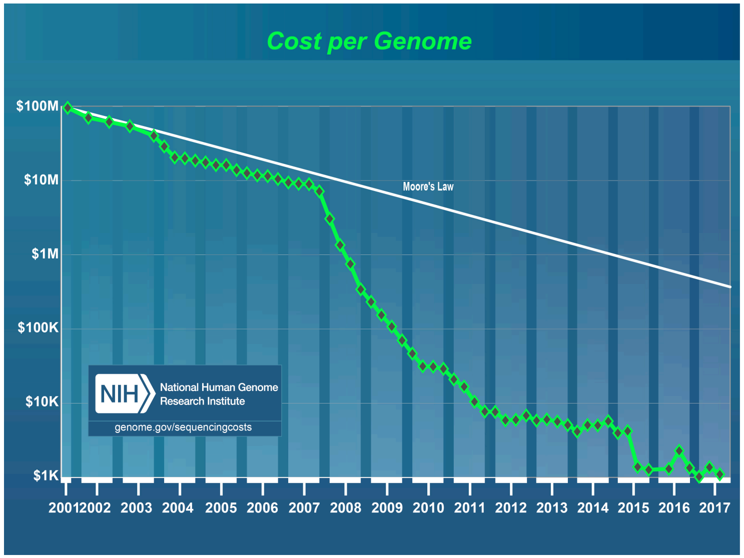
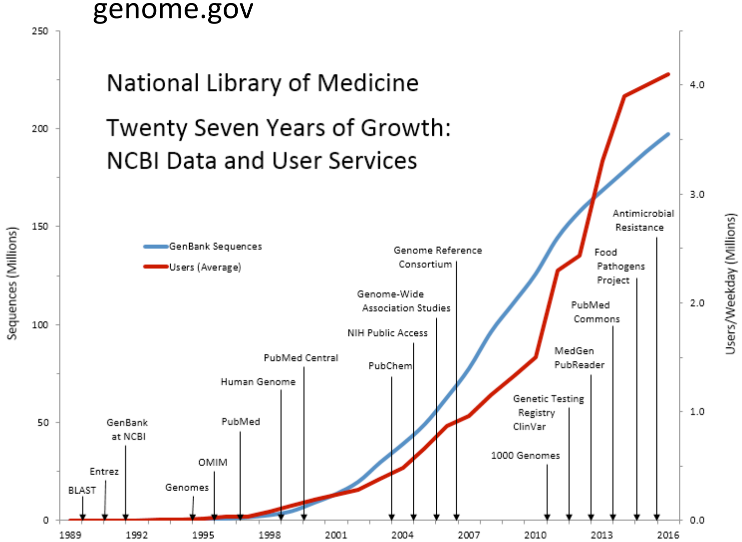
Improved Statistical Methods for Analyzing Circadian Rhythms in High-Throughput Data

Alan L. Hutchison, Ph.D.
M.D. Candidate 2019
University of Chicago
Ph.D. Advisor: Aaron R. Dinner

The challenge of properly analyzing massive amounts of data



genome.gov



datascience.nih.gov

Statistical analysis of circadian rhythms

↓

Identified errors in premier methods

- Hutchison & Dinner, 2017 *bioRxiv* 10.1101/118547

↓

Developed improved and correct methods

- Hutchison *et al.* 2015 *PLoS Comp. Biol.* 11(3)
- Hutchison *et al.* 2018 *J. Biol. Rhythms* 33(4)

↓

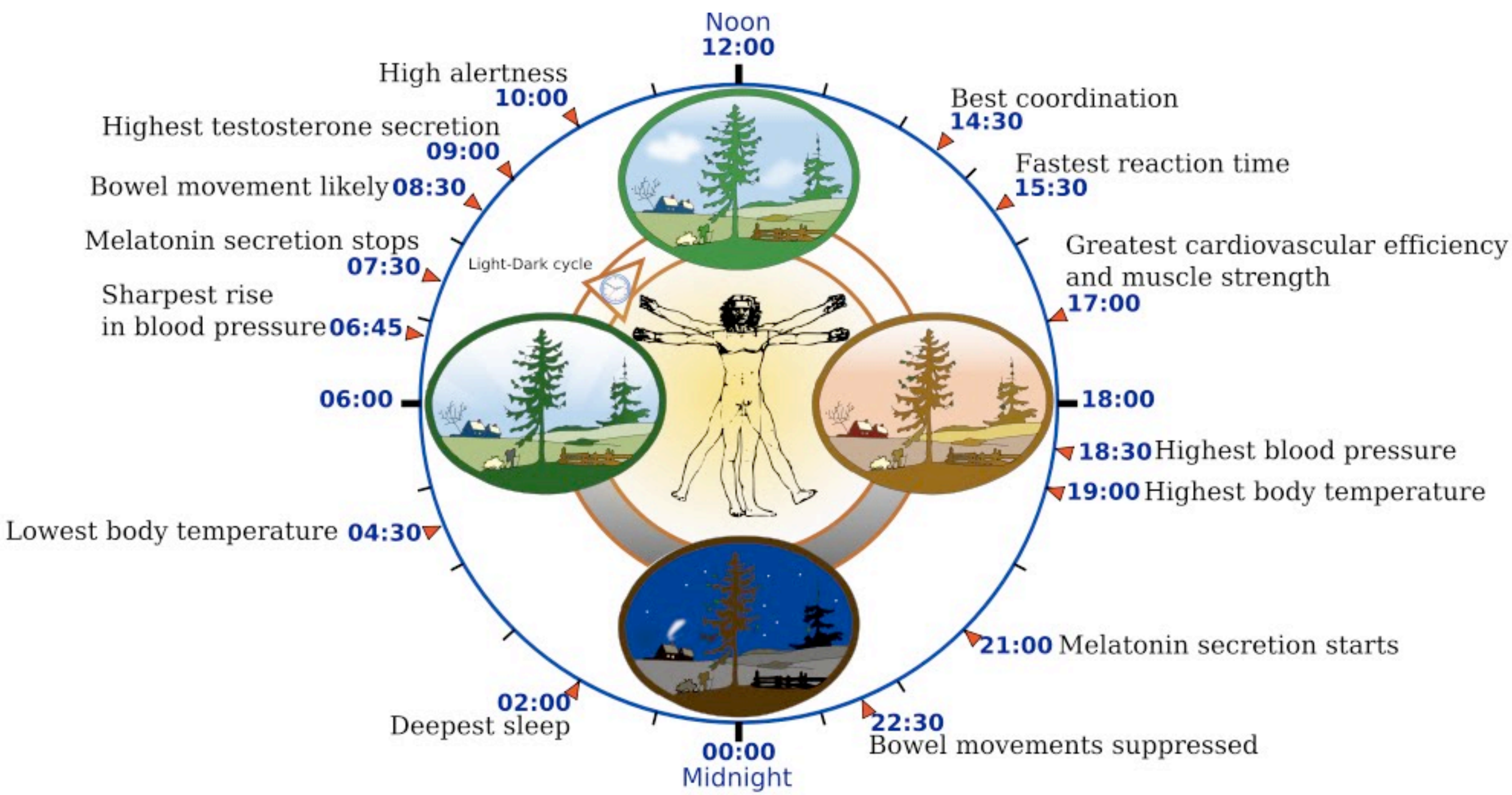
Applied methods to diverse circadian questions (fruit flies, beta-cells, microbiome)

- Perelis *et al.* 2015 *Science.* (350) 6261
- Flourakis *et al.* 2015. *Cell* 162
- Leone *et al.* 2015 *Cell Host-Microbe* 17

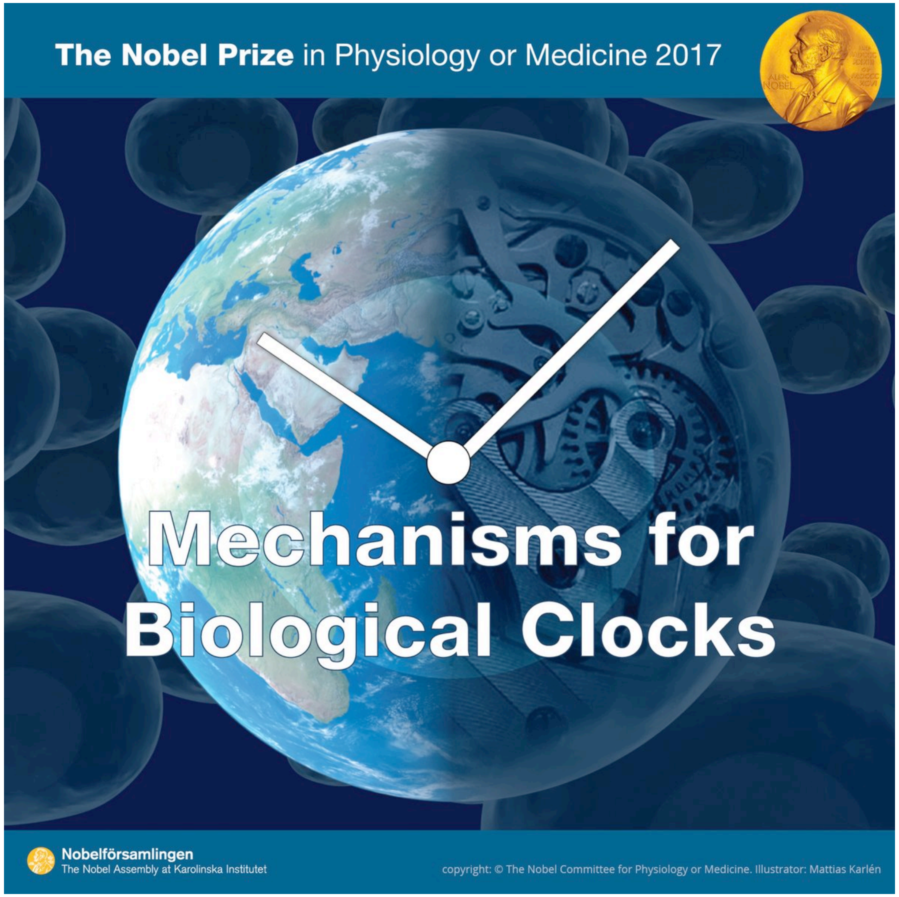
↓

Interest in methodology and circadian biology of gut microbiome

Circadian Rhythms are physiological rhythms regulated by an internal clock



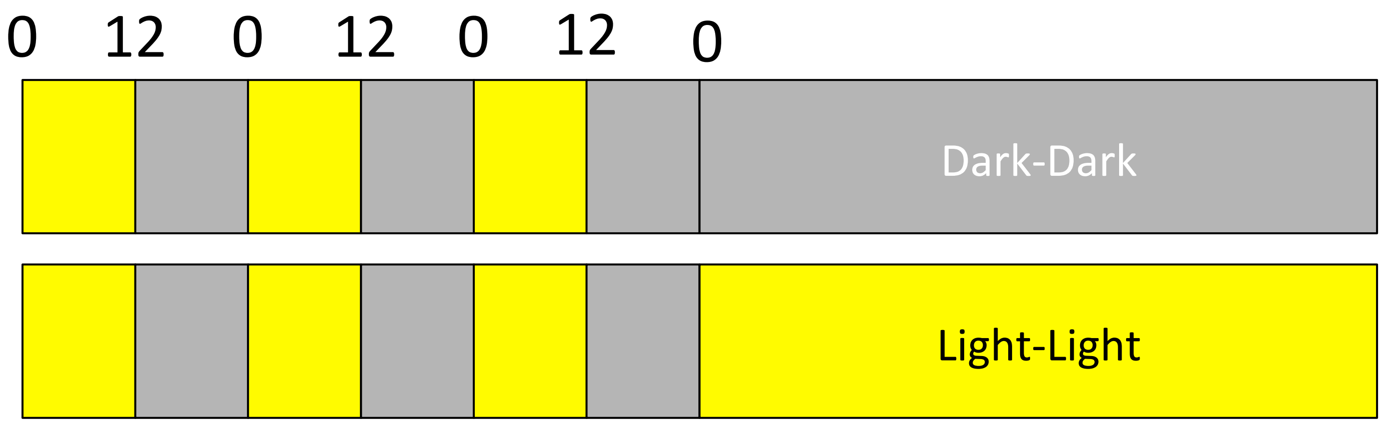
Dis-regulation of circadian processes can cause physiological changes and disease



- Long *et al.* “Morning vaccination enhances antibody response over afternoon vaccination: A cluster-randomised trial” *Vaccine* 2016 34(24)
- Scheer *et al.* “Adverse metabolic and cardiovascular consequences of circadian misalignment.” *PNAS* 2009 106(11)

Circadian experiment

12 h light
12 h dark

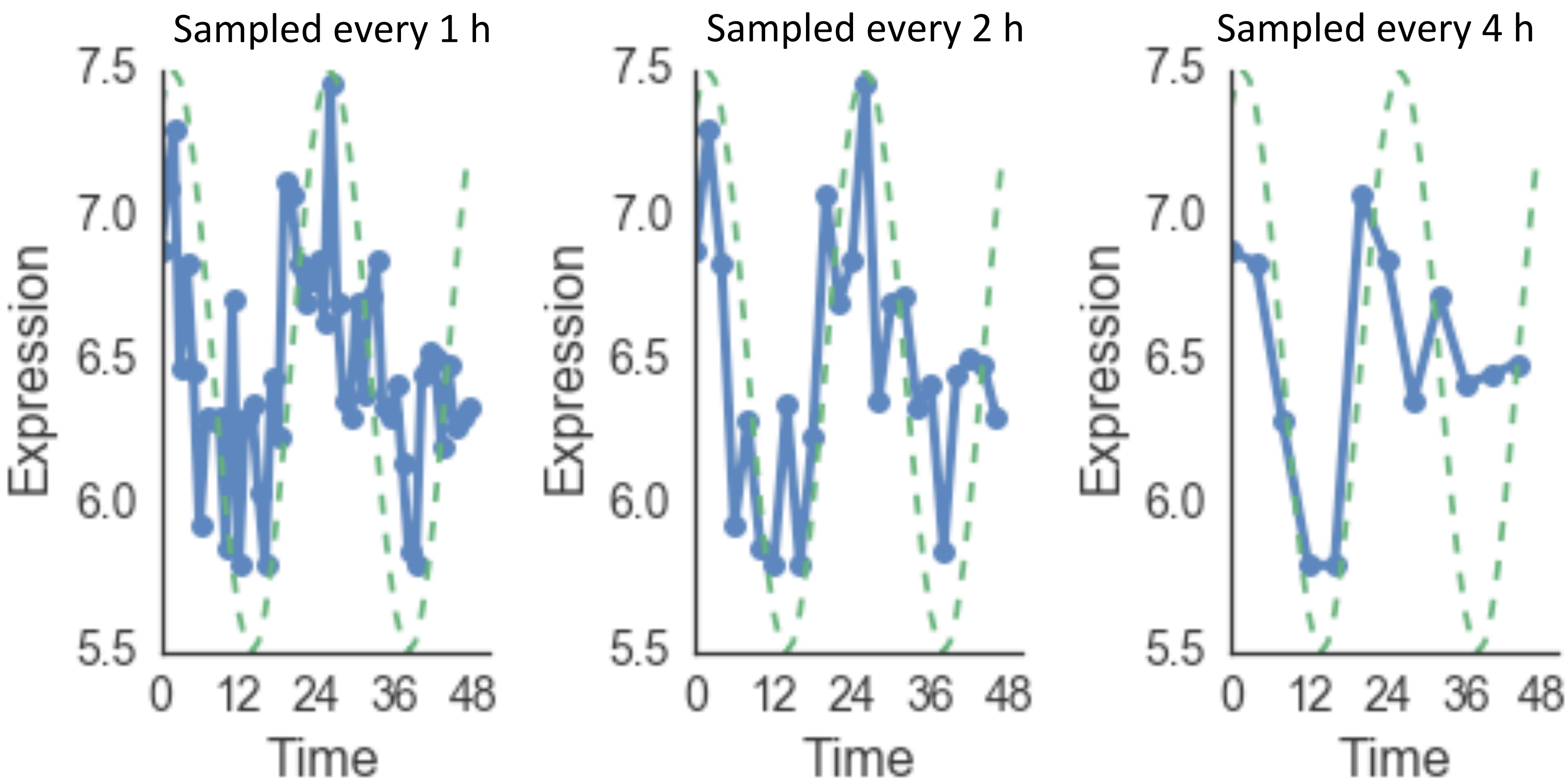


Molecular circadian experiment

12 h light
12 h dark

Every time point

Three challenges of rhythm detection



1. Sparse sampling of data
2. High noise of measurements
3. High false positive rate

Time series data from
Hughes *et al. PLoS Gen.* 2009 5(4)

Incorrect methods can lead to incorrect identification of rhythmicity

Science

RESEARCH ARTICLES

Cite as: L. S. Mure *et al.*, *Science* 10.1126/science.aao0318 (2018).

