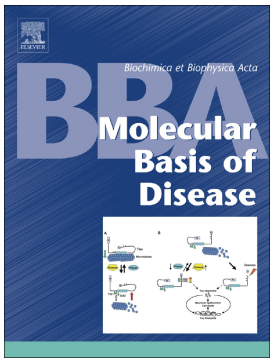
Accepted Manuscript



Identification of human circadian genes based on time course gene expression profiles by using a deep learning method

Peng Cui, Tingyan Zhong, Zhuo Wang, Tao Wang, Hongyu Zhao, Chenglin Liu, Hui Lu

PII: S0925-4439(17)30453-2
DOI: doi:10.1016/j.bbadis.2017.12.004
Reference: BBADIS 64980

To appear in:

Received date: 18 September 2017
Revised date: 28 November 2017
Accepted date: 3 December 2017

Please cite this article as: Peng Cui, Tingyan Zhong, Zhuo Wang, Tao Wang, Hongyu Zhao, Chenglin Liu, Hui Lu , Identification of human circadian genes based on time course gene expression profiles by using a deep learning method. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Bbadis(2017), doi:10.1016/j.bbadis.2017.12.004

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Identification of human circadian genes based on time course gene expression profiles by using a deep learning method

Peng Cui[1,2], Tingyan Zhong[1,2], Zhuo Wang[1], Tao Wang[1,2], Hongyu Zhao[2,3], Chenglin Liu[1,*], Hui Lu[1,2,*]

[1]School of Life Science and Biotechnology, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China

[2]SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China

[3]Department of Biostatistics, Yale University, New Heaven, USA

*Correspondence should be addressed to: tocliu@126.com or huilu@sjtu.edu.cn.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Abstract

Circadian genes express periodically in an approximate 24-hour period and the identification and study of these genes can provide deep understanding of the circadian control which plays significant roles in human health. Although many circadian gene identification algorithms have been developed, large numbers of false positives and low coverage are still major problems in this field. In this study we constructed a novel computational framework for circadian gene identification using deep neural networks (DNN) ─ a deep learning algorithm which can represent the raw form of data patterns without imposing assumptions on the expression distribution. Firstly, we transformed time-course gene expression data into categorical-state data to denote the changing trend of gene expression. Two distinct expression patterns emerged after clustering of the state data for circadian genes from our manually created learning dataset. DNN was then applied to discriminate the aperiodic genes and the two subtypes of periodic genes. In order to assess the performance of DNN, four commonly used machine learning methods including k-nearest neighbors, logistic regression, naïve Bayes, and support vector machines were used for comparison. The results show that the DNN model achieves the best balanced precision and recall. Next, we conducted large scale circadian gene detection using the trained DNN model for the remaining transcription profiles. Comparing with JTK_CYCLE and a study performed by Möller-Levet et al (*doi: 10.1073/pnas.1217154110*), we identified 1132 novel periodic genes. Through the functional analysis of these novel circadian genes, we found that the GTPase superfamily exhibits distinct circadian expression patterns and may provide a molecular switch of circadian control of the functioning of the immune system in human blood. Our study provides novel insights into both the circadian gene identification field and the study of complex circadian-driven biological control.

Keywords: circadian genes, deep learning, deep neural network, classification, functional analysis, GTPase

Abbreviations

DNN, deep neural networks; k-NN, k-nearest neighbors; SVM, support vector machine; ROC curve, receiver operating characteristic curve; AUC, area under the curve; ReLU, rectified linear unit; GO, gene ontology

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

1. Introduction

Circadian rhythms, which control a wide range of important biological behaviors in a twenty-four-hour periodical manner, have been observed in many organisms including human, animals, plants, and prokaryotes [1], [2]. This circadian oscillator helps organisms better adapt to predictable environmental changes by regulating thousands of periodically expressed genes with various distinct biological functions [3], [4]. At the molecular level, the circadian core clock is driven by a genetically encoded transcriptional feedback loop involving about a dozen genes, such as CLOCK, BMAL1, PER and CRY [5]. This core clock locates in the brain's suprachiasmatic nucleus (SCN) and interacts with subsidiary clocks from other tissues [6]. In humans, the circadian clock exhibits close relationships with nutrition [7], metabolism [8], [9], and the functioning of the immune system [10], [11]. Furthermore, the disruption of the circadian system is implicated in a number of health problems [12] ranging from depression, cancer, obesity, diabetes, and ageing [13]–[17]. Nowadays, chronotherapy which refers to the use of circadian rhythms to maximize effectiveness and minimize side effects of the therapy, is becoming increasingly popular in various clinical fields [18]. Francis Levi reported that chronotherapy achieved significant long-term survival for colorectal cancer patients [19]. Hermida et al. investigated the circadian variation of blood pressure which could develop potential chronotherapy of hypertension [20]. And Zhang R et al. also discovered that the majority of best-selling drugs target the products of circadian genes [21]. In consideration of the significant importance of biological circadian control, identification and study of such periodically regulated genes can help gain deeper understanding of circadian rhythm mechanisms and drive advancement towards precision circadian medicine [22].

