An evolutionary conservation based benchmark of seven methods detecting nycthemeral rhythms in gene expression shows that only strong rhythmic signals are reliably detected.

David Laloum[1,2], Marc Robinson-Rechavi[1,2,*]

[1] Department of Ecology and Evolution, Batiment Biophore, Quartier UNIL-Sorge, Université de Lausanne, 1015 Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics, Batiment Génopode, Quartier UNIL-Sorge, Université de Lausanne, 1015 Lausanne, Switzerland

* marc.robinson-rechavi@unil.ch

Abstract

The nycthemeral transcriptome embodies all genes displaying a rhythmic variation of their mRNAs periodically every 24 hours, including but not restricted to circadian genes. In this study, we show that the nycthemeral rhythmicity at the gene expression level is biologically functional and that this functionality is more conserved between orthologous genes than between random genes. We used this conservation of the rhythmic expression to assess the ability of seven methods (ARSER, Lomb Scargle, RAIN, JTK, empirical-JTK, GeneCycle, and meta2d) to detect rhythmic signal in gene expression. We have contrasted them to a naive method, not based on rhythmic parameters. By taking into account the tissue-specificity of rhythmic gene expression and different species comparisons, we show that no method is strongly favored. The results show that these methods designed for rhythm detection, in addition to having quite similar performances, are consistent only among genes with a strong rhythm signal. Rhythmic genes defined with a standard p-value threshold of 0.01 for instance, could include genes whose rhythmicity is biologically irrelevant. Although these results were dependent on the datasets used and the evolutionary distance between the species compared, we call for caution about the results of studies reporting or using large sets of rhythmic genes. Furthermore, given the analysis of the behaviors of the methods on real and randomized data, we recommend using primarily ARS, empJTK, or GeneCycle, which verify expectations of a classical distribution of p-values. Experimental design should also take into account the circumstances under which the methods seem more efficient, such as giving priority to biological replicates over the number of time-points, or to the number of time-points over the quality of the technique (microarray vs RNAseq). GeneCycle, and to a lesser extent empirical-JTK, might be the most robust method when applied to weakly informative datasets. Finally, our analyzes suggest that rhythmic genes are mainly highly expressed genes.

Introduction

The nycthemeral transcriptome is characterized by the set of genes that display a rhythmic change in their mRNAs levels with a periodicity of 24 hours. These include, but are not limited to, circadian genes whose rhythm is endogenous and entrainable. In baboon, 82% of protein-coding genes have been reported to be rhythmic in at least one tissue [27]. The nycthemeral rhythmicity of these transcripts can be driven by the internal oscillator clock or by other circadian input such as food-intake, the light-dark cycle, sleep-wake behavior, or social activities. Moreover, the nycthemeral transcriptome is tissue-specific [43]. Given the importance of biological rhythms in understanding biology and medicine, many algorithms have been proposed to detect such rhythms. Some were developed specifically for biological data, while others were adapted from other fields where periodicity is important, such as Lomb Scargle (LS). Most methods are based on non-parametric models that search for referenced patterns, classically sinusoid, called time-domain methods, while some are frequency-domain methods based on spectral analysis [41]. Some of them have been designed to detect more diverse waveforms, including asymmetric patterns, such as RAIN [37] or empirical_JTK (empJTK) [19]. For instance, RAIN outperformed the original JTK_CYCLE algorithm for simulated data consisting of sinusoidal and ramp waveforms [19]. Thus, methods differ in the conception of their algorithm and in how they take into account features of the dataset such as curve shapes, period, noise level, presence of missing data, phase shifts, sampling rates [10], asymmetry of the waveform, or the number of cycles (total period length of the experiment). Each method has in principle different strengths and weaknesses for some features of the dataset. In *Arabidopsis*, HAYSTACK identified 45% more cycling transcripts than COSOPT, mainly due to the inclusion of a 'spike' pattern in its model [25]. Deckard et al. [10] studied how four methods (LS, JTK_CYCLE, de Lichtenberg, and persistent homology) performed across a variety of organisms and periodic processes. Based on synthetic data, they investigated the algorithms' ability to distinguish periodic from non-periodic profiles, to recover period, phase and amplitude, and they evaluated their performance for different signal shapes, noise levels, and sampling rates. They proposed a decision tree to recommend one of these four algorithms based on these features of datasets [10].

The performance of algorithms to identify such periodic signal has been assessed so far based on synthetic (i.e., simulated) data, or on benchmark sets of known cycling genes. Hughes et al. [17] recently published guidelines for the analysis of biological rhythms and proposed a web-based application (CircaInSilico) for generating synthetic genome biology data to benchmark statistical algorithms for studying biological rhythms. While such benchmarks are useful to explore the behavior of methods in a set of cases, the applicability of results to real data is limited. For example, simulations need to impose an a priori fluctuation pattern, typically cosine. The fluctuation of transcript abundance of core clock genes does seem to follow a cosine shape [21], but sometimes follows non-sinusoidal periodic patterns in mouse liver (e.g., Nr1d1 or Arntl) [7] (based on the data from [26]). The fluctuations of the nycthemeral transcriptome are entrained by a complex network involving external cues [42] [4] [28] [14], as simplified in Figure 1a, which might yield non-sinusoidal periodic patterns among rhythmic genes even if circadian genes were sinusoidal. But the biological relevance of these waveforms is still not clear. This raises two issues: benchmarks based on simulations are biased towards methods that detect the same types of patterns as simulated; and when an algorithm detects more rhythmic genes, it could be more true positives or more false positives. When pattern constraints are released this increases the number of genes detected as rhythmic, but is not necessarily informative on the capacity of the algorithm to detect genes whose rhythmicity is biologically relevant.

Using real data and randomization tests, we compared seven methods: JTK_CYCLE (**JTK**) [18], **LS** [24]& [33], ARSER (**ARS**) [41], and **meta2d** (Fisher integration of LS, ARS, and JTK), are frequently used by many studies and are all included within the MetaCycle R package [39]. We also included empirical_JTK (**empJTK**) [19] and **RAIN** [37], which have been recently developed to deal with more non cosine patterns and with asymmetric waveforms. empJTK and RAIN aim to improve the original JTK algorithm which assumed that any underlying rhythms have symmetric waveforms [37]. Finally, robust.spectrum [2] extents a robust rank-based spectral estimator to the detection of periodic signals. It is integrated in the GeneCycle R package [1] and called **GeneCycle** in this paper. We excluded de Lichtenberg [9], Persistent Homology [8], COSOPT [35], Fisher's G test [12], MAPES, Capon, and other algorithms for reasons such as i. difficult accessibility of the software which limit their use by researchers, ii. their higher sensitivity to certain features of the data such as the sampling density, the number of replicates and/or periods, noise level, and waveform, iii. their weaker efficiency on simulated data or known cycling genes, or iv. their previously reported less good detection of non-sinusoidal periodic patterns [10] [41] [44] [11] [40] [19]. We first analysed the behavior of these seven methods applied to a variety of real datasets in animals, and within each dataset, we compared results between representative gene subsets such as highly and lowly expressed genes, known cycling genes, and randomized data. Contrary to real data, randomized data is not expected to show any signal of rhythmicity, which we used to test proper statistical behavior under the null hypothesis. Secondly, as function tends to be conserved between orthologs [13], true rhythmic genes are expected to be enriched in orthologs that are themselves rhythmic in other species. Indeed, evolutionary conservation provides a valuable filter through which to highlight functional biological networks, notably for clock-controlled functions [15]. The biological relevance of rhythmic genes is expected to be higher for rhythmic orthologs. On the other hand, errors in the prediction of rhythmicity by each method are not expected to be conserved between orthologs. Thus, the best methods are expected to report rhythmic genes with a high proportion of rhythmic orthologs. We used this approach to compare the algorithms based on their ability to capture the evolutionary conservation signal within nycthemeral genes, and compared them to a Naive method.

Results

We used gene expression time-series datasets that come from circadian experiments and kept the data from "normal" individuals (apparently healthy, wild-type, no treatment) for seven species (Additional Table 1), allowing comparisons among vertebrates and among insects. We benchmarked methods on animal data since organ homology allowed to compare datasets for which we expect conservation of functional patterns (tissue-specific rhythms). For readability, we present vertebrate results in the main figures and insect results in supplementary results (Additional file 5). Apart from the rat and Anopheles datasets, data with several biological replicates were obtained already normalized over replicates (one value per time-point).

We define a rhythmic gene as a gene which displays a nycthemeral change in its mRNA abundance, i.e. occurring over 24 hours and repeated every 24 hours. All these rhythmic genes represent the nycthemeral transcriptome. Different organs have been reported to have transcriptomes which are more or less rhythmic [43]. The rhythmic expression of these genes can be entrained directly by the internal clock or indirectly by external inputs, such as the light-dark cycle or food-intake [42] [4] [28] [14] (Fig. 1a). We consider the entirety of these rhythms to be a biologically relevant signal to detect. That is why we preferred data from light-dark and ad-libitum experimental conditions whenever

possible (Additional Table 1), as providing a better representation of wild conditions. Some methods are distinguished by their higher sensitivity to alternative patterns such as peak, box, or asymmetric profiles. A visual inspection of the KEGG "Circadian entrainment" gene set (see Methods) provides indeed informal confirmation that such patterns can be observed among known cycling genes, such as Npas2, Nr1d1, or Bhlhe41 (Additional file 1: Fig. S2).

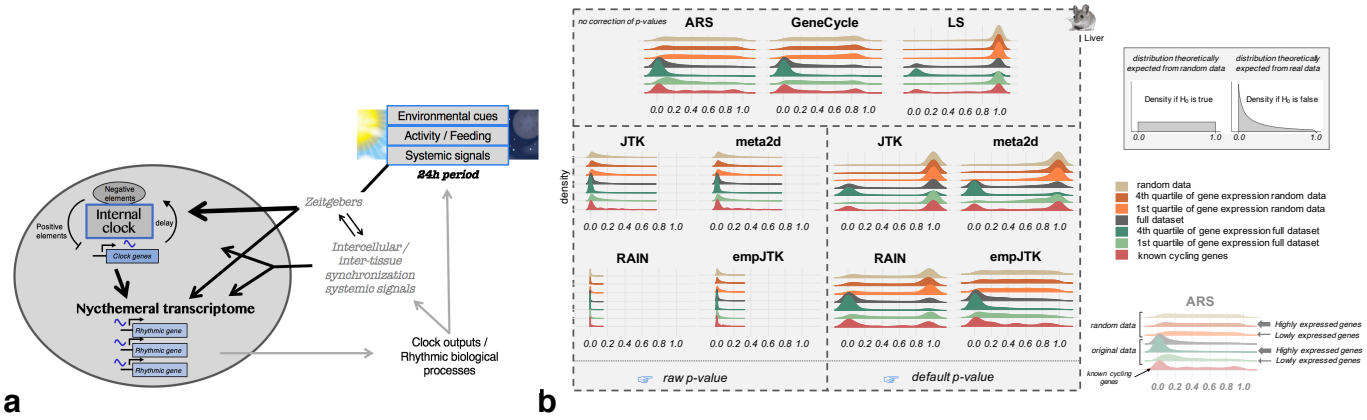

Figure 1. The nycthemeral transcriptome is the group of genes whose mRNAs have periodic variations with a 24h period, called rhythmic genes. To detect these rhythmic genes, we applied seven methods to time-series datasets that produced different density distribution of p-values.
a) Simplified diagram of the entrainment of nycthemeral gene expression. Environmental cues include the light-dark cycle, food-intake, sleep-wake behavior, social activities, or any other 24h periodic event.
b) Density distribution of p-values obtained before (raw) and after the default correction (software) for the seven methods applied to mouse liver data (microarray) sub-categorized in: i. randomized data which represents the null hypothesis; ii. randomized data restricted to the first and fourth quartiles of the median gene expression level, to check for the impact of expression level under the null; iii. the full original dataset; iv. the first and fourth quartiles of the median gene expression level of the original data; and v. a subset of known cycling genes (99 genes from KEGG "circadian entrainment"). The default p-values of ARS, GeneCycle, and LS are uncorrected.

Analysis of statistical behaviors of methods applied to real data

p-values distribution analysis

First, a good method should produce a uniform distribution of p-values when there is no structure in the data, in contrast to the distribution obtained from empirical data, which is expected to be skewed towards low p-values because of the presence of rhythmic genes. We investigated the properties of the different methods applied to randomized vs real data. We also investigated to what extent the density distribution of p-values of each method was affected by gene expression levels. Indeed, higher expression provides more power for detecting rhythmic patterns - highly expressed genes have more chance to shape rhythmic patterns because the variations of expression levels are relative to the general expression level - but this should not be the main driver of results. I.e., a method to detect rhythmicity should not be essentially reporting high expression levels. Even if true rhythmic genes were enriched in high expressed genes, we expect a good method to report both high and low p-values, at each expression level.
Figure 1b shows the density distribution of raw p-values obtained for the seven methods applied to mouse liver data (microarray) sub-categorized in: i. randomized data which

represents the null hypothesis; ii. randomized data restricted to the first and fourth quartiles of the median gene expression level, to check for the impact of expression level under the null; iii. the full original dataset; iv. the first and fourth quartiles of the median gene expression level of the original data; and v. a subset of known cycling genes (8 to 99 genes according to species, see Methods). Results from the other datasets are provided in Additional file 2. Surprisingly, only ARS and GeneCycle displayed close to the expected uniform raw p-value distribution for randomized data (Fig. 1b). The adjustment by default of empJTK (minimum of the p-value calculated from an empirical null distribution, and of Bonferroni) recovered the expected uniform distribution, suggesting that this correction allows recovering proper p-values (Fig. 1b). We used each software output "p-values" for calls, which we call "default p-value". In some software, these values result from an internal p-value adjustment, so we also analysed "raw p-values" (uncorrected, see Methods and Table 1 for JTK). For ARS, GeneCycle, and LS, the default p-values are uncorrected. Under the null hypothesis, LS has an abnormal peak near p-value=1 (Fig. 1b), implying an issue with its definition of the null hypothesis, or maybe a one-sided test when a two-sided test would be appropriate. The three other algorithms (RAIN, JTK, and meta2d) seem to have issues with false positives, displaying large proportions of low p-values even for randomized data.

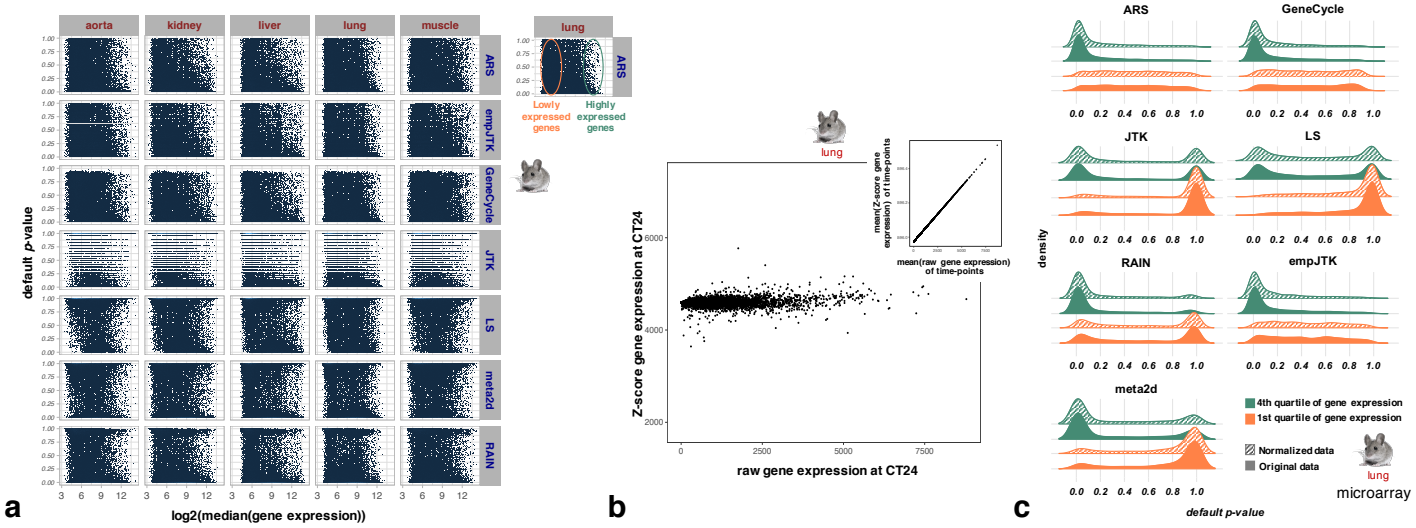

Figure 2. Detected rhythmic genes are enriched in highly expressed genes.
a) Higher expression levels imply a higher power to detect rhythmic patterns.
b) Control of the normalization by Z-score of gene expression values at a given time-point (CT24).
c) Methods applied to original vs normalized gene expression values produce the same distributions of p-values within highly expressed genes, or within lowly expressed genes. Particularly, the normalization of gene expression values does not allow to recover rhythmicity within lowly expressed genes.

Before analysing the impact of expression levels, we checked that the data follow a typical bimodal density distribution of gene expression (Additional file 1: Fig. S3a). We also checked that using the median of time-points for gene expression gives similar results to using the minimum or the mean value (Additional file 1: Fig. S3b). Unsurprisingly, higher expression levels imply a higher power to detect rhythmic patterns (Fig. 2a). Only empJTK, ARS, and GeneCycle presented broad distributions of the default p-values, from 0 to 1, for genes with low levels of expression; the other algorithms showed a large excess of high p-values for lowly expressed genes (Fig. 2a). The p-values distributions

imply that most methods detect almost all highly expressed genes, and almost no lowly expressed genes, as "rhythmic" (Fig. 1b). This led us to be concerned about the risk of a very high false-positive rate for highly expressed genes. This risk could be produced by the high sensitivity to large variations of gene expression levels. To assess this, we normalized gene expression values by a Z-score transformation so that they all had the same mean and variance (see Methods), and applied methods after this transformation (Fig. 2b). This normalization of gene expression values did not change the p-values distributions within highly expressed genes, and particularly did not recover rhythmicity within lowly expressed genes (Fig. 2c). This was not due to sampling biases of microarray data since results are consistent with RNAseq data (Additional file 1: Fig. S3c). Thus, the differences obtained between highly and lowly expressed genes could be explained either by a truly larger number of highly expressed genes among the rhythmic genes, or by experimental noise which masks circadian signal for lowly expressed genes. Overall, ARS, empJTK, and GeneCycle had the best behavior, producing a uniform distribution under the null hypothesis, and a skew towards p-value=0 for all empirical data. Figure 2a indicates that less rhythmic organs, such as the muscle or the aorta, generated a wider distribution of p-values with ARS and GeneCycle, and to a lesser degree with empJTK, for both the lowly and highly expressed genes.

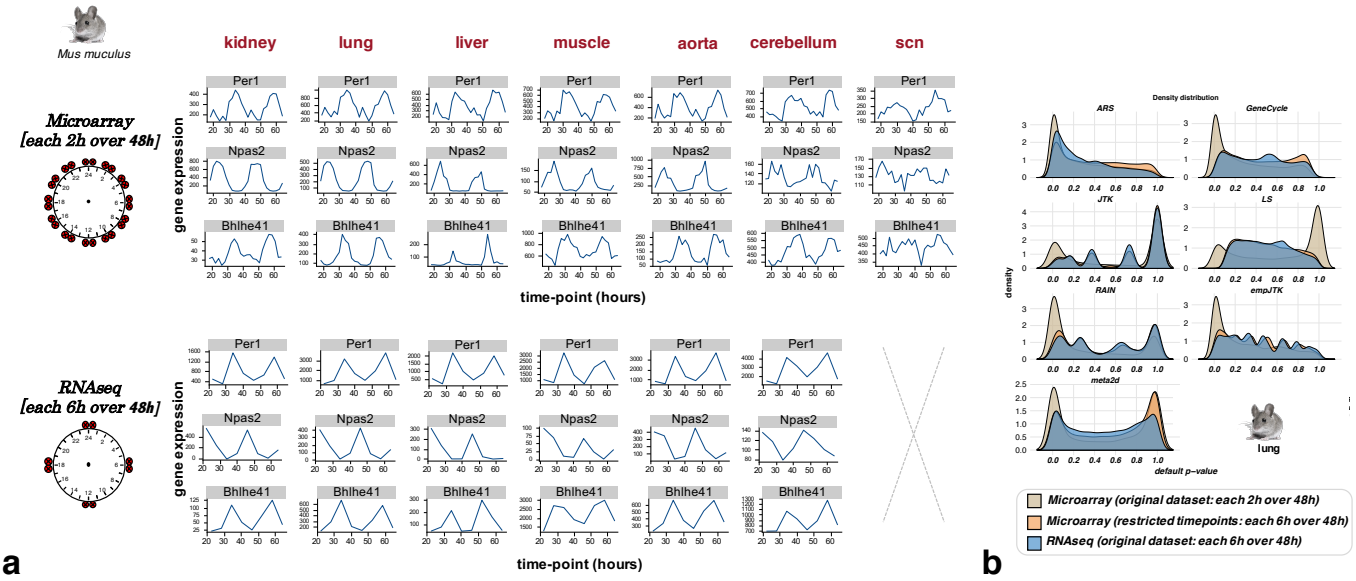

Figure 3. Less time points per cycle lead to a weaker detection of rhythmic patterns even if the transcriptome profiling quality is better.
a) Bhlhe41, Npas2, and Per1 expression over time from data of the same mouse experiment [43] using two transcriptome profiling techniques: microarray vs RNAseq. The number of time-points with data is 24 for microarray and 8 for RNAseq. **b)** The restriction of microarray time-series to the same time-points as in the RNAseq series produces similar p-value distributions to those obtained with RNAseq. This supports a major role of the temporal resolution for method results, relative to a minor role for the difference between RNAseq and microarrays.

From data of the same mouse experiment [43], we observed differences of p-value density distributions between microarray and RNAseq, with the skew towards p-value=0 less marked for RNAseq data (Additional file 1: Fig. S4a and S4b). This can be due to the more precise temporal resolution of the microarray time-series dataset, or to differences in the detection of gene expression by RNAseq vs microarrays (Fig. 3a). When we

restricted the microarray time-series to the same time-points as in the RNAseq series, we obtained a p-value distribution very similar to that of the RNAseq data (Fig. 3b). The same time-series restriction applied to known cycling genes produced comparable results (Additional file 1: Fig. S4c). This supports a major role of the temporal resolution for method results, relative to a minor role for the difference between RNAseq and microarrays. That is why for the next steps, we only considered the microarray dataset for the mouse.

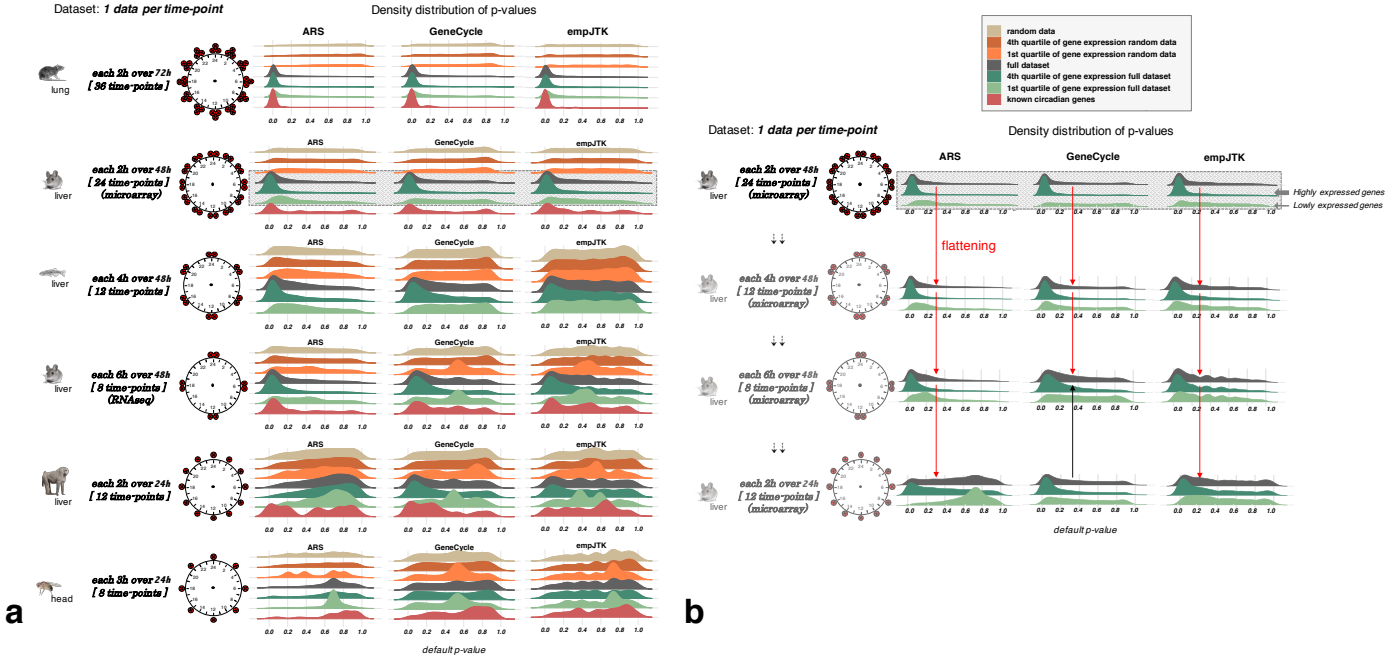

Figure 4. Datasets with one replicate per time-point over a unique cycle of 24 hours do not provide enough information to detect rhythmicity.
Methods lose in efficiency for detecting rhythmic patterns in gene expression when the number of 24h cycles decreases, or when the number of time-points sampled decreases.
a) Default p-value distributions obtained for ARS, GeneCycle, and empJTK applied to different datasets and sub-categorized in: i. randomized data which represents the null hypothesis; ii. randomized data restricted to the first and fourth quartiles of the median gene expression level, to check for the impact of expression level under the null; iii. the full original dataset; iv. the first and fourth quartiles of the median gene expression level of the original data; and v. a subset of known cycling genes (8 to 99 genes according to species, see Methods). For each dataset, the number of time-points with data and the temporal resolution is illustrated around a 24h clock. For the same number of time-points, performance seems better with two cycles than only one cycle (zebrafish vs baboon).
b) The reduction of the number of time-points of the mouse liver microarray dataset shows increasingly weak rhythm detection by ARS, GeneCycle, and empJTK, shown by a flattening of the p-value distribution on the full dataset (red arrow). GeneCycle showed no difference between a few time-points over two cycles or more time-points over a single cycle (black arrow).

This observation can be generalized to diverse datasets. We see that each method loses in efficiency when the number of 24h cycles decreases, or when the number of time-points sampled decreases (Fig. 4a). We show only results for ARS, GeneCycle, and empJTK because they were the only methods with correct behavior in their p-value distributions (Fig. 1b). For the same number of time-points, performance seems better with two cycles than only one cycle, as shown comparing zebrafish and baboon data

which have both twelve time-points (Fig. 4a). But this observation could be confused by the comparison of different species or different samples' quality. ARS performed better with a smaller total number of time-points but over two cycles than with more total time-points over a single cycle (mouse RNAseq vs baboon in Fig. 4a), indicating that ARS is very dependant on the repetitive nature of profiles. The reduction of the number of time-points of the mouse microarray dataset shows similar effects on the rhythm detection by ARS, GeneCycle, and empJTK (Fig. 4b). Of note, GeneCycle presented more or less no differences between having a few time-points over two cycles and having more time-points over a single cycle (black arrow Fig. 4b).

Overlap between methods

Among genes called rhythmic, we analysed the number of those called in common by the different methods. For p-value thresholds of 0.05 or 0.01, we found a large proportion of genes called rhythmic by only one or few methods (Fig. 5a and additional file 1: Fig. S5). Nevertheless, the overlap between all methods was the largest category for the mouse liver data (Fig. 5a). If we ignore p-value thresholds and consider the first 6000 genes detected rhythmic for each method, the overlap becomes stronger (Fig. 5d). We obtained similar results from the most informative dataset (Additional file 1: Figure S6b). Indeed, the rat lung dataset has 36 time-points spread over three 24h cycles (Fig. 4a). Thus, the same genes seem to be called rhythmic by all methods but the threshold of significance appear inconsistent. These results suggest an issue with the significance of p-value thresholds for the methods. With a smaller number of top rhythmic genes, the overlap between methods was weaker (Fig. 5c and 5d). Thus the methods agree on a large number of rhythmic genes, but not necessarily on the order of significance among them. Finally, for baboon liver data there was less overlap of methods (Fig. 5b; Additional file 1: Fig. S6b and S6c), which might be due to the low information in that data (Fig. 4a).

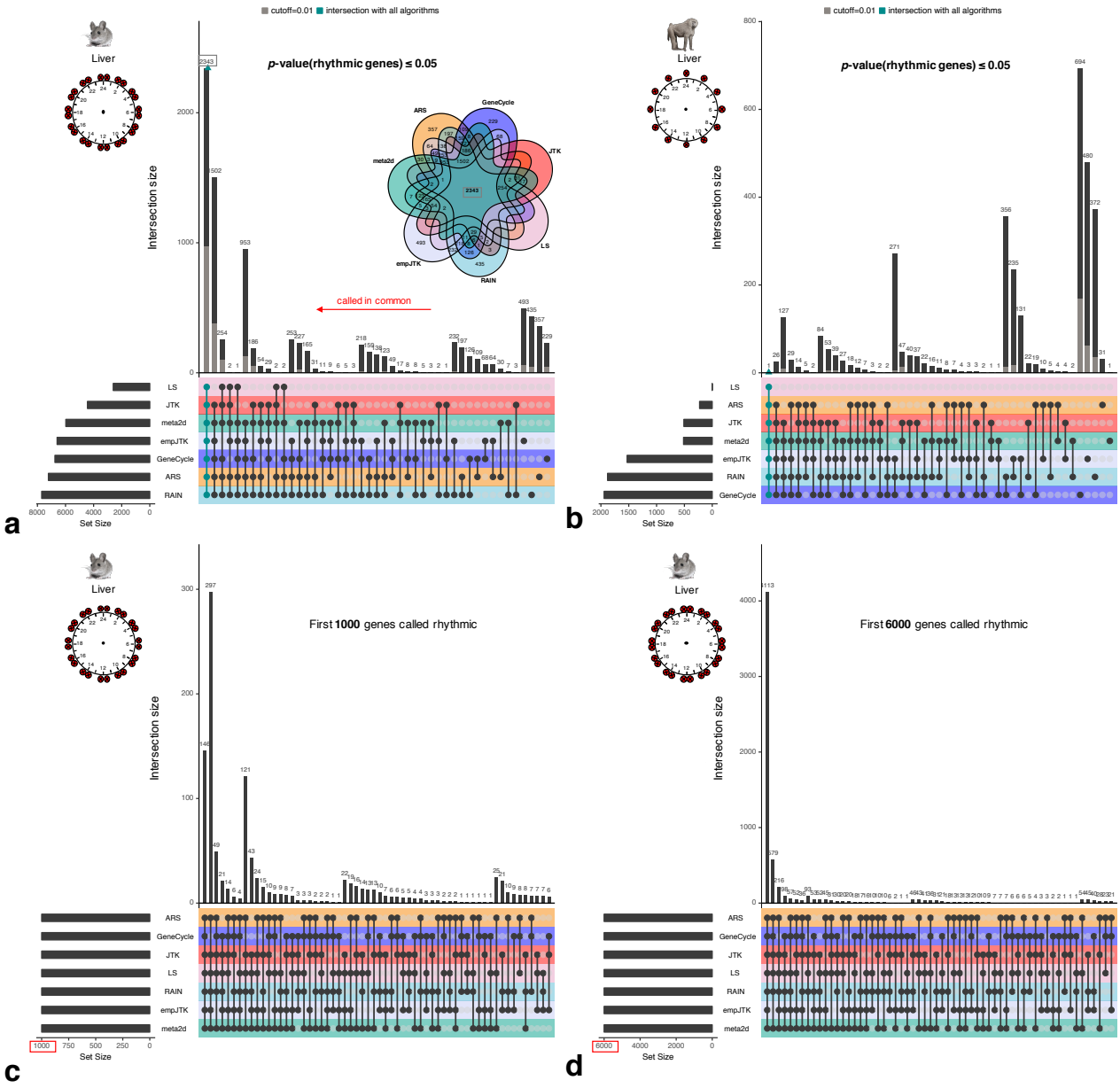

Figure 5. Methods detect the same first top rhythmic genes, but with inconsistencies in the meaning of their p-values.