Diurnal transcriptome atlas of a primate across major neural and peripheral tissues

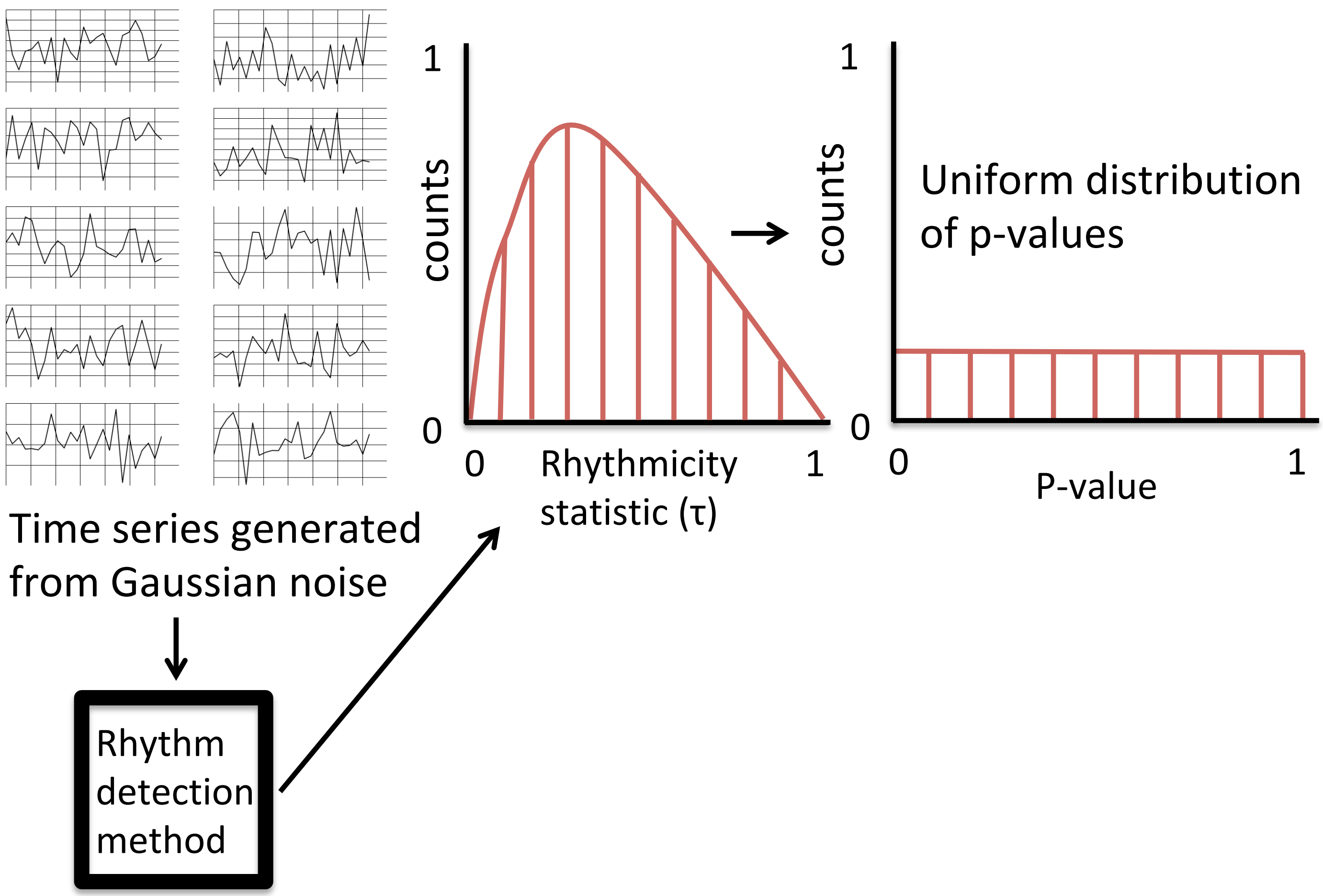
Ludovic S. Mure,[1] Hiep D. Le,[1] Giorgia Benegiamo,[1] Max W. Chang,[1,2] Luis Rios,[1] Ngalla Jillani,[3] Maina Ngotho,[3] Thomas Kariuki,[3] Ouria Dkhissi-Benyahya,[4] Howard M. Cooper,[4*] Satchidananda Panda[1*]

[1]Regulatory Biology Laboratory, Salk Institute for Biological Studies, 10010, North Torrey Pines Road, La Jolla, CA 92037, USA. [2]Department of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA. [3]Institute of Primate Research, National Museums of Kenya, Nairobi, Kenya. [4]Université Lyon, Universite Claude Bernard Lyon 1, Inserm, Stem Cell and Brain Research Institute U1208, 69500 Bron, France.

*Corresponding author. Email: howard.cooper@inserm.fr (H.M.C.); satchin@salk.edu (S.P.)

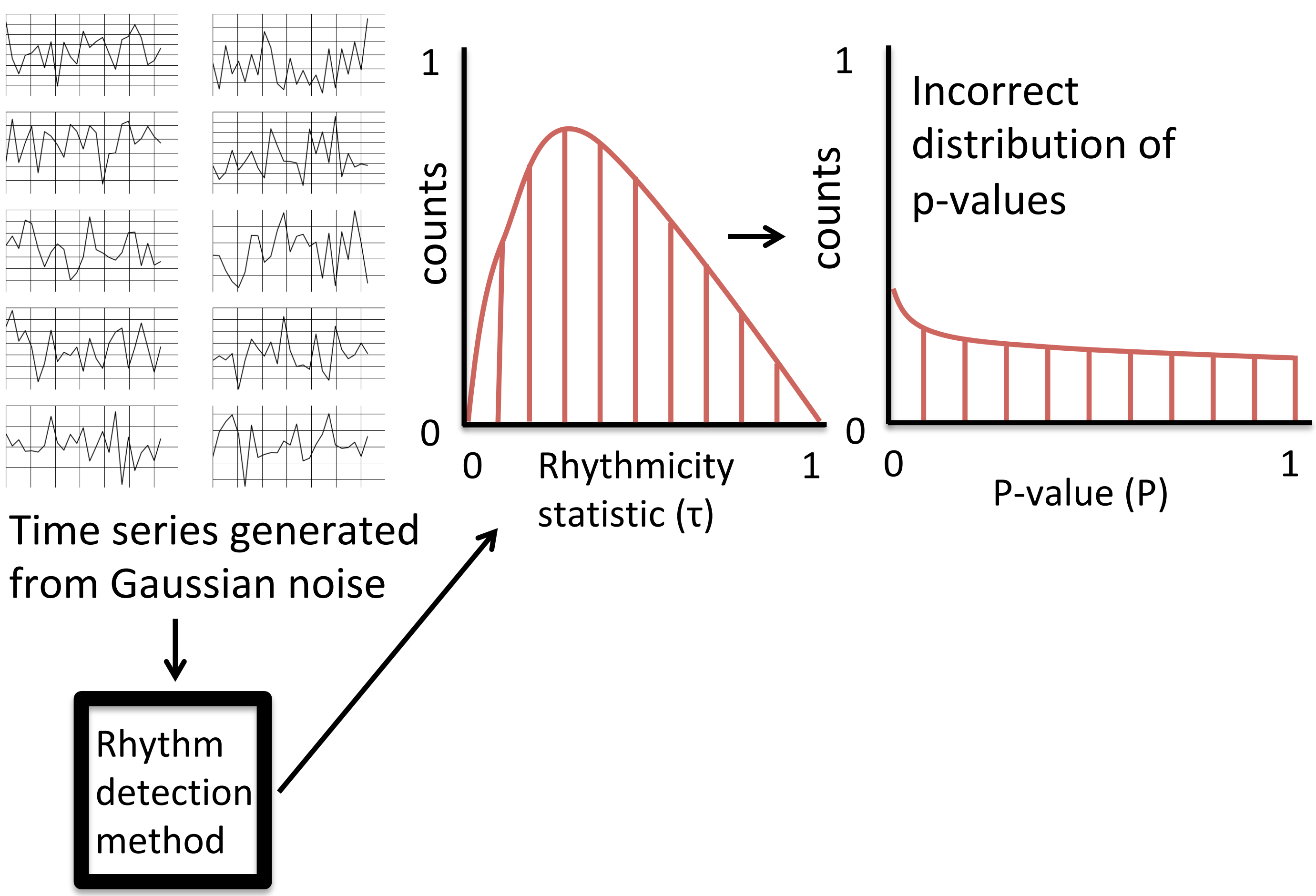
Diurnal gene expression patterns underlie time of the day-specific functional specialization of tissues. However, available circadian gene expression atlases of a few organs are largely from nocturnal vertebrates. We report the diurnal transcriptome of 64 tissues, including 22 brain regions, sampled every 2 hours over 24 hours, from the primate *Papio anubis* (baboon). Genomic transcription was highly rhythmic with up to 81.7% of protein-coding genes showing daily rhythms in expression. In addition to