The first large scale identification of periodically expressed genes was conducted in Arabidopsis in 2000 by Harmer et al. using microarray technology [23]. This work analyzed the transcripts in the time domain and compared the expression pattern of each transcript to a cosine function. Over the years, a dozen algorithms have been proposed in the detection of circadian genes and such algorithms can roughly be grouped into two categories: spectral analysis and sinusoidal fits [24]. The spectral analysis method first determines the spectral composition of each transcript using Fourier transform, and then identifies as circadian such transcripts which have a strong spectrum signal corresponding to a period of about 24 hours. Spectral based methods, such as Fisher's G-test (GeneTS) [25] have proved to be effective. However, these methods are better suited for higher sampling frequency datasets and can only be used in evenly spaced data [26]. The sinusoidal fits methods compare the correlation between every transcript and sinusoidal curves predefined by the user by using a linear regression. The representative method of this category is COSOPT (COS) [26]. COS measures the goodness of fit between expression data and cosine curves, and has been applied to gene expression profiles in mammals and Drosophila. These algorithms greatly advance towards the goal of identifying periodically expressed transcripts.

Currently, JTK_CYCLE (JTK) [27], which is a non-parametric algorithm using

ACCEPTED MANUSCRIPT

Jonckheere-Terpstra and Kendall's tau tests, has been widely adopted in this field. Recently, Wu et al. developed MetaCycle as JTK successor [28]. Although the JTK algorithm has made great improvements, certain challenges still exist. We must note that the critical assumption of sinusoidal expression pattern limits the detection ability because of the existence of many circadian genes with asymmetry or saw tooth-shaped rhythms [29]. For circadian experiments with replicates, MetaCycle analyzes data for every individual and then integrates the results from multiple individuals. This strategy cannot take advantage of replicated experiments because every individual is affected by the noise of microarray technology, so the integrated results may be worse than the true picture. Based on the above considerations, we suggest that JTK_CYCLE may potentially miss a portion of true circadian genes.

Nowadays, deep learning methods have advanced rapidly and become the start-of-the-art algorithms in many fields, ranging from image classification, speech recognition, and bioinformatics [30], [31]. In contrast to frequently-used machine learning methods which need hand-designed features, deep learning can automatically extract features from the raw data and learn complex features directly by combining simpler features through deep multi-layer architectures [32]. The standard deep learning architectures can be categorized into four groups including deep neural networks (DNN) [33], [34], convolutional neural networks (CNN) [35], recurrent neural networks (RNN) [36], and emergent architectures [37]. Different architectures can be used in solving diverse problems. In the bioinformatics field, various deep learning algorithms have been applied to answer significant biological questions. Liang et al. employed deep belief network to enhance cancer patients clustering [38]. Xiong et al. applied convolutional neural network to predict defects in RNA splicing and discovered novel clinical variants [39]. Xu et al conducted drug-induced liver injury prediction analysis using recursive neural network [40]. Deep learning methods have shown powerful performance in these studies, and more works using deep learning will emerge in the biological research field.

In this study, we developed a novel machine learning pipeline for the detection of periodically expressed genes in human blood transcriptome based on deep neural networks (DNN). Benefiting from the deep learning algorithm, the DNN model can capture the raw form of expression patterns regardless of the uneven sampling issues and the critical assumption of sinusoidal expression distribution. The time-course gene expression data was obtained from Möller-Levet et al. [41], which included human whole blood transcriptome. In general, deep learning methods require learning data to generate a data representation. However, due to the absence of ground truth labels, we manually confirmed the expression patterns for 2000 genes, and created a high-quality subset of periodic/aperiodic labels including 270 positive cases and 1730 negative cases. Next, in order to discover the underlying distinct expression patterns of the generated circadian genes, we first transformed the raw expression data to (-1,0,1) states denoting the decreasing, stable, and increasing expression changes, and then conducted a heatmap clustering using the transformed data. We obtained two distinct clusters. After completing the benchmark process, a DNN model was constructed based on the time-course

ACCEPTED MANUSCRIPT

expression data to distinguish the periodically expressed genes with two distinct subtypes from the negative cases. Compared with four conventional machine learning methods, the DNNs model achieved the best performance. Finally, we conducted large scale circadian gene prediction and found 1132 novel circadian genes compared to genes found by JTK_CYCLE and the original study by Möller-Levet et al. The functional analysis of these novel circadian genes reveal that the GTPase superfamily exhibits distinct circadian expression patterns and may regulate the circadian control of immune system in human blood.

2. Material and methods

2.1. Datasets

Time series gene expression data of human blood transcriptome was downloaded from NCBI Gene Expression Omnibus (GEO) through accession number GSE39445 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39445 [41]. The data included insufficient sleep group and a corresponding control group. Because the purpose of our study was to construct a more reliable circadian gene identification algorithm, for the sake of simplification, we only adopted the corresponding control dataset for our analysis. The control group dataset contains the expression profiles of 18541 genes at 10 time points (from 7.5 hours after the experimental baseline to 34.5 hours) with 3-hour intervals each, for 24 subjects. The time of melatonin secretion, which is between about 15-25 hours after the baseline, is considered the biological night phase. In order to construct a high quality periodic/aperiodic label set, we manually confirmed the expression patterns for 2000 genes according to their expression distributions which resulted in labeling of 270 circadian genes and 1730 negative cases. The distributions of some non-circadian/circadian genes are presented in Supplementary Material S1. The other 16541 genes were left to be identified by the learned model.

2.2. Data transformation, clustering, and learning dataset construction

Firstly the expression data was transformed into (-1, 0, 1) categorical state data. The (-1, 0, 1) labels denoted respectively the decrease, stability, and increase of expression trends compared to that of previous time point. Then a heatmap from hierarchical clustering based on Ward's method was created using the transformed data for 270 circadian genes. This resulted in two distinct expression patterns. We also displayed the average expression tendency for these two clusters of genes using the averaged cumulative summation of state data for these genes. Then a learning dataset was constructed based on the 270 circadian genes and 1730 negative ones. The 270 circadian genes were labeled as class 1 or class 2, and the negative 1730 cases were labeled as class 0. Finally, machine learning algorithms were applied to classify periodic/aperiodic genes (including the subclasses of periodic genes) using the time series expression profiles of all subjects.

