycling time series by better estimating the statistical significance of periodic patterns and, in turn, by searching for a wider range of patterns than traditionally investigated. We apply these methods to a compilation of data on gene expression in fruit flies, an important model organism. We find that our method allows us to discover rhythmic biological activities that the other methods tested are unable to reveal.

2.3 Introduction

Rhythmic behavior is ubiquitous across the spectrum of life [28, 1, 83, 104]. Diverse fundamental biological functions such as cell division, energy metabolism, and sleep are periodic, and a growing body of evidence implicates temporal dysregulation as a contributing factor to depression, neurodegeneration, cardiovascular disease, and metabolic disorders in higher organisms [116, 9, 95, 40, 74]. Arguably the most well-studied periodic patterns are circadian rhythms: oscillatory changes in gene expression, metabolism, physiology, and behavior with approximately 24 hour (24 h) periods that enable organisms to anticipate and respond to daily changes in their environment, such as nutrient accessibility, temperature, and light [5, 24, 47, 112].

Circadian rhythms arise from innate clocks. The components of the core clock are well characterized and are strongly conserved across a wide range of species [79, 75]. However, it remains to be determined how this clock couples to other molecular processes. Moreover,

these interactions are likely to depend on tissue type and environmental conditions [24, 95, 123, 12, 1]. There is thus a need to identify molecular profiles that cycle and to characterize them as a function of conditions. The advent of high-throughput methods for measuring gene expression now makes transcriptome-wide studies of this nature possible. Previous work suggests that hundreds, possibly thousands, of genes are regulated by circadian clocks [5, 75, 19].

Despite the decreasing cost of measuring transcript levels, profiling time series genome-wide continues to present formidable challenges: tissue-specific samples are difficult to collect, and, in contrast to imaging, measuring transcript levels is destructive in nature, requiring separate samples for each time point. As a result, gene expression time series are typically sparsely sampled (e.g., every 2-4 hours (h) in circadian studies), often without multiple measurements per time point, which we refer to here as *replicates*. These experimental limitations result in low signal-to-noise ratios that prevent straightforward identification of cycling gene expression.

Quantitative methods are thus needed to identify rhythmic time series from minimal data with statistical confidence. These methods can aid researchers in assessing the tradeoffs between the amount of data acquired, statistical precision, and breadth of biological discovery. While a number of different methods have been proposed for identifying cycling time series [23, 77, 78, 29, 3, 26, 20, 52, 59, 22, 122, 99], further analysis is needed to guide selection of the best method(s) for a given situation and to aid in design of improved computational methods and further experiments.

In this chapter, we improve on the JTK_CYCLE method [52]. The original method uses a conservative estimation for its p-values and a cosine as its only reference waveform. Here, we introduce a procedure, empirical JTK_CYCLE with asymmetry search, that provides accurate empirically-calculated p-values for arbitrary waveforms. We test its performance for detecting rhythms in simulated data and a circadian metadataset [59] against other algo-

rithms: Cyclohedron Test [77, 78], Address Reduction [29, 3], Stable Persistence [26, 20], F24 [111, 110], and one-way analysis of variance (ANOVA) [59]. The simulated data allow us to examine how performance varies with sampling density, number of replicates and/or periods, noise level, and waveform. Most methods provide accurate rhythm detection when sampling density is high and noise is low. However, we find that the choice of method significantly affects rhythm detection when data are limited and/or noisy. In particular, JTK_CYCLE, F24, and ANOVA consistently outperform the other methods and offer distinct advantages for certain types of data. Our improved method, empirical JTK_CYCLE with asymmetry search, performs best of all for data that include asymmetric waveforms. Application of our improved method, empirical JTK_CYCLE with asymmetry, to a metadataset of whole head *D. melanogaster* circadian microarrays [59] reveals a strong lights-on peak in expression for genes involved in glutathione metabolism, high enrichment for genes involved in oxidation reduction, many more metabolic genes cycling than previously appreciated, and rhythmic genes with transcripts that have alternative splicings.

2.4 Methods

2.4.1 Overview

The methods that we consider are general and can be applied to detecting periodic behavior in any context, but we describe them here in terms of searching for circadian rhythms in gene expression for clarity. The methods that we test are Cyclohedron Test [77, 78], Address Reduction [29, 3], Stable Persistence [26, 20], F24 [111, 110], one-way analysis of variance (ANOVA) [59], and JTK_CYCLE [52]. We describe each briefly below and note specific features; additional details can be found in the references introducing the methods.

The methods can be broadly categorized as tests with and tests without reference waveforms. Cyclohedron test, Address Reduction, Stable Persistence, and ANOVA seek to iden-

tify patterns without specifying the waveform *a priori*. Address reduction, Cyclohedron Test, and Stable Persistence test for monotonicity. ANOVA compares the means of different time points with their variances to determine if differences are significant.

In contrast, F24 and JTK_CYCLE compare the time series in question to a reference waveform, which is typically sinusoidal. These methods also test for a specific period. As mentioned above, here we assume a period of 24 h, but the period of the reference can be varied, in the same manner that the phase can be varied, to search for rhythms on other time scales.

Cyclohedron Test

Cyclohedron test [77, 78] maps data to a cyclohedron and joins data points into sets according to their adjacency in rank-ordering. Monotonicity is quantified by how the sizes of the sets scale as more data points are included. Cyclohedron test has an exact null distribution computable from a generating function. The domain of test statistics increases very quickly with the number of data points, however, so Monte Carlo (MC) sampling, in which representations of the null model are randomly generated and evaluated, is required to estimate p-values if there are more than approximately twenty time points due to the computational cost of the generating function.

Address Reduction

Address reduction [29, 3] measures the entropy of the dataset by comparing the rank-ordering of adjacent time points. Low entropy data are monotonic and score higher in the method. The null distribution for Address Reduction is estimated by MC sampling.

Stable Persistence

In Stable Persistence [26, 20], local minima are paired with surrounding local maxima, and the "persistences" of these features are characterized by the differences in ranks of the paired extrema. A hierarchy of such features is established, and the test compares the global persistence to local ones. In this way, Stable Persistence tries to robustly assess overall monotonicity of a time series. The null distribution for Stable Persistence is estimated by MC sampling.

Analysis of Variance (ANOVA)

One-way ANOVA is a standard statistical test of the equivalence of means in several groups. In this case, each time point is a different group, and ANOVA is equivalent to testing for any statistically significant variation across the time points. Because expression measurements are averages over many cells and different time points come from different samples (as the measurement is destructive), only synchronized, consistent variation across all samples can generate a statistically significant trend. By this reasoning, significant changes in expression across time points can be attributed to time-dependent expression within the population, such as circadian rhythms. ANOVA tests for these time-dependent changes in expression. ANOVA has an exact null distribution derived from an assumption of normally distributed data; unlike the other five methods, however, ANOVA requires replicates to estimate the variance of experimental measurements at each time point.

24 Hour Fourier Projection (F24)

F24 [111, 110] assesses periodicity by focusing on the 24 h period of the Fourier transform of the data. The test statistic for F24 is the projection of the data onto the 24 h Fourier basis function, and the null distribution is obtained by recomputing the test statistic over repeated random permutations of the data. The phase is determined by projecting the data

onto the cosine part of the Fourier basis function and finding the optimal phase for the projection. We find that the null distribution can be modeled by the Gamma distribution (Supplementary Fig. 2.S1), which we parameterize from the mean and variance of the null distribution. We estimate the null distribution from a small number of permutations (usually 100). This allows more rapid and precise computation of p-values than can be obtained by standard permutation. Testing periods other than 24 h is accomplished simply by changing the period of the Fourier basis function used to compute the test statistic.

Jonckheere-Terpstra-Kendall Cycle (JTK_CYCLE)

JTK_CYCLE [52] computes the Kendall τ rank correlation coefficient between the data and a reference function over a range of possible reference function phases. For two time series $\vec{x} = (x_1, x_2, \ldots, x_n)$ and $\vec{y} = (y_1, y_2, \ldots, y_n)$,

$$\tau(\vec{x}, \vec{y}) = \frac{\sum_{1 \leq i < j \leq n} \operatorname{sgn}(x_j - x_i) \cdot \operatorname{sgn}(y_j - y_i)}{\frac{1}{2} n(n-1)} \tag{2.1}$$

where $\operatorname{sgn}(x)$ is –1 if $x < 0$ and +1 if $x > 0$. The numerator is the number of pairs that vary concordantly between the two time series minus the number that vary discordantly (Fig. 2.1A). Every possible pair is included, not just ones between neighboring points in the time series. The denominator is the total number of pairs, which normalizes the value of τ to the range $[-1, 1]$. Perfectly correlated series score $\tau = 1$, perfectly anti-correlated series score $\tau = -1$, and uncorrelated series score $\tau = 0$. Like the Cyclohedron Test, the null distribution for JTK_CYCLE can be computed exactly from a generating function [44], although again the test statistic domain grows quickly with time series size (becoming impractically large at 100-200 time points with present computing power). However, for large time series the JTK_CYCLE null distribution is approximately normal, allowing for a convenient, fast p-value estimate.

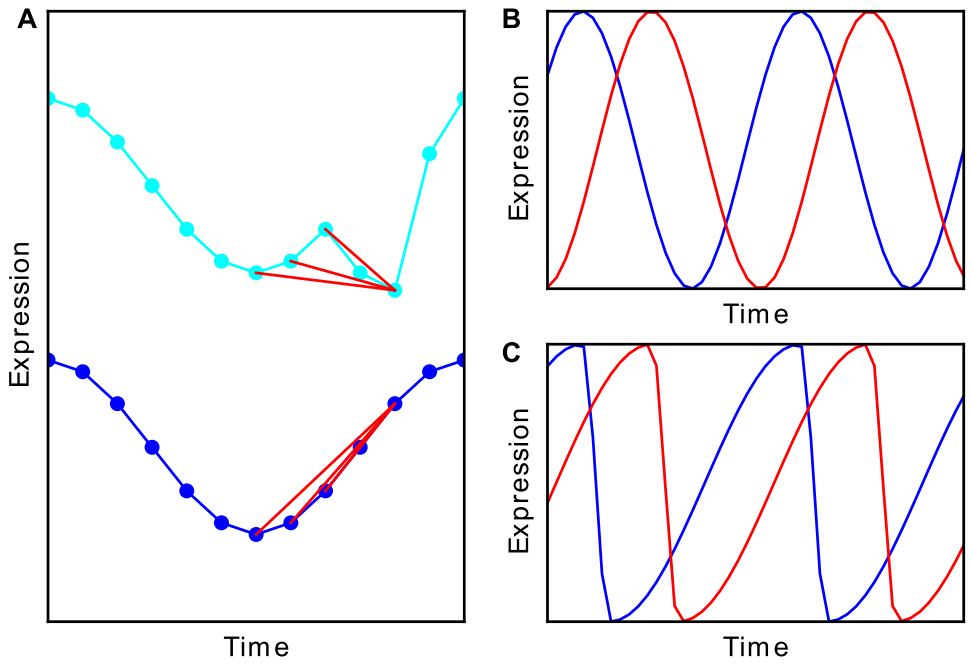

Figure 2.1: JTK_CYCLE compares all possible pair relations for a time series to those for a reference waveform. (A) JTK_CYCLE tests for pairwise agreement between a reference (blue) and a signal time series (cyan). Three discordant pairwise relationships are indicated by red lines. (B) The previous implementation compared a time series to a set of phase-shifted cosines. (C) We add a set of asymmetric waveforms to the reference. An example is shown here with the same phases as in A.

2.4.2 Improvements to the JTK_CYCLE Method

It is important to note that τ (Eq. 2.1) is calculated for a specific reference time series, and thus JTK_CYCLE typically tests against a family of curves (e.g., to consider the possible phases of a waveform, as illustrated in Fig. 2.1B). It is thus necessary to account for multiple hypothesis testing across reference waveforms in assessing the significance of the results. Hughes *et al.* [52] employed the Bonferroni correction [109] in their original formulation and implementation of the method. This method is known to be conservative [109], and we illustrate this fact here explicitly for JTK_CYCLE (Fig. 2.2). These considerations motivate a new procedure for estimating the significance of the results, which we describe. We end this section by discussing the comparison of time series to reference waveforms (Fig. 2.1C) other than the cosine waveform that was used originally. Together, our improvements allow for the JTK_CYCLE method to include additional reference waveforms in its rhythm detection without compromising sensitivity and specificity.

Empirical p-values

By definition, a p-value is the likelihood of obtaining a test statistic equal to or more extreme than the value that is observed if the null hypothesis is true—it increases cumulatively as one progresses through a set of rank ordered test statistics. In the case of JTK_CYCLE, under the null hypothesis, time series values are independent (and generated by the same noise distribution) and so the rank ordering of time series values is random. For a dataset generated from this null model, the p-values should be uniformly distributed from 0 to 1, exclusive: the highest Kendall's τ out of N tests should have a p-value of $1/(N+1)$, the second highest test statistic has a p-value of $2/(N+1)$, and the i^{th} highest test statistic has a p-value of $i/(N+1)$ [82]. Restated, the p-values should be a linear function of the ranks (black lines in Fig. 2.2).

JTK_CYCLE computes the Kendall τ values for all the reference time series against the

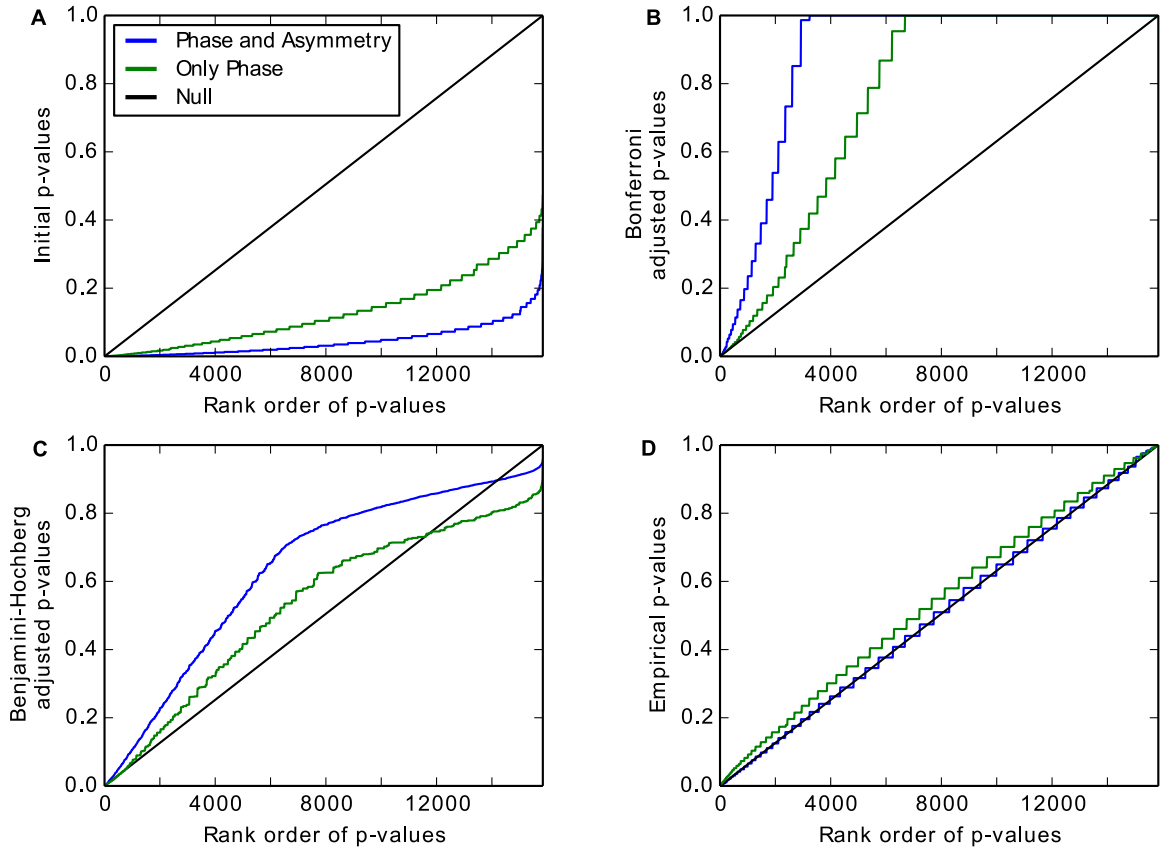

Figure 2.2: Empirical p-values are uniformly distributed for the null model of JTK_CYCLE. P-values versus their ranks from lowest (most significant) to highest (least significant) for JTK_CYCLE testing phases at 2 h intervals (green line) or phases and asymmetries at 2 h intervals (blue line) with time series consisting of Gaussian noise. Unbiased estimates should follow the black line (see text). (A) "Initial" p-values from JTK_CYCLE with multiple hypothesis testing underestimate the true p-values. (B) The Bonferroni correction results in p-values that are too high (less significant). (C) The Benjamini-Hochberg correction performs better than the Bonferroni correction but still results in p-values that are generally too high. (D) Empirical p-values that we calculate by permutation are close to uniformly distributed, as desired; their correspondence to the null model improves as the number of hypotheses tested increases.

signal of interest and then performs a selection step for the lowest p-value (i.e., the highest τ), which we refer to here as the "initial" p-value. This procedure biases the p-values (the blue and green lines in Fig. 2.2A) such that they underestimate the true probability of obtaining test statistics by chance (the black line in Fig. 2.2A). The previous version of JTK_CYCLE corrects for underestimating the p-values with the Bonferroni correction, which controls the family-wide error rate (FWER) by multiplying the p-values by the number of hypothesis tests being performed. The FWER is the probability that there is at least one false positive for a given threshold. Therefore, a threshold of 0.01 means that there is a 1% chance that the list of time series with a Bonferroni adjusted p-value below 0.01 contains a false positive. This method, while rigorous, is overly conservative and overcompensates for the bias that comes from selecting the lowest p-value (blue and green lines in Fig. 2.2B). The likelihood of false positives is greatly reduced, but so is the likelihood of identifying true positives.

A common alternative to the Bonferroni correction is the Benjamini-Hochberg procedure [13], which seeks to control the false discovery rate (FDR). The FDR is the fraction of the time series that are identified as cycling that are false. For example, a Benjamini-Hochberg adjusted p-value cutoff of 0.05 means that 5% of the positives are false. This is a less stringent constraint than the FWER. In this procedure, the p-values are also multiplied by the number of hypotheses tested, as in the Bonferroni procedure. However, the p-values are additionally ordered from lowest to highest and then divided by their rank order (the lowest p-value has rank order 1, the second highest p-value has rank order 2, and so on). The p-values are also adjusted such that there is no change in ordering: if the originally lowest p-value is adjusted so that it is higher than the originally second lowest p-value, the lowest p-value takes the value of the adjusted second p-value so that the ordering is not violated. The same holds for the relationship between the second and the third lowest p-values and so forth. While the Benjamini-Hochberg procedure is a reasonable approach to multiple hypothesis testing in general, it does not account for the selection step in JTK_CYCLE; it still is thus overly

conservative (Fig. 2.2C).

Consequently, we instead numerically compute the null distribution by applying the full JTK_CYCLE procedure to time series in which the values have random rank orders. Since we test a family of curves (e.g., spanning phases), we focus on positive correlations and compute one-sided p-values. In the present study, these "empirical" p-values are based on 2×10^6 random time series and are nearly uniformly distributed, as desired (Fig. 2.2D). Though this empirical calculation is more computationally expensive than the application of the Bonferroni correction or the Benjamini-Hochberg correction, we show that empirically calculating the p-values results in better rhythmic detection and biological insight. We term this improved method *empirical* JTK_CYCLE, as we empirically calculate the p-values after selecting the highest τ value for each time series.

Below, we compare the Bonferroni adjusted p-values, Benjamini-Hochberg adjusted p-values, and empirical p-values directly. These have been adjusted on the basis of correcting for multiple hypothesis testing across different waveforms for a single time series. When we compare different time series to each other, we have to correct again for multiple hypothesis testing, this time across time series. To do this we use the Benjamini-Hochberg correction, as in the original implementation of the method [52]. When we refer to the Bonferroni, Benjamini-Hochberg, or empirical method, we refer to corrections across different waveforms for a given time series; all corrections across time series are with the Benjamini-Hochberg method. While the Bonferroni adjusted p-values, Benjamini-Hochberg adjusted p-values, and empirical p-values represent different quantities (the FWER, the FDR, and the p-values, respectively) they are all at least as conservative as the "true" p-values in the range that we are examining (compare blue and green lines with black lines in Figs. 2.2B, C, and D). This means that the FDRs that result from the Benjamini-Hochberg correction between time series are more conservative than the true FDRs.

Asymmetric Waveforms

There is no *a priori* reason biological time series need be sinusoidal [118, 103], so it is of interest to test additional waveforms. In this regard, it is important to keep in mind that for JTK_CYCLE the rank order of the points in the reference matters, so we can represent a wide range of simple waveforms (e.g., Fig. 2.1C) by a triangle function with a specified separation between the maximum and the minimum. This allows us to avoid functionally defining an asymmetric cosine waveform. For the time series that we examine in this paper the difference is insignificant (Fig. 2.S2). We term the size of the interval from the maximum to the minimum the "asymmetry", and we express the asymmetry here in units corresponding to a 24 h period. In this notation, a cosine has an asymmetry of 12 h, while a time series with an asymmetry of 8 h has a more rapid fall than rise (the values decrease over 8 h and increase over 16 h). The triangle reference waveforms have the same monotonicity as a cosine, and we keep the convention that the peak value corresponds to the phase of the time series.

To parse the effects of empirically calculating the p-values from those of including asymmetric waveforms, we test our form of JTK_CYCLE with and without asymmetric reference time series. In the former case, we denote searching over asymmetry values in steps of 2 h "by 2 h", in steps of 4 h "by 4 h", etc. We expect these additional waveforms to be more sensitive to asymmetric patterns of gene expression, resulting in discovery of additional rhythmic time series. It is important, however, to be cognizant of the fact that we are increasing the total number of hypotheses that we test, resulting in a greater need for the empirical correction procedure. Fig. 2.2 shows the different correction methods for the minimum p-values for JTK_CYCLE with searching over 12 phases (every 2 h, green line) or searching over 12 phases and 11 asymmetries (every 2 h, blue line). The added hypotheses for searching across asymmetries result in larger selection bias when choosing the highest τ value (Fig. 2.2A) as well as larger correction biases (Fig. 2.2B and C) than when only

searching across phases. The empirical calculation of the p-values improves as the number of tests increases as well (Fig. 2.2), further justifying its use.

2.5 Results

2.5.1 Simulated Data Benchmarks

To assess the performance of our empirical form of JTK_CYCLE against the original form as well as other methods, we utilized two simulated datasets. We employ the first simulated dataset to understand the sensitivity of each method to different shapes of time series. It comprises four types of waveforms: sine, ramp (a triangle with maximum asymmetry), impulse, and step, as well as an equal number of time series consisting of Gaussian noise. We compare all the precision-recall curves for all the methods on these data via the area under the receiver operating characteristic (AUROC), a measure of the sensitivity and specificity of the rhythm detection methods that does not depend on the proportions of positives and negatives in the dataset. The second simulated dataset contains 10% rhythmic time series of triangle waveform with uniformly distributed phases and asymmetries and 90% time series consisting solely of Gaussian noise. We use it to further assess the importance of considering asymmetric waveforms, and we explore how multiple hypothesis correction impacts the results when the true positives represent a relatively small fraction of the simulated time series, as we expect to be the case in genome-wide studies.

F24, ANOVA, and JTK_CYCLE outperform other methods

To construct the first dataset described immediately above, for each of the four waveforms in Fig. 2.3A we generated 10,000 time series with uniformly distributed random phase shifts (always with a 24 h period) and added Gaussian noise to each point with a standard deviation of 25% or 50% of the total waveform amplitude, examples of which can be seen in Fig. 2.3B.

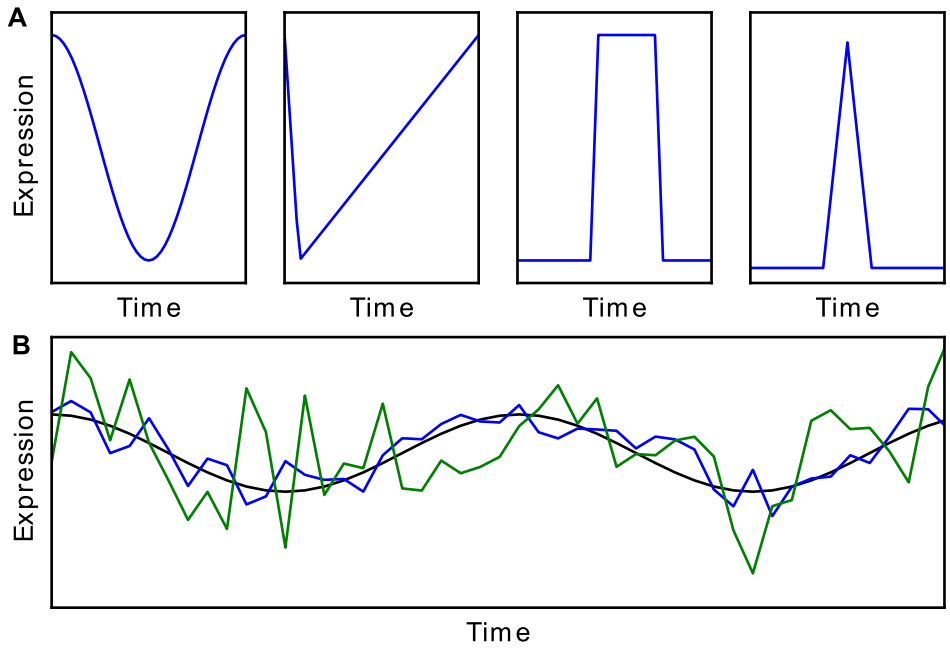

Figure 2.3: Examples of simulated data. (A) Different waveforms simulated with a 24 h period. From left to right, cosine, ramp, step, and impulse (width at half-max is 2 h). Waveforms in figure may not be to scale. (B) Cosine in black, with Gaussian noise with standard deviation of 25% (blue) or 50% (green) of amplitude.

We tested data with 4, 6, 8, or 12 evenly spaced points per 24 h period, and 1, 2, 3, or 4 replicates per time point (which is the equivalent of 1, 2, 3, or 4 periods per time series). At each spacing and replicate count we also generated 10,000 time series of Gaussian noise to serve as true negatives. The Cyclohedron Test, Address Reduction, Stable Persistence, and F24 are designed for single-replicate data, so we treated replicates as subsequent days of data, yielding multiple-period time series.

We scored each method by computing the area under the receiver operating characteristic curve (AUROC). The receiver operating characteristic (ROC) curve plots the true positive rate (TPR) as a function of the false positive rate (FPR) as the threshold for calling a time series as a positive is varied. The TPR and FPR are the fractions of the 10,000 simulated or Gaussian noise time series determined to be rhythmic at a threshold, respectively, and the threshold is varied over the entire range of false positive scores, such that the FPR ranges

from 0 to 1. The AUROC is the integral of this curve; perfect classifiers have an AUROC of 1.0, while random classifiers have an AUROC of 0.5. An advantage of the AUROC as a metric is that it does not depend on the proportions of positives and negatives in the dataset because the TPR and FPR are calculated separately, i.e., they are normalized by the total number of positives and negatives, respectively. For Stable Persistence, the Cyclohedron Test, and Address Reduction, we calculate the AUROC from the test statistics themselves as opposed to the p-values, which we use for the latter three methods. Although the AUROC for JTK_CYCLE can be computed directly from the Kendall's τ statistic, we include the multiple hypothesis testing correction because it impacts the TPR and FPR in practice; in particular, aggressive correction can lead to a loss of rank information because p-values must be less than or equal to 1.

The performance of the different methods at 50% noise can be seen in Fig. 2.4 (performance at 25% noise can be seen in Fig. 2.S3). ANOVA requires multiple measurements at each time point to determine the variance, so we define ANOVA to have an AUROC of 0.5 (performs no better than random guessing) when there is only one replicate. At 25% noise and high sampling rate, the Cyclohedron Test, Address Reduction, and Stable Persistence all perform roughly equivalently to the JTK_CYCLE methods, F24, and ANOVA. However, the former do noticeably worse than the latter at 50% noise. Empirical JTK_CYCLE outperforms original JTK_CYCLE, F24, and ANOVA for the sine and ramp waveforms, while ANOVA generally outperforms the other methods for the step and impulse waveforms.

While the empirical calculation approximates the null model well, it does not fully prevent multiple hypothesis testing from weakening the ability to identify rhythmic time series. Therefore, we do not sample phases and asymmetries more densely than the resolution of the data (e.g., if the experimental time points are spaced every 4 h, then we do not test phase values spaced every 2 h). We break this rule in Fig. 2.4 for the time series with 4-8 points for consistency of the figure. Sampling phases and asymmetries more densely than

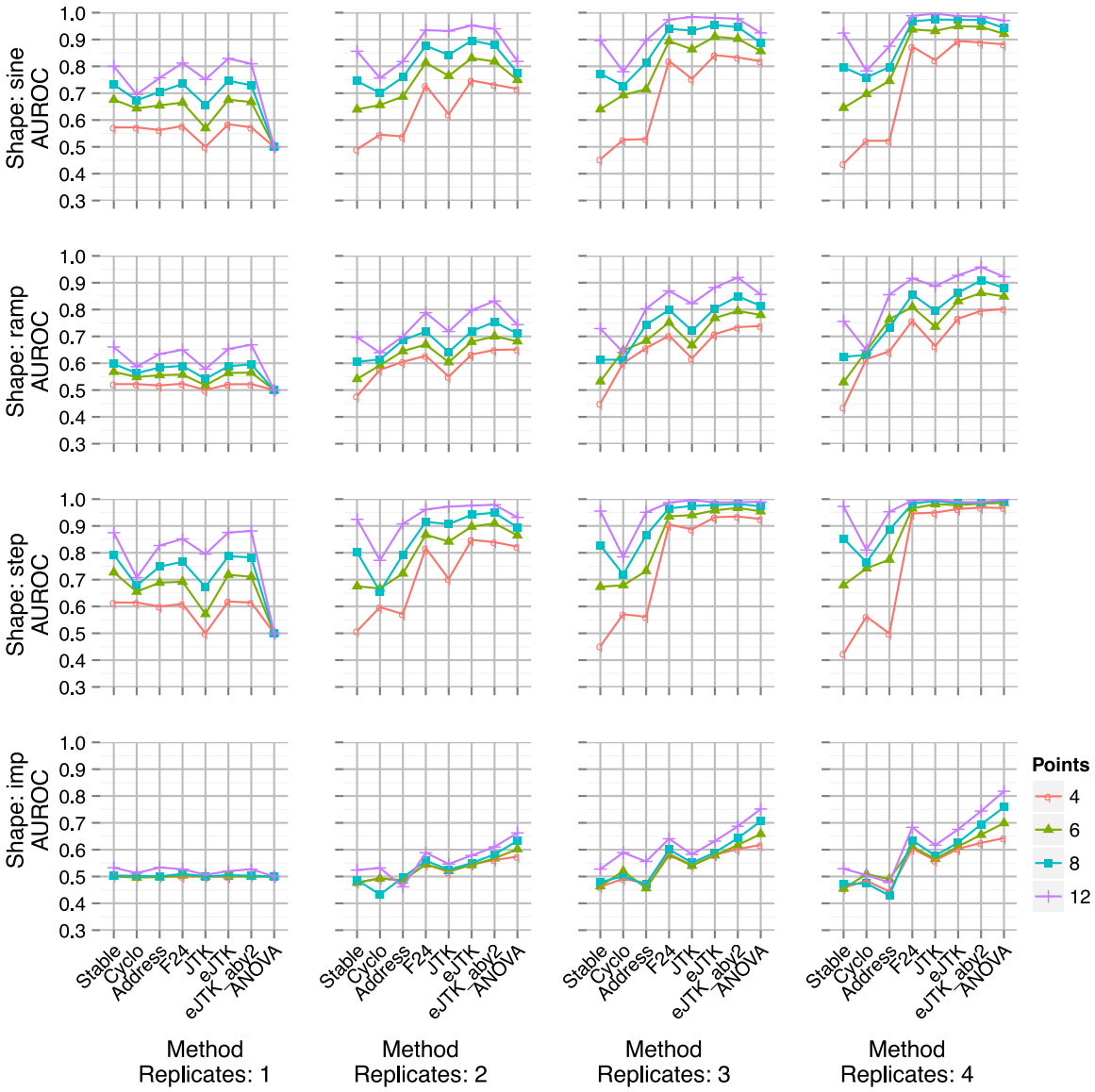

Figure 2.4: AUROCs for simulated data with 50% noise (standard deviation of Gaussian noise as a percent of amplitude). An AUROC value of 1 represents perfect discrimination between rhythmic and arrhythmic time series, and a value of 0.5 corresponds to random guessing. In each panel, the number of replicates increases from 1 to 4 replicates from left to right, and the number of sampled points per period is indicated by color. AUROC for single-replicate ANOVA (for which the method is undefined) is set at 0.5 exactly. Imp: impulse waveform, Cyclo: Cyclohedron Test, Address: Address Reduction, Stable: Stable Persistence, JTK: original JTK_CYCLE with Bonferroni correction, JTK_BH: JTK_CYCLE with Benjamini-Hochberg correction with symmetric triangle reference, eJTK: empirical JTK_CYCLE with symmetric triangle reference, JTK_BH_aby2: JTK_CYCLE with Benjamini-Hochberg correction and triangle references with asymmetries from 2 to 22 h by 2 h, eJTK_aby2: empirical JTK_CYCLE with triangle references with asymmetries from 2 to 22 h by 2 h.

the resolution of the data needlessly reduces the power of our test but does not affect the analysis in Fig. 2.4.

The JTK_CYCLE with Benjamini-Hochberg correction (JTK_BH) has AUROC values that are in between the AUROC values for the original JTK_CYCLE with Bonferroni correction (JTK) and empirical JTK_CYCLE. This is to be expected since the Benjamini-Hochberg method is more conservative than the empirical method but less conservative than the Bonferroni method. An additional detail is that the original JTK_CYCLE here uses a cosine as a reference waveform, in comparison to the triangle used by the other JTK_CYCLE methods. The methods that use the triangle waveform do not do significantly worse than the methods that use the cosine waveform in any of the cases, justifying the use of the triangle waveform for rhythm detection.

The Cyclohedron Test, Address Reduction, and Stable Persistence fail to improve as the number of replicates increases and perform worse for low sampling rates. For example, a sine wave sampled at 4 time points per period for multiple periods has extrema at every other time point. Because Cyclohedron Test, Address Reduction, and Stable Persistence are essentially tests of monotonicity, they fail to detect the sparse periodic pattern in such data. In fact sparsely sampled data sometimes results in scores consistently lower than expected by chance, leading to the AUROC values less than 0.5 for these methods on some datasets in Fig. 2.4.

In summary, we find that all the methods tested can identify rhythmic expression patterns when the sampling density, replicate number, and signal-to-noise ratios are high. If data are sparse or noisy, however, method choice can significantly impact rhythm detection. In such cases, we find that ANOVA, F24, and JTK_CYCLE consistently better distinguish true and false positives. Empirical JTK_CYCLE outperforms ANOVA, F24, and original JTK_CYCLE for sine and ramp waveforms, but ANOVA performs better for impulse waveforms.

Increasing replicate number for a fixed number of total measurements improves sensitivity

The total number of samples required for an experiment is the product of the number of time points and the number of replicates. Consequently, it is important to consider how best to apportion resources. To this end, in Fig. 2.5 we compare possible combinations of numbers of time points and replicates that give rise to 12 or 24 total samples (see Fig. 2.S4 for additional waveforms and numbers of samples). For this comparison, we consider sampling at a density of 4 points per period to be the minimal requirement for the identification of rhythmicity. Furthermore, we focus on genome-wide experiments where the experimental design is such that there is no meaningful difference between data collected over multiple periods and data collected at the same sampling rate in replicate over a single period. This assumption does not hold for experiments that follow the response to a synchronization event or other perturbation because time points from successive periods are not equivalent.

To optimize the performance of ANOVA, it is best to maximize the number of replicates at the expense of the number of time points, which is not surprising given the importance of accurately estimating the variance in this test. For JTK_CYCLE and F24, the choice is less clear, but greater improvement is obtained with replicate increases in the case of the step and impulse waveforms (Fig. 2.S4). By contrast, original JTK_CYCLE performs slightly better for sinusoidal waveforms with higher numbers of time points, but empirical JTK_CYCLE does not. Overall, the results suggest that limited resources are better directed at increasing replicate numbers than the density of time points.

Interpolated pseudo-replicates improve ANOVA sensitivity

Given the importance of replicates in improving sensitivity, we also explored interpolating neighboring time points to create pseudo-replicates, which would double the number of time points in the data (Fig. 2.S5). However, this requires recomputing null distributions via MC

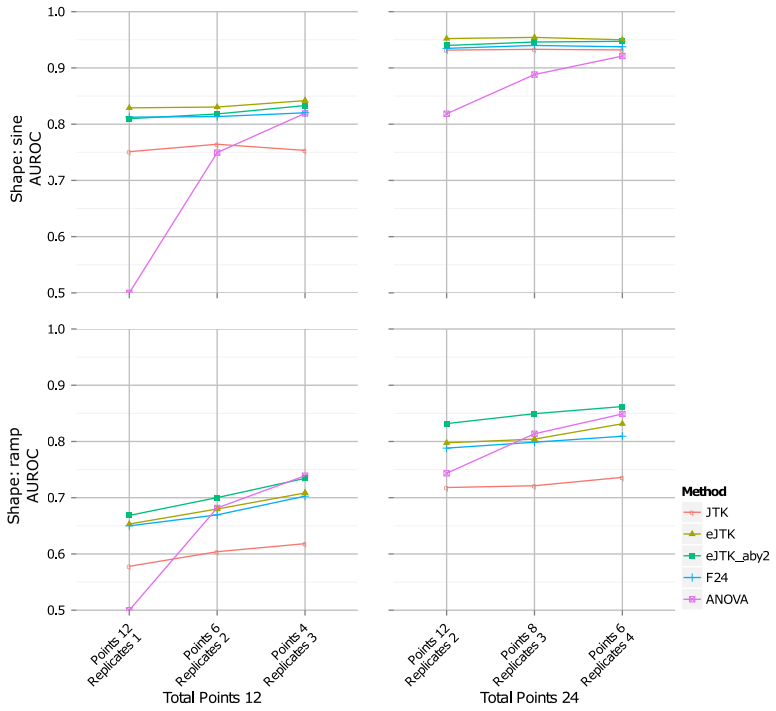

Figure 2.5: Higher numbers of replicates provide greater sensitivity compared to increased density of time points for the same number of samples. Results shown are AUROC values for sine and ramp simulated data with 50% noise (see Fig. 2.S4 for additional waveforms and sample numbers). "Points" refers to the number of time points per period ("Points 12" refers to 12 points per period) and "Replicates" refers to the number of replicates per time series ("Replicates 2" refers to 2 samples per time point). Together, "Points 12 Replicates 2" refers to a time series that consists of 12 time points per period with 2 replicates per time point. Abbreviations are the same as in Fig. 2.4.

sampling because the construction procedure introduces correlations between data points, resulting in p-value underestimates if not corrected. We found that the pseudo-replicates improved the performance mainly of ANOVA when the replicate number was low (e.g., 1 or 2); in particular, they allowed ANOVA to be applied and give good results for the single replicate case (Fig. 2.S6). We stress that two or more biological replicates should be obtained if at all possible, and we do not recommend using the pseudo-replicate approach if sufficient data are available.

Empirical JTK_CYCLE outperforms other methods after correcting for multiple hypothesis testing

Our second benchmark comprises 15,840 times series, which was chosen to allow equal numbers of time series with different phases and asymmetries. 10% of the time series were generated from a triangle waveform with noise added and 90% were generated entirely from Gaussian noise. This composition was chosen to be reflective of a genome-wide dataset. The rhythmic time series were 24 points long, with 2 periods, each with 12 time points. Here, we analyze two such datasets: one with only asymmetry of 12 h (analogous to a cosine) and one with a uniform sampling of possible asymmetries (by 2 h from 2 to 22 h). In both cases, phases (peak values) were uniformly distributed over the possible discrete values. We added Gaussian noise with a standard deviation of either 25% or 50% of the amplitude of the time series, as previously described. We tested these data against the empirical JTK_CYCLE method with various asymmetries as well as original JTK_CYCLE, ANOVA, F24, and Benjamini-Hochberg adjusted JTK_CYCLE with various asymmetries for comparison. In all cases the JTK_CYCLE methods used the triangle waveform as the reference waveform, as it was the waveform used to generate the data.

We show cumulative histograms of the number of cycling time series identified for a given significance cutoff in Fig. 2.6. The methods shown yield comparable numbers for p-values

greater than 0.05, a reasonable threshold (Fig. 2.6A and B). However, in reporting total cycling numbers, it is important to correct for the fact that we are testing many time series (as opposed to testing many waveforms for a single time series, as previously). The p-values of empirical JTK_CYCLE are approximately uniformly distributed (Fig. 2.2D), as are those of ANOVA and F24, which satisfies the assumptions of the Benjamini-Hochberg correction [13], described above, so we use it for this purpose. We also apply the Benjamini-Hochberg correction to the original JTK_CYCLE with the intra-time series Bonferroni and Benjamini-Hochberg corrections discussed previously, which results in underestimates of the true FDR since their adjusted p-values are conservative (Fig. 2.2B and C).

In Figs. 2.6C and D, we see that the performance of the methods differs considerably when controlling for the false discovery rate (FDR). For these curves, the proportion of false positives identified as cycling matches the FDR. Specifically, the Benjamini-Hochberg correction (for many time series) penalizes methods with many p-values clustered at relatively high values (corresponding to a rapid rise toward the right of Figs. 2.6A and B, as for ANOVA and F24). Thus despite the fact that ANOVA and F24 perform comparably to JTK_CYCLE in the AUROC analysis (Fig. 2.4), their p-values provide less discrimination between time series, and thus they provide less sensitivity for a given FDR. In addition to looking at AUROC scores in Fig. 2.4 and time series identification in Figs. 2.6C and D, we also computed the Matthews Correlation Coefficient [72], which quantifies the quality of a binary classification. A score of 1 indicates that a method correctly identified all true positives and true negatives, while a score of –1 indicates that a method yielded all false positives and false negatives. Fig. 2.S7 shows that the JTK_CYCLE methods have higher-quality classification ability than F24 and ANOVA for these simulated data. Furthermore, the figure shows that empirical JTK_CYCLE with asymmetry search performs equally well with and without asymmetric time series, whereas the JTK_CYCLE methods without asymmetry search perform worse when the dataset includes asymmetric time series.

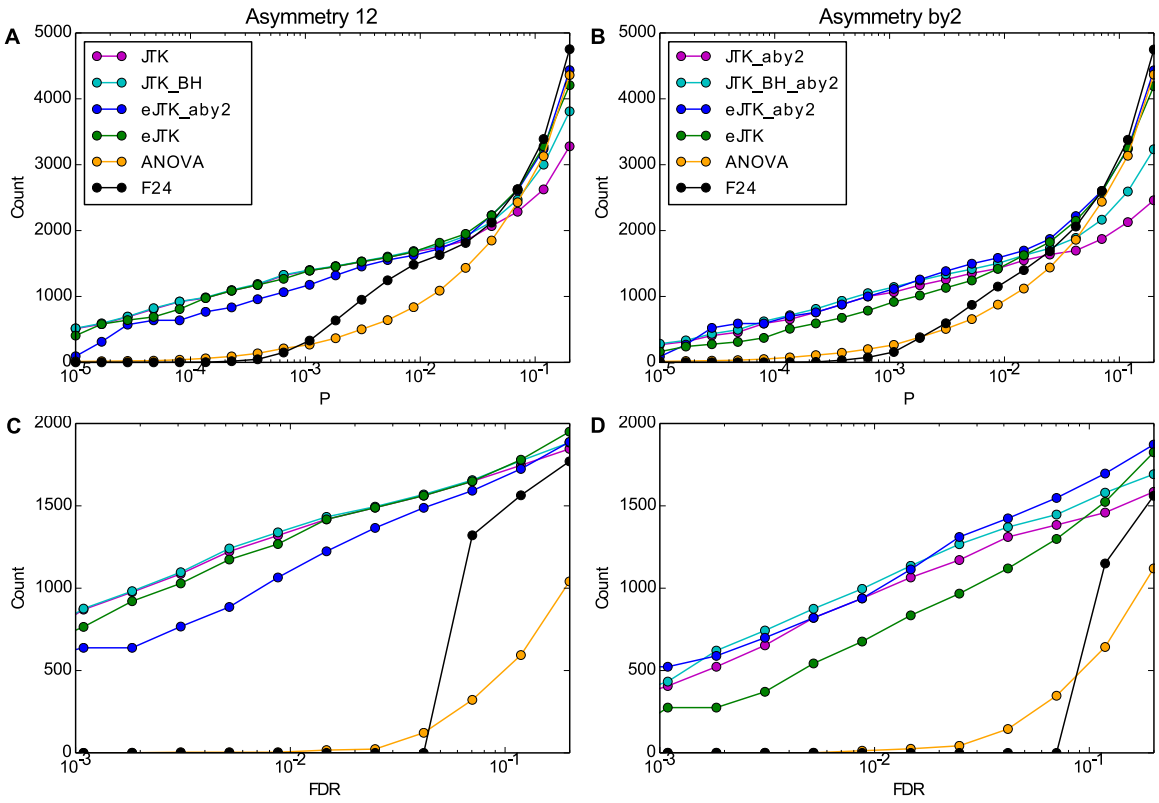

Figure 2.6: Empirical JTK_CYCLE outperforms the other methods in the presence and absence of asymmetric time series. Simulated data with rhythmic time series without asymmetry (left, A and C) or with evenly distributed asymmetry (right, B and D) were tested with different methods. The cumulative histograms are plotted before (A and B) and after (C and D) Benjamini-Hochberg multiple hypothesis correction across time series. The vertical axis shows the number of time series with a p-value (P) (A and B) or false discovery rate (FDR, the Benjamini-Hochberg adjusted p-value) (C and D) below or equal to a significance threshold, shown on the horizontal axis. Results shown are for the second simulated dataset with 25% noise, but the effects of Benjamini-Hochberg correction are significantly greater at 50% noise (not shown). The method abbreviations are the same as those in Fig. 2.4. The legends of A and B correspond to C and D, respectively. The rightmost point on the horizontal axis is 0.2.