Upset diagrams show the number of rhythmic genes called in common by the methods.

a)b) Upset diagram for mouse liver dataset (microarray) (**a**) and baboon liver dataset (**b**) for the p-value thresholds of 0.05 (black) or 0.01 (grey) for calling genes rhythmic. The Venn diagram (**a**) illustrates the upset diagram with, for instance, 2343 genes called rhythmic by all methods.

c)d) Upset diagram for mouse liver dataset (microarray) for the first 1000 (**c**) or 6000 (**d**) genes detected rhythmic for each method. With a smaller number of top rhythmic genes, the overlap between methods is weaker.

Use of evolutionary conservation as a benchmark

Signal of evolutionary conservation

We expect biologically relevant rhythmic activity of genes to be more conserved between species than putative false positives from detection methods. For each condition (species and tissue), we defined the group of genes whose orthologs are called rhythmic in the homologous tissue of another species (Fig. 6a). For example, starting with all mouse genes, we only kept mouse-zebrafish one-to-one orthologs. Considering the liver, these orthologs were separated into two groups: genes for which the ortholog is detected as rhythmic in zebrafish liver, called rhythmic orthologs; and the remaining one-to-one orthologs (Fig. 6b). To detect zebrafish liver rhythmic genes, we used the ARS method following the results presented above and fixed a p-value threshold of 0.01. Figure 6d shows the density distribution of p-values of rhythmic orthologs and non-rhythmic orthologs obtained for the seven methods applied to mouse liver data. Mouse-zebrafish orthologs, that are detected rhythmic in zebrafish liver, were significantly more enriched in small p-values in mouse liver, for all methods (Kolmogorov-Smirnov test p-values $<$ 0.001; Fig. 6d). Similar results were obtained using other methods and/or a different threshold to call orthologs as rhythmic in zebrafish liver (Additional file 1: Figure S7). This result obtained for distant species (Additional file 1: Fig. S1) shows that the conservation of rhythmicity is strong and should be informative. Similar results were obtained in other species comparisons (Additional file 1: Fig. S9), with a stronger signal for evolutionarily close species such as mouse and rat (Additional file 1: Fig. S8). Thus, for the same homologous organ, rhythmic orthologs have a stronger statistical signal of rhythmicity than non-rhythmic orthologs. We are going to use this evolutionary conservation of the rhythmicity of gene expression in order to compare the performance of methods.

Only strong rhythmic signals of gene expression are reliable

In this last part, we compared the performances of methods to detect the rhythmic orthologs. For a given dataset, the best method is expected to report rhythmic genes with the highest proportion of rhythmic orthologs. It should be noted that this does not imply that we expect all rhythmic behavior to be conserved between orthologs, but rather that true rhythmic genes should have more rhythmic orthologs than false-positive predictions. For a given p-value threshold, each method detects a certain number of rhythmic genes (genes with p-value threshold). At each threshold we calculated the proportion of orthologs rhythmic in species2 among one-to-one species1-species2 orthologs, as defined in Figure 7a. This proportion allows assessing how each method is able to detect the conservation of rhythmicity and can be calculated for each p-value threshold. The benchmark set is composed of orthologs detected rhythmic in the second species, called rhythmic orthologs. To define this set, we chose a rhythmicity detection method among ARS, empJTK, and GeneCycle, in agreement with results of previous sections, and a p-value threshold of 0.01.

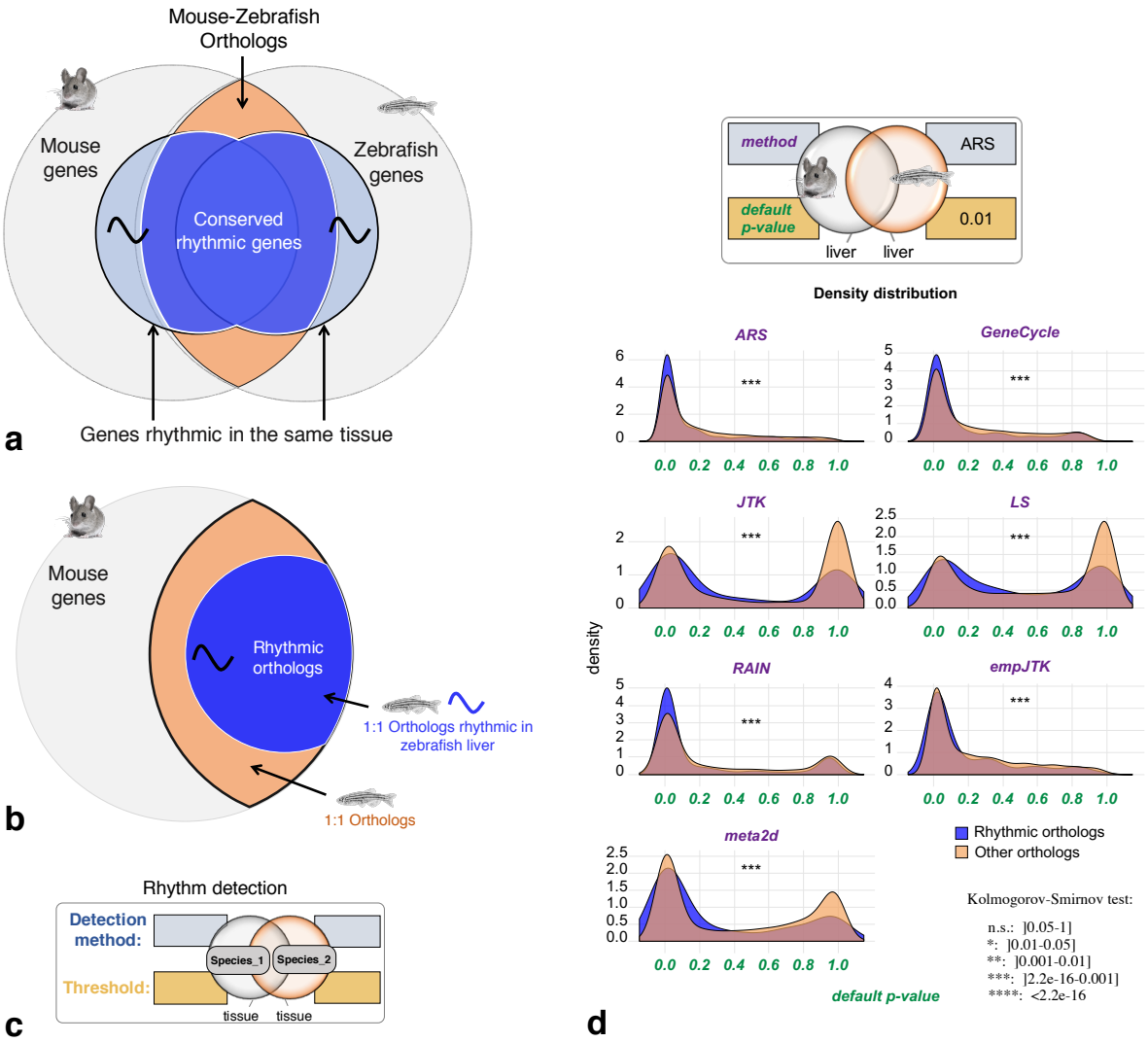

Figure 6. Signal of evolutionary conservation of rhythmic gene expression. Orthologous genes detected as rhythmic in the same organ of two species have a stronger statistical signal of rhythmicity than those not detected as rhythmic in at least one species.
a) Mouse and zebrafish share orthologous genes, some of which are rhythmic in the homologous tissues.
b) Method used for ortholog benchmarking, as in panel d: From all mouse genes, only mouse-zebrafish one-to-one orthologs are kept. Considering the liver, these orthologs are separated into two groups: genes for which the ortholog is detected as rhythmic in zebrafish liver, called rhythmic orthologs; and the remaining one-to-one orthologs.
c) Chart providing the legends to inform about the method and the threshold used to call genes rhythmic for each condition (species and tissue).
d) p-values density distribution of rhythmic orthologs vs non-rhythmic orthologs obtained for the seven methods applied to mouse liver data. Mouse-zebrafish orthologs, that are detected rhythmic in zebrafish liver, are significantly more enriched in small p-values in mouse liver, for all methods (Kolmogorov-Smirnov test p-values < 0.001.

