CIRCUST: A NOVEL METHODOLOGY FOR RECONSTRUCTION OF TEMPORAL ORDER OF MOLECULAR RHYTHMS; VALIDATION AND APPLICATION TOWARDS A HUMAN CIRCADIAN GENE EXPRESSION ATLAS

Yolanda Larriba
Department of Statistics and Operational Research, University of Valladolid, Valladolid, Spain.
Mathematics Research Institute of the University of Valladolid, University of Valladolid, Valladolid, Spain.
yolanda.larriba@uva.es

Ivy Mason
Medical Chronobiology Program, Division of Sleep and Circadian Disorders, Departments of Medicine and Neurology, Brigham and Women's Hospital, Boston, MA, US.
Division of Sleep Medicine, Harvard Medical School, Boston, MA, US.
imason@bwh.harvard.edu

Richa Saxena
Center for Genomic Medicine and Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, US.
Division of Anesthesia, Harvard Medical School, Boston, MA, US.
Division of Sleep Medicine, Harvard Medical School, Boston, MA, US.
Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, US.
rsaxena@partners.org

Frank A.J.L. Scheer
Medical Chronobiology Program, Division of Sleep and Circadian Disorders, Departments of Medicine and Neurology, Brigham and Women's Hospital, Boston, MA, US.
Division of Sleep Medicine, Harvard Medical School, Boston, MA, US .
Broad Institute, Cambridge, MA, US
fscheer@bwh.harvard.edu

Cristina Rueda
Department of Statistics and Operational Research, University of Valladolid, Valladolid, Spain.
Mathematics Research Institute of the University of Valladolid, University of Valladolid, Valladolid, Spain.
cristina.rueda@uva.es

December 21, 2022

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 display 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 timepoints has several challenges, including the reconstruction of temporal order or highly noisy data. Previous methodologies have attempted to address these challenges in one or a small number of tissues for which clock gene evolutionary conservation is preserved. Here we propose CIRCUST, a novel CIRCular-robUST methodology for analyzing molecular rhythms, that relies on circular statistics, is highly robust against noise and requires fewer assumptions than existing methodologies. First, we describe the method, then we validate it against two 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 human circadian expression atlas. The validation and application shown here indicate that CIRCUST provide 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/`.

***K*eywords** Gene Expression · Biological Rhythms · Circadian Rhythms · Time of Death · GTEx · Oscillatory Signal · Temporal Order Estimation · Circular Principal Component Analysis · Frequency Modulated Möbius (FMM) model

1 Introduction

Circadian clocks orchestrate metabolic, endocrine and behavioral functions. The molecular clock drives tissue-specific rhythms in gene expression [1]. More than $\sim$50$\%$ 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, 3, 4]. A comprehensive human temporal atlas of 24-h rhythms in gene expression across tissues is therefore of great potential value. Obviously, 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]. 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 due to 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 human circadian gene expression atlas 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. As a vast collection of unordered RNA samples from postmortem individuals are typically available for one or more tissues, the first challenge to be solved, when analyzing molecular rhythms with imprecise sample collection times, and/or unknown biological times, is to estimate the 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. Next, gene expression rhythms and peak time relationships between genes and between tissues were analyzed towards a human circadian expression atlas.

The temporal order estimation problem can be mathematically formulated as that of looking for a m-dimensional vector that provides what is known as a *circular order* $\boldsymbol{o} = (o_1, \ldots, o_m)'$, where m denotes the number of sample collection times to be ordered, see Figures 1 and 2 as illustration. In practice, a circular ordering represents until $2m$ distinct sample collection time configurations along the 24-h of the day, depending on the choice of the starting point and the orientation (clockwise or counter-clockwise). The choice of directionality is not trivial, and plays a key role in correctly locating gene biological functions along 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, being Oscope [10], reCAT [11] or CYCLOPS [12] among the most extended 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 framework overcomes these drawbacks, but it requires rhythmicity evolutionary information of homologs of genes from other species which is not always available. Even so, CYCLOPS-based solutions to this problem have proliferated, but for a single tissue, or for limited and preselected collections of tissues [13, 14]. In [15] a non-parametric framework is proposed to mathematically formulate and efficiently solve the temporal order estimation problem without any additional genome information, but for the case of equally-spaced timepoints, which may be a strict assumption in postmortem gene studies. Hence, none of these methods are entirely suited to the problem at hand.

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

Cosinor [16] is the classical rhythmicity model widely utilized in chronobiology [7, 8, 12, 13]. 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 Figure 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 [17]. In addition, the use of a large number of components may result in serious overfitting issues.

Alternative rhythmicity models have emerged from different subjects. In [18], 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 depend on the process under study. Within a non-parametric perspective, [15] 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. Nevertheless, pattern comparison, in this case, is not straightforwardly derived, as done for parametric models. To overcome these drawbacks, [17] presented FMM, a flexible five parametric model that allows deformations to sinusoidal shape to accommodate commonly seen asymmetries in applications (see Figure 3). This is because FMM is formulated in terms of the phase, an angular variable that represents the periodic movement of the oscillation. Moreover, FMM model parameters are easy to estimate providing meaningful interpretations. An overview of the FMM model is given in Section 3.1 of the Supplementary Materials.

This work proposes CIRCUST, a general methodology that solves temporal order estimation problem, as well as identifies and characterizes a wide variety of circadian genes, including those with asymmetric expression patterns. The method makes use of the underlying CIRCular structure of the molecular rhythms [15], 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 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 many human tissues, most human studies have been limited to blood (e.g., [19]), or to one other tissue with low sampling frequency [20, 21]. Additionally, inter-individual variability increases uncertainty [22]. This paper shows that CIRCUST is a sound framework to analyze molecular rhythms from two controlled experiments. 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 [14]. The second validation set consists of a large set of different tissues collect at known timepoints across a 24-h timeframe from baboons, a primate closely related to humans [23]. Following the main aim of this paper, the Genotype-Tissue Expression (GTEx) dataset, the postmortem gene expression dataset across the largest number of human tissues, was also considered [24]. GTEx provides annotated times of death (TODs) aimed to order subject samples across tissues. However, such TODs may provide inaccurate information about patients' biological death times, see 2 and Figure S1. This latter because of the large inter-individual differences in the timing of the central circadian pacemaker, even in healthy patients [25, 26, 27]. Thus, finally, CIRCUST is applied to GTEx to develop an atlas of human 24-h expression rhythms across a wide range of tissues that may provide novel insights into the molecular clock networks.

2 Methods

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

2.1 CIRCUST solution to approach 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 [28, 29]. 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. Two different sets of rhythmic genes are considered in this paper: a set of 12 well-established core clock 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-based solution starts with the computation of the two first *eigengenes* from a sub-matrix of rhythmic gene expressions. Eigengenes are gene-like expression patterns across samples obtained as a linear combination of the expressions in the matrix [30]. Despite the initial unordered expression patterns of these two eigengenes, its mapping reveals the underlying circular structure over the samples, as illustrated in Section 3.2 of the Supplementary Materials. 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 core clock genes' peak phase relations in mammals are fulfilled see Section 3.2 in the Supplementary Materials for details). These assumptions can be user-refined, in terms of peak phases' order restrictions, in case that the molecular clock network of the specie 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 Section 3.2 of the Supplementary Material. Figures 1 and 2 illustrate a CPCA solution to approach the temporal order identification problem.

2.2 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. Figure S2 shows an outline of the methodology.

$$[\boldsymbol{R}] \xrightarrow{\substack{Prepro\\cessing}} [\boldsymbol{N}] \xrightarrow{\substack{Prelimi\\nary\,Order}} [\boldsymbol{X}] \xrightarrow{\substack{TOP\\rhythmic\\orderings}} [\boldsymbol{X}_k^{TOP}] \xrightarrow{\substack{Robust\\Estima\\tion}} [\boldsymbol{M}^{TOP}],$$

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 Section 3.1 in the Supplementary Materials. CIRCUST steps are described below.

