RESEARCH ARTICLE

CIRCUST: A novel methodology for temporal order reconstruction of molecular rhythms; validation and application towards a daily rhythm gene expression atlas in humans

Yolanda Larriba[1,2]*, **Ivy C. Mason**[3,4], **Richa Saxena**[4,5,6,7], **Frank A. J. L. Scheer**[3,4,8], **Cristina Rueda**[1,2]

1 Department of Statistics and Operational Research, University of Valladolid, Valladolid, Spain, **2** Mathematics Research Institute of the University of Valladolid, University of Valladolid, Valladolid, Spain, **3** Medical Chronobiology Program, Division of Sleep and Circadian Disorders, Departments of Medicine and Neurology, Brigham and Women's Hospital, Boston, Massachusetts, United States of America, **4** Division of Sleep Medicine, Harvard Medical School, Boston, Massachusetts, United States of America, **5** Center for Genomic Medicine and Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America, **6** Division of Anesthesia, Harvard Medical School, Boston, Massachusetts, United States of America, **7** Program in Medical and Population Genetics, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, United States of America, **8** Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, United States of America

* yolanda.larriba@uva.es

Abstract

The circadian system drives near-24-h oscillations in behaviors and biological processes. The underlying core molecular clock regulates the expression of other genes, and it has been shown that the expression of more than 50 percent of genes in mammals displays 24-h rhythmic patterns, with the specific genes that cycle varying from one tissue to another. Determining rhythmic gene expression patterns in human tissues sampled as single time-points has several challenges, including the reconstruction of temporal order of highly noisy data. Previous methodologies have attempted to address these challenges in one or a small number of tissues for which rhythmic gene evolutionary conservation is assumed to be preserved. Here we introduce CIRCUST, a novel CIRCular-robUST methodology for analyzing molecular rhythms, that relies on circular statistics, is robust against noise, and requires fewer assumptions than existing methodologies. Next, we validated the method against four controlled experiments in which sampling times were known, and finally, CIRCUST was applied to 34 tissues from the Genotype-Tissue Expression (GTEx) dataset with the aim towards building a comprehensive daily rhythm gene expression atlas in humans. The validation and application shown here indicate that CIRCUST provides a flexible framework to formulate and solve the issues related to the analysis of molecular rhythms in human tissues. CIRCUST methodology is publicly available at https://github.com/yolandalago/CIRCUST/.

Citation: Larriba Y, Mason IC, Saxena R, Scheer FAJL, Rueda C (2023) CIRCUST: A novel methodology for temporal order reconstruction of molecular rhythms; validation and application towards a daily rhythm gene expression atlas in humans. PLoS Comput Biol 19(9): e1011510. https://doi.org/10.1371/journal.pcbi.1011510

Editor: Christian I. Hong, University of Cincinnati College of Medicine, UNITED STATES

Received: February 12, 2023

Accepted: September 12, 2023

Published: September 28, 2023

Data Availability Statement: Data are publicly available at:https://www.gtexportal.org/home/datasets CIRCUST methodology is publicly available at https://github.com/yolandalago/CIRCUST/.

Funding: Y.L. gratefully acknowledges the financial support received by the IMFAHE Foundation for the research stay in the Division of Sleep Medicine of Harvard Medical School. C.R. and Y.L gratefully acknowledge the financial support received by the

Spanish Ministerio de Ciencia e Innovaci´on [PID2019-106363RB-I00]. I.C.M. thanks the support received from the National Institute of Health (NIH) for grants R01 HL140574 and T32 HL7901-20 and the American Heart Association for grant 19POST34380188. R.S. gratefully acknowledges the financial support received by the NIH for grants R01-DK102696, R01-DK105072, R01-DK107859, and R01-HL146751.F.A.J.L.S. thanks the NIH for the support received for grants R01-DK102696, R01-DK105072, R01-HL140574, and R01-HL153969. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: F.A.J.L.S. served on the Board of Directors for the Sleep Research Society and has received consulting fees from the University of Alabama at Birmingham. F.A.J.L.S. interests were reviewed and managed by Brigham and Women's Hospital and Partners HealthCare in accordance with their conflict of interest policies. F. A.J.L.S. consultancies are not related to the current work. R.S. is a founder of Magnet Biomedicine, which is not related to the current work. Y.L., C.R. and I.M. declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author summary

Rhythmic gene expressions determine tissue-specific functional activity regulating processes such as metabolism, endocrine function, and immune function. Expression patterns of rhythmic genes usually display oscillatory shapes with the timing of the peak closely related to the organ's function. The knowledge of these rhythmic expression patterns may be important in the prevention, diagnosis, and treatment of disease. Yet, human molecular rhythm analysis usually relies on post-mortem samples collected from many people because repeated biopsies within an individual across multiple tissues are impractical. Because the biological time in each donor generally is unknown, the sample order and temporal direction need to be reconstructed. This paper describes and validates CIRCUST, a robust, extensible, and open-source statistical framework to address, separately across organs, the molecular analysis of human rhythms with unknown sampling times; see yolandalago/CIRCUST (github.com). In particular, CIRCUST's application to GTEx, a post-mortem human gene expression database, provides the largest human rhythmic gene expression atlas to date.

Introduction

Circadian clocks orchestrate metabolic, endocrine, and behavioral functions. The molecular clock drives tissue-specific rhythms in gene expression [1]. More than ∼50% of mammalian genes exhibit daily rhythmic expression patterns, although the specific genes that are rhythmic in one tissue may be non-rhythmic in another, and *vice versa*. Based on these fundamental insights, the importance of biological timing has become increasingly recognized in basic research and medicine, with potential implications for the effectiveness of cancer treatments, heart surgery, and pharmacodynamics [2–4]. A comprehensive human temporal atlas of 24-h rhythms in gene expression across tissues is therefore of great potential value. Due to the invasive nature, repeat human biopsies are limited to very few tissues, and human gene expression rhythms across tissues rely critically on human postmortem tissue banks [5, 6]. Indeed, human postmortem gene studies are very valuable in circadian biology [7, 8]. However, there are a number of challenges when trying to reconstruct 24-h molecular rhythmicity from postmortem datasets, where each donor only provides one timepoint, including among others, possible uncertainty regarding the actual time of death, postmortem delay and its effect on RNA degradation [9], or inter-individual differences in the alignment of tissue rhythms relative to local clock-time.

The goal of this paper is to describe, validate and apply a method for the estimation of rhythmicity of gene expression given noisy data in order to build a daily rhythm gene expression atlas in humans from postmortem samples. In particular, our interest is focused on the identification and analysis of tissue-specific molecular rhythms and clock genes phase relationships in the human body. Because of imprecisions in estimations of time of death and/or unknown underlying biological times, the first challenge is to estimate temporal order among the samples. This problem is known as the temporal order estimation problem and addressing this problem was the first step in our analysis.

The temporal order estimation problem can be mathematically formulated as that of looking for an m-dimensional vector that provides what is known as a *circular order* $\boldsymbol{o} = (o_1,. . ., o_m)'$, where m denotes the number of sample collection times to be ordered, see Figs 1 and 2 as illustration. In practice, a circular ordering represents up to $2m$ distinct sample collection time configurations along the 24-h day, depending on the choice of the

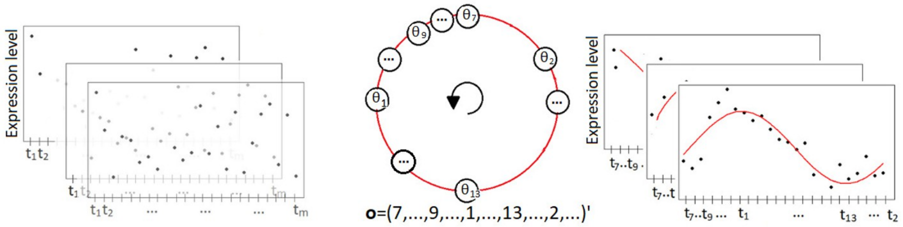


Fig 1. Illustrative outline of CIRCUST solution to temporal order estimation conducted at each tissue. Left: Unordered gene expression data across m samples registered at arbitrary clock times $t_1, \ldots, t_m$ along the 24-h day. Superimposed rectangles are different genes of the tissue. Dots are the gene expression data. Middle: Circular order $\boldsymbol{o}$ obtained from $\boldsymbol{\theta}$, where $0 \preceq \theta_7 \preceq \ldots \preceq \theta_9 \preceq \ldots \preceq \theta_1 \preceq \ldots \preceq \theta_{13} \preceq \ldots \preceq \theta_2 \prec 2\pi$. Starting point and direction are fixed so the assumptions considered are fulfilled. Right: Ordered gene expression data, as a function of CIRCUST estimated times, across m samples registered at clock times $t_1, \ldots, t_m$ along the 24-h day, where $o_j = k \Leftrightarrow t_{(j)} = t_k, \forall j = 1, \ldots, m, k \in \{1, \ldots, m\}$ and $t_{(j)}$ is the j-th element in the vector of ordered timepoints. Superimposed rectangles are the different genes of the tissue. Dots are gene expression data.

https://doi.org/10.1371/journal.pcbi.1011510.g001

starting point and the orientation (clockwise or counter-clockwise). The choice of directionality is not trivial and plays a key role in correctly identifying the timing of biological processes across the day, see the Methods section for details.

This problem has recently garnered a lot of interest within circadian biology, and several methodological approaches have emerged depending on the problem at hand, including Oscope [10], reCAT [11] and CYCLOPS [12] among the most extensively used in practice. Oscope and reCAT were specifically developed to recover cell-cycle dynamics from unsynchronized single-cell transcriptome data, and are highly sensitive to inter-subject variability, as those observed in human gene studies. CYCLOPS, based on a neural network approach overcomes these drawbacks, but it requires evolutionary conservation in mice to establish a consistent set of clock genes from which to estimate reliable orderings. This latter may be inconsistent for particular cases when circadian rhythms are disrrupted [13]. Even so, CYCLOPS has been used widely, however only for one or a limited number of tissues [14, 15]. In the last years, novel methodologies to address this problem have emerged. In [16], we

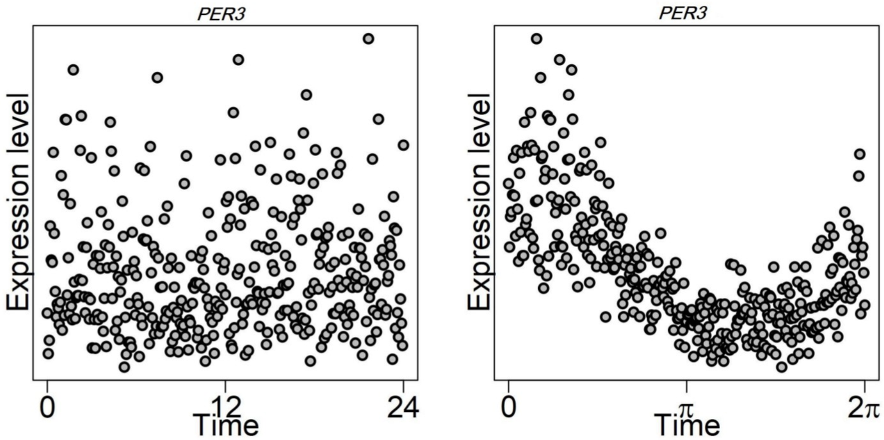


Fig 2. TOD versus CPCA sampling time estimates for gene *PER3* on Skin Not Sun Exposed (Suprapubic) tissue from GTEx. Left: Gene expression as a function of TOD times. Right: Gene expression as a function of CIRCUST estimated times.

https://doi.org/10.1371/journal.pcbi.1011510.g002

introduced a non-parametric framework to mathematically formulate and efficiently solve the temporal order estimation problem without any additional genomic information, but for the case of equally-spaced timepoints, which is an assumption that is not met in postmortem gene studies. More recently, [17] proposed a methodology to infer circadian phases using CHIRAL algorithm [18]. Unlike the previous works, in [17] it is assumed that circadian phases are conserved across tissues, but this algorithm can be also applied to single tissues. Specifically, in [18] is claimed that CHIRAL outperforms CYCLOPS in a human biopsy dataset from skeletal muscle. A comparison of CIRCUST with CYCLOPS and CHIRAL algorithms is outlined in S1 Text (see Section 5).

