UNIVERSITY OF CALIFORNIA,
IRVINE

Bioinformatic Software Library for Circadian Analysis

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Computer Science

by

Nicholas Ceglia

Thesis Committee:
Professor Pierre Baldi, Chair
Professor Paolo Sassone-Corsi
Professor Marco Levorato

2018

DEDICATION

To mom and dad.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to sincerely thank my advisor, Pierre Baldi, who has shown tremendous support and understanding, I am extremely grateful.

I would like to thank my labmates Yu Liu, Vishal Patel, and Mike Zeller for support and friendship.

I would like to thank collaborators Kristin Eckel-Mahan, Selma Masri, Paola Tognini, Ken Kenichiro, and Paolo Sassone-Corsi.

I would like to thank Camilla Favaretti for pretty much everything else.

I thank the editors and publishers of Nature Methods, Cell, Journal of Biological Chemistry, Bioinformatics, Cell Reports, Nature Communications, and Nucleic Acid Research for publishing parts of this work.

Section 2.1 adapted from Agostinelli et al. (2016).
Section 2.2-2.7 & 3.1 adapted from Ceglia and Liu et al. (2018).

ABSTRACT OF THE THESIS

Bioinformatic Software Library for Circadian Analysis

By

Nicholas Ceglia

Master of Science in Computer Science

University of California, Irvine, 2018

Professor Pierre Baldi, Chair

Circadian oscillations play a fundamental role in many biological processes including cell metabolism and cell cycle. As such, interest in understanding these molecular oscillations has generated an increasingly large collection of circadian omic data. These studies have demonstrated the remarkable plasticity with which the set of oscillating molecular species within a cell are selected. These large shifts in oscillating species are known as circadian reprogramming events. These events have been observed across experimental condition, tissue, and species. While many of these studies have made tissue or condition specific conclusions, a consolidated framework of software tools and a central repository of data has become necessary to answer questions about the orchestration of these reprogramming events.

Chapter 1

Introduction

Circadian rhythms are a ubiquitous biological phenomenon that has been etched into cell biology over trillions of day and night cycles. Many omic studies have found that hundreds of thousands of individual molecular feedback loops operating at a 24 hour frequency are present in almost every cell in almost any tissue. It is well known that the core clock is a robust and prolific regulator of this oscillation. However, the core clock as a sole regulator of these oscillations cannot account for large reprogramming events found in diverse experimental conditions including diet challenge, drug treatment, and disease. The goal of this manuscript is to present a software framework for the analysis of circadian reprogramming. These methods make use of the large database of circadian omic datasets compiled on CircadiOmics and improved methods for identifying oscillating transcripts using BIO_CYCLE.

1.1 Motivation

Circadian oscillations in the concentrations of molecular species play a fundamental role in many biological processes from metabolism, to cell cycle, and to neuronal function [4, 12,

16, 41]. To study the role of these oscillations, an increasing amount of high through-put circadian omic data is being generated under diverse genetic, epigenetic, and environmental conditions. In any single circadian transcriptomic experiment, roughly 10% of measured transcripts are found to oscillate in a circadian manner [14, 13, 38, 37, 43]. However, the intersection of oscillating transcripts between any two experiments is typically small, only about 2% [48]. This small overlap between experiments suggests that the union of all oscillating transcripts across all experiments is large. Remarkably, we calculate that over 95% of all of protein coding transcripts in mouse are found to oscillate in at least one condition [8]. Previous studies have demonstrated specific mechanisms by which a cell can select different oscillating subsets of transcripts, an event known as circadian reprogramming [33, 34, 63, 43]. However, the question of how almost every transcript is capable of oscillating in a circadian manner remains unanswered. The body of this research is devoted to the development of informatic tools and the analysis of specific reprogramming events to generate an understanding of the mechanisms behind these events. Finally, this research aims to identify a model for the transcriptomic organization of circadian rhythms.

To address this problem, it must be noted first that the concentration of any molecular species cannot oscillate in isolation [6]. The fundamental unit of any such oscillation is a feedback loop of molecular interactions, such as transcriptional regulation, post-trascriptional modification, and protein-protein interactions [38, 51, 48], causing all species in the loop to oscillate at the same frequency. A very large number of such regulatory loops have been identified using informatics methods and large omic repositories [48, 10, 62, 58, 24, 53]. The empirically observed pervasiveness of circadian oscillations implies that a significant fraction of these loops is capable of oscillating with a 24 hour period. This 24 hour common period is most likely due to evolution given the importances of the differences between night and day for all biological life, the ~ 2 trillion night-day transitions that have occurred since the origin of life 3.5 billion years ago, and the inherently circadian nature of the molecular circuitry of early photosynthetic life (cyanobacteria) [47]. Thus, in short, modern cells contain entire

networks of circadian coupled oscillators. The question again is how specific subsets of oscillators are selected under specific genetic, epigenetic, and environmental conditions.

A key element of the answer to this question is the circadian core clock. The circadian core clock is genetically implemented by a relatively small set of genes whose transcripts are consistently found to oscillate in most circadian experiments [26, 27, 59]. The core clock regulates an extensive number of transcripts through a set of transcription factors (TF) including CLOCK-BMAL [50]. CLOCK-BMAL binds to E-box motifs that are found abundantly throughout the genome [45, 65]. A possible centralized model of organization is that the core clock directly orchestrates the selection of oscillators in the coupled network. While the importance of the core clock is undeniable [60, 50, 28, 54], additional findings have shown that knocking out elements of the core clock (including CLOCK-BMAL) does not lead to a complete loss of circadian oscillations [29, 35, 2, 64, 11]. Thus, at the other extreme, a completely decentralized model of circadian oscillations is also conceivable where oscillators compete and self organize. Here we seek to find where in this spectrum, from centrally orchestrated to completely decentralized, the cellular network of coupled-oscillators operates.

Chapter 2

Software Framework

2.1 Identification of Oscillation

High-throughput transcriptomic or metabolomic experiments [14, 20, 37, 47], have revealed that typically on the order of 10% of all transcripts or metabolites in the cell are oscillating in a circadian manner. Furthermore, the oscillating transcripts and metabolites differ by cell, tissue type, or condition [47]. Genetic, epigenetic and environmental perturbationssuch as a change in dietcan lead to cellular reprogramming and profoundly influence which species are oscillating in a given cell or tissue [7, 13, 37]. When results are aggregated across tissues and conditions, a very large fraction approaching 100%, of all transcripts is capable of circadian oscillations under at least one set of conditions [48, 8].

In a typical circadian experiment, high-throughput omic measurements are taken at multiple timepoints along the circadian cycle. The first fundamental problem that arises in the analysis of such data is the problem of detecting periodicity, in particular circadian periodicity, in these time series. The problem of detecting periodic patterns in time series is of course not new. However, in the cases considered here the problem is particularly challenging for

several reasons, including:

- The sparsity of the measurements.
- The noise in the measurements and the well known biological variability.
- The related issue of small sample sizes.
- The issue of missing data.
- The issue of uneven sampling in time.
- The large number of measurements and the associated multiple-hypothesis testing problem.

We developed and apply deep learning methods for robustly assessing periodicity in high-throughput circadian experiments, and systematically compare the deep learning approach to the previous, non-machine learning, approaches [20]. While this is useful for circadian experiments, the vast majority of all high-throughput expression experiments have been carried, and continue to be carried, at single timepoints. This can be problematic for many applications, including applications to precision medicine, precisely because circadian variations are ignored creating possible confounding factors. This raises the second problem of developing methods that can robustly infer the approximate time at which a single-time high-throughput expression measurement was taken. Such methods could be used to retrospectively infer a time stamp for any expression dataset, in particular to improve the annotations of all the datasets contained in large gene expression repositories, such as the Gene Expression Omnibus (GEO) [15], and improve the quality of all the downstream inferences that can be made from this wealth of data. There may be other applications of such a method, for instance in forensic sciences, to help infer a time of death. In any case, to address the second problem we also develop and apply deep learning methods to robustly infer time or phase for single-time high-throughput gene expression measurements.

