circafinder: Time-of-day classification using single cell transcriptomics

Lihan Zhong
The Rockefeller University
Laboratory of Single-Cell
Genomics and Population Dynamics
lzhong01@rockefeller.edu

Ahmet Doymaz
The Rockefeller University
Laboratory of Single-Cell
Genomics and Population Dynamics
adoymaz@rockefeller.edu

Abstract

Circadian rhythms are intrinsic 24-hour cycles governing physiological processes, with disruptions linked to sleep disorders and metabolic syndromes. Traditional bulk RNA sequencing obscures cell-type-specific circadian dynamics by averaging gene expression across entire cell populations. Single-cell RNA sequencing (scRNA-seq) provides the resolution necessary to explore these dynamics at an individual cell level. Our application, circafinder, employs an Artificial Neural Network (ANN) trained on publicly available datasets to predict the circadian phase of individual cells. This user-friendly Streamlit tool enables researchers and clinicians to investigate circadian regulation across various cell types, supporting targeted chronotherapeutic strategies.

1. Introduction

Motivation: Circadian rhythms are 24-hour cycles governing metabolic and gene regulatory mechanisms. Disruptions in these rhythms are linked to numerous health issues, including sleep disorders and metabolic syndromes. Traditional methods, such as bulk RNA sequencing, provide averaged gene expression profiles, obscuring cell-type-specific circadian dynamics. Single-cell RNA sequencing (scRNA-seq) offers the resolution needed to explore these dynamics at the individual cell level. Our application, *Circafinder*, predicts the circadian phase of individual cells using scRNA-seq data, providing insights into how aging and other factors influence circadian regulation at the cellular level.

We employed supervised machine learning methods, including Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs), to predict circadian phases from scRNA-seq data. SVMs classified cells into discrete circadian time intervals based on gene expression profiles. ANNs captured complex, non-linear relationships in gene expression data, proving particularly effective in addressing the high dimensionality and sparsity characteristic of scRNA-seq data.

Prior Efforts: Previous computational methods, such as CYCLOPS(2), ZeitZeiger(3), and Tempo(4), have been developed to infer circadian phases from transcriptomic data. CYCLOPS utilizes an autoencoder neural network to order samples based on circadian phase, while ZeitZeiger applies sparse principal component analysis for phase prediction. Tempo, a Bayesian variational inference approach, addresses challenges like data sparsity and estimation uncertainty in single-cell transcriptomics. However, these methods often lack the resolution to capture cell-type-specific circadian dynamics and do not account for variability introduced by different conditions. Our approach addressed these gaps by integrating supervised machine learning techniques capable of handling the complexities of scRNA-seq data.

Machine Learning Approach: Our pipeline included the following steps:

- **Data Exploration:** Assessed the quality and characteristics of scRNA-seq data using violin plots and UMAP visualizations.

- **Preprocessing:** Normalized data, handled missing values, and selected highly variable genes as relevant features.

- **Model Training:** Implemented and trained SVMs and ANNs to learn patterns in gene expression data.

- **Evaluation:** Assessed model performance using Mean Absolute Error (MAE), Circular Correlation Coefficient, and Prediction Interval Coverage Probability (PICP).

- **Deployment:** Developed a user-friendly Streamlit application allowing researchers and clinicians to explore the scRNA-seq data and obtain circadian phase predictions for individual cells and cell types.

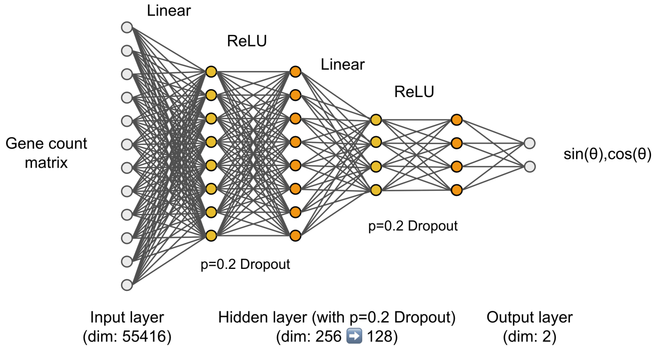


Figure 1: Illustration of machine learning pipeline

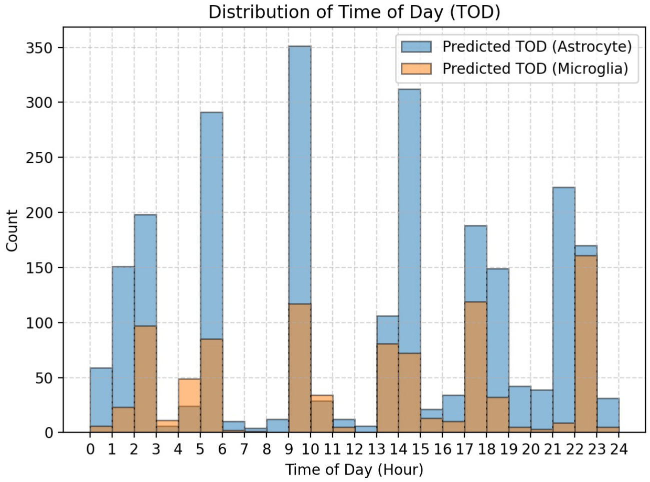


Figure 2: Cell counts across time of day, comparing ANN-predicted cell phases from Microglial cells and Astrocytes.

Impact: *Circafinder* enhances our understanding of how aging impacts circadian regulation across different cell types. By providing a robust framework for predicting circadian phases at the single-cell level, this project informs the development of targeted chronotherapeutic strategies aimed at restoring or maintaining healthy circadian rhythms, especially in aging-related vulnerable cell types. Ethical considerations, such as data privacy and informed consent, have been addressed in accordance with relevant guidelines and regulations.

2. Background

Project Scope: Circadian rhythms regulate numerous biological processes through intrinsic 24-hour cycles, influencing metabolism, immune function, and sleep-wake behavior. Disruptions in circadian rhythms have been linked to significant health concerns, such as metabolic disorders, neuro-degenerative diseases, and impaired immune responses. Traditional methods of studying these rhythms have relied largely on bulk RNA sequencing, which averages gene expression across cell populations, potentially obscuring cell-type-specific dynamics. Our project specifically addresses this limitation by developing machine learning models to accurately predict circadian phases at single-cell resolution using scRNA-seq data, with a particular focus on microglial and astrocyte cell populations.

Prior work: Several computational approaches have been previously established to study circadian rhythms using transcriptomic data. In particular, methods such as ZeitZeiger (Hughey et al., 2016), CYCLOPS (Anafi et al., 2017), and Tempo (Auerbach et al., 2022) have been widely used. CYCLOPS employs autoencoder neural networks to infer circadian phases from bulk RNA-seq data, whereas ZeitZeiger utilizes supervised dimensionality reduction and sparse principal component analysis for circadian phase prediction. Tempo addresses single-cell transcriptomic data challenges using Bayesian inference methods. Although these methods represent significant advances, they often lack the specificity required to capture nuanced circadian dynamics at the individual cell level across diverse cell types and conditions.

Common knowledge: A key challenge in the field of circadian biology remains the accurate prediction of circadian phase at single-cell resolution. There is ongoing debate about the extent to which circadian gene expression varies between individual cells within the same tissue, especially under varying physiological conditions such as aging or disease states. Recent single-cell studies, such as the Nature Neuroscience publication by Wen et al. (2020), have demonstrated significant variability in circadian gene expression across individual cells, highlighting the need for cell-specific predictive models. Our approach aims to resolve these disputes by implementing machine learning methodologies specifically designed to handle the inherent complexity and sparsity of single-cell RNA sequencing data, thus refining our understanding of circadian rhythms at a cellular level.

Knowledge Gaps: Despite advances in single cell computational methodologies and circadian prediction algorithms focused on bulk sequencing data, accurately determining the circadian phase of individual cells remains challenging, particularly in the context of aging. Existing methods often lack the resolution to capture the circadian dynamics specific to cell types or do not account for the variability introduced by aging processes. Our approach aims to fill these gaps by integrating supervised machine learning techniques capable of handling the complexities of scRNA-seq data.

3. End-to-End ML Pipeline

3.1. Offline Model Training and Evaluation

3.1.1 Data Collection, Exploration, and Processing

We employed the single-cell RNA sequencing dataset published by Wen et al. (2020) in Nature Neuroscience, obtained from the NIH Gene Expression Omnibus (GEO). This dataset includes numerical gene expression matrices from approximately 50,000 cells, each characterized by about 5,000 unique molecular identifiers (UMIs). The dataset, released in 2020, was specifically used to train predictive models for determining the circadian phase of microglial and astrocyte cell types based on gene expression.

Ground truth labels provided with the dataset include the annotated cell type and the exact time of day each cell was extracted. These labels served as benchmarks to evaluate the accuracy of our predictive models.

Data exploration included visualization techniques such as violin plots and UMAP scatterplots. Violin plots illustrated the distribution of UMIs and gene counts per cell, facilitating the exclusion of low-quality cells. UMAP scatterplots were employed to confirm clear clustering patterns by cell type, thereby validating the dataset's suitability for circadian phase prediction.

Preprocessing steps involved:

- Data cleaning by removing cells with unusually low or high gene counts.
- Normalization and scaling of gene expression data.
- Feature selection, focusing on highly variable genes.
- Dimensionality reduction using UMAP to address data sparsity and facilitate meaningful clustering.

3.1.2 Methods and Model Training

We implemented two machine learning algorithms: Support Vector Machines (SVM) and Artificial Neural Networks (ANN). Both algorithms were implemented from scratch without utilizing external machine learning libraries.

Support Vector Machines (SVM) solve classification problems by finding an optimal hyperplane that separates circadian phases into discrete time intervals. Given training vectors x_i and labels $y_i \in \{-1, 1\}$, the SVM optimization problem is formulated as:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y_i(w \cdot x_i + b) \geq 1, \quad \forall i \tag{1}$$

Artificial Neural Networks (ANN) model complex, nonlinear relationships between gene expression data and continuous circadian phases. ANN neuron activations are computed using the following equation:

$$a_j^{(l)} = \sigma\left(\sum_k w_{jk}^{(l)} a_k^{(l-1)} + b_j^{(l)}\right) \tag{2}$$

Here, $a_j^{(l)}$ represents the activation of neuron j in layer l, $w_{jk}^{(l)}$ and $b_j^{(l)}$ are weights and biases connecting neurons across layers, and σ is the Rectified Linear Unit (ReLU) activation function defined as:

$$\sigma(x) = \max(0, x)$$

The ANN model inputs are raw RNA count matrices from each cell, while the outputs represent predicted circadian phases encoded as sine and cosine components.

The ANN was ultimately selected over SVM due to its superior capability in modeling continuous, nonlinear circadian phase predictions at minute-level resolution. Although we initially employed an SVM due to its simplicity and reduced risk of overfitting, we found the ANN model more informative and better suited for predicting a continuous variable.

3.1.3 Model Evaluation

Model evaluation employed the following three metrics:

- **Mean Absolute Error (MAE)** quantifies the average magnitude of prediction errors:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3}$$

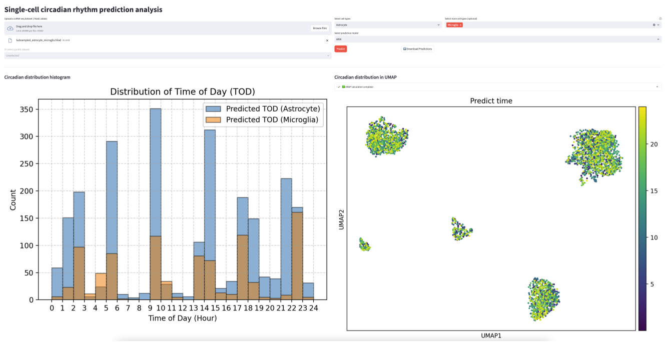
- **Circular Correlation Coefficient** assesses the cyclical correlation between predicted and observed circadian phases:

$$r = \frac{\sum_{i=1}^{n} \sin(y_i - \bar{y}) \sin(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n} \sin^2(y_i - \bar{y}) \sum_{i=1}^{n} \sin^2(\hat{y}_i - \bar{\hat{y}})}} \tag{4}$$

- **Prediction Interval Coverage Probability (PICP)** evaluates reliability by measuring the fraction of observed values within their predicted intervals:

$$\text{PICP} = \frac{1}{n}\sum_{i=1}^{n} c_i, \quad \text{where} \quad c_i = \begin{cases} 1, & y_i \in [L_i, U_i] \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

These metrics were selected to comprehensively assess prediction accuracy, robustness, and reliability for circadian phase data. Training involved an 80/20 randomized train-test split. Hyperparameter tuning was performed using grid search to optimize learning rate, batch size, and hidden layers. Dropout regularization and cross-validation across replicates were used to mitigate overfitting.



Single-cell circadian rhythm prediction analysis

Upload a scRNA-seq dataset (.h5ad/.adata): Subsampled_astrocyte_microglia.h5ad

Or select a public dataset: Unselected

Select cell types: Astrocyte

Select more cell types (optional): Microglia

Select prediction model: ANN

UMAP calculation complete!