Therefore, in terms of genome-wide studies, empirical JTK_CYCLE with asymmetric waveforms is the method of choice for identifying rhythmic genes. Fig. 2.S8 examines how the inclusion of different asymmetries affects rhythm detection.

2.5.2 Microarray Metadataset

Keegan *et al.* [59] previously assembled a metadataset comprised of data from four DNA microarray studies of *Drosophila melanogaster* under light-dark (LD) conditions (from Ceriani [18], Claridge-Chang [19], Lin [65], and Ueda [105]). We do not include a fifth dataset from that study [75], because it was limited to dark-dark (DD) conditions. Here, we discuss issues that arise from merging data from different laboratories and use the resulting metadataset to test the methods. We find that empirical JTK_CYCLE with asymmetry search identifies a larger number of rhythmic genes and, in turn, enriched annotations among those genes, such as oxidation reduction, glutathione metabolism, and alternative splicing.

Z-score-based procedure for preparing the metadataset

All of the measurements in the contributing studies are at intervals of 4 h. Time points for circadian LD time series are referenced as zeitgeber time points (ZT); the beginning of the light period is ZT0. Under 12 hours of light and 12 hours of dark, ZT24 is the equivalent of ZT0. Three studies sampled at ZT0, 4, 8, 12, 16, and 20, and the fourth (Ueda [105]) sampled at ZT1, 5, 9, 13, 17, and 21. We found that the differences in sampling protocols, together with variations from one laboratory to another, consistently gave rise to a jagged structure in the time series of known cycling genes (Fig. 2.7A). Microarray-wide normalization techniques such as quantile normalization were unable to produce curves consistent with independently measured profiles. Instead, we found that the best approach was to convert the values in each time series to Z-scores—i.e., for each gene in each dataset, we subtract its mean expression level and divide by its standard deviation. Then we pool the Z-scores to generate

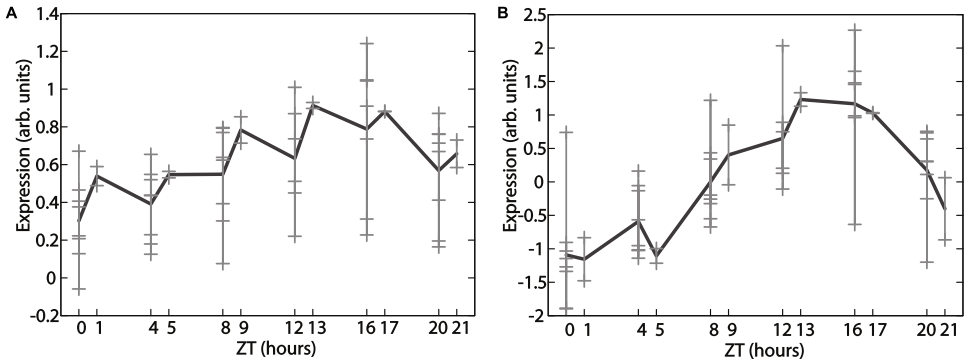

Figure 2.7: Z-score normalization allows combining of time series from different datasets into smooth time series. *Pdp1* gene expression from metadata before (A) and after (B) Z-score normalization. Light gray crosses indicate individual replicates, and the black curve is the mean.

the metadataset. Fig. 2.7 illustrates the effect of the processing step on *Pdp1*, a known cycling gene. This method is equivalent to treating measurements from the same zeitgeber time point as replicates. For probes that corresponded to the same gene, we chose the probe with the highest mean expression value to use in the analysis. This reduced 14,010 probes to 11,625 genes.

The metadata have many places where the values are not available (NA). To prevent the need to recalculate the null distribution for every pattern of NAs in the data for empirical JTK_CYCLE (there were 5005 unique NA permutations in the data), the NAs were replaced by random noise drawn from a Gaussian distribution with mean and standard deviation that match those of the data on the whole. While this adds noise to the time series, it should not have a large effect given that each time series has 57 points. To mitigate the impact of this procedure on our study, however, time series that had more then half their points as NA were discarded from the dataset, leaving 9,313 out of 11,625 genes. We consistently used the dataset resulting from these preprocessing steps for all our analysis to ensure that comparisons between methods were fair; where comparisons with and without NA substitution were possible, we found that NA substitution led to slight increases in cycling numbers in all cases except ANOVA (114 vs. 101). However, these differences did not change any of

the ontological results (discussed below).

Analysis of Microarray Metadataset

To evaluate the methods against genes for which we know the rhythmicity *a priori*, we compared the p-values for six positive examples (*per*, *tim*, *vri*, *Pdp1*, *cry*, and *Clk*) and four negative examples (*cam*, *RpL32*, *cyc*, and *dco*). Supplementary Fig. 2.S9 shows the performance of the different methods for the known positive and negative examples. Stable persistence, the Cyclohedron Test, and Address Reduction all have false negatives. The JTK_CYCLE methods, ANOVA, and F24, however, detect all of the known cycling genes and none of the non-cycling genes as rhythmic.

Having again established F24, ANOVA, and JTK_CYCLE as the better methods, we now apply them to the full dataset (Fig. 2.8). As in Fig. 2.6, the Benjamini-Hochberg correction decreases the sensitivity of ANOVA and F24 relative to JTK_CYCLE for a given FDR (compare Fig. 2.8A with Fig. 2.8B). Choosing a Benjamini-Hochberg adjusted p-value cutoff of 0.05 (i.e., 5%), the number of genes and overlap between methods can be seen in Figs. 2.8C and D. All the JTK_CYCLE methods outperform F24 and ANOVA. Empirical JTK_CYCLE with asymmetry search by 4 h (eJTK_aby4) identified the most genes, showing a clear improvement over the Bonferroni (JTK) and Benjamini-Hochberg (JTK_BH) methods with asymmetry search by 4 h (aby4); eJTK_aby4 also identified more cycling genes than methods without asymmetry search, and the genes were distinct.

Interestingly, among the JTK_CYCLE methods without asymmetry search the Bonferroni and Benjamini-Hochberg methods identified more genes than the empirical method did. For JTK_CYCLE without asymmetry search, there were only 6 hypotheses tested per gene time series (for each of the 6 phases searched), for which the Bonferroni and Benjamini-Hochberg correction across waveforms is very slight. For JTK_CYCLE with asymmetry search every 4 h, the number of hypotheses tested becomes 30, for 6 different phases paired

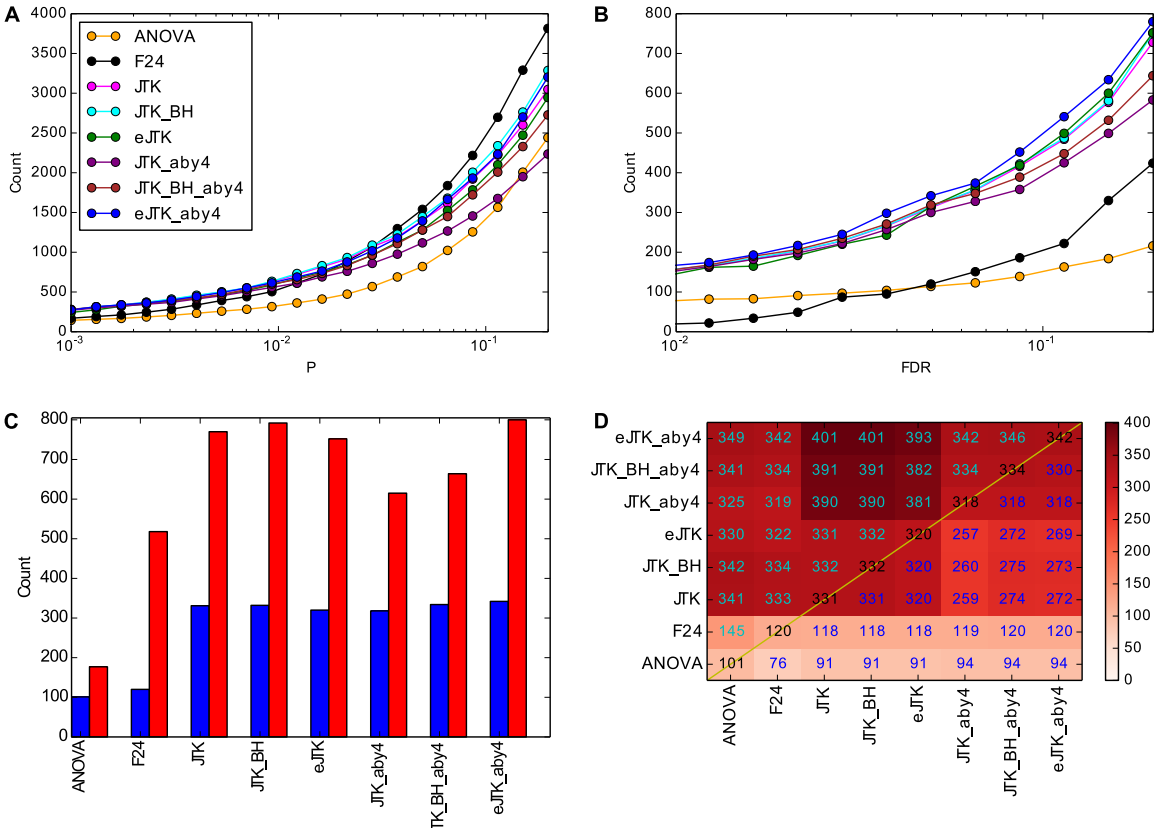

Figure 2.8: Empirical JTK_CYCLE with asymmetry search of 4 h (eJTK_aby4) identifies more genes than ANOVA, F24, and the other JTK_CYCLE methods. (A) The vertical axis shows the number of genes with a p-value below or equal to the horizontal axis for the methods indicated. The rightmost point on the horizontal axis is 0.2. (B) The Benjamini-Hochberg correction for testing multiple genes impacts the relative performance of the different methods. The rightmost point on the horizontal axis is 0.2. The colors are the same as in A. (C) The number of genes with Benjamini-Hochberg adjusted p-values below 0.05 (blue) and below 0.20 (red) with the different methods is shown. (D) A comparison of the intersection (below the diagonal) and union (above the diagonal) of genes identified as rhythmic with Benjamini-Hochberg adjusted p-values less than 0.05 for the different methods. JTK: the original JTK_CYCLE method with Bonferroni correction. JTK_BH: the JTK_CYCLE method with Benjamini-Hochberg correction. eJTK: the JTK_CYCLE method with empirical calculation of the p-values. "_aby4" refers to an asymmetry search every 4 h (at 4, 8, 12, 16, and 20 h).

with 5 different asymmetries, which results in a more stringent correction by the Bonferroni and Benjamini-Hochberg methods. As experimental sampling rates and sampling densities enable more extensive searching of phases, periods, and asymmetries, we expect the advantage for empirical JTK_CYCLE relative to the original formulation to grow because the Bonferroni correction strongly penalizes adding hypothesis tests. Provided that sufficient permutations are performed, empirical JTK_CYCLE provides the more robust identification of rhythmic genes. Another way of viewing this difference between the inclusion and exclusion of asymmetry search is by examining the distributions of the Bonferroni-adjusted p-values against the empirical p-values, as in Fig. 2.S10. With asymmetry search, the empirical p-values are significantly lower than the Bonferroni-adjusted p-values, a pattern that is less pronounced without asymmetry search.

We examine the effect of searching multiple asymmetries with empirical JTK_CYCLE further in Fig. 2.S11. Searching for rhythmic genes with asymmetries of 8 and 16 h alone yielded 4 more genes than searching for rhythmic genes with asymmetries of 4, 8, 12, 16, and 20 h, with an overlap of 293 genes and approximately 50 genes each that were separately called rhythmic by each method. Comparing the two sets of cycling genes in Fig. 2.S12, we find that searching by 4 h excludes genes with asymmetries 8 to 16 h that are barely below the adjusted p-value of 0.05 (upper left quadrant), while searching at asymmetries of 8 and 16 h excludes genes that have extreme asymmetries.

We also examined how our results depended on using a triangle vs. a cosine for the reference waveform. Figs. 2.S13 and 2.S14 show that there is no substantial difference in genes identified as cycling or in ontological results (discussed below). This can be attributed to the fact that across many time points (57 in the case of the metadataset), the differences between the cosine and triangle waveform are slight (Fig. 2.S2).

Comparison with Keegan *et al.*

We compared our results to those of Keegan *et al.* [59], an earlier analysis of this metadataset. There were two main differences in the way we constructed the dataset: we excluded time series that had more than half their values as NA, and we excluded the dark-dark (DD) McDonald dataset, as discussed above. Of the 200 genes identified as cycling by Keegan *et al.*, 169 remained after pre-processing to remove time series with more than half of their values as NAs. Of those 169, 111 had Benjamini-Hochberg adjusted p-values of less than 0.05 for the empirical JTK_CYCLE with asymmetry search by 4 h (eJTK_aby4). 58 genes that were identified as cycling by Keegan *et al.* were not identified by eJTK_aby4. Fig. 2.S15 compares the cycling genes identified by Keegan *et al.* with the cycling genes identified by eJTK_aby4. Keegan *et al.* identified genes as cycling primarily on the basis of scoring well ($p < 0.05$) on several tests following pre-screening by ANOVA. Fig. 2.S15A shows a comparison of the number of tests passed after the ANOVA pre-screening with the Benjamini-Hochberg adjusted p-value from eJTK_aby4. While there appears to be a weak relation between the number of tests passed and the p-value, there is not a clear pattern that would enable one to predict the cycling genes common to both Keegan *et al.* and eJTK_by4. Fig. 2.S15B shows the maximum amplitude measurements (after Z-scoring) for the genes identified as cycling by Keegan *et al.*, organized by whether they are identified as cycling by eJTK_aby4 as well. The genes identified by Keegan *et al.* but not by eJTK_aby4 tend to have larger maximum amplitudes than the ones identified by both. The ANOVA pre-screening in Keegan *et al.* can account for this difference; our results with empirical JTK_CYCLE suggest that there are many cycling genes with lower amplitudes. Fig. 2.S15C shows the asymmetries of the genes identified by Keegan *et al.* as cycling, as determined by eJTK_aby4. A large number of genes identified by Keegan *et al.*, but not by eJTK_aby4, have asymmetry of 16 h. The bias in the earlier study may reflect the fact that one of the tests that Keegan *et al.* employs is based on correlation with the gene *per*, which has an asymmetry of 16 h. More generally, Keegan *et*

al. fail to identify 231 genes as cycling that eJTK_aby4 identifies with Benjamini-Hochberg adjusted p-values below 0.05. Of these 231, 82 have Benjamini-Hochberg adjusted p-values below 0.01, 65 have values below 0.005, and 16 have values below 0.001.

Comparison with Wijnen *et al.*

In addition to comparing our results to those of Keegan *et al.*, we also compared our results to those of Wijnen *et al.* [110], who identified 336 genes as rhythmic using an F24-based method. Again, we excluded time series that had more than half their values as NA, and we excluded the dark-dark (DD) McDonald dataset. Fig. 2.S16 shows a comparison of the genes that are identified as rhythmic by eJTK_aby4, Keegan *et al.*, and Wijnen *et al.* Whereas 31 genes that Keegan *et al.* identified as rhythmic were removed by the empirical JTK_CYCLE analysis pre-processing, 57 genes that were identified by Wijnen *et al.* were removed by the empirical JTK_CYCLE analysis pre-processing due to more than half their time points being NA. These genes are "unassigned" in Fig. 2.S16A because an asymmetry estimate is not available. Wijnen *et al.* and eJTK_aby4 jointly identified 120 genes as rhythmic, of which Keegan *et al.* identified 59 as well. Wijnen *et al.* uniquely identified 177 genes as rhythmic, whereas eJTK_aby4 uniquely identified 167 genes. A comparison of the asymmetry distributions for all the genes (Fig. 2.S16A) shows that they are similar for eJTK_aby4 and Wijnen *et al.*

Validation with dataset-independent literature citations

As a first step toward validation the new genes that eJTK_aby4 exclusively identified as rhythmic (i.e., those genes not previously identified by Keegan *et al.* or Wijnen *et al.*), we examined the literature for references that independently suggest that these genes are cycling. Specifically, for each gene, we identified the references in FlyBase that mention the gene. Of those references, those that have the term "circadian" in their title or abstract were

identified. Fig. 2.S16B shows the distribution of genes based on their citation in FlyBase by a "circadian" paper, by the original five dataset papers [18, 19, 65, 105, 75], or by neither. Genes identified by "circadian" papers but not by the original five dataset papers represent further validation that the genes that we select as rhythmic are related to circadian processes.

Among these references, there were some that discussed several genes. Kadener *et al.* [58] assayed for genes regulated by the gene *Clk* and referenced 6 of the genes not mentioned by the original five papers out of a total of 32 genes, which has less than 1.6% probability of occurring by chance (Fisher's Exact Test unadjusted $p < 0.016$ [30]). One gene referenced by Kadener *et al.* as well as Abruzzi *et al.*, who also assayed for genes regulated by *Clk*, is *cabut* (*cbt, CG4427, FBgn0043364*, referred to as EP2237 by Kadener *et al.*). The gene *cbt* was previously unidentified as having rhythmic expression. The average time series from the metadata can be seen in Fig. 2.S17. The gene *cbt* is a metal-ion binding transcription factor downstream of the JNK cascade and is involved in morphogenesis [35, 15, 81, 80]. It has an asymmetry of 4 h, potentially explaining why it was missed by previous methods but identified by eJTK_aby4. Abruzzi *et al.* [2] also discuss another *Clk*-regulated gene that was uniquely identified by eJTK_aby4 as rhythmic, *twins* (*tws, CG6235, FBgn0004889*), seen in Fig. 2.S17. It has an asymmetry of 20 h, which explains how, like *cbt*, it could have been missed by previous methods. These genes, though previously unidentified as rhythmic, are strong candidates for having roles in circadian regulation and processes based on our identification of them as rhythmic and the work of Kadener *et al.* and Abruzzi *et al.* This warrants further experimental studies of these genes in a circadian context as well as the other genes that we have identified.

The gene *cbt* is also referenced by another study that discusses several genes identified as rhythmic by eJTK_aby4. Fujikawa *et al.* [38] identified 114 genes that are up-regulated and down-regulated in the head of *D. melanogaster* following 24 h of starvation. 16 of these genes are not mentioned by the original five papers but are identified as rhythmic by eJTK_aby4,

which has less than 0.3% probability of occurring by chance (Fisher's Exact Test unadjusted $p < 0.003$). Fujikawa *et al.* refer to several genes from the circadian dataset papers that also appear in their lists of differentially expressed genes, but they do not associate rhythmic behavior with all the genes that they describe. In addition to the gene *cbt*, Fujikawa *et al.* reference other genes that were previously unidentified as rhythmic: *Esterase-Q* (*Est-Q, CG7529, FBgn0037090*) and *1,4-Alpha-Glucan Branching Enzyme* (*AGBE, CG33138, FBgn0053138*). Both have asymmetries of 16 h, which is also outside the range of standard symmetric-waveform detection (Fig. 2.S17). The identification of these genes as rhythmic reinforces the connection between metabolism and circadian regulation and indicates other potential areas of experimental exploration.

To further understand the relationship between circadian regulation that we see in the genes eJTK_aby4 has identified as rhythmic and biological processes, we examined the enrichment of functional annotations in the identified genes.

Functional classification of cycling genes

We used DAVID [49, 50] to analyze the ontological enrichment of the genes contributing to Fig. 2.8 separately for each rhythm detection method. Because many of the annotation terms are obviously related (e.g., "oxidoreductase" and "oxidation reduction"), we manually grouped them. The grouped terms enriched with Benjamini-Hochberg adjusted p-values of less than 0.05 for the different methods can be seen in Fig. 2.9. Genes that are identified as rhythmic by F24 and ANOVA are enriched in the fewest terms. They are mainly in rhythm/light/circadian categories, corroborating the selection of these genes as cycling. The JTK_CYCLE methods without asymmetry search identify sets of genes that are enriched for different terms in addition to the rhythm/light/circadian ones found by ANOVA and F24, such as glutathione and oxidation reduction annotation terms. The JTK_CYCLE methods with asymmetry search identify sets of genes enriched in the most terms of all the methods.

The original JTK_CYCLE method with Bonferroni correction and empirical JTK_CYCLE method identify sets of enriched genes known to have alternative splice forms of their RNA; JTK_CYCLE with the Benjamini-Hochberg correction and empirical JTK_CYCLE identify sets of genes that are enriched for genes involved in biosynthetic pathways.

Because eJTK_aby4 captures all the annotation terms of interest, we focus on its results for the remainder of this section. The individual annotation terms that are enriched in the rhythmic genes found by eJTK_aby4 can be seen with their adjusted p-values and phase distributions in Fig. 2.10A. Fig. 2.10B shows the total phase distribution of the genes, and Fig. 2.10C shows the total asymmetry distribution. Functionally, these genes fall into several annotation categories, each of which we discuss in turn.

Many of the rhythmic genes involved in glutathione metabolism are also involved in drug metabolism. The peak expressions of these genes are focused around ZT4, and these genes mainly have asymmetries close to 12 h (Figs. 2.10A and 2.S18). Glutathione and drug metabolism are known to be circadian [48, 11, 43]; possible links to aging are suggested by the role of glutathione metabolism in clearing reactive oxygen species [86]. Other oxidation-reduction related terms peak at either ZT4 or ZT16-20 (Figs. 2.10A and 2.S19). These genes have a broader distribution of asymmetries, with several with extreme values of 4 or 20 h.

A subset of the genes involved with oxidation-reduction are also associated with iron and have a bimodal distribution of phases, with peaks at ZT4 and ZT16-20 (Fig. 2.10A). Various iron-related genes have been implicated as important in circadian rhythms. Recent studies, however, have only looked at the effect of iron-related genes on whole organism activity, or on particular circadian genes, such as *per* or *tim* [69, 37]. These studies have shown that individual iron-related genes affect circadian rhythms. To our knowledge, no studies to date have found as many iron-related genes displaying rhythmic behavior as we have described here.

Genes that have multiple protein forms due to alternative splicing peak at times that

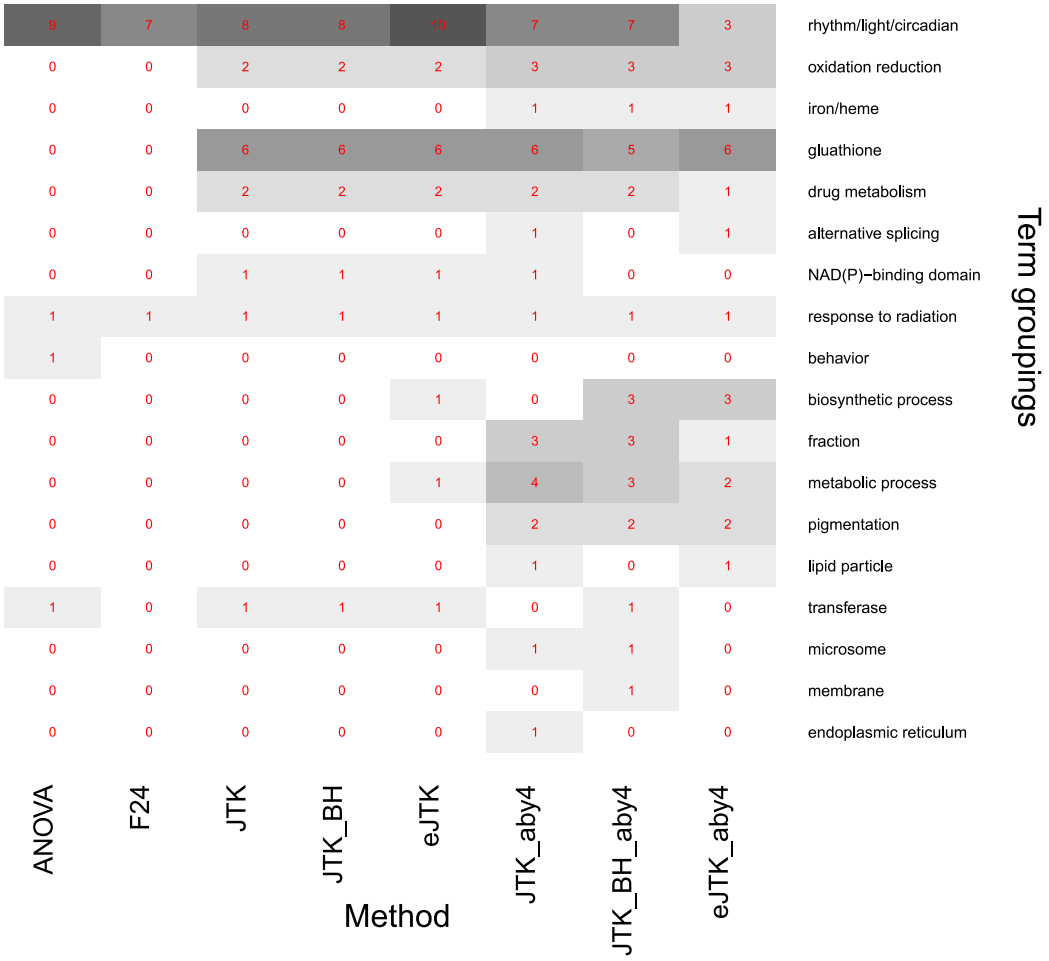

Figure 2.9: Manual grouping of annotation terms identified as enriched by DAVID. The number of annotation terms enriched in the genes with Benjamini-Hochberg adjusted p-values less than 0.05 that are identified by each method are shown in grey shading and red numbers. Annotation terms were enriched with Benjamini-Hochberg adjusted p-values below 0.05 as identified by the DAVID web tool [49, 50]. Abbreviations are the same as in Fig. 2.8.

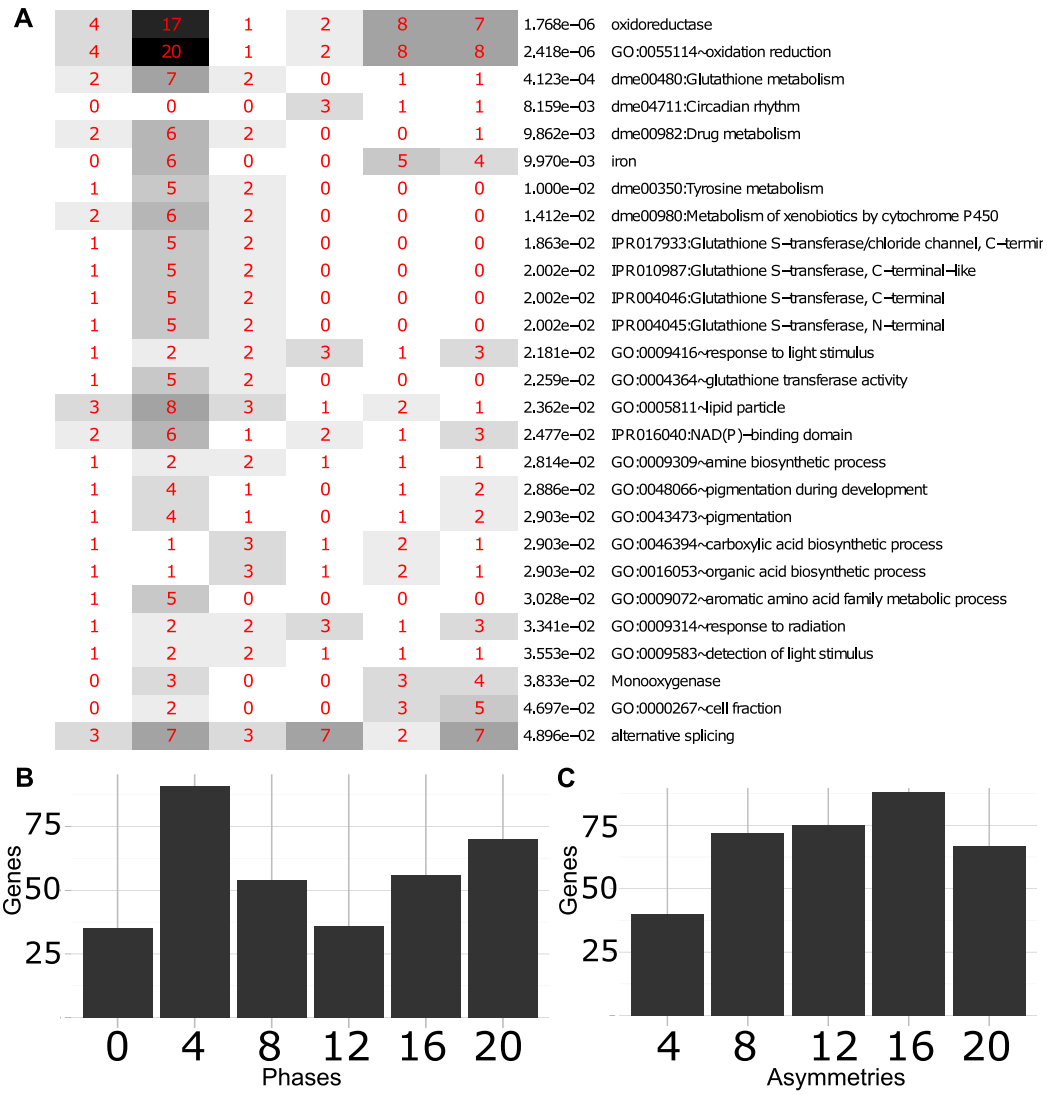

Figure 2.10: Annotation terms identified by DAVID as enriched for rhythmic genes. Rhythmic genes shown are those that are identified with eJTK_aby4 with a Benjamini-Hochberg adjusted p-value of less than 0.05. The terms shown are those identified by the DAVID web tool [49, 50] as enriched with a Benjamini-Hochberg adjusted p-value of less than 0.05. (A) The individual annotation terms are shown with their adjusted p-values and phase distributions. The red numbers refer to the number of genes in that annotation term with that phase. The horizontal axis of A is the same as that of B. (B) Total phase distribution of the cycling genes. (C) Total asymmetry distribution of the cycling genes.

are evenly distributed throughout the day (Figs. 2.10A and 2.S20). They have a broad distribution of asymmetries as well, with several genes with extreme values of 4 h and 20 h, such as *tws*, the newly discovered cycling gene previously mentioned as having an asymmetry of 20 h and as a regulatory target of *Clk*. The existence of these alternatively spliced genes with extreme asymmetries explains why "alternative splicing" was only found to be enriched in the genes identified as rhythmic by methods searching for asymmetric waveforms. Alternative splicing has been found to be important in circadian rhythms in *Drosophila* as well as in other species. Most studies, however, have focused on specific experimental findings that discovered particular genes that modulate circadian rhythms [10, 87]. No studies exist that have found that so many genes with alternative splicing are rhythmic.

2.6 Discussion

In this paper, we compare several rhythm detection methods. These approaches are general and can be applied to detecting periodic behavior in a wide range of contexts, but we focus on time series representative of genome-wide expression data. Deckard *et al.* [22] recently reviewed a number of earlier studies of rhythm detection methods and selected four algorithms for comparison (de Lichtenberg, Lomb-Scargle, JTK_CYCLE, and persistent homology) based on their mathematical properties and applicability to genome-wide expression data. They test the methods with simulated data and experimental data for the metabolic cycle in yeast, circadian rhythms in mouse, and the root clock in the flowering plant *Arabidopsis thaliana* (see [22] for references). They find that there is no all-around best method and construct a decision tree for picking an algorithm based on the expected nature of the data. For increasing noise and decreasing sampling rate, they favor JTK_CYCLE and Lomb-Scargle, a Fourier-like method. These recommendations are consistent with our own findings that JTK_CYCLE and F24 were consistently more accurate than the monotonicity tests (represented by persistent homology in [22]) and justifies our focus on improving

JTK_CYCLE.

Also recently, Zielinski *et al.* [122] reviewed six different Fourier-like methods for their ability to estimate periods in periodic time series. They focus on time series that are well-sampled (36 and 72 time points were the lowest-sampled time series they examined, 720 was among the highest), as might be obtained from tracking a luminescence reporter for a single gene. The fundamentally different nature of their data from ours highlights the fact that it is important to match the computational tool with the task of interest. Zielinski *et al.* [122] seek precise period determination for genes already known to cycle. By contrast, here we focus on discovering rhythmic time series that represent only a fraction of a genome-wide dataset. JTK_CYCLE can provide estimates of a periodic time series' phase, period, and asymmetry, but it resolves these parameters only to the level of the sampling (or search) depth. Likely there is a tradeoff between robust separation of rhythmic and arrhythmic time series and precise estimation of the cycling parameters; presently, no algorithm achieves both these goals simultaneously.

Zielinski *et al.* [122], as well as earlier studies [118, 103], note that biological oscillations are expected to have asymmetric patterns of expression. Thaben and Westermark recently published one approach to this problem [99]. Their method, RAIN, employs the Mann-Whitney U test [70] between different time points to look for a rising pattern followed by a falling pattern. They show that RAIN outperforms the original JTK_CYCLE method as well as a cosine-fitting method [21] for simulated data consisting of sinusoidal and ramp waveforms. They also analyze genomic and proteomic data for the mouse liver. Their work reinforces our finding that searching for asymmetric waveforms can produce better rhythm detection sensitivity and validates our efforts. Thaben and Westermark suggest that modifying JTK_CYCLE to allow for asymmetric waveforms would provide a useful complement to their approach. In particular, in contrast to RAIN, JTK_CYCLE can search for specific waveforms, including arbitrary shapes with multiple peaks. We meet that need

here.

We were able to expand JTK_CYCLE to search for asymmetric waveforms without degrading sensitivity because we empirically calculate p-values, which yields much more accurate significance estimates than the Bonferroni correction employed in the original formulation of JTK_CYCLE [52]. Our analysis of two different simulated datasets and the fly head metadataset clearly shows the importance of accurate significance estimates. As sequencing costs continue to decrease, we expect sampling density to increase. This trend should favor use of empirical JTK_CYCLE over alternative means of correcting for multiple hypothesis testing because the increase in data will enable more phases, asymmetries, and periods to be examined. Our analysis shows that certain gene ontologies have many genes with highly asymmetric patterns of expression. It will be interesting to determine the prevalence of different waveforms in additional biological datasets and to understand how their features depend quantitatively on genotype, tissue, and environmental conditions.

Conclusions

In this paper, we compare methods for detecting rhythmic time series in genome-wide expression data. With regard to experimental design, we find that increasing the number of replicates is more important than increasing the sampling density for achieving greater sensitivity. A key aspect of our study is that we improve the estimation of p-values in JTK_CYCLE. This enables control of the false discovery rate and testing waveforms beyond sinusoidal ones. For both simulated data and a circadian metadataset [59] the resulting empirical JTK_CYCLE with asymmetry search exhibits the greatest sensitivity among the methods that we evaluated. The annotation terms that are enriched among the genes that we identify as cycling include rhythm/light/circadian, glutathione/drug metabolism, oxidation-reduction, iron metabolism, and alternative splicing. These findings are consistent with known circadian biology but also suggest new investigations.

2.7 Supplementary Figures

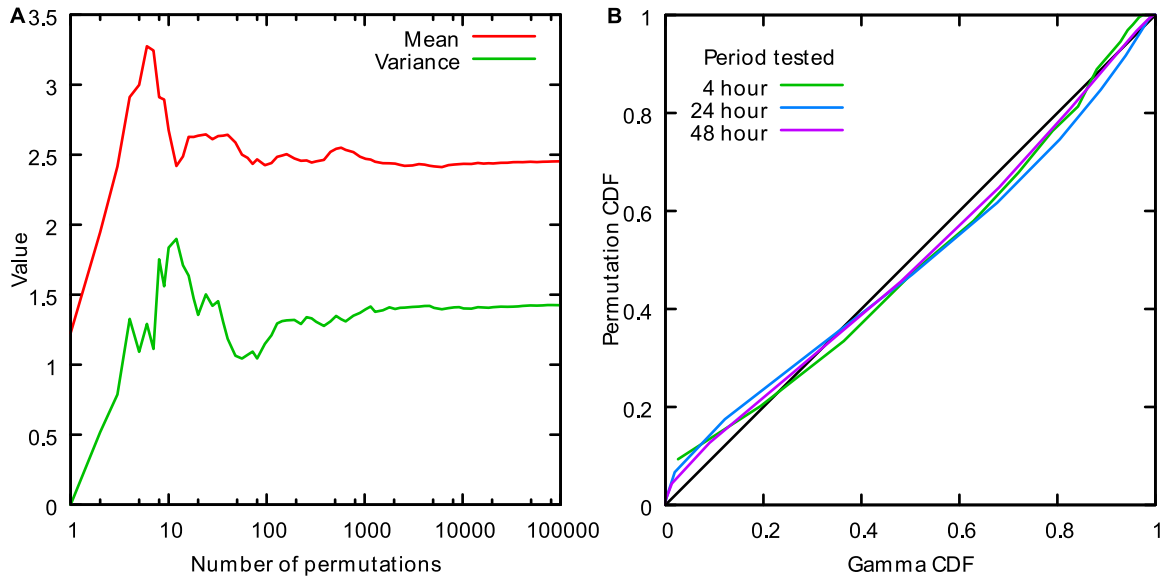

Figure 2.S1: Evaluation of Gamma distribution modeling for the F24 null distribution. The time series used for this example was a 24 h sine wave sampled every 2 h for 1 period (no replicates); noise was added at 25% of the amplitude. (A) Convergence of the mean and variance estimates, used to parameterize the Gamma distribution, as a function of the number of permutations performed, for testing the 24 h period (blue curve in A; convergence for 4 h and 48 h periods were similar, data not shown). (B) The cumulative distributions obtained by random permutation fit to the Gamma distribution, as shown by their proximity to the diagonal (black). Shown are fits for testing a 24 h period, plus a 4 h period and a 48 h period (i.e., F4 and F48). For these fits, 100 permutations were used.

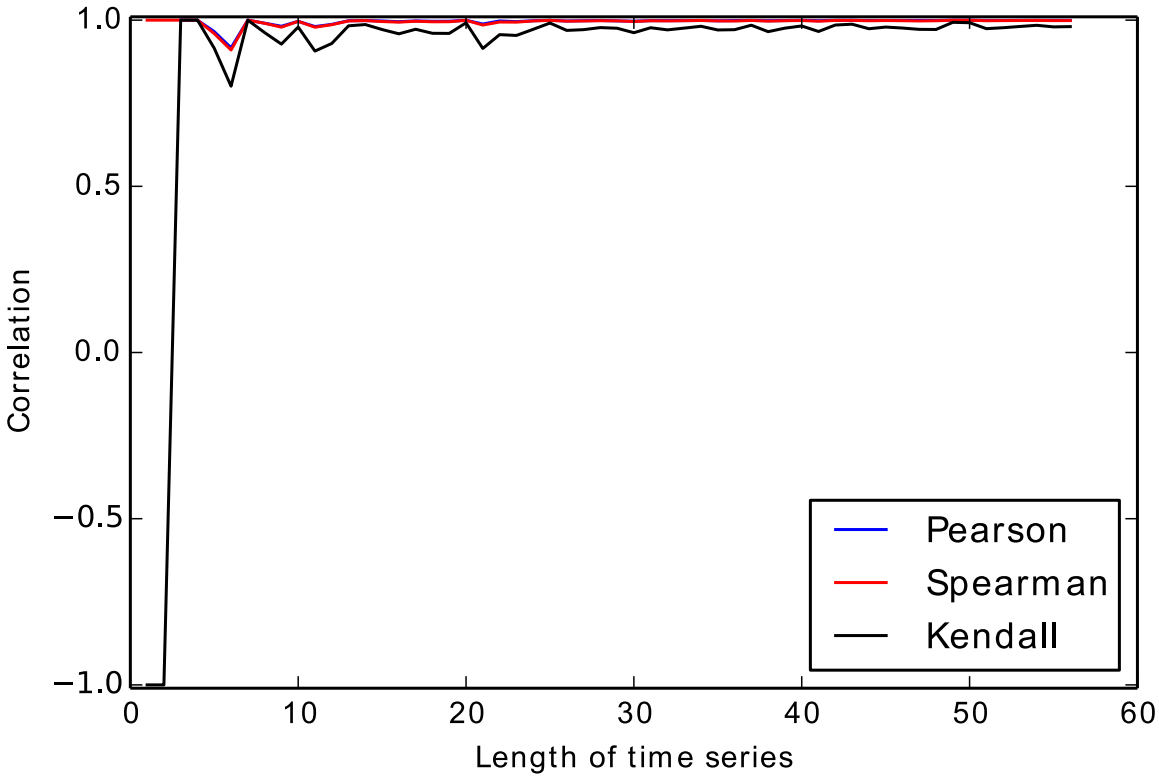

Figure 2.S2: The triangle waveform is highly correlated with the cosine waveform. The correlation between triangle and cosine waveforms are compared for time series of different lengths for three different correlation metrics: Pearson, Spearman, and Kendall. Correlations can range from –1 (completely anti-correlated) to +1 (completely correlated).

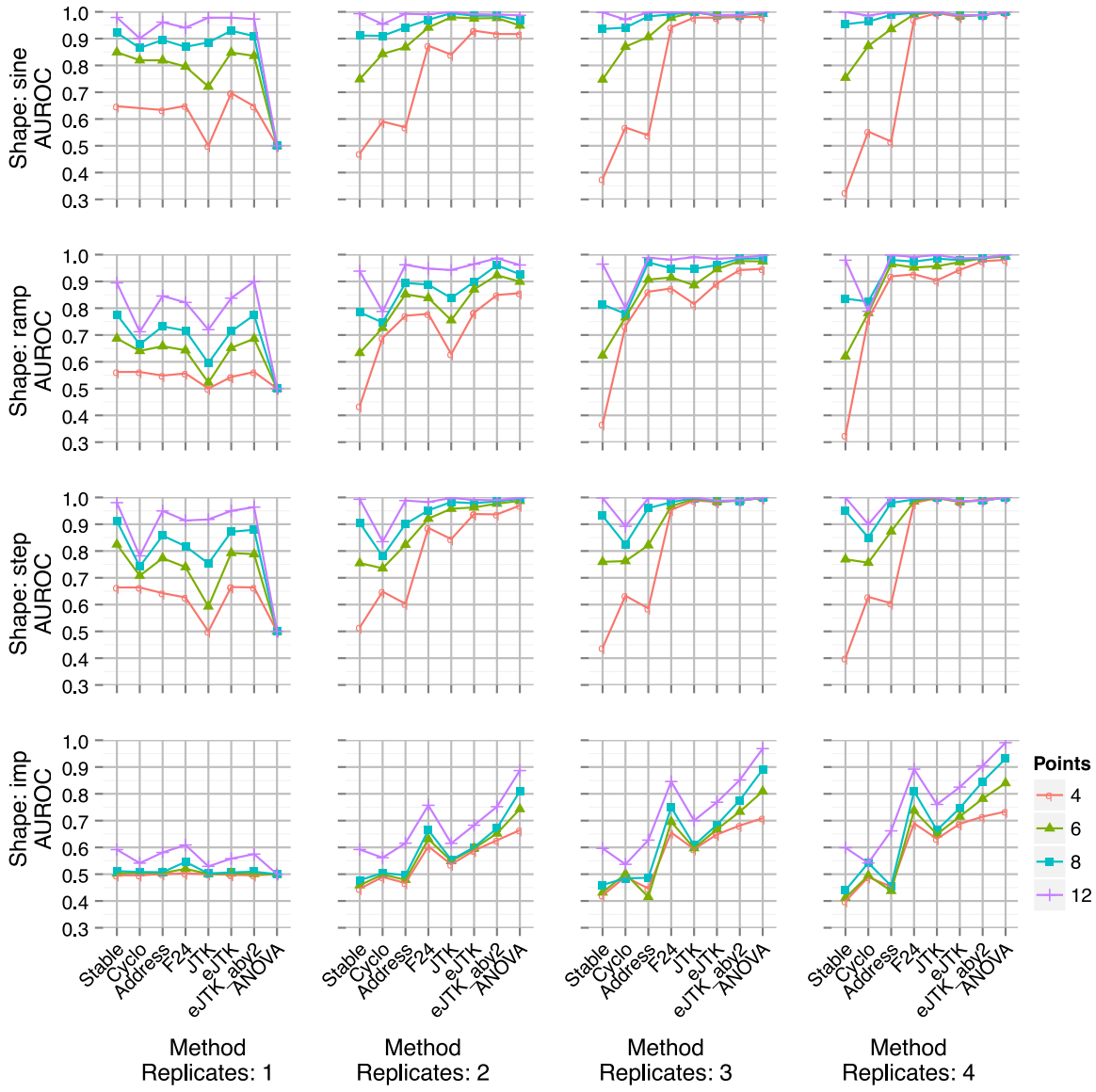

Figure 2.S3: AUROCs for simulated data with 25% noise (standard deviation of Gaussian noise as a percent of amplitude). Layout and abbreviations are the same as in Fig. 2.4.