2.3. Deep neural network construction

ACCEPTED MANUSCRIPT

A traditional neural network depends on shallow nets which are composed of one input layer, one output layer, and at most one hidden layer in between. However, DNN are distinguished from the common single-hidden-layer neural networks by the depth. Neural networks with more than three layers (including input and output) can be viewed as "deep" neural networks. In DNN, each layer of nodes trains on a different set of features based on the previous layer's output, and the latter layer can recognize more complex features because of the aggregation and recombination from the previous layer. To classify signals as periodic (including class 1 and class 2) or aperiodic, we trained a multi-classes DNN which can automatically perform feature engineering using Tensorflow [42]. For each gene, the input of the DNN is the time series expression profiles of all 24 subjects and the output is the particular class (0, 1, or 2) of this gene. We experimented with many hyper-parameters, especially the number of hidden layers, until no significant improvements could be made. We found the structure of a DNN with 10 hidden layers of size 100 nodes achieved better performance. The rectified linear unit (ReLU) was used for activation function. The learning rate was set at 0.0001 using an Adam gradient-based optimizer [43]. To avoid an overfitting problem, we also adopted a dropout strategy [44]. Specifically, during the training process, the DNN randomly dropped 5% of units along with their connections. This strategy reduced overfitting and made improvements over other regularization methods. The structure of our DNN and hyper-parameters were consistent across all our analysis.

2.4. Performance comparison between DNN and other methods

In order to evaluate the power of our DNN for detecting periodically expressed genes, we also trained four frequently-used machine learning models, namely, k-nearest neighbors (k-NN, k=3 in our analysis), logistic regression, naïve Bayes, and support vector machines (SVM). K-NN is a non-parametric method which classifies the objects by the majority votes of its neighbors [45]. Logistic regression is a regression model determining the relationship between the categorical responses and multiple predictors by the logistic function [46]. Naïve Bayes classifier is a probabilistic classifier taking advantage of Bayes's theorem to predict the category of a sample [47]. SVM is a discriminative classifier defined by a large-margin separating hyper-plane [48]. We implemented these machine learning models using Python (Scikit-learn).

2.5. Evaluation metrics

To compare the performance of the developed DNN with the four machine learning methods, we performed ten-fold cross-validation and calculated a ROC curve. Furthermore, four additional metrics were also used for comparison: accuracy score, precision, recall, and F1 score. In classification problems, accuracy score is the ratio of true instances (both true positives and true negatives) among the total number of instances examined. However, it cannot reflect the underlying class distribution. Precision is the ratio of true instances from the retrieved instances, while recall is the ratio of true

ACCEPTED MANUSCRIPT

instances that are successfully retrieved. The F1 score combines both precision and recall, and often acts as a balanced metric. We implemented all these metrics which can give a more complementary comparison. The definition of these four metrics are presented in the following formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP}, \quad (2)$$

$$Recall = \frac{TP}{TP + FN}, \quad (3)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (4)$$

Here, TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative.

2.6. Circadian gene detection with DNN and functional analysis

Utilizing the trained DNN model, the remaining 16541 genes were predicted as periodic or aperiodic transcripts (including the subclasses of periodic genes). We also compared the detection results of our DNN with detection by JTK_CYCLE and the detection results from the Möller-Levet et al study. Regarding JTK_CYCLE, the algorithm was implemented using MetaCycle (p-value < 0.1 as circadian genes) [28], and regarding the Möller-Levet et al study (which also adopted a sinusoidal fit method) - we directly utilized their identification results [41]. Next, in order to figure out the potential distinct biological functions of the novel circadian genes only identified by our DNN, we conducted gene ontology (GO) and pathway enrichment analysis using clusterProfiler [49]. The Benjamini and Hochberg corrected p-values which were less than 0.05 are selected as enrichment.

3. Results

3.1. Transformed data reveals two distinct clusters of expression patterns among circadian genes

The goal of our study was to uncover the potential distinct changing genomic expression patterns. Thus as a first step, expression profiles were transformed into (-1, 0, 1) states reflecting a decreasing, stable, and increasing trend in comparison to a previous time point. From the heatmap in Figure 1, derived from the two-side clustering of (-1, 0, 1)-transformed data of manually identified 270 circadian genes, we observed that the periodically expressed genes are roughly grouped into two clusters – cluster 1 and cluster 2. 155 genes are classified into cluster 1 and the other 115 genes are in cluster 2. Furthermore, the 10 time points are also categorized into 2 types, suggesting that these two clusters of genes tend to actively express in a distinct time domain. Next, to elucidate

ACCEPTED MANUSCRIPT

the expression patterns of these two distinct clusters of genes across time, we displayed the average expression tendency for these genes by using the averaged cumulative summation of state data for each class. We can see that genes in cluster 1 are inclined to actively express in the night (about 15~25 hours after the baseline) (Figure 2). On the contrary, genes in cluster 2 exhibit the exact opposite pattern whose expression trend declines directly after melatonin secretion (about 15 hours after the baseline). In contrast, only small fluctuations could be observed across time in the average expression pattern of non-circadian genes (class 0) (Figure 2).

