

SCIENTIFIC REPORTS



OPEN

circRNADb: A comprehensive database for human circular RNAs with protein-coding annotations

Xiaoping Chen¹, Ping Han², Tao Zhou³, Xuejiang Guo³, Xiaofeng Song¹ & Yan Li^{3,4}

Received: 31 May 2016
Accepted: 21 September 2016
Published: 11 October 2016

It has been known that circular RNAs are widely expressed in human tissues and cells, and play important regulatory roles in physiological or pathological processes. However, there is lack of comprehensively annotated human circular RNAs database. In this study we established a circRNA database, named as circRNADb, containing 32,914 human exonic circRNAs carefully selected from diversified sources. The detailed information of the circRNA, including genomic information, exon splicing, genome sequence, internal ribosome entry site (IRES), open reading frame (ORF) and references were provided in circRNADb. In addition, circRNAs were found to be able to encode proteins, which have not been reported in any species. 16328 circRNAs were annotated to have ORF longer than 100 amino acids, of which 7170 have IRES elements. 46 circRNAs from 37 genes were found to have their corresponding proteins expressed according mass spectrometry. The database provides the function of data search, browse, download, submit and feedback for the user to study particular circular RNA of interest and update the database continually. circRNADb will be built to be a biological information platform for circRNA molecules and related biological functions in the future. The database can be freely available through the web server at <http://reprod.njmu.edu.cn/circrnadb>.

Unlike linear RNA, circular RNA is a special group of non-coding RNA which forms a covalently closed continuous loop from exon circularization. In classical molecular biology, precursor RNA produced from DNA template strand by transcription can be processed into mature linear messenger RNA by canonical RNA splicing, in which introns are removed, while exons connect together in genomic order. However, non-canonical splicing can make exons scrambled to form a circle^{1,2}. The first circular RNA was recognized in the 1970s. In 1979, the researcher suggested that RNAs could exist in circular form in the cytoplasm of eukaryotic cells³. Ten years later, it was reported that human cytoplasmic RNA contained very low levels of transcripts of the DCC gene with scrambled exons⁴. For the next few decades, due to the specificity of the structure and the low expression level of circRNA, only a few genes were identified to express circRNAs, including DCC, EST-1, SRY etc. Recently, with the development of high throughput sequencing technology, a large number of circRNAs has been discovered across species^{5–8}. These circRNA molecules were found to be evolutionary conservative, stable, and specifically expressed across tissues or developmental stages^{9–12}. It has been shown that they play important roles in gene regulation^{9,13}. Therefore circular RNA has become the hotspots in the current transcriptomics research field.

Recently, as researchers put a lot of efforts into the study of circRNA, building a comprehensive circular RNA database become imperative. Several databases of the circRNA have been published, such as circBase, circRNA-Base and Circ2Trait^{14–16}. The circBase merged and unified data sets of circRNAs from public references, with the evidence supporting their expression within the genomic context¹⁴. Circ2Traits is a comprehensive database for circRNA potentially associated with disease and traits¹⁵, which has only 1954 circRNAs. circRNABase is designed for decoding miRNA-circRNA interaction networks from thousands of circRNAs and 108 CLIP-Seq (HITS-CLIP, PAR-CLIP, iCLIP, CLASH) datasets¹⁶, however it does not provide the genomic information of circRNAs.

In order to further study the circRNA and related biological functions, we build a comprehensive reference database, named as circRNADb. We collected dataset of circRNAs from relevant literatures, together with the

¹Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China.

²Department of Gynecology and Obstetrics, The First Affiliated Hospital with Nanjing Medical University, Nanjing 210029, China. ³State Key Laboratory of Reproductive Medicine, Department of Histology and Embryology, Nanjing Medical University, Nanjing 210029, China. ⁴Center of Pathology and Clinical Laboratory, Sir Run Run Hospital Affiliated with Nanjing Medical University, Nanjing 211166, China. Correspondence and requests for materials should be addressed to X.S. (email: xfsong@nuaa.edu.cn) or Y.L. (email: yanli@njmu.edu.cn)

circRNAs dataset identified from the Gliomas RNA-Seq dataset by our research group¹⁷. However, the primary data may have false positives (circRNAs with two ends from different genes) and redundancy, so we filtered the dataset according to gene annotation GTF file, and obtained a total of 32,914 human exonic circRNAs. Its detailed genomic information are also listed in the database, including its best matched transcript and the corresponding exon splicing information, genome sequences, in addition to all the possible isoforms and the corresponding exon splicing information.