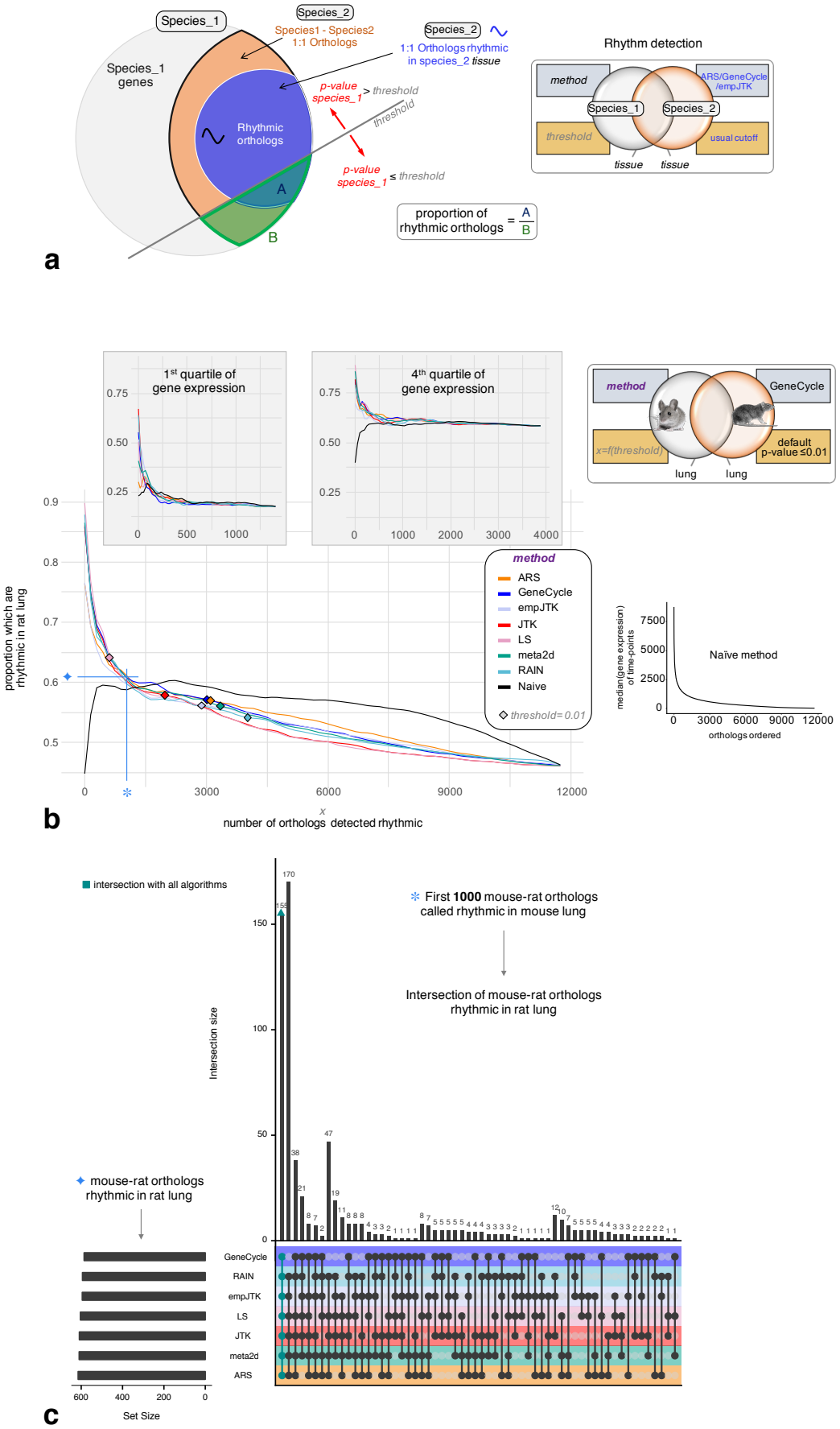
Species_1
Species_2
Species1 - Species2
1:1 Orthologs
Species_1
genes
Rhythmic
orthologs
A
B
Species_2
1:1 Orthologs rhythmic
in species_2 tissue
p-value
species_1
> threshold
threshold
p-value
species_1
≤ threshold
proportion of
rhythmic orthologs
= A/B
a
Rhythm detection
method
ARS/GeneCycle
/empJTK
Species_1
Species_2
threshold
usual cutoff
tissue
tissue
1st quartile of
gene expression
4th quartile of
gene expression
method
GeneCycle
x=f(threshold)
default
p-value ≤0.01
lung
lung
proportion which are
rhythmic in rat lung
number of orthologs detected rhythmic
x
method
ARS
GeneCycle
empJTK
JTK
LS
meta2d
RAIN
Naive
threshold= 0.01
median(gene expression)
of time-points
Naïve method
orthologs ordered
b
intersection with all algorithms
First 1000 mouse-rat orthologs
called rhythmic in mouse lung
Intersection of mouse-rat orthologs
rhythmic in rat lung
Intersection size
mouse-rat orthologs
rhythmic in rat lung
GeneCycle
RAIN
empJTK
LS
JTK
meta2d
ARS
Set Size
c

Figure 7. Only strong rhythmic signals of gene expression are reliable
Methods designed for rhythm detection in gene expression show an advantage only for the genes with a strong rhythmic signal. For a fixed number of top genes called rhythmic, all the methods, despite their design differences, retrieve approximately the same proportion of biologically functional rhythmic genes and the same genes themselves.
a) Method to obtain figure b: For a given p-value threshold, each method detects a certain number of rhythmic genes (genes with p-value threshold). At each threshold, we calculate the proportion of orthologs rhythmic in species2 (A) among one-to-one species1-species2 orthologs (B). The benchmark set is composed of one-to-one orthologs detected rhythmic in the second species (using method ARS, GeneCycle, or empJTK), called rhythmic orthologs.
b) Variation of the proportion rhythmic orthologs/all orthologs in mouse as a function of the number of mouse orthologs detected rhythmic, for each method applied to the mouse lung dataset. The benchmark gene set is composed of mouse-rat orthologs, detected rhythmic in rat lung by the GeneCycle method with default p-value ≤ 0.01. The black line is the Naive method which orders genes according to their median expression levels (median of time-points), from highest expressed to lowest expressed gene, then, for each gene, calculates the proportion of rhythmic orthologs among those with higher expression. The proportion of the benchmark set among one-to-one orthologs is higher for highly expressed genes (4th quartile) than for lowly expressed genes (1st quartile) ($\sim$60% vs $\sim$20% respectively). Diamonds correspond to a p-value threshold of 0.01.
c) Upset diagram showing the number of rhythmic orthologs (figure **a**) called in common by the methods among the first 1000 mouse-rat orthologs that are called rhythmic in mouse lung.

A risk is that orthologs have conservation of gene expression levels and that there is a bias towards calling highly expressed genes "rhythmic". To control for this in the benchmarking, we added a "Naive" method based only on expression levels. This Naive method simply orders genes (orthologs here) according to their median expression levels (median of time-points), from highest expressed to lowest expressed gene, then, for each gene, we calculated the proportion of rhythmic orthologs among those with higher expression. We also present results for subsets obtained from the division in four quartiles of expression levels. Figure 7b shows the variation of the proportion defined above as a function of the number of orthologs detected rhythmic, obtained for each method applied to the mouse lung dataset. The benchmark gene set was defined by mouse-rat orthologs, detected rhythmic in rat lung by the GeneCycle method (default p-value ≤ 0.01). Genes are given by order of their detection by the methods. The genes with small p-values, i.e. with a strong signal of rhythmicity, had a high proportion of rhythmic orthologs. Importantly, for all methods, this proportion was higher than that obtained from the Naive method (Fig. 7b). Results are consistent in almost all species comparisons, with exceptions for cerebral tissues (Additional file 3). However, the thresholds of 0.01 are to the right of the cross between the curves of rhythm detection methods and the Naive, except for LS. This means that, for an apparently reasonable threshold (p-value ≤ 0.01), ranking genes by expression level performed "better" than all methods designed specially for rhythm detection. These methods performed better only for genes with very high signal of rhythmicity. Finally, for the top 1000 mouse-rat orthologs detected as rhythmic in mouse lung, all the methods reported a similar proportion of rhythmic orthologs, around 62%, mainly highly expressed genes (fourth quartile of gene expression) (Fig. 7b). And the overlap between these orthologs was largely detected by all methods (Fig. 7c). Thus, for genes with a high signal of rhythmicity, all methods performed similarly to detect the tissue-specific conservation of gene expression rhythmicity. Similar results were obtained for other species comparisons (Additional file 4).

Discussion

The methods designed for rhythm detection in gene expression perform similarly and only for strong rhythmic signal.

In this study, we show that orthologous genes detected as rhythmic in the same organ of two species have a stronger statistical signal of rhythmicity than those detected as not-rhythmic in at least one species. These results support our hypothesis that the nycthemeral rhythmicity at the gene expression level is biologically functional, and that this functionality is more conserved between orthologous genes than between random genes. We define the nycthemeral transcriptome as all genes displaying a rhythmic expression repeated every 24 hours. In order to assess the performance of seven methods to detect these rhythms, we used this concept of conservation of the rhythmicity between species for benchmarking. We employed genes whose orthologs had a rhythmic expression called in the same homologous organ as a proxy for a true positive set, as done in some previous benchmarks. For instance, [31] assessed the quality of microarrays quality control methods based on evolutionary conservation of expression profiles, and [22] benchmarked tissue-specificity methods in the same way. This approach, based on real data, avoids relying on simulations which tend to favor methods using the same model, e.g. the same patterns, and has the advantage of not being based on specific assumptions, other than general evolutionary conservation of function. By taking into account the tissue-specificity of rhythmic gene expression and different species comparisons, we show that no method is strongly favoured. For instance, one would have expected that the added features of RAIN and empJTK allowing then to detect more diverse patterns than a classical sinusoidal, would have favored them. But this flexibility did not provide them any advantage in the benchmark. Furthermore, the comparison of the methods with a 'Naive' one, uninformed about rhythmicity, shows an advantage for informed methods only for the genes with a strong rhythmic signal. Thus, only genes with a strong rhythmic signal, i.e. the top genes called rhythmic, can be considered as relevant. Even if the threshold of "relevance" of these genes is dependant on the evolutionary distance of the species compared, these results suggest a call for caution about the results of previous studies reporting or based on large sets of rhythmic genes. For the same number of genes called rhythmic, all the methods, despite their design differences, retrieved approximately the same proportion of biologically functional rhythmic genes (Fig. 7b) and the same genes themselves (Fig. 7c).

The issue of significance

For the same p-value threshold, the number of genes called rhythmic is different from one method to another, with a large proportion of these genes detected rhythmic by only one or a few methods. But, if we consider the top genes called rhythmic for each method, without taking into account any p-value threshold, the overlap of rhythmic genes become strong between the methods (Fig. 5). This highlights an issue with the meaning of the p-value and the associated thresholds used. This is directly related to the issue of correction that needs to be improved in this field. When a smaller number of top rhythmic genes is used, the overlap between methods becomes weaker (Fig. 5c and 5d). Thus, the order of calling genes rhythmic is different from one method to another. Finally, since methods performed better than a Naive method only for genes with a strong rhythmic signal, we can not conclude for the relevance of the other genes called rhythmic, even when they have very low nominal p-values.

