ARTICLES
https://doi.org/10.1038/s41587-021-00842-6
nature biotechnology

Check for updates

# Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long

Jinyang Zhang[1,2,6], Lingling Hou[1,6], Zhenqiang Zuo[1,6], Peifeng Ji[1], Xiaorong Zhang[3], Yuanchao Xue[3] and Fangqing Zhao[1,2,4,5] ✉

**Reconstructing the sequence of circular RNAs (circRNAs) from short RNA sequencing reads has proved challenging given the similarity of circRNAs and their corresponding linear messenger RNAs. Previous sequencing methods were unable to achieve high-throughput detection of full-length circRNAs. Here we describe a protocol for enrichment and full-length sequencing of circRNA isoforms using nanopore technology. Circular reverse transcription and size selection achieves a 20-fold higher enrichment of circRNAs from total RNA compared to previous methods. We developed an algorithm, called circRNA identifier using long-read sequencing data (CIRI-long), to reconstruct the sequence of circRNAs. The workflow was validated with simulated data and by comparison to Illumina sequencing as well as quantitative real-time RT–PCR. We used CIRI-long to analyze adult mouse brain samples and systematically profile circRNAs, including mitochondria-derived and transcriptional read-through circRNAs. We identified a new type of intronic self-ligated circRNA that exhibits special splicing and expression patterns. Our method takes advantage of nanopore long reads and enables unbiased reconstruction of full-length circRNA sequences.**

CircRNAs, a large class of RNAs with covalent circular structure, are involved in the regulation of many biological processes[1,2]. Most circRNAs investigated to date were proposed to act as microRNA (miRNA) sponges[3–5], whereas others were suggested to act as RNA-binding protein (RBP) sponges[6], to enhance protein function[7,8], to encode peptides[9], to form RNA duplex structures[10] and to regulate transcriptional pausing by binding to their host gene locus[11]. A small proportion of the millions of circRNAs profiled in eukaryotic organisms are derived from the mitochondrial genome and have been associated with the progression and prognosis of diseases[12–14]. The emerging roles of circRNAs indicate the importance of sequencing these circular transcripts.

Several computational methods were recently developed to sequence circRNAs[15–18]. Most existing tools rely on the alignment of short Illumina RNA sequencing (RNA-seq) reads, and their detection ability is considerably limited by the relatively short length of Illumina sequencing reads. Considering that most circRNAs are derived from exonic regions[18], these alignment-based methods are unable to distinguish circular reads from the overlapping regions of corresponding linear transcripts. Long-read sequencing technologies, including PacBio and Oxford Nanopore, have provided insights into transcriptome complexity and allow ultra-long sequencing of complementary DNA (cDNA) molecules, enabling easy reconstruction of transcript isoforms[19]. However, in most studies, cDNA was sequenced using oligo(dT) primers, which is not suitable for circRNA detection because circRNAs do not contain the poly(A) sequence. Thus, only a few attempts have been made to apply long-read sequencing technology to circRNA studies. In a recent study, PacBio sequencing was used to identify full-length sequences of circRNAs from reverse transcription polymerase chain reaction (RT–PCR) products[20]. However, the specific PCR primers were designed to target a subset of candidate circRNAs and were able to detect only full-length sequences of selected circRNAs one at a time. Another study used nanopore sequencing with fragmented circRNAs and identified circRNAs using a BLAT-based pipeline[21]. However, the accuracy and sensitivity of these methods remain unexamined, and the methods were unable to achieve high-throughput, direct sequencing of circRNAs.

Here we present an experimental and computational method (CIRI-long) for extensive profiling of full-length circRNAs using nanopore sequencing technology. We demonstrate that our method has considerable advantages in efficiency and reliability for circRNA detection and reconstruction over currently available approaches and provides novel insights into the diversity and biogenesis of circRNAs.

