The FEBS Journal

FEBS PRESS science publishing by scientists

VIEWPOINT

Venn diagram analysis overestimates the extent of circadian rhythm reprogramming

Anne Pelikan[1], Hanspeter Herzel[2], Achim Kramer[3] and Bharath Ananthasubramaniam[1,2]

1 Institute for Theoretical Biology, Humboldt Universität zu Berlin, Germany
2 Institute for Theoretical Biology, Charité Universitätsmedizin Berlin, Germany
3 Institute for Medical Immunology, Charité Universitätsmedizin Berlin, Germany

Keywords
differential rhythmicity; high-throughput data; metabolism; reproducibility; statistical models

Correspondence
B. Ananthasubramaniam, Institute for Theoretical Biology, Humboldt Universität zu Berlin, 10115 Berlin, Germany
Tel: +49 30 2093 98410
E-mail: bharath.ananthasubramaniam@hu-berlin.de

(Received 29 January 2021, revised 18 May 2021, accepted 28 June 2021)

doi:10.1111/febs.16095

The circadian clock modulates key physiological processes in many organisms. This widespread role of circadian rhythms is typically characterized at the molecular level by profiling the transcriptome at multiple time points. Subsequent analysis identifies transcripts with altered rhythms between control and perturbed conditions, that is, are differentially rhythmic (DiffR). Commonly, Venn diagram analysis (VDA) compares lists of rhythmic transcripts to catalog transcripts with rhythms in both conditions, or that have gained or lost rhythms. However, unavoidable errors in rhythmicity detection propagate to the final DiffR classification resulting in overestimated DiffR. We show using artificial experiments on biological data that VDA indeed produces excessive *false* DiffR hits both in the presence and absence of *true* DiffR transcripts. We review and benchmark hypothesis testing and model selection approaches that instead compare circadian amplitude and phase of transcripts between the two conditions. These methods identify transcripts that 'gain', 'lose', 'change', or have the 'same' rhythms; the third category is missed by VDA. We reanalyzed three studies on the interplay between metabolism and the clock in the mouse liver that used VDA. We found not only fewer DiffR transcripts than originally reported, but VDA overlooked many relevant DiffR transcripts. Our analyses confirmed some and contradicted other conclusions in the original studies and also generated novel insights. Our conclusions equally apply to circadian studies using other omics technologies. We believe that avoiding Venn diagrams and using our convenient R-package COMPARERHYTHMS will improve the reliability of analyses in chronobiology.

Introduction

Circadian or near-24 h rhythms are present in all kingdoms of life [1]. These rhythms regulate critical physiological processes in many species [2]. In eukaryotes, a gene-regulatory feedback network involving a small group of *clock genes* generates cell-autonomous circadian rhythms. Transcription factors among the clock genes subsequently drive transcript rhythms in target *clock-controlled genes* (CCGs) [3]. Effects of genetic or environmental perturbations on CCGs and their outputs provided insights into the widespread role of circadian rhythms at the molecular level [4]. Transcripts, which are the proximal clock output, are easily

Abbreviations
AIC, Akaike information criterion; BIC, Bayesian information criterion; CCG, clock-controlled genes; FDR, false discovery rate; HFD, high-fat diet; KD, ketogenic diet; VDA, Venn diagram analysis.

quantified using high-throughput techniques (microarray and bulk RNA sequencing). Therefore, almost all studies focused on the effects of altered *Zeitgebers*, such as light regime and feeding, or genotype on the circadian transcriptome.

In experiments of this kind, one or more periods of the rhythm are sampled at regular intervals under the two conditions of interest. The samples themselves might consist of pools of individuals or biological replicates. The datasets obtained are subjected to statistical analyses to identify transcripts that are *differentially rhythmic* (DiffR) between the two conditions.

DiffR transcripts are commonly identified using Venn diagram analysis (VDA), as we term it: A list of rhythmic transcripts is compiled under each condition using one of many popular methods (JTKcycle [5], RAIN [6], harmonic regression/cosinor [7]). The two lists are compared for overlaps and differences, and the results are visualized using Venn diagrams.

This approach is inappropriate for two reasons. First, VDA seemingly finds DiffR features that are rhythmic in one condition and arrhythmic in the other. This is, of course, not all we want to know. For example, VDA overlooks transcripts that remain rhythmic but have altered circadian parameters (amplitude, phase). Second, the analysis even fails to accurately find transcripts that are rhythmic in one condition but not the other. One test of rhythmicity in each of the two conditions is necessary in VDA, but any statistical test for rhythmicity is inherently imperfect.

Statistical tests make two kinds of errors in classifying transcripts as rhythmic or arrhythmic; false positives (arrhythmic transcript classified as rhythmic) and false negatives (rhythmic transcript classified as arrhythmic). The corresponding correct classifications are true positives and true negatives. These errors can result in non-DiffR transcripts incorrectly tagged as *hits* in the DiffR analysis and vice versa. For example, a true DiffR transcript that is a false positive in one condition and true positive in the other will be considered a DiffR miss by VDA. Similarly, a transcript that is true negative and false positive in the two datasets will be considered a DiffR hit, when it is not. ('Hit' and 'miss' refer to the algorithm's prediction of DiffR transcripts).

Any statistical test involves trade-offs between the number of false positives and false negatives. Consequently, no choice of threshold (on the *P*-value or test statistic) will alleviate this misclassification. Moreover, in standard situations, false positives are stringently controlled, while false negatives are tolerated. Thus, many false negatives can be expected in both conditions. Even if some of these false negatives are true positives in the other condition, VDA results in overestimating the 'reprogramming'. Our conclusions based on transcriptomic data also hold true for other high-throughput datasets (proteomics, metabolomics) measured under two conditions. This paper examines whether VDA overestimates the number and identity of DiffR features in circadian studies and whether the misclassification of DiffR features affects the interpretation of those studies.

We illustrate using artificial scenarios constructed from real data that the VDA does indeed perform poorly and produces too many false DiffR hits. Next, we present the two different approaches to directly compare rhythms between the two conditions and identify the four categories (and not just the three categories depicted in a Venn diagram) of pertinent rhythmic transcripts. We provide complete pipelines of the available approaches in an easy-to-use R-package COMPARERHYTHMS. We reevaluate the number and identity of DiffR transcripts in three public circadian transcriptomic datasets that used VDA and compare and contrast our interpretation with theirs. We found that the extent of 'remodeling' is indeed much smaller than suggested in the original studies, and despite this overestimation, the VDA analyses overlooked several DiffR transcripts that our analysis identified. This discrepancy altered some conclusions, but confirmed others and our analysis often generated novel insights.

Results

VDA overestimates the number of DiffR features

We illustrate the shortcomings of VDA using two artificial scenarios constructed from the circadian transcriptome of the mouse liver. We use these same artificial scenarios to benchmark the approaches in our R-package COMPARERHYTHMS and compare them to VDA in the next section. Mouse liver transcripts were quantified every hour for 48 h in ref. [8].

First scenario

We compared two datasets comprising the odd and even time points (Fig. 1A). Without any measurement or experimental noise, the odd and even time points would follow the same rhythmic pattern and we expect no DiffR transcripts under these conditions. However, ever-present noise could occasionally cause the two sets of time points to differ resulting in false DiffR hits. We can thus benchmark using this scenario with no expected DiffR transcripts whether an approach limits