3.2. Circadian gene classification with DNN and the four machine learning methods

To demonstrate the classification accuracy of DNN, we compared our method with k-NN (k=3), logistic regression, naïve Bayes, and support vector machine (SVM) classifiers. Specifically, we first performed periodic/aperiodic gene classification using these five methods based on the created learning dataset. Ten-fold cross-validation was performed and the corresponding averaged ROC curve was displayed for comparison. As shown in Figure 3, the DNN model achieves the best prediction accuracy with the highest AUC score of 0.92. Logistic regression and SVM also get good performance with AUC scores of 0.81. The prediction accuracy of K-NN and naïve Bayes is not sufficient, with corresponding AUC scores of 0.74 and 0.73, respectively.

Next, as the periodic genes can be classified into 2 distinct clusters, we created multi-classes classifiers using these five methods and the performance was evaluated based on accuracy, precision, recall, and F1 score. Among these evaluation metrics, F1 score is the most desirable index which incorporates both precision and recall levels. From Figure 4, we can see that DNN, k-NN, and logistic regression achieve the maximal accuracy score of about 0.9. For precision and recall, DNN and logistic regression get better precision, while SVM and DNN achieve better recall compared with other methods. The performance of naïve Bayes was the worst among all these four evaluation metrics. By considering both precision and recall, the DNN model showed the maximal F1 score of about 0.8. In summary, our method achieved the best balanced performance on circadian genes prediction.

3.3. Large scale circadian gene detection using the learned DNN model

The above described analysis showed that the DNN model preforms better than its competitors in the current application. Now, based on the trained DNN model, we directly predicted circadian genes from the remaining 16541 genes. For comparison, we also inferred circadian genes using JTK_CYCLE and the method adopted in the Möller-Levet et al study. The numbers and the overlap of identified periodically-expressed genes by the three methods are shown in Figure 5. The DNN model identified a maximum of 2007 circadian genes, including 873 genes from class 1 and the other 1134 genes from class 2. In contrast, JTK_CYCLE and the Möller-Levet et al study identified 921 and 1167 genes

ACCEPTED MANUSCRIPT

as circadian respectively. For the overlap, 394 genes were identified as periodically expressed genes by the three methods, and 1007 genes were identified by at least two methods. For the identified orphan genes, 1132 genes were identified as circadian only by the DNN model. The detailed results for all these three methods are listed in the Supplementary Material S2.

Next, to see whether the identified orphan genes of DNN are true circadian genes or false positives, we manually checked the expression distributions of genes whose p-values given by JTK_CYCLE are greater than 0.5. We found that most of them represent distinct circadian expression patterns. As shown in Figure 6, the expression distributions of IGF1R and KIAA1211L are skewed sinusoidal waves. CWC27 and DNAJC2 exhibit triangle expression shapes and the distributions of SNHG15 and NOSIP are similar to sharp peaks. Many of these shapes are non-regular sinusoidal waves. We also found that for a particular gene, although it could not represent circadian expression distribution for only one subject because of the experimental noise inherent to microarray technology, the integrated distribution for all 24 subjects shows obvious circadian patterns. These results suggest that the conventional circadian gene identification methods might have missed many circadian genes due to their critical sinusoidal assumptions and less powerful analytical strategy for replicate microarray samples. The complete 1132 novel circadian genes identified only by DNN are listed in the Supplementary Material S2.

3.4. Functional analysis of the detected novel circadian genes

There are two classes of circadian genes with distinct expression patterns identified by our analysis. For genes in class 1, the expression levels are continuously increasing after melatonin secretion (about 24:00) until midday of the next day (Figure 2) which suggest that these genes may play important regulation roles during the night. On the other hand, genes in class 2 are expressed in an opposite fashion suggesting a day regulation roles for these genes. In order to find out the potential distinct biological functions of these two classes of genes, we conducted gene ontology and pathway enrichment analysis for the novel circadian genes detected only by the DNN model.

From Figure 7, we can see that genes in class 1 are associated with kinase activities, phosphorylation regulation, and vacuolar transport. All these activities are associated with the functioning of the immune system which has been confirmed as significant circadian function in the circulation during the resting phase [50], [51]. These activities tend to occur in the membrane region because many immune functions need vacuolar transport which happens in the membrane such as endocytosis of some immune complexes. Genes in class 2 conduct quite different functions - mainly including ribosome biogenesis, rRNA, ncRNA, protein localization, and oxidation-reduction reactions. The cellular components where the resulting effect of these genes tends to occur are mainly in the cytosol and the ribosomes. These functions are consistent with previous studies that circadian genes which are active in the daytime are associated with nutrition and energy metabolism [52], [53]. The gene ontology analysis also suggests that our detected novel circadian genes are

ACCEPTED MANUSCRIPT

true positives. The detailed GO enrichment results can be found in Supplementary Material S3.

In addition to already known circadian functions, we also found that the GTPase superfamily exhibits distinct circadian expression distributions. As shown in Figure 7, the molecular functions of circadian genes in class 1 are associated with GTPase binding, small GTPase binding, GTPase regulator activity, Ras GTPase binding etc. Although some studies argued that GTPases have important regulation roles for the circadian clock, such as the small GTPase LIP1 in Arabidopsis and RHO1 GTPase in Drosophila, very few studies have reported that they are associated with circadian rhythms in human blood [54], [55]. We know that genes in class 1 are closely related with the functioning of the immune system, so there is an interesting question whether GTPases have significant circadian regulation roles related to immune system activities. In previous studies such as, Johnson et al, it has been argued that small GTPases especially Ras superfamily exert significant impacts in immunity and inflammation [56]. Bokoch has reported that Rho GTPases have essential roles in regulating particular aspects of immune cell differentiation [57] and Saoudi et al demonstrated that Rho-GTPases act as key regulators of T lymphocyte development [58]. In addition, GTPases often function as molecular switches in significant number of pathways such as cytokinesis, apoptosis and cell motility [59]. Integrating all the above arguments, we hypothesize that GTPase superfamily may provide molecular switching of the circadian clock of the immune system in human blood.