After estimating temporal order, it is needed to identify tissue-specific molecular rhythms, as well as to assess peak phase relationships across tissues. Several models have been proposed in the literature for the analysis of oscillatory rhythms, referred to hereafter as *rhythmicity models*.

Cosinor [19] is the classical rhythmicity model widely utilized in chronobiology [7, 8, 12, 14]. It is a parametric model that consists of three parameters and captures rhythmic patterns using a sinusoid. Yet, Cosinor may be too rigid for the analysis of transcriptome data exhibiting asymmetric patterns (see Fig 3). Cosinor can be extended to a multi-component model by including multiple sinusoidal harmonics to gain flexibility. Even in this case, it may be unsuitable for the analysis of molecular rhythms as was shown in [20]. In addition, the use of a large number of components may result in serious overfitting issues.

Alternative rhythmicity models have emerged. In [21], and references therein, models based on ordinary differential equations are proposed to describe circadian clock dynamics. However, the type of equations and model parameters are arbitrary and highly dependent on the process under study. Within a non-parametric perspective and in the context of the Order Restricted Inference, we developed ORI, a computationally efficient and versatile model, that formulates rhythmicity (up-down-up pattern) by using mathematical inequalities covering a wide range of rhythmic patterns [16]. However, comparing rhythmic patterns with this model is not straightforward, as it is for parametric models. To overcome these drawbacks, we presented Frequency Modulated Möbius (FMM) model, a flexible five-parametric model that allows deformations to sinusoidal shape to accommodate commonly seen asymmetries in applications (see Fig 3) [20]. This is because FMM is formulated in terms of the phase, an angular variable that represents the intrinsic rhythmicity of the oscillation that periodically repeats every 24-h. Moreover, FMM model parameters are easy to estimate providing meaningful interpretations. An overview of the FMM model is given in S1 Text (see Section 3.1).

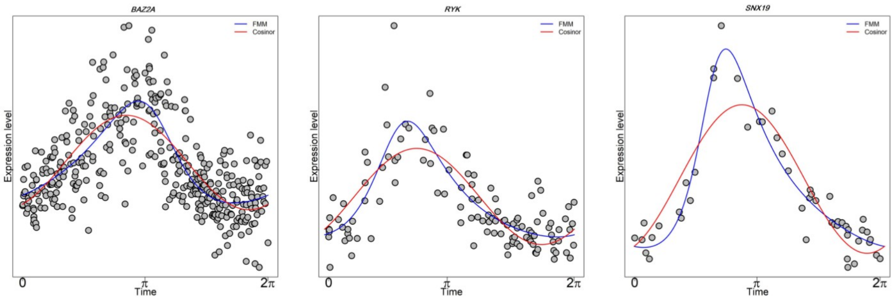


Fig 3. FMM versus Cosinor performance on selected TOP genes from different GTEx tissues. *BAZ2A* (left), *RYK* (middle), and *SNX19* (right) gene expression as a function of CIRCUST estimated times from Lung, Small Intestine, and Kidney, respectively. FMM predictions are shown as blue solid lines. Cosinor predictions are shown as red solid lines.

https://doi.org/10.1371/journal.pcbi.1011510.g003

This work proposes CIRCUST, a general methodology that solves the temporal order estimation problem, as well as identifies and characterizes a wide variety of genes that express 24-h biological rhythmicity, including those with asymmetric expression patterns. The method makes use of the underlying CIRCular structure of the molecular rhythms [16], and the robUSTness of the mathematical procedure to cope with the high noise levels and inter-individual variability that characterize human postmortem gene studies. Specifically, the temporal order reconstruction problem is addressed by a circular dimensionality reduction approach called Circular Principal Component Analysis (CPCA), see Methods section, while the use of the FMM rhythmicity model provides precise estimates of the rhythmicity parameters such as phase.

There is no gold-standard dataset with repeated sampling across multiple human tissues and most human studies have been limited to blood (e.g., [22]) or to another tissue with low sampling frequency [23, 24]. Additionally, inter-individual variability increases uncertainty of estimation of biological timing [25]. This paper shows that CIRCUST is a sound framework based on the analysis of molecular rhythms from four controlled experiments and a simulation study. The first validation dataset consists of human epidermis, a tissue with robust circadian oscillations, repeatedly collected at known and unknown timepoints across a 24-h timeframe from healthy adults [15]. The second one corresponds to a time-labeled human biopsy dataset from skeletal muscle already used for validation in previous works [18]. The third is a postmortem dataset that contains expression data from autopsies in the human prefrontal cortex and well-curated TODs [26]. The fourth validation set consists of a large set of different tissues collected at known timepoints across a 24-h timeframe from baboons, a primate closely related to humans [27]. Next, the Genotype-Tissue Expression (GTEx) dataset, a postmortem gene expression dataset across the largest number of human tissues was independently analysed [28]. GTEx provides annotated times of death (TODs) estimates. However, such TODs may give inaccurate information, see Fig 2 and Fig A in S1 Text, because of the large inter-individual differences in the timing of the central circadian pacemaker, even in healthy patients [29–31]. CIRCUST was conducted on GTEx towards developing an atlas of human 24-h expression rhythms across a wide range of tissues that may provide novel insights into the molecular clock networks.

Materials and methods

The CIRCUST methodology includes reconstruction of temporal order followed by estimation of rhythmic parameters. The details are described below.

CIRCUST solution to temporal order estimation

For each tissue, CIRCUST addresses temporal order reconstruction based on Circular Principal Component Analysis (CPCA), a simple and efficient approach to the sampling time estimation problem. CPCA is a nonlinear dimensionality reduction method that describes the potential circular structure of the molecular rhythms by its projection onto the unit circle [32, 33]. CPCA is often computed from a sub-matrix of a reduced number of tissue-specific rhythmic genes, instead of considering the raw gene expression matrix. This latter ensures the preservation of the underlying rhythmicity signature achieving rhythmic eigengene's patterns which are biologically interpretable [34]. Two different sets of rhythmic genes are considered in this paper: a set of 12 well-established seed genes for an early stage; and subsets of tissue-specific markedly rhythmic genes, called TOP rhythmic genes, at later stages, see below for details.

The CPCA solution starts with the computation of two *eigengenes* from a sub-matrix of rhythmic gene expressions following the lines described in S1 Text (see Section 3.2). Eigengenes are gene-like expression patterns across samples obtained as a linear combination of the expressions in the matrix [34]. Despite the initial unordered expression patterns of these two eigengenes, its mapping reveals the underlying circular structure over the samples, as is illustrated in Section S1 Text (see Section 3.2). Next, eigengenes are projected onto the unit circle $[0, 2\pi)$, computing the arctan of these projections which allows defining the angular vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ that represents the temporal position of the m samples in the raw gene expression sub-matrix onto the unit circle. The increased order of these angles sets the circular order $\boldsymbol{o} = (o_1, \ldots, o_m)'$ which provides a circular arrangement of the timepoints. Finally, for the given order, there exist $2m$ sample time configurations according to the starting point and the (clockwise/counterclockwise) direction selection. In general, this choice is made so that two standard assumptions concerning the seed genes' peak phase relations in mammals are fulfilled, see S1 Text (Section 3.2) for details. These assumptions can be user-refined, in terms of peak phases' order restrictions, in case the molecular clock network of the species is (partially) known *prior*, yielding more reliable sampling time estimates. We refer to this particular case as CIRCUST$_{prior}$. Full details regarding temporal order estimation are given in S1 Text (see Section 3.2). Figs 1 and 2 illustrate a CPCA solution to approach the temporal order identification problem.

CIRCUST methodology

Let $[\boldsymbol{R}]$ denote the matrix of raw and unordered expressions data that serves as input. For each tissue, CIRCUST is sequenced as follows. Fig B in S1 Text shows an outline of the methodology.

$$[\boldsymbol{R}] \xrightarrow{Preprocessing} [\boldsymbol{N}] \xrightarrow{\substack{Preliminary\\ Order}} [\boldsymbol{X}] \xrightarrow{\substack{TOP\ \ rhythmic\\ orderings}} [\boldsymbol{X}_k^{TOP}] \xrightarrow{\substack{Robust\\ Estimation}} [\boldsymbol{M}^{TOP}],$$

where $[\boldsymbol{N}]$ is the matrix of preprocessed, normalized (and unordered) expression data. $[\boldsymbol{X}]$ is a preliminary ordered gene expression matrix, and $[\boldsymbol{X}_k^{TOP}]$ is the k–th expression matrix with the ordered expression data of the tissue-specific *TOP* genes, i.e. the highly rhythmical genes of each tissue, $k = 1, \ldots, K$ with K a prefixed integer value (see below). To define these two latter (ordered) matrices the temporal order problem must be addressed. The output of CIRCUST is $[\boldsymbol{M}^{TOP}]$, a matrix that contains robust (Median) of the main FMM parameter estimates computed for the *TOP* genes in $[\boldsymbol{X}_k^{TOP}]$, $k = 1, \ldots, K$. FMM parameters are meaningfully interpretable and characterize rhythmicity, see S1 Text (Section 3.1). CIRCUST steps are described below.

1. ***Preprocessing***
   Genes with zero read counts in more than 30% of samples are discarded [35]. Gene expressions are one by one normalized into [-1, 1] by using a min-max normalization [16]. The preprocessed expression matrix is denoted by $[\boldsymbol{N}]$.

