CropCircDB: a comprehensive circular RNA resource for crops in response to abiotic stress

Authors: Kai Wang, Kai Wang

Date: 2023-02-09T00:00:00+00:00

Abstract

Circular RNAs (circRNAs) may mediate mRNA expression as miRNA sponges. As the research community has increasingly focused on circRNAs, numerous circRNA databases have been developed for plants. However, a comprehensive collection of circRNAs involved in crop responses to abiotic stress is still lacking. In this work, we employed a big-data approach to leverage large-scale sequencing data and developed a comprehensive circRNA resource, CropCircDB, initially for maize and rice, with subsequent extension to additional crop species. We also designed a metric, the stress detection score, specifically for identifying circRNAs under stress conditions. In summary, we systematically analyzed 244 and 288 RNA-seq samples for maize and rice, respectively, identifying 38,785 circRNAs in maize and 63,048 circRNAs in rice. This resource not only supports a user-friendly JBrowser for easy genome visualization but also provides an elegant view of circRNA structures and dynamic expression profiles across all samples. Collectively, this database will host all predicted and validated crop circRNAs responsive to abiotic stress.

Full Text

Preamble

CropCircDB: a comprehensive circular RNA resource for crops in response to abiotic stress

Kai Wang[1],†, Chong Wang[1],†, Baohuan Guo[2], Kun Song[1], Chuanhong Shi[1], Xin Jiang[1], Keyi Wang[1], Yacong Tan[1], Lequn Wang[1], Lin Wang[2], Jiangjiao Li[1], Ying Li[1], Yu Cai[1], Hongwei Zhao[2],* and Xiaoyong Sun[1],*

[1]Agricultural Big-Data Research Center, College of Information Science and Engineering, Shandong Agricultural University, Taian 271018, China
[2]Department of Plant Pathology, Nanjing Agricultural University, Nanjing 210095, China

Corresponding authors:
Hongwei Zhao: Tel.: +86-25884399552; Fax: +86-25884399552; Email: hzhao@njau.edu.cn
Xiaoyong Sun: Tel.: +86-5388249879; Fax: +86-5388241878; Email: johnsunx1@yahoo.com

†These authors contributed equally to this article.

Abstract

Circular RNAs (circRNAs) can mediate mRNA expression by acting as miRNA sponges. While numerous circRNA databases have been developed for plants, a comprehensive collection of circRNAs specifically involved in crop responses to abiotic stress is still lacking. In this work, we applied a big-data approach to leverage large-scale sequencing data and developed a comprehensive circRNA resource: CropCircDB for maize and rice, with plans to extend it to incorporate additional crop species. We also designed a novel metric, the stress detection score, specifically for detecting circRNAs under stress conditions. In summary, we systematically investigated 244 and 288 RNA-Seq samples for maize and rice, respectively, identifying 38,785 circRNAs in maize and 63,048 circRNAs in rice. This resource not only provides a user-friendly JBrowser for easy genome visualization but also offers elegant views of circRNA structures and dynamic profiles of circRNA expression across all samples. Together, this database will serve as a central repository for all predicted and validated crop circRNAs responding to abiotic stress.

Database URL: http://deepbiology.cn/crop/

Introduction

Circular RNA (circRNA) was first reported as sub-viral agents encoding viral genomes in plants. In 2012, circRNAs were discovered to exist widely in eukaryotes, including fungi, protists, and plants, where they function as microRNA sponges and thus mediate mRNA expression (1, 2). Later studies further reported their widespread presence across eukaryotic kingdoms (3). Recently, several teams have characterized circRNAs in plants, including Arabidopsis (4, 5, 6) and rice (7), and verified their important roles in alternative splicing (8). Although substantial work has been done, the functions of circRNAs remain incompletely understood. To date, circRNAs have been reported to mediate mRNA expression as miRNA sponges (1, 2), control protein translation processes (9), or produce proteins directly via translation (10, 11).