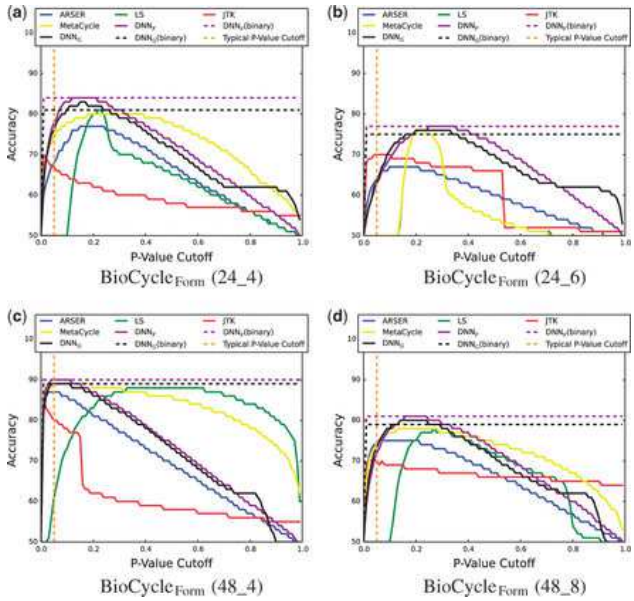

Figure 2.1: Accuracy of periodic/aperiodic classification at different p-value cutoffs on the *BioCycleForm* dataset

To classify signals as periodic or aperiodic, we train deep neural networks (DNNs) using standard gradient descent with momentum [57]. We train separate networks for data sampled over 24 and 48hours. The input to these networks are the expression time-series levels of the corresponding gene (or metabolite). The output is computed by a single logistic unit trained to be 1 when the signal is periodic and 0 otherwise, with relative entropy error function. We experimented with many hyperparameters and learning schedules. In the results reported, the learning rate starts at 0.01, and decays exponentially. The training set consists of 1 million examples, a size sufficient to avoid overfitting. The DNN uses a mini-batch size of 100 and is trained for 50,000 iterations. Use of dropout [3], or other forms of regularization, leads to no tangible improvements. The best performing DNN found has 3 hidden layers of size 100. We are able to obtain very good results by training BIO_CYCLE on synthetic data alone and report test results obtained on biological data shown in Figure 2.1.

In a way similar to how we train DNNs to classify between periodic and aperiodic signals, we can also train DNNs to estimate the period of a signal classified as periodic. During

training, only periodic time series are used as input to train these regression DNNs. The output of the DNNs are implemented using a linear unit and produce an estimated value for the period. The error function is the squared error between the output of the network and the true period of the signal, which is known in advance with synthetic data. Except for the difference in the output unit, we use the same DNNs architectures and hyperparameters as for the previous classification problem.

To calculate p-values, the distribution of the null hypothesis must first be obtained. To do this, N aperiodic signals are generated from one of the two *BioCycleSynth* datasets. Then we calculate the N output values *V(i)* (i=1,,N) of the DNN on these aperiodic signals. The p-value for a new signal s with output value V is now $\frac{1}{N}\sum N_i l(V > V(i))$, where l is the indicator function. This equation provides an empirical frequency estimate for the probability of obtaining an output of size V or greater, assuming that the signal s comes from the null distribution (the distribution of aperiodic signals). Therefore, the smaller the p-value, the more likely it is that s is periodic. The q-values are obtained through the Benjamini and Hochberg procedure. We also compute a posterior probability of periodic expression (PPPE), which models the distribution of p-values as a mixture of beta distributions.

2.2 Web Server

The CircadiOmics web application is constructed as a three-tier Model View Controller architecture. The web server is implemented with the Flask Python library. The interface is generated dynamically with Twitter Bootstrap and Google Charts. Fast query response times are accomplished by caching JSON serialized datasets on disk as the server is started. The interface loads with an example search of ARNTL (CLOCK-BMAL) in a sample liver control dataset. Dynamic filtering of the available datasets is provided based on tissue and experimental perturbations. Examples of filtering options are provided in the documentation

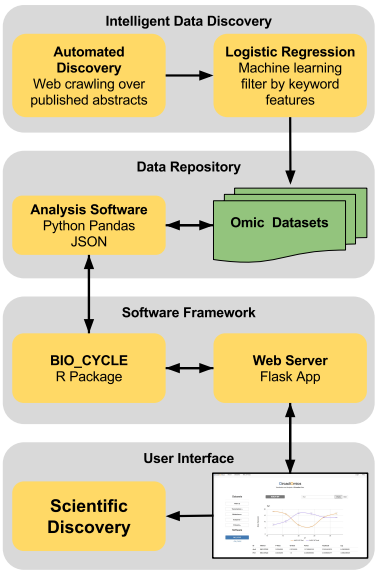

Figure 2.2: Three-tier Model-View-Controller architecture of the CircadiOmics web portal. Intelligent data discovery supplies candidate datasets for inclusion in the repository using a machine learning filter applied to key word features derived from web crawling published abstracts. BIO_CYCLE results are obtained and stored for all datasets. The user interface sends requests and displays results from the web server allowing for interactive hypothesis generation and scientific discovery.

on the main web server in the context of various sample workflows. Downloadable results for each search include high resolution images in PNG or SVG format, and an excel table of BIO_CYCLE reported statistics. Dataset documentation includes a short technical description as well as a link to the corresponding article in PubMed. At last, additional help information on the features of CircadiOmics is provided through a link on the main page of the web server. 2.2 shows a simplified view of the web server architecture.

2.3 Data Repository

The omic datasets available on CircadiOmics are compiled from project collaborations, automated discovery and manual curation. Over 6400 individual time points spanning 227 separate circadian experiments are available for search and visualization. In aggregate, these

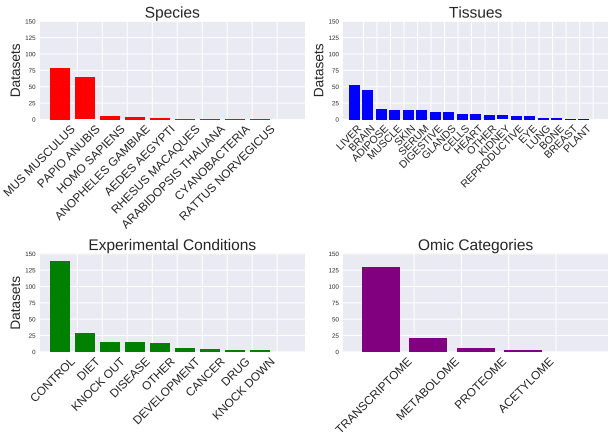

Figure 2.3: Dataset Collection by Species, Tissues, Experimental Conditions, and Omic Categories.

datasets form the largest single repository of circadian data available, including all datasets from other repositories including CircaDB [49]. Eight species are currently available on CircadiOmics. The majority are collected from Mus musculus and Papio anibus.

Over 62 tissues grouped into 18 categories are represented in the database. Within these categories, liver and brain experiments comprise the majority. Diverse experimental conditions grouped into nine broad categories are available for comparison. Unique conditions include chronic and acute ethanol consumption, high-fat diet, traumatic brain injury, fibroblast undergoing myogenic reprogramming and several cancer-specific datasets. At last, CircadiOmics is the only tool that includes transcriptome, metabolome, acetylome and proteome experiments. The full table of datasets is available, with a short description and experimental details such as number of replicates, on the CircadiOmics web portal. 2.3 quantifies the number of datasets by category.