2. ***Preliminary order***
   A core information set consisting of the 12 genes: *PER1, PER2, PER3, CRY1, CRY2, ARNTL, CLOCK, NR1D1, RORA, DBP, TEF* and *STAT3*. In the following, we refer to them as seed genes. There is no a gold-standard for seed genes selection, though gene expression patterns of this choice, generally display marked circadian signals in most of the mammalian tissues and were also considered as circadian benchmarks in previous works [1, 12, 14,

15, 27]. Particularly, the gene *STAT3* is included as it has been identified as rhythmic for several human tissues [36, 37]. The stability of the results regarding seed gene selection has been assessed in S1 Text (see Section 3.4).

The role of CPCA at this point is twofold. CPCA is computed on the sub-matrix of the 12 seed genes from [$\boldsymbol{N}$]. First, CPCA allows detecting outlier samples following the lines described in S1 Text (see Section 3.3). Outliers samples are deleted from all the genes in [$\boldsymbol{N}$], and the expression data are normalized again. Second, CPCA provides a solution for the temporal order identification problem (setting starting point and direction), from the sub-matrix of the 12 seed genes from [$\boldsymbol{N}$], as was detailed above. Then, [$\boldsymbol{N}$] is ordered with regard to the circular order obtained as the solution of CPCA. We refer to this matrix by [$\boldsymbol{X}$]. In case the median of the R^2_{FMM} from the seed genes after preliminary order is lower than 0.3, the subsequent analysis may be inaccurate.

3. ***TOP rhythmic orderings***

Rhythmicity models are used at this stage to predict gene expression patterns. First, the ORI model's [16] computational efficiency allows discarding potentially non-rhythmic genes, with $R^2_{ORI} < 0.5$, in [$\boldsymbol{X}$]. R^2 is a rhythmicity model's goodness of fit measure taking values from 0 to 1; the closer to 1, the higher the rhythmicity. Details are given in S1 Text (see Section 3.5). Then, the tissue-specific *TOP rhythmic genes* are defined, based on the FMM model predictions, as those which are: i) non-spiked ($\hat{\omega} > 0.1$); ii) with the highest rhythmicity ($R^2_{FMM} > 0.5$); and iii) whose peak phases ($\hat{t}_U$) cover all the quarters of the unit circle ($[0, 2\pi)$). This definition results from the meaningful interpretation of the FMM parameters: ω, t_U, see S1 Text (Section 3.1) and [20] for details. The 12 seed genes are usually among the TOP genes, if not, they are forced to be included. [$\boldsymbol{X}^{TOP}$] denotes the sub-matrix of TOP genes once they are filtered from [$\boldsymbol{X}$].

Next, random selections of size 2/3 of the genes in the TOP are considered. CPCA solution for temporal order estimates is recomputed for each of these sub-matrices resulting from filtering the selected genes of [$\boldsymbol{X}^{TOP}$]. The process is repeated until obtaining a prefixed number of K random gene collections verifying that: (a) angular values in $\boldsymbol{\theta}$ are distributed along with more than half of the unit circle; (b) and the maximum distance between two consecutive angular values in $\boldsymbol{\theta}$, does not exceed the observed distances for any pair of consecutive angular values with regard to the preliminary order given by the vector $\boldsymbol{\theta}$ considered in step. Conditions (a) and (b) pursue robustness on peak's estimations and avoid spurious gaps not detected from the seed genes, respectively. Both improve the quality of the orders. Hence, $\boldsymbol{o}_k$, $k = 1, \ldots, K$ circular orders are defined. For each of them, [$\boldsymbol{X}^{TOP}$] is reordered, obtaining [$\boldsymbol{X}^{TOP}_k$], that denotes the k-th matrix of TOP genes ordered by $\boldsymbol{o}_k$, $k = 1, \ldots, K$.

4. ***Robust Estimation***

FMM predictions for the TOP genes in [$\boldsymbol{X}^{TOP}_k$], $k = 1, \ldots, K$, are computed. For each gene at the TOP, there are K FMM parameter estimates, and K rhythmicity measures (R^2_{FMM}). Robust FMM parameter estimates, in terms of the medians, are computed. [$\boldsymbol{M}^{TOP}$] is the matrix that contains for the genes in the TOP the median of the FMM features: R^2, t_U and ω which are key to assess and compare rhythmicity across tissues.

Results

In this section, CIRCUST is validated and compared to related methods across four different studies. The first and second experiment compares CIRCUST performance with CYCLOPS and CHIRAL on human biopsy timed sample studies, respectively. Next, CIRCUST is applied to autopsy (postmortem) data from the human cortex comparing the results with those obtained using well-curated TODs. Moreover, CIRCUST flexibility is shown on timed sample baboons dataset covering a large number of tissues. In addition to that, we conducted a simulation study to assess CIRCUST performance when faced with symmetric and asymmetric patterns, non-rhythmic confounds, and equally and non-equally sampling distributions, see S1 Text (Section S4). Finally, we illustrate the application of CIRCUST to GTEx towards developing a daily rhythm gene expression atlas in humans.

CIRCUST validation on human biopsies epidermis samples

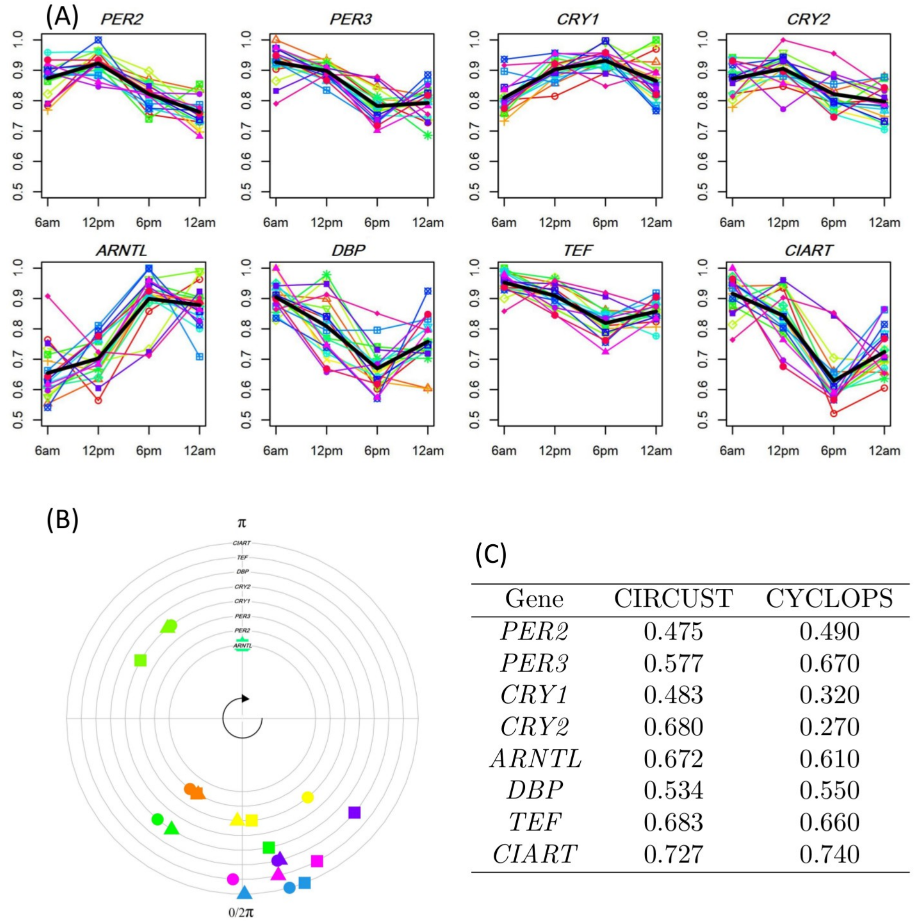
This validation relies on the hybrid human gene expression dataset from epidermis tissue (GEO accession number GSE139301) [38]. On the one hand, this dataset contains gene expressions for a set of 19 participants for which biopsies were collected at 6 am, 12 pm, 6 pm, and 12 am. On the other hand, it includes the gene expression for 533 epidermis samples for which sample collection times were unrecorded.

We apply CIRCUST on the set of 533 unordered samples in order to compare the results with those obtained for the 19 participants where clock times were known. This latter mimics what was done in [38] to validate CYCLOPS. For comparison purposes, the analyses refer to the set of clock-associated genes in [38] which are among those at the TOP genes of CIRCUST for epidermis tissue see Fig 4. Gene expression data from biopsies at the four timepoints for the 19 participants are displayed in Fig 4A, see thin color lines. Due to the low sampling frequency and the noise inherent in the experiment, for each gene, the averaged expression pattern is computed (blue thick line). FMM model predictions for the average expression patterns are computed, assuming that the estimated FMM peaks ($\hat{t}_U$) as the peak phase values for these genes. Fig 4B compares these peak values (triangles) derived from the 19 participants, where clock times are known, with the estimated peak phases derived from CIRCUST (circles) and CYCLOPS (squares) for the 533 samples, when sampling times are unknown. The circular correlation [39] between participant peak phases and estimated phases from CIRCUST and CYCLOPS for these genes are 0.862 and 0.819, respectively, revealing coherence between the CIRCUST peak phases' estimates and the peak phases given for known clock times. In particular, the differences between CIRCUST and peak phases from the participants are, in general, less than ~ 2 hours (~ 0.52 radians), being especially low for the genes *PER2*, *CRY1*, *CRY2* and *Ddp*. Except for *PER3*, such differences tend to be lower for the CIRCUST than for the CYCLOPS. Moreover, the orders among peak phases defined by CIRCUST match with those observed from the human biopsies across the 19 participants: {*PER3, DBP, TEF, CIART*}$\preceq$ {*CRY2, PER2*}$\preceq${*CRY1, ARNTL*}$\preceq${*PER3, DBP, TEF, CIART*}, $\preceq$ is read as "before than". Finally, the rhythmicity measures R^2_{FMM} for the eight TOP genes (see Fig 4C) appear to be more consistent with the oscillatory expression patterns observed in Fig 4A than those given by CYCLOPS in [38], with *CIART, TEF* or *ARNTL* being among those that display the strongest rhythmicity.

CIRCUST validation on human biopsies skeletal muscle samples

For this validation, we examine a set of 54 labeled samples from human skeletal muscle biopsies around 24 hours. Specifically, biopsies were taken every 4 hours from 10 healthy volunteers. Sample collection was performed under controlled laboratory conditions. A total of 13377 genes were quantified. The dataset is publicly available, one may refer to [40] for further details.