As the research community has paid increasing attention to circRNAs, many circRNA databases for human and animal systems have been developed. For

example, Circ2Traits links circRNAs with human diseases and traits (12), while circBase collects thousands of circRNAs from nine independent studies across human, mouse, nematode, and latimeria. CircNet was the first circRNA database derived from large-scale sequencing data. More recently, TSCD was developed as a tissue-specific circRNA database for human and mouse, hosting 302,853 tissue-specific circRNAs (13). Finally, CSCD reports 272,152 cancer-specific circRNAs and 950,962 circRNAs from normal samples (14).

Simultaneously, three plant circRNA databases have been reported to date. PlantcircBase collects 77,595 publicly available circRNAs from rice, Arabidopsis, maize, tomato, and barley (15). PlantCircNet hosts circRNAs from eight plant species and provides plant circRNA-miRNA-gene regulatory networks (16). AtCircDB, developed by our group in 2016 based on large-scale sequencing data (27), hosts 30,648 tissue-specific circRNAs for Arabidopsis derived from 87 independent studies. However, to the best of our knowledge, a comprehensive and systematic collection of circRNAs for crops in response to abiotic stress is still lacking. Building on our previous work (27), we applied a big-data approach to fully leverage large-scale sequencing data and developed a rich stress-specific circRNA resource: CropCircDB (http://genome.sdau.edu.cn/crop/ or http://deepbiology.cn/crop/) for maize and rice, with plans to extend it to additional crop species. This database currently hosts 38,785 circRNAs in maize and 63,048 circRNAs in rice, which are freely available for download.

Materials and Methods

Data Collection

On November 12, 2017, we searched the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra) using keywords including 'drought' , 'cold' , 'heat' , 'salt' , 'flood' , and 'high wind' to identify RNA-Seq datasets for two crops: maize and rice. We retained only three abiotic stresses ('drought' , 'cold' , 'salt') with more than 20 samples for analysis. We will extend to other abiotic stresses as more samples become publicly available. These samples represent diverse plant tissues, including root, leaf, flower, and shoot. Additionally, we selected sequencing data without 'PolyA' selection in sample preparation. Finally, we retained samples meeting three criteria: (i) sequencing on the Illumina platform, (ii) file size >1 GB, and (iii) identifiable circRNAs. Detailed information about the sequencing samples is available on the website.

We also searched PubMed (https://www.ncbi.nlm.nih.gov/pubmed) using the terms 'rice, circular RNA' and 'maize, circular RNA' . One maize (17) and three rice (7, 4, 18) articles provided detailed circRNA lists, which we collected and annotated. This circRNA collection is also publicly available on our website.

circRNA Identification

To detect circRNAs, we simultaneously utilized two algorithms—CIRCexplorer2 (19) and CIRI2 (20) with default parameters—to increase prediction accuracy. In the CIRCexplorer2 pipeline, TopHat (21) aligned raw sequencing data to the reference genome with parameters: '–max-multihits 1 -a 6 –microexon-search -m 2' . Unmapped bam files were converted to fastq format using bam2fastx. TopHat then processed fastq files with parameters: '-p 15 –fusion-search –keep-fasta-order –bowtie1 –no-coverage-search' . Finally, CIRCexplorer2 analyzed the results with default parameters.

In the CIRI2 pipeline, sequencing data was first aligned to the reference genome with BWA-MEM using parameter '-T 19' (22). CIRI2 was then applied to the alignment file (SAM format) to detect circRNAs. All detected circRNAs were annotated using SplicingTypesAnno (23) and the Bioconductor package GenomicAlignments (24). We extracted all circRNA sequences using BEDTools (25) and used the Bioconductor package Biostrings (26) to predict amino acid sequences from spliced sequences following the approach of (10).

Detection Score and Stress Detection Score

Following our previous approach (27), we used 'detection score' to measure the robustness of circRNA detection across samples. To further understand circRNA existence under abiotic stress, we designed a new metric: 'stress detection score' as follows:

$$\text{detection score} = \frac{\text{number of samples containing a circRNA}}{\text{total number of samples}} \times 100$$

$$\text{stress detection score} = \frac{\text{number of stressed samples containing a circRNA}}{\text{total number of stressed samples}} \times 100$$

If this score is 100, the circRNA was detected in all related samples; if 0, it was not found in any related samples. This metric helps experimental biologists rank circRNAs and design further analyses.