For the enriched pathways, as shown in Table 1 and Table2, genes in class 1 are associated with B cell and T cell receptor signaling pathway, EGFR tyrosine kinase inhibitor resistance, autophagy, ErbB signaling pathway, and some carcinomas such as acute myeloid leukemia and glioma. All these pathways are related to immune system activity which is a finding consistent with previous studies. On the contrary, genes in class 2 are mainly involved in the ribosome, spliceosome, oxidative phosphorylation, protein and RNA transport. Interestingly, these genes are also enriched in three diseases including Alzheimer, Parkinson and Huntington's disease. Previous studies have shown that circadian dysfunction has significant impacts in neurodegenerative disorders [60]–[62]. Our findings provide novel insights about the close relationship between neurodegeneration and circadian rhythms. Detailed information about the pathway analysis can be found in Supplementary Material S4.

4. Discussion

In this work, we implemented a computational framework using deep learning for the identification of periodically expressed genes. Based on a manually curated learning dataset, circadian genes were grouped into two distinct classes according to expression patterns. For the purposes of selecting the best performing machine learning method, we compared the classification performance of DNN with four frequently-used methods: k-NN, logistic

ACCEPTED MANUSCRIPT

regression, naïve Bayes, and support vector machines. Five evaluation metrics including ROC curves, accuracy score, precision, recall, and F1 score were selected for comparison. The results show that the DNN model outperforms all the other methods and achieves the best balanced precision and recall.

Next we detected periodically expressed transcripts from the remaining 16541 genes using the learned DNN model. For comparison, we also detected circadian genes using JTK_CYCLE and the method used in the Möller-Levet et al study. The detection results show that the DNN model identified 2007 circadian genes and it showed the most complete coverage. These results reveal that the strong sinusoidal assumptions and improper analytical strategy for replicating experimental samples used by the other methods are possible reasons for the missing of many non-particular sinusoidal expressed circadian genes and limit the detection ability [28]. The DNN method transcends these limitations regardless of any prior assumptions and achieves the best complete performance. We also manually examined some orphan genes that were only identified by the DNN model and found that most of them have distinct circadian expression patterns such as skewed peaks or triangles. Although the DNN model achieves the best performance, we must note that preprocessing the learning dataset (the 'seed' circadian genes and negative ones) maybe the biggest limitations because of the potential requirement of large upfront human effort. As an extension, future work could design more elegant algorithms which can retain the coverage of identified circadian genes with different expression distributions and avoid manual operations.

For the identified 1132 novel circadian genes, different expression patterns may impact and guide distinct biological functions, so we further performed gene ontology and pathway enrichment analyses for the two identified classes of genes. We found that genes in class 1 which are actively expressed in the night are associated with kinase activity, phosphorylation regulation, and vacuolar transport. Notable here is that these functions are related to the functioning of the immune system. This finding is consistent with existing hypotheses that frequent late sleep may disrupt the circadian system and reduce immunity. Detailed discussion about the circadian control of the immune system can be found in the Scheiermann et al review [50]. Genes in class 2 continuously express in the daytime, conducting functions related to ribosome biogenesis, rRNA, ncRNA, protein localization, and oxidation-reduction reactions. This finding is not surprising due to the fact that humans perform many energy-consuming behaviors and consumes food in the daytime which precipitates the increased expression of genes involved in the synthesis and degradation of macromolecules [63].

When studying the close relationships between circadian rhythms and the immune system, one important question to address is regarding the source of the circadian immune regulation. In this study, we found that the GTPase superfamily which often acts as molecular switch for many important biological functions exhibits distinct circadian distribution. In addition, many previous studies argued that the GTPase superfamily exerts significant regulation roles pertaining to the immune system including immune cell differentiation, T lymphocyte development and cell migration [56], [58]. Integrating this information, we hypothesize that

ACCEPTED MANUSCRIPT

the GTPase superfamily may control the circadian regulation of the immune system in human blood. Future work could explore the in-depth relationships between the circadian rhythms, the GTPase superfamily, and the functioning and regulation of the immune system.

In conclusion, the deep learning method shows excellent results in the circadian gene identification problem and our results and functional analysis provide novel insights into the complex circadian regulation of the human blood transcriptome.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

We are grateful to Dr. Jing Li and Dr. Georgi Genchev at Shanghai Jiao Tong University for helpful discussions. This work was supported by Center for HPC, Shanghai Jiao Tong University and National Natural Science Foundation of China (11601326).

References

[1] J. C. Dunlap, "Molecular Bases for Circadian Clocks," *Cell*, vol. 96, no. 2, pp. 271–290, Jan. 1999.

[2] S. Panda *et al.*, "Coordinated Transcription of Key Pathways in the Mouse by the Circadian Clock," *Cell*, vol. 109, no. 3, pp. 307–320, May 2002.

[3] E. Boikoglou, Z. Ma, M. von Korff, A. M. Davis, F. Nagy, and S. J. Davis, "Environmental memory from a circadian oscillator: the Arabidopsis thaliana clock differentially integrates perception of photic vs. thermal entrainment.," *Genetics*, vol. 189, no. 2, pp. 655–64, Oct. 2011.

[4] C. Choi and M. N. Nitabach, "Circadian Biology: Environmental Regulation of a Multi-Oscillator Network," *Curr. Biol.*, vol. 20, no. 7, pp. R322–R324, Apr. 2010.

[5] H. A. Duong *et al.*, "A Molecular Mechanism for Circadian Clock Negative Feedback," *Science (80-.).*, vol. 332, no. 6036, pp. 1436–1439, Jun. 2011.

[6] C. Dibner, U. Schibler, and U. Albrecht, "The Mammalian Circadian Timing System: Organization and Coordination of Central and Peripheral Clocks," *Annu. Rev. Physiol.*, vol. 72, no. 1, pp. 517–549, Mar. 2010.

[7] S. Aoyama and S. Shibata, "The Role of Circadian Rhythms in Muscular and Osseous Physiology and Their Regulation by Nutrition and Exercise.," *Front. Neurosci.*, vol. 11, p. 63, 2017.

[8] G. Asher and P. Sassone-Corsi, "Time for Food: The Intimate Interplay between Nutrition, Metabolism, and the Circadian Clock," *Cell*, vol. 161, no. 1, pp. 84–92, Mar. 2015.

[9] J. D. Johnston, "Physiological links between circadian rhythms, metabolism and nutrition," *Exp. Physiol.*, vol. 99, no. 9, pp. 1133–1137, Sep. 2014.

[10] K. Adams, O. Castanon-Cervantes, and A. J. Davidson, "Circadian Rhythms in Immune Function," in *Circadian Medicine*, Hoboken, NJ: John Wiley & Sons, Inc, 2015, pp. 211–219.

ACCEPTED MANUSCRIPT

[11] C. Scheiermann, Y. Kunisaki, and P. S. Frenette, "Circadian control of the immune system.," *Nat. Rev. Immunol.*, vol. 13, no. 3, pp. 190–8, Mar. 2013.

[12] C. Feillet, G. T. J. van der Horst, F. Levi, D. A. Rand, and F. Delaunay, "Coupling between the Circadian Clock and Cell Cycle Oscillators: Implication for Healthy Cells and Malignant Growth.," *Front. Neurol.*, vol. 6, p. 96, 2015.

[13] J. Mattis and A. Sehgal, "Circadian Rhythms, Sleep, and Disorders of Aging," *Trends Endocrinol. Metab.*, vol. 27, no. 4, pp. 192–203, Apr. 2016.

[14] D. Landgraf *et al.*, "The role of disturbed circadian clocks in the development of depression-like behavior and metabolic comorbidity in mice," *Eur. Psychiatry*, vol. 41, p. S531, Apr. 2017.

[15] O. Pluquet, N. Dejeans, and E. Chevet, "Watching the clock: endoplasmic reticulum-mediated control of circadian rhythms in cancer," *Ann. Med.*, vol. 46, no. 4, pp. 233–243, Jun. 2014.

[16] J. L. Broussard and E. Van Cauter, "Disturbances of sleep and circadian rhythms," *Curr. Opin. Endocrinol. Diabetes Obes.*, vol. 23, no. 5, pp. 353–359, Oct. 2016.

[17] K. Rakshit, A. P. Thomas, and A. V. Matveyenko, "Does Disruption of Circadian Rhythms Contribute to Beta-Cell Failure in Type 2 Diabetes?," *Curr. Diab. Rep.*, vol. 14, no. 4, p. 474, Apr. 2014.

[18] J. Sajan, T. Cinu, A. Chacko, J. Litty, and T. Jaseeda, "Chronotherapeutics and Chronotherapeutic Drug Delivery Systems," *Trop. J. Pharm. Res.*, vol. 8, no. 5, Nov. 2009.

[19] "Circadian chronotherapy for human cancers," *Lancet Oncol.*, vol. 2, no. 5, pp. 307–315, May 2001.

[20] "Circadian variation of blood pressure: The basis for the chronotherapy of hypertension," *Adv. Drug Deliv. Rev.*, vol. 59, no. 9–10, pp. 904–922, Aug. 2007.

[21] R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch, "A circadian gene expression atlas in mammals: Implications for biology and medicine," *Proc. Natl. Acad. Sci.*, vol. 111, no. 45, pp. 16219–16224, Nov. 2014.

[22] S. Dulong, A. Ballesta, A. Okyar, and F. Lévi, "Identification of Circadian Determinants of Cancer Chronotherapy through In Vitro Chronopharmacology and Mathematical Modeling," *Mol. Cancer Ther.*, vol. 14, no. 9, 2015.

[23] S. L. Harmer *et al.*, "Orchestrated Transcription of Key Pathways in Arabidopsis by the Circadian Clock," *Science (80-.).*, vol. 290, no. 5499, 2000.

[24] C. J. Doherty and S. A. Kay, "Circadian control of global gene expression patterns.," *Annu. Rev. Genet.*, vol. 44, pp. 419–44, 2010.

[25] S. Wichert, K. Fokianos, and K. Strimmer, "Identifying periodically expressed transcripts in microarray time series data.," *Bioinformatics*, vol. 20, no. 1, pp. 5–20, Jan. 2004.

[26] M. Straume, "DNA Microarray Time Series Analysis: Automated Statistical Assessment of Circadian Rhythms in Gene Expression Patterning," in *Methods in enzymology*, vol. 383, 2004, pp. 149–166.

[27] M. E. Hughes, J. B. Hogenesch, and K. Kornacker, "JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets.," *J. Biol. Rhythms*, vol. 25, no. 5, pp. 372–80, Oct. 2010.

ACCEPTED MANUSCRIPT

[28] G. Wu, R. C. Anafi, M. E. Hughes, K. Kornacker, and J. B. Hogenesch, "MetaCycle: an integrated R package to evaluate periodicity in large scale data," *Bioinformatics*, vol. 32, no. 21, pp. 3351–3353, Nov. 2016.