## Results

**A modified nanopore sequencing protocol for efficient identification of circRNAs.** To efficiently enrich circRNAs, a customized approach for RNA-seq library preparation was modified (Fig. 1a). In brief, ribosomal RNA (rRNA) was removed from extracted total RNA using a RiboZero kit. Then additional poly(A)-tailing treatment was applied before RNase R digestion to increase the efficiency of linear RNA degradation according to a previous study[22]. Afterwards, reverse transcription was performed with random primers and SMARTer reverse transcriptase to amplify circRNAs by producing long cDNA molecules containing multiple copies of full-length circRNA sequences. In this step, SMARTer sequencing adapters were added to both ends of cDNA molecules to enable effective amplification of these long cDNAs without poly(A) tails. Finally, nanopore sequencing libraries were constructed after fragment size selection and sequenced using the MinION platform.

[1]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]Key Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China. [4]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China. [5]Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou, China. [6]These authors contributed equally: Jinyang Zhang, Lingling Hou, Zhenqiang Zuo. ✉e-mail: zhfq@biols.ac.cn

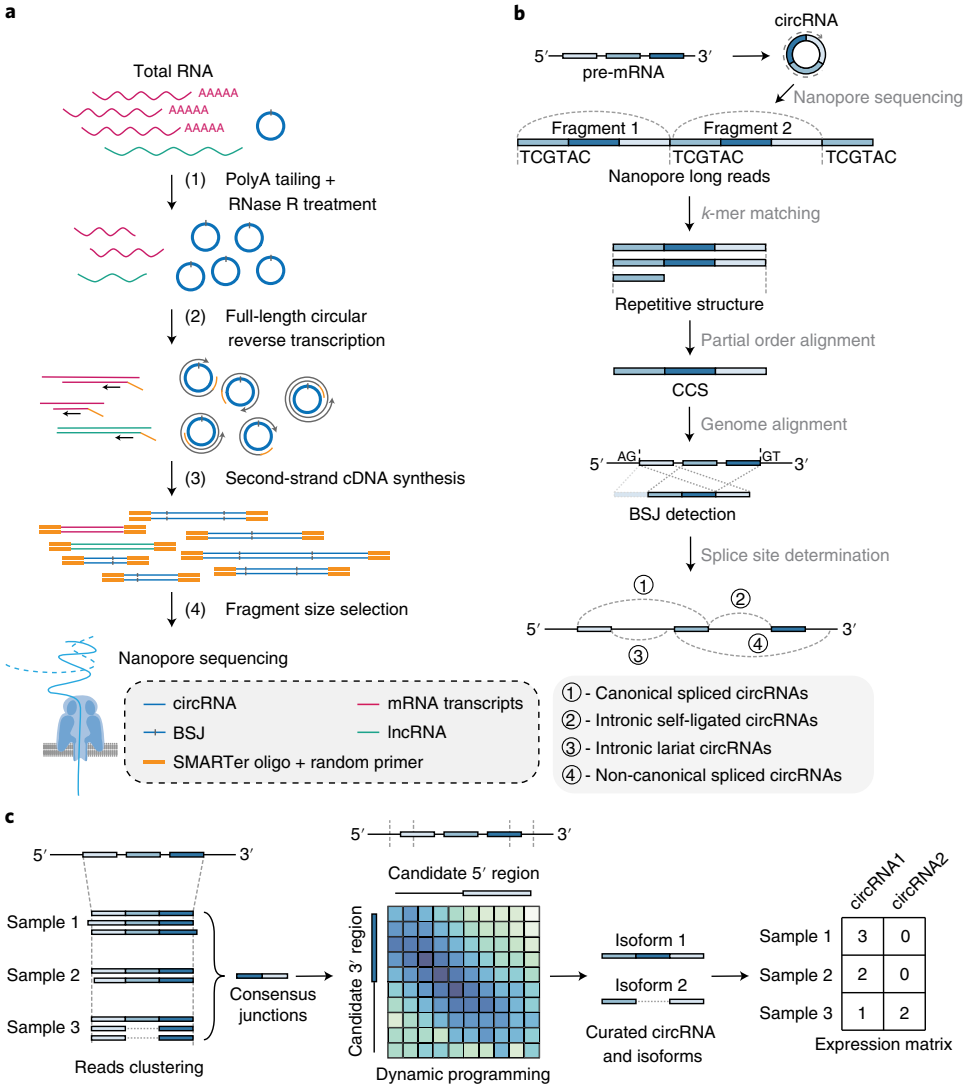


**Fig. 1 | Method overview. a**, Construction of nanopore sequencing libraries. Total RNA was extracted and subjected to poly(A) tailing and RNase R treatment to degrade linear RNAs. The remaining RNAs were amplified using rolling circular reverse transcription, and the 1-kb size was selected with magnetic beads. The cDNA libraries were sequenced using the MinION platform. **b**, Workflow of circRNA identification. The raw nanopore reads are split into repetitive fragments using the occurrence of identical *k*-mers, and a CCS for each read is constructed using partial order alignment. The CCS sequences are aligned using Python interfaces of minimap2 and bwa, and BSJs are identified from the alignment results. The type of circRNAs is determined based on the coordinates of the BSJ and annotation of the reference genome. **c**, Aggregation of the results from