OXFORD

Gene expression

Differential rhythmicity: detecting altered rhythmicity in biological data

Paul F. Thaben and Pål O. Westermark*

Institute for Theoretical Biology, Charité – Universitätsmedizin Berlin, D-10115 Berlin, Germany

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Abstract

Motivation: Biological rhythms, such as rhythms in gene expression controlled by the cell cycle or the circadian clock, are important in cell physiology. A common type of experiment compares rhythmicity in tissues or cells either kept under different conditions or having different genotypes. Such investigations provide insights into underlying mechanisms as well as functions of rhythms.

Results: We present and benchmark a set of statistical and computational methods for this type of analysis, here termed differential rhythmicity analysis. The methods detect alterations in rhythm amplitude, phase and signal to noise ratio in one set of measurements compared to another. Using these methods, we compared circadian rhythms in liver mRNA expression in mice held under two different lighting conditions: constant darkness and light-dark cycles, respectively. This analysis revealed widespread and reproducible amplitude increases in mice kept in light-dark cycles. Further analysis of the subset of differentially rhythmic transcripts implied the immune system in mediating ambient light-dark cycles to rhythmic transcriptional activities. The methods are suitable for genome- or proteome-wide studies, and provide rigorous P values against well-defined null hypotheses.

Availability and Implementation: The methods were implemented as the accompanying R software package DODR, available on CRAN.

Contact: pal-olof.westermark@charite.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biological rhythms in metabolites or mRNA and protein expression include circadian (daily) rhythms, cell cycle dependent rhythms and various ultradian rhythms, the latter with periods of at most a few hours (Ferrell *et al.*, 2011; Hazlerigg and Wagner, 2006; Lin *et al.*, 2015; Westermark and Lansner, 2003). Our research focus is on circadian rhythms, which have been extensively studied at the whole transcriptome level especially in mouse liver and *Drosophila*. In mouse liver, thousands of transcripts exhibit rhythmic abundances, which lead to rhythmic variations in numerous activities and aspects of cell physiology (Lück and Westermark, 2016).

When investigating mRNA or protein expression on a genome-wide scale, rhythmic or not, a task is often to compare samples harvested from subjects kept under different conditions, undergoing

different treatments, or having different genotypes (all these hereafter termed groups). From each sample, thousands of transcript or protein abundances may be estimated using technologies such as microarray expression profiling, RNA sequencing, or mass spectrometry. On the data analysis side, this routinely leads to the well-known differential expression problem (Anders and Huber, 2010; Ideker *et al.*, 2000). In a basic differential expression problem, the researcher compares transcriptome-, metabolome- or proteome-wide expression profiles or abundances of two groups, each consisting of a few biological replicates. The objective is to identify metabolites, transcripts or proteins whose concentrations or expression levels are significantly different between the two groups.

For studies on biological rhythms on a transcriptome-, metabolome- or proteome-wide scale, samples are typically harvested from

animals at different time points over a time span covering one or more period lengths. When studying circadian rhythms, this entails, for instance, harvesting samples every 2, 3 or 4 h over a time span of 24 or 48 h. Samples are often taken from animals either synchronized to 12 h light/12 h dark cycles, or they may be taken from animals kept in constant darkness. The former condition allows detection of rhythms that are generated by the cellular core circadian clock (a small network of transcription factors influencing their own expression present in all cells), by the central circadian system acting upon the tissue (in mammals the central system consists of the suprachiasmatic nuclei in the hypothalamus), or directly by the light cues via various mechanisms (Husse *et al.*, 2015). The latter condition allows the isolation of rhythms generated by the intrinsic circadian clock, either the cellular clock or the central circadian system.

In this way, thousands of time series of features are obtained. These time series have to be analyzed statistically to isolate features whose levels vary in a circadian fashion. Suitable statistical tests, such as parametric harmonic regression method (Cornelissen, 2014), spectral based methods (Wijnen *et al.*, 2005), or nonparametric methods such as JTK_CYCLE (Hughes *et al.*, 2010) or RAIN (Thaben and Westermark, 2014), are employed for this purpose. Many other methods exist, as reviewed previously (Doherty and Kay, 2010). These tests result in a *P* value for each feature (e.g. transcript) for the null hypothesis that its expression level does not vary in a rhythmic fashion. *P* values are then corrected for multiple testing, and the investigator may then isolate a set of rhythmic transcripts of interest with a specified false discovery rate (FDR). Parametric methods are able to detect weaker rhythms (they are more powerful) than non-parametric methods. However, parametric methods rely on stricter assumptions: often, measurement errors must have Gaussian distributions, and the methods are sensible to outliers. Non-parametric methods sacrifice some degree of power in order to gain robustness against outliers, and they make fewer assumptions about measurement errors (Hettmansperger and McKean, 2011).

Often, transcriptome- or metabolome-wide circadian rhythm studies compare two groups, which leads to the subject matter of the present article. One example is comparison of wild type animals to mutant animals with various deletions of core clock transcription factor genes (Hughes *et al.*, 2012; Kornmann *et al.*, 2007; Masri *et al.*, 2014; McCarthy *et al.*, 2007). Such studies may allow isolating the contributions to the circadian transcriptome from various sources. Another example is comparative studies of animals subjected to various nutritional challenges (Adamovich *et al.*, 2014; Eckel-Mahan *et al.*, 2013; Vollmers *et al.*, 2009). These studies all have a certain analytical task in common: to detect changes in rhythmicity in many metabolites or transcripts. Here, we will refer to this task as the differential rhythmicity problem.

To detect e.g. differentially rhythmic transcripts, researchers have the typically isolated rhythmic transcripts in both groups, using *P* value or FDR cutoffs against the null hypothesis of no rhythmicity in each group. Transcripts with *P* values falling under the cutoff in one group and over the cutoff in the other group are then considered differentially rhythmic (Hughes *et al.*, 2012; Masri *et al.*, 2014). In principle, the two cutoffs may also be chosen differently for each group, in order for the procedure to be more conservative. For example, in order to detect transcripts whose rhythms have been lost in mutant animals, one may require a *P* value smaller than 0.001 for wild type animals, and a *P* value greater than 0.3 in mutant animals. Additional criteria might be employed too, such as fold changes between peaks and troughs. Another approach to the differential rhythmicity problem was taken by investigators researching core

clock transcripts in two different rat strains (Miyazaki *et al.*, 2011). Amplitudes were estimated with 95% confidence intervals for each transcript and strain using a non-parametric method (Hughes *et al.*, 2010; Miyazaki *et al.*, 2011). Transcripts for which the confidence intervals did not overlap were classified as differentially rhythmic. Unfortunately, FDRs or *P* values for these approaches against reasonable null hypotheses (i.e. no change in rhythmicity between the groups) have not been presented or computed routinely, since a comprehensive framework for this task has been lacking in the field. Changes of phase (peak time) may also be of interest, as was the case for a study on the effects of high fat diet on the circadian transcriptome and metabolome (Eckel-Mahan *et al.*, 2013), although a rigorous method to estimate and assign *P* values to phase shifts is lacking.