The rationale behind the choice of this dataset for CIRCUST validation is that is exactly the dataset proposed by [18] to validate and compare CHIRAL's performance against CYCLOPS



Gene	CIRCUST	CYCLOPS
PER2	0.475	0.490
PER3	0.577	0.670
CRY1	0.483	0.320
CRY2	0.680	0.270
ARNTL	0.672	0.610
DBP	0.534	0.550
TEF	0.683	0.660
CIART	0.727	0.740

Fig 4. CIRCUST consistency for human epidermis (GSE139301) dataset. (A): Expression patterns across the 19 participants, obtained from GEO accession number GSE139301 [38], for the eight genes under analysis at four biopsies times. Thin color lines represent participants' gene expression. Thick black lines represent the average expression profile across participants. (B): CIRCUST peak phase estimates (with the 533 samples) compared to CYCLOPS estimates using the peaks from 19 participants as reference for the eight genes under analysis. Participant peak phases (triangles), estimated phases from CIRCUST (circles) and CYCLOPS (squares). Biopsies times given along the 24-h interval are read into $[0, 2\pi)$. For comparison purpose, π is fixed at *ARNTL*'s peak. (C): The goodness of fit measures for predicted expression patterns of the eight genes under analysis when CIRCUST and CYCLOPS are applied to the 533 samples. The measurements considered are R^2_{FMM} for CIRCUST and *rsq* [38] for CYCLOPS. The higher values correspond to stronger rhythmicity, but the scales are different.

https://doi.org/10.1371/journal.pcbi.1011510.g004

on solving the temporal order estimation problem. CHIRAL algorithm implementation was supported by using the CHIRAL R package publicly available at GitHub (https://github.com/naef-lab/CHIRAL/tree/master/Pkg/CHIRAL). For a fair comparison, we used CIRCUST seed gene set in both cases as well as the validation metrics considered in [18].

Results shown in Fig 5 display a similar and competitive performance of both algorithms. Both methods underestimate the times of the two first time samples. While for the rest of the sample times, CIRCUST presents slightly more accurate results. This latter is confirmed by two metrics, also employed in [18], to evaluate temporal order reconstruction which are the circular correlation between the true and the estimated order and the median absolute deviance (MAD) of the estimations with regard to the true times. The circular correlation is 0.752 in CHIRAL and 0.798 in CIRCUST; while the MAD is 0.872 in CIRCUST and 0.875 in

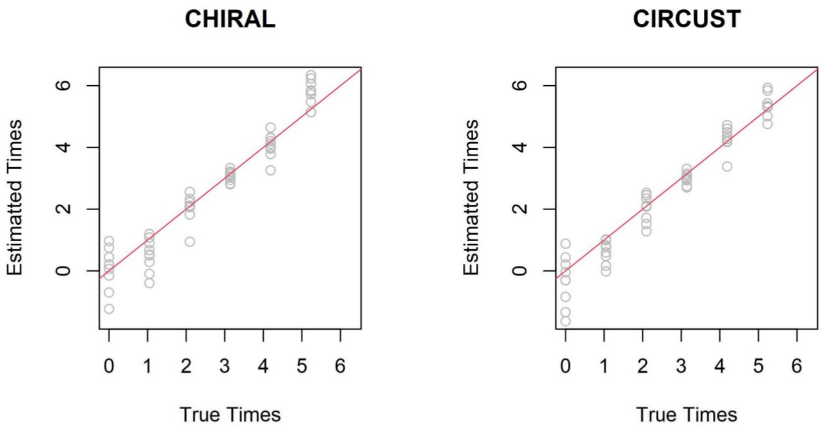


Fig 5. Comparison of CHIRAL and CIRCUST on human muscle biopsies. True times are plotted in X-axes. Estimated phases on Y-axes were adjusted for illustrative purposes in both algorithms using the functions given in CHIRAL R package.

https://doi.org/10.1371/journal.pcbi.1011510.g005

CIRCUST. These numbers provide a slightly better performance of the CIRCUST against CHIRAL algorithm. In [18] CHIRAL outperformed CYCLOPS using this same dataset and metrics.

CIRCUST validation on human postmortem brain samples

This validation considers a well-curated time-stamped samples autopsy gene expression dataset from the human prefrontal cortex (Brodmann's area (BA) 47) with GEO accession number GSE71620) [26]. This dataset contains expression data, TOD, and demographic variables across 210 subjects. Despite being a postmortem dataset, in this case, well-annotated TODs are provided [26].

CIRCUST is applied on the set of 210 autopsy samples in order to confirm its ability to order clinical collections with realistic non-circadian confounds. The results compare seed and TOP temporal expressions ordered by CIRCUST to those directly obtained from TOD ordering. First, a simple visual inspection of Fig 6A and 6B illustrates the similarity among the temporal patterns of the seed genes for both procedures. Specifically, the rhythmicity signature, measured in terms of R^2_{FMM}, is maintained across orderings being generally higher in CIRCUST. Moreover, there is a concordance among the genes with the higher and lower R^2 of both orders. For example, *PER1, PER2* and *PER3* are among those with the highest R^2_{FMM}, while *CLOCK* and *RORA* are among those with the lowest ones (see Fig 6A and 6B). In addition, CIRCUST and TOD display a high concordance on peak estimates for the seed genes, see Fig 6C. In particular, the difference between TOD and CIRCUST for the seed genes is, on median, lower than 1.671 hours. Finally, CIRCUST identifies a set of 70 TOP genes in BA47. Among these genes, there were found well-known rhythmic genes such as *CIART* or *NIFL3* displaying both sinusoidal and asymmetric patterns (see Fig 6D). It is worth noting that these results are consistent with those given in [26], regardless of CIRCUST synchronization. CIRCUST was also conducted on cortex region BA11 obtaining similar results.

CIRCUST validation on multiple baboons tissues

The fourth validation is driven by the baboon gene expressions dataset (GEO accession number GSE98965). Data were collected, under controlled conditions, every 2 hours (ZT0,

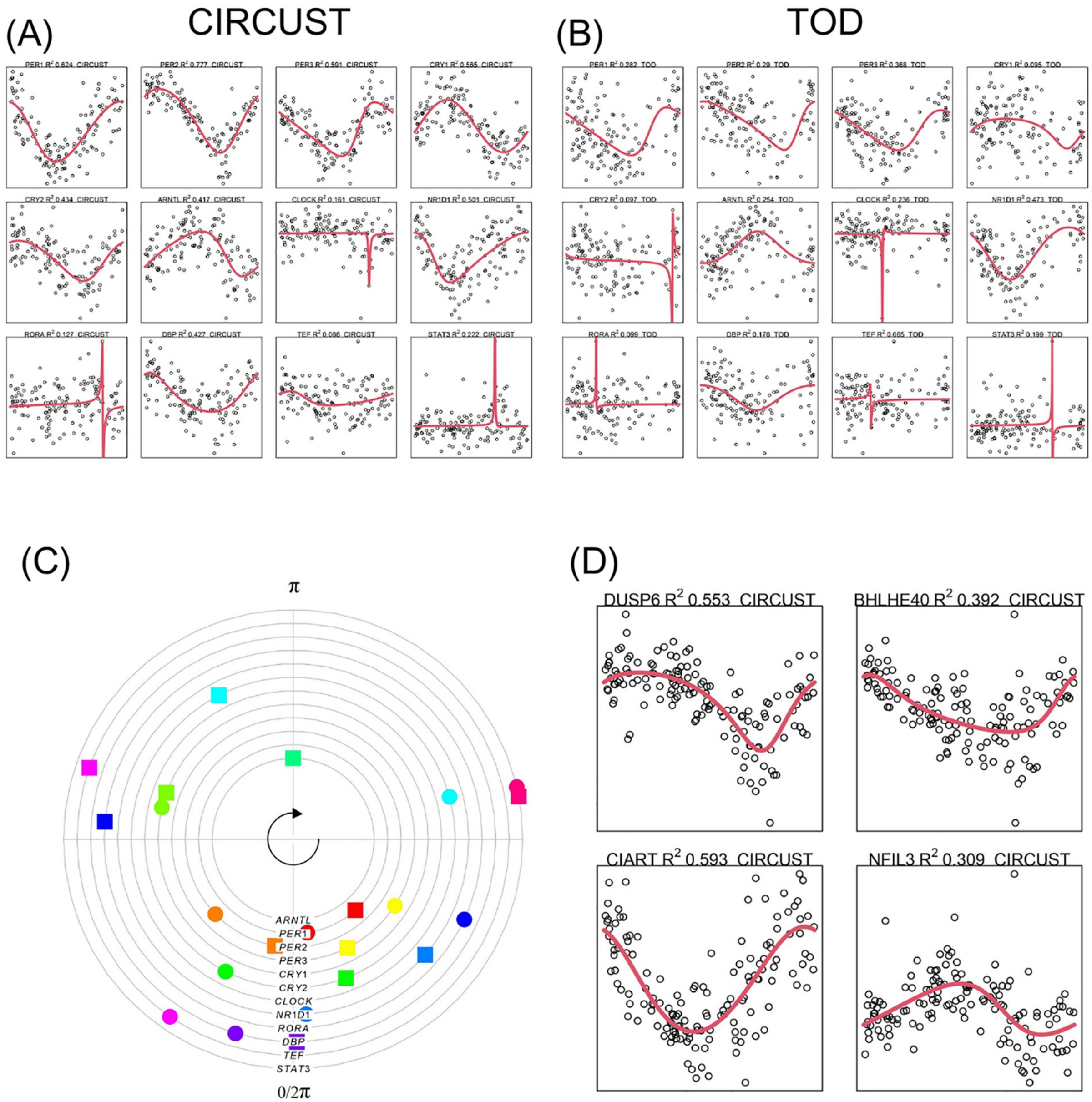


Fig 6. CIRCUST and TOD performance in autopsies dataset (GSE71620). (A) and (B) show seed genes reordered according to CIRCUST and TOD orderings, respectively. FMM fitting is displayed in red. R^2_{FMM} is given in the title. (C) Peak seed estimates according to CIRCUST (circle) and TOD (square) orders. (D) Selected TOP different rhythmic genes with different patterns derived from CIRCUST.

https://doi.org/10.1371/journal.pcbi.1011510.g006

ZT2,. . ., ZT22) over the 24-h day across 64 different tissues, which are aggregated into 13 functional groups [27]. In order to guarantee the consistency of the results, analyses are restricted to the 47 baboons' tissues for which the rhythmicity measure (R^2_{FMM}) for the 12 seed genes is, on average, higher than 0.7, see Table A in S1 Text for details. Among these tissues, there are representatives of 12 out of 13 of the functional groups, all except for the male genitals. The baboon is a well-studied mammalian species in circadian biology with well-established prior knowledge regarding its molecular clock network. $CIRCUST_{prior}$ allows incorporating such information into the method in terms of order peak relationships (inequalities) improving its