2.4 Data Discovery

Increased interest in circadian rhythms is driving a continuous increase in publicly available omic datasets. Automated discovery of datasets has become necessary to maintain the most current and comprehensive repository. A Python framework built with scholarly and geotools Python packages is used to continuously search the literature for new circadian omic studies and datasets. Automated discovery based on keyword searches in published abstracts is filtered using several features including publishing journal, author and provided supplementary materials. A logistic regression step is used to classify datasets that are good candidates for inclusion in CircadiOmics. Results produced by this automated pipeline are then manually inspected for quality, based primarily on the time point resolution of the dataset. The minimum sampling density for any dataset in the repository is every eight hours over a 24-h cycle. Additionally, the CircadiOmics team and collaborating biologists periodically search recent publications for new datasets that qualify for inclusion in CircadiOmics.

2.5 Visualization

The main functionality of CircadiOmics is the search, comparison and visualization of oscillation trends. The user can search any molecular species in the omic datasets within the repository and overlay multiple searches together to initiate a comparative study. A typical work flow may consist of comparing a set of specific transcripts, metabolites or proteins among several datasets. Intelligent auto-completion facilitates user queries within the currently selected dataset. Searches can be performed individually or in batch on a selected dataset. When datasets do not have the same time course, results are displayed from the minimum to the maximum time point over all selected datasets. Documentation available on the web server illustrates common query tasks and results. Datasets with large difference

in intensity values at each time point can be dynamically scaled for easy visual comparison. Minimum and maximum values are normalized to zero and one, respectively.

A table of statistics is compiled and displayed beneath the main search window after each query. Statistics can be updated dynamically to reflect results obtained with BIO_CYCLE. The table can be downloaded in several formats compatible with Excel. Individual searches can be removed from both the search view and the statistics table. 2.4 highlights an example query and accompanying results.

With a rapidly expanding dataset collection, filtering candidate dataset within the interface has become necessary. The filtering menu allows the user to limit the scope of datasets displayed under drop-down menus for each dataset type. Filtering can be done by species, tissues and experimental conditions. Similar experimental conditions are categorically grouped together in the filtering menu. These include knock-downs, knock-outs, diet changes and drug treatments. The search interface uses an abbreviated dataset identification. Upon selection of a dataset, the user can quickly verify the source of the data through a corresponding literature citation. Additional details for each dataset can be found in tabular form under the dataset tab. These details include a brief description of the experimental protocol.

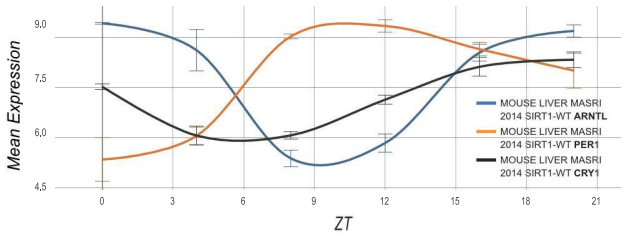

Figure 2.4: Visualization of queries for ARNTL, PER1, and CRY1 in a control mouse dataset. Any number of queries, across any number of datasets, can be displayed simultaneously.

2.6 Metabolomic Atlas

The Metabolic Atlas web portal (http://circadiomics.ics.uci.edu/metabolicatlas) is also available under the CircadiOmics umbrella. In addition to metabolite time series, interactive metabolic networks can be generated and visualized. These networks are derived in part from the KEGG database [24] and can be filtered using BIO_CYCLE statistics.

2.7 Circadian Regulatory Control

We formulated the Circadian Regulatory Control (CRC) method for identification of regulatory edges in circadian feedback loops. Two CRC scores, *B-Score* and *E-score*, incorporate multiple sources of evidence including statistical significant of transcript oscillation and high quality predicted binding sites. These scores further take into account the delay between transcript and protein abundance using available proteomic datasets included in CircadiOmics. CRC scores provide evidence for circadian regulation from a TF or RBP to a specific target. Additionally, an aggregated score provides a measure of the regulatory influence of a TF or RBP by combining the scores of all outgoing edges.

The CRC graph can be seen as representation of the structure of circadian transcriptomic regulation based on the evidence presented in previous results. Here we analyze this representation by combining results from both nodes and regulatory edges. The following results were generated from both individual dataset CRC graphs and the aggregate CRC graph. The CRC *B-score* was used in place of a weighted *E-score* to discretely determine the presence of a regulatory edge.

Regulatory distance was computed as the length of the shortest directed path in the CRC graph between a source TF or RBP and a target transcript. The set of oscillating transcripts

that are found to have a regulatory distance-one from the core clock were considered to be directly regulated by the core clock. The mean percentage of distance-one transcripts across all dataset CRC graphs is roughly 35%. While the majority of transcripts are not found to be directly regulated by the core clock, almost any transcript can be connected through a regulatory path in the CRC graph to the core clock. On average, greater than 80% of oscillating transcripts in a dataset CRC graph can be connected within distance-three from the core clock.

2.8 PyCircadiOmics

PyCircadiOmics is a modular software library written in Python designed to facilitate analysis of circadian high-throughput sequencing experiments through integration of common bioinformatic tools including BIO_CYCLE [1]. The library is composed of three layers described in Figure 2.5. Layers are organized by granularity of analysis including processed data from experiments (e.g. Microarray and RNA-Seq), pair-wise comparison of experimental conditions to a control, and integrated analysis over a large repository of data collected on the webserver CircadiOmics [8].

Dataset class is the base object for all PyCircadiOmics analysis. This class defines objects that work on a processed tab delimited file containing experimental measurements for each timepoint and sample replicate or those datasets that are included in the CircadiOmics repository. Data is stored internally in a Pandas data frame object [40]. The class provides simple methods for instantiating objects using data from either case. Methods exist can be used for annotating trends for each molecular species in the dataset with oscillation statistics. Currently, these methods include JTK_CYCLE [20] and BIO_CYCLE. Additional methods can be easily written to append statistics based on desired third party software. The class also provides utilities for managing identification symbols and basic lookup for trends and

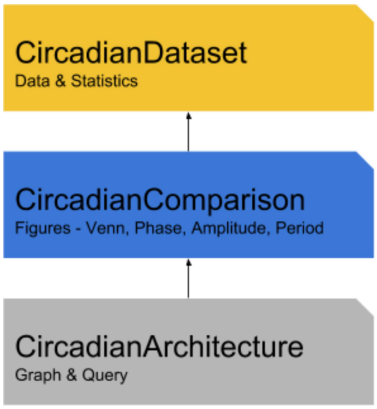

Figure 2.5: PyCircadiOmics Abstract Class Hierarchy

statistics within the class data.

The *CircadianDataset* can be instantiated from a *Dataset* object. This class provides higher level analysis tools related to the statistical results obtained from BIO_CYCLE (or JTK_CYCLE). Statistics for the instantiated object can be summarized into histograms as seen in Figure 3.2. Additionally, it is possible to query transcriptomic datasets by symbol to retrieve CRC graph downstream and upstream results. Downstream results are defined as targets of TF and RBP queries that meet user defined CRC criteria and respective *E-scores*. Finally, molecular species with trends correlated to a user querying can be retrieved using significance scores calculated from *spearmanr*, *pearsonr*, or DTW. While this functionality is available for any dataset, combining this correlation with the KEGG database [24] provides the functionality for network visualization in Metabolomic Atlas.

The *CircadianComparison* class exposes functionality for comparing two datasets. The primary comparison is the set of molecular species (e.g. transcript, metabolite, or protein) found to be uniquely oscillating in each condition and the set that are found to be oscillating in both conditions. These sets can be visualized as venn diagrams as shown in Figure 2.7.