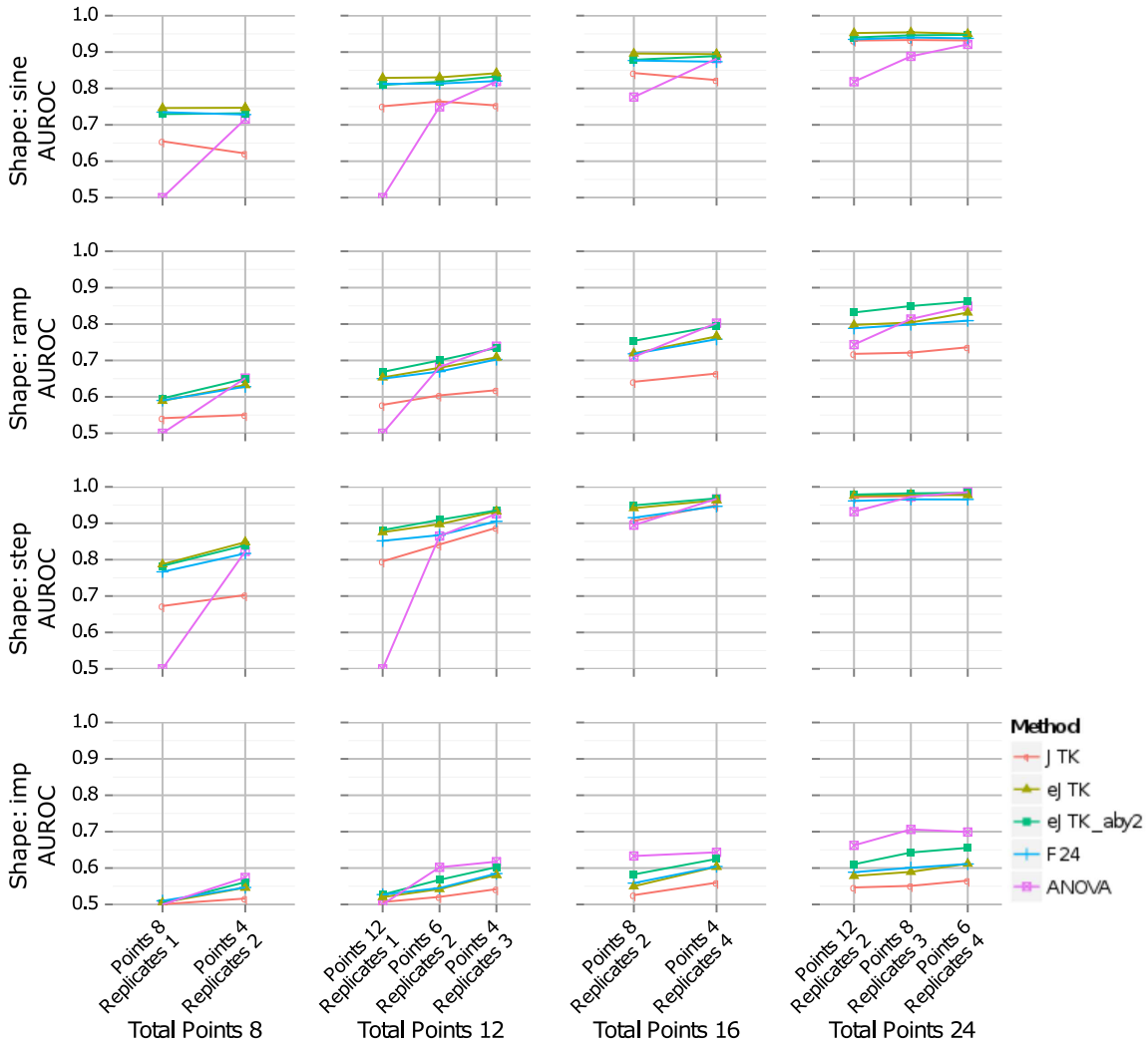

Figure 2.S4: Full set of comparisons used to evaluate the trade-off between increased numbers of replicates and increased densities of time points per period. Layout and abbreviations are the same as in Fig. 2.5.

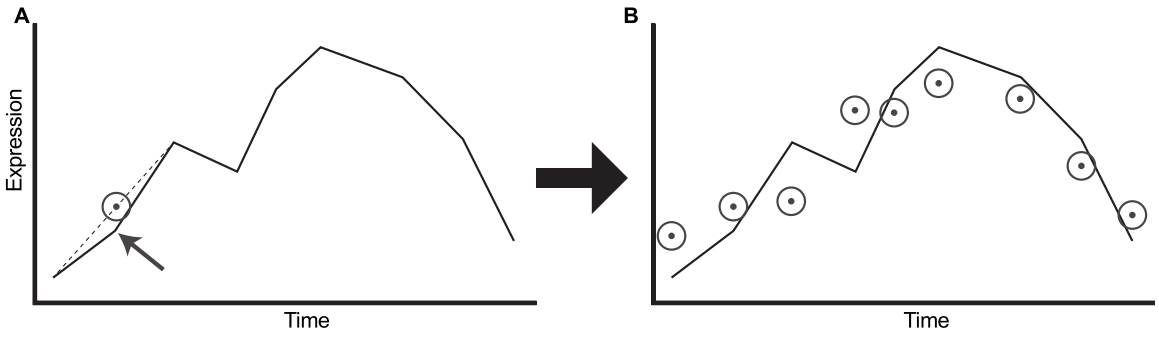

Figure 2.S5: Interpolation scheme for increasing replicate counts. (A) A pseudo-replicate ($\odot$) for time t_i (indicated by the arrow) is obtained by linearly interpolating between time points t_{i-1} and t_{i+1} (dashed line). (B) Repeating this procedure for each time point (modulo 24 h) generates a new time series ($\odot$ symbols).

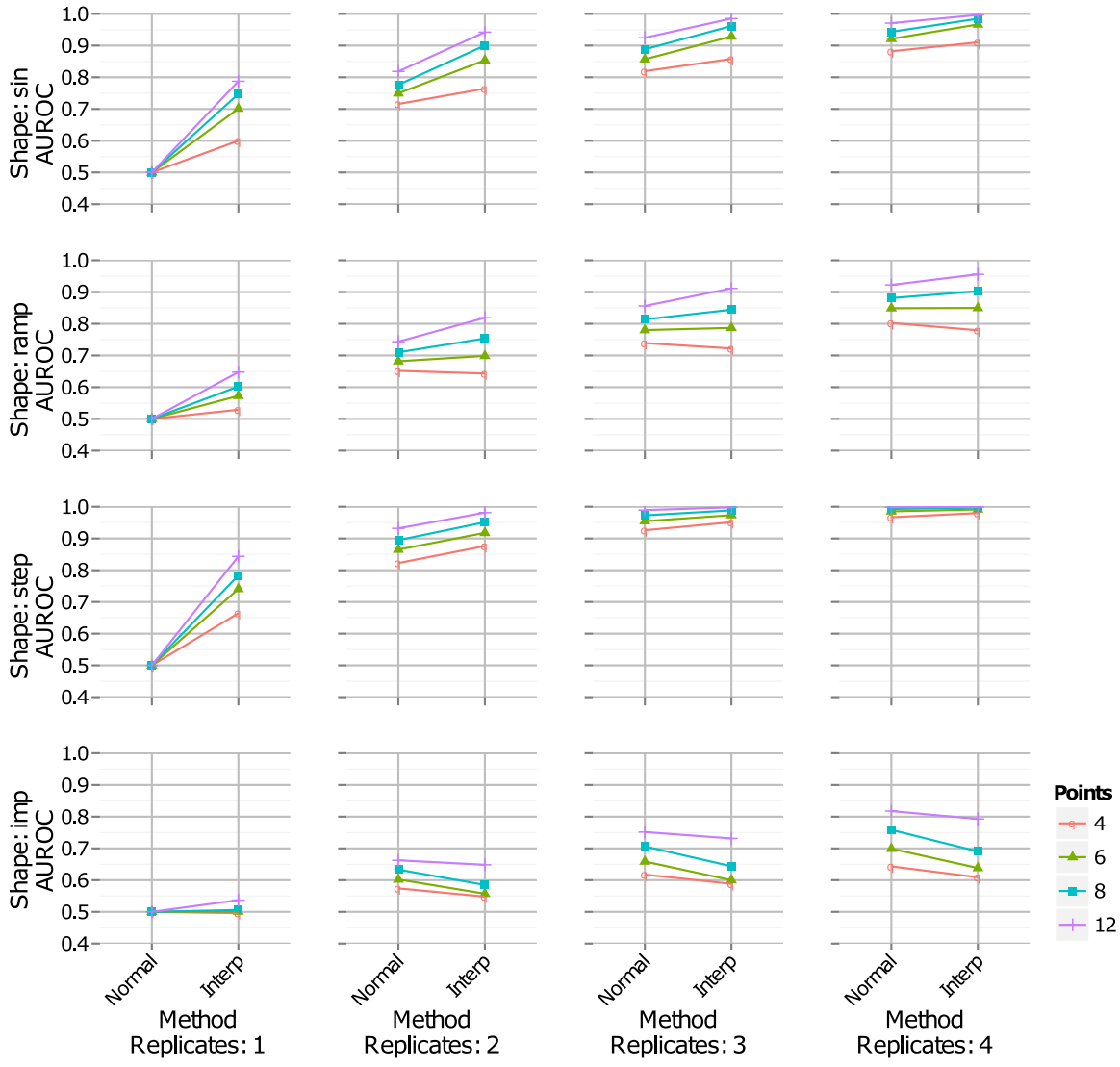

Figure 2.S6: Interpolating the data points to generate pseudo-replicates improves AUROCs when the number of actual replicates is low. We compare performance with (Interp) and without (Normal) pseudo-replicates for the first simulated dataset with 50% noise.

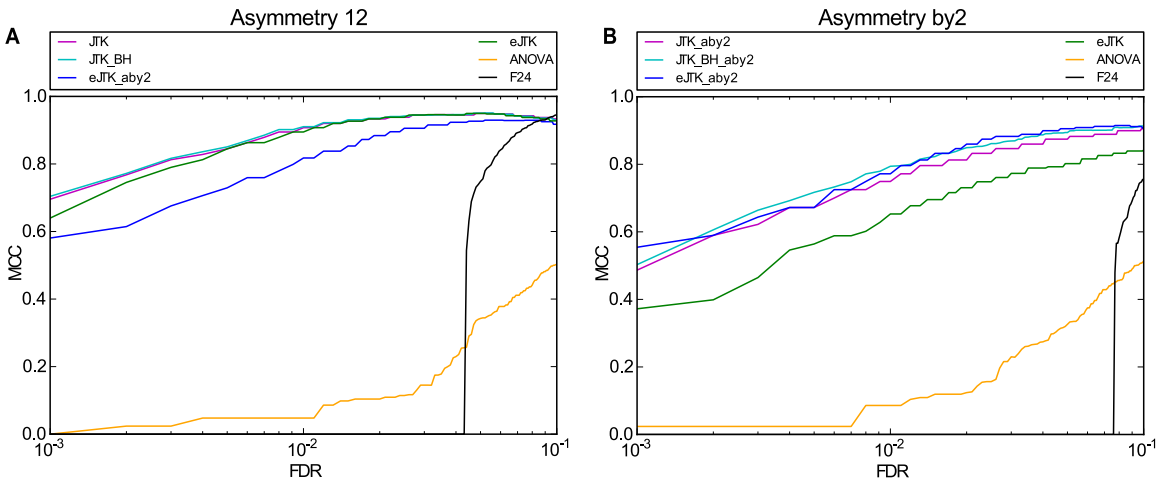

Figure 2.S7: Matthews Correlation Coefficient shows that JTK_CYCLE methods outperform ANOVA and F24 in the presence and absence of asymmetric time series. Simulated data with rhythmic time series without asymmetry (A) or with evenly distributed asymmetry (B) was tested with different methods. The vertical axis shows the Matthews Correlation Coefficients (MCC) [72] for different Benjamini-Hochberg adjusted p-value cutoffs (FDR) along the x-axis. These data are with 25% noise, but the effects of Benjamini-Hochberg correction are significantly greater at 50% noise (not shown). The method abbreviations are the same as those in Fig. 2.4.

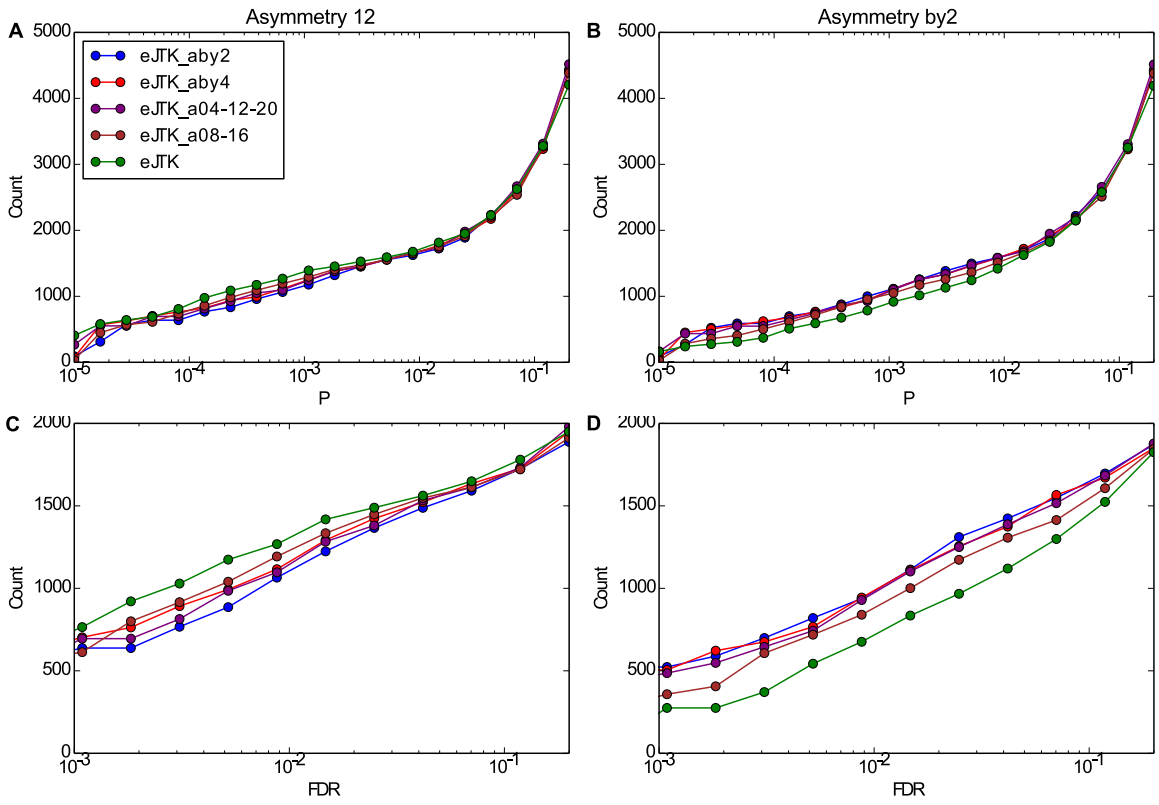

Figure 2.S8: Searching for asymmetric waveforms is detrimental if none are present, but is otherwise advantageous. Simulated data with rhythmic time series without asymmetry (left, A and C) or with evenly distributed asymmetry (right, B and D) was tested with different asymmetries. The cumulative histograms are plotted before (A and B) and after (C and D) Benjamini-Hochberg multiple hypothesis correction across time series. The vertical axis shows the number of genes with a p-value (P) (A and B) or false discovery rate (FDR, the Benjamini-Hochberg adjusted p-value) (C and D) below or equal to a significance threshold, shown on the x-axis. These data are with 25% noise, but the effects of Benjamini-Hochberg correction are significantly greater at 50% noise (not shown). The legend in A applies to B, C, and D as well as A. The rightmost point on the horizontal axis is 0.2. eJTK_aby2: asymmetries sampled every 2 h, from 2 h to 22 h, eJTK_aby4: asymmetries sampled every 4 h, from 4 h to 20 h, eJTK_a04-12-20: asymmetries sampled at 4 h, 12 h and 20 h, eJTK_a08-16: asymmetries sampled at 8 h and 16 h, eJTK: no asymmetry (i.e. asymmetry of 12 h, equivalent to a cosine).

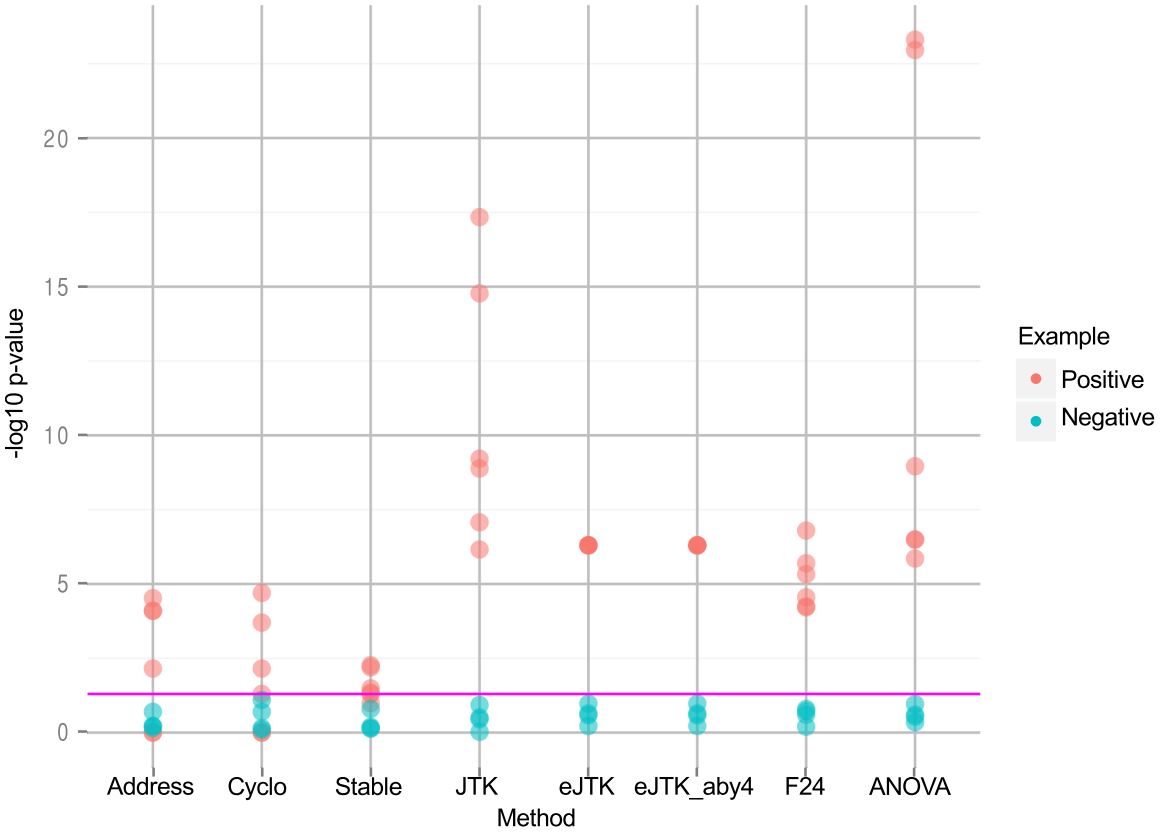

Figure 2.S9: Metadata results for known positive and negative examples. The positive examples are known cycling genes *per*, *tim*, *vri*, *Pdp1*, *cry*, and *Clk*. The negative examples are known non-cycling genes *cam*, *RpL32*, *cyc*, and *dco*. As plotted, large values for the positive examples and small values for the negative examples are desirable. The magenta line marks a p-value of 0.05 ($-\log_{10} 0.05 = 1.3$). Since 2×10^6 permutations were used to generate the empirical JTK_CYCLE p-values, they cannot be lower than 5×10^{-7}. Abbreviations are the same as in Fig. 2.8.

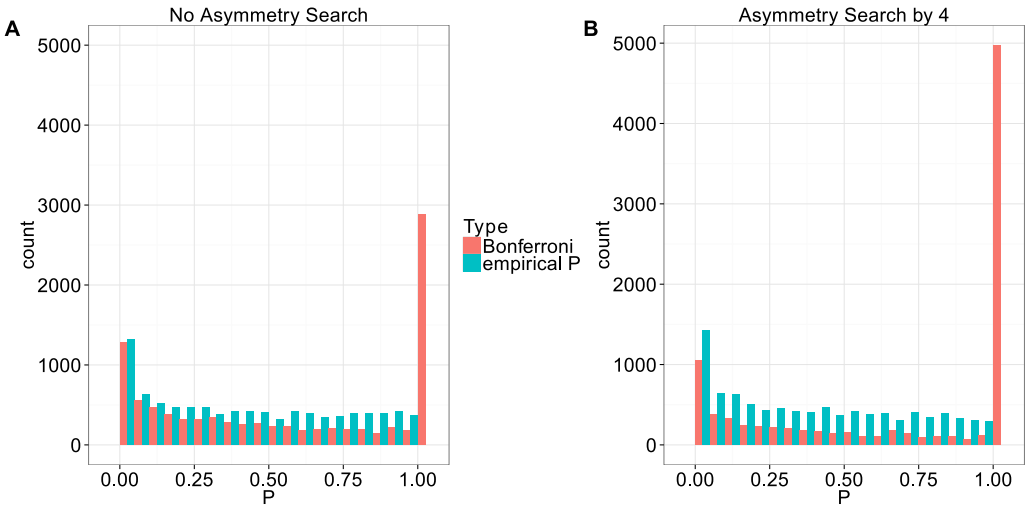

Figure 2.S10: Comparison of the p-value distributions of the original JTK_CYCLE method (with Bonferroni correction) with the empirical JTK_CYCLE method without (A) and with (B) asymmetry search.

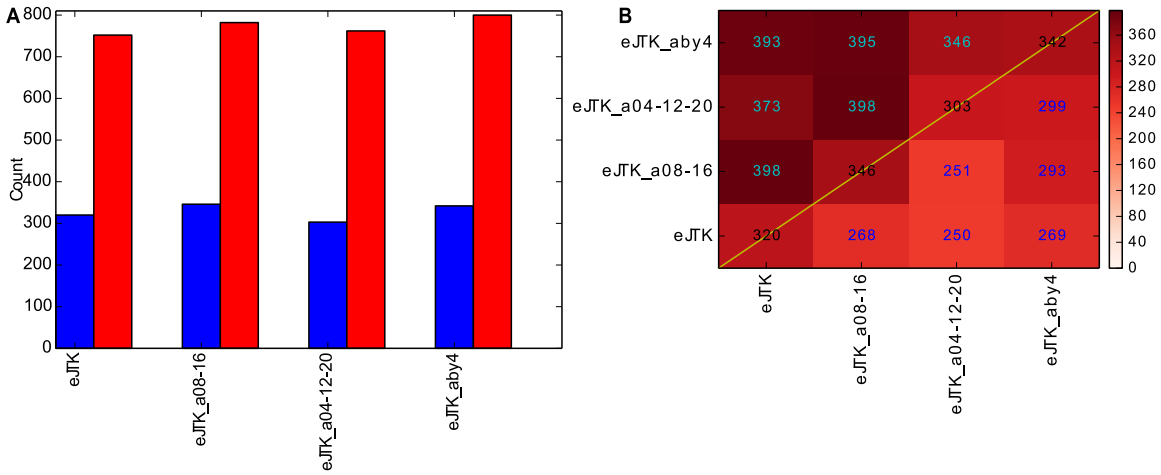

Figure 2.S11: Comparison of the intersection and union of genes identified as rhythmic with Benjamini-Hochberg adjusted p-values less than 0.05 (blue bars) and 0.20 (red bars) for empirical JTK_CYCLE with different asymmetry searches. (A) The number of genes with a Benjamini-Hochberg adjusted p-value (FDR) below 0.05 (blue) and 0.20 (red) are shown. (B) A comparison of the intersection (below the diagonal) and union (above the diagonal) of genes identified as rhythmic with Benjamini-Hochberg adjusted p-values less than 0.05 for the different methods. Abbreviations are the same as in Fig. 2.S8.

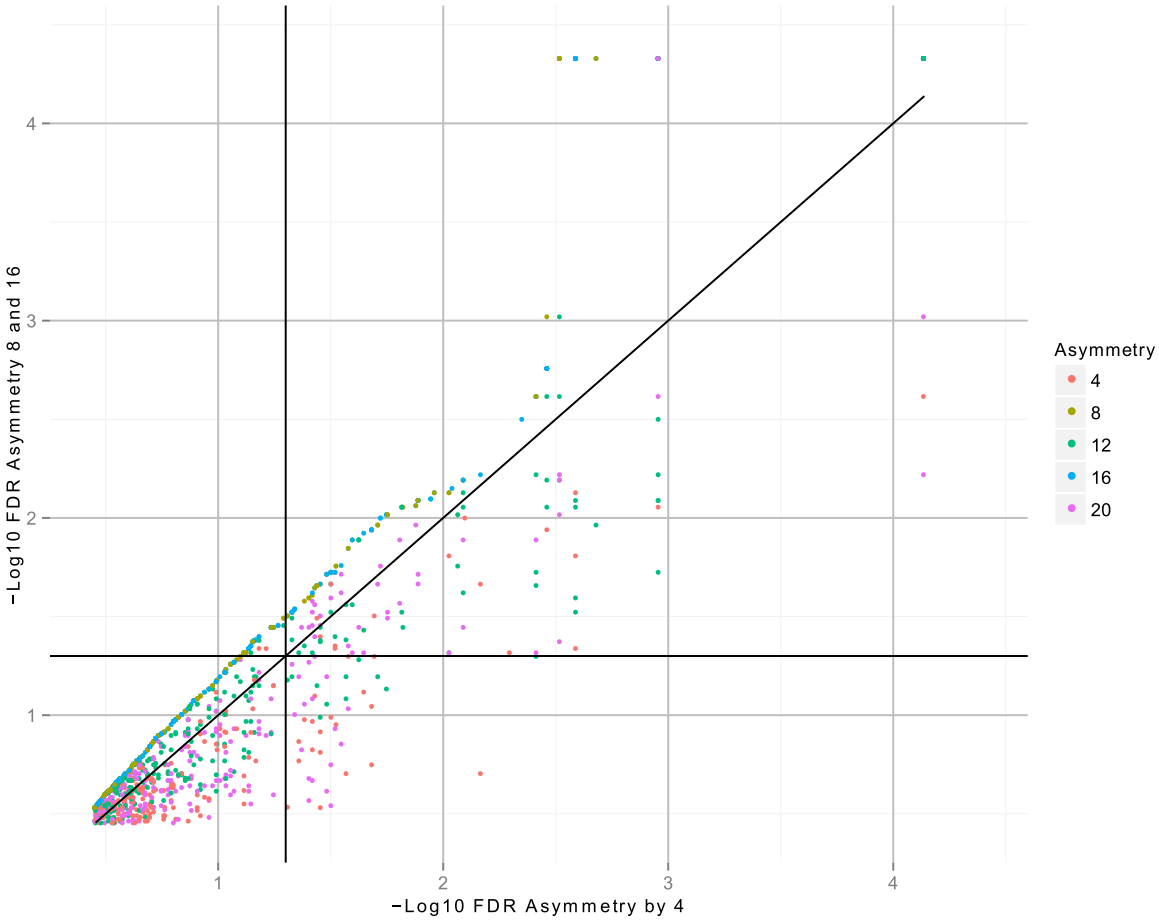

Figure 2.S12: Comparison of JTK_CYCLE asymmetry search depths. Points represent genes, colored by the asymmetry search by 4 h-estimated asymmetries. The black vertical and horizontal lines mark a FDR of 0.05 ($-\log_{10} 0.05 \approx 1.30$). Genes to the the right of the vertical line pass the threshold cutoff for eJTK_aby4, while genes above the horizontal line pass the threshold cutoff for eJTK with asymmetry search of 8 and 16 h. Genes that are above the horizontal line but left of the vertical line barely pass the threshold and have asymmetries in the range of 8 to 16 h. Genes that are right of the vertical line but below the horizontal line pass the threshold much more significantly than the previously mentioned genes and have asymmetries that are more extreme.

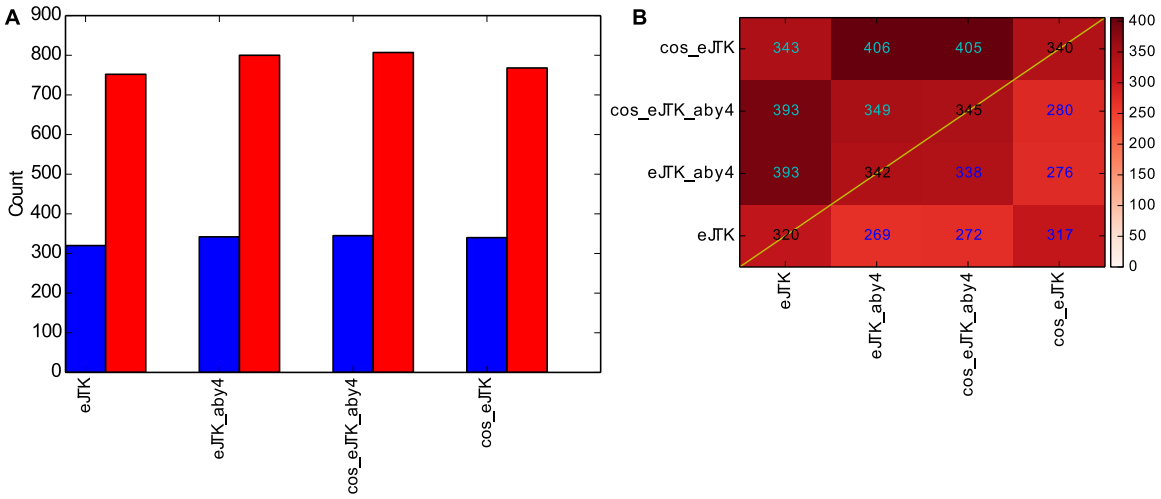

Figure 2.S13: Using a cosine as a reference waveform instead of a triangle does not produce substantially different results in genes identified as cycling. A comparison of the intersection and union of genes identified as rhythmic with Benjamini-Hochberg adjusted p-values less than 0.05 (blue bars) or 0.20 (red bars) for empirical JTK_CYCLE without asymmetry (eJTK), and empirical JTK_CYCLE with asymmetry search of 4, 8, 12, 16, and 20 h (eJTK_aby4) calculated with a reference waveform of a triangle (no prefix) or with a reference waveform of a cosine (prefix "cos"). (A) The number of genes with a Benjamini-Hochberg adjusted p-value below 0.05 (blue) and 0.20 (red) are shown. (B) A comparison of the intersection (below the diagonal) and union (above the diagonal) of genes identified as rhythmic with Benjamini-Hochberg adjusted p-values less than 0.05 for the different methods.

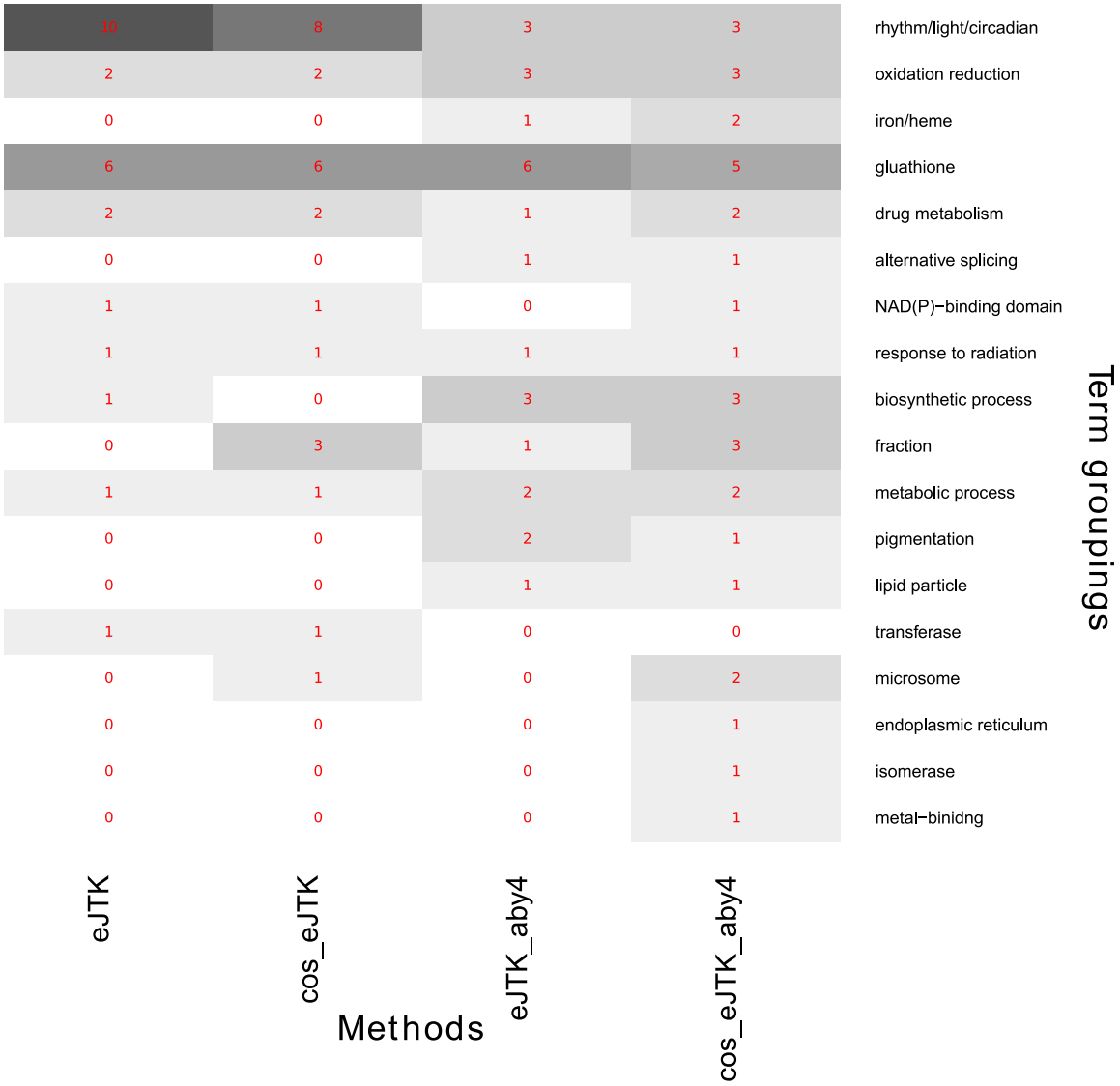

Figure 2.S14: Using a cosine as a reference waveform instead of a triangle does not produce substantially different results in annotation terms enriched for in genes identified as cycling. Annotation terms identified as enriched by DAVID share many similarities and were therefore grouped. The number of annotation terms enriched in the genes discovered by each method are shown in grey shading and red numbers. Empirical JTK_CYCLE methods with and without asymmetry search ("_aby4" and no suffix, respectively) and with a triangle as a reference waveform or cosine as a reference waveform (no prefix or "cos", respectively). The annotation terms displayed are enriched with Benjamini-Hochberg adjusted p-values below 0.05.

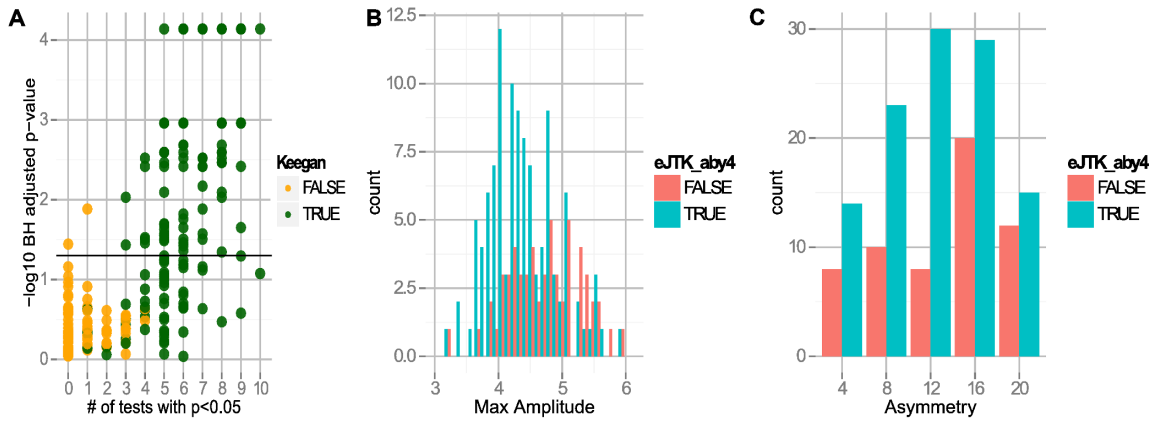

Figure 2.S15: Comparison of genes identified as cycling by Keegan *et al.* and empirical JTK_CYCLE with asymmetry search of 4 h (eJTK_by4). (A) All the genes shown passed the ANOVA pre-screen, but only the green ones are identified as cycling by Keegan *et al.* [59]. Higher negative logarithms of p-values are more significant than lower ones: the horizontal black line indicates a Benjamini-Hochberg adjusted p-value for eJTK_aby4 of 0.05. (B) All the genes shown were identified as cycling by Keegan *et al.* The mean and variance of the genes identified as cycling by Keegan *et al.* and eJTK_aby4 (blue), are 4.34 and 0.54, respectively. The mean and variance of the genes identified as cycling by Keegan *et al.* and but not eJTK_aby4 (red), are 4.75 and 0.56, respectively. (C) All the genes shown were identified as cycling by Keegan *et al.* The asymmetry of the genes was determined by eJTK_aby4.

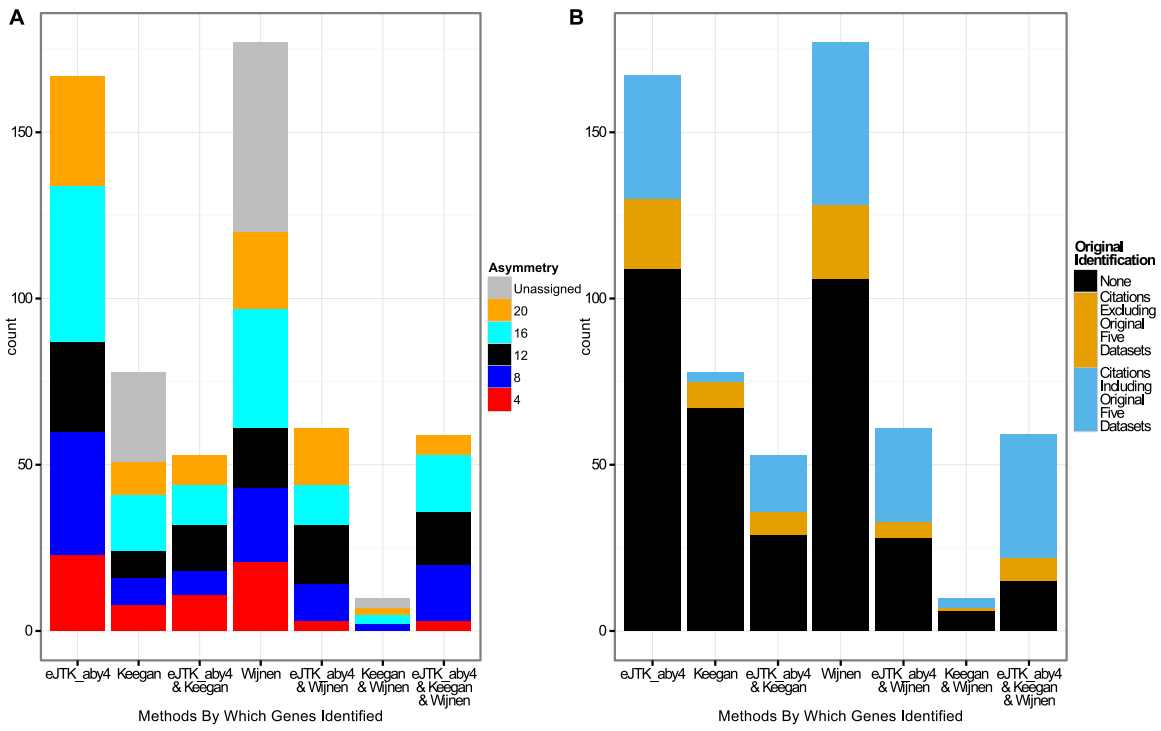

Figure 2.S16: Comparison of genes identified as cycling by Keegan *et al.*, Wijnen *et al.*, and empirical JTK_CYCLE with asymmetry search by 4 h (eJTK_by4). (A) Comparison of genes identified as rhythmic by Keegan *et al.* [59], Wijnen *et al.* [110], and eJTK_by4. Stacked bars are colored to represent the asymmetry, as determined by eJTK_aby4. eJTK_aby4 identifies more genes with non-12 h asymmetries than the other methods. "Unassigned" refers to genes that were excluded from the empirical JTK_CYCLE analysis. (B) For each gene, the references on FlyBase that mention the gene were identified. The genes identified by eJTK_by4, Keegan *et al.*, and Wijnen *et al.* are shown in a histogram with stacked bars colored to represent the genes being cited by references with "circadian" in the title or abstract, genes cited in the original five dataset papers, or neither. While there are more genes uniquely identified by Wijnen *et al.*, there are more total genes identified by eJTK_by4, as well as more genes that are cited in papers that have "circadian" in their title or abstract.

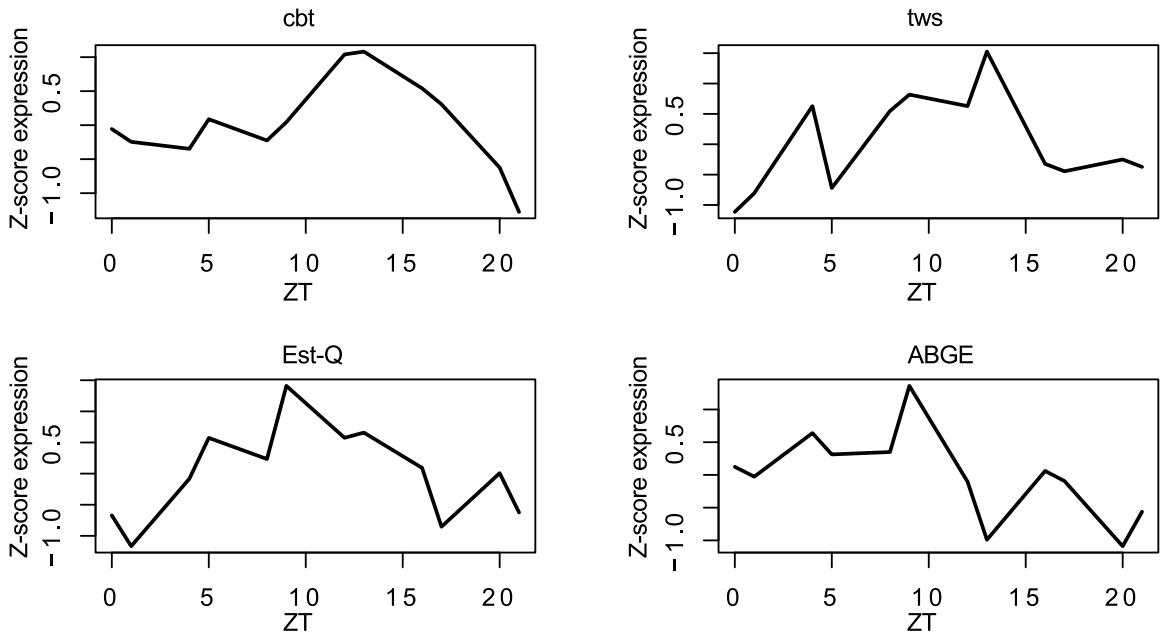

Figure 2.S17: Z-score expression time series of *cbt*, *tws*, *Est-Q*, and *ABGE* averaged across replicate time points.

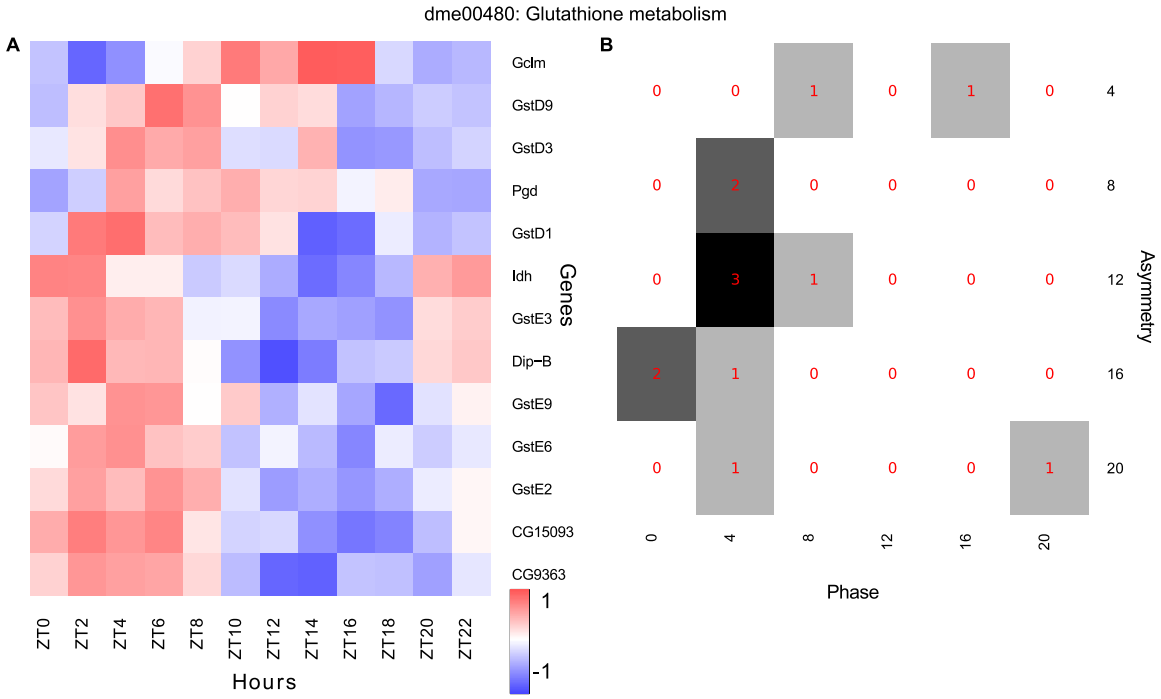

Figure 2.S18: KEGG pathway "dme00480: Glutathione metabolism" is enriched in genes identified as rhythmic by eJTK_aby4. Peak expression (phase) of these genes is mainly in the light period. (A) Z-scored gene expression of genes from the metadataset involved in glutathione metabolism averaged across 24 h and interpolated to every 2 h. (B) Phase and asymmetry distribution of the genes from the metadataset involved in glutathione metabolism.