multiple samples. Candidate circRNAs from multiple samples are collapsed together and clustered based on the alignment positions. The consensus junction sequence of each cluster is constructed, and the corrected BSJ sites are calculated using dynamic programming. All candidate sequences are assigned to curated circRNA isoforms, and the expression matrices of high-confidence circRNAs are generated as the final output. lncRNA, long non-coding RNA.

To obtain clean sequences that contain only circRNAs, reads were first base-called and de-multiplexed using ont-guppy. Then barcodes and primers were trimmed using Porechop with a modified primer library (Methods). A novel algorithm, CIRI-long, is proposed for circRNA characterization and isoform quantification from the nanopore sequencing data (Fig. 1b). Initially, a set of *k*-mers was used to search for repetitive patterns and detect boundaries of circRNA fragments in the nanopore reads (Supplementary Fig. 1). Next, partial order alignment[23] was performed to generate a cyclic consensus sequence (CCS) for each nanopore read using the SPOA library[24], and a threshold of 80% similarity between repetitive segments and CCS was used to screen high-confidence candidates (Supplementary Fig. 2). Afterwards, CCSs were aligned to the reference genome for detection of back-spliced junction (BSJ) patterns, and annotated splice sites and canonical splice signals were used to determine the boundaries and types of circRNAs.

To further improve the accuracy of circRNA detection and quantification, all candidate circRNA sequences from the nanopore reads were aggregated together in the next step (Fig. 1c). These sequences were clustered based on their alignment positions, where each cluster represents a putative circRNA locus. Then a consensus sequence for each cluster was generated, and the precise location of the BSJ site was determined using dynamic programming. Finally, after filtering the clusters with fewer supporting reads or ambiguous splice signals, the expression matrices of remaining circRNAs were generated as the final output.

**Experimental optimization for capture of full-length circRNAs.** To comprehensively evaluate the effects of various experimental

