

RESEARCH

Open Access



Dual-approach co-expression analysis framework (D-CAF) enables identification of novel circadian co-regulation from multi-omic timeseries data

Joshua Chuah^{1,2*}, Carmalena V. Cordi³, Juergen Hahn^{2,4} and Jennifer M. Hurley^{2,3*}

*Correspondence:
chuahj@union.edu; hurlej2@rpi.edu

¹ Department of Electrical, Computer, and Biomedical Engineering, Union College, 807 Union St, Schenectady, NY 12308, USA

² Department of Biomedical Engineering, Rensselaer Polytechnic Institute, 110 8th St, Troy, NY 12180, USA

³ Department of Biological Sciences, Rensselaer Polytechnic Institute, 110 8th St, Troy, NY 12180, USA

⁴ Department of Chemical and Biological Engineering, Rensselaer Polytechnic Institute, 110 8th St, Troy, NY 12180, USA

Abstract

Background: The circadian clock is a central driver of many biological and behavioral processes, regulating the levels of many genes and proteins, termed clock controlled genes and proteins (CCGs/CCPs), to impart biological timing at the molecular level. While transcriptomic and proteomic data has been analyzed to find potential CCGs and CCPs, multi-omic modeling of circadian data, which has the potential to enhance the understanding of circadian control of biological timing, remains relatively rare due to several methodological hurdles. To address this gap, a dual-approach co-expression analysis framework (D-CAF) was created to perform co-expression analysis that is robust to Gaussian noise perturbations on time-series measurements of both transcripts and proteins.

Results: Applying this D-CAF framework to previously gathered transcriptomic and proteomic data from mouse macrophages gathered over circadian time, we identified small, highly significant clusters of oscillating transcripts and proteins in the unweighted similarity matrices and larger, less significant clusters of oscillating transcripts and proteins using the weighted similarity network. Functional enrichment analysis of these clusters identified novel immunological response pathways that appear to be under circadian control.

Conclusions: Overall, our findings suggest that D-CAF is a tool that can be used by the circadian community to integrate multi-omic circadian data to improve our understanding of the mechanisms of circadian regulation of molecular processes.

Keywords: Co-expression, Multi-omics, Transcriptomics, Proteomics, Robustness, Clustering, Co-regulation, Circadian, Network analysis

Introduction

Circadian rhythms are molecular rhythms that cycle once approximately every 24 h in order to time organismal physiology to the day/night cycle of earth [1]. These rhythms occur broadly in organisms that live in the Photic Zone [2] and play a crucial role in regulating physiological and behavioral processes, such as the sleep-wake cycle, autophagy,



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

immune system function, hormone production, and metabolism [3–7]. Given their wide-ranging regulation of human physiology, disruption of circadian rhythms has the ability to enhance the rates of human diseases, such as Alzheimer’s disease, diabetes, cancer, and cardiovascular disease [8–11].

Because of the significant effect circadian rhythms have over physiology, much research has been devoted to understanding the drivers of these rhythms. The molecular mechanism that generates circadian rhythms is referred to as the molecular circadian clock and is comprised of a transcription-translation negative feedback loop that coordinates the rhythmic transcription/expression of a host of genes and proteins (called clock controlled genes/proteins, CCGs/CCPs) [12, 13]. It is these CCGs/CCPs that are responsible for driving the 24-h biological phenotypes that are collectively referred to as circadian physiology. With CCG/CCPs numbering in the hundreds to thousands depending on cell type, omics level data is often collected to characterize the effect the molecular circadian clock has on the transcriptome and proteome in an attempt to identify the mechanisms that link circadian disruption to disease [14–17].

Although many investigations into circadian rhythms use single omics approaches, like transcriptomics or proteomics, there is a growing recognition of the importance of modeling cellular functions using multiple omics, or multi-omics, datasets [18, 19]. Multi-omic approaches are crucial in understanding the regulation of circadian physiology as oscillations in one data set (e.g. transcription factors in the proteome) can affect oscillations in another data set (e.g. rhythmic levels of expression in the transcriptome), which multi-omics can help to explore. One analysis technique that has been used independently in both multi-omics analysis and circadian biology is co-expression analysis, which finds clusters of analytes with similar expression patterns to define regulatory relationships within and between pathways [20, 21]. Co-expression network analysis represents large scale data as a collection of nodes and interactions between nodes, which provides researchers with the ability to study the structure of complex systems using omics level data, enabling the clustering and exploration of modules of genes, transcripts, or proteins [22, 23].

While co-expression may be highly useful in defining the network of the circadian clock, to the best of our knowledge, co-expression analyses that integrate multiple omics datasets of circadian data simultaneously have only been used in a limited fashion [22]. This lack of multi-omic co-expression analyses of circadian data is likely due to a plethora of hurdles to multi-omic analysis, many of which involve modeling of the relationships between multi-omics data from time series data. In fact, while there are several techniques that exist for the integration of multiple omics datasets, few have been proven to be useful for longitudinal data [24]. Those that can be applied to longitudinal data either involve data transformation or necessitate modeling step pre-integration [25–27]. These limit the ability to perform co-expression analysis, as transformation of the original expression patterns leads to a loss of direct interpretability, making it challenging to assess the similarities of the original data over time.

One way to mitigate these hurdles is to implement a strategy known as early integration. Early integration is the simplest multi-omic data integration method and involves concatenating multiple omics datasets together. However, this method is not frequently used as raw transcript and protein measurements can vary by orders of magnitude

[28]. Our suite of software packages, the Pipeline for Amplitude Integration of Circadian Exploration (PAICE) suite, including the Extended Circadian Harmonic Oscillator (ECHO) program, allows us to overcome this as it fits relative expression curves to longitudinal biological data that are comparable between data sets [29]. Given that ECHO allows for early integration, we hypothesized that we would be able to perform true multi-omics co-expression analysis on circadian multi-omics data.

Therefore, the aim of this study was to (1) develop a framework to perform co-expression analysis of circadian multi-omic data and (2) use this framework to identify circadian rhythms and rhythm-regulated functions in mouse macrophages. This framework would, in turn, enhance understanding of the impact of the circadian clock on biological processes by identifying proteins and mRNAs that oscillate with similar expression patterns, and therefore have similar points of co-regulation beyond their shared circadian regulation. To test and develop this framework, circadian transcriptomic and proteomic data from previously collected mouse macrophages were first preprocessed to enable the comparison of transcript and protein time-series using our ECHO program. Subsequently, multiple network analysis models were applied to (1) concurrently compute rhythms with the most similarities to known circadian rhythms and (2) create clusters that are then utilized to construct protein-protein interaction networks. The clusters we identified using this framework showed that circadian rhythms drive aspects of the immune system in mouse macrophages. We found our co-expression framework to be useful in identifying densely connected, communities robust to Gaussian noise perturbations, and larger, biologically relevant clusters. Moreover, we have designed a framework for others to apply this approach to their own longitudinal multi-omic datasets. In total, our dual-approach co-expression analysis framework (D-CAF) builds a foundation that can be used broadly to analyze circadian multi-omic data.

Materials and methods

A description of the application of the Dual-approach co-expression analysis framework (D-CAF) to a multi-omic (transcriptomic and proteomic) circadian time-series dataset is given below. A flowchart representing the overall steps in the framework can be found in Fig. 1. In short, the framework can be broken down into a data preprocessing step, where early integration of the transcriptomic and proteomic datasets are performed, a modeling step, where models are generated to group rhythms with similar expression patterns over time, and a validation step, where the models are validated using various methods.

Data collection

Data from [30] were used as source data as model data for our framework development. In this previously published data set, macrophages were collected from the bone marrow of *Per2::Luc* or *Per1/Per2* knockout mice aged 3–6 months old. Macrophages were collected every 2 h over a 48 h window across three (independent) biological replicates. Transcriptomic (RNA-sequencing data) and proteomic measurements were then taken and computed from these macrophages. Protein concentrations were measured via multiplex tandem mass tag mass spectrometry. Further technical details on this data set and data collection can be found in the source paper for the data [30].