Figure 3: Screenshot depicting actual UI deployment for cell types in Figure 2

3.2. Results

The ANN model demonstrated robust performance in predicting circadian phases for astrocytes, achieving a validation circular Mean Absolute Error (MAE) of 1.4335 hours at epoch 08 in Training cycle 01. For microglia, the ANN model yielded a higher validation circular MAE of 2.3850 hours at epoch 16 in Training cycle 03, indicating slightly reduced prediction accuracy compared to astrocytes.

The Support Vector Machine (SVM) algorithm performed poorly, achieving only median 52 percent accuracy in correctly classifying circadian phases in discrete-time bins for astrocytes, substantially underperforming relative to the ANN. Across evaluated metrics, the ANN consistently outperformed the SVM, reinforcing the suitability of the ANN for capturing the complex, continuous circadian phase relationships present in single-cell RNA sequencing data.

3.2.1 Model Deployment

The ANN model was chosen for deployment based on superior granularity (minute-level prediction) and balanced computational efficiency. Performance trade-offs were evaluated considering MAE, circular correlation, and PICP. Ultimately, the ANN was deployed on a GPU-supported high-performance computing cluster, maximizing prediction accuracy and computational performance.

4. Front-End (Streamlit)

Circafinder features a user-friendly web interface built with Streamlit, enabling users to upload scRNA-seq datasets, select cell types, and generate circadian phase predictions. Visualization tools, including downloadable UMAP plots of predicted phases, allowing for data exploration and to observe cell populations that exhibit phase differences.

5. Conclusion

The Circafinder project successfully developed an ANN-based machine learning framework capable of accurately predicting circadian phases at the single-cell level using scRNA-seq data. Our approach leveraged comprehensive data preprocessing and rigorous evaluation metrics, demonstrating strong predictive accuracy and reliability.

The implications of this work are significant, as an improved understanding of cell-specific circadian dynamics could lead to improved treatments for circadian rhythm disorders and related metabolic diseases, thus benefiting researchers, clinicians, and society at large.

Future plans include refining the ANN architecture further, incorporating larger and more diverse datasets to enhance generalizability, and extending the web-based user interface to include condition comparisons.

6. Team Member Contribution

6.1. Technical Components

Ahmet Doymaz - dataset cleanup, algorithm development, and comparison to other existing packages

Lihan Zhong - dataset cleanup and QC, algorithm development, UI design

6.2. Writing Components

Ahmet Doymaz - Background and research, Introduction, editing

Lihan Zhong - End to End ML pipeline, Risk Mitigation

7. Code Availability

Our package and test data can be found here: https://github.com/Cornell-Tech-PAML-Course-2025/circafinder

8. References

1. Takahashi, J. S., Hong, H.-K., Ko, C. H., McDearmon, E. L. (2008). The genetics of mammalian circadian order and disorder: Implications for physiology and disease. Nature Reviews Genetics, 9(10), 764–775. https://doi.org/10.1038/nrg2430

2. Anafi, R. C., Francey, L. J., Hogenesch, J. B., Kim, J. (2017). CYCLOPS reveals human transcriptional rhythms in health and disease. Proceedings of the National Academy of Sciences, 114(20), 5312–5317. https://doi.org/10.1073/pnas.1619320114

3. Hughey, J. J., Hastie, T., Butte, A. J. (2016). ZeitZeiger: Supervised learning for high-dimensional data from an oscillatory system. Nucleic Acids Research, 44(8), e80. https://doi.org/10.1093/nar/gkw030

4. Auerbach, B. J., FitzGerald, G. A., Li, M. (2022). Tempo: An unsupervised Bayesian algorithm for circadian

phase inference in single-cell transcriptomics. Nature Communications, 13(1), 6580. https://doi.org/10.1038/s41467-022-34287-0

5. Wen, S., Ma, D., Zhao, M., Xie, L., Wu, Q., Gou, L., Zhu, C., Fan, Y., Wang, H., Yan, J. (2020). Spatiotemporal single-cell analysis of gene expression in the mouse suprachiasmatic nucleus. Nature Neuroscience, 23(4), 456–467. https://doi.org/10.1038/s41593-020-0586-x

6. Droin, C., El Kholtei, J., Bahar Halpern, K., Hurni, C., Rozenberg, M., Muvkadi, S., Itzkovitz, S., Naef, F. (2021). Space-time logic of liver gene expression at sub-lobular scale. Nature Metabolism, 3(1), 43–58. https://doi.org/10.1038/s42255-020-00323-1