CYCLOPS reveals human transcriptional rhythms in health and disease

Ron C. Anafi[a,b,c,1], Lauren J. Francey[d,e,f], John B. Hogenesch[d,e,f], and Junhyong Kim[g,h]

[a]Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104; [b]Center for Sleep and Circadian Neurobiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104; [c]Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104; [d]Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229; [e]Center for Chronobiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229; [f]Perinatal Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229; [g]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104; and [h]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104

Edited by Joseph S. Takahashi, Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX, and approved March 20, 2017 (received for review November 23, 2016)

Circadian rhythms modulate many aspects of physiology. Knowledge of the molecular basis of these rhythms has exploded in the last 20 years. However, most of these data are from model organisms, and translation to clinical practice has been limited. Here, we present an approach to identify molecular rhythms in humans from thousands of unordered expression measurements. Our algorithm, cyclic ordering by periodic structure (CYCLOPS), uses evolutionary conservation and machine learning to identify elliptical structure in high-dimensional data. From this structure, CYCLOPS estimates the phase of each sample. We validated CYCLOPS using temporally ordered mouse and human data and demonstrated its consistency on human data from two independent research sites. We used this approach to identify rhythmic transcripts in human liver and lung, including hundreds of drug targets and disease genes. Importantly, for many genes, the circadian variation in expression exceeded variation from genetic and other environmental factors. We also analyzed hepatocellular carcinoma samples and show these solid tumors maintain circadian function but with aberrant output. Finally, to show how this method can catalyze medical translation, we show that dosage time can temporally segregate efficacy from dose-limiting toxicity of streptozocin, a chemotherapeutic drug. In sum, these data show the power of CYCLOPS and temporal reconstruction in bridging basic circadian research and clinical medicine.

gene expression | biological rhythms | machine learning | autoencoder | circadian rhythms

Circadian rhythms are nearly ubiquitous in nature. In animals, much of physiology and behavior is under circadian control. Body temperature, hormonal rhythms, blood pressure, and locomotor activity are just a few of the processes displaying daily rhythms. In circadian model systems (e.g., cyanobacteria, *Neurospora*, *Arabidopsis*, *Drosophila*, and mice), high-resolution time sampling is straightforward, and experiments show that a substantial fraction of the transcriptome is under clock control. For example, in mice, a majority of genes are clock regulated in at least 1 of 12 different organs (1).

Circadian rhythms are also critical for humans. Shift work-induced circadian misalignment is associated with higher rates of metabolic, cardiovascular, and neoplastic disease. Clinical experience suggests time of day can have a marked effect on disease severity (2–4). Indeed, the majority of the best-selling prescription drugs and World Health Organization essential medicines target molecules that oscillate in mice (1). However, translation of these findings to clinical medicine remains slow. How does human molecular physiology change with circadian time? In mice, and presumably humans, circadian output genes are markedly different in each tissue. Obviously, repeated sampling from most human organs is not possible. As a result, we have limited ability to study human molecular rhythms and relate them to either normal or disease physiology.

One approach is to analyze temporally annotated clinical samples, where time of sample collection is recorded. There are >1 million human gene expression samples in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) repository. Unfortunately, the sample collection time is almost never reported. Ueda et al. (5) first used transcriptional "time-stamping" to reconstruct the circadian phase of tissue samples from mouse liver, and supervised learning methods continue to improve (6, 7). However, supervised learning requires a training library of samples with known circadian time. With the exception of blood (8, 9) and brain (10), temporally annotated human samples are lacking. Although theoretically possible, scheduling people for internal organ biopsies every 2 h for 2 d is both dangerous and impractical.

Alternatively, in single-cell biology, unsupervised algorithms are being used to reconstruct the relative temporal order of samples, for example, in cellular development and differentiation (11). Orderings that minimize the distance between adjacent samples or maximize the smoothness of the trajectories connecting them are calculated directly from gene expression data. For example, Oscope is designed to extract oscillatory (cell cycle) dynamics from single-cell data (12). To do this, Oscope compares every gene-by-gene pairing in the genome to identify those that best approximate an ellipse. In addition to being computationally taxing, this approach is highly sensitive to systematic (nonrhythmic) intersubject variation found in clinical samples.

Here, we describe a method, cyclic ordering by periodic structure (CYCLOPS), that uses global descriptors of expression structure, unsupervised machine learning, and evolutionary conservation, to order periodic data. We show CYCLOPS is robust by analyzing legacy mouse and human data, where time is known. We

Significance

Circadian rhythms influence most aspects of physiology and behavior. However, how do we apply this knowledge in medicine? Identifying molecular mechanisms in humans is challenging as existing large-scale datasets rarely include time of day. To address this problem, we combine understanding of periodic structure, evolutionary conservation, and unsupervised machine learning to order unordered human biopsy data along a periodic cycle. We show this works using ordered mouse and human data and that it gives consistent results when applied to populations on different continents. Then, we investigate molecular rhythms in normal human lung and liver and cancerous liver. Finally, we demonstrate proof of concept by finding the best time to administer a chemotherapeutic drug in an animal model.

Author contributions: R.C.A., J.B.H., and J.K. designed research; R.C.A. and L.J.F. performed research; R.C.A. and J.B.H. contributed new reagents/analytic tools; R.C.A., L.J.F., J.B.H., and J.K. analyzed data; and R.C.A., L.J.F., J.B.H., and J.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

See Commentary on page 5069.

[1]To whom correspondence should be addressed. Email: ron.anafi@uphs.upenn.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1619320114/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1619320114

also demonstrate remarkably consistent results when analyzing unordered human data from different geographical populations. We report the cycling of hundreds of human disease genes and drug targets. We also analyze the altered circadian function of hepatocellular carcinoma (HCC) samples. Finally, for proof of concept, we used this information to design a dosing scheme that temporally segregates efficacy from toxicity for streptozocin (STZ), a cytotoxic chemotherapeutic agent.