Analysis of miRNA-circRNA Interaction

To understand the relationship between miRNA and circRNA, we downloaded microRNA information from miRBase (http://mirbase.org/). We extracted all miRNA and circRNA sequences using the Bioconductor package Biostrings and BEDTools. To predict interactions between miRNAs and circRNAs, we utilized psRNATarget (28) and uploaded circRNA sequences to the website (https://plantgrn.noble.org/psRNATarget/analysis?function=2). After analyzing results with default scoring parameters, we extracted potential miRNA-circRNA interactions using R. The final results were annotated and deposited into the database.

Super circRNA Region

To help group circRNAs, we followed our previous approach (27) and used super circRNA regions to cluster circRNAs originating from the same genomic locus. The pipeline is as follows: first, we collapsed all overlapped circRNAs into one region and defined the number of circRNAs in one region as Ci. Second, we analyzed Ci using five-number summary (min, Q1, median, Q3, max). We then calculated super circRNA regions as those containing more circRNAs than CQ1 + 1.5(CQ3 − CQ1). The final results were annotated and deposited into the database.

Database Construction

We used PHP and MySQL to develop the database. Genomic visualization was accomplished through JBrowse (29). Two tracks—including gene and circRNA annotation—were implemented using JavaScript. Gene tracks were imported using GFF3 files downloaded from EnsemblPlants (http://plants.ensembl.org/). CircRNA tracks were imported using annotations from the circRNA identification pipeline. Additionally, circRNA structure and expression visualization were developed using D3 (https://d3js.org/). CircRNA structure was generated dynamically with exon annotation for each circRNA. Expression visualization was generated based on circRNA expression (RPM: reads per million mapped reads) labeled with sample ID from the NCBI SRA database. To improve website accessibility, we developed both an institutional website (http://genome.sdau.edu.cn/crop/) and a mirror site at a commercial organization (http://deepbiology.cn/crop/).

Results and Discussion

In this study, we systematically investigated 244 maize samples and 288 rice samples from diverse tissues including leaf, root, and shoot. All samples were downloaded from the NCBI SRA database, and circRNAs were detected using two algorithms: CIRCexplorer2 and CIRI2. The circRNAs were further processed by psRNATarget to predict potential miRNA target sites [Figure 1: see original paper]. In total, we identified 38,785 circRNAs in maize and 63,048 circRNAs in rice. The median circRNA length was approximately 261 nt for maize and 260 nt for rice, with 27% and 38% of genes hosting circRNAs, respectively. More than half of circRNAs originated from single exons, suggesting that circRNAs are generated from fewer exons, consistent with our previous findings (27). Notably, 85% and 75% of circRNAs overlapped with exon boundaries, while 4% and 3% originated from intergenic regions in maize and rice, respectively.

To investigate environmental effects on circRNAs, we systematically analyzed 111 stress-related maize samples and 148 stress-related rice samples . For maize, we collected 85 drought samples versus 73 control samples and 23 salt samples

versus 4 control samples. For rice, we collected 60 drought samples versus 47 control samples, 29 salt samples versus 29 control samples, and 46 cold samples versus 16 controls. We identified 12,643 and 15,588 circRNAs in control samples for maize and rice, respectively. For maize, we found 11,206 drought-specific circRNAs and 6,770 salt-specific circRNAs. For rice, we found 824 drought-specific circRNAs, 6,313 salt-specific circRNAs, and 5,724 cold-specific circRNAs. All stress-related circRNAs and tissue information were deposited in CropCircDB.

To understand the relationship between circRNAs and proteins, we extracted all circRNA sequences using SplicingTypesAnno and BEDTools, removed intron sequences, and retained spliced sequences. Following the approach of (11), we translated all RNA sequences into amino acid sequences. Sequences without stop codons were stored as predicted proteins (10). Consequently, the database hosts not only full sequences from start to end of circRNAs but also spliced sequences without introns and predicted proteins.

