Briefings in Bioinformatics, 23(1), 2022, 1–10

https://doi.org/10.1093/bib/bbab404
Problem Solving Protocol

Characterizing viral circRNAs and their application in identifying circRNAs in viruses

Mengting Niu, Ying Ju, Chen Lin and Quan Zou

Corresponding authors. Quan Zou, Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China; Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, Zhejiang, China. E-mail: zouquan@nclab.net; Chen Lin, School of Informatics, Xiamen University, Xiamen, China. E-mail: chenlin@xmu.edu.cn

Abstract

Circular RNAs (circRNAs) are non-coding RNAs with a special circular structure produced formed by the reverse splicing mechanism, which play an important role in a variety of biological activities. Viruses can encode circRNA, and viral circRNAs have been found in multiple single-stranded and double-stranded viruses. However, the characteristics and functions of viral circRNAs remain unknown. Sequence alignment showed that viral circRNAs are less conserved than circRNAs in animal, indicating that the viral circRNAs may evolve rapidly. Through the analysis of the sequence characteristics of viral circRNAs and circRNAs in animal, it was found that viral circRNAs and animals circRNAs are similar in nucleic acid composition, but have obvious differences in secondary structure and autocorrelation characteristics. Based on these characteristics of viral circRNAs, machine learning algorithms were employed to construct a prediction model to identify viral circRNA. Additionally, analysis of the interaction between viral circRNA and miRNAs showed that viral circRNA is expected to interact with 518 human miRNAs, and preliminary analysis of the role of viral circRNA. And it has been also found that viral circRNAs may be involved in many KEGG pathways related to nervous system and cancer. We curated an online server, and the data and code are available: http://server.malab.cn/viral-CircRNA/.

Key words: viral circRNA; conservation; characteristic; identify; function

Introduction

Circular RNAs (circRNAs) were first detected in plant viral RNA pathogens [1], called viroids [2, 3]. Studies have shown that circRNA is synthesized from pre-mRNA through a nuclear reverse splicing mechanism [4, 5]. Some circRNAs contain scrambled exons, which are produced by abnormal splicing mechanisms [6]. Most circRNAs come from exons with inverted intron sequences on both sides, and RNA circularization is achieved by participating in base pair interactions. There are approximately 5000–25,000 circRNAs in each cell, and 20% of the transcribed genes produce unique circRNAs, which are expressed in a tissue-specific manner [7–9]. Initially, it was thought that circRNA was a byproduct of splicing errors and had few functions [6, 10]. Recent studies have shown that circRNAs can regulate gene splicing or transcription, such as the sponge effect of miRNAs, which relieves the inhibitory effect of miRNA on its target genes [11], transcription regulation, and adsorption of miRNAs to affect the expression of it and its target genes [12–14]. Moreover, RNA sequencing (RNA-seq) has revealed that circRNA is widespread [15, 16]. More than 10,000 circRNAs have been identified in various animals such as humans, mice, nematodes, macaques, fruit flies and marlinus [12]. In addition, circRNAs have also been shown to exist in plants, such as rice, and in protists [13]. With the discovery of more circRNAs in animals and plants, researchers have begun to ask whether viruses produce circRNAs.

Recent studies have shown that the virus encodes the entire sequence of the circRNA, and some viruses express large amounts of circRNA. Viral circRNA is a nonreplicating, noninfectious RNA transcript obtained through reverse splicing of viral genes [17, 18]. It has been identified and discovered in

Mengting Niu is currently doctor degree candidate in University of Electronic Science and Technology of China. Her research interests include circRNAs research and machine learning.
Ying Ju is doing research at Xiamen University. Her research interests include bioinformatics and machine learning.
Chen Lin is an associate professor in Tianjin University. Her main research interests include computational biology and algorithm.
Quan Zou is is a professor in University of Electronic Science and Technology of China. His research interests include bioinformatics and machine learning.
Submitted: 14 July 2021; **Received (in revised form):** 23 August 2021

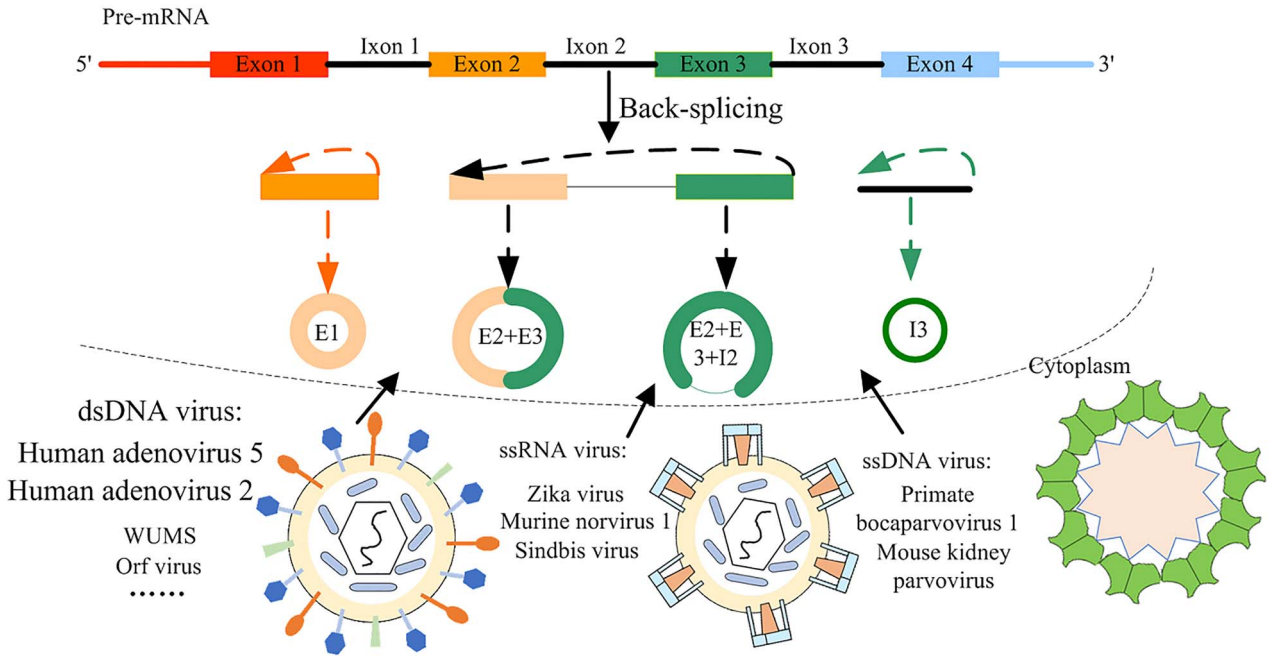

Figure 1. The biogenies of the viral circRNAs.