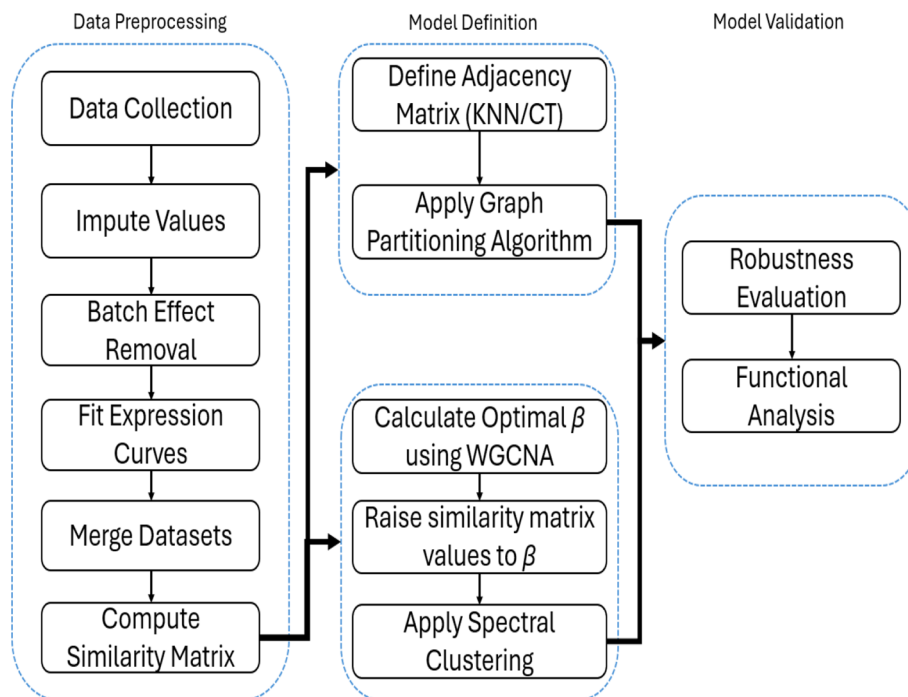


Fig. 1 A flowchart for the general D-CAF workflow. Data was first preprocessed to integrate the individual omics datasets, after which both weighted and unweighted models were generated, and finally validation was run on the optimal models

Data preprocessing

Early integration of the proteomic and transcriptomic datasets were performed to merge both datasets into one multi-omic dataset. For early integration, both data sets had to be transformed and normalized such that there was minimal heterogeneity between data sets. Since raw expression values of transcripts and proteins can vary by factors of hundreds to thousands, our approach consisted of employing the LIMBR and ECHO software packages to create comparable expression curves. LIMBR first performed missing value imputation and batch effect removal of each individual replicate for both the proteomic and the transcriptomic datasets individually. After LIMBR, the data was processed through the ECHO software package in order to fit a smoothed relative expression curve for all three biological replicates [29, 31]. It should be noted that while a data from a 48 h time series is used in this study, ECHO only requires that each time series is larger than 24 h, and is not so sparse that LIMBR removes the majority of data. More details about the requirements of LIMBR and ECHO can be found in their source papers [29, 31]. In summary, D-CAF utilizes and builds on the pre-processing steps LIMBR and ECHO to identify groups of rhythms that are highly similar, which in turn suggest a shared mechanism or function.

Missing value imputation and batch normalization

To impute missing data and batch normalize the individual omics datasets, values in the transcriptomic (transcripts per million) and proteomic (relative peptide

concentration) datasets were log2 normalized to reduce the skew and range of values at each time point while still remaining representative of the proportional differences at each time point. Next, unique peptides or reads from a given peptide or transcript respectively were summed into one measurement, so each row in the dataset was a unique peptide or transcript at a given time. Imputation was then performed using the K-nearest neighbors algorithm with $K = 10$, with only rows with less than 30% values missing imputed to limit over-imputation, as these are the parameters recommended by LIMBR. Next, batch effect removal was performed to center the measurements around the same distribution to eliminate noise using surrogate variable analysis. While efforts can be made to mitigate batch effects, utilizing LIMBR for data-wise batch effect removal is still useful to transform the data to account for any unforeseen batch effects. Once the data was batch effect normalized, measurements from different peptides belonging to the same protein were averaged together. Further information on this process can be found in the source paper [31].

Expression curve fitting

To create a smoothed relative expression curve for each protein or transcript, we employed our ECHO program [29]. ECHO is a software that fits a curve to rhythmic longitudinal data with replicates and evenly spaced time points, and the time series is at least 24 h long. The goal of using this is to then be able to compare the fitted rhythms of each analyte using network analysis methods, mentioned in Sects. 2.2.1, 2.2.2 and 2.3. ECHO analysis normalizes biological replicates such that the transcript or protein mean is 0 and the row standard deviation is 1. This normalized data is then fitted to a regression model, resulting in a relative expression profile for each transcript and protein that is a weighted combination of the data from the three replicates. ECHO also computes several parameters of the curve, such as the initial amplitude, amplitude change coefficient, rhythm period, phase shift between the curve fitted by ECHO and the original data, and the goodness of fit p value, which were all useful properties to verify the similarity of two rhythms. Further information on this process can be found in the source paper [29].

Using the combined LIMBR and ECHO approach, the rhythms of 36,000 transcripts and 6000 proteins were fitted. ECHO computes goodness-of-fit of the fitted rhythms by computing the Kendall's Tau test p value between the original data and the fitted rhythm it computed. These p values were then Benjamini-Hochberg (B-H) adjusted to account for potential false positives that occur in multiple hypothesis tests [32]. To ensure that the vast majority of rhythms analyzed were representative of the original data, only transcripts or proteins with a goodness-of-fit B-H adjusted p value less than 0.01 were considered for further analysis [32]. After B-H score filtering, there were 10092 transcripts and 1986 proteins left for analysis. Since relative expression curves of both the transcriptomic and proteomic data were generated using the same LIMBR + ECHO workflow, this data was merged into one multi-omic dataset consisting of 12,078 rhythms. This resulted in a final dataset of 12,078 rhythms (rows) and 24 columns (time points).

Network analysis

Once we created our merged multi-omic dataset, we next sought to characterize similarities between the expression profiles of transcripts and proteins in the multi-omic dataset using network analysis. To build a network, first an $N \times N$ similarity matrix, where N is the total number of analytes (12,078) in the dataset, was constructed [33]. Each element of this matrix, $s[i, j]$ corresponded to the similarity between analytes i and j , where i and j were between 1 and 12,078. The similarity metric used in this study was the Pearson correlation between each pair of rhythms. As such, the similarity between rhythm i and rhythm j , e.g., $s[i, j]$, was always between 1 (perfectly in-phase, overlapping) and -1 (perfectly antiphase). Networks were then defined by converting the similarity matrices into adjacency matrices using a dual-approach network analysis consisting of an unweighted network and a weighted network. Unweighted networks were represented by discrete adjacency matrices where each value in the similarity matrix was either converted to a 1, indicating two rhythms are “connected” in the network, or a 0, indicating no connection [34]. These networks are computationally efficient and effective in modeling specific outcomes or circumstances due to their binary nature. However, they may lack information about the resulting model. Weighted analysis, conversely, uses a continuous adjacency matrix that involves transforming the values in the similarity matrix such that larger correlation values have a larger separation from smaller correlation values [33]. In this way, weighted networks are more accurately able to capture groups of highly correlated rhythms, but they are computationally more expensive. Given their strengths and weaknesses, we used both approaches to capture local similarities between rhythms through unweighted network analysis and larger functional groups through weighted network analysis.

Unweighted network analysis

In our framework, unweighted network models were developed to find small communities of similar rhythms that show related patterns over time as compared to known core clock genes and proteins. To create these unweighted networks, we computed adjacency matrices using two methods: K-nearest neighbors (KNN) and correlation thresholding (CT). These methods are defined further below. Groups of similar expression patterns were then computed via several graph partitioning methods. These models were then validated with a robustness evaluation technique to determine if the computed communities are resilient to noise, e.g., if the similarity between two rhythms is stable in the presence of noise. The optimal unweighted network model was then selected and analyzed based on an assessment of all computed models’ robustness.

Correlation thresholding

Correlation thresholding (CT) is a method to generate adjacency matrices that is often used in biological co-expression analyses as the results are relatively easy to interpret due to the use of a simple threshold of the similarity matrix (correlation matrix) [35]. To convert the correlation matrix to an adjacency matrix, correlations above a threshold are marked as adjacent (1), while the rest are marked as non-adjacent (0). In circadian studies, the absolute value of the correlation in CT is typically used to group together

in-phase and anti-phase rhythms. However, as the goal of this study is to identify similar expression patterns among circadian rhythms, adjacency matrices were built off a correlation matrix, which was generated by computing the Pearson correlation coefficient between each pair of rhythms instead. In addition, as there are relatively few time points, we necessitated a high correlation between rhythms to consider two rhythms sufficiently similar to be considered related.

To determine a suitable correlation threshold, the percent of shared nearest neighbors was plotted against the correlation between circadian rhythms. If there were very few shared nearest neighbors at a correlation threshold, then this threshold would likely be too low to consider two rhythms co-expressed. To set the parameters for this procedure, we found the average number of overlapping nearest neighbors at several correlation levels among transcripts and proteins with a circadian period (Fig. 3). In general, there were few analytes that had many nearest neighbors with correlation less than 0.9. This is shown in Fig. 3A, as over 66% of the rhythms had an average correlation of 0.9 or greater with their 100-nearest neighbors. This demonstrated that setting a CT above 0.9 was an appropriate threshold for network analysis and the minimum CT used to create adjacency matrices was therefore set to 0.9. To garner a better understanding of the correlation, three adjacency matrices were computed in this work, using a CT of 0.9, 0.95, and 0.99.