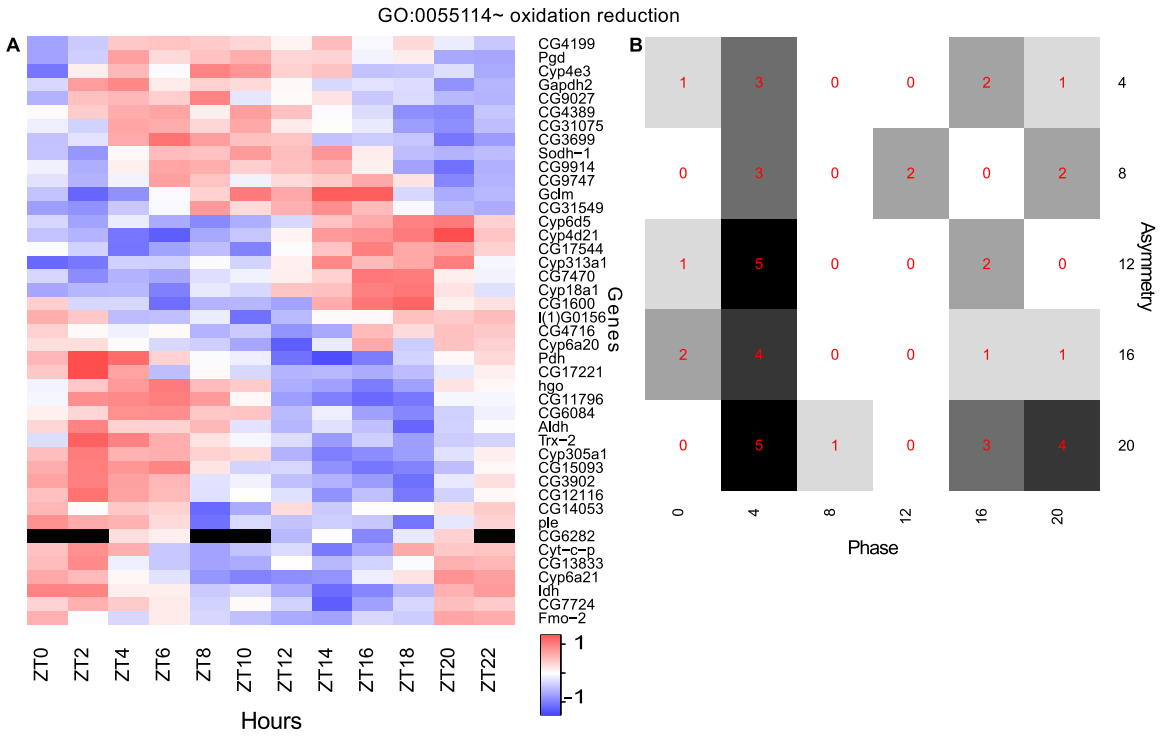

Figure 2.S19: Gene ontology "GO:0055114 oxidation reduction" is enriched in genes identified as rhythmic by eJTK_aby4. Peak expression (phase) of these genes is distributed over 24 h. (A) Z-scored gene expression of genes from the metadataset involved in oxidation reduction averaged across 24 h and interpolated to every 2 h. Black indicates time points where data were not available (NA). (B) Phase and asymmetry distribution of the genes from the metadataset involved in oxidation reduction.

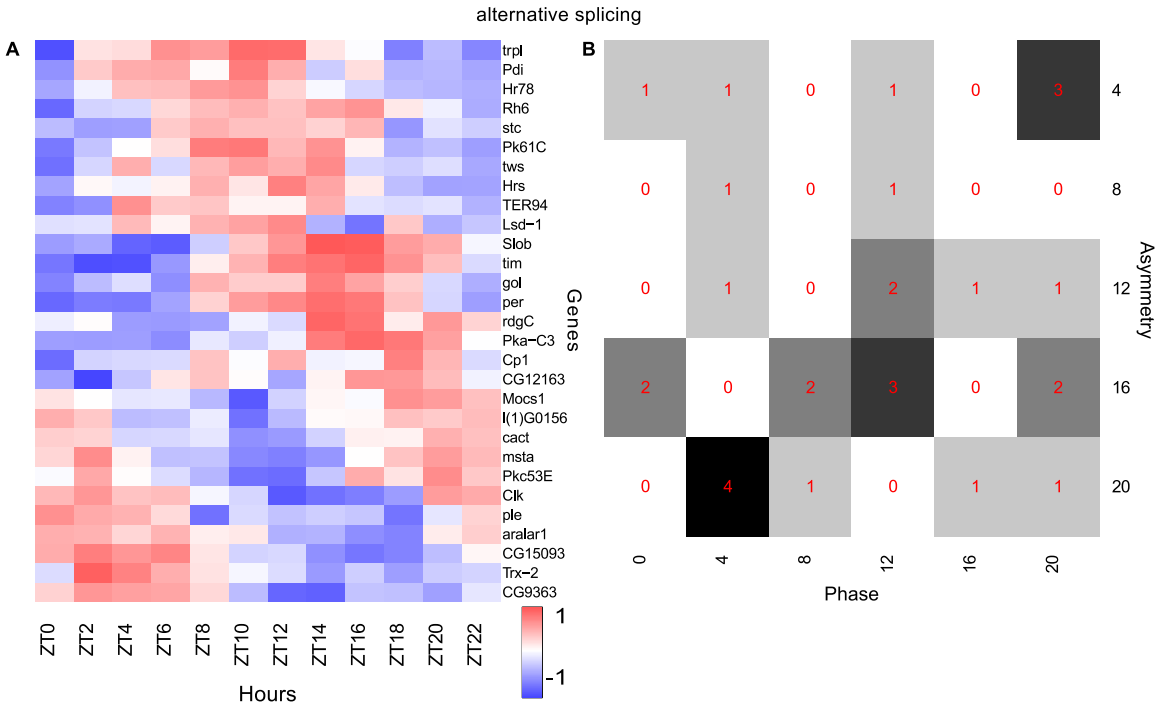

Figure 2.S20: PIR keyword "alternative splicing" is enriched in genes identified as rhythmic by eJTK_aby4. Peak expression (phase) of these genes is distributed over 24 h. (A) Z-scored gene expression of genes from the metadataset involved in alternative splicing averaged across 24 h and interpolated to every 2 h. (B) Phase and asymmetry distribution of the genes from the metadataset involved in alternative splicing.

CHAPTER 3

CORRECTING FOR DEPENDENT P-VALUES IMPROVES ACCURACY OF LEADING RHYTHM DETECTION METHODS

3.1 Abstract

There is much interest in using genome-wide expression time series to identify circadian genes. Several methods have been developed to detect 24 hour rhythmicity in these data, which would be suggestive of circadian regulation. Two recent examples are RAIN [99] and MetaCycle [113]. While these methods are powerful, they can be improved by more accurately calculating their p-values. Here, we identify and suggest solutions the common pitfalls that result in inaccurate p-values. We use as our benchmark the requirement of uniformity of p-values under the null hypothesis, and detail how to ensure this when it is not it not initially the case. Both these methods, though they take divergent approaches, make incorrect assumptions of mutual p-value independence. We discuss how to adjust these methods to incorporate the p-value dependence and improve their accuracy.

3.2 Introduction

Periodic patterns (rhythms) are pervasive in biology at molecular, cellular, organismal, and ecological scales. Many statistical methods have been developed to identify genes with cycling expression patterns from genome-wide time series [113, 53, 114, 99, 22, 117, 52, 59]. The improvements in rhythm detection methods recently has done much to allow the field of circadian biology to take full advantage of the high-throughput data regarding rhythms at the tissue and single-cell level. These methods could be further improved, however, by making adjustments to improve the accuracy of their p-values, which is the main metric

of rhythmicity which is used for downstream analysis. We present two ways to improve the accuracy of p-values for two different methods. The first uses an empirical approach to calculating p-values which ensures that the p-values are uniform under the null hypothesis, a definitive aspect of accurate p-values. The second addresses recent methodological advances which assume the mutual independence of p-values in situations in which they are actually dependent on one another. One method, RAIN [99], uses a variety of reference waveforms for rhythm detection, but incorrectly assumes that p-values from different reference waveforms correlated against an experimental time series are independent: this results in p-values which are under-estimates for very low p-values and over-estimates for very high p-values. The second method, MetaCycle [113] combines the power of different rhythm detection methods into a single method, but incorrectly assumes that p-values from different methods run on the same experimental times series are independent: this results in p-values which are under-estimates of their true values.

We show that both these methods can be improved by realizing that the p-values being manipulated in both cases are not independent. RAIN and the underlying methods behind MetaCycle can be improved by empirically calculating p-values, and MetaCycle can be further improved by using a p-value integration approach that incorporates the dependence of the p-values.

3.3 Methods

3.3.1 Empirically calculating p-values from simulated data

A p-value provides the probability that a test statistic is observed in the case that the null hypothesis is true. For rhythm detection methods, the null hypothesis is generally that a time series consists of independent draws from a normal distribution. The test statistic is the scalar output of some test, such as a correlation or R^2 of best fit. For the purposes of

this discussion we will work within this rhythm detection framework and we will treat test statistics as being one-sided: higher test statistics indicate greater rhythmicity than lower test statistics, P-values are often derived analytically for various test statistics, they can also be empirically derived from simulated data. Running a rhythm detection method on null hypothesis-simulated time series (independent draws from a normal distribution) will produce a distribution of test statistics, which we will refer to as the null distribution. Using Eq. 3.1, we can compare a chosen test statistic (from running a method on a time series, be it experimental or simulated) to the null distribution to obtain a p-value for that test statistic, for a null distribution T_{null} of size N.

$$p = \frac{1 + \sum_{i=1}^{N} (T_{null,i} \geq T_{obs})}{1 + N} \tag{3.1}$$

Using the null distribution test statistics as the observed test statistics, we can generate a p-value for each test statistic. Assuming a null distribution of size 10 where we remove each test statistic and compare it to the remaining $N = 9$ and no ties in the test statistics, the resulting p-values will be distributed evenly from 0.1 to 1. For example, the largest test statistic will be larger than the other 9 values (0 test statistics will be larger than it), resulting in a p-value of 0.1. The second largest test statistic will have 1 value larger than it, resulting in a p-value of 0.2. The smallest test statistic will have 9 values larger than it, resulting in a p-values of 1. This even distribution of p-values from the null distribution is referred to as the uniformity of p-values under the null hypothesis, and is by definition true of correct p-values. This means that for any rhythm detection method we use, if we run it on data generated from independent draws from a normal distribution, the resulting p-values should be evenly spaced and therefore uniform. P-values generated from this procedure that are not uniform are not accurate.

Though we have introduced this topic with general application to rhythm detection meth-

ods, this approach has been taken in the recent literature by us, in the development of our empirical JTK_CYCLE method (eJTK) [53], where we correct the over-conservative p-values of the original JTK_CYCLE method [52] via simulation of the null distribution, which allows us to empirically calculate accurate p-values. We will demonstrate the generality of this approach as we address improving methods in which p-values are treated as independent when it would be more accurate to treat them as dependent.

3.4 Results

3.4.1 Dependence of p-values from comparing reference waveforms with experimental time series

JTK_CYCLE [52], RAIN [99], and eJTK [53] all use reference waveforms with different phases (and in RAIN and eJTK's case, different peak to trough and trough to peak times) to find the reference waveform that best matches the experimental waveform, and use a comparison between the reference waveforms and the time series (Kendall's τ for JTK_CYCLE and eJTK, Mann-Whitney U for RAIN) to obtain p-values. The three methods, however, take different approaches to obtaining a single p-value for the time series from the many p-values obtained for each reference waveform comparison. JTK_CYCLE applies the Bonferroni correction, which multiplies the best-matching p-value by the number of comparisons, producing an adjusted p-value that represents the Family-Wide Error Rate: a value that indicates the probability of one false positive existing at that threshold. This results in p-values that are overly conservative, and not accurate for correction across genes or probes, as is normally done with the Benjamini-Hochberg FDR [13] or q-value approach, [96], as is described in [53].

RAIN takes a less conservative approach by using the Benjamini-Hochberg method to adjust the p-values from the different reference waveforms, then selecting the smallest ad-

justed p-value. This correction produces p-values that are closer to uniform under the null distribution. However, the Benjamini-Hochberg method is designed specifically for p-values that are independent from one another. Since the reference waveforms that are compared to the experimental time series are all correlated with one another, the application of the Benjamini-Hochberg method should result in inaccurate p-values. To test this, using the criteria described above, we ran RAIN on simulated data generated from independent draws from the normal distribution (i.e., the null hypothesis). The resulting p-values were non-uniformly distributed, and were in fact under-estimates for low p-values and over-estimates for high p-values (Figs. 3.4A) and B). Given that low p-values are used to detect rhythmic genes worthy of further study, under-estimating these values will increase the rate of false positives, the incorrect identification of genes as rhythmic. Using the method detailed above, we can make the RAIN p-values accurate. Using 10^6 simulated time series from Gaussian noise as a reference, we obtained the raw p-values from the RAIN algorithm in R and corrected these p-values using the procedure outlined in Hutchison *et al.* [53].

3.4.2 Dependence of p-values from comparing different rhythm detection results for the same time series

Given that there are a variety of rhythm detection methods available that focus on different aspects of the experimental time series, it would be advantageous to be able to combine the strengths of these different methods. One approach has been taken by Wu *et al.* [113], where they use Fisher Integration [31] to integrate the p-values from JTK_CYCLE [52], Lomb-Scargle [67], and ARSER [117].

The Fisher Integration method was developed for combining several independent tests with the same null hypothesis. Unfortunately, the p-values being combined in this case are not independent. Since all the methods under consideration about being performed on the same time series, and they have some overlap in the characteristics that they detect. For

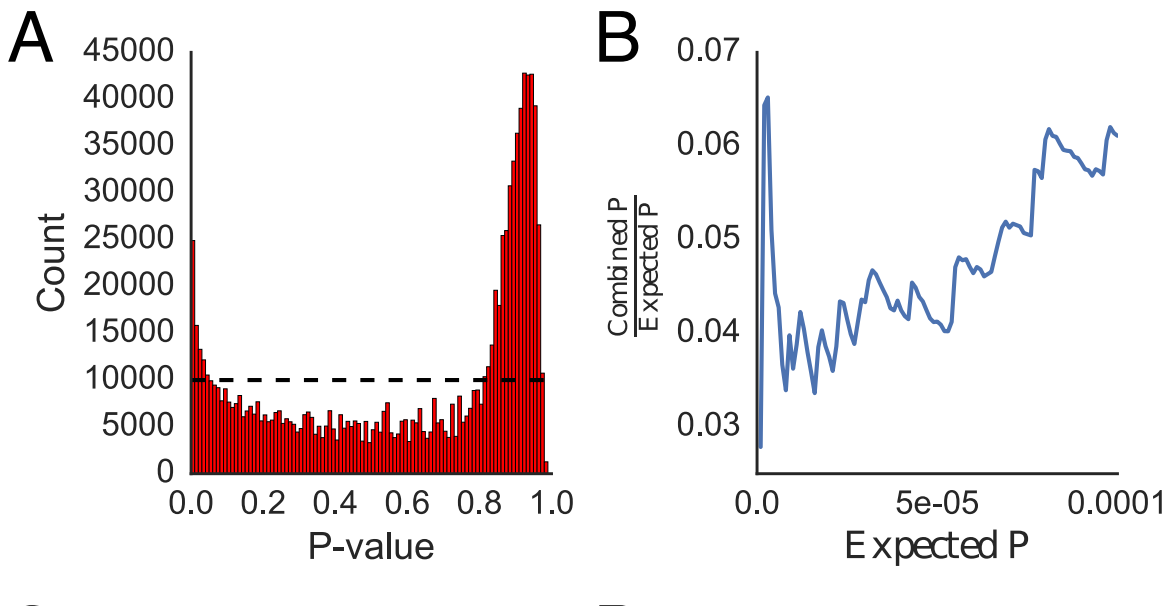

Figure 3.1: RAIN does not produce p-values that have a uniform distribution under the null hypothesis. (A) The p-values produced by RAIN run on 10^6 Gaussian noise time series (the null distribution) are not uniform (the dashed line). Instead, low p-values are smaller than expected and high p-values are larger than expected. (B) For p-values below 10^{-4}, the RAIN p-values from the Gaussian noise time series in (A) are underestimates by between 1/25 and 1/16 of their expected values.

example, a perfect cosine wave will be identified as strongly rhythmic by all three methods. This means that the p-values are correlated from the different methods are not actually independent of one another. Applying Fisher Integration to these p-values will result in under-estimates of low p-values and over-estimates of high p-values.

We can see this by comparing the effect of Fisher method on three sets of p-values generated in two different ways. In the first case, the 3 sets of 1000 p-values are generated randomly drawing from a uniform distribution from 0 to 1. As these p-values are independent, using Fisher method (dashed line, Fig. 3.2A) produces p-values which are uniformly distributed from 0 to 1. Fig. 3.2A compares the expected p-value (which we know since we are working under the null hypothesis) against the ratio of the observed p-value to the expected p-value, which should be 1 and are for Fisher Integration of these p-values in Fig. 3.2A. In the second case, the 3 sets of 1000 p-values are each identical, making the correlation between the sets 1. Using Fisher method in this case (dashed line, Fig. 3.2B) results in p-values whose observed:expected ratio is far below 1, indicating p-values that underestimate the true p-values. Therefore, while the Fisher Integration method works in cases where p-values are independent, when p-values are dependent it will not produce accurate p-values.

To correct for this p-value dependence, we adopt the Brown method, a modification of the Fisher method which applies it to non-independent, one-sided tests of significance [16]. The Brown method uses the covariance between the different p-value sets to adjust the χ^2 distribution and produce p-values that account for any non-independence between p-values. In the case where the p-values are independent, it provides the same results as the Fisher method (Fig. 3.2A). When the p-values are non-independent, however, and the Fisher method does not produce accurate p-values, the Brown method produces accurate integrated p-values that are uniform from 0 to 1 (Fig. 3.2B).

Having identified a preferable method for integrating p-values, we can observe the difference between the Fisher method and the Brown method for integrating the p-values for

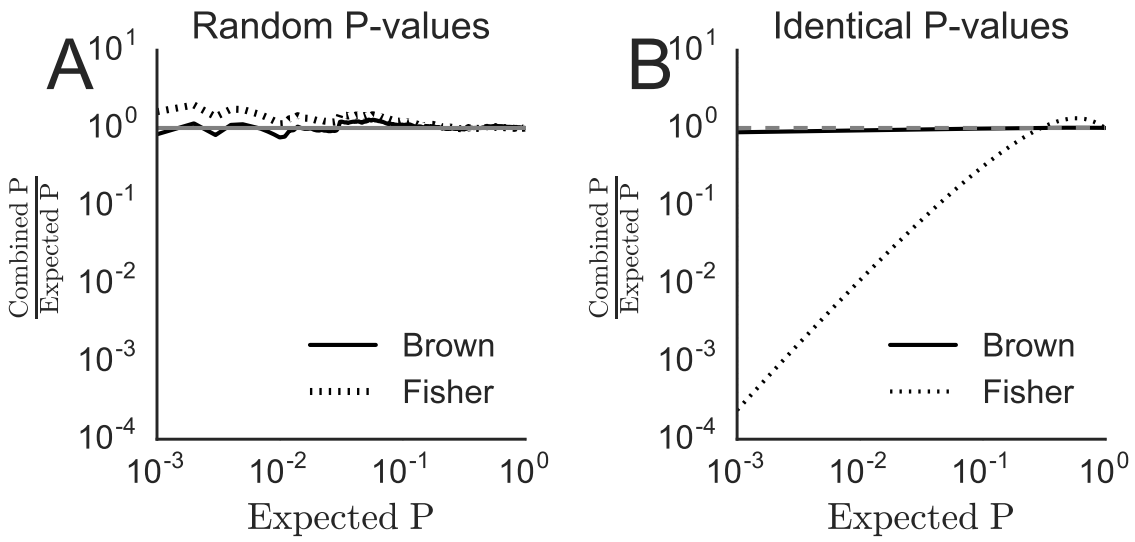

Figure 3.2: The p-values from the Brown correction of Fisher integration recapitulate the uniform distribution under the null regardless of the dependency of the p-values. (A) P-values from integration of three sets of randomly generated p-values, combined using the Fisher method (dashed line) or the Brown method (solid line). (B) P-values from the integration of three sets of identical evenly distributed p-values, combined using the Fisher method (dashed line) or the Brown method (solid line). The horizontal grey line in (A) and (B) indicates a ratio between combined p-value and expected p-value of 1.

the different rhythm detection methods. We generated 1000 time series drawn from the normal distribution and ran MetaCycle on them. We ensured that none of the data had replicate time points, causing the algorithm to run Lomb-Scargle (LS), ARSER (ARS), and JTK_CYCLE (JTK) on the data and then integrate the resulting p-values (Meta2D). Observing the resulting p-values, we saw that in addition to the p-values produced by JTK being overly conservative, as previously discussed [53], that the p-values from LS were also overly conservative (Fig. 3.3A). The p-values from ARSER, however, match the uniform distribution, accounting for some deviation due to the nature of using randomly generated data. The resulting integration using the Fisher method (Meta2D), is likewise non-normal, with underestimates of the p-vales for low values and overestimates for high values (Fig. 3.3A).

To make these data more accurate, we took the same approach as empirical JTK_CYCLE (eJTK) [53], which we discussed in our introduction of p-values above. We ran MetaCycle on 1,000,000 randomly generated time series and used those results to adjust the p-values from the 1000 time series for the four different methods (LS, JTK, ARSER, and Meta2D), generating p-value distributions that were uniform while accounting for some deviation due to the nature of using randomly generated data (Fig. 3.3B). One deviation from this uniformity was for the case of JTK, where many of the p-values produced by MetaCycle were 1. This is due to the ceiling on the Bonferroni correction, where p-values that may have had initially different values were all corrected to 1. This caused a deviation in the p-value adjustment, as no difference could be discerned among these p-values. This deviation can be seen as a spike in the JTK curve in Fig. 3.3.

Having discussed how to adjust the methods from MetaCycle to produce accurate p-values and how to use Brown's method for to generate accurate integrated p-values, we now explore the effects of the Fisher method and the Brown method on combining different methods.

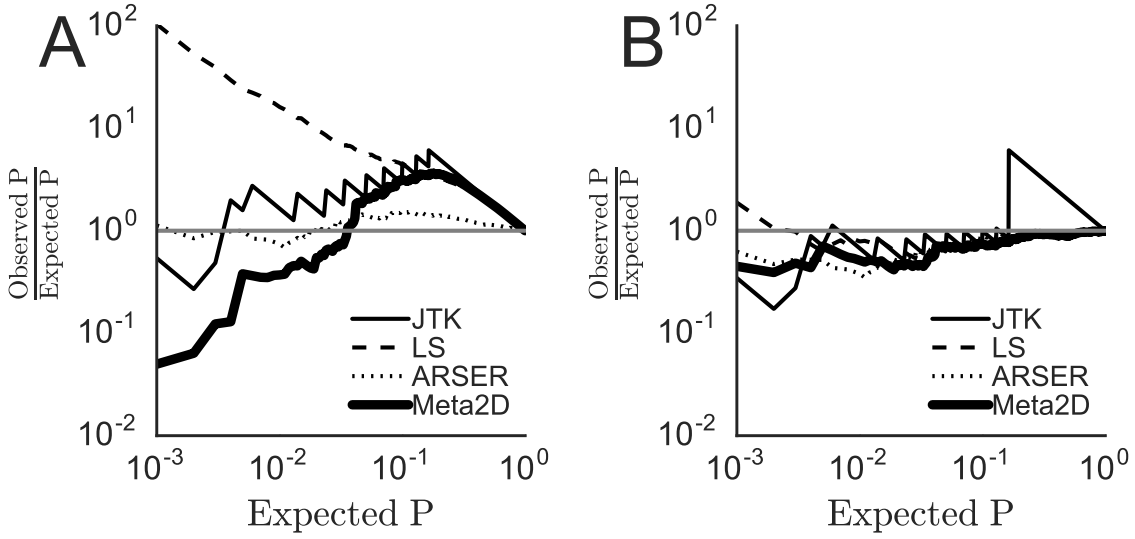

Figure 3.3: MetaCycle was run on 1000 time series generated randomly from a normal distribution, producing p-values for JTK_CYCLE (JTK), Lomb-Scargle (LS), ARSER, and the Fisher method integration of all three of them, Meta2D. (A) The p-values generated by the methods from MetaCycle, with the exception of those from ARSER, deviate from the uniform distribution. (B) Using the empirical method detailed in the methods and in [53], we corrected the p-values of the methods from MetaCycle such that they more closely match the uniform distribution. JTK_CYCLE sets many p-values equal to 1, leading to the deviation from 1 for high p-values as the empirical method cannot uniformly redistribute those p-values.

Given that the correction that the Brown method incorporates accounts for correlation between the methods, we expected that when more dissimilar methods were combined using the Fisher method the error would be less than when combing two methods that were very similar. In application to the adjusted methods from MetaCycle, we find that all three methods are roughly equal in their correlation, producing low p-values roughly 1/100th to 1/1000th of the accurate p-values (Fig. 3.4). When all three methods are combined using the Fisher method, the resulting p-values are 1/10,000th of the accurate p-values. When using the Brown method, these underestimates, which will result in false positives, disappear, and the resulting p-values are accurate.

In addition to looking at the methods from MetaCycle, we also chose to compare two other methods, eJTK [53] and ANOVA [59], to ARSER [117], which is also used by MetaCycle. Using the Fisher method, we find that the underestimates as not as extreme, perhaps because ANOVA is less similar to ARSER and eJTK. When the Brown method is used, these underestimates likewise are reduced to deviations to be expected of data generated from noise.

3.5 Discussion

In this paper, we have identified some two pitfalls when working with p-values and discussed methods by which that can be corrected. Both of these regard making corrections to and combining p-values which are not independent.

In the original JTK_CYCLE method [52] the multiple hypothesis testing across multiple waveforms is corrected with the Bonferroni method, a method which results in overly conservative p-values that control for the family-wide error rate. This method uses no information about the other p-values being corrected. The RAIN method [99] uses the Benjamini-Hochberg method to correct the p-values across multiple waveforms, controlling for the false discovery rate. It integrates information about the other p-values by taking their order into

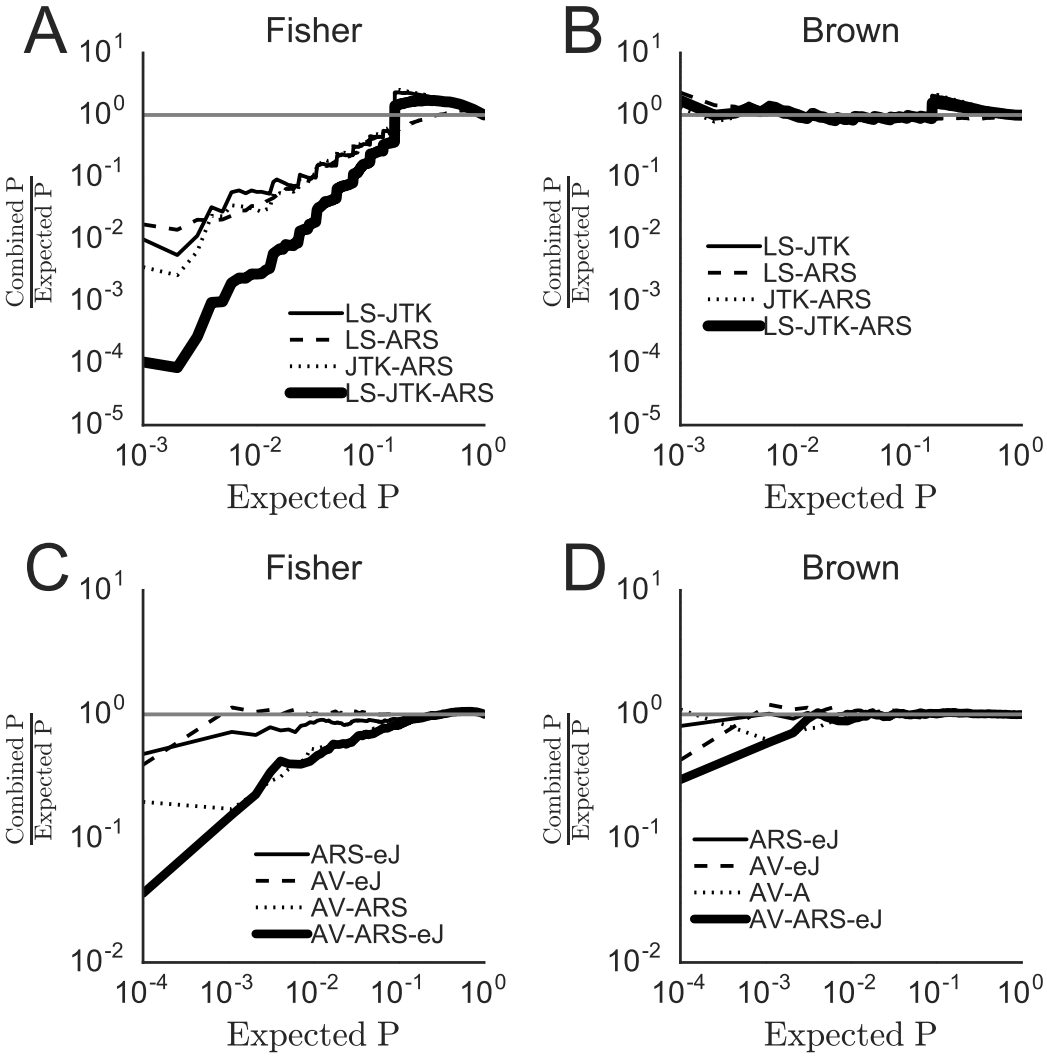

Figure 3.4: The Brown method for p-value integration provides more accurate results for rhythm detection method combination than the Fisher method. Various rhythm detection methods were tested on 1000 time series generated from random noise and then integrated using either the Fisher method ((A) and (C)) or the Brown method ((B) and (D)). (A) P-values from the integration via the Fisher method of the different methods (Lomb-Scargle (LS), ARSER (ARS), and JTK_CYCLE (JTK)) incorporated in MetaCycle after adjusting their p-values to be uniform. (B) P-values from the integration via the Brown method of the different methods incorporated in MetaCycle after adjusting their p-values to be uniform. (C) P-values from the integration via the Fisher method of ARSER (ARS), eJTK (eJ), and ANOVA (AV) results. (D) P-values from the integration via the Brown method of ARSER (ARS), eJTK (eJ), and ANOVA (AV) results. Expected p-values are determined knowing that the p-values should be uniformly distributed under the null hypothesis. The horizontal grey line in indicates a ratio between combined p-value and expected p-value of 1.

account when making the correction. This method makes the assumption that the p-values under consideration are independent. This is not the case for p-values matching one time series to a set of shifted and transformed waveforms. Since the waveforms are correlated, the p-values are correlated as well. This means that the Benjamini-Hochberg method produces p-values that are too small when applied to p-values from multiple waveform tests, as seen in Fig. 3.4A.

The second pitfall we identified was a method combining p-values from different rhythm detection methods using Fisher integration [31, 113]. Fisher integration is designed for independent p-values: for dependent p-values a correction is suggested by Brown [16]. Since the different rhythm detection methods search for similar features when applied to the same time series, the resulting p-values for each method are correlated. Therefore, using Brown's correction of Fisher integration results in more accurate p-values, though the improvement gained differs based on the methods being combined (Fig. 3.4C).

CHAPTER 4

BOOTSTRAPPING AND EMPIRICAL BAYES METHODS IMPROVE RHYTHM DETECTION IN SPARSELY SAMPLED DATA

4.1 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 limits 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 dataset with high time resolution to show that BooteJTK provides more consistent rhythm detection than existing methods at typical sampling frequencies. Then, we apply it 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 .

4.2 Introduction

Periodic patterns (rhythms) are pervasive in biology at molecular, cellular, organismal, and ecological scales. However, it can be challenging to detect these patterns with confidence of

their significance because biological dynamics are intrinsically noisy, and often it is feasible to obtain only a few samples of a process. To address this issue, several statistical methods have recently been developed to identify genes with cycling expression patterns from genome-wide time series [53, 99, 22, 117, 52, 59].

Empirical JTK_CYCLE (eJTK) [52, 53] and RAIN [99] are non-parametric 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 the measurement uncertainty. ANOVA [59] 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 [53]. ARSER [117] 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 non-sinusoidal time series.

Either explicit or implicit in these methods is comparison of the variation in measurements at each (presumed equivalent) time point (i.e., across replicates/periods) to the variation from one time point to another over the period of interest. The cost and effort of sample preparation and measurement limits the number of replicates/periods obtained. As a result, the observed variation at each time point may poorly represent the variation that would be obtained from many samples. In particular, if the data yield an estimate of the variation 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 variation has the potential to provide substantial gains in accuracy of rhythm detection and, in turn, aid in understanding periodic biological processes.

To this end, we introduce an empirical Bayes (eBayes) procedure [94, 68]. In this approach, which is commonly employed in differential expression analysis [101, 89], information from across a dataset is combined to estimate a prior distribution for the standard error, 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 (Fig. 4.S1) [68]. We incorporate the empirical Bayes variance estimates by applying a rhythm detection algorithm to parametric bootstrap time series samples. The parametric bootstrap [27] is also established in bioinformatics, and is applied in packages for RNA-Seq quantification [14] and differential expression [85]. To the best of our knowledge, this is its first application in rhythm detection.

While the strategy is general, we focus on its implementation with the empirical JTK_CYCLE (eJTK) method, which we have demonstrated outperforms most other algorithms [53]. eJTK compares time series to a set of reference waveforms varying in peak expression (phase) and distance from peak to trough using a non-parametric rank order correlation, Kendall's τ. Selecting the best waveform presents a multiple hypothesis testing problem, which eJTK solves by empirically calculating the null distribution to assign p-values to resulting rhythmicity scores. This approach is accurate but relatively computationally costly because the null distribution must be re-evaluated for each set of measurements (distinguishing time series 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 two circadian genome-wide expression datasets. The first dataset is densely sampled with measurements of gene expression in mouse liver samples at 1 h intervals for 2 periods [51]; this dataset allows us to examine the performance (self-consistency) of the method as fewer time points are included. The second dataset comprises gene expression measurements every 2 h for 2 periods for 12 mouse tissues in continuous darkness [121]. This dataset allows us to look at

the consistency of rhythm detection across tissues. We find that fewer genes are rhythmic than previously believed, due to 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 (CCT) [61], we find no decrease in CCT enrichment between BooteJTK and eJTK. At the same time, we find increases in the overlap of rhythmic genes across tissues. 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.

4.3 Methods

4.3.1 Empirical Bayes variance estimation

Empirical Bayes methods are an established part of many workflows for differential expression analysis [94, 101, 89]. These methods combine information, here standard deviation estimates, from all the time points in the data to 'shrink' the spread of standard deviations and improve the estimates. As a result, low standard deviation estimates are increased and high standard deviation estimates are decreased (Fig. 4.S1). Given that we have low replicate numbers, it is beneficial to use empirical Bayes methods. In particular, we use *voom* (specifically, *vooma* for microarrays) [89] to obtain initial estimates of the standard deviation that take mean-variance relationships into account. We initially used *limma* [89], which resulted in over-dispersion and over-estimates of rhythmicity p-values, partially due to adjustment of small standard errors away from zero. Instead, we found that using *vash* [68] to adjust our standard deviations did not result in over-dispersion.

4.3.2 Bootstrapping eJTK

In bootstrapping, data are resampled with replacement to create a distribution of simulated measurements that can in turn be used to compute statistics [27]. Genome-wide circadian time series generally consist of two or three replicates per time point, minimizing the benefits offered by resampling the data directly (i.e., non-parametric bootstrapping). Instead, we use parametric bootstrapping. Specifically, we log-transform the expression measurements [102] and model the resulting data at each time point as normally distributed with the mean directly calculated from the replicates and the variance modeled by *voom* [89] and the empirical Bayes procedure *vashr* [68], both implemented in R (Fig. 4.S1).

We generated time series for this model and analyzed them with eJTK to determine their circadian characteristics: rhythmicity score (τ), phase (peak), and best-matching waveform. We averaged each of these 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 non-integer, 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 τ 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 τ 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 τ tending to decrease as the noise becomes larger in comparison with the amplitude of the oscillation. As a result, the mean of the distribution of τ 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 τ distribution is skewed when τ 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 τ: $\tilde{\tau} = \text{arctanh}(\tau)$, truncating the values to $\pm\text{arctanh}(0.99)$ for $\pm\tau > 0.99$ to ensure that the $\tilde{\tau}$ values are finite.

4.3.3 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 [53]. We repeated this procedure for each number of measured time points in the experimental time series. Because this numerical procedure represents most of the computational expense of eJTK and, in turn, BooteJTK, we sought an approximate analytical form for the null distribution, and we found the Gamma distribution, which we previously used to model the F24 null distribution [53], to be a reasonable choice (Figs. 4.1 and 4.S2). 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 Gamma-distribution-generated p-values to the empirical p-values is only slightly larger than 1 at very low p-values and closer to 1 for more moderate p-values (Figs. 4.1B and 4.S2B). This means that the Gamma-distribution-based methods favor the null hypothesis slightly.

4.3.4 BooteJTK Outperforms Alternative Rhythm Detection Methods

To test the ability of BooteJTK to identify rhythmic time series, we generated 100 time series from a cosine with a 24 h period sampled every 2 h in duplicate with Gaussian noise added to each point. 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

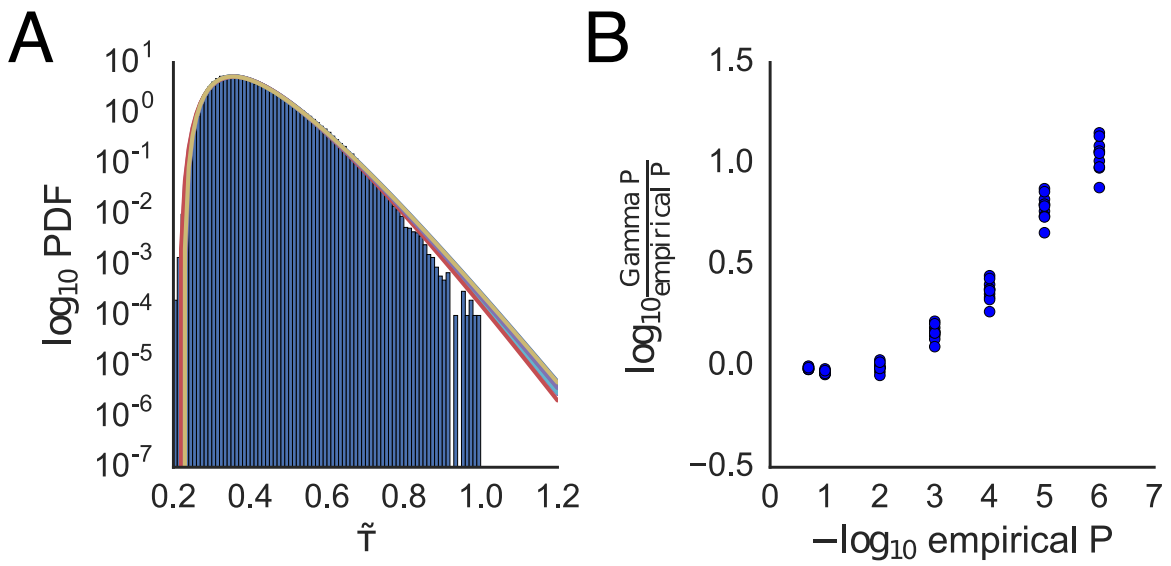

Figure 4.1: The BooteJTK $\tilde{\tau}$ null distribution can be modeled by a Gamma distribution. (A) Comparison between Gamma distributions parameterized by the Python scipy.stats.gamma.fit [56] function of $\tilde{\tau}$ values for 10^3 time series comprised of 24 Gaussian noise-generated time points and a histogram of 10^6 $\tilde{\tau}$ values. 10 different such Gamma distributions are shown. (B) The $\log_{10}$ ratios of p-values estimated from the Gamma distributions in (A) (Gamma P) to the p-value calculated directly from the cumulative distribution function for the 10^6 $\tilde{\tau}$ values used to construct the histogram (empirical P).

only from the Gaussian distribution. We compared BooteJTK to eJTK, approximating the null distribution as a Gamma distribution in both cases. In each case, we compared the simulated time series against cosine reference waveforms with 24-h periods, testing phases every 2 h from 0 h to 22 h and asymmetries every 2 h from 2 h to 22 h (132 total reference waveforms). Figure 4.2 shows that BooteJTK outperforms eJTK: for different noise levels the True Positive Rate (TPR) is higher for a given False Positive Rate (FPR) (Figs. 4.2A and 4.S3A), and the Matthews Correlation Coefficient (MCC) is higher for all p-values (Fig. 4.2B and 4.S3B). The MCC is 1 if a classifier is perfect and 0 if it performs no better than random guessing.

To gain a better sense of the differences between BooteJTK and eJTK, we selected two time series from the simulated data with noise level 0.30 that had the same τ value for eJTK ($\tau = 0.57$, $p = 0.002$) but different values for BooteJTK ($\tilde{\tau} = 0.67$ and 1.08, $p = 0.008$ and 0.001) (Figs. 4.S4A and B). With Benjamini-Hochberg correction to control for the false discovery rate when testing many time series [13], the first time series would likely be considered arrhythmic, while the second time series would likely be considered rhythmic. In Fig. 4.S4A, the average time series matches the original well, but the standard deviations around the mean points are large and overlap substantially. The overlap results in a lower BooteJTK rhythmicity score. In Fig. 4.S4B, the original time series matches the average and the standard deviations around the mean points are small and distinct from one another. The tighter uncertainties result in a higher BooteJTK rhythmicity score.

We also compared BooteJTK to RAIN [99], another non-parametric 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 had previously found that RAIN underestimates p-values, so we adjusted RAIN using 10^6 simulations of the null distribution as described in [55]. We found that BooteJTK outperforms RAIN (Fig. 4.2). As mentioned in the Introduction, parametric

methods do account for the size of the noise relative to the amplitude of the time series, and ARSER [117] is likely the best such method presently. Because ARSER fits a time series to a sinusoidal curve, we expect it to outperform nonparametric methods when detecting time series that are approximated well by that waveform. However, we expect that many biological time series deviate from sinusoidal [53]. For this reason, we compared BooteJTK, ARSER, as well as a reference free parametric method, ANOVA [59], for their abilities to detect 24 h time series with peak to trough intervals of 20 h before adding noise. BooteJTK outperforms both methods as shown by TPR vs. FPR and MCC (Fig. 4.S3C and D).

We thus see that BooteJTK combines eJTK's non-parametric aspects with ANOVA-like comparison of the relative measurement uncertainty (variance between replicate points) to the amplitude of the entire time series (variance across the time series). We thus wondered if a combination of ANOVA and eJTK p-values would perform similarly to BooteJTK. A new method, MetaCycle, was recently developed that combines ARSER, the original JTK_CYCLE, and Lomb-Scargle by integrating their p-values using Fisher integration to increase rhythm detection ability [31, 113]. We showed that their method underestimates p-values and can be improved by using the Brown correction of the Fisher integration method [16] as well as empirically calculating their p-values [55]. We used the Brown correction of Fisher integration to combine ANOVA and eJTK and compared it to BooteJTK. eJTK-ANOVA and adjusted MetaCycle are both outperformed by BooteJTK (Fig. 4.2). The combination of eJTK and ANOVA may not be suitable for improved rhythm detection and this result affirms that the classification strength of BooteJTK is greater than a combination of eJTK and ANOVA.

4.3.5 Computational Expense

Having established the sensitivity and specificity of BooteJTK, we sought to minimize the computational cost of our method. For the simulated data described above, we found no

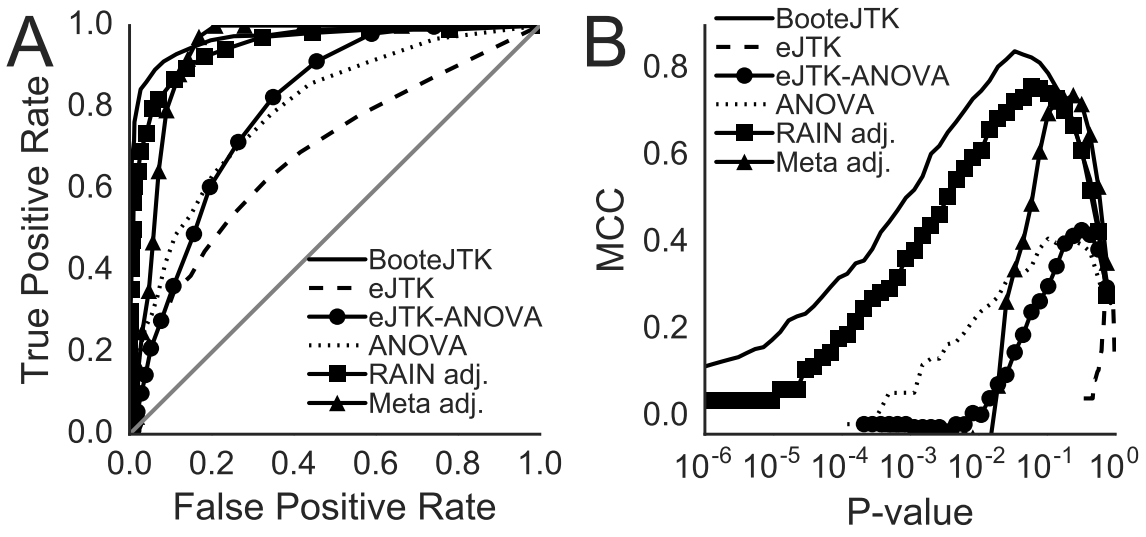

Figure 4.2: BooteJTK outperforms alternative methods in detecting simulated sinusoidal times as measured by the (A) True Positive Rate (TPR) against False Positive Rate (FPR) and (B) Matthews Correlation Coefficient. As indicated in the legend, we compare BooteJTK with 25 bootstrap samples to RAIN, empirical JTK (eJTK), ANOVA, and a Brown-Fisher method p-value integration citeFisher1925,Brown1975 of eJTK and ANOVA (eJTK-ANOVA). The RAIN p-values are adjusted empirically to ensure that they are uniform under the null model. Simulated data are generated with noise-to-amplitude ratio of 0.50 as discussed in the text.

discernible difference in $\tilde{\tau}$ scores for 100, 50, 25, and 10 bootstrap samples. We use 25 bootstrap samples throughout the rest of this study. With 25 bootstrap samples, we obtain timings on a Late-2013 iMac Desktop 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 paper, the computational cost of eJTK is several orders of magnitude less than in Hutchison *et al.* [53].