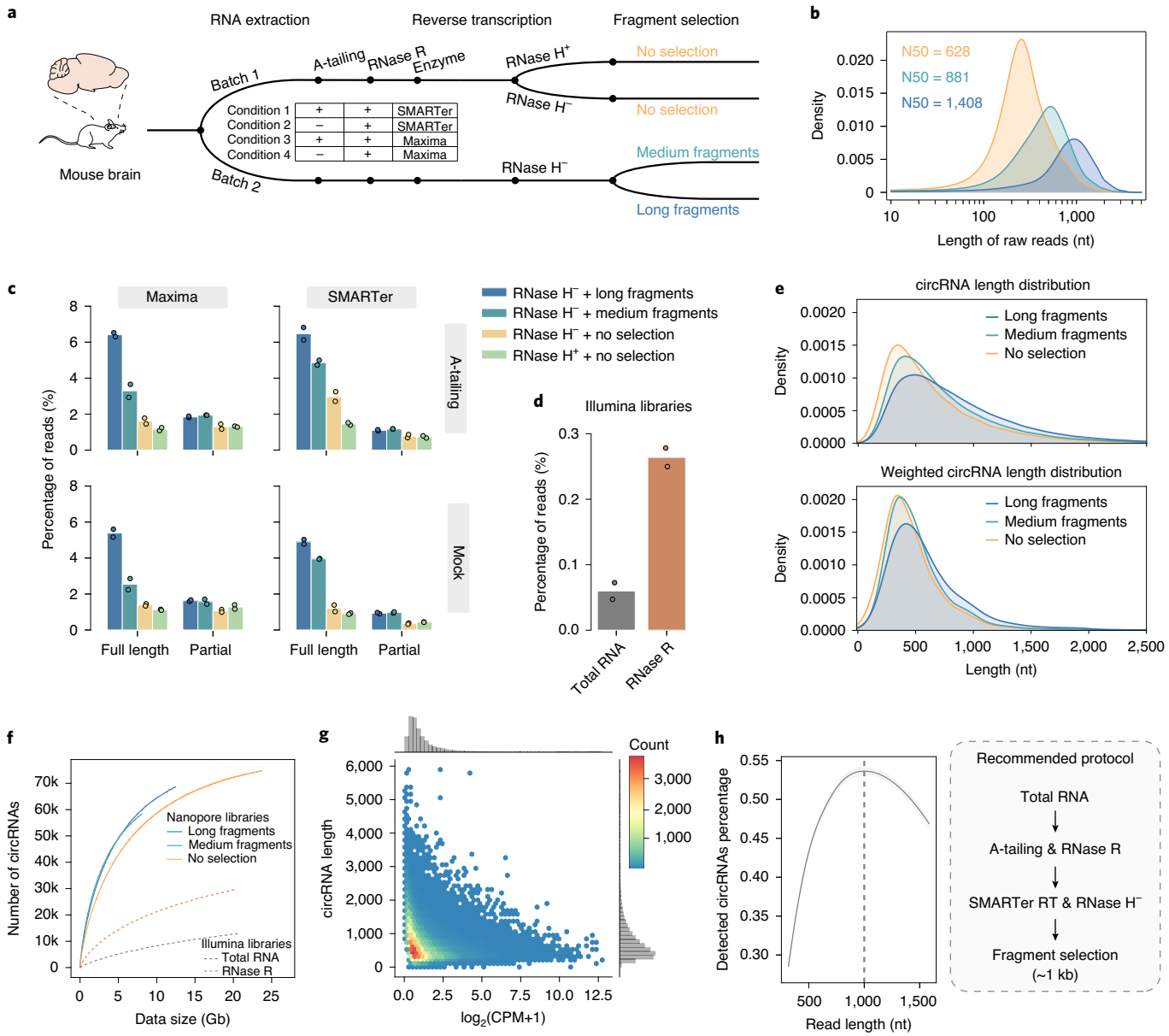


**Fig. 2 | Experimental optimization of capture of full-length circRNAs. a**, Schematic view of optimization of experimental conditions. Total mouse brain RNA was extracted, divided into two batches and treated using a combination of various reverse transcription strategies (RNase R treatment with or without A-tailing, SMARTer or Maxima RT, with or without RNase H treatment). Various fragment sizes (400 bp, 600 bp and 1 kb) of the cDNA products were selected and then sequenced on a MinION platform. **b**, N50 of raw nanopore reads. The probability density function of read length from the libraries with various fragment sizes is shown in yellow (400 bp), green (600 bp) and blue (1 kb). **c**, Percentage of full-length and partial supporting reads under various experimental conditions. Different bars on the *x* axis represent libraries with different experimental treatments ($n = 2$ independent experiments). **d**, The percentage of BSJ reads from mutually supported circRNAs in Illumina RNA-seq data ($n = 2$ independent experiments). CircRNAs were identified from the total RNA and RNase R-treated Illumina RNA-seq libraries using CIRI2, and the number of circular reads was calculated by summing all BSJ reads for each circRNA. **e**, Length distribution of detected circRNAs from libraries with different size selections. Upper, the distribution of circRNA lengths was calculated using Gaussian kernel density estimation. Bottom, the expression level of each circRNA was used as weight in Gaussian kernel density estimation. **f**, Saturation curve of circRNA identification. For each dataset, different numbers of reads were randomly sampled from real sequencing data, and the corresponding numbers of circRNAs and data size were calculated. **g**, The distribution of circRNA length and expression levels. The hexagonal heat map was plotted using the hexbin function of matplotlib[42], and the filled color represents the number of circRNAs in each hexagonal bin. **h**, Efficiency of circRNA detection based on various fragment sizes. Datasets with identical data size (1 Gb) but different read lengths were simulated using all circRNAs detected in the Illumina data, and the *y* axis represents the recall rate of the simulated circRNAs. Right, the recommended protocol for nanopore library construction. Total RNA was subjected to A-tailing and RNase R treatment, reverse transcribed using SMARTer RT without RNase H treatment and size selected to approximately 1 kb. nt, nucleotides.

conditions on our methods, 32 nanopore sequencing libraries were constructed to determine the optimized protocol (Fig. 2a). In brief, total RNA was extracted from two biological replicates of mouse brains, and each replicate was divided into two batches. In each batch, several reverse transcription strategies (with or without A-tailing, SMARTer or Maxima reverse transcription, with or