Although circRNA is classified as a non-coding RNA, researchers have reported that eukaryotic ribosome can initiate translation on circRNA, but only when the RNA contains internal ribosome entry site (or IRES) elements¹⁸. In 1995, Chen and colleagues showed that a synthetic circRNA containing IRES elements could recruit the ribosome to initiate translation, whereas those circRNAs without IRES did not¹⁸. Although the tested circRNA was a purely artificial construct, Chen and colleagues stated in their paper that they would be interested to see whether natural circRNAs contain IRES elements. So in this work we annotated the internal ribosome entry site and open reading frame (ORF) for the circRNA with protein-coding potential. Their protein expression evidences by mass spectrometry were also provided. Besides, we analyzed the features of the proteins translated from circRNAs, included domains, N-Glycosylation sites, mucin type O-Glycosylation sites and phosphorylation sites. Users can also employ “SProtP Human” to recognize those short-lived proteins based on sequence-derived features¹⁹.

Finally, human circRNA data sets, along with its genomic features, protein coding potential and protein features were integrated into circRNADb.

Materials and Methods

Raw circRNAs dataset in circRNADb were collected from related literatures^{9–12,17}. We filtered those circRNAs supported by only one read, and only included those circRNAs supported by two or more reads. Basic genomic information about the circRNAs from the above dataset was extracted in BED format, including chromosome name, start position, end position and the cell (or tissue) type. Aided by gene annotation GTF file, we obtained the circRNA genomic information, such as strand, gene symbol, all possible transcription and the corresponding exon splicing information. The longest possible spliced transcript was taken as the best candidate sequence of the circRNAs.

It has been known that canonical ribosome-based translation needs 5'-cap structure¹⁸. The alternative mechanism to initiate translation in eukaryotic mRNAs is that the RNA has internal ribosome entry site¹⁸. As circRNAs don't have 5'-cap structure, another structure used for ribosome entry is internal ribosome entry site. An IRES element is a nucleotide sequence that allows the ribosome initiate translation directly in the middle of the mRNA sequence, instead of reading from the 5' head to 3' end²⁰. If a circRNA contains at least one IRES element, it may be able to encode a protein. Thus, in order to annotate the protein-coding potential of all the circRNAs, we employed the method of VIPS depending on RNA structure similarity as proposed by Hong J. J. *et al.* to predict the IRES element in the spliced sequence of each circRNA²¹. VIPS has three key steps: RNA folding, RNA secondary structure comparison and pseudoknot prediction program²¹. RNALfold (from ViennaRNA package, version 2.1.9), RNA Align (provided by corresponding author of VIPS) and pknotsRG (version 1.3) were used in the IRES prediction system. RNALfold was used to predict local stable RNA secondary structures of long RNA by minimum free energy method²², while pknotsRG employed the newest Turner energy rules for finding the structure of minimal free energy, so it was able to predict a restricted class of pseudoknots²³.

In addition to IRES elements, the longest open reading frame was predicted for each circRNA. Any frame that starts with a start codon, and ends with a stop codon with a length longer than 300 bp were considered as an ORF for protein coding. Furthermore, features of the proteins translated from circRNAs were analyzed. SMART was used to detect domains. NetNGlyc 1.0, NetOGlyc 3.1 and NetPhos 3.1 were employed to predict N-Glycosylation sites, Mucin type GalNAc O-Glycosylation sites and phosphorylation sites respectively^{24–28}. “SProtP Human” was also used to distinguish those short-lived proteins¹⁹.

To verify the potential coding ability of circRNA, two public proteomic datasets were used to search for the existence of fragmental peptides encoded by circRNAs. The raw data of human brains were downloaded from the PRoteomics IDentifications (PRIDE) database (data identifies: PXD000458 and PXD002528)²⁹. The parameters including quantification and modifications for database search were the same as the original studies. However, to obtain reliable identification results, an integrated database of the whole human protein sequences (UniProt release: 2013_07; 133806 sequences) and the predicted protein sequences exclusively encoded by circRNAs was applied to identify unique peptides encoded by circRNAs.

Finally, the circRNA dataset, together with its protein-coding potential, protein features were merged into circRNADb. The database was configured in the typical WAMP (Windows+Apache+Mysql+PHP) integrated environment, and built by integrating HTML5, CSS3 and Javascript programming languages. The data sources and the structure of circRNADb are shown in Fig. 1.

Results

As a comprehensive, user-friendly, interactive database, circRNADb provides the following main functions to users, including simple search, advanced search, browse, resource download, and information feedback. It includes 32,914 human exonic circRNAs after filtering for redundancy.

Search function. There is a search text box in the top right corner of each webpage, users can enter the search terms as required, such as chromosome name, gene symbol, transcript, and other keywords to query the circRNA, then the results that matched the query keywords will be listed in the result page. For example, if the user want to search by gene symbol (e.g., “SDF4”), input the keyword of “SDF4” or “SDF”, circRNADb will return circRNAs from genes matching those keywords (ID: has_circ_21644, chr1:1158623:1159348). The users can also customize

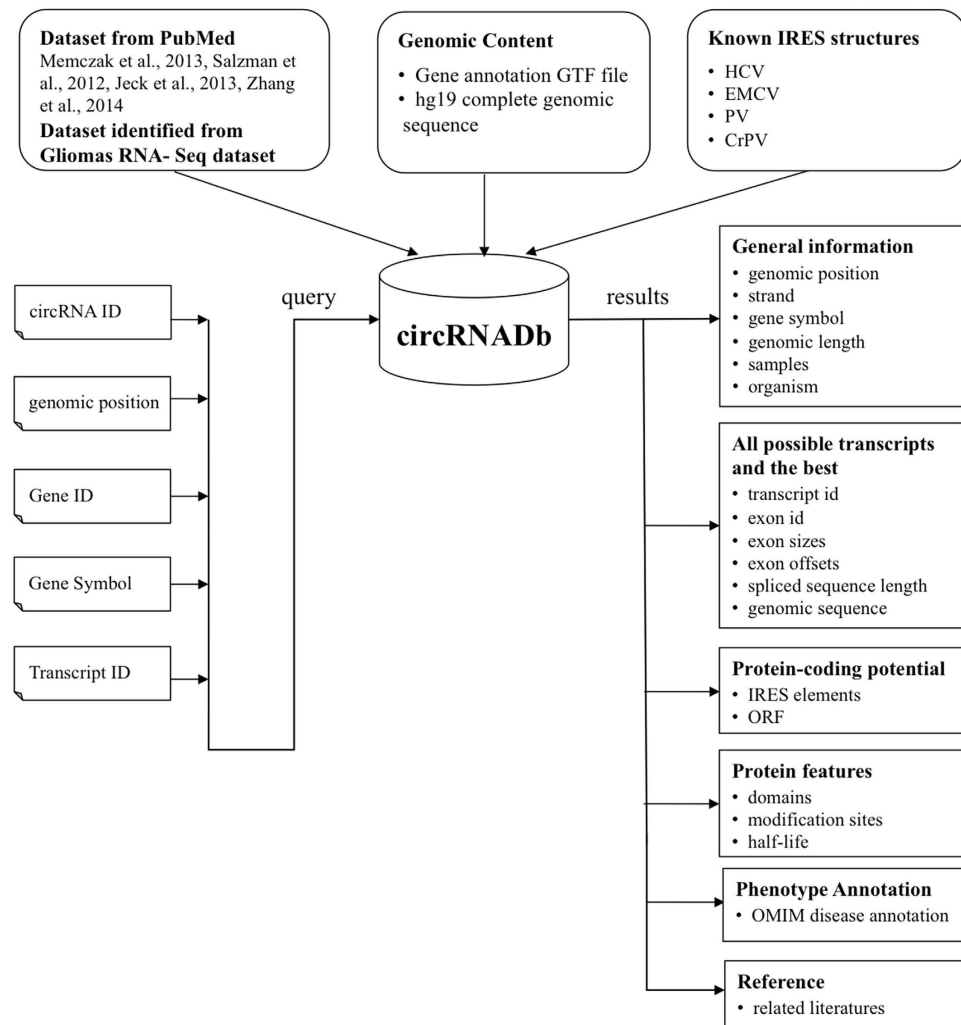


Figure 1. Data sources and the structure of circRNADb.

searching fields by clicking the “CUSTOMIZE” button. After clicking the circRNA ID in the first column, a web page with detailed information of the circRNA will be loaded.


In addition, users can restrict search terms through advanced search. As many as six fields combined with ‘AND’, ‘OR’ and ‘NOT’ in “Advanced Search” page could be used to retrieve specific circRNAs.

Data browse and download. The dataset in circRNADb can be browsed in three options: (1) browse by gene symbol, (2) browse by PubMed ID, and (3) browse by cell type.

In “Browse by Gene Symbol” page, all gene symbols and the total number of circular isoforms produced by parental genes are listed in form of a table. Users can view the detailed information of each parental gene including general information of gene symbol and all circular RNA isoforms by clicking the “Counts” on the right column. In addition, if users want to query a specific gene, the gene symbol or gene ID in NCBI can be used as query terms in the search box for targeted results. This feature allows user to simply and effectively view the information of a parental gene that they are interested in, including all circular transcripts produced.

circRNADb can also be browsed by cell (or tissue) type. In the result page, data are grouped by cell (or tissue) types, which are listed in a table with 11 types of cells and tissues. The total number of circRNAs for each cell or tissue is also listed, and users can click the number to view the detailed list of all the circRNAs identified in the cell or tissue type. This function is quite effective to retrieve all circRNAs expressed in a specific cell or tissue. Finally, all dataset of circRNA in circRNADb can also be downloaded at “Resources” page.

Submit a new circRNA and feedback. To make circRNADb more comprehensive and updated, it is important to maintain and update the database. So we designed the submission page for the users to submit their own data to circRNADb and the feedback page for the users to report problems or suggestions in circRNADb. When a user submits a new circular RNA, they need to provide the corresponding information including chromosome name, start position, end position, strand, gene symbol, best transcript, user’s name and email.


Search

circRNAdb - A Database for Human Circular RNAs

[Home](#) | [View All RNAs](#) | [Advanced Retrieval](#) | [Resources](#) | [Interaction](#) | [Tutorial](#)

Detail Information

Circ ID: hsa_circ_07894

General Information

Circ ID	hsa_circ_07894
Location (hg19)	chr9 : 107645319-107651476
Strand	-
Gene Symbol	ABCA1
Genomic length	6157
Samples	oligodendroma, Hs68
Organism	Homo sapiens (human)

Detail Information

Best Transcript

Transcript id	NM_005502			
Exon	Exon Number	Exon id (total)	Exon Sizes	Exon Offsets
	3	46, 47, 48 (50)	119, 142, 94	0, 1388, 6063
Spliced seq length	355 nt			
Sequence	TGTCAGCTGC TGCTGGAAGT GGCTGGCCT CTATTTATCT TCCTGATCCT GATCTCTGTT CGGCTGAGCT ACCCACCTA TGAACAACAT GAATGCCATT TTCCAAATAA AGCCATGCC TCTGCAGGAA CACTTCTTG GGTTCAGGGG ATTATCTGTA ATGCCAACAA CCCCTGTTTC CGTTACCCGA CTCCTGGGGA GGCTCCCGGA GTTGTGGAA ACTTTAACAA ATCCATTGTG GCTCGCCTGT TCTCAGATGC TCGGAGGCTT CTTTATACA GCCAGAAAGA CACCAGCATG AAGGCATCG GCAAAGTTCT GAGAACATTA CAGCAGATCA AGAAATCCAG CTCAA			

All Possible Transcripts

Transcript id	Exon information			
Spliced_jen	Exon Number	Exon Sizes	Exon Offsets	
NM_005502	355	3	119, 142, 94	0, 1388, 6063

Protein coding potential

IRES Elements	Parameter Index		
	Position (start--end)	R Score	With Pseudoknot (Y/N)
	262--349	1.498811	Y
195--248	1.442485	Y	
Open Reading Frame (ORF)	Start Position		
	End Position	Protein Length	
	115	1*83	107 aa
MPSAGTLPW QGICNANNP CFRYPTGPEA PGVWGNFNS IVARLFS DAR RLLLYSQDT SMKDMRKLVR TLQIQIKSSS MSAAGSGLA STYLPDPLC SAELPTL*			
Note: (1). nr represents n rounds(n<3); (2). * represents a stop codon.			
Protein Features	The possibility of encoding protein is relatively low(R<1.6 or it has no open reading frame), so no protein features was predicted!		

Phenotype Annotation (Parental Gene)

OMIM Disease Annotation	CORONARY HEART DISEASE IN FAMILIAL HYPERCHOLESTEROLEMIA, PROTECTION AGAINST, INCLUDED;;HIGH DENSITY LIPOPROTEIN CHOLESTEROL LEVEL QUANTITATIVE TRAIT LOCUS 13, INCLUDED; HDLQ14, INCLUDED (OMIM:60046)
-------------------------	---

Reference

PMID: 26873924	Song X, Zhang N, Han P, et al. (2016) Circular RNA profile in gliomas revealed by identification tool UROBORUS. Nucleic acids research. [Epub ahead of print]
PMID: 23249747	Jeck WR, Sorrentino JA, Wang K, et al. (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA. 19:141-157.

Figure 2. Detailed information web page of a circRNA (ID: hsa_circ_07894).

Detailed information of each circRNA. As mentioned above, users are allowed to view the detailed information of each circRNA by clicking the ID on the first column in the result page. To ensure the accuracy of the database, all the information of each entry was carefully checked manually. As shown in Fig. 2, the page is divided into two

major sections: basic information and detailed information. In the basic information section, ID, genomic location, strand, gene symbol, sample name and species of each circRNA are displayed. And the detailed information section provides detailed information for each circRNA.

The detailed information section provides the best matched transcript of the circRNA, its exons information and spliced sequences. According to gene annotation GTF file, all possible circular isoforms and related information are also displayed. Next, in order to study the protein-coding potential of the circRNA, IRES elements and potential ORF longer than 300 bp in each circRNA were predicted. IRES elements with top two highest scores were provided, with details including position, parameter index (R Score and the existence of Pseudoknot). The longest ORF that starts with a start codon (ATG), ends with a stop codon (TAA, TAG or TGA) is shown in the below. There are 11,423 and 16,328 circRNAs predicted to contain IRES element and ORF respectively. 7,010 circRNAs contain both IRES and ORF, approximately accounting for 21.3% of all circRNAs, which are considered as potential protein-coding circRNAs. If the circRNA has the potential to code a protein, protein features including domains, post-translational modification sites and half-life prediction are provided. In addition, we annotated circRNA parental gene associated with human disease (OMIM). At the bottom, literature sources of the circRNA are provided, including PubMed IDs and the detailed references.

The expression evidences were also provided for the circRNA-coding proteins. Two mass spectrometry datasets of human brains from PRIDE database were used to identify proteins exclusively coded by circRNAs. In total, 45 peptides mapped to 46 circRNA-encoding proteins, corresponding to 37 genes, were identified by mass spectrometry. These peptides didn't belong to any known UniProt human proteins, and were unique to circRNA-encoding proteins. Some peptides were mapped to two different circRNAs, we found that all these are different splicing isoforms from the same gene. Thus, circRNAs from the 37 genes were confidently identified to encode proteins in human brain (Supplementary Table S1). Among the 46 protein coding circRNAs with mass spectrometry evidences at protein expression level, 22 circRNAs were annotated to have IRES element. The representative spectra of one example peptide ("LLQCYPPEDPAVR", encoded by has_circ_25375) indicated the high quality of proteomic identification (Supplementary Figure S1). The expression of these circRNA-encoding proteins in human brain indicated that circRNAs might perform functions by translation into proteins.

Discussions and Conclusions

As a comprehensive human exonic circRNA database with protein-encoding feature annotation, circRNADb is designed to provide a rich data resource for the circRNAs research. circRNADb has collected circRNA dataset from relevant literatures and the brain RNA-seq dataset from our work. In total, we obtained 32,914 non-redundant human exonic circRNAs. circRNADb may facilitate circRNA studies by (1) providing users with detailed genomic information of each circRNA; (2) annotating protein-coding potential of each circRNA; (2) including protein expression evidences of circRNA by mass spectrometry; (3) providing convenient interfaces to retrieve the data.

We will update the newly identified circRNAs, and incorporate their potential functions of encoding proteins or regulating gene expressions. With the improvement of circRNADb, it is expected to become a powerful tool and provide a foundation for further study on circRNA.