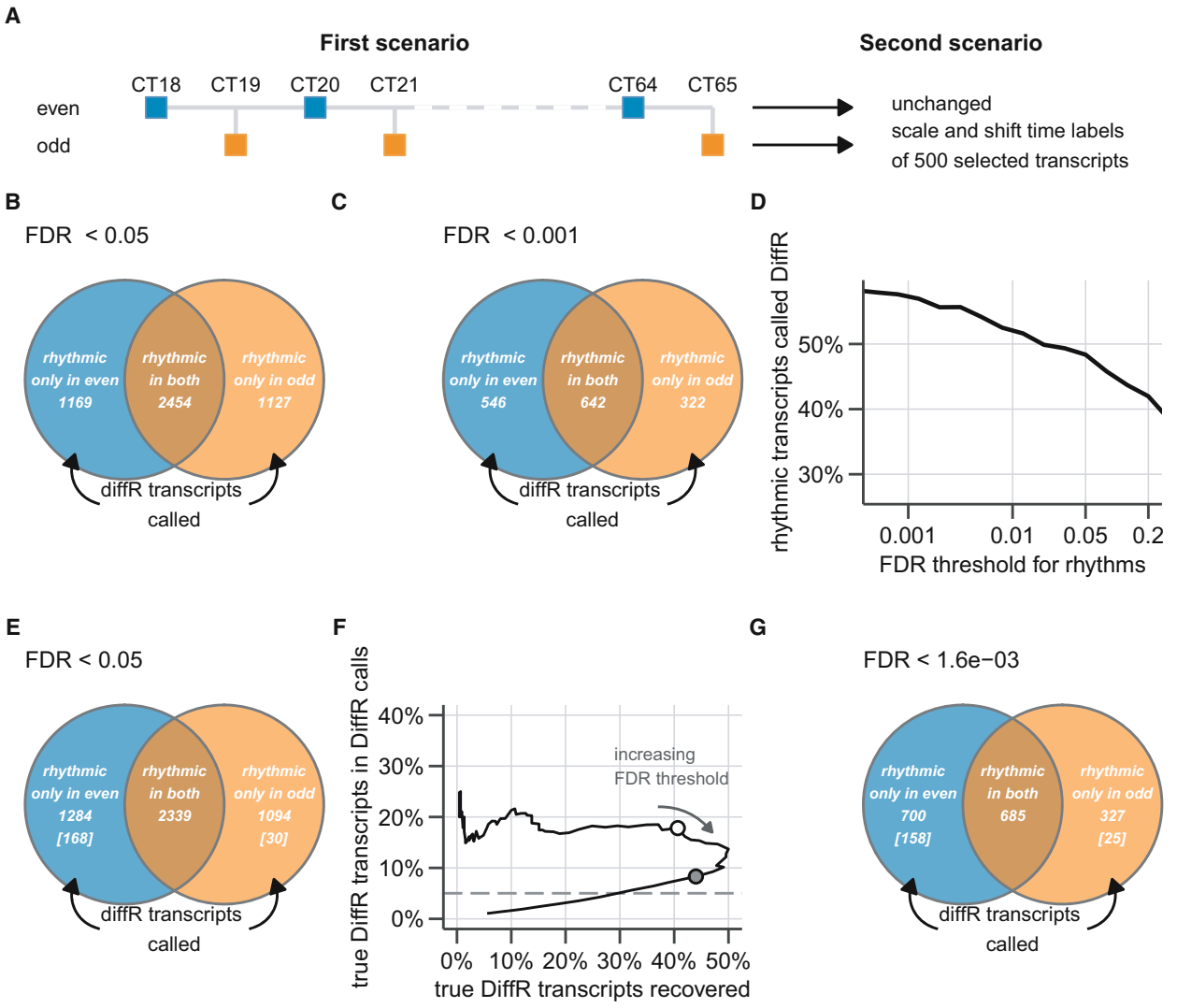


Fig. 1. Venn diagram analysis (VDA) of two artificial circadian studies created from Hughes *et al.* [8] data. (A) Construction of two scenarios from the high-resolution time series of Hughes *et al.* [8]. VDA applied to the first scenario comparing odd and even time points for two different false discovery rate (FDR) thresholds: (B) 0.05 and (C) 0.001. (D) The fraction of rhythmic transcripts incorrectly identified as DiffR for different FDR thresholds under the first scenario. (E) The result of the VDA of the second scenario, where a known set of 450 transcripts is truly DiffR with changes in both amplitude and phase. (F) Precision-recall curve of the overall performance of VDA under the second scenario. The circles are two possible operating points (FDR threshold = 1.6×10^{-3} (white fill), 0.05 (gray fill)). (G) VDA results for the best precision-recall performance point (open circle in (F)). The number of true DiffR transcripts identified in each group is given within square brackets.

the number of false DiffR hits and at what level. VDA with 0.05 false discovery rate (FDR) threshold called 2296 DiffR hits or 48% of all rhythmic transcripts (Fig. 1B). With a more stringent FDR threshold for rhythmicity detection, fewer DiffR hits were called at the cost of fewer detected rhythmic transcripts (Fig. 1C). In fact, more than 40% of rhythmic transcripts were incorrectly called DiffR hits across a range of thresholds (Fig. 1D).

Second scenario

We next created a scenario where a known set of transcripts had altered amplitudes and/or phases of rhythms between the odd and even time points, that is, were truly DiffR (Fig. 1A). To that end, we altered the dataset of odd time points from the first scenario for an initial set of 500 rhythmic transcripts. We randomly altered the amplitude and phase of these

transcripts and also added noise (Fig. S1A). Only 450 of the initial set of altered transcripts were rhythmic in either the odd or even time points. Transcripts rhythmic in both even and odd time points had differences in amplitude and/or phases (Fig. S1B). VDA called 2378 DiffR hits among 4717 rhythmic transcripts (or 50%) (Fig. 1E). Only a small fraction of hits (8.3%) were true DiffR transcripts. If we chose a random set of rhythmic transcripts and called them DiffR, we expect on average about $500/4717 \approx 11\%$ to be true DiffR transcripts in the second scenario. Thus, this random approach would outperform VDA with a standard choice of FDR threshold.

Precision-recall curves characterize the overall performance of DiffR classification. Precision-recall curves are particularly informative when true DiffR transcripts make up only a small fraction of all transcripts in the data [9], which we expect is this case. Precision is the fraction of DiffR calls that are correct. Recall is the fraction of true DiffR transcripts recovered. We desire DiffR detection with *both* high precision and high recall, or in other words, that is trustworthy and thorough, respectively.

Precision-recall performance of VDA is poor for all choices of FDR threshold. This performance metric can only be computed for the second scenario that contains a known set of true DiffR transcripts. VDA never recovered more than 50% of true DiffR transcripts, and no more than 20% of the DiffR hits were truly DiffR (Fig. 1F). The best performance achievable (white filled circle in Fig. 1F) is the recovery of 40% of DiffR transcripts with only 20% true DiffR transcripts among the hits—133 true DiffR transcripts along with 595 false DiffR hits (Fig. 1G). This stringent FDR threshold for rhythms (1.6×10^{-3}) greatly reduces the rhythmic transcripts considered in VDA.

An amplitude threshold does not improve the performance of VDA. Restricting attention to rhythmic transcripts with a minimum amplitude helps select biologically important results [10] and also considers the effect size in addition to a *P*-value [11]. An amplitude threshold of 0.5 $\log_2$ expression reduced the false DiffR hits called by VDA to ~ 40% of rhythmic transcripts under the first scenario (Fig. S1C,D), but did not eliminate them for any choice of FDR threshold (Fig. S1E). Under the second scenario, VDA with an amplitude threshold recovered fewer true DiffR transcripts and also called fewer false DiffR hits (Fig. S1F). Moreover, an amplitude threshold does not improve the best precision-recall performance achievable (Fig. S1G). Note, the fewer false DiffR hits produced with an amplitude threshold comes at the cost of fewer rhythmic transcripts considered for DiffR analysis (Fig. S1H), similar to Fig. 1G.

Thus, the presence of a large number of false DiffR hits, both in data with and without true DiffR transcripts, confirms our expectation that VDA overestimates the true number of DiffR features.

compareRhythms reliably and thoroughly finds DiffR transcripts

Clearly, a better approach than VDA is needed to identify DiffR transcripts. Two types of approaches have been proposed in the literature—hypothesis testing and model selection (Fig. 2).

Hypothesis testing assesses whether circadian parameters (amplitude and phase) are different between the two conditions by defining a null hypothesis for DiffR analysis [12]. Rejecting this null hypothesis produces results closer to one's intuitive understanding of DiffR features. Hypothesis testing produces four groups of transcripts (Fig. 2, bottom) and not just the three in VDA. Let us term the two conditions A and B. Among the prefiltered transcripts rhythmic in either A or B, there are transcripts that are (i) only rhythmic in A (and are DiffR hits), (ii) only rhythmic in B (and are DiffR hits), (iii) rhythmic in A and B (and are DiffR hits because they have different amplitude and/or phase), (iv) rhythmic in A and B (and are not DiffR hits). If A is the control condition, we could equally term these as (i) *loss* of rhythms (ii) *gain* of rhythms (iii) *change* of rhythms, and (iv) *same* rhythms. The distinction between (iii) and (iv) is nonexistent in VDA. Note, hypothesis testing is performed on the same set of prefiltered rhythmic transcripts used in VDA. Hence, a Venn diagram visualization is best avoided or must be altered to accommodate the fourth category.

The model selection approach [13] forgoes the choice of a null hypothesis (Fig. 2). Instead, the four rhythm categories identified by hypothesis testing along with an arrhythmic category are fit using a nested collection of harmonic regression (cosinor) models [7]. The 'best' model (rhythm category) is chosen based on an information-theoretic criterion, such as Akaike information criterion (AIC) or the Bayesian Information Criterion (BIC). Furthermore, the best model needs to be significantly better (determined by user-defined threshold set) than the second best model based on the same criterion in order to have confidence in the classification. Otherwise, the transcript is left unclassified.

We have implemented both these approaches in an R-package COMPARERHYTHMS (https://github.com/bharathananth/compareRhythms). The hypothesis testing