without RNase H treatment) were used, and the capture efficiency of full-length circRNAs was evaluated.

To verify the effect of RNase H treatment in circRNA enrichment, libraries in Batch 1 were subdivided into two groups, one of which was treated with RNase H during reverse transcription to hydrolyze the template RNA behind the polymerase. Variable size of library selection (600 bp and 1 kb; Fig. 2b) was used in Batch 2, where longer libraries can reduce the proportion of linear undegraded fragments. All libraries were sequenced across three MinION flow cells, generating a total of 73,168,913 nanopore reads with average size ranging from 500 Mb to 3 Gb (Supplementary Table 1). Additionally, one total RNA library and one RNase R library were constructed for each biological replicate and were sequenced using an Illumina sequencer with paired-end 150-bp reads for further comparison and validation.

To rigorously evaluate the performance of each treatment, the percentage of full-length and partially supported circRNA reads was calculated to represent the enrichment coefficient of circRNAs (Fig. 2c and Supplementary Fig. 3). As expected, a significant improvement was observed after fragment size selection. The average fraction of full-length reads was approximately 1% in libraries without size selection. However, there was a three-fold and six-fold increase in libraries with medium and long fragment sizes, respectively, indicating that library size selection is the most relevant factor for capture of full-length circRNAs. The A-tailing treatment also improved the detection sensitivity and increased the efficiency of RNase R digestion of linear transcripts containing G-quadruplex structure, whereas using SMARTer reverse transcriptase induced only small differences in the A-tailing-treated group compared to that detected using Maxima reverse transcriptase (Supplementary Figs. 4 and 5). Comparison of the RNase H group (light green bars) and the other groups indicated that RNase H treatment has a negative influence on circRNA detection. Overall, these results demonstrate that the length of reverse transcription is the major factor influencing the detection efficiency of circRNAs. Compared to 6% of circular reads obtained in nanopore libraries under optimized conditions, the percentage of circular reads identified in Illumina RNA-seq data was only 0.06% and 0.27% in total RNA- and RNase R-treated libraries, respectively (Fig. 2d), in agreement with the data of previous studies[25]. This result suggests that nanopore sequencing is more efficient in circRNA detection than traditional short-read sequencing technologies.