Results

Data generated by a common periodic process have a defined structure. Analyzing the yeast cell cycle, Alter, Brown, and Botstein (13) used singular value decomposition to reduce the dimensionality of the data and identify "eigengenes," characteristic expression patterns, that span the global expression profiles. Alter et al. recognized the first eigengenes as out-of-phase sinusoidal oscillations. When plotted in expression space, they form an ellipse. Importantly, this result is independent of the annotated collection time and can be used to determine the relative order of samples in the dataset (Fig. S1).

With human data, confounds such as genetic differences, age, gender, exercise, diet, etc., all add significant noise and limit this approach. Circadian and noncircadian patterns can be mixed and distributed among the various eigengenes. CYCLOPS optimally weights and combines the eigengenes patterns to reveal underlying elliptical structure, and then uses this structure to order the data. CYCLOPS couples our prior knowledge of rhythms in model organisms with use of a circular node autoencoder (Fig. S1*D*). Autoencoders are feedforward neural networks trained so that the network's output reproduces its input (14). By constraining the size of the intervening "bottleneck layer," the network is forced to encode the data in a reduced number of dimensions. Here, we combine linear encoding and decoding neurons with a circular bottleneck node (15). The outputs of the two coupled circular bottleneck nodes represent a single angular phase. CYCLOPS linearly projects the data and encodes it on a simple elliptical curve (15). In this way, CYCLOPS identifies a closed curve that best represents the characteristic expression patterns. An angular phase represents the position of each sample on the ellipse and its temporal phase in the reconstructed periodic cycle. Circular autoencoders have been used to generate nonlinear models of periodic processes in nature (16, 17). To our knowledge, their use in ordering these data are novel.

We first applied CYCLOPS to mouse time course expression data (1, 18). With no prior knowledge, CYCLOPS correctly ordered the samples from mouse liver (Fig. 1*A*). The circular correlation (ρ_c) (19) and the circular rank correlation(η_c) (19) between the CYCLOPS-estimated phases and true circadian times were both greater than 0.9. CYCLOPS also ordered data from other highly rhythmic organs (e.g., lung, kidney, and adrenals) but failed to correctly order data from tissues with weaker circadian signals (e.g., skeletal muscle, cerebellum, and brainstem; Fig. S2). Reasoning that prior biological knowledge could increase the signal-to-noise ratio and improve ordering, we restricted the analysis to either a list of transcripts that cycled in that tissue or a list of transcripts found to cycle in >75% of other tissues. With this method, CYCLOPS was able to correctly order samples for all mouse tissues (Fig. S2).

CYCLOPS was developed to analyze data without an annotated order. Thus, assessing the quality of CYCLOPS orderings when the true order is unknown is important. CYCLOPS computes a quickly interpretable smoothness metric, Met^{smooth}, and a more computationally intensive error statistic, $Stat^{err}$, the significance of which is assessed by bootstrap. Met^{smooth} compares the smoothness of the reconstructed circular trajectory in expression space to the smoothness of a linear ordering based on the first principal component. $Stat^{err}$ describes the improvement in the residual sum of squares error when encoding the data onto a closed, one-dimensional elliptical manifold compared with the residual error when encoding the data onto a one-dimensional linear manifold. In the cases where $Met^{smooth} < 1$ and $Stat^{err}$ differed from background ($P < 0.05$), the ordering was generally well correlated to ground truth (Fig. S2).

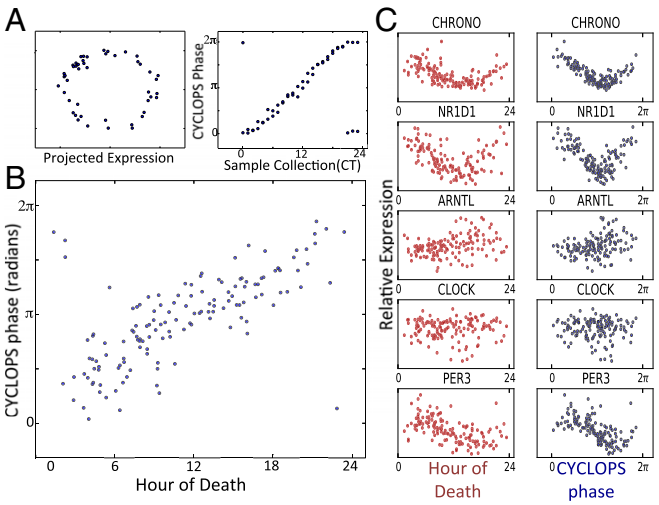

Fig. 1. Validation of CYCLOPS. Time course expression data from the mouse liver (18) were encoded with CYCLOPS. (*A*, *Left*) The linear encoding is visualized as a projection onto a plane where the data approximates an ellipse. (*A*, *Right*) Sample collection phase is plotted along the horizontal axis, whereas the CYCLOPS-estimated phase is plotted on the vertical axis. (*B*) Expression data from 146 human prefrontal cortex samples (10) encoded with CYCLOPS. The hour of death for each sample is plotted on the horizontal axis. The CYCLOPS-derived phases are plotted on the vertical axis. Time 0 is the same as 24 and phase 0 is the same as 2π; samples plotted near the corners of the graph are actually "near" the diagonal line of identity. (*C*) Expression of select transcripts is plotted as a function of both TOD (red) and CYCLOPS phase (blue).