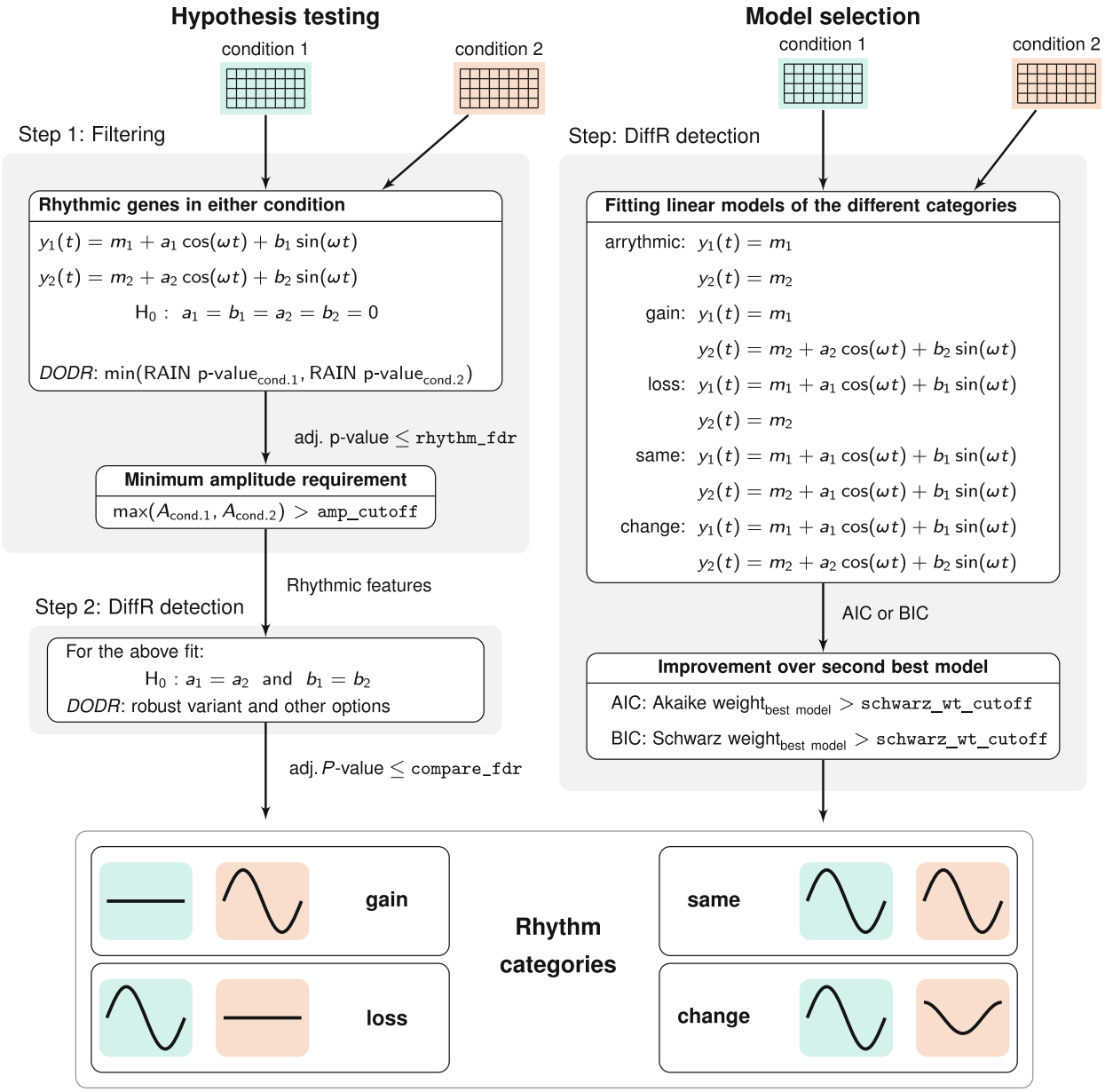


Fig. 2. The two approaches for DiffR identification implemented in COMPARERHYTHMS. Rhythmic transcripts (features) are classified into four categories: loss, gain, change, and same rhythm in condition 2 with respect to condition 1. 'rhythm_fdr', 'compare_fdr', 'amp_cutoff', and 'schwarz_wt_cutoff' are parameters controlling different thresholds within the analysis.

approach can be implemented for microarray data (using *limma* [14]), RNA-seq data (using *DESeq2* [15], *edgeR* [16] or *limma-voom*), or generic prenormalized data (using *RAIN* and *DODR* [12]). There are often two ways of analyzing the same data using hypothesis testing. For example, microarray data can be either normalized and analyzed using *DODR* or directly analyzed accounting for microarray properties using *limma*. We evaluate the effect of this choice of 'implementation' on our conclusions as appropriate. The model selection approach can be directly applied to any generic prenormalized data. A single command performs the standard analysis on the two datasets with the chosen approach.

Both methods in COMPARERHYTHMS (without amplitude thresholds) called negligible number of false

DiffR hits in the first scenario with no true DiffR transcripts (Fig. 1A). Hypothesis testing called the 'same' rhythms in all 4750 rhythmic transcripts (DiffR test with FDR < 0.05) and no DiffR transcripts between odd and even time points. On the other hand, model selection called 189 *false* DiffR hits and 4244 transcripts with 'same' rhythms (compare with Fig. 1B). Different implementations of hypothesis testing in COMPARERHYTHMS also did not call any DiffR transcripts.

Hypothesis testing called significantly fewer false DiffR transcripts than model selection to recover ~ 75% of the true DiffR transcripts in the second scenario. The analysis without an amplitude threshold aims to recover all true DiffR transcripts (Fig. 3, left). Hypothesis testing almost perfectly recalled 70% of the true DiffR transcripts or 80% of the true DiffR transcripts with precision above 80% (independent of its implementation in COMPARERHYTHMS as described above (Fig. S2)). On the other hand, model selection suffered from a poor 50% precision in recovering 75% of the true DiffR transcripts. Nevertheless, both these methods performed significantly better than VDA.

Both methods were equally good at finding DiffR transcripts with biological relevance based on rhythm amplitude. If only rhythmic transcripts with sufficiently large amplitude are considered relevant, we can run both methods with a suitable amplitude threshold, which is default in COMPARERHYTHMS. Both model selection and hypothesis testing recalled 80% of true DiffR transcripts at about 80% precision in the second scenario (Fig. 3, right). The amplitude threshold deteriorated the performance of hypothesis testing slightly, but improved model selection. Note, the default setting in COMPARERHYTHMS achieves the best trade-off between precision and recall (circles in Fig. 3, right).

Although VDA clearly overestimates DiffR transcripts, its impact on studies that used VDA is unclear. To that end, we reanalyzed three studies using COMPARERHYTHMS to re-assess changes in rhythmicity and whether novel biological insights might have been overlooked.

High-fat diet mainly affects the core circadian clock in the liver

We reanalyzed an early high-resolution study that characterized changes in the circadian liver transcriptome in response to a nutritional challenge, that is, high-fat diet (HFD) [17]. In order to validate our reanalysis, we analyzed DiffR in response to HFD

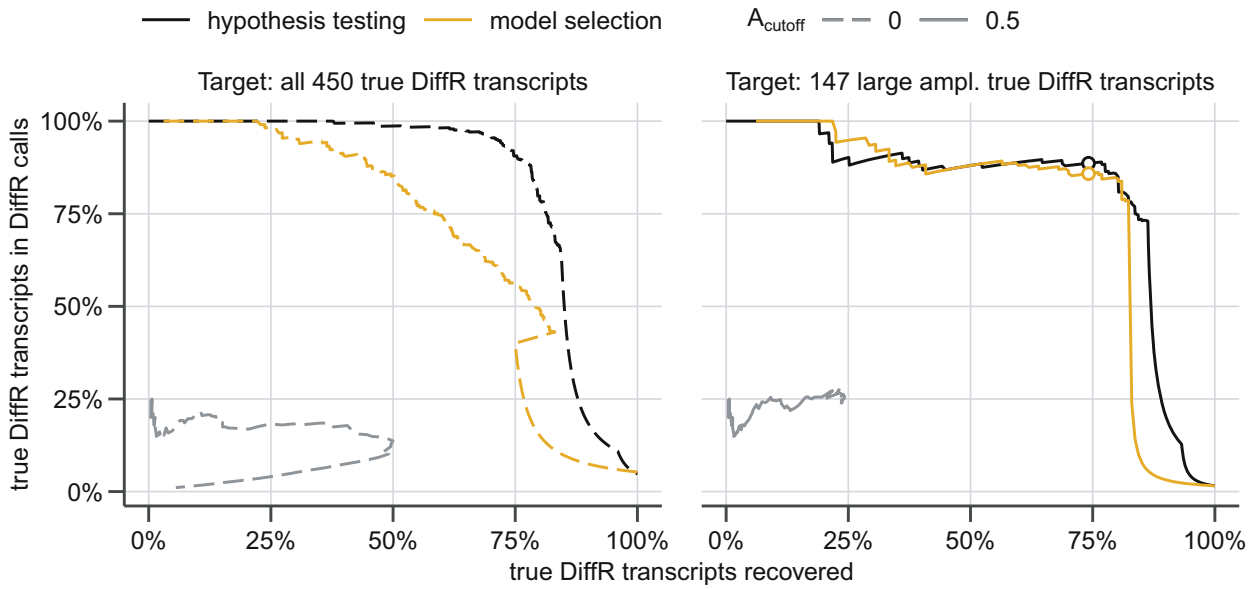


Fig. 3. Precision-recall performance of different COMPARERHYTHMS approaches applied to the second scenario. (*Left*) Performance of hypothesis testing and model selection on the data analyzed in Fig. 1E–G without an amplitude threshold (A_{cutoff}) for rhythmic transcripts. (*Right*) Performance of the two approaches with an amplitude threshold on data analyzed in Fig. 1E–G aimed at recovering the true DiffR transcripts with biologically relevance (with amplitudes > 0.5 $\log_2$ expression). Performance at the default setting in COMPARERHYTHMS is marked with circles. The corresponding performance of VDA (from Fig. 1F, Fig. S1G) is shown in gray for reference. The curves were constructed by varying 'compare_fdr' for hypothesis testing and 'schwarz_wt_cutoffE for model selection.

quantified using high-throughput sequencing from an independent study, that neither used VDA nor performed DiffR analyses [18].

Hypothesis testing and model selection called less than a tenth and about a third of rhythmic transcripts, respectively, to be DiffR across both studies (Fig. 4A, Table S1). Hypothesis testing called only 90 and 160 DiffR hits in the microarray and RNA-seq data respectively. Model selection, on the other hand, called 328 and 791 DiffR hits in the two studies. The higher number of DiffR transcripts called by model selection is consistent with the second scenario, where model selection called more false DiffR hits compared to hypothesis testing in some situations (Fig. 3). Many more rhythmic transcripts were detected in the RNA-seq data by both methods (Fig. 4A, Fig. S3A). Interestingly, the fraction of rhythmic transcripts called DiffR by hypothesis testing is higher with than without an amplitude threshold mainly due to an increase in the detected number of rhythmic transcripts in the latter (Fig. S3A).

DiffR hits called by hypothesis testing were a subset of DiffR hits called by model selection in both studies. All but 9 transcripts called DiffR by hypothesis testing were also called DiffR by model selection in the microarray data (Fig. S3B, Table S1). These 9 transcripts could not be classified by model selection, since multiple rhythm categories matched the expression pattern equally well. The excess DiffR hits in model selection were mostly (203) called rhythmic but non-DiffR and a few (44) were called arrhythmic by hypothesis testing. In the RNA-seq data, the DiffR hits from hypothesis testing not also called DiffR by model selection were almost all left unclassified (Fig. S3C, Table S1). Again, the excess DiffR hits in model selection were mostly categorized 'same' by hypothesis testing.