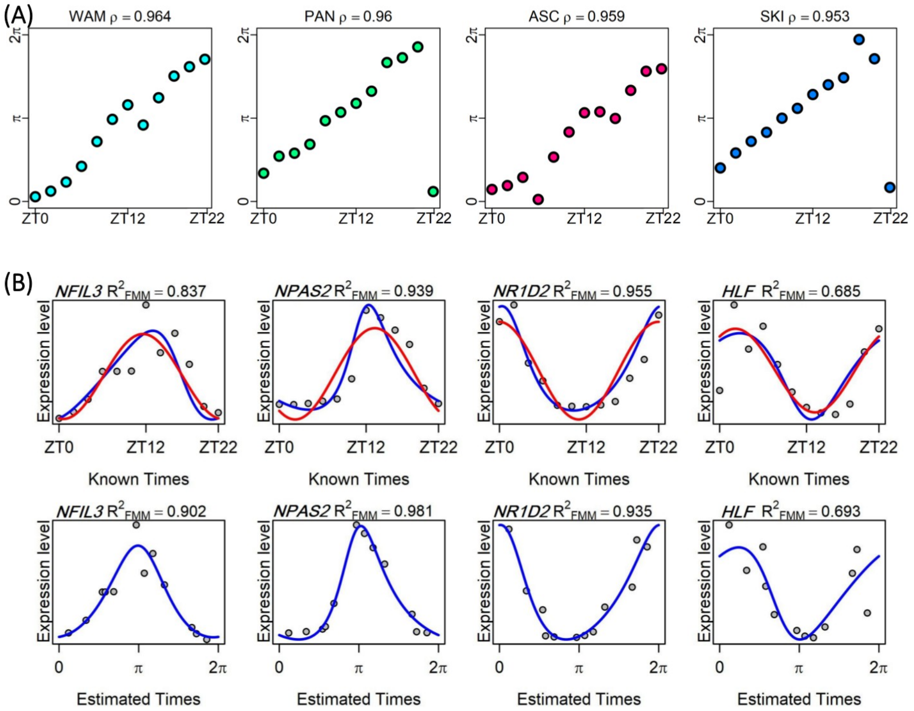


Fig 7. CIRCUST validation based on baboon dataset (GSE98965). (A) Estimated ($[0, 2\pi)$) vs clock times (ZT0, ZT2,. . .ZT22) in baboons' tissues (GSE98965): White Adipose Mesenteric (WAM), Pancreas (PAN), Ascending Colon (ASC) and Skin (SKI). Horizontal axis: sampling times along 24-h, where ZT0 is the time when light is on and ZT12 is when light is off [14]. Vertical axis: CIRCUST estimated times in $[0, 2\pi)$, where 0 is the time when light is on and π is when light is off. Time 0-h is the same as 24-h and phase 0 is the same as 2π. The diagonal line observed for most of the tissues is used as a marker of the coherence between the orders. (B) Expression of selected clock genes *NFIL3*, *NPAS2*, *NR1D2* and *HLF* in Pancreas (PAN) tissue from baboons (GSE98965). Top panels: expressions as a function of known times ZT0, ZT2,. . .,ZT22. Bottom panels: expressions as a function of CIRCUST estimated times. FMM predictions are shown as blue solid lines. Cosinor predictions are shown as red solid lines.

https://doi.org/10.1371/journal.pcbi.1011510.g007

performance (see Methods section). Specifically, [1, 12, 41, 42] reported that baboons' peak phases usually fulfil: $\{DBP\} \preceq \{CRY1, CRY2\} \preceq \{ARNTL\}$ or $\{Nrd1\} \preceq \{PER1, PER2, PER3\} \preceq \{ARNTL\}$. In case one of the relationships above increases the number of seed genes with their peaks within the active period ($[0, \pi)$) with regard to the standard order peak time assumption (2), described in S1 Text (see Subsection 3.2), it will be replaced by the specific relation given for baboons.

Circular association between CIRCUST estimated times in $[0, 2\pi)$ and the real times (ZT0, ZT2,. . ., ZT22) along the periodic scale of 24-h, which can be represented as points on a circle, is assessed. Both variables can be considered as angular, then a circular-circular regression problem [43], similar to the linear regression when both variables are euclidean, is solved. For each tissue, the goodness of fit measure ρ, defined as an analog of residual sums of squares in a linear regression model, is computed to assess the coherence among both orders [44, 45]. A closer ρ to 1 indicates a better correspondence between the orders. $CIRCUST_{prior}$ performs well in ordering the samples across the 47 tissues, see Fig C in S1 Text. The interquartile boundary (P_{25}, P_{75}) for the values of ρ across the 47 baboons' tissues is: $(P_{25}, P_{75}) = (0.729, 0.895)$, see Table A in S1 Text for details. The estimated order is very close to the real temporal order for highly rhythmic organs such as White Adipose ($\rho = 0.964$); Pancreas ($\rho = 0.960$); Colon ($\rho = 0.959$); or Skin ($\rho = 0.953$), see Fig 7A. In addition, Fig 7B and Figs D, E, F and G in

S1 Text reveal, that CIRCUST conserves rhythmicity across selected genes for the four tissues mentioned above. From mere visual inspection, the gene expression patterns in the baboons at times ZT0, ZT2,. . .ZT22, (top panels of Fig 7B) are closely tracked by expressions obtained as a function of CIRCUST estimated times for these same genes (bottom panels of Fig 7B). Moreover, these plots show that the FMM model accommodates a wide variety of rhythmic patterns with high (close to 1) and similar rhythmicity strength values quantified by R^2_{FMM}, across the selected clock genes, even in those with non-sinusoidal gene pattern, e.g. *NPAS2* in Fig 7B and more in Figs D, E, and F in S1 Text.

CIRCUST application to GTEx

This section reports the analysis of the molecular rhythms and clock network from GTEx (V7) database. Only tissues with more than 40 samples were included in the analysis. In addition, two cell lines and thirteen brain tissues were discarded [28, 46]. Cell lines may not capture the molecular complexity of the tissue [47]; the brain tissues usually evince intra-tissue heterogeneity and they are often considered as independent molecular networks [48, 49]. Hence, the CIRCUST methodology was separately applied to 34 tissues with a fixed number $K = 5$ of random selections given from the genes at the TOP for each of the tissues. With this criteria, our analysis considers 621 donors characterized by a mix of ages, sex, and health status (see Table B in S1 Text). Specifically, the results below report, for each tissue, the analyses of the medians of the FMM estimated parameters (R^2_{FMM}, $\hat{t}_U$ and $\hat{\omega}$) of the TOP genes obtained as outputs (at Step 4) from CIRCUST, see the Methods section for details.

GTEx Molecular rhythm analysis. The molecular rhythms for the TOP genes in each of the 34 tissues from GTEx were analyzed. TOP genes, defined in the Methods section, display non-spike, and heterogeneous rhythmic patterns whose peaks are distributed along the 24-h day, as is seen in Fig 3. The number of TOP genes varies among the analyzed tissues (see Fig H in S1 Text). Muscle-Skeletal, Testis, and Lung are among the tissues with the highest number of TOP genes; while Pancreas or Thyroid are among those with a lower number of them. Moreover, most of the TOP genes belong to non-intersecting sets (see Fig I in S1 Text). In particular, for Artery-Tibial and Nerve-Tibial, which are the tissues with the highest number of TOP genes, there are only 5.32% (5 out of 94) shared between both tissues, apart from the 12 seed genes considered. Moreover, in other rhythmic organs like Testis, 81.61% (71 out of 87) of the genes at the TOP are exclusively rhythmic of this tissue. These latter findings evince tissue-specific rhythmicity in human gene studies.

The heterogeneity observed regarding rhythmicity persists even for seed genes. Fig 8 illustrates R^2_{FMM} distribution for the genes in the TOP of the 34 organs analyzed. The R^2_{FMM} of the 12 seed genes are shown as different coloured dots. As seen, seed genes do not always rank among the most highly rhythmic genes of the tissue. Even when analyzing highly rhythmic organs, several scenarios are shown. For example, in Kidney-Cortex, most of the seed genes are distributed among the TOP genes. On the contrary, the seed genes in Whole-Blood are not among the TOP genes of this tissue. This latter does not mean that the seed genes are not rhythmic, but that there are other circadian genes among those in the TOP that, regarding the tissue-specific variability, present a stronger rhythmic signature. In general, for the vast majority of the tissues, at least a quarter of the seed gene expression oscillations persist across wide inter-individual variability, with seed genes such *PER3* being among those with the highest rhythmicity for more than the 75% of the tissues. This observation supports CIRCUST's potential to detect novel tissue-specific molecular rhythms in humans, such as *Snx19* in the Kidney, see Fig 3.

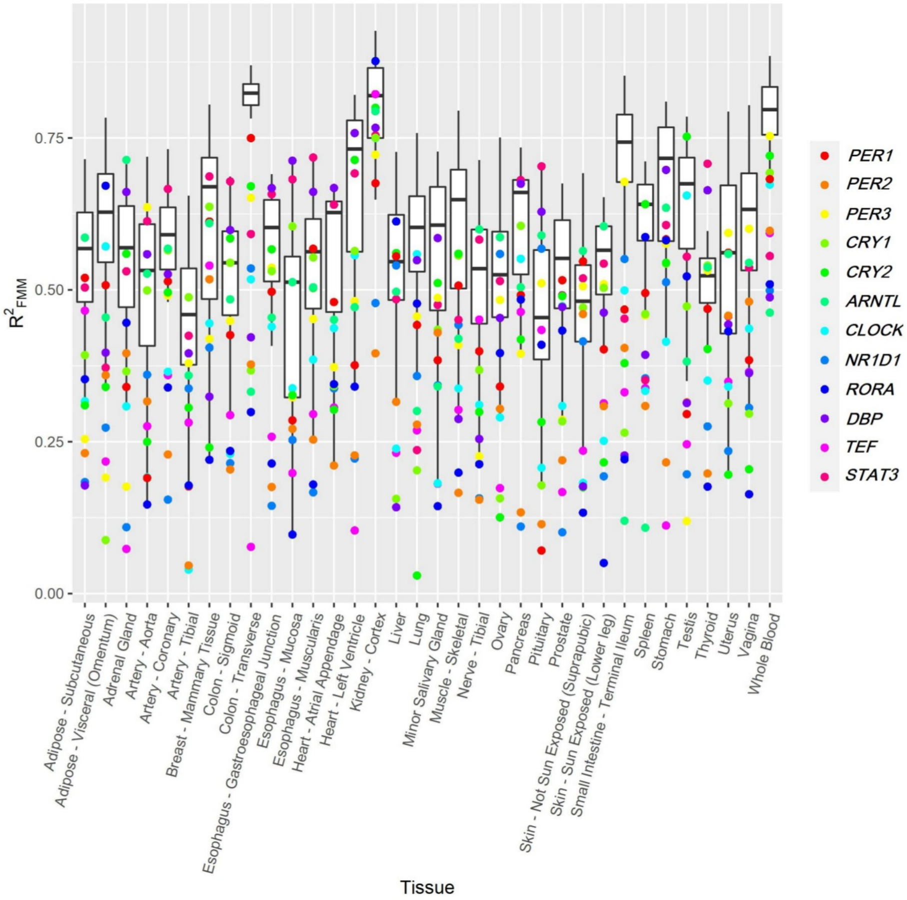


Fig 8. R^2_{FMM} distribution across the 34 organs in GTEx. Dots denote the R^2_{FMM} of the 12 seed genes. Each seed gene is represented using a different color. Tissues are alphabetically sorted.

https://doi.org/10.1371/journal.pcbi.1011510.g008

Finally, the atlas of robust human molecular rhythms for the 34 human tissues is provided in the S1 Data. For each tissue, the atlas includes the list of TOP genes, ranked from the highest to the lowest rhythmicity, based on the rhythmicity measure (R^2_{FMM}), the estimated amplitude ($\hat{A}$), peak phase ($\hat{t}_U$), and the timing of the peak phase relative to *ARNTL*: corresponding to active/lightened, if $\hat{t}_U \in [0, \pi)$, or to inactive/darkness $\hat{t}_U \in [\pi, 2\pi)$. All of these values are derived from Step 4 of CIRCUST methodology. This is the largest rhythmic gene characterization across human tissues to date assuming inter-individual tissue variability. This work represents important advances towards a human rhythmic gene expression atlas.

