Computational and Structural Biotechnology Journal 23 (2024) 1844–1853

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Database article

SUDAZFLNC – a curated and searchable online database for zebrafish lncRNAs, mRNAs, miRNAs, and circadian expression profiles

Shital Kumar Mishra [a,b], Han Wang [a,b,*]

[a] Center for Circadian Clocks, Soochow University, Suzhou 215123, Jiangsu, China
[b] School of Biology & Basic Medical Sciences, Suzhou Medical College, Soochow University, Suzhou 215123, Jiangsu, China

ARTICLE INFO

Keywords:
LncRNAs
MiRNAs
MRNAs
Database
Zebrafish
Bioinformatics

ABSTRACT

The zebrafish (*Danio rerio*) has emerged as a model organism for investigating lncRNAs-driven fundamental biological processes, such as circadian rhythms, physiology, metabolism, and various diseases. While state-of-the-art sequencing technologies have identified an increasing number of lncRNAs in zebrafish, their annotations are far from complete. In this study, we collect 28,925 lncRNAs from both the published studies and our own RNA-seq analyses and establish a novel webserver-based database called SUDAZFLNC (https://sudarna.website/). The database, containing 28,925 lncRNAs, 25,432 mRNAs, and 368 miRNAs, provides several crucial features and annotations for the zebrafish RNAs, such as sequence identifiers (IDs), sequence length, hexamer score, coding probabilities, GO and KEGG annotations, and micropeptides. SUDAZFLNC also includes time-course expression profiles of 3288 lncRNAs, 25,432 mRNAs, and 342 miRNAs generated from our RNA-seq experiments, and 149, 4407, and 43 rhythmically expressed lncRNAs, mRNAs, and miRNAs, respectively. Based on the peak expression patterns, we classified these RNAs into morning RNAs, evening RNAs, and night RNAs. Users of the database can access the RNA sequences and their expression profiles by searching the corresponding IDs from the Graphical User Interface (GUI) of the database. The database supports several features to investigate RNA sequences and expression profiles, including BLAST, search of sequence and data, ID conversion, and RNA-RNA interaction prediction. This is the largest curated database of zebrafish RNAs and their expression profiles to date.

1. Introduction

Zebrafish have been employed to investigate a variety of crucial biological processes, including circadian rhythms [1–4]. Zebrafish provide a suitable model to study the light induction effects, locomotor activity, circadian mechanisms, and sleep-wake cycle [5–9]. LncRNAs, defined as noncoding RNAs longer than 200 nucleotides, are known to regulate numerous biological processes [10], including orchestrating biological rhythms [11,12], tissue and organ repair, and the function of the immune system [13]. A growing number of studies have revealed the involvement of the lncRNAs in transcription and regulation of cell fate determination [14,15].

Various noncoding RNA databases have been developed, including RNAcentral (https://rnacentral.org/) of lncRNAs from a broad range of organisms, LncBook 2.0 (https://ngdc.cncb.ac.cn/lncbook/) of human lncRNAs, MONOCLdb (https://www.monocldb.org/) of mouse lncRNAs, and deepBase (https://rna.sysu.edu.cn/deepbase3/) of lncRNAs from the cancer samples. However, there is still a lack of a comprehensive database for zebrafish lncRNAs that includes sequences, experimental profiles, and analytical tools. Although recent studies have cataloged thousands of zebrafish lncRNAs, they often lack crucial annotation information, such as the coding abilities of the lncRNAs. For example, the NONCODE database [16] contains 4852 zebrafish lncRNAs, ZFLNC database [17] contains 21,128 lncRNAs, and Ensembl contains 8115 zebrafish lncRNAs. However, these datasets require additional curation owing to the lack of crucial annotation information. For instance, the NONCODE database lacks curated proper identifiers for the lncRNAs. The ZFLNC database has more than 7000 lncRNAs without appropriate identifiers. Further, several lncRNAs are shared by the NONCODE, ZFLNC, and Ensembl databases with different IDs in each database, making it difficult for the researchers to treat them as the same lncRNAs. We noted that ZFLNC shared 1549 lncRNAs with Ensembl, and 332 lncRNAs with NONCODE. The same lncRNA is named ZFLNCT00001 in ZFLNC database, while it has an Ensembl ID of

* Corresponding author at: Center for Circadian Clocks, Soochow University, Suzhou 215123, Jiangsu, China.
E-mail address: wanghan@suda.edu.cn (H. Wang).

https://doi.org/10.1016/j.csbj.2024.04.026
Received 30 December 2023; Received in revised form 29 March 2024; Accepted 9 April 2024
Available online 12 April 2024

ENSDART00000152494. Another lncRNA, named ZFLNCT00002, is called NONDRET000440 in the NONCODE database. Although lncRNAs are known to contribute to circadian rhythms [1–4], none of these databases provides a comprehensive experimental profile of the RNAs, let alone the circadian profiles. In particular, there is a lack of an integrative zebrafish RNA expression dataset. These studies and their limitations inspired us to develop a new database of zebrafish lncRNAs that supplements previous studies and provides novel annotations for the zebrafish lncRNAs.

In this study, we collected tens of thousands of zebrafish RNAs from the available literature sources and our own RNA-seq studies and develop a new online query-based searchable database of 28,925 lncRNAs, 25,432 mRNAs, and 368 miRNAs (Fig. 1). The database also provides time-course RNA-seq profiles of 3288 lncRNAs, 25,432 mRNAs, and 342 miRNAs generated from our own whole transcriptome sequencing experiments. Rhythmicity analysis of these expression profiles revealed 149, 4407, and 43 rhythmically expressed lncRNAs, mRNAs, and miRNAs, respectively. Interestingly, we found that tens of thousands of these zebrafish lncRNAs harbor small Open Reading Frames (smOFRs) with coding potentials. The user-friendly web

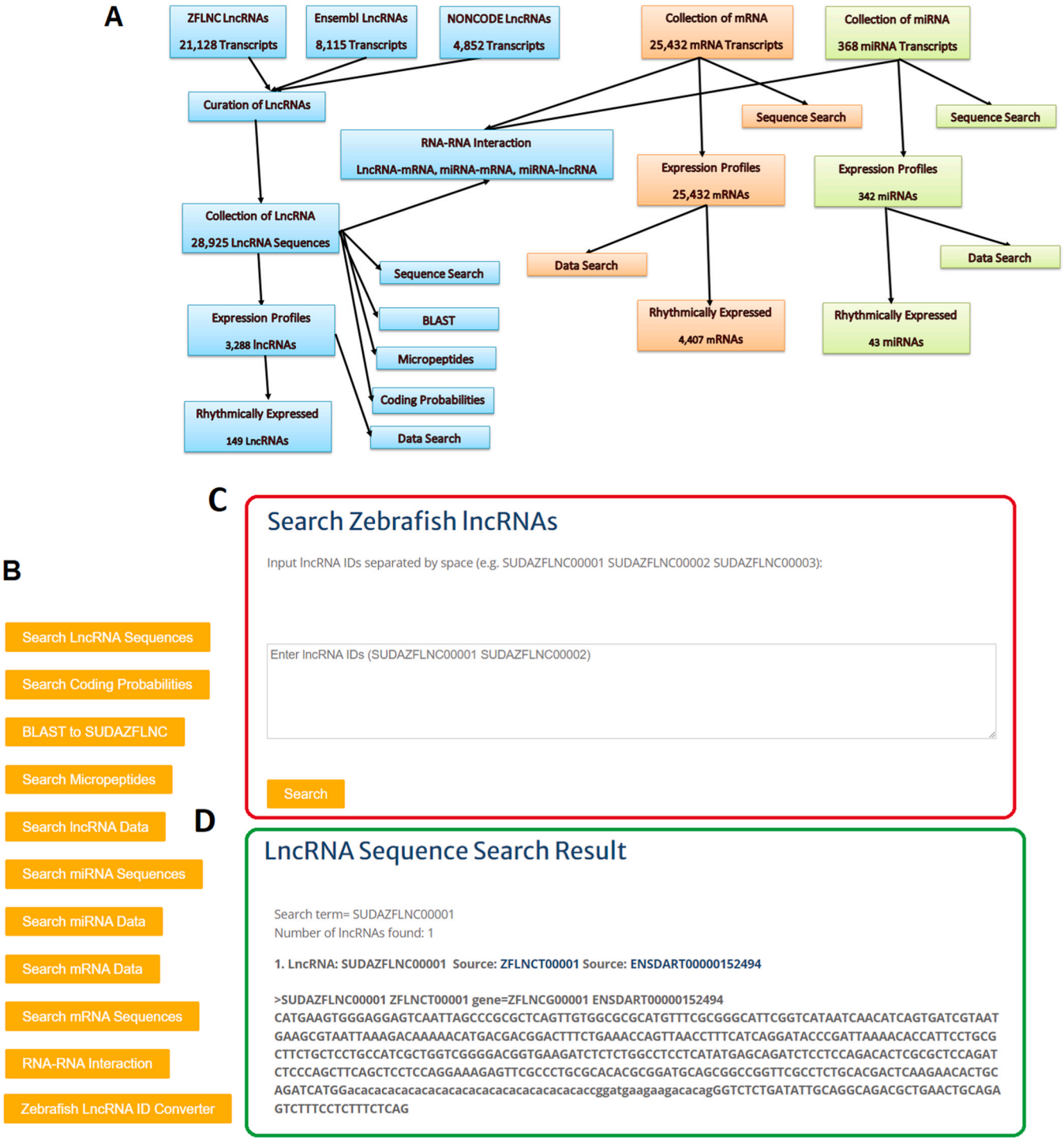

Fig. 1. Database development flowchart, features, and searching zebrafish lncRNA sequences. **(A)** Flow chart of the development of the SUDAZFLNC database in the study. **(B)** List of all features supported by the database. **(C)** Input of identifiers to search the lncRNA sequence. **(D)** Output of lncRNA search query.

interface (https://sudarna.website/) of the database provides several methods for the users to access the database features (Fig. 1A), such as sequence search, expression profile search, BLAST to the lncRNAs, lncRNAs' coding potential search, and RNA-RNA binding interaction search. The database enables a user to search the RNAs by their unique identifiers in both single and batch queries. A local nucleotide BLAST operation is supported by the database where a user can BLAST any random nucleotide sequence to find the sequence similarities with the known lncRNAs in the database. To the best of our knowledge, SUDAZFLNC is the largest database of the zebrafish RNAs and their expression profiles and should help biologists select interesting candidate zebrafish lncRNAs, mRNAs, and miRNAs, and identify their novel functions.

2. Data Sources and Implementation

2.1. Collection, curation, and searching zebrafish lncRNAs and their GO and KEGG analyses

We collected 21,128 zebrafish lncRNAs from the ZFLNC database (Supplementary Table 1, Fig. 1 A), 4852 zebrafish lncRNAs from NONCODE (Supplementary Table 2), and 8115 zebrafish lncRNAs from Ensembl (Supplementary Table 3). Subsequently, we curated these lncRNAs to derive a unified set of a total of 28,925 zebrafish lncRNAs. These lncRNAs were assigned unique identifiers from SUDAZFLNC00001 to SUDAZFLNC28925. Out of the 21,128 ZFLNC lncRNAs, only 13,417 lncRNAs (Supplementary Table 4) have standard identifiers, including Ensembl IDs, and/or NONCODE IDs. A total of 7711 ZFLNC lncRNAs (Supplementary Table 5) have no standard Identifiers. Further, 332 NONCODE lncRNAs were already part of the ZFLNC database. The 13,417 lncRNAs with standard identifiers were assigned identifiers from SUDAZFLNC00001 to SUDAZFLNC13417. While curating the lncRNAs, we preserved the original identifier information. For example, the header of SUDAZFLNC00001 includes all corresponding header information from the ZFLNC database, *i.e.*, "ZFLNCT00001 gene=ZFLNCG00001 ENSDART00000152494."

In order to find identifiers of the remaining 7711 ZFLNC lncRNAs, we performed local BLAST of these 7711 lncRNAs to the known lncRNAs from NONCODE and Ensembl databases. We employed a high expected value of E-100 as a threshold cut-off score to detect matching lncRNA identifiers. Out of the 7711 ZFLNC lncRNAs, 1133 matched with 332 NONCODE lncRNAs (Supplementary Table 6). These 1133 ZFLNC lncRNAs were given unique identifiers from SUDAZFLNC13418 to SUDAZFLNC14550. Another set of 332 NONCODE lncRNAs were already part of ZFLNC lncRNAs. Further, 24 NONCODE lncRNAs were part of both 1133 and 332 NONCODE lncRNAs. Hence, the remaining 4212 NONCODE lncRNAs (= 4852 – 332 - 332 + 24), computed with standard set theory formula, were included in the new database as SUDAZFLNC14551 to SUDAZFLNC18762. The remaining 6578 ZFLNC lncRNAs without identifiers were further analyzed to find standard identifiers. Ensembl database had a downloadable file containing 8115 zebrafish lncRNAs. We performed local BLAST of these 6578 lncRNAs to the 8115 Ensembl lncRNAs. A total of 553 ZFLNC lncRNAs match with 250 Ensembl lncRNAs (Supplementary Table 7). These 553 ZFLNC lncRNAs were given unique IDs from SUDAZFLNC18763 to SUDAZFLNC19315. Further, a new set of 1549 Ensembl lncRNAs were already part of ZFLNC lncRNAs. No lncRNAs were common between these 553 and 1549 Ensembl lncRNAs. Hence, the remaining 6316 lncRNAs (= 8115 – 250 – 1549), were included in our database with IDs ranging from SUDAZFLNC19316 to SUDAZFLNC25631. Despite these analyses, 6025 ZFLNC lncRNAs were still left without IDs.

We performed NCBI BLAST of these 6025 ZFLNC lncRNAs and revealed that 2576 lncRNAs match with mRNAs. Hence, we excluded these 2576 ZFLNC lncRNAs from our database. We analyzed the remaining 3449 (= 6025 - 2576) with CPAT [18] to detect coding labels and found 170 lncRNAs with coding labels, and these lncRNAs were excluded from the database. Out of the remaining 3279 ZFLNC lncRNAs, 217 have standard NCBI IDs. These 217 lncRNAs were included in the database with IDs ranging from SUDAZFLNC25632 to SUDAZFLNC25848. The remaining 3062 ZFLNC lncRNAs were included in the database with IDs ranging from SUDAZFLNC25849 to SUDAZFLNC28910. Since new lncRNAs are continuously being identified, we hypothesized that the list of 8115 lncRNAs from Ensembl might not be complete. With an additional manual search in the Ensembl database, we found 15 additional Ensembl lncRNAs. As such, we collected 18,165 lncRNAs (13,417 + 1133 + 553 + 3062) from ZFLNC, 4212 lncRNAs from NONCODE, 6331 lncRNAs (6316 + 15) from Ensembl, and 217 lncRNAs from NCBI. Together, we have 28,925 zebrafish lncRNAs (Supplementary Table 8).

The database supports searching lncRNAs using unique identifiers (https://sudarna.website/search-zebrafish-lncrna-sequences/). The user can simply input a particular lncRNA ID, such as SUDAZFLNC00001, to find the corresponding sequences (Fig. 1B-1D). The database also supports batch search of the lncRNAs. For instance, to search the sequence of three lncRNAs SUDAZFLNC00001, SUDAZFLNC00002, and SUDAZFLNC00003, the user should enter these sequence IDs separated by spaces, *i.e.*, "SUDAZFLNC00001 SUDAZFLNC00002 SUDAZFLNC00003." The search result includes fasta format sequences as well as hyperlinks to relevant source databases such as ZFLNC, Ensembl, NCBI, and NONCODE.

The SUDAZFLNC database also supports analyzing GO (https://sudarna.website/go-annotation-search/) and KEGG (https://sudarna.website/kegg-annotation-search/) annotations of the zebrafish lncRNAs. The search of GO and KEGG functions of the lncRNAs provides the corresponding hyperlinked GO and KEGG terms.

2.2. Creation of local BLAST database and performing BLAST to the database

We employed the standard NCBI *makeblastdb* command (https://ftp.ncbi.nih.gov/blast/executables/LATEST/) on Linux operating system to create the local database. The fasta format of 28,925 lncRNA sequences was given as input to *makeblastdb* command. The *dbtype* and *index* parameters of *makeblastdb* command were set to *nucl*, and *hash_index*, respectively. The database supports both single sequence BLAST and batch BLAST (Supplementary Figure ^{1}A–^{1}B). To perform BLAST to the SUDAZFLNC, the user can input fasta format nucleotide sequence in the query input box (https://sudarna.website/lncrna-blast/). Further, the users also can select a desired *p-value* and *wordsize* input parameters from the dropdown menu. The BLAST feature supports a maximum query length up to 1000,000 bp. The output of the BLAST search contains several statistical measures, such as sequence alignment, *Score*, *Expect value*, *Identities*, *Gaps*, *Strand*, *Gap Penalties*, and *Lambda*.

2.3. Computation of coding probabilities and micropeptides of lncRNAs

We employed the CPAT to identify small Open Reading Frames (smORFs) and their corresponding coding probabilities because CPAT has an outstanding sensitivity (0.96) and specificity (0.97) [18]. The computationally predicted measures included ORF_strand, ORF_frame, ORF_start, ORF_end, ORF size, Fickett score, Hexamer score, and Coding_prob. The analyses revealed that 25,411 out of the 28,925 (87.85%) lncRNAs harbor smORFs. The total number of smORFs harbored by 25,411 lncRNAs is 276,035 (Supplementary Table 9). For the lncRNAs with multiple smORFs, the highest coding probability of the corresponding smORF was assigned as the coding probabilities (Supplementary Table 10). Intriguingly, 3514 lncRNAs have no smORFs, and hence, these 3514 lncRNAs have no coding potentials (Supplementary Table 11).

2.4. Searching lncRNAs' coding probabilities and smORFs

The coding probabilities and smORFs of the lncRNAs can be assessed using the IDs of the lncRNAs (https://sudarna.website/sudazflnc-coding-probabilities/). The database supports both individual and batch queries. The output of the ORF search includes *ORF ID*, *RNA size*, *ORF_strand*, *ORF_frame*, *ORF_start*, *ORF_end*, *ORF size*, *Fickett score*, *Hexamer score*, and *Coding_prob* (Supplementary Figure 2A–2B). The smORFs are accessible using the lncRNA IDs from the interface of the site (https://sudarna.website/micropeptide-search/). The query output includes all corresponding smORFs and their sequences (Supplementary Figure 3A–3B).

2.5. Time-course RNA-seq dataset of zebrafish lncRNAs, their rhythmicity analysis, and searching zebrafish lncRNAs' expression profiles

We generated time-course expression profiles for 3288 wild-type lncRNAs for consecutive two days with a four-hour interval, as

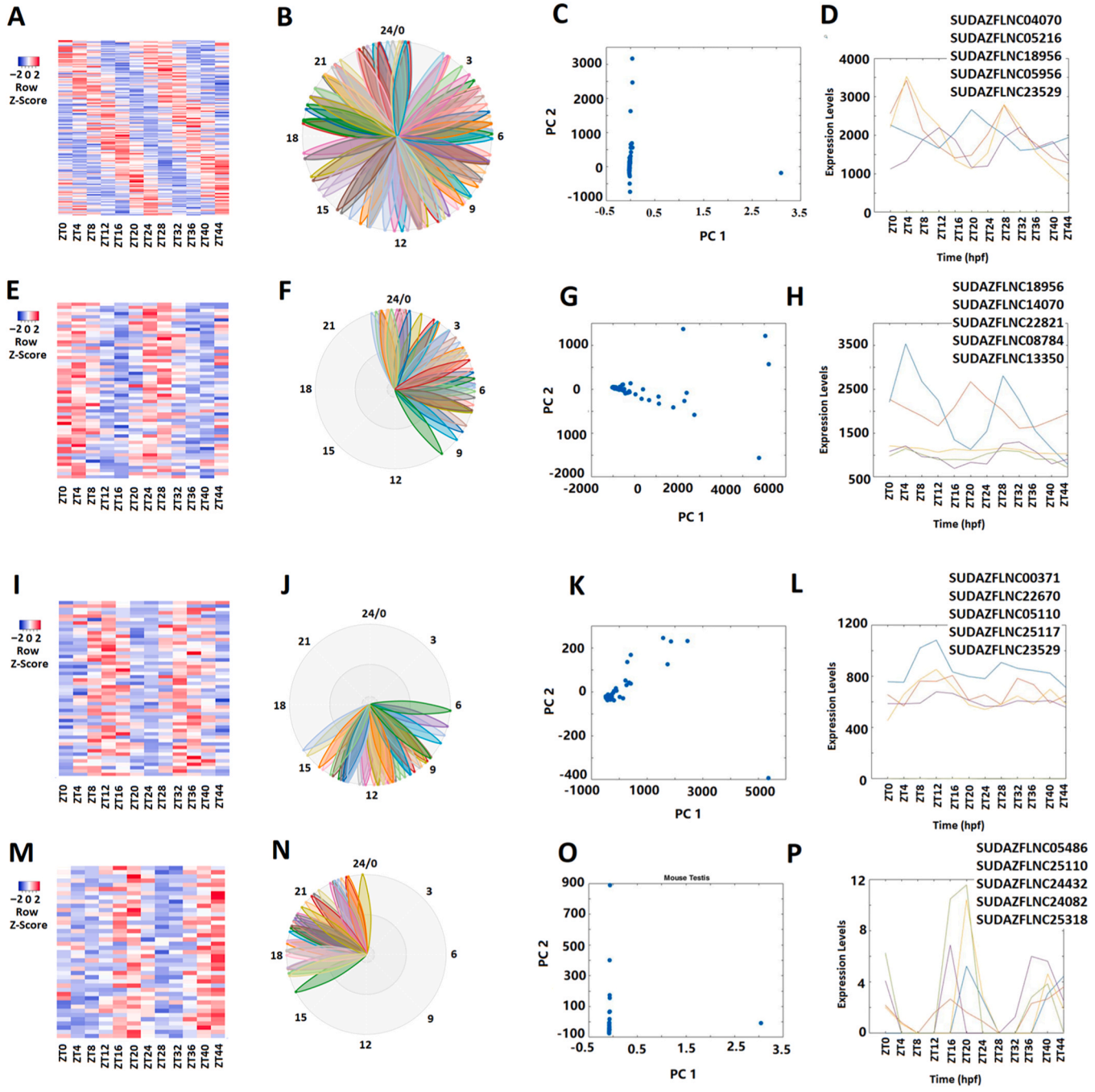

Fig. 2. Analyses of 149 rhythmically expressed zebrafish lncRNAs. (A-D) Analysis of all the 149 rhythmically expressed zebrafish lncRNAs: Heat map (A) of all the 149 rhythmically expressed lncRNAs, BioDare2 phase plots of all lncRNAs (B), PCA analyses of all lncRNAs with variances of PC1 99.96% and PC2 0.02% (C), and expression profiles of the representative lncRNAs (D). (E-H) Analysis of 56 morning lncRNAs: Heat map of the 56 morning lncRNAs (E), BioDare2 phase plots of morning lncRNAs (F), PCA analyses of the morning lncRNAs with variances of PC1 93.94% and PC2 3.97% (G), and expression profiles of the representative lncRNAs (H). (I-L) Heat map of the 52 evening lncRNAs (I), BioDare2 phase plots of evening lncRNAs (J), PCA analyses of the evening lncRNAs with variances of PC1 98.53% and PC2 0.79% (K), and expression profiles of the representative lncRNAs (L). (M-P) Heat map of the 41 night lncRNAs (M), BioDare2 phase plots of the night lncRNAs (N), PCA analyses of the night lncRNAs with variances of PC1 99.99% and PC2 0.01% (O), and expression profiles of the representative lncRNAs (P).

described in Section 2.5 (Supplementary Table 12). The expression profiles for the first day were measured at six time points, namely WT96, WT100, WT104, WT108, WT112, and WT116. The expression profile for the second day included another set of six time points, namely WT120, WT124, WT128, WT132, WT136, and WT140. The rhythmicity analysis of these lncRNAs with MetaCycle [19] (Supplementary Table 13) was conducted with parameters set as meta2d (infile="data_file.txt", filestyle="txt", outdir="metacycle analysis", minper=18, maxper=30, timepoints=seq (0, 44, by=4), outIntegration= "onlyIntegration", outRawData=TRUE). LncRNAs with a *p-value* threshold of statistical significance of 0.05 revealed 149 (4.53%) lncRNAs were rhythmically expressed, whereas 3139 lncRNAs were not rhythmically expressed (Supplementary Table 14). The expression profiles of these 3288 lncRNAs and their rhythmicity analysis can be accessed *via* the database

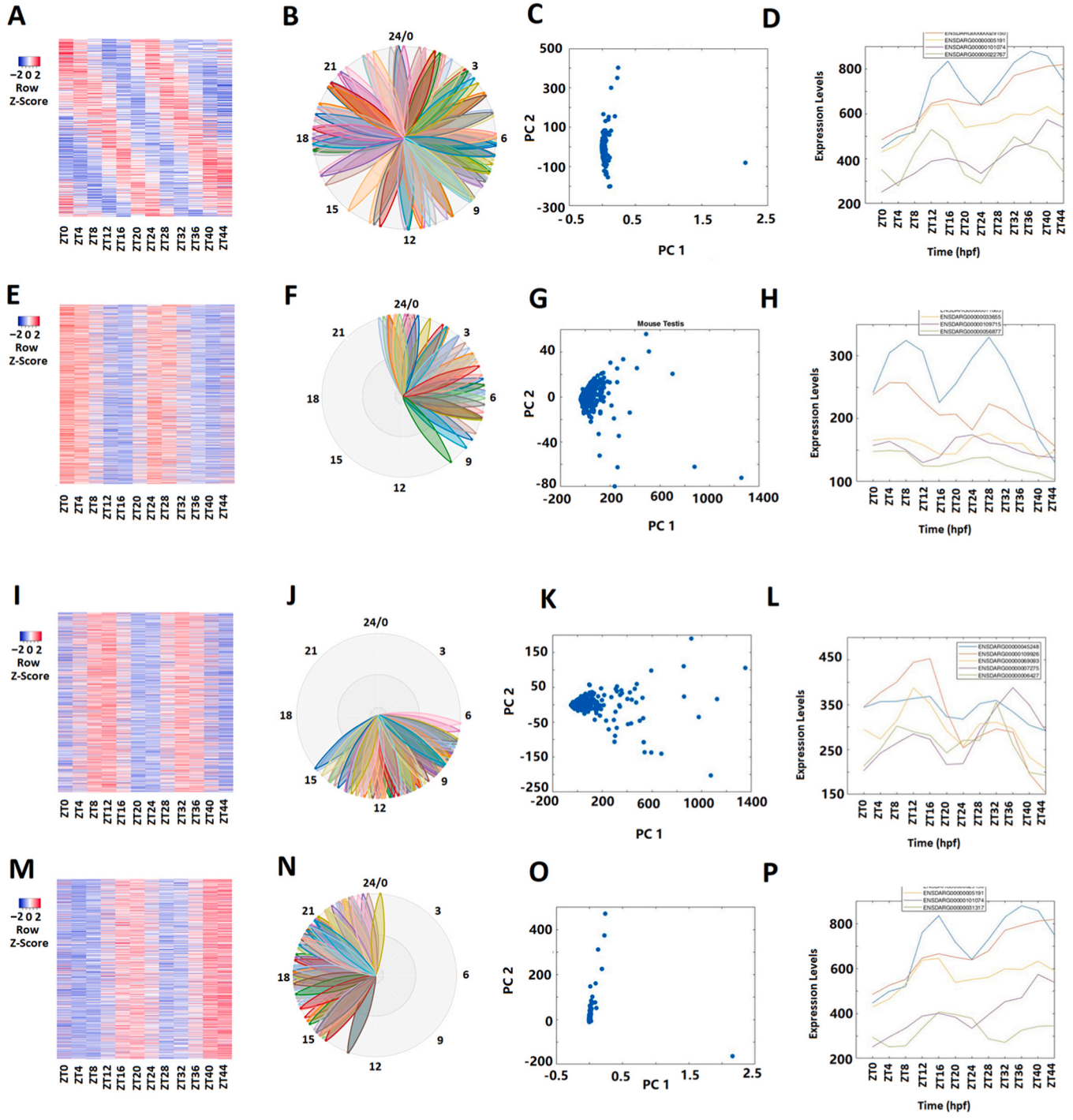

Fig. 3. Analyses of 4047 rhythmically expressed zebrafish mRNAs. (A-D) Analysis of all the 4047 rhythmically expressed zebrafish mRNAs: Heat map (A) of all the 4047 rhythmically expressed mRNAs, BioDare2 phase plots of all mRNAs (B), PCA analyses of all mRNAs with variances of PC1 99.56% and PC2 0.19% (C), and expression profiles of the representative mRNAs (D). (E-H) Analysis of 1817 morning mRNAs: Heat map of the 1817 morning mRNAs (E), BioDare2 phase plots of morning mRNAs (F), PCA analyses of the morning mRNAs with variances of PC1 97.82% and PC2 1.04% (G), and expression profiles of the representative mRNAs (H). (I-L) Heat map of the 1128 evening mRNAs (I), BioDare2 phase plots of evening mRNAs (J), PCA analyses of the evening mRNAs with variances of PC1 97.18% and PC2 1.75% (K), and expression profiles of the representative mRNAs (L). (M-P) Heat map of the 1426 night mRNAs (M), BioDare2 phase plots of the night mRNAs (N), PCA analyses of the night mRNAs with variances of PC1 99.82% and PC2 0.13% (O), and expression profiles of the representative mRNAs (P).

search interface page by searching the corresponding identifiers. The database supports querying both individual identifiers and space-delimited identifiers in the form of batch search. The search results include lncRNAs' time-course expression profiles and their rhythmicity analyses by MetaCycle (Supplementary Figure ^{4}A–^{4}B).

Previous studies have shown gene expression patterns peaking in the morning, evening, and night [20,21]. Therefore, we classified these rhythmically expressed lncRNAs into morning lncRNAs, evening lncRNAs, and night lncRNAs (Fig. 2). Further, we performed the Principal Component Analysis (PCA) [22] on the expression profiles of the rhythmically expressed lncRNAs to select representative lncRNAs for visualization. The PCA analyses identified the two most important principal components of the expression data. The analyses assigned PCA scores to each of the lncRNAs, and we selected the lncRNAs with the highest absolute PCA score of the first principal component and visualized them. Each point on the PCA plot represents a particular lncRNA. The phases of these lncRNAs were visualized with BioDare2 online application (https://biodare2.ed.ac.uk/) [23].

The heat map and BioDare2 plots of these 149 lncRNAs are shown in Fig. 2A-2B. The PCA analyses of these lncRNAs are shown in Fig. 2C, and the representative lncRNAs are shown in Fig. 2D. We further analyzed the peak expression patterns of 149 lncRNAs. The classification of the gene expression profiles revealed (Supplementary Table 14) that out of the 149 lncRNAs, 56 lncRNAs peaked in the morning (ZT0, ZT4, ZT24, and ZT28) (Fig. 2E), 52 lncRNAs peaked in the evening (ZT8, ZT12, ZT32, and ZT36) (Fig. 2I), and 41 lncRNAs peaked in the night (ZT16, ZT20, ZT40, and ZT44) (Fig. 2M). We performed PCA analysis of these morning lncRNAs (Fig. 2G), evening lncRNAs (Fig. 2K), and night lncRNAs (Fig. 2O). The corresponding lncRNAs were ranked based on the corresponding absolute PCA scores from the first principal components. The representative lncRNAs are shown in Fig. 2 for the morning lncRNAs (Fig. 2H), evening lncRNAs (Fig. 2L), and night lncRNAs (Fig. 2P). As observed by the heat maps (Figs. 2E, 2I, and 2M) and BioDare2 plots (Figs. 2F, 3N, and 3N), the morning lncRNAs, evening lncRNAs, and night lncRNAs exhibited different expression patterns.

Subsequently, we performed the GO and KEGG analyses of the rhythmically expressed lncRNAs with known Ensembl IDs (Supplementary Table 15). However, due to the lack of IDs of annotation information of the lncRNAs, we could find annotations for only 11 lncRNAs. The analyses revealed that several lncRNAs are involved in crucial biological processes, such as lncRNAs ENSDARG00000104200 and ENSDARG00000098435 are involved into Serine protease inhibitor, and calcium ion binding, respectively.

2.6. Generation of time-course RNA-seq dataset by whole transcriptome sequencing analysis

2.6.1. Collection of zebrafish larvae and RNA extraction

Approximately 60 wild-type (WT) zebrafish larvae were collected under the DD (constant darkness) condition at the 12 time points with a four-hour interval from 96 to 140 hpf (hours postfertilization), each with two duplicate samples, respectively. Total RNAs were extracted with TRIzol reagent (Invitrogen) from each zebrafish larval sample, respectively. The amount and purity of RNA in each sample were assessed with 1% agarose electrophoresis, Nanodrop2000 spectrophotometer (Wilmington, DE, United States), Qubit® 4.0 Fluorometer (Life Technologies, CA, USA), and Agilent 2100 RNA Nano 6000 Assay Kit (Agilent Technologies, CA, USA).

2.6.2. Sequencing library preparation

After the rRNAs were removed, a fragmentation buffer was added to the extracted RNAs to fragment them into short fragments, and then the first cDNA strand was synthesized with six-base random hexamers and these fragment RNAs. The second strand of cDNA was synthesized by adding buffer, dNTPs (dUTP), RNase H, and DNA polymerase I, purified by QiaQuick PCR kit, and eluted with EB buffer. The cDNA chain containing U was digested with UNG enzyme and amplified by PCR to complete the preparation of the whole library.

2.6.3. Whole transcriptome sequencing and its assembly analysis

Qubit4.0 was employed for preliminary quantification of the sequencing library, which was diluted to 1 ng/ul. Then, Agilent 2100 was used to detect the insert size of the library and accurately quantify the effective concentration of the library (effective concentration of the library $>$ 10 nM) to ensure the quality of the library. The whole transcriptome library was then sequenced with Illumina Novaseq6000 using the PE150 (double-ended 150 bp) sequencing process at Origingene (http://www.origin-gene.com/).

Whole transcriptome sequencing generated a large amount of clean sequence data. Standard HISAT2 [24] was used to align the clean data with the zebrafish reference genome (GRCz11) for subsequent transcript assembly and analysis. StringTie was used to calculate the FPKM value of each gene/transcript in the sample [25]. The FPKM value was regarded as the expression level of the gene/transcript in the sample. The time-course RNA-seq dataset includes 12-time-point 3288 lncRNAs, 25,432 mRNAs, and 314 miRNAs.

2.7. Collection of zebrafish mRNA sequences and querying mRNA sequences

The database contains cDNA sequences of 25,432 mRNAs (Supplementary Table 16). All mRNA sequences with Ensembl IDs were downloaded from the Ensembl Biomart database (http://www.ensembl.org/info/data/biomart/index.html). The 25,432 mRNA sequences can be retrieved by searching the corresponding individual or a batch of Ensembl IDs (Supplementary Table 16) from the database interface (https://sudarna.website/search-zebrafish-mirna-sequences/). The search results display the fasta format of mRNA sequences along with the hyperlinks to the corresponding sequence source from Ensembl database (Supplementary Figure ^{5}A–^{5}B).

2.8. Time-course RNA-seq dataset of zebrafish mRNAs, their rhythmicity analysis, and searching zebrafish mRNAs' expression profiles

We conducted the whole transcriptome sequencing experiment to generate time-course expression profiles for 25,432 mRNAs at 12 time points spread over consecutive two days with a four-hour interval, as described in Section 2.5 (Supplementary Table 17). Specifically, we generate expression profiles for the first day at WT96, WT100, WT104, WT108, WT112, and WT116 time points. The expression profiles for the second day were measured at WT120, WT124, WT128, WT132, WT136, and WT140 (Supplementary Table 18). Rhythmicity analysis found 4407 (17.33%) rhythmically expressed mRNAs and 21,025 arrhythmically expressed mRNAs, respectively (Supplementary Table 19). The expression profiles of 25,432 mRNA can be queried from the SUDAZFLNC interface using Ensembl IDs. The query can be performed for either a single identifier or a group of identifiers. The search results include the expression profile at each of the six time points for two days and their rhythmicity analyses (Supplementary Figure ^{6}A–^{6}B).

Subsequently, we visualized the expression patterns of the rhythmically expressed mRNAs. The heat map, BioDare2 plots, PCA analyses, and representative mRNAs of these 4407 mRNAs are shown in the Fig. 3A-3D. Next, we analyzed the peak expression patterns of 4407 mRNAs. Their classification revealed (Supplementary Table 19) that out of the 4407 mRNAs, 1817 mRNAs peaked in the morning (ZT0, ZT4, ZT24, and ZT28) (Fig. 3E), 1128 mRNAs peaked in the evening (ZT8, ZT12, ZT32, and ZT36) (Figs. 3I), and 1,462 mRNAs peaked in the night (ZT16, ZT20, ZT40, and ZT44) (Fig. 3M). The PCA analyses were performed to identify the representative mRNA among the morning mRNAs (Fig. 3G), evening mRNAs (Fig. 3K), and night mRNAs (Fig. 3O). The representative mRNAs are shown in Fig. 3 for the morning mRNAs (Fig. 3H), evening mRNAs (Fig. 3L), and night mRNAs (Fig. 3P). As

depicted in the heat maps (Figs. 3E, 3I, and 3M) and BioDare2 plots (Figs. 3F, 3N, and 3N), the morning mRNAs, evening mRNAs, and night mRNAs exhibited different expression patterns.

The GO and KEGG pathway analyses were performed for the rhythmically expressed mRNAs. The KEGG pathway enrichment revealed several mRNAs are involved in a variety of pathways (Supplementary Figure 7A–7B), including p53 signaling pathway (Supplementary Figure 8 A), DNA replication (Supplementary Figure 8B), and Cell cycle (Supplementary Figure 8 C). The GO pathway analysis revealed GO terms for numerous mRNAs, including a number of genes involved in biological_process (GO:0008150) (Supplementary Figure 9 A), cellular_component (GO:0005575) (Supplementary Figure 9B), and molecular_function (GO:0003674) (Supplementary Figure 9 C).

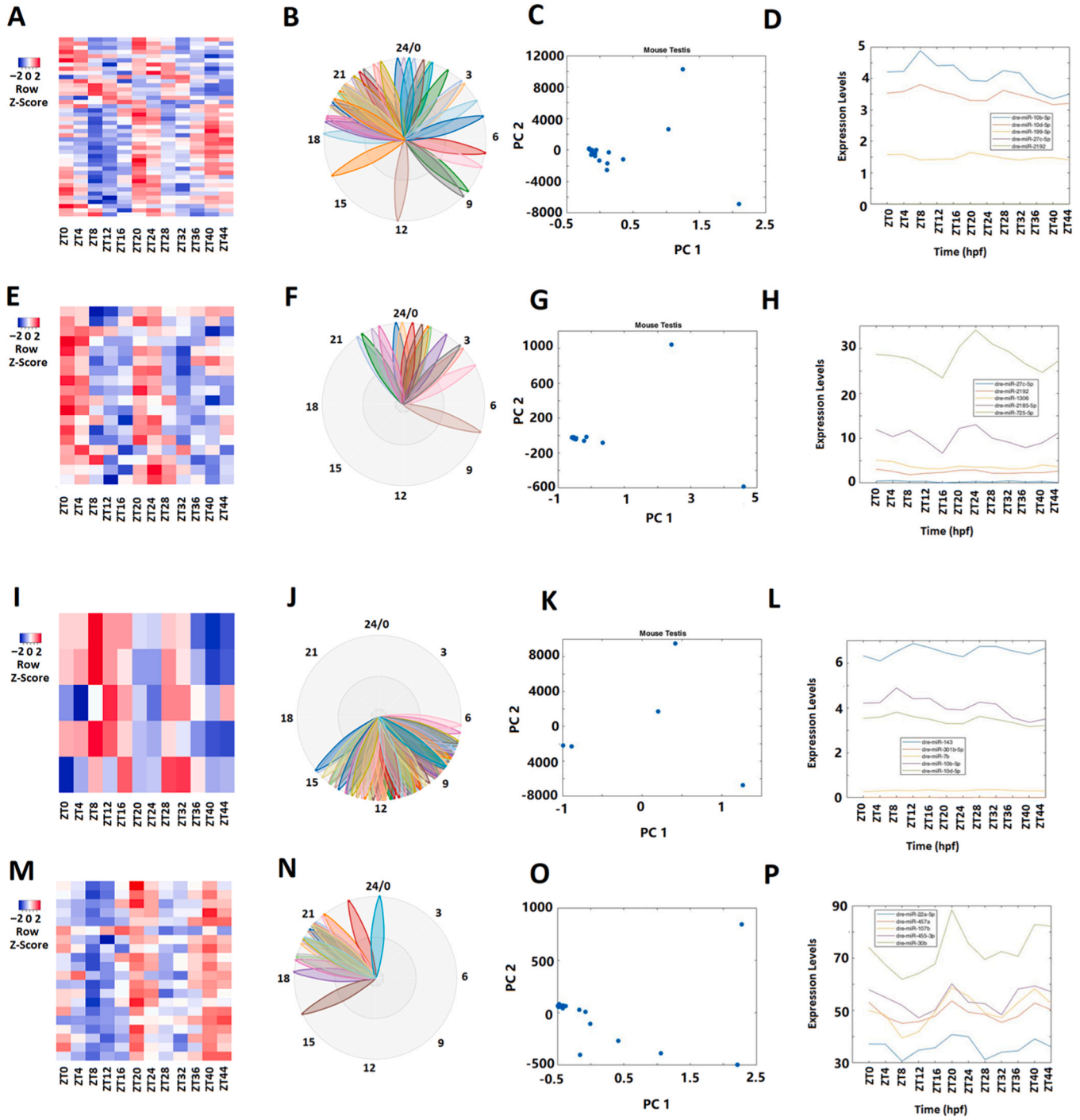

Fig. 4. Analyses of 43 rhythmically expressed zebrafish miRNAs. (A-D) Analysis of all the 4047 rhythmically expressed zebrafish miRNAs: Heat map (A) of all the 43 rhythmically expressed miRNAs, BioDare2 phase plots of all miRNAs (B), PCA analyses of all miRNAs with variances of PC1 99.73% and PC2 0.21% (C), and expression profiles of the representative miRNAs (D). (E-H) Analysis of 18 morning miRNAs: Heat map of the 18 morning miRNAs (E), BioDare2 phase plots of morning miRNAs (F), PCA analyses of the morning miRNAs with variances of PC1 99.94% and PC2 0.04% (G), and expression profiles of the representative miRNAs (H). (I-L) Heat map of the 5 evening miRNAs (I), BioDare2 phase plots of evening miRNAs (J), PCA analyses of the evening miRNAs with variances of PC1 99.56% and PC2 0.41% (K), and expression profiles of the representative miRNAs (L). (M-P) Heat map of the 20 night miRNAs (M), BioDare2 phase plots of the night miRNAs (N), PCA analyses of the night miRNAs with variances of PC1 99.85% and PC2 0.10% (O), and expression profiles of the representative mRNAs (P).

2.9. Collection of zebrafish miRNA sequences and querying miRNA sequences

The SUZFLNC database contains 368 miRNA sequences (Supplementary Table 20). The sequences were downloaded from miRBase database (https://www.mirbase.org/) [26]. The fasta sequences of the 368 miRNAs can be retrieved by searching their identifiers (Supplementary Table 20) from the database interface page. The search results provide the nucleotide sequences for each of the query miRNAs and hyperlinks to their source database (Supplementary Figure ^{10}A–^{10}B).

2.10. Time-course RNA-seq dataset of zebrafish miRNA and their rhythmicity analyses and querying miRNAs' expression profiles

We generated time-course RNA-seq data for 314 miRNAs for consecutive two days with a four-hour interval spread over 12 time points, as described in **Section 2.5** (Supplementary Table 21). Rhythmicity analyses of the miRNA expression profiles with MetaCycle (Supplementary Table 22) revealed 43 (13.69%) rhythmically expressed miRNAs and 299 arrhythmically expressed miRNAs (Supplementary Table 23). The expression profile of each of the 368 miRNAs can be obtained by performing an online query using miRNA sequence identifiers. The search results include two-day time-course expression profiles for each of the queried miRNAs and their rhythmicity analyses with MetaCycle (Supplementary Figure ^{11}A–^{11}B).

Subsequently, the time-course expression profiles of 43 rhythmically expressed miRNAs were visualized (Fig. 4A-4D) with heat map, BioDare2 plots, PCA analyses, and representative miRNAs. We also analyzed the peak expression patterns of 43 miRNAs by classifying them into morning miRNAs, evening miRNAs, and night miRNAs. Their classification revealed (Supplementary Table 23) that out of the 43 miRNAs, 18 miRNAs peaked in the morning (ZT0, ZT4, ZT24, and ZT28) (Figs. 4E), 5 miRNAs peaked in the evening (ZT8, ZT12, ZT32, and ZT36) (Fig. 4I), and 18 miRNAs peaked in the night (ZT16, ZT20, ZT40, and ZT44) (Fig. 4M). The PCA analyses were performed to identify the representative miRNAs among the morning miRNAs (Fig. 4G), evening miRNAs (Fig. 4K), and night miRNAs (Fig. 4O). The representative miRNAs are shown in Fig. 4 for the morning miRNAs (Fig. 4H), evening miRNAs (Fig. 4L), and night miRNAs (Fig. 4P). The heat maps (Figs. 4E, 4I, and 4M) and BioDare2 plots (Figs. 4F, 4N, and 4N), the morning miRNAs, evening miRNAs, and night miRNAs depicted distinct peak expression patterns.

2.11. Computational prediction of mechanisms of RNA-RNA interactions

The SUDAZFLNC database supports the prediction of broadly two types of RNA-RNA interactions, namely base-pairing of RNA-lncRNAs and RNA-mRNA. In particular, users can find the binding sites of any RNA sequences to the 28,925 lncRNA and 25,432 mRNA in the database, including lncRNA-lncRNA interaction (Fig. 5A), lncRNA-mRNA interaction (Fig. 5B), miRNA-lncRNA interaction (Fig. 5C), and miRNA-mRNA interaction (Fig. 5D). The RNA target prediction is implemented in the form of a reverse sequence alignment problem. The input nucleotide sequences are reversed and compared with the BLAST program suite to the target nucleotide database of lncRNA or mRNA. The target search interface allows for the search of the target of input

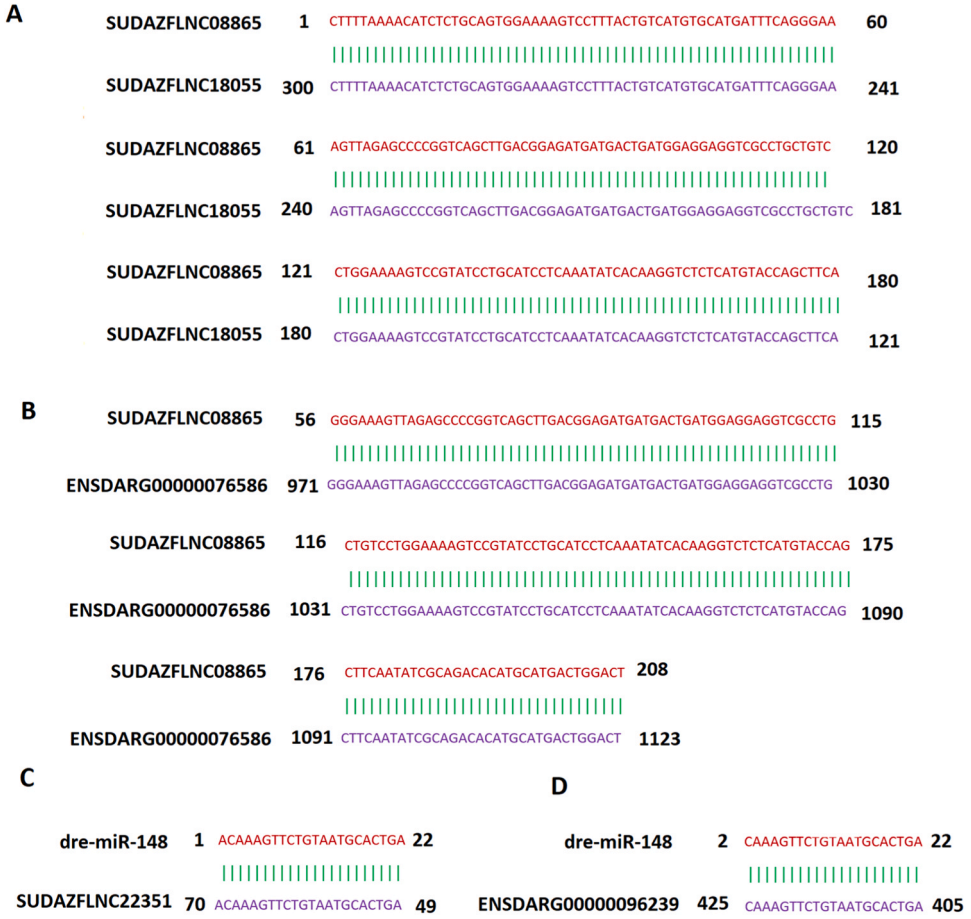

Fig. 5. RNA-RNA interaction and target prediction. LncRNA-lncRNA target prediction: binding of lncRNA SUDAZFLNC08865 to lncRNA SUDAZFLNC18055 (A), lncRNA-mRNA target prediction: binding of lncRNA SUDAZFLNC08865 to mRNA ENSDARG00000076586 (B), miRNA-lncRNA target prediction: binding of miRNA dre-miR-148 to lncRNA SUDAZFLNC22351 (C), and miRNA-mRNA target prediction: binding of miRNA dre-miR-148 to mRNA ENSDARG00000096239 (D).

nucleotide sequences on either 28,925 lncRNAs or 25,432 mRNAs. Further, the checkbox on the query interface page enables target prediction on either strand of lncRNA and mRNA. A suitable threshold value of the expected value and wordsize can be selected from the dropdown list. The fasta format input sequence can be of short RNA sequences, such as miRNAs, or longer sequences, such as lncRNAs (Supplementary Figure ^{12}A–^{12}B). The output of the target binding site prediction query includes nucleotides (C-G) binding sites as well as the statistical significance of the sequence binding.

2.12. Zebrafish lncRNA ID conversion tool

SUDAZFLNC database supports the conversion of lncRNA IDs from Ensembl, NONCODE, NCBI, and ZFLNC to SUDAZFLNC lncRNA IDs (https://sudarna.website/zebrafish-lncrna-id-converter/). A user can input known identifiers such as ENSDART00000152494 NONDRET000440 ZFLNCG00012 separated by space in the search input box, and the tool will display the corresponding SUDAZFLNC IDs and sequences (Fig. 1 and Supplementary Figure ^{13}A–^{13}B).

3. Conclusions

In this study, we establish an online user-friendly database of tens of thousands of zebrafish RNAs and enable several features to query the sequences and experimental profiles. The SUDAZFLNC database provides several advantages over the previously published datasets. For example, SUDAZFLNC includes tens of thousands of curated RNA sequences with unique IDs for all lncRNA sequences, circadian RNA-seq experimental profiles, rhythmicity analysis of the datasets, a user-friendly web interface, and an ID conversion tool. The database includes 28,925 lncRNAs, 25,432 mRNAs, and 368 miRNAs. Through whole transcriptome sequencing experiments, we measure circadian experimental profiles of 3288 lncRNAs, 25,432 mRNAs, and 342 miRNAs, allowing for identifying 149, 4407, and 43 rhythmically expressed lncRNAs, mRNAs, and miRNAs, respectively. Together, our database, integrating experimental profiles and state-of-the-art bioinformatic tools, is the largest repository of zebrafish RNAs and their time-course circadian experimental profiles.

Funding

This work was supported by grants from the National Key Research and Development Program of China (2019YFA0802400), the National Natural Science Foundation of China (NSFC) (#31961133026 and #31871187), Natural Science Foundation of Jiangsu Province (BK20130302), A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PARD).

CRediT authorship contribution statement

Shital Kumar Mishra: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Han Wang:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors have declared that no competing interest exists.

Acknowledgments

We are sincerely grateful to the authors and developers of ZFLNC, NONCODE, and Ensembl databases for providing zebrafish RNA sequences. We also wish to thank the members of the Han Wang laboratory for their invaluable comments and suggestions during the early stages of this study.

Institutional Review Board Statement

All procedures were approved by the Soochow University Animal Care and Use Committee (#SUDA20211013A01).

Informed Consent Statement

Not applicable.

Data Availability Statement

The data supporting the reported results can be found in the supplementary materials.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.04.026.

References

[1] Ben-Moshe Z, Foulkes NS, Gothilf Y. Functional development of the circadian clock in the zebrafish pineal gland. BioMed Res Int 2014;2014. **235781**..
[2] Mishra SK, Liu T, Wang H. Identification of rhythmically expressed LncRNAs in the zebrafish pineal gland and testis. *Int J Mol Sci* 2021;22(15):**7810**.
[3] Zhong, Z., Wang, M., Huang, G., Zhang, S., Wang, H. 2017 Molecular Genetic and Genomic Analyses of Zebrafish Circadian Rhythmicity. Springer India. 193–209.
[4] Mahan, K.E., Corsi, P.S. 2015 Phenotyping Circadian Rhythms in Mice. *Curr Protoc Mouse Biol.*,
[5] Young MW. Life's 24-hour clock: molecular control of circadian rhythms in animal cells. Trends Biochem Sci 2000;25(12):**601**–**6**.
[6] Breed, M.D. 2017 Conceptual Breakthroughs in Ethology and Animal Behavior, Chapter 5 - 1729 Biological Clocks. Academic Press. 15–16.
[7] Ishida N, Kaneko M, Allada R. Biological clocks. PNAS 1999;96:8819–20.
[8] Vatine G, Vallone D, Gothilf Y, Foulkes NS. It's time to swim! Zebrafish and the circadian clock. FEBS Lett 2011;585(10):177–95.
[9] Hirayama J., K.M., Cardone L., Cahill G., Sassone-Corsi P. 2005 Analysis of circadian rhythms in zebrafish. Methods Enzymol 186–204.
[10] Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. The Transcriptional Landscape of the Mammalian Genome. Science 2005;309:1559–63.
[11] Cui M, Zheng M, Sun B, Wang Y, Ye L, Zhang X. A long noncoding RNA perturbs the circadian rhythm of hepatoma cells to facilitate hepatocarcinogenesis. Neoplasia 2015;17:79–88.
[12] Fan Z, Zhao M, Joshi PD, Li P, Zhang Y, Guo W, Xu Y, Wang H f, Zhao Z, Yan J. A class of circadian long non-coding RNAs mark enhancers modulating long-range circadian gene regulation. Nucleic Acids Res 2017;45:5720–38.
[13] Valenzuela-Muñoz, V., Pereiro, P., Álvarez-Rodríguez, M., Gallardo-Escárate, C., Novoa, A.F.B. 2019 Comparative modulation of lncRNAs in wild-type and rag1-heterozygous mutant zebrafish exposed to immune challenge with spring viraemia of carp virus (SVCV). Scientific Reports. 14174,
[14] Chen J, Wang Y, Wang C, Hu J-F, Li W. LncRNA Functions as a New Emerging Epigenetic Factor in Determining the Fate of Stem Cells. Front Genet 2020.
[15] Zhaoa T, Xub J, Liu L, Bai J, Wang L, Xiao Y, Li X, Zhang L. Computational identification of epigenetically regulated lncrnas and their associated genes based on integrating genomic data. FEBS Lett 2015;589(4):521–31.
[16] Zhao L, Wang J, Li Y, Song T, Wu Y, Fang S, Bu D, Li H, Sun L, Pei D. NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. Nucleic Acids Res 2021;49.
[17] Hu, X., Chen, W., Li, J., Huang, S., Xu, X., Zhang, X., Xiang, S., Liu, C. 2018 ZFLNC: a comprehensive and well annotated database for zebrafish lncRNA. Database (doi: 10.1093/database/bay114).
[18] Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: coding-potential assessment tool using an alignment-free logistic regression model. Nucleic Acids Res 2013;41(6).
[19] Wu G, Anafi RC, Hughes ME, Kornacker K, Hogenesch JB. MetaCycle: an integrated r package to evaluate periodicity in large scale data. Bioinformatics 2016;32:3351–3.
[20] Barik S. Molecular Interactions between pathogens and the circadian clock. Int J Mol Sci 2019;20(23).
[21] Doherty CJ, Kay SA. Circadian control of global gene expression patterns. Annu Rev Genet 2010;44:**419**–**44**.
[22] Lever, J., Krzywinski, M., Altman, N. 2017 Principal component analysis. NATURE METHODS |. 14 (7), 641–642.
[23] Zielinski T, Moore AM, Troup E, Halliday KJ, Millar AJ. Strengths and limitations of period estimation methods for circadian data. PLoS One 2014.

[24] Kim Daehwan, Paggi Joseph M, Park Chanhee, Bennett Christopher, Salzberg L, S. Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 2019;37(8):**907**–**15**.

[25] Mihaela Pertea GMP, Antonescu Corina M, Chang Tsung-Cheng, Mendell Joshua T, Salzberg Steven L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;33(3):**290**–**5**.

[26] Ana Kozomara M.B. Sam Griffiths-Jones miRBase: from microRNA sequences to function 47 D1 2019 **D155 D162** .