K-nearest neighbors

K-Nearest Neighbors (KNN) is a widely used approach to identify similarities in biological data that, as opposed to CT, ranks each correlation to a given rhythm, sets the highest K to 1, and sets all other relationships to 0 [36, 37]. This results in an adjacency matrix where, given a rhythm, the rhythms that have the highest K correlation with that rhythm are considered “connected”. While this can mean that some barely correlated rhythms are marked as related, KNN benefits from being able to use directional relationships (e.g., i can be the nearest neighbor of j , but j does not have to be the nearest neighbor of i).

To determine the parameters of the adjacency matrix using KNN, we found the values of K that would be used for comparison in this study by calculating the standard deviation of the period, phase shift, and amplitude change coefficient parameters (computed from ECHO) among the K nearest neighbors, ranging from 10 to 150 nearest neighbors, in intervals of 10. Based on the low standard deviation of parameters, K values of 20, 40, 80 and 100 were found to have small standard deviations in ECHO-derived rhythmicity parameters such as phase shift, amplitude coefficient and period among each rhythm’s K-nearest neighbors while not having the exact same results as each other (i.e., non-redundant). and were therefore used to construct the adjacency matrices (Fig. 3B–D).

Model generation

Using both a CT and KNN approach, we build our D-CAF framework to generate models that identify groups of similarly expressed rhythms over time by partitioning a co-expression network into communities. These models then needed to be validated by evaluating their robustness to noise. To complete the partitioning step, graph partitioning (GP) algorithms, also known as community detection (CD)

methods, were applied to partition the network into densely connected communities (clusters) of rhythms [38]. In this framework, four CD methods were used: the original Leiden algorithm that used Modularity as a quality function [39], and 3 adaptations of this algorithm to use the RB Potts (RBP), Asymptotic Surprise (AS), and Significance quality functions respectively [40–43]. Essentially, each of these methods attempt to find communities, or clusters of nodes in the network that maximize each of the respective quality functions. Each of these methods were employed using the 'leidenalg' Python package. More details on each of these can be found in Supplementary File 2. [44] Each model was applied to all 7 (3 CT, 4 KNN) computed unweighted networks (defined by adjacency matrices). In total, 28 models were

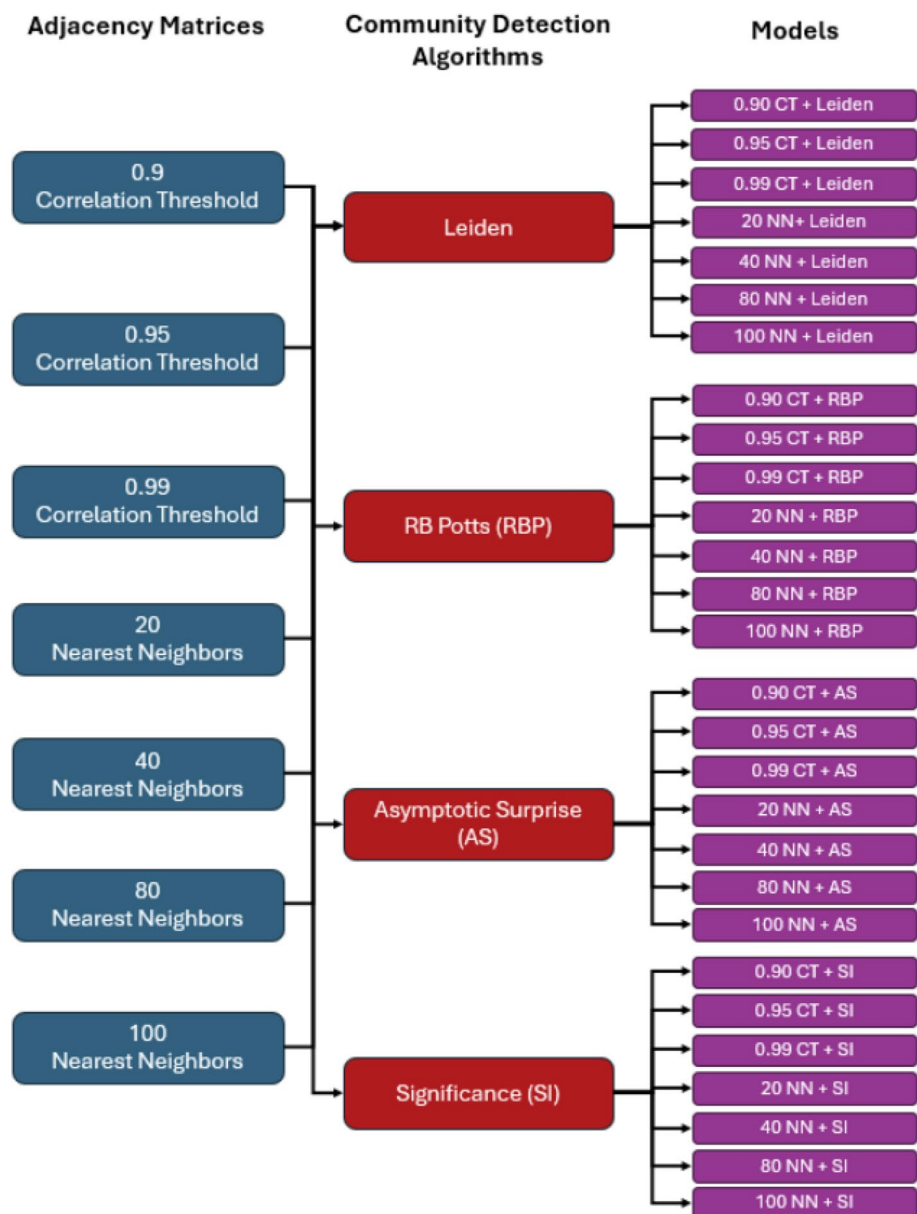


Fig. 2 A graphical representation of the origin of the 28 unweighted models used in this study. 7 adjacency matrices were clustered using 4 community detection algorithms to create a total of 28 distinct models