Next, we applied CYCLOPS to expression data derived from human prefrontal cortex samples obtained at autopsy (10). Following the CYCLOPS methodology, we used evolutionary conservation and knowledge of murine rhythms to sharpen the expected circadian signature. We restricted the list of transcripts used for temporal reconstruction to human homologs of genes found to cycle in >75% of mouse tissues. CYCLOPS produced a high-quality ordering ($Met^{smooth} < 1$, $P < 0.05$) that provides an excellent estimate of time of death (TOD) ($\rho_c = 0.68$, $\eta_c = 0.55$, median absolute error = 1.69 h) (Fig. 1*B*). When the expression of individual transcripts is plotted as a function of either CYCLOPS phase or TOD (Fig. 1*C*), *CHRONO* (20) was found to have the strongest circadian cycling. Known clock genes *NR1D1* and *PER3* also showed clear rhythms. More generally, transcripts that cycled as a function of TOD also cycled as a function of CYCLOPS phase, whereas nonrhythmic transcripts by TOD were also nonrhythmic by CYCLOPS phase. Sinusoidal fits to CYCLOPS phase were slightly better than sinusoidal fits to TOD (Fig. 1*C*). We hypothesize that CYCLOPS better accounts for interindividual differences in circadian entrainment to the terrestrial day, for example, due to shift work, biological variation, or the poor entraining conditions of hospitals.

Then we applied CYCLOPS to biopsy data describing the normal human pulmonary transcriptome (21). Human pulmonary physiology demonstrates clear circadian rhythms. However, to our knowledge, molecular rhythms in the human lung remain unexamined. We confined the CYCLOPS reconstruction to human homologs of genes that cycle in the mouse lung. We independently analyzed data from Groningen and Quebec City (22) and used modified cosinor regression to identify transcripts well described by a sinusoidal function of CYCLOPS phase in both datasets (23) (Dataset S1). The phase of peak expression of each transcript was remarkably consistent between research sites ($\rho_c = 0.66$, median absolute discrepancy = 0.32 radians ~1.2 h) (Fig. 2*A*). Known circadian genes, including *CLOCK*, *CRY1*, and *CRY2* were periodic with phase relationships similar to those seen in mouse (Fig. 2*B*).

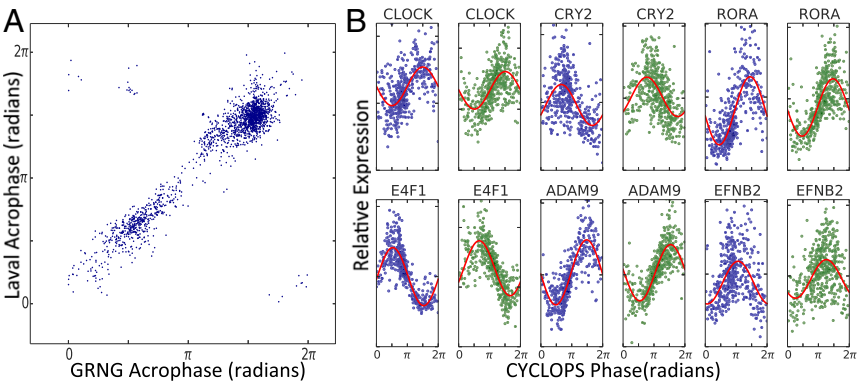

Fig. 2. CYCLOPS analysis of circadian transcriptome in human lung. Using independent biopsy data sets (21) from the University of Groningen (GRNG) (Groningen, The Netherlands) and the University of Laval (Quebec City, QC, Canada), we used CYCLOPS to generate two reconstructions of the circadian transcriptome in the human lung. Modified cosinor regression was then used to identify cycling transcripts. (*A*) Results from the transcripts found to cycle in both datasets are shown. For each transcript, the acrophase in the Laval dataset is plotted against the transcript acrophase as determined from the Groningen data. (*B*) CYCLOPS-ordered expression data from Groningen and Quebec City are plotted in blue and green, respectively.

Clinically important transcripts also showed strong cycling (Fig. 2*B* and Fig. S3). For example, *ADAM9* is implicated in lung cancer and is a risk marker for distant metastases (24). *EFNB2*, a receptor tyrosine kinase (TK), also cycled strongly and may have prognostic significance in both small cell lung cancer and nonpulmonary cancers (25). We used the Drug Signatures Database to identify rhythms in drug targets (Dataset S2) (26). Several drug target classes in asthma treatment were rhythmic, including β-adrenergic receptors (targeted by β-agonists) and glucocorticoid receptors (targeted by inhaled and systemic steroids). Various TKs cycled (e.g., *MAP4K1*, *MAP4k3*, *SLK*, *FYN*, *KDR*, *PKN2*, *TAOK*, and *TAOK2*). Several of these are targeted in the treatment of non–small-cell lung cancer and pulmonary fibrosis (22, 27).

Drugs used for nonrespiratory conditions that act via the pulmonary system also target rhythmic molecules. Angiotensin-converting enzyme (ACE) inhibitors are used in the treatment of hypertension and heart failure. Inhibiting ACE reduces the production of the potent vasoconstrictor Angiotensin II (28). ACE is predominantly localized to the pulmonary and renal vasculatures and, per CYCLOPS, demonstrates a marked diurnal fluctuation in human lung. Night-time dosing of ACE inhibitors improves nocturnal blood pressure control without sacrificing daytime efficacy (29). The cycling of pulmonary ACE may provide the underlying molecular mechanism for this findings.

To identify biological pathways and processes that show circadian coordination in the human lung, we applied phase set enrichment analysis (PSEA) (30). As in the mouse (30), pathways describing cell cycle regulation, adaptive immune function, and channel-mediated transport demonstrate phase-synchronized expression (Fig. S4). These data are consistent with clinical evidence demonstrating diurnal variation in the symptoms of asthma (31) and the efficacy of cell cycle-targeting chemotherapeutic agents (32). The SMAD and TGF-β pathways were among those that demonstrated the strongest phase clustering. Both have recently been highlighted in the pathogenesis of pulmonary fibrosis and nonsmall cell lung cancer (33, 34).