1. ***Preprocessing.***
   Genes with zero read counts in more than 30% of samples are discarded. Gene expressions are one by one normalized into [-1,1] by using a min-max normalization [15]. The preprocessed expression matrix is denoted by $[\boldsymbol{N}]$.
2. ***Preliminary order.***
   A core information set consisting of the 12 clock genes: *Per1, Per2, Per3, Cry1, Cry2, Arntl, Clock, Nr1d1, Rora, Dbp, Tef* and *Stat3* is considered. There is no a gold-standard for core clock genes selection, though gene expression patterns of this choice, generally display marked circadian signals in most of the human tissues [1] and were also considered as circadian benchmarks in previous works [12, 23, 13, 14].

   The role of CPCA at this point is twofold. CPCA is computed on the sub-matrix of the 12 core clock genes from $[\boldsymbol{N}]$. First, CPCA allows detecting outlier samples, see Section 3.3 in the Supplementary Materials for details. Outliers are deleted from $[\boldsymbol{N}]$, 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 core 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 $[\boldsymbol{X}]$.

3. ***TOP rhythmic orderings.***
   Rhythmicity models are used at this stage to predict gene expression patterns. First, the ORI model's [15] 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, see Section 3.4 of the Supplementary Material. 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 ($\hat{t}_U$) coverage over 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 Section 3.1 in the Supplementary Materials and [17] for details. The 12 core clock 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 are filtered from $[\boldsymbol{X}]$.
   Next, random selections of size $2/3$ of the genes in the TOP are considered. CPCA-based 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 2.
   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 asses and compare rhythmicity across tissues.

3 Results

In this section, CIRCUST is validated using ordered data from humans and baboons. We also illustrate the application of CIRCUST to GTEx towards developing a human circadian gene expression atlas.

3.1 CIRCUST validation on human epidermis

This validation relies on the hybrid human gene expression dataset from epidermis tissue (GEO accession number GSE139301) [31]. 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 times are known. This latter mimics which was those done in [31] to validate CYCLOPS. For comparison purposes, the analyses refer to the set of clock-associated genes in [31] which are among those at the TOP genes of CIRCUST for epidermis tissue see Figure 4. Gene expression data from biopsies at the four timepoints for the 19 participants are displayed in Figure 4 (A), 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. Figure 4 (B) compares these peak values (triangles) derived from the 19 participants, where 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 [32] between participant peak phases and estimated phases from CIRCUST and CYCLOPS for these genes are 0.862 and 0.819, respectively, revealing a higher coherence of CIRCUST estimates with the peak phases when times are known. 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 are 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 Figure 4 (C)) are more consistent with the oscillatory expression patterns observed in Figure 4 (A) than those given by CYCLOPS in [31], being *Ciart, Tef* or *Arntl* among those with display strongest rhythmicity.

3.2 CIRCUST validation on multiple baboons tissues

The second validation is driven by the baboon gene expressions dataset (GEO accession number GSE98965). Data were collected, under controlled conditions, every 2 hours (ZT0, ZT2,..., ZT22) over the 24-h day across 64 different tissues, which are aggregated into 13 functional groups [23]. 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 core clock genes is, on average, higher than 0.7, see Table S1 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 mammals specie 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 performance (see Methods Section). Specifically, in [1, 12, 33, 34] is claimed 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 relations above increases the number of core clock genes with their peaks within the active period ($[0, \pi)$) with regard to the standard order peak times assumption (2) (see Subsection 2.2 in the Supplementary Materials), it will be replaced by the specific relation given for baboons.

Circular association between CIRCUST estimated times in $[0, 2\pi)$ and the real circadian 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 [35], 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 [36, 37]. The closer ρ to 1 indicates a better correspondence between the orders. $CIRCUST_{prior}$ correctly orders the samples across the 47 tissues, see Figure S3. 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 S1 for details. The true circadian order is almost exactly discovered 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 Figure 5 (A). In addition, Figures 5 (B), S4, S5, and S6 reveal, that CIRCUST conserves rhythmicity across selected clock 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 Figure 5 (B)) are mimicked by expressions obtained as function of CIRCUST estimated times for these same genes (bottom panels of Figure 5 (B)). Moreover, these plots exhibit that the FMM model accommodates a wide variety of circadian patterns with high (closely to 1) and similar rhythmicity strength values, R^2_{FMM}, across the selected clock genes, even in those with non-sinusoidal gene pattern, see *Npas2* in Figure 5 (B) and more in Figures S4, S5, and S6.

3.3 CIRCUST application to GTEx

This section comprises the molecular rhythms' and clock networks' analyses from GTEx (V7) database. Only tissues with more than 40 samples were selected. In addition, two cell lines and thirteen brain tissues were discarded [24, 38]. Cell lines may not capture the molecular complexity of the tissue [39]; while the brain tissues usually evince intra-tissue heterogeneity and they are often considered as independent molecular networks [40, 41]. 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. According to that, our analysis considers 621 donors characterized by a mix of ages, sex, and health status (see Table S2). Specifically, the results below involve, 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.

3.3.1 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, as is seen in Figure 3. The number of TOP genes varies among the analyzed tissues (see Figure S7). 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 Figure S8). In particular, for Artery-Tibial and Nerve-Tibial, which are the tissues with the highest number of TOP genes, there are only 5.319% (5 out of 94) of shared between both tissues, apart from the 12 core clock genes considered. Moreover, in other rhythmic organs like Testis, 81.609% (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 core clock genes. Figure 6 illustrates R^2_{FMM} distribution for the genes in the TOP of the 34 organs analyzed. The R^2_{FMM} of the 12 core clock genes are shown as colored dots. As seen, core clock 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 core clock genes are distributed among the TOP genes. On the contrary, the core clock genes in Whole-Blood are not among the TOP genes of this tissue. This latter does not mean that the core clock genes are not rhythmic, but 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 core clock gene expression oscillations persist across wide inter-individual variability, with clock genes such *Per3* being among the highest rhythmic 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*, a prognostic marker in renal activity [42, 43], see Figure 3.

Finally, the atlas of robust human molecular rhythms for the 34 human tissues is provided as a Supplementary data file. For each tissue, the atlas includes the list of TOP genes, sorted according to the higher rhythmicity, the rhythmicity measure (R^2_{FMM}), the estimated peak phases ($\hat{t}_U$), and their peak timing's periods relative to *Arntl*: corresponding to active/lightened, if $\hat{t}_U \in [0, \pi)$, or to inactive/darkness $\hat{t}_U \in [\pi, 2\pi)$. The estimated amplitudes ($\hat{A}$) are also provided. All of these values are derived from Sept 4 of CIRCUST methodology. This is the largest rhythmic gene characterization across human tissues to date. This work represents a significant advance towards a human rhythmic gene expression atlas.

3.3.2 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. It is the first time that a large number of molecular clock networks are simultaneously analyzed across a large set of human tissues, until now, no more than 13 human organs were considered [12, 13, 14].