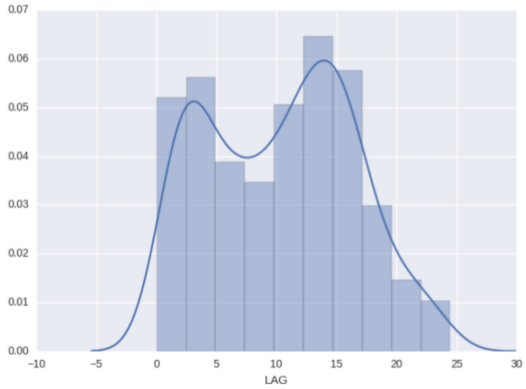

Figure 2.6: Histogram of BIO_CYCLE computed lags.

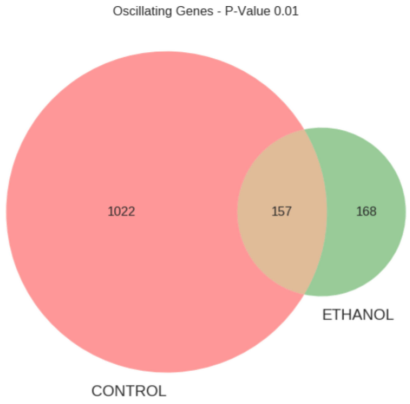

Figure 2.7: Comparison of oscillating genes in a control and chronic ethanol consumption condition within mouse liver.

Additionally, several other tools are provided including the generation of heatmaps and a comparison of enriched GO terms in each of the oscillating sets described above obtained using Goatools Python library [25].

CircadianArchitecture class provides implementations for integrative analysis over all datasets contained within CircadiOmics repository. CircadianDataset objects for each dataset are created and stored in memory along with accompanying CRC graphs. The superimposition of individual CRC graphs can potentially lead to a deeper understanding of the underlying transcriptomic architecture necessary for circadian reprogramming.

Chapter 3

Applications

3.1 Collaborations

Central to the study of circadian rhythms are large-scale reprogramming events. Understanding these events at the molecular level critically depends on being able to access and compare significant amounts of high-throughput circadian omic data. CircadiOmics, with its advanced search features and unprecedented amount of high quality circadian data, is a primary enabling tool for such studies. In a circadian reprogramming event, changes in oscillation of one molecular species can often be related to changes in other molecular species [48, 19]. One of the main qualities of CircadiOmics is the flexibility of the comparative analyses it enables. For instance, a user can compare transcripts across species, or relate metabolites to proteins and transcripts and identify underlying oscillatory trends. An important example can be seen in the loss of oscillation in the metabolite NAD+ as a response to changes in the transcriptomic oscillatory landscape [14]. As a result, CircadiOmics has proven to be highly effective for hypothesis generation in new studies. To date, the web server has contributed to multiple studies that have been published in high impact journals.

The server has been accessed more than 250000 times in total traffic in 2017 alone.

Eckel-Mahan et al. [14] utilized CircadiOmics to analyze three related omic datasets in mouse liver. They found that core clock genes regulate the acetylation of the enzyme AceCS1. AceCS1 is responsible for changes in the oscillation of the metabolite acetyl-CoA, a key metabolite involved in fatty acid synthesis. Similarly, [37] compared liver transcriptomic data with metabolomic data in mice afflicted with cancer using CircadiOmics. They discovered that a distal tumor-bearing lung can reprogram the liver circadian transcriptome through inflammatory pathways and insulin related metabolic pathways [37]. More recently, CircadiOmics has been used to examine the role of circadian regulation in myogenic reprogramming of fibroblast . It was observed that the core clock is completely disrupted during this process. However, exogenous MYOD1 gains rhythmicity during transition to muscle cell. As a result, MYOG and a majority of critical transcription factors related to muscle development known to be regulated by MYOD1 synchronize oscillation. This behavior was identified in CircadiOmics through visualization and confirmed by BIO_CYCLE reported phase lag. At last, aggregating all mouse transcriptomic datasets confirms and amplifies the notion that circadian oscillations are pervasiveness: 93.5% of all possible protein coding transcripts exhibit circadian oscillations in at least one tissue or experiment (up from about 67% in [48]). The large number of datasets in CircadiOmics facilitates these kinds of integrative analyses.

The latest release of CircadiOmics is the largest single repository of circadian omic data available. Updates in server architecture and data mining ensure that CircadiOmics will continue to maintain and grow as new data is published. Improvement in features for search and visualization expand the possibilities for study of circadian rhythms in omic datasets. These possibilities include generating specific hypothesis for individual experiments and answering larger questions about the organization of oscillation within a cell.

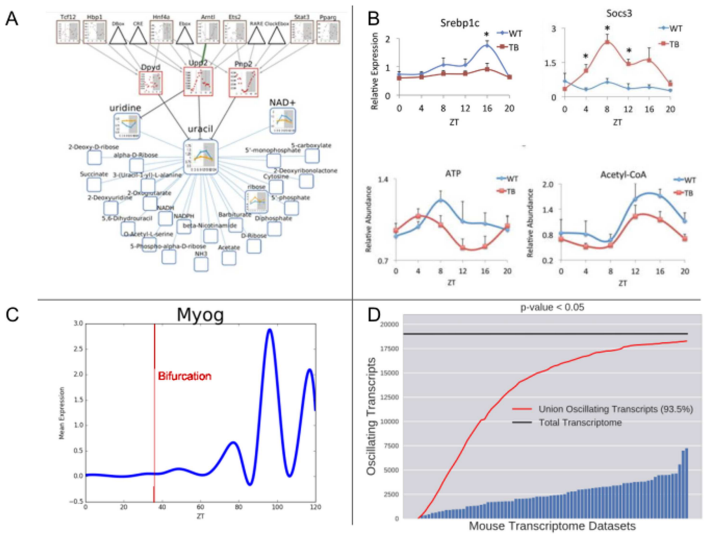

Figure 3.1: Selected Examples of the Impact Of CircadiOmics. **(A)** CircadiOmics was used to link a multitude of circadian metabolites with functionally related circadian transcripts. Figure taken from Figure 5A of [14]. **(B)** CircadiOmics was used to discover reprogrammed circadian transcripts and metabolites related to inflammatory and energy pathways. Figure taken from Figure 2E, 4B and 5D of [37]. **(C)** Exogenous MYOD1, during MEF myogenic reprogramming, entrains oscillation in MYOG and related targets in absence of oscillation of the core clock. **(D)** Bar heights show the ordered number of oscillating protein coding transcripts with a $p \leq 0.05$ in each mouse transcriptomic experiment in the repository. The trend is the cumulative union of oscillating transcripts. Over 93% of possible protein coding transcripts are found to oscillate in at least one tissue or condition across all mouse datasets.