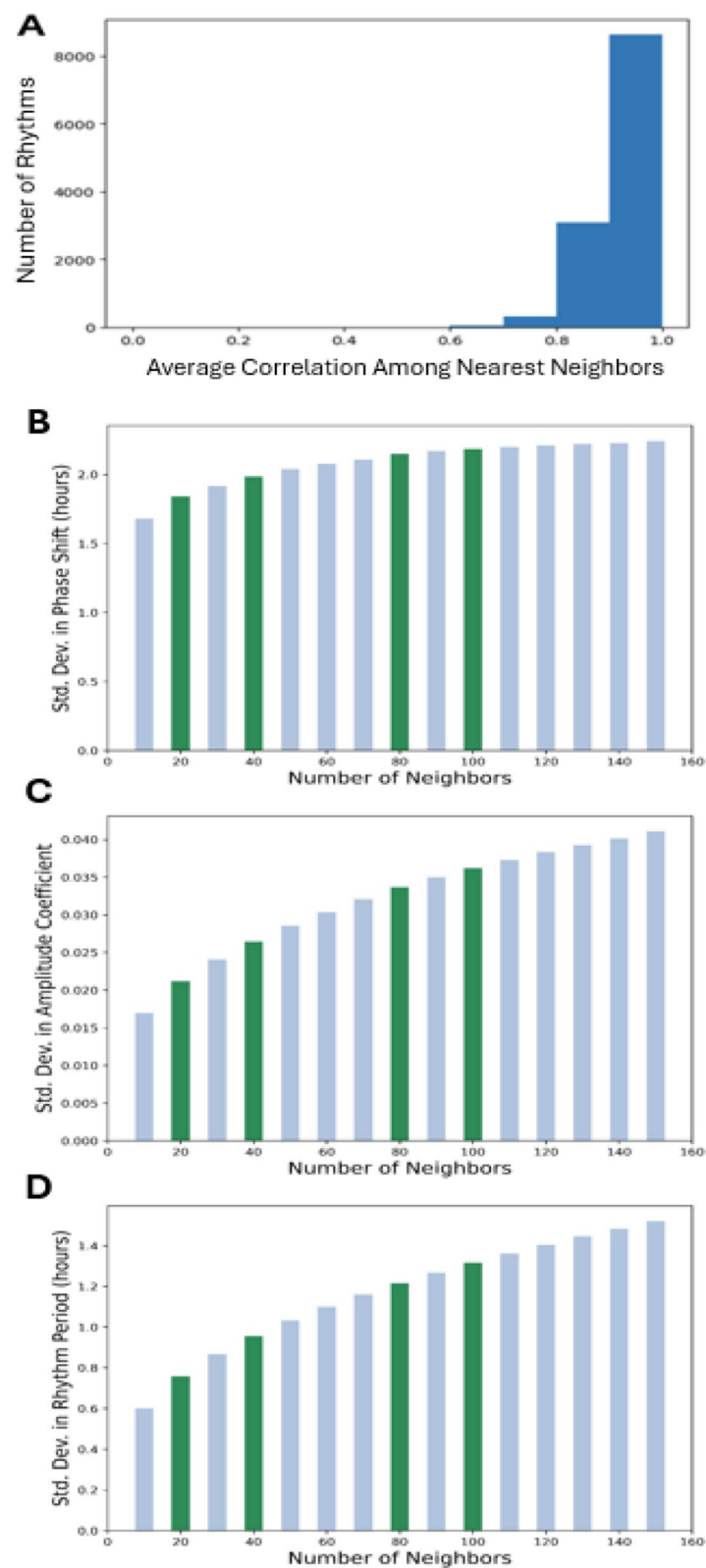


Fig. 3 Parameter selection for adjacency matrices used in unweighted model construction. **A** Shows the average correlation among each rhythm’s 100 nearest neighbors. **B–D** show the average standard deviation in phase shift, period, and amplitude coefficient respectively among a model’s K-nearest neighbors. Bars in green indicate values of K selected to create adjacency matrices for unweighted model construction

generated in this study by applying each of the 4 community detection algorithms to each of the 7 adjacency matrices. A graphical representation of this process is shown in Fig. 2. These models partitioned sets of rhythms into communities, resulting in a set of cluster labels for each model that indicated, for each rhythm, which cluster that rhythm belongs to. From these 28 models, the optimal model was then selected via robustness evaluation, e.g. testing how a model's cluster labels change when noise is introduced to the data.

As the purpose of this study was to specifically find circadian clusters, the clusters produced by the models were further filtered by identifying clusters with an average circadian period (20–28 h) and small standard deviation in period. The overall goal was to develop clusters with high within-cluster similarity and low between-cluster similarity. To this end, standard deviation in the ECHO-derived oscillation parameters (phase shift and period) were computed to determine the similarity of the rhythms contained within a cluster. If the standard deviation of these parameters is less than the data collection resolution (e.g. less than 2 h for the dataset used in this study), then the rhythms within a cluster were considered to have highly similar expression patterns. Since ECHO fits rhythmic curves and therefore computes parameters such as period with the available data, a standard deviation of these parameters larger than the data collection resolution would indicate rhythms that have different expression patterns are being clustered together, which would result in poorer interpretability of those clusters. Furthermore, the modularity of these clusters were computed to ensure the overall partitioned graph had many connections within each cluster and few connections between clusters.

Weighted network analysis

While unweighted community detection finds small communities of rhythms, weighted network analysis can identify large functional groups of co-expressed rhythms. To do so, weighted networks were generated by weighting factor β of the similarity matrices computed by the Weighted Gene Co-Expression Analysis package (WGCNA) based on the truncated R^2 value [33]. The weighted similarity matrix was then generated by raising each element of the original similarity matrix to a power β . This is because raising values less than 1 to any positive power causes the smaller values to decrease more than the values closer to 1. By doing this, high correlations remained high and low correlations converged towards zero as was shown in the original WGCNA study [33]. Spectral clustering was applied to create clusters from this weighted similarity matrix, primarily because it (1) was faster for continuous similarity matrices and (2) allowed for the specification of the number of clusters to compute to avoid clusters that are too small. 30 clusters were chosen from the WGCNA analysis as the standard deviation of the within-cluster period plateaued for greater amounts of clusters, indicating larger numbers of clusters did not create significantly more tight-knit clusters (Fig. S1.1). Essentially, as the number of clusters increased beyond 30, the standard deviation of the within-cluster period plateaued, meaning this was the peak similarity of transcripts and proteins within each cluster. The robustness of the resulting model was then computed using the same procedure as the unweighted models to determine the reliability of the model.

Functional enrichment analysis

Circadian clusters were identified by selecting only clusters with average periods that were approximately 24 h (between 20 and 28 h). Biological functions in clusters were analyzed via protein-protein interaction (PPI) networks modeled by StringDB [45]. Since the number of transcripts in this data outnumbered the amount of proteins, there are many instances where each transcript does not have its corresponding protein in the dataset. As such, analysis via methods that would analyze transcript-protein pairs such as Reactome or PathviewDB analysis prove difficult with this dataset. Therefore, PPI networks were chosen for biological analysis of clusters, as it enables comparison of these transcripts with the proteins that were detected. By inputting the members of each cluster, StringDB outputs a protein-protein interaction network where nodes are analytes and edges indicated interactions. To analyze this network, the PPI enrichment p value was measured to determine if the number of connections within the cluster was significantly greater than the number of connections that would be expected in a random clustering of analytes. If there were significantly more ($p < 0.05$) edges than expected in the PPI network, then the clusters identified had some functional significance. Finally, gene ontology and reaction pathway analysis was performed to determine potential pathways and functions that the molecules in each cluster were related to by utilizing StringDB to analyze the members of each cluster. This was performed by StringDB database analysis, which searches several databases such as gene ontology, KEGG, PubMed publications, Pfam, and SMART domains to find enriched functions among the analytes in a specific cluster [45].

Robustness evaluation

After defining both the weighted and unweighted models, the models then had to be validated to ensure quality. To do so, model robustness was evaluated by determining how stable a model's clustering output, referred to as C , was when noise was introduced to the data. This was implemented as biological data can be inherently noisy, and slightly different measurements in replicates could result in completely different networks and clustering outputs. If a community detection model is robust to noise, it means the relationship between the rhythms that were clustered are sufficiently strong to ignore small amounts of noise. To do so, noise was simulated by adding or subtracting a number randomly generated by a Gaussian distribution with mean = 0 and standard deviation = σ to each value in the data post-ECHO transformation. The values of σ used in this study were 0.05–0.5 in increments of 0.05, with greater values of σ indicating greater severity of noise. This range was chosen to test how reliable a community detection model was with only a small amount of noise (0.05) and severe noise that would be expected to cause a decrease in robustness (0.5).

After applying Gaussian noise, the adjacency matrix was then re-created for each network and community detection was reapplied to generate C^* , the recomputed cluster labels. The adjusted mutual information (AMI) score between C and C^* was used to measure the difference between these original cluster labels and the recomputed cluster labels. The formula to compute AMI given two sets of cluster labels X, Y

is defined by 1, where MI denotes mutual information, $E(x)$ is the expected value of x , and $H(x)$ is the entropy of x .

$$AMI(X, Y) = \frac{[MI(X, Y) - E(MI(X, Y))]}{[avg(H(X), H(Y)) - E(MI(X, Y))]} \quad (1)$$

When comparing two sets of cluster labels for a series of data points, we used an AMI score of 1 to represent total overlap, and an AMI score of 0 to represent total randomness [46]. This was repeated 50 times for each model per σ . This noise application was repeated because, while the same distribution was used to sample the noise added or subtracted to each element, the exact amount of noise was randomly sampled from the distribution $N(0, \sigma)$. As such, several repetitions of noise application would provide a better overall understanding of each model's robustness. The average AMI score among these 50 repetitions was recorded as the model's robustness to the severity of noise (σ). Models with greater AMI were deemed more robust to noise, and therefore were a higher quality model.

Results

Unweighted network analysis of macrophage transcripts and proteins

Once we had a sound pipeline for D-CAF, we subjected our previously analyzed data from Collins et al. to D-CAF unweighted analysis, with the goal of identifying densely connected clusters of rhythmic elements that were also robust to Gaussian noise perturbations [30]. We measured the robustness of the original 28 candidate models (Fig. 2) and computed the adjusted mutual information (AMI) score during repeated perturbation analysis. Heatmaps containing the average AMI score of each model when perturbed by small ($\sigma=0.05$) and severe ($\sigma=0.5$) noise are presented in Fig. 4A and B. A higher AMI (close to 1) indicates a more stable model, whereas a lower AMI indicates that the model was not robust to noise, and is therefore not reliable [47].

At low noise severity ($\sigma = 0.05$), we found the most robust models were those that were generated using the Asymptotic Surprise (AS) and Significance community detection methods. In general, the Leiden and RB Potts community detection methods were less robust than the AS and Significance community detection methods, likely because the AS and Significance methods are known to be able to be better at forming small communities, which is the goal of unweighted network analysis. In investigating the matrices, we found only the 0.99 CT Adjacency matrix was not as robust as all the other matrices. This is likely due to the conservative correlation threshold, where even slight noise would cause a rhythm to fall below the correlation threshold of 0.99. At severe noise ($\sigma = 0.5$), we found the maximum AMI was achieved when the 80 and 100-NN matrices were paired with the AS community detection (Fig. 4B). Interestingly, all CT-based clustering models generated low AMI scores when challenged with severe noise, indicating that these cluster models were not very reliable. This data reinforces the idea that candidate models should be evaluated at multiple severities of noise to ensure cluster model robustness when performing co-expression analysis.