GTEx Molecular clock networks. This section describes and compares the molecular clock networks across the 34 human tissues. The estimated peak phases ($\hat{t}_U$) of the TOP rhythmic genes are assessed and compared across the tissues. Here we present data on molecular clock networks simultaneously analyzed across 34 human tissues [12, 14, 15].

Fig 9 shows the peak phase distributions of the 12 seed genes across the 34 tissues. Non-rhythmic seed genes were discarded from this analysis, see Tables C and D in S1 Text. Distributions varied across organs, but they were not randomly distributed throughout the 24-h day. The peak phase estimates are generally in one or two clusters, with one of them usually preceding the presumed inactive/darkness phase in mammals ($[\pi, 2\pi)$) [14]. For seed genes such as

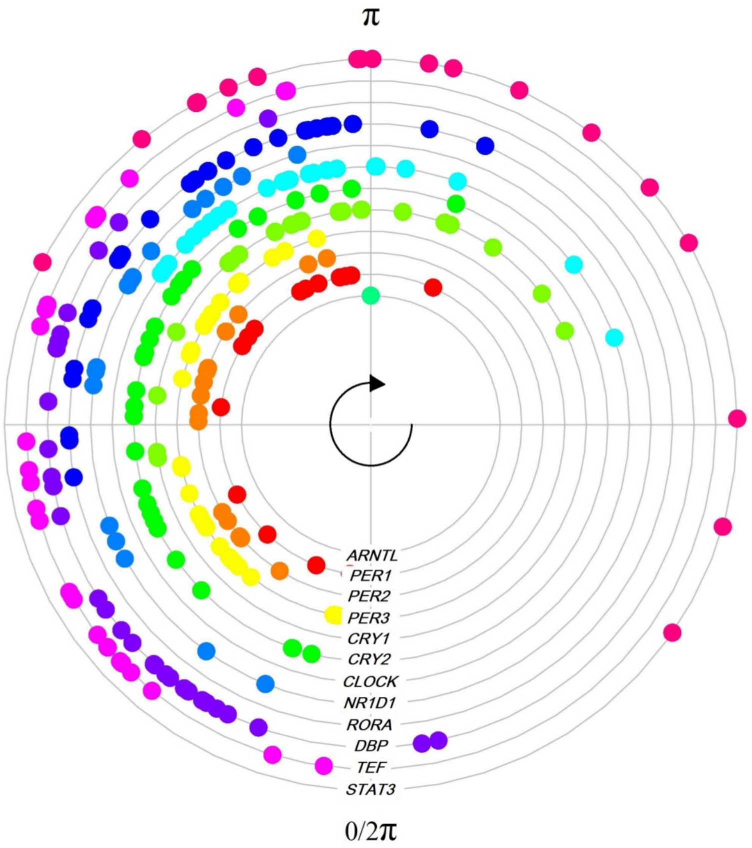


Fig 9. Peak phase estimate distributions of the 12 rhythmic seed genes across the 34 tissues in GTEx. *ARNTL* is known to peak in anticipation of the inactive/darkness $[\pi, 2\pi)$ period in mammals and was set to π for comparisons [14]. The R^2 and t_U estimated values are given in Tables C and D in S1 Text, respectively.

https://doi.org/10.1371/journal.pcbi.1011510.g009

CLOCK or *PER1*, peak phase distribution was mainly restricted to a $\sim$6-hour interval. However, for most of the other seed genes, the peaks were distributed along $\sim$12-hour, matching with the presumed active period ($[0, \pi)$) or light day hours. This reveals human inter-tissue variability and the heterogeneous behavior of the molecular clock networks across the variety of organs analyzed. Moreover, we observe that seed peaks' variability is reduced when the the analyses limit to the highest tissue-specific rhythmic genes. But, for the particular case of highly rhythmic organs such as Skin (epidermis), molecular clock networks are maintained across species, as shown in Fig 10. There, the estimated seed genes' peak phases in the skin for humans, from GTEx database, were similar to those estimated for the baboons and both are close to those obtained as a function of the true clock times.

Finally, the distributions of the peak phases of the TOP rhythmic genes across the human tissues were explored. TOP peaks estimates for nearly all organs display different distributions with one, two, or even three-phase clusters, see Fig 11. In tissues such as Artery-Tibial, Heart tissues, Pancreas, or Stomach, most of the TOP genes peaked within a narrow interval, whereas TOP genes in Colon-Transverse, Spleen, Small Intestine-Terminal Ileum, or Whole-Blood peaked within two distinct time intervals. Three modes are displayed in the Vagina or Testis. Despite human inter-tissue variability, anatomically adjacent tissues showed phased clusters that are temporally close, see for example Esophagus-Gastroesophageal Junction and Colon-

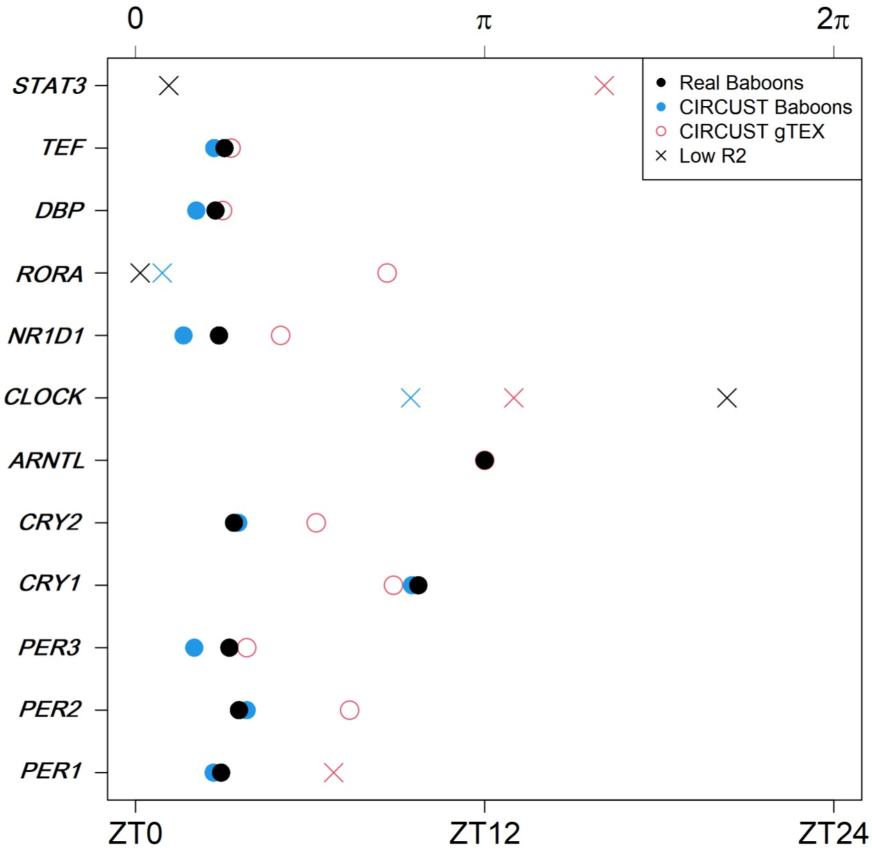


Fig 10. Seed genes' peaks in the human epidermis (Tissue: Skin—Sun exposed from GTEx) and baboon epidermis (Tissue: SKI from GSE98965) tissue. Estimated phases are derived from the FMM model. Black and blue dots match with the peaks obtained as a function of the true clock times and as a function of the CIRCUST estimated times in baboons, respectively. Red dots match with the peaks obtained as a function of the CIRCUST estimated times in the human epidermis GTEx dataset. Non-rhythmic genes ($R^2 < 0.5$) are marked with a cross.

https://doi.org/10.1371/journal.pcbi.1011510.g010

Sigmoid, both of which belong to the digestive tract. A compilation of phases across the TOP rhythmic genes revealed that, for the vast majority of tissues, presumed early afternoon major peak anticipating the inactive phase, and a quiescent zone is also observed for many of the tissues which are considered distinctive features of rhythmicity in the diurnal primate [27].

Discussion

CIRCUST methodology presented in this paper efficiently formulates and solves, based on circular statistics, the temporal order estimation problem arising in gene studies for which the biological time of sample collection is unknown or imprecise. The robustness of CIRCUST against the characteristic noise of postmortem gene studies together with the flexibility of the rhythmicity model FMM expands our knowledge regarding the rhythmicity of genes within and across tissues (see Fig 3). These strengths of methodology give rise to the building of a comprehensive daily rhythm gene expression atlas in humans from the GTEx database that represents rhythmicity analysis across the largest number of human tissues to date assuming inter-individual-tissue variability. Moreover, validation experiments conducted confirm that CIRCUST outperforms CYCLOPS and CHIRAL temporal order estimation algorithms for single tissue analyses.

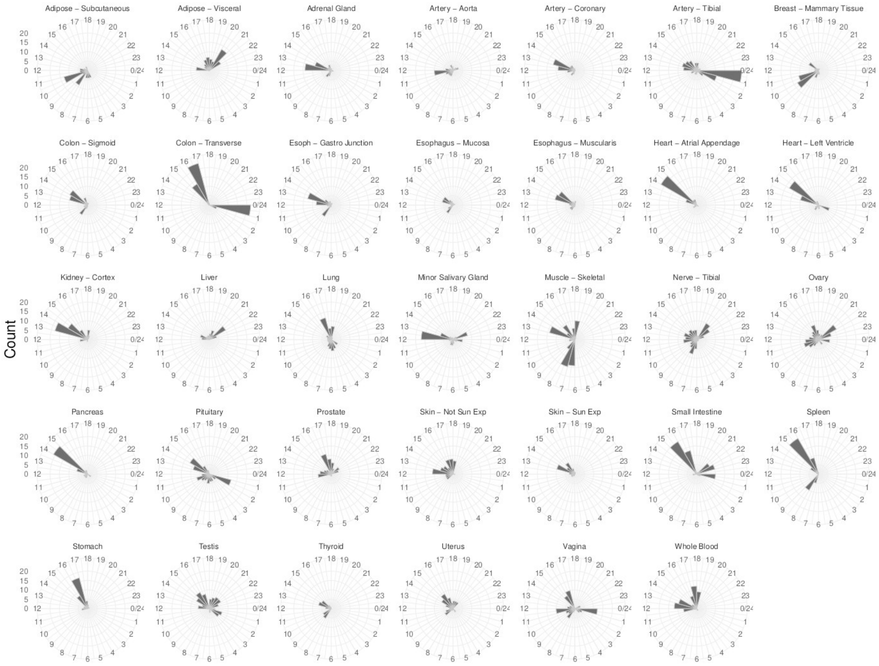


Fig 11. Radial plot of the distribution of the CIRCUST estimated peak phases of the TOP genes along the 34 tissues analyzed from GTEx. Active/light period (0, π] is identified with [6am,6pm) and the inactive/dark period (π, 2π] is done with [6pm,6am) with π corresponding with 6pm.

https://doi.org/10.1371/journal.pcbi.1011510.g011

CIRCUST has a number of advantages compared to existing order reconstruction models [12, 14, 15, 18]. First, CIRCUST allows for a more transparent assessment of the influence of outlier samples and eigengene selection than machine learning based-models such as CYCLOPS. Moreover, the simulation and real application confirm that CIRCUST eigengenes are not dominated by non-rhythmic confounds. Second, CIRCUST seed gene selection may change to be adapted to the particularities of the study (e.g. cardiovascular disease, cancer, or diabetes, among others). Such flexibility allows for a more effective exploration of diverse research contexts and enhances its applicability. Third, CIRCUST can be easily adapted to obtain sub-atlases across covariates like age or sex improving its performance. Although we plan to address this issue as a future extension of this work, in a preliminary analysis, we have observed that the seed genes' peaks estimates maintain across age groups in many tissues (e.g. Whole-Blood), while they change in other tissues like the testis or prostate which may be affected by hormonal changes. An in-depth assessment regarding sex and age influence is required to fairly compare the results with those given in [17]. Fourth, CIRCUST requires only limited assumptions of the temporal order of the seed or other genes, in that for solving the directionality problem (clockwise or counterclockwise choice), only one comparison between two seed genes is required in the base model (i.e., *ARNTL* and *DBP*). Note that in this work it is assumed that both genes are core components of the circadian molecular oscillator and therefore rhythmic. Moreover, we state that the median of the R^2_{FMM} across the seed genes in the preliminary order must be higher than 0.3 to guarantee consistent subsequent analyses. In line with mentioned above, not requiring further assumptions of the order of rhythmic genes is beneficial because they may not be conserved between model species and humans. In particular, and contrary to what is done by similar works [14], CIRCUST does not use evolutionary conservation as inclusion criteria from tissue analysis which increases the applicability of the method. As a result a larger number of human tissues as well as application to more diverse species is possible. For example, in [14] only 13 tissues out of the 51 from the GTEx collection were analyzed, in part because in these tissues the molecular clock network is preserved and the ordering conserves evolutionary conservation with mice across a set of clock genes. This contrasts with the 34 tissues considered in this paper. Beyond the base model, CIRCUST adaptability enables the incorporation of known peak phase relationships between other genes to further enhance the precision of the model.

CIRCUST application can be extended to any tissue of mammal species, regardless of whether or not the molecular clock network is known. Moreover, CIRCUST can be easily adapted to obtain sub-atlases across age, sex, or other variables. To enable wider use, the methodology is open source and publicly available to the scientific community on GitHub https://github.com/yolandalago/CIRCUST/.

It is worth noting that comparison with prior works is challenging as molecular rhythms and clocks network analyses differ across the methods employed, the covariates considered, and the tissues or species studied. Still, we found consistency regarding the tissues with the higher (Artery tibial, Skeletal-Muscle or Lung) and lower (Liver) number of rhythmic genes, as well as the tissues with the higher number of intersecting rhythmic genes (Artery Tibial and Nerve Tibial) with those reported in [1, 14, 27]. Moreover, even though the species considered were different, CIRCUST molecular rhythms analyses in the human GTEx dataset share certain similarities with the results in baboons (a closely related species), where sample collection times were known. In line with that stated in [27], the seed genes detected by CIRCUST do not systematically rank among the TOP rhythmic genes. This latter finding suggests that seed gene expression patterns are tissue-specific and their rhythmicity depends on the variability of the tissue and on the rhythmicity strength of the rest of the genes in the TOP. The number of rhythmic TOP genes is relatively lower than in previous works, mainly because of the more stringent requirements of the CIRCUST model. These findings are in line with [49] which reports that the specific genes expressing rhythmicity, including seed genes, are very tissue-specific. In addition to that, and similar to that described in [27], we discovered that well-established circadian-associated transcripts, such as the recently described *CIART*, are among the TOP rhythmic genes in more than one-third of the analyzed tissues.

There are substantial differences in the observed daily rhythms in gene expression in the current work compared among species. For example, as observed here, the ranking of highly rhythmic tissues previously documented in humans [14] shows that the number of rhythmic genes in the liver is very low as compared to other tissues such as visceral adipose tissue or tissues in the heart, while in mice, the liver has the highest number of rhythmic genes [1]. In addition to species differences, other differences that may explain different results in the literature may relate to genetic heterogeneity, environmental or behavioral factors. As described, our work is more similar to that from [14], in showing that the liver has fewer rhythmic genes in humans as compared to in mice [1].

CIRCUST methodology provides a new insight with regard to the inherent presence of variability in seed genes' peak phase timing across tissues. This evinces that there is no reason for considering the average of peak expressions for the analysis of clock molecular networks as done by [14]. Despite the heterogeneity between tissues shown by CIRCUST, tissue-specific molecular clock networks' analyses show peak phases' of TOP gene expressions clustered around dawn and dusk, with a quiescent period in-between as usually happens for diurnal primates as explained in [27].

CIRCUST presents several limitations to be considered. The first is regarding the assumption that the relative phase angle between the two selected clock genes is maintained across tissues. We address this limitation by testing a modified model built on an alternative second clock gene of choice (*DBP* primary and *CRY2* secondary). In addition, CIRCUST can derive the sequence and directionality of the temporal order of samples derived from a single-sample database to a high degree, but it does not make any assumptions or predictions regarding the phase angle between the circadian clock gene rhythms and local clock time. Finally, a limitation in this case of the GTEx dataset is that its population is heterogeneous in many ways including disease state, medication use, and environmental exposures. In addition, most tissues are heterogeneous with respect to cell type composition, and different cell types may have different properties regarding rhythmicity [49]. Even in a homogeneous tissue, individual cells may not be synchronized.

CIRCUST represents a step forward towards the building of a daily rhythm gene expression atlas in humans. Among the future directions, further systematic head-to-head comparisons of CIRCUST with other analytical methods are needed to determine the relative performance of each method under different conditions and in different populations and species. In particular, the study of covariates in the GTEx dataset to develop atlas comparisons regarding different demographic or clinical populations and conditions is important. We also plan future research to investigate the biological assumption regarding the invariance of circadian phases across tissues for a given subject. CIRCUST methodology can be easily adapted to be conducted under this assumption as is detailed in S1 Text (see Section 5.2), but an exhaustive validation is required. Also, there is a need for future analysis in larger datasets because seasonality may interact with geographic location, the season could be an important covariate. This may be especially relevant for the sun-exposed skin from lower leg tissue. Finally, pathway analyses to follow up on the hits derived from our CIRCUST analyses are needed to advance understanding of tissue-specific and across-tissue rhythmic biological processes.

Supporting information

S1 Text. Supporting information. Supplementary material for this paper including figures, tables, additional methodological details, simulations, and CIRCUST comparisons with other methods.
(PDF)

S1 Data. Daily gene expression atlas in humans. Atlas of robust human molecular rhythms for 34 human tissues.
(XLSX)

Author Contributions

Conceptualization: Yolanda Larriba, Richa Saxena, Frank A. J. L. Scheer, Cristina Rueda.

Data curation: Yolanda Larriba.

Formal analysis: Yolanda Larriba, Cristina Rueda.

Funding acquisition: Richa Saxena, Frank A. J. L. Scheer, Cristina Rueda.

Investigation: Yolanda Larriba, Richa Saxena, Frank A. J. L. Scheer, Cristina Rueda.

Methodology: Yolanda Larriba, Cristina Rueda.

Resources: Yolanda Larriba, Richa Saxena, Frank A. J. L. Scheer, Cristina Rueda.

Software: Yolanda Larriba.

Supervision: Yolanda Larriba, Richa Saxena, Frank A. J. L. Scheer, Cristina Rueda.

Validation: Yolanda Larriba.

Visualization: Yolanda Larriba.

Writing – original draft: Yolanda Larriba.

Writing – review & editing: Yolanda Larriba, Ivy C. Mason, Richa Saxena, Frank A. J. L. Scheer, Cristina Rueda.

References

1. Zhang R, Lahens N, Ballance H, Hughes M, Hogenesch J. A circadian gene expression atlas in mammals: Implications for biology and medicine. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(45):16219–16224. https://doi.org/10.1073/pnas.1408886111 PMID: 25349387
2. Antoch MP, Kondratov RV. Pharmacological modulators of the circadian clock as potential therapeutic drugs: focus on genotoxic/anticancer therapy. Handbook of experimental pharmacology. 2013; 217. https://doi.org/10.1007/978-3-642-25950-0_12 PMID: 23604484
3. Sulli G, Manoogian E, Taub PR, Panda S. Training the Circadian Clock, Clocking the Drugs, and Drugging the Clock to Prevent, Manage, and Treat Chronic Diseases. Trends in pharmacological sciences. 2018; 39(9). https://doi.org/10.1016/j.tips.2018.07.003 PMID: 30060890
4. Hesse J, Martinelli J, Aboumanify O, Ballesta A, Relógio A. A mathematical model of the circadian clock and drug pharmacology to optimize irinotecan administration timing in colorectal cancer. Computational and Structural Biotechnology Journal. 2021; 19:5170–5183. https://doi.org/10.1016/j.csbj.2021.08.051 PMID: 34630937
5. Brown SA, Fleury-Olela F, Nagoshi E, Hauser C, Juge C, Meier CA, et al. The Period Length of Fibroblast Circadian Gene Expression Varies Widely among Human Individuals. PLOS Biology. 2005; 3(10). https://doi.org/10.1371/journal.pbio.0030338 PMID: 16167846
6. Mavroudis PD, Jusko WJ. Mathematical modeling of mammalian circadian clocks affecting drug and disease responses. Journal of Pharmacokinet and Pharmacodyn. 2021; 48:375–386. https://doi.org/10.1007/s10928-021-09746-z PMID: 33725238
7. Chen CY, Logan RW, Ma T, Lewis DA, Tseng GC, Sibille E, et al. Effects of aging on circadian patterns of gene expression in the human prefrontal cortex. Proceedings of the National Academy of Sciences. 2016; 113(1):206–211. https://doi.org/10.1073/pnas.1508249112 PMID: 26699485
8. Seney ML, Cahill K, Enwright JF, Logan RW, Huo Z, Zong W, et al. Diurnal rhythms in gene expression in the prefrontal cortex in schizophrenia. Nature Communications. 2019; 10 (3355). https://doi.org/10.1038/s41467-019-11335-1 PMID: 31399567
9. Zhu Y, Wang L, Yin Y, Yang E. Systematic analysis of gene expression patterns associated with postmortem interval in human tissues. Scientific Reports. 2017; 7:5435. https://doi.org/10.1038/s41598-017-05882-0 PMID: 28710439
10. Leng N, Chu LF, Barry C, Li Y, Choi J, Li X, et al. Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. Nature methods. 2015; 12(10):947–950. https://doi.org/10.1038/nmeth.3549 PMID: 26301841
11. Liu Z, Lou H, Xie K, et al. Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. Nature Communications. 2017; 8(22). https://doi.org/10.1038/s41467-017-00039-z PMID: 28630425
12. Anafi R, Francey L, Hogenesch J, Kim J. CYCLOPS reveals human transcriptional rhythms in health and disease. Proceedings of the National Academy of Sciences of the United States of America. 2017; 114(20):5312–5317. https://doi.org/10.1073/pnas.1619320114 PMID: 28439010
13. Laloum D, Robinson-Rechavi M. Methods detecting rhythmic gene expression are biologically relevant only for strong signal. PLOS Computational Biology. 2020; 16(3):e1007666. https://doi.org/10.1371/journal.pcbi.1007666 PMID: 32182235
14. Ruben MD, Wu G, Smith DF, Schmidt RE, Francey LJ, Lee YY, et al. A database of tissue-specific rhythmically expressed human genes has potential applications in circadian medicine. Science Translational Medicine. 2018; 10(458). https://doi.org/10.1126/scitranslmed.aat8806 PMID: 30209245
15. Wu G, Ruben MD, Schmidt RE, Francey LJ, Smith DF, Anafi RC, et al. Population-level rhythms in human skin with implications for circadian medicine. Proceedings of the National Academy of Sciences. 2018; 115(48):12313–12318. https://doi.org/10.1073/pnas.1809442115 PMID: 30377266
16. Larriba Y, Rueda C, Fernández M, Peddada S. Order restricted inference in chronobiology. Statistics in Medicine. 2020; 39(3):265–278. https://doi.org/10.1002/sim.8397 PMID: 31769057

17. Talamanca L, Gobet C, Naef F. Sex-dimorphic and age-dependent organization of 24-hour gene expression rhythms in humans. Science. 2023; 379(6631):478–483. https://doi.org/10.1126/science.add0846 PMID: 36730411

18. Talamanca L. Statistical physics of periodic biological processes. EPFL; 2023.

19. Cornelissen G. Cosinor-based rhythmometry. Theoretical biology & medical modelling. 2014; 11:16. https://doi.org/10.1186/1742-4682-11-16 PMID: 24725531

20. Rueda C, Larriba Y, Peddada S. Frequency Modulated Möbius Model Accurately Predicts Rhythmic Signals in Biological and Physical Sciences. Scientific Reports. 2019; 9:18701. https://doi.org/10.1038/s41598-019-54569-1 PMID: 31822685

21. Olmo Md, Grabe S, Herzel H. Mathematical Modeling in Circadian Rhythmicity. In: Circadian Regulation: Methods and Protocols. New York, NY: Springer US; 2022. p. 55–80.

22. James FO, Boivin DB, Charbonneau S, Bélanger V, Cermakian N. Expression of clock genes in human peripheral blood mononuclear cells throughout the sleep/wake and circadian cycles. Chronobiology international. 2007; 24(6):1009–1034. https://doi.org/10.1080/07420520701800736 PMID: 18075796

23. Christou S, Wehrens SMT, Isherwood C, Möller-Levet CS, Wu H, Revell VL, et al. Circadian regulation in human white adipose tissue revealed by transcriptome and metabolic network analysis. Scientific Reports. 2019; 9(1):2641. https://doi.org/10.1038/s41598-019-39668-3 PMID: 30804433

24. Watanabe M, Hida A, Kitamura S, Enomoto M, Ohsawa Y, Katayose Y, et al. Rhythmic expression of circadian clock genes in human leukocytes and beard hair follicle cells. Biochemical and biophysical research communications. 2012; 425(4):902–907. https://doi.org/10.1016/j.bbrc.2012.08.008 PMID: 22902636

25. Fair BJ, Blake LE, Sarkar A, Pavlovic BJ, Cuevas C, Gilad Y. Gene expression variability in human and chimpanzee populations share common determinants. eLife. 2020; 9:e59929. https://doi.org/10.7554/eLife.59929 PMID: 33084571

26. Chen CY, Logan RW, Ma T, Lewis DA, Tseng GC, Sibille E, et al. Effects of aging on circadian patterns of gene expression in the human prefrontal cortex. Proceedings of the National Academy of Sciences. 2016; 113(1):206–211. https://doi.org/10.1073/pnas.1508249112 PMID: 26699485

27. Mure LS, Le HD, Benegiamo G, Chang MW, Rios L, Jillani N, et al. Diurnal transcriptome atlas of a primate across major neural and peripheral tissues. Science. 2018; 359 (6381). https://doi.org/10.1126/science.aao0318 PMID: 29439024

28. Consortium G. The Genotype-Tissue Expression (GTEx) project. Nature genetics. 2013; 45(6):580–585. https://doi.org/10.1038/ng.2653

29. Burgess HJ, Fogg LF. Individual differences in the amount and timing of salivary melatonin secretion. PLOS one. 2008; 3(8):e3055. https://doi.org/10.1371/journal.pone.0003055 PMID: 18725972

30. Sack RL, Auckley D, Auger RR, Carskadon MA, Wright KP Jr, Vitiello MV, et al. Circadian rhythm sleep disorders: part I, basic principles, shift work and jet lag disorders. Sleep. 2007; 30(11):1460–1483. https://doi.org/10.1093/sleep/30.11.1460 PMID: 18041480

31. Sack RL, Auckley D, Auger RR, Carskadon MA, Wright KP, Vitiello MV Jr, et al. Circadian rhythm sleep disorders: part II, advanced sleep phase disorder, delayed sleep phase disorder, free-running disorder, and irregular sleep-wake rhythm. An American Academy of Sleep Medicine review. Sleep. 2007; 30 (11):1484–1501. https://doi.org/10.1093/sleep/30.11.1484 PMID: 18041481

32. Kirby MJ, Miranda R. Circular Nodes in Neural Networks. Neural Computation. 1996; 8(2):390–402. https://doi.org/10.1162/neco.1996.8.2.390 PMID: 8581887

33. Scholz M. Analysing Periodic Phenomena by Circular PCA. In: Proceedings of the Conference on Bioinformatics Research and Development. vol. 4414; 2007. p. 38–47.

34. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences. 2000; 97(18):10101–10106. https://doi.org/10.1073/pnas.97.18.10101 PMID: 10963673

35. Jiang R, Sun T, Song D, Li JJ. Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome biology. 2022; 23(1):1–24. https://doi.org/10.1186/s13059-022-02601-5

36. Lech K, Ackermann K, Revell VL, Lao O, Skene DJ, Kayser M. Dissecting daily and circadian expression rhythms of clock-controlled genes in human blood. Journal of Biological Rhythms. 2016; 31(1):68–81. https://doi.org/10.1177/0748730415611761 PMID: 26527095

37. Bozek K, Relógio A, Kielbasa SM, Heine M, Dame C, Kramer A, et al. Regulation of clock-controlled genes in mammals. PLOS one. 2009; 4(3):e4882. https://doi.org/10.1371/journal.pone.0004882 PMID: 19287494

38. Wu G, Ruben M, Francey L, Smith D, Sherrill J, Oblong J, et al. A population-based gene expression signature of molecular clock phase from a single epidermal sample. Genome medicine. 2020; 12(1):73. https://doi.org/10.1186/s13073-020-00768-9 PMID: 32825850

39. Rao JS, Sengupta A. Topics in circular statistics. vol. 5; 2001.

40. Perrin L, Loizides-Mangold U, Chanon S, Gobet C, Hulo N, Isenegger L, et al. Transcriptomic analyses reveal rhythmic and CLOCK-driven pathways in human skeletal muscle. Elife. 2018; 7:e34114. https://doi.org/10.7554/eLife.34114 PMID: 29658882

41. El-Athman R, Knezevic D, Fuhr L, Relógio A. A Computational Analysis of Alternative Splicing across Mammalian Tissues Reveals Circadian and Ultradian Rhythms in Splicing Events. International Journal of Molecular Sciences. 2019; 20:3977. https://doi.org/10.3390/ijms20163977 PMID: 31443305

42. Wucher V, Sodaei R, Amador R, Irimia M, Guigó R. Day-night and seasonal variation of human gene expression across tissues. PLOS Biology. 2023; 21(2):1–30. https://doi.org/10.1371/journal.pbio.3001986 PMID: 36745672

43. Fisher NI. Statistical Analysis of Circular Data. Cambridge University Press; 1993.

44. Adzhar R. Outlier detection in circular data and circular-circular regression model. Universiti Malaya; 2011.

45. Agostinelli C, Lund U. R package circular: Circular Statistics (version 0.4-95); 2022. Available from: https://r-forge.r-project.org/projects/circular/.

46. Hughey J, Butte A. Differential Phasing between Circadian Clocks in the Brain and Peripheral Organs in Humans. Journal of biological rhythms. 2016; 31(6):588–597. https://doi.org/10.1177/0748730416668049 PMID: 27702781

47. Lopes-Ramos CM, Paulson JN, Chen CY, Kuijjer ML, Fagny M, Platig J, et al. Regulatory network changes between cell lines and their tissues of origin. BMC Genomics. 2017; 18. https://doi.org/10.1186/s12864-017-4111-x PMID: 28899340

48. Brinkmeyer-Langford CL, Guan J, Ji G, Cai JJ. Aging Shapes the Population-Mean and -Dispersion of Gene Expression in Human Brains. Frontiers in Aging Neuroscience. 2016; 8:183. https://doi.org/10.3389/fnagi.2016.00183 PMID: 27536236

49. Donovan M, D'Antonio-Chronowska A, D'Antonio M, Frazer K. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. Nature Communications. 2020; 11:955. https://doi.org/10.1038/s41467-020-14561-0 PMID: 32075962