4.4 Results

4.4.1 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 expression cycling 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, here we applied BooteJTK to microarray data collected every 1 h for 48 h from mouse liver tissue under constant conditions [51]. As the original analysis of the dataset was performed with JTK_CYCLE [52], we analyzed the dataset with eJTK as well for comparison. 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 [121] or 4 h [33, 84], we parsed the data (from CT18-CT65) into two datasets with measurements every two hours (denoted 2a: CT18, CT20, etc. and 2b: CT19, CT21, etc.) and into four datasets with measurements every four hours (denoted 4a, 4b, 4c, and

4d, starting at CT18, CT19, CT20, and CT21, respectively).

Using the R package *gcrma* [115] to normalize the data (GEO GSE11923) and removing probes with constant expression, we compared the Benjamini-Hochberg adjusted p-values (BH) from BooteJTK to eJTK on the full dataset using a significance threshold of 0.05 (Fig. 4.3A). BooteJTK identifies more genes as rhythmic than eJTK does. When BooteJTK identifies a gene as rhythmic, it tends to assign it a lower p-value than eJTK, which leads to lower Benjamini-Hochberg adjusted p-values. BooteJTK can provide lower p-values because it checks whether the values of a time series have the right rank order (the sole criterion for eJTK) *and* whether the differences between pairs of points are large compared to the uncertainties in those measurements. The additional requirement makes it harder for a rhythmic pattern to arise by chance.

Downsampling the data to measurements every 2 h reduces the number of rhythmic genes identified by both methods using a BH-corrected p-value threshold of 0.05 (Fig. 4.3B), though a greater fraction of BooteJTK-identified genes remain. Downsampling the data to measurements every 4 h prevents eJTK (and adjusted RAIN, Fig. 4.S6) from finding any rhythmic genes at this significance threshold, while BooteJTK identifies several thousand of the genes that it originally considered rhythmic. As noted above, BooteJTK incorporates more information, and the results are thus less sensitive to Benjamini-Hochberg correction.

The results can also be analyzed using the overlap between downsampling results. The overlap between results obtained with different levels of downsampling can be quantified by the probability that a probe is rhythmic in one dataset (a row in Fig. 4.4B) if it is rhythmic in another (a column in Fig. 4.4B). For example, the probabilities are 0.78 and 0.62 ((row 3, column 2) and (row 2, column 3), respectively) that BooteJTK considers a time series downsampled to measurements every 2 h to be rhythmic if it identified that time series as rhythmic in the other dataset downsampled to measurements every 2 h. For eJTK, these values are 0.63 and 0.68 ((9,10), and (10,9), respectively) and for RAIN they are

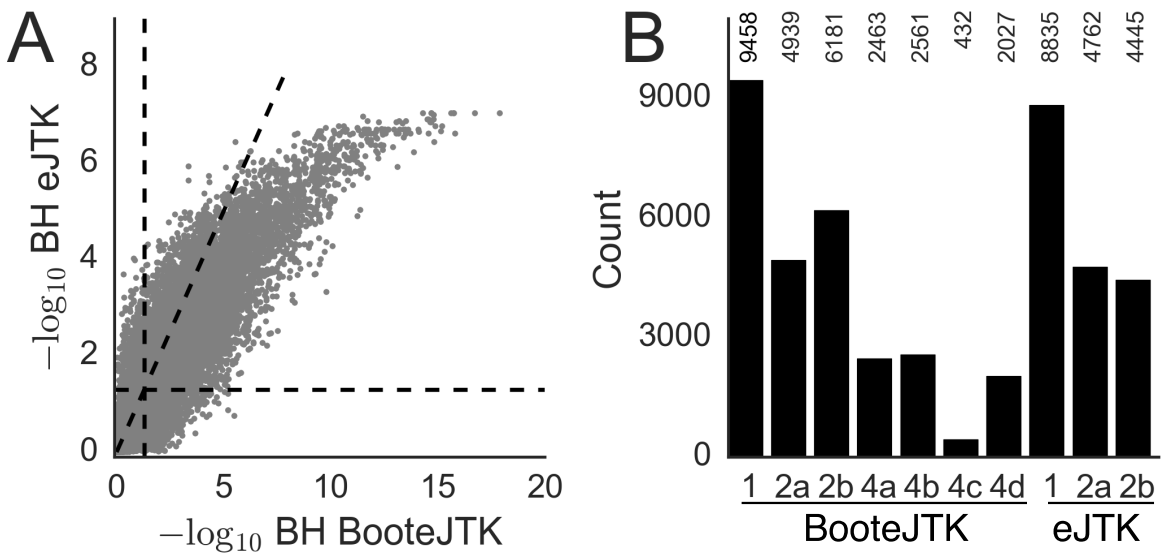

Figure 4.3: BooteJTK identifies rhythmic genes more consistently than eJTK as data are downsampled. Data shown are from Hughes *et al.* [51] and are originally sampled every 1 h for 48 h; the datasets are downsampled to measurement intervals of 2 h (denoted 2a and 2b) and 4h (denoted 4a, 4b, 4c, and 4d). (A) Comparison of BooteJTK and eJTK BH values for the full dataset. Diagonal line has slope of 1; horizontal and vertical lines indicate $-\log_{10}(\text{BH} = 0.05) \approx 1.3$. (B) Number of rhythmic probes at BH < 0.05 for the indicated methods and datasets downsampled to measurements every two hours (denoted 2a: CT18, CT20, etc. and 2b: CT19, CT21, etc.) and to measurements every four hours (denoted 4a, 4b, 4c, and 4d, starting at CT18, CT19, CT20, and CT21, respectively); the full dataset sampled every hour is denoted 1. eJTK identifies zero genes as rhythmic when the data are downsampled by 4 h.

0.61 and 0.68 (Fig. 4.S7B, (12,13) and (13,12), respectively). These results indicate greater consistency for the BooteJTK results and provide insight into how much inconsistency we should expect from experimental uncertainty because no biological differences should exist between downsampled sets.

4.4.2 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.* [121]. The twelve tissues are adrenal (Adr),

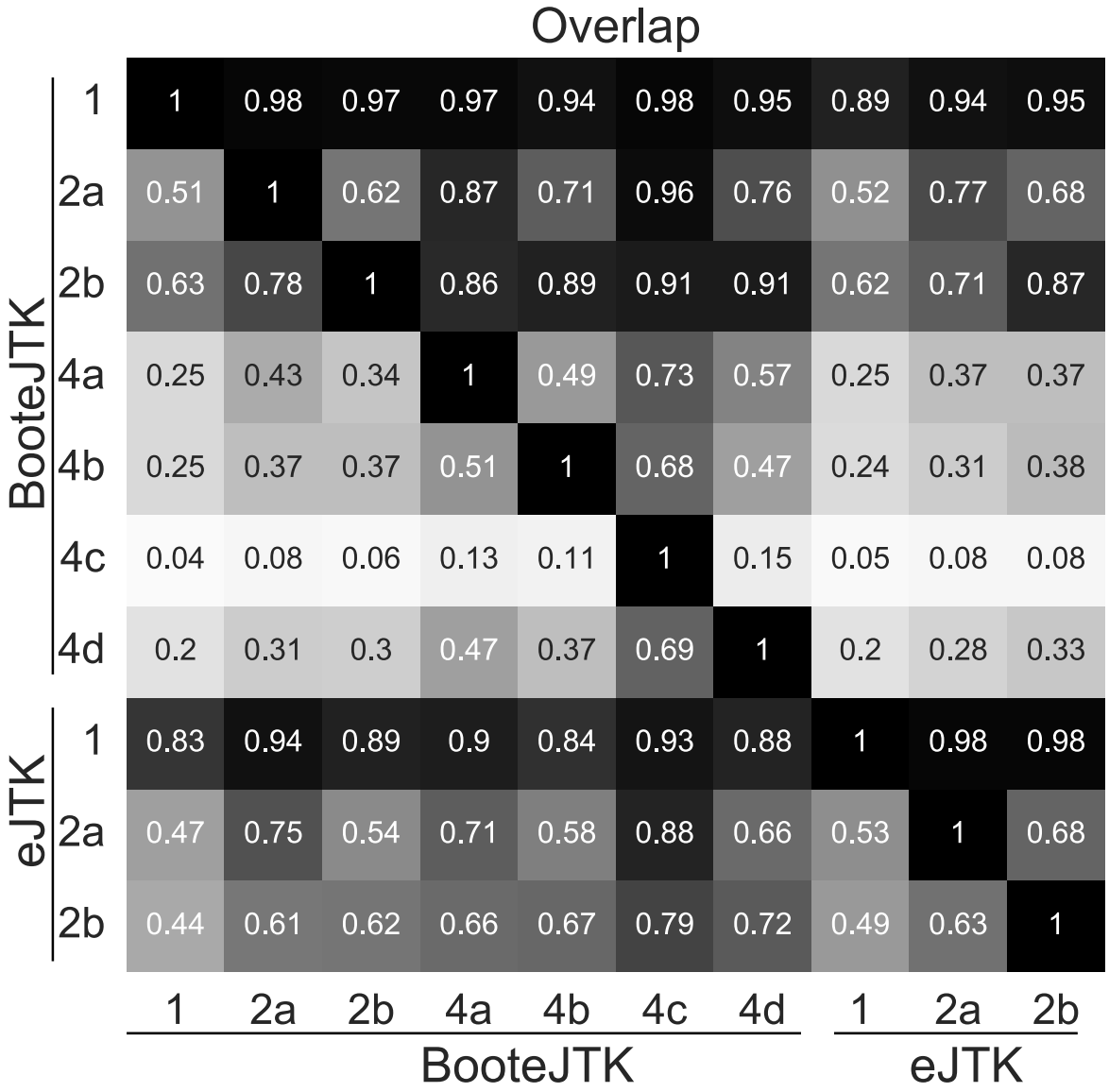

Figure 4.4: BooteJTK provides more consistent rhythm detection than eJTK between downsampled datasets. We quantified the overlap between results with different levels of downsampling by the probability that a probe is rhythmic in one dataset (a row) if it is rhythmic in another (a column). As no probes are found to be rhythmic when using eJTK on data downsampled to every 4 h, rows and columns for that conditions are not shown.

aorta, brown fat (BFAT), brainstem (BS), cerebellum (Cere), heart, hypothalamus (Hypo), kidney, liver, lung, muscle (Mus), and white fat (WFAT). Expression was sampled every 2 h for 48 h, which we again treated as duplicate measurements over 24 h. We also applied eJTK, mirroring the application of the original JTK_CYCLE method [52] by Zhang *et al.* [121].

BooteJTK is more stringent than eJTK but exhibits comparable enrichment for core clock targets

When applied to the 12-tissue microarray dataset from Zhang *et al.* [121], 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) (Figs. 4.5A and B). 12,261 of these BooteJTK-rhythmic probes are contained within the four tissues with the highest number of rhythmic genes: liver, kidney, lung, and brown fat (Fig. 4.5B). 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. However, we were concerned that the greater stringency might exclude actual rhythmic time series. To evaluate this possibility, we corroborated our rhythmic genes with core clock transcription factor targets (CCTs) identified by ChIP-Seq in mouse liver [61]. Across the 12 tissues and between the two methods, we found no meaningful difference in the fraction of CCT genes relative to the number of genes identified as rhythmic (Fig. 4.S8). This result persists as we increase the requirements for a gene to qualify as a CCT: on the vertical axis of Figure 4.S8 we increase the number of core clock transcription factors that need to target a gene for it to be classified as a CCT.

Relationships between tissues

We examined the relationships between different tissues through the probability that a gene is rhythmic in one tissue conditioned on it being rhythmic in another (π). In Fig. 4.6A, row tissues are conditioned on column tissues, 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 0.46 that a probe is rhythmic in the liver, but being rhythmic in the liver leads to only a probability of 0.07 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.* [51]: the π values for two datasets from identical conditions downsampled to measurements every 2 h were 0.62 and 0.78. We note that a similarly high value is observed for the probability that a probe is rhythmic in brown fat conditioned on being rhythmic in the aorta (0.70). This result is interesting, but we feel further study is warranted as this point is an outlier in Fig. 4.6B, and the presence of brown fat around the aorta can easily lead to contamination of aorta samples [32].

In Fig. 4.6C, we show the differences between the π relationships from BooteJTK and eJTK. The columns corresponding to the brain tissues show marked differences. Zhang *et al.* [121] 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 regarding other tissues relative is due to an decrease in probes identified as rhythmic by BooteJTK relative to eJTK (Figs. 4.6C and D).

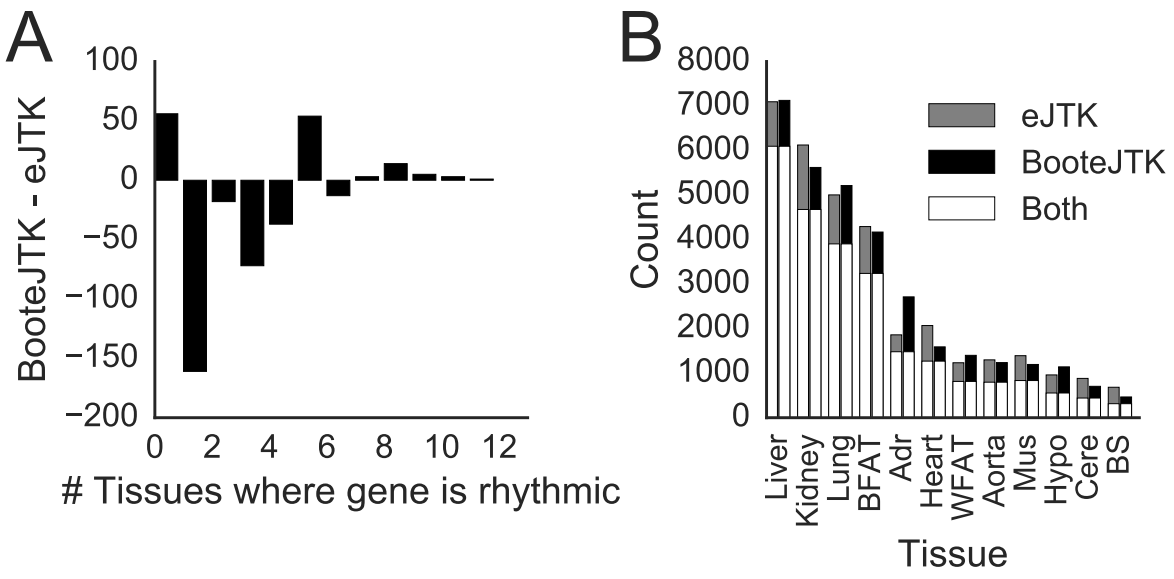

Figure 4.5: BooteJTK reveals fewer genes with circadian rhythmic expression across 12 mouse tissues than eJTK for the Zhang *et al.* [121] dataset. (A) Fewer probes are rhythmic in multiple tissues under BooteJTK than under eJTK (with BH < 0.05 for both methods). (B) Fewer probes per tissue are identified as rhythmic with BooteJTK than with eJTK. Dashed lines are cumulative sums of unique probes from left to right.

Genes that are rhythmic in most tissues

Thirteen genes are identified as rhythmic in all 12 tissues by BooteJTK: Arntl (Bmal), Nr1d1 (Rev-erbA), Nr1d2 (Rev-erbB), Dbp, Per1, Per2, Per3, Ciart (Chrono), Bhlhe41 (Dec2), Tns2, Tsc22d3, Usp2, and Tspan4. Many of these are genes involved in the core clock machinery [106, 41]. 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 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.* [121] 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 rhythmic in 9 or more

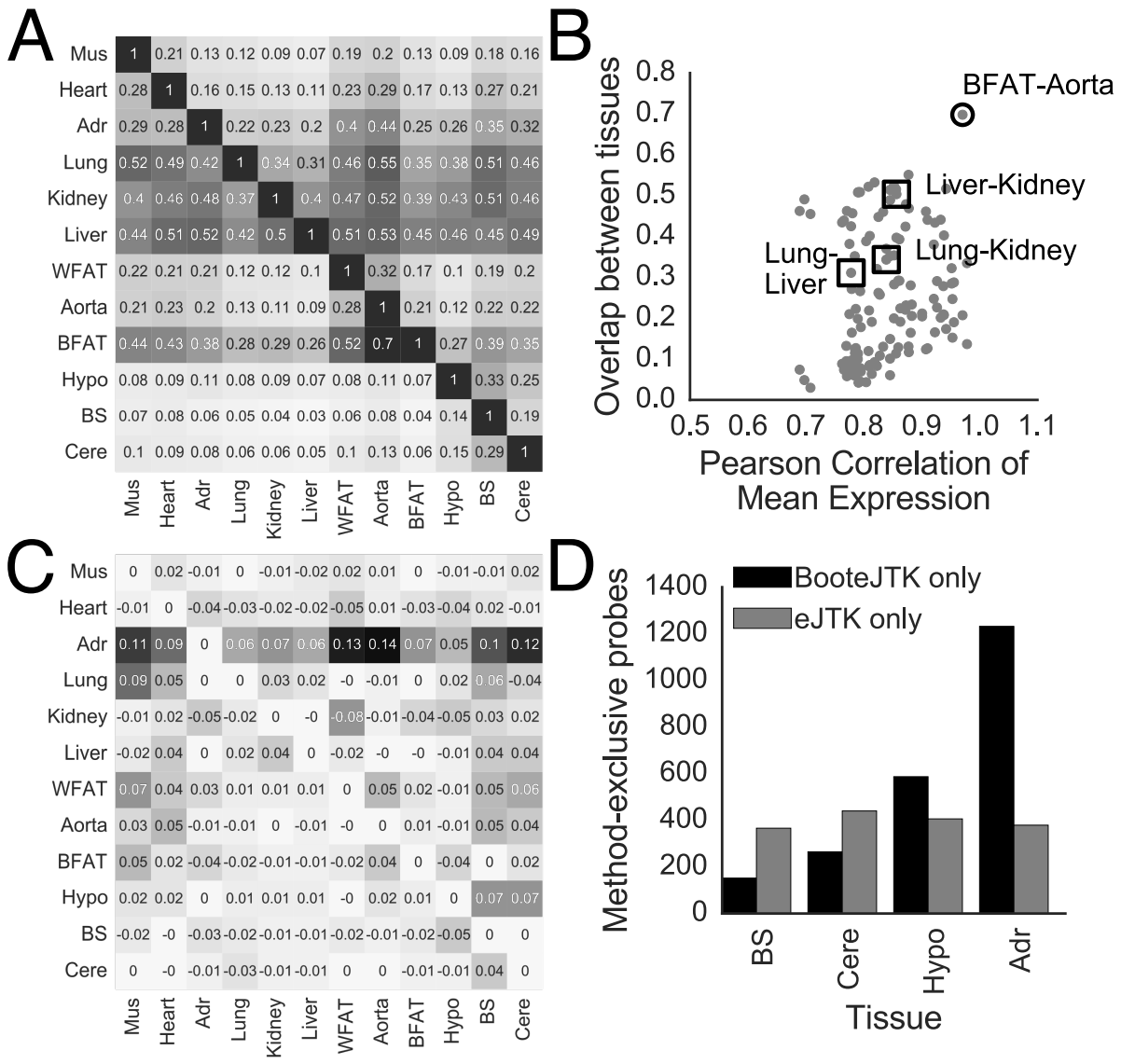

Figure 4.6: (A) The probability that a probe as rhythmic in a row tissue if it is rhythmic in a column tissue (π) for BooteJTK. (B) Comparing correlation of mean probe expression between tissues with rhythmic overlap reveals that brown fat and aorta tissue have high correlation of gene expression and high rhythmic overlap relative to other tissue pairs, such as liver, kidney, and lung. (C) The difference in π probabilities from BooteJTK to eJTK. A positive value indicates a higher π for BooteJTK than eJTK. Tissues are clustered by column π vectors from BooteJTK. (D) Numbers of rhythmic genes specific to each method for the tissues showing large differences in (C).

of the tissues. Most of these genes were not rhythmic in the brainstem, hypothalamus, or cerebellum (Fig. 4.S9).

Examining the functional annotation of these genes revealed many ontologies to be expected of consistently rhythmic genes, such as rhythmic processes and transcription regulation (Tab. 4.4.2). However, additional functional annotations were identified, such as genes involved in the stress response, endoplasmic reticulum, pigment granules, and heat shock. In addition to these genes, a few interesting other genes stand out. Wee1 is rhythmic in 10 tissues (absent from hypothalamus and brain stem). Wee1 regulates cellular division by inhibiting entry into mitosis [60] and is known to be regulated by the core clock [71]; more generally, it has been suggested that the cell cycle is under circadian control [92]. 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 Gstt2 are identified as rhythmic in 10 tissues, while Fmo2 is identified as rhythmic in 11 tissues. These genes are identified as genes involved in drug metabolism by the DAVID webtool [49, 50]. Given the increasing interest in chronotherapeutics [121], further research into these genes is warranted in order to better understand their involvement in circadian processes.

4.5 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 [53] by a parametric bootstrap. Because eJTK itself is nonparametric, BooteJTK maintains sensitivity for arbitrarily shaped and scaled waveforms but accounts for the experimental uncertainty when comparing measurements. We demonstrated that the method provides

Functional Annotation	Fold Enrich.	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

Table 4.1: Select functional annotations for the 119 genes identified as rhythmic in 9 or more tissues by BooteJTK with a Benjamini-Hochberg adjusted p-value threshold of 0.05 from the Zhang *et al.* dataset [121] analyzed the with DAVID webtool [49]. Fold Enrichment refers to how enriched the functional annotation is in the set of genes relative to what would be expected from a set of randomly selected genes. Abbreviations: BH, Benjamini-Hochberg adjusted p-value SP-PIR, Swiss-Prot Protein Information Resource keywords GO, Gene Ontology keywords

improved accuracy in identifying simulated rhythmic time series and improved consistency across related experimental datasets. More generally, we expect framework that we have built around JTK_CYCLE [52, 53]—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 fast, accurate p-values as we do here.

Our method uses replicates to estimate the variance in expression, which is then propagated to the rhythmicity estimate. For time series data where replicate time points do not exist, a different approach is needed. We suggest that the standard deviation of arrhythmic time series is a reasonable approximation of the standard deviation of the time points of rhythmic time series. In Fig. 4.S10, we show that the mean of the standard deviation of the arrhythmic ($p > 0.8$) time series slightly overestimates the standard deviation for the Hughes *et al.* dataset sampled every 1 h [51]. While ideally replicate time points would be collected experimentally, we suggest that this approximation be used when replicates are lacking.

Our analysis of the mouse liver microarray dataset from Hughes *et al.* [51] emphasizes the importance of sampling time series frequently and understanding the expected consistency of results. While BooteJTK, in contrast to eJTK, was 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. Given this, sampling every 2 h is much better than the commonly used 4 h sampling rate, especially when comparisons are made across tissues or conditions. We show, in application to the 12-tissue Zhang *et al.* dataset, that comparing overlap results to this benchmark can be informative, as we demonstrated in identifying the aorta tissue data as potentially contaminated by brown fat. We suggest that future studies comparing rhythmicity between datasets use these values to better understand the overlap and consistency that should be expected from rhythm detection methods for these types of data.

Multiple studies have discussed the tissue-specificity of circadian rhythms [83, 95, 121]. Our analysis with BooteJTK supports these claims but suggests that a slightly greater fraction of genes are rhythmic in multiple tissues than was previously appreciated (Fig. 4.5B). One limitation of current rhythm detection methods (including eJTK and BooteJTK) is that the null hypothesis states that we have no prior belief or knowledge about the rhythmicity of any given gene. However, genes such as Arntl (Bmal), Nr1d1 (Rev-erbA), Nr1d2 (Rev-erbB), Dbp, Per1, Per2, and Per3 are well-established as circadian and are identified as rhythmic in all 12 tissues. Given that we have information from so many tissues, we can begin to think about how to integrate that information. Performing a simple combination of BooteJTK p-values via Fisher's integration method, which would ignore all tissue-specific knowledge to be gained from sampling 12 tissues, results in 14,598 probes with a Benjamini-Hochberg p-value below 0.05. This provides a pooled estimate for the total number of rhythmic probes in the mouse, ignoring tissue-specific effects.

Methods have been recently developed, however, that can combine information across tissues while allowing for heterogeneity in the context of identifying single-nucleotide polymorphisms (SNPs) that are expression quantitative trait loci (eQTLs) [34]. These methods tend to increase the number of eQTLs that are common to multiple tissues, relative to single-tissue approaches. We expect that applying a similar strategy in rhythm detection would have an analogous effect. For example, Npas2, Hlf, and Tef are identified as rhythmic in 11 of the 12 tissues profiled. Given that these are documented circadian genes [106], it is more likely that these genes are rhythmic in the tissue in which they have not been identified as rhythmic than their p-values suggest. Therefore the current estimates of genes that are rhythmic across several tissues may be underestimates. By the same token, 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 [61], with the caveat that

the ChIP-Seq data was only performed in the mouse liver. It would be useful to be able to systematically integrate multiple sources and types of evidence for rhythm detection.

4.6 Supplementary Figures

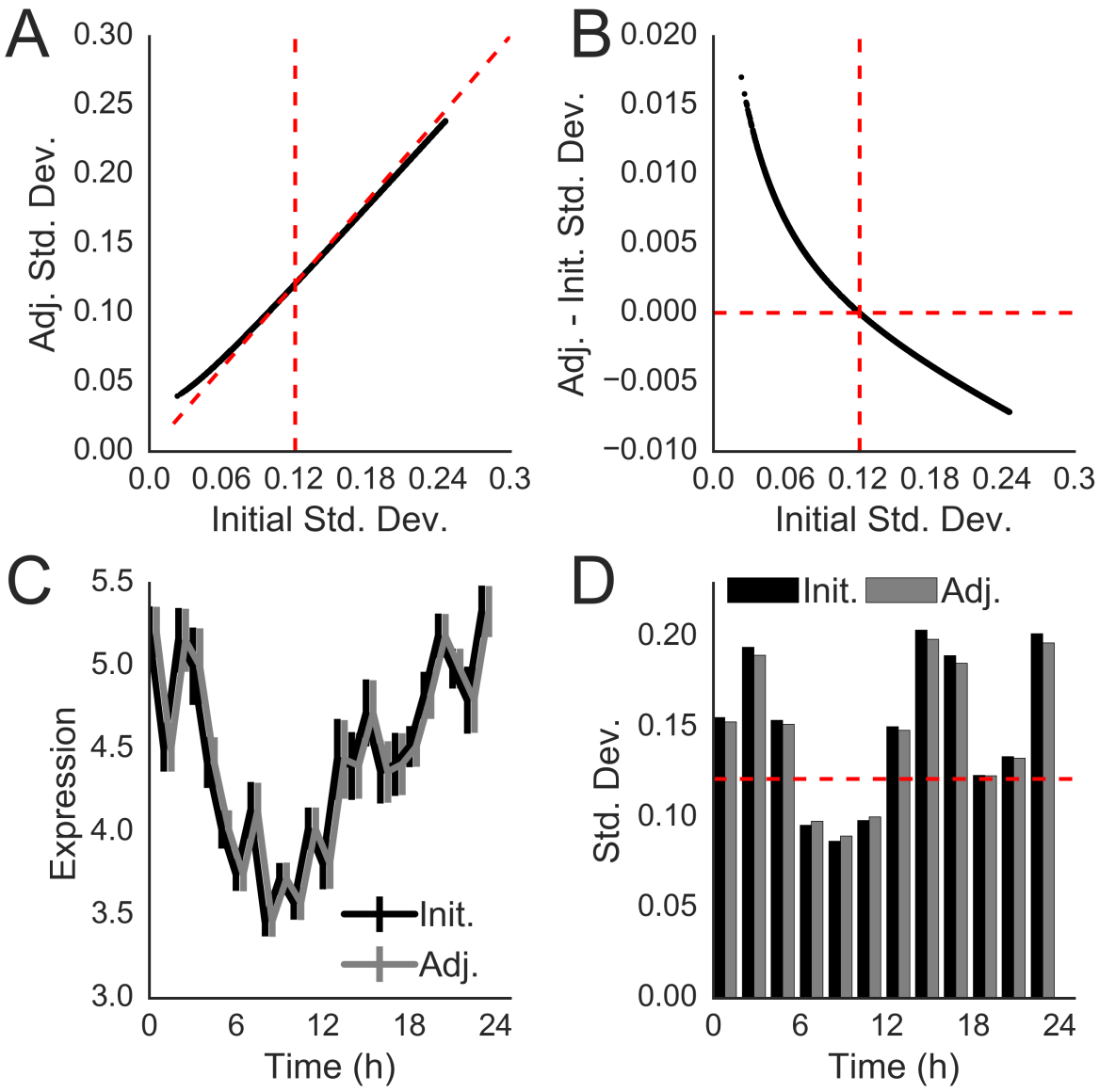

Figure 4.S1: (A) Empirical Bayes adjustment of RNA expression from Hughes *et al.* [51]. Empirical Bayes adjusts the standard deviation estimates to 'shrink' the distribution towards the global estimate of 0.1214 (vertical line). The diagonal line has slope 1; points below it have been reduced by the eBayes method while points above it have been increased. Shrinkage was performed using *voom* [89] and *vash* [68] in R . (B) The difference between adjusted standard deviation and initial standard deviation is shown against the initial standard deviation. (C) Time series of Clock expression from Hughes *et al.* shown with initial and adjusted standard deviations as error bars. (D) Initial and adjusted standard deviations for even time points Clock RNA expression (C) shown as bars.

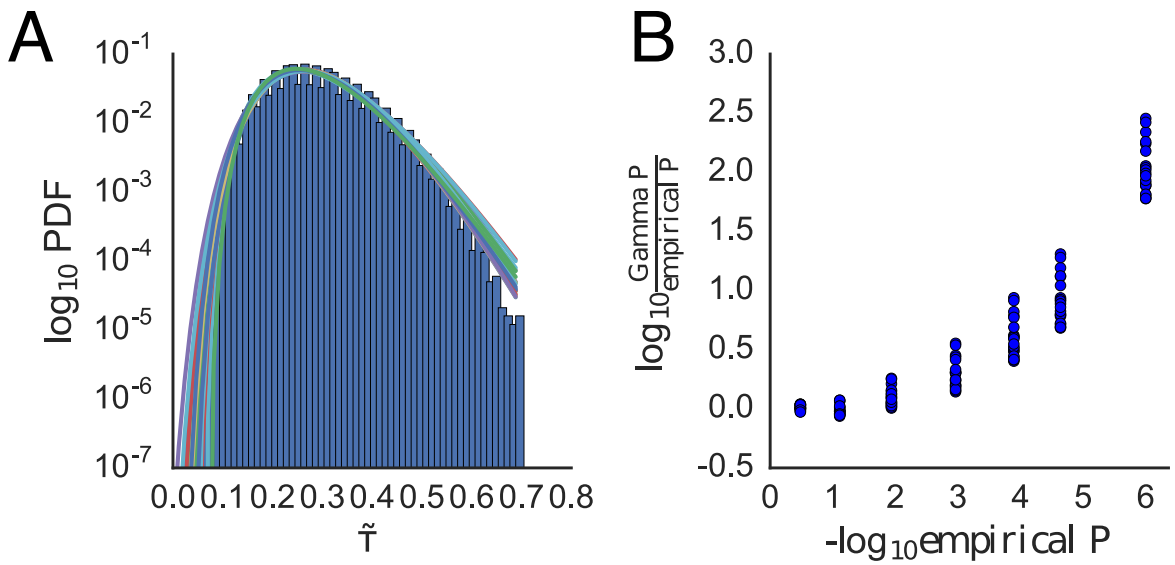

Figure 4.S2: The eJTK τ null distribution can be modeled by a Gamma distribution. (A) Comparison between Gamma distributions parameterized by the Python scipy.stats.gamma.fit [56] function of $\tilde{\tau}$ values for 10^3 time series comprised of 24 Gaussian noise time points and a histogram of 10^6 $\tilde{\tau}$ values. 20 different such Gamma distributions are shown. (B) The $\log_{10}$ ratios of p-values estimated from the Gamma distributions in (A) (Gamma P) to the p-value calculated directly from the cumulative distribution function for the 10^6 $\tilde{\tau}$ values used to construct the histogram (empirical P).

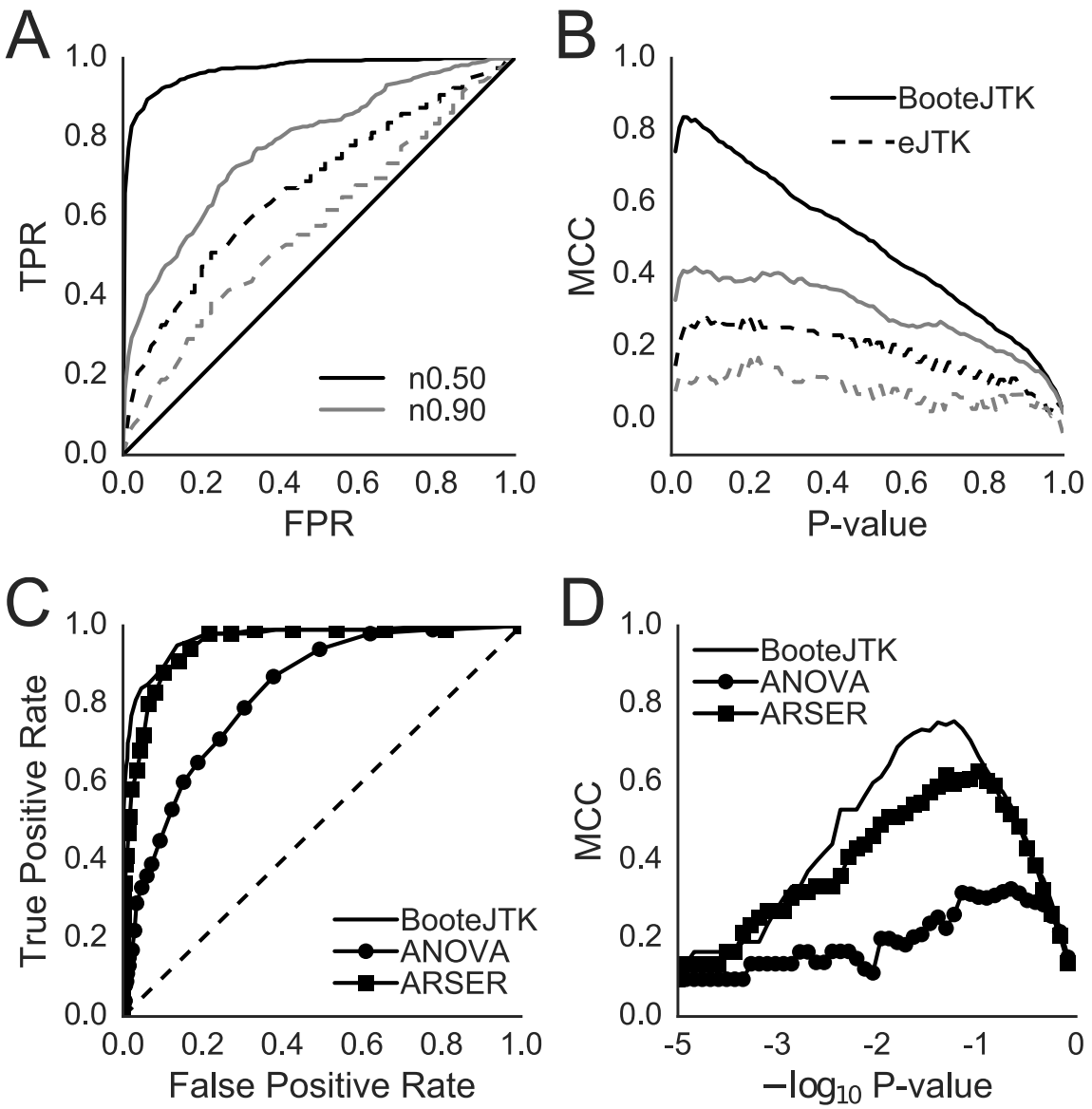

Figure 4.S3: (A & B) BooteJTK outperforms eJTK for simulated data generated from a cosine with noise-to-amplitude ratio of 0.50 and 0.90. (A) True Positive Rate (TPR) vs. False Positive Rate (FPR) and (B) Matthews Correlation Coefficient (MCC). (C & D) BooteJTK outperforms ANOVA and ARSER for simulated data generated from a cosine modified so time from peak to trough is 4 h and the time from trough to peak is 20 h with noise-to-amplitude ratio of 0.50 as shown for (C) TPR vs. FPR and (D) MCC.

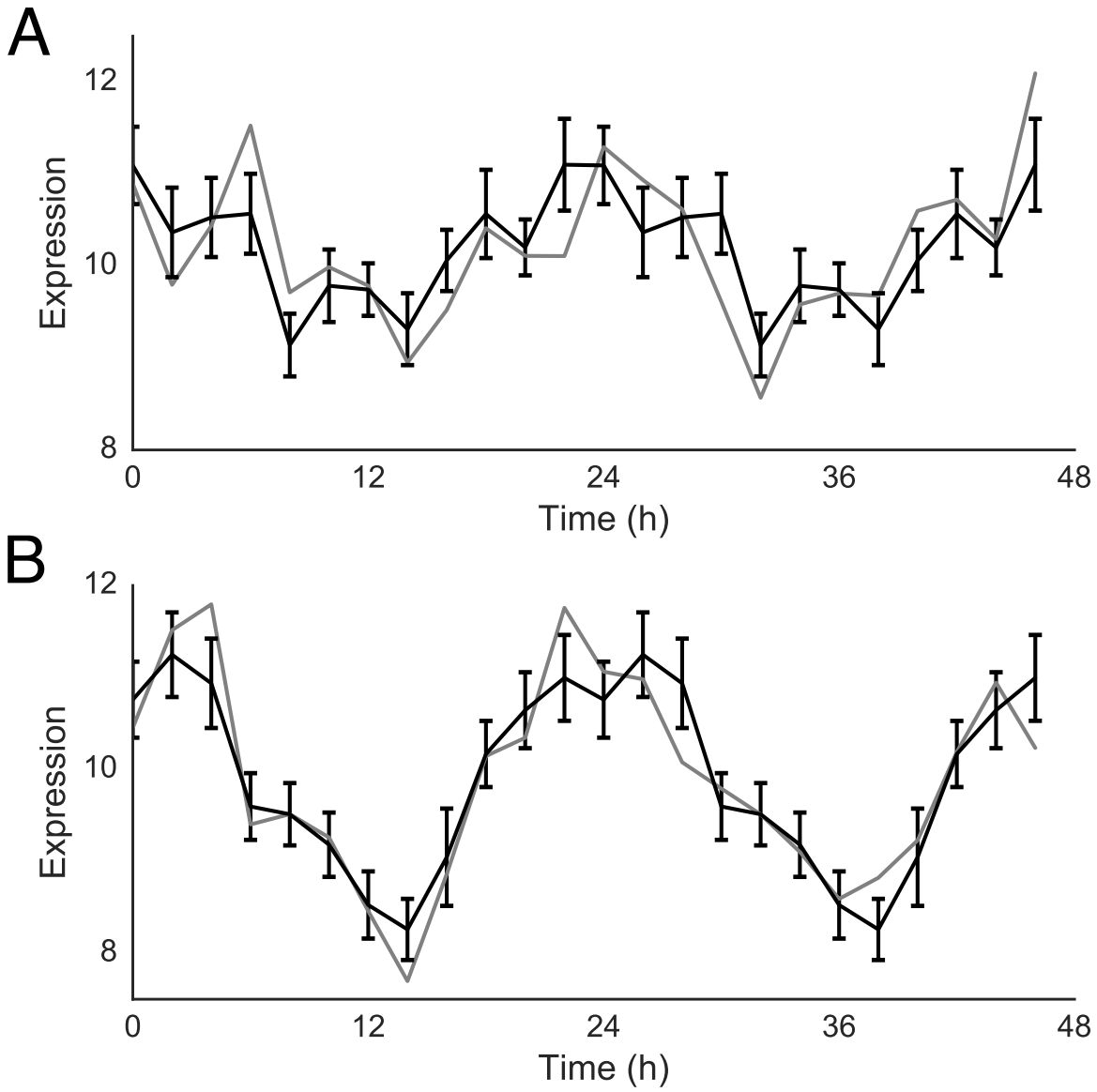

Figure 4.S4: (A) and (B) Examples of time series that have the same eJTK Example of two time series with the same eJTK $\tilde{\tau}$ but different BooteJTK $\tilde{\tau}$ values (as indicated). The green line is the original time series, the blue is the averaged time series with error bars, double plotted to the length of the original time series, and the black dashed line is the best-matching cosine. The BooteJTK $\tilde{\tau}$ values for (A) and (B) are 0.66 and 0.97 which correspond to p-values of 10^{-3} and 10^{-6}. The eJTK $\tilde{\tau}$ score was $\tilde{\tau} = 0.57$, which corresponds to a p-value of 0.002.

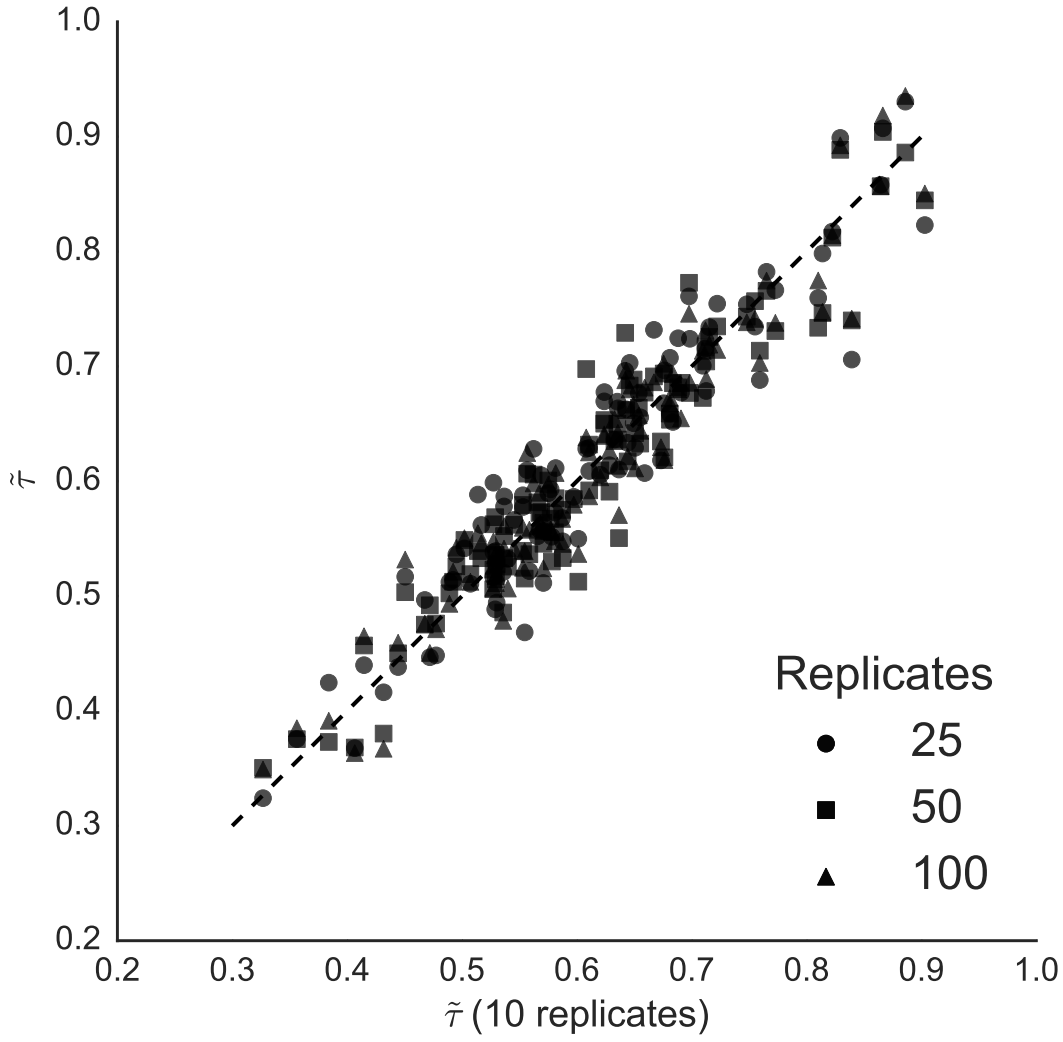

Figure 4.S5: The BooteJTK $\tilde{\tau}$ scores for the same time series dataset show no substantial difference for 10, 25, 50, or 100 bootstrap samples. The diagonal line is of slope 1.