Given that it was among the most robust models to both high and low severities of noise, the 80-NN + AS model was selected and used for the unweighted co-expression analysis of the data from Collins et al. [30]. The 80-NN + AS model partitioned

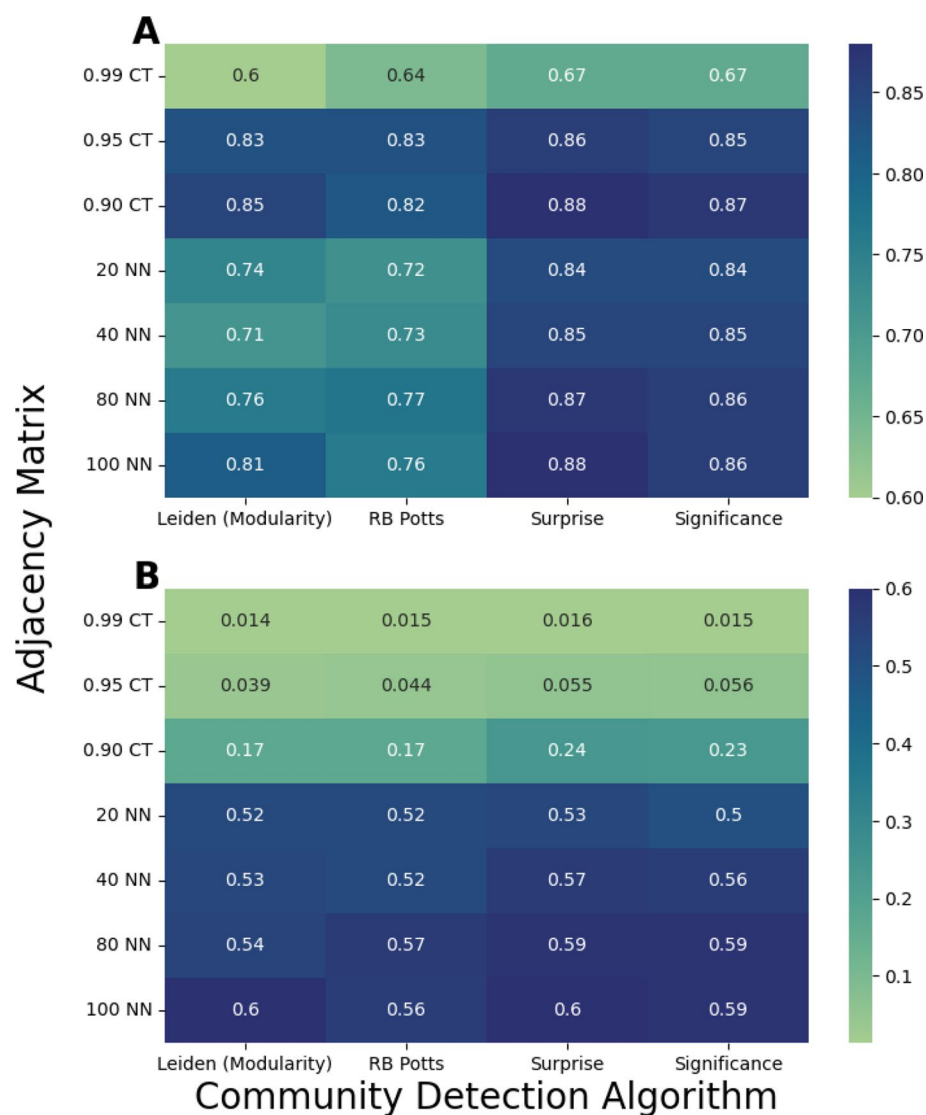


Fig. 4 AMI scores identify optimal models to be used for Unweighted Network Analysis. Heatmaps displaying the adjusted mutual information (AMI) of clusters predicted by unweighted analysis when data was perturbed by Gaussian distributed noise with a standard deviation of 0.05 **(A)** and 0.5 **(B)**. Values closer to 1 indicate higher similarity between the perturbed and original clusters and implies greater robustness

the data into 84 clusters of rhythms. Of these, 18 were found to have a circadian period (i.e. the average period \pm standard deviation for these clusters was between 20 and 28 h). Within each circadian cluster, the average standard deviation in period was 1.49 h, indicating that the rhythms in each cluster oscillated at approximately the same frequency. Furthermore, the average standard deviation in phase shift within each cluster was 1.86 h, indicating that rhythms in the each cluster shared the same circadian phase. A cluster-by-cluster breakdown of the cluster periods and phase shifts is presented in Table S1.1. The low standard deviation of rhythm parameters within each cluster demonstrated that the 80-NN + AS model successfully grouped similar rhythms without requiring explicit information about the rhythm parameters (e.g., phase shift and period).

We next computed the modularity score of the clusters generated by the 80-NN + AS model. The modularity of a cluster model score can fall between -1 and 1 , and a score closer to 1 means there are many connections within a cluster and few connections between clusters (i.e., clusters are dense and distinct), which is optimal for a clustering model [48]. Previous research has indicated that a modularity above 0.3 indicates that the structure of the network is not random [49]. The computed modularity of the 80-NN + AS model used in this study was 0.640 , demonstrating that the clusters found were both densely connected within each cluster and sparsely connected between clusters, further supporting that the 80-NN + AS model successfully clustered our data.

When analyzing what transcripts and proteins were identified within the clusters, we noted that transcripts/proteins that are a part of the core circadian transcription-translation negative feedback loop were found within clusters identified by the 80-NN + AS model that oscillated with a circadian period (Fig. 5). For example, *Per1* and *Cry2*, two genes that encode for components that regulate the core clock mechanism, were found in the cluster represented in Fig. 5A, while *Cry1* and *Npas2* were found individually in other clusters (Fig. 5B and C) [12]. Notably, the cluster with *Per1* and *Cry2* also contained the immune proteins *Tm9sf1*, *Ptcra*, and *NTAL*, all genes/proteins not previously

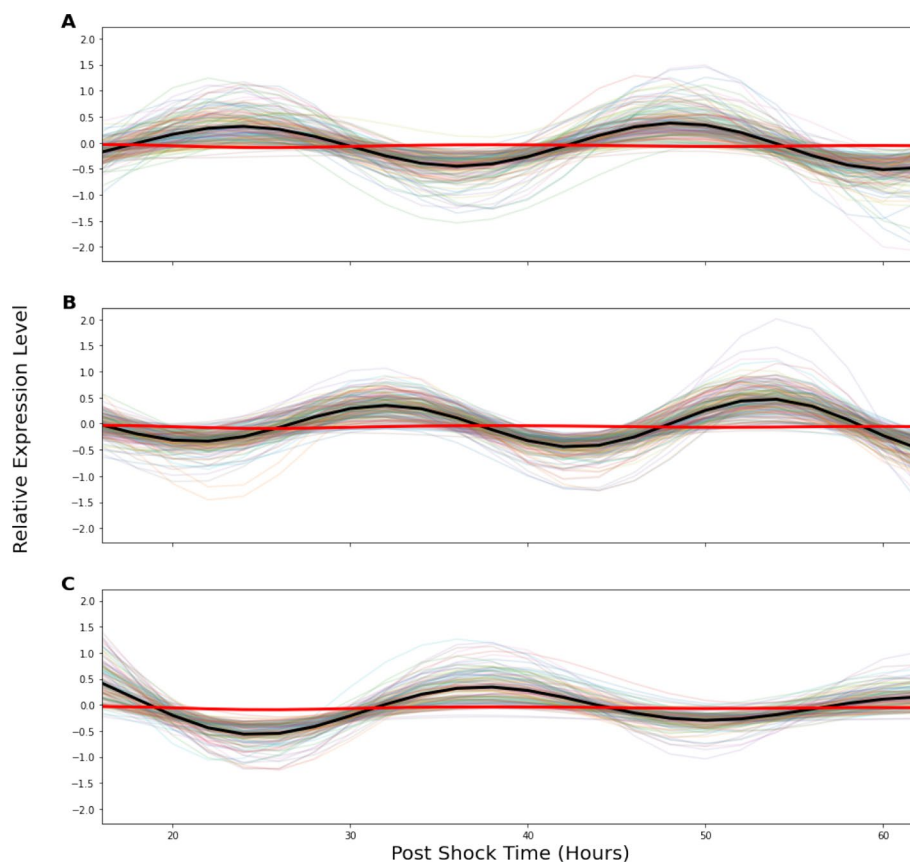


Fig. 5 Expression profiles of circadian clusters identified by Unweighted Network Analysis demonstrate successful clustering. Each displayed cluster was found to have a circadian period, with the average expression profile for the cluster bolded in black. As a comparison, the average expression profile for the whole dataset is displayed bolded in red. The clusters contained 173 genes/proteins (A), 189 genes/proteins (B), 137 genes/proteins (C)

shown to have a relationship with the circadian clock or circadian output [50–52]. These genes/proteins have a correlation score of 0.884, 0.995, and 0.676 with respect to *Per1*, indicating that the unweighted analysis was able to identify and cluster novel genes/proteins with known circadian clock genes. As such, the clusters generated by our unweighted analysis present novel avenues of investigation for circadian co-regulation.