Well-behaved methods will have uniformly distributed p-values in null conditions



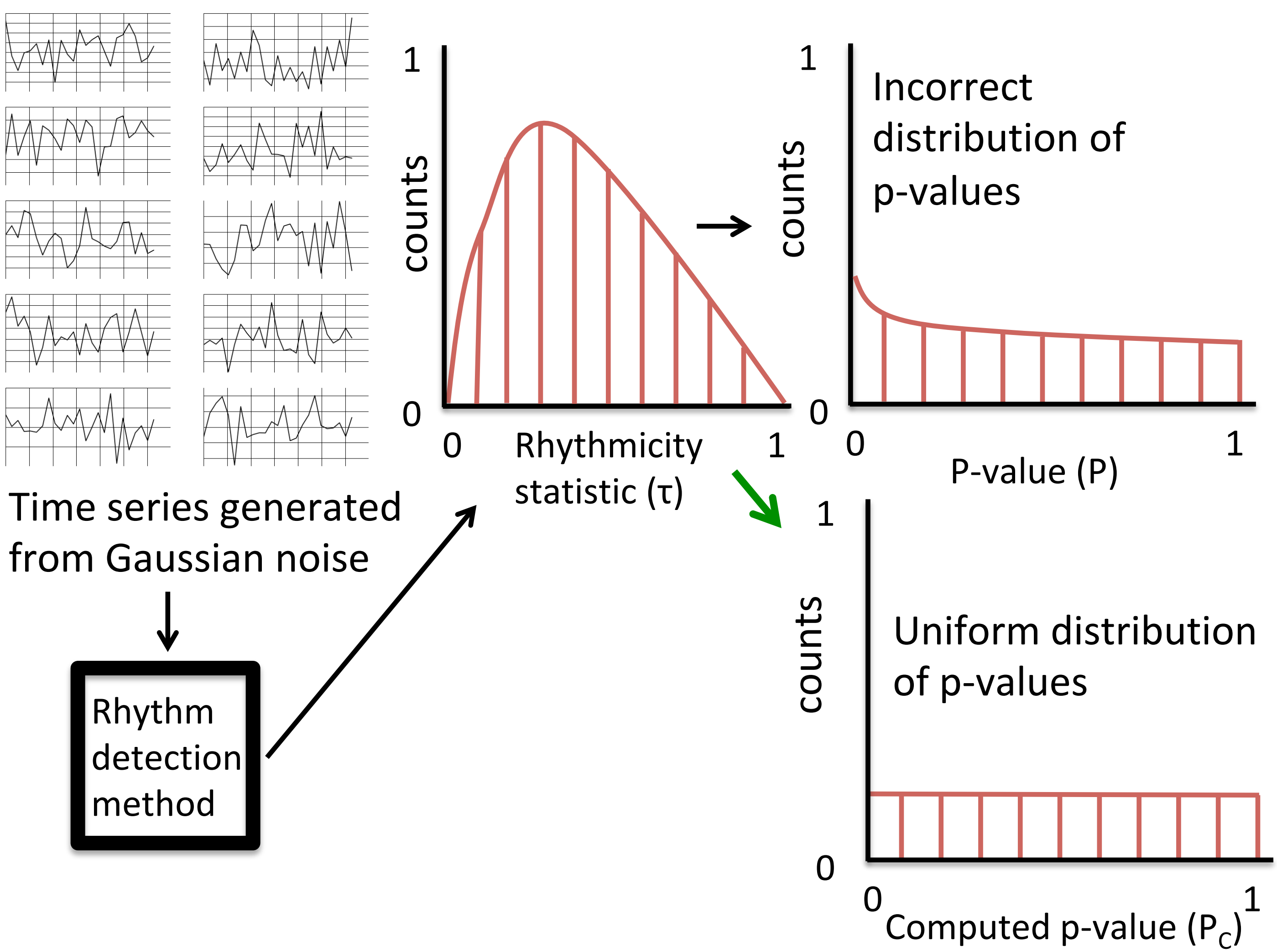
Hutchison *et al.* 2015

Well-behaved methods will have uniformly distributed p-values in null conditions



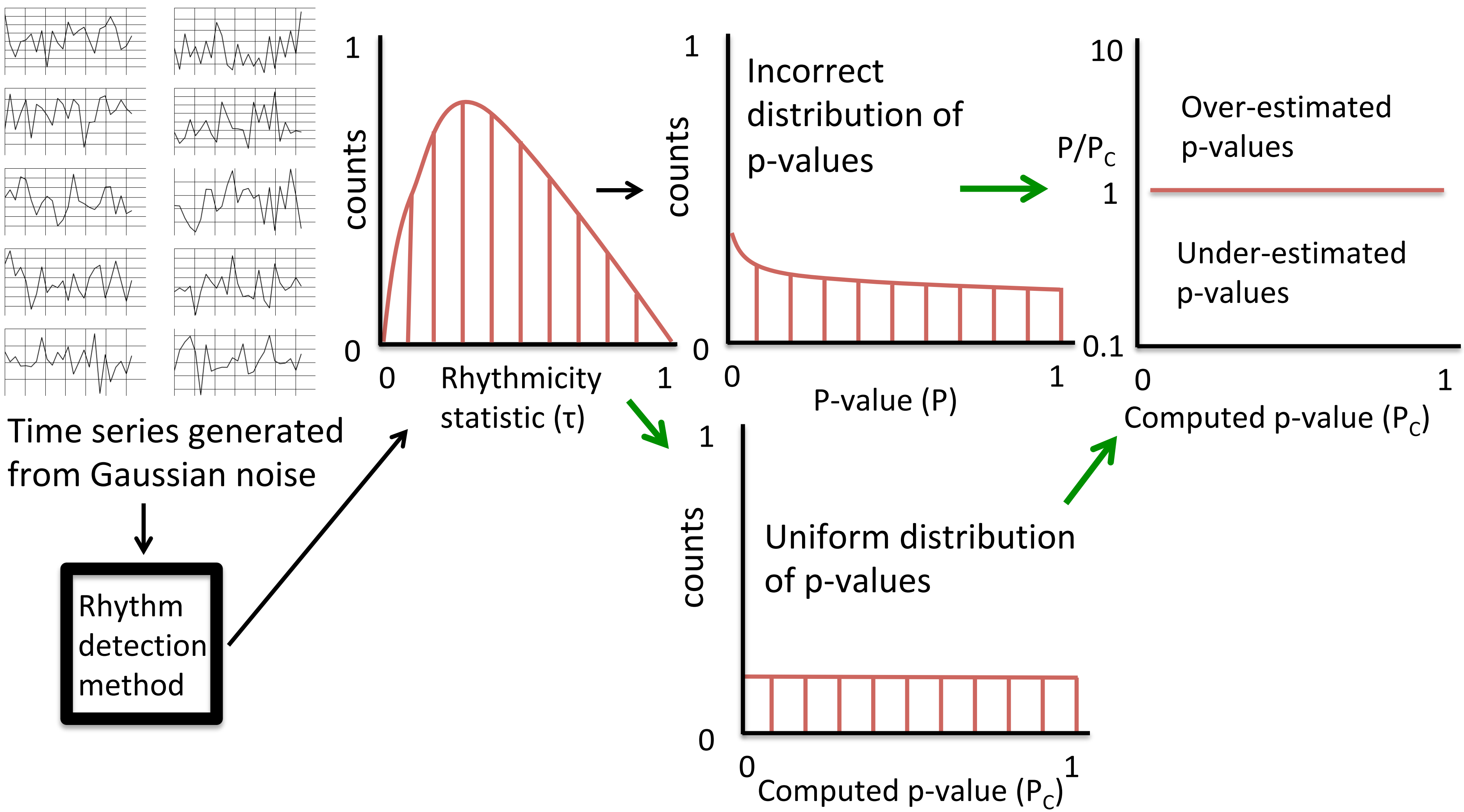
Hutchison *et al.* 2015

Simulation of null data allows for the computation of the correct p-values



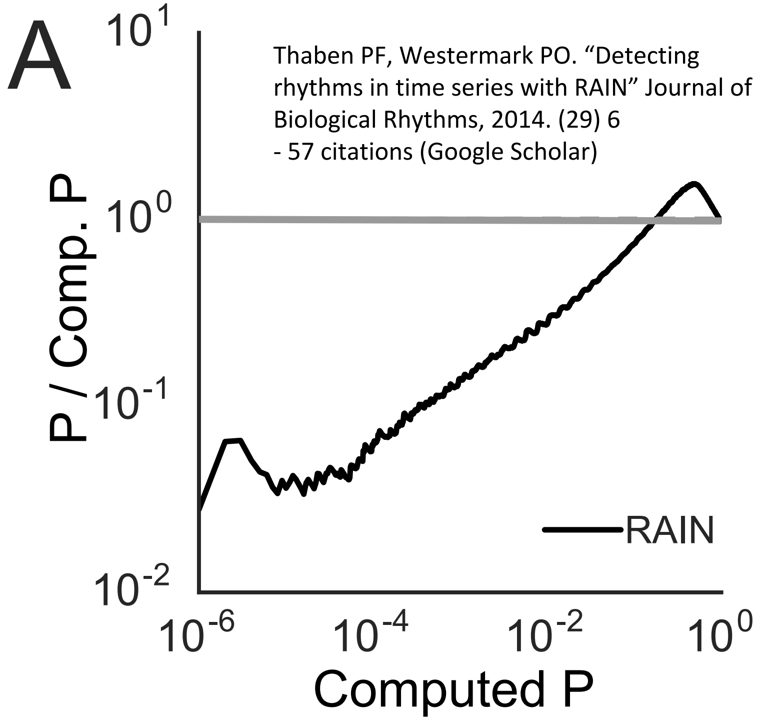
Hutchison *et al.* 2015

Simulation of null data allows for the computation of the correct p-values

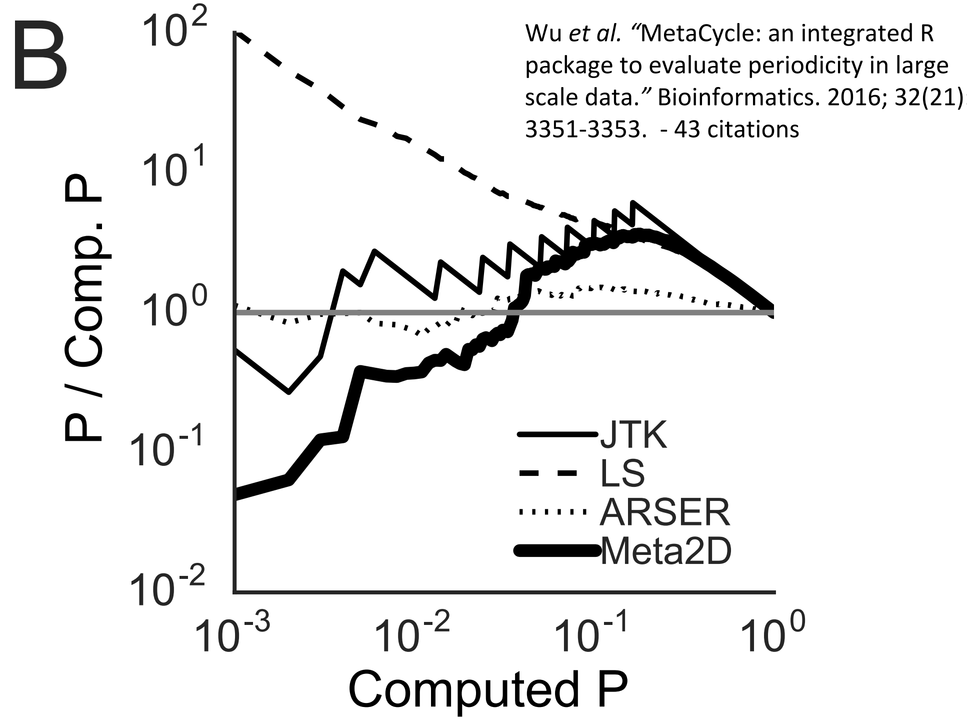


Hutchison *et al.* 2015

Leading methods in the field have artifactually low p-values



RAIN fails to account for correlation of p-values from different waveforms

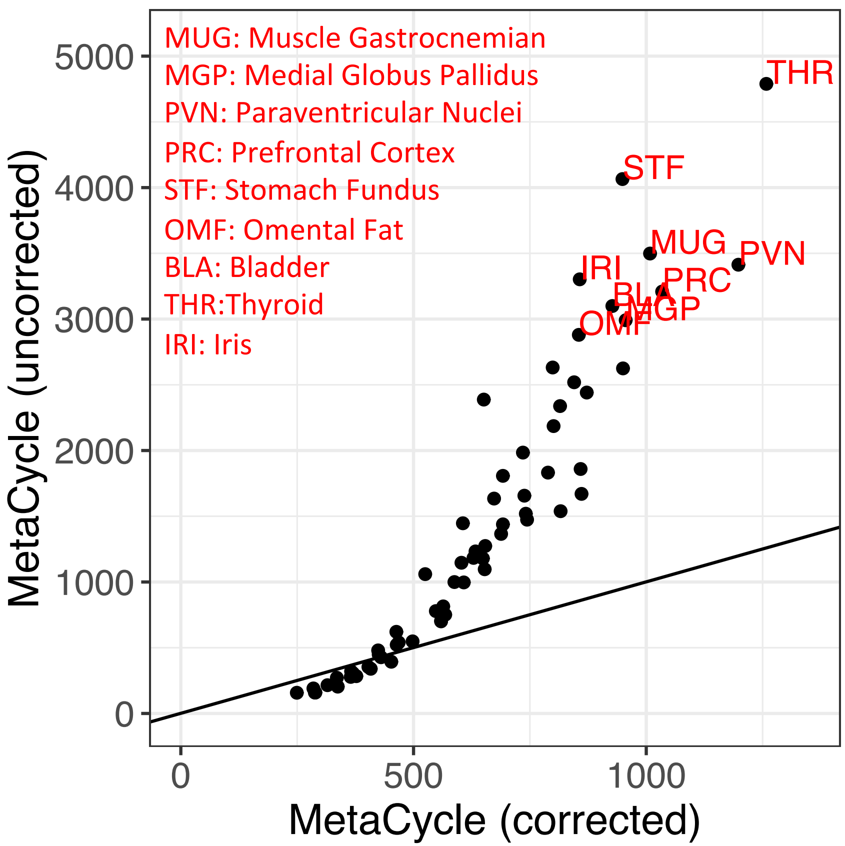


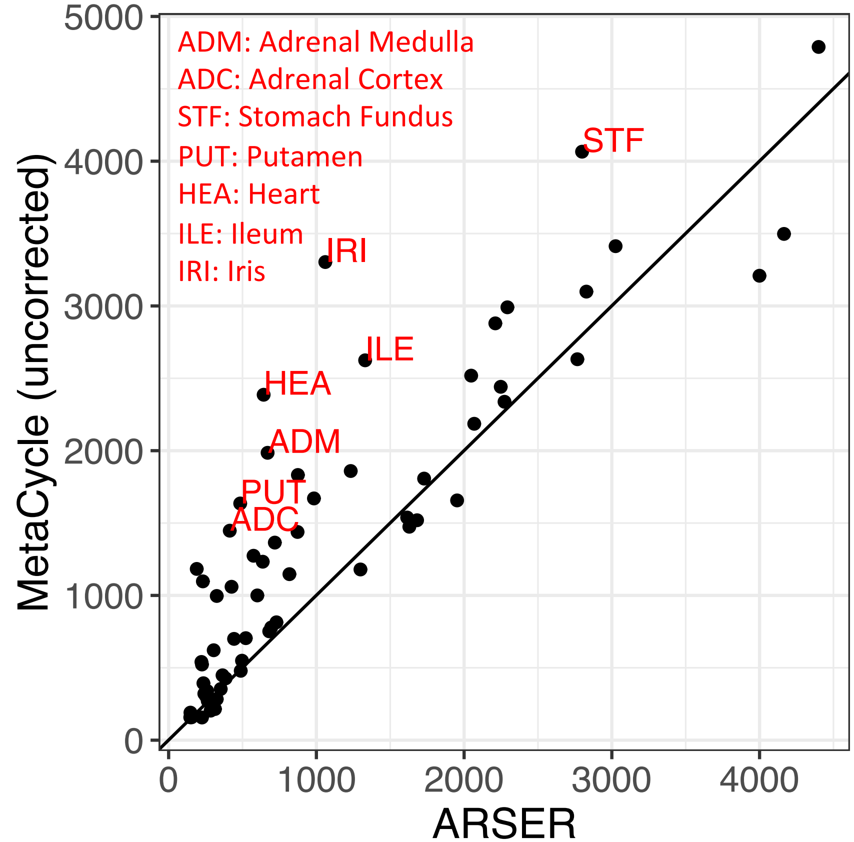
MetaCycle fails to account for correlation of p-values from different methods

Null time series (generated from Gaussian noise)

Hutchison & Dinner 2017

Different gene rhythmicity is found when accurately calculating p-values





Hutchison & Dinner, *in preparation*

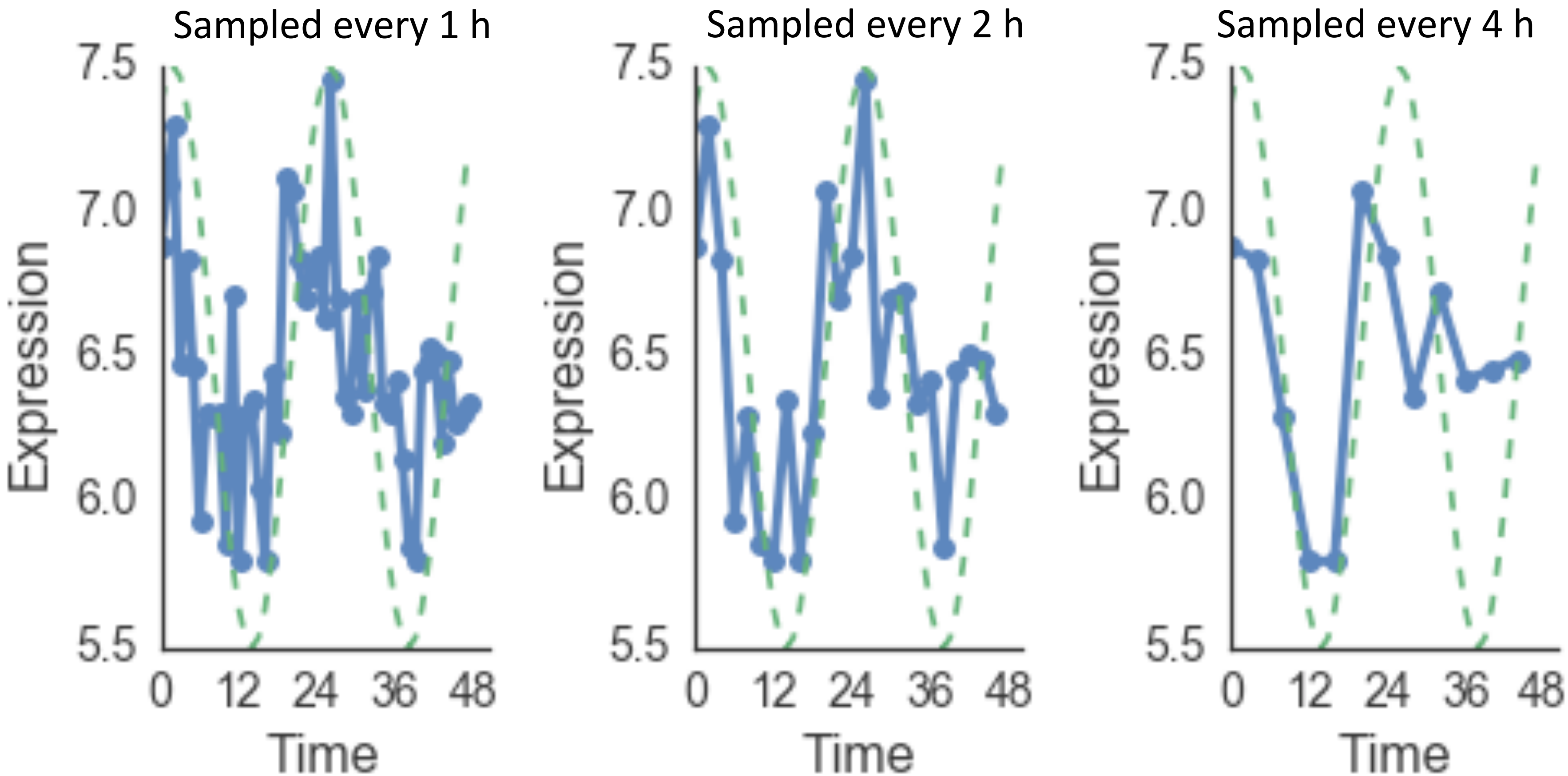
Outline

- Introduction
- **Statistical inaccuracies in p-value calculation**
  - **Computational correction**
    - **Hutchison *et al.* 2015**
  - **Identification of errors**
    - **Hutchison *et al.* 2015**
    - **Hutchison & Dinner 2017**
- Methodological improvements in rhythm detection
- Future directions

Outline

- Introduction
- Statistical inaccuracies in p-value calculation
- **Methodological improvements in rhythm detection**
  - **Adapting empirical Bayesian methods to rhythm detection**
    - **Hutchison *et al.* 2018**
- Future directions

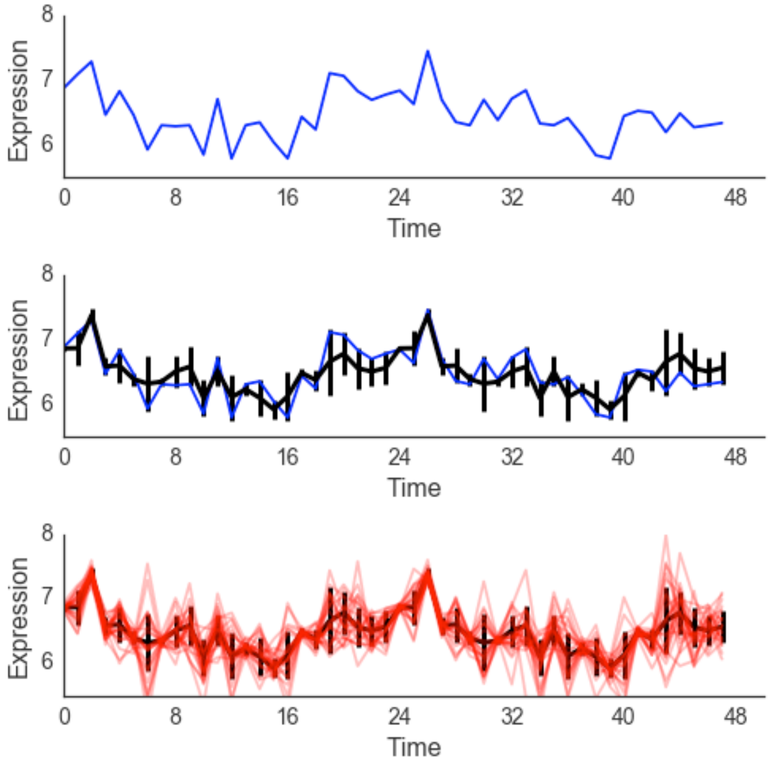
Three challenges of rhythm detection

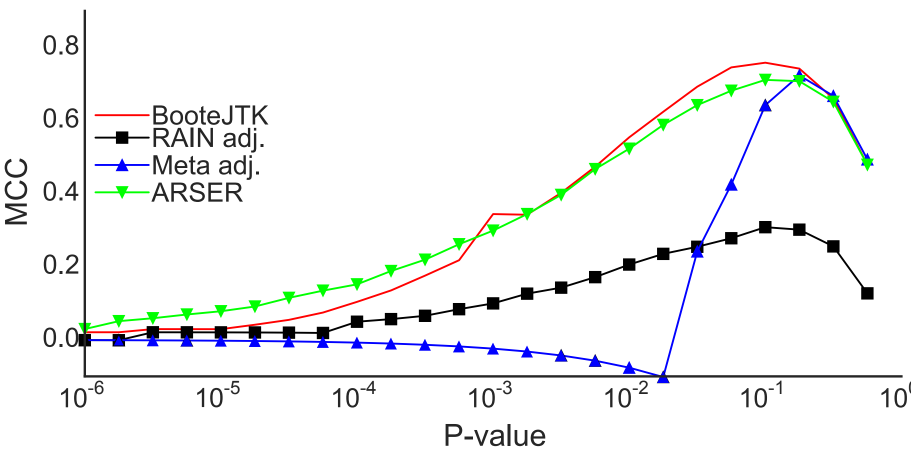


- Sparse sampling of data
- High noise of measurements
- High false positive rate

Time series data from Hughes *et al. PLoS Gen.* 2009

Introducing empirical Bayes variance estimation via bootstrapping improves rhythm detection



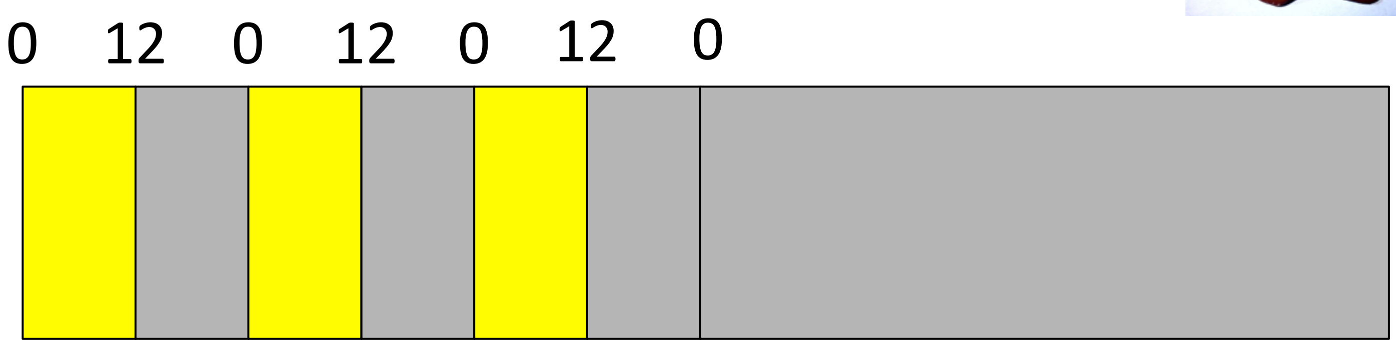


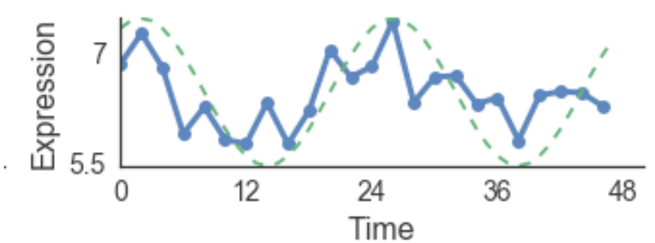
1100 time series, 11 asymmetries, cosine with Gaussian noise added to each point with noise-to-amplitude ratio of 1