Next, we determined whether size selection in nanopore sequencing library preparation causes a bias in circRNA detection. As shown in Fig. 2e, an increase in the fragment size induced a corresponding shift in the length distribution of detected circRNAs to the right. However, when the expression levels of circRNAs were considered, nanopore libraries with medium fragment size and without size selection had a similar length distribution of identified circRNAs, which is consistent with the characteristics of circRNAs detected in the Illumina sequencing dataset (Supplementary Fig. 6). Nanopore libraries with long fragment sizes preferentially detect longer circRNAs, suggesting that improper selection of fragment size might lead to biased detection of circRNAs.

To explore the optimized length of fragment selection, the saturation curve of circRNA detection was calculated for the nanopore and Illumina RNA-seq libraries. As shown in Fig. 2f, the saturation curve of the nanopore sequencing libraries was significantly higher than that of the Illumina libraries. However, nanopore libraries with long and medium fragments showed a similar performance in detection of circRNAs, indicating that size selection over a certain threshold does not increase detection sensitivity. To explain this result, the correlation between the circRNA length and expression levels was visualized as a hexagonal heat map. As shown in Fig. 2g, most circRNAs are shorter than 1 kb, and longer circRNAs tend to have lower expression levels, indicating that a fragment size of approximately 1 kb might be already sufficient for the detection of most circRNAs. Subsequently, we simulated datasets with the same data size but different fragment lengths ranging from 400 bp to 1,500 bp. As expected, the percentage of detected circRNAs was dramatically decreased concomitant to an increase in fragment sizes over 1 kb, indicating that circRNA detection using fragment selection around 1 kb has the highest efficiency (Fig. 2h). Thus, optimal procedures for circRNA detection using nanopore sequencing technology should include total RNA treatment using A-tailing and RNase R digestion and subsequent reverse transcription with SMARTer RT under RNase $H^-$ conditions. The optimal fragment size selection should be approximately 1 kb to balance the tradeoff between the detection efficiency and sensitivity for full-length circRNAs.

**Validation of circRNAs using simulated and Illumina RNA-seq datasets.** To evaluate the reliability of CIRI-long, NanoSim[26] was used to generate the simulated nanopore sequencing datasets (Methods). Then CIRI-long was used to characterize circRNAs in the simulated data. For read-level analysis, CIRI-long achieved an $F_1$ score of 0.92, indicating high sensitivity and accuracy of circular read detection (Supplementary Table 2). The length of CCSs detected by CIRI-long also highly correlated to the simulated length (Fig. 3a). The predicted and simulated coordinates of each circRNA were evaluated to compare circRNA levels. As shown in Fig. 3b, for the vast majority (96.57%) of identified circRNAs, CIRI-long was able to accurately determine the BSJ site. In addition, a small fraction (1.07%) of reads were observed to be falsely assigned to BSJ sites within a 50-bp distance, which was apparently due to incorrect estimation of splice sites caused by sequencing errors. Considering the relatively high error rate in nanopore sequencing data, we assessed the accuracy of raw sequencing reads and CCSs, respectively. With the increase of read length, the coverage of full-length CCS increased accordingly, which resulted in an accuracy of CCS ranging from 92.6% to 98.1% (≥5 copies of full-length CCS; Supplementary Figs. 7 and 8). Taken together, these results demonstrate that our approach can effectively reduce sequencing errors by constructing CCSs and, thus, improve the accuracy of circRNA detection.