Of note, temporal reconstruction with CYCLOPS did not uniformly distribute samples across the circadian cycle (Fig. S5). Biopsies are obtained during surgical working hours (~6:00 AM to 6:00 PM). However, samples obtained from shift workers during the terrestrial day likely provide data describing the circadian night (sleep period). The phase distribution of samples is consistent with US data that ~15–20% of the population are shift workers (35). Of course, the effect of shift work on local tissue clocks remains incompletely understood. It is possible that circadian perturbations alter local molecular timekeeping in a tissue-dependent manner, resulting in intertissue (36) or intratissue (30) desynchrony.

Next, we wanted to examine circadian rhythms in a cancerous and paired normal organ. We applied CYCLOPS to expression data from 249 patient biopsies of noncancerous (NC) liver tissue (37). The vast majority (n = 243) were of the "normal margin" adjacent to tumor. Using homologs of the transcripts that cycle in the mouse liver (1), CYCLOPS was able to order the samples ($Met^{smooth} < 1$, $P < 0.05$). Core clock components showed similar phase relationships to those observed in mouse (Fig. 3*A*). A full list of transcripts and pathways found to cycle in NC human liver are presented in (Datasets S3 and S4). Pathways describing metabolism, lipid and cholesterol processing, and cell cycle regulation all demonstrated strong circadian cycling.

We used data from biopsies of HCC to explore transcriptional rhythms in an intact solid human tumor (37). HCC is the most common primary liver cancer. We initially analyzed the HCC data as we did the normal margin data, seeding the ordering on the human homologs of mouse cycling genes (1). However, we were not able to generate a quality ordering in this way. We reasoned that HCC might compromise clock function or that the increased interindividual variation between neoplastic samples may have confounded CYCLOPS. To reduce the influence of neoplastic variability and emphasize circadian variation, HCC expression data were projected onto the eigenvectors established by the NC samples. Applying CYCLOPS to these data produced a high-quality fit ($P < 0.05$). We then used cosinor regression analysis to identify cycling transcripts.

Surprisingly, most "core clock" components continued to cycle in HCC samples. Notable exceptions were *PER1* and *CRY1* (Fig. S6). Nearly one-half of the genes cycling in NC samples were not well fit by cosinor regression in the HCC data. Again, we wondered whether this might reflect increased "noise" among HCC samples rather than a true change in circadian expression. We used a nested modeling approach to better distinguish these possibilities. Pooled, ordered expression data from both HCC and NC samples were first fit with a single (sinusoidal) model. We then tested whether adding additional sinusoidal terms dependent on histological status significantly improved fit. The combined modeling framework allowed us to identify transcripts that cycled in NC samples but (*i*) were not well fit by a sinusoidal function when HCC samples were fit in isolation, (*ii*) were significantly better fit by a nested model with different circadian parameters for HCC and NC samples, and (*iii*) had at least a twofold reduction in amplitude among HCC samples in the pooled model (Fig. 3 *B* and *C*). Based on these combined criteria, we estimate that ~15% of the transcripts that cycled in NC samples lost rhythmic expression in HCC.

Using DAVID (38), we identified pathways overrepresented among genes that lost rhythmicity in HCC. In a related analysis, we ranked all circadian genes in NC samples by the reduction of their amplitude in HCC. The ranked list was analyzed with gene set enrichment analysis (GSEA) (39). Reassuringly, these analyses yielded overlapping results (Table S1). There was temporal deregulation of key circadian outputs including overlapping apoptotic pathways and JAK–STAT signaling. We also find evidence for reduced cycling among transcripts related to hypoxia and redox metabolism. Of note was loss of rhythmicity in TKs targeted by several latest-generation antineoplastic agents. Also notable was a loss of cycling in *ARNTL2*, which has been implicated in several neoplastic diseases (40, 41).

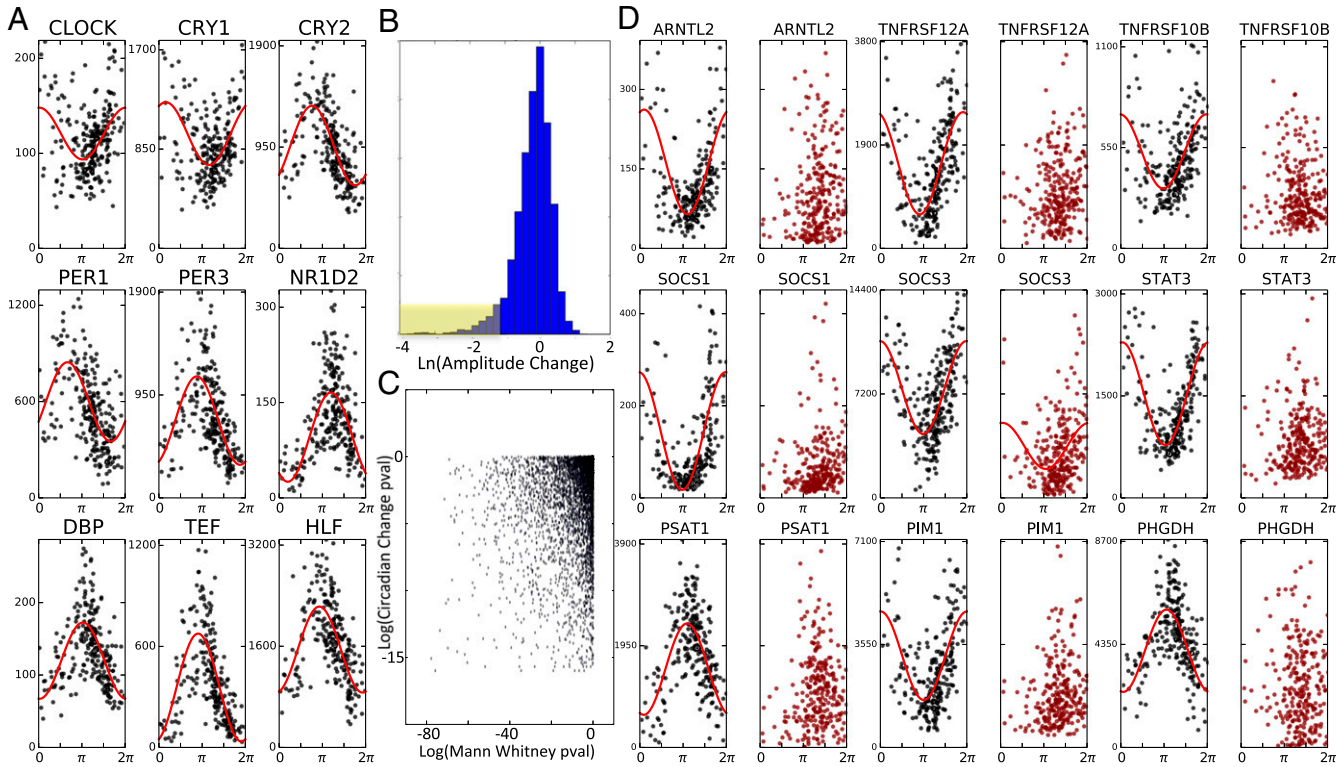

Fig. 3. CYCLOPS analysis of noncancerous (NC) and cancerous (HCC) human liver. Expression data from biopsy-derived NC tissue was processed using CYCLOPS. (*A*) Reconstructed expression profiles of selected clock genes are plotted as a function of CYCLOPS phase. Expression data from samples with HCC were projected onto the eigenvectors established in the NC samples before CYCLOPS ordering. (*B*) Histogram of circadian amplitude differences between NC and HCC samples. A long tail, highlighted in yellow, shows transcripts with reduced amplitude in HCC. (*C*) A scatter plot compares the statistical significance of testing for a change in mean expression (Mann–Whitney test) with the statistical significance of testing for a circadian expression change. (*D*) Expression of selected genes as a function of CYCLOPS phase in both NC (black) and HCC (red) samples.

Chronotherapy is an immediate area of interest for clinical translation. Earlier, we proposed that drugs that target rhythmic, high-amplitude gene products represent a path for mechanism-driven chronotherapy. With CYCLOPS, we can now identify drug targets that oscillate in humans. Among the many transcripts with high-amplitude oscillations in normal human liver was *SLC2A2* (Fig. 4*A*). Murine *Slc2a2* cycles with similar temporal phasing in both the liver and kidney (1). *SLC2A2* encodes GLUT2, a glucose transporter highly expressed in pancreas, liver, and kidney. STZ is a GLUT2 substrate and is standard of care in patients with locally advanced pancreatic neuroendocrine tumors (pNETs) (42). Although pNETS are rare, the incidence has nearly doubled in the last decades (43). STZ is cytotoxic to GLUT2-expressing cells, including islet cells and pNETS, with renal and hepatic toxicity being dose-limiting and potentially lethal (42).

As STZ has a remarkably short half-life (<15 min), it is an excellent candidate for chronotherapy. We reasoned dosing STZ during the nadir of hepatic SLC2A2 abundance could preserve STZ efficacy while minimizing renal and hepatic toxicity. The same dose of STZ was administered in the morning [Zeitgeber time (ZT) 0] or evening (ZT 12) to DBA/2J mice (44) for 5 consecutive days. We measured blood glucose levels as a surrogate marker for the efficacy of STZ in killing islet cells. Body weight was used as a simple measure of animal health and gross toxicity. Mice treated with STZ at either time were equally susceptible to hyperglycemia (Fig. 4*B*). However, mice administered STZ in the morning, when *Slc2a2* transcript expression is low and GLUT2 protein abundance is high (1, 45), had a much greater loss in body mass compared with mice receiving STZ in the evening (−19.8 g vs. −12.9 g, $P = 0.015$). Thus, we temporally separated apparent efficacy (hyperglycemia) from toxicity (loss of body weight).

Discussion

Much of the molecular mechanics underlying circadian rhythms has been revealed in the last two decades. Much less progress has been made in converting these findings into actionable clinical knowledge. The lack of human time course data has presented a key barrier to translation. CYCLOPS aims to address this deficiency, using global descriptors of gene expression, evolutionary conservation, and machine learning to order unordered data within a periodic cycle. CYCLOPS builds on the foundation of Alter et al. (13) and the computational structure of Kirby and Miranda (15) to order high-throughput data and identify latent periodic oscillations in transcription. We validated CYCLOPS using ordered mouse and human data. We also demonstrated the consistency of CYCLOPS using human lung data from two distinct patient populations on separate continents.

CYCLOPS has advantages and disadvantages compared with existing methods. Supervised methods (e.g., ZeitZeiger) continue to improve but require time course training data. Obtaining blood and skin samples is straightforward. Serial biopsies of internal human organs are not practical. Unsupervised methods, like Oscope, have recovered cell cycle rhythms from unordered single-cell data. However, Oscope works on the single transcript level and requires thousands more computations than does CYCLOPS. Furthermore, Oscope is highly sensitive to the inter-subject variability inherent to human data. Supervised methods are tissue specific and are similarly sensitive to biologic variability (as might be expected in cancer), as they have been optimized to use only a small number of highly informative transcripts. CYCLOPS uses global descriptors of expression structure, making it both robust and efficient for population-based human data. However, as with other high-dimensional bioinformatics methods, the particular

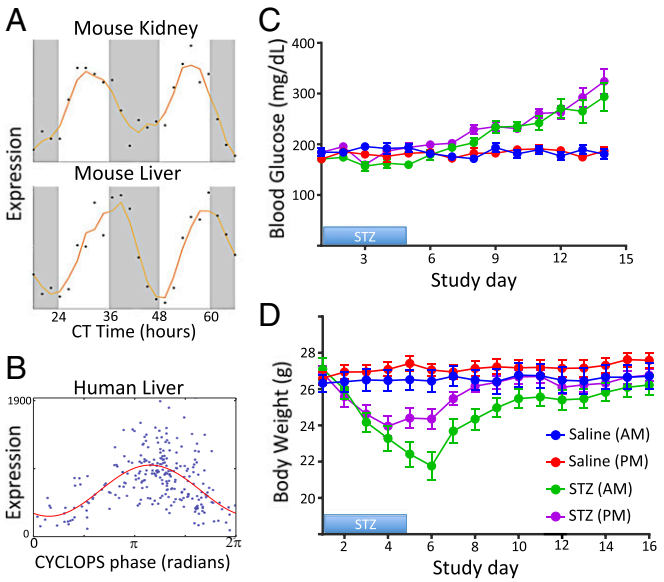

Fig. 4. Prospective chronotherapy for streptozocin (STZ). STZ is a cytotoxic agent used to treat pancreatic neuroendocrine tumors. STZ is actively transported into cells by the protein product of *SLC2A2* and is associated with renal and hepatic toxicity. (*A*) The expression of *Slc2a2* in mouse kidney and liver (1) is plotted as a function of circadian time. (*B*) Expression of *SLC2A2* in human liver samples is plotted as a function of CYCLOPS phase. (*C*) Eleven-week-old male mice were dosed with STZ (green and purple) or saline (blue and red) at 7:00 AM (blue and green) or 7:00 PM (red and purple). Dosing time did not significantly impact the induction of hyperglycemia and expected treatment efficacy. (*D*) Body weight was used as a measure of gross toxicity. There was less weight loss among mice administered STZ at 7:00 PM.

data normalization scheme and descriptors of expression structure used can influence the final results.

CYCLOPS also has several limitations. It requires data from the entire periodic cycle to form an ellipse. Biopsies are almost exclusively obtained during the day. A large patient population, including shift workers, is necessary to fill in underrepresented times of day. Our experience suggests that >250 samples are required to order biopsy samples (Table S2). We also leveraged evolutionary conservation and mouse data to focus the genes used for human temporal reconstruction. CYCLOPS does not require that rhythms in mice and men are identical but does assume that the human homologs of mouse cycling genes are more likely to cycle. Importantly, CYCLOPS identifies features that are consistent with oscillations with respect to a latent variable, assumed to be time. Several findings lend confidence to our reconstructions. First, we recovered oscillations consistent with known circadian biology (e.g., phase relationships of core clock genes). We also recovered sample collection phases consistent with biopsy collection times and smooth orderings that well explain the data. CYCLOPS orderings are also relative. Additional information is needed to assign a circadian time to any particular CYCLOPS phase. In ordering the human lung and liver transcriptomes, we used the average acrophase of the PAR bZip transcription factors to fix time "π." In the lung and liver of nocturnal mice, these factors show peak expression near ZT12 (1), the beginning of the peak activity period.

Circadian rhythms persist in the absence of environmental cues. The observation of rhythms under normal conditions is not sufficient to classify a rhythm as circadian. Pending further study, the human transcriptional oscillations identified by CYCLOPS are more properly labeled as diurnal.

A final caveat lies in the identification of periodic transcripts from CYCLOPS-ordered data. Regression and other rhythm detection methods are predicated on time as a variable independent of expression. CYCLOPS phases are derived from gene expression. As a result, standard statistical significance tests tend to be too liberal. To mitigate this concern, we have imposed an unusually strict numerical cutoff for statistical significance. We also require cycling with sufficient amplitude to suggest physiologic importance.

Despite these limitations, we have successfully used CYCLOPS to explore diurnal rhythms in human lung, liver, and HCC. Our analyses of normal lung and liver present clear translational opportunities. We found strong circadian cycling of the cell cycle and immune pathways in human lung. ACE, well expressed in the pulmonary vasculature and a key drug target for hypertension, appeared rhythmic. We also found cycling in members of the SMADs and the JAK–STAT pathways along with various TKs, many of which are important targets in idiopathic pulmonary fibrosis.

In liver, *PPARA*, *DDC*, and *XDH*, targets of the fibrates, dopamine decarboxylase inhibitors, and xanthine oxidase inhibitors, respectively, all display high-amplitude rhythms (Fig. S7). *SLC2A2*, the target of STZ, also displayed strong cycling in human liver. In a proof-of-concept experiment, we leveraged these data to time STZ administration and segregate gross toxicity from efficacy. In sum, this approach presents a straightforward path from genome-scale human data to hypothesis-driven opportunities in chronotherapy.

An important aspect of chronotherapy is the accurate circadian assessment or "phasing" individual patients. However, how accurate must this be? The answer likely depends on the kinetics of the drug and the dynamics of its target. For STZ and other fast-acting drugs that target molecules with high-amplitude rhythms, there may be a broad window of acceptable dosing times. For other drugs, more temporal precision might be required.

CYCLOPS is an algorithm that temporally reconstructs population-based human organ data. Applying CYCLOPS to over 2,000 human samples, we observe clear, high-amplitude molecular rhythms in lung, liver, brain, and HCC. Despite disparities in patient age, gender, genetics, diet, and environment, CYCLOPS extracted significant periodic signatures. For a large subset of genes, circadian variability in expression was larger than the variability attributable to these aggregated genetic and environmental variables. By implication, circadian control may offer a powerful tool for precision medicine.

Finally, we investigated the state of circadian rhythms in a human cancer, HCC. The circadian clock is believed to gate the cell cycle. In HCC, we find that, despite continued oscillator function, there is circadian deregulation of JAK–STAT, apoptotic, and metabolic pathways. To catalyze the further pursuit of translational chronobiology, we have posted the CYCLOPS program and associated scripts on GitHub. We hope this and related approaches will propel investigation into the role of circadian biology in clinical medicine.