The proportion of transcripts in the different rhythm categories was similar across studies, but overlap of DiffR hits was poor. The two methods categorized DiffR transcripts similarly as 'loss', 'change', and 'gain' in each study (Fig. S3B,C). Differences in classification of common DiffR transcripts resulted from different assignment between 'gain' or 'loss' and 'change' in the two methods. Comparing the studies, hypothesis testing predicted similar fractions of DiffR transcripts in the 'loss', 'change' and 'gain' categories (Fig. 4B). Differences in assays and annotations limited the number of commonly rhythmic transcripts to 736 (Fig. S3D). Of these, only 10 transcripts were called DiffR in both, while 635 were called not DiffR in both. The remaining DiffR hits in each study were called 'same' or arrhythmic in the other.

DiffR estimates from VDA greatly exceeded the estimates from hypothesis testing, but still missed relevant DiffR transcripts. VDA in the original microarray study [17] called about three-quarters of rhythmic transcripts DiffR (Fig. S3E), and our recomputed VDA was only slightly smaller (Table 1). The number and fraction of DiffR transcripts from VDA exceeded the estimates from hypothesis testing both with and without amplitude thresholds. Only 25 of 90 DiffR hits from hypothesis testing were also DiffR hits in the original VDA (Fig. S3F). As expected, VDA missed 31 (of 90) DiffR transcripts that showed altered circadian parameters and classified them as rhythmic in 'both'. 70% of the excess DiffR hits from VDA were called arrhythmic by our reanalysis and 19% were classified as 'same'. The RNA-seq study [18] that we included to validate our DiffR estimates did not report any DiffR analysis including VDA.

DiffR transcripts showed a consistent phase advance in HFD; however, DiffR hits were not significantly enriched for any process. We observed a consistent phase advance of between 2 and 4 h in almost all the DiffR transcripts across both studies (Fig. 4C). Gene enrichment analysis typically follows DiffR analysis in order to generate hypotheses. The small DiffR transcript set (Fig. 4B) was expected to make enrichment analysis less statistically powerful. Nevertheless, circadian rhythms and metabolism-related terms constituted the top five enriched KEGG categories among the DiffR transcripts (Fig. 4D); we always used all the rhythmic transcripts in either group as the background.

The individual transcript time courses were also remarkably similar between the studies (Fig. 4E). All the core clock genes in Fig. 4E except *Nr1d2* and *Per1* were DiffR in at least one study and even the two exceptions showed a trend toward earlier phases under HFD seen for the DiffR transcripts (Fig. 4C).

In summary, all the core clock genes and a few CCG transcripts were DiffR under HFD with a very consistent phase advance.

A ketogenic diet significantly activates circadian immune response in the mouse liver

We next reanalyzed microarray data on the effect of a ketogenic diet (KD) on the mouse liver transcriptome [19] (Fig. 5).

DiffR hits from model selection contained hits from hypothesis testing as with HFD, and both methods predicted that most DiffR transcripts gained rhythms under KD in the liver. Hypothesis testing and model selection called 271 and 457 DiffR transcripts in

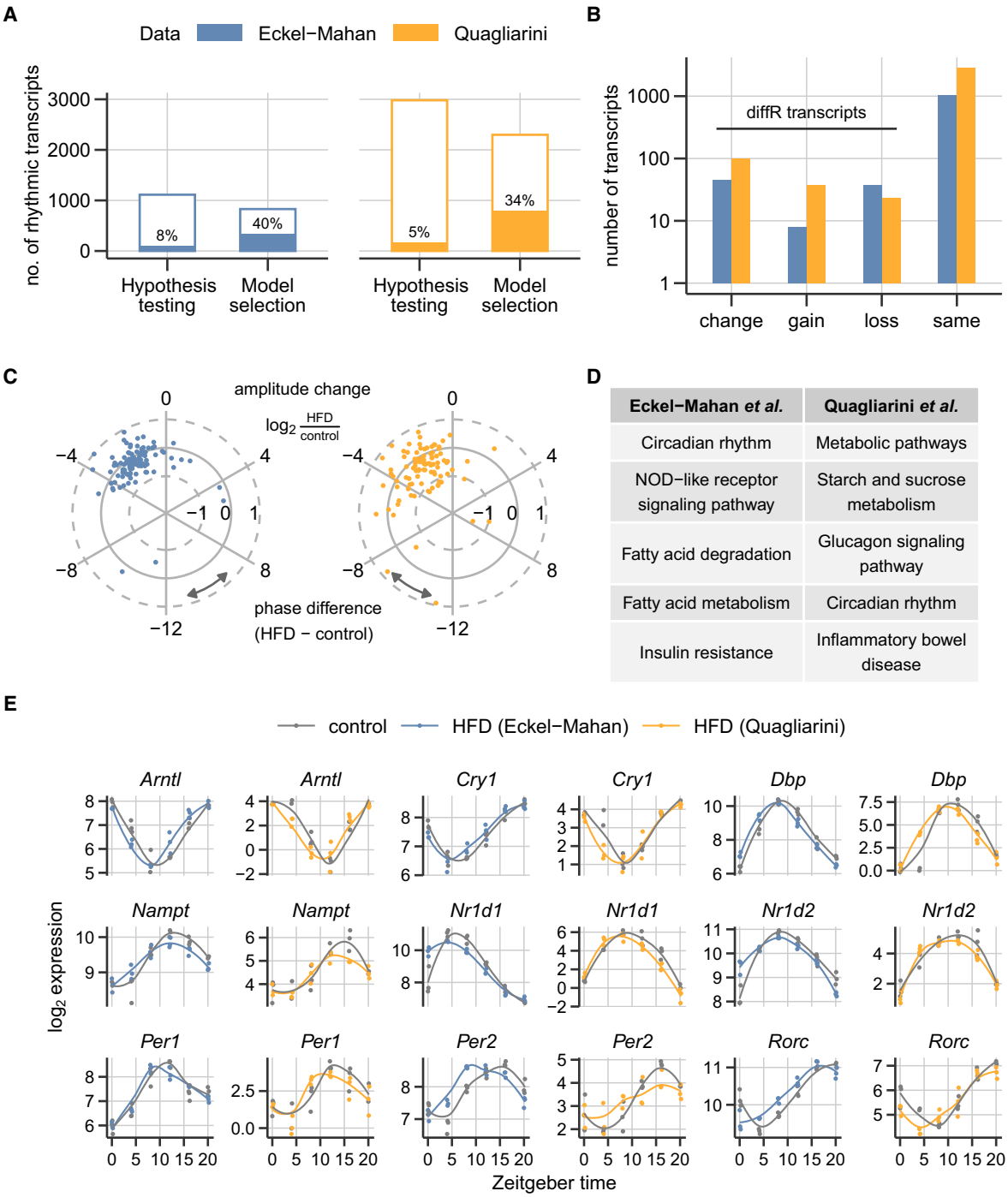


Fig. 4. Effect of a high-fat diet (HFD) on the mouse liver clock. (A) The number of rhythmic transcripts (open bars) and DiffR transcripts (filled bars) called by the two approaches in the microarray data [17] and the analogous RNA-seq data [18] using COMPARERHYTHMS with default parameters. The percentage of rhythmic transcripts called DiffR is displayed within the bars. (B) The classification of DiffR hits predicted by hypothesis testing into those that 'change', 'gain', or 'lose' rhythms. Rhythmic DiffR misses have the 'same' rhythms in the two groups. (C) Circular plot representing the phase and amplitude change in the DiffR transcripts between control and HFD. Amplitude changes are represented as radial deviations from the solid gray circle and angular phase (in h) are positive for delays and negative for advances. (D) The top five KEGG enrichment categories for DiffR transcripts in each study. (E) The raw data as $\log_2$ expression for the 9 core clock genes, out of which 7 are DiffR in either the microarray or RNA-seq datasets. The lines are the mean LOESS-smoothed expression profiles for visual comparison.

Table 1. DiffR transcripts called in mouse liver by the original studies and our reanalyses. VDA results are presented as reported in the original studies. Results from hypothesis testing are provided both without an amplitude threshold (directly comparable to reanalyzed VDA) and with an amplitude threshold (default in COMPARERHYTHMS). We recomputed the VDA numbers using the same filtered rhythmic transcripts and rhythmicity detection used in hypothesis testing (without an amplitude threshold). 'rhy.' Is the number of transcripts rhythmic in *either* condition and 'diffR' is the total diffR transcripts in the 'gain', 'loss', and 'change' categories.

	VDA (reported)			VDA (recomputed)			Hypothesis testing (no min amp.)			Hypothesis testing (with min amp.)		
Study	diffR	rhy.	%	diffR	rhy.	%	diffR	rhy.	%	diffR	rhy.	%
Eckel-Mahan *et al.* [17]	2048	2826	72	2042	3103	66	57	3059	2	90	1113	8
Quagliarini *et al.* [18]	–	–	–	3534	4728	75	120	4728	3	160	2981	5
Tognini *et al.* [19]	3058	3859	79	2973	4034	74	508	4034	13	271	1203	23
Pei *et al.* [21]	2083	2503	83	3189	5346	60	424	5346	8	414	4473	9

response to KD (Fig. 5A, Table S2). Except for 32 transcripts that could not be reliably categorized by model selection, all DiffR hits from hypothesis testing were called DiffR by model selection (Fig. S4A). Moreover, the common DiffR hits were categorized identically. The additional DiffR transcripts called by model selection were considered to be mostly non-DiffR and the remaining arrhythmic by hypothesis testing. The proportions of different categories, however, of DiffR transcripts were similar between the two methods.