Here, we report a unified framework for the estimation and detection of changes in rhythmicity, both amplitudes and phases. We present both parametric methods for the task, as well as rank based methods that are robust to outliers and non-Gaussian variations and measurement error. These methods were implemented as an accompanying R package called DODR. The procedures were thoroughly benchmarked, and compared to *P* value cutoff methods. We also report an application of DODR to mRNA expression profiles of livers of mice kept in 12 h light/12 h dark conditions, and in constant darkness, respectively. By doing this, we were able to quantify the light-governed circadian transcriptome, and to implicate the transcription factor IRF7 in relaying light cues to the liver.

2 Methods

2.1 Simulated time series
Time series were generated by estimates based on mouse liver mRNA expression data (Supplementary methods and Fig. S1).

2.2 HANOVA
Summed squared *e* residuals (*SS*) for the null hypothesis (SS_0), and for the alternative hypothesis (SS_1), respectively (Equation 2), are combined to an *F* statistic:

$$\frac{(SS_0 - SS_1)/2}{SS_1/(n-5)} \sim F_{2,n-5}.$$

Here, the number 5 is the number of parameters for the alternative model, and the number 2 is the difference between the parameter numbers of the alternative and the null hypotheses.

2.3 Robust differential rhythmicity
The problem formulation (Equation 2) remains the same for this case, but rather than *SS*, a rank-based dispersion measure *D* is minimized in order to yield a fit of the models to data:

$$D(e_g) = \sqrt{\frac{1}{12}} \frac{1}{n+1} \sum_{i=1}^{n} \left(R(e_{gi}) - \frac{n+1}{2} \right) e_{gi}.$$

Here, *R* denotes the rank. Let D_0 and D_1 be the dispersions obtained for fitting the null hypothesis and the alternative hypothesis models as above. Then,

$$\frac{D_0 - D_1}{\hat{\tau}} \sim F_{2,n-5},$$

where $\hat{\tau}$ is an estimation of the scale parameter defined as $\tau = \sqrt{\frac{1}{12} \left(\int_{-\infty}^{\infty} f^2(t)dt \right)^{-1}}$. Here, *f(t)* is the probability density function of the of the error terms e_{gi}. The estimated value $\hat{\tau}$ is based on the empirical distribution function of the error terms (Koul *et al.*, 1987).

The R package Rfit (Kloke and McKean, 2014) was leveraged to adapt these robust methods to differential rhythmicity analysis.

2.4 Scale tests

For the standard non-robust linear least squares approach, summed squared residuals for the separate model fits to the two groups (i.e. the HANOVA alternative hypothesis model) may be combined and obey an *F* distribution when errors are independently identically distributed with a Gaussian distribution:

$$\frac{\sum_{i=1}^{n_1} e_{1i}^2 / (n_1 - 3)}{\sum_{i=1}^{n_2} e_{2i}^2 / (n_2 - 3)} \sim F_{n_1-3, n_2-3}.$$

As a robust alternative, we used the Fligner–Killeen test, which is a nonparametric test for difference in scale between the residuals of the two groups. Formally, if residuals of group 1 have the distribution function $F(x)$, then the residuals of group 2 are assumed to have distribution function $F((y - \Delta)/\eta)$, and the test is against the null hypothesis that $\eta = 1$. We leveraged the R package npsm (Kloke and McKean, 2014) to adapt the Fligner–Killeen test to the differential rhythmicity problem.

2.5 mRNA expression data for differential expression analysis

Mouse liver mRNA expression data for LD and DD conditions were gathered and processed as described in the Supplementary methods section.

2.6 mRNA differential rhythmicity analysis

In order to avoid counting the same sequencing reads, probe sets and transcription start sites (TSSs, below) more than once, a subset of 14 253 non-overlapping RefSeq transcripts were used for the analysis, where the shortest transcript associated with a given TSS was retained. 7171 transcripts with a MAS5 present call for at least half of the time points in both datasets (Hughes *et al.*, 2009; Jouffe *et al.*, 2013) were considered to be expressed and were used for further analysis. For the study performed under constant darkness (dark-dark, DD) conditions (Hughes *et al.*, 2009), only every second time point was used (starting with the first sample at CT 18) in order to match the sampling resolution of the study performed under light-dark (LD) conditions (Jouffe *et al.*, 2013). For the validation data, transcripts with Affymetrix expression level >64 in the DD dataset (Zhang *et al.*, 2014) were considered further. *P* values for circadian rhythms were computed using the R package RAIN (Thaben and Westermark, 2014). For the main LD/DD datasets (Hughes *et al.*, 2009; Jouffe *et al.*, 2013), a meta-analysis of *P* values was performed using the minP method (Wang *et al.*, 2012), in order to combine the *P* values of the two studies: the resulting combined *P* values represents for transcript the alternative hypothesis that at least one of the DD or LD time series exhibits oscillations. These *P* values were Benjamini–Hochberg adjusted for multiple testing. Amplitudes and phases were estimated using the R package Harmonic Regression (Lück *et al.*, 2014). The robust DODR method was used on the first LD/DD dataset pair (Hughes *et al.*, 2009; Jouffe *et al.*, 2013). An adjusted *P* value cutoff of 0.05 and relative amplitude threshold of 0.15 (Lück and Westermark, 2016) was used to isolate transcripts exhibiting a biologically significant circadian rhythm in at least one of the LD or DD conditions. These transcripts were considered for further differential rhythmicity analysis: for this, their

robust DODR *P* values were Benjamini–Hochberg adjusted. Phase differences between LD and DD conditions are only meaningful when there are well-defined rhythms in both conditions. Therefore, a subset of 69 differentially rhythmic transcripts with relative amplitudes >0.15 in both conditions were used for phase comparisons. These were classified as having pure phase shifts if the absolute log2 fold amplitude change was smaller than 0.1. Amplitude gains were declared for differentially rhythmic transcripts with log2 fold amplitude change >0.5, amplitude differences were declared for differentially rhythmic transcripts with log2 fold amplitude change either >0.5 or < −0.5. For comparisons with the validation dataset, a meta-analysis of the two LD datasets (Jouffe *et al.*, 2013; Menet *et al.*, 2012) with the maxP method (Wang *et al.*, 2012) was performed using the RAIN *P* values. Resulting *P* values represent the alternative hypothesis that both LD time series exhibit oscillations, these *P* values were Benjamini–Hochberg adjusted and a *P* value cutoff of 0.05 defined transcripts with circadian rhythms in both LD datasets. As before, phase comparisons were made only for transcripts exhibiting relative amplitudes >0.15 in both conditions examined.