CIRI-long was tested using 32 nanopore sequencing datasets described above. Initially, the robustness of circRNA detection in two biological replicates was verified. All sequencing reads from each biological replicate were merged into a single dataset, and counts per million (CPM) reads were used to estimate the levels of expression. As shown in Fig. 3c, the expression levels of circRNAs detected in the replicate samples were highly correlated (Pearson correlation coefficient = 0.91), indicating high robustness of our experimental and computational methods. Next, circRNAs identified by nanopore sequencing were compared to those detected by Illumina RNA-seq (Methods) or archived in the public circRNA database (circAtlas 2.0 (ref. [27])). All circRNAs detected in the nanopore sequencing datasets were pooled, and a threshold of two back-spliced reads was set to filter high-confidence circRNAs. A total of 32,223 and 140,588 circRNAs were detected in the Illumina and nanopore datasets, respectively (Fig. 3d); 15,673 circRNAs were detected in both datasets and have also been recorded in the circAtlas database. Only half of circRNAs (50.29%) in the Illumina datasets were present in the nanopore sequencing data; however, these shared circRNAs had higher expression levels than that of the other half of circRNAs and constituted 86.42% of the BSJ reads in the Illumina RNA-seq data (Fig. 3e). For circRNAs detected by nanopore sequencing, 65% and 78% of BSJ reads are confirmed by the Illumina data and the circAtlas database, respectively (Supplementary Fig. 9), and over 80% of medium- or high-abundance circRNAs (≥20 supported reads) are validated in the Illumina RNA-seq or circAtlas database (Fig. 3f). Notably, a large number of circRNAs were present only in the nanopore datasets, and these nanopore-specific circRNAs were generally expressed at low levels and had the length distribution similar