Currently, CropCircDB provides the following information: (i) **circRNA name**. The naming system follows previous work (30) and incorporates species abbreviation, circRNA IDs, and gene names, facilitating convenient linking of circRNAs with genes and enabling comparison, querying, retrieval, and storage of circRNA clusters; (ii) **circRNA information**, including chromosome, start, end, strand, length, and antisense information; (iii) **detection score and stress detection score**. Detection score measures the probability of finding a circRNA in samples, calculated as the number of samples with detected circRNA divided by total number of samples. A high detection score suggests high probability of circRNA presence. Similarly, stress detection score measures robustness in stressed samples, calculated as the number of stressed samples with detected circRNAs divided by total number of stressed samples; (iv) **experimental evidence**. All validated circRNAs are annotated as 'validated' , and the website accepts community submissions, with new circRNAs deposited within 24 hours; (v) **potential miRNA-circRNA interactions**. Since circRNAs function as miRNA sponges, we analyzed all circRNAs using psRNATarget. We found 96 miRNAs interacting with 327 circRNAs in maize and 518 miRNAs interacting with 5,475 circRNAs in rice; (vi) **super circRNA regions**. These regions contain highly enriched circRNAs as described in our previous study (27). We extracted them with all related circRNAs, obtaining 3,030 super circRNA regions for maize and 5,813 for rice.

The web interface of CropCircDB includes tutorial, browser, search, download, publication, team, and news information [Figure 2A: see original paper]. The search portal is the main function [FIGURE:2B and C], supporting three key features: (i) **genome visualization** [Figure 2E: see original paper]. JBrowse provides annotation for all linear transcripts, including exons, transcripts, and genes. CircRNAs are highlighted from start to end, enabling easy comparison with other genomic features; (ii) **circRNA structure visualization** [Figure 2F: see original paper]. By inserting circRNAs into the splicing schema, users can compare circRNA structure with related exons from the same isoform; (iii)

circRNA expression visualization [Figure 2D: see original paper]. The platform supports pie charts, scatterplots, and boxplots to visualize the proportion of circRNAs in control versus stressed samples and detailed expression values (RPM) per sample.

Finally, we compared our database with two other plant circRNA resources: PlantcircBase (15) and PlantCircNet (16). These three databases share a few thousand circRNAs and are complementary to each other [Figure 3: see original paper].

Conclusions

CropCircDB provides a comprehensive platform for circRNAs in maize and rice responding to abiotic stress. It holds detailed circRNA information including genomic locus, gene name, and isoform name, along with extended services such as detection scores, super circRNA regions, miRNA interactions, experimental evidence, and predicted proteins. The platform supports user-friendly genome visualization through JBrowse and offers elegant views of circRNA structures relative to exons. Finally, it provides dynamic profiles of circRNA expression across all samples, enabling users to explore back-splicing and canonical splicing in relation to exons and introns, examine stress-specificity, and unravel functional and regulatory roles. We believe this resource will help the community gain deeper understanding of circRNAs in crops.

Author Contributions

X.S. and H.Z. designed and supervised the project; X.S., K.W., K.S., C.S., K.W., R.L. and Y.T. analyzed the data; C.W., X.J., J.L. and Y.L. developed the database; X.S. wrote the manuscript.

Supplementary Data

Supplementary data are available at Database Online.

Funding

This work was supported by the National Natural Science Foundation of China (grant number 31571306 to X.S.) and Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase) under grant no. U1501501 to X.S. We thank the National Supercomputer Center in Jinan for technical support.

Conflict of Interest

None declared.