2.7 GO analysis

The DAVID resource (Huang *et al.*, 2009) was used to compute enriched GO biological process terms for differentially rhythmic transcripts compared to circadian transcripts; the resulting functional annotation chart is supplied as Supplementary Table S1.

2.8 TF binding sites predictions

SwissRegulon (Pachkov *et al.*, 2007) TF binding site predictions were considered for promoter regions ±300 bp of RefSeq TSSs.

3 Results

3.1 The limits of the threshold based method

To set the stage, we considered the simple *P* value threshold (PVT) method used previously in the circadian rhythms field (Hughes *et al.*, 2012; Masri *et al.*, 2014). The setting involves time resolved measurements of two groups, e.g. mRNA expression levels in wild type and mutant animals, respectively. The PVT method then entails calling differential rhythmicity when the *P* value for a rhythmicity detection test falls under a threshold for one group and above a threshold for the other group (loss of rhythmicity) or vice versa (gain of rhythmicity). We aimed primarily to quantify the FDRs of PVT approaches, which to the best of our knowledge has not previously been done.

To do this, we estimated circadian amplitudes and random variability of mouse liver mRNA expression, using a published dataset (Hughes *et al.*, 2012; Section 2). The amplitude measure used in the present work is relative amplitude: amplitude in absolute units divided by mean. Then, we simulated millions of time series drawing from the estimated amplitude distribution, and obtained *P* values for each time series using harmonic regression (Section 2). Simulations were performed for experimental settings with 6, 12, 16 and 24 time points sampled over two oscillation periods with constant sampling interval, designs common in the circadian rhythms field. We then asked two questions regarding FDRs. First (Q1): what fraction of the simulated time series with a true amplitude above a given amplitude threshold yielded *P* values below a given *P* value threshold? Second (Q2): what fraction of the simulated time series with a true amplitude below a given amplitude threshold yielded *P* values above a given *P* value threshold? Both these questions are necessary for the PVT method: the task is to detect time

series that oscillate in one group (Q1) and to not oscillate in the second group (Q2): for both these questions, the FDRs need to be controlled. The amplitude cutoff might be set collectively to e.g. 0 or 0.1. It might also be different for the two groups, but then the amplitude cutoff used to detect rhythmic time series should be larger than the cutoff used to detect non-rhythmic time series. If not, results are ambiguous: for instance, a feature might be classified as rhythmic with amplitude greater than 0.05 in group 1, and non-rhythmic with amplitude less than 0.15 in group 2, both with $P < 0.001$. We note at this point that the FDRs obtained for groups 1 and 2 should be combined to yield a final P value for differential rhythmicity. Such a combination could amount to simply retaining the largest of the P values, which is a so-called intersection-union procedure (Berger and Hsu, 1996).

The results of this computational experiment showed that as the amplitude cutoff grows larger, the FDRs for Q1 deteriorate quite substantially. For example in the situation with 16 samples, using a cutoff $P < 10^{-5}$ resulted in an FDR of ~ 0.01 for the amplitude threshold 0.1 (Fig. 1, blue dotted curve). The correct interpretation of these numbers is that 7% of the time series with $P < 10^{-5}$ have an amplitude that is less than 0.1. Such an FDR for a single feature is obviously far too high for genome scale experiments because of the multiple testing correction needed for such settings, which often entails a factor increase of P values of 100–1000. For an amplitude cutoff of 0 (blue solid curve), the Q1 FDR and P value cutoff are identical by definition, and the $P < 10^{-5}$ cutoff may be reasonable for genome wide studies. For an amplitude cutoff of 0.2 on the other hand, the P value cutoff has to be on the order of 10^{-10} to achieve an FDR of ~ 0.0001. Here, one should note that the amplitude pertains to the signal only: the noise contains (typically) components of myriad periods including a comparatively insignificant 24 h component.

However, a disturbing result was that the inverse problem Q2, that of detecting non-rhythmic time series, is not feasible for small amplitude cutoffs. This is due to the fact that 85% of the amplitudes estimated from the mouse liver dataset were larger than 0, (although most of them very small), raising the P value cutoff to almost 1 for Q2 still results in a large FDR because of spillover from the large
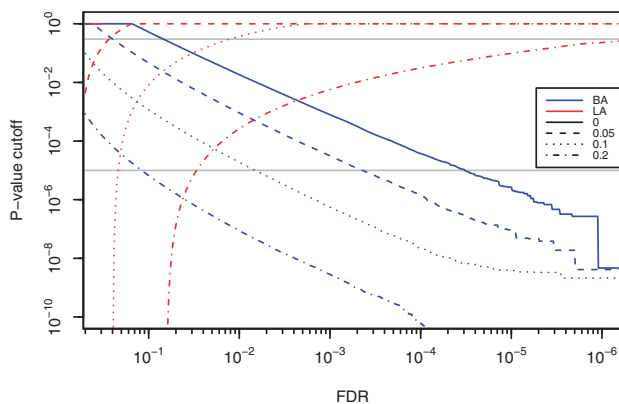
population with amplitudes just slightly larger than 0. This suggests that a substantial amplitude cutoff is necessary for PVT methods in general. In fact, to achieve FDRs of around 0.0001 for Q2 with 16 samples, the amplitude cutoff needs to be set to 0.2 and a threshold $P > 0.1$ (Fig. 1, red dashed-dotted curve). In general, the blue curves in Figure 1 get more useful as they are pushed to the upper right, while the red curves get more useful as they are pushed to the lower right. Hence, there is a tradeoff: as the amplitude cutoff increases, the detection of time series with rhythms below the cutoff becomes easier, but detection of time series above the cutoff becomes harder.

For other numbers of samples the situation was, naturally, different; the same analysis is presented for 6, 12 and 24 samples in Supplemental Figure S2. In particular, for an experimental design with only 6 samples, amplitude cutoffs need to be well over 0.1 for this method to be useful even when investigating only a single feature.

The main lesson to be learned is that a single P value cutoff generally works very poorly, in particular due to false positives when detecting non-rhythmic or sub-amplitude threshold time series. In practice, the investigator needs to apply much larger P value thresholds for Q2 than for Q1. Furthermore, for experimental settings with hundreds or thousands of features, the analysis showed that PVT methods work in practice only if interpreted with amplitude cutoffs around 0.1–0.2. This is due to the fact that unadjusted P values typically need to lie below 0.001 since multiple testing corrections have to be performed. When investigating a single feature (e.g. rhythmic abundance of one protein) the threshold based method may work, but even then, amplitude cutoffs need to be high if the number of samples is small. The researcher has to interpret results in terms of features changing amplitude across a well-defined amplitude threshold, rather than switching between being fundamentally rhythmic and non-rhythmic, and that simulation studies such as the one presented here need to be performed.