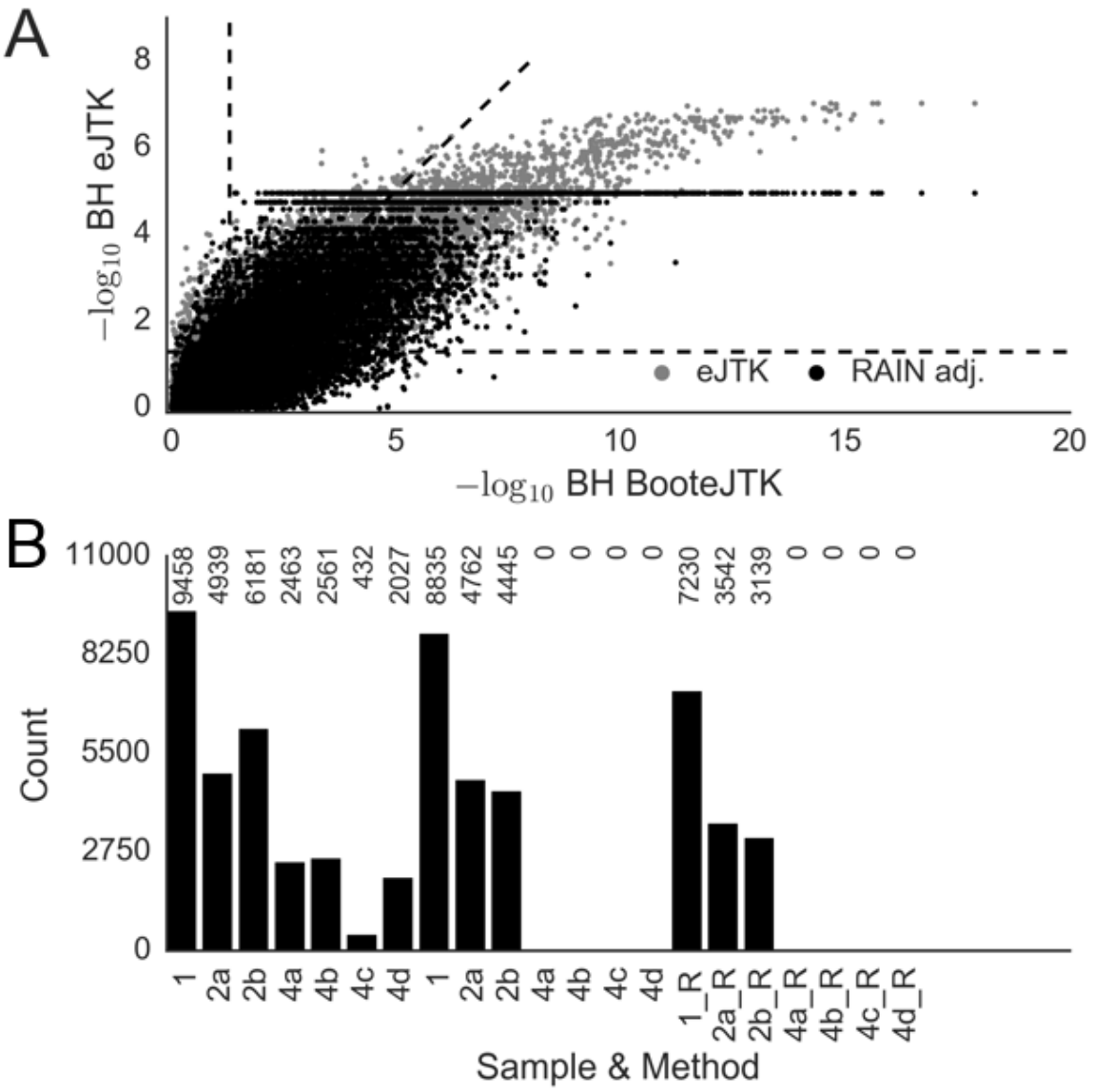

Figure 4.S6: BooteJTK identifies rhythmic genes more consistently than eJTK and RAIN as data are downsampled. Data shown are from [51] and are originally sampled every 1 h for 48 h; the datasets are downsampled to measurement intervals of 2 h (denoted 2a and 2b) and 4h (denoted 4a, 4b, 4c, and 4d). (A) Comparison of BooteJTK (B) and eJTK (eJ) BH values. Diagonal dashed line has slope of 1; horizontal and vertical dashed lines indicate BH = 0.05, as $-\log_{10}(0.05) \approx 1.3$. (B) Number of rhythmic probes at BH < 0.05 for different methods and downsamplings.

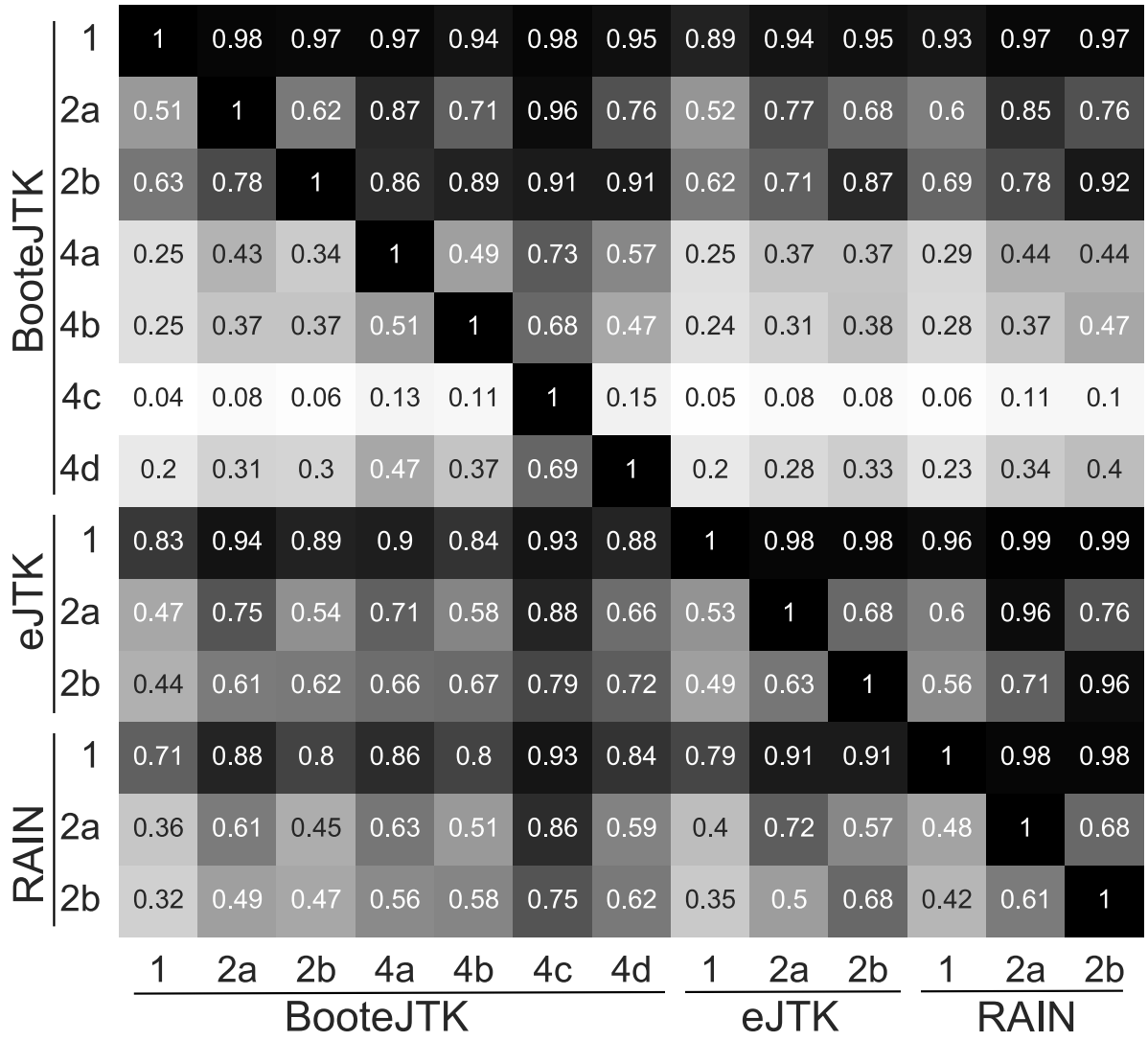

Figure 4.S7: For the Hughes *et al.* [51] dataset, downsampled data analyzed with BooteJTK (B) have greater overlap in genes identified as rhythmic between samples as when analyzed with eJTK (eJ) or with RAIN (R). The downsampled data are parsed into two datasets with measurements every two hours (denoted 2a: CT18, CT20, etc. and 2b: CT19, CT21, etc.) and into four datasets with measurements every four hours (denoted 4a, 4b, 4c, and 4d, starting at CT18, CT19, CT20, and CT21, respectively), while the full dataset sampled every hour is denoted by '1'. The overlap between results obtained with different levels of downsampling can be quantified by the probability that a probe is rhythmic in one dataset (a row) if it is rhythmic in another (a column). As no probes are found to be rhythmic when using eJTK or RAIN on data downsampled to every 4 h, rows and columns for those conditions are not shown.

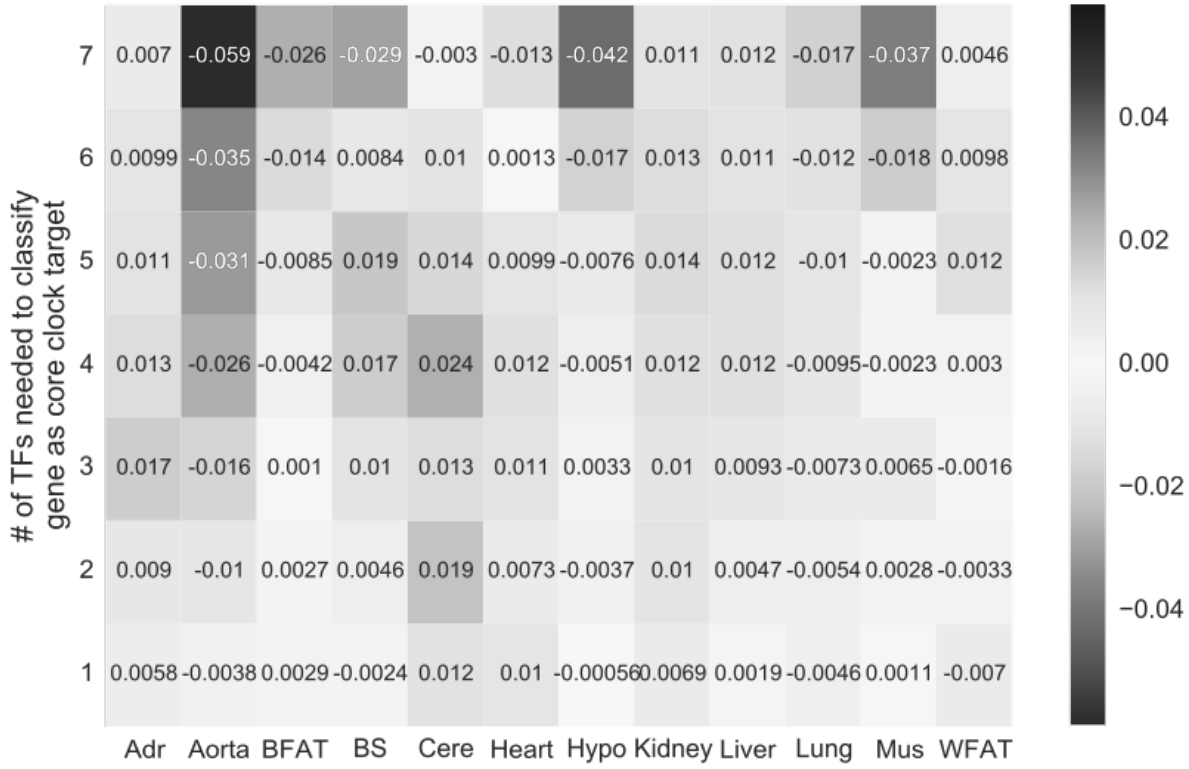

Figure 4.S8: No net difference exists between fraction of core clock genes rhythmic across tissues analyzed with BooteJTK as compared to eJTK (BooteJTK - eJTK). The cells indicate the difference in fractional presence of core clock genes. Core clock target genes are defined as being targets of core clock transcription factors Clock, Bmal1, Per1, Per2, Per3, Cry1, or Cry2 as defined in Koike *et al.* [61]. The vertical axis indicating the number of core clock transcription factors that need to target a gene in order for a gene to be defined as a core clock target.

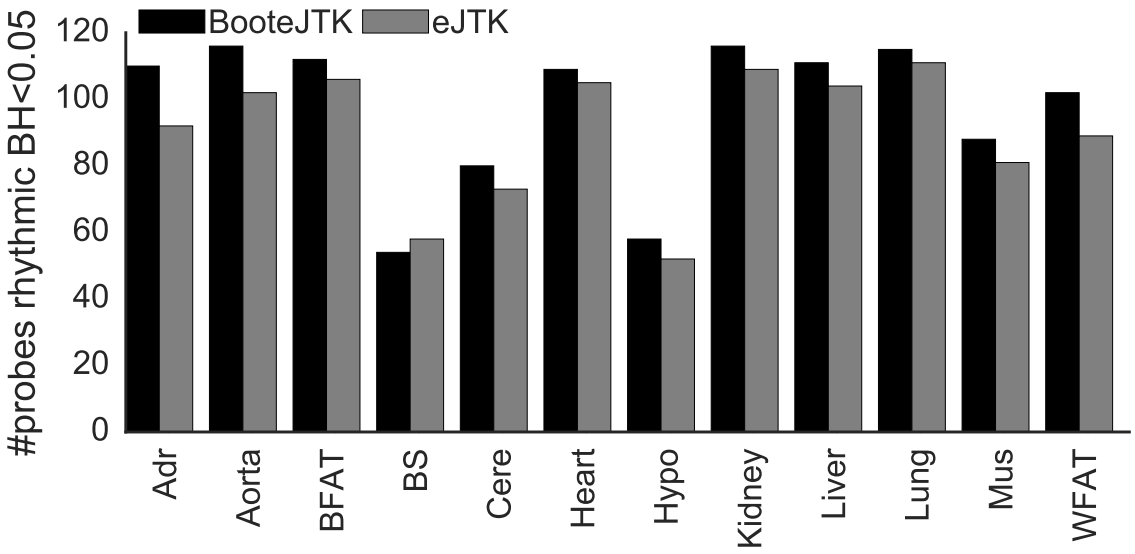

Figure 4.S9: (A) Number of probes rhythmic in each tissue out of the 78 probes rhythmic in 9 or more tissues. Of the probes rhythmic in more than 9 tissues, many of them are not found to be rhythmic in the brain tissues.

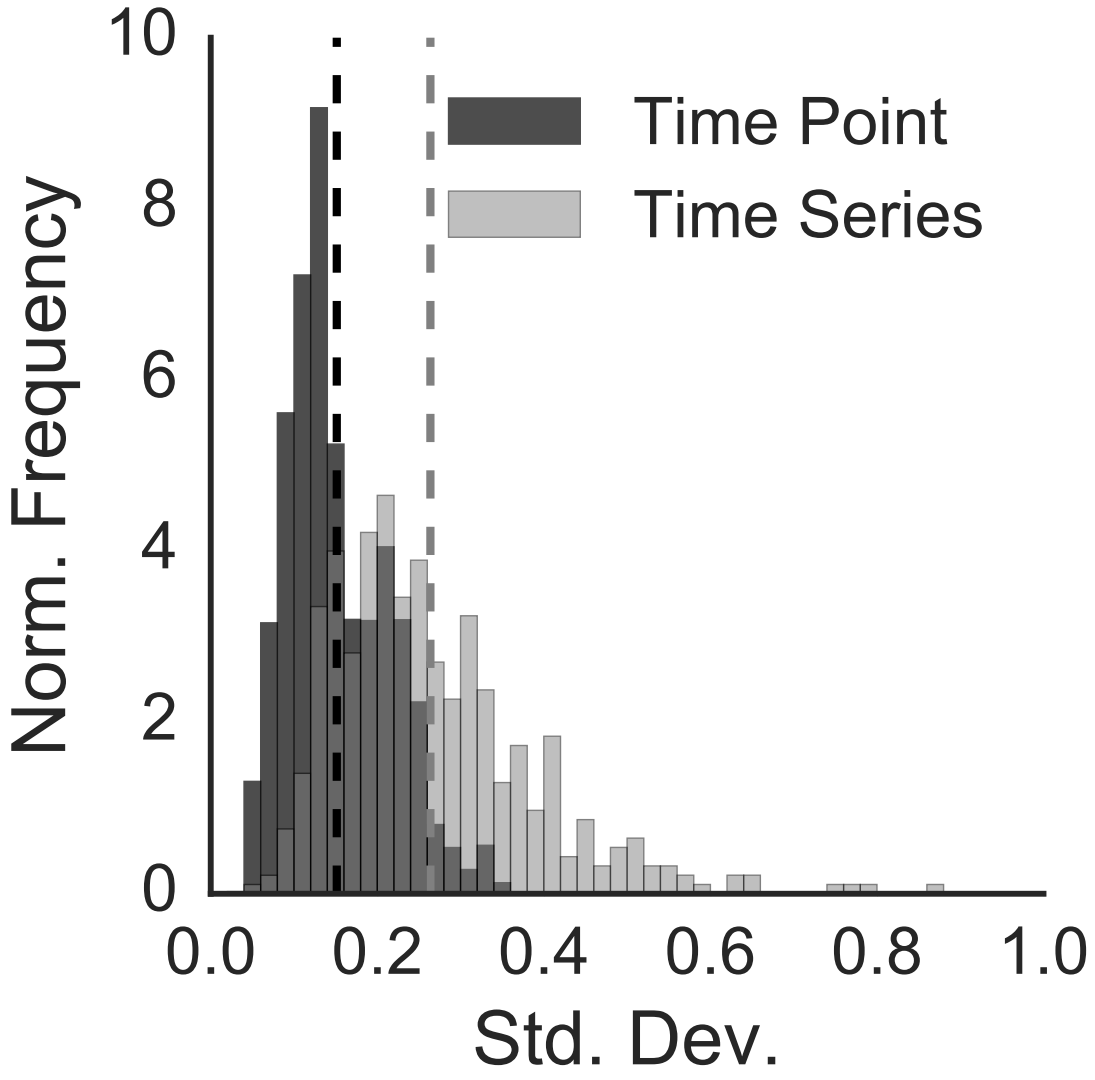

Figure 4.S10: The standard deviation of arrhythmic time series provides an approximation of the standard deviation of time points, as shown for the Hughes *et al.* dataset sampled every 1 h [51]. Vertical lines represent the means of the two distributions.

CHAPTER 5

DIFFERENTIAL RHYTHMICITY

5.1 Abstract

Circadian rhythms are found in organisms from all kingdoms of life and have been implicated in many diseases. Circadian rhythms are often studied by exploring the rhythmic behavior of gene expression. Many studies focus on differences in rhythmic gene expression due to perturbations such as genetic manipulation, exposure to light, or modification of feeding. Differences in rhythmicity and peak expression (phase) currently are determined by differences in point estimates of p-value significance and peak location. This results, however, in inaccurate classification. Ideally, a method for comparing rhythmicity between two conditions could propagate the experimental uncertainty found at the expression level to provide estimates of the uncertainty in the rhythmicity and phase. We statistically compare rhythmicity and phase measurements generated via parametric bootstrapping of the observed time series. Using ANOVA to assess noisiness of the time series, we use simulated data to assess the expected uncertainty in these measurements, allowing for significance testing of differential rhythmicity and phase. This method, Bootstrap Differential Rhythmicity (BDR), more robustly identifies differentially rhythmic or phase shifted transcripts than comparing rhythmicity p-values to a threshold. We show that the use of BDR results in more sensible biological interpretation and more accurately reflects the noise and error present in the experimental data. We compare to Light-Dark and Dark-Dark mouse liver RNA expression, showing that BDR provides insight into differential rhythmicity and phase shifts, while an alternative method, DODR, provides information on changes in amplitude. We show in comparison of circadian protein and RNA datasets that increased measurement of circadian protein levels is necessary to identify rhythmic protein levels independent of RNA expression rhythmicity, for which we find no evidence. As circadian comparisons across conditions and

tissues become increasingly important, use of BDR will allow for the effective identification of differential rhythmicity and phase. DiffRhyth is implemented in Python, and is available on GitHub at http://github.com/alanlhutchison/DiffCyc.

5.2 Introduction

Biological rhythms are recognized as critical aspects of cellular, physiological, organismal, and ecological function. Cell division cycles are often periodic, hearts pump consistently using electric rhythms, circadian behavior is enforced by the rotation of the Earth, and interconnected populations of predator and prey periodically vary across seasons and decades. These rhythms adjust and respond to environmental cues and external stimuli, such as temperature and availability of food. Being able to identify these rhythms, as well as detecting differences between these rhythms, is critical to understanding these diverse biological systems.

For concreteness, we focus on circadian rhythms, which are physiological oscillations found in organisms from all kingdoms of life that allow them to anticipate environmental changes from day to night. Circadian rhythms are entrainable to external stimuli and able to run endogenously in the absence of stimuli. Their periods are roughly 24 hours long and weakly sensitive to temperature. These rhythms are reflected at multiple levels of physiology, from organismal behavior to levels of gene expression. Many types of external stimuli can perturb these rhythms, such as light, temperature, feeding, sleep pattern, or genetic manipulation [6]. The impact of these stimuli is often observed via gene expression to understand their effects on cellular and physiological processes.

For time series that are highly temporally sampled, such as luciferase or fluorescence traces, where the standard error in measurements of a time point is small in comparison the difference between time points, identification of rhythmicity is relatively straightforward. It is increasingly common to perform genome-wide high-throughput time series measuring

transcript levels for every gene and even every known isoform of every gene [84, 33, 25]. Experimental and financial limitations, however, limit the frequency with which data may be collected, leading to the need for sophisticated techniques for statistically assessing rhythmicity to overcome noise and observer bias.

Many statistical methods exist that can extract evidence of periodic biological rhythms from time series data, often in conditions of low temporal resolution and high noise [54, 53, 122, 22]. These methods produce a p-value or other equivalent statistic indicating the rhythmicity of a time series, and thresholds are used to classify a time series as rhythmic or arrhythmic. This determination allows for further experiments to focus more specifically on these time series of interest to confirm the finding and understand the role of the biological rhythms.

In comparative circadian studies, it is common to compare rhythmicity test statistics and phase estimates across conditions, using differences between them to assess changes in rhythmicity or phase [33, 84, 25]. These point estimates, however, can fail to account for the uncertainty in these measurements of rhythmicity and phase. Greater classification power for determining changes in rhythmicity and phase can be gained by incorporating this noise.

A method was recently developed for identifying changes in differential rhythmicity, known as DODR (Detection of Differential Rhythmicity) [100]. It assesses if the difference between two time series can be attributed to the addition of a scaled and phase shifted sinusoid to an initial sinusoid. DODR fits one time series to a sine wave and then adds to it another sine wave to attempt to fit the second time series. If the amplitude and phase parameters of the added sine wave are significantly different from 0 as assessed by an ANOVA-based method, then differential rhythmicity is declared. The developers show that this method is preferable to the rhythmicity p-value comparison method (p-value threshold, PVT) for identifying changes in amplitude.

Differences in amplitude and phase are important features to identify between time se-

ries. However, detecting a difference in amplitude or phase is not necessarily equivalent to detecting a change in rhythmicity in the sense defined by existing rhythm detection methods [99, 53]. Here, we present a method that directly compares rhythmicity test statistics, takes into consideration relative noise-to-amplitude of time series, and allows for asymmetric time series that do not fit well to a sinusoid. Our method also has the ability to compare phase shifts. We only do this in the case where differential rhythmicity is not identified, since an arrhythmic time series cannot have a clearly defined phase.

For detecting and assigning rhythmicity, we use the recently developed Bootstrap eJTK (BooteJTK) method [54], although the framework we present here can be applied to any rhythm detection method. BooteJTK incorporates an empirical Bayes method to better estimate the measurement variance and runs eJTK [53] on parametric bootstrap resamples of the time series to propagate the uncertainty in measurement into uncertainty in rhythmicity and phase. eJTK finds the best non-parametric correlation (Kendall's τ) between a time series of interest and reference waveforms differing in phase and time from peak to trough (width or asymmetry). eJTK uses other reference waveforms in addition to a sinusoid, giving it and, in turn, our method, improved sensitivity to non-sinusoid time series.

We present a method, Bootstrap Differential Rhythmicity (BDR), that quantifies the uncertainty in our ability to determine that rhythmicity and phase has significantly changed in a gene's expression between two conditions, allowing for more nuanced comparison of gene expression rhythmicity and phase. We introduce the use of ANOVA not as a rhythm detection method itself, but as a means to judge the noisiness of a time series, which allows us to determine the variance we should expect from differences in rhythmicity and phase between two time series. This allows us to perform significance testing and obtain a p-value for differential rhythmicity and differential phase. We can combine these results with those derived by the application of DODR to determine changes in amplitude to better compare changes in rhythmicity across conditions and tissues.

We use our method to revisit a comparison of Light-Dark (LD) mouse liver RNA expression [57] and Dark-Dark (DD) mouse liver RNA expression [51] performed by Thaben *et al.* [100]. We find that combining BDR with DODR allows for greater understanding of the difference between the two conditions, with BDR providing insight into changes in rhythmicity and phase and DODR providing insight into changes in amplitude.

We next consider three studies comparing circadian protein levels in LD and DD conditions in mouse liver and in LD conditions in the cyanobacteria *Synechococcus elongatus*. While previous analyses employing the naïve PVT method had found 20%-50% of genes with rhythmic protein levels had no corresponding RNA expression rhythmicity, we find with our more rigorous method that the data support only 10% of genes having protein levels that are rhythmic without corresponding RNA expression rhythmicity. We do find substantial differences in phase between rhythmic RNA expression and rhythmic protein levels, indicating post-transcriptional and post-translational regulation, including in genes involved in the heat shock stress response and xenobiotic detoxification. We conclude that greater sampling rates and replicate numbers are necessary for protein level rhythmicity to be interrogated with the same rigor of RNA expression rhythmicity under current high-throughput protein quantification technology. Our method, BDR, provides a rigorous significance testing method to better understand changes of rhythmicity and phase which will become of increasing importance as studies explore circadian differences across conditions and tissues.

5.3 Methods

5.3.1 Naïve Comparison Approach

Currently, the field of circadian biology relies on very basic comparisons of point estimates to determine differential rhythmicity or phase shifts. For example, the expression of gene X is measured in two separate conditions. In condition A the p-value of rhythmicity is 0.041 and

in condition B it is 0.051. The gene is considered rhythmic in A but not in B. Similarly, for gene Y the p-values are 0.039 and 0.049. The gene is considered rhythmic in both conditions. A similarly simple method is used for assessing phase shifts. For the same gene Y the peak expression (phase) is estimated to occur at zeitgeber time (ZT)0 in condition A and at ZT2 in condition B. Using this naïve method, a phase shift would be considered to have occurred between the two conditions, regardless of how broad or narrow the peaks of A or B were. This approach is used in many studies in the circadian literature [25, 63]. We refer to this approach as the naïve method, although it is also referred to as the p-value threshold, PVT, by Thaben *et al.*, so we will use that terminology as well. We use BooteJTK to perform the naïve comparison of rhythmicity empirical p-values and phase estimates, as it has been shown to outperform other rhythm detection methods in our previous work [54, 53]. For the data sampled every 2 h, we search for phases from 0 h to 22 h every 2 h, and we search for asymmetries from 2 h to 22 h every 2 h.

5.3.2 Differential Rhythmicity Method

Identification of Differential Rhythmicity

Our comparison of time series is based on the difference in the $\tilde{\tau}$ statistic from BooteJTK ($\Delta\tilde{\tau}$). To obtain a p-value for this difference, we need to know the expected variance in the $\Delta\tilde{\tau}$ and ΔPh test statistics in the case where there is no change in rhythmicity and no change in phase. We find that the variance in the $\Delta\tilde{\tau}$ test statistics is dependent on the noisiness of the time series (Figs. 5.S1). To quantify noisiness, we use the ANOVA F-statistic, which compares the variance between replicate time points to the variance across the entire time series. To construct a model for the expected variance as a function of the ANOVA score, we generate two pairs of 1000 time series datasets with the sampling rate matching the experimental time series dataset. Each pair of time series has noise added at a randomly chosen noise level from 0.1 to 0.9. We then fit the ANOVA $-\log_{10}$ p-value and

the expected variance of the test statistics (Figs. 5.S1B and D). This allows us to determine the expected variance in $\Delta\tilde{\tau}$ if we compare that time series to one with similar noise level and phase. Repeating this procedure on a second time series, we can assess the expected variance for that time series as well. We then average the two variances to obtain the expected variance (and standard deviation) if the two time series had similar rhythmicity. By dividing the expected standard deviation by the observed $\Delta\tilde{\tau}$ for the two time series, we obtain a Z-score, which we then fit to a normal distribution of mean 0 and variance 1 to obtain a p-value for the observed $\Delta\tilde{\tau}$. We can similarly use this process to obtain a Z-score and p-value for the difference in phase ΔPh. When performed on time series data for which no differences in rhythmicity or phase exist (cosine time series with the same level of noise added and no phase shift), the $\Delta\tilde{\tau}$ and ΔPh Z-scores fit to a normal distribution and the p-values are uniformly distributed from 0 to 1. The ΔPh p-value deviates toward being overly conservative around $p = 0.4$, which is beyond the general range of interest for determining significance from p-values.

5.3.3 Comparison of Naïve and Differential Rhythmicity Methods

There are several potential ways to compare the rhythmicity test statistics and phase estimates from BooteJTK. The most direct is to subtract the values directly ($\Delta\tilde{\tau}, \Delta\bar{Ph}$). An alternative would be a T-statistic (T_τ, T_{Ph}) to incorporate the variance of the estimates that are provided by the individual bootstrap resampling scores. If the normality of the individual bootstrap resampling scores was in question, the non-parametric Mann-Whitney U-statistic (U_τ, U_{Ph}) could be used instead. We can use the above method to obtain p-values for T-statistics, but not for Mann-Whitney U-statistics, since they are not symmetrically distributed in the null case.

We compared our BDR $\Delta\tilde{\tau}$ test statistics against BooteJTK rhythm detection p-values to determine which was better at classifying differential rhythmicity. We model time series

of different rhythmicities by adding noise to evenly spaced time points (2 per time point, 12 time points per period) from a cosine with noise drawn from a normal distribution to each point. The standard deviation of the normal distribution is varied to change the overall rhythmicity of the modeled time series. The ratio of the standard deviation to the amplitude of the underlying cosine waveform, referred to here as the 'noise level', scales inversely with the rhythmicity of the modeled time series. True negatives were classified as time series with the same noise level, whereas true positives were classified as time series with different noise levels. We present our comparison of methods in two ways. The first is using the receiver operating characteristic (ROC), which shows how the relationship between the False Positive Rate and the True Positive Rate changes for a given threshold for a method (Figs. 5.1A and C). Ideally, the False Positive Rate would be low while the True Positive Rate would be high, meaning the results would be in the upper left area of the plot. If the method were no better than random, the results would have a slope of 1 (the diagonal line in Figs. 5.1A and C). The second way with which we present our results is the Matthews Correlation Coefficient (MCC) [72] for a given threshold (Figs. 5.1B and D). The MCC provides a single metric from -1 to 1 measuring the quality of a classifier. An MCC of 1 indicates perfect classification, while an MCC of 0 indicates no better than random.

For distinguishing time series with a noise level of 0.05 from those with a noise level of 0.10, the $\Delta\tilde{\tau}$ statistic outperforms the BooteJTK p-values as well as T-statistics and Mann-Whitney U statistics based on $\Delta\tilde{\tau}$ (Figs. 5.1A and B). For distinguishing phase shifts, we first required that the absolute difference in phase $|\Delta\bar{P}h|$ be greater than 1. That requirement alone was our naïve comparison. The additional condition was the use of the p-value threshold of the $|\Delta\bar{P}h| > 1$ and T-statistic. Across a range of shifts for a noise level of 0.1, $\Delta\bar{P}h$ p-values outperform the naïve method. It also outperforms the phase T-statistics and Mann-Whitney U statistics (Figs. 5.1C and D).

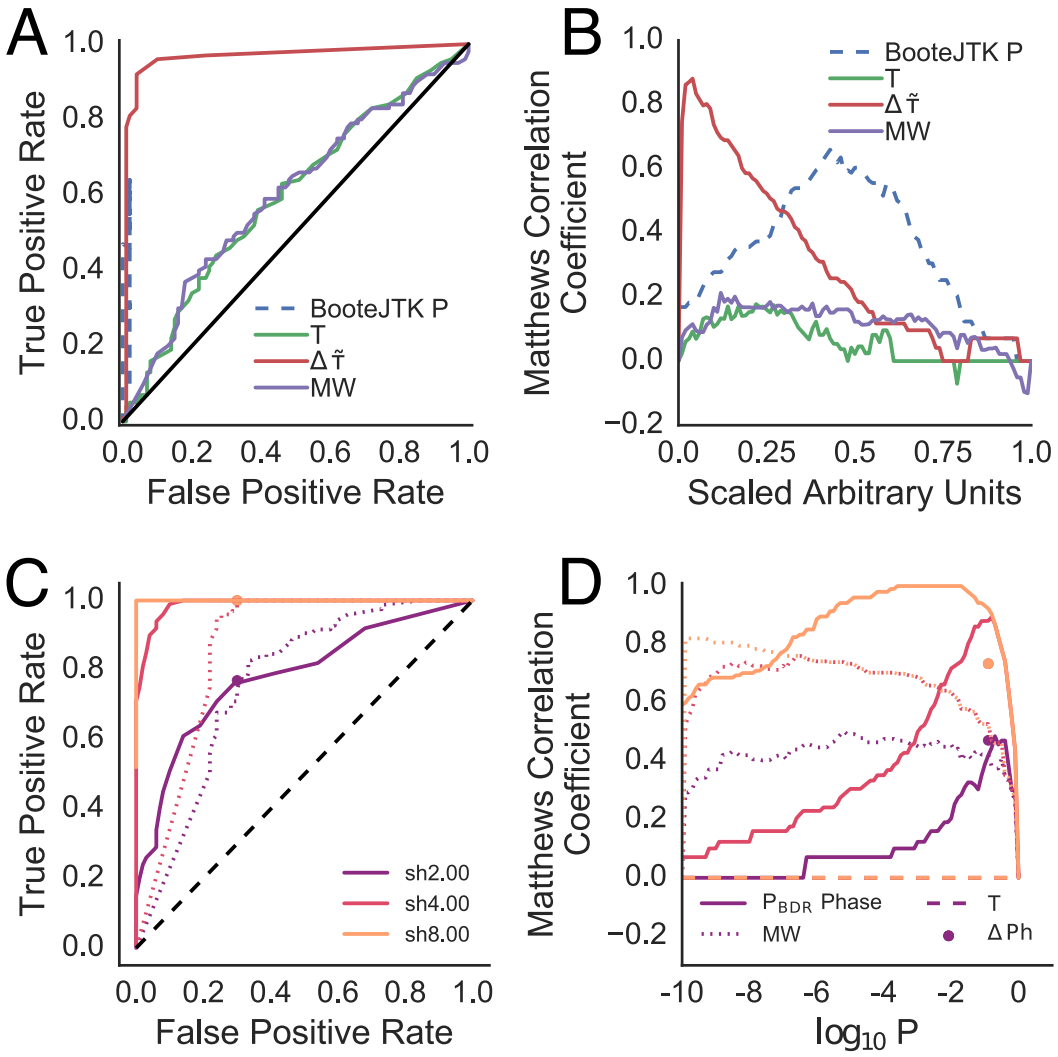

Figure 5.1: (A)-(B) Comparison of differential rhythmicity classification ability of BDR $\Delta\tilde{\tau}$, BDR $\Delta\tilde{\tau}$ T-statistic (T), BDR $\Delta\tilde{\tau}$ Mann-Whitney U statistic (MW), and BooteJTK p-value comparison comparing two datasets of 100 time series each with a noise level of 0.05 and 0.10. The BDR $\Delta\tilde{\tau}$ statistic outperforms the other methods as seen by the comparison of False Positive Rate and True Positive Rate in the receiver operating characteristic (ROC) (A) and the Matthews Correlation Coefficient (B), where the metric is defined from the minimum of the statistic to the maximum for each statistic. (C)-(D) Comparison of phase shift classification ability between BDR $\Delta\bar{P}h$ p-value ($P_{\bar{B}DR}$ Phase), BDR $\Delta\bar{P}h$ T-statistic (T), BDR $\Delta\bar{P}h$ Mann-Whitney U statistic (MW) and $|\Delta\bar{P}h| > 1$ (points) comparing datasets with 100 time series each with a noise level of 0.1 and phase shifts. While the BDR phase p-value is plotted as shown by the horizontal axis, the other two statistics are plotted from their minima to their maxima. The BDR phase p-value outperforms the other methods across a range of phase shifts as seen by ROC (C) and MCC (D).

5.4 Results

5.4.1 Comparison of Light-Dark and Dark-Dark Circadian RNA Expression with BDR and DODR

To better understand the abilities of the differential rhythmicity methods, we re-examine the comparison performed by [100] of mouse liver RNA expression under Light-Dark and Dark-Dark conditions.

We compared total mouse liver RNA expression under Light-Dark (LD, ZT) conditions from Jouffe *et al.* [57] with mouse liver RNA expression under Dark-Dark (DD, CT) conditions from Hughes *et al.* [51], downsampling the data to make it the same sampling rate as the Jouffe dataset (every 2 h, on even ZT or CT hours).

Using the R package *gcrma* [115] to normalize the data (GEO GSE11923 and GSE33726) and removing probes from each data set with constant expression, we set a mean expression threshold in the Hughes dataset of 7.5 units to exclude low expression probes. This brought the total number of probes under consideration to 7537, which is in line with the analysis performed by Thaben *et al.* [100]. Of these probes, 5531 were rhythmic with FDR of 0.05 in either condition using BooteJTK.

First, we compared the results with BDR to the naïve approach for rhythmicities and used the absolute difference in phase for phases (Fig. 5.2A, Tab. 5.1). An initial analysis of the noisiness of the time series (measured by the ANOVA p-value) in each dataset shows that the datasets are fairly equivalent in terms of the measurement noise relative to amplitudes of the time series (Fig. 5.S4). BDR identifies fewer probes as significantly differentially rhythmic than the naïve method does and finds fewer genes total and proportionally that are rhythmic only under the DD condition (Hughes) than under the LD condition (Jouffe). That more genes are rhythmic in the LD condition than the DD condition is not surprising given that LD cycling is a strong driver of rhythmicity. Of the probes that maintain rhythmicity

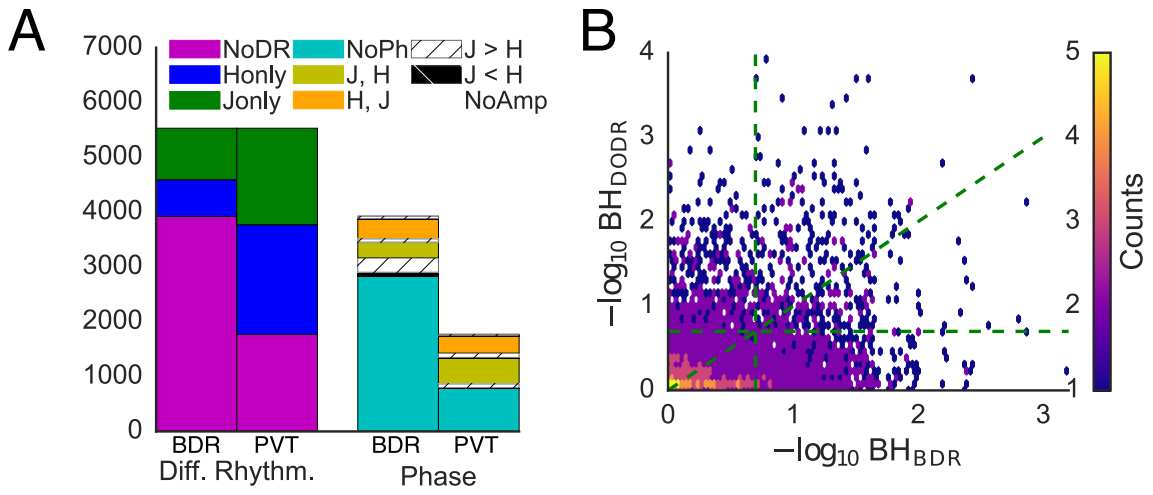

Figure 5.2: (A) Comparison of BDR (left column) and naïve method (right column) on comparison of Jouffe and Hughes datasets for Differential Rhythmicity (Diff. Rhythm.) and Phase. The probes compared for differential phase are those which have no differential rhythmicity. The values for this subfigure are in Table 5.1. Rhythm detection was performed with BooteJTK with a threshold of 0.05. Differential rhythmicity and differential phase thresholds for BDR and DODR were set at 0.2. Abbreviations: NoDR: no differential rhythmicity; OnlyH: probes only rhythmic in Hughes DD dataset; OnlyJ: probes only rhythmic in Jouffe LD dataset; NoPh: no differential phase; J, H; probe peaks in Jouffe *et al.* LD dataset before it peaks in Hughes *et al.* DD dataset; H, J: probe peaks in Hughes *et al.* DD dataset before it peaks in Jouffe *et al.* LD dataset; J > H: probe has higher Std. Dev. in Jouffe *et al.* LD dataset than in Hughes *et al.* LD dataset and $\mathrm{BH}_{DODR} < 0.2$ H > J: probe has lower Std. Dev. in Jouffe *et al.* LD dataset than in Hughes *et al.* LD dataset and $\mathrm{BH}_{DODR} < 0.2$, NoAmp: $\mathrm{BH}_{DODR} > 0.2$. (B) BDR $-\log_{10}$ BH-values plotted against DODR $-\log_{10}$ BH-values. The slope of the diagonal line is 1, and the vertical and horizontal lines indicate BH-values of 0.2. Numbers of probes in each quadrant indicated in Table 5.2.

across conditions, BDR finds fewer probes differ significantly in phase. BDR finds a more even distribution of DD probes phase advanced relative to LD probes and LD probes phase advanced relative to DD probes than the PVT method does.

While some overlap between the BDR and DODR results exists, there are differences in the probes that are classified as significantly distinct from the Jouffe *et al.* dataset and the Hughes *et al.* dataset (Fig. 5.2B, Tab. 5.2). For each intersection of BDR and DODR results, example time series can be found in Figure 5.S5. Out of 5531 probes rhythmic in either the Jouffe *et al.* dataset or in the Hughes *et al.* dataset, 460 probes are differentially

	BDR	J<H	J>H	NoAmp	PVT	J<H	J>H	NoAmp
Jouffe only	940				1761			
Hughes only	666				1997			
No Diff. Rhyth.	3925				1773			
No DR, No Phase Diff.	3162	76	258	2828	866	15	69	782
No DR, H, J	413	16	44	353	345	10	23	312
No DR, J, H	350	9	59	282	560	18	76	466
Hughes rhythmic	3770							
Jouffe rhythmic	3534							

Table 5.1: Comparison of Jouffe *et al.* dataset and Hughes *et al.* dataset with BDR and PVT. Results are shown graphically in Figure 5.2A. Rhythm detection was performed with BooteJTK with a threshold of 0.05. Differential rhythmicity and differential phase thresholds for BDR and DODR were set at 0.2. Abbreviations: BDR: Bootstrap Differential Rhythmicity DR, differentially rhythmic; H>J: Hughes *et al.* Std. Dev. > Jouffe *et al.* Std. Dev. and $\mathrm{BH}_{DODR} < 0.2$; J>H: Hughes *et al.* Std. Dev. < Jouffe *et al.* Std. Dev. and $\mathrm{BH}_{DODR} < 0.2$; H, J: probe peaks in Hughes *et al.* before Jouffe *et al.* J, H: probe peaking in Jouffe *et al.* before Hughes *et al.* NoAmp: $\mathrm{BH}_{DODR} > 0.2$; PVT: p-value threshold naïve method,

rhythmic by both BDR and by DODR (Fig. 5.S5A). 1328 probes are differentially rhythmic under BDR but have no distinct amplitude or phase change as identified by DODR. These probes have lost rhythmicity due to the variability in replicate time points increasing, the distance between some points in the series decreasing, or the order of the points relative to one another changing as to become less rhythmic (Fig. 5.S5B). 540 probes are not differentially rhythmic under BDR exhibiting changes in phase or amplitude and so are identified by DODR. These probes maintain distinct separation of their time points such that they continue to be rhythmic though their amplitude or phase change (Fig. 5.S5C). 3203 probes are not differentially rhythmic under BDR or DODR (Fig. 5.S5D).

The probes that are not identified as differentially rhythmic by BDR could potentially have differences in phase and/or amplitude. Probes identified as differentially rhythmic by definition cannot have differences in phase, since one time series is arrhythmic and therefore does not have a clearly defined phase. We show these in the right set of columns in Figure 5.2A, with the bars filled in to represent their phase classification by BDR and classification

	BH_{DODR} <0.2	BH_{DODR} >0.2	BH_{DODR} <0.2	BH_{DODR} >0.2
BH_{BDR} <0.2	460	1328		
BH_{BDR} >0.2	540	3203		
Phase BH_{BDR} <0.2			149	614
Phase BH_{BDR} >0.2			418	2744

Table 5.2: Comparison of BDR and DODR on Jouffe *et al.* and Hughes *et al.* data. Comparison of BDR differential phase is done for probes where the differential rhythmicity BH_{BDR} >0.2.

by DODR.

Out of the 3203 probes not identified as differentially rhythmic by BDR, 149 show differential phases by BDR and differences by DODR. 614 probes show differential phase by BDR but no differences by DODR. Of these probes, their average phase shift is 10.9 h with a standard deviation of 5 h, making it surprising they were not observed by DODR (Fig. 5.S6A). We found that, of different ways of measuring change in amplitude, such as minimum expression subtracted from or divided from maximum expression, logarithmic difference of the standard deviations provided the best proxy for the DODR BH score. Using the $\log_2$ ratio of the difference in standard deviations between the Jouffe time series and the Hughes time series as a proxy for amplitude differences, we find these probes have an average log ratio of 0.3 with a standard deviation of 0.5 (Fig. 5.S6B). 418 probes show no differential phase by BDR but differences by DODR. The average phase shift is -1.0 h with standard deviation of 2.9 h (Fig. 5.S6C), which suggests the differences DODR is detecting are in amplitude. We find these probes have average log ratio of 0.4 and standard deviation of 0.7 (Fig. 5.S6D). 2744 of the probes remaining show neither differential phase by BDR nor differences by DODR. These values can be seen in Table 5.2.

To better understand the differences between BDR and DODR, we can look at the distributions of rhythmicity p-values, standard deviations of the time series (a proxy for differences in amplitude), and phases relative to the three differential rhythmicity FDR values we use: the BDR BH, the DODR BH, and the BDR Phase BH. In Figure 5.3, we

compare these metrics for the probes from the Jouffe *et al.* dataset and Hughes *et al.* dataset. Probes with no differences in metrics appear near the diagonal, whereas probes with large differences between the conditions appear far from the diagonal. The points are color-coded by the BDR and DODR BH scores, to show how these scores relate to the differences in the metrics.