Figure 7 shows the peak phase distributions of the 12 core clock genes across the 34 tissues. Non-rhythmic core clock genes were discarded from this analysis, see Tables S3 and S4. Distributions vary across organs, but they are not random throughout the 24-h day. The peak phase estimates are generally in one or two clusters, with one of them usually preceding the inactive/darkness period in mammals ($[\pi, 2\pi)$) [13]. For core clock genes such as *Clock* or *Per1*, peak phase distribution is mainly restricted to a $\sim$6-hour interval. However, for most of the core clock components, the peaks of the core clock gene expression are distributed along $\sim$12-hour, matching with the active period ($[0, \pi)$) or light day hours. This reveals human inter-tissue variability that is characteristic of GTEx database and the heterogeneous behavior of the molecular clock networks across the variety of organs analyzed. But, for the particular case of highly rhythmic organs such as Skin (epidermis), molecular clock networks are maintained across species, as is shown in Figure 8. There, the estimated core clock peak phases in the skin for humans, from GTEx database, are similar to those estimated for the baboons and both are close to those obtained as a function of the true circadian times.

Finally, the distribution 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 Figure 9. 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 tissue show phased clusters that are temporally close, see for example Esophagus-Gastroesophageal Junction and Colon-Sigmoid, both belonging to the digestive tract. A compilation of phases across the TOP rhythmic genes reveals that for the vast majority of tissues, an early afternoon major peak anticipating the inactive period, and a quiescent zone are also observed for many of them which are considered a distinctive feature of rhythmicity in the diurnal primate [23].

4 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 circadian 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 evinces a new insight regarding rhythmicity including the identification of novel rhythmic genes (see Figure 3). These strengths of methodology give rise to the building of a comprehensive human circadian expression atlas from the GTEx database which is the largest rhythmic gene characterization across human tissues to date.

The circular-based statistical methodology that sustains CIRCUST is its strong point, as its adequacy to the problem at hand provides a wide versatility compared to the CYCLOPS-based alternative methodologies in the literature

[12, 13, 14]. Alternatives are based on machine learning procedures and work as a black box. Indeed, none of them explicitly define the temporal order estimation problem as a circular order identification problem, and nor describe the directionality problem (clockwise or counterclockwise choice). This is because CYCLOPS-based methods are nurtured from evolutionary information of homolog genes from other species, and as a result, the circular order identification problem is reduced to that of finding the order among sampling times that is closest to that given for the other specie. However, supervised information from other species is not always available (or accurate), and notably restricts the number of tissues considered in the analysis. To be more concrete, in [13] only 13 tissues out of 54 from GTEx collection, where phase relationships between humans and mice were conserved, were analyzed, which is a significantly lower size in comparison with the size of 34 organs considered in this paper. In addition, as is claimed by [23] even for core clock gene expressions such as *Tef, Dbp*, or *Per2* and when sample collection times are known, peak expression phases between mice and primates are distinct. CIRCUST, on contrary, overcomes these drawbacks. It mathematically formulates, solves, and validates the temporal order estimation problem as a circular ordering identification problem, including the solution for directionality. To do so, CIRCUST does not need genetic information from homologs, with the base model just requiring directionality derived from standard assumptions regarding clock genes in mammals. What is more, in case the molecular clock network is (partially) known, CIRCUST adaptability enables the incorporation of the knowledge of peak phase relationships in terms of order restrictions, as shown in this work. Specifically, the genuine definition of TOP rhythmic genes, the proposal of a novel and flexible method like FMM to predict rhythmicity, and the independence of genetic information from homolog species set and add value to the human circadian gene expression atlas provided.

The potential of CIRCUST, sustained by a comprehensive and adaptative statistical methodology, goes beyond this first work. CIRCUST application usage can be extended to any tissue of mammal species, regardless molecular clock network is or is not known. Moreover, CIRCUST can be easily adapted to obtain sub-atlas across ages, sex, or other variables improving its performance. Specifically, we have observed that core clock peak phase 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, this finding will be addressed as a future extension of this work. Finally, the methodology is open source and it is 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 the tissues or the species which leads to finding controversial results in the literature. Despite that the species considered are different, CIRCUST molecular rhythms analyses in the human GTEx dataset are generally in line with those given for baboons (a closely related specie), where sample collection times are known [23]. First, the tissues with the higher and lower number of rhythmic genes (see the Results Section); as well as the tissues with the higher number of intersecting rhythmic genes (Artery and Tibial Nerve) are the same in both studies. Second, we found that the core clock genes detected by CIRCUST do not systematically rank among the TOP rhythmic genes, which is also in line with the results given in [23]. This latter suggests that core clock 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. It is also observed in the number of rhythmic TOP genes, which is relatively lower than in previous works, mainly as any homolog gene expression information is considered. These findings are in line with [41] where it is claimed that the heterogeneity across human tissues from GTEx database, even in the case of core clock genes, rhythmicity signature is very tissue-specific. In addition to that, 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 organs similar to those observed by [23].

There are substantial differences in the observed daily rhythms in gene expression in the current work compared among species and prior works too. For example, the ranking of highly rhythmic tissue previously documented in humans [13] 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]. There are at least four main differences to consider in comparing rodent and human studies that may explain differences observed between rodent and human results: (1) species differences: (2) genetic heterogeneity; (3) environmental; and (4) behavioral differences. With regard to this issue, our work is more similar to those from [13], in showing that the liver is showing fewer rhythmic genes in humans as compared to in mice [1]. However, on the other hand, tissues that are generally thought to have a large number of cycling genes do not show this high rhythmicity, e.g., the liver. But, this notion appears to be based primarily on previous studies being done in rodents, and indeed the number of rhythmic genes in the liver of baboons has been shown to be very low as well [23], suggesting a role of species differences in tissue-specific rhythmicity.

Discrepancies mentioned are specially marked for non-controlled and highly noisy experiments such as the GTEx database [41]. Indeed, the main novelty of the CIRCUST methodology with regard to the previous works is the variability and differences in core clock peak phase locations across tissues. This fact provides new insight into this

issue and evinces that there is no reason for considering the average of peak expressions for the analysis of clock molecular networks as done by [13, 14, 31]. Despite this fact, tissue-specific molecular clock networks' analyses show peak phases' of TOP gene expressions clustered around dawn and dusk, with a quiescent period as usually happens for diurnal primates as explained in [13].

CIRCUST presents several limitations to be considered. The first is regarding the assumption that the relative phase angle between clock genes is maintained across tissues and between people. 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.

CIRCUST establishes a first approach towards the building of a human circadian expression atlas. 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, we are working on the study of covariates in the GTEx dataset to develop atlas comparisons regarding different demographic or clinical variables. Also, for future analysis in larger data sets, as the seasonal factor may interact with geographic location, the season could be a 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.

5 Figures

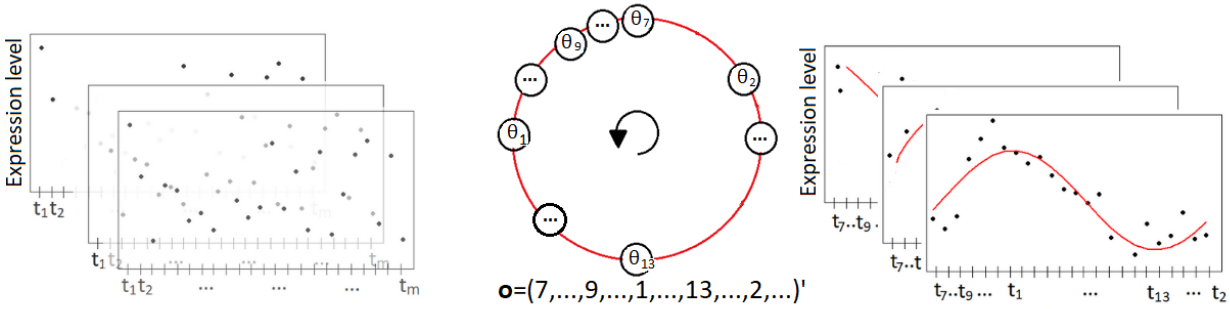


Figure 1: Illustrative outline of CIRCUST solution to temporal order estimation conducted at each tissue. Left: Unordered gene expression data across m samples registered at arbitrary clock times $t_1, \ldots, t_m$ along the 24-h day. Superposed rectangles are different genes of the tissue. Dots are the gene expression data. Middle: Circular order $\boldsymbol{o}$ obtained from $\boldsymbol{\theta}$, where $0 \preceq \theta_7 \preceq \cdots \preceq \theta_9 \preceq \cdots \preceq \theta_1 \preceq \cdots \preceq \theta_{13} \preceq \cdots \preceq \theta_2 \prec 2\pi$. Starting point and direction are fixed so the assumptions considered are fulfilled. Right: Ordered gene expression data, as 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. Superposed rectangles are the different genes of the tissue. Dots are gene expression data.