Hutchison *et al.* 2018

Hughes *et al.* 1h liver dataset

12 h light
12 h dark



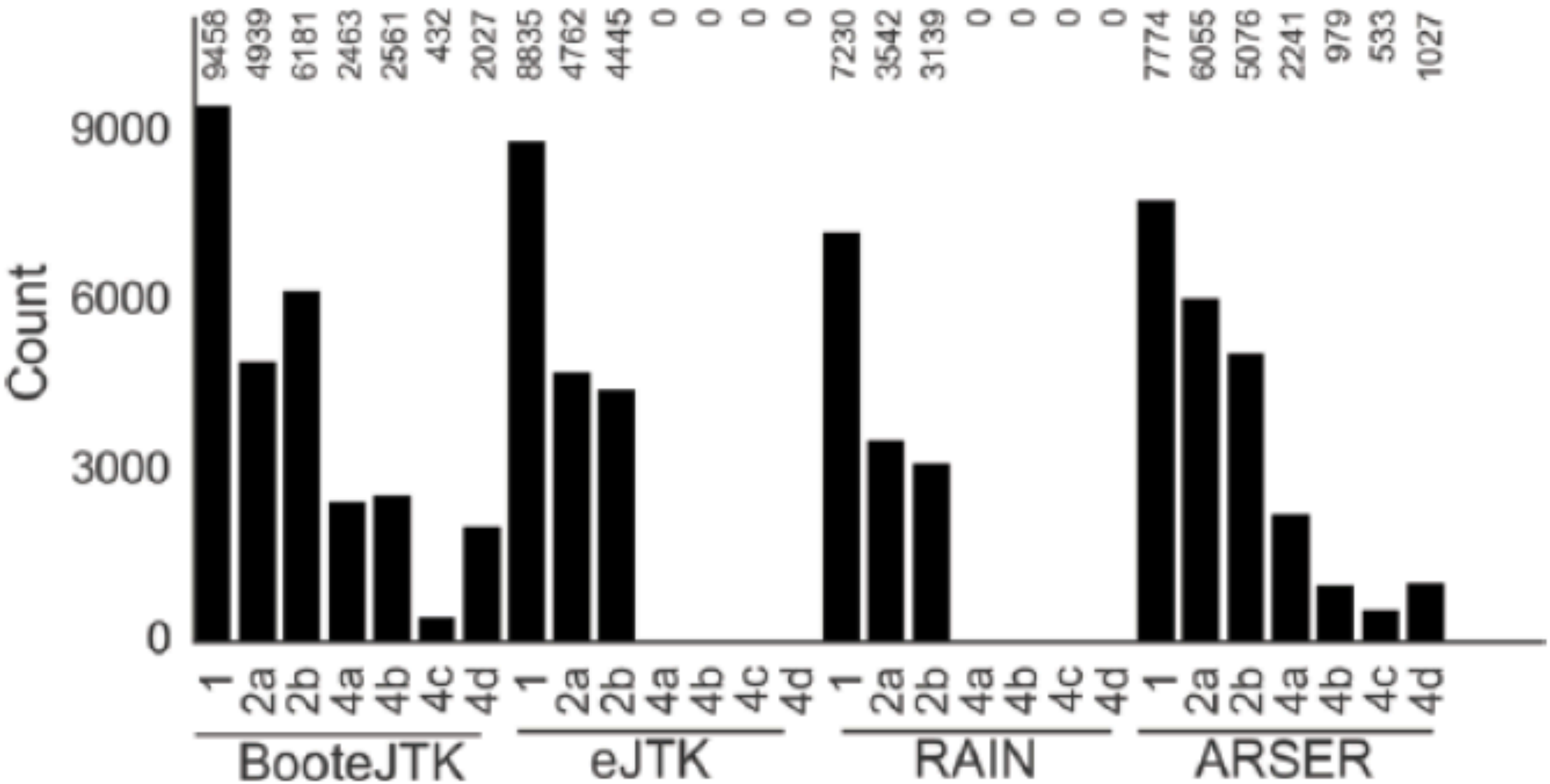


Hutchison *et al.* 2018

1: 0, 1, 2, 3, 4...
2a: 0, 2, 4, 6...
2b: 1, 3, 5, 7...
4a: 0, 4, 8, 12...
4b: 1, 5, 9, 13...

Hughes *et al.* (2009) "Harmonics of Circadian Gene Transcription in Mammals." *PLoS Genetics,* 2009. 5(4): e1000442

Bootstrap eJTK performs better on sparse data than other methods

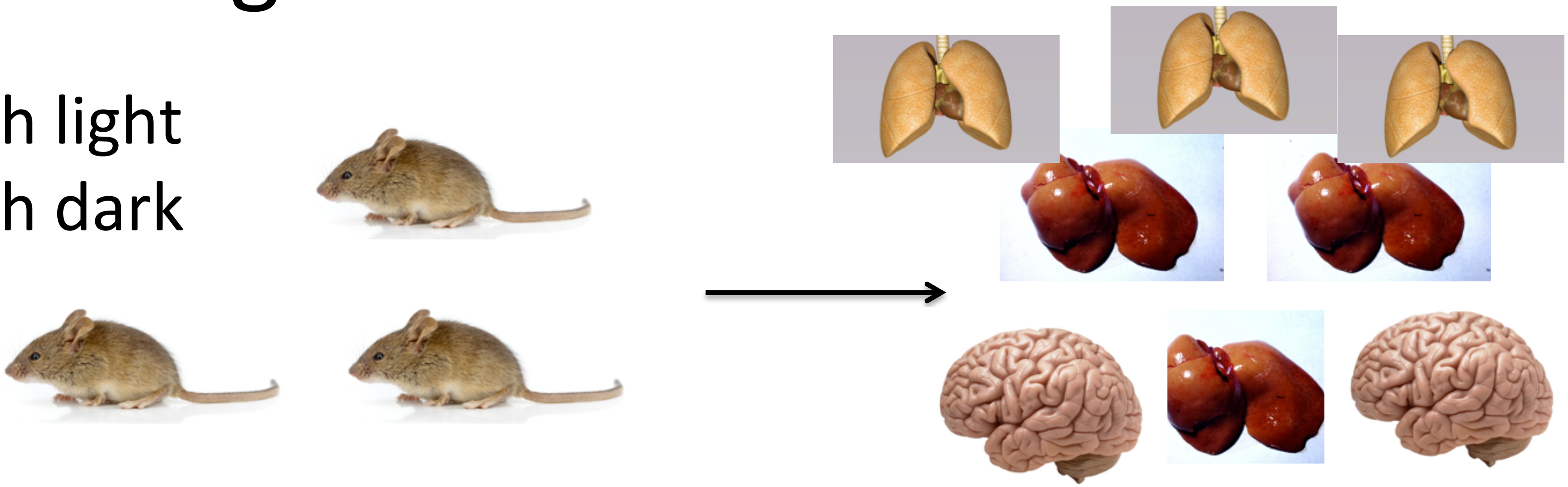


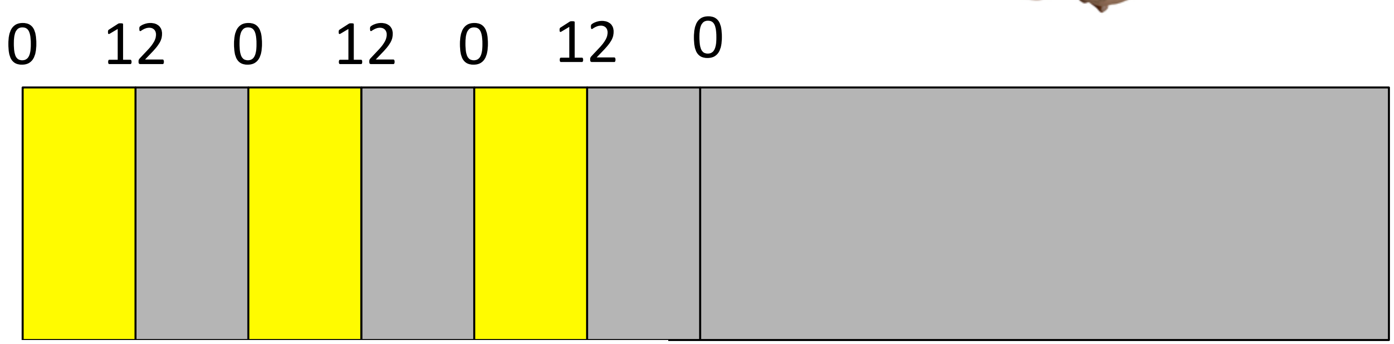
Rhythmicity criteria: Benjamini-Hochberg adjusted p-value <0.05

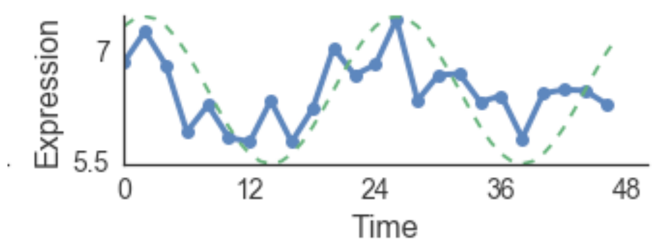
Hutchison *et al.* 2018

Zhang *et al.* 2h 12 tissue dataset

12 h light
12 h dark



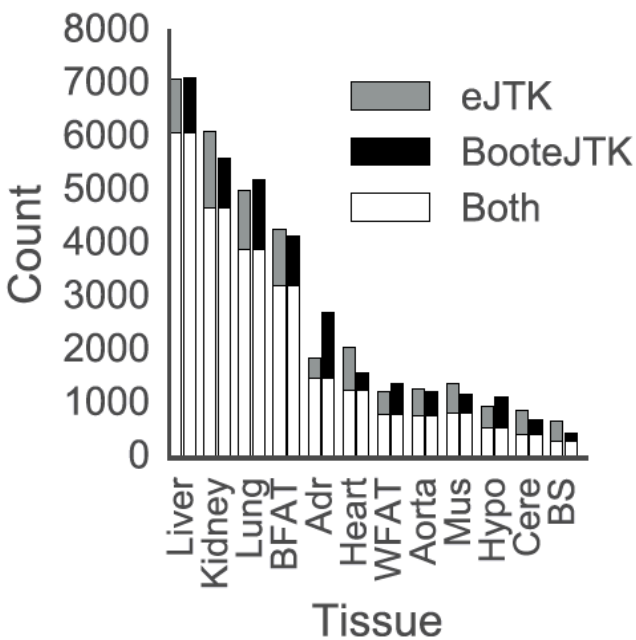




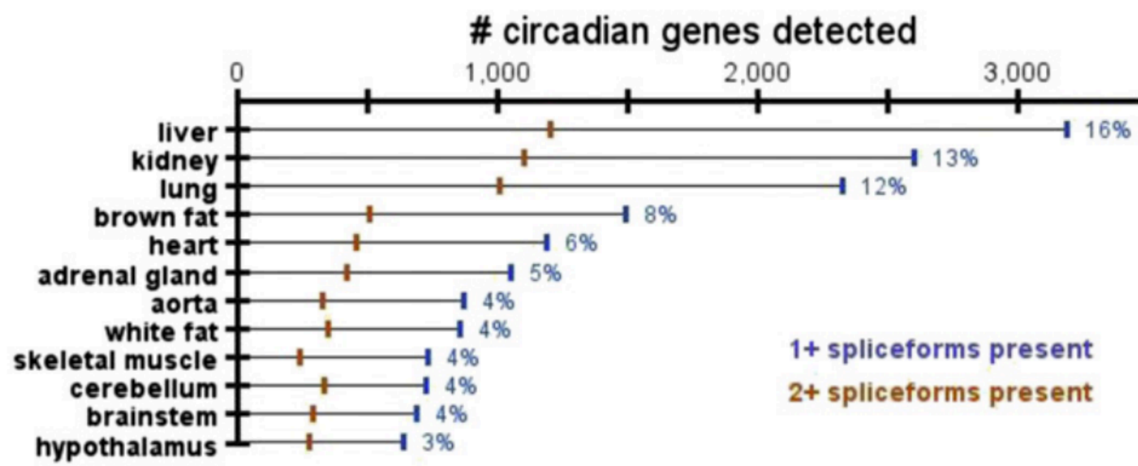
Zhang *et al.* (2014) “A circadian gene expression atlas in mammals: Implications for biology and medicine.” PNAS (111) 45

Hutchison *et al.* 2018

Bootstrap eJTK reveals greater rhythmicity across tissues



11,731/20,038 (55%) rhythmic genes

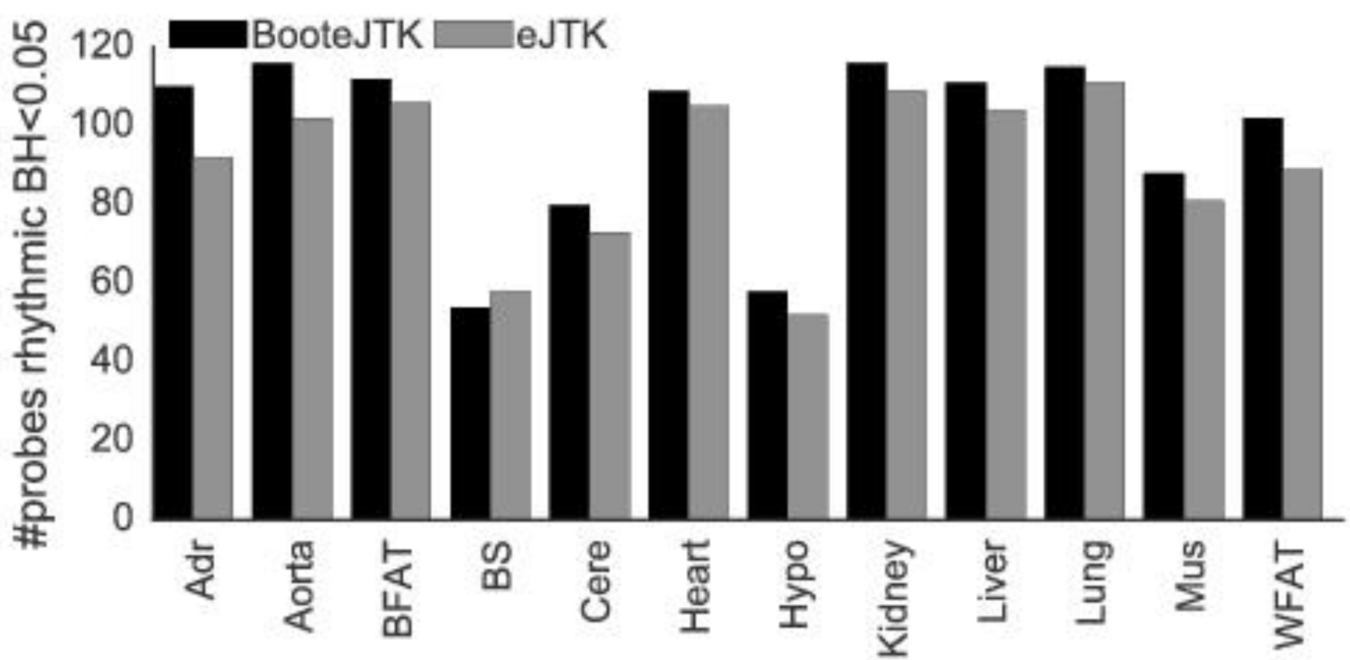


Zhang *et al.* (2014) "A circadian gene expression atlas in mammals: Implications for biology and medicine." PNAS (111) 45

Hutchison *et al.* 2018

Looking at commonly rhythmic genes reveals novel rhythmic pathways

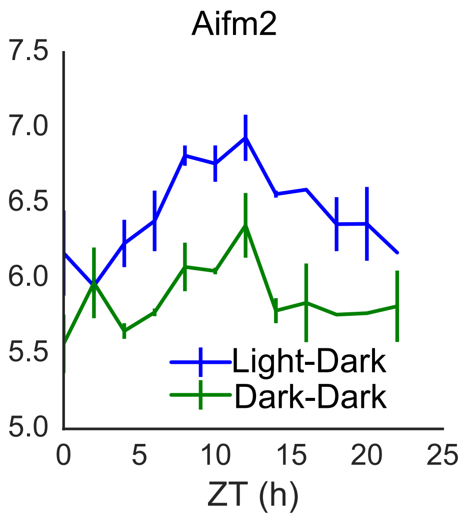
Probes rhythmic in >9 tissues



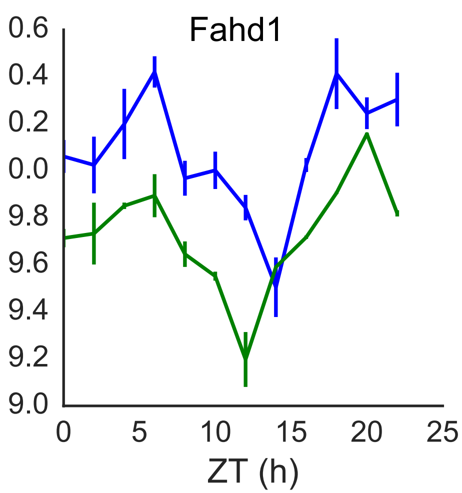
- Stress response
- Heat Shock Protein 70
- Endoplasmic reticulum