ARS, empJTK, and GeneCycle produce consistent p-values

ARS, empJTK, and GeneCycle were the methods that showed the best behavior on real and randomized data (single species tests). They were the only methods displaying both a uniform distribution of their p-values under the null hypothesis, and a left-skewed distribution when applied to real data. For empJTK, its default correction allowed to produce these expected results. However, each of these three methods is conceptually completely different, which indicates that there is not one conceptual framework which dominates rhythmic gene detection. ARS combines time-domain and frequency-domain analyses. GeneCycle, which is the robust spectrum function of the R package, is based on a robust spectral estimator which is incorporated into the hypothesis testing framework using a so-called g-statistic together with correction for multiple testing. And, empJTK improves the original JTK including additional reference waveforms in its rhythm detection. The other methods all presented major issues. LS has a right-skewed distribution of its initial uncorrected p-values suggesting an invalid null hypothesis. JTK, RAIN, and meta2d had also issues with their null hypothesis displaying left-skewed distributions of their uncorrected p-values. Their default adjustment was excessive, favoring high p-values obtained after correction.

Biological insight into gene rhythmicity in animal tissues

The separation of genes according to their expression levels showed differences in the p-values distributions, with more rhythmic signal among highly expressed genes, and almost none among lowly expressed genes. This led us to be concerned about the risk of a very high false-positive rate for highly expressed genes. However, the normalization of gene expression values (by Z-score, see above and methods) did not change the p-values distributions within highly expressed genes and particularly did not recover rhythmicity within lowly expressed genes (same results for microarray and RNAseq data). Thus, it is possible that the rhythmic genes are largely enriched in highly expressed genes. Experimental noise that would mask the rhythmic signal of lowly expressed genes could also explain this result in part, especially considering that the datasets with good sampling used microarray technology.

The observation of known cycling genes in different organs seems to indicate different profiles of rhythmicity possible for the same gene. For instance, Npas2 displays a cosine shape in kidney and lung, and a peak/box shape in liver and muscle (Fig. 3a). This observation suggests that methods might perform differently depending on the organ studied. This is also one of the reasons why all our analyses were made for homologous organs.

In mouse-baboon comparisons, there were no significant differences of p-value density between rhythmic and non-rhythmic orthologs in cerebral tissues: brain stem, cerebellum, and supra-chiasmatic nucleus, except for the hypothalamus (Additional file 3: Fig. S4-S5). This could be explained by the fact that there are only low amplitudes of expression of clock genes and few rhythmic genes in almost all brain regions. This is assumed to be due to an inefficient synchronization of individual cellular oscillators in brain cells to avoid noise into the synchronizator element [34]. In addition, it could also be an essential aspect for intrinsic brain processes which could require a constant expression of most genes.

The importance of having an informative dataset

Because of the cost and complexity of circadian experiments, time-series datasets of gene expression in animals are rare, especially in the same experimental conditions. Algorithms must be able to deal with little data, but importantly experiments should take into account the algorithms' sensitivity. All algorithms appeared to produce relatively poor p-values distributions when applied to the available Drosophila or baboon datasets, and, for the baboon dataset, were almost always less efficient than the Naive method (Additional file 1: Fig. S10). This baboon dataset is probably not very informative, which raises questions about the biological conclusions from the associated study [27]. With only one replicate per time-point, over only one cycle of 24 hours, the algorithms are unable to detect repetitive patterns. Variations over a single 24 hours cycle appear to be insufficient to detect rhythmic signal, when there is no evidence of repetition over several cycles. Moreover, each data comes from different outbred individuals. The variations of gene expression between two time-points can be due to individual variations or real oscillation within the population. It is possible that sinusoidal patterns with a continuous trend over successive time-points could be detected without replicates, although power will be lacking, but patterns such as the peak pattern will be extremely sensitive to inter-individual variation. Figure 4 generally suggests that datasets with one replicate per time-point over a unique cycle of 24 hours do not provide enough information that would allow to correctly detect the rhythmicity. It seems that ARS in peculiar is very sensitive to the repetitive nature of profiles. Thus, one cycle with several biological replicates should be favored. Our results support the conclusions of Hutchison et al. [19] who indicate that for a fixed number of samples, better sensitivity and specificity are achieved with higher numbers of replicates than with higher sampling density. We propose that future experiments should produce data with two biological replicates per time-points as a strict minimum. Obviously, we suggest considering biological replicates as new cycles within one replicate, as proposed in recent guidelines [17]. GeneCycle, and to a lesser extent empJTK, were the most robust methods when applied to weakly informative datasets. Thus, the performance of the algorithms is dependent on techniques and experimental designs used for the experiments. The optimization of experimental plans (see section Recommendations) could improve the methods' performance for the detection of rhythmically expressed genes. Finally, contrary to the mouse experiment, the rat experiment has been done under zeitgeber conditions which have most likely resulted in more genes being expressed rhythmically, so in proportion, more periodic patterns. This might explain the higher density of small p-values obtained for the rat dataset (Fig. 4a).

Limitations and improvement of methods

ARS and GeneCycle need complete chronological data and cannot deal with biological replicates. Except for LS and RAIN, all methods studied here assume equally spaced time-points. Furthermore, ARS needs an integer sampling interval with regular time-series datasets and cannot deals with missing values, or with several replicates per time-point. In this study, ARS appeared to be efficient only for the dataset with at least two cycles of data. Indeed it produced aberrant p-value distributions when applied to datasets restricted to one cycle of 24 hours. But, for these datasets, all algorithms behaved poorly. The improvement of JTK by empJTK produced much better results than the original JTK algorithm. We believe that LS could be a very interesting method if its null hypothesis could be clarified and would thus provide p-values with proper behavior. LS has advantages that other algorithms don't. For instance, it can deal with irregular intervals, missing data, and has been shown to stay efficient on small sample

size [44], which constitutes one of the big issues of circadian transcriptomic data. On the other hand, relative to JTK, ARS, or MICOP methods, LS has also been shown to be highly sensitive to the increasing of sampling intervals and to noise for proteomic data [20].

A good method must, at least, display a uniform distribution under the null hypothesis, and a classic skewed distribution when applied to full dataset or even more to known cycling genes. It should also be able to detect efficiently rhythmic orthologs, which represent an important part of the functionally relevant nycthemeral rhythmicity. In this study, we did not assess the amplitudes, phases, and precise period provided by the algorithms. We only analysed the performance of methods for nycthemeral or circadian rhythms in gene expression data, and cannot conclude directly for ultradian or seasonal rhythms, and for other types of datasets which are not gene expression data.

Recommendations

Experimental design

1. Always use at least 2 biological replicates per time-point.
2. One full period sampled is the minimum required.
3. Favor time-points number (small temporal resolution) over transcriptome profiling quality (e.g., Microarray vs RNAseq).
4. Favor regular sampling because only few algorithms can deal with irregular interval time-series.
5. For a fixed number of samples, favor higher numbers of replicates over higher sampling density (see also [19]).

Recommendations about the choice of rhythm detection method, the arrangement of the time-series dataset, and the interpretation of results based on these seven methods studied.

1. Only genes with a strong rhythmic signal should be considered as relevant. By "strong" we mean the top genes called rhythmic, knowing that the threshold of p-value ≤ 0.01 is already not stringent enough for some methods.
2. Take into account that detected rhythmic genes are strongly enriched in highly expressed genes.
3. LS could be a good candidate to improve.
4. Favor ARS, GeneCycle, or empJTK with default parameters.
5. Consider biological replicates as new cycles with one replicate.
6. Check by eye for rhythms of known circadian genes.
7. Never duplicate and concatenate data before running algorithms [17].
8. Never consider technical replicates as biological replicates [17].

Tables

-

JTK	description	R
raw p-value	No Bonferroni correction	-
default p-value	Bonferroni correction of raw p-values	p.adjust(raw.pvals, method="bonf")
BH.Q (this paper)	Benjamini-Hochberg correction of raw p-values	p.adjust(raw.pvals, method="BH")
BH.Q (software)	Benjamini-Hochberg correction of default p-values	p.adjust(default.pvals, method="BH")

Table 1. Raw, default, and BH.Q in JTK algorithm.

Species		*Papio anubis*	*Mus musculus*		*Rattus norvegicus*	*Danio rerio*	*Drosophila melanogaster*	*Aedes aegypti*	*Anopheles gambiae*
Transcriptome profiling technique		RNA-seq	microarray	RNA-seq	microarray	microarray	RNA-seq	microarray	microarray
Tissue	Aorta	★	★	★					
	Brain stem	★	★	★					
	Cerebellum	★	★	★					
	Heart	★	★	★			★		
	Liver	★	★	★		★			
	Lung	★	★	★	★				
	Muscle	★	★	★					
	SCN	★	★						
	White adipose	★	★	★					
	Head						★	★	★
	Body						★	★	
Experimental conditions	Total period length	24h	48h	48h	72h	48h	24h	48h	48h
	Exp. conditions	LD	DD	DD	LD	LD	LD	LD	LD
	Regular sampling each	2h	2h	6h	2h	4h	3h	4h	4h
	Number of biological replicates per time-point	x1	x3	x3	x1	x3	x1	x1	x2
References		Mure, 2018	Zhang, 2014		Sukumaran, 2011	Boyle, 2017	Gill, 2017	Leming, 2014	Rund, 2011

Additional Table 1. Gene expression time-series datasets.
Gene expression time-series datasets that come from circadian experiments. We kept data from "normal" individuals (apparently healthy, wild-type, no treatment) for these seven species, allowing comparisons among vertebrates and among insects. We preferred data from light-dark (LD) and ad-libitum experimental conditions whenever possible as providing a better representation of wild conditions. LD for regular alternation of light and darkness each 24h; and DD for continuous darkness after an entrainment to a 12h:12h light:dark schedule for one week.

Additional files

Additional file 1 — Supplementary results

Additional file 2 — Density distribution of raw and default p-values obtained for the seven rhythm detection methods applied to vertebrate datasets.

Density distribution of p-values obtained before (raw) and after the default correction (software) for the seven methods applied to each vertebrate dataset, sub-categorized in: i. randomized data which represents the null hypothesis; ii. randomized data restricted to the first and fourth quartiles of the median gene expression level, to check for the impact of expression level under the null; iii. the full original dataset; iv. the first and fourth quartiles of the median gene expression level of the original data; and v. a subset of known cycling genes when such data was available (8 to 99 genes according to species). The default p-values of ARS, GeneCycle, and LS are uncorrected.