VDA overestimated the number of DiffR hits relative to hypothesis testing and missed several DiffR hits in our reanalysis. VDA (reported and recomputed) predicted about 75–80% of rhythmic transcripts were DiffR (Table 1). 75% of the DiffR transcripts called by VDA also showed novel rhythms under KD (Fig. S4B, compare with Fig. 5A). Nonetheless, only 151 of the 271 DiffR transcripts overlapped between the reanalysis and the original VDA (Fig. S4C). Hypothesis testing placed about 40% of DiffR hits missed by VDA (as rhythmic in 'both') in the 'change' category, a category absent in VDA. The fraction of rhythmic transcripts called DiffR by hypothesis testing is smaller without compared to with an amplitude threshold due to increases in the number of DiffR hits but also the number of rhythmic transcripts (Table 1).

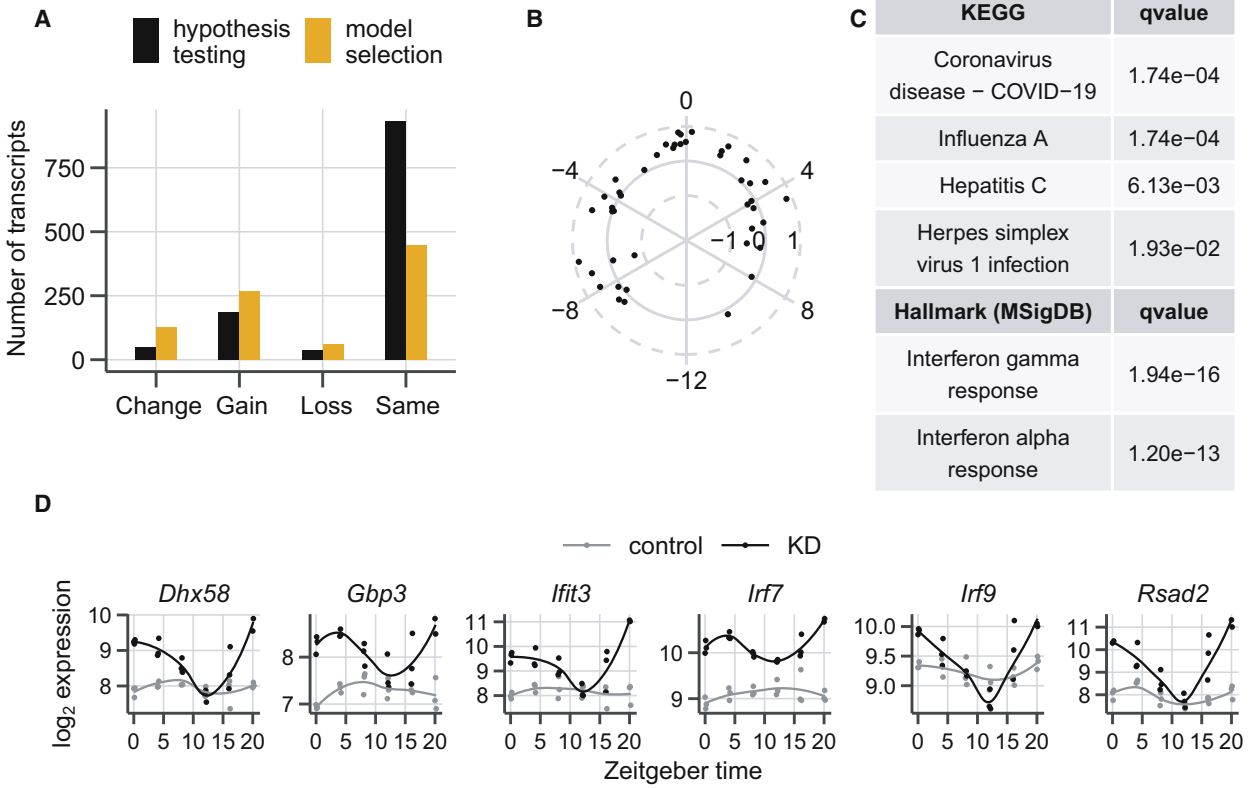
DiffR transcripts generally increased amplitude under KD and were enriched for immune response pathways. A majority of DiffR transcripts placed in the 'gain' category and DiffR transcripts in the 'change' category showed a rhythm amplitude increase but no clear trend in the phase change (Fig. 5B). Rhythm parameter changes cannot be estimated accurately when the transcript is arrhythmic in one condition ('gain' and 'loss' categories). The significantly enriched KEGG pathways were all associated with responses to infectious disease—COVID-19, influenza A, hepatitis C, and herpes (Fig. 5C). In agreement, the DiffR transcripts were highly over-represented for hallmark genes upregulated in response to Interferon α and γ according to the MSigDB database [20]. Moreover, most immune response associated DiffR transcripts acquired rhythmicity under KD (Fig. 5D and Table S2).

In a nutshell, KD induced more rhythm changes in the liver than HFD including *de novo* rhythms in a large subset of DiffR transcripts. These DiffR transcripts were closely associated with immune response pathways.

Disruption of endogenous H_2O_2 rhythms activates circadian oncogenic signaling

We reanalyzed finally the RNA-seq data on the effect of disrupting endogenous H_2O_2 rhythms by knocking-out (KO) $p66^{Shc}$ on the circadian liver transcriptome [21].

Both methods similarly categorized DiffR transcripts with hypothesis testing hits included in model selection hits yet again. Hypothesis testing and model selection called 414 and 1522 DiffR transcripts (Fig. 6A, Table 1, Table S3). This result was insensitive to the implementation of hypothesis testing in COMPARERHYTHMS (not shown). As before, all but 88 DiffR transcript calls by hypothesis testing agreed with model selection; 71 of these could not be reliably classified by model selection (Fig. S5A). About 70% of the excess DiffR hits from model selection were classified as having 'same' rhythms by hypothesis testing. Nevertheless, the proportion of the different DiffR categories was conserved across approaches, with most DiffR hits in 'change' followed by 'gain' and then 'loss'.



KEGG	qvalue
Coronavirus disease − COVID−19	1.74e−04
Influenza A	1.74e−04
Hepatitis C	6.13e−03
Herpes simplex virus 1 infection	1.93e−02
Hallmark (MSigDB)	**qvalue**
Interferon gamma response	1.94e−16
Interferon alpha response	1.20e−13

Fig. 5. Effect of a ketogenic diet (KD) on the mouse liver clock. (A) The number of transcripts in the four categories resulting from DiffR analysis of microarray data [19] using hypothesis testing and model selection. (B) Circular plot representing the phase and amplitude change in the transcripts in the 'change' category in (A) between control and KD. Amplitude changes are represented as radial deviations from the solid gray circle and angular phase (in h) are positive for delays and negative for advances. (C) KEGG and Molecular Signatures hallmark gene set enrichment of all the DiffR transcripts with the set of transcripts rhythmic in either control or KD as background. (D) Raw $\log_2$ expression time courses of selected transcripts involved in interferon response under control and KD. The lines are the mean LOESS-smoothed expression profiles for visual comparison.

We observed a large mismatch between VDA in the original study and hypothesis testing. VDA called about 84% of the rhythmic transcriptome DiffR in the original study (Fig. S5B, Table 1). Our recomputed VDA predicted a lower fraction (60%) to be DiffR, since we identified many more rhythmic transcripts by accounting for a batch effect (see Methods) unnoticed in the original analysis. Two-thirds of the hypothesis testing DiffR hits were overlooked (Fig. S5C) and over 80% of these were considered arrhythmic by VDA. Unlike the previous two studies, the amplitude threshold had no effect on the hypothesis testing results (Table 1). Surprisingly, more than 2500 transcripts with the same rhythms in both conditions were considered not expressed or arrhythmic in the original VDA.

DiffR transcripts were enriched in vasculature development. The phase and amplitude shifts in the 'change' DiffR transcripts did not show a specific trend (Fig. 6B). However, there appeared to be two cluster of phase shifts: those that are phase advanced by ~ 4 h and those that are phase delayed by ~ 8 h. The DiffR set was significantly over-represented for the GO categories 'angiogenesis' and 'blood vessel morphogenesis' (Fig. 6C). 'Genes upregulated by KRAS signaling' were also significantly enriched in the DiffR set. Most genes that overlapped with this hallmark set gained rhythms in the knockout (Fig. 6D).

To sum up, altering the endogenous H_2O_2 rhythms cause gain, loss, and change in rhythms in the liver and these DiffR transcripts are associated with the circulatory system development and oncogenic KRAS signaling.