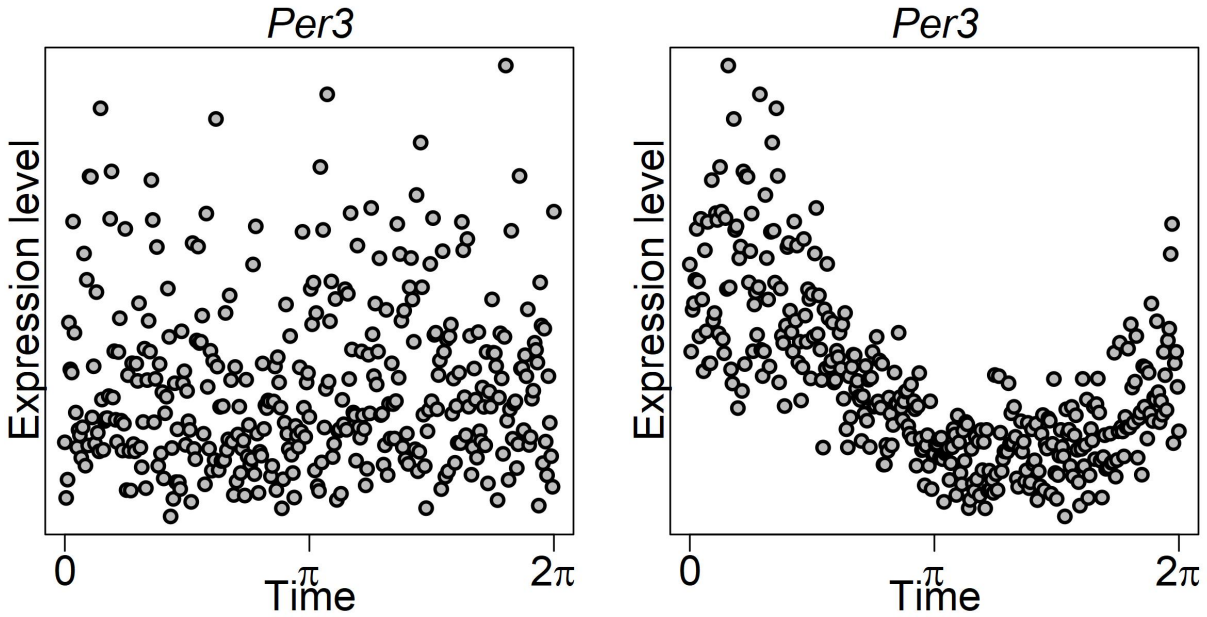


Figure 2: TOD versus CPCA sampling time estimates for gene *Per3* on Skin Not Sun Exposed (Suprapubic) tissue from GTEx. Left: Gene expression as function of TOD times. Right: Gene expression as function of CIRCUST estimated times.

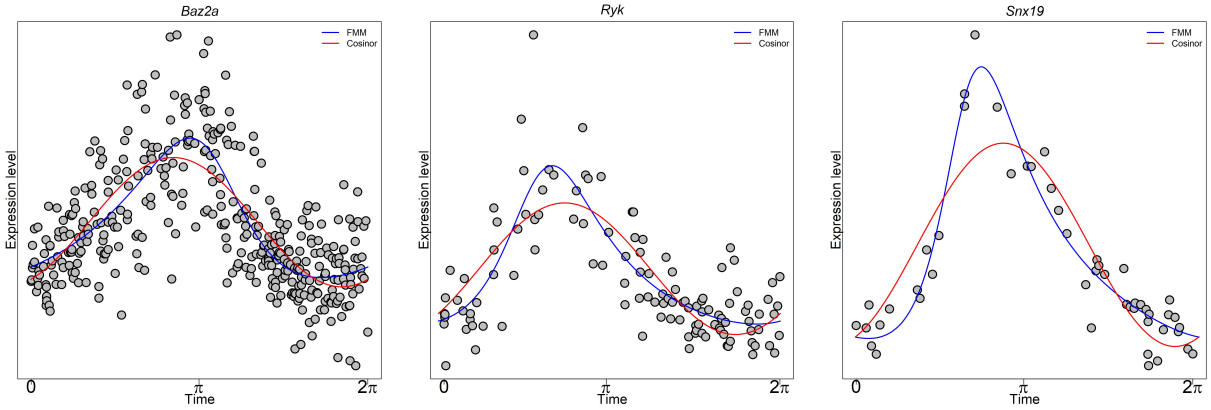
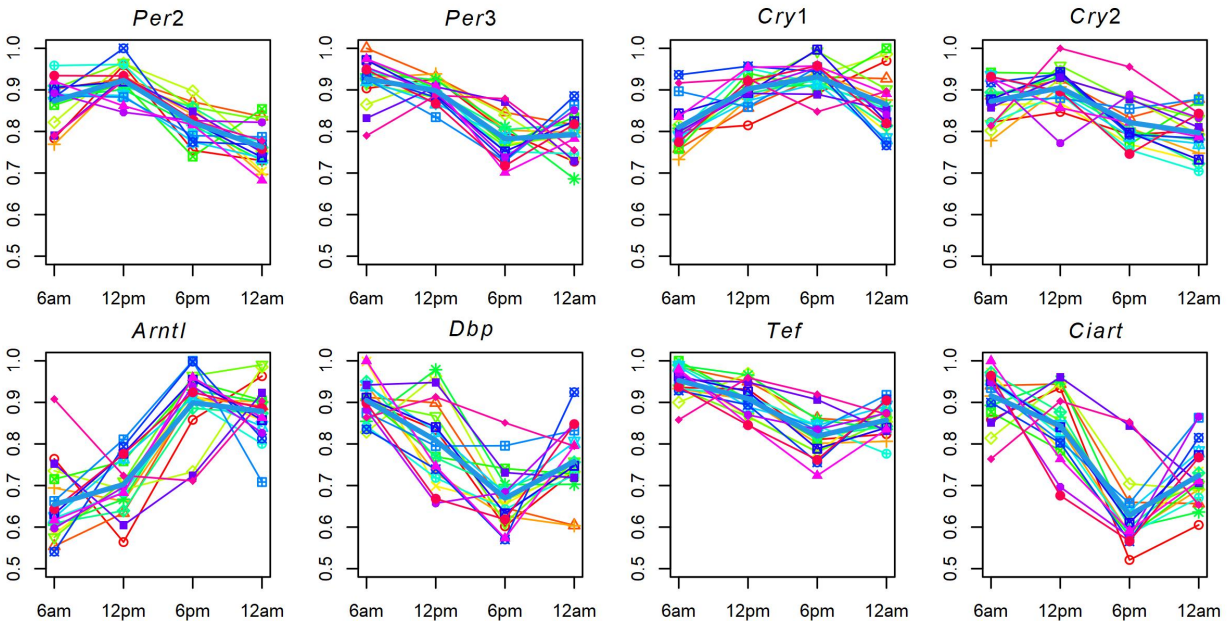
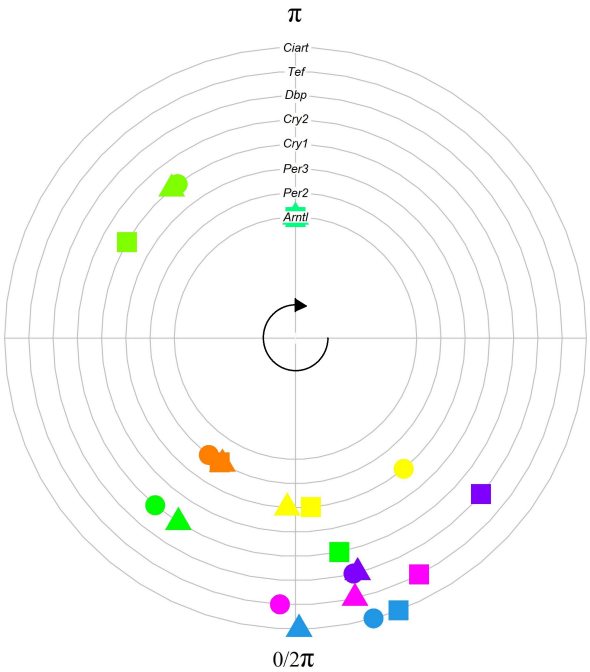