Weighted network analysis of macrophage transcripts and proteins

Although the small deviation in rhythm parameters from the unweighted network clusters indicated high similarity between individual rhythms within the same cluster, deriving global biological conclusions from these clusters was difficult as gene ontology or pathway analysis requires a larger sample size to find enriched functions. For this reason, we next employed the weighted network analysis function from D-CAF to analyze the data from Collins et al. [30] to develop clusters that could elucidate more broad, large-scale network information within the multi-omics dataset. To this end, fewer, larger clusters were generated by applying spectral clustering to the weighted network (SC + WN) model as described above. Of the 30 clusters identified, 5 were found to have a circadian period (i.e. the average period \pm standard deviation for these clusters was between 20 and 28 h). The average standard deviation in period among these clusters was 2.28 h, and the average standard deviation among the phase shifts was 1.95 h. Although these standard deviations were significantly larger than the standard deviation of parameters among the circadian clusters generated by using an unweighted network approach, they were still close to the resolution of the time series used to generate the data (2 h).

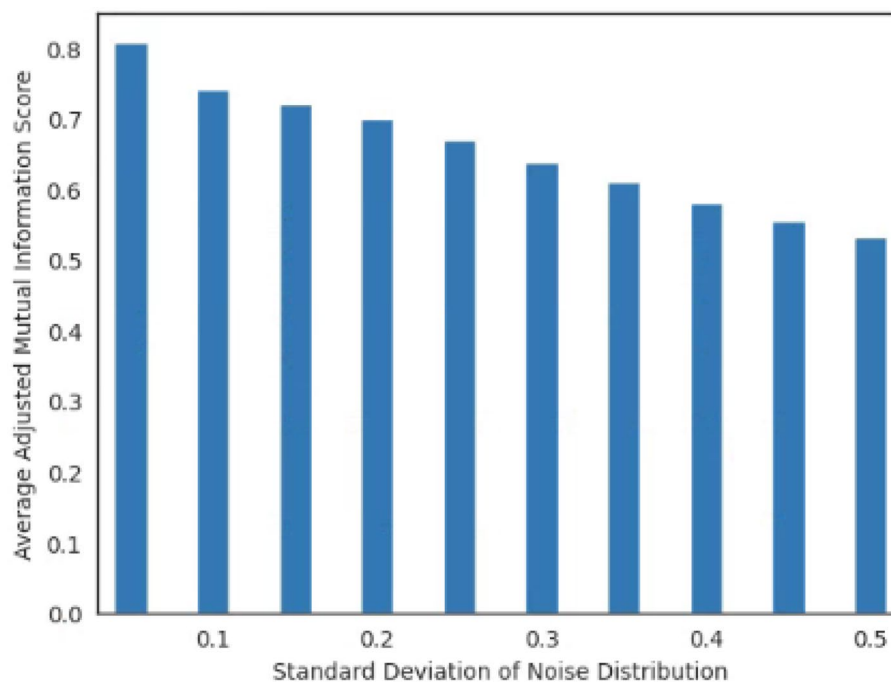


Fig. 6 AMI scores for Weighted Network Analysis are similar to scores for the Unweighted Network Analysis. Higher AMI indicates higher overlap between the perturbed clustering and the original clustering. The severity of noise is represented by the standard deviation of the noise distribution. AMI for each severity of noise is averaged over 50 repetitions

StringDB was used to find biological functions regulated by analytes with similar expression profiles to known clock transcripts and proteins. To this end, PPI enrichment networks were constructed from each of the SC + WN circadian clusters and we found that three of these networks had a PPI enrichment p value less than 0.05, indicating that the cluster contained more biologically related analytes than would be expected in a random list of analytes. PPI enrichment p values were calculated using the StringDB software, which compares the difference between the number of connected nodes in a random network to the number of connected nodes identified in the network from the data [45]. A list of p values for each of the PPI networks computed from the SC + WN circadian clusters is available in Table S1.2. One such PPI network showed that rhythms from the IL-27 signaling pathway were contained in the same cluster, and therefore had highly similar expression patterns to known clock genes in mouse macrophages (Fig. 7A). In addition to finding functionally related clusters,

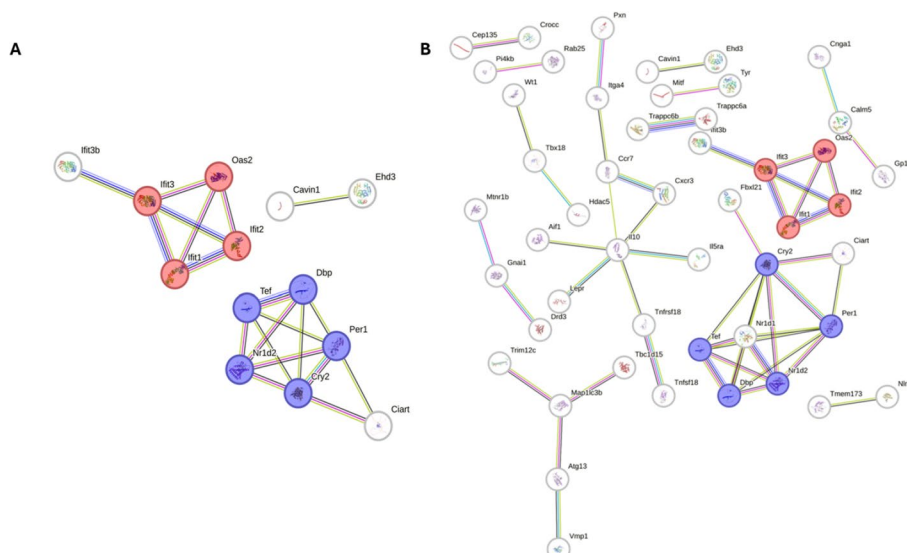


Fig. 7 PPI networks derived from **A** the 80-NN + AS model and **B** the SC + WN model. These networks show a cluster containing several core clock components with only edges with a minimum interaction score of 0.7 (high confidence) included for ease of visualization. Core clock components are denoted in purple and IL-27 signaling components are denoted in red

weighted clusters uncover more connections than the unweighted clusters (Fig. 7B). When we created PPI enrichment networks from the unweighted and weighted modeled data, we found that, though the same module of IL-27 analytes were clustered with core clock genes, the weighted modeled data identified more auxiliary connections (Fig. 7A and B), demonstrating while both the unweighted and weighted models used in this study identify the same core information, the SC + WN model creates a more comprehensive functional network.

Discussion

Analysis of multi-omics data has long been a hurdle in the circadian field. Existing methods to integrate longitudinal multi-omics data generally use latent variable methods, which make analysis of rhythmic properties difficult, or require an even number of transcripts and proteins in the dataset. D-CAF does neither of these, allowing the user to use and compare all rhythms contained in a dataset. Our goal in creating the D-CAF pipeline was to enable simultaneous co-expression analysis of circadian, or any type, of transcript and protein time series measurements. This, in turn, would identify potential analytes of interest for future research. To this end, we utilized our LIMBR and ECHO software to successfully generate comparable relative expression curves for raw transcriptomic and proteomic time series. The successful integration is evidenced by the lack of clusters formed solely of isolated transcript or protein data types after unweighted analysis (Figure S1.2). Following this, co-expression analysis pathways for the multi-omic data were created using two methods: unweighted and weighted network analysis, both of which partitioned the data into densely connected clusters that were robust to Gaussian noise (Fig. 4). This robustness, combined with the small deviation in rhythm parameters within each cluster, indicated this approach identified small and large groups that may have regulatory relationships, with the unweighted networks identifying small cluster groups that are highly related (Figs. 5 and 7A) and the weighted networks granting insights into less related clusters that granted a more global view of the circadian network (Fig. 7B). In total, the combination of clustering approaches were useful to find robust local and global relationships between known clock genes/proteins and other genes/proteins not previously found to be related to the circadian clock. With this, our D-CAF framework can be used by others to similarly analyze their own multi-omic circadian data.