3.2 Differential rhythmicity: general methods
The initial analysis of PVT methods shows three serious shortcomings of such approaches. First, PVT methods must rely upon simulations to yield rigorous P values, which is costly and require estimations of amplitude distributions from prior data. Second, the PVT approach puts a serious cap on the detection power for experimental designs with 12 or fewer samples. Third, the approach detects only amplitude changes over a pre-defined amplitude threshold. Neither arbitrary amplitude changes, such as an amplitude decrease from very high to moderate, nor phase changes are detected. Thus, we investigated alternatives.

We started with a simple description of rhythms in terms of their fundamental harmonic (first Fourier coefficient). We divide each time series with its mean value to obtain scaled measurements. If y_{gi} is the scaled measurement at time point t_i, we have

$$y_{gi} = 1 + a_g\cos(\omega t_i) + b_g\sin(\omega t_i) + e_{gi}, \qquad (1)$$

where e_{gi} is the differences between this model and scaled measured readouts, and g symbolizes possible subdivision in groups. This is equivalent to $y_{gi} = 1 + A_g\cos\left(\omega t_i - \varphi_g\right)$, so that the relative amplitude is $A_g = \sqrt{a_g^2 + b_g^2}$ and the phase satisfies $\tan\left(\varphi_g\right) = b_g/a_g$. We have two groups (indexed 1 and 2) that we want to compare, with a total of $n = n_1 + n_2$ samples. The basic alternative is that we let Equation 1 describe both groups, i.e. assume no difference between the parameters a_g and b_g for the different groups and thus no differential rhythmicity. This is the null hypothesis H_0. Values y are fitted

Fig. 1. Specificity and power of the cutoff based method. FDR (*x* axis) should here be interpreted as a *P* value for the probability that using a particular harmonic regression *P* value cutoff (*y* axis) results in: (blue curves) a misclassification of a time series with an actual amplitude smaller than a given value as specified in the legend as having a bigger amplitude (BA) than the value, or (red curves) a misclassification of a rhythmic time series with an actual amplitude bigger than a given value as having a lower amplitude (LA) than the value. Two guidelines (grey) at $P = 0.3$ and 10^{-5} may aid the eye to get a quick quantitative overview

using the standard linear least squares procedure: harmonic regression (Cornelissen, 2014). We set up an alternative (H_1) to describe the situation where there is a change in the rhythmicity between the groups and also fit this extended model to the y values. This translates into:

$$H_1 : \begin{cases} y_1 = & 1+ & x\beta_1+ & 0\beta_2+ & e_1 \\ y_2 = & 1+ & x\beta_1+ & x\beta_2+ & e_2 \end{cases}, \quad \beta_2 \neq 0, \quad (2)$$

$$H_0 : \quad \beta_2 = 0$$

with $x = [\cos(\omega t) \quad \sin(\omega t)]$ and $\beta_g = \begin{bmatrix} a_g \\ b_g \end{bmatrix}$ with t as a column vector of all time points. The alternative hypothesis may describe both a change in amplitude and in phase, quantified by β_2. Estimated phases and amplitudes using the least squares method may be represented as sticks in a plane, where the length of the stick represents the amplitude and the direction represents the phase. The procedure also yields standard error estimates for a and b represented by a confidence ellipse, and amplitude, phase, and confidence ellipses may thus be represented as balls on sticks ('lolliplots', Fig. 2). The differences between the groups is then quantified by the distance between the two lolliplots for the separate fits (Fig. 2). We use an ANOVA procedure to test the whether the quality of fit gained by fitting two rhythmic models (H_1), as compared to just one (H_0), is statistically significant (Fig. 2 and Section 2). The quality of fit is in this case quantified as the sum of the squared differences e_{gi}, which are assumed to be independent and identically (for both groups g) distributed with (approximately) a Gaussian distribution. This particular ANOVA procedure is sometimes termed Chow test.

Finally, we noted that even if rhythms do not change, biological noise levels may differ between groups. In this way, rhythms may be more buried in random variability in one group than in the other: the signal to noise ratios might differ. From a biological viewpoint, this effect may in some settings be functionally equivalent to differential rhythmicity. Equations 1 allow a straightforward testing procedure also for this problem: whether the variance of $e1$ equals that of $e2$. For the null hypothesis of equal variances, a scale test of the ratio of the sum-squared residuals of group 1 and 2 will follow an F distribution (Section 2) and thus addresses this problem. In practice, researchers may want to combine the differential rhythmicity ANOVA test and the scale test.

With this, we had a general procedure for quantifying differential rhythmicity and assigning a P value to the measured rhythmicity difference, as well as a general test for differential signal to noise ratios. We implemented these harmonic ANOVA (hereafter: HANOVA) procedures as an R package named DODR (for Detection Of Differential Rhythmicity), and aimed to assess its performance in settings typical for transcriptome-wide circadian rhythm studies.

To gauge the specificity and power of the DODR HANOVA procedure, we first designed artificial experiments with 500 000 simulated pairs of time series. In one class of experiment, we tested detection of phase shifts. Simulated time series consisted of a mean value of 1, a rhythmic component with a period of 24 h, a relative amplitude 0.2, and an added Gaussian random noise term with standard deviation 0.1. These values are typical of mRNA expression levels detected in tissue samples (Thaben and Westermark, 2014)). We generated 12, 16, or 24 samples over two periods (48 h). Phase difference between the two time series in each pair were either 0 (true negatives) or 4.5 h (true positives, reflecting biologically relevant phase shifts that are non-trivial to detect). The HANOVA procedure was applied to each pair, yielding P values for the null hypothesis of no difference in rhythms (0 h phase difference).

In practice, only pairs with low P values would be retained as candidate hits likely to have a real phase difference. For each P value cutoff (ideally equal to the false positive rate, FPR), we computed the FPR (fraction of true negatives misclassified as positives) and the recall (fraction of true positives correctly classified). A graph of the recall against FPR is called a ROC (receiver operator characteristic) curve, and should be oriented close to the upper left corner (i.e. recall $\gg$ FPR, with the area under the curve (AUC) approaching 1) if the P values of the HANOVA methods are meaningful. This was indeed the case (Fig. 3a, black solid curve); the 4.5 h phase shift was detected with high accuracy (AUC = 0.9995).

We investigated a pure amplitude shift with simulated time series with parameters as above, forming pairs with no phase shift but with either the same amplitude, or with amplitudes of 0.2 and 0.1, respectively. Although this is a significant two-fold amplitude change, the lolliplots show that the detection of amplitude changes in general are harder than phase changes: the balls on the sticks tend to be closer. Indeed, while still acceptable, the ROC curve for this task reflected the general greater difficulty of detecting amplitude changes (Fig. 3a, black dashed curve, AUC = 0.8807).

Finally, the scale test was benchmarked for the case of a noise increase of from 0.1 to 0.2 at constant amplitude 0.2. The ROC curve for this scale test setting showed a power lying between those of the phase shift and amplitude change situations (Supplemental Fig. S3A, AUC = 0.9699).