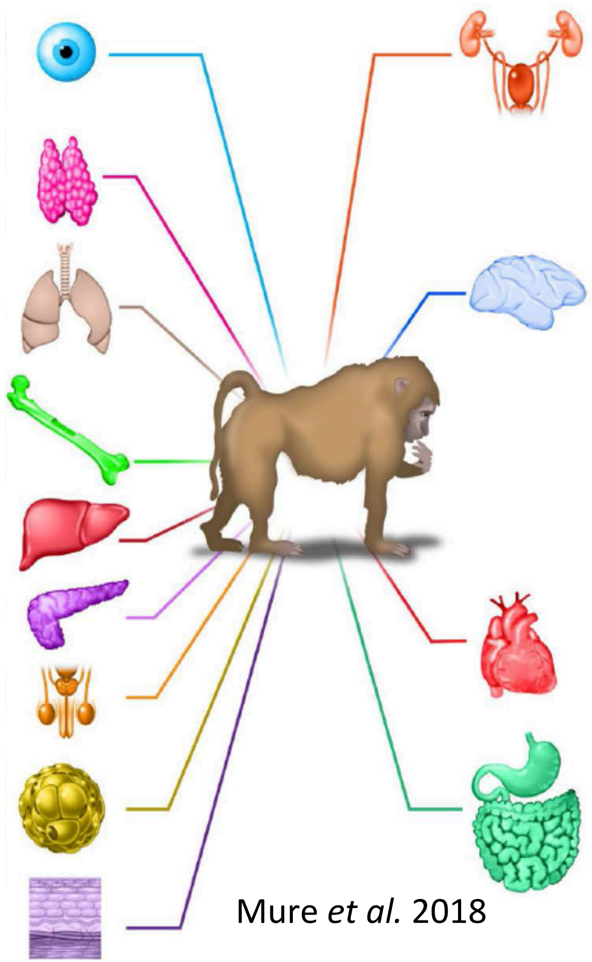
Hutchison *et al.* 2018

Circadian future directions



Jouffe *et al.* 2013
Hughes *et al.* 2009





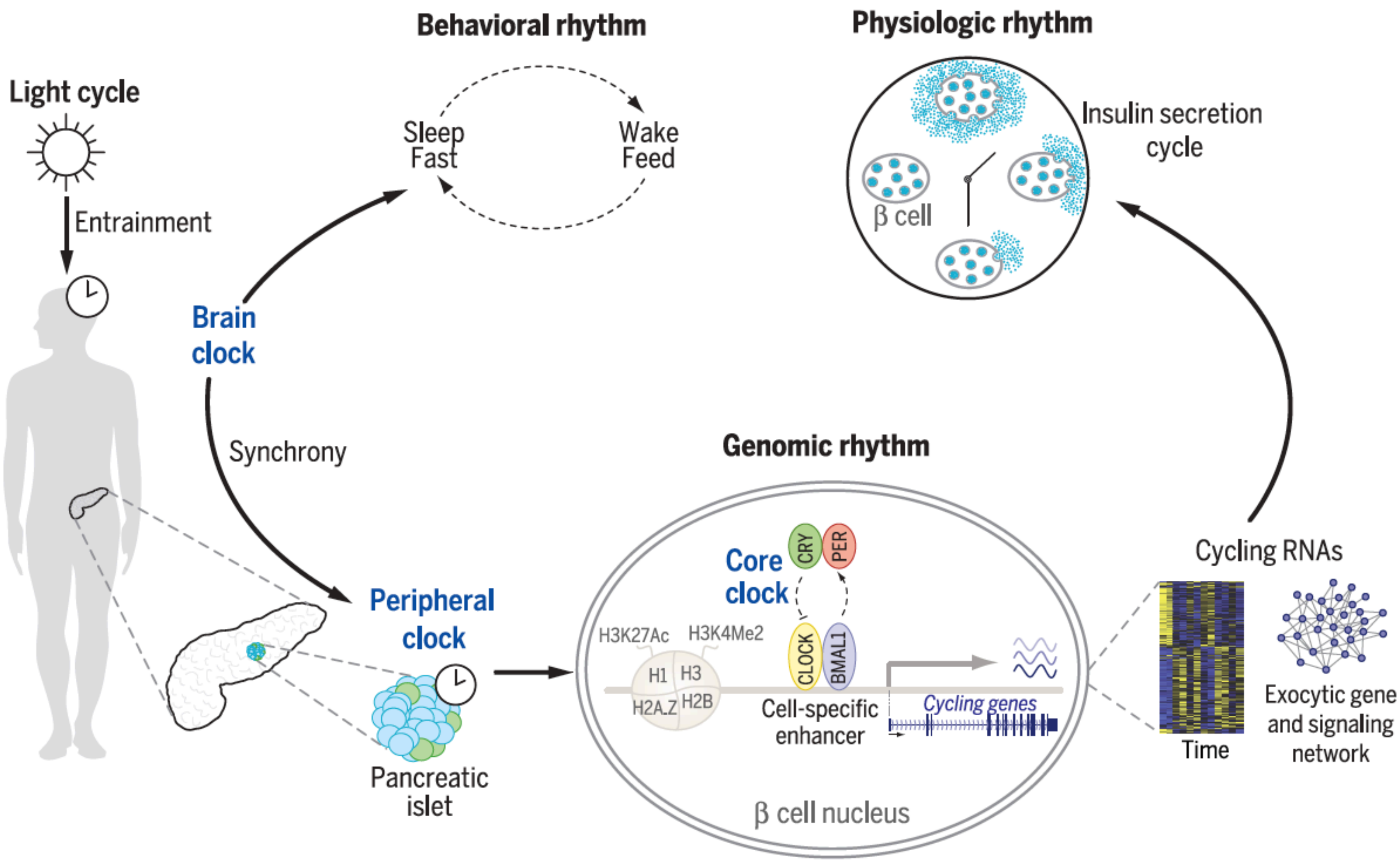
Mure *et al.* 2018

Hutchison & Dinner, *in preparation*

Collaborations

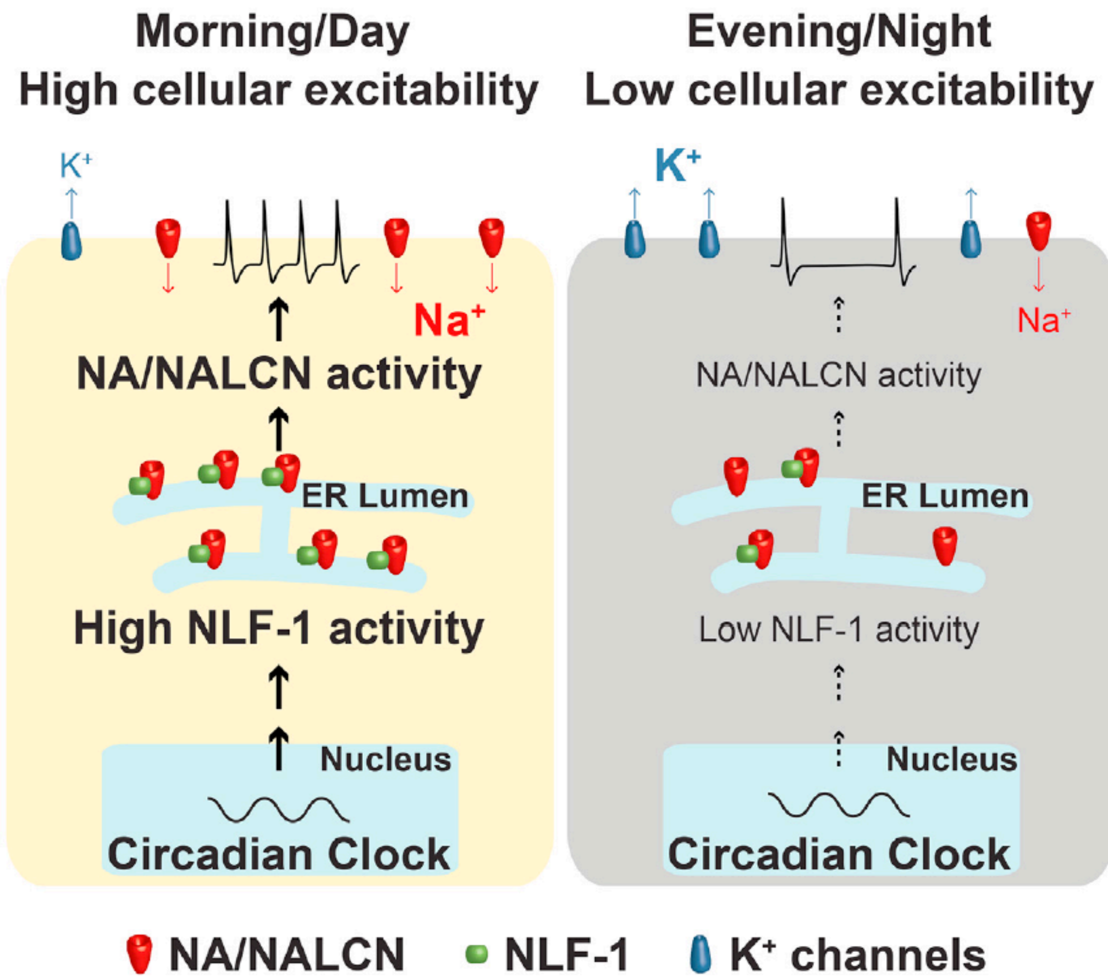
- Perelis *et al.* 2015 *Science.* (350) 6261
  - Joseph Bass group at Northwestern U
- Flourakis *et al.* 2015. *Cell* 162
  - Ravi Allada group at Northwestern U
- Leone *et al.* 2015 *Cell Host-Microbe* 17
  - Eugene Chang group at UChicago

Pancreatic beta-cell enhancers regulate rhythmic transcription of genes controlling insulin secretion



Perelis *et al.* 2015

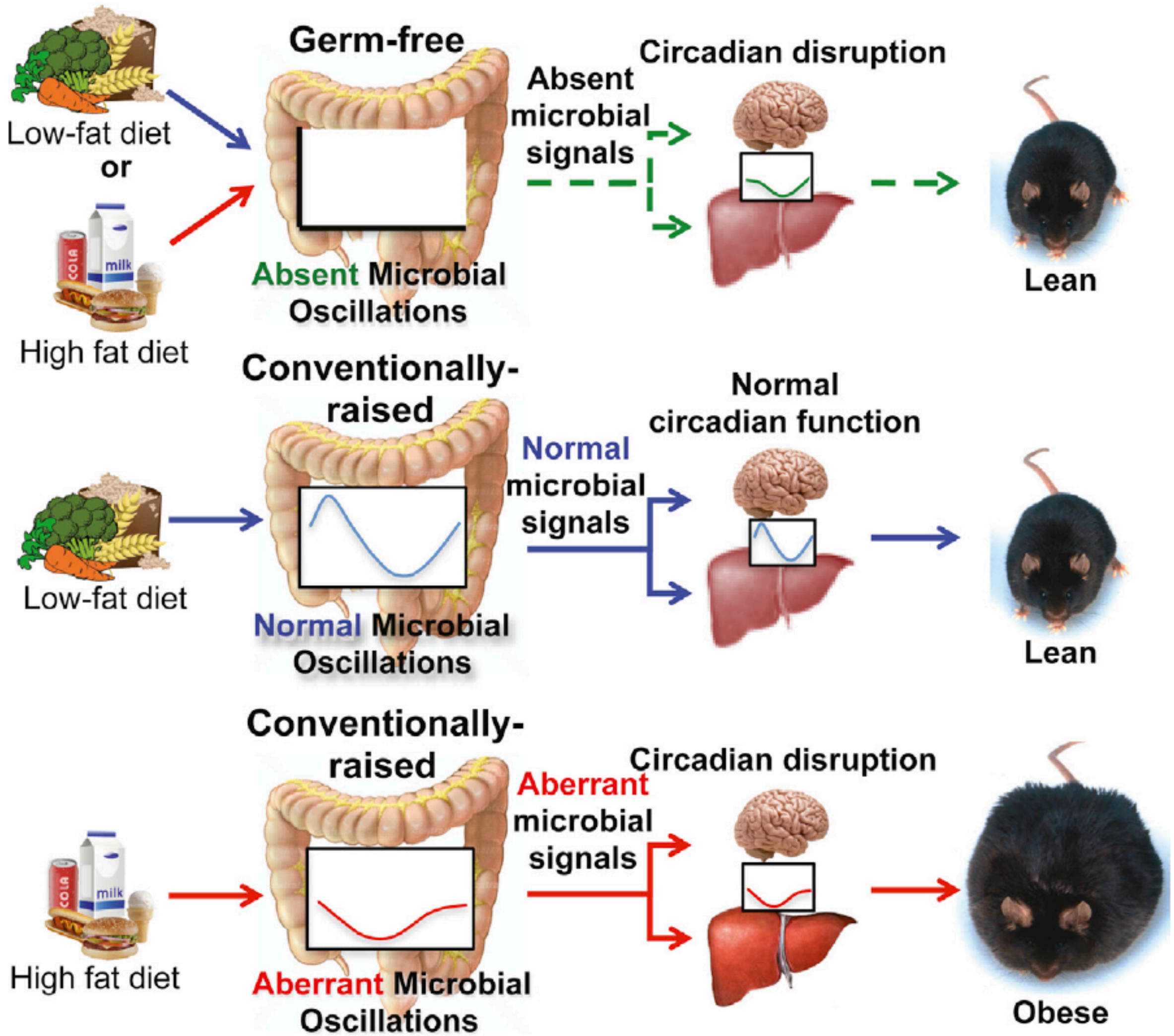
A Conserved Bicycle Model for Circadian Clock Control of Membrane Excitability



Two distinctly timed sodium and potassium electrical drives collaborate to directly control membrane excitability and neuronal function in a circadian manner.

Flourakis *et al.* 2015

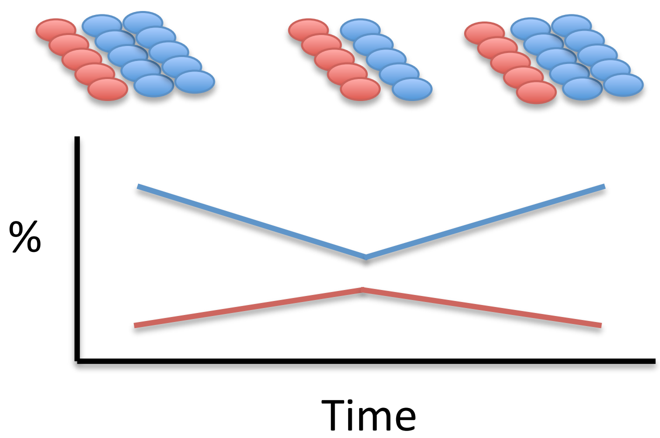
Effects of Diurnal Variation of Gut Microbes and High-Fat Feeding on Host Circadian Clock Function and Metabolism

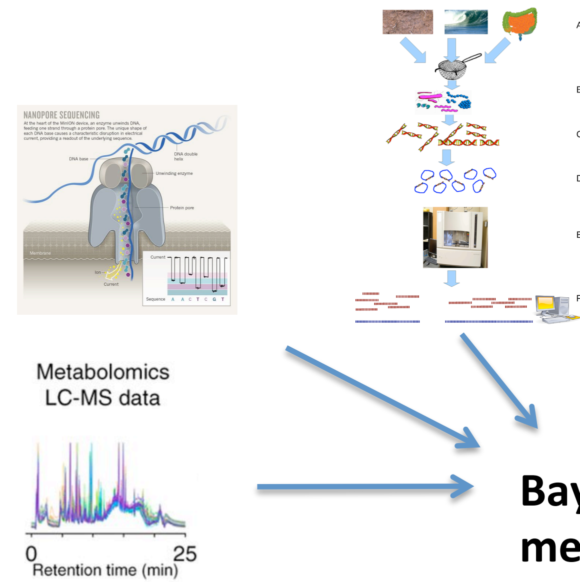


Leone *et al.* 2015

Future Interests

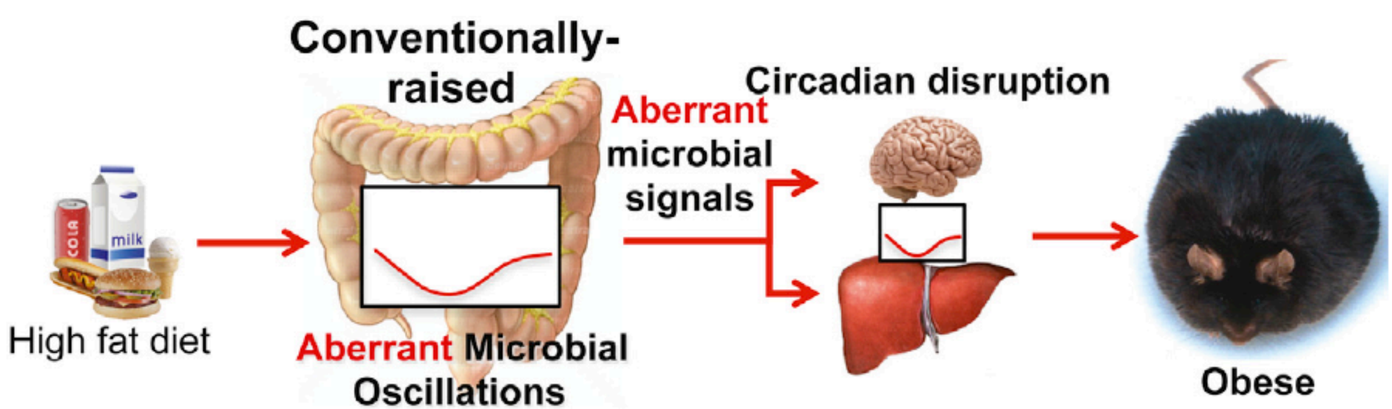
Exploration of 16S quantification limitations





Bayesian methods to improve quantification

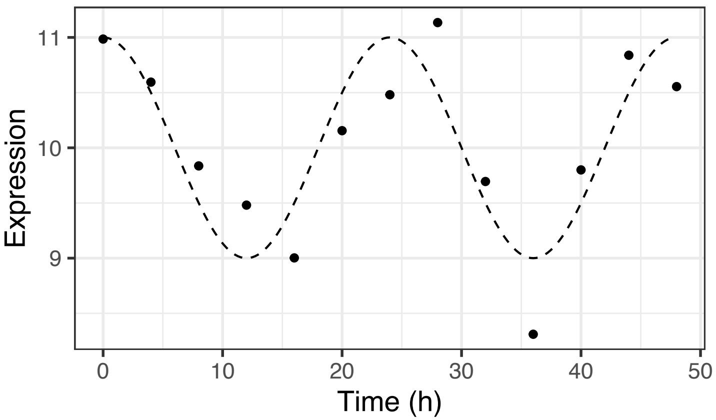
Circadian host-microbiome interactions



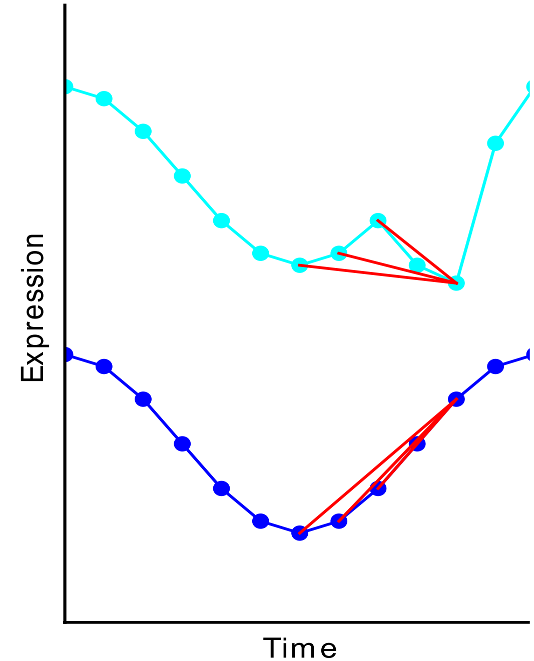
Fin.