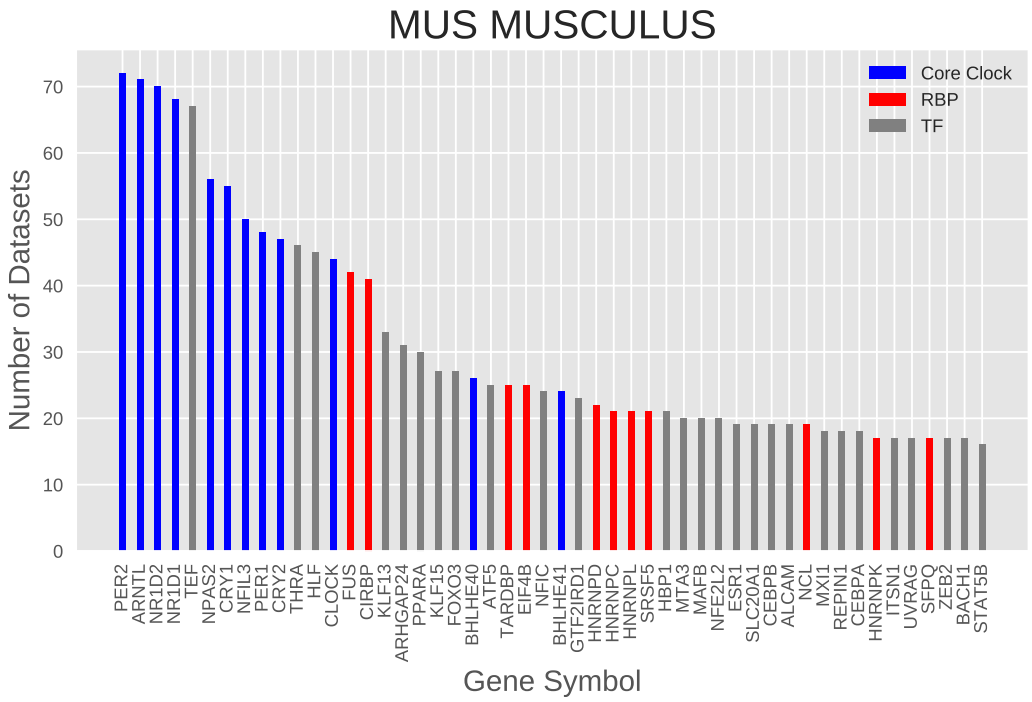

Figure 3.2: Most Frequent Oscillating TFs and RBPs

3.2 Integrated Analysis

The frequency at which a TF or RBP is found to oscillate in a collection of datasets provides a simple metric for estimating its consistency in circadian oscillation. Figure 3.2 illustrates this frequency distribution for mouse at a BIO_CYCLE p-value < 0.01. Additionally, 64 datasets from Papio anubis (baboon) were used for comparison to validate the methods. Both analyses show that TFs involved in the circadian core clock are found to be the most frequently oscillating. This purely data-driven approach automatically discovers the circadian core clock. Furthermore, it identifies additional TFs and RBPs that must play an important role in circadian oscillation.

Transcript frequency is defined as the total number of datasets where a given protein coding transcript is found to be oscillating at a BIO_CYCLE predicted p-value < 0.01. Protein-coding transcripts were identified from BioMart ENSEMBL gene database [55].

Measuring the circadian regulatory influence of the TFs and RBPs identified in the previous analysis requires further investigation using more sophisticated computational methods. To this end, a novel computational method was used to identify and score directed regulatory edges in oscillating loops. The Circadian Regulatory Control (CRC) method can be understood as a proxy for circadian regulation between a TF or RBP (source) and a transcript (target). There are three major components of the CRC method. First, as a prerequisite, the source and target must be oscillating, as assessed by BIO_CYCLE. Second, the source must have at least one high quality binding site on the target for transcriptional or post-transcriptional regulation, as assessed by MotifMap and MotifMap-RNA [62, 10, 36]. For a TF, binding sites were assessed at the promoter region of the target transcript. For an RBP, binding sites were assessed at the introns or UTRs of the target transcript. Third, there must be a correlative relationship between the phases of the source and the target. Recent studies have shown a significant lag between the transcript expression and the concentration of the corresponding protein [52]. We addressed this issue by computing and modeling the distribution of this lag, using transcriptomic and proteomic datasets produced from the same study on CircadiOmics.

After filtering on p-value for the first criteria, the remaining two criteria were combined into two different CRC scores. The *B-score* is a binary indicator of circadian regulation at various filtering thresholds for the number of high quality binding sites and the likelihood of phase correlation. The *E-score* is an exponentially weighted combination of these two criteria. In general, results generated using both scores tend to agree. However, *B-score*, as a binary indicator, is more convenient for large scale analysis of graph structures. In contrast, *E-score*, as a real valued metric, has more sensitivity and is used for ranking nodes and edges. For each source TF or RBP, a CRC score was computed by aggregating all the CRC E-scores from all its outgoing edges either in all experiments or in tissue-specific experiments. The highest scoring TFs and RBPs are shown in Table 3.3.

Mouse All (n=81)		
TF/RBP	All Score	All Ranking
FUS	8.83	9
THRA	8.44	10
BHLHE40	8.30	11
NFIC	8.09	12
HNRPDL	7.56	15
CIRBP	7.42	16
MXI1	7.35	17
EIF4B	7.34	18
CEBPB	6.65	20
HNRNPK	5.61	21
TARDBP	5.40	22
KLF13	4.75	23
PPARA	4.73	24
FOXO3	4.64	25
RAD21	4.56	26
KHDRBS1	4.38	27
ALCAM	4.30	28
MTA3	4.25	29
ARHGAP24	4.23	30
YY1	4.22	31
HNRNPL	4.15	32
NFYA	4.14	33
MAFK	4.12	34
KLF15	4.12	35
PCBP4	4.06	36
HNRNPC	4.06	37
ESR1	4.00	38
SREBF2	3.92	39
BACH1	3.77	40

Mouse Brain (n=13)		
TF/RBP	Brain Score	Brain Rank
CIRBP	1.54	2
SFPQ	1.20	6
KLF15	0.96	8
FUS	0.94	9
ZC3H11A	0.88	10
MXI1	0.86	11
RBM28	0.81	14
EGR1	0.75	15
CHD1	0.71	16
CREB1	0.71	17
HIF1A	0.67	18
HNRPDL	0.66	19
CEBPG	0.65	20
BHLHE40	0.64	21
HNRNPK	0.63	22
SP2	0.63	23
RAD21	0.62	24
NFE2L2	0.60	25
EGR2	0.59	26
GTF2I	0.58	28
KLF12	0.58	29
HCFC1	0.58	30
CHD2	0.57	31
GTF2F1	0.56	32
SRPR	0.54	33
CEBPD	0.53	34
ETV1	0.49	35
GABPA	0.49	36
A1CF	0.49	37

Mouse Skin (n=14)		
TF/RBP	Skin Score	Skin Rank
NFIC	3.44	6
E2F1	2.68	7
MXI1	2.34	10
RUNX1	1.59	13
BRCA1	1.55	14
TCF4	1.46	15
HCFC1	1.41	16
MEF2A	1.32	17
ETV5	1.27	18
THRA	1.17	19
CHD1	1.16	20
FOXM1	1.12	21
NFATC1	1.12	22
FUS	1.12	23
ALCAM	1.11	24
HNRPDL	1.08	25
NFYA	1.03	26
ZFP161	1.00	27
CHD2	0.98	28
RBM5	0.97	29
TCF12	0.95	30
SREBF2	0.94	31
KHDRBS1	0.91	32
EIF4B	0.87	33
KLF13	0.85	34
ELK4	0.85	35
RFX5	0.83	36
SF1	0.81	37
ELAVL1	0.79	38

Mouse Liver (n=31)		
TF/RBP	Liver Score	Liver Rank
CEBPB	4.25	4
BHLHE40	4.22	5
FUS	3.92	7
HNRNPK	3.58	10
EIF4B	3.51	12
PCBP4	3.45	14
THRA	3.30	15
MXI1	2.83	18
YY1	2.80	19
MAFK	2.79	20
ATF5	2.75	21
MTA3	2.75	22
PPARA	2.63	23
BACH1	2.59	24
RXRA	2.58	25
ESR1	2.57	26
RFX4	2.54	27
HNRPDL	2.53	28
FOXO3	2.52	29
CIRBP	2.52	30
STAT5B	2.49	31
CRP	2.43	32
TARDBP	2.37	33
HNRNPC	2.35	34
ARHGAP24	2.33	35
DCTN2	2.28	36
NFIC	2.27	37
KLF1	2.25	38
USF2	2.25	39

Figure 3.3: Tables showing the ranking of circadian TFs and RBPs by CRC E-score in different tissue types. The leftmost table shows ranking in mouse transcriptome across all datasets. RBPs are labeled in red, while TFs are labeled in black. Core clock TFs have been removed from the listing.