Discussion

High-throughput time series profiling of a control and an experimental group is the standard approach to characterize the response of the circadian clock system

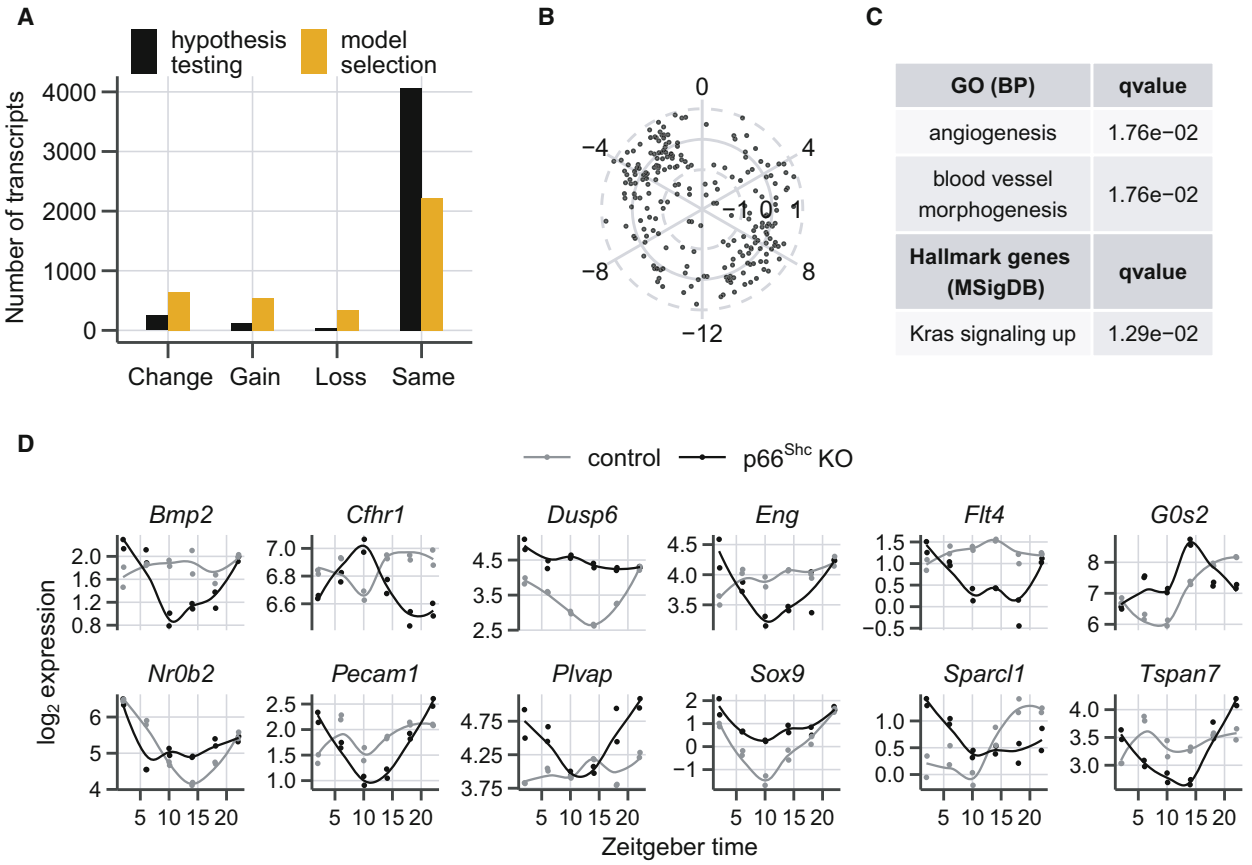


Fig. 6. Effect of *p66*^{Shc} knockout (KO) on the mouse liver clock. (A) The number of transcripts in the four categories resulting from DiffR analysis of the data [21] using hypothesis testing and model selection. (B) Circular plot representing the phase and amplitude change between control and KO in the transcripts in the 'change' category in (A). Amplitude changes are represented as radial deviations from the solid gray circle and angular phase (in h) are positive for delays and negative for advances. (C) KEGG and Molecular Signatures hallmark gene set enrichment of all the DiffR transcripts with the set of transcripts rhythmic in either control or *p66*^{Shc} KO as background. (D) Raw $\log_2$ expression time courses under control and *p66*^{Shc} KO of selected transcripts upregulated in KRAS signaling. The lines are the mean LOESS-smoothed expression profiles for visual comparison.

to a treatment. Many such studies in high-impact journals discovered large-scale circadian changes resulting from the treatment and often described this phenomena as 'circadian reprogramming' or 'circadian remodeling'. Most of these studies used VDA to determine the number and identity of DiffR transcripts. VDA finds rhythmic transcripts in the two groups separately and then compares these lists to predict DiffR and non-DiffR transcripts. Analyses of high-throughput data stringently control for false discoveries at the cost of missing many true discoveries. In this work, we questioned the validity of the VDA due to the propagation of false discoveries and missed true discoveries from two individual rhythmicity tests to the determination of DiffR transcripts.

We showed using two artificial experiments designed on real liver transcriptomic data that VDA produces excessive false DiffR hits (Fig. 1, Fig. S1). VDA called DiffR hits even when DiffR transcripts are unexpected. When true DiffR transcripts were present, not only were more than three-quarters of the returned hits incorrect, but VDA never recovered more than half the true DiffR transcripts. This failing of VDA could not be overcome by being very conservative (very low FDR threshold) or being permissive (larger FDR threshold) with rhythm detection in either group. This issue with VDA generalizes to high-throughput data beyond transcriptomes and, in fact, to any effect (not just rhythmicity) measured independently in two datasets.

We thus argue that VDA is defective and ought to be avoided. First, VDA is the high-throughput counterpart to a well-known statistical error—see 'Interpreting comparisons between two effects (here

rhythms) without directly comparing them' in ref. [3]. Second, VDA leaves to chance the fraction of returned DiffR hits that are false-positive akin to failing to correct for multiple testing in high-throughput data analyses (Fig. 1D). Third, in our benchmarking (Fig. 1F), VDA recovered < 50% of the DiffR transcripts. Fourth, VDA estimated at least 2–3 times the number of DiffR transcripts estimated by hypothesis testing (which was superior in our benchmarking), even assuming the latter conservatively recovers only 50% of true DiffR transcripts. The goal of DiffR analysis is to find gene sets and so we cannot be certain about individual hits. Nevertheless, we infer many individual hits from VDA are incorrect, since DiffR transcripts from hypothesis testing are much smaller in number than VDA's prediction.

We presented model selection and hypothesis testing to identify DiffR features (Fig. 2) with implementations for different datatypes in a convenient to use R-package COMPARERHYTHMS. Both approaches address two key drawbacks of VDA. First, these methods explicitly (hypothesis testing) or implicitly (model selection) control false DiffR hits in the analysis. In other words, we can set a desired significance cutoff for DiffR analysis just as we set one for rhythmicity analysis. For instance, we compared (not shown) WT control time series (from [17] and [19]) performed under identical conditions (age, mouse strain, lighting, feeding), using the same assay in the same laboratory. In this situation where no DiffR is expected, VDA and hypothesis testing predicted 75% and 6% DiffR transcripts among rhythmic transcripts. While the former does not control false hits, the latter controls them at the chosen 5% threshold. Both methods in COMPARERHYTHMS recovered at least 75% of the true DiffR transcripts, as benchmarked by the second scenario (Fig. 3). While > 80% of hypothesis testing calls were correct, only 50% of model selection calls were correct for some parameter settings. The default settings in our implementation correspond to the best performance trade-off between recovery of true DiffR transcripts and reliability of the DiffR calls.

Second, by directly comparing amplitudes and phases between control and experimental groups, the methods INCOMPARERHYTHMS also call transcripts with changed rhythms. Model selection and hypothesis testing further classify rhythmic transcripts into four categories: transcripts that gain rhythms, lose rhythms, whose rhythms have changed (amplitude and/or phase), or have the same rhythms between the control and experimental groups. VDA analysis completely disregards the third category of transcripts. The seemingly intuitive set-theoretic approach underlying VDA only demarcates three groups from two lists of rhythmic transcripts (sets).

To evaluate our conclusions further, we reanalyzed three studies that used VDA (and one that did not report such an analysis for validation) to quantify the effect of metabolic changes on the mouse liver circadian transcriptome. Across all studies, hypothesis testing and model selection called between 8–23% and between 34–51% of rhythmic transcripts to be DiffR, respectively (Table 1). VDA (as reported) estimated 72–85% of rhythmic transcripts were DiffR, while our recomputed VDA (which is matched to hypothesis testing) estimated 60–75% of rhythmic transcripts were DiffR. The absolute number of DiffR hits was also much higher with VDA. The difference between VDA and hypothesis testing was not due to differences in the analyzed set of rhythmic transcripts (Table 1). KD elicited DiffR in a larger fraction of rhythmic transcripts compared to H_2O_2 rhythm disruption and HFD, which were equal, according to our reanalysis. The fraction of rhythmic transcripts called DiffR in response to HFD was also consistent across two identical studies performed in different laboratories and on different assays (Fig. 4). Assuming a conservative 50% recovery of DiffR transcripts (Fig. 3), the true DiffR number is at most twice our hypothesis testing estimates, which has high precision. This reinforces our conclusion that VDA overestimates the true extent of circadian 'reprogramming'.

DiffR analysis produces sets of transcripts that can be functionally interpreted using pathway, gene ontology, or reactome analysis. We might expect from VDA's tendency to overestimate DiffR transcripts and call many false DiffR hits that the DiffR hits in the reanalysis are contained among the VDA hits. Surprisingly, between 43–73% of DiffR hits called by hypothesis testing were absent in the DiffR hits called by the originally reported VDA. As expected, many of the misses were DiffR transcripts deemed in the 'change' category that were classified as non-DiffR by VDA. However, the hits from hypothesis testing were contained in the hits from model selection. This is consistent with our insight from benchmarking that model selection produces more false positives than hypothesis testing. In summary, VDA neglects all DiffR hits with altered rhythm parameters, since VDA is not designed to find these.

We next explored whether changes in the number and identity of DiffR hits affected the conclusions of the original studies. Under HFD, we confirmed that most DiffR hits lost rhythms or amplitude and were phase advanced (Fig. 4B,E), albeit in a small DiffR transcript set. But, we found no evidence for 'massive

induction of de novo oscillating transcripts' [17]. Under KD, we could corroborate the gain of rhythms in a plurality of DiffR transcripts (Fig. 5A). However, we found neither KEGG enrichment of 'metabolism' or 'PPAR signaling' among DiffR transcripts (Fig. 5C) nor trends in the phase changes (Fig. 5B). On altering H_2O_2 rhythms, we found a bias in DiffR hits toward 'gained' or 'change' categories (Fig. 6A) not present in the original study. We also found no enrichment of 'oxidation-reduction process' or 'metabolic process' among DiffR transcripts (Fig. 6C). To sum up, some conclusions held up under the reanalysis, some did not and others could not be evaluated. Therefore, high-throughput circadian studies using VDA must be re-assessed individually.