a variety of single- and double-stranded viruses (Figure 1) [19]. Erik Flemington's research team found that viral circRNA can be expressed between the latency and lysis cycles of Epstein–Barr virus (EBV), and spanned cell lines in different latency states [20]. Huang *et al.* performed RNA-seq on the RNA consumed by ribosomal RNA in various latent cell line models of EBV infection and confirmed that the circRNA encoded by EBV is in cell lines and tumor tissues (type I latent EBV-related gastric cancer and type II latent nasopharyngeal carcinoma) [20, 21]. Two other reports have found that EBV-circrpms1, circlmp2 and EBV-circbhlf1 can encode circRNAs [21, 22]. Studies have found that Burkitt's lymphoma, EBV-associated gastric cancer, nasopharyngeal carcinoma and AIDS-related lymphoma were expressed in EBV-positive cell lines and tissues [23, 24]. And Toptan's research shows that besides EBV virus, Kaposi's sarcoma herpesvirus (KSHV) can also encode circRNA. Moreover, recent studies have proved that DNA viruses have found viral circRNAs, for instance, herpesviruses and papillomaviruses [25]. Interestingly, the circRNAs in EBV virus are encoded by latent genes, and the expression of a part of the viral circRNA after lysis activation is up-regulated [26]. Moreover, the biological function and significance of viral circRNA have been explored and discovered [27–29]. CircRNA can show antiviral effects. A special upregulated circRNA, circPSD3, has shown significant effects on viral RNA abundance in cells infected with hepatitis C virus and dengue virus [2]. It has been found that the expression levels of some viral circRNAs (such as circRPMS1_E4_E3a and circBHLF1) are not much different from the host circRNA levels, or even higher, which indicates the potential biological significance of viral circRNAs [30]. Furthermore, researchers have found that cells infected with EBV and KSHV not only change the cell morphology of circRNA but also show evidence of viral circRNA [10, 20, 21]. The discovery of latent related circRNAs has increased the repertoire of potential future therapeutic targets. EBV ring RNA has other non-microRNA sponge functions [31]. The circRNA in tumor viruses is a long-lived and unique tumor biomarker. Because of its characteristics, this provides new research direction for understanding how these viruses cause cancer [32–34].

At present, although research on viral circRNA has achieved certain results [35, 36], there are still many problems and challenges. These include: (1) How does the virus synthesize circRNA? (2) What is the function of the viral circRNA? (3) What is the difference between viral circRNA and circRNA of animal and plant? (4) Whether all viruses can encode circRNA, or whether circRNA only exists specifically. To better understand the functions of viral circRNAs and study the characteristics of viral circRNAs, in this paper, we first analyzed the conservation of viral circRNA, then explored the main characteristic relationship between viral circRNA and animal and plant circRNA, and employed these characteristics to predict new circRNAs in viruses. After that, we analyzed viral circRNA. Subsequently, machine learning methods were applied to build a model for identifying viral circRNA. Next, the interactions between circRNA and miRNA were evaluated. Furthermore, gene ontology (GO) and kyoto encyclopedia of genes and genomes (KEGG) analyses were carried out for the target genes of circRNA to deeply explore the function of circRNA (Figure 2).

Results

Sequence conservation of viral circRNA

First, the Clustal X method was used for multiple sequence alignment of the 1592 viral circRNA sequences. The results revealed that HAdV5_circ_Homo_sapiens_5253 and HAdV2_circ_Homo_sapiens_6330 have the highest similarity, which is 99.524%. The similarity of EBV_circ_Homo_sapiens_1742 and macacine herpesvirus 4 was 88.5%. The similarity between EBV_circ_Homo_sapiens_1683 and macacine herpesvirus 4 was 88.1%. Both of these viruses belong to the herpesvirus family; HAdV5_circ_Homo_sapiens_5255/7–6, EBV_circ_Homo_sapiens_circle 13912_HS1_PR327/HSV_circ_Homo_sapiens_1683-623 and HSV1_circ_

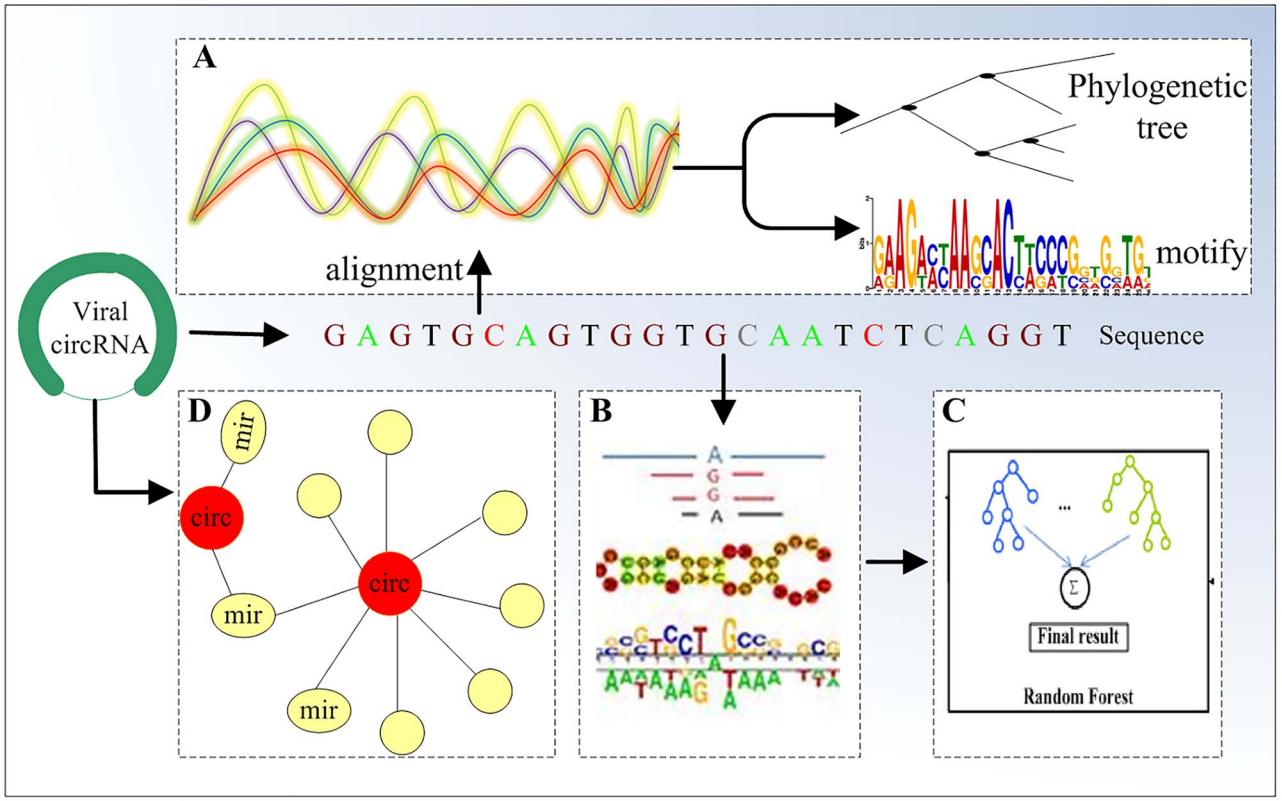

Figure 2. Flowchart of viral circRNA research. **A**. Conservative analysis. **B**. Analysis of sequence characteristics. **C**. Construction of a prediction model of viral circRNAs. **D**. Analysis of interaction between viral circRNAs and miRNAs.