While the BDR BH value clearly classifies increasing differences in rhythmicity (Fig. 5.3A), the DODR BH value does not do so as distinctly (Fig. 5.3B). Namely, the color ordering relates to the distance from the diagonal in Fig. 5.3A but not Fig. 5.3B. The DODR BH value, however, is better at classifying differences in time series standard deviation (Fig. 5.3D) than the BDR BH value (Fig. 5.3C), which can be seen by looking at the points around (1.0, 0.1). The BDR Phase BH values are much better at classifying differences in phase (Fig. 5.3E) than the DODR BH values are (Fig. 5.3F), which can be seen for the points around (5, 15).

These results make sense in the context of how BDR and DODR operate. DODR fits a sine wave to one time series and then adds a second sine wave, using the significance of the difference of the amplitude and phase parameters from 0 as a proxy for difference in rhythmicity. BDR uses the rhythmicity test statistic and mean phase estimate themselves to generate p-values, but amplitude is not explicitly considered. Given these results, we suggest that DODR be used for identifying changes in amplitude, but BDR be used for identifying changes in rhythmicity and changes in phase.

5.4.2 Comparison of Circadian Protein and RNA Time Series Rhythmicity

Having compared BDR and DODR, we now focus on a dataset where changes in amplitude are not applicable but identifying changes in rhythmicity and phase still are: comparison of RNA and protein expression rhythmicity.

An open question in the field of circadian biology is the extent to which rhythms in

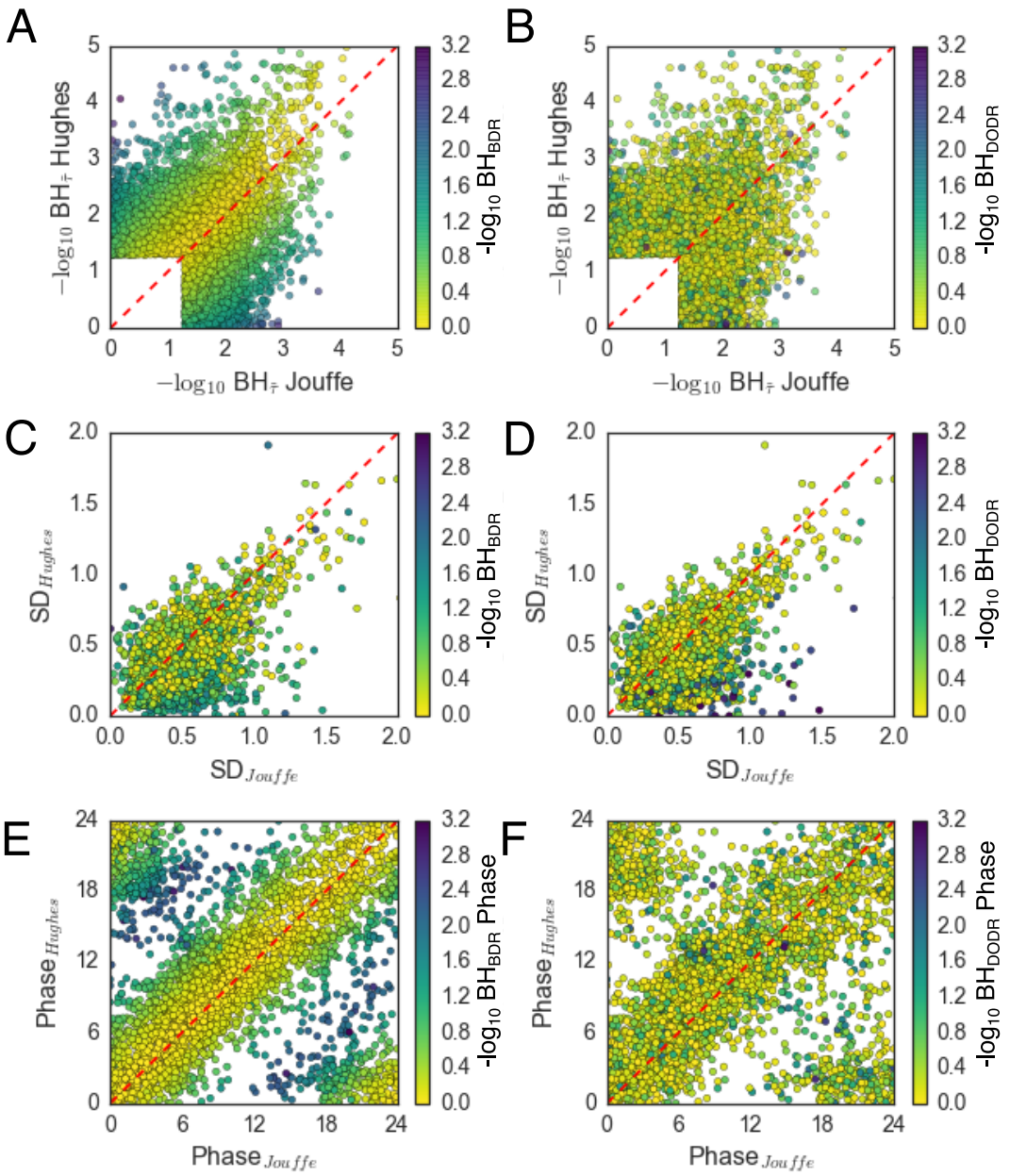

Figure 5.3: Rhythmicity measures from the Jouffe and Hughes dataset plotted against each other with the colors indicating differential rhythmicity measures. (A) BooteJTK BH-values for Jouffe and Hughes probes, colors indicate $-\log_{10}$ BDR BH-values. (B) BooteJTK BH-values for Jouffe and Hughes probes, colors indicate $-\log_{10}$ DODR BH-values. (C) Standard deviation of time series for Jouffe and Hughes probes, colors indicate $-\log_{10}$ BDR BH-values. (D) Standard deviation of time series for Jouffe and Hughes probes, colors indicate $-\log_{10}$ DODR BH-values. (E) Phases for Jouffe and Hughes probes, colors indicate $-\log_{10}$ BDR PhDiff BH-values. (F) Phases for Jouffe and Hughes probes, colors indicate $-\log_{10}$ DODR BH-values.

RNA gene expression, which are easily measurable via amplification and next-generation sequencing, are reflected in rhythmicity of protein levels and activity. Several studies have profiled protein rhythmicity levels and compared them with RNA rhythmicity levels under similar conditions. Mauvoisin *et al.* and Robles *et al.* use *in vivo* Stable Isotope Labeling by Amino acids in Cell culture (SILAC) technology with mass spectrometry to obtain protein levels in mouse liver tissue, which they then compare to RNA expression. Mauvoisin *et al.* perform their study in LD conditions, finding that 4% of measured proteins are rhythmic and 50% of those did not have observable RNA rhythmicity [73]. Robles *et al.* conduct their study in DD conditions, finding that 6% of measured proteins are rhythmic and 20% of those do not have observable RNA rhythmicity [91]. In a separate system and method, Guerreiro *et al.* use tandem mass spectroscopy to compare the protein and RNA levels in the cyanobacteria *Synechococcus elongatus*, finding 5% of the identified proteins to be rhythmic and 35% of those proteins to lack RNA expression rhythmicity [42]. In all three of these studies the comparisons were made by naïvely comparing p-values between RNA and protein time series, and could benefit from application of the BDR method.

Before examining the results of BDR, we can gain insight into the datasets by examining the ANOVA p-values for the probes in the different datasets (Fig. 5.S7). In all three datasets the sampling rate of the protein time points is lower than the sampling rate of the RNA time points. Additionally, it is difficult to quantify the degree to which the accuracy of protein measurements matches the accuracy of RNA measurement. The ANOVA p-values provide information regarding both factors together; in all three datasets the $-\log_{10}$ p-values are often higher for the RNA than for the protein, which means that the uncertainty in measurement is lower for the RNA datasets than for the protein datasets. This can be seen by comparing the p-values for each probe in Figs. 5.S7A, C, or E and for a distribution of p-values in Figs. 5.S7B, D, and F.

For each dataset, we used high rhythmicity BH-value thresholds matching those of the

original studies: 0.33 for Robles *et al.*; 0.25 for Mauvoisin *et al.*, and 0.09 for Guerreiro *et al.* (which is equivalent to a p-value of 0.05), including probes that were rhythmic in at least RNA or protein. We do expect genes identified as rhythmic to be different in our analysis compared to those studies, however, since we are analyzing the rhythmicity and phase of the data with BooteJTK, which was not used in the original studies. Furthermore, we set thresholds for minimal standard deviation necessary to be considered for rhythmic analysis: 0.10 for Robles *et al.*; 0.12 *et al.* for Mauvoisin; 0 for Geuerriro *et al.*). For all three datasets, 0.05 was used as the BH-value threshold for differential rhythmicity and differential phase.

Broadly, the BDR results show fewer differences between RNA and protein than the naïve method does (Fig. 5.4 and Tab. 5.3). Out of 2323 rhythmic genes in Mauvoisin *et al.*, 126 genes are rhythmic in RNA only, no differential rhythmicity is found for 2170 genes, and 15 genes are rhythmic in protein only. Similar results are found for Robles *et al.*, where, out of 718 rhythmic genes, 91 genes are rhythmic in RNA only, no differential rhythmicity is found for 607 genes, and 20 genes are found to be rhythmic only in protein, and Guerreiro *et al.*, where out of 612 genes found to be rhythmic 35 are rhythmic only in RNA, no differential rhythmicity is found for 577 genes, and 0 genes are rhythmic only in protein (Fig. 5.4A). In all three cases, fewer cases are found of proteins rhythmic without their accompanying RNAs being rhythmic as well, while 0.5%-14% of genes are found to be rhythmic in RNA only. Two potential explanations could account for these results. First, it could be that fewer genes have protein rhythmicity independent of RNA rhythmicity than previously appreciated. Given the different levels of post-transcriptional and post-translational regulation between prokaryotes and eukaryotes, the lack of proteins whose levels are rhythmic independent of RNA expression in Guerreiro *et al.* could be a feature of prokaryotic circadian systems. Second, the noisiness of the protein time series is significantly higher than that of the RNA time series (Fig. 5.S7). This increases the chances that rhythmicity would be identified at the RNA level but not the protein level as well as the chances that rhythmicity only at the protein level would be

missed.

Next we examined the genes not found to be differentially rhythmic for phase shifts between RNA and protein. Such shifts could suggest post-transcriptional regulation that would delay the translation of mRNA into protein or post-translational regulation that would change the rates of degradation to vary the timing of peak protein levels. Mauvoisin *et al.* and Robles *et al.* show similar patterns of phase shifts, with more genes having the RNA expression peak precede the protein level peak than vice versa (606 vs. 533 for Mauvoisin *et al.*, 214 vs. 146 for Robles *et al.*) (Fig. 5.S8). A greater proportion of genes from the Mauvoisin *et al.* have no change in phase as well as no change in rhythmicity than in Robles *et al.*. In the Guerreiro *et al.* dataset, equal numbers of genes had the RNA expression peak precede the protein level peak as vice versa (163 in both cases), while slightly more genes showed no difference in phase as well as no difference in rhythmicity (251). The similarity between the Guerreiro *et al.* phase shift results and the Mauvoisin *et al.* and Robles *et al.* results is perhaps surprising given the differences in prokaryotic vs. eukaryotic regulation, though differences in the two systems appear to yield similar results regarding the relative circadian behavior of RNA and protein.

Comparing the Mauvoisin *et al.* and Robles *et al.* studies, we identify genes whose behavior from RNA to protein matches in both studies (Fig. 5.5 and Tab. 5.4). 378 genes are found to be rhythmic in at least one condition in each dataset. Of those, 6 are rhythmic in only RNA in both, 291 are not found to be differentially rhythmic, and 0 are found to be rhythmic in protein only. Of those genes not differentially rhythmic, 47 show no change in phase. These genes are enriched for many functional annotations as assessed by the DAVID webtool [49, 50]. Particularly interesting are the genes Cyp17a1, Cyp2b10, Cyp2c70, Cyp2f2, Cyp39a1, and Cyp2c37, which are related to cytochrome P450 (BH=2.0×10^{-4}), heme binding (BH=1.7×10^{-3}), and the metabolism of xenobiotics (BH=6.5×10^{-3}), as well as related to the endoplasmic reticulum along with Bcap29, Ykt6, Zw10, and Pdia6

	Robles		Mauvoisin		Guerreiro	
	BDR	PVT	BDR	PVT	BDR	PVT
NoDR	607	76	2170	256	577	350
Prot. only	20	32	15	58	0	59
RNA only	91	610	126	1998	35	203
NoDR, NoPh	247	17	1035	41	251	45
NoDR RNA, Prot.	214	49	606	115	163	159
NoDR Prot., RNA	146	10	533	100	163	135
Protein-rhythmic	108		314		409	
RNA-rhythmic	686		2253		553	

Table 5.3: Results for differential rhythmicity and phase analysis for Robles *et al.*, Mauvoisin *et al.*, and Guerreiro *et al.*. Rhythm detection thresholds for BH were 0.33 for Robles, 0.25 for Mauvoisin, and 0.09 for Guerreiro. Differential rhythmicity and phase thresholds for BH were 0.05.

(BH=9.7 $\times$ 10$^-$3). 26 genes have the RNA peak before the protein peak but show no enrichment for functional annotation. For 15 genes the protein level peak occurs before the RNA expression peak, and these genes are enriched for secondary metabolite biosynthesis (BH=0.053).

5.5 Discussion

Our method, Bootstrapping Differential Rhythmicity (BDR), provides a significance test to assess whether two time series have changed rhythmicity or phase. BDR uses the bootstrapped $\tilde{\tau}$ value from Bootstrap eJTK, which propagates the uncertainty in expression levels into uncertainty in the rhythmicity and phase of the time series. This gives us sensitivity to differential rhythmicity between non-sinusoidal waveforms. We introduce the concept of using ANOVA not to detect rhythmicity itself, as has been done previously [59], but instead as a method to measure the noisiness of time point measurements relative to the amplitude of the full time series. By identifying the noisiness of the two time series in this manner and comparing them to time series simulated to reflect the experimental time series, we can assess the expected uncertainties in the differences between the rhythmicities and phases of

	#	Genes overlapping in differential behavior between Mauvoisin *et al.* and Robles *et al.*
NoDR	291	
Prot. only	0	
RNA only	6	Fmo2, Hsd3b5, Evi5, Serpinb1a, Mcu, Tbc1d15
NoDR, NoPh	47	
NoDR, RNA, Prot.	26	
NoDR, Prot., RNA	15	Nudt8, Bet1, Ptprd, Cyp4f13, Fbxo3, Bsdc1, Ostc, Anp32e, Me1, Zc3hav1, Atp6v1g1, Mrps10, Cyp3a25, Asap1, F9
PVT, NoDR	18	Tdo2, Gabarapl1, Clpx, Cyp2e1, Lgals3bp, Alas1, Golim4, Pck1, Cyp3a25, Abcc2, Tat, Slc38a4, Dnaja1, Slc38a3, Crot, Slc7a2, Tfrc, Dap
PVT, Prot. only	0	
PVT, RNA only	271	
PVT, NoDR, NoPh	0	
PVT, NoDR, RNA, Prot.	9	Clpx, Tdo2, Gabarapl1, Crot, Slc7a2, Cyp2e1, Dnaja1, Dap, Tfrc
PVT, NoDR, Prot., RNA	1	Cyp3a25
Protein-rhythmic	20	Pck1, Gabarapl1, Slc38a4, Cyp2e1, Lgals3bp, Alas1, Golim4, Tdo2, Abca1, Cyp3a25, Abcc2, Tat, Clpx, Dnaja1, Slc38a3, Crot, Slc7a2, Vim, Tfrc, Dap
RNA-rhythmic	353	

Table 5.4: Genes overlapping in differential behavior between Mauvoisin *et al.* and Robles *et al.*. For categories with no more than 20 genes the gene names are shown.
Abbreviations: PVT: p-value threshold method; DR: differential rhythmicity; X, Y: peak in dataset X is followed by peak in dataset Y

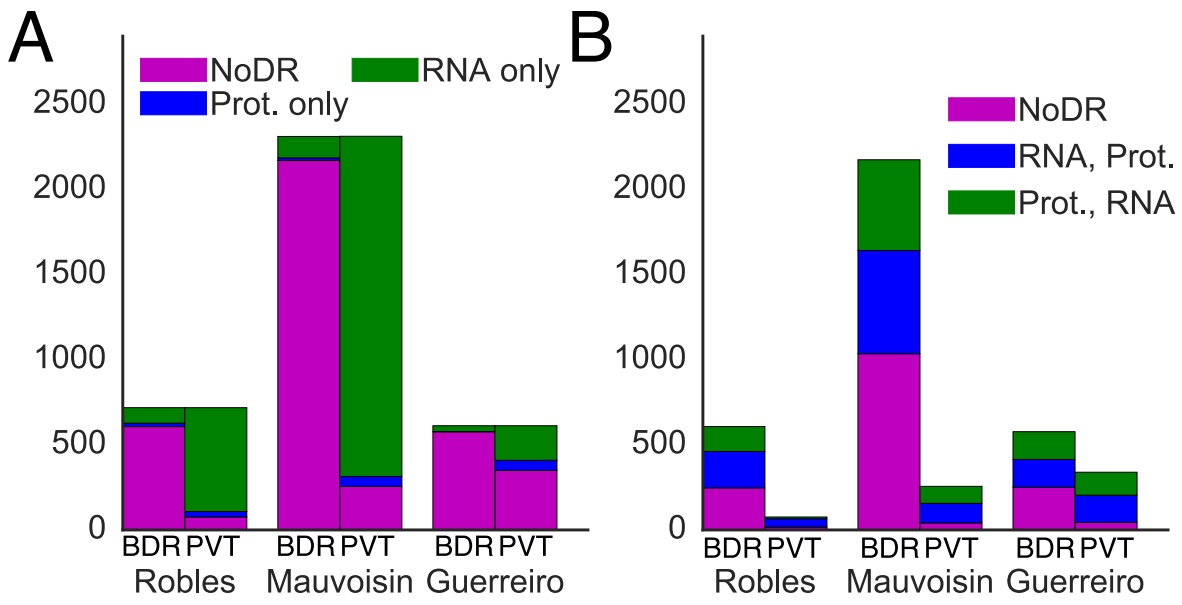

Figure 5.4: (A) Differential rhythmicity results comparing RNA and protein using BDR (left columns) or the naïve method (right columns) across the three different protein-RNA comparison studies. (B) Differential phase results in the same form as (A). The height of the full stacked bars for each study and method match the number of 'NoDR' condition (magenta) of (A).

the time series. This allows us to accurately apply significance testing and calculate p-values for the differences in rhythmicity and phase.

We have shown, in application to simulated data, that our method outperforms simply comparing rhythmicity p-values and provides separate information from an alternative method which fits the time series to sinusoids, DODR [100]. Comparing LD and DD mouse liver RNA time series datasets, we find application of BDR provides greater insight into differences in rhythmicity and phase in combination with DODR, which provides information regarding amplitude changes that BDR does not provide.

We apply BDR to three separate studies comparing protein and RNA circadian characteristics, and we find that the data do not support the existence of many genes whose protein levels are rhythmic but whose RNA expression is not rhythmic. These results could indicate the rarity of exclusive RNA rhythm-independent protein level rhythms. Alternatively, it could be that the high-throughput protein quantification technology used, *in vivo*

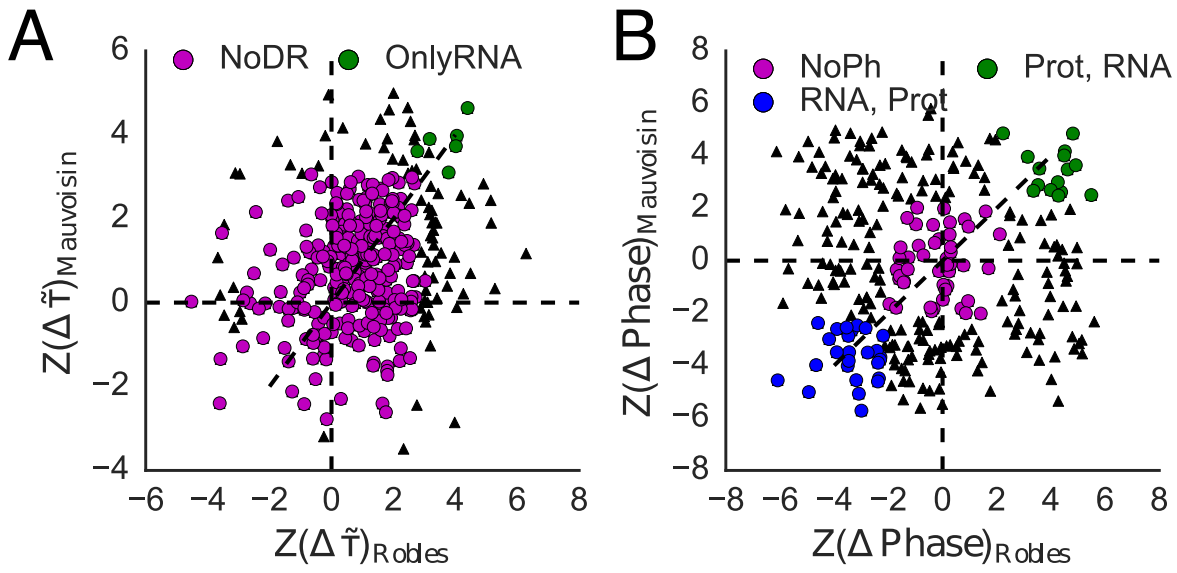

Figure 5.5: 378 genes had similar relationships between RNA and protein behavior in Mauvoisin *et al.* and Robles *et al.* The axes are the Z-scores of the differences, which have been corrected by the expected differences in mean normalized by the expected standard deviation under the hypothesis that no difference exists for time series of that level of noise. (A) 6 genes are rhythmic in RNA expression only in both the Robles *et al.* and Mauvoisin *et al.* datasets, while 291 genes are found to have no differential rhythmicity in both datasets. None of the genes found to be rhythmic only at the protein level overlapped between Robles *et al.* and Mauvoisin *et al.* (B) Of the genes found to have no differential rhythmicity in both datasets, 47 show no change in phase, 26 show an RNA expression peak before a protein level peak, and 15 show the protein peak before the RNA peak.

SILAC and tandem mass spectrometry, is not yet precise enough to provide clear evidence supporting RNA-rhythm independent protein level rhythms. We showed that the protein level data is much noisier than the RNA expression data and for many genes RNA expression rhythmicity was found to exist while protein level expression rhythmicity did not. Higher rates of sampling and increased replicate numbers for the protein quantification are necessary in order to obtain better measures of protein level rhythmicity. While current studies have protein level sampling rates below that of the RNA sampling rate, these results indicate that it might be necessary for the protein sampling to be higher than the RNA sampling to achieve equal levels of noise under current experimental methods.

We do find significant and large shifts in phase between the RNA expression and protein levels, which are indicative of potential post-transcriptional and post-translational regulation that are impacting circadian protein behavior. A recent study in mouse fibroblasts showed that BMAL, known mainly as a transcription factor in the core clock machinery, plays a role in post-transcriptional regulation as well [66]. The phase shifts between protein and RNA identified in the data suggest BMAL could play a role in the protein level peaks that occur long after the RNA expression peaks.

BDR is designed for determining differential rhythmicity and phase between two time series. The simulation procedure generates the expected variance in the difference of rhythmicity scores in the case where the time series have the same rhythmicity. Neither BDR nor DODR is currently suited, however, for a comparison between more than two time series. For example, for comparisons made between the 12 tissues surveyed in the mouse under DD conditions in Zhang *et al.* [121], BDR and DODR could do no better than pairwise comparisons between all the tissues. An ANOVA-like method for assessing differential rhythmicity across more than two tissues remains an outstanding problem.

We have developed our method to focus on aspects of differences specific to circadian biology. Since a growing body of literature shows the detrimental impact of circadian-specific

disruption, such as timing of feeding, jet lag, and shift work on overall health [36, 17], we believe this method, which focuses specifically on circadian-related characteristics, will be useful.

5.6 Supplementary Figures

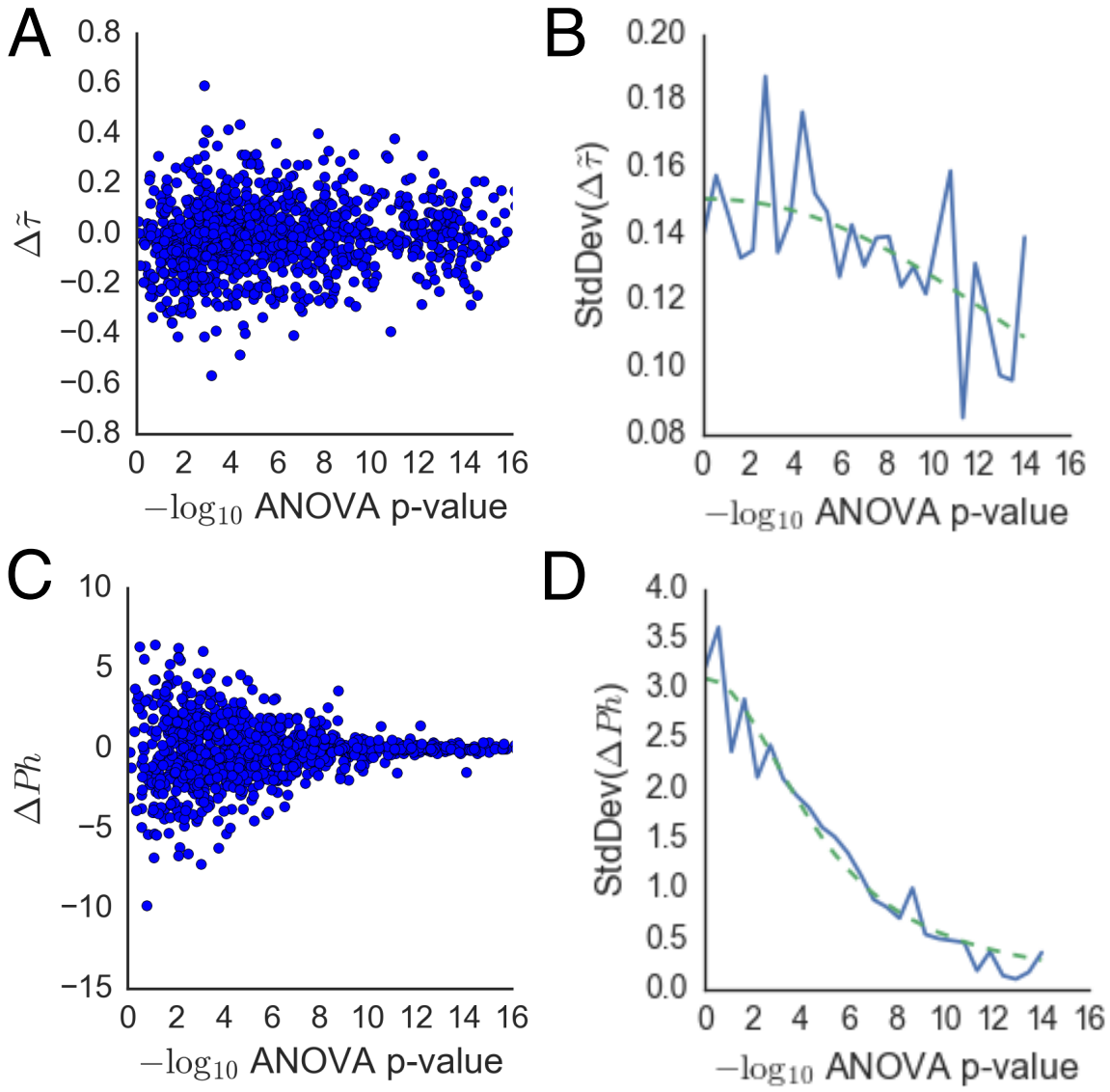

Figure 5.S1: (A) Mean Tau Difference ($\Delta\bar{\tau}$) between time series with the same noise as a function of ANOVA p-value of one of the time series. (B) Standard deviation of $\Delta\bar{\tau}$ as a function of ANOVA p-value. (C) Mean Phase difference ($\Delta\bar{Ph}$) between time series with the same noise and pre-noise phase as a function of ANOVA p-value of one of the time series. (D) Standard deviation of $\Delta\bar{\tau}$ as a function of ANOVA p-value. Legends in (B) and (D) indicates number of bins into which p-values are separated.

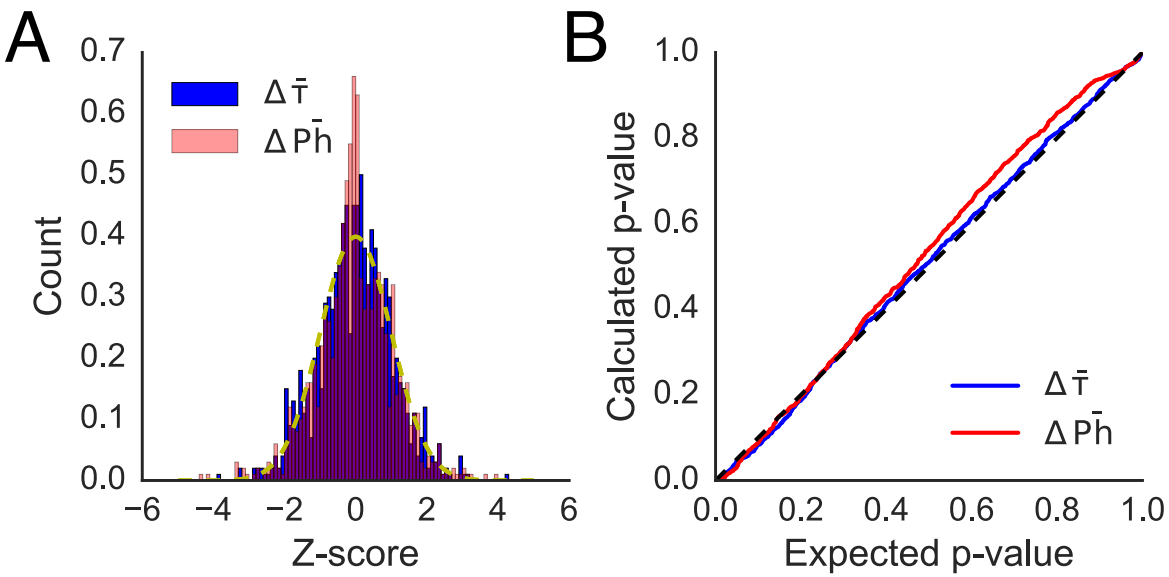

Figure 5.S2: (A) Z-scores of differences in rhythmicity $\Delta\tilde{\tau}$ and phase $\Delta\bar{P}h$ for 1000 same-noise comparisons of simulated cosine time series. Yellow dashed line indicates a normal distribution of mean 0 and variance 1. (B) P-values of z-scores shown in (A) for rhythmicity difference and phase.

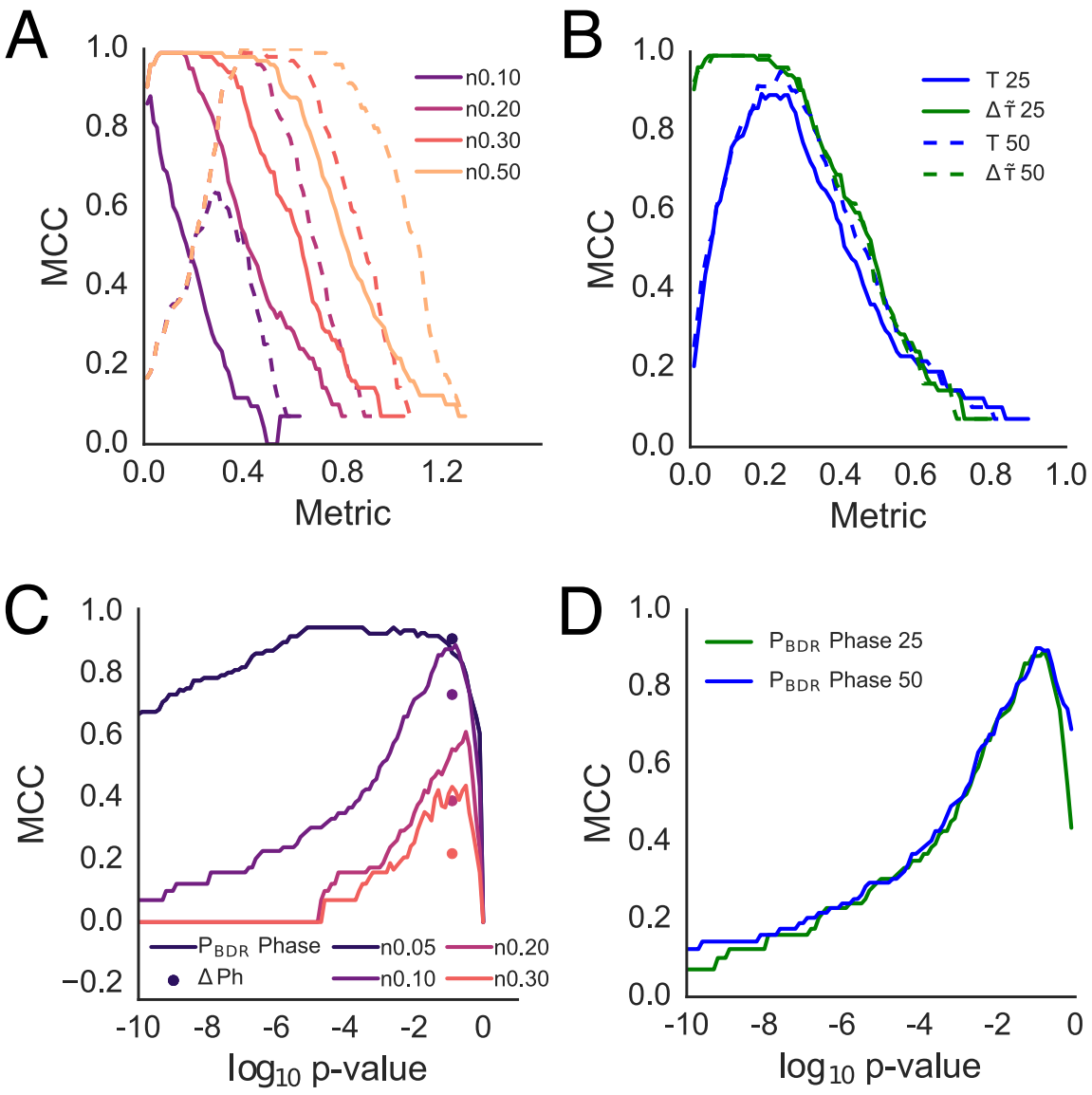

Figure 5.S3: Bootstrap Differential Rhythmicity differential rhythmicity and differential phase p-values outperforms BooteJTK p-values, Phase Difference measurements, and T-statistic p-values in classification ability. 25 bootstrap replicates are found to perform equivalently to 50 bootstrap replicates. (A) Comparison of differential rhythmicity classification ability of BDR $\Delta\tilde{\tau}$ p-value (solid line), and the BooteJTK p-value (dashed line) across several noise level conditions. (B) Comparison of differential rhythmicity classification ability between the BDR $\Delta\tilde{\tau}$ p-value and $\Delta\tilde{\tau}$ T-statistic p-value (T) for 25 bootstrap replicates (solid line) and 50 bootstrap replicates (dashed line). (C) Comparison of BDR Phase Difference $\Delta\bar{P}h$ p-value (lines) against the $\Delta\bar{P}h$ measurement (points) for different noise levels and a phase shift of 4 hours. (D) Comparison of BDR $\Delta\bar{P}h$ p-values for 25 bootstrap replicates and 50 bootstrap replicates.

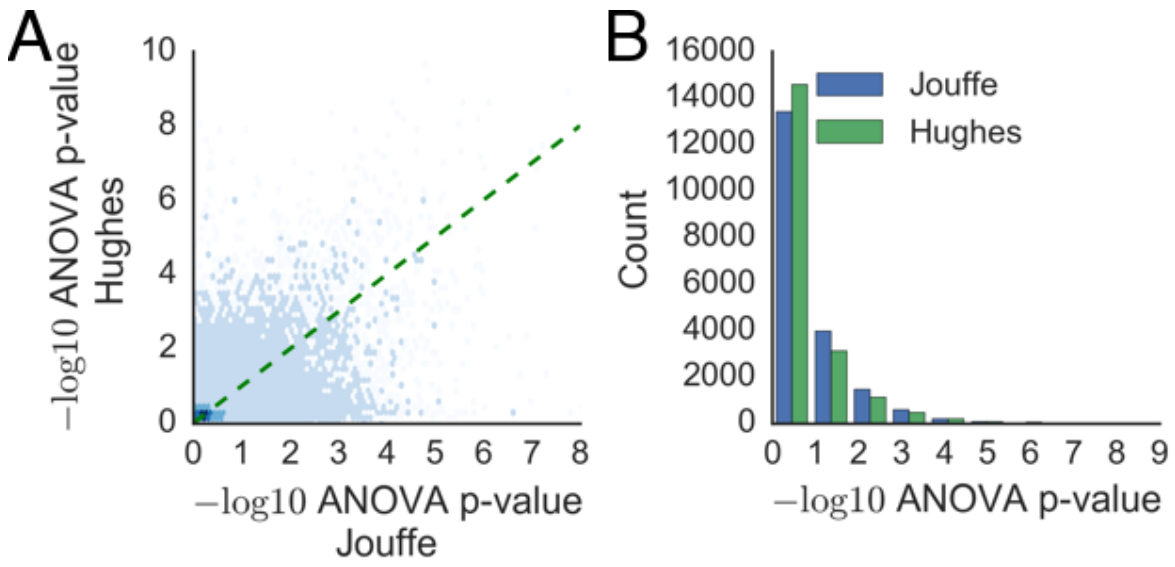

Figure 5.S4: Comparison of RNA expression time series noisiness from Jouffe *et al.* and Hughes *et al.* datasets using ANOVA finds equivalent levels of noisiness in gene-by-gene comparison (A) as well as overall distribution comparison (B). The slope of the green dashed line in (A), (C), and (E) is 1. Higher $-\log_{10}$ ANOVA p-values indicate less noisy time series.

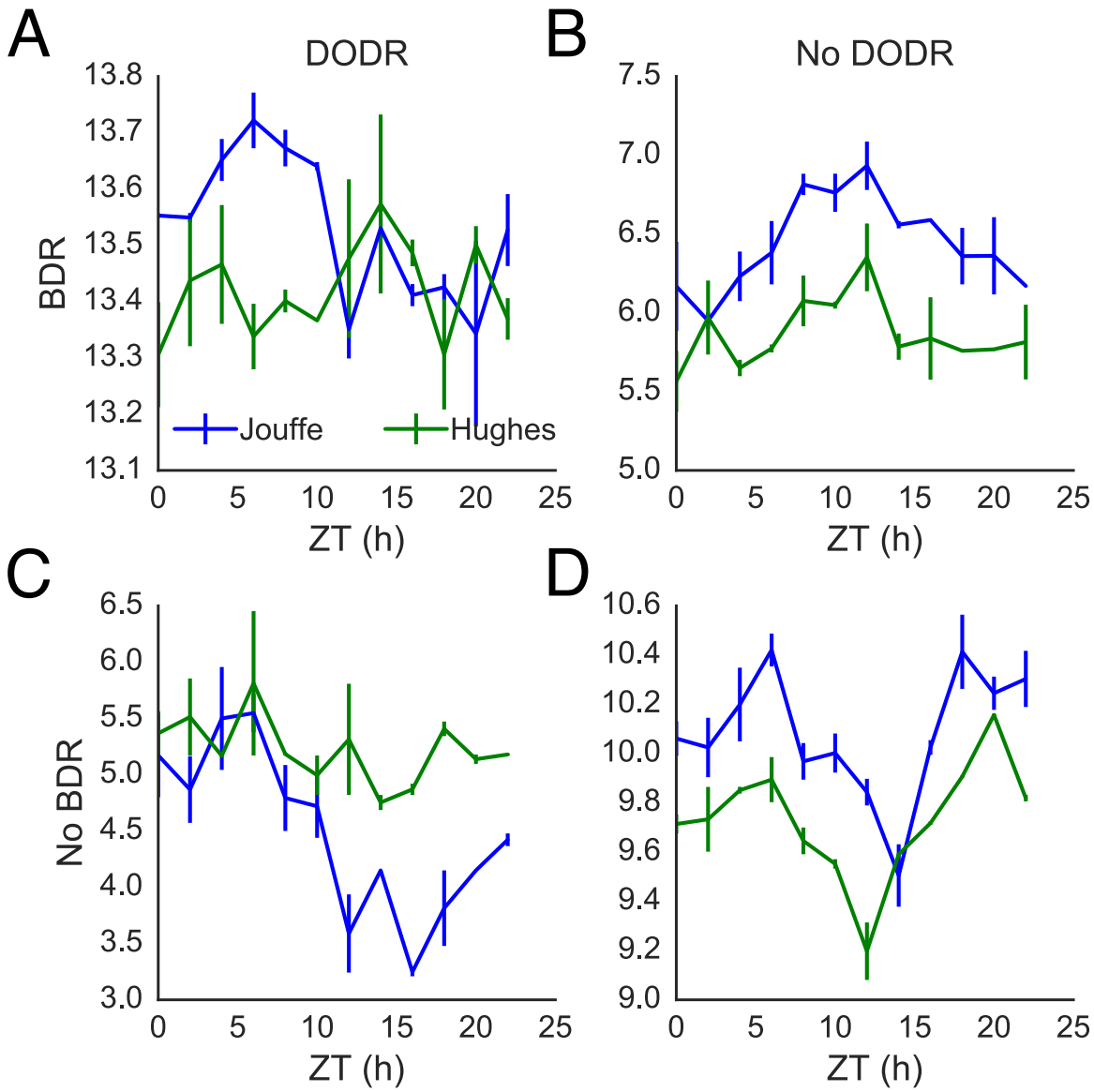

Figure 5.S5: Examples of genes with RNA expression time series from Jouffe *et al.* and Hughes *et al.* which are found to have converging or diverging differential rhythmicity results by BDR (vertical) and DODR (horizontal). In all panels Jouffe *et al.* time series are blue and Hughes *et al.* time series are green. Probes shown are (A) 1437902_s_at (Rarres2), (B) 1455112_at (Aifm2), (C) 1443671_x_at (Odf3b), and (D) 1419547_at (Fahd1).

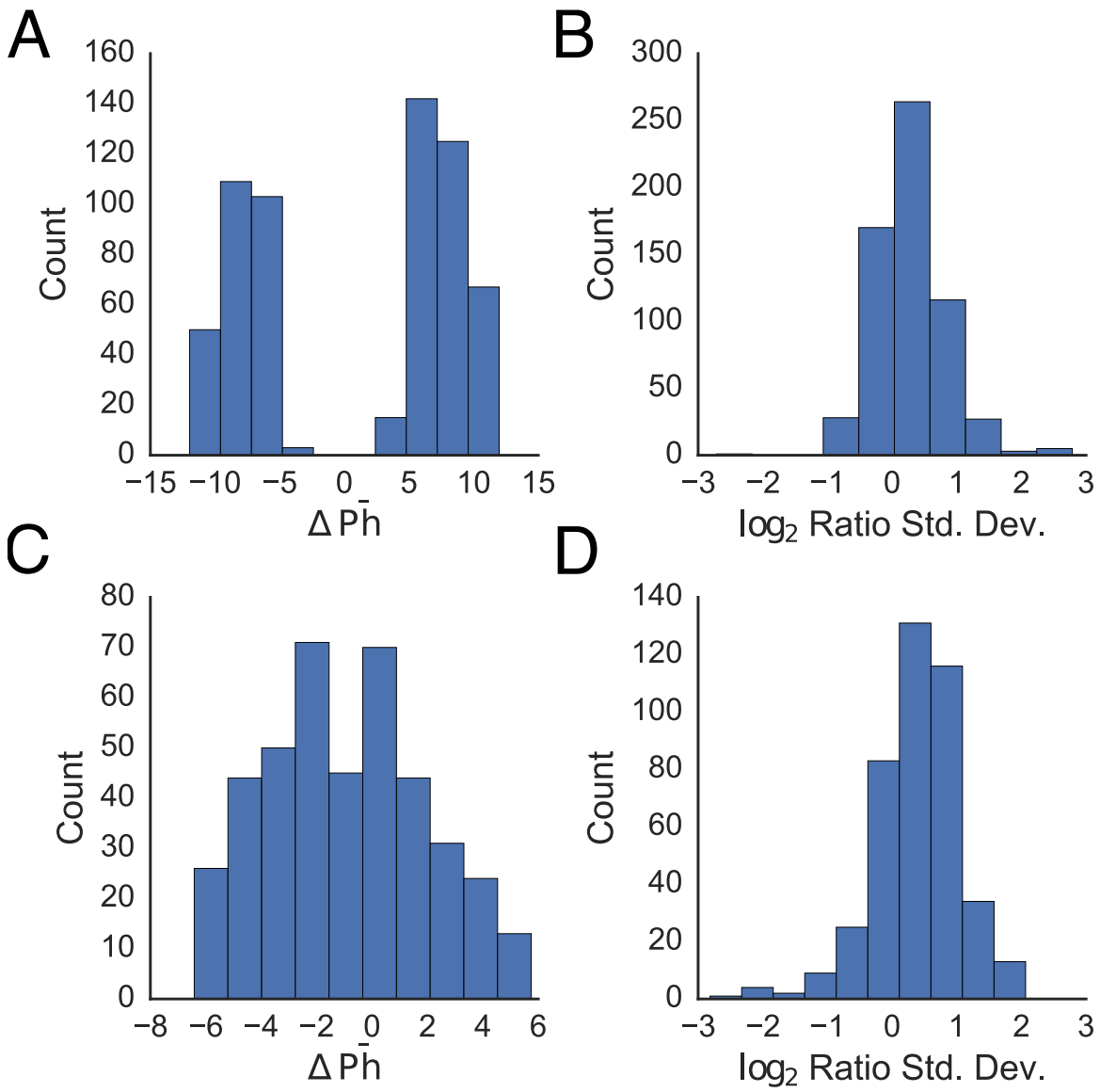

Figure 5.S6: (A) Phase shift distribution of 345 probes identified as not differentially rhythmic by BDR, identified as having no differences by DODR, and as have differential phases by BDR. The probes have an average phase shift of 11.7 h with a standard deviation of 3.4 h (modulo 24 h). (B) Phase shift distribution of 209 probes identified as not differentially rhythmic by BDR, identified as having no differential phases by BDR, and identified as having changes by DODR. The probes have an average phase shift of -2 h with standard deviation of 3.5 h (modulo 24 h). (C) Log_2 ratio of standard deviations of time series of 209 probes identified as not differentially rhythmic by BDR, identified as having no differential phases by BDR, and identified as having changes by DODR. The probes have an average $\log_2$ ratio of standard deviations of 0.4 with a standard deviation of 0.7.