We wondered next whether our reanalysis revealed novel insights overlooked in the original studies. The effect of HFD (based on both studies) on the circadian system is rather modest and restricted to the core clock and a few additional transcripts that show a consistent ~ 4 h phase advance of rhythms (Fig. 4E). We conjecture that *Nampt* alone (Fig. 4C), with no role played by PPARγ, drives a similar pattern in the metabolome (The metabolome is also easily reanalyzed using COMPARERHYTHMS). Enrichment analysis strongly suggested that KD activates viral defense (including against COVID-19) and interferon α,γ response pathways by inducing *de novo* transcript rhythms in these pathways (Fig. 5D). Recently, KD was shown to provide protection against influenza infection [22] and this effect is likely in part mediated by the circadian system. Finally, DiffR transcripts in response to redox changes overlapped significantly with transcripts upregulated in KRAS signaling (Fig. 6C,D). We propose that the complex interaction between redox balance and cancer [23] is also circadian clock mediated with a possible role for blood vessel development. Our novel insights thus involve circadian clock modulation of the interaction between physiological processes.

Analysis using COMPARERHYTHMS uses an amplitude threshold by default. We recommend using an amplitude threshold to filter transcripts considered for DiffR analysis for four reasons. First, statisticians recommend considering the effect size (amplitude in this case) in addition to statistical significance (*P*-values) [11]. Second, the subsequent DiffR analysis compares circadian parameters that are unreliably estimated from low amplitude rhythms. Third, in high-throughput analyses often less is more, since multiple testing corrections penalize *P*-values in relation to the size of the search set. This is seen in the complex changes in the number of DiffR hits on removal of the amplitude threshold (Table 1). Four, it is arguable that only rhythms with sufficiently large amplitude are *biologically relevant*. In our benchmarking, an amplitude threshold improved model selection performance to match hypothesis testing (Fig. 3). In the reanalyses, amplitude threshold removal increased numbers of rhythmic transcripts more than the number of DiffR hits and did not affect our general conclusions.

Our conclusions must be nonetheless viewed within the context of the presented approaches. We identify DiffR hits using changes in circadian parameters estimated assuming sinusoidal rhythm patterns, which might be unsuitable in certain situations (e.g., long/short photoperiods). The DiffR hits in response to HFD from the microarray and RNA-seq data showed limited overlap (Fig. S3D) despite well-matched experimental conditions (mouse strain, age, food, lighting), since the superior RNA-seq detected many more rhythmic transcripts than the microarray for the default settings in COMPARERHYTHMS. Combining data across assays is beyond the scope of this work andCOMPARERHYTHMS. All bioinformatic analyses involve choosing multiple thresholds, which requires considerable thought. We measured DiffR in terms of hits as a fraction of rhythmic transcripts, since they were less sensitive to thresholds than absolute numbers. Biologically, it is unclear, which of the two is more relevant.

The choice of approach in COMPARERHYTHMS must consider, beyond performance, the nature of the data (transcriptomic vs. nontranscriptomic data), covariates, effects of experimental batches, waveforms of interest (sinusoidal vs. nonsinusoidal), size of the datasets/speed, and experimental design complexity. We summarize the trade-offs involved in Table 2. We focused on transcriptomic data due to their sheer abundance in public archives, but our tool can be easily applied to any normalized (according to the particular datatype) data, such as metabolomic and proteomic data, from any organism. Our tool can currently only compare two groups with one categorical variable to account for covariates, such as batch or sex. We also disregard changes in mean expression between the two groups. We lent toward simplicity and performance in COMPARERHYTHMS at the cost of including many different experimental designs (e.g., [24,25]).

Although COMPARERHYTHMS provides multiple approaches for different datatypes in one convenient package, the shortcomings of VDA can be combated using other approaches too. However, none of these approaches use an amplitude threshold, that is, consider the size of the effect in DiffR identification. *LimoRhyde*, *CircaCompare*, *diffCircadian,* and *CosinorPy* are all cosinor-based [7] and fall within the hypothesis testing framework in Fig. 2. *LimoRhyde*

Table 2. Advantages and disadvantages of the analysis pipelines in COMPARERHYTHMS. Linear modeling encompasses all implementations of hypothesis testing other than *DODR*, that is, *limma*, *voom*, *DESeq2*, *edgeR*.

Hypothesis testing		
DODR	Linear modeling	Model selection
+ straightforward	+ very fast	+ simple, elegant and fast
+ rhythm detection using RAIN	+ blends easily into standard pipelines for transcriptomic data	+ directly provides all DiffR categories
+ works on any normalized data	+ can include other covariates and batch variables	+ works on any normalized data
– mixes parametric (for DiffR detection) and nonparametric methods (for rhythm detection)	– rhythm and DiffR detection based on sinusoidal rhythm pattern	+ can include other covariates and batch variables
– slow		– exponential growth in models with rhythm categories of interest
– cannot include other covariates in pipeline		– more false hits in some situations

[26] allows for more complex experimental designs than COMPARERHYTHMS (such as changes in mean expression) at the cost of simplicity for DiffR analysis of transcriptomics datasets. An alternative formulation, fit by nonlinear regression, is used in *CircaCompare* to analyze DiffR in any normalized data, and it can additionally provide statistical significance for differences in particular circadian parameters [27]. However, nonlinear regression is not robust to violations of assumptions, does not account for particular properties of transcriptomic data, and the analysis does not handle multiple testing needed in high-throughput datasets. *diffCircadian* [28] present a likelihood-ratio test-based DiffR analysis for generic prenormalized data, which our implementations (except *DODR*) already use for transcriptomic datasets. *CosinorPy* [29] is the only package that allows for simple DiffR analysis in Python. *dryR* [30] is a newer version of model selection approach that specifically accounts for properties of RNA-seq data and allows more complex designs with the drawback of having a combinatorial explosion of rhythm categories. Finally, *MOSAIC* [31] uses hierarchical modeling, nonlinear regression, and assumptions regarding rhythms in the transcriptome and proteome to address DiffR analysis in a specific multi-omics context.

Conclusion

Problems with statistics and experimental design are often cited as one of the main causes for the reproducibility crisis in science [32]. The deficiency of the common approach to DiffR analysis is related to a common mistake of comparing two experimental effects without directly comparing them [33,34] and afflicts many more studies than those we reanalyzed. Venn diagrams that are symptomatic of VDA ought to serve as a warning flag. We trust that chronobiologists will find our tool an easy way to avoid this pitfall and generate reliable hypotheses to best utilize their resources.

Methods

All analysis and statistics were performed using R 3.6.3 [35].

Data sources

All data used in this study were gathered from Gene Expression Omnibus (GEO) database [36] or the Short Read Archive. The accession numbers for the different studies are listed in Table 3.

Data preprocessing

Raw microarray data were loaded using custom chip definition files from Brainarray (v24.0.0) [37] with probes arranged and annotated according to Ensembl gene ID. They were subsequently normalized using the RMA algorithm in the OLIGO package (v1.50.0) [38] to obtain final $\log_2$ expression values. The included subset of gene IDs had a minimum $\log_2$ expression of 5 in at least 70% of the samples in each condition.

Table 3. Accession numbers of the public data from GEO analyzed in this study.

Publication	Accession number	Data type
Hughes *et al.* [8]	GSE11923	Microarray
Eckel-Mahan *et al.* [17]	GSE52333	Microarray
Quagliarini *et al.*[18]	GSE108688	RNA-seq
Tognini *et al.* [19]	GSE87425	Microarray
Pei *et al.* [21]	PRJNA449625	RNA-seq

Raw RNA-seq reads were quantified using SALMON (v1.1.0) [39] with *Mus musculus* reference genome GENCODE build M24 [40]. The *salmon*-quantified transcript expression was converted to gene expression using TXIMPORT package (v1.14.2) [41]. We retained for the analysis all genes that had at least 10 reads per 1 million mapped reads in at least 70% of samples in each condition.

Exploratory data analysis was performed on all datasets using principal component analysis (PCA) to identify potential outlier samples and batch effects (see Vignette in COMPARERHYTHMS for a practical guide). A batch effect was identified in one dataset [21], where replicates one and two separated into two different clusters after PCA.

Artificial scenarios

The 450 true DiffR transcripts for the second scenario were created as follows: we selected 500 transcripts with harmonic regression (cosinor) adjusted *P*-values below 0.05 for the complete 48 time point dataset and peak-to-trough amplitudes above 0.26 (the median amplitude of rhythmic transcripts). We altered the amplitudes in the odd dataset by scaling mean-subtracted expression for each chosen transcript by a random number in [0,1]. We then shifted time labels of odd dataset for each transcript by a number of places chosen randomly in [1,24]; there are 24 time points in the odd dataset. Finally, we added Gaussian noise with standard deviation of 0.25 times the transcript amplitude to the odd dataset (see example in Fig. S1A).

50 of the 500 transcripts were no longer rhythmic in either group leaving 450 true DiffR transcripts. Since the transcripts were chosen based on all 48 time points, chosen transcripts were occasionally rhythmic in only one or the other dataset even without sample alteration (i.e., the first scenario). 198 of the 450 chosen transcripts were rhythmic in only one of the two datasets in the second scenario. The remaining 252 chosen transcripts were rhythmic in both datasets and had differences in amplitude and/or phase of expression (Fig. S1B).

DiffR identification

The hypothesis testing and model selection approaches are outlined in Fig. 2 and in the text and were implemented in the package COMPARERHYTHMS v0.99.0 (https://github.com/bharathananth/compareRhythms). Expression values of the expressed genes were processed with default parameter values using *limma* [14], *DODR* [12] or model selection [14] pipelines for microarray and voom [42] or DESeq2 [15] pipelines for RNA-seq data. For the hypothesis testing-based approaches, all *P*-values were false discovery adjusted using Benjamini–Hochberg correction and adjusted *P*-values were thresholded at 0.05 (unless otherwise mentioned). The amplitude threshold was set at 0.5 $\log_2$ expression with amplitude measured peak-to-trough.