Homo_sapiens_4/1471-1470, HSV1_circ_Homo_sapiens_825/499–498, HSV1_circ_Homo_sapiens_7329/3878-3877 did not match any of the above. Therefore, we deleted the nonconserved sequences and the low-similarity sequences, obtained 121 sequences and then used DNAMAN for sequence alignment and MEGA 5.0 software to construct a phylogenetic tree (Figure 3A). The analysis found that 121 viral circRNAs clustered into five subgroups, and each subgroup contained a different number of viral circRNAs. Among them, the IV subgroup contained the least number of sequences, and the III subgroup contained the most number of sequences. It can be seen from the phylogenetic tree that there are very close evolutionary relationships and both close and long genetic distances within the phylogenetic tree. The distribution of viral circRNAs in different viruses on the phylogenetic tree was clustered, suggesting that viral circRNAs derived from the same or similar genetic relationship may have the same or similar characteristics. This provided a basis for the study of the characteristics of the circRNA of the new virus.

We used three tools, MEME, Homer, and STREME, to predict motif. The seven motifs contain nucleotide sequences of different lengths. The motif and *P* value results are shown in Figure 3B (In order to compare the motifs obtained by the three methods more intuitively, we changed the display order of the motifs, which took the motif of MEME as the reference basis, and intuitively changed the order of Homer, and STREME). Although the lengths of the obtained motifs were not the same, it can also be found that there are similarities in the probability of the occurrence of bases at some sites. For example, in the third motif, there are fragments of ATAAA in the results obtained by the three methods; in the fifth motif, there is an obvious high probability distribution of base A. There is an AAA fragment in the second motif and a CTC fragment in the seventh motif. These predicted motifs have not been discovered by any early studies and cannot be verified with existing knowledge, and we hope that they can be further verified in future experiments.

Sequence characteristics analysis of viral circRNA

We explored the similarities and differences between the characteristics of viral circRNA and animal circRNA. To make a more intuitive graphical expression, a dimensionality reduction algorithm—uniform manifold approximation and projection (UMAP) for dimension reduction [37] was used to visualize it in a two-dimensional space. Figure 4A shows the distribution map of the two datasets simplified by UMAP for all the learned features.

From Figure 4A, it can be seen that the divergence in the distribution of feature points based on *k*-mer ($k=2$ and 3) is imperceptible. There is a phenomenon of repeated distribution of data points, which shows that in sequence features, *k*-mer indicates that the sequence frequency characteristics of the viruses are not significantly different between circRNAs. In terms of nucleotide composition characteristics and structural characteristics, the distribution of characteristic points is quite dissimilar, with apparent differences. This shows that at the sequence level, the similarity between structural features and nucleotide composition features is less ambiguous. We can also conclude that viral circRNAs and animal circRNAs are similar in their distribution of nucleotides. Viruses may change the structure of circRNAs when encoding them. This may also provide new direction for the study of the structure of viral circRNA.

After analyzing the performance of each feature, all the feature of the datasets were merged and used t-distributed random neighbor embedding (t-SNE) [38] for feature selection

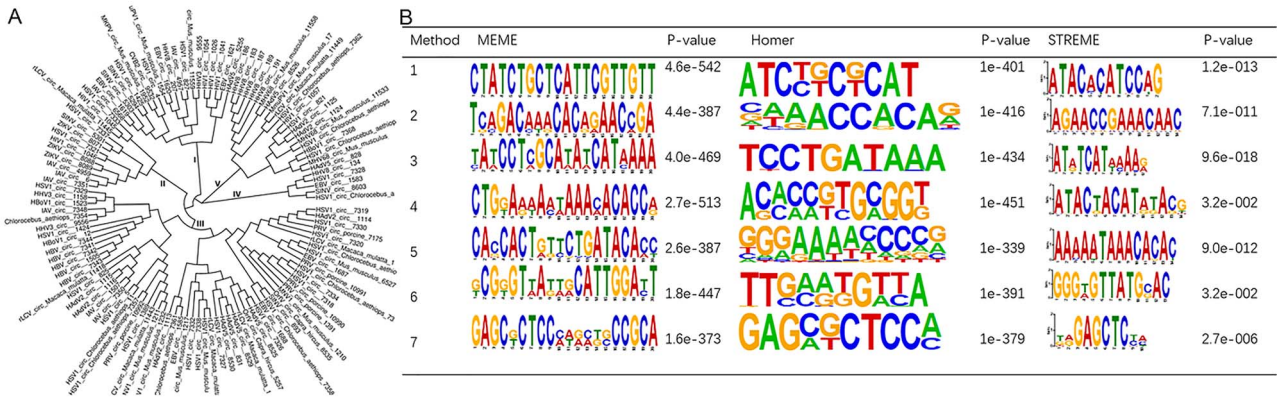

Figure 3. Conservative analysis results. **A**. Phylogenetic tree. **B**. conserved motif.

and visualization in a two-dimensional space. Figure 4B shows the distribution curves of the two datasets simplified by t-SNE for all the learned features.

As shown in Figure 4B, the distribution boundaries of viral circRNAs and animal circRNAs were relatively explicit. The characteristics after t-SNE can clearly distinguish the two, indicating that the distinction in viral circRNAs and animal circRNAs are quite evident. The positive and negative datasets have a significant correlation with the data displayed on the horizontal axis, that is, the normalized basic polar coordinate vector.

Therefore, we drew a violin graph to visualize the horizontal vector of t-SNE results of viral circRNAs and animal and plant circRNAs (Figure 4C). There were still noticable differences in the distribution of characteristic data between viral circRNAs and animal circRNAs, further demonstrating the characteristic diverse between the two.

Predicting viral circRNAs using sequence features

Based on the above analysis of the sequence characteristics of viral circRNA, we tried to use generally used machine learning algorithms [MLAs, such as NaiveByes (NB), support vector machine (SVM) and Random forest (RF)] to analyze viral circRNA (positive example) and non-viral circRNA (negative example, mentioned in the Materials and Methods) in human. We constructed a predictive model and used the 10-fold cross-validation for model verification. The accuracy of using each single feature, the feature after feature fusion (written as all feature) are summarized in Figure 5A.

From the prediction results, we identified different prediction effects of each feature. Moreover, these analyses highlighted the features that were more important. Structural features achieved the best results among all single features, and the accuracy of the three classifiers was 75.107, 75.130, and 78.433%, respectively. The opposite sequence composition characteristics achieved the worst results, especially the worst when $k = 2$, and the accuracy rates were 64.391, 59.782 and 67.391%. After feature fusion, the prediction accuracy improved, and among the three classifiers, RF achieved the best results. These results also verified the feasibility of using machine learning to study viral circRNA. The sequence features related to viral circRNA shown in this paper will also be applied to the next step of viral circRNA research.

To verify the robustness and generalization of RF, independent verification is entailed. Consequently, we used 80% of the datasets to train a model, and the left-over 20% were used for independent test set validation (Figure 5B). From Figure 5B, it can know that RF achieve the best ACC (83.032%) than other algorithms (NB is 76.195%, SVM is 75.858%). It demonstrated that the prediction performance of RF is preferable than others. Through the independent test set verification, it can beared out that the classification performance of RF is not accidental and has a certain stability, and RF effectively recognizes new viral circRNA. Tools developed based on RNA-seq data have been widely used. Although tools developed based on sequence data have achieved good results on training and test data, they still hope to be widely used in future practical applications and prove their effectiveness.