[29] P. F. Thaben and P. O. Westermark, "Detecting Rhythms in Time Series with RAIN," *J. Biol. Rhythms*, vol. 29, no. 6, pp. 391–400, Dec. 2014.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning.," *Nature*, vol. 521, no. 7553, pp. 436–44, May 2015.

[31] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov, "Applications of Deep Learning in Biomedicine," *Mol. Pharm.*, vol. 13, no. 5, pp. 1445–1454, May 2016.

[32] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics.," *Brief. Bioinform.*, p. bbw068, Jul. 2016.

[33] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[34] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemom. Intell. Lab. Syst.*, vol. 39, no. 1, pp. 43–62, Nov. 1997.

[35] S. Lawrence, C. L. Giles, A. C. Ah Chung Tsoi, and A. D. Back, "Face recognition: a convolutional neural-network approach," *IEEE Trans. Neural Networks*, vol. 8, no. 1, pp. 98–113, 1997.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory.," *Neural Comput.*, vol. 9, no. 8, pp. 1735–80, Nov. 1997.

[37] R. Hadsell *et al.*, "Learning long-range vision for autonomous off-road driving," *J. F. Robot.*, vol. 26, no. 2, pp. 120–144, Feb. 2009.

[38] M. Liang, Z. Li, T. Chen, and J. Zeng, "Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 12, no. 4, pp. 928–937, Jul. 2015.

[39] H. Y. Xiong *et al.*, "The human splicing code reveals new insights into the genetic determinants of disease," *Science (80-.).*, vol. 347, no. 6218, 2015.

[40] Y. Xu, Z. Dai, F. Chen, S. Gao, J. Pei, and L. Lai, "Deep Learning for Drug-Induced Liver Injury," *J. Chem. Inf. Model.*, vol. 55, no. 10, pp. 2085–2093, Oct. 2015.

[41] C. S. Möller-Levet *et al.*, "Effects of insufficient sleep on circadian rhythmicity and expression amplitude of the human blood transcriptome.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 12, pp. E1132–41, Mar. 2013.

[42] L. Rampasek and A. Goldenberg, "TensorFlow: Biology's Gateway to Deep Learning?," *Cell Syst.*, vol. 2, no. 1, pp. 12–14, Jan. 2016.

[43] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Dec. 2014.

[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[45] L. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[46] S. H. Walker and D. B. Duncan, "Estimation of the Probability of an Event as a Function of Several Independent Variables," *Biometrika*, vol. 54, no. 1/2, p. 167, Jun. 1967.

[47] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," Springer, Berlin, Heidelberg, 1998, pp. 4–15.

ACCEPTED MANUSCRIPT

[48] V. Jeyakumar, G. Li, and S. Suthaharan, "Support vector machine classifiers with uncertain knowledge sets via robust optimization," *Optimization*, vol. 63, no. 7, pp. 1099–1116, Jul. 2014.

[49] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters," *Omi. A J. Integr. Biol.*, vol. 16, no. 5, pp. 284–287, May 2012.

[50] C. Scheiermann, Y. Kunisaki, and P. S. Frenette, "Circadian control of the immune system.," *Nat. Rev. Immunol.*, vol. 13, no. 3, pp. 190–8, Mar. 2013.

[51] A. M. Curtis, M. M. Bellet, P. Sassone-Corsi, and L. A. J. O'Neill, "Circadian Clock Proteins and Immunity," *Immunity*, vol. 40, no. 2, pp. 178–186, Feb. 2014.

[52] G. Asher and P. Sassone-Corsi, "Time for Food: The Intimate Interplay between Nutrition, Metabolism, and the Circadian Clock," *Cell*, vol. 161, no. 1, pp. 84–92, Mar. 2015.

[53] G. D. M. Potter, J. E. Cade, P. J. Grant, and L. J. Hardie, "Nutrition and the circadian system.," *Br. J. Nutr.*, vol. 116, no. 3, pp. 434–42, Aug. 2016.

[54] É. Kevei *et al.*, "Arabidopsis thaliana Circadian Clock Is Regulated by the Small GTPase LIP1," *Curr. Biol.*, vol. 17, no. 17, pp. 1456–1464, Sep. 2007.

[55] N. V. Rao, "Role of the Rh01 GTPase signaling pathway in regulating the circadian clock in Drosophila melanogaster," University of Virginia, Charlottesville, VA, 2013.

[56] D. S. Johnson and Y. H. Chen, "Ras family of small GTPases in immunity and inflammation.," *Curr. Opin. Pharmacol.*, vol. 12, no. 4, pp. 458–63, Aug. 2012.

[57] G. M. Bokoch, "Regulation of innate immunity by Rho GTPases," *Trends Cell Biol.*, vol. 15, no. 3, pp. 163–171, Mar. 2005.

[58] A. Saoudi, S. Kassem, A. Dejean, and G. Gaud, "Rho-GTPases as key regulators of T lymphocyte biology.," *Small GTPases*, vol. 5, 2014.

[59] A. Delprato, D. Galas, T. Dewey, T. Mansfield, and R. Judson, "Topological and Functional Properties of the Small GTPases Protein Interaction Network," *PLoS One*, vol. 7, no. 9, p. e44882, Sep. 2012.

[60] E. S. Musiek, D. D. Xiong, and D. M. Holtzman, "Sleep, circadian rhythms, and the pathogenesis of Alzheimer disease.," *Exp. Mol. Med.*, vol. 47, no. 3, p. e148, Mar. 2015.