Figure 3: FMM versus Cosinor performance on selected TOP genes from different GTEx tissues. *Baz2a* (left), *Ryk* (midle) and *Snx19* (right) gene expression as function of CIRCUST estimated times from Lung, Small Intestine and Kidney, respectively. FMM predictions are blue solid lines. Cosinor predictions are red solid lines.



(A) Expression patterns across the 19 participants for the eight genes under analysis at four biopsies times. Thin color lines represent participants' gene expression. Thick blue lines represent the average expression profile across participants.

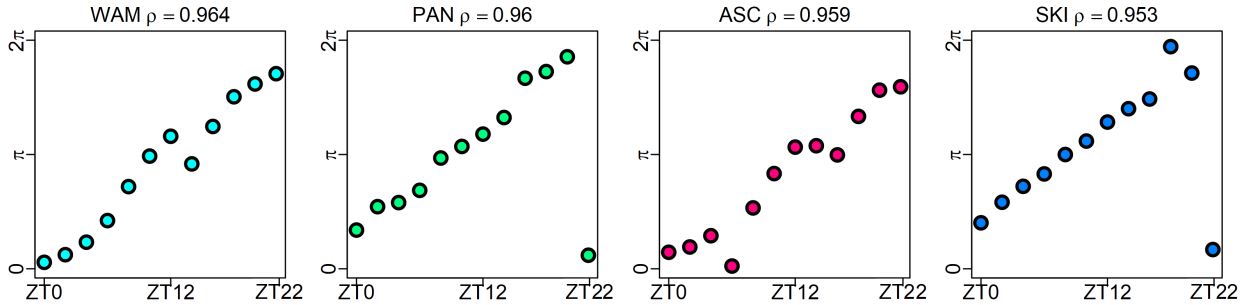


(B) CIRCUST peak phase estimates (with the 533 samples) compared to CYCLOPS ones using the peaks from 19 participants as reference for the eight genes under analysis. Participant peak phases (triangles), estimated phases from CIRCUST (circles) and CYCLOPS (squares). Biopsies times given along the 24-h interval are read into $[0, 2\pi)$. For comparison purpose, π is fixed at *Arntl*'s peak.

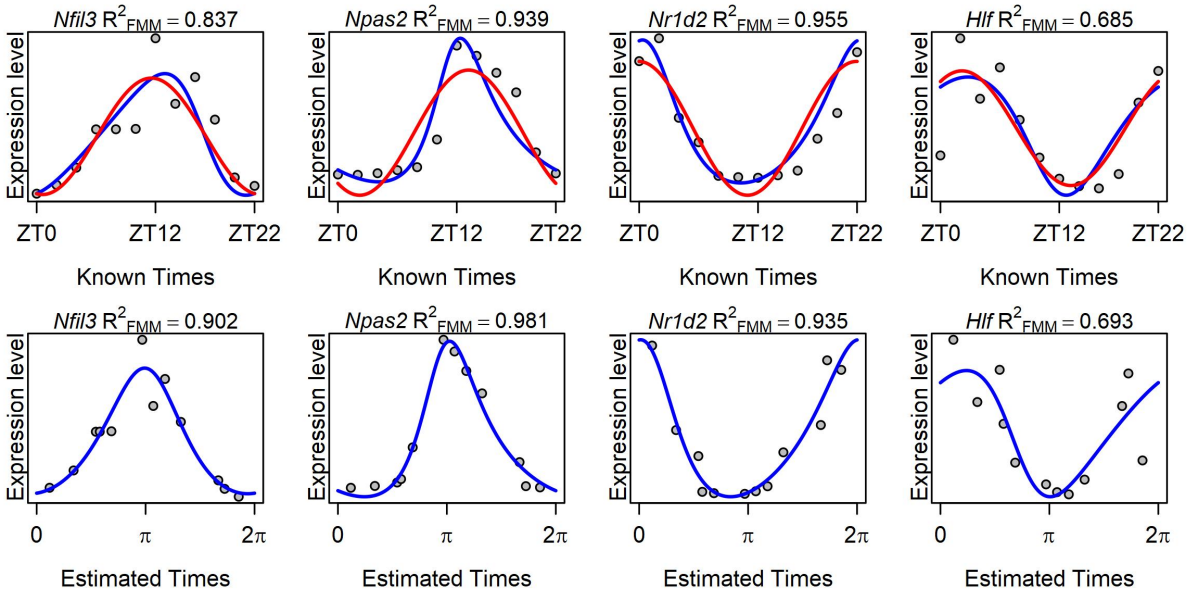
Gene	CIRCUST	CYCLOPS
Per2	0.475	0.490
Per3	0.577	0.670
Cry1	0.483	0.320
Cry2	0.680	0.270
Arntl	0.672	0.610
Dbp	0.534	0.550
Tef	0.683	0.660
Ciart	0.727	0.740

(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 [31] for CYCLOPS. The higher values the stronger the rhythmicity, but the scales are different.

Figure 4: CIRCUST consistency for human epidermis (GSE139301) dataset



(A) Estimated ($[0, 2\pi)$) vs circadian 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 [13]. 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 function of known times ZT0, ZT2,...,ZT22. Bottom panels: expressions as function of CIRCUST estimated times. FMM predictions are blue solid lines. Cosinor predictions are red solid lines.

Figure 5: CIRCUST validation based on baboon dataset (GSE98965).