Analysis of the interactions of viral circRNA and miRNA

CircRNA can be used as a miRNA sponge that indirectly regulates gene expression. Does viral circRNA interact with miRNA? Furthermore, most viral circRNAs are found in human cell lines or tissues infected by virus. Thus, we used viral circRNA and human miRNAs target prediction tools to analyze the interaction between viral circRNA and human miRNAs and explored the biological characteristics of viral circRNA (Figure 4D). Viral circRNAs are expected to interact with 518 human miRNAs. We found that chi-miR-103-5p_R-7, chi-miR-26b-3p, chi-miR-92a-5p and chi-miR-122-R_1 can bind at least four circRNAs. PC_3p-10316_124 of microRNA is shared by circRNA-186 and circRNA-10,457, whereas chi-miR-2335 and circRNA8386 have a unique combination. Some viral circRNAs, such as EBV_circ_Homo_sapiens_1693, EBV_circ_Homo_sapiens_1757 and HHV8_circ_Homo_sapiens_1972, are predicted to interact with more than 260 human miRNAs, and EBV_circ_Homo_sapiens_6528 interacts with 50 miRNAs. Moreover, hsa-miR-3912-3p, hsa-miR-6732-5p, hsa-miR-181a-5p, hsa-miR-23c, hsa-miR-23b-3p and hsa-miR-23a-3p were also found. Human miRNAs are predicted to interact with more than 320 viral circRNAs. In this study, we predicted the interaction target of viral circRNA and miRNA. We can guess that one function of viral circRNAs is that regulate viral cells by affecting cell growth, apoptosis and/or DNA replication. Nevertheless, more experiments are necessitated to prove its function. We believe that with a better understanding of its functions, better methods can be designed to heal viral infections of humans and animals.

Analysis of the functions of viral circRNA

To further explore the potential functions of viral circRNA, we analyzed the GO and KEGG pathways using all hypothetical

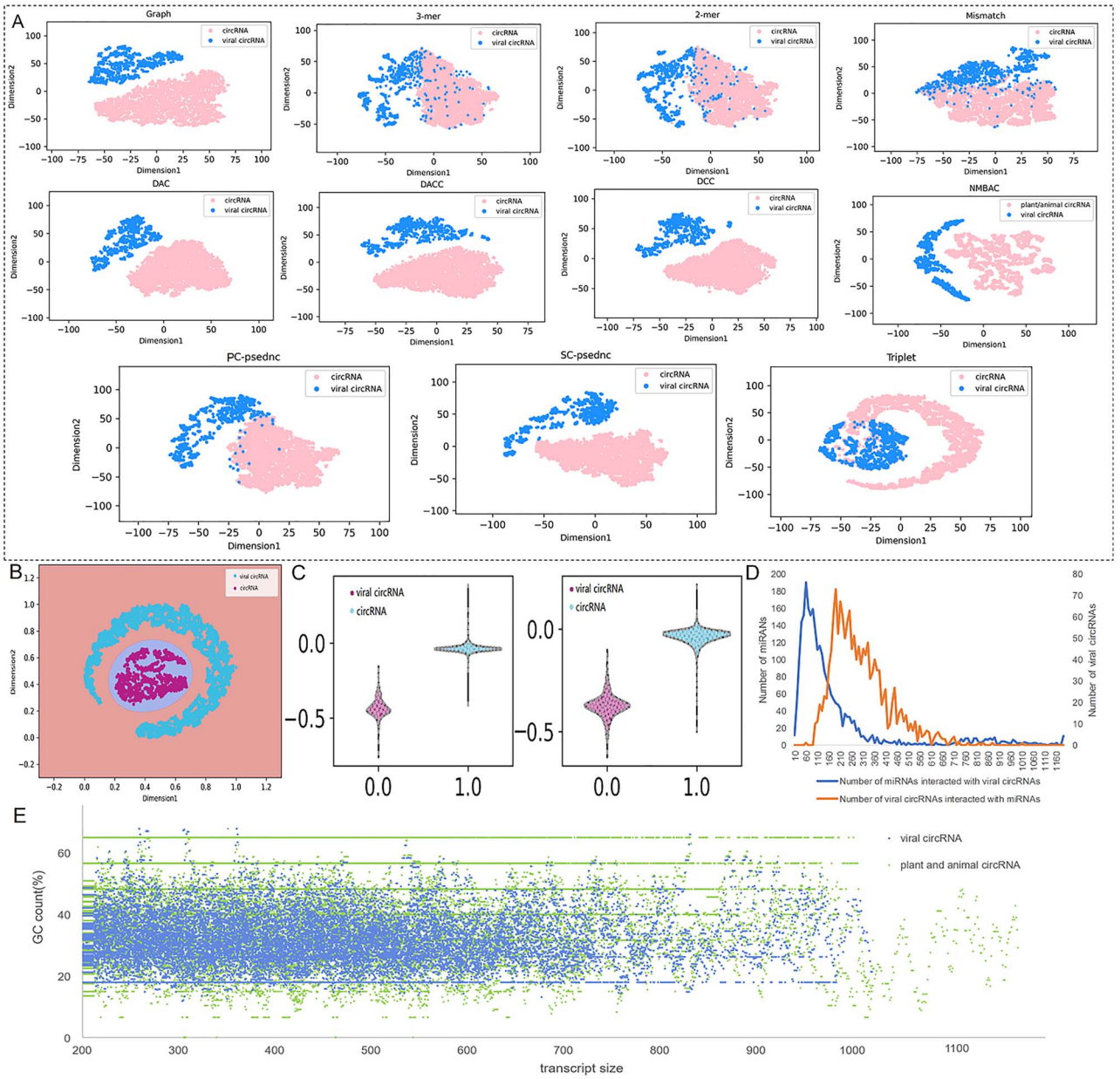

Figure 4. **A**. Diagram of the two-dimensional distribution of each sequence feature. **B**. Diagram of distribution after T-SNE feature selection. **C**. Diagram of the violin with important features. **D**. The interactions between viral circRNAs and human miRNAs. **E**. GC content and transcript size comparison on between viral circRNAs and animal and plant circRNAs.

target genes of miRNA. Figure 6 shows the top 10 enrichment of GO and KEGG pathways analysis.

In terms of biological process (BP), the top 10 are enriched in protein-mediated transport, exogenous apoptosis, nervous system development, cell regulation, virus cycle, etc.; in terms of cell composition (CC), it is mainly in cell connections, nerve growth cones and transcription factors, etc. On molecular function (MF), it mainly includes RNA polymerase binding, DNA binding protein transcription activity, calmodulin-dependent protein kinase activity, cadherin binding, chromatin binding, etc. It can be found that 5 of them are related to binding, and 4 Related to kinase activity. The top 10 pathways obtained by KEGG pathway analysis are: RNA transport, viral infection pathway, tuberculosis pathway, MAPK signaling pathway, neurotrophic factor signaling pathway, dopamine synapse, etc. It can be found that these pathways are widely involved in cell growth, differentiation, stress, inflammation and other physiological and pathological effects, and it can also indicate that viral circRNA is likely to participate in a variety of important physiological/pathological effects such as the nervous system and cancer.