The implementation of VDA for the two artificial scenarios was based on rhythm detection using RAIN [6] with the resulting *P*-values adjusted using the Benjamini–Hochberg (BH) approach. For VDA with an amplitude threshold, the peak-to-trough amplitudes were determined using harmonic regression of the normalized expression values. The VDA reported in the studies (Table 1) used an unadjusted *P*-value significance threshold of 0.01. We recomputed VDA numbers using the rhythmicity calls from the prefiltering step in hypothesis testing applied to that dataset.

Gene enrichment

Gene enrichment analysis was performed using the CLUSTERPROFILER package (v3.14.3) [43] with KEGG database [44] and MSIGDBR (v7.1.1) [45] package translation of Molecular Signatures database [20].

Acknowledgements

We dedicate this manuscript to the memory of Michael E. Hughes, who lead the charge to improve high-throughput data analysis in chronobiology [4]. The authors thank Marta del Olmo for useful comments on the manuscript. This work was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) grant AN 1553/2-1 to Bharath Ananthasubramaniam and DFG - Project-ID 278001972 - TRR 186 to Hanspeter Herzel and Achim Kramer.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

BA designed the study. BA and AP performed the analyses and analyzed the results. BA wrote the manuscript. BA, AP, AK, and HH edited and approved the manuscript.

Data accessibility

The R code and associated data to perform all the analyses in the manuscript are available in the Appendix S1.

References

1 Dunlap JC & Loros JJ (2017) Making time: conservation of biological clocks from fungi to animals. *Microbiol Spectr* **5** 3:5.3.05. https://doi.org/10.1128/microbiolspec.FUNK-0039-2016

2 Rijo-Ferreira F & Takahashi JS (2019) Genomics of circadian rhythms in health and disease. *Genome Med* **11**, 82.
3 Takahashi JS (2017) Transcriptional architecture of the mammalian circadian clock. *Nat Rev Genet* **18**, 164–179.
4 Hughes ME, Abruzzi KC, Allada R, Anafi R, Arpat AB, Asher G, Baldi P, de Bekker C, Bell-Pedersen D, Blau J *et al.* (2017) Guidelines for genome-scale analysis of biological rhythms. *J Biol Rhythms* **32**, 380–393.
5 Hughes ME, Hogenesch JB & Kornacker K (2010) JTK cycle: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J Biol Rhythms* **25**, 372–380.
6 Thaben PF & Westermark PO (2014) Detecting rhythms in time series with RAIN. *J Biol Rhythms* **29**, 391–400.
7 Cornelissen G (2014) Cosinor-based rhythmometry. *Theor Biol Med Model* **11**, 16.
8 Hughes ME, DiTacchio L, Hayes KR, Vollmers C, Pulivarthy S, Baggs JE, Panda S & Hogenesch JB (2009) Harmonics of circadian gene transcription in mammals. *PLoS Genet* **5**, e1000442.
9 Saito T & Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, e0118432.
10 Lück S & Westermark PO (2016) Circadian mRNA expression: insights from modeling and transcriptomics. *Cell Mol Life Sci* **73**, 497–521.
11 Nuzzo R (2014) Scientific method: statistical errors. *Nature* **506**, 150–152.
12 Thaben PF & Westermark PO (2016) Differential rhythmicity: detecting altered rhythmicity in biological data. *Bioinformatics* **32**, 2800–2808.
13 Atger F, Gobet C, Marquis J, Martin E, Wang J, Weger B, Lefebvre G, Descombes P, Naef F & Gachon F (2015) Circadian and feeding rhythms differentially affect rhythmic mRNA transcription and translation in mouse liver. *Proc Natl Acad Sci USA* **112**, E6579–E6588.
14 Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W & Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47.
15 Love MI, Huber W & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550.
16 Robinson MD, McCarthy DJ & Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140.
17 Eckel-Mahan K, Patel V, de Mateo S, Orozco-Solis R, Ceglia N, Sahar S, Dilag-Penilla S, Dyar K, Baldi P & Sassone-Corsi P (2013) Reprogramming of the circadian clock by nutritional challenge. *Cell* **155**, 1464–1478.
18 Quagliarini F, Mir AA, Balazs K, Wierer M, Dyar KA, Jouffe C, Makris K, Hawe J, Heinig M, Filipp FV *et al.* (2019) Cistromic reprogramming of the diurnal glucocorticoid hormone response by high-fat diet. *Mol Cell* **76**, 531–545.e5.
19 Tognini P, Murakami M, Liu Y, Eckel-Mahan KL, Newman JC, Verdin E, Baldi P & Sassone-Corsi P (2017) Distinct circadian signatures in liver and gut clocks revealed by ketogenic diet. *Cell Metab* **26**, 523–538.e5.
20 Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP & Tamayo P (2015) The molecular signatures database hallmark gene set collection. *Cell Syst* **1**, 417–425.
21 Pei JF, Li XK, Li WQ, Gao Q, Zhang Y, Wang XM, Fu JQ, Cui SS, Qu JH, Zhao X *et al.* (2019) Diurnal oscillations of endogenous H2O2 sustained by p66Shc regulate circadian clocks. *Nat Cell Biol* **21**, 1553–1564.
22 Goldberg EL, Molony RD, Kudo E, Sidorov S, Kong Y, Dixit VD & Iwasaki A (2019) Ketogenic diet activates protective γδ T cell responses against influenza virus infection. *Sci Immunol* **4**, eaav2026.
23 Reczek CR & Chandel NS (2017) The two faces of reactive oxygen species in cancer. *Annu Rev Cancer Biol* **1**, 79–98.
24 Ananthasubramaniam B, Diernfellner A, Brunner M & Herzel H (2018) Ultradian rhythms in the transcriptome of *Neurospora crassa*. *iScience* **9**, 475–486.
25 Kervezee L, Cuesta M, Cermakian N & Boivin DB (2018) Simulated night shift work induces circadian misalignment of the human peripheral blood mononuclear cell transcriptome. *Proc Natl Acad Sci USA* **115**, 5540–5545.
26 Singer JM & Hughey JJ (2019) LimoRhyde: a flexible approach for differential analysis of rhythmic transcriptome data. *J Biol Rhythms* **34**, 5–18.
27 Parsons R, Parsons R, Garner N, Oster H & Rawashdeh O (2019) CircaCompare: a method to estimate and statistically support differences in mesor, amplitude and phase, between circadian rhythms. *Bioinformatics* **36**, 1208–1212.
28 Ding H, Meng L, Liu AC, Gumz ML, Bryant AJ, Mcclung CA, Tseng GC, Esser KA & Huo Z (2021) Likelihood-based tests for detecting circadian rhythmicity and differential circadian patterns in transcriptomic applications. *Brief Bioinform* bbab224. https://doi.org/10.1093/bib/bbab224
29 Moškon M (2020) CosinorPy: a python package for cosinor-based rhythmometry. *BMC Bioinform* **21**, 485.
30 Weger BD, Gobet C, David FPA, Atger F, Martin E, Phillips NE, Charpagne A, Weger M, Naef F & Gachon F (2021) Systematic analysis of differential rhythmic liver gene expression mediated by the circadian clock and feeding rhythms. *Proc Natl Acad Sci USA* **118**, e2015803118.

31 De los Santos H, Bennett KP & Hurley JM (2021) MOSAIC: a joint modeling methodology for combined circadian and non-circadian analysis of multi-omics data. *Bioinformatics* **37**, 767–774.
32 Baker M (2016) 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454.
33 Nieuwenhuis S, Forstmann BU & Wagenmakers EJ (2011) Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat Neurosci* **14**, 1105–1107.
34 Makin TR & Orban de Xivry JJ (2019) Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife* **8**, e48175.
35 R Core Team (2020) R: A Language and Environment for Statistical Computing.
36 Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* **41**, D991–D995.
37 Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* **33**, e175.
38 Carvalho BS & Irizarry RA (2010) A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363–2367.
39 Patro R, Duggal G, Love MI, Irizarry RA & Kingsford C (2017) Salmon provides fast and bias aware quantification of transcript expression. *Nat Methods* **14**, 417–419.
40 Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766–D773.
41 Soneson C, Love MI & Robinson MD (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521.
42 Liu R, Holik AZ, Su S, Jansz N, Chen K, Leong HS, Blewitt ME, Asselin-Labat ML, Smyth GK & Ritchie ME (2015) Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res* **43**, e97.
43 Yu G, Wang LG, Han Y & He QY (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287.
44 Kanehisa M, Furumichi M, Tanabe M, Sato Y & Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353–D361.
45 Dolgalev I (2020) msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Venn diagram analysis (VDA) of two artificial circadian studies with a minimum amplitude requirement for rhythmic transcripts.

Fig. S2. Precision-recall performance of different *compareRhythms* implementations of hypothesis testing applied to the second scenario.

Fig. S3. Comparison of the DiffR hits in response to HFD predicted by different approaches including VDA in the original study.

Fig. S4. Comparison of the DiffR hits called in response to KD by different approaches including VDA in original study.

Fig. S5. Comparison of the DiffR hits called in response to *p66Shc* KO by different approaches including VDA in the original VDA study.

Table S1. DiffR analysis results of the effect of high fat diet on the liver circadian transcriptome.

Table S2. DiffR analysis results of the effect of ketogenic diet on the liver circadian transcriptome.

Table S3. DiffR analysis results of the effect of p66Shc KO on the liver circadian transcriptome.

Appendix S1. The R code and associated data to perform all the analyses in the manuscript.