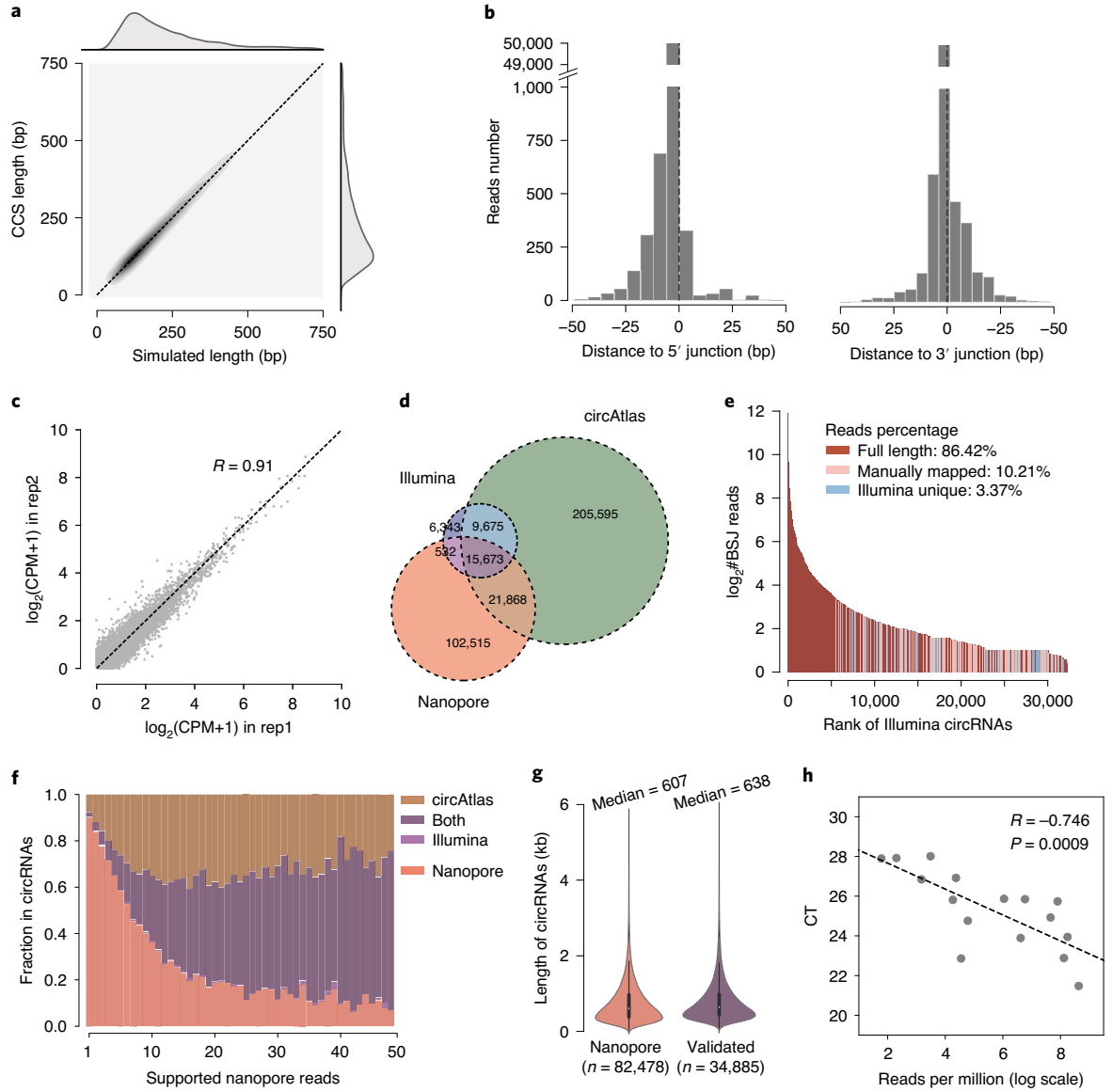


**Fig. 3 | Validation of circRNAs using simulated and Illumina datasets. a**, Correlation between estimated and simulated length of CCSs. The length of CCSs was calculated from the output of circRNA sequences of CIRI-long. **b**, Distance between the predicted and expected splice sites. The *y* axis represents the number of simulated reads, and the *x* axis represents the distance from the expected BSJ sites. **c**, Correlation between the expression levels of circRNAs in nanopore sequencing libraries of two biological replicates. The *x* axis and *y* axis represent the log-scaled expression level (measured by CPM) in two corresponding biological replicates. **d**, Overlap of circRNAs detected in the nanopore data, Illumina data and circAtlas. All circRNAs detected in the mouse samples from circAtlas 2.0 are included. **e**, The expression levels of shared circRNAs and Illumina-specific circRNAs. All circRNAs are ranked according to the expression levels in the Illumina RNA-seq data, and the *y* axis represents the number of BSJ reads in log-scale. Red, circRNAs that are mutually detected in the nanopore and Illumina data. Pink, circRNAs that can be manually aligned to the nanopore reads using minimap2 but lack full-length supporting reads. Blue, circRNA uniquely detected in Illumina data. **f**, Stacked bar plot of the fraction of various types of circRNAs in the nanopore sequencing data with various numbers of supporting reads. **g**, Length distribution of circRNAs with more than two supported reads is represented as violin plots where the center of the inner boxes is the median, and the lower and upper bounds are the first and third quartiles. The upper and lower whiskers represent the proportion of 1.5 interquartile ranges. Pink, nanopore-specific circRNAs ($n = 82{,}748$). Purple, circRNAs that are also detected in the Illumina RNA-seq data ($n = 34{,}885$). **h**, Experimental validation of the quantification results by CIRI-long. Divergent primers were specifically designed targeting 16 randomly selected circRNAs. The *y* axis represents the CT values using RT–PCR, and the *x* axis represents the CPM values in log-scale. The Pearson correlation coefficient (*R*) and *P* values were calculated using linear least squares regression in scipy[43].

to that of the shared circRNAs (Fig. 3g), thus demonstrating high sensitivity of detection of low-abundance circRNAs using nanopore sequencing technology.

To experimentally validate the accuracy of circRNA detection and quantification by CIRI-long, 16 quantitative real-time RT–PCR assays were performed using randomly selected circRNAs with various expression levels (Supplