circMine: a comprehensive database to integrate, analyze and visualize human disease–related circRNA transcriptome

Wenliang Zhang[1,2,9,*,†], Yang Liu[2,3,10,†], Zhuochao Min[4,†], Guodong Liang[12], Jing Mo[9], Zhen Ju[2,6,7], Binghui Zeng[8,9], Wen Guan[9,11], Yan Zhang[2], Jianliang Chen[1], Qianshen Zhang[1], Hanguang Li[1], Chunxia Zeng[2,6,7], Yanjie Wei[2,6,7,*] and Godfrey Chi-Fung Chan[1,5,*]

[1]Department of Pediatrics, The University of Hong Kong-Shenzhen Hospital, Shenzhen, Guangdong 518053, China, [2]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China, [3]Department of Gastroenterology and Hepatology, The University of Hong Kong-Shenzhen Hospital, Shenzhen, Guangdong 518053, China, [4]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China, [5]Department of Pediatrics and Adolescent Medicine, LKS Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Hong Kong 999077, China, [6]Center for High Performance Computing, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China, [7]CAS Key Laboratory of Health Informatics, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China, [8]Hospital of Stomatology, Guangdong Provincial Key Laboratory of Stomatology, Guanghua School of Stomatology, Sun Yat-sen University, Guangzhou 510055, China, [9]Department of Bioinformatics, Outstanding Biotechnology Co., Ltd.-Shenzhen, Shenzhen 518053, China, [10]Experimental Training Management Center, Jilin Business and Technology, Jilin Province 130507, China, [11]Guangdong Key Laboratory of Animal Conservation and Resource Utilization, Institute of Zoology, Guangdong Academy of Sciences, Guangzhou 510260, China and [12]Department of Colorectal and Stomach Cancer Surgery, Jilin Cancer Hospital, Changchun, Jilin 130000, China

Received July 19, 2021; Revised September 01, 2021; Editorial Decision September 04, 2021; Accepted September 07, 2021

ABSTRACT

Many circRNA transcriptome data were deposited in public resources, but these data show great heterogeneity. Researchers without bioinformatics skills have difficulty in investigating these invaluable data or their own data. Here, we specifically designed circMine (http://hpcc.siat.ac.cn/circmine and http://www.biomedical-web.com/circmine/) that provides 1 821 448 entries formed by 136 871 circRNAs, 87 diseases and 120 circRNA transcriptome datasets of 1107 samples across 31 human body sites. circMine further provides 13 online analytical functions to comprehensively investigate these datasets to evaluate the clinical and biological significance of circRNA. To improve the data applicability, each dataset was standardized and annotated with relevant clinical information. All of the 13 analytic functions allow users to group samples based on their clinical data and assign different parameters for different analyses, and enable them to perform these analyses using their own circRNA transcriptomes. Moreover, three additional tools were developed in circMine to systematically discover the circRNA–miRNA interaction and circRNA translatability. For example, we systematically discovered five potential translatable circRNAs associated with prostate cancer progression using circMine. In summary, circMine provides user-friendly web interfaces to browse, search, analyze and download data freely, and submit new data for further integration, and it can be an important resource to discover significant circRNA in different diseases.

*To whom correspondence should be addressed. Tel: +86 13560313976; Email: zhangwl25@mail3.sysu.edu.cn
Correspondence should also be addressed to Yanjie Wei. Email: yj.wei@siat.ac.cn
Correspondence should also be addressed to Godfrey Chi-Fung Chan. Email: gcfchan@hku.hk
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

© The Author(s) 2021. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab809/6371375 by guest on 17 September 2021

INTRODUCTION

Circular RNA (circRNA) is covalently closed endogenous RNA produced by back-splicing of precursor messenger RNA (mRNA) in eukaryotes, and they are largely discovered by deep sequencing techniques (1,2). Studies showed circRNAs play critical roles in important biological processes through acting as a microRNA (miRNA) sponge, competing endogenous RNA (ceRNA), protein regulator or even by translating themselves to produce peptides and proteins (1,2). They are emerging as a novel type of biomarker for human disease (1,2). Although studies on circRNA are accumulating rapidly in the last decade, our knowledge on their clinical and biological significance in human diseases remains elusive.

Databases and computational tools have been developed for identifying and annotating the transcribed circRNA (3–6), tissue- and cell type-specific circRNA (7–10), circRNA interactor (such as miRNA and protein) (11,12), potential translatable circRNA (13–15) and experimentally validated disease associated circRNA (16–19). For example, the circBase (20), CircBank (21) and circAtlas (22) databases provide comprehensive genomics and functional annotations for circRNA. The CSCD (7) and CircRiC (8) databases were designed to annotate cancer-specific circRNAs. In addition, exoRBase (10) was developed to provide a landscape of human blood exosome circRNAs derived from RNA-seq data analyses. Recently, Li *et al.* has designed riboCIRC (13) to specifically host and investigate translatable circRNAs from ribosome sequencing data analyses. However, there is a lack of database to integrate human disease–related circRNA transcriptome datasets and provide comprehensive analyses to investigate these invaluable datasets online for identifying and discovering the clinical and biological significance of circRNA.

Over the past 5 years, lots of human circRNA transcriptome datasets have been generated and deposited in Gene Expression Omnibus (GEO) (23,24). However, it is very challenging for researchers without bioinformatics skills to investigate these invaluable data or their own data, and it is also difficult to find the specific circRNA data with designated physiological and pathological conditions from the massive data in GEO. Moreover, the circRNA transcriptome data generated by different high-throughput platforms in GEO show great heterogeneity, which further hinders the application. Thus, there is an urgent need to design a specific database platform that can integrate and provide comprehensive analytical functions to investigate the circRNA transcriptome data, which can facilitate the study and understanding of circRNA in human disease.

To overcome all these challenges, we specifically designed circMine (http://hpcc.siat.ac.cn/circmine and http://www.biomedical-web.com/circmine) to integrate human circRNA transcriptome data from GEO and developed comprehensive web applications to investigate these data (Figure 1). Currently, circMine provides 1 821 448 entries that formed by 136 871 circRNAs, 87 diseases and 120 circRNA transcriptome datasets of 1107 samples across 31 human body sites (Supplementary Table S1). To eliminate the data heterogeneity, each dataset was standardized and manually annotated with specific physiological and pathological conditions for accurate data retrieval (Figure 1A). Moreover, to discover and identify the significant circRNAs in human diseases, we developed 13 different analytical functions to investigate the integrated data individually (Figure 1B, upper left panel). The Web Server application on circMine further provides opportunities for researchers to conduct the 13 analyses on their own circRNA transcriptome data (Figure 1B, upper left panel).

In addition, circMine provides three additional tools to discover and identify the biological significance of circRNA, including the circRNA-miRNA prediction, circRNA IRES prediction and ribo-circRNA location (Figure 1B, lower left panel). Moreover, circMine provides user-friendly web interfaces to browse, search, download data openly and submit new circRNA transcriptome data for further integration (Figure 1B, right panel). By using our circMine, we systematically identified five potential translatable circRNAs associated with prostate cancer progression through translating themselves and interacting with thousands of miRNAs. In summary, circMine can significantly improve our insight in discovering and identifying the significance of circRNA in human diseases.

MATERIALS AND METHODS

Data collection and processing

To collect the human circRNA transcriptome data, we retrieved the GEO resource (23,24) by searching keywords of '[(circRNA OR circular RNA) AND Homo sapiens]'. About 358 candidate datasets had been retrieved, which were made public before 10 April 2021. Moreover, all of these datasets were manually curated by at least two professional curators based on two criteria: (i) the datasets providing circRNA expression data of human samples (including tissue, plasma, exosome and cell line) were included regardless of the high-throughput platforms and (ii) the sample information and expression data of the datasets could be downloaded for further integration. After manual curation based on these criteria, 120 datasets were selected and downloaded, which showed great heterogeneity for they were generated from 21 different high-throughput platforms. Therefore, to overcome the data heterogeneity, we further developed a systematic pipeline to standardize and normalize the circRNA IDs, sample IDs, and junction read counts in the datasets by using the related annotation files in the circBase (20) and GEO (23,24) resources. About 82.50% (99/120), 28.33% (34/120) and 17.50% (21/120) of the datasets were standardized including circRNA ID, sample ID and junction read count to circBase ID, GEO accession and spliced reads per billion mapping (SRPBM), respectively. In addition, we assigned each dataset with a unique ID (e.g., HSACM000001) in circMine. We further manually annotated each dataset with specific physiological and pathological conditions for fast and accurate data retrieval, such as disease grade and stage, drug resistance, metastasis, virus infection, age and gender.

Differential expression module and co-expression module

To comprehensively investigate the integrated data in circMine through customized grouping and setting, we designed seven differential analytical functions in the dif-

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab809/6371375 by guest on 17 September 2021

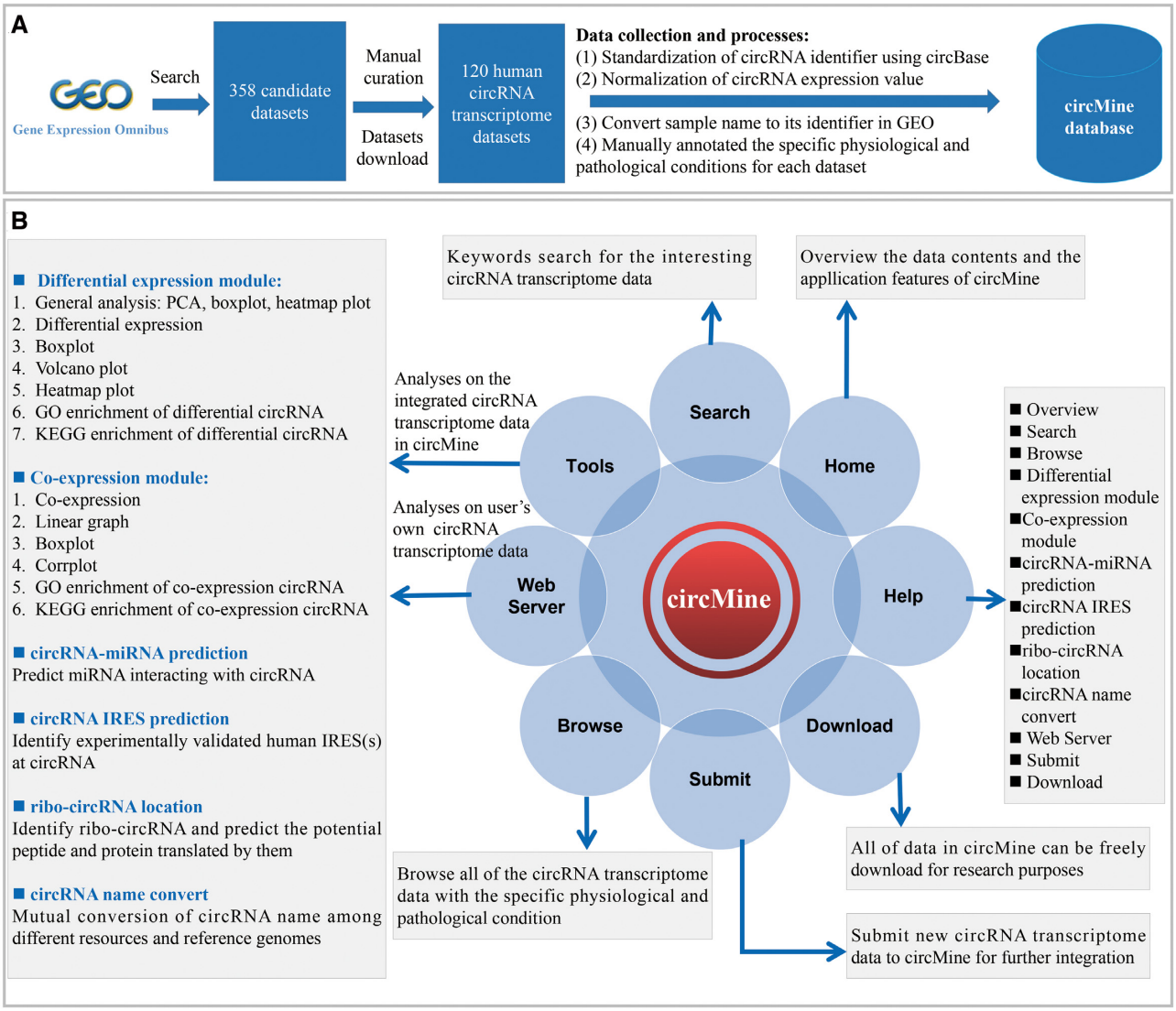


Figure 1. The scheme for data collection and manual curation (**A**) and the web application framework of circMine (**B**); GO, Gene Ontology; IRES, internal ribosome entry site element; KEGG, Kyoto Encyclopedia of Genes and Genomes; PCA, Principal Component Analysis.

ferential expression module and six co-expression analytical functions in the co-expression module based on the R project (https://www.r-project.org/) (Figure 1B, upper left panel, Table 1 and Supplementary Figure S1). In addition, Supplementary Table S2 and the 'Help' web-page on the database have the details of these 13 analytical functions, including the major R packages and functions used for their implementation, data processing, analyzing and visualization.

circRNA-miRNA prediction

Knowing that circRNA can act as a miRNA sponge and ceRNA to regulate many biological pathways in human diseases, we developed the circRNA-miRNA prediction tool to identify putative circRNA-miRNA interactions based on the miRanda (25), miRBase (26) and circBase (20) resources. We first extracted the human miRNA seed region sequences and circRNA sequences in the miRBase and circBase resources, respectively. After installing miRanda locally, a shell script has been designed to encapsulate miRanda with the extracted sequence data.

circRNA IRES prediction

Considering that the internal ribosome entry site elements (IRESs) at circRNA are able to drive its translation (2), we developed the circRNA IRES prediction to identify the experimentally validated human IRESs at circRNA based on the BLAST (27), IRESbase (28) and circBase (20) resources. First, we installed BLAST (ncbi-blast-2.11.0+-x64-linux), and then we extracted the experimentally validated human IRES sequences from IRESbase, and further constructed a human IRES indexes database for the BLAST. Moreover, we extracted human circRNA sequences in circBase. Finally, we wrote a shell script to encapsulate the BLAST (including the IRES indexes database) with the extracted circRNA sequences to implement this tool.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab809/6371375 by guest on 17 September 2021

Table 1. The description of the 13 analytical functions in the differential expression and co-expression modules

Web analysis	Description
Differential expression module	
General analysis	To conduct a heatmap plot, principal component analysis and box plot on the data
Differential expression	To identify all of significant differential circRNA between two conditions
Boxplot	To present the expression difference of a circRNA on different conditions as a box plot
Volcano plot	To depict the expression difference of one or more circRNAs between two conditions as a volcano plot
Heatmap plot	To depict the circRNA expression pattern on different conditions as a heatmap
GO enrichment	GO enrichment on the host genes of the differential circRNAs between two conditions
KEGG enrichment	KEGG enrichment on the host genes of the differential circRNAs between two conditions
Co- expression module	
Co-expression	To present correlation analysis of a circRNA with all of circRNAs on specific conditions
Linear graph	To present the correlationship of two circRNAs on specific conditions
Boxplot	To conduct paired correlation analysis of two circRNAs on specific conditions
Corrplot	To depict the correlationships among multiple circRNAs on specific conditions
GO enrichment	GO enrichment on the host genes of the co-expression circRNAs on specific conditions
KEGG enrichment	KEGG enrichment on the host genes of the co-expression circRNAs on specific conditions

ribo-circRNA location tool

To identify ribosome-associated circRNA (ribo-circRNA) and predict the subcellular localization of the putative peptide and protein translated by it, we implemented the ribo-circRNA location tool based on riboCIRC (13), DeepLoc (29) and circBase (20). First, we extracted ribo-circRNAs and the amino acids sequences of their putative peptides and proteins from the riboCIRC database. Second, we annotated the ribo-circRNA with the circRNA ID in different resources such as circBase, CircBank and RefSeq. Third, we installed the DeepLoc (version 1.0) tool. The DeepLoc with default parameters was used to differentiate ten subcellular localizations of those putative peptides and proteins based on their amino acids sequences. Finally, we implemented the ribo-circRNA location tool in the R project and encapsulated it by using shell scripts.

circRNA name convert tool

To achieve the mutual conversion of the circRNA IDs of different resources, we first constructed an annotation file that annotates circRNAs with various IDs in different resources and reference genomes by using circBase (20), CircBank (21), riboCIRC (13), UCSC liftOver (30), and the annotation files of the GPL19978 and GPL21825 platforms in GEO (23,24). The GPL19978 and GPL21825 are two common platforms for human circRNA expression profile. Moreover, we implemented the circRNA name convert tool in the R project and encapsulated it with the constructed annotation file by shell scripts.

Web Server

Web Server is a systematic pipeline to automatically integrate and standardize the circRNA transcriptome data and its corresponding sample information uploaded by users, and further it assigns a temporary ID to the uploaded data so that they can perform analyses and remove their uploaded data in the database. The Web Server application was implemented in the R project based on the above constructed annotation file. The annotation file is used to convert the circRNA IDs in the uploaded data to the circBase IDs. Moreover, we designed the application so as to handle various types of circRNA expression value in the uploaded data, including junction read counts, transcripts per kilobase of exon model per million mapped reads (TPM), fragments per kilobase of exon model per million mapped fragments (FPKM), SRPBM and normalization value with or without log2 transformation. In addition, taking into account the data security, the application allows users to remove private data uploaded by themselves at any stage using the assigned temporary ID of the data, and we regularly clean up the uploaded data monthly.

Data storage and web implementation

circMine is freely available at the websites of http://hpcc.siat.ac.cn/circmine/ and http://www.biomedical-web.com/circmine. All of the annotation data are stored and managed in the MySQL database. The R version 4.0.3 (https://www.r-project.org/) was installed to run the analyses. In addition, Supplementary Tables S2 and S3 described the R packages and the resources used in the database implementation, respectively. Moreover, the database was implemented with a separated back-end and front-end web framework. The back-end was built with the web framework of Spring Boot (https://spring.io/projects/spring-boot/), while the front-end was built with Vue3 (https://vuejs.org/), JQuery (https://jquery.com/) and bootStrap4 (https://getbootstrap.com/). The programs for data processing and application operation were written in Java. Finally, the database was deployed in the Apache Tomcat Server.

RESULTS

A comprehensive resource for human circRNA transcriptome data with specific pathological and physiological conditions

circMine contains 1 821 448 entries that formed by 136 871 circRNAs, 87 diseases and 120 circRNA transcriptome datasets of 1107 samples across 31 human body sites (Supplementary Table S1). These diseases include various cancers, infection and immune inflammatory diseases, and diseases of heart, brain, digestive system, renal, spine, oral, bone, lung, vascular etc. and the samples include tissue (75.83%, 91/120), plasma (13.33%, 16/120), exosome

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab809/6371375 by guest on 17 September 2021

(2.50%, 3/120) and cell line (8.34%, 10/120). To eliminate the heterogeneity of the datasets produced by different high-throughput platforms, circRNA ID, sample ID and expression value in the dataset have been standardized and normalized using circBase (20) and the annotation files in GEO (23,24). For facilitating fast and accurate retrieval, each dataset has been manually annotated with specific pathological and physiological conditions. The pathological and physiological conditions include disease grade and stage, genotype, drug resistance, metastasis, immune, lifestyle, virus infection, age and gender. The 'Home' webpage (http://hpcc.siat.ac.cn/circmine) provides a landscape of the data contents and features of the database. For example, the top six body sites with the largest number of datasets and samples are brain, heart, bone and bone marrow, blood and blood vessel, intestine and liver. Moreover, most of the datasets are associated with 37 cancer sub-types, which account for about 57.5% (69/120) of the datasets. The top three cancer sub-types with the largest number of datasets are gastric cancer, hepatocellular carcinoma and colorectal carcinoma, accounting for about 6.7% (8/120), 6.7% (8/120) and 4.2% (5/120), respectively. In addition, the statistic results on the 'Home' web-page showed that the number of circRNA transcriptome data has been accumulating rapidly in recent 5 years and 91.7% of the data are from China, while the remaining are from Germany, USA, Australia, Spain, Sweden, Japan and Netherlands.

Furthermore, circMine provides user-friendly web interfaces to browse, search, access data openly, as well as to download and submit new data for further integration. For example, the search function in the download web-page enables users to quickly browse and download the specific data for their research. circMine also provides a variety of common data formats for users to download, including text table, CSV and JSON. In order to allow more flexibility, the 'Submit' application page has been developed to allow researchers to submit their new circRNA transcriptome data with its sample information to the database for future integration. After careful data evaluation (such as data quality and ethical approval) by our submission committee, the submitted data will be included in the future release.

Comprehensive analysis to investigate the integrated and researchers own human circRNA transcriptome data

To analyze the integrated disease-related circRNA transcriptome data individually for discovering the clinical and biological significance of circRNA, we implemented seven differential expression and six co-expression analytical functions in the differential expression and co-expression modules, respectively (Figure 1B, upper left panel and Table 1). All of these 13 analytical functions allow users to customize grouping samples based on their clinical metadata and setting parameters for individual analysis (Supplementary Figure S1). We enumerate the details of the 13 analytical functions in Table 1 and on the 'Help' web-page of the database.

To better serve the community, we developed the Web Server application for users to upload their own circRNA transcriptome data, and further conduct the 13 different analyses on the uploaded data by themselves only. Given that the circRNA expression profiles from different researchers were generated from different platforms, the Web Server application has been designed to handle various types of circRNA IDs (such as circBase and CircBank IDs) and their expression value (including junction read counts, TPM, FPKM, SRPBM and normalization with or without log2 transformation) in the uploaded data. In addition, in order to better protect the security of the uploaded data, the Web Server allows users to delete the uploaded data using the assigned temporary ID at any stage, and we regularly clean up the uploaded data monthly.

To enhance the application efficiency of circMine in the circRNA research community, circMine also provides three additional tools to study the biological mechanisms of circRNA in human disease, including circRNA-miRNA prediction, circRNA IRES prediction and ribo-circRNA location (Figure 1B, lower left panel). The detailed descriptions of these three tools are as follows:

(i) *circRNA-miRNA prediction*. This tool aims to discover the putative interaction between miRNA and circRNA. It allows users to enter a list of circBase IDs, the full length of interesting circRNAs sequences in FASTA format or upload a text file containing the circBase IDs or the full length of circRNAs sequences. Moreover, the score cutoff and energy cutoff parameters are provided to filter the significant interacting miRNA.

(ii) *circRNA IRES prediction*. In order to discover the translatability of circRNA, the circRNA IRES prediction tool has been designed to identify the experimentally validated human IRES at circRNA. The circRNA IRES prediction tool enables users to enter a list of circBase IDs or the full length of interesting circRNAs sequences in FASTA format or upload a text file containing the circBase IDs or the full length of circRNA sequences. Moreover, the *E*-value parameter is provided to set a threshold for the significant IRES.

(iii) *ribo-circRNA location tool.* This tool aims to identify ribo-circRNA and predict the subcellular localization of its putative peptide and protein. The tool can differentiate ten subcellular localizations such as nucleus, extracellular, cytoplasm, mitochondrial, cell membrane and endoplasmic reticulum. In addition, the tool allows users to enter a list of circRNAs with various nomenclatures such as circBase IDs, CircBank IDs, RefSeq IDs and the chromosome positions of hg19 and hg38.

circMine also provides a tool called circRNA name convert to automatically annotate circRNAs among different resources and reference genomes, including circBase, CircBank, riboCIRC, RefSeq, GenBank, hg19 and hg38, and the common platforms of GPL19978 and GPL21825 in GEO. Given that circRNA coordinates probably caused by 0-based/1-based errors of coordinate systems or potential sequencing errors, the circRNA name convert and ribo-circRNA location tools are designed to allow 2 bp mismatch for the circRNA coordinate conversion.

Finally, the results from the above analyses and tools are presented as graph or table. The results in graph format can be freely downloaded and saved as a PDF file in high resolution, while the result tables offer filtering and sorting func-

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab809/6371375 by guest on 17 September 2021

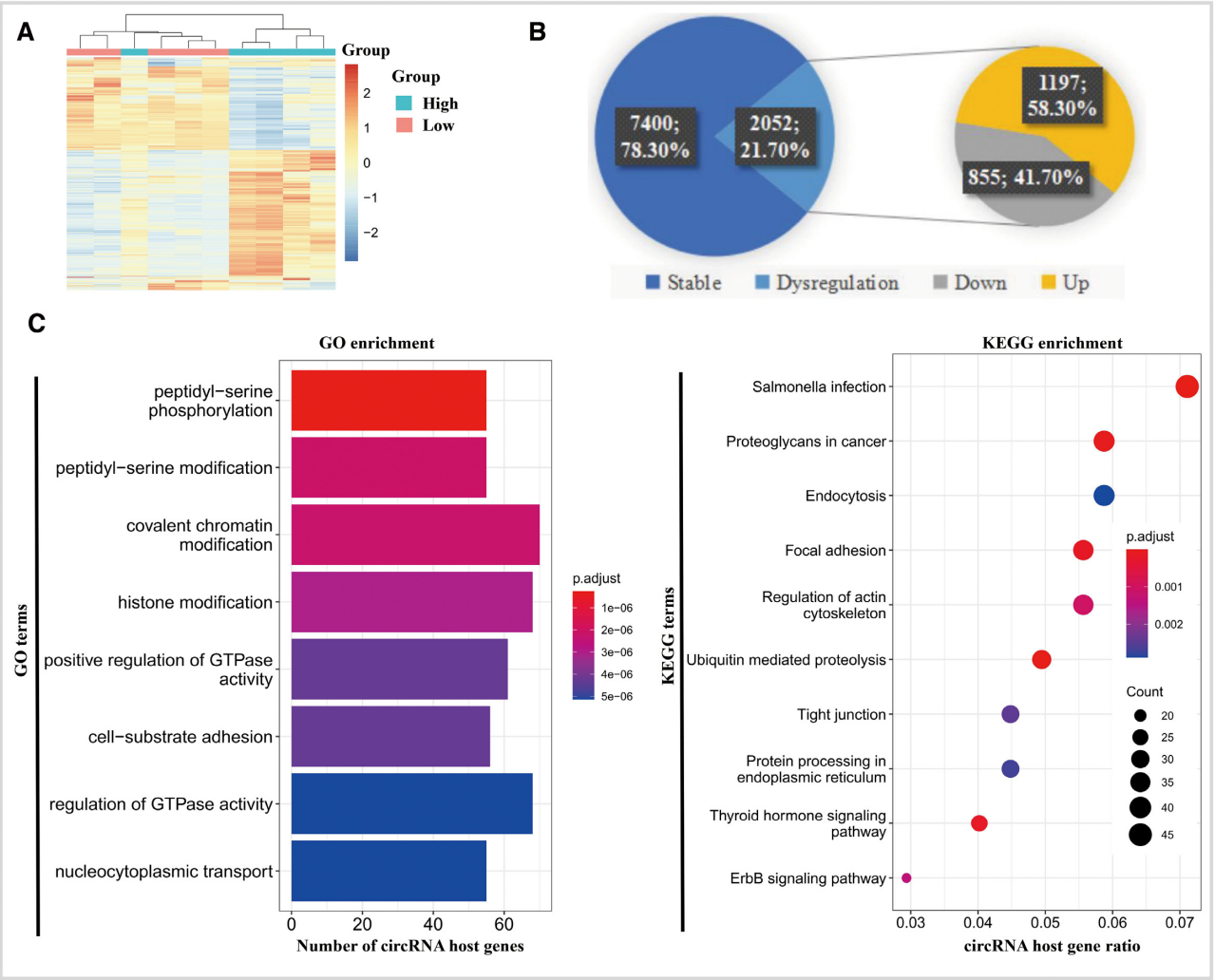


Figure 2. The circRNA differential expression analysis results from five low (Gleason < 6) and five high (Gleason > 8) grade prostate cancer tissues. (**A**) The heatmap plot shows significantly different circRNA expression profile between the two groups. (**B**) The pie chart indicates the numbers and percentages of deregulated, up-regulated and down-regulated circRNAs. (**C**) The bar and dot plot to show the GO and KEGG enrichment results from the 2052 deregulated circRNAs (|logFC| ≥ 1.0 and *P*-value ≤ 0.05), respectively.

tions that allow users to easily search for the interesting data (Supplementary Figure S2). A menu at the table header allows users to set a table with columns and rows to show the interesting data (Supplementary Figure S2). Moreover, by right clicking on the table body, an instruction form will be shown to copy and export the table results for further analyses (Supplementary Figure S2). In addition, the resulted tables provide links to the related resources for the detailed information about the circRNA, ribo-circRNA, interacting miRNA and human IRES. A comprehensive tutorial is available for these applications on the database 'Help' webpage.

Case study: comprehensive analysis to discover potential translatable circRNAs associated with prostate cancer progression using circMine

To identify key circRNA associated with prostate cancer progression, we comprehensively investigated the circRNA transcriptome dataset (circMine ID: HSACM000016) (31) by using circMine. First, we classified the samples into two groups named low (Gleason < 6) and high (Gleason > 8) based on the sample information of the dataset (Supplementary Figure S1). We further performed the seven analyses in the differential expression module. The results from the general analysis showed significantly different circRNA expression profiles between the two groups (Figure 2A and Supplementary Figure S3A,B). Moreover, from the differential expression analysis, 2052 deregulated circRNAs were identified (|logFC| ≥ 1.0 and *P*-value ≤0.05), including 1197 up-regulated and 855 down-regulated (Figure 2B). The GO and KEGG enrichment results (Figure 2C and Supplementary Figure S3C–F) suggested that these up-regulated and down-regulated circRNAs were enriched on the critical biological functions and pathways associated with prostate cancer progression (32–35). These biological functions include protein modification, cell adhesion, actin cytoskeleton, proteoglycans, GTPase activity, ErbB signaling, endocytosis, thyroid hormone signaling, and RNA and DNA processing (Figure 2C and Supplementary Figure S3C–F).

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab809/6371375 by guest on 17 September 2021

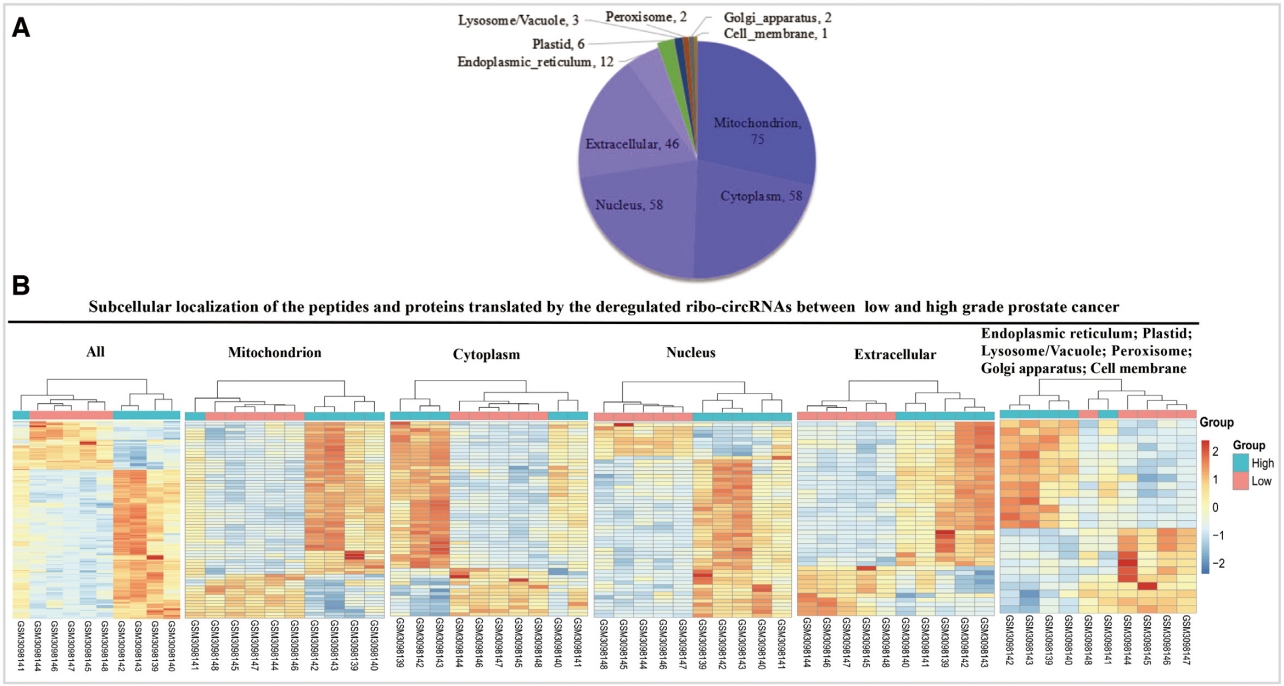


Figure 3. Deregulated ribo-circRNA significantly distinguished different grades of prostate cancer regardless of the subcellular localization of their putative peptides and proteins. (**A**) The distribution of subcellular localization of the 263 putative peptides and proteins translated by the 190 deregulated ribo-circRNAs. (**B**) The heatmaps show that the 190 ribo-circRNAs significantly distinguish different grade prostate cancer regardless of the subcellular localization of their putative peptides and proteins.

Consistent with recent studies (29,36–39), our pilot test-run results supported that circRNA plays critical roles in regulating prostate cancer progression.

We further identified 190 ribo-circRNAs from the 2052 deregulated circRNAs, and 263 putative peptides and proteins translated by these ribo-circRNAs used the ribo-circRNA location tool (Supplementary Table S4). The results from the ribo-circRNA location tool showed that the 263 peptides and proteins have variable subcellular localizations, including mitochondrion (28.53%, 75/263), cytoplasm (22.05%, 58/263), nucleus (22.05%, 58/263), extracellular (17.49%, 46/263), endoplasmic_reticulum (4.56%, 12/263), plastid (2.28%, 6/263), lysosome/vacuole (1.14%, 3/263), peroxisome (0.76%, 2/263), golgi apparatus (0.76%, 2/263) and cell membrane (0.38%, 1/263) (Figure 3A). Furthermore, the heatmap results suggested that the 190 ribo-circRNAs can significantly distinguish different grades of prostate cancer regardless of the subcellular localization of their putative peptides and proteins (Figure 3B).

Next, circRNA IRES prediction tool was used to identify 79 circRNAs that have at least one experimentally validated human IRES (*E*-value $\leq$ 1E-5) from the 2052 deregulated circRNAs (Figure 4A and Supplementary Table S5). Moreover, from the 79 circRNAs, we identified five ribo-circRNAs associated with prostate cancer progression with best translation potential, including hsa_circ_0003700, hsa_circ_0003458, hsa_circ_0001112, hsa_circ_0008351 and hsa_circ_0003643 (Figure 4A and B; Supplementary Figure S4A and Supplementary Table S6). In addition, the boxplot and corrplot analysis in the co-expression module showed that they are significantly correlated with each other (Figure 4C and Supplementary Figure S4B–K). The circRNA-miRNA prediction discovered 1532 miRNAs (score $\geq$ 140 and energy $\leq$ -7.0) that may be able to interact with the five circRNAs (Supplementary Table S7). Finally, the function enrichment results from the co-expression circRNAs of the five circRNAs are consistent with those enrichment results of the 2052 deregulated circRNAs, suggesting that these five circRNAs may play a vital role in regulating prostate cancer progression through translating themselves and interacting with the miRNAs, but future validation *in vitro* and *in vivo* are needed (Figures 2C and 4D–H). In summary, the case study demonstrated that circMine can serve as an important resource to improve our insight in understanding the significance of circRNA in human diseases.

DISCUSSION

circMine is the first comprehensive database designed to systematically integrate, analyze and visualize human circRNA transcriptome data on specific physiological and pathological conditions. Compared with other circRNA databases (7–11,13–15,21,22,40,41), which only provide genomics, expression patterns, and functional and structure annotations for common circRNA and tissue specific circRNA, circMine can provide unique data and significant function to fill some of the service gaps. For example, in contrary to the disease-related databases such as CircR2Disease (16) and circRNADisease (17) that provide a limited amount of experimentally validated circRNA-disease association data through manual curation on publication, circMine provides 1 821 448 entries of 136 871 circR-

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab809/6371375 by guest on 17 September 2021

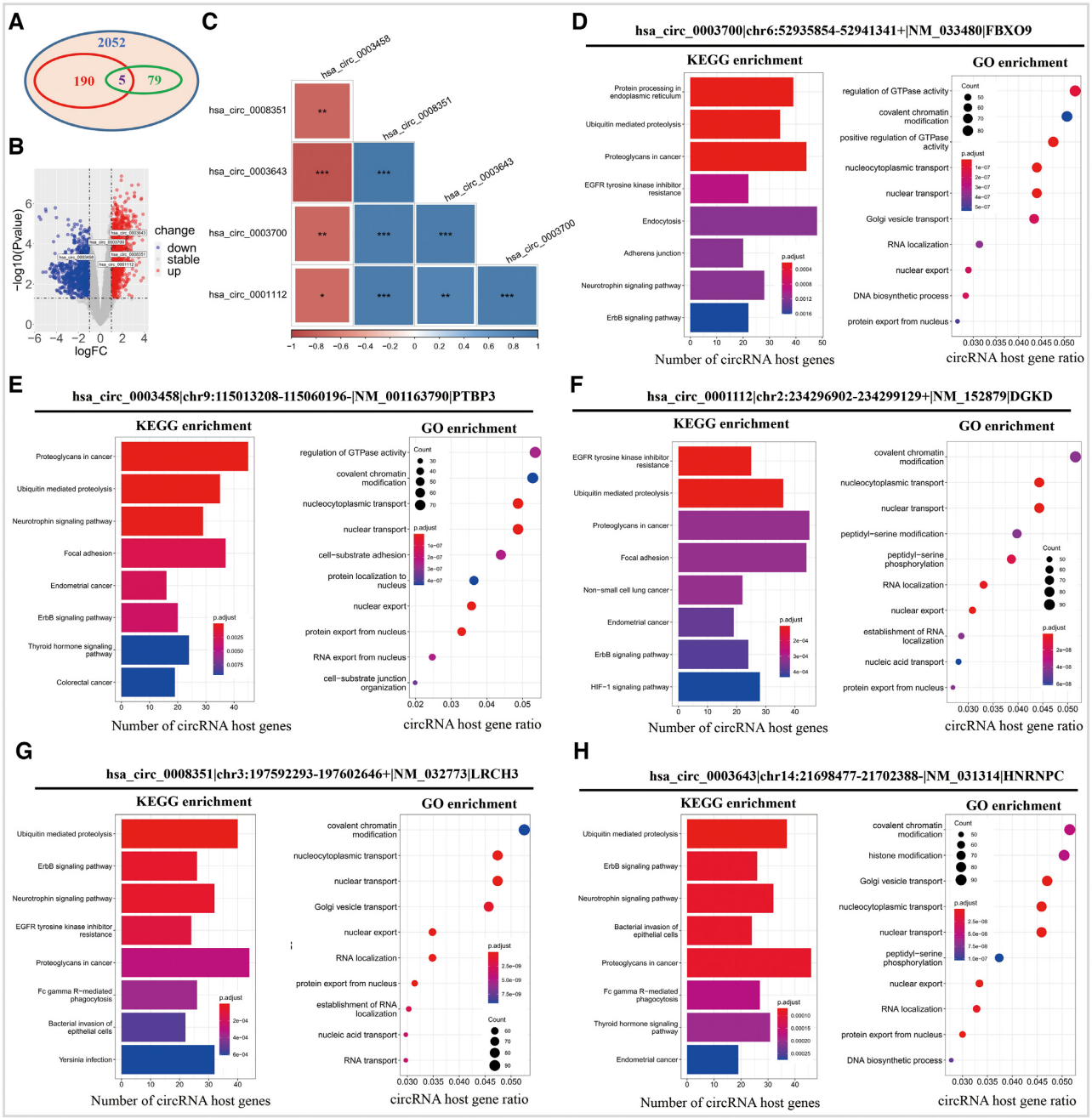


Figure 4. Comprehensive investigation to discover potential translatable circRNAs and identify their biological and clinical significance in prostate cancer progression. (**A**) Of the 2052 deregulated circRNAs, 190 circRNAs are ribo-circRNA, 79 circRNAs have human IRES(s) (*E*-value ≤ 1E-5), and five circRNAs are ribo-circRNA and have experimentally validated human IRES(s), including hsa_circ_0003700, hsa_circ_0003458, hsa_circ_0001112, hsa_circ_0008351 and hsa_circ_0003643. (**B**) The volcano plot shows that the five circRNAs are significantly different between low- and high- grade prostate cancers. (**C**) The corrplot diagram shows that the expression patterns of the five circRNAs are significantly correlated with each other. *: *P*-value ≤ 0.05, **: *P*-value ≤ 0.01 and ***: *P*-value ≤ 0.001. (**D**–**H**) The GO and KEGG enrichment of the five circRNAs based the annotations of the host genes of their co-expression circRNAs (|Correlation| ≥ 0.8 and *P*-value ≤ 0.05).

NAs, 87 diseases and 120 circRNA transcriptome datasets (Supplementary Table S1). Moreover, circMine provides 13 different analytical functions with customized grouping and setting features to analyze and visualize these integrated circRNA transcriptome data for identifying the circRNA biomarkers in human diseases (Table 1, Figure 1B, upper left panel and Supplementary Table S1). And both the disease-related circRNA transcriptome datasets and 13 analytical functions are not provided in the existing circRNA databases such as circAtlas (22), circRNADb (15), Circbank (21), CircR2Disease (16) and circRNADisease (17) (Supplementary Table S1). Furthermore, circMine provides the Web Server tool and allows researchers to study their own circRNA transcriptome datasets by using

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab809/6371375 by guest on 17 September 2021

the aforementioned 13 analytical functions, while only circAtlas (22) supports two of those functions without customized grouping and setting features (Table 1, Figure 1B, upper left panel and Supplementary Table S1). Although circAtlas (22), circRNADb (15) and Circbank (21) provide biological annotations (such as interaction and translatability) of circRNA, circMine also provides three additional tools to predict biological functions of circRNA to enhance the application efficiency of circMine in the circRNA research community, including circRNA-miRNA prediction, circRNA IRES prediction and ribo-circRNA location (Figure 1B, lower left panel and Supplementary Table S1). The ribo-circRNA location function in circMine is not provided in those available databases (Supplementary Table S1). Different from the online tools such as CIRCexplorer (6) that is for the upstream analysis on the circRNA sequencing data to identify and annotate circRNA and quantify its expression, circMine aims for the downstream analysis on the disease-related circRNA transcriptome data to identify the clinical and biological significance of circRNA in different human diseases (Figure 1 and Supplementary Table S1). Thus, circMine is significantly different from all the existing circRNA databases and tools, and it provides a new data and function platform that is not currently available.

As a test-run pilot study, circMine systematically identified five circRNAs associated with prostate cancer progression. The results suggest that they may modify several critical biological functions and pathways through translating themselves and interacting with miRNAs, but it needs to be further validated *in vitro* and *in vivo*. Thus, as a unique resource, circMine can serve as an invaluable tool for bench and computational researchers to conduct in-depth investigation about the role of circRNA in human diseases.

In the coming future, more and more human circRNA transcriptome data are expected to be generated by different high-throughput platforms. To better serve the research community, circMine will continue to enrich its data resource and analytical power by integrating both public and newly shared private data. It will provide new functionalities for analyzing and visualizing circRNA data periodically at every 6 months. More data features can also be included through integrating the circRNA data related to genomics, proteomics, epigenetics and even the circRNA data of other organisms from the public resources, which include Catalogue of Somatic Mutations in Cancer (42), GWAS Catalog (43), The Cancer Genome Atlas (44), International Cancer Genome Consortium (45) and GEO (23,24). For example, we plan to integrate datasets that include both the circRNA expression and linear RNA expression data, and further add the co-expression function of circRNA-linear RNA in the co-expression module when there is enough accessible datasets available in the public resource. We believe that all these additional data and functionalities will enhance the application efficiency of circMine in the circRNA research community.

DATA AVAILABILITY

The circMine database platform is accessible at websites of http://hpcc.siat.ac.cn/circmine and http://www.biomedical-web.com/circmine.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the supports from The Clinical, Translational and Basic Research Laboratory of The University of Hong Kong - Shenzhen Hospital.

FUNDING

National Key R&D Program of China [2018YFB0204403 to Y.W.]; Guangdong Basic and Applied Basic Research Foundation, China [2020A1515110528 to W.Z.]; National Natural Science Foundation of China [32100513 to W.Z.]; China Postdoctoral Science Foundation [2021M693302 to W.Z.]; Strategic Priority CAS Project [XDB38000000 to Y.W.]; Shenzhen Basic Research Fund [RCYX20200714114734l9, KQTD20200820113106007, JSGG20201102163800001 to Y.W.]; Sanming Project of Medicine (Shenzhen) [SZSM201911016 to Q.Z.]. Funding for open access charge: China Postdoctoral Science Foundation [2021M693302 to W.Z.].
Conflict of interest statement. None declared.

REFERENCES

1. Kristensen,L.S., Andersen,M.S., Stagsted,L., Ebbesen,K.K., Hansen,T.B. and Kjems,J. (2019) The biogenesis, biology and characterization of circular RNAs. *Nat. Rev. Genet.*, **20**, 675–691.
2. Chen,L.L. (2020) The expanding regulatory mechanisms and cellular functions of circular RNAs. *Nat. Rev. Mol. Cell Biol.*, **21**, 475–490.
3. Zhang,X.O., Dong,R., Zhang,Y., Zhang,J.L., Luo,Z., Zhang,J., Chen,L.L. and Yang,L. (2016) Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.*, **26**, 1277–1287.
4. Gao,Y., Zhang,J. and Zhao,F. (2018) Circular RNA identification based on multiple seed matching. *Brief. Bioinform.*, **19**, 803–810.
5. Cheng,J., Metge,F. and Dieterich,C. (2016) Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics*, **32**, 1094–1096.
6. Ma,X.K., Xue,W., Chen,L.L. and Yang,L. (2021) CIRCexplorer pipelines for circRNA annotation and quantification from non-polyadenylated RNA-seq datasets. *Methods*, **21**, S1046–S2023.
7. Xia,S., Feng,J., Chen,K., Ma,Y., Gong,J., Cai,F., Jin,Y., Gao,Y., Xia,L., Chang,H. *et al.* (2018) CSCD: a database for cancer-specific circular RNAs. *Nucleic Acids Res.*, **46**, D925–D929.
8. Ruan,H., Xiang,Y., Ko,J., Li,S., Jing,Y., Zhu,X., Ye,Y., Zhang,Z., Mills,T., Feng,J. *et al.* (2019) Comprehensive characterization of circular RNAs in ∼ 1000 human cancer cell lines. *Genome Med*, **11**, 55.
9. Xia,S., Feng,J., Lei,L., Hu,J., Xia,L., Wang,J., Xiang,Y., Liu,L., Zhong,S., Han,L. *et al.* (2017) Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. *Brief. Bioinform.*, **18**, 984–992.
10. Li,S., Li,Y., Chen,B., Zhao,J., Yu,S., Tang,Y., Zheng,Q., Li,Y., Wang,P., He,X. *et al.* (2018) exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucleic Acids Res.*, **46**, D106–D112.
11. Dudekula,D.B., Panda,A.C., Grammatikakis,I., De,S, Abdelmohsen,K. and Gorospe,M. (2016) CircInteractome: a web tool for exploring circular RNAs and their interacting proteins and microRNAs. *Rna Biol*, **13**, 34–42.
12. Liu,Y.C., Li,J.R., Sun,C.H., Andrews,E., Chao,R.F., Lin,F.M., Weng,S.L., Hsu,S.D., Huang,C.C., Cheng,C. *et al.* (2016) CircNet: a database of circular RNAs derived from transcriptome sequencing data. *Nucleic Acids Res.*, **44**, D209–D215.
13. Li,H., Xie,M., Wang,Y., Yang,L., Xie,Z. and Wang,H. (2021) riboCIRC: a comprehensive database of translatable circRNAs. *Genome Biol.*, **22**, 79.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab809/6371375 by guest on 17 September 2021

14. Huang,W., Ling,Y., Zhang,S., Xia,Q., Cao,R., Fan,X., Fang,Z., Wang,Z. and Zhang,G. (2021) TransCirc: an interactive database for translatable circular RNAs based on multi-omics evidence. *Nucleic Acids Res.*, **49**, D236–D242.
15. Chen,X., Han,P., Zhou,T., Guo,X., Song,X. and Li,Y. (2016) circRNADb: a comprehensive database for human circular RNAs with protein-coding annotations. *Sci. Rep.*, **6**, 34985.
16. Yao,D., Zhang,L., Zheng,M., Sun,X., Lu,Y. and Liu,P. (2018) Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. *Sci. Rep.*, **8**, 11018.
17. Zhao,Z., Wang,K., Wu,F., Wang,W., Zhang,K., Hu,H., Liu,Y. and Jiang,T. (2018) circRNA disease: a manually curated database of experimentally supported circRNA-disease associations. *Cell Death Dis.*, **9**, 475.
18. Zhang,W., Zeng,B., Yang,M., Yang,H., Wang,J., Deng,Y., Zhang,H., Yao,G., Wu,S. and Li,W. (2021) ncRNAVar: a manually curated database for identification of noncoding RNA variants associated with human diseases. *J. Mol. Biol.*, **433**, 166727.
19. Zhang,W., Yao,G., Wang,J., Yang,M., Wang,J., Zhang,H. and Li,W. (2020) ncRPheno: a comprehensive database platform for identification and validation of disease related noncoding RNAs. *RNA Biol.*, **17**, 943–955.
20. Glazar,P., Papavasileiou,P. and Rajewsky,N. (2014) circBase: a database for circular RNAs. *RNA*, **20**, 1666–1670.
21. Liu,M., Wang,Q., Shen,J., Yang,B.B. and Ding,X. (2019) Circbank: a comprehensive database for circRNA with standard nomenclature. *RNA Biol.*, **16**, 899–905.
22. Wu,W., Ji,P. and Zhao,F. (2020) CircAtlas: an integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. *Genome Biol.*, **21**, 101.
23. Sayers,E.W., Beck,J., Bolton,E.E., Bourexis,D., Brister,J.R., Canese,K., Comeau,D.C., Funk,K., Kim,S., Klimke,W. *et al.* (2021) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **49**, D10–D17.
24. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
25. Miranda,K.C., Huynh,T., Tay,Y., Ang,Y.S., Tam,W.L., Thomson,A.M., Lim,B. and Rigoutsos,I. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.
26. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
27. Boratyn,G.M., Camacho,C., Cooper,P.S., Coulouris,G., Fong,A., Ma,N., Madden,T.L., Matten,W.T., McGinnis,S.D., Merezhuk,Y. *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.*, **41**, W29–W33.
28. Zhao,J., Li,Y., Wang,C., Zhang,H., Zhang,H., Jiang,B., Guo,X. and Song,X. (2020) IRESbase: a comprehensive database of experimentally validated internal ribosome entry sites. *Genomics Proteomics Bioinformatics*, **18**, 129–139.
29. Almagro,A.J., Sonderby,C.K., Sonderby,S.K., Nielsen,H. and Winther,O. (2017) DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, **33**, 3387–3395.
30. Kuhn,R.M., Haussler,D. and Kent,W.J. (2013) The UCSC genome browser and associated tools. *Brief. Bioinform.*, **14**, 144–161.
31. Yang,Z., Qu,C.B., Zhang,Y., Zhang,W.F., Wang,D.D., Gao,C.C., Ma,L., Chen,J.S., Liu,K.L., Zheng,B. *et al.* (2019) Dysregulation of p53-RBM25-mediated circAMOTL1L biogenesis contributes to prostate cancer progression through the circAMOTL1L-miR-193a-5p-Pcdha pathway. *Oncogene*, **38**, 2516–2532.
32. De Piano,M., Manuelli,V., Zadra,G., Otte,J., Edqvist,P.D., Ponten,F., Nowinski,S., Niaouris,A., Grigoriadis,A., Loda,M. *et al.* (2020) Lipogenic signalling modulates prostate cancer cell adhesion and migration via modification of Rho GTPases. *Oncogene*, **39**, 3666–3679.
33. Aksoy,O., Pencik,J., Hartenbach,M., Moazzami,A.A., Schlederer,M., Balber,T., Varady,A., Philippe,C., Baltzer,P.A., Mazumder,B. *et al.* (2021) Thyroid and androgen receptor signaling are antagonized by mu-Crystallin in prostate cancer. *Int. J. Cancer*, **148**, 731–747.
34. Edwards,I.J. (2012) Proteoglycans in prostate cancer. *Nat Rev Urol*, **9**, 196–206.
35. Miller,D.R., Ingersoll,M.A. and Lin,M.F. (2019) ErbB-2 signaling in advanced prostate cancer progression and potential therapy. *Endocr. Relat. Cancer*, **26**, R195–R209.
36. Feng,Y., Yang,Y., Zhao,X., Fan,Y., Zhou,L., Rong,J. and Yu,Y. (2019) Circular RNA circ0005276 promotes the proliferation and migration of prostate cancer cells by interacting with FUS to transcriptionally activate XIAP. *Cell Death Dis.*, **10**, 792.
37. Wu,Y.P., Lin,X.D., Chen,S.H., Ke,Z.B., Lin,F., Chen,D.N., Xue,X.Y., Wei,Y., Zheng,Q.S., Wen,Y.A. *et al.* (2020) Identification of prostate cancer-related circular RNA through bioinformatics analysis. *Front. Genet.*, **11**, 892.
38. Wang,S., Su,W., Zhong,C., Yang,T., Chen,W., Chen,G., Liu,Z., Wu,K., Zhong,W., Li,B. *et al.* (2020) An eight-circRNA assessment model for predicting biochemical recurrence in prostate cancer. *Front. Cell Dev. Biol.*, **8**, 599494.
39. Vo,J.N., Cieslik,M., Zhang,Y., Shukla,S., Xiao,L., Zhang,Y., Wu,Y.M., Dhanasekaran,S.M., Engelke,C.G., Cao,X. *et al.* (2019) The landscape of circular RNA in cancer. *Cell*, **176**, 869–881.
40. Dong,R., Ma,X.K., Li,G.W. and Yang,L. (2018) CIRCpedia v2: an updated database for comprehensive circular RNA annotation and expression comparison. *Genomics Proteomics Bioinformatics*, **16**, 226–233.
41. Feng,J., Xiang,Y., Xia,S., Liu,H., Wang,J., Ozguc,F.M., Lei,L., Kong,R., Diao,L., He,C. *et al.* (2018) CircView: a visualization and exploration tool for circular RNAs. *Brief. Bioinform.*, **19**, 1310–1316.
42. Tate,J.G., Bamford,S., Jubb,H.C., Sondka,Z., Beare,D.M., Bindal,N., Boutselakis,H., Cole,C.G., Creatore,C., Dawson,E. *et al.* (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **47**, D941–D947.
43. Buniello,A., MacArthur,J., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
44. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
45. Zhang,J., Bajari,R., Andric,D., Gerthoffert,F., Lepsa,A., Nahal-Bose,H., Stein,L.D. and Ferretti,V. (2019) The international cancer genome consortium data portal. *Nat. Biotechnol.*, **37**, 367–369.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab809/6371375 by guest on 17 September 2021