Web server

As a bioinformatics analysis about viral circRNAs, we developed an online web server. Web is built on eclipse, and developed in JAVA language, and extensively tested using several commonly used web browsers. Through the web server, the data used in

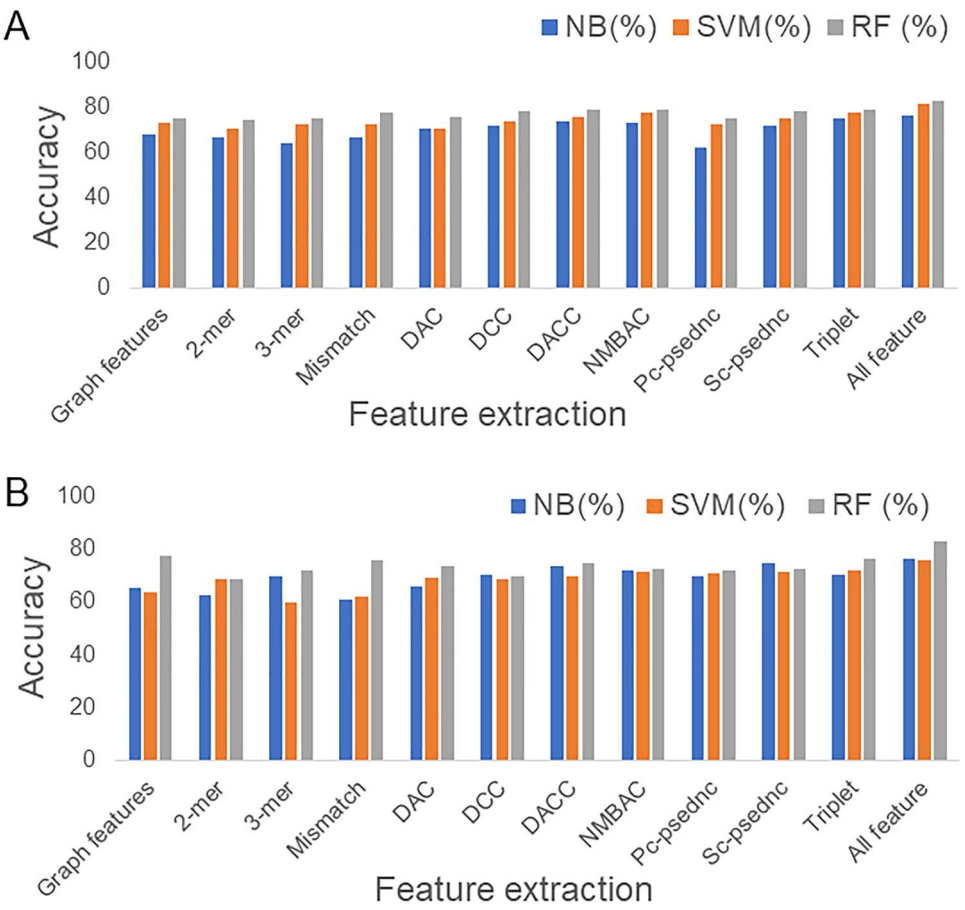

Figure 5. **A**. The result of using different feature representation methods on cross-validation. **B**. The result of using different feature representation methods on external validation.

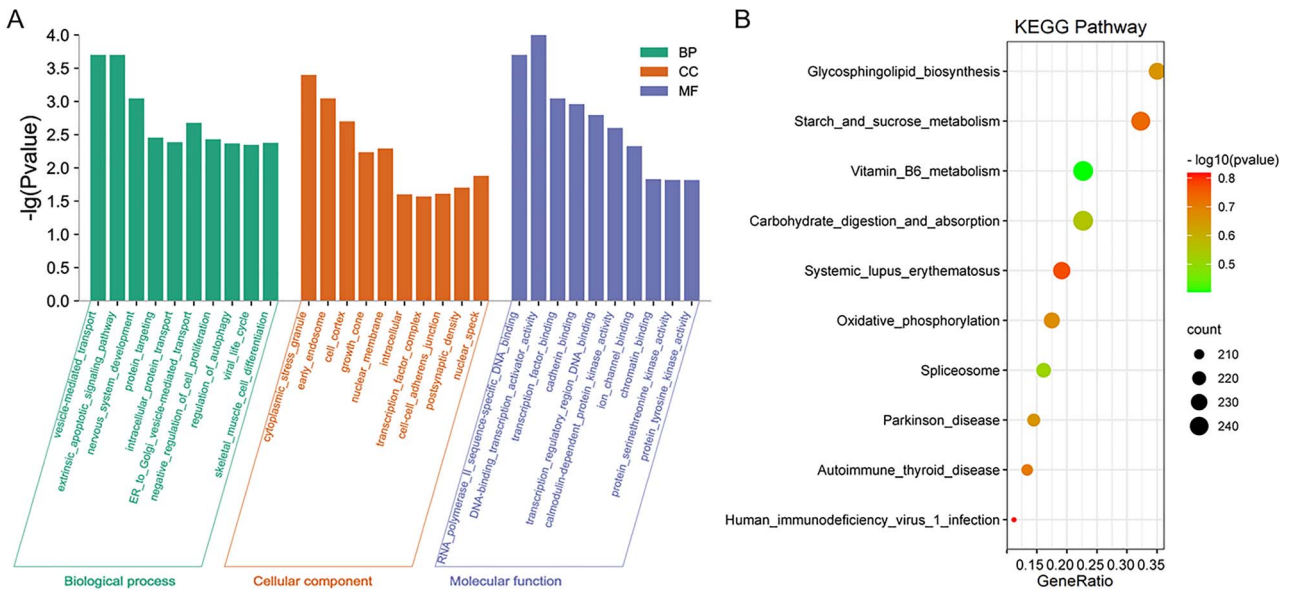

Figure 6. GO and KEGG analysis of viral circRNAs. **A**. Top ten enrichment GOs. **B**. Top ten enrichment pathways.

this study can be accessed online; the viral circRNA can be predicted online, and the supplementary intermediate data of the experiment can be browsed. The web server access URL is: http://server.malab.cn/viral-CircRNA/. Among them, when predicted viral circRNA, users can provide viral circRNA sequences, and web server performs feature extraction. Then based on the trained model, it gives the probability value of the predicted viral circRNA.

Discussion

As more and more viral circRNAs are found to play an important role in single-stranded and double-stranded viruses, we attempted to perform mechanical analyses on viral circRNAs. The evolution of circRNA from one species to another in the same biological kingdom is highly conserved. In animals, circRNAs are highly conserved [21]. In plants, some circRNAs are conserved in highly flowering plants (such as Arabidopsis rice) [18, 20]. Compared with animal and/or plant circRNAs, viral circRNAs are rarely conserved. Among the viral circRNAs currently stored in the viral circRNA database, most have recently been reported. Most viral circRNAs come from two related viruses, EBV and HSV. Viral circRNAs may evolve faster than circRNAs in plant/animals. As for circRNAs in plant/animals, after evolution, they were almost identical in mature form. For viral circRNAs, it may be tough to detect homologous genes in distantly related species. Although some homologous genes have been found in related species, such as EBV and HSV, there is no obvious homology between them.