3.3 Methods robust to outliers

Although the HANOVA and scale test procedures seemed promising for detecting differential rhythmicity, the presence of outliers (measurements that markedly deviate from other measurements) might present a serious problem. In practice, outliers may be present both in RNA sequencing data, whose error distributions are often markedly non-Gaussian, and for microarray based measurements, where array or hybridization quality may fluctuate in complex manners.

When including one outlier in each simulated time series (Section 2), the ROC curves indeed indicated significantly reduced power of the HANOVA method for the detection of changes in phase, amplitude and noise levels (Fig. 3B and S3B, AUCs now only 0.9533, 0.6867 and 0.8402 for phase, amplitude and noise, respectively). We therefore next aimed at devising a robust method for the differential rhythmicity problem, capable of better handling outliers.
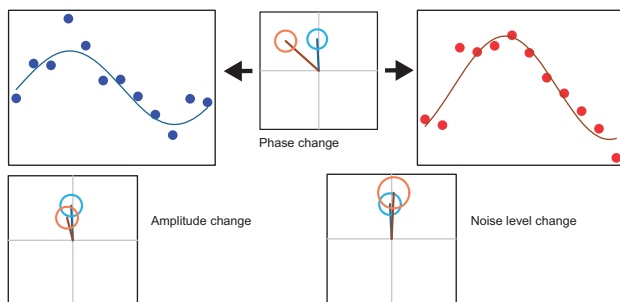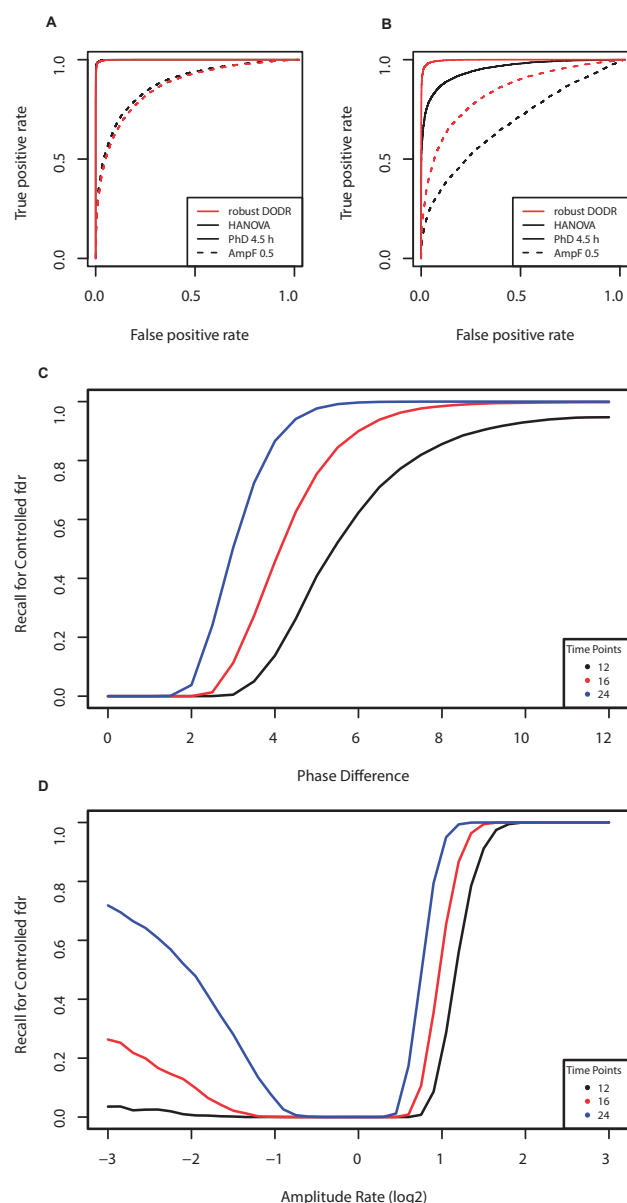

Fig. 2. Quantification of rythms and definition of differential rhythmicity. Shown are two rhythmic time series where the red samples have a larger amplitude and later phase than the blue samples. Harmonic regression results in confidence ellipses, which together with the amplitude (length) and phase (angle) form a ball on a stick (lolliplot) representation in a plane. Amplitude and phase change may be detected by the HANOVA or the robust DODR procedure, whereas noise changes (right) may be detected by scale tests

Fig. 3. Systematic assessment of power and specificity of differential rhythmicity methods. Systematic assessment of power and specificity of differential rhythmicity methods. (**A**) ROC curves for simulated time series without outliers (Section 2). PhD = phase difference, AmpF = amplitude factor: one amplitude was 0.2 and the other 0.1. AUCs 0.9995 and 0.8807 for HANOVA PhD and AmpF, and 0.9991 and 0.8681 for robust DODR. (**B**) ROC curves as in panel A, except that time series had outliers (Section 2). AUCs 0.9533 and 0.6867 for HANOVA PhD and AmpF, and 0.9972 and 0.8389 for robust DODR. (**C**) Power (or recall) for the amplitude change setting from a reference value of 0.2, for various numbers of samples. Prevalence and FDR were fixed at 5% and 0.1, respectively. (**D**) Power for the phase change setting, where only phase differed with 4.5 between the time series, amplitudes were constant at 0.2. Other parameters as for panel C

For the general regression problems, robust rank based methods have been developed that work quite well when there are outliers in the data. Therefore, we implemented a robust rank based method for the differential rhythmicity problem. For this, we stayed with the general formulation (Equation 1), but switched to rank-based methods for quantifying and comparing the qualities of the different fits (Hettmansperger and McKean, 2011; Section 2). This was done

both for the amplitude and phase change problem, as well as for the scale (noise change) problem. The main difference to the HANOVA procedures is that instead of sum of squared differences *egi*, a score function that includes the ranks of the *egi* values is used instead. As before, *egi* are assumed to be independent and identically distributed, but here not necessarily with a Gaussian distribution, which endows the method with robustness to outliers.

Applying this robust method indeed resulted in marked improvements of the ROC curves for the cases with outliers (Fig. 3B and S3B red curves, AUCs 0.9972, 0.8389 and 0.9411 for phase, amplitude and noise, respectively). As is the hallmark of sound robust methods, there were only very marginal losses of power when comparing the robust method to HANOVA for the cases without outliers (Fig. 3B, red curves, AUCs 0.9991, 0.8681 and 0.9549 for phase, amplitude and noise, respectively). We thus concluded that the robust DODR method is a promising method for application to biological data.

We proceeded to more systematically investigate the power of the robust DODR method for detecting various phase shifts and amplitude differences in typical experimental settings. Researchers are often interested in how large the recall is for a given FDR. For instance, a recall of 0.5 at an FDR of 0.1 means that 50% of the true positives are captured, when using a *P* value cutoff that results in at most 10% of the hits being false discoveries (the *P* value cutoff is usually determined by using Benjamini Hochberg-adjusted *P* values). We investigated the recall for various FDRs, various prevalences (fraction of true positives among the simulated time series pairs) and various numbers of samples often taken in genome-wide experiments for investigating circadian rhythms: 12, 16 and 24 time points, respectively. We first fixed the FDR to 0.1 and the prevalence to 0.05 (i.e. 5% of the pairs had true differences in rhythmicity), and included outliers in the simulated time series. Phase difference were varied between 0 and 12 h, while the amplitude change was varied between −3 and 3 log2 fold changes, with a reference amplitude of 0.2 as above.

We found that a 4-fold (2 log2 fold) amplitude change, i.e. between 0.2 and 0.8, is readily detectable with a recall of nearly 1 (Fig. 3C). However, for the 2-fold amplitude change between 0.2 and 0.4, the recall had sunk to only around 0.5 if the number of samples was 12. For 16 samples, the recall was still around 80%, and for 24 samples, the recall was around 95%. The recall then rapidly sank as the amplitudes became more similar. The situation was not symmetric: detecting amplitude losses from relative amplitudes of 0.2 to lower amplitudes was harder (Fig. 3C). This is because the signal-to-noise ratio was approached (the noise level was constantly at 0.1). Phase change detection was nearly perfect for all numbers of samples for phase differences >8 h, at which point the detection power started to deteriorate for the 12 samples case (Fig. 3C). This point of deterioration decreased with ∼2 h for each of the 16 and 24 samples cases, so that phase shifts of ∼4 h still are almost perfectly detected for the 24 samples case. Phase shifts < 2 h were in all cases very hard to detect. Phase shift detection became easier for higher amplitudes than the default 0.2, with phase shifts >2 h almost perfectly detected for 16 samples per group when the amplitude was 0.8 (Supplementary Fig. S3C). The prevalence and FDR did not have as dramatic effects as the number of samples, although naturally, higher FDRs and prevalences increased detection power (Supplementary Fig. S3D–G).

These results provide useful practical guidelines about what to expect in terms of detection power and false discoveries for various numbers of samples, which should aid future experimental planning. In particular, we note that for 12 samples per group, the recall is

already limited and we thus would not recommend fewer samples for differential rhythmicity analysis.

3.4 Application: the light-induced circadian transcriptome

In order to demonstrate the usefulness of differential rhythmicity analysis to gain biological insight, we applied the method to mRNA expression data collected from mouse liver to address a specific biological question.

In mammals, circadian rhythms in mRNA expression of cells in tissues depend on the cellular clocks in the tissue, as well as on circadian systemic cues. Such systemic cues, which may be humoral or neuronal, are orchestrated by the suprachiasmatic nuclei (SCN) in the hypothalamus. The SCN become synchronized to external light-dark cycles via innervation from the retinohypothalamic tract (Dibner *et al.*, 2010). For a metabolic organ such as the liver, there are also rhythmic cues that are not directly emanating from the SCN, such as feeding-related cues (Menaker *et al.*, 2013). In addition, there appears to be pathways that relay cues from external light-dark cycles directly to peripheral organs such as the liver, without involving the circadian clock in the SCN (Husse *et al.*, 2015).

There are two standard protocols when performing assays on peripheral tissues of mice: Either animals are kept in 12 h–12 h ambient light-dark (LD) cycles, or they are kept in constant darkness (DD) (Dunlap *et al.*, 2004). Animals are then harvested at different time points (time is measured Zeitgeber time, ZT, for LD, or circadian time, CT, for DD). The DD protocol grants a greater certitude that measured rhythms originate from the endogenous circadian clock, rather than from direct light cues: strictly, rhythms in LD are sometimes referred to as diurnal rather than circadian in order to reflect this (here, however, we use the term circadian throughout). On the other hand, the LD protocol may be closer to natural conditions.

Especially with regard to circadian mRNA expression, liver is particularly well studied, exhibiting thousands of transcripts whose abundances cycle in a circadian manner (Zhang *et al.*, 2014). There are differences in circadian rhythms in mouse liver mRNA expression levels as measured under DD and LD conditions, respectively (Korenčič *et al.*, 2014). However, to the best of our knowledge, these differences have not been quantified on a transcriptome-wide scale.

Thus, we asked whether there are widespread differences in the mouse liver circadian transcriptome between animals kept under DD and LD conditions. To investigate this, we collected microarray-based transcriptome data from two different studies, one in which mice were held in DD (Hughes *et al.*, 2009) and one where mice were held in LD (Jouffe *et al.*, 2013). In both cases, C57BL/6J mice were studied, and the same microarray platform was employed.

Out of 7171 expressed transcripts, we obtained 1580 transcripts that were cycling in either condition (see Section 2 for criteria). Using DODR with a FDR of 0.1, we found that 262 of these transcripts exhibit differential rhythmicity. Of these, only 10 were pure phase shifts (see Section 2 for criteria and Figure S4 for time series plots of these transcripts). There was a broad tendency for increased amplitudes in the LD compared to DD mice, 183 transcripts exhibited amplitude gains (Fig. 4A and Supplementary Fig. S5). These transcripts tended to have late phases with a median phase at ZT 17.2 (Fig. 4B). This late phase distribution was significantly different from the rest of the LD circadian transcriptome ($P < 0.001$, Watson's two-sample test, Supplementary Fig. S6A). Furthermore, a gene ontology (GO) analysis of these transcripts strongly suggested that they are involved in immune response functions (Supplementary Table S1 and Section 2).
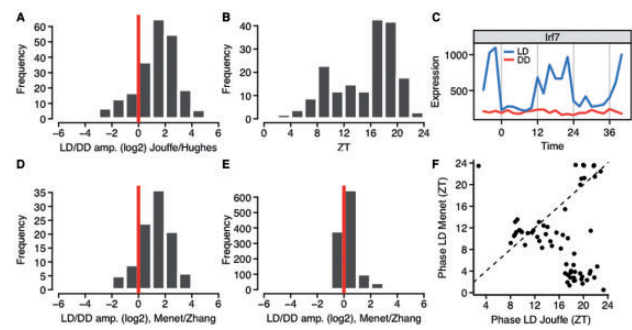

Fig. 4. Application of DODR differential rhythmicity analysis to mouse liver DD and LD mRNA expression data. Application of DODR differential rhythmicity analysis to mouse liver DD and LD mRNA expression data. (**A**) Amplitude log2 fold changes from DD to LD, for DODR *P* value < 0.1. (**B**) LD phases of transcripts with larger amplitudes in LD, circular median ZT 17.2. (**C**) Time series for the *Irf7* transcript in LD and DD, respectively. (**D, E**) Amplitude log2 fold changes from DD to LD for the validation dataset, (D) for transcripts already determined to be differentially rhythmic with an amplitude change in the original data (Section 2), and (E) for transcripts not determined to have an amplitude change in the original data. (**F**) Phases of original and validation LD transcripts that were already determined to be differentially rhythmic with an LD amplitude gain in the original data. Only transcripts with amplitudes larger than 0.15 in both LD datasets were used, to obtain clean phase estimations