Methods

All animal studies were done under Charles River Laboratories study number 20091523 under Institutional Animal Care and Use Committee protocol P01182016A.

Microarray Processing. CEL files containing raw data were downloaded from NIH GEO and processed with RMA in R (version 3.2.3) Bioconductor.

Computational Methods. The CYCLOPS autoencoder and downstream analysis were implemented in Julia 0.3.10. The associated files are available for download on GitHub.

Data Scaling and Normalization. For temporal reconstruction, we first restricted the list of probes used to the top 10,000 highest expressed probes (as sorted by mean probe value). For each probe, we impute extreme expression values at the top/bottom 2.5th percentile. The expression $X_{i,j}$ of each probe i in sample j was scaled as follows:

$$S_{i,j} = \frac{(X_{i,j} - M_{i,j})}{M_i},$$

where M_i is the mean expression of probe i across samples: $M_i = (1/N_j)\sum_j X_{i,j}$.

The $S_{i,j}$ data were expressed in eigengene coordinates $E_{i,j}$ following the methods of Alter et al. The number of eigengenes N_E (singular values) retained was set so as to preserve 85% of the variance of the data. The autoencoder was applied to these characteristic expression patterns for the purposes of temporal reconstruction.

CYCLOPS Autoencoder. The activated value of neuron j in layer l is denoted by a_j^l and for linear neurons is given by $a_j^l = \sum_k w_{j,k}^l a_k^{l-1} + b_j^l$, where weight from the kth neuron in layer $l-1$ to the jth neuron in layer l is represented by $w_{j,k}^l$. The bias in jth neuron in layer l is denoted b_j^l (46).

A single, circular node was used in the bottleneck layer. The single circular neuron was implemented as two coupled neurons (15). The preactivation values of these neurons o_j^l and o_{j*}^l are given by the following:

$$o_j^l = \sum_k w_{j,k}^l a_k^{l-1} + b_j^l \; o_{j*}^l = \sum_k w_{j*,k}^l a_k^{l-1} + b_{j*}^l.$$

Activated values are obtained by mapping these onto the unit circle:

$$a_j^l = \frac{o_j^l}{\sqrt{\left(o_j^l\right)^2 + \left(o_{j*}^l\right)^2}} \; a_{j*}^l = \frac{o_{j*}^l}{\sqrt{\left(o_j^l\right)^2 + \left(o_{j*}^l\right)^2}},$$

with phase

$$\theta_j = \tan^{-1}\left(a_{j*}^l / a_j^l\right).$$

N_E linear neurons were used in both the encoding and decoding steps. The autoencoder was trained by backpropagation using stochastic batch gradient descent with momentum (46). Default training parameters were set as set batch size = 10, rate = 0.3, and momentum = 0.5.

Training is repeated multiple times (default = 40) starting at different, randomly set initial weighting conditions. The result with minimal sum of squares output error is used.

The fully trained autoencoder was used to encode the characteristic expression data $E_{i,j}$. The value of the circular node assigned each sample j ($\varnothing_j$) was the phase assigned to that sample.

The same autoencoder training parameters were used for all reconstructions.

Additional methodological details can be found in *SI Methods*.

ACKNOWLEDGMENTS. We thank Gang Wu, Robert Schmidt, and Marc Ruben for their critical reading of the manuscript and testing of the CYCLOPS program. We are grateful to researchers who generated the original datasets. This work is supported by Defense Advanced Research Projects Agency Grants D17AP00003 (to R.C.A.) and in part by D12AP00025, National Institute of Neurological Disorders and Stroke Grant 5R01NS054794-08 (to J.B.H.), in part by National Institute on Aging Grant 2P01AG017628-11, and the Penn Genome Frontiers Institute under a Health Research Formula Fund grant with the Pennsylvania Department of Health.