Rhythm detection approaches

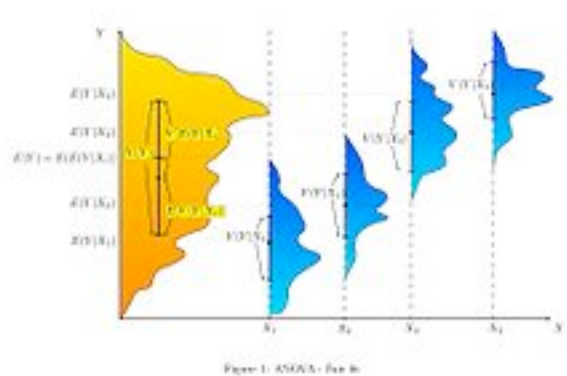
Cosine-fitting
- ARSER
- Fourier methods



Reference waveform matching
- JTK_CYCLE, **eJTK, Bootstrap eJTK**
- RAIN

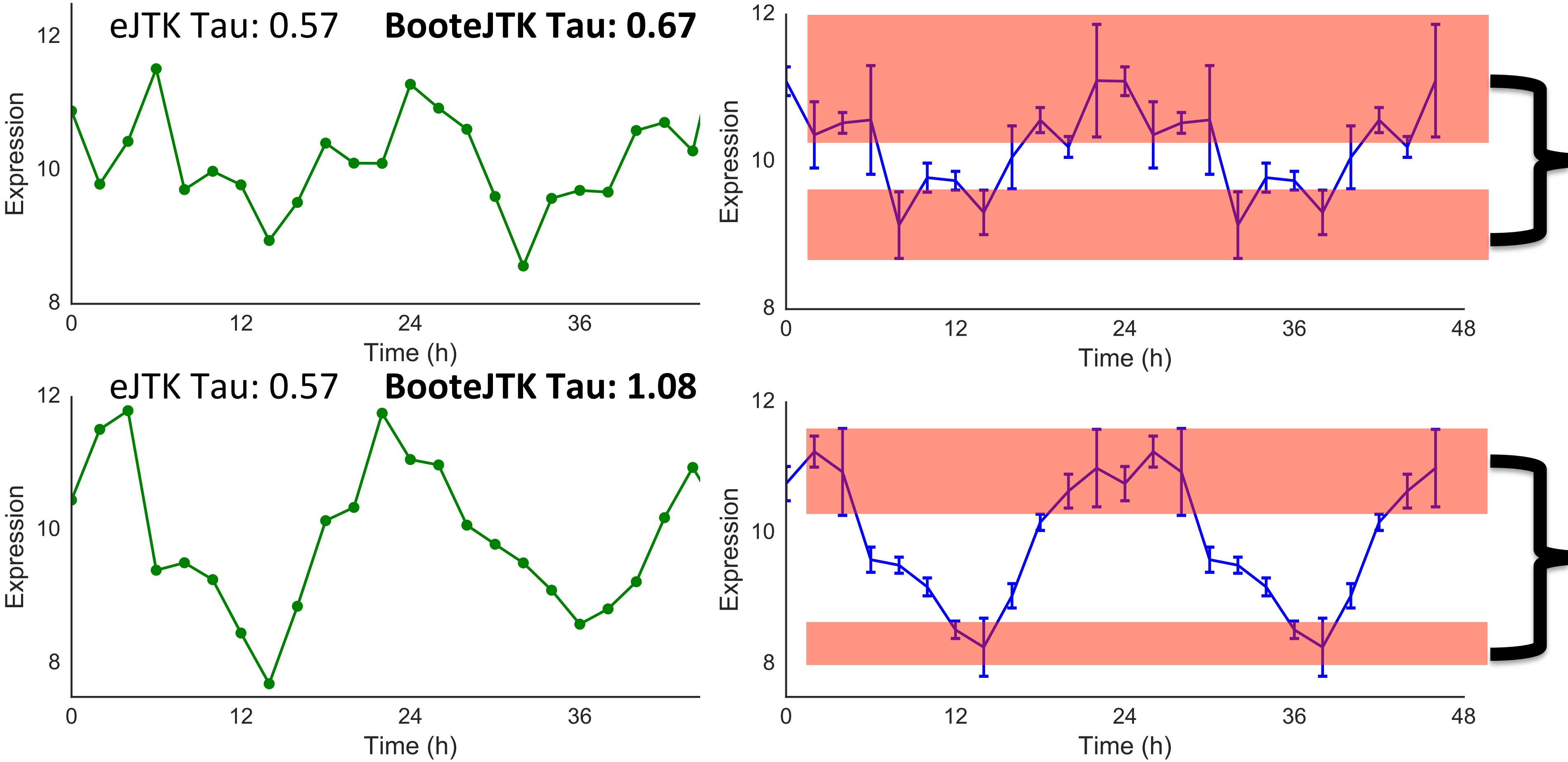


Reference-free methods
- ANOVA



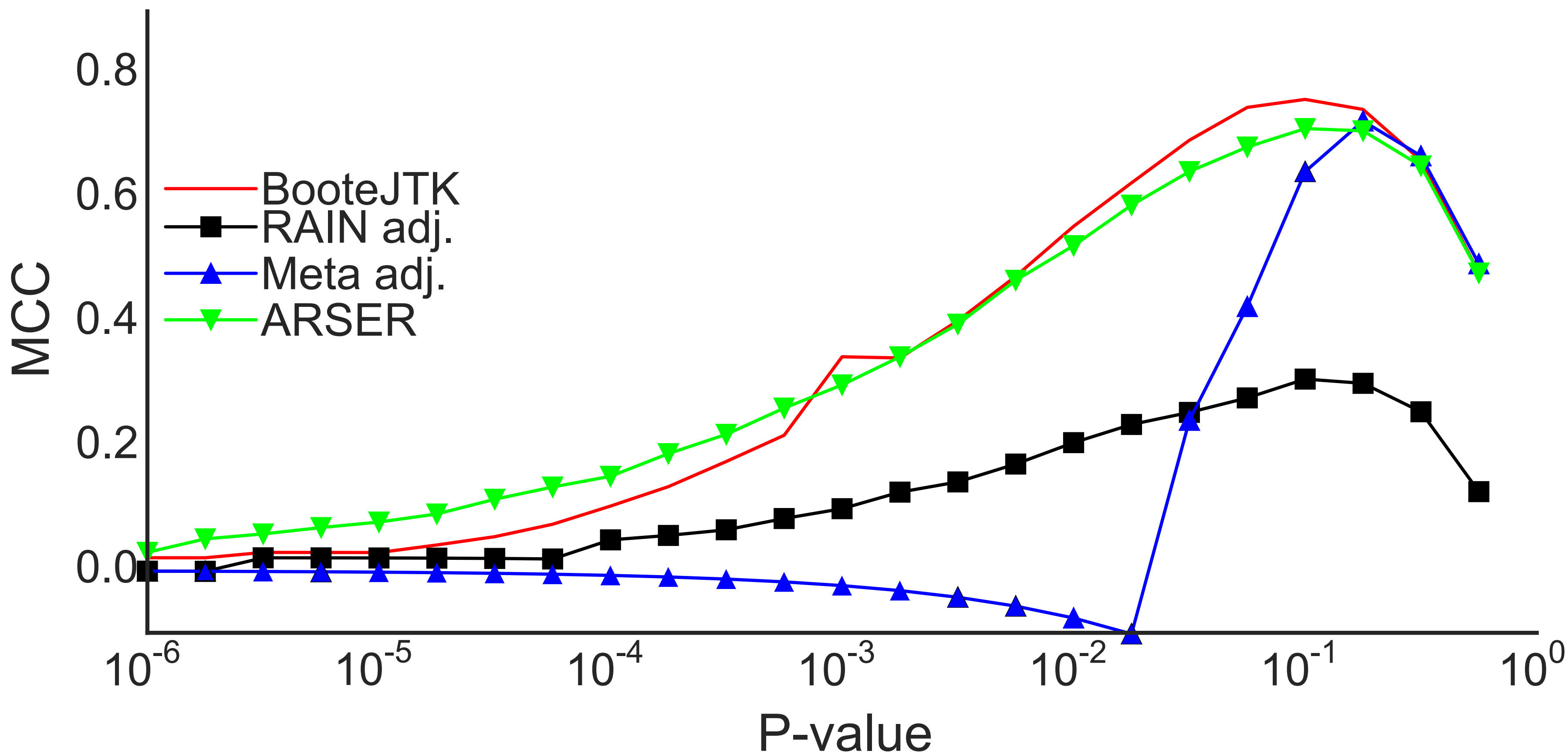
Combination methods
- MetaCycle (JTK_CYCLE, Lomb-Scargle, ARSER)

Non-parametric methods avoid arbitrary amplitude thresholds, but we want amplitude relative to measurement uncertainty



Hutchison *et al.* 2018

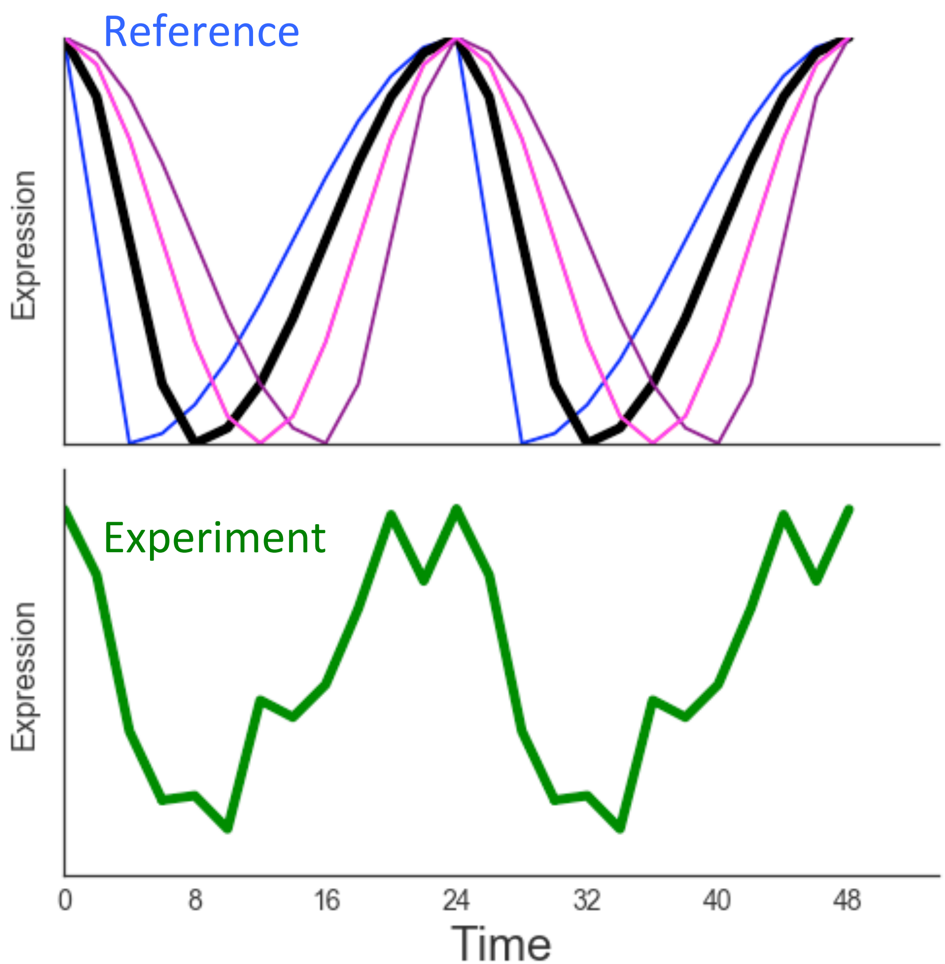
Bootstrap eJTK outperforms other rhythm detection methods



1100 time series, 11 asymmetries, cosine with Gaussian noise added to each point with noise-to-amplitude ratio of 1

Hutchison *et al.* 2018

Asymmetric waveforms improve rhythm detection



Period	Phase	Asym.	τ
24	10	4	0.6
24	**10**	**8**	**1.3**
24	10	12	1.0
24	10	16	0.9

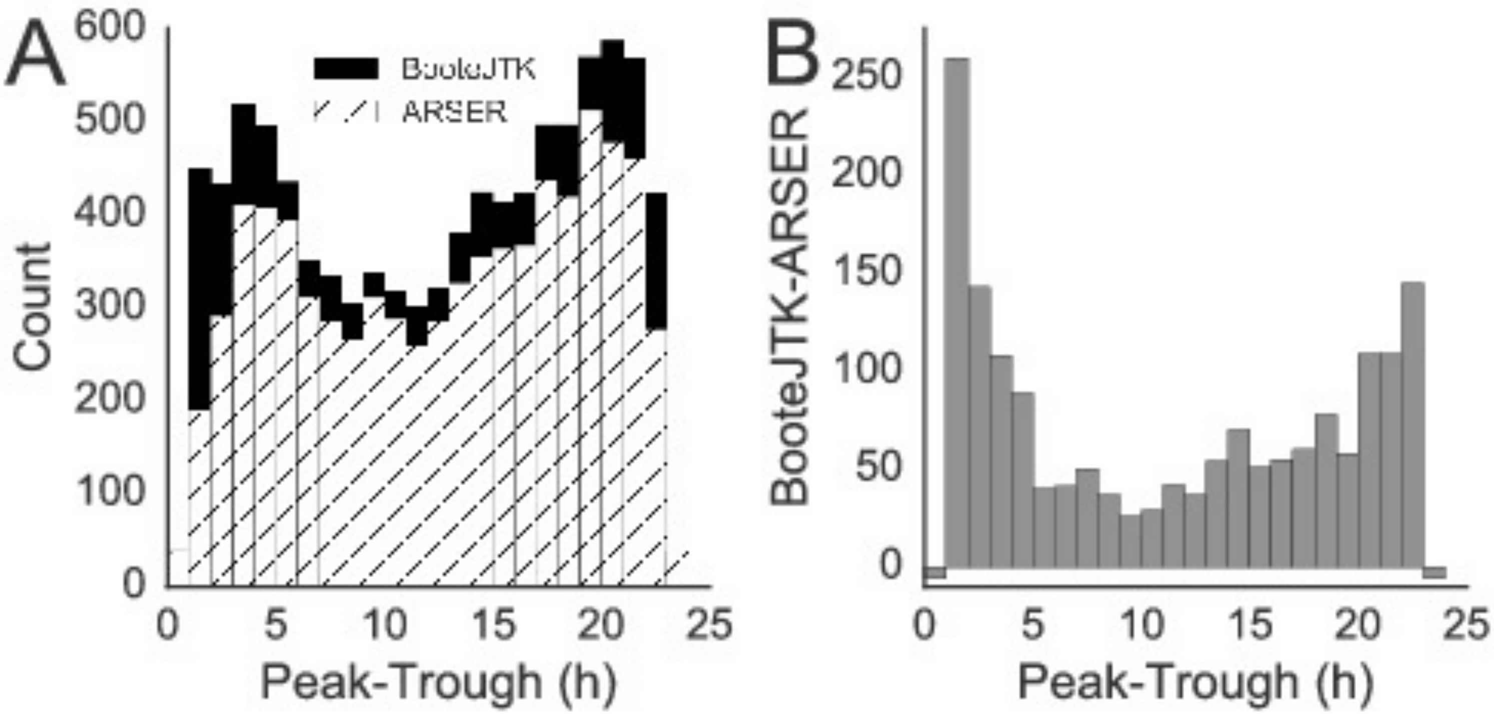
Data sampled every 2 h over 24 h:

12 possible phases

11 possible asymmetries

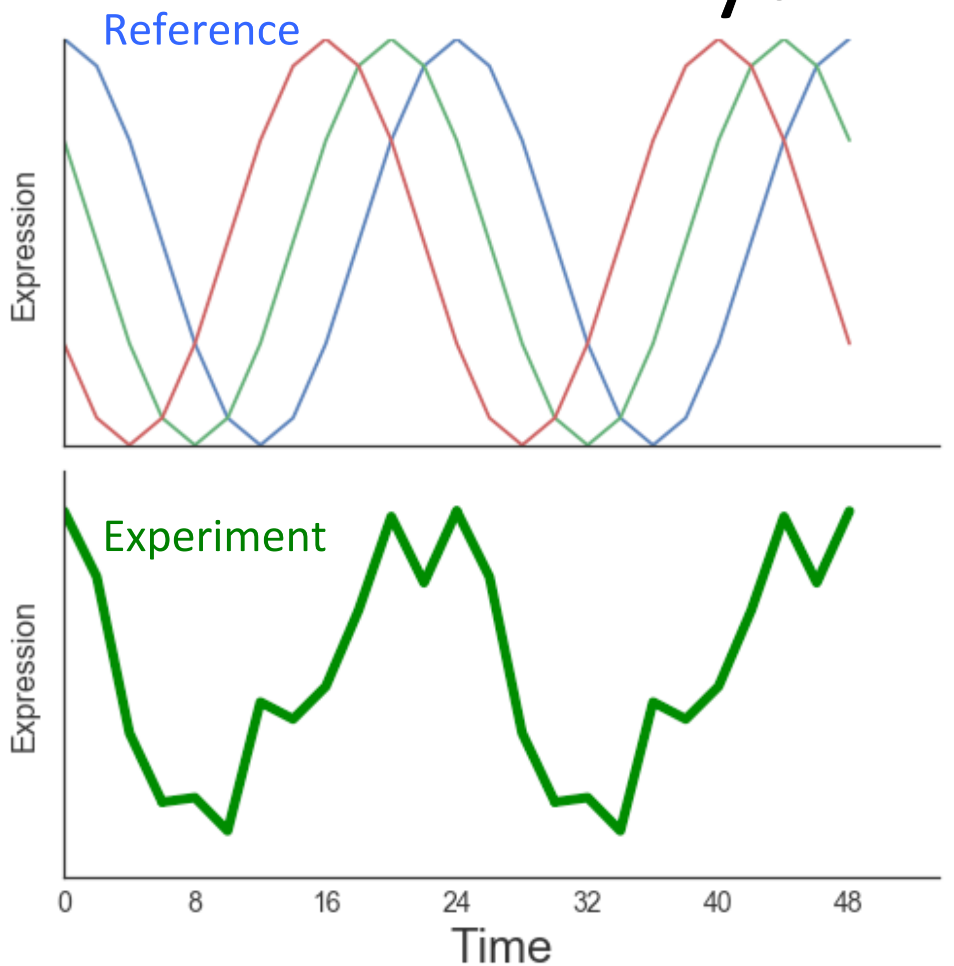
132 reference waveforms

Asymmetric waveforms improve rhythm detection



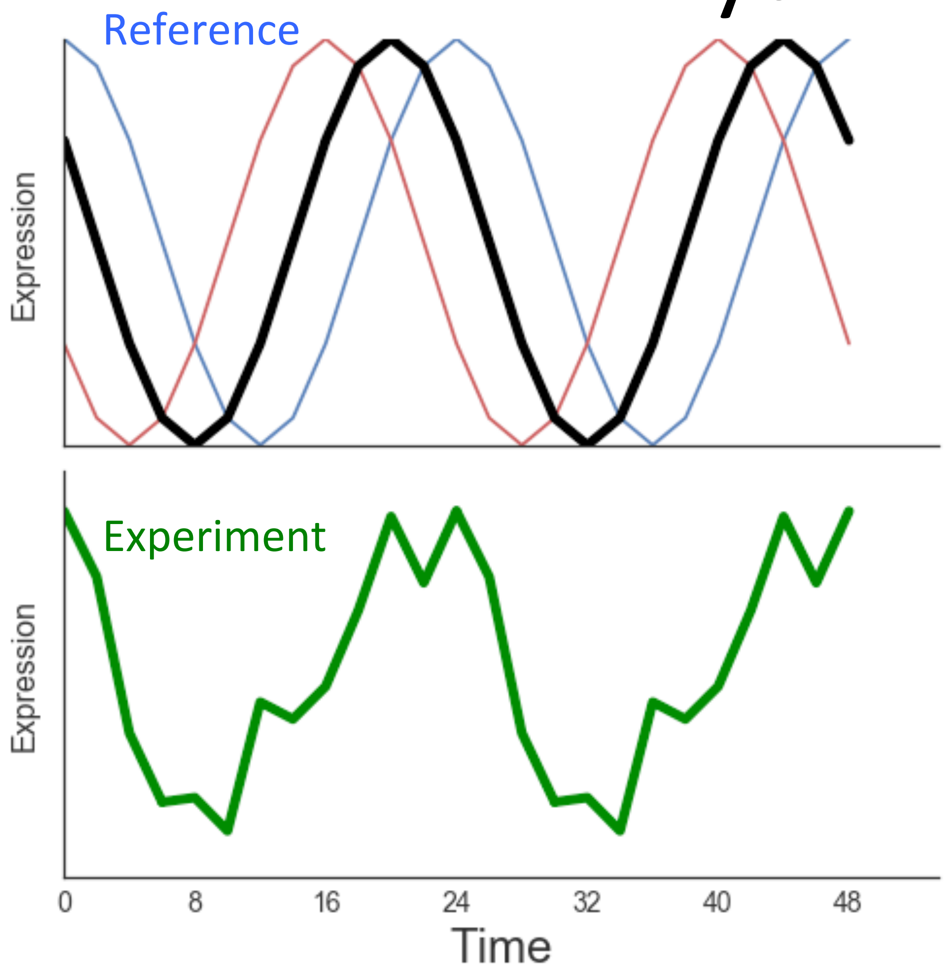
Hutchison *et al.* 2018

JTK_CYCLE picks the best reference waveform match as its measure of rhythmicity



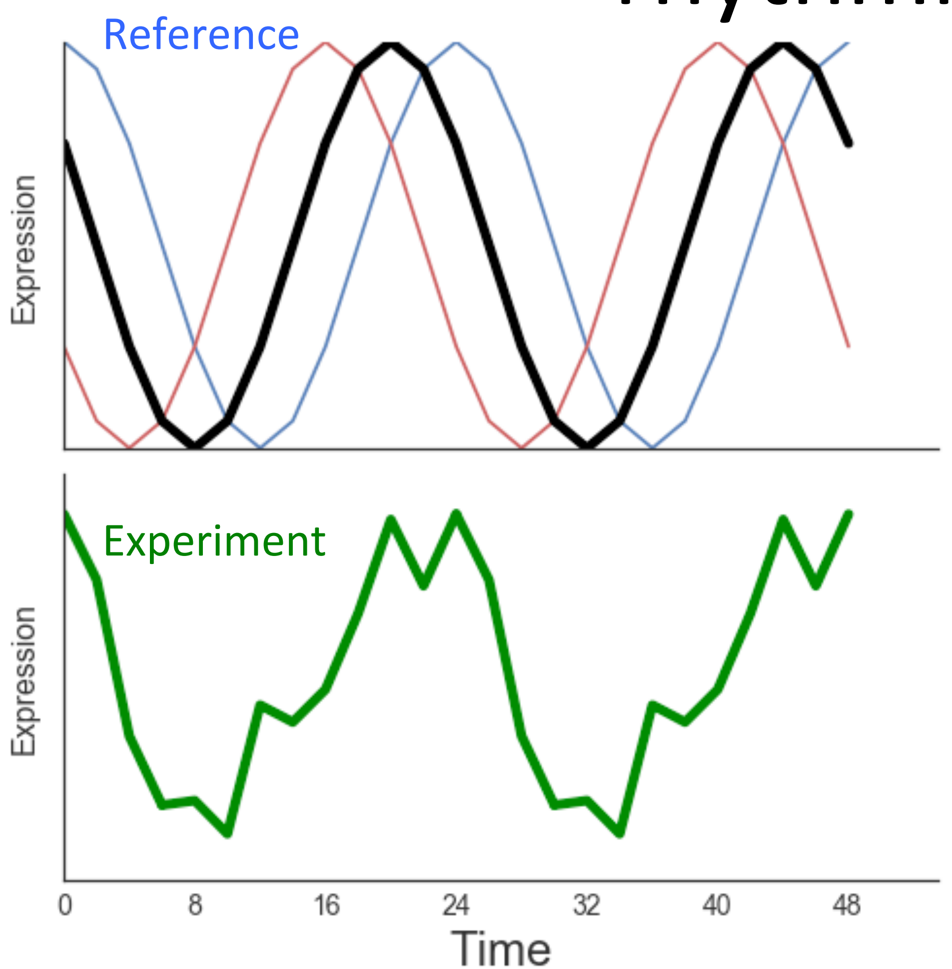
Period	Phase	τ
24	4	0.4
24	8	1.1
24	12	0.8

JTK_CYCLE picks the best reference waveform match as its measure of rhythmicity



Period	Phase	τ
24	4	0.4
24	**8**	**1.1**
24	12	0.8

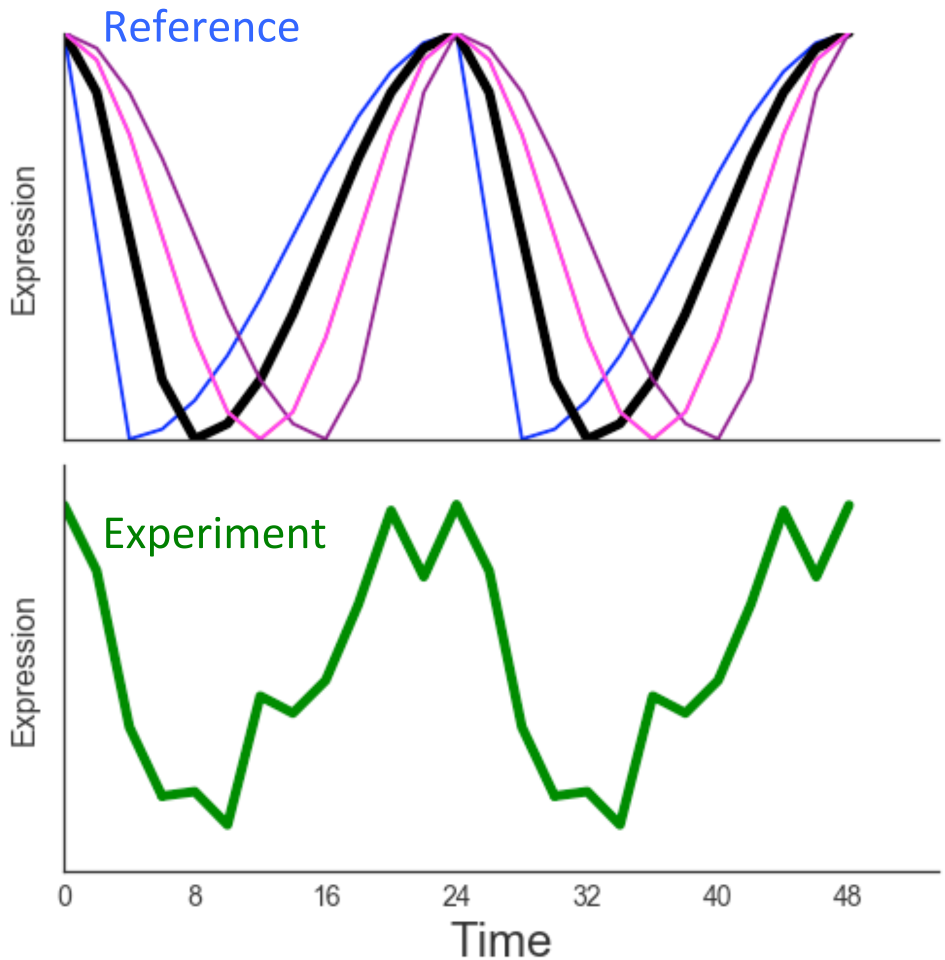
JTK_CYCLE picks the best reference waveform match as its measure of rhythmicity



Period	Phase	τ
24	4	0.4
24	**8**	**1.1**
24	12	0.8

Data sampled every 2 h over 24 h:
12 possible phases

Asymmetric waveforms improve rhythm detection

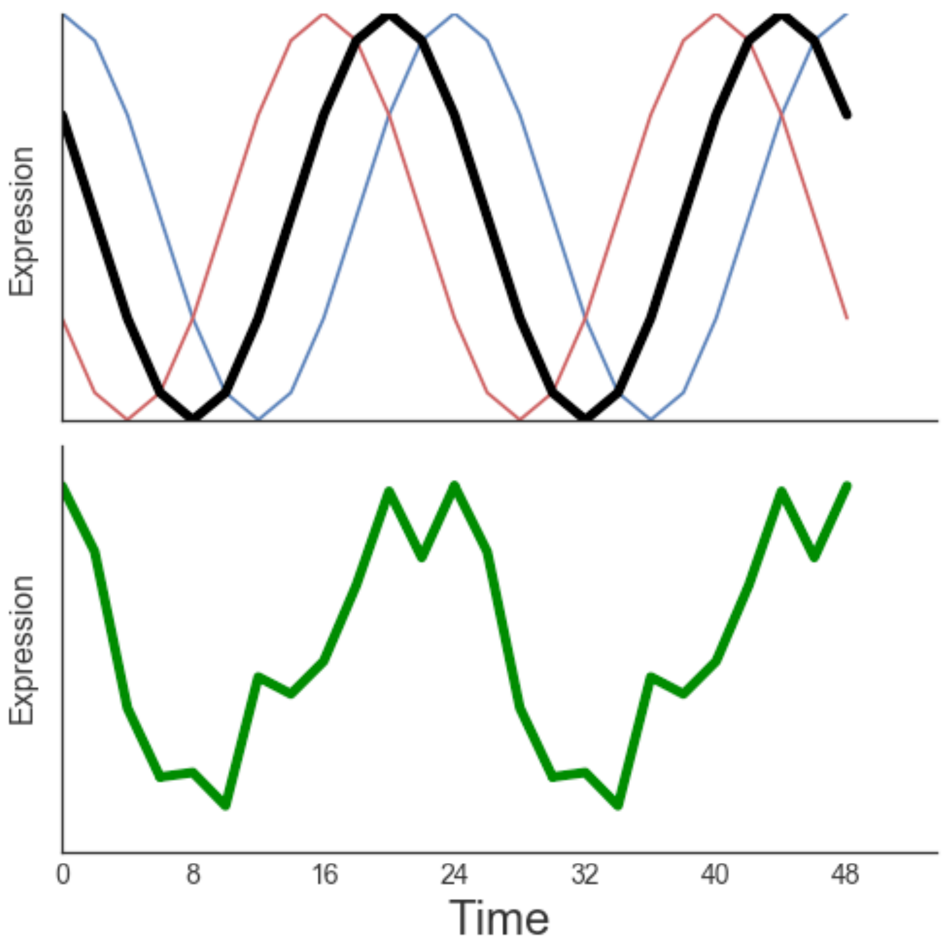


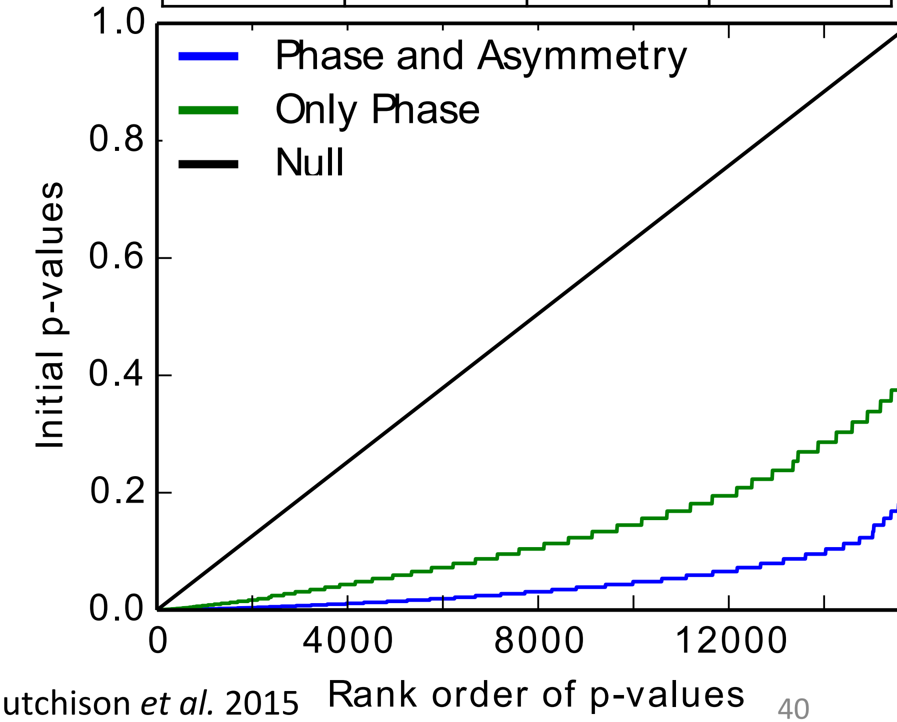
Period	Phase	Asym.	τ
24	10	4	0.6
24	**10**	**8**	**1.3**
24	10	12	1.0
24	10	16	0.9

Data sampled every 2 h over 24 h:
12 possible phases
11 possible asymmetries
132 reference waveforms