Which mechanisms might underlie these amplitude gains? Transcription factors (TF) with circadian patterns of abundance are an important factor behind circadian rhythms in mRNA expression. Overrepresentation patterns of binding sites for TFs in promoters of genes expressed in a circadian fashion can provide clues to TFs and mechanisms behind circadian regulation of transcription (Lück and Westermark, 2016). In order to isolate candidate TFs involved in the differentially rhythmic LD transcriptome, we searched for TF binding sites enriched in the promoter regions within 300 bp of the transcription start sites of the differentially rhythmic transcripts (Section 2). Enrichment was computed against the background of all circadian transcripts. The top candidate binding site was for IRF1, 2 and 7, with adjusted *P* value $<10^{-5}$ (two-sided Fisher's exact test for enrichment). Notably, the *Irf7* transcript was itself differentially rhythmic, with a strong circadian rhythm in the LD mice and a flat expression profile in the DD mice (Fig. 4C). Also, the phase of the *Irf7* transcript was in the evening, at ZT 18.4, just as differentially rhythmic transcripts with amplitude gain in general. A transcript list with the results of this analysis is provided in Supplementary Table S2.

Thus, IRF7 emerged as a candidate TF for involvement in circadian transcriptional regulation limited to the LD condition. This is consistent with the strong enrichment of functions related to the immune system among the transcripts with stronger rhythmicity in the LD condition, as revealed by the GO analysis. To validate the findings, we turned to two other mouse liver datasets (hereafter: validation data). These were microarray-based data for the DD conditions with 24 samples taken every 2 h (Zhang *et al.*, 2014), and RNA sequencing data for the LD condition with 12 samples taken every 4 h (Menet *et al.*, 2012). Because of the differences in technology with different dynamic ranges, rhythms should here not be compared directly using DODR: spurious amplitude differences due to technology would bias the analysis. However, we may start with the transcripts identified as differentially rhythmic in the previous analysis, and ask whether the tendencies for amplitudes and phases are the same in the validation data.

The time resolution of the validation data is poorer, but 88 of the 183 transcripts previously found to have amplitude gains had

circadian rhythms in both LD datasets (see Section 2 for criteria). The vast majority of these exhibited amplitude gains in LD compared to DD also in the validation dataset (Fig. 4D and S7). For the transcripts found to be rhythmic in the validation dataset, but not found to be differentially rhythmic in the original dataset, amplitude gains were much more moderate (Fig. 4E, $P < 10^{-15}$, rank sum test), so that technical artifacts may be excluded. Thus, amplitude gains under LD compared to DD conditions were widespread and highly reproducible. The transcripts exhibiting pure phase shift in the original dataset exhibited very similar phase shifts in the validation dataset, which underscored the reproducibility (Supplementary Fig. S8). However, phases of transcripts with increased amplitude in the original analysis tended to have early to mid-day rather than late phases in the validation mice (Supplementary Fig. S6B). This discrepancy was visible in a phase comparison between the two LD datasets of transcripts with amplitude gains in LD (Fig. 4F). However, the discrepancy was absent among cycling transcripts without amplitude gains (Supplementary Fig. S6C). Thus, there seems to be a group of transcripts that cycle more strongly in LD, but whose phase might be rather plastic.

Since the transcripts with plastic phase had evening phases in the original dataset, IRF7 could be involved in the transcriptional regulation of these plastic transcripts. Indeed, the *Irf7* transcript was rhythmic in the validation LD but not DD mice, exhibiting a morning rather than an evening phase (Supplementary Fig. S6D). The putative IRF7-regulated transcripts (i.e. with predicted binding site(s) in the promoters) also exhibited larger amplitude gains than other circadian transcripts in the validation data ($P = 0.0023$, rank sum test). Thus, IRF7 binding sites were predictive of larger amplitude in LD as compared to DD also in the validation dataset. Furthermore, transcripts with promoters having IRF7 binding sites clearly tended to have plastic phases, with many transcripts exhibiting evening phases in the original dataset and morning phases in the validation data (Supplementary Fig. S6E).

Among possible explanations for the discrepancy could be different feeding conditions: in one of the studies, mice were subjected to restricted feeding during night, to reduce confounding influence from feeding-induced cues (Jouffe *et al.*, 2013). Another explanation could be differing lighting strengths in the mouse facilities. This explanation would imply a strong phase sensitivity of LD-specific rhythms in mRNA expression to lighting strength.

4 Discussion

We presented general methods for detecting changes in rhythm, both in amplitude and phase. This works for a variety of types of paired measurements, such as pairs of single or thousands of time series measured under two different conditions. We showed that generally, at least $\sim$12 samples in each group are needed, and also quantified the improvements that can be expected for larger number of samples. The methods were implemented as the freely available R package DODR, appearing on the R package repository CRAN together with the present paper.

When the main interests of an investigation are pure amplitude gains or losses, the threshold based method might be used, and the performance measures that we computed (Fig. 1) may be used as guideline. However, a more robust strategy is probably to use DODR, and to infer the nature of the changes *post hoc*, by amplitude and phase estimates (also provided by the DODR program).

We applied differential rhythmicity analysis to quantify differences in the circadian transcriptomes in livers of mice held in DD or LD cycles, respectively. This analysis showed that the primary difference between the LD and DD is an amplitude boost for a few hundred transcripts in the LD condition. This effect was robustly reproducible, as shown by analysis of an independent validation dataset. Further, the analysis suggested that IRF7 is a putative factor involved in generating circadian rhythms in transcription under LD conditions. The *Irf7* transcript was shown previously to exhibit a circadian rhythm in mouse retina, even in mice with a mutation that disrupts the core cellular circadian clock (Storch *et al.*, 2007), supporting a role for light cues directly driving rhythms in the abundance of this transcript. Possibly, the immune system is involved in relaying light-induced rhythmic cues to *Irf7* transcription, since transcription of this gene is regulated by the JAK-STAT pathway (Ning *et al.*, 2005). Although the set of transcripts that gained rhythmicity in LD was reproducible and seen in the validation dataset as well, curiously these transcripts tended to have different phases in the two different LD experiments considered. We can only speculate about the reasons for this, which may include laboratory conditions. In any case, this phase plasticity is a notable phenomenon and probably a general feature of the circadian program (van der Vinne *et al.*, 2015).

Several research paths might be followed to develop differential rhythmicity analysis further. The present methods does not address possible period changes. In principle, such an analysis would not present serious difficulties, other than that nonlinear fitting methods would have to be employed. Standard likelihood ratio tests for problem formulations similar to Equation 1 could then be performed. More challenging future developments could be development of methods for detecting changes in wave form or underlying trends or oscillation envelopes.

In summary, the methods presented here will be useful for research on biological rhythms such as circadian rhythms when different conditions are investigated. There are several examples: circadian rhythms are often studied in various mutants or in other conditions such as aging organisms (Chen *et al.*, 2016; Eckel-Mahan *et al.*, 2013; Sancar *et al.*, 2011; Wijnen *et al.*, 2006). Rigorous differential rhythmicity analysis could routinely be employed to assess alterations or loss of rhythmicity.

References

Adamovich,Y. *et al.* (2014) Circadian clocks and feeding time regulate the oscillations and levels of hepatic triglycerides. *Cell Metab.*, **19**, 319–330.

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Berger,R.L. and Hsu,J.C. (1996) Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Stat. Sci.*, **11**, 283–319.

Chen,C.Y. *et al.* (2016) Effects of aging on circadian patterns of gene expression in the human prefrontal cortex. *Proc. Natl. Acad. Sci. USA*, **113**, 206–211.

Cornelissen,G. (2014) Cosinor-based rhythmometry. *Theor. Biol. Med. Model.*, **11**, 16.

Dibner,C. *et al*. (2010) The mammalian circadian timing system: organization and coordination of central and peripheral clocks. *Annu. Rev. Physiol*., **72**, 517–549.

Doherty,C.J. and Kay,S.A. (2010) Circadian control of global gene expression patterns. *Annu. Rev. Genet*., **44**, 419–444.

Dunlap,J.C. *et al*. eds. (2004) *Chronobiology: Biological Timekeeping Sinauer Associates*, Sunderland, MA, USA.

Eckel-Mahan,K.L. *et al*. (2013) Reprogramming of the circadian clock by nutritional challenge. *Cell*, **155**, 1464–1478.

Ferrell,J.E.Jr *et al*. (2011) Modeling the cell cycle: why do certain circuits oscillate? *Cell*, **144**, 874–885.

Hazlerigg,D.G. and Wagner,G.C. (2006) Seasonal photoperiodism in vertebrates: from coincidence to amplitude. *Trends Endocrinol. Metab*., **17**, 83–91.

Hettmansperger,T.P. and McKean,J.W. (2011) *Robust Nonparametric Statistical Methods*. 2nd edn. CRC Press, Boca Raton, FL.

Huang,D.W. *et al*. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc*., **4**, 44–57.

Hughes,M.E. *et al*. (2012) Brain-specific rescue of Clock reveals system-driven transcriptional rhythms in peripheral tissue. *PLoS Genet*., **8**, e1002835.

Hughes,M.E. *et al*. (2009) Harmonics of circadian gene transcription in mammals. *PLoS Genet*., **5**, e1000442.

Hughes,M.E. *et al*. (2010) JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J. Biol. Rhythms*, **25**, 372–380.

Husse,J. *et al*. (2015) Synchronization of the mammalian circadian timing system: light can control peripheral clocks independently of the SCN clock. *BioEssays*, **37**, 1119–1128.

Ideker,T. *et al*. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol*., **7**, 805–817.

Jouffe,C. *et al*. (2013) The circadian clock coordinates ribosome biogenesis. *PLoS Biol*., **11**, e1001455.

Kloke,J. and McKean,J.W. (2014) *Nonparametric Statistical Methods Using R*. CRC Press.

Korenčič,A. *et al*. (2014) Timing of circadian genes in mammalian tissues. *Sci. Rep*., **4**, 5782.

Kornmann,B. *et al*. (2007) System-driven and oscillator-dependent circadian transcription in mice with a conditionally active liver clock. *PLoS Biol*., **5**, e34.

Koul,H.L. *et al*. (1987) An estimator of the scale parameter for the rank analysis of linear models under general score functions. *Scand. J. Stat*, **14**, 131–141.

Lin,X.W. *et al*. (2015) Clocks within the master gland hypophyseal rhythms and their physiological significance. *J. Biol. Rhythms*, **30**, 263–276.

Lück,S. *et al*. (2014) Rhythmic degradation explains and unifies circadian transcriptome and proteome data. *Cell Rep*., **9**, 741–751.

Lück,S. and Westermark,P.O. (2016) Circadian mRNA expression: insights from modeling and transcriptomics. *Cell. Mol. Life Sci*., **73**, 497–521.

Masri,S. *et al*. (2014) Partitioning circadian transcription by SIRT6 leads to segregated control of cellular metabolism. *Cell*, **158**, 659–672.

McCarthy,J.J. *et al*. (2007) Identification of the circadian transcriptome in adult mouse skeletal muscle. *Physiol. Genomics*, **31**, 86–95.

Menaker,M. *et al*. (2013) Central control of peripheral circadian oscillators. *Curr. Opin. Neurobiol*., **23**, 741–746.

Menet,J.S. *et al*. (2012) Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. *eLife*, **1**, e00011.

Miyazaki,M. *et al*. (2011) Age-associated disruption of molecular clock expression in skeletal muscle of the spontaneously hypertensive rat. *PLoS ONE*, **6**, e27168.

Ning,S. *et al*. (2005) Regulation of the transcriptional activity of the IRF7 promoter by a pathway independent of interferon signaling. *J. Biol. Chem*., **280**, 12262–12270.

Pachkov,M. *et al*. (2007) SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res*., **35**, D127–D131.

Sancar,G. *et al*. (2011) A global circadian repressor controls antiphasic expression of metabolic genes in Neurospora. *Mol. Cell*, **44**, 687–697.

Storch,K.F. *et al*. (2007) Intrinsic circadian clock of the mammalian retina: importance for retinal processing of visual information. *Cell*, **130**, 730–741.

Thaben,P.F. and Westermark,P.O. (2014) Detecting rhythms in time series with RAIN. *J. Biol. Rhythms*, **29**, 391–400.

van der Vinne,V. *et al*. (2015) Food reward without a timing component does not alter the timing of activity under positive energy balance. *Neuroscience*, **304**, 260–265.

Vollmers,C. *et al*. (2009) Time of feeding and the intrinsic circadian clock drive rhythms in hepatic gene expression. *Proc. Natl. Acad. Sci. USA*, **106**, 21453–21458.

Wang,X. *et al*. (2012) An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. *Bioinformatics*, **28**, 2534–2536.

Westermark,P.O. and Lansner,A. (2003) A model of phosphofructokinase and glycolytic oscillations in the pancreatic β-cell. *Biophys. J*., **85**, 126–139.

Wijnen,H. *et al*. (2006) Control of daily transcript oscillations in Drosophila by light and the circadian clock. *PLoS Genet*., **2**, e39.

Wijnen,H. *et al*. (2005) Molecular and statistical tools for circadian transcript profiling. *Methods Enzymol*., **393**, 341–365.

Zhang,R. *et al*. (2014) A circadian gene expression atlas in mammals: implications for biology and medicine. *Proc. Natl. Acad. Sci*., **111**, 16219–16224.