CirGRDB: a database for the genome-wide deciphering circadian genes and regulators

Xianfeng Li1, Leisheng Shi2,3, Kun Zhang2,3, Wenqing Wei2, Qi Liu4, Fengbiao Mao5, Jinchen Li1, Wanshi Cai2, Huiqian Chen3, Huajing Teng2,*, Jiada Li1,* and Zhongsheng Sun2,*

1State Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan 410078, China, 2Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China, 3Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou, Zhejiang 325035, China, 4State Key Laboratory of Tree Genetics and Breeding, Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China and 5Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA

Received August 15, 2017; Revised September 21, 2017; Editorial Decision October 02, 2017; Accepted October 03, 2017

ABSTRACT

Circadian rhythms govern various kinds of physiological and behavioral functions of the living organisms, and disruptions of the rhythms are highly detrimental to health. Although several databases have been built for circadian genes, a resource for comprehensive post-transcriptional regulatory information of circadian RNAs and expression patterns of disease-related circadian RNAs is still lacking. Here, we developed CirGRDB (http://cirgrdb.biols.ac.cn) by integrating more than 4936 genome-wide assays, with the aim of fulfilling the growing need to understand the rhythms of life. CirGRDB presents a friendly web interface that allows users to search and browse temporal expression patterns of interested genes in 37 human/mouse tissues or cell lines, and three clinical disorders including sleep disorder, aging and tumor. More importantly, eight kinds of potential transcriptional and post-transcriptional regulators involved in the rhythmic expression of the specific genes, including transcription factors, histone modifications, chromatin accessibility, enhancer RNAs, miRNAs, RNA-binding proteins, RNA editing and RNA methylation, can also be retrieved. Furthermore, a regulatory network could be generated based on the regulatory information. In summary, CirGRDB offers a useful repository for exploring disease-related circadian RNAs, and deciphering the transcriptional and post-transcriptional regulation of circadian rhythms.

INTRODUCTION

Circadian rhythms represent evolutionary adaptations of living organisms to the daily cyclical changes in their environments (1,2). In mammals, the rhythms are generated by cell-autonomous transcriptional feedback loops composed of positive transcriptional activators and negative feedback elements, resulting in rhythmic expression of several core clock genes and numerous clock-controlled genes (CCGs) (3–5). Genome-wide transcriptional profiling has revealed thousands of circadian-expressed transcripts, the majority of which are essential for mammalian physiology and behavior (6–10). Thus, through the CCGs, circadian rhythms govern a wide variety of behavioral, physiological and metabolic functions (11–14). Persistent abnormalities in circadian rhythms are highly detrimental to health, possibly leading to aberrant metabolism and cellular proliferation, and then increasing the risk of cancers, obesity, diabetes, accelerated aging and mental illnesses (15–21). Therefore, identifying the genes that are expressed with a circadian cycle and uncovering their regulators will not only improve the understanding of the rhythms of life, but will also provide valuable insights into potential treatments and therapies for clinical disorders, which are associated with disturbances of circadian rhythms (22–24).

Genome-wide oscillations of CCGs require the coordinated control of numerous regulators at multiple levels (7). Besides the core circadian feedback elements, numerous transcriptional and post-transcriptional regulators have been shown to be involved in the oscillations of core clock genes and CCGs (25,26). For example, CLOCK, a central component of the circadian pacemaker, can work as a histone acetyltransferase and directly alters histone acetylation of specific lysine residues around their DNA binding sites (27,28). Recently, histone deacetylase 3 (HDAC3) is suggested to be another critical component for the cir-
circadian system through regulating the activation and repression processes in a deactylase activity-independent manner (29). Furthermore, data obtained by chromatin immunoprecipitation sequencing (ChIP-Seq) demonstrated the circadian fluctuation of various histone modifications, such as H3K4me3, H3K9ac, H3K27ac, H3K36me3 and H3K4me1, correlated with the rhythmic expression of their co-localized CCGs (9). Rhythmic enhancer RNAs (eRNAs) have also been speculated to harbor enhancers to form chromatin loops, which facilitates the rhythmic binding of transcription factors (TFs) to regulate the expression of CCGs at the transcriptional level (30–32). However, only ~20% of cycling mRNAs have corresponding rhythmic nascent RNA (6,7,33), indicating the existence of another type of regulation for the cycling of the remaining rhythmic mRNAs. Following the development of diverse sequencing technologies, the results of several studies have strongly suggested that post-transcriptional regulation may be implicated in 70~80% of cycling of RNAs. For example, two rhythmic RNA-binding proteins (RBPs), CIRBP and RBM3, can regulate several clock genes in response to temperature fluctuations by binding to 3'UTRs of these genes (34,35). Additionally, miRNA and RNA editing are both viewed as being the key mechanisms in the circadian clockwork (9,36). To acquire an overall view of the regulation of genes involved in circadian expression, it is imperative to elucidate the regulators of RNA cycling at both the transcriptional and post-transcriptional level (37).

As an increasing number of cycling genes and transcriptional and post-transcriptional regulators have been identified due to multiple ‘omic’ approaches, it is imperative to build a platform to collect and integrate the multi-layer information. Several databases have been built for circadian genes, such as CGDB (http://cgdb.biocuckoo.org/index.php) (38), CircaDB (http://circadb.hogeneschlab.org/) (39), and CircadiOmites (http://circadiomics.igb.uci.edu/) (40). CGDB hosts circadian genes curated from published small-scale or high-throughput experimental data of multiple organisms; CircaDB focuses on gene expression profiles of different mouse tissues; CircadiOmites integrates genomic, transcriptomic, proteomic and metabolomic data sets from the livers of wild-type and Clock mutant mice. However, disease-related circadian RNAs and their expression patterns, and comprehensive post-transcriptional regulatory information of circadian RNAs such as histone modifications, eRNAs, RBPs and miRNA, are still lacking from these databases. In the present study, we developed CirGRDB by integrating >4936 genome-wide assays from 96 circadian transcriptome/regulome profiling datasets, with the aim of fulfilling the growing need to understand the rhythms of life. The database offers a user-friendly web interface to search, visualize and interpretate circadian genes and multiple transcriptional and post-transcriptional regulatory events in both humans and mice. In general, CirGRDB is designed to provide bench researchers substantial convenience to explore the underlying regulatory mechanism for oscillating genes and disease-related circadian RNAs.

DATA COLLECTION AND PROCESSING

Data collection

CirGRDB focuses on the identification and annotation of circadian RNAs and their regulators from the available human and mouse transcriptome/regulome datasets. Firstly, 2249 assays from 10 circadian datasets of human sleep disorder, aging, epidermal stem cell and cancer cell line were obtained from the ArrayExpress/SRA/GEO database (Supplementary Table S1). We also collected 2259 assays from 50 genome-wide circadian transcriptional profiling datasets in 25 mouse tissues/cell lines of wild type, loss of function of core clock genes, tumor-bearing, sleep restriction, aging or fasting strains under normal light-dark cycle or constant darkness. Moreover, we incorporated 60 regulatome assays from six human cell lines. In addition, we included 368 assays from 30 genome-wide circadian datasets of eight kinds of potential regulators, including TFs, histone modifications, chromatin accessibility, eRNAs, miRNAs, RNA editing, RBPs and RNA methylation, from mouse livers, muscles, and two cell lines, respectively. Detailed information is listed in Supplementary Table S1. Due to the lack of direct regulatory information, we adopted the latest published methods to predict the potential role of rhythmic miRNA in RNA cycling.

Identification of the expression pattern of coding genes

For microarray data sets, we performed standard analysis process including quality control, pre-processing, normalization by using simpleaffy, genefilter, oligo (41) and limma (42) packages across different platforms. To avoid distortion of the results by noise, we filtered out uninformative probes (low variance, expressed uniformly close to background detection levels, without Entrez Gene identifiers, or have duplicated Entrez Gene identifiers). Finally, normalized log2-transformed expression values were obtained. For RNA-Seq data sets, after removing adapters and low quality bases of RNA-Seq datasets, we implemented STAR (43) to map reads to mouse genome mm10 or human genome hg38. Data sets with high mapping rate (more than 70%) were selected for further analysis. Then, featureCounts (44) was used to count the reads according to the coordinates of genes in GENCODE (human V24 and mouse V1) (45). Gene expression levels were normalized using the TMM method by edgeR (46). Genes with RPKM (reads per kilobase of exon model per million mapped reads) value >0.5 in at least one third of the samples (33) were used for further analysis. Rhythmically expressed genes were identified using LSPR (47). Genes expressed with period between 22 and 26 h and a cutoff of P<0.05 and amplitude >0.10 were defined as rhythmic-expressed genes.

Identification and prediction of rhythmic binding regions, histone modifications and DNase I hypersensitive sites

ChIP-Seq datasets of 23 TFs related to circadian rhythm and 8 histone modifications, and datasets of DNase I hypersensitive sites were mapped to human genome hg38 or mouse genome mm10. MACS_1.4.0rc2 (48) was employed to identify peaks, and subdivided by Peaksplitter (49). The
highest peak was selected as the master peak, and signals of master peaks were normalized into counts per million (CPM) (7). LSPR (47) was then implemented to identify regions of rhythmic binding or histone modifications. For ChIP-Seq datasets only detected two-time points within 24-hour, differential regions were identified using diffReps (50). The regions obtained above were then annotated to human or mouse genome using HOMER (51).

Identification and collection of RNA editing and RBPs binding sites

RNA editing means the deamination of adenosine into inosine (A-to-I) catalyzed by the ADAR family of proteins. ADARB1, a key player in cycling RNA editing and mRNA rhythms, could control the speed of circadian oscillations (36). We collected 132 rhythmic editing sites, mediated by the rhythmic expression of ADARB1 in mouse liver (36). In addition, we analyzed RNA editing sites in seven circadian transcriptome datasets from 14 mouse tissues and two human tissues using RED-ML (53). Rhythmic editing sites were identified using LSPR (47) and annotated using HOMER (51). RBPs can form ribonucleoprotein (RNP) complexes with RNAs, and play a critical role in RNA biogenesis, stability, transport and cellular localization at the post-transcriptional level (54). Genome-wide binding regions of two rhythmic expressed RNA-binding proteins (RBPs), CIRBP and RBM3 (34,35), were also integrated into our database.

Analysis of regulatory network and functional effect

For each rhythmic genes, we constructed a comprehensive regulatory network using Cytoscape Web (55) based on the collected hierarchical gene regulatory networks from PTHGRN (56), which contained the information of post-translational modifications (PTMs), TFs, epigenetic modi-
fications and gene expression. The functional effect of specific gene was annotated to human disease database, DISEASES (http://diseases.jensenlab.org).

DATABASE CONTENT AND THE WEB INTERFACE

Database content and statistics
To fulfill the unprecedented need for decoding circadian oscillations from multiple levels, we developed CirGRDB by integrating more than 4936 genome-wide assays. All the metadata of CirGRDB were stored in a MySQL database, and the gene symbols and aliases of rhythmic genes were used as the primary keys to index by MySQL (Figure 1). Open source JavaScript frameworks, jQuery.js, highcharts.js and Cytoscape Web, were employed to display tables and to visualize the oscillations of regulatory elements or gene expression. By employing this database, users can search, browse and annotate circadian genes and their potential regulators, including: (i) temporal expression patterns of circadian genes in 12 human tissues/cell lines and 25 mouse tissues/cell lines under normal light-dark conditions or in constant darkness; (ii) cycling genes related to 3 clinical disorders and corresponding abnormal expression patterns; (iii) eight kinds of potential transcriptional and post-transcriptional regulators involved in the rhythmic expression of the specific genes, such as RNA editing site, binding regions of miRNA and histone modification; (iv) abnormal expression patterns of cycling genes under the loss function of 17 regulators; (v) functional effect and network of target genes. Data retrieved using CirGRDB could be utilized, based on the genes of interest, and users may access the data freely through the web interface (http://cirgrdb.biols.ac.cn).

Web interface
A user-friendly web interface for CirGRDB was constructed, and users can input the gene symbol or alias of interest to retrieve its rhythmic expression patterns and eight kinds of regulatory information involved in its expression (Figure 2). Briefly, user can retrieve three groups of transcriptome profiles including Group1: Expression pattern of rhythmic gene in normal tissues or cell lines; Group2: Expression pattern of rhythmic gene under different conditions; Group3: Effects of knock-out/down or over-expression of specific gene (Figure 2A). Then, user can also retrieve 8 kinds of regulatory information including transcriptional factors, histone modifications, enhancer RNAs, chromatin accessibility, miRNAs, RNA binding proteins, RNA editing and 6 mA (Figure 2B). Furthermore, a regulatory network and disease information relating to rhythmic genes are shown in the figures and tables below (Figure 2C). In addition, user can visualize these 3 groups of transcriptome profiles and regulatory signals in CirGRDB.
Figure 3. An example of data access in CirGRDB. ‘CirGRDB’ search (A, B) enables integrative viewing of the expression patterns (C–F), multiple potential regulatory information (G–J) and network (K) of Pparg. User can retrieve ‘Expression’, ‘Regulation’ and ‘Network’ information related to Pparg (B). Expression patterns of Pparg with raw expression level or normalized (Z score) in eight mouse tissues (C), and in livers of wild type, Sirt1−/−, and Sirt6−/− mice (D) at different circadian time (CT) or zeitgeber time (ZT). Different expression patterns of PPARG in human breast cancer cell line (MCF7) and normal breast cell line (MCF10A) were observed (E). Knockout of Ncor1/Ncor2 results in increased expression of Pparg in liver, whereas knockout of Per2 results in decreased expression of Pparg in liver (F). Differential binding of Rev-erba between rhythmic expression time points (G). Peak time represents time point with the highest binding of Rev-erba, while trough time represents time point with the lowest binding of Rev-erba. H3K27ac peaks nearby Pparg show similar expression pattern with Pparg (H). Enhancer RNA (eRNA) nearby Pparg shows rhythmic expression pattern (I). Knockout of SRC2 results in deactivation of DNase I hypersensitive sites (J). Network of input gene was constructed based on PTHGRN database (K).
with JBrowse (Figure 2D). For example, taking the gene Pparg as the input (Figure 3A, B) to the search web interface will instantly get the ‘Expression’ (Figure 3C–F), ‘Regulation’ (Figure 3G–J), and ‘Network’ information (Figure 3K) related to Pparg. In more detail, Pparg was expressed rhythmically in several tissues (Figure 3C), and different expression patterns of Pparg in human breast cancer cell line and normal breast cell line were observed (Figure 3E). Knockout of Neor1/Neor2 results in increased expression of Pparg in liver (Figure 3E), whereas knockout of Per2 results in decreased expression of Pparg in liver (Figure 3F). H3K27ac peaks nearby Pparg show similar expression pattern with Pparg (Figure 3H), and eRNA nearby Pparg shows rhythmic expression pattern (Figure 3I). The regulatory network of Pparg was also constructed and visualized based on regulatory information collected from PTHGRN (Figure 3K).

**DISCUSSION AND PERSPECTIVES**

Several databases including CGDB (38) and CircaDB (39) have been built for circadian genes curated from published small-scale or high-throughput experimental data, but disease-related circadian RNAs, and their post-transcriptional regulators are not included into these databases. CircadiOmics (40) is another database that integrates genomic, transcriptomic, proteomic and metabolomic data sets from the livers of wild-type and Clock mutant mice, and several public data sets of ChIP-Seq/ChIP-ChIP. However, information on circadian time point, which is crucial for circadian researches, are still absent from these ChIP-Seq/ChIP-ChIP data sets. Thus, disease-related circadian RNAs and their expression patterns, and comprehensive post-transcriptional regulatory information of circadian RNAs such as eRNAs, RBPs and miRNA, are still lacking from the published databases. This study builds the first platform for genome-wide deciphering both transcriptional and post-transcriptional regulation of circadian RNAs using 96 circadian transcriptome/regulome profiling datasets. In this study, eight kinds of regulatory elements were collected and predicted, including TFs, histone modifications, chromatin accessibility, eRNAs, miRNAs, RBPs, RNA editing and RNA methylation, to construct a regulatory network of rhythmic genes at multiple regulatory layers. An integrated analysis revealed that circadian clock was endogenously oscillated by interconnected hierarchies consisting of various transcriptional and post-transcriptional regulators (32). The present study provides a highly useful resource for the further analysis of temporal patterns of gene expression and functionality. Currently, there is a lack of sets of regulatory data relating to circadian rhythms in mice and human tissues other than liver, such as the SCN, kidney, colon and heart, hence CirGRDB is limited to regulatory information in this well-studied mouse tissue. In the nucleus, the chromatin remodelers, dynamic transcriptional factors, non-coding RNAs and epigenetic modifications might control 3–30% of oscillating transcripts at the transcriptional level through the formation of a three-dimensional circadian interactome (32,57). With increasing numbers of studies into the regulatory mechanisms of circadian clocks and the prevalence of high-throughput sequencing, more regulatory sequencing datasets, such as three-dimensional interactions, rhythm TFs and RBP bindings, and histone modifications in different tissues, will become available. Emerging mass spectrometry methodologies will also provide unprecedented opportunities for the identification of the rhythmic expression of proteins and subsequent post-translational modifications. We will keep CirGRDB updated as above information emerges. It is anticipated that CirGRDB will be a useful repository for deciphering the regulation of circadian rhythms, and that it will provide valuable insights into novel therapies for circadian-related disorders of humans.

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR online.

**FUNDING**

National Natural Science Foundation of China [31371187, 31200888, 31271266]. Funding for open access charge: National Natural Science Foundation of China [31371187]. Conflict of interest statement. None declared.

**REFERENCES**

1. Bass, J. and Takahashi, J.S. (2010) Circadian integration of metabolism and energetics. Science, 330, 1349–1354.
2. Green, C.B., Takahashi, J.S. and Bass, J. (2008) The meter of metabolism. Cell, 134, 728–742.
3. Albrecht, U., Sun, Z.S., Eichele, G. and Lee, C.C. (1997) A differential response of two putative mammalian circadian regulators, mper1 and mper2, to light. Cell, 91, 1055–1064.
4. Hogenesch, J.B. and Ueda, H.R. (2011) Understanding systems-level properties: timely stories from the study of clocks. Nat. Rev. Genet., 12, 407–416.
5. Sun, Z.S., Albrecht, U., Zhuchenko, O., Bailey, J., Eichele, G. and Lee, C.C. (1997) RIGUI, a putative mammalian ortholog of the Drosophila period gene. Cell, 90, 1003–1011.
6. Menet, J.S., Rodriguez, J., Abruzzi, K.C. and Rosbash, M. (2012) Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. eLife, 2012, 1–25.
7. Koike, N., Yoo, S.H., Huang, H.C., Kumar, Y., Lee, C., Kim, T.K. and Takahashi, J.S. (2012) Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. Science, 338, 349–354.
8. Rey, G., Cesbron, F., Rougemont, J., Reinke, H., Brunner, M. and Naef, F. (2011) Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. PLoS Biol., 9, e1000595.
9. Vollmers, C., Schmitz, R.J., Nathanson, J., Yeo, G., Ecker, J.R. and Panda, S. (2012) Circadian oscillations of protein-coding and regulatory RNAs in a highly dynamic mammalian liver epigenome. Cell Metab., 16, 833–845.
10. Zhang, E.E. and Kay, S.A. (2010) Clocks not winding down: unravelling circadian networks. Nat. Rev. Mol. Cell. Biol., 11, 764–776.
11. Hood, S. and Amir, S. (2017) The aging clock: circadian rhythms and later life. J. Clin. Invest., 127, 437–446.
12. Sahar, S. and Sassone-Corsi, P. (2009) Metabolism and cancer: the circadian clock connection. Nat. Rev. Cancer, 9, 886–896.
13. Takahashi, J.S. (2016) Transcriptional architecture of the mammalian circadian clock. Nat. Rev. Genet., 18, 164–179.
14. Zhou, J., Yu, W. and Hardin, P.E. (2016) CLOCKWORK ORANGE Enhances PERIOD Mediated rhythms in transcriptional repression by antagonizing E-box binding by CLOCK–CYCLE. PLoS Genet., 12, e1006430.