Then, we analyzed the characteristics of viral circRNAs and animal circRNAs and analyzed the differences in characteristics. Viral circRNAs and animal circRNAs have relatively subtle distinctions in sequence composition and obvious differences in structural characteristics and autocorrelation characteristics. Moreover, excellent feature descriptors provide a basis for better prediction of viral circRNAs, and the results also indicate that viruses may change the structure of circRNAs when encoding them, which is also the structure of viral circRNAs. The results of this research provide direction for future studies. Second, some targets have been predicted through the interactions of viral circRNA and miRNA. And it has been also found that viral circRNAs may be involved in many KEGG pathways related to nervous system and cancer. Although the development of circRNA based on RNA-seq data has achieved good applications [39–42], the application of MLAs based on sequence feature information to identify new viral circRNA is also one of the current research hotspots in bioinformatics [19, 36, 43, 44]. Moreover, machine learning analysis of viral circRNA provides new ideas for similar machine learning analyses. To a certain extent, the developed prediction model can help biologists improve the efficiency of predicting viral circRNA. In future work, we will also improve prediction algorithms to increase prediction performance and use more data for verification. At the same time, regarding questions such as whether all viruses can encode circRNA, we will also collect more data from the GEO database and cooperate with relevant research laboratories to explore and solve more practical problems.

Materials and methods

Dataset

The viral circRNA sequence data were downloaded from the VirusCircBase database [35]. The VirusCircBase database was built by Cai *et al*. and included the circRNA sequences of 23 viruses [the first version of the database (2019-08-23)]. All viral circRNA sequences were identified using circRNA_finder [45], find_circ [8] and CIRI [46]. Some studies have shown that circRNAs with a sequence length of less than 200 bp cannot show RNase R resistance and will cause be false positives [8]. Moreover, the lengths of most human circRNAs range from 200 bp to 1 kbp [47]. Therefore, sequences with lengths $\leq$200 bp and $\geq$1 kbp were removed, and then a total of 1592 high-confidence sequences were obtained. Among the 1592 sequences, the virus types included were Zika virus, EBV, rLCV, KSHV, mouse gamma herpesvirus, herpes simplex virus and monkey virus 40, influenza A virus, and Zaire Ebola virus. The original sequence data were obtained from NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) and Sequence Read Archive, so we have added all relevant Accession number of virus reference genome. Moreover, it can be downloaded from web server and github.

The circRNA sequence data came from the circBase database [7], which is a collection of datasets released by multiple laboratories, all of which have been verified by experiments. We downloaded 2082 circRNA transcripts from circBase. For circRNAs, we deleted transcripts smaller than 200 bp and overlapping circRNA transcripts and finally generated a set of 1884 circRNAs, which are also used in Niu *et al*. [22] and Pan *et al*. [48].

As described in the steps for obtaining viral circRNAs, we can also obtain sequences that are predicted by three software as non-viral circRNAs. Then, using the same processing steps as the viral circRNA sequence, a total of 4080 sequences were obtained. To use machine learning to predict viral circRNAs, we used viral circRNAs sequences as positive examples and non-viral circRNAs as negative examples. Because of the potential redundancy in the viral circRNA dataset, we use CD-HIT to eliminate redundancy. CD-HIT is a program based on sequence similarity clustering to remove redundant sequences, which is widely used in large-scale biological sequence data clustering [22]. According to the empirical value, the similarity thresholds of positive and negative examples are set as 0.8 and 0.4, respectively, and the number of negative and positive samples were 1359 and 2772, respectively.

Analysis of conservation of viral circRNA

First, we used the Clustal X [49] method to do multiple sequence alignment on the viral circRNAs sequences, and the sequences with particularly poor gap conservation were deleted. Then, DNAMAN software was used to perform multiple sequence alignment. Based on the results of the alignment, we used MEGA 5.0 to draw a phylogenetic tree using the neighbor-joining algorithm, and the test parameter (bootstrap) was set to 1000. The MEME online tool [50], Homer [25] and STREME [26] was used to analyze conserved domains of the obtained viral circRNA sequences, and the parameters were set as follows: motif length was limited to 6–100, and the total number of recognized motifs was limited to 7.

Analysis of sequences characteristics of viral circRNA

As shown in Figure 4E, simple characteristics such as GC content and transcript size cannot clearly distinguish viral circRNA from animal and plant circRNA. To achieve a more elaborate analysis, we extracted comprehensive features from the sequence, including Graph characteristic [51], Nucleotide composition

characteristic, autocorrelation characteristic, Pseudo ribonucleic acid composition and structural features.

Graph characteristic

Graph is a method to encode information about viral circRNA sequence and structure in a natural way. Graph encoded the RNA sequence and its folded structure in the form of a graph, where nodes represent nucleotides, and edges represent the relationship between backbones or bonds. In addition, in order to model the high-level features of the structure, we have added an extra layer of secondary structure annotations, which summarizes specific nucleotide information and describes the general shape of the substructure, similar to the shape abstraction in RNAshapes [17], such as stem (S), multilayer ring (M), hairpin (H), inner ring (I), protrusion (B) and outer area (E). A detailed introduction can be obtained from https://github.com/dmaticzka/GraphProt.

Nucleotide composition characteristic

This feature is composed of two parts: (i) k-mer and (ii) Mismatch. (i) K-mer expresses RNA sequences by counting the frequency of k adjacent nucleic acids. We used $k = 2,3$. For example, when $k = 2$, we obtain the frequency features for four nucleotides that are adjacent to each other, such as 'AA', 'AC'. (ii) For a sequence of k length, Mismatch counts the occurrences of adjacent nucleic acids that differ by at most m mismatches. For example, there is a sequence of 3 lengths 'CCA', and max 1 mismatch, so there are three cases: '-CA','C-A' and 'CC-' ('-' can be replaced by any nucleic acid).

Autocorrelation characteristic

This feature is composed of four parts: (i) DAC, (ii) DCC, (iii) DACC and (iv) NMBAC. They are all based on the physicochemical attribute matrix of nucleotides to extract features. (i) DAC calculates the correlation between two dinucleotides separated by λ along the sequence of the same physicochemical index. (ii) DCC calculates the correlation of two dinucleotides separated by λ distance under different physicochemical properties. (iii) DACC combines the DAC and DCC methods. (iv) NMBAC calculates the correlation between two nucleotides separated by a distance under the same physical and chemical properties. The calculation formula is as Equations (1–3).

$$\mathrm{DAC}(u,\gamma) = \sum_{i=1}^{L-\gamma} \left(P_u(D_i) - \overline{P_u}\right)\left(P_u(D_{i+\gamma}) - \overline{P_u}\right) / (L-\gamma) \quad (1)$$

$$\mathrm{DCC}(u_1,u_2,\gamma) = \sum_{i=1}^{L-\gamma} \left(P_{u_1}(D_i) - \overline{P_{u_1}}\right)\left(P_{u_2}(D_{i+\gamma}) - \overline{P_{u_2}}\right) / (L) \quad (2)$$

$$\mathrm{NMBAC}(u,\gamma) = \sum_{i=1}^{L-\gamma} \left(P_u(x_i) \times P_u(x_{i+\gamma})\right)^2 \quad (3)$$

$$\overline{P_u} = \sum_{i=1}^{L-1} \left(P_u(D_i)/(L-1)\right. \quad (4)$$

where u indicates the physicochemical properties index; L is the length of the viral circRNA sequence;$D_i \in \{AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, UU\}$, $1 \le \gamma < L$, $\overline{P_u}$ represents the mean value of the ith row of the above physical and chemical attribute matrix.