Additional file 3 — Signal of evolutionary conservation of rhythmic gene expression in vertebrates.

p-values density distribution of rhythmic orthologs vs non-rhythmic orthologs obtained for the seven methods applied to different vertebrate datasets. Orthologous genes detected as rhythmic in the same organ of two species have a stronger statistical signal of rhythmicity than those detected as not-rhythmic in at least one species. From all species_1 genes, only species_1-species_2 one-to-one orthologs are kept. Considering homologous tissues, these orthologs are separated into two groups: genes for which the ortholog is detected as rhythmic in this tissue of species_2, called rhythmic orthologs; and the remaining one-to-one orthologs.

Additional file 4 — Variation of the proportion A/B as a function of the number of orthologs detected rhythmic, obtained for each method applied to different vertebrate datasets

The benchmark gene set is composed of species_1-species_2 orthologs, detected rhythmic in the homologous tissue of species_2 by the ARS, GeneCycle, or empJTK method with default p-value ≤ 0.01 or 0.05. See Figure 7 for definitions of sets A and B. The black line is the Naive method which orders genes according to their median expression levels (median of time-points), from highest expressed to lowest expressed gene, then, for each gene, calculates the proportion of rhythmic orthologs among those with higher expression.

Additional file 5 — Results in insects

Methods

Pre-processing

For each time-series dataset, only protein coding genes were kept. For microarrays, we removed probIDs which were assigned to several GeneIDs. Genes with no expression (= 0) at all time-points were also removed. For each species dataset, we only kept comparable conditions to other species of reference. Tissues separated in sub-tissues such as adrenal gland in adrenal cortex and adrenal medulla in baboon experiment were removed.
For each condition (species and tissue), several datasets have been built: i. the full original dataset; ii. the first and fourth quartiles of the median gene expression level of the original data; iii. randomized data (time-points redistributed randomly); iv. randomized data restricted to the first and fourth quartiles of the median gene expression level; and v. a subset of known cycling genes when such data was available (8 to 99 genes according to species).

Statistical analysis of rhythmic gene expression

All the rhythm detection methods (See Materials) were applied to each pre-processed dataset, producing a list of p-values as output. Then, for each gene having several results (ProbIDs or transcripts), we combined p-values by Brown's method using the EmpiricalBrownsMethod R package. Thus, for each dataset, we obtained a unique p-value per gene. Whenever the per-gene normalization was not necessary (unique data for all genes), we obtained the original p-value for each gene.

Normalization by Z-score

The normalization of gene expression values by Z-score transforms the pre-processed data such that for gene i with the original expression value at time-point j is $gene.ij$, we have:

$$gene.ij.normalized = gene.ij - xi$$

with $xi = mi$ - $\frac{Zi}{j}$. mi is the mean expression of gene i: $mi = \frac{\sum gene.ij}{j}$; and $Zi = \frac{mi-m}{sd}$; m and sd being the mean and the standard deviation of the original full dataset.

Orthology relationships

For each species comparison, orthologs relationships have been downloaded from OMA [3]. For simplicity, we only considered one-to-one orthologs. In species comparisons, we only kept orthologous genes that had available data in both species.

Naive method

The Naive method is only based on expression levels of genes and is not informed about rhythm detection. It simply orders genes according to their median expression levels (median of time-points), from highest expressed to lowest expressed gene. Then, for

each gene i, we calculate the proportion of rhythmic orthologs among those with higher expression, i.e. among the genes from the highest expressed one to the gene i.

Materials

Datasets

Mus musculus (13 tissues)

Raw microarray and RNA-seq data, from [43], was downloaded from GEO accession (GSE54652). Microarray gc-rma normalized data was sent by Katharina Hayer from CircaDB database [30]. Expression values were already normalized between biological replicates to average out both biological variance between individual animals and technical variance between individual dissections. RNA-seq data was already normalized using DESeq2. Data was obtained for adrenal gland, aorta, brain stem, brown adipose, cerebellum, heart, hypothalamus, kidney, liver, lung, muscle, SCN (only microarray), and white adipose. Probesets on the Affymetrix MoGene-1.0-ST-V1 array were cross-referenced to best-matching gene symbols by using Ensembl BioMart software.

Papio anubis [Olive baboon] (11 tissues used)

RNA-seq data from [27] was downloaded already normalized by using DESeq2. Read counts per gene were calculated using FeatureCounts. We kept data for aorta, brain stem, cerebellum, heart, hypothalamus, kidney, liver, lung, muscle, SCN, and white adipose tissues. Data were already provided with Ensembl gene symbols.

Rattus norvegicus (lung)

Raw microarray data from [36] was downloaded from GEO accession (GSE25612). Over 3 days, 54 samples were extracted in light-dark condition with a temporal resolution closer for some time-points (See paper for more details). Contrary to the study, we still considered the 3 successive days samples as successive days measurements. ARS, JTK and RAIN methods don't operate with irregular time-series. We normalized time-series by calculating the mean value of irregular time-points to obtain regular time-series. rma normalization was performed using the rma R-package. Probesets on the Affymetrix 230-2-probe array were cross-referenced to best-matching gene symbols by using Ensembl BioMart software.

Dano rerio (liver)

Raw microarray data from [5] was downloaded from GEO accession (GSE87659). Data was already rma-normalized, averaged gene-level signal intensity, and already cross-referenced to best-matching transcript symbols.

Anopheles gambiae (head and body)

Raw microarray data from [32] was downloaded from GEO accession (GSE22585). Non-blood fed female mosquito heads and bodies were collected under light dark and constant

dark conditions. We only used data collected in LD condition. We normalized data using the rma R package and cross-referenced to best-matching gene symbols by using VectorBase software.

Aedes aegypti (head)

Raw microarray data from Ľeming was downloaded from GEO accession (GSE60496). Non-blood fed female mosquito heads were collected under light dark and constant dark conditions. We only used data collected in LD condition. NimbleGen Aedes aegypti 12plex array already rma normalized were provided with VectorBase geneIDs.

Drosophila melanogaster (head and body)

RNA-seq data from [6] was downloaded from GEO accession (GSE64108). They measured RNA concentrations in the head and body of 3-, 5-, and 7-week-old adult flies in ad libidum feeding or 12-hour time-restricted feeding conditions. We only used data from ad libidum feeding condition of 5-week-old adult flies with best temporal resolution.

Cross-referenced gene IDs and known cycling genes

GeneID and protein coding status were downloaded from VectorBase [16] using BioMart. Known cycling genes were obtained from the KEGG database [29] or [38] as:
- KEGG circadian entrainment entry pathway for the mouse (mmu04713) and the rat (rno04713). This is the pathway by which light activates SCN neurons and the resulting signaling cascade that leads to a phase resetting of the circadian rhythm generated in these neurons. Most of these genes are not involved in generating the rhythm itself and as such cannot be called 'clock genes'.
- KEGG circadian rhythm entry pathway for the baboon (human hsa04710), and Anopheles (aga04711)
- FlyBase circadian rhythm entry pathway for Drosophila (GO:0007623)

Algorithms and packages

`MetaCycle` R package was performed with parameters: minper=20h and maxper=28h. This package incorporates the 3 algorithms to detect rhythmic signals from time-series datasets: ARSER (ARS), JTK_CYCLE (JTK), and Lomb-Scargle (LS). It also provides meta2d that integrates analysis results from multiple methods based on an implementation strategy (see "Introduction to implementation steps of MetaCycle" in MetaCycle documentation for more details). ARS does not deal with several replicates per time-point. To not introduce biases, we only kept one replicate for ARS performing when the dataset was provided with several replicates per time-point. `Rain` R package was performed with parameters: period=24h, period.delta=4h (width of period interval), and method='independent'. In order to obtain unadjusted pvalues as output, we modified the source code of the rain and MetaCycle R packages.

`Empirical-JTK` (empJTK) was executed by running bash commands with parameters: cosine waveform, 24 hours' period, look for phases every 2 hour from 0 to 22 hours and look for asymmetries every 2 hour from 2 to 22 hours (GitHub alanlhutchison/empirical-JTK_CYCLE-with-asymmetry). It is important to run empJTK with python version

2.7.11. Raw p-value correspond to `P` output (P-value corresponding to Tau, uncorrected for multiple hypothesis testing), and default p-value correspond to `empP` output (min(p-value calculated from empirical null distribution, Bonferroni)).

`GeneCycle` R package [1] was downloaded from CRAN. We used the robust.spectrum function developped by [2] that computes a robust rank-based estimate of the periodogram/correlogram.

Plots have been created using ggplot2 R library (version 3.1.0); Upset diagrams using UpSetR R package (version 1.3.3) [23]; and Venn diagram using venn R package (version 1.7).

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Paul Franken for useful discussions, as well as all members of the Robinson-Rechavi lab.

Ethical considerations

We had ethical issues to use olive baboon data since these data needed the sacrifice of twelve baboons. We would like to remind that such data would have been impossible to get in Switzerland where the primate research is prohibited. We still support Switzerland ethical considerations in matter of animal research and think that the scientific knowledge can not justify an irresponsible employment of life on earth. While being aware that our results would have been less robust without these data and that these considerations on primate could also be generalized to other living organisms.

References

1. M. Ahdesmaki, K. Fokianos, and K. Strimmer. *GeneCycle: Identification of Periodically Expressed Genes*, 2012. R package version 1.1.4.
2. M. Ahdesmäki, H. Lähdesmäki, R. Pearson, H. Huttunen, and O. Yli-Harja. Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, 6(1):117, May 2005.
3. A. M. Altenhoff, G. H. Gonnet, C.-M. Train, D. Dylus, N. M. Glover, T. M. de Farias, C. Dessimoz, A. Warwick Vesztrocy, K. Zile, C. Stevenson, K. Kaleb, H. Redestig, and J. Long. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Research*, 46(D1):D477–D485, 11 2017.
4. C. E. Boothroyd, H. Wijnen, F. Naef, L. Saez, and M. W. Young. Integration of light and temperature in the regulation of circadian gene expression in drosophila. *PLOS Genetics*, 3:1–16, 04 2007.
5. G. Boyle, K. Richter, H. D. Priest, D. Traver, T. C. Mockler, J. T. Chang, S. A. Kay, and G. Breton. Comparative analysis of vertebrate diurnal/circadian transcriptomes. *PLOS ONE*, 12(1):1–18, 01 2017.

6. B. Christopher, S. Gill, G. Melkani, and S. Panda. Transcriptomic changes in drosophila tissues under time-restricted feeding, 2015. GSE64108.

7. D. Chudova, A. Ihler, K. Lin, B. Andersen, and P. Smyth. Bayesian detection of non-sinusoidal periodic patterns in circadian expression data. *Bioinformatics (Oxford, England)*, 25:3114–20, 09 2009.

8. D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko. Lipschitz functions have lp-stable persistence. *Foundations of Computational Mathematics*, 10(2):127–139, Apr 2010.

9. U. de Lichtenberg, R. Wernersson, T. S. Jensen, H. B. Nielsen, A. Fausbøll, P. Schmidt, F. B. Hansen, S. Knudsen, and S. Brunak. New weakly expressed cell cycle-regulated genes in yeast. *Yeast*, 22(15):1191–1201, 2005.

10. A. Deckard, R. C Anafi, J. Hogenesch, S. Haase, and J. Harer. Design and analysis of large-scale biological rhythm studies: A comparison of algorithms for detecting periodic signals in biological data. *Bioinformatics (Oxford, England)*, 29, 09 2013.

11. M.-L. Dequéant, S. Ahnert, H. Edelsbrunner, T. M. A. Fink, E. F. Glynn, G. Hattem, A. Kudlicki, Y. Mileyko, J. Morton, A. R. Mushegian, L. Pachter, M. Rowicka, A. Shiu, B. Sturmfels, and O. Pourquié. Comparison of pattern detection methods in microarray time series of the segmentation clock. *PLOS ONE*, 3(8):1–9, 08 2008.

12. K. Fokianos, K. Strimmer, and S. Wichert. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20, 01 2004.

13. T. Gabaldón and E. V. Koonin. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14:360 EP –, 04 2013.

14. A. Gerber, C. Saini, T. Curie, Y. Emmenegger, G. Rando, P. Gosselin, I. Gotic, P. Gos, P. Franken, and U. Schibler. The systemic control of circadian gene expression. *Diabetes, Obesity and Metabolism*, 17(S1):23–32, 2015.

15. Z. Gerhart-Hines and M. A. Lazar. Circadian Metabolism in the Light of Evolution. *Endocrine Reviews*, 36(3):289–304, 06 2015.

16. G. I. Giraldo-Calderón, S. J. Emrich, R. M. MacCallum, G. Maslen, E. Dialynas, P. Topalis, N. Ho, S. Gesing, the VectorBase Consortium, G. Madey, F. H. Collins, and D. Lawson. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Research*, 43(D1):D707–D713, 12 2014.

17. M. E. Hughes, K. C. Abruzzi, R. Allada, R. Anafi, A. B. Arpat, G. Asher, P. Baldi, C. de Bekker, D. Bell-Pedersen, J. Blau, S. Brown, M. F. Ceriani, Z. Chen, J. C. Chiu, J. Cox, A. M. Crowell, J. P. DeBruyne, D.-J. Dijk, L. DiTacchio, F. J. Doyle, G. E. Duffield, J. C. Dunlap, K. Eckel-Mahan, K. A. Esser, G. A. FitzGerald, D. B. Forger, L. J. Francey, Y.-H. Fu, F. Gachon, D. Gatfield, P. de Goede, S. S. Golden, C. Green, J. Harer, S. Harmer, J. Haspel, M. H. Hastings, H. Herzel, E. D. Herzog, C. Hoffmann, C. Hong, J. J. Hughey, J. M. Hurley, H. O. de la Iglesia, C. Johnson, S. A. Kay, N. Koike, K. Kornacker, A. Kramer, K. Lamia, T. Leise, S. A. Lewis, J. Li, X. Li, A. C. Liu, J. J. Loros, T. A. Martino, J. S. Menet, M. Merrow, A. J. Millar, T. Mockler, F. Naef, E. Nagoshi, M. N. Nitabach, M. Olmedo, D. A. Nusinow, L. J. Ptáček, D. Rand, A. B. Reddy, M. S. Robles, T. Roenneberg, M. Rosbash, M. D. Ruben, S. S. Rund, A. Sancar, P. Sassone-Corsi, A. Sehgal, S. Sherrill-Mix, D. J. Skene, K.-F. Storch, J. S. Takahashi, H. R. Ueda, H. Wang, C. Weitz, P. O. Westermark, H. Wijnen, Y. Xu, G. Wu, S.-H. Yoo, M. Young, E. E. Zhang, T. Zielinski, and J. B. Hogenesch. Guidelines for genome-scale analysis of biological rhythms. *Journal of Biological Rhythms*, 32(5):380–393, 2017. PMID: 29098954.

18. M. E. Hughes, J. B. Hogenesch, and K. Kornacker. Jtk_cycle: An efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *Journal of Biological Rhythms*, 25(5):372–380, 2010. PMID: 20876817.

19. A. L. Hutchison, M. Maienschein-Cline, A. H. Chiang, S. M. A. Tabei, H. Gudjonson, N. Bahroos, R. Allada, and A. R. Dinner. Improved statistical methods enable greater sensitivity in rhythm detection for genome-wide data. *PLOS Computational Biology*, 11(3):1–29, 03 2015.

20. H. Iuchi, M. Sugimoto, and M. Tomita. Micop: Maximal information coefficient-based oscillation prediction to detect biological rhythms in proteomics data. *BMC Bioinformatics*, 19(1):249, Jun 2018.

21. A. Korenčič, G. Bordyugov, R. Košir, D. Rozman, M. Goličnik, and H. Herzel. The interplay of cis-regulatory elements rules circadian rhythms in mouse liver. *PLOS ONE*, 7(11):1–13, 11 2012.

22. N. Kryuchkova-Mostacci and M. Robinson-Rechavi. A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, 18(2):205–214, 02 2016.

23. A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, Dec 2014.

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

25. T. P. Michael, T. C. Mockler, G. Breton, C. McEntee, A. Byer, J. D. Trout, S. P. Hazen, R. Shen, H. D. Priest, C. M. Sullivan, S. A. Givan, M. Yanovsky, F. Hong, S. A. Kay, and J. Chory. Network discovery pipeline elucidates conserved time-of-day–specific cis-regulatory modules. *PLOS Genetics*, 4(2):1–17, 02 2008.

26. B. H. Miller, E. L. McDearmon, S. Panda, K. R. Hayes, J. Zhang, J. L. Andrews, M. P. Antoch, J. R. Walker, K. A. Esser, J. B. Hogenesch, and J. S. Takahashi. Circadian and clock-controlled regulation of the mouse transcriptome and cell proliferation. *Proceedings of the National Academy of Sciences*, 104(9):3342–3347, 2007.

27. L. S. Mure, H. D. Le, G. Benegiamo, M. W. Chang, L. Rios, N. Jillani, M. Ngotho, T. Kariuki, O. Dkhissi-Benyahya, H. M. Cooper, and S. Panda. Diurnal transcriptome atlas of a primate across major neural and peripheral tissues. *Science*, 359(6381), 2018.

28. E. Nagoshi, C. Saini, C. Bauer, T. Laroche, F. Naef, and U. Schibler. Circadian gene expression in individual fibroblasts: Cell-autonomous and self-sustained oscillators pass time to daughter cells. *Cell*, 119(5):693 – 705, 2004.

29. H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34, 01 1999.

30. A. Pizarro, K. Hayer, N. F Lahens, and J. Hogenesch. Circadb: A database of mammalian circadian gene expression profiles. *Nucleic acids research*, 41, 11 2012.

31. M. Rosikiewicz and M. Robinson-Rechavi. IQRray, a new method for Affymetrix microarray quality control, and the homologous organ conservation score, a new benchmark method for quality control metrics. *Bioinformatics*, 30(10):1392–1399, 01 2014.

32. S. S. C. Rund, T. Y. Hou, S. M. Ward, F. H. Collins, and G. E. Duffield. Genome-wide profiling of diel and circadian gene expression in the malaria vector anopheles gambiae. *Proceedings of the National Academy of Sciences*, 108(32):E421–E430, 2011.

33. J. Scargle. Studies in astronomical time series analysis. ii - statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263, 01 1983.

34. U. Schibler. The daily rhythms of genes, cells and organs. *EMBO reports*, 6(S1):S9–S13, 2005.

35. M. Straume. Dna microarray time series analysis: Automated statistical assessment of circadian rhythms in gene expression patterning. In *Numerical Computer Methods, Part D*, volume 383 of *Methods in Enzymology*, pages 149 – 166. Academic Press, 2004.

36. S. Sukumaran, W. J. Jusko, D. C. DuBois, and R. R. Almon. Light-dark oscillations in the lung transcriptome: implications for lung homeostasis, repair, metabolism, disease, and drug action. *Journal of Applied Physiology*, 110(6):1732–1747, 2011. PMID: 21436464.

37. P. F. Thaben and P. O. Westermark. Detecting rhythms in time series with rain. *Journal of Biological Rhythms*, 29(6):391–400, 2014. PMID: 25326247.

38. the FlyBase Consortium, J. Thurmond, J. L. Goodman, T. C. Kaufman, V. B. Strelets, B. R. Calvi, G. Millburn, G. Antonazzo, H. Attrill, S. J. Marygold, V. Trovisco, B. B. Matthews, and L. S. Gramates. FlyBase 2.0: the next generation. *Nucleic Acids Research*, 47(D1):D759–D765, 10 2018.

39. G. Wu, J. B. Hogenesch, R. C. Anafi, M. E. Hughes, and K. Kornacker. MetaCycle: an integrated R package to evaluate periodicity in large scale data. *Bioinformatics*, 32(21):3351–3353, 07 2016.

40. G. Wu, J. Zhu, J. Yu, L. Zhou, J. Z. Huang, and Z. Zhang. Evaluation of five methods for genome-wide circadian gene identification. *Journal of Biological Rhythms*, 29(4):231–242, 2014. PMID: 25238853.

41. R. Yang and Z. Su. Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics*, 26(12):i168–i174, 06 2010.

42. S.-H. Yoo, S. Yamazaki, P. L. Lowrey, K. Shimomura, C. H. Ko, E. D. Buhr, S. M. Siepka, H.-K. Hong, W. J. Oh, O. J. Yoo, M. Menaker, and J. S. Takahashi. Period2::luciferase real-time reporting of circadian dynamics reveals persistent circadian oscillations in mouse peripheral tissues. *Proceedings of the National Academy of Sciences*, 101(15):5339–5346, 2004.

43. R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proceedings of the National Academy of Sciences*, 111(45):16219–16224, 2014.

44. W. Zhao, K. Agyepong, E. Serpedin, and E. R. Dougherty. Detecting periodic genes from irregularly sampled gene expressions: A comparison study. *EURASIP Journal on Bioinformatics and Systems Biology*, 2008(1):769293, Jun 2008.