[61] D. P. Breen *et al.*, "Sleep and circadian rhythm regulation in early Parkinson disease.," *JAMA Neurol.*, vol. 71, no. 5, pp. 589–595, May 2014.

[62] D. Kuljis, A. M. Schroeder, T. Kudo, D. H. Loh, D. L. Willison, and C. S. Colwell, "Sleep and circadian dysfunction in neurodegenerative disorders: insights from a mouse model of Huntington's disease.," *Minerva Pneumol.*, vol. 51, no. 3, pp. 93–106, Sep. 2012.

[63] C. Jouffe *et al.*, "The Circadian Clock Coordinates Ribosome Biogenesis," *PLoS Biol.*, vol. 11, no. 1, p. e1001455, Jan. 2013.

ACCEPTED MANUSCRIPT

Tables

Table 1 The enriched pathways for genes in class 1 using clusterProfiler ($p < 0.05$). Genes in class 1 tend to actively express in the night. P.adjust values are the Benjamin and Hochberg corrected p values.

Table 2 The enriched pathways for genes in class 2 using clusterProfiler ($p < 0.05$). Genes in class 2 tend to actively express in the daytime. P.adjust values are the Benjamin and Hochberg corrected p values.

Figure Legends

Figure 1 A heatmap with two side clustering for 270 circadian genes based on the transformed state data. The 10 time points are from 7.5 hours after the experimental baseline to 34.5 hours with 3 hours intervals. The red and cyan colors indicate increasing and decreasing trends compared with the previous time point.

Figure 2 Three distinct expression patterns of the periodic/aperiodic genes in human blood. The dots and bars represent the mean and 95% confidence interval of the cumulative summation of the state data for each class at every time point. Circadian genes are classified into class 1 and class 2 and non-circadian genes are denoted as class 0.

Figure 3 The ROC curve for circadian gene classification of DNN and other four machine learning methods. The x-axis and y-axis stand for false positive and true positive rate respectively. AUC (area under the curve) scores are listed in the bottom right corner for every method.

Figure 4 The performance of the five algorithms for multi-classes classification (class 0, 1 and 2) using ten-fold cross-validation. The evaluation metrics include accuracy, precision, recall and F1 score. DNN, Knn, Logit, Nb, Svm stand for deep neural network, k-nearest neighbor (k=3), logistic regression, naïve Bayes and support vector machine.

Figure 5 Comparison of circadian genes identified by the trained DNN model, JTK_CYCLE and the Möller-Levet et al method. The Venn diagram describes the overlaps of genes reported by these methods.

Figure 6 The expression distributions of some orphan circadian genes identified only by DNN. The p-values of all these eight genes given by JTK_CYCLE are greater than 0.5.

Figure 7 The top 10 enriched gene ontology terms for biological process, molecular function and cellular component. Circadian genes are predicted into two classes by DNN model. Color bars stand for p-values computed by clusterProfiler indicating the enrichment levels ($p < 0.05$).

ACCEPTED MANUSCRIPT

Table 1

ID	Description	pvalue	p.adjust
hsa04662	B cell receptor signaling pathway	1.62E-06	0.000334356
hsa01521	EGFR tyrosine kinase inhibitor resistance	3.18E-05	0.003294597
hsa04660	T cell receptor signaling pathway	0.000255037	0.011724368
hsa04140	Autophagy - animal	0.000291	0.011724368
hsa04012	ErbB signaling pathway	0.00036882	0.011724368
hsa05221	Acute myeloid leukemia	0.00038703	0.011724368
hsa04650	Natural killer cell mediated cytotoxicity	0.000396476	0.011724368
hsa05214	Glioma	0.000606558	0.015694688
hsa01522	Endocrine resistance	0.000775057	0.016449165
hsa04510	Focal adhesion	0.000794646	0.016449165
hsa04611	Platelet activation	0.000943	0.017749301
hsa05220	Chronic myeloid leukemia	0.001069	0.018435687
hsa05215	Prostate cancer	0.002027	0.031456532
hsa05211	Renal cell carcinoma	0.002279	0.031456532
hsa05223	Non-small cell lung cancer	0.002279	0.031456532
hsa04658	Th1 and Th2 cell differentiation	0.00279	0.036090713
hsa04917	Prolactin signaling pathway	0.003075	0.037447147
hsa05212	Pancreatic cancer	0.004343	0.049938937

Table 2

ID	Description	pvalue	p.adjust
hsa03010	Ribosome	1.66E-47	3.82E-45
hsa00190	Oxidative phosphorylation	9.87E-05	0.00705084
hsa03040	Spliceosome	0.000108	0.00705084
hsa03008	Ribosome biogenesis in eukaryotes	0.000123	0.00705084
hsa03060	Protein export	0.001252	0.04815085
hsa03013	RNA transport	0.001482	0.04815085
hsa05010	Alzheimer's disease	0.001482	0.04815085
hsa05016	Huntington's disease	0.001881	0.04815085
hsa05012	Parkinson's disease	0.002069	0.04815085
hsa03018	RNA degradation	0.002094	0.04815085

ACCEPTED MANUSCRIPT

Highlights

- We design a novel circadian gene identification framework using deep neural network (DNN) in human blood transcriptome.
- The DNN model achieves the best performance compared with other machine learning methods.
- The DNN model outperforms previous circadian gene identification methods, and detects 1132 novel circadian genes.
- The novel circadian genes reveal that the GTPase superfamily exhibits distinct circadian expression patterns.
- The GTPase superfamily may provide molecular switch of circadian control of immune system in human blood.

ACCEPTED MANUSCRIPT