Pseudo ribonucleic acid composition

This feature uses sequence local sequence information and sequence global sequence information to represent RNA sequences, which is composed of two parts: (i) General parallel correlation pseudo dinucleotide composition (PC-PseDNC), (ii) general series correlation PC-PseDNC (SC-PseDNC). The PC-PseDNC method considers the parallel correlation between two dinucleotides under certain physical and chemical properties. SC-PseDNC method considers the continuous correlation of two dinucleotides under certain physical and chemical properties.

Structural features

Local structure-sequence triplet element (Triplet) express RNA sequence by counting the status of secondary structure. Triplet is an early method that uses structural information from RNA sequences and can calculate secondary structure characteristics. Triplet calculates its secondary structure through the Vienna RNA software package (version 2.1.6), and each nucleotide has two states (paired or unpaired), which are represented by brackets '("or")' and dots '.', respectively.

Identification of viral circRNA based on sequence features

Just as we used three tools developed based on RNA-Seq data (circRNA_finder, find_circ and CIRI) when processing the dataset. However, on the other hand, many pieces of research on RNA recognition and site detection are based on machine learning. Moreover, with the increase of circRNA sequences, there are more and more researches using machine learning methods to predict circRNA [19, 22, 36, 43, 48]. Based on the characteristics of sequence data, the use of MLAs to develop viral circRNA prediction models is another important direction for predicting circRNA. For another, viral circRNA sequences are gradually increasing. Therefore, how to use the characteristics of viral circRNA sequences to identify viral circRNA more accurately is a problem that needs to be studied. First, we constructed a new dataset, including the sequence data of viral circRNA (positive example) and circRNA (negative example). Then used MLAs (NB, SVM and RF) to analyze the effectiveness of the sequence feature descriptors of the viral circRNA analyzed above and predictive performance. RF is an important bagging-based ensemble learning method, which can be used for classification, regression and other problems. Its composition is composed of multiple weak learners, and it is relatively simple to implement. The parameters in the SVM, RF and NB algorithms directly use the default parameters, and the programming environment used is python3.7.4. We performed 10-fold cross-validation and independent test set validation and used accuracy as an evaluation indicator.

Analysis of the function between viral circRNA and miRNA

TargetScan 7.0 [52], miRanda [53] and CCmiR [54] software programs were used to predict the circRNA-miRNA interaction, and the intersection was performed to obtain a more accurate target gene. TargetScan 7.0 is a method for predicting miRNAs targets based on the homology of seed regions. MiRanda is a method of miRNA target prediction based mainly on the free energy combination of miRNA and its target genes. Moreover, the binding strength of the target gene is inversely proportional to the free energy. The lower the free energy, the stronger the binding. CCmiR is a method based on a hidden Markov model, a statistical method that models the observations generated from a hidden Markov chain. CCmiR has an optional file input, miRNA

expression information, which considers the expression level of miRNAs and the competition and cooperation of multiple miRNAs. The percentile of the TargetScan score was less than 50, and the maximum free energy value of miRanda was less than 10, and a Miranda score greater than 140 was defined as the cutoff point for target prediction. Target prediction takes the intersection of the three results.

Functional enrichment analysis of viral circRNAs

Functional enrichment of viral circRNA source genes or potential target genes to determine its potential biological functions is an important method for studying circRNA. Therefore, based on the 1592 high-confidence circRNA from the virus, we used The Database for Annotation, Visualization and Integrated Discovery (DAVID) [55] to perform GO and KEGG enrichment analysis of miRNA target genes to further understand the potential functions of circRNA. DAVID is a biological information database that integrates biological data and analysis tools to provide systematic and comprehensive biological function annotation information for large-scale gene or protein lists. GO analysis evaluates from three aspects: BP, MF and CC. We also analyzed related pathways through KEGG.

Key Points

- We built the first recognition model for viral circRNAs using MLAs.
- Viral circRNAs and animal circRNAs have relatively subtle distinctions in sequence composition and obvious differences in structural characteristics and autocorrelation characteristics.
- Viral circRNAs may be involved in many KEGG pathways related to nervous system and cancer.

Authors' contributions

Zou Q and Lin C designed the research; Niu MT carried out the studies, participated in the sequence alignment and drafted the manuscript. Ju Y participated in the design of the study and performed the statistical analysis. Zou Q and Lin C conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Data and code availability

https://github.com/nmt315320/virus-circRNA.git.

Funding

The work was supported by the National Natural Science Foundation of China (No.61922020, No.61771331), and the Special Science Foundation of Quzhou (2020D003).

References

1. Kos A, Dijkema R, Arnberg A, *et al*. The hepatitis delta (δ) virus possesses a circular RNA. *Nature* 1986;**323**(6088):558–60.
2. Chen T-C, Tallo-Parra M, Cao QM, *et al*. Host-derived circular RNAs display proviral activities in Hepatitis C virus-infected cells. *PLoS Pathog* 2020;**16**(8):e1008346.
3. Sanger HL, Klotz G, Riesner D, *et al*. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. *Proc Natl Acad Sci* 1976;**73**(11):3852–6.
4. Hsu M-T, Coca-Prados M. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. *Nature* 1979;**280**(5720):339–40.
5. Arnberg A, Van Ommen G-J, Grivell L, *et al*. Some yeast mitochondrial RNAs are circular. *Cell* 1980;**19**(2):313–9.
6. Nigro JM, Cho KR, Fearon ER, *et al*. Scrambled exons. *Cell* 1991;**64**(3):607–13.
7. Glažar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. *RNA* 2014;**20**(11):1666–70.
8. Hansen TB, Venø MT, Damgaard CK, *et al*. Comparison of circular RNA prediction tools. *Nucleic Acids Res* 2016;**44**(6):e58.
9. Szabo L, Salzman J. Detecting circular RNAs: bioinformatic and experimental challenges. *Nat Rev Genet* 2016;**17**(11):679.
10. J-T H, J-n C, L-P G, *et al*. Identification of virus-encoded circular RNA. *Virology* 2019;**529**:144–51.
11. Hansen TB, Jensen TI, Clausen BH, *et al*. Natural RNA circles function as efficient microRNA sponges. *Nature* 2013;**495**(7441):384–8.
12. Chen L-L. The biogenesis and emerging roles of circular RNAs. *Nat Rev Mol Cell Biol* 2016;**17**(4):205–11.
13. Wang PL, Bao Y, Yee M-C, *et al*. Circular RNA is expressed across the eukaryotic tree of life. *PLoS One* 2014;**9**(3):e90859.
14. Kelly S, Greenman C, Cook PR, *et al*. Exon skipping is correlated with exon circularization. *J Mol Biol* 2015;**427**(15):2414–7.
15. Broadbent KM, Broadbent JC, Ribacke U, *et al*. Strand-specific RNA sequencing in plasmodium falciparum malaria identifies developmentally regulated long non-coding RNA and circular RNA. *BMC Genomics* 2015;**16**(1):1–22.
16. Lu T, Cui L, Zhou Y, *et al*. Transcriptome-wide investigation of circular RNAs in rice. *RNA* 2015;**21**(12):2076–87.
17. Steffen P, Voß B, Rehmsmeier M, *et al*. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* 2006;**22**(4):500–3.
18. Wang Z, Liu Y, Li D, *et al*. Identification of circular RNAs in kiwifruit and their species-specific response to bacterial canker pathogen invasion. *Front Plant Sci* 2017;**8**:413.
19. Chen L, Zhang Y-H, Huang G, *et al*. Discriminating cirRNAs from other lncRNAs using a hierarchical extreme learning machine (H-ELM) algorithm with feature selection. *Mol Genet Genomics* 2018;**293**(1):137–49.
20. Ye CY, Chen L, Liu C, *et al*. Widespread noncoding circular RNA s in plants. *New Phytol* 2015;**208**(1):88–95.
21. Kristensen LS, Andersen MS, Stagsted LV, *et al*. The biogenesis, biology and characterization of circular RNAs. *Nat Rev Genet* 2019;**20**(11):675–91.
22. Niu M, Zhang J, Li Y, *et al*. CirRNAPL: a web server for the identification of circRNA based on extreme learning machine. *Comput Struct Biotechnol J* 2020;**18**:834–42.
23. Liu Q, Shuai M, Xia Y. Knockdown of EBV-encoded circRNA circRPMS1 suppresses nasopharyngeal carcinoma cell proliferation and metastasis through sponging multiple miRNAs. *Cancer management and research* 2019;**11**:8023.
24. Lp G, Jn C, Dong M, *et al*. Epstein–Barr virus-derived circular RNA LMP 2A induces stemness in EBV-associated gastric cancer. *EMBO Rep* 2020;**21**(10):e49689.
25. Heinz S, Benner C, Spann N, *et al*. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;**38**(4):576–89.