When looking at aggregated results, core clock TFs such as CLOCK and BMAL1 were found to have the largest scores, a finding consistent with both the frequency of oscillation and previous literature [60]. Extended members of the core clock were also identified in the ranking including THRA and BHLHE40 [56, 32].

In the results across all datasets, additional TFs and RBPs were identified that seem to have a much broader regulatory role than what is reported in the literature. For instance, FUS and CIRBP have been reported to affect the core circadian factor PER2 via alternative splicing, but only in the mouse liver [30, 42, 46]. In contrast, we find that FUS and CIRBP are found to be high scoring also in both brain and skin. EIF4B has been identified in the circadian regulation of translation in mouse liver [23]. We find that EIF4B is also top scoring in skin. HNRPDL is listed as a potential target of circadian regulation via microRNA in the brain [9]. Strikingly, these RBPs and TFs are found to have very high CRC scores across all mouse datasets. This suggest that they play a broader, previously uncharacterized role

in circadian regulation.

When looking at tissue specific results, many additional TFs and RBPs with high CRC scores are discovered. Although literature evidence has shown that these factors interact with circadian pathways, they are not known to be regulators of oscillation. These TFs may explain tissue specific circadian reprogramming. Within brain tissue, SFPQ is functionally involved in the cell cycle pathway, which also includes NONO and PER2 [17]. EGR1 has been found to oscillate and regulated by the core clock [31]. Within our results, EGR1 potentially regulates a large number of downstream transcripts in the brain. CHD1 is known to be involved in circadian chromatin remodeling in brain [5]. KLF15 is well known to be regulated by the peripheral clock in relation to circadian nitrogen homeostatis in liver and muscle [22]. Within skin tissue, RUNX is a top TF and is known to be regulated in a circadian fashion in epidermal cells [21]. E2F1 is regulated by circadian factors SIRT1 and CLOCK[44]. BRCA1 is known to interact with core clock TFs such as PER2 [61]. Within liver tissue, CEBPB is top ranking excluding core clock TFs. This agrees with the literature finding that it interacts with the core clock through REV-ERB [18]. PCBP4 is known to be involved in circadian alternative splicing in the liver [39].

Additionally, there are many other novel findings that have been linked to very few circadian studies. These findings include: NFIC, RAD21, MXI1, and TARDBP across all tissues; ZC3H11A, RBM28, and CEBPG in brain; HCFC1 and ETV5 in skin; and HNRNPK, ATF5, and BACH1/MAFK in liver, and may provide leads for investigations of previously unknown circadian regulatory mechanisms.

3.3 Summary

Further analysis is required to understand the organization of transcriptomic oscillation. The software framework presented in this paper provides a foundation for such inquiry. The latest release of CircadiOmics is the largest single repository of circadian omic data available. Updates in server architecture and data mining ensure that CircadiOmics will continue to maintain and grow as new data is published. Improvement in features for search and visualization expand the possibilities for study of circadian rhythms in omic datasets. These possibilities include generating specific hypothesis for individual experiments and answering larger questions about the organization of oscillation within a cell. PyCircadiOmics is currently being used in ongoing collaborations within the Institute for Genomics and Bioinformatics. Finally, the Circadian Regulatory Control (CRC) provides a foundation for deeper study of the transcriptomci organization of circadian reprogramming events.

Bibliography

[1] F. Agostinelli, N. Ceglia, B. Shahbaba, P. Sassone-Corsi, and P. Baldi. What time is it? Deep learning approaches for circadian rhythms. *Bioinformatics*, 32(12):i8–i17, 2016. [PubMed:27307647] [PubMed Central:PMC4908327] [doi:10.1093/bioinformatics/btw243].

[2] J. Alvarez, A. Hansen, T. Ord, P. Bebas, P. E. Chappell, J. M. Giebultowicz, C. Williams, S. Moss, and A. Sehgal. The circadian clock protein bmal1 is necessary for fertility and proper testosterone production in mice. *Journal of biological rhythms*, 23(1):26–36, 2008.

[3] P. Baldi and P. J. Sadowski. Understanding dropout. In *Advances in neural information processing systems*, pages 2814–2822, 2013.

[4] J. Bass. Circadian topology of metabolism. *Nature*, 491(7424):348–356, 2012. [PubMed:23151577] [doi:10.1038/nature11704].

[5] W. J. Belden, Z. A. Lewis, E. U. Selker, J. J. Loros, and J. C. Dunlap. Chd1 remodels chromatin and influences transient dna methylation at the clock gene frequency. *PLoS genetics*, 7(7):e1002166, 2011.

[6] D. Bell-Pedersen, V. M. Cassone, D. J. Earnest, S. S. Golden, P. E. Hardin, T. L. Thomas, and M. J. Zoran. Circadian rhythms from multiple oscillators: lessons from diverse organisms. *Nature Reviews Genetics*, 6(7):544, 2005.

[7] M. M. Bellet, E. Deriu, J. Z. Liu, B. Grimaldi, C. Blaschitz, M. Zeller, R. A. Edwards, S. Sahar, S. Dandekar, P. Baldi, et al. Circadian clock regulates the host response to salmonella. *Proceedings of the National Academy of Sciences*, 110(24):9897–9902, 2013.

[8] N. Ceglia, Y. Liu, S. Chen, F. Agostinelli, K. Eckel-Mahan, P. Sassone-Corsi, and P. F. Baldi. Circadiomics: Circadian omic web portal. *Nucleic Acid Research*, 2018.

[9] H.-Y. M. Cheng, J. W. Papp, O. Varlamova, H. Dziema, B. Russell, J. P. Curfman, T. Nakazawa, K. Shimizu, H. Okamura, S. Impey, et al. microrna modulation of circadian-clock period and entrainment. *Neuron*, 54(5):813–829, 2007.

[10] K. Daily, V. R. Patel, P. Rigor, X. Xie, and P. Baldi. MotifMap: integrative genome-wide maps of regulatory motif sites for model species, 2011.

[11] J. P. DeBruyne, D. R. Weaver, and S. M. Reppert. Clock and npas2 have overlapping roles in the suprachiasmatic circadian clock. *Nature neuroscience*, 10(5):543, 2007.

[12] C. Dibner, U. Schibler, and U. Albrecht. *The Mammalian Circadian Timing System: Organization and Coordination of Central and Peripheral Clocks*, volume 72. 2010. [PubMed:20148687] [doi:10.1146/annurev-physiol-021909-135821].

[13] K. L. Eckel-Mahan, V. R. Patel, S. De Mateo, R. Orozco-Solis, N. J. Ceglia, S. Sahar, S. A. Dilag-Penilla, K. A. Dyar, P. Baldi, and P. Sassone-Corsi. Reprogramming of the circadian clock by nutritional challenge. *Cell*, 155(7):1464–1478, 2013. [PubMed:24360271] [PubMed Central:24360271] [doi:10.1016/j.cell.2013.11.034].

[14] K. L. Eckel-Mahan, V. R. Patel, R. P. Mohney, K. S. Vignola, P. Baldi, and P. Sassone-Corsi. Coordination of the transcriptome and metabolome by the circadian clock. *Proceedings of the National Academy of Sciences*, 109(14):5541–5546, 2012. [PubMed:22431615] [PubMed Central:PMC3325727] [doi:10.1073/pnas.1118726109].

[15] R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[16] J. R. Gerstner, L. C. Lyons, K. P. Wright, D. H. Loh, O. Rawashdeh, K. L. Eckel-Mahan, and G. W. Roman. Cycling Behavior and Memory Formation. *Journal of Neuroscience*, 29(41):12824–12830, oct 2009. [PubMed:19828795] [PubMed Central:PMC4077269] [doi:10.1523/JNEUROSCI.3353-09.2009].