As an example of these novel relationships, by unweighted analysis we found that the transcripts *Ptcra* and *Tm9sf1*, and protein *NTAL*, were contained in the same cluster as known clock genes (Fig. 5A). Given the high degree of similarity in expression patterns and high correlation between rhythms in the same cluster, this suggests that these rhythms are likely co-regulated by factors that also influence the phase timing of the clock. *Ptcra* regulates early T-cell development, *Tm9sf1* has been shown to be involved in the inflammatory response to acute lung injury, and *NTAL* is important for the signaling of mast cells that control the immune system, suggesting further pathways by which circadian rhythms tune the immune system [50, 53, 54]. Using weighted analysis, we highlighted components of the interferon (IFN) signaling pathway (*Ifit1*, *Ifit2*, *Ifit3*, and *Oas2*), which are a major subclass of IFN-stimulated genes (ISGs) known to coordinate innate immune signaling pathways and directly inhibit viral protein synthesis.

[55]. These ISGs were significantly represented (P value = 0.0142) in a cluster along with core circadian clock components *Cry2* and *Per1*, as well as accessory components *Ciart*, *Nr1d1*, *Nr1d2*, *Tef*, *Dbp*, and *Hdac5*.

Circadian rhythms have been implicated in ISG responses across various tissues, including the skin, lungs, and liver. In a study using pharmacological activation on murine skin, ISG expression was found to vary depending on the treatment time within the circadian cycle, primarily eliciting responses from epidermal immune cells, particularly monocytes [56]. Additionally, systemic *Bmal1* knockout mice exhibited an amplified ISG response in both skin and isolated epidermis, suggesting that mice lacking circadian rhythms may experience heightened activation of the IFN pathway. Transcriptomic data from primary human hepatocytes revealed that inflammatory ISGs exhibit circadian-regulated expression patterns. However, when *Bmal1* expression is suppressed, a down-regulation of ISGs is observed [57]. This suggests that the induction of ISGs varies depending on the time of stimulation, potentially leading to innate responses that fluctuate throughout the circadian day. SARS-CoV-2 has been demonstrated to also induce IFNs and ISGs, both in *in vivo* and *in vitro* airway mucosa [58, 59]. ISGs, particularly members of the *Ifit* family, have been shown to exhibit antiviral activity against SARS-CoV-2 [60]. In *Bmal1*-silenced cells, the induction of ISGs resulted in a decreased expression of the viral receptor angiotensin-converting enzyme 2 (ACE2) and reduced viral entry into lung epithelial cells [61]. This increase in ISG transcripts could be regulated at various stages of the IFN sensing and signaling pathways through direct or indirect mechanisms. These instances in skin, liver, and lung, along with the relationships predicted by our networks, suggest that the ISG antiviral response varies depending on the time of day.

There are limitations associated with this work. First, the network parameters were chosen based on the properties of the multi-omic data specifically used in this study. Due to the nature of unsupervised learning methods such as clustering and community detection, if the trends in the data change, the models themselves will also change. These changes include using only one of the individual omics datasets, using a different range of time points, or attempting to analyze additional proteins and transcripts post-hoc. Furthermore, there are many steps between mRNA transcription and protein translation, and as such phenomena such as translational modulation and protein degradation may be overlooked by only looking at the clustering data. As such, it is recommended that this framework should precede experimental follow-up, not replace experimental work. In addition, a very conservative BH adjusted p value threshold was used to filter the data, which may result in too many rhythms being filtered out for smaller datasets. As such, the properties of datasets should be evaluated and used when tuning the parameters used to apply the D-CAF framework.

Applications of D-CAF in future work should note that this framework is applicable to relatively large (greater than 100 analytes) datasets of longitudinal, multi-omic data, where A) the data is continuous, B) there are several replicates of each measurement, and C) the majority of analytes have fewer than 30% missing data. Additionally, the time series should have enough measurements that the correlation computed is informative. Future work should note that D-CAF is primarily meant to be a screening tool that can aid research into mechanisms of circadian regulation of biological

processes by finding analytes that are similarly co-expressed, thereby identifying potential analytes of interest and PPI networks that are co-regulated with processes that modulate the circadian clock. As such, even when the sample size is too small to yield clusters that provide meaningful function enrichment results, D-CAF can still identify individual analytes that are co-expressed, and thus potentially co-regulated with circadian processes, as was shown with the clustering results of the 80-NN+AS model. Finally, transcripts and the proteins they transcribe with significant phase differences that result in them belonging to two different clusters. Thus, it may be of interest to future studies as to why the transcript may share similar rhythmic properties with circadian clock components, but the protein it transcribes does not.

In summary, we found D-CAF created a framework to successfully integrate multi-omic data to find clusters of highly similar rhythms with biological relevance using a validated and multifaceted analysis. The pairing of unweighted and weighted network analysis facilitated the identification of relationships between clock genes/proteins and novel genes/proteins with a high correlation while simultaneously allowing for the broad analysis of the circadian network. While any model generated must be validated, D-CAF presents a novel method to perform multi-omic co-expression analysis of longitudinal omics data to determine previously unknown relationships.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-025-06089-1>.

Supplementary file 1.
Supplementary file 2.

Acknowledgements

The authors would like to thank Sharleen Buel and Meaghan Jankowski for their contributions to this study. This work was supported by a NIH-National Institute of Aging T32 Fellowship (T32AG078123) (to J.C. and C.C.), and a NIH-National Institute of Biomedical Imaging and Bioengineering Grant (U01EB022546), a NIH-National Institute of General Medical Sciences Grant (R35GM128687), and an NSF CAREER Award 2045674 (to J.M.H.).

Author contributions

Conceptualization: J.C., J.H., J.M.H.; Formal Analysis: J.C.; Funding Acquisition: J.C., C.C., J.H., J.M.H.; Investigation: J.C., J.H., J.M.H.; Methodology: J.C., J.H., J.M.H.; Software: J.C.; Visualization: J.C.; Writing- original draft: J.C., J.M.H.; Writing- review and editing: J.C., C.C., J.H., and J.M.H.

Funding

This work was supported by a NIH-National Institute of Aging T32 Fellowship (T32AG078123) (to J.C. and C.C.), and a NIH-National Institute of Biomedical Imaging and Bioengineering Grant (U01EB022546), a NIH-National Institute of General Medical Sciences Grant (R35GM128687), and an NSF CAREER Award 2045674 (to J.M.H.).

Availability of data and materials

The source code and preprocessed data are available at <https://github.com/ChuahResearchGroup/DCAF>. The original data can be found at <http://www.genome.org/cgi/doi/10.1101/gr.263814.120>. [30]

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors have no Conflict of interest to declare.

Received: 18 September 2024 Accepted: 18 February 2025

Published online: 04 March 2025

References

1. Vitaterna MH, Takahashi JS, Turek FW. Overview of circadian rhythms. *Alcohol Res Health*. 2001;25(2):85–93.
2. Tett P. The photic zone. *Light and Life in the Sea*, 1990;59–87
3. Huang W, Ramsey KM, Marcheva B, Bass J, et al. Circadian rhythms, sleep, and metabolism. *J Clin Investig*. 2011;121(6):2133–41.
4. Ma D, Li S, Molusky MM, Lin JD. Circadian autophagy rhythm: a link between clock and metabolism? *Trends Endocrinol Metab*. 2012;23(7):319–25.
5. Scheiermann C, Kunisaki Y, Frenette PS. Circadian control of the immune system. *Nat Rev Immunol*. 2013;13(3):190–8.
6. Gnocchi D, Bruscalupi G. Circadian rhythms and hormonal homeostasis: pathophysiological implications. *Biology*. 2017;6(1):10.
7. Panda S. Circadian physiology of metabolism. *Science*. 2016;354(6315):1008–15.
8. Musiek ES, Xiong DD, Holtzman DM. Sleep, circadian rhythms, and the pathogenesis of Alzheimer disease. *Exp Mol Med*. 2015;47(3):148–148.
9. Gale JE, Cox HI, Qian J, Block GD, Colwell CS, Matveyenko AV. Disruption of circadian rhythms accelerates development of diabetes through pancreatic beta-cell loss and dysfunction. *J Biol Rhythm*. 2011;26(5):423–33.
10. Gery S, Koeffler HP. Circadian rhythms and cancer. *Cell Cycle*. 2010;9(6):1097–103.
11. Thosar SS, Butler MP, Shea SA, et al. Role of the circadian system in cardiovascular disease. *J Clin Investig*. 2018;128(6):2157–67.
12. Ko CH, Takahashi JS. Molecular components of the mammalian circadian clock. *Hum Mol Genet*. 2006;15(suppl-2):271–7.
13. Reppert SM, Weaver DR. Molecular analysis of mammalian circadian rhythms. *Annu Rev Physiol*. 2001;63(1):647–76.
14. Archer SN, Oster H. How sleep and wakefulness influence circadian rhythmicity: effects of insufficient and mistimed sleep on the animal and human transcriptome. *J Sleep Res*. 2015;24(5):476–93.
15. Takahashi JS. Transcriptional architecture of the mammalian circadian clock. *Nat Rev Genet*. 2017;18(3):164–79.
16. Mauvoisin D, Gachon F. Proteomics in circadian biology. *J Mol Biol*. 2020;432(12):3565–77.
17. Reddy AB, Karp NA, Maywood ES, Sage EA, Deery M, O'Neill JS, Wong GK, Chesham J, Odell M, Lilley KS, et al. Circadian orchestration of the hepatic proteome. *Curr Biol*. 2006;16(11):1107–15.
18. Scarpa JR, Elemento O. Multi-omic molecular profiling and network biology for precision anaesthesiology: a narrative review. *Br J Anaesth*. 2023;131:26.
19. Scandola S, Mehta D, Li Q, Rodriguez Gallo MC, Castillo B, Uhrig RG. Multi-omic analysis shows reveille clock genes are involved in carbohydrate metabolism and proteasome function. *Plant Physiol*. 2022;190(2):1005–23.
20. Na Y-J, Sung JH, Lee SC, Lee Y-J, Choi YJ, Park W-Y, Shin HS, Kim JH. Comprehensive analysis of microRNA-mRNA co-expression in circadian rhythm. *Exp Mol Med*. 2009;41(9):638–47.
21. Suravajhala P, Kogelman LJ, Kadarmideen HN. Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare. *Genet Sel Evol*. 2016;48(1):1–14.
22. Serin EA, Nijveen H, Hilhorst HW, Ligterink W. Learning from co-expression networks: possibilities and challenges. *Front Plant Sci*. 2016;7:444.
23. Noort V, Snel B, Huynen MA. Predicting gene function by conserved co-expression. *TRENDS Genet*. 2003;19(5):238–42.
24. Vasaikar SV, Savage AK, Gong Q, Swanson E, Talla A, Lord C, Heubeck AT, Reading J, Graybuck LT, Meijer P, et al. A comprehensive platform for analyzing longitudinal multi-omics data. *Nat Commun*. 2023;14(1):1684.
25. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, Lê Cao K-A. Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*. 2019;35(17):3055–62.
26. Rohart F, Eslami A, Matigian N, Bougeard S, Le Cao K-A. Mint: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinform*. 2017;18(1):1–13.
27. Argelaguet R, Velten B, Arnod D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. 2018;14(6):8124.
28. Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J*. 2021;19:3735–46.
29. Santos H, Collins EJ, Mann C, Sagan AW, Jankowski MS, Bennett KP, Hurley JM. Echo: an application for detection and analysis of oscillators identifies metabolic regulation on genome-wide circadian output. *Bioinformatics*. 2020;36(3):773–81.
30. Collins EJ, Cervantes-Silva MP, Timmons GA, O'Siorain JR, Curtis AM, Hurley JM. Post-transcriptional circadian regulation in macrophages organizes temporally distinct immunometabolic states. *Genome Res*. 2021;31(2):171–85.
31. Crowell AM, Greene CS, Loros JJ, Dunlap JC. Learning and imputation for mass-spec bias reduction (LIMBR). *Bioinformatics*. 2019;35(9):1518–26.
32. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc Ser B (Methodol)*. 1995;57(1):289–300.
33. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform*. 2008;9(1):1–13.
34. Fekete J-D. Visualizing networks using adjacency matrices: progresses and challenges. In: 2009 11th IEEE international conference on computer-aided design and computer graphics. IEEE; 2009 pp. 636–638.
35. Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinform*. 2012;13(1):1–21.
36. Okun O, Priisalu H. Dataset complexity in gene expression based cancer classification using ensembles of k-nearest neighbors. *Artif Intell Med*. 2009;45(2–3):151–62.
37. Ye X, Wu Y, Pi J, Li H, Liu B, Wang Y, Li J. Deepgmd: a graph-neural-network-based method to detect gene regulator module. *IEEE/ACM Trans Comput Biol Bioinform*. 2021;19(6):3366–73.
38. Bichot C-E, Siarry P. Graph partitioning. Hoboken: Wiley; 2013.

39. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep*. 2019;9(1):5233.
40. Reichardt J, Bornholdt S. Statistical mechanics of community detection. *Phys Rev E*. 2006;74(1): 016110.
41. Leicht EA, Newman ME. Community structure in directed networks. *Phys Rev Lett*. 2008;100(11): 118703.
42. Traag VA, Aldecoa R, Delvenne J-C. Detecting communities using asymptotical surprise. *Phys Rev E*. 2015;92(2): 022816.
43. Traag VA, Krings G, Van Dooren P. Significant scales in community structure. *Sci Rep*. 2013;3(1):2930.
44. Traag V, Zanini F, Gibson R, van Kuppevelt D. vtraag/leidenalg 0.8.1 (0.8.1). Zenodo. 2020. <https://doi.org/10.5281/zenodo.4015087>
45. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, et al. The string database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2021;49(D1):605–12.
46. Vinh N, Epps J, Bailey J. Information theoretic measures for clusterings comparison: variants. In: Properties, normalization and correction for chance. 2009;18
47. Vinh N.X, Epps J, Bailey J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th annual international conference on machine learning; 2009. pp. 1073–1080.
48. Newman ME. Modularity and community structure in networks. *Proc Natl Acad Sci*. 2006;103(23):8577–82.
49. Clauset A, Newman ME, Moore C. Finding community structure in very large networks. *Phys Rev E*. 2004;70(6): 066111.
50. Xiao J, Shen X, Chen H, Ding L, Wang K, Zhai L, Mao C. Tm9sf1 knockdown decreases inflammation by enhancing autophagy in a mouse model of acute lung injury. *Heliyon*. 2022;8(12): e12092.
51. Von Boehmer H. Unique features of the pre-t-cell receptor alpha-chain: not just a surrogate. *Nat Rev Immunol*. 2005;5(7):571–7.
52. Brdiccka T, Imrich M, Angelisova P, Brdicckovaa N, Horvath O, SSpicka J, Hilgert I, Petra L, Draber P, Novak P, et al. Non-T cell activation linker (NTAL) a transmembrane adaptor protein involved in immunoreceptor signaling. *J Exp Med*. 2002;196(12):1617–26.
53. Dik WA, Pike-Overzet K, Weerkamp F, De Ridder D, De Haas EF, Baert MR, Van Der Spek P, Koster EE, Reinders MJ, Van Dongen JJ, et al. New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling. *J Exp Med*. 2005;201(11):1715–23.
54. Volna P, Lebduska P, Draberova L, Simova S, Heneberg P, Boubelik M, Bugajev V, Malissen B, Wilson BS, Horejsi V, et al. Negative regulation of mast cell signaling and function by the adaptor LAB/NTAL. *J Exp Med*. 2004;200(8):1001–14.
55. Schneider WM, Chevillotte MD, Rice CM. Interferon-stimulated genes: a complex web of host defenses. *Annu Rev Immunol*. 2014;32(1):513–45.
56. Greenberg EN, Marshall ME, Jin S, Venkatesh S, Dragan M, Tsoi LC, Gudjonsson JE, Nie Q, Takahashi JS, Andersen B. Circadian control of interferon-sensitive gene expression in murine skin. *Proc Natl Acad Sci*. 2020;117(11):5761–71.
57. March S, Nerurkar N, Jain A, Andrus L, Kim D, Whittaker CA, Tan EK, Thiberge S, Fleming HE, Mancio-Silva L, et al. Autonomous circadian rhythms in the human hepatocyte regulate hepatic drug metabolism and inflammatory responses. *Sci Adv*. 2024;10(17):9281.
58. Mick E, Kamm J, Pisco AO, Ratnasiri K, Babik JM, Castañeda G, DeRisi JL, Detweiler AM, Hao SL, Kangelaris KN, et al. Upper airway gene expression reveals suppressed immune responses to SARS-CoV-2 compared with other respiratory viruses. *Nat Commun*. 2020;11(1):5854.
59. Ravindra NG, Alfajaro MM, Gasque V, Huston NC, Wan H, Szigeti-Buck K, Yasumoto Y, Greaney AM, Habet V, Chow RD, et al. Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium identifies target cells, alterations in gene expression, and cell state changes. *PLoS Biol*. 2021;19(3):3001143.
60. Martin-Sancho L, Lewinski MK, Pache L, Stoneham CA, Yin X, Becker ME, Pratt D, Churas C, Rosenthal SB, Liu S, et al. Functional landscape of SARS-CoV-2 cellular restriction. *Mol Cell*. 2021;81(12):2656–68.
61. Zhuang X, Tsukuda S, Wensch F, Wing PA, Schilling M, Harris JM, Borrmann H, Morgan SB, Cane JL, Mailly L, et al. The circadian clock component bmal1 regulates SARS-CoV-2 entry and replication in lung epithelial cells. *IScience*. 2021;24(10): 103144.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.