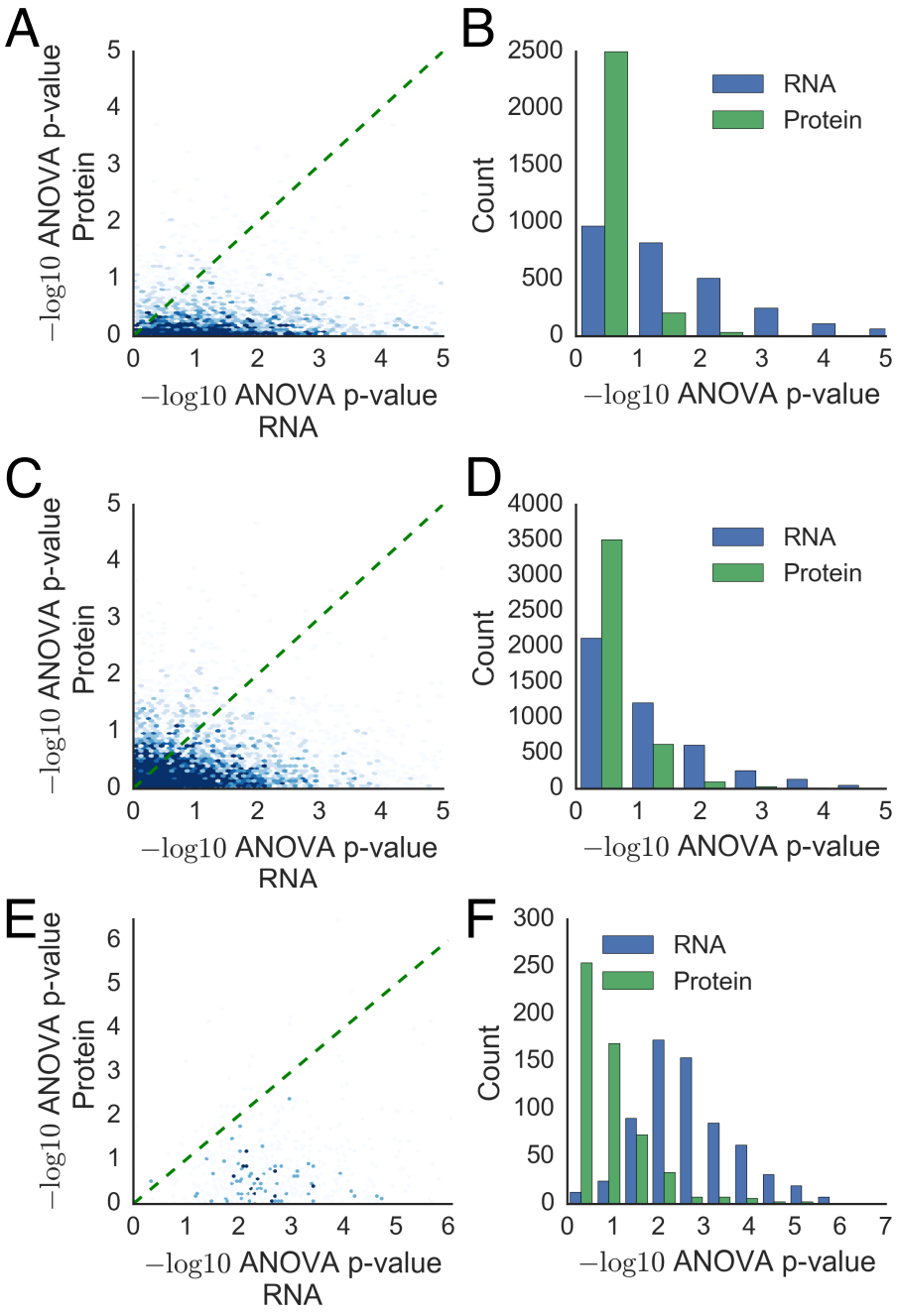

Figure 5.S7: Comparison of RNA expression noisiness to protein level noisiness as measured by ANOVA at a gene-by-gene comparison in (A), (C), and (E) and a distribution level comparison in (B), (D), and (F). The datasets compared are from Robles *et al.* (A) and (B), Mauvoisin *et al.* (C) and (D), and Guerreiro *et al.* (E) and (F). The slope of the green dashed line in (A), (C), and (E) is 1. Higher $-\log_{10}$ ANOVA p-values indicate less noisy time series.

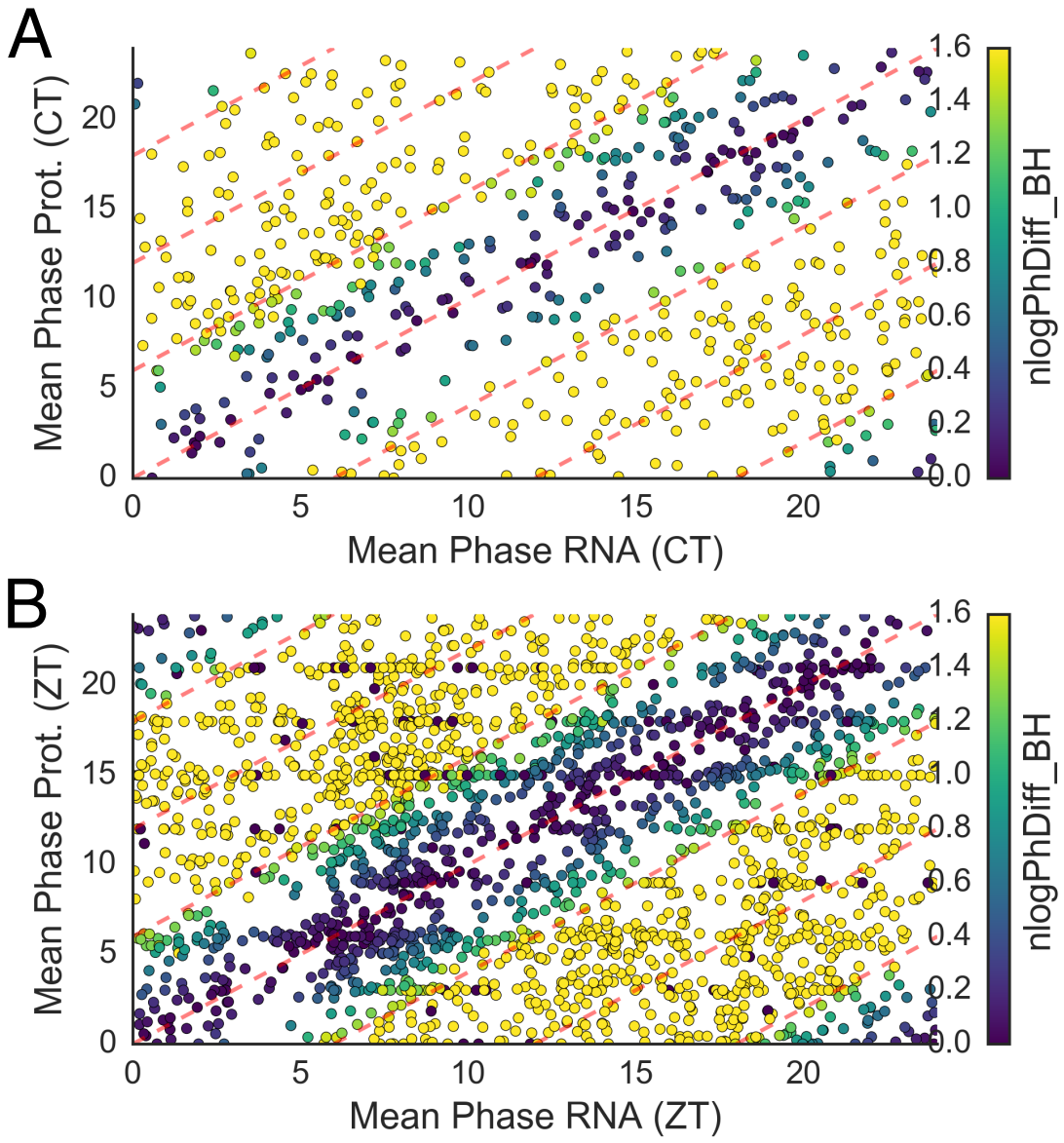

Figure 5.S8: Comparison of phases for genes whose RNA expression and protein levels are found not to be differentially rhythmic by BDR for (A) Robles *et al.* and (B) Mauvoisin *et al.*. The colors indicate the $-\log_{10}$ BDR differential phase BH-value, with a ceiling of 1.6 (BH-value≈0.025) for clarity of visualization. Red dashed lines have slope 1 and are offset to indicate phase shifts every 6 h.

CHAPTER 6

CONCLUSION

In this dissertation, I have presented three methods to better understand and interpret circadian high-throughput time series experiments.

In Chapter 2, I presented improvements to JTK_CYCLE, a non-parametric method matching the best reference waveform to an experimental time series. By empirically calculating the p-values that result from the multiple waveform comparison instead of using the Bonferroni correction as JTK_CYCLE does or the Benjamini-Hochberg correction as RAIN does, empirical JTK_CYCLE (eJTK) generates accurate p-values which are neither conservative nor anti-conservative. This allows for the use of reference waveforms with different trough to peak and peak to trough times, which would otherwise lead to harsher corrections under the previous two methods.

In Chapter 3, I identified incorrect assumptions regarding the independence of p-values in two leading rhythm detection methods, RAIN [99] and MetaCycle [113] and provided approaches to correct and improve these methods. The application of the Benjamini-Hochberg method to dependent p-values in RAIN results in underestimation of p-values, which I correct with empirical calculation of p-values based off of a null distribution generated from simulated data (as in Chapter 2). The application of Fisher integration to p-values from similar rhythm detection methods in MetaCycle, which are dependent since the rhythm detection methods measure similar aspects of the time series, results in underestimation of p-values, which can be corrected by the use of Brown's method for dependent p-values [16].

In Chapter 4, I applied eJTK to parametric bootstrap replicates of the experimental time series, where I utilized an empirical Bayes method to better estimate the variance of each time point measurement in a method referred to as Bootstrap eJTK (BooteJTK). This method incorporates the non-parametric avoidance assumptions about the shape of the rhythmic waveform with the ability to assess whether time point measurements are

sufficiently distinct to ensure the observed signal is consistent.

In Chapter 5, I presented a method, Bootstrap Differential Rhythmicity (BDR) which builds on BooteJTK to assess the presence of differential rhythmicity between two time series. This method introduces the use of ANOVA to measure the noisiness of a time series, which can be used to determine the expected uncertainty in measurements of rhythmicity and phase. This expected uncertainty can be compared with the observed differences to perform null hypothesis significance testing to generate accurate p-values. I showed that current studies of protein level and RNA expression rhythmicity do not provide support for the existence of widespread protein circadian rhythmicity independent of concurrent RNA expression rhythmicity.

Having developed these methods for detecting rhythmicity in high-throughput circadian time series, we may ask is what it means to be rhythmic, and how the methods presented match this definition. From a biological point of view, an RNA transcript or protein being rhythmic implies that its expression has a consistent pattern over 24 hours that repeats itself and there are biological processes whose evolved functioning would be changed if the expression of that transcript or protein ceased to be rhythmic or were shifted by several hours. One way of viewing these two aspects is as a identification of statistical significance and effect size, only with the effect size incorporating several other metrics separate from the strength of the observed rhythmicity, which varies with the method used. Methods for rhythm detection are generally geared towards identifying the consistent pattern over 24 hours. That the rhythms are biologically meaningful requires the expression levels for a given transcript and protein to be sufficiently high and that the change from trough to peak be sufficiently large to impact a biological system. This second part is a requirement that depends on many external factors, such as the measurement method being used and the biological system in question; many decisions about the minimum required mean expression and fold changes from trough to peak need to be made prior to or separate from the application of these methods. Of the

rhythm detection methods, however, few adequately attempt to incorporate both parts of the definition. The cosine-fitting and Fourier-based methods have very narrow definitions of rhythmicity (matching a cosine) while also arbitrarily defining the progression from trough to peak necessary to be biologically relevant. The reference-based non-parametric methods have a broader definition of rhythmicity to capture most cases but completely ignore amplitude levels (being rank based) and therefore provide little insight into biological relevance. Of the reference-free methods, ANOVA provides a means to interpret whether the points of a time series are sufficiently separate from one another that rhythmicity could be identified by a biological system. BooteJTK combines the broad definition of rhythmicity of reference-based non-parametric methods with the noise-measuring features of ANOVA, clearly addressing the first part of the definition while providing as much assumption-free insight to the second part of the definition as possible, as it is solely based on the time series in question.

This difficulty in defining rhythmicity is also apparent when thinking about differential rhythmicity, where the rhythmicity between two time series may be different. In the naïve approach, where p-values are compared, the question being asked is whether rhythmicity is lost or gained from one condition or tissue to another. An interpretation of this question is "has the reliability of the expected pattern of the transcript or protein expression changed between conditions?" Again, it is difficult to ensure that this question has a biological relevance. Does Per2 being considerably rhythmic in one condition and extremely rhythmic in another condition lead to a biological difference? Does a rhythmicity p-value 10^{-4} mean something different biologically than a rhythmicity p-value of 10^{-6}? If the amplitude of a protein's expression doubles (which DODR would identify) but its rhythmicity decreases (which BDR would identify), then what is the biological implication? Does the protein now have a greater role in the molecular machinery and biological processes in which it is involved due to its increased expression, or does its loss of temporal resolution mean that its role in the molecular machinery is compromised, or both? Bootstrap Differential

Rhythmicity (BDR) improves upon previous methods in addressing these issues, but its main strength in these situations is its ability to be more stringent than the naïve p-value threshold approach in addressing claims that differences between time series exist, literally by determining the probability that such a difference between time series would exist when there was in actuality no difference. In Chapter 5, we show this strength by comparing RNA expression with protein level data from several studies, finding that there is not sufficient evidence to conclude that protein level rhythmicity exists where RNA expression rhythmicity does not.

As RNA-Seq and microarray technology becomes less expensive, the density of time points sampled and number of replicates obtained for these circadian experiments will increase. This might suggest that advanced rhythm detection methods will become obsolete under improvements in the data amount and quality. While this may hold true for RNA-Seq and microarray technology, experimental methods sampling new aspects of circadian systems will most likely be expensive and noisy when first introduced. This means that while the type of data may change, the rhythm detection methods and differential rhythmicity methods developed here will still be required.

We have built the three methods presented here in the framework of null hypothesis significance testing: we have made no assumptions about the genes we are studying, assumed no dependences between genes, included no prior information about what we know biologically to likely be the case. If we were to use BooteJTK to analyze a mouse liver circadian dataset under Light-Dark conditions in a wild-type genetic background and did not find Per2 to have rhythmic gene expression, we would more likely doubt the experimental results and ask our collaborators to repeat the experiment than conclude that maybe Per2 is not always rhythmic in mouse liver. Though we expect that Per2 is rhythmic in mouse liver tissue in a wild-type genetic background in Light-Dark conditions, this information is not incorporated into our methods. Likewise, in the Zhang *et al.* dataset we discuss in

Chapter 4, though we would expect knowledge of the rhythmicity of Bhlhe41 in 11 tissues to influence how we interpret results in the one tissue where Bhlhe41 is found not to be rhythmic, instead these results remain in isolation from one another. Methods to combine information across tissues and conditions have been successful in a related problem, assessing whether single nucleotide polymorphisms (SNPs) in DNA are expression Quantitative Trait Loci (eQTLs) (i.e., responsible for changes in gene expression). Several methods have been developed which use hierarchical models to allow evidence for the presence or absence of eQTLs to be shared across tissues and conditions while still allowing for heterogeneity (i.e., the possibility that a SNP is an eQTL in one tissue but not in another). These methods have found that sharing of eQTLs is underestimated across tissues when these hierarchical models are not used. Similar methods for circadian rhythm data do not exist currently, but their development would provide a critical tool for comparing circadian rhythmicity across conditions and tissues. In Chapter 4, we examine genes which are rhythmic in 9 or more tissues, under the assumption we may be underestimating the rhythmicity of genes across all tissues. Hierarchical models that could integrate information across tissues in a study, as well as combine information across studies and integrate other known features of the data, such as regulatory relationships, would be an important addition to the armory of methods for understanding circadian data.

With this body of work, I have added to the field of statistical rhythm detection in circadian biology by

1. increasing the number of features of interest in rhythm detection,
2. incorporating measurement uncertainty and noise in the data in a novel way,
3. generating accurate p-values for my methods and identifying cases where p-values were not accurate in other methods,

4. applying more rigor to the common comparisons made between time series than had been done previously.

In Chapters 2 and 4, I develop methods to detect rhythmicity that show increased sensitivity for asymmetric waveforms (eJTK) and identify time series whose time point uncertainties are small relative to their amplitudes (BooteJTK), both features of interest. In Chapter 3, I build on methods to obtain accurate p-values efficiently, and identify and correct two cases where methods for independent p-values are improperly used on dependent p-values. In Chapter 4, I account for the high levels of noise found in high-throughput time series to increase the likelihood that the rhythms detected are biologically distinguishable, meaningful, and unlikely to be due to noise. In Chapter 5, I develop a method that provides a more rigorous means of comparing the rhythmicity and phase of two sparse and noisy time series than was previously available.

I have worked to implement the methods I developed so that they are easily accessible and useable. I hope that the ideas behind them, such as I have presented here, are equally useful to the research community that builds upon them in order to produce more rigorous and powerful methods for applications both within and beyond circadian biology.

REFERENCES

[1] Michikazu Abe, Erik D Herzog, Shin Yamazaki, Marty Straume, Hajime Tei, Yoshiyuki Sakaki, Michael Menaker, and Gene D Block. Circadian Rhythms in Isolated Brain Regions. *The Journal of Neuroscience*, 22(1):350–356, 2002.

[2] K C Abruzzi, J Rodriguez, J S Menet, J Desrochers, A Zadina, W Luo, S Tkachev, and M Rosbash. Drosophila CLOCK target gene characterization: implications for circadian tissue-specific gene expression. *Genes & Development*, 25(22):2374–2386, nov 2011.

[3] S E Ahnert, K Willbrand, F C S Brown, and T M A Fink. Unbiased pattern detection in microarray data series. *Bioinformatics*, 22(12):1471–1476, 2006.

[4] Ruth A Akhtar, Akhilesh B Reddy, Elizabeth S Maywood, Jonathan D Clayton, Verdun M King, Andrew G Smith, Timothy W Gant, Michael H Hastings, Charalambos P Kyriacou, and Leicester Le. Circadian Cycling of the Mouse Liver Transcriptome, as Revealed by cDNA Microarray, Is Driven by the Suprachiasmatic Nucleus. *Current Biology*, 12(02):540–550, 2002.

[5] Urs Albrecht. Orchestration of gene expression and physiology by the circadian clock. *Journal of Physiology-Paris*, 100(5–6):243–251, 2006.

[6] Ravi Allada and Brian Y. Chung. Circadian organization of behavior and physiology in Drosophila. *Annual Review of Physiology*, 72:605–624, 2010.

[7] David B Allison, Xiangqin Cui, Grier P Page, and Mahyar Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7(January):55–65, 2006.

[8] Simon N Archer, Emma E Laing, Carla S Möller-levet, Daan R Van Der Veen, Giselda Bucca, Alpar S Lazar, Nayantara Santhi, Ana Slak, Renata Kabiljo, Malcolm Von

Schantz, Colin P Smith, and Derk-jan Dijk. Mistimed sleep disrupts circadian regulation of the human transcriptome. 2014.

[9] R Armitage. Sleep and circadian rhythms in mood disorders. *Acta Psychiatrica Scandinavica*, 115:104–115, 2007.

[10] Osnat Bartok, Charalambos P Kyriacou, Joel Levine, Amita Sehgal, and Sebastian Kadener. Adaptation of molecular circadian clockwork to environmental changes: a role for alternative splicing and miRNA. *Proceedings of the Royal Society B: Biological Sciences*, 280(1765):20130011–20130011, jul 2013.

[11] L M Beaver, L A Hooven, S M Butcher, N Krishnan, K A Sherman, E S Y Chow, and J M Giebultowicz. Circadian Clock Regulates Response to Pesticides in Drosophila via Conserved Pdp1 Pathway. *Toxicological Sciences*, 115(2):513–520, jun 2010.

[12] Deborah Bell-Pedersen, Vincent M Cassone, David J Earnest, Susan S Golden, Paul E Hardin, Terry L Thomas, and Mark J Zoran. Circadian rhythms from multiple oscillators: lessons from diverse organisms. *Nature Reviews Genetics*, 6(7):544–556, 2005.

[13] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[14] Nicolas L Bray, Harold Pimentel, Pall Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech*, 34(5):525–527, may 2016.

[15] J L Brown. An Sp1/KLF binding site is important for the activity of a Polycomb group response element from the Drosophila engrailed gene. *Nucleic Acids Research*, 33(16):5181–5189, sep 2005.

[16] Morton B Brown. A Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics*, 31(4):987–992, 1975.

[17] Raquel Canuto, Anderson S Garcez, and Maria T A Olinto. Metabolic syndrome and shift work: a systematic review. *Sleep Medicine Reviews*, 17(6):425–31, dec 2013.

[18] M Fernanda Ceriani, John B Hogenesch, Marcelo Yanovsky, Satchidananda Panda, Martin Straume, and Steve A Kay. Genome-Wide Expression Analysis in Drosophila Reveals Genes Controlling Circadian Behavior. *The Journal of Neuroscience*, 22(21):9305–9319, 2002.

[19] Adam Claridge-Chang, Herman Wijnen, Felix Naef, Catharine Boothroyd, Nikolaus Rajewsky, and Michael W Young. Circadian Regulation of Gene Expression Systems in the Drosophila Head. *Neuron*, 32(4):657–671, 2001.

[20] David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Yuriy Mileyko. Lipschitz functions have Lp-stable persistence. *Foundations of Computational Mathematics*, 10(2):127–139, 2010.

[21] Germaine Cornelissen. Cosinor-based rhythmometry. *Theoretical Biology and Medical Modelling*, 11(1):16, 2014.

[22] Anastasia Deckard, Ron C Anafi, John B Hogenesch, Steven B Haase, and John Harer. Design and analysis of large-scale biological rhythm studies: a comparison of algorithms for detecting periodic signals in biological data. *Bioinformatics*, 29(24):3174–80, dec 2013.

[23] Mary-Lee Dequéant, Sebastian Ahnert, Herbert Edelsbrunner, Thomas M A Fink, Earl F Glynn, Gaye Hattem, Andrzej Kudlicki, Yuriy Mileyko, Jason Morton, Arcady R Mushegian, Lior Pachter, Maga Rowicka, Anne Shiu, Bernd Sturmfels, and Olivier Pourquié. Comparison of Pattern Detection Methods in Microarray Time Series of the Segmentation Clock. *PLoS ONE*, 3(8):e2856, 2008.

[24] Charna Dibner, Ueli Schibler, and Urs Albrecht. The Mammalian Circadian Timing System: Organization and Coordination of Central and Peripheral Clocks. *Annual Review of Physiology*, 72(1):517–549, 2010.

[25] Kristin L. Eckel-Mahan, Vishal R. Patel, Sara De Mateo, Ricardo Orozco-Solis, Nicholas J. Ceglia, Saurabh Sahar, Sherry A. Dilag-Penilla, Kenneth A. Dyar, Pierre Baldi, and Paolo Sassone-Corsi. Reprogramming of the circadian clock by nutritional challenge. *Cell*, 155(7):1464–1478, 2013.

[26] H Edelsbrunner, D Letscher, and A Zomorodian. Topological persistence and simplification. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, FOCS '00, pages 454—-, Washington, DC, USA, 2000. IEEE Computer Society.

[27] Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994.

[28] Paul D Etter and Mani Ramaswami. The ups and downs of daily life: Profiling circadian gene expression in Drosophila. *BioEssays*, 24(6):494–498, 2002.

[29] T M A Fink, K Willbrand, and F C S Brown. 1-D random landscapes and non-random data series. *EPL (Europhysics Letters)*, 79(3):38006, 2007.

[30] R A Fisher. On the Interpretation of Chi2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87, jan 1922.

[31] R A Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.

[32] Timothy P Fitzgibbons, Sophia Kogan, Myriam Aouadi, Greg M Hendricks, Juerg Straubhaar, and Michael P Czech. Similarity of mouse perivascular and brown adipose tissues and their resistance to diet-induced inflammation. *American journal of physiology. Heart and circulatory physiology*, 301(4):H1425–37, oct 2011.

[33] Matthieu Flourakis, Elzbieta Kula-Eversole, Alan L. Hutchison, Tae Hee Han, Kimberly Aranda, Devon L. Moose, Kevin P. White, Aaron R. Dinner, Bridget C. Lear, Dejian Ren, Casey O. Diekman, Indira M. Raman, and Ravi Allada. A Conserved Bicycle Model for Circadian Clock Control of Membrane Excitability. *Cell*, 162(4):836–848, 2015.

[34] Timothée Flutre, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLoS Genetics*, 9(5):1–8, 2013.

[35] FlyBase Curators, SwissProt Project Members, and InterPro Project Members. Gene Ontology annotation in FlyBase through association of InterPro records with GO terms. 2004.

[36] Laura K Fonken and Randy J Nelson. The effects of light at night on circadian clocks and metabolism. *Endocrine Reviews*, 35(4):648–70, aug 2014.

[37] Amanda A H Freeman, Konstantinos Mandilaras, Fanis Missirlis, and Subhabrata Sanyal. An emerging role for Cullin-3 mediated ubiquitination in sleep and circadian rhythm: Insights from Drosophila. *Fly*, 7(1):39–43, jan 2013.

[38] Kazuyo Fujikawa, Aya Takahashi, Azusa Nishimura, Masanobu Itoh, Toshiyuki Takano-Shimizu, and Mamiko Ozaki. Characteristics of genes up-regulated and down-regulated after 24 h starvation in the head of Drosophila. *Gene*, 446(1):11–17, oct 2009.

[39] Michael J Gardner, Katharine E Hubbard, Carlos T Hotta, Antony N Dodd, and Alex A R Webb. How plants tell the time. *Biochem J.*, 24:15–24, 2006.

[40] Anne Germain and David J Kupfer. Circadian rhythm disturbances in depression. *Human Psychopharmacology: Clinical and Experimental*, 23(7):571–585, 2008.

[41] Akihiro Goriki, Fumiyuki Hatanaka, Jihwan Myung, Jae Kyoung Kim, Takashi Yoritaka, Akio Matsubara, Daniel Forger, and Toru Takumi. A Novel Protein, CHRONO, Functions as a Core Component of the Mammalian Circadian Clock. *PLoS Biology*, 12(4), 2014.

[42] Ana C L Guerreiro, Marco Benevento, Robert Lehmann, Bas van Breukelen, Harm Post, Piero Giansanti, a F Maarten Altelaar, Ilka M Axmann, and Albert J R Heck. Daily rhythms in the cyanobacterium synechococcus elongatus probed by high-resolution mass spectrometry-based proteomics reveals a small defined set of cyclic proteins. *Molecular & Cellular Proteomics*, 13(8):2042–55, 2014.

[43] Kelly A Hamby, Rosanna S Kwok, Frank G Zalom, and Joanna C Chiu. Integrating Circadian Activity and Gene Expression Profiles to Predict Chronotoxicity of Drosophila suzukii Response to Insecticides. *PLoS ONE*, 8(7):e68472, jul 2013.

[44] E F Harding. An Efficient, Minimal-Storage Procedure for Calculating the Mann-Whitney U, Generalized U and Similar Distributions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(1):1–6, 1984.

[45] Stacey L Harmer, John B Hogenesch, Marty Straume, Hur-song Chang, Bin Han, Tong Zhu, Xun Wang, Joel A Kreps, and Steve A Kay. Orchestrated Transcription of Key Pathways in Arabidopsis by the Circadian Clock. *Science*, 290, 2000.

[46] Sibah Hasan, Nayantara Santhi, Alpar S Lazar, Ana Slak, June Lo, Malcolm Von Schantz, Simon N Archer, Jonathan D Johnston, and Derk-jan Dijk. Assessment of circadian rhythms in humans: comparison of real-time fibroblast reporter imaging with plasma melatonin. *FASEB J.*, 26:2414–2423, 2012.

[47] Kevin Hayes, Julie Baggs, and John Hogenesch. Circadian clocks are seeing the systems biology light. *Genome Biology*, 6(5):219, 2005.

[48] Louisa A Hooven, Katherine A Sherman, Shawn Butcher, and Jadwiga M Giebultowicz. Does the Clock Make the Poison? Circadian Variation in Response to Pesticides. *PLoS ONE*, 4(7):e6469, jul 2009.

[49] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, jan 2009.

[50] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, jan 2009.

[51] Michael E. Hughes, Luciano DiTacchio, Kevin R. Hayes, Christopher Vollmers, S. Pulivarthy, Julie E. Baggs, Satchidananda Panda, and John B. Hogenesch. Harmonics of Circadian Gene Transcription in Mammals. *PLoS Genet*, 5(4):e1000442, apr 2009.

[52] Michael E Hughes, John B Hogenesch, and Karl Kornacker. JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *Journal of Biological Rhythms*, 25(5):372–80, oct 2010.

[53] Alan L. Hutchison, Mark Maienschein-Cline, Andrew H. Chiang, S. M Ali Tabei, Herman Gudjonson, Neil Bahroos, Ravi Allada, and Aaron R. Dinner. Improved Statistical Methods Enable Greater Sensitivity in Rhythm Detection for Genome-Wide Data. *PLoS Comput Biol*, 11(3):e1004094, 2015.

[54] Alan Louis Hutchison, Ravi Allada, and Aaron R. Dinner. Bootstrapping and Empirical Bayes Methods Improve Rhythm Detection in Sparsely Sampled Data. *In preperation*, 2016.

[55] Aaron R. Hutchison, Alan L., Dinner. Correcting for Dependent P-values Improves Accuracy of Leading Rhythm Detection Methods. *In preparation*, 2016.

[56] Eric Jones, Travis Oliphant, and Pearu Peterson. Scipy: Open source scietific tools for Python, 2001.

[57] Celine Jouffe, Gaspard Cretenet, Laura Symul, Eva Martin, Florian Atger, and Felix Naef. The Circadian Clock Coordinates Ribosome Biogenesis. *PLoS Biology*, 11(1), 2013.

[58] Sebastian Kadener, Dan Stoleru, Michael McDonald, Pipat Nawathean, and Michael Rosbash. Clockwork Orange is a transcriptional repressor and a new Drosophila circadian pacemaker component. *Genes and Development*, 21(13):1675–1686, jul 2007.

[59] Kevin P. Keegan, Suraj Pradhan, Ji Ping Wang, and Ravi Allada. Meta-analysis of Drosophila circadian microarray studies identifies a novel set of rhythmically expressed genes. *PLoS Computational Biology*, 3(11):2087–2110, 2007.

[60] Douglas R Kellogg. Wee1-dependent mechanisms required for coordination of cell growth and cell division. *Journal of Cell Science*, 116(24):4883–4890, nov 2003.

[61] Nobuya Koike, Tae-kyung Kim, and Joseph S Takahashi. Transcriptional Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals. *Science*, 338(August):1–10, 2012.

[62] Ronald J Konopka and Seymour Benzer. Clock Mutants of Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 68(9):2112–2116, 1971.

[63] Vanessa Leone, Sean M Gibbons, Kristina Martinez, Alan L Hutchison, Edmond Y Huang, Candace M Cham, Joseph F Pierre, Aaron F Heneghan, Anuradha Nadimpalli, Nathaniel Hubert, Elizabeth Zale, Yunwei Wang, Yong Huang, Betty Theriault, Aaron R Dinner, Mark W Musch, Kenneth A Kudsk, Brian J Prendergast, Jack A Gilbert, and Eugene B Chang. Effects of Diurnal Variation of Gut Microbes and High-

Fat Feeding on Host Circadian Clock Function and Metabolism. *Cell Host & Microbe*, 17:1–9, 2015.

[64] Andrew S P Lim, Gyan P. Srivastava, Lei Yu, Lori B. Chibnik, Jishu Xu, Aron S. Buchman, Julie A. Schneider, Amanda J. Myers, David A. Bennett, and Philip L. De Jager. 24-Hour Rhythms of DNA Methylation and Their Relation with Rhythms of RNA Expression in the Human Dorsolateral Prefrontal Cortex. *PLoS Genetics*, 10(11), 2014.

[65] Yiing Lin, Mei Han, Brian Shimada, Lin Wang, Therese M Gibler, Aloka Amarakone, Tarif A Awad, Gary D Stormo, Russell N Van Gelder, and Paul H Taghert. Influence of the period-dependent circadian clock on diurnal, circadian, and aperiodic gene expression in Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 99(14):9562–9567, 2002.

[66] Jonathan O Lipton, Elizabeth D Yuan, Lara M Boyle, Darius Ebrahimi-fakhari, Erica Kwiatkowski, Ashwin Nathan, Fred Davis, John M Asara, and Mustafa Sahin. The Circadian Protein BMAL1 Regulates Translation in Response to S6K1-Mediated Phosphorylation. *Cell*, 161:1138–1151, 2015.

[67] N R Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39(2):447–462, 1976.

[68] Mengyin Lu and Matthew Stephens. Variance Adaptive Shrinkage (vash): Flexible Empirical Bayes estimation of variances. *bioRxiv*, (July):048660, 2016.

[69] Konstantinos Mandilaras and Fanis Missirlis. Genes for iron metabolism influence circadian rhythms in Drosophila melanogaster. *Metallomics*, 4(9):928, 2012.

[70] H B Mann and D R Whitney. On a Test of Whether one of Two Random Variables is

Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, mar 1947.

[71] Takuya Matsuo, Shun Yamaguchi, and Shigeru Mitsui. Control Mechanism of the Circadian Clock for Timing of Cell Division In Vivo. *Science*, 302(October):255–260, 2003.

[72] B.W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, oct 1975.

[73] Daniel Mauvoisin, Jingkui Wang, Céline Jouffe, Eva Martin, Florian Atger, Patrice Waridel, Manfredo Quadroni, Frédéric Gachon, and Felix Naef. Circadian clock-dependent and -independent rhythmic proteomes implement distinct diurnal functions in mouse liver. *Proceedings of the National Academy of Sciences*, 111(1):167–72, 2014.

[74] Colleen A McClung. Circadian genes, rhythms and the biology of mood disorders. *Pharmacology & Therapeutics*, 114(2):222–232, 2007.

[75] Michael J McDonald and Michael Rosbash. Microarray Analysis and Organization of Circadian Gene Expression in Drosophila. *Cell*, 107(5):567–578, 2001.

[76] C S Moller-Levet, S N Archer, G Bucca, E E Laing, A Slak, R Kabiljo, J C Y Lo, N Santhi, M von Schantz, C P Smith, and Et al. Effects of insufficient sleep on circadian rhythmicity and expression amplitude of the human blood transcriptome. *Proceedings of the National Academy of Sciences*, 110(12):E1132–E1141, mar 2013.

[77] Jason Morton, Lior Pachter, Anne Shiu, and Bernd Sturmfels. The Cyclohedron Test for Finding Periodic Genes in Time Course Expression Studies. *Statistical Applications in Genetics and Molecular Biology*, 6(1):1–25, 2007.

[78] Jason Morton, Lior Pachter, Anne Shiu, Bernd Sturmfels, and Oliver Wienand. Geometry of rank tests, may 2006.

[79] Kathryn Moynihan Ramsey, Biliana Marcheva, Akira Kohsaka, and Joseph Bass. The Clockwork of Metabolism. *Annual Review of Nutrition*, 27(1):219–240, 2007.

[80] S Munoz-Descalzo, M C Llobell, and N Paricio. Cabut, a new gene involved in multiple processes during Drosophila melanogaster development. *Abstracts. Eighteenth European Drosophila Research Conference, Gottingen, 2003.*, page P04, 2003.

[81] S Munoz-Descalzo, M C Llobell, and N Paricio. Cabut encodes a C2H2 zinc finger transcription factor required during Drosophila embryogenesis. *Program and Abstracts. 45th Annual Drosophila Research Conference, Washington, DC, 2004*, page 528C, 2004.

[82] B V North, D Curtis, and P C Sham. A Note on the Calculation of Empirical P-Values from Monte Carlo Procedures. *The American Journal of Human Genetics*, 71(2):439–441, aug 2002.

[83] Satchidananda Panda, Marina P. Antoch, Brooke H. Miller, Andrew I. Su, Andrew B. Schook, Marty Straume, Peter G. Schultz, Steve A. Kay, Joseph S. Takahashi, and John B. Hogenesch. Coordinated Transcription of Key Pathways in the Mouse by the Circadian Clock. *Cell*, 109(3):307–320, 2002.

[84] Mark Perelis, Biliana Marcheva, Kathryn Moynihan Ramsey, Matthew J. Schipma, Alan L. Hutchison, Akihiko Taguchi, Clara Bien Peek, Heekyung Hong, Wenyu Huang, Chiaki Omura, Amanda L. Allred, Christopher A. Bradfield, Aaron R. Dinner, Grant D. Barish, and Joseph Bass. Pancreatic β cell enhancers regulate rhythmic transcription of genes controlling insulin secretion. *Science*, 350(6261):aac4250, 2015.

[85] Harold J Pimentel, Nicolas Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Differential analysis of RNA-Seq incorporating quantification uncertainty. *bioRxiv*, jun 2016.

[86] A Pompella, A Visvikis, A Paolicchi, V De Tata, and A F Casini. The changing faces of glutathione, a cellular protagonist. *Biochem Pharmacol*, 66:1499–1503, 2003.

[87] Marco Preußner, Ilka Wilhelmi, Astrid-Solveig Schultz, Florian Finkernagel, Monika Michel, Tarik Möröy, and Florian Heyd. Rhythmic U2af26 Alternative Splicing Controls PERIOD1 Stability and the Circadian Clock in Mice. *Molecular Cell*, 54(4):651–662, may 2014.

[88] Roberto Refinetti. Laboratory Instrumentation and Computing: Comparison of Six Methods for the Determination of the Period of Circadian Rhythms. *Physiology & Behavior*, 54:869–875, 1993.

[89] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. {limma} powers differential expression analyses for {RNA}-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[90] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1), 2013.

[91] Maria S. Robles, Jürgen Cox, and Matthias Mann. In-Vivo Quantitative Proteomics Reveals a Key Contribution of Post-Transcriptional Mechanisms to the Circadian Regulation of Liver Metabolism. *PLoS Genetics*, 10(1), 2014.

[92] Oded Sandler, Sivan Pearl Mizrahi, Noga Weiss, Oded Agam, Itamar Simon, and Nathalie Q Balaban. Lineage correlations of single cell division time as a probe of cell-cycle dynamics. *Nature*, 519(7544):468–471, mar 2015.

[93] J. D Scargle. Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *APJ*, 263(Dec):835–853, 1982.

[94] Gordon K Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology Volume*, 3(1):1–26, 2004.

[95] Kai-Florian Storch, Ovidiu Lipan, Igor Leykin, N Viswanathan, Fred C Davis, Wing H Wong, and Charles J L B Weitz. Extensive and divergent circadian gene expression in liver and heart. *Nature*, 417(6884):78–83, 2002.

[96] John D. Storey. The Positive False Discovery Rate : A Bayesian Interpretation and the q-Value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

[97] Martin Straume. DNA Microarray Time Series Analysis: Automated Statistical Assessment of Circadian Rhythms in Gene Expression Patterning Introduction. *Methods in Enzymology*, 383(2001):149–166, 2004.

[98] Joseph S Takahashi, Hee-Kyung Hong, Caroline H Ko, and Erin L McDearmon. The genetics of mammalian circadian order and disorder: implications for physiology and disease. *Nature Review Genetics*, 9(10):764–775, 2008.

[99] Paul F Thaben and Pål O Westermark. Detecting rhythms in time series with RAIN. *Journal of Biological Rhythms*, 29(6):391–400, 2014.

[100] Paul F Thaben and Pål O Westermark. Differential rhythmicity: detecting rhythmicity in biological data. *Bioinformatics*, 2016.

[101] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene

and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–78, mar 2012.

[102] John W Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[103] John J Tyson, Katherine C Chen, and Bela Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current Opinion in Cell Biology*, 15(2):221–231, 2003.

[104] Hiroki R Ueda, Wenbin Chen, Akihito Adachi, Hisanori Wakamatsu, Satoko Hayashi, Tomohiro Takasugi, Mamoru Nagano, Ken-ichi Nakahama, Yutaka Suzuki, Sumio Sugano, Masamitsu Iino, Yasufumi Shigeyoshi, and Seiichi Hashimoto. A transcription factor response element for gene expression during circadian night. *Nature*, 418(6897):534–539, 2002.

[105] Hiroki R Ueda, Akira Matsumoto, Miho Kawamura, Masamitsu Iino, Teiichi Tanimura, and Seiichi Hashimoto. Genome-wide Transcriptional Orchestration of Circadian Rhythms in Drosophila. *J. Biol. Chem*, 16(277):14048–14052, 2002.

[106] Hideki Ukai and Hiroki R Ueda. Systems Biology of Mammalian Circadian Clocks. *Annual Peview of Physiology.*, 72:579–603, 2010.

[107] Martha Hotz Vitaterna, David P King, Anne-Marie Chang, Jon M Kornhauser, Phillip L Lowrey, J David Mcdonald, William F Dove, Lawrence H Pinto, Fred W Turek, and Joseph S Takahashi. Mutagenesis and Mapping of a Mouse Gene, Clock, Essential for Circadian Behavior. *Science*, 264(5159):719–725, 1994.

[108] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, jan 2009.

[109] Eric W Weisstein. Bonferroni Correction, From MathWorld–A Wolfram Web Resource, 2014.

[110] Herman Wijnen, Felix Naef, Catharine Boothroyd, Adam Claridge-Chang, and Michael W. Young. Control of daily transcript oscillations in Drosophila by light and the circadian clock. *PLoS Genetics*, 2(3):0326–0343, 2006.

[111] Herman Wijnen, Felix Naef, Michael W Young, and Michael W Young. *Molecular and Statistical Tools for Circadian Transcript Profiling*, volume 393, pages 341–365. Academic Press, 2005.

[112] Herman Wijnen and Michael W Young. Interplay of Circadian Clocks and Metabolic Rhythms. *Annual Review of Genetics*, 40(1):409–448, 2006.

[113] Gang Wu, Ron C Anafi, Michael E Hughes, Karl Kornacker, and John B Hogenesch. MetaCycle: an integrated R package to evaluate periodicity in large scale data. *Bioinformatics*, 1-3(July):040345, 2016.

[114] Gang Wu, Jiang Zhu, Jun Yu, Lan Zhou, Jianhua Z Huang, and Zhang Zhang. Evaluation of five methods for genome-wide circadian gene identification. *Journal of biological rhythms*, 29(4):231–42, 2014.

[115] Jean Wu, James MacDonald, Jeff Gentry, and Rafael Irizarry. *gcrma: Background Adjustment Using Sequence Information*, 2016.

[116] Katharina Wulff, Silvia Gatti, Joseph G Wettstein, and Russell G Foster. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. *Nature Review Neuroscince*, 11(8):589–599, 2010.

[117] Rendong Yang and Zhen Su. Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics*, 26(12), 2010.

[118] Nir Yosef and Aviv Regev. Impulse Control: Temporal Dynamics in Gene Transcription. *Cell*, 144(6):886–896, 2011.

[119] William A Zehring, David A Wheeler, Michael Rosbash, and Jeffrey C Hall. P-Element Transformation with period Locus DNA Restores Rhythmicity to Mutant, Arrhythmic Drosophila melanogaster. *Cell*, 39(December):369–376, 1984.

[120] Bin Zhang, Chris Gaiteri, Liviu-gabriel Bodea, Zhi Wang, Joshua Mcelwee, Alexei A Podtelezhnikov, Chunsheng Zhang, Tao Xie, Linh Tran, Radu Dobrin, Eugene Fluder, Bruce Clurman, Stacey Melquist, Manikandan Narayanan, Christine Suver, Hardik Shah, Milind Mahajan, Tammy Gillis, Jayalakshmi Mysore, Marcy E Macdonald, John R Lamb, David A Bennett, Cliona Molony, David J Stone, Vilmundur Gudnason, Amanda J Myers, Eric E Schadt, Harald Neumann, Jun Zhu, and Valur Emilsson. Resource Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer' s Disease. *Cell*, 153(3):707–720, 2013.

[121] Ray Zhang, Nicholas F Lahens, Heather I Ballance, Michael E Hughes, and John B Hogenesch. A circadian gene expression atlas in mammals: implications for biology and medicine. *Proceedings of the National Academy of Sciences*, 111(45):16219–24, 2014.

[122] Tomasz Zielinski, Anne M. Moore, Eilidh Troup, Karen J. Halliday, and Andrew J. Millar. Strengths and Limitations of Period Estimation Methods for Circadian Data. *PLoS ONE*, 9(5):e96462, may 2014.

[123] Sanjin Zvonic, Andrey A Ptitsyn, Steven A Conrad, L Keith Scott, Z Elizabeth Floyd, Gail Kilroy, Xiying Wu, Brian C Goh, Randall L Mynatt, and Jeffrey M Gimble. Characterization of Peripheral Circadian Clocks in Adipose Tissues. *Diabetes*, 55(4):962–970, 2006.