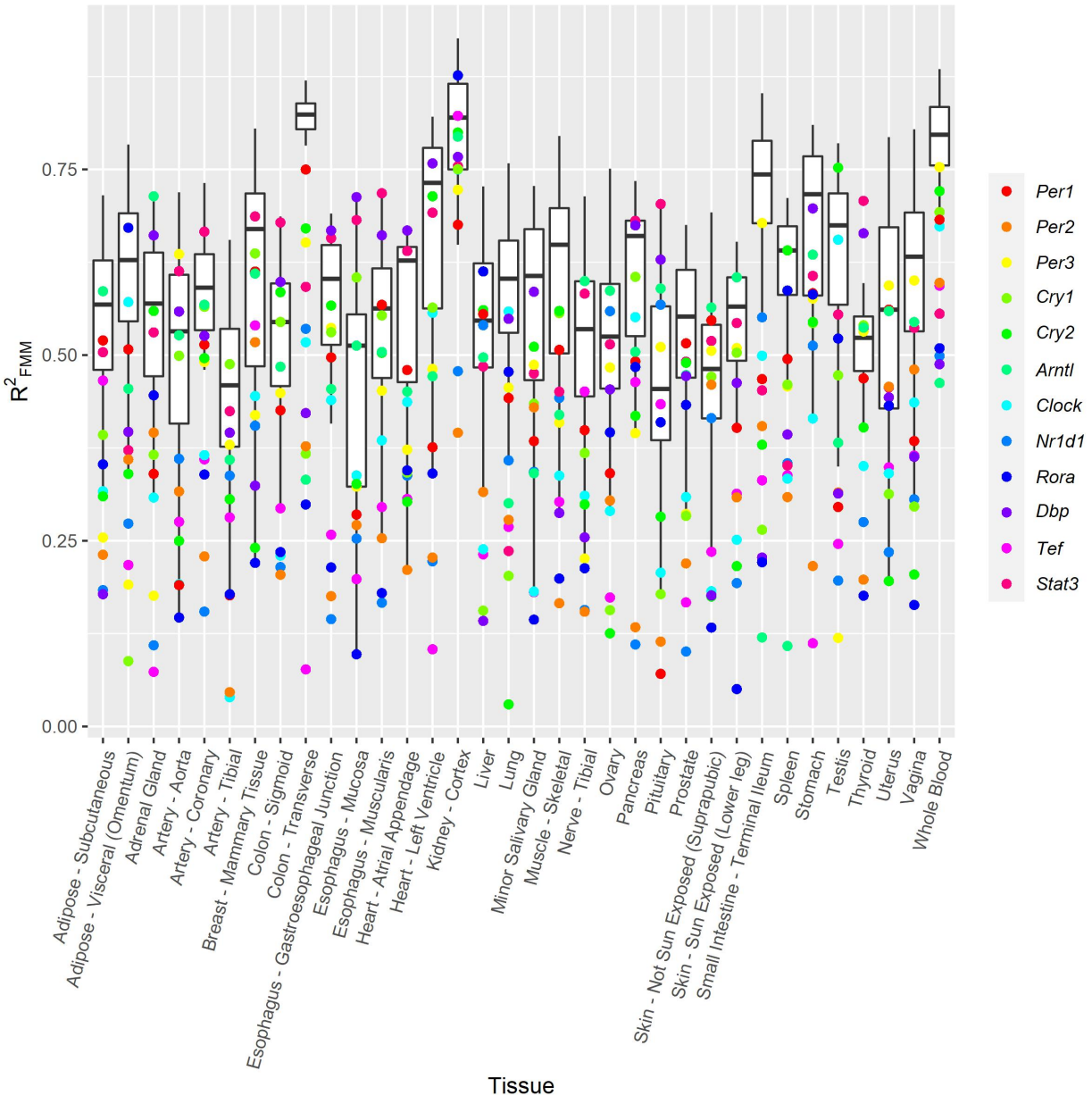


Figure 6: R^2_{FMM} distribution across the 34 organs in GTEx. Coloured dots denote the R^2_{FMM} of the 12 core clock genes considered. Tissues are alphabetically sorted.

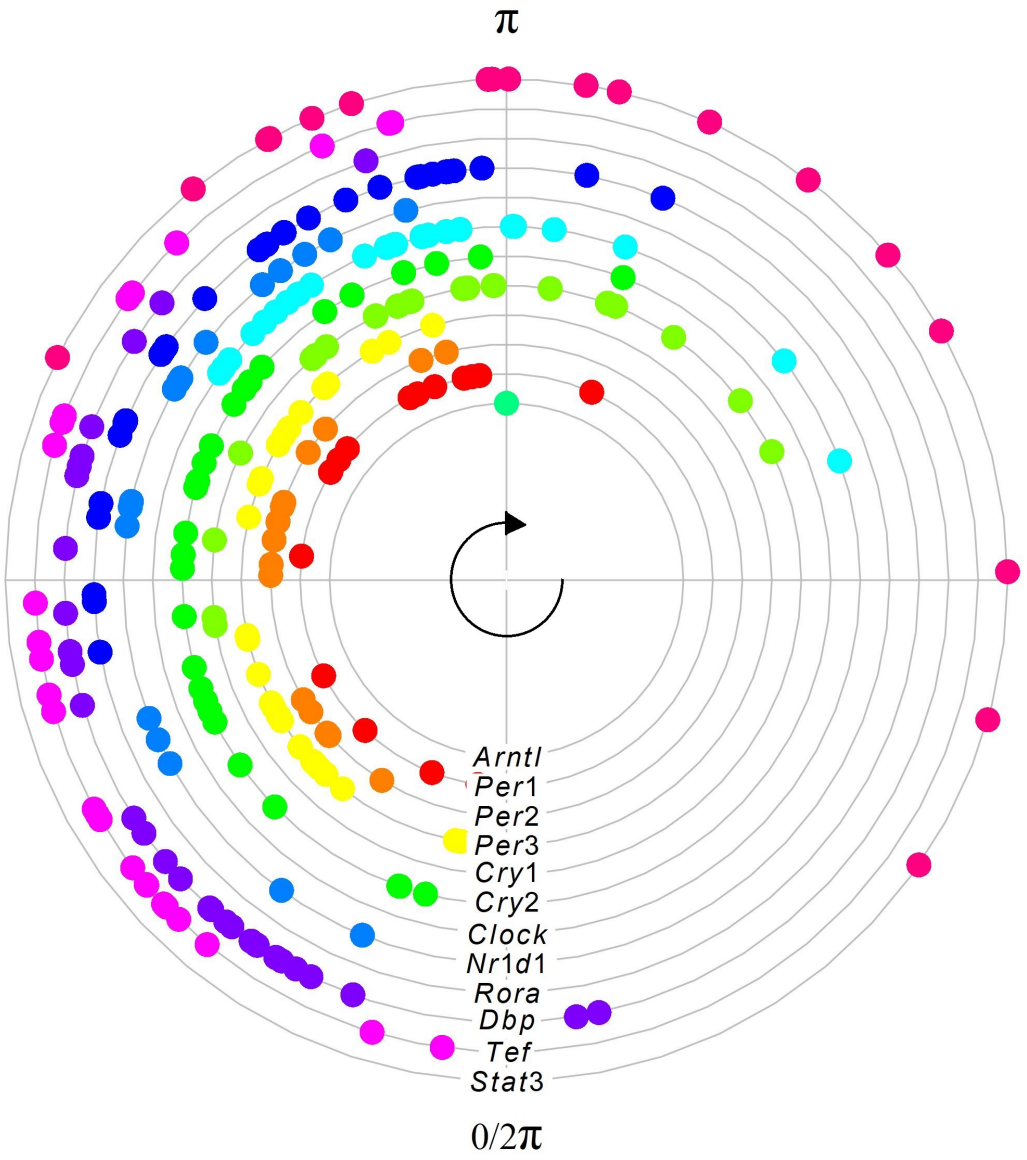


Figure 7: Peak phase estimate distributions of the 12 rhythmic core clock genes across the 34 tissues in GTEx. *Arntl* is known to peak in anticipation of the inactive/darkness $[\pi, 2\pi)$ period in mammals and was set to π for comparisons [13]. The R^2 and t_U estimated values are given in Tables S3 and S4, respectively.

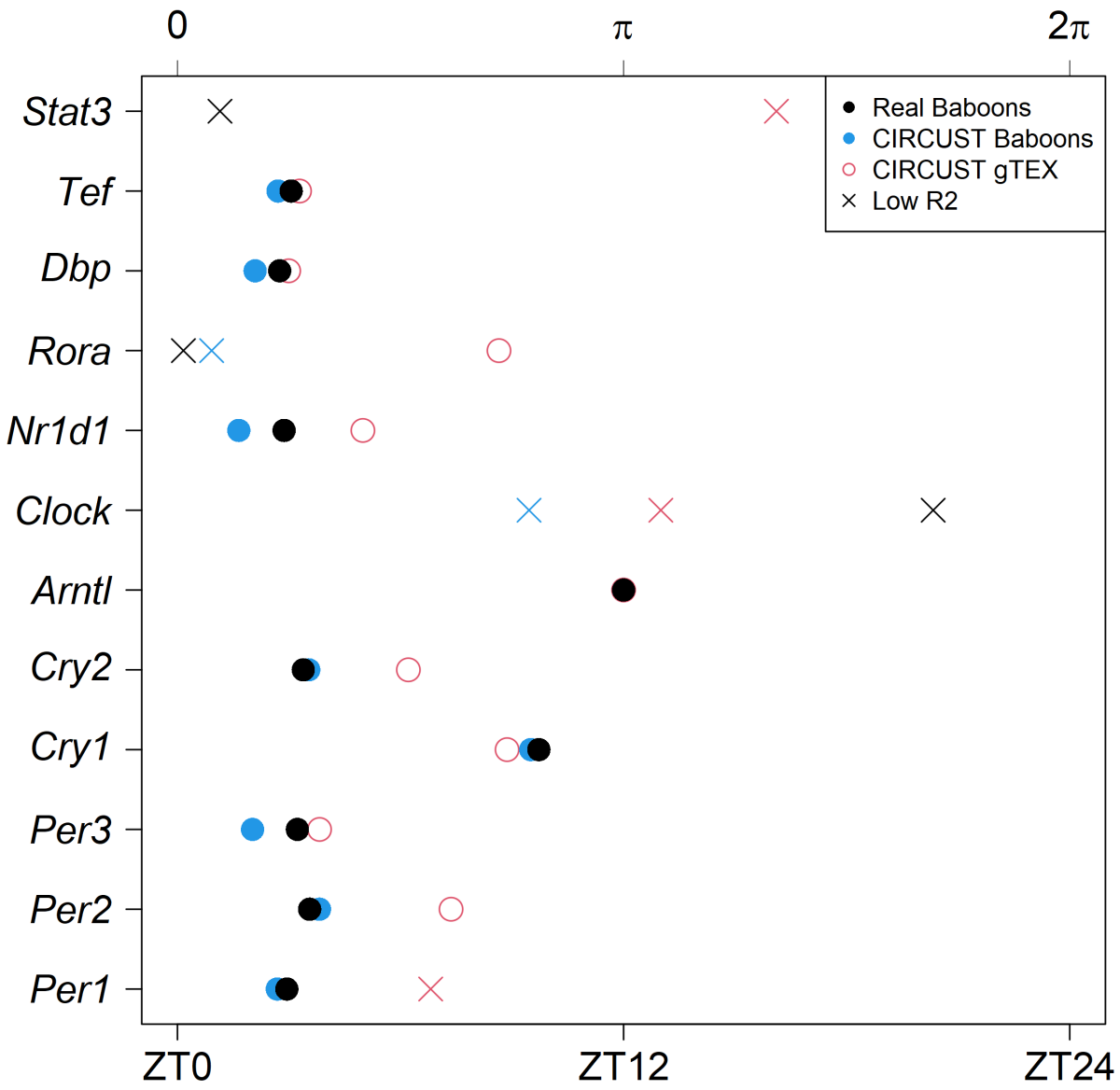


Figure 8: Core clock peaks in the human epidermis (Tissue: Skin - Sun exposed from GTEx) and baboon epidermis (Tissue: SKI from GSE98965) tissue. Estimated phases are derived from the FMM model. Black and blue dots match with the peaks obtained as a function of the true circadian times and as a function 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.

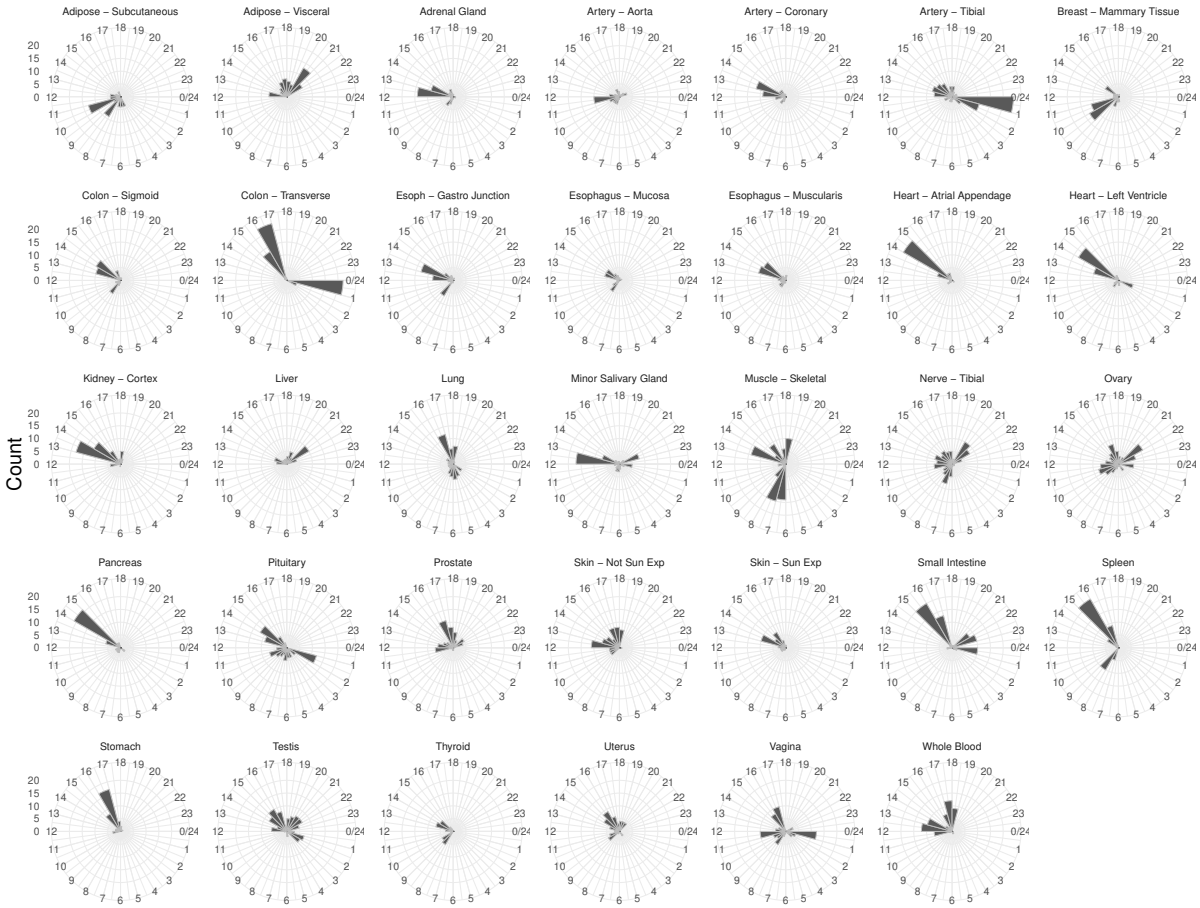


Figure 9: Radial plot of the distribution of the CIRCUST estimated peak phases of the TOP genes along the 34 tissues analyzed from GTEx. Active/lightened period $(0, \pi]$ is identified with [6am,6pm) and the inactive/darkness period $(\pi, 2\pi]$ is done with [6pm,6am) with π corresponding with 6pm.

Data and Code availability

CIRCUST methodology is publicly available at `https://github.com/yolandalago/CIRCUST/`.

Supplemental Materials

Supplemental material for this paper is available online.

Author Contribution

Y.L., C.R., R.S. and F.A.J.L.S conceived of the presented idea. R.S. provided partial data. Y.L. and C.R. developed the theoretical proposal, and conceptual design and analyzed the results. Y.L. developed computational code, processed the data, performed the computations, and design the analysis to validate the methodology. I.M., R.S. and F.A.J.L.S assisted with the discussion. Y.L. wrote the manuscript with input from all authors. All authors approved the final manuscript.

Declaration of conflicting interests

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. 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.

Funding

C.R. and Y.L gratefully acknowledge the financial support received by the Spanish Ministerio de Ciencia e Innovación [PID2019-106363RB-I00]. I.M. thanks the support received from the National Institutes of Health (NIH) for grants R01 HL140574 and T32 HL7901-20 and American Heart Association grant 19POST34380188. R.S. gratefully acknowledge the financial support received by the NIH for grants R01-DK102696, R01-DK105072, R01-DK107859, and R01-HL146751. F.A.J.L.S. thanks to the NIH for the support received for grants R01-DK102696, R01-DK105072, R01-HL140574, and R01-HL153969.

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* **111**, 16219–16224 (2014).

[2] Antoch, M. P. & Kondratov, R. V. Pharmacological modulators of the circadian clock as potential therapeutic drugs: focus on genotoxic/anticancer therapy. *Handbook of experimental pharmacology* **217** (2013).

[3] Sulli, G., Manoogian, E., Taub, P. R. & Panda, S. Training the circadian clock, clocking the drugs, and drugging the clock to prevent, manage, and treat chronic diseases. *Trends in pharmacological sciences* **39** (2018).

[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* **19**, 5170–5183 (2021).

[5] Brown, S. A. *et al.* The period length of fibroblast circadian gene expression varies widely among human individuals. *PLOS Biology* **3** (2005).

[6] Mavroudis, P. & Jusko, W. Mathematical modeling of mammalian circadian clocks affecting drug and disease responses. *J Pharmacokinet Pharmacodyn* **48**, 375–386 (2021).

[7] Chen, C.-Y. *et al.* Effects of aging on circadian patterns of gene expression in the human prefrontal cortex. *Proceedings of the National Academy of Sciences* **113**, 206–211 (2016).

[8] Seney, M. L. *et al.* Diurnal rhythms in gene expression in the prefrontal cortex in schizophrenia. *Nature Communications* **10** (2019).

[9] Zhu, Y., Wang, L., Yin, Y. & Yang, E. Systematic analysis of gene expression patterns associated with postmortem interval in human tissues. *Scientific Reports* **7**, 5435 (2017).

[10] Leng, N. *et al.* Oscope identifies oscillatory genes in unsynchronized single-cell rna-seq experiments. *Nature methods* **12**, 947–950 (2015).

[11] Liu, Z., Lou, H., Xie, K. *et al.* Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nature Communications* **8** (2017).

[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* **114**, 5312–5317 (2017).

[13] Ruben, M. D. *et al.* A database of tissue-specific rhythmically expressed human genes has potential applications in circadian medicine. *Science Translational Medicine* **10** (2018).

[14] Wu, G. *et al.* Population-level rhythms in human skin with implications for circadian medicine. *Proceedings of the National Academy of Sciences* **115**, 12313–12318 (2018).

[15] Larriba, Y., Rueda, C., Fernández, M. & Peddada, S. Order restricted inference in chronobiology. *Statistics in Medicine* **39**, 265–278 (2020).

[16] Cornelissen, G. Cosinor-based rhythmometry. *Theoretical biology & medical modelling* **11**, 16 (2014).

[17] Rueda, C., Larriba, Y. & Peddada, S. Frequency modulated möbius model accurately predicts rhythmic signals in biological and physical sciences. *Scientific Reports* **9**, 18701 (2019).

[18] Olmo, M. d., Grabe, S. & Herzel, H. *Circadian Regulation: Methods and Protocols*, chap. Mathematical Modeling in Circadian Rhythmicity, 55–80 (Springer US, New York, NY, 2022).

[19] James, F. O., Boivin, D. B., 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* **24**, 1009–1034 (2007).

[20] Christou, S. *et al.* Circadian regulation in human white adipose tissue revealed by transcriptome and metabolic network analysis. *Scientific reports* **9**, 2641 (2019).

[21] Watanabe, M. *et al.* Rhythmic expression of circadian clock genes in human leukocytes and beard hair follicle cells. *Biochemical and biophysical research communications* **425**, 902–907 (2012).

[22] Fair, B. J. *et al.* Gene expression variability in human and chimpanzee populations share common determinants. *eLife* **9**, e59929 (2020).

[23] Mure, L. S. *et al.* Diurnal transcriptome atlas of a primate across major neural and peripheral tissues. *Science* **359** (2018).

[24] Consortium, G. The genotype-tissue expression (gtex) project. *Nature genetics* **45**, 580–585 (2013).

[25] Burgess, H. J. & Fogg, L. F. Individual differences in the amount and timing of salivary melatonin secretion. *PloS one* **3**, e3055 (2008).

[26] Sack, R. L. *et al.* Circadian rhythm sleep disorders: part i, basic principles, shift work and jet lag disorders. *Sleep* **30**, 1460–1483 (2007).

[27] Sack, R. L. *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* **30**, 1484–1501 (2007).

[28] Kirby, M. J. & Miranda, R. Circular nodes in neural networks. *Neural Comput.* **8**, 390–402 (1996).

[29] Scholz, M. Analysing periodic phenomena by circular pca. In *Proceedings of the Conference on Bioinformatics Research and Development*, vol. 4414, 38–47 (2007).

[30] Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences* **97**, 10101–10106 (2000).

[31] Wu, G. *et al.* A population-based gene expression signature of molecular clock phase from a single epidermal sample. *Genome medicine* **12**, 73 (2020).

[32] Rao, J. & Sengupta, A. *Topics in circular statistics*, vol. 5 (2001).

[33] 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* **20**, 3977 (2019).

[34] Wucher, V., Sodaei, R., Amador, R., Irimia, M. & Guigó, R. Day-night and seasonal variation of human gene expression across tissues. *bioRxiv: the preprint server for biology* (2021).

[35] Fisher, N. I. *Statistical Analysis of Circular Data* (Cambridge University Press, 1993).

[36] Adzhar, R. *Outlier detection in circular data and circular-circular regression model/Adzhar Rambli*. Ph.D. thesis, Universiti Malaya (2011).

[37] Agostinelli, C. & Lund, U. *R package `circular`: Circular Statistics (version 0.4-95)*. CA: Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University, Venice, Italy. UL: Department of Statistics, California Polytechnic State University, San Luis Obispo, California, USA (2022). URL `https://r-forge.r-project.org/projects/circular/`.

[38] Hughey, J. & Butte, A. Differential phasing between circadian clocks in the brain and peripheral organs in humans. *Journal of biological rhythms* **31**, 588–597 (2016).

[39] Lopes-Ramos, C. M. *et al.* Transcriptional landscape of cell lines and their tissues of origin. *bioRxiv* (2016).

[40] Brinkmeyer-Langford, C. L., Guan, J., Ji, G. & Cai, J. J. Aging shapes the population-mean and -dispersion of gene expression in human brains. *Frontiers in Aging Neuroscience* **8**, 183 (2016).

[41] 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* **11**, 955 (2020).

[42] Yang, J. *et al.* Abstract 296: Sorting nexin 19: A novel regulator of renal dopamine d 1 receptor. *Hypertension* **64** (2014).

[43] Saric, A. *et al.* Snx19 restricts endolysosome motility through contacts with the endoplasmic reticulum. *Nature Communications* **12**, 4552 (2021).