[17] F. Guillaumond, B. Boyer, D. Becquet, S. Guillen, L. Kuhn, J. Garin, M. Belghazi, O. Bosler, J.-L. Franc, and A.-M. François-Bellan. Chromatin remodeling as a mechanism for circadian prolactin transcription: rhythmic nono and sfpq recruitment to hltf. *The FASEB Journal*, 25(8):2740–2756, 2011.

[18] G. Huang, F. Zhang, Q. Ye, and H. Wang. The circadian clock regulates autophagy directly through the nuclear hormone receptor nr1d1/rev-erbα and indirectly via cebpb/(c/ebpβ) in zebrafish. *Autophagy*, 12(8):1292–1309, 2016.

[19] M. E. Hughes, K. C. Abruzzi, R. Allada, R. Anafi, A. B. Arpat, G. Asher, P. Baldi, C. De Bekker, D. Bell-Pedersen, J. Blau, et al. Guidelines for genome-scale analysis of biological rhythms. *Journal of biological rhythms*, 32(5):380–393, 2017. [PubMed:29098954],[PubMed Central:PMC5692188][doi:10.1177/0748730417728663].

[20] M. E. Hughes, J. B. Hogenesch, and K. Kornacker. JTK_CYCLE: An Efficient Nonparametric Algorithm for Detecting Rhythmic Components in Genome-Scale Data Sets. *Journal of Biological Rhythms*, 25(5):372–380, oct 2010. [PubMed:19772347] [PubMed Central:PMC2761978] [doi:10.1177/0748730410379711].

[21] P. Janich, K. Toufighi, G. Solanas, N. M. Luis, S. Minkwitz, L. Serrano, B. Lehner, and S. A. Benitah. Human epidermal stem cell function is regulated by circadian oscillations. *Cell stem cell*, 13(6):745–753, 2013.

[22] D. Jeyaraj, F. A. Scheer, J. A. Ripperger, S. M. Haldar, Y. Lu, D. A. Prosdocimo, S. J. Eapen, B. L. Eapen, Y. Cui, G. H. Mahabeleshwar, et al. Klf15 orchestrates circadian nitrogen homeostasis. *Cell metabolism*, 15(3):311–323, 2012.

[23] C. Jouffe, G. Cretenet, L. Symul, E. Martin, F. Atger, F. Naef, and F. Gachon. The circadian clock coordinates ribosome biogenesis. *PLoS biology*, 11(1):e1001455, 2013.

[24] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[25] D. Klopfenstein, L. Zhang, B. S. Pedersen, F. Ramírez, A. W. Vesztrocy, A. Naldi, C. J. Mungall, J. M. Yunes, O. Botvinnik, M. Weigel, et al. Goatools: A python library for gene ontology analyses. *Scientific reports*, 8(1):10872, 2018.

[26] C. H. Ko and J. S. Takahashi. Molecular components of the mammalian circadian clock. *Human Molecular Genetics*, 15(SUPPL. 2):271–277, 2006. [PubMed:16987893] [doi:10.1093/hmg/ddl207].

[27] N. Koike, S.-H. Yoo, H.-C. Huang, V. Kumar, C. Lee, T.-K. Kim, and J. S. Takahashi. Transcriptional Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals. *Science*, 338(6105):349–354, oct 2012. [PubMed:22936566] [PubMed Central:PMC3694775] [doi:10.1126/science.1226339].

[28] R. V. Kondratov, A. A. Kondratova, V. Y. Gorbacheva, O. V. Vykhovanets, and M. P. Antoch. Early aging and age-related pathologies in mice deficient in bmal1, the core componentof the circadian clock. *Genes & development*, 20(14):1868–1873, 2006.

[29] R. V. Kondratov, A. A. Kondratova, V. Y. Gorbacheva, O. V. Vykhovanets, and M. P. Antoch. Early aging and age-related pathologies in mice deficient in bmal1, the core componentof the circadian clock. *Genes & development*, 20(14):1868–1873, 2006.

[30] B. Kornmann, O. Schaad, H. Bujard, J. S. Takahashi, and U. Schibler. System-driven and oscillator-dependent circadian transcription in mice with a conditionally active liver clock. *PLoS biology*, 5(2):e34, 2007.

[31] J. H. Lee and A. Sancar. Circadian clock disruption improves the efficacy of chemotherapy through p73-mediated apoptosis. *Proceedings of the National Academy of Sciences*, 108(26):10668–10672, 2011.

[32] J. Z. Li, B. G. Bunney, F. Meng, M. H. Hagenauer, D. M. Walsh, M. P. Vawter, S. J. Evans, P. V. Choudary, P. Cartagena, J. D. Barchas, et al. Circadian patterns of gene expression in the human brain and disruption in major depressive disorder. *Proceedings of the National Academy of Sciences*, 110(24):9950–9955, 2013.

[33] X.-M. Li, F. Delaunay, S. Dulong, B. Claustrat, S. Zampera, Y. Fujii, M. Teboul, J. Beau, and F. Lévi. Cancer inhibition through circadian reprogramming of tumor transcriptome with meal timing. *Cancer research*, 70(8):3351–3360, 2010.

[34] X. M. Li, F. Delaunay, S. Dulong, B. Claustrat, S. Zampera, Y. Fujii, M. Teboul, J. Beau, and F. Lévi. Cancer inhibition through circadian reprogramming of tumor transcriptome with meal timing. *Cancer Research*, 70(8):3351–3360, 2010. [PubMed:20395208] [doi:10.1158/0008-5472.CAN-09-4235].

[35] A. C. Liu, D. K. Welsh, C. H. Ko, H. G. Tran, E. E. Zhang, A. A. Priest, E. D. Buhr, O. Singer, K. Meeker, I. M. Verma, et al. Intercellular coupling confers robustness against mutations in the scn circadian clock network. *Cell*, 129(3):605–616, 2007.

[36] Y. Liu, S. Sun, T. Bredy, M. Wood, R. C. Spitale, and P. Baldi. MotifMap-RNA: a genome-wide map of RBP binding sites. *Bioinformatics*, 33(13):2029–2031, jul 2017.

[37] S. Masri, T. Papagiannakopoulos, K. Kinouchi, Y. Liu, M. Cervantes, P. Baldi, T. Jacks, and P. Sassone-Corsi. Lung Adenocarcinoma Distally Rewires Hepatic Circadian Homeostasis. *Cell*, 165(4):896–909, 2016. [PubMed:27153497] [PubMed Central:PMC5373476] [doi:10.1016/j.cell.2016.04.039].

[38] S. Masri, V. R. Patel, K. L. Eckel-Mahan, S. Peleg, I. Forne, A. G. Ladurner, P. Baldi, A. Imhof, and P. Sassone-Corsi. Circadian acetylome reveals regulation of mitochondrial metabolic pathways. *Proceedings of the National Academy of Sciences*, 110(9):3339–3344, 2013. [PubMed:23341599] [PubMed Central:PMC4908327] [doi:10.1073/pnas.1217632110].

[39] N. J. McGlincy, A. Valomon, J. E. Chesham, E. S. Maywood, M. H. Hastings, and J. Ule. Regulation of alternative splicing by the circadian clock and food related cues. *Genome biology*, 13(6):R54, 2012.

[40] W. McKinney. pandas: a foundational python library for data analysis and statistics.

[41] J. S. Menet and M. Rosbash. When brain clocks lose track of time: cause or consequence of neuropsychiatric disorders. *Current Opinion in Neurobiology*, 21(6):849–857, dec 2011. [PubMed:21737252] [PubMed Central:PMC3252427] [doi:10.1016/j.conb.2011.06.008].

[42] J. Morf, G. Rey, K. Schneider, M. Stratmann, J. Fujita, F. Naef, and U. Schibler. Cold-inducible rna-binding protein modulates circadian gene expression posttranscriptionally. *Science*, 338(6105):379–383, 2012.