Picking the best Kendall Tau p-value underestimates the true p-value

Period	Phase	τ	p-value
24	4	0.4	0.3
24	**8**	**1.1**	**0.001**
24	12	0.8	0.02

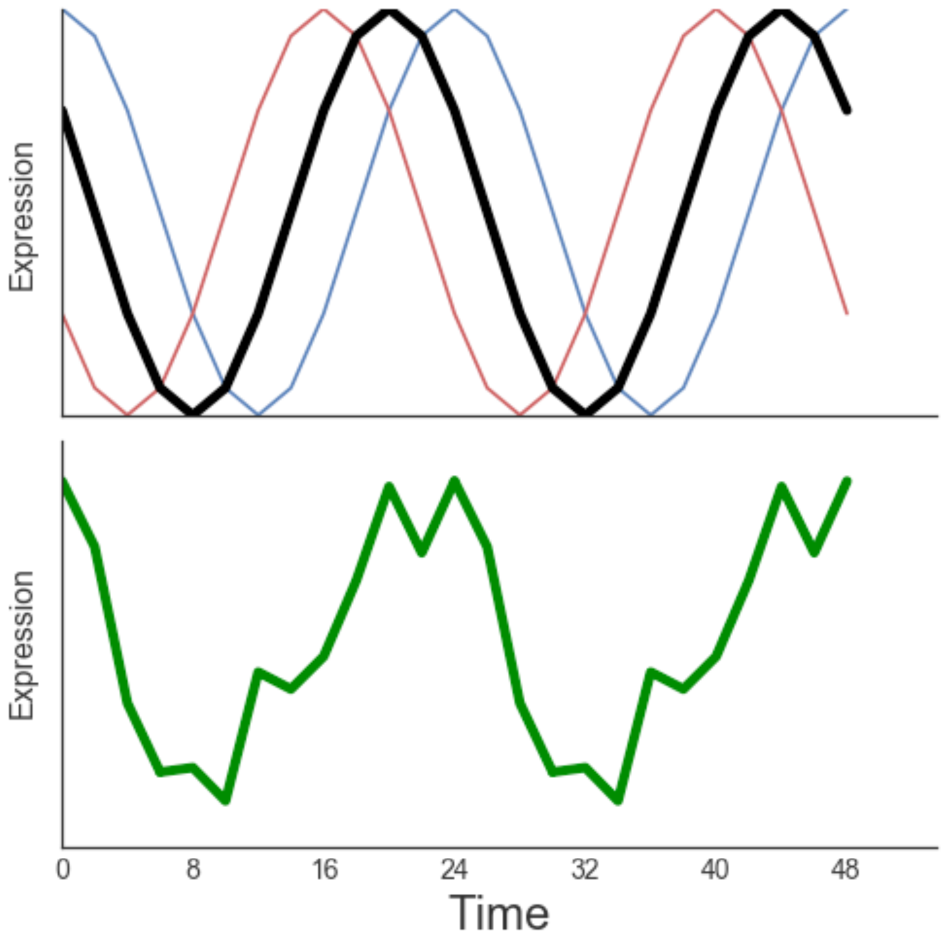


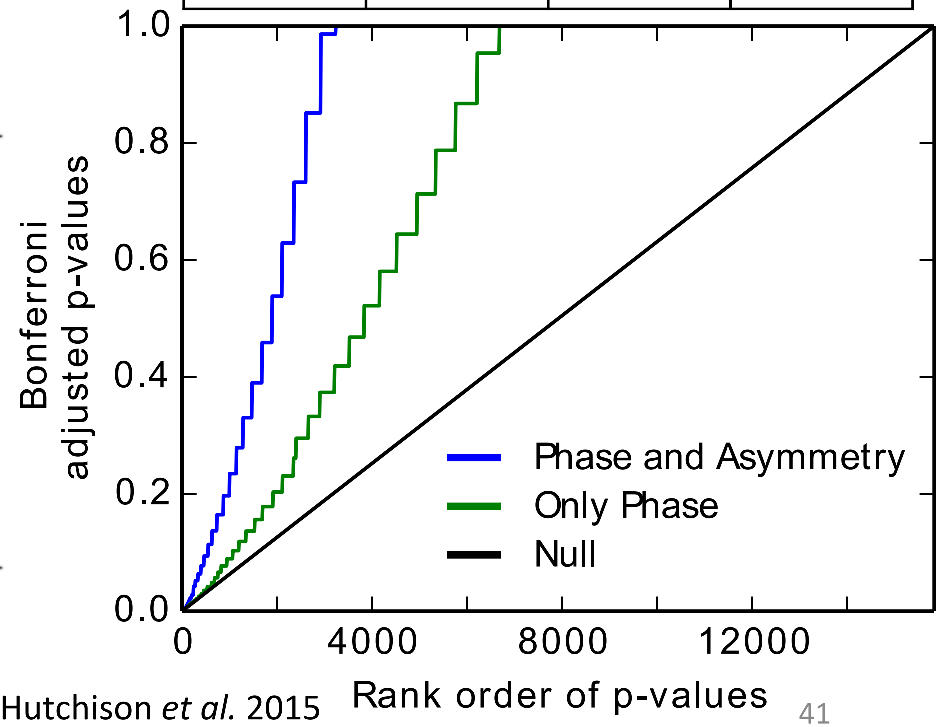


Hutchison *et al.* 2015

The Bonferroni correction results in overestimates of p-values

τ	p-value	# ref.	Bonf.
0.4	0.3	12	1
1.1	**0.001**	**12**	**0.012**
0.8	0.02	12	0.24

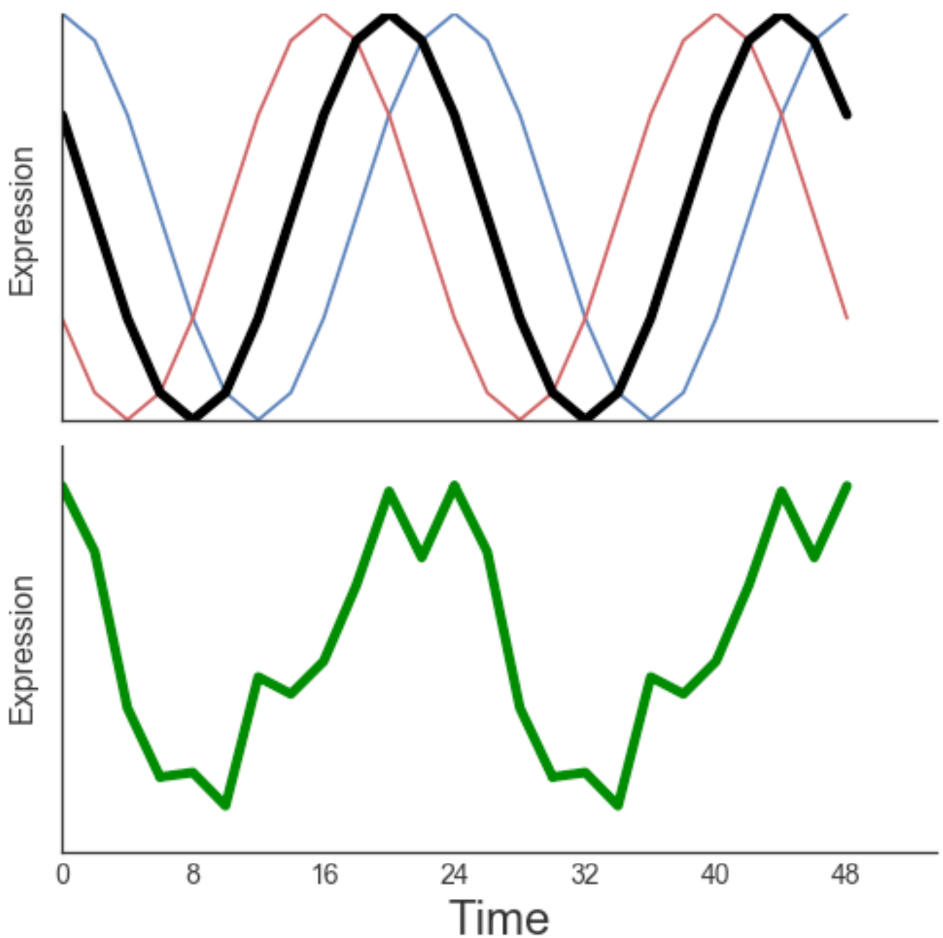


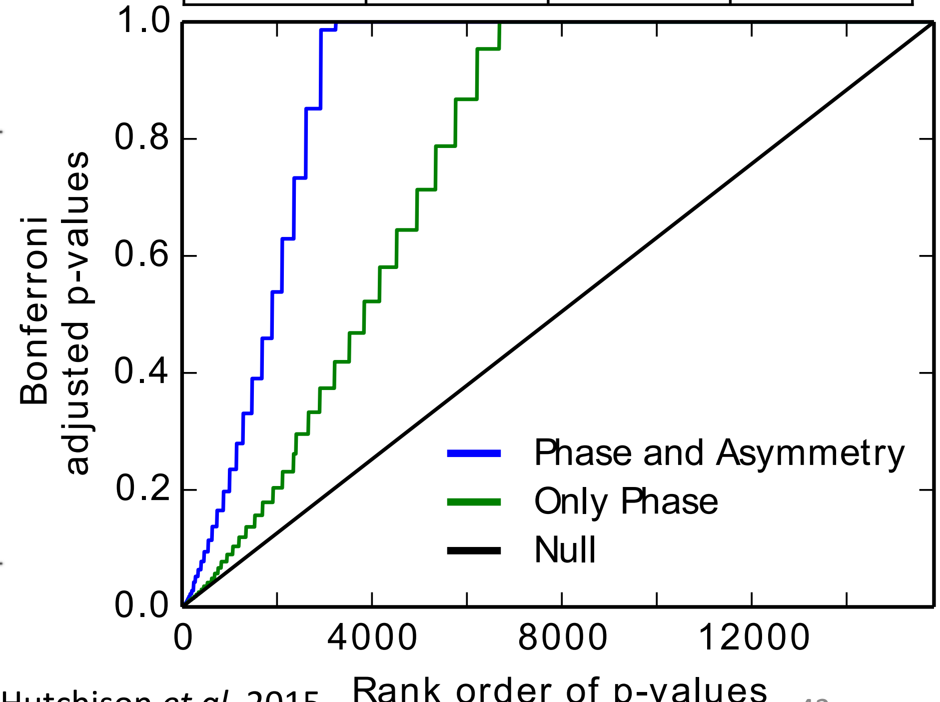


Hutchison *et al.* 2015

The Bonferroni correction results in overestimates of p-values

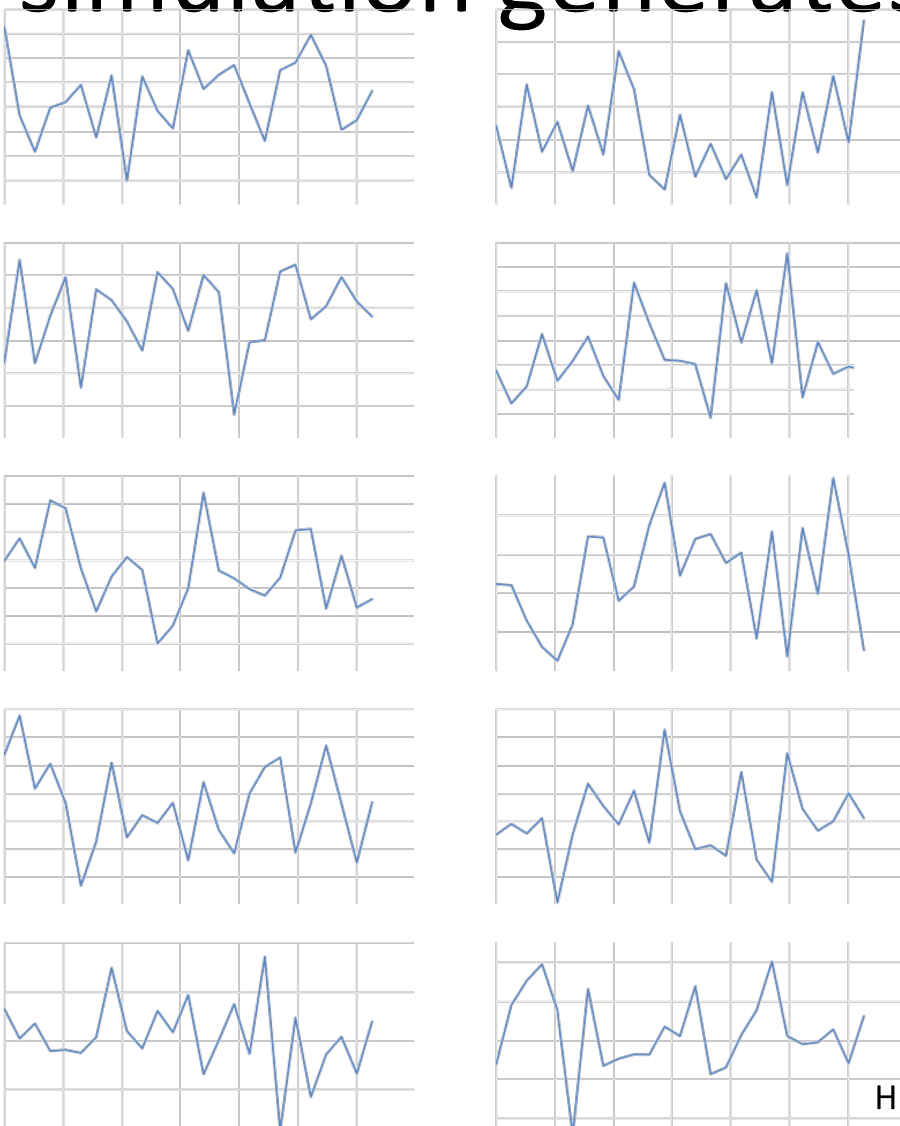
τ	p-value	# ref.	Bonf.
0.4	0.3	132	1
1.1	**0.001**	**132**	**0.132**
0.8	0.02	132	1





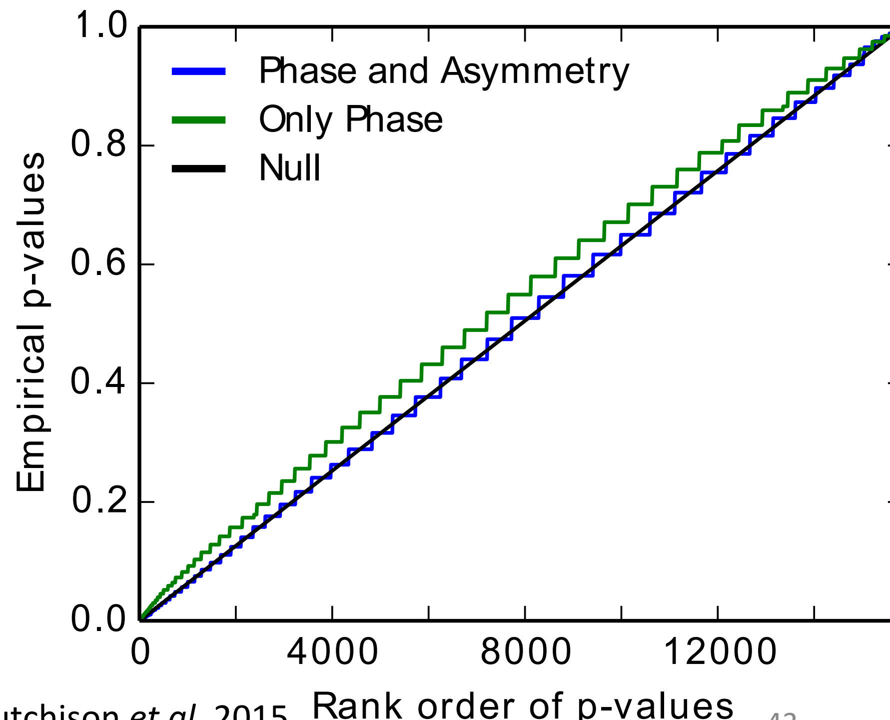
Hutchison *et al.* 2015

Empirically calculating the p-values via simulation generates accurate p-values



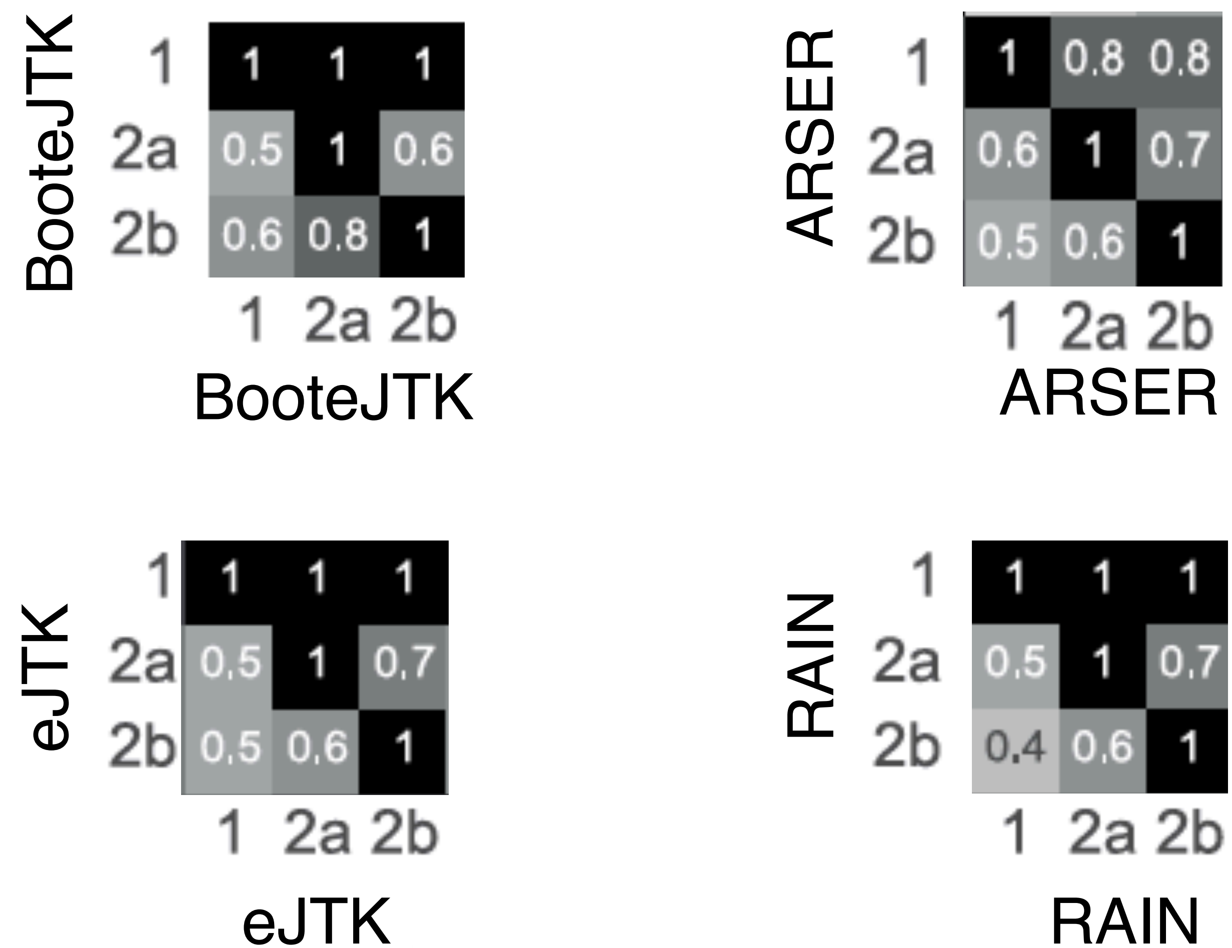
Simulate 1 million time series from noise to get empirical distribution of null p-values

Can fit distribution using fitting of 1000 time series



Hutchison *et al.* 2015

Bootstrap eJTK shows increased consistency compared to other methods



Hutchison *et al.* 2018