26. Bailey TL, Elkan C. *Fitting a mixture model by expectation maximization to discover motifs in bipolymers. International Conference on Intelligent Systems for Molecular Biolog*, 1994;2:28–36.
27. Abere B, Li J, Zhou H, *et al*. Kaposi's sarcoma-associated herpesvirus-encoded circRNAs are expressed in infected tumor tissues and are incorporated into virions. *MBio* 2020;**11**(1):e03027–19.
28. Berman TA, Schiller JT. Human papillomavirus in cervical cancer and oropharyngeal cancer: one cause, two diseases. *Cancer* 2017;**123**(12):2219–29.
29. Torresi J, Tran BM, Christiansen D, *et al*. HBV-related hepatocarcinogenesis: the role of signalling pathways and innovative ex vivo research models. *BMC Cancer* 2019;**19**(1):1–14.
30. Wu F, Cheng W, Zhao F, *et al*. Association of N6-methyladenosine with viruses and related diseases. *Virol J* 2019;**16**(1):1–10.
31. Ungerleider N, Concha M, Lin Z, *et al*. The Epstein Barr virus circRNAome. *PLoS Pathog* 2018;**14**(8):e1007206.
32. Chen X, Yang T, Wang W, *et al*. Circular RNAs in immune responses and immune diseases. *Theranostics* 2019;**9**(2): 588.
33. Zhu K, Zhan H, Peng Y, *et al*. Plasma hsa_circ_0027089 is a diagnostic biomarker for hepatitis B virus-related hepatocellular carcinoma. *Carcinogenesis* 2020;**41**(3):296–302.
34. Ghorbani A, Izadpanah K, Peters JR, *et al*. Detection and profiling of circular RNAs in uninfected and maize Iranian mosaic virus-infected maize. *Plant Sci* 2018;**274**:402–9.
35. Cai Z, Fan Y, Zhang Z, *et al*. VirusCircBase: a database of virus circular RNAs. *Brie Bioinform*. 2020;**22**(2):2182–90.
36. Stricker M, Asim MN, Dengel A, *et al*. CircNet: an encoder-decoder-based convolution neural network (CNN) for circular RNA identification. *Neural Comput Appl* 2021;1–12. 10.1007/s00521-020-05673-1.
37. McInnes L, Healy J, Melville J. *Umap: Uniform manifold approximation and projection for dimension reduction*. arXiv preprint arXiv:180203426. 2018.
38. Shen Z, Zou Q. Basic polar and hydrophobic properties are the main characteristics that affect the binding of transcription factors to methylation sites. *Bioinformatics* 2020;**36**(15):4263–8.
39. Gao Y, Zhang J, Zhao F. Circular RNA identification based on multiple seed matching. *Brief Bioinform* 2018;**19**(5): 803–10.
40. Memczak S, Jens M, Elefsinioti A, *et al*. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 2013;**495**(7441):333–8.
41. Westholm JO, Miura P, Olson S, *et al*. Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep* 2014;**9**(5):1966–80.
42. Zhang X-O, Dong R, Zhang Y, *et al*. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res* 2016;**26**(9):1277–87.
43. Chaabane M, Williams RM, Stephens AT, *et al*. circDeep: deep learning approach for circular RNA classification from other long non-coding RNA. *Bioinformatics* 2020;**36**(1):73–80.
44. Zhang G, Deng Y, Liu Q, *et al*. Identifying circular rna and predicting its regulatory interactions by machine learning. *Front Genet* 2020;**11**:655.
45. Fu X, Liu R (eds). CircRNAFinder: a tool for identifying circular RNAs using RNA-Seq data. In: *Proceedings of the 6th International Conference on Bioinformatics and Computational Biology*. BICOB, 2014.
46. Gao Y, Wang J, Zhao F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol* 2015;**16**(1):1–16.
47. Sebastian M, Panagiotis P, Oliver P, *et al*. Identification and characterization of circular RNAs as a new class of putative biomarkers in human blood. *Plos One* 2015;**10**(10):e0141214.
48. Pan X, Xiong K. Predcirc RNA: computational classification of circular RNA from other long non-coding RNA using hybrid features. *Mol Biosyst* 2015;**11**(8):2219–26.
49. Thompson JD, Gibson TJ, Frédéric P, *et al*. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 1997;**24**:4876–82.
50. Timothy Bailey CE (ed). In: *Fitting a mixture model by expectation maximization to discover motifs in biopolymers*, 1994.
51. Maticzka D, Lange SJ, Costa F, *et al*. Graph Prot: modeling binding preferences of RNA-binding proteins. *Genome Biol* 2014;**15**(1):R17.
52. Vikram A, Bell GW, Jin-Wu N, *et al*. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 2015;**4**:e05005.
53. Friedman RC, Farh Kyle K-H, *et al*. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 2009;**19**(1):8.
54. Ding J, Li X, Hu H. CCmiR: a computational approach for competitive and cooperative microRNA binding prediction. *Bioinformatics* 2018;**2**:2.
55. Dennis G, Sherman BT, Hosack DA, *et al*. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 2003;**4**(9):1–11.