[43] M. Murakami, P. Tognini, Y. Liu, K. L. EckelMahan, P. Baldi, and P. SassoneCorsi. Gut microbiota directs PPARγdriven reprogramming of the liver circadian clock by nutritional challenge. *EMBO reports*, 17(9):1292–1303, 2016. [PubMed:27418314] [PubMed Central:PMC5007574] [doi:10.15252/embr.201642463].

[44] Y. Nakahata, M. Kaluzova, B. Grimaldi, S. Sahar, J. Hirayama, D. Chen, L. P. Guarente, and P. Sassone-Corsi. The nad+-dependent deacetylase sirt1 modulates clock-mediated chromatin remodeling and circadian control. *Cell*, 134(2):329–340, 2008.

[45] K. Oishi, K. Miyazaki, K. Kadota, R. Kikuno, T. Nagase, G.-i. Atsumi, N. Ohkura, T. Azama, M. Mesaki, S. Yukimasa, et al. Genome-wide expression analysis of mouse liver reveals clock-regulated circadian output genes. *Journal of Biological Chemistry*, 278(42):41519–41527, 2003.

[46] H. Oster, S. Damerow, R. A. Hut, and G. Eichele. Transcriptional profiling in the adrenal gland reveals circadian regulation of hormone biosynthesis genes and nucleosome assembly genes. *Journal of biological rhythms*, 21(5):350–361, 2006.

[47] S. Panda, J. B. Hogenesch, and S. A. Kay. Circadian rhythms from flies to human. *Nature*, 417(6886):329–335, 2002. [PubMed:12015613] [doi:10.1038/417329a].

[48] V. R. Patel, N. Ceglia, M. Zeller, K. Eckel-Mahan, P. Sassone-Corsi, and P. Baldi. The pervasiveness and plasticity of circadian oscillations: The coupled circadian-oscillators framework. *Bioinformatics*, 31(19):3181–3188, 2015. [PubMed:26049162] [PubMed Central:PMC4592335] [doi:10.1093/bioinformatics/btv353].

[49] A. Pizarro, K. Hayer, N. F. Lahens, and J. B. Hogenesch. CircaDB: A database of mammalian circadian gene expression profiles. *Nucleic Acids Research*, 41(D1):1009–1013, 2013. [PubMed:23180795] [PubMed Central:PMC3531170] [doi:10.1093/nar/gks1161].

[50] J. A. Ripperger and U. Schibler. Rhythmic clock-bmal1 binding to multiple e-box motifs drives circadian dbp transcription and chromatin transitions. *Nature genetics*, 38(3):369, 2006.

[51] M. S. Robles, J. Cox, and M. Mann. In-Vivo Quantitative Proteomics Reveals a Key Contribution of Post-Transcriptional Mechanisms to the Circadian Regulation of Liver Metabolism. *PLoS Genetics*, 10(1), 2014. [PubMed:24391516] [PubMed Central:PMC2840478] [doi:10.1371/journal.pgen.1004047].

[52] M. S. Robles, J. Cox, and M. Mann. In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLoS genetics*, 10(1):e1004047, 2014.

[53] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173, 2005.

[54] R. D. Rudic, P. McNamara, A.-M. Curtis, R. C. Boston, S. Panda, J. B. Hogenesch, and G. A. FitzGerald. Bmal1 and clock, two essential components of the circadian clock, are involved in glucose homeostasis. *PLoS biology*, 2(11):e377, 2004.

[55] D. Smedley, S. Haider, S. Durinck, L. Pandini, P. Provero, J. Allen, O. Arnaiz, M. H. Awedh, R. Baldock, G. Barbiera, P. Bardou, T. Beck, A. Blake, M. Bonierbale, A. J. Brookes, G. Bucci, I. Buetti, S. Burge, C. Cabau, J. W. Carlson, C. Chelala, C. Chrysostomou, D. Cittaro, O. Collin, R. Cordova, R. J. Cutts, E. Dassi, A. D. Genova, A. Djari,

A. Esposito, H. Estrella, E. Eyras, J. Fernandez-Banet, S. Forbes, R. C. Free, T. Fujisawa, E. Gadaleta, J. M. Garcia-Manteiga, D. Goodstein, K. Gray, J. A. Guerra-Assuno, B. Haggarty, D.-J. Han, B. W. Han, T. Harris, J. Harshbarger, R. K. Hastings, R. D. Hayes, C. Hoede, S. Hu, Z.-L. Hu, L. Hutchins, Z. Kan, H. Kawaji, A. Keliet, A. Kerhornou, S. Kim, R. Kinsella, C. Klopp, L. Kong, D. Lawson, D. Lazarevic, J.-H. Lee, T. Letellier, C.-Y. Li, P. Lio, C.-J. Liu, J. Luo, A. Maass, J. Mariette, T. Maurel, S. Merella, A. M. Mohamed, F. Moreews, I. Nabihoudine, N. Ndegwa, C. Noirot, C. Perez-Llamas, M. Primig, A. Quattrone, H. Quesneville, D. Rambaldi, J. Reecy, M. Riba, S. Rosanoff, A. A. Saddiq, E. Salas, O. Sallou, R. Shepherd, R. Simon, L. Sperling, W. Spooner, D. M. Staines, D. Steinbach, K. Stone, E. Stupka, J. W. Teague, A. Z. DayemUllah, J. Wang, D. Ware, M. Wong-Erasmus, K. Youens-Clark, A. Zadissa, S.-J. Zhang, and A. Kasprzyk. The biomart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, 43(W1):W589–W598, 2015.

[56] K.-F. Storch, O. Lipan, I. Leykin, N. Viswanathan, F. C. Davis, W. H. Wong, and C. J. Weitz. Extensive and divergent circadian gene expression in liver and heart. *Nature*, 417(6884):78, 2002.

[57] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[58] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452, 2014.

[59] J. S. Takahashi, H.-K. Hong, C. H. Ko, and E. L. McDearmon. The genetics of mammalian circadian order and disorder: implications for physiology and disease. *Nature Reviews Genetics*, 9(10):764, 2008.

[60] H. R. Ueda, S. Hayashi, W. Chen, M. Sano, M. Machida, Y. Shigeyoshi, M. Iino, and S. Hashimoto. System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nature genetics*, 37(2):187, 2005.

[61] S. L. Winter, L. Bosnoyan-Collins, D. Pinnaduwage, and I. L. Andrulis. Expression of the circadian clock genes pert, per2 in sporadic, familial breast tumors. *Neoplasia*, 9(10):797–800, 2007.

[62] X. Xie, P. Rigor, and P. Baldi. MotifMap: A human genome-wide map of candidate regulatory motif sites. *Bioinformatics*, 25(2):167–174, 2009.

[63] K. Yagita, K. Horie, S. Koinuma, W. Nakamura, I. Yamanaka, A. Urasaki, Y. Shigeyoshi, K. Kawakami, S. Shimada, J. Takeda, and Y. Uchiyama. Development of the circadian oscillator during differentiation of mouse embryonic stem cells in vitro. *Proceedings of the National Academy of Sciences*, 107(8):3846–3851, 2010. [PubMed:20133594] [PubMed Central:PMC2840478][doi:10.1073/pnas.0913256107].

[64] S. Yamaguchi, S. Mitsui, L. Yan, K. Yagita, S. Miyake, and H. Okamura. Role of dbp in the circadian oscillatory mechanism. *Molecular and cellular biology*, 20(13):4773–4781, 2000.

[65] E. E. Zhang, A. C. Liu, T. Hirota, L. J. Miraglia, G. Welch, P. Y. Pongsawakul, X. Liu, A. Atwood, J. W. Huss III, J. Janes, et al. A genome-wide rnai screen for modifiers of the circadian clock in human cells. *Cell*, 139(1):199–210, 2009.