1. Zhang R, Lahens NF, Ballance HI, Hughes ME, Hogenesch JB (2014) A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proc Natl Acad Sci USA* 111:16219–16224.
2. Hetzel MR, Clark TJ (1980) Comparison of normal and asthmatic circadian rhythms in peak expiratory flow rate. *Thorax* 35:732–738.
3. Straub RH, Cutolo M (2007) Circadian rhythms in rheumatoid arthritis: Implications for pathophysiology and therapeutic management. *Arthritis Rheum* 56:399–408.
4. Ferrell JM, Chiang JYL (2015) Circadian rhythms in liver metabolism and disease. *Acta Pharm Sin B* 5:113–122.
5. Ueda HR, et al. (2004) Molecular-timetable methods for detection of body time and rhythm disorders from single-time-point genome-wide expression profiles. *Proc Natl Acad Sci USA* 101:11227–11232.
6. Hughey JJ, Hastie T, Butte AJ (2016) ZeitZeiger: Supervised learning for high-dimensional data from an oscillatory system. *Nucleic Acids Res* 44:e80.
7. Agostinelli F, Ceglia N, Shahbaba B, Sassone-Corsi P, Baldi P (2016) What time is it? Deep learning approaches for circadian rhythms. *Bioinformatics* 32:i8–i17.
8. Möller-Levet CS, et al. (2013) Effects of insufficient sleep on circadian rhythmicity and expression amplitude of the human blood transcriptome. *Proc Natl Acad Sci USA* 110: E1132–E1141.
9. Arnardottir ES, et al. (2014) Blood-gene expression reveals reduced circadian rhythmicity in individuals resistant to sleep deprivation. *Sleep* 37:1589–1600.
10. Chen C-Y, et al. (2016) Effects of aging on circadian patterns of gene expression in the human prefrontal cortex. *Proc Natl Acad Sci USA* 113:206–211.
11. Trapnell C, et al. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 32:381–386.
12. Leng N, et al. (2015) Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat Methods* 12:947–950.
13. Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* 97:10101–10106.
14. Kramer MA (1991) Nonlinear principal component analysis using autoassociative neural networks. *AIChE J* 37:233–243.
15. Kirby MJ, Miranda R (1996) Circular nodes in neural networks. *Neural Comput* 8:390–402.
16. Scholz M (2007) Analysing periodic phenomena by circular PCA. Bioinformatics *Research* and *Development* (Springer, Berlin), pp 38–47.
17. Hsieh WW (2001) Nonlinear principal component analysis by neural networks. *Tellus A* 53:599–615.
18. Hughes ME, et al. (2009) Harmonics of circadian gene transcription in mammals. *PLoS Genet* 5:e1000442.
19. Jammalamadaka SR, Sengupta A (2001) *Topics in Circular Statistics* (World Scientific, Singapore).
20. Anafi RC, et al. (2014) Machine learning helps identify CHRONO as a circadian clock component. *PLoS Biol* 12:e1001840.
21. Bossé Y, et al. (2012) Molecular signature of smoking in human lung tissues. *Cancer Res* 72:3753–3763.
22. Sgambato A, et al. (2012) The role of EGFR tyrosine kinase inhibitors in the first-line treatment of advanced non small cell lung cancer patients harboring EGFR mutation. *Curr Med Chem* 19:3337–3352.
23. Refinetti R, Lissen GC, Halberg F (2007) Procedures for numerical analysis of circadian rhythms. *Biol Rhythm Res* 38:275–325.
24. Lin C-Y, et al. (2014) ADAM9 promotes lung cancer metastases to brain by a plasminogen activator-based pathway. *Cancer Res* 74:5229–5243.
25. Brantley-Sieders DM (2012) Clinical relevance of Ephs and ephrins in cancer: Lessons from breast, colorectal, and lung cancer profiling. *Semin Cell Dev Biol* 23:102–108.
26. Yoo M, et al. (2015) DSigDB: Drug signatures database for gene set analysis. *Bioinformatics* 31:3069–3071.
27. Richeldi L, et al. (2011) Efficacy of a tyrosine kinase inhibitor in idiopathic pulmonary fibrosis. *N Engl J Med* 365:1079–1087.
28. Bader M (2010) Tissue renin-angiotensin-aldosterone systems: Targets for pharmacological therapy. *Annu Rev Pharmacol Toxicol* 50:439–465.
29. Hermida RC, Ayala DE (2009) Chronotherapy with the angiotensin-converting enzyme inhibitor ramipril in essential hypertension: Improved blood pressure control with bedtime dosing. *Hypertension* 54:40–46.
30. Zhang R, Podtelezhnikov AA, Hogenesch JB, Anafi RC (2016) Discovering biology in periodic data through phase set enrichment analysis (PSEA). *J Biol Rhythms* 31: 244–257.
31. Mehra R (2014) Understanding nocturnal asthma. The plot thickens. *Am J Respir Crit Care Med* 190:243–244.
32. Lévi F, Okyar A, Dulong S, Innominato PF, Clairambault J (2010) Circadian timing in cancer treatments. *Annu Rev Pharmacol Toxicol* 50:377–421.
33. Warburton D, Shi W, Xu B (2013) TGF-β-Smad3 signaling in emphysema and pulmonary fibrosis: An epigenetic aberration of normal development? *Am J Physiol Lung Cell Mol Physiol* 304:L83–L85.
34. Jeon H-S, Jen J (2010) TGF-beta signaling and the role of inhibitory Smads in non-small cell lung cancer. *J Thorac Oncol* 5:417–419.
35. McMenamin TM (2007) Time to work: Recent trends in shift work and flexible schedules, A. *Monthly Lab Rev* 130:3.
36. Archer SN, et al. (2014) Mistimed sleep disrupts circadian regulation of the human transcriptome. *Proc Natl Acad Sci USA* 111:E682–E691.
37. Lamb JR, et al. (2011) Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. *PLoS One* 6:e20090.
38. Huang W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44–57.
39. Subramanian A, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102: 15545–15550.
40. Ha N-H, Long J, Cai Q, Shu XO, Hunter KW (2016) The circadian rhythm gene Arntl2 is a metastasis susceptibility gene for estrogen receptor-negative breast cancer. *PLoS Genet* 12:e1006267.
41. Brady JJ, et al. (2016) An Arntl2-driven secretome enables lung adenocarcinoma metastatic self-sufficiency. *Cancer Cell* 29:697–710.
42. Chan JA, Kulke M, Clancy TE (2016) Metastatic well-differentiated pancreatic neuroendocrine tumors: Systemic therapy options to control tumor growth and symptoms of hormone hypersecretion. *UpToDate*. Available at www.uptodate.com/index. Accessed August 10, 2016.
43. Hallet J, et al. (2015) Exploring the rising incidence of neuroendocrine tumors: A population-based analysis of epidemiology, metastatic presentation, and outcomes. *Cancer* 121:589–597.
44. Furman BL (2015) Streptozotocin-induced diabetic models in mice and rats. *Curr Protoc Pharmacol* 70:5.47.1–5.47.20.
45. Lamia KA, Storch K-F, Weitz CJ (2008) Physiological significance of a peripheral tissue circadian clock. *Proc Natl Acad Sci USA* 105:15172–15177.
46. Bishop CM (2007) *Pattern Recognition and Machine Learning* (Springer, New York), 20th Ed.
47. Baldi P, Hornik K (1989) Neural networks and principal component analysis: Learning from examples without local minima. *Neural Netw* 2:53–58.
48. Hughes ME, Hogenesch JB, Kornacker K (2010) JTK_CYCLE: An efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J Biol Rhythms* 25:372–380.
49. Liberzon A, et al. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27:1739–1740.