riboCIRC: a comprehensive database of translatable circRNAs

Huihui Li[1,2], Mingzhe Xie[1,2], Yan Wang[1,2], Ludong Yang[1], Zhi Xie[1,3], Hongwei Wang[1,3]

[1] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China

[2] These authors contributed equally to this work

[3] Correspondence: bioccwhw@126.com or xiezhi@gmail.com

Abstract

riboCIRC is a translatome data-oriented circRNA database specifically designed for hosting, exploring, analyzing, and visualizing translatable circRNAs from multi-species. The database provides a comprehensive repository of computationally predicted ribosome-associated circRNAs, a manually curated collection of experimentally verified translated circRNAs, an evaluation of cross-species conservation of translatable circRNAs, a systematic *de novo* annotation of putative circRNA-encoded peptides, including sequence, structure, and function, and a genome browser to visualize the context-specific occupant footprints of circRNAs. It represents a valuable resource for the circRNA research community and is publicly available at http://www.ribocirc.com.

Keywords

CircRNAs; Translatable circRNAs; Ribosome profiling; circRNA-encoded peptides

Background

Circular RNAs (circRNAs) are an abundant class of covalently closed endogenous RNA molecules generated by back-splicing of pre-mRNAs. Recent advances in computational analysis and high-throughput RNA sequencing (RNA-seq) have unveiled a detailed view of circRNA biogenesis, regulatory mechanisms and cellular functions [1]. With the development of various computational and experimental approaches to effective identification of circRNAs, many dedicated databases for circRNAs were constructed, such as circBase and circAtlas for vertebrate circRNAs [2,3], CSCD and TSCD for disease/tissue-specific circRNAs [4,5], and Circ2Disease and Circ2Traits for circRNA-disease associations [6,7]. These transcriptome data-oriented databases provide essential information about circRNAs, facilitating the current understanding of circRNAs related to their biological importance and clinical relevance. It becomes increasingly clear that circRNAs can regulate multiple biological processes via a variety of mechanisms. For instance, circRNAs can act as ‘sponges’ or ‘decoys’ for microRNAs or RNA-binding proteins to modulate gene expression or mRNA translation [8–10].

circRNAs are generally considered as ‘non-coding’ elements; however, circRNAs can in fact serve as templates for protein translation. Using ribosome profiling (Ribo-seq) that enables genome-wide investigation of *in vivo* translation at a sub-codon resolution [11,12], a subset of circRNAs have recently been identified to be associated with translating ribosomes [13,14]. Furthermore, by performing *in vivo* and *in vitro* translation assays, circRNAs have been shown to enable cap-independent translation and generate functional proteins. Of these proteins, some have been demonstrated to play vital roles under a number of pathophysiological conditions, such as muscle-enriched circRNA circ-ZNF609 [10] and brain-ubiquitously expressed circRNA circAβ-a [15]. In addition, several mechanisms for circRNAs translation have been proposed. For instance, internal ribosome entry site (IRES)- and N6-methyladenosines (m6A)-mediated cap-independent translation initiation are potential mechanisms for circRNA translation [16,17].

Although translation of circRNAs has attracted considerable attention and a large number of the Ribo-seq datasets have been generated in the past several years [18], there is no translatome data-oriented database that aims to provide direct *in vivo* translation evidence for multi-species circRNAs to date. To fill the gap, we analyzed the 3,168 publicly available Ribo-seq and 1,970 matched RNA-seq datasets from 314 studies covering 21 various species to determine the prevalence of circRNA translation. We further provided a dedicated multi-species translatable circRNA database, riboCIRC, towards a comprehensive repository of computationally predicted and experimentally verified translatable circRNAs. Overall, the riboCIRC database provides an important resource for the circRNA research community and can serve as a useful starting point for further investigation of the details of circRNA function and their involvement in cellular processes and diseases.

Construction and content

The backend of riboCIRC is powered by MySQL and accessed using the PHP framework as the middleware. The MySQL includes three large tables: the first table stores all information about the available computationally predicted ribosome-associated circRNAs in the database; the second table stores all information about the manually curated experimentally verified circRNAs; and the third table stores all annotation information about the putative circRNA-encoded peptides. The front-end of riboCIRC is a multi-page web application built using HTML, JavaScript, and CSS code that consists of numerous pages with static information, such as text and image. The entire database is deployed and running on Amazon AWS EC2 platform. The current version of riboCIRC records a total of 2,247 computationally predicted ribosome-associated circRNAs and 216 experimentally verified translated circRNAs from six different species. The web application of riboCIRC is designed for hosting, exploring, analyzing and visualizing these translatable circRNAs (**Figure 1**). The home page provides general information about riboCIRC and quick links to access translatable circRNAs, circRNA-encoded peptides, and footprint visualization of circRNAs. Other primary features of riboCIRC can be accessed using various buttons of the website.

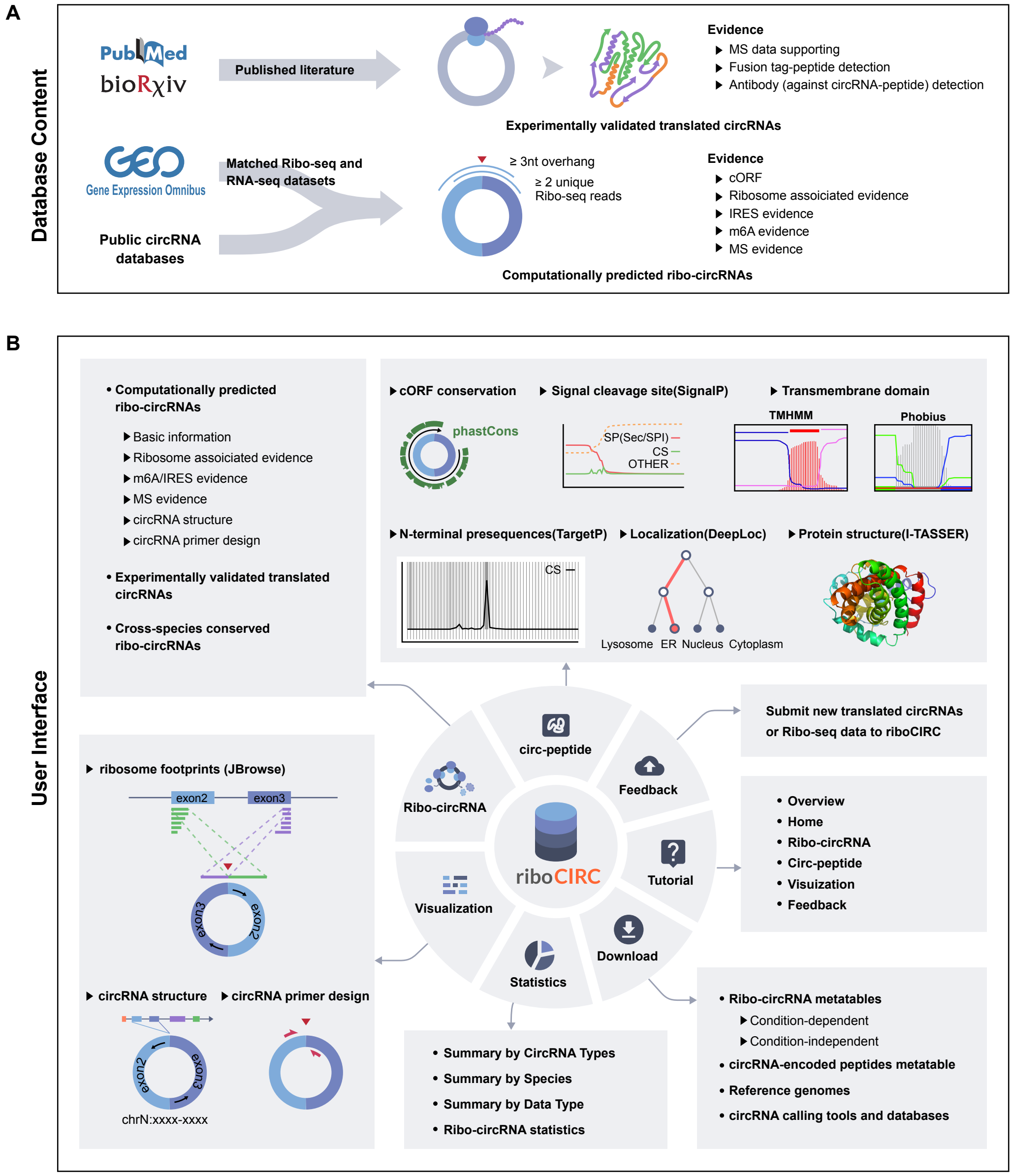

Figure 1. Overview of the riboCIRC database. (a) Schematic view of the database content. (b) Schematic view of the database interface.

Data collection and preprocessing

We collected 3,168 publicly available Ribo-seq datasets and 1,970 matched RNA-seq datasets of the same samples from 314 studies covering 21 species, including *Arabidopsis*, *Caenorhabditis elegans*, *Caulobacter crescentus*, *Cryptococcus neoformans*, Chinese hamster, *Drosophila*, *Escherichia coli*, *Halobacterium salinarum*, human, mouse, *Plasmodium falciparum*, *Pseudomonas aeruginosa*, rat, *Saccharomyces cerevisiae*, *Salmonella enterica*, *Schizosaccharomyces pombe*, *Streptomyces coelicolor*, *Staphylococcus aureus*, *Trypanosoma brucei*, *Vibrio vulnificus*, and zebrafish (see **Additional file 1: Table S1**). After downloading the raw data files from the NCBI SRA database [19], we applied a unified pipeline to perform preprocessing of the Ribo-seq and RNA-seq data. Briefly, the 3'-end adapters were clipped using Cutadapt (version 1.8.1) [20]; Low-quality bases were trimmed using Sickle (version 1.33) [21]; and the retained reads that mapped to rRNAs or tRNAs were removed.

Detection of transcribed and ribosome-associated circRNAs

We combined three different detection tools to identify transcribed circRNAs in each RNA-seq dataset, namely, CIRCexplorer2, CIRI2, and DCC [22–24]. The full-length sequence of each identified circRNA was assembled by the CIRI-full pipeline [25] or extracted from the circAtlas database [2,3] when RNA-seq data were unavailable. Taking advantage of these full sequences, we generated a pseudo circRNA reference for each species by initial extraction of the 23-base pair (bp) sequences on either side of the backsplice junction (BSJ) site of each transcribed circRNA with subsequent concatenation of the two-sided sequences. To identify ribosome-associated circRNAs (ribo-circRNAs), we first eliminated sequence reads corresponding to nonribosomal RNA-protein complexes in each Ribo-seq dataset using Rfoot (version 1.0) [26], considering that ribosomes are not specifically selected during the biochemical isolation procedure of ribosome profiling experiment. After removal of footprints from nonribosomal complexes, all the ribosome-protected footprints were then mapped with Tophat2 (version 2.1.1) [27] to the corresponding linear reference genome, and further the resulting unmapped.bam files were remapped to the pseudo circRNA reference using Tophat2 (version 2.1.1) [27] with default parameters except N, which was set to 0 (the default is 2). Finally, a circRNA was defined to be associated with translating ribosomes only when it met all of the following three criteria simultaneously: (1) at least two unique backsplice junction-spanning Ribo-seq reads, (2) a minimum read-junction overlap of three nucleotides (nt) on either side of the backsplice junction site, and (3) a typical range of read lengths of 25-35 nt (see **Additional file 2: Figure S1**).

Two different strategies were here used to characterize ribo-circRNAs: (1) condition-dependent detection for Ribo-seq and perfectly matched RNA-seq datasets and (2) condition-independent detection for previously reported circRNAs and Ribo-seq datasets. The former strategy was applied to the initial genome-wide characterization of transcribed circRNAs using 1,922 RNA-seq datasets with subsequent examination of ribosome associations of these circRNAs using 1,970 Ribo-seq datasets from the same samples. In total, 278 out of the 91,143 transcribed circRNAs were identified as ribo-circRNAs, involving four different species (*Drosophila*, human, mouse, and rat). The latter strategy was applied to the systemic examination of ribosome associations of the circRNAs reported in the public databases using 3,168 Ribo-seq datasets. To accomplish this task, we selected nine out of the 18 examined public circRNA databases, including circAtlas, circBank, circBase, CIRCpedia, circRNADb, CSCD, exoRBase, TSCD and Circ2Disease [2–6,28–31], and obtained 1,411,865 unique circRNAs after conversion of their coordinates using LiftOver [32]. Notably, the other public circRNA databases were excluded from this analysis due to lack of a batch download link, incomplete annotation or inaccessible webpage (see **Additional file 3: Table S2**). Among these well-documented circRNAs, a total of 1,969 circRNAs were finally identified as ribo-circRNAs, involving six different species (*C. elegans*, *Drosophila*, human, mouse, rat, and zebrafish).

Cross-species conservation analysis of translatable circRNAs

To evaluate translatable circRNA conservation among different species, we first annotated the parental genes of ribo-circRNAs using the GTF files, and then identified orthologous gene pairs expressing these circRNAs using a pairwise orthologous gene list downloaded from the OMA orthology database (http://omabrowser.org) [33]. After that, we extracted 50-bp fragments on either side of the ribo-circRNA BSJ site from the reference genome and further concatenated both fragments to represent the ribo-circRNA BSJ sequence. Next, all ribo-circRNA BSJ sequences in one species were aligned to those of the other species using BLAT with default parameters [34], followed by a reciprocal best hit strategy to find the orthologous ribo-circRNAs.

Finally, a pair of conserved ribo-circRNAs were defined based on their sequence alignment length ≥ 80 and alignment bit-score ≥ 150.

Prediction of circRNA-derived ORFs

We predicted putative circRNA-derived ORFs (cORFs) for each ribo-circRNA using the cORF_prediction_pipeline with some modifications [13]. Briefly, the full-length sequence of each ribo-circRNA was retrieved and multiplied four times to allow for rolling circle translation. All cORFs beginning with an AUG initiation codon were identified separately for each circRNA, and further filtered based on the requirements of a minimum length of 20 amino acids (aa) and of spanning the backsplice junction site. Notably, those cORFs terminating without an in-frame stop codon were defined as INF (infinite)-cORFs, representing that the corresponding circRNAs could be translated via a rolling circle amplification mechanism. Finally, only the longest cORF was kept for each one of the three reading frames, considering that circRNA with a long ORF would have a better chance of undergoing translation.

Annotation of IRES elements and m6A sites in circRNAs

Given that previous studies have shown the ability of IRES elements and m6A modification to drive circRNA translation [13,16], we predicted potential IRES elements and m6A sites in circRNAs by using publicly available IRES sequences and m6A modification data. To identify IRES elements in circRNAs, we extracted experimentally validated IRES sequences from the IRESbase database [35], and then aligned them to circRNA sequences using BLASTN (version 2.7.1+) [36] with at least 80% sequence identity and a cutoff 30 nucleotides alignment length. To identify potential m6A sites in circRNAs, we extracted m6A modification peaks detected by three different peak calling tools (exomePeak, MeTPeak, and MACS2) from the REPIC database [37], followed by aligning them to circRNA sequences and the presence of m6A consensus motif ‘RAC’ (where R is any purine) in the aligned positions.

Annotation of cORF-encoded peptides

We constructed a semi-automated bioinformatic workflow system to perform *de novo* annotation of all putative cORF-encoded peptides, including sequence conservation, transmembrane topology, signal cleavage site, subcellular localization, folding structure, potential function, etc.

Specifically, sequence conservation of each putative cORF-encoded peptides was computed by an in-house Python script based on the phastCons score files at the University of California Santa Cruz (UCSC) [38]. The presence or absence of the signal peptide cleavage sites was predicted by SignalP (version 5.0b) with default parameters [39]. Transmembrane helical topology was predicted by TMHMM (version 1.1) [40] and Phobius (version 1.01) [41] with default parameters. The N-terminal presequences, such as signal peptide (SP), mitochondrial transit peptide (mTP), chloroplast transit peptide (cTP) or thylakoid luminal transit peptide (luTP), were predicted by TargetP (version 2.0) [42] with default parameters. Subcellular localization was predicted by DeepLoc (version 1.0) [43] with default parameters, which can differentiate between 10 different localizations, including nucleus, cytoplasm, extracellular, mitochondrial, cell membrane, endoplasmic reticulum, chloroplast, Golgi apparatus, lysosome/Vacuole, and peroxisome. The 3D structure was predicted by I-TASSER (version 5.1) [44] that also provided other information, such as secondary structure, solvent accessibility, normalized B-factor and Top 10 threading templates.

Detection of cORF-encoded peptides by mass spectrometry

We used public proteomics data to find protein evidence of putative cORF-encoded peptides. Briefly, the raw files were of 26 datasets downloaded from the PRIDE database [45] (see **Additional file 4: Table S3**) and analyzed using MaxQuant software (version 1.6.15.0) [46] against a custom-tailored database separately for each species (the respective size for human: n=22,113; mouse: n=18,308; rat: n=8,159; *Drosophila*: n=3,629; *C.elegans*: n=4,156; and zebrafish: n=3,165), which combined all documented sequences from UniProt/Swiss-Prot with additional sequences derived from circRNA translation, based on the target-decoy strategy (Reverse) with the standard search parameters with the following exceptions: (1) the peptide-level FDR was set to 5%, and the protein-level FDR was excluded; (2) the minimal peptide length was set to seven amino acids; and (3) a maximum of two missed cleavages were allowed. For each search, fixed modifications and variable modifications were customized according to different proteomics data. In total, 719 cORF-encoded proteins from 669 circRNAs were evidenced by at least one unique junction-spanning peptide.

Primer design and structure representation of circRNAs

Based on the sequence of each circRNA, we performed circRNA-specific primer design. Divergent primer sets spanning the backsplice junction sequence were generated using circtools [47]. The graphical representations of circRNAs and their linear host transcripts were constructed using circView [48].

Collection of experimentally verified translated circRNAs

Experimentally verified translated circRNAs were manually curated from the literature. To accomplish this task, we searched the PubMed literature database using the keyword '(circRNA [MeSH terms]) AND (translation [MeSH terms])' and the bioRxiv preprint server using the keyword '("circRNA"+"translation")', and found a total of 65 relevant published or preprint references. After retrieving the full text of these references, we reviewed the studies to manually collect the circRNA entries, which generated the peptides and were validated by various experiments. Strict screening identified 216 translated circRNAs with mass spectrometry-derived detection of the corresponding peptides, tag-peptide fusion system detection, or/and antibody (against circRNA-peptide) detection evidence and incorporated all information into the riboCIRC database. Additional basic information on these experimentally verified circRNAs was also collected, including the circRNA name, circBase id, genomic coordinates, strand, host gene, transcript, species/condition, circRNA-encoded peptide sequence, peptide length, experimental method, and reference information.

Utility and discussion

Exploration of translatable circRNAs

The ribo-circRNA page provides a comprehensive repository of translatable circRNAs, including computationally predicted ribosome-associated circRNAs and experimentally verified translated circRNAs. Users can click the 'Ribo-circRNA' button on the navigation bar and then select one of the dropdown-menu options (including 'Computationally predicted ribo-circRNAs', 'Experimentally verified translated circRNAs', and 'Cross-species conserved ribo-circRNAs') for a quick query.

Selection of 'Computationally predicted ribo-circRNAs' returns the result page containing all predicted ribo-circRNAs identified using Ribo-seq data, including 1,969 condition-independent and 278 condition-dependent ribo-circRNAs. Brief descriptions of these circRNAs are shown in this results page, including riboCIRC id, chromosome position, best transcript, host gene symbol, and circRNA length. A built-in search box can narrow the results down to a particular subject by entering additional search terms. Furthermore, clicking the riboCIRC id in the second column opens a separate page for every circRNA that displays detailed information on the matching circRNA, including cORF annotation such as the location of the junction-spanning cORF in the genome, total number of junction-spanning footprints, unique number of junction-spanning footprints, translation conditions, involved dataset, cORF sequence, cORF-encode peptide, and length of cORF-encode peptide, evidence for translation of circRNAs such as IRES element, m6A site, and mass spectrometric proof, graphical representation of the linear and circular RNA structure, and designed circRNA primer sets. Clicking the chromosome position in the third column opens a separate page for visualization the host gene track, genomic features and aligned junction-spanning ribosome footprints of the circRNA in the JBrowse [49]. Selection of 'Experimentally verified translated circRNAs' returns the result page with all 216 experimentally verified translated circRNAs that have been validated by various experiments to generate peptides. All information collected on these circRNAs is shown with some additional relevant information on the validated circRNAs accessible via the hyperlinks provided on the result page. In addition, selection of 'Cross-species conserved ribo-circRNAs' returns the result page with all cross-species inference of conserved translatable circRNA pairs that involve a total of 90 evolutionarily conserved ribo-circRNAs.

Comprehensive analysis of circRNA-encoded peptides

The circ-peptide page provides a systematic annotation of putative circRNA-encoded peptides, including their sequence, structure and function. Users can click the 'Circ-peptide' button on the navigation bar to quickly browse the putative circRNA-encoded peptides. A dropdown menu shows a list of the available circRNA-encoded peptide options, and users can select one of the options to retrieve additional information, including basic information on the given peptide (sequence, and conservation), summary of peptide characteristics (signal cleavage site, transmembrane domain, and N-terminal presequence), and location and topology of the peptide (subcellular localization, secondary structure, and structural conformation).

Intuitive visualization of ribosome-associated circRNAs

The visualization page provides an intuitive view of ribo-circRNAs, including visualization of the host gene track, genomic features and aligned junction-spanning ribosome footprints of the circRNAs. Users can click the 'Visualization' button on the navigation bar to visualize the data on the features of ribosome-associated circRNAs in JBrowse browser [49] embedded in the result page. A cascading dropdown menu consists of three independent selection dropdown buttons for quick navigation to a circRNA of interest, interactive exploration of the data, and intuitive comparison of the data originating from various datasets.

Data download, statistics, user guide, and feedback

The download page provides access to a convenient tabular data format. Users can click the 'Download' button on the navigation bar to easily access the data. Tabular list of metadata for all computationally predicted ribo-circRNAs, circRNA sequences, nucleotide sequences of all cORFs and their corresponding protein sequences, as well as customized protein sequence databases for proteomics search can be freely downloaded for nonprofit and academic purposes. In addition, the Ribo-circRNA page also provides the download buttons for download of computationally predicted ribosome-associated or experimentally verified translated circRNAs in various formats, including JSON, XML, CSV, TXT, SQL and MS-Excel. The statistics page provides a summary that summarizes the data in all accessible records of the database. The tutorial page provides step-by-step instructions for users to familiarize themselves with the database. The feedback page provides a feedback form for translatable circRNAs and Ribo-seq datasets, making it easy for users to provide feedback.

Potential Limitation

It should be noted here that traditional approaches based on the properties of active translation such as three-nucleotide periodic subcodon pattern are not feasible for identifying active translated circRNAs due to the difficulty in distinguishing circular and linear Ribo-seq reads. In this database, we also adopt a similar strategy of translatable circRNA detection as previously described [13,14], where ribosome-associated circRNAs were identified only by Ribo-seq reads spanning a head-to-tail splice junction. However, computational prediction through this strategy does not necessarily mean that the circRNA is being actively translated into a detectable micropeptide, even though it is associated with translating ribosomes. This database is just a starting point for bench scientists and computational biologists to pursue translatable circRNAs. The translated details of individual circRNAs still have to be further experimentally validated.

Future directions

In the future, riboCIRC will be periodically updated. Increasing availability of new high-throughput Ribo-seq data will be used to characterize the putative translation of circRNAs and expand the size of computationally predicted ribo-circRNAs. We will continue to fill the database with new reported experimentally verified translated circRNAs. In addition, we will continue to extend our collection of public proteomics data and to further enhance the identification rate of cORF-encoded peptides. These additions are anticipated to enhance efficiency of the applications of riboCIRC in the circRNA research community.

Conclusions

To the best of our knowledge, riboCIRC is the first database for hosting, exploring, analyzing, and visualizing translatable circRNAs for multi-species. The database provides a comprehensive repository of computationally predicted ribo-circRNAs, together with multiple lines of evidence supporting their translation, and experimentally verified translated circRNAs. It also provides an evaluation of cross-species conversed translatable circRNAs, a systematic functional annotation of the putative circRNA-encoded peptides, a flexible visualization framework for ribosome-associated circRNAs, and a user-friendly web interface for easy data access and exploration. Thus, riboCIRC will serve as a valuable resource for bench scientists and computational biologists to explore translatable circRNAs and to drive functional investigation of the circRNA translation.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

riboCIRC is available at http://www.ribocirc.com to all users without any login or registration restrictions. All public Ribo-seq, RNA-seq, and mass spectrometry datasets used during the current study are available in **Additional file 1: Table S1 and Additional file 4: Table S3**. All translatable circRNAs can be downloaded from the riboCIRC data download page.

Competing interests

The authors declare no competing financial interests.

Funding

This work was supported, in part, by the National Natural Science Foundation of China under Award Number 31871302 to Z.X. and the Joint Research Fund for Overseas Natural Science of China under Award Number 31829002 to Z.X.

Authors' contributions

H.W.W. and Z.X. directed the project and wrote the manuscript. H.H.L, Y.W. and L.D.Y. performed the data analyses and result presentation. H.H.L. and M.Z.X. designed and constructed the database. All named authors read and approved the final manuscript.

Acknowledgments

We would like to thank our team members Congying Chen and Liang Yi for their assistance in collecting translated circRNAs. We also thank the support from the Center for Precision Medicine at Sun Yat-sen University.

References

1. Kristensen LS, Andersen MS, Stagsted LVW, Ebbesen KK, Hansen TB, Kjems J. The biogenesis, biology and characterization of circular RNAs. Nat Rev Genet. 2019;20:675-91.

2. Glažar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA. 2014;20:1666-70.

3. Wu W, Ji P, Zhao F. CircAtlas: an integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. Genome Biol. 2020;21:101.

4. Xia S, Feng J, Chen K, Ma Y, Gong J, Cai F, et al. CSCD: a database for cancer-specific circular RNAs. Nucleic Acids Res. 2018;46:D925-9.

5. Xia S, Feng J, Lei L, Hu J, Xia L, Wang J, et al. Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. Brief Bioinform. 2017;18:984-92.

6. Yao D, Zhang L, Zheng M, Sun X, Lu Y, Liu P. Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. Sci Rep. 2018;8:11018.

7. Ghosal S, Das S, Sen R, Basak P, Chakrabarti J. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. Front Genet. 2013;4:283.

8. Piwecka M, Glažar P, Hernandez-Miranda LR, Memczak S, Wolf SA, Rybak-Wolf A, et al. Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. Science. 2017;357:eaam8526.

9. Abdelmohsen K, Panda AC, Munk R, Grammatikakis I, Dudekula DB, De S, et al. Identification of HuR target circular RNAs uncovers suppression of PABPN1 translation by CircPABPN1. RNA Biol. 2017;14:361-9.

10. Legnini I, Di Timoteo G, Rossi F, Morlando M, Briganti F, Sthandier O, et al. Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis. Mol Cell. 2017;66:22-37.

11. Brar GA, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat Rev Mol Cell Biol. 2015;16:651-64.

12. Ingolia NT. Ribosome Footprint Profiling of Translation throughout the Genome. Cell. 2016;165:22-33.

13. Pamudurti NR, Bartok O, Jens M, Ashwal-Fluss R, Stottmeister C, Ruhe L, et al. Translation of CircRNAs. Mol Cell. 2017;66:9-21.

14. van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, et al. The Translational Landscape of the Human Heart. Cell. 2019;178:242-60.

15. Mo D, Li X, Raabe CA, Rozhdestvensky TS, Skryabin BV, Brosius J. Circular RNA Encoded Amyloid Beta peptides-A Novel Putative Player in Alzheimer's Disease. Cells. 2020;9:2196.

16. Yang Y, Fan X, Mao M, Song X, Wu P, Zhang Y, et al. Extensive translation of circular RNAs driven by N6-methyladenosine. Cell Res. 2017;27:626-41.

17. Fan X, Yang Y, Wang Z. Pervasive translation of circular RNAs driven by short IRES-like elements. bioRxiv. 2019; doi:10.1101/473207.

18. Wang H, Yang L, Wang Y, Chen L, Li H, Xie Z. RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. Nucleic Acids Res. 2019;47:D230-4.

19. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40:D54-6.

20. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10-2.

21. Joshi NA, Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. 2011. https://github.com/najoshi/sickle. Accessed 14 Feb 2020.

22. Zhang X-O, Dong R, Zhang Y, Zhang J-L, Luo Z, Zhang J, et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res. 2016;26:1277-87.

23. Gao Y, Zhang J, Zhao F. Circular RNA identification based on multiple seed matching. Brief Bioinform. 2018;19:803-10.

24. Cheng J, Metge F, Dieterich C. Specific identification and quantification of circular RNAs from sequencing data. Bioinformatics. 2016;32:1094-6.

25. Zheng Y, Ji P, Chen S, Hou L, Zhao F. Reconstruction of full-length circular RNAs enables isoform-level quantification. Genome Med. 2019;11:2.

26. Ji Z, Song R, Huang H, Regev A, Struhl K. Transcriptome-scale RNase-footprinting of RNA-protein complexes. Nat Biotechnol. 2016;34:410-3.

27. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.

28. Liu M, Wang Q, Shen J, Yang BB, Ding X. Circbank: a comprehensive database for circRNA with standard nomenclature. RNA Biol. 2019;16:899-905.

29. Dong R, Ma X-K, Li G-W, Yang L. CIRCpedia v2: An Updated Database for Comprehensive Circular RNA Annotation and Expression Comparison. Genomics Proteomics Bioinformatics. 2018;16:226-33.

30. Chen X, Han P, Zhou T, Guo X, Song X, Li Y. circRNADb: A comprehensive database for human circular RNAs with protein-coding annotations. Sci Rep. 2016;6:34985.

31. Li S, Li Y, Chen B, Zhao J, Yu S, Tang Y, et al. exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. Nucleic Acids Res. 2018;46:D106-12.

32. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013;14:144-61.

33. Altenhoff AM, Glover NM, Train C-M, Kaleb K, Warwick Vesztrocy A, Dylus D, et al. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Res. 2018;46:D477-85.

34. Kent WJ. BLAT-The BLAST-Like Alignment Tool. Genome Res. 2002;12:656-64.

35. Zhao J, Li Y, Wang C, Zhang H, Zhang H, Jiang B, et al. IRESbase: A Comprehensive Database of Experimentally Validated Internal Ribosome Entry Sites. Genomics Proteomics Bioinformatics. 2020;18:129-39.

36. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res. 2008;36:W5-9.

37. Liu S, Zhu A, He C, Chen M. REPIC: a database for exploring the N6-methyladenosine methylome. Genome Biol. 2020;21:100.

38. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20:110-21.

39. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019;37:420-3.

40. Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305:567-80.

41. Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. Nucleic Acids Res. 2007;35:W429-32.

42. Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, et al. Detecting sequence signals in targeting peptides using deep learning. Life Sci Alliance. 2019;2:e201900429.

43. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 2017;33:3387-95.

44. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Res. 2015;43:W174-81.

45. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019;47:D442-50.

46. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26:1367-72.

47. Jakobi T, Uvarovskii A, Dieterich C. circtools-a one-stop software solution for circular RNA research. Bioinformatics. 2019;35:2326-8.

48. Feng J, Xiang Y, Xia S, Liu H, Wang J, Ozguc FM, et al. CircView: a visualization and exploration tool for circular RNAs. Brief Bioinform. 2019;20:745-51.

49. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 2016;17:66.

Additional files

Additional file 1: Table S1. Summary of Ribo-seq and matched RNA-Seq datasets used in this study.

Species	No. of studies	No. of RNA-seq samples	No. of Ribo-seq samples
Human	84	621	636
	GSE21992,GSE69047,GSE69602,GSE69906,GSE59820,GSE63591,GSE70211,GSE62247,GSE35469,GSE73565,GSE77292,GSE77315,GSE77317,GSE77347,GSE78959,GSE78960,GSE79664,GSE79804,GSE81802,GSE82232,GSE83493,GSE41605,GSE42509,GSE87328,SRA492656,GSE94454,GSE45785,GSE96643,GSE96714,GSE96716,GSE97140,GSE97384,GSE46613,GSE100007,GSE101760,GSE102040,GSE48785,GSE48933,GSE49339,SRA096542,GSE49716,SRA099816,GSE51584,GSE52447,GSE52809,GSE55195,GSE56148,GSE56887,GSE56924,GSE59817,GSE59818,GSE59819,GSE60426,GSE51424,GSE61375,GSE63570,GSE64962,GSE65778,GSE65885,GSE65912,GSE66809,GSE67902,GSE133111,GSE112705,GSE121391,GSE125086,GSE129869,GSE112085,GSE125114,GSE127713,GSE121952,GSE113695,GSE123564,GSE112305,GSE115647,GSE123539,GSE122071,GSE112295,GSE113171,GSE118239,GSE111866,GSE133925,GSE106483,GSE119615		
Mouse	57	551	565
	PRJEB12126,PRJEB17636,PRJEB7207,PRJEB7276,GSE22001,GSE30839,GSE68265,GSE69699,GSE69800,GSE71333,GSE72064,GSE72066,GSE36892,GSE37111,GSE74537,GSE74683,GSE60930,GSE80156,GSE81283,GSE83332,GSE83351,GSE84112,GSE41246,GSE41785,GSE50983,GSE89011,GSE89108,GSE89184,GSE94385,GSE99787,GSE102659,GSE102890,GSE51424,GSE52809,GSE53743,GSE58423,GSE60426,GSE67305,GSE112185,GSE108331,GSE116221,GSE114064,GSE110618,GSE105147,GSE112223,GSE97286,GSE123919,GSE120762,GSE112502,GSE120097,GSE119365,GSE112766,GSE116233,GSE125725,GSE110866,GSE115526,GSE119567		
Saccharomyces cerevisiae	39	384	391
	GSE13750,GSE69414,GSE70259,GSE34082,GSE34438,GSE74393,GSE76117,GSE61753,GSE812		

	69,GSE81932,GSE81966,GSE84746,GSE85036,GSE85198,GSE85944,GSE87614,GSE87892,GSE91068,GSE100626,GSE52119,GSE50049,GSE108334,GSE51532,GSE52968,GSE53313,GSE55400,GSE56622,PRJNA245106,PRJNA254353,GSE63789,GSE66411,GSE67387,GSE125038,GSE114892,GSE109734,GSE122039,GSE115366,GSE102837,GSE104506		
Escherichia coli	13	78	81
	PRJDB2960,PRJEB7301,GSE68762,GSE72899,GSE77617,GSE85540,GSE88725,GSE90056,GSE53767,GSE56372,GSE51052,GSE58637,GSE119454		
Arabidopsis	7	27	27
	GSE69802,GSE81332,GSE86581,GSE43703,GSE98610,GSE50597,GSE109122		
Schizosaccharomyces pombe	5	27	27
	PRJEB21099,PRJEB5150,PRJEB5263,GSE98934,GSE52809		
Caenorhabditis elegans	4	35	35
	SRA049309,PRJNA170771,GSE48140,GSE67387		
Drosophila	4	31	31
	GSE83616,GSE99920,GSE49197,GSE52799		
Rat	4	43	43
	PRJEB7498,GSE60752,GSE66715,GSE129924		
Trypanosoma brucei	3	20	22
	PRJEB4801,GSE72463,GSE57336		
Zebrafish	3	30	33
	GSE34743,GSE52809,GSE53693		
Caulobacter crescentus	2	8	8
	GSE68200,GSE54883		
Staphylococcus aureus	2	12	16
	GSE74197,GSE57175		
Salmonella enterica	2	6	6
	GSE87871,GSE91066		
Crytococcus neoformans	1	6	6
	GSE133125		
Halobacterium salinarum	1	12	12
	PRJNA413990		
Chinese hamster	1	6	6
	GSE79512		
Pseudomonas aeruginosa	1	12	12
	PRJNA379630		

Plasmodium falciparum	1	5	5
	GSE58402		
Streptomyces coelicolor	1	4	4
	GSE69350		
Vibrio vulnificus	1	4	4
	GSE111991		

Additional file 2: Figure S1. Flow diagram of processing pipeline for translatable circRNAs. Two different strategies were used to characterize ribosome-associated circRNAs: condition-dependent detection for Ribo-seq and perfectly matched RNA-seq datasets and condition-independent detection for previously reported circRNAs and Ribo-seq datasets.

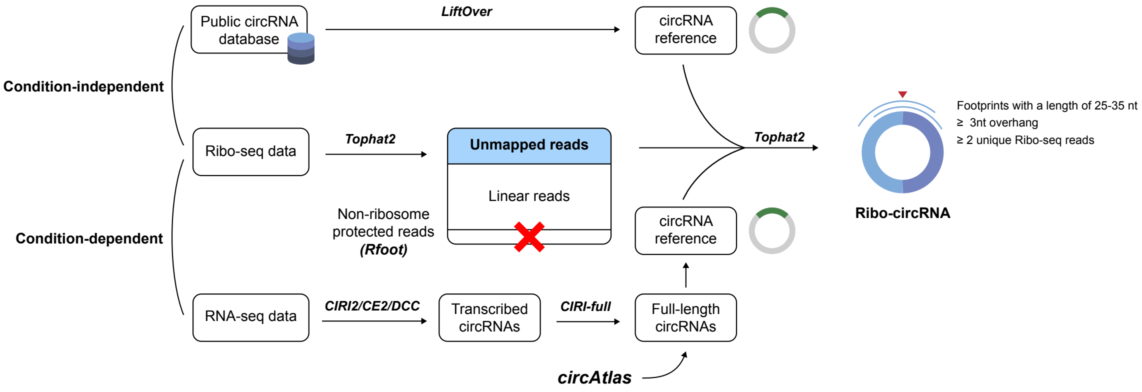

Figure S1. Flow diagram of processing pipeline for translatable circRNAs. Two different strategies were used to characterize ribosome-associated circRNAs: condition-dependent detection for Ribo-seq and perfectly matched RNA-seq datasets and condition-independent detection for previously reported circRNAs and Ribo-seq datasets.

Additional file 3: Table S2. Summary of circRNAs reported in the public databases.

Database	Species	Genome	LiftOver	Number of circRNAs	Pubmed ID	Note
circAtlas	Human	hg38	hg38	580,718	32345360	-
	Mouse	mm10	mm10	252,811		
	Rat	rn6	rn6	97,378		
circBank	Human	hg19	hg38	138,414	31023147	-
circBase	Human	hg19	hg38	138,477	25234927	-
	Mouse	mm9	mm10	16,420		
	C.elegans	ce6	ce11	718		
	Drosophila	dm3	dm6	3,406		
CIRCpedia	Human	hg38	hg38	183,993	30172046	-
	Mouse	mm10	mm10	55,312		
	Rat	rn6	rn6	10,197		
	C.elegans	ce10	ce11	3,840		
	Drosophila	dm6	dm6	8,560		
	Zebrafish	danRer10	danRer11	891		
circRNADb	Human	hg19	hg38	32,472	27725737	-
CSCD	Human/cancer	hg38	hg38	507,197	29036403	-
exoRBase	Human/blood exosomes	hg38	hg38	58,330	30053265	-
TSCD	Human	hg38	hg38	128,470	27543790	-
	Mouse	mm10	mm10	5,038		
Circ2Disease	Human	hg19	hg38	248	30030469	-
MiOncoCirc	-	-	-	-	30735636	incomplete annotation
Circad	-	-	-	-	32219412	lack of a download link
CircFunBase	-	-	-	-	30715276	lack of a download link
CircR2disease	-	-	-	-	29741596	incomplete annotation
CircRNADisease	-	-	-	-	29700306	incomplete annotation
LncRNADisease	-	-	-	-	23175614	incomplete annotation
CircInteractome	-	-	-	-	26669964	lack of a download link
CircNet	-	-	-	-	26450965	Inaccessible webpage

Additional file 4: Table S3. Summary of public proteomics datasets used in this study.

Species	PXD	Tissue/Cell line	Samples (.raw)	No.of cORFs with peptide hits	No. of ribo-circRNAs with peptide hits
Human	PXD018569	Lung cell line	216	51	484
	PXD018570	Lung cell line	214	55	
	PXD018571	Lung cell line	239	55	
	PXD018572	Lung cell line	216	54	
	PXD018573	Lung cell line	214	52	
	PXD018574	Lung cell line	212	126	
	PXD001406	LCLs	42	4	
	PXD002389	HEK293	100	10	
	PXD002395	11 human cell lines	198	42	
	PXD016999	32 normal human tissues	672	87	
	PXD007203	Foreskin fibroblasts	6	2	
	PXD017159	Blood/T lymphocyte	210	177	
	PXD021391	Multiple healthy human tissues	723	168	
Mouse	PXD013502	Brain	151	20	184
	PXD013892	Kidney inner medulla	20	0	
	PXD019880	Spleen/Liver/Lung/Muscle/Kidney/Brain /Heart	8	0	
	PXD020091	Lymphoid and myeloid populations	192	37	
	PXD023256	Lymph node/T cell	120	40	
	PXD014512	Liver	40	26	
	PXD000867	Liver	177	112	
Rat	PXD015427	Spleen/Liver/Cell culture/Lung/Kidney/Testis	1145	0	0
	PXD006349	Brain	20	0	
Drosophila	PXD007669	S2 cell line	48	0	0
	PXD000455	Whole fly	119	0	
C.elegans	PXD004561	Whole body	82	2	1
Zebrafish	PXD000479	Eye/Brain/Liver/Spleen/Intestine-pancreas/Ovary/Testes/ Muscle/Heart/Head	64	0	0
Total	26	-	5,548	719 (unique cORFs)	669 (unique circRNAs)