References

1. Hansen, T.B., Jensen, T.I., Clausen, B.H. et al. (2013) Natural RNA circles function as efficient microRNA sponges. Nature, 495, 384-388.
2. Memczak, S., Jens, M., Elefsinioti, A. et al. (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. Nature, 495, 333-338.
3. Wang, P.L., Bao, Y., Yee, M.C. et al. (2014) Circular RNA is expressed across the eukaryotic tree of life. PLoS One, 9, e90859.
4. Ye, C.Y., Chen, L., Liu, C. et al. (2015) Widespread noncoding circular RNAs in plants. New Phytol., 208, 88-95.
5. Sun, X., Wang, L., Ding, J. et al. (2016) Integrative analysis of Arabidopsis thaliana transcriptomics reveals intuitive splicing mechanism for circular RNA. FEBS Lett., 590, 3510-3516.
6. Dou, Y., Li, S., Yang, W. et al. (2017) Genome-wide discovery of circular RNAs in the leaf and seedling tissues of Arabidopsis thaliana. Curr. Genom., 18, 360-365.
7. Lu, T., Cui, L., Zhou, Y. et al. (2015) Transcriptome-wide investigation of circular RNAs in rice. RNA, 21, 2076-2087.
8. Conn, V.M., Hugouvieux, V., Nayak, A. et al. (2017) A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation. Nat. Plants., 18, 17053.
9. Holdt, L.M., Stahringer, A., Sass, K. et al. (2016) Circular noncoding RNA ANRIL modulates ribosomal RNA maturation and atherosclerosis in humans. Nat. Commun., 7, 12429.
10. Pamudurti, N.R., Bartok, O., Jens, M. et al. (2017) Translation of CircRNAs. Mol. Cell., 66, 9-21.e7.
11. Yang, Y., Fan, X., Mao, M. et al. (2017) Extensive translation of circular RNAs driven by N6-methyladenosine. Cell Res., 27, 626-641.
12. Ghosal, S., Das, S., Sen, R. et al. (2013) Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. Front. Genet., 4, 283.

13. Xia, S., Feng, J., Lei, L. et al. (2017) Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. Brief Bioinform., 18, 984–992.

14. Xia, S., Feng, J., Chen, K. et al. (2018) CSCD: a database for cancer-specific circular RNAs. Nucleic Acids Res., 46, D925–D929.

15. Chu, Q., Zhang, X., Zhu, X. et al. (2017) PlantcircBase: a database for plant circular RNAs. Mol. Plant., 10, 1126–1128.

16. Zhang, P., Meng, X., Chen, H. et al. (2017) PlantCircNet: a database for plant circRNA-miRNA-mRNA regulatory networks. Database. doi: 10.1093/database/bax089.

17. Chen, L., Zhang, P., Fan, Y. et al. (2018) Circular RNAs mediated by transposons are associated with transcriptomic and phenotypic variation in maize. New Phytol., 217, 1292–1306.

18. Ye, C.Y., Zhang, X., Chu, Q. et al. (2017b) Full-length sequence assembly reveals circular RNAs with diverse non-GT/AG splicing signals in rice. RNA Biol., 14, 1055–1063.

19. Zhang, X.O., Dong, R., Zhang, Y. et al. (2016) Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res., 26, 1277–1287.

20. Gao, Y., Wang, J. and Zhao, F. (2015) CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. Genome Biol., 13, 4.

21. Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25, 1105–1111.

22. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754–1760.

23. Sun, X., Zuo, F., Ru, Y. et al. (2015) SplicingTypesAnno: annotating and quantifying alternative splicing events for RNA-Seq data. Comput. Methods Programs Biomed., 119, 53–62.

24. Lawrence, M., Huber, W., Pagès, H. et al. (2013) Software for computing and annotating genomic ranges. PLoS Comput. Biol., 9, e1003118.

25. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26, 841–842.

26. Pagès, H., Aboyoun, P., Gentleman, R. et al. (2017) String objects representing biological sequences, and matching algorithms. R package version 2.44.2.

27. Ye, J., Wang, L., Li, S. et al. (2017) AtCircDB: a tissue-specific database for Arabidopsis circular RNAs. Brief Bioinform. doi: 10.1093/bib/bbx089.

28. Dai, X. and Zhao, P.X. (2011) psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res., 39, W155–W159.

29. Skinner, M.E., Uzilov, A.V., Stein, L.D. et al. (2009) JBrowse: a next-generation genome browser. Genome Res., 19, 1630–1638.

30. Liu, Y.C., Li, J.R., Sun, C.H. et al. (2016) CircNet: a database of circular RNAs derived from transcriptome sequencing data. Nucleic Acids Res., 44, D209–D215.

Note: Figure translations are in progress. See original paper for figures.