References

- de la Mata, M., Lafaille, C. & Kornblihtt, A. R. First come, first served revisited: factors affecting the same alternative splicing event have different effects on the relative rates of intron removal. *RNA*. **16**, 904–912 (2010).
- Schindewolf, C., Braun, S. & Domdey, H. *In vitro* generation of a circular exon from a linear pre-mRNA transcript. *Nucleic Acids Res.* **24**, 1260–1266 (1996).
- Hsu, M. T. & Coca-Prados, M. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. *Nature*. **280**, 339–340 (1979).
- Nigro, J. M. *et al.* Scrambled exons. *Cell*. **64**, 607–613 (1991).
- Danan, M., Schwartz, S., Edelheit, S. & Sorek, R. Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Res.* **40**, 3131–3142 (2012).
- Zhang, Y. *et al.* Circular intronic long noncoding RNAs. *Mol. Cell*. **51**, 792–806 (2013).
- Zhang, X. O. *et al.* Complementary sequence-mediated exon circularization. *Cell*. **159**, 134–147 (2014).
- Jeck, W. R. & Sharpless, N. E. Detecting and characterizing circular RNAs. *Nat Biotechnol.* **32**, 453–461 (2014).
- Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*. **495**, 333–388 (2013).
- Jeck, W. R. *et al.* Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*. **19**, 141–157 (2013).
- Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L. & Brown, P. O. Cell-type specific features of circular RNA expression. *PLoS Genet.* **9**, e1003777 (2013).
- Salzman, J., Gawad, C., Wang, P. L., Lacayo, N. & Brown, P. O. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One*. **7**, e30733 (2012).
- Hansen, T. B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature*. **495**, 384–388 (2013).
- Glažar, P., Papavasileiou, P. & Rajewsky, N. CircBase: a database for circular RNAs. *RNA*. **20**, 1666–1670 (2014).
- 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.* **4**, 283 (2013).
- Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. StarBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **42(Database issue)**, D92–7 (2014).
- Song, X. *et al.* Circular RNA profile in gliomas revealed by identification tool UROBORUS. *Nucleic Acids Res.* **44**, e87 (2016).
- Chen, C. Y. & Sarnow, P. Initiation of protein synthesis by the eukaryotic translational apparatus on circular RNAs. *Science*. **268**, 415–417 (1995).
- Song, X. *et al.* SProtP: a web server to recognize those short-lived proteins based on sequence-derived features in human cells. *PLoS One*. **6**, e27836 (2011).
- Vaklavas, C. *et al.* Small molecule inhibitors of IRES-mediated translation. *Cancer Biol. Ther.* **16**, 1471–1485 (2015).
- Hong, J. J., Wu, T. Y., Chang, T. Y. & Chen, C. Y. Viral IRES prediction system - a web server for prediction of the IRES secondary structure in silico. *PLoS One*. **8**, e79288 (2013).

22. Hofacker, I. L., Priwitzer, B. & Stadler, P. F. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*. **20**, 186–190 (2004).
23. Jens, R., Peter, S. & Robert, G. pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids Res.* **35**(Web Server issue), W320–W324 (2007).
24. Joshi, H. J. & Gupta, R. Eukaryotic glycosylation: online methods for site prediction on protein sequences. *Methods Mol. Biol.* **1273**, 127–37 (2015).
25. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a Simple Modular Architecture Research Tool: Identification of Signaling Domains. *Proc. Natl. Acad. Sci. USA* **95**, 5857–5864 (1998).
26. Ivica, L., Tobias, D. & Peer, B. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.* **43**(Database issue), D257–60 (2015).
27. Julenius, K., Molgaard, A. R. & Brunak, S. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology*. **15**, 153–164 (2005).
28. Blom, N., Gammeltoft, S. & Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294**, 1351–1362 (1999).
29. Vizcaino, J. A. *et al.* update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**(D1), D447–56 (2016).

Acknowledgements

The study was supported by grants from National Key Research and Development Program of China (2016YFA0503300, X.G), the 973 program (2013CB911400, X.G), the National Natural Science Foundation of China (No. 61571223, 61171191, X.S; No. 81270700, P.H; No. 31471403, X.G), the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20133218110016, X.S), the “Six Talent Peak” Project of Jiangsu Province under Grant No. SWYY-021(X.S).

Author Contributions

X.C., P.H. and T.Z. implemented the study, and developed the web server, X.S., X.G. and Y.L. conceived study, and wrote the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Chen, X. *et al.* circRNADb: A comprehensive database for human circular RNAs with protein-coding annotations. *Sci. Rep.* **6**, 34985; doi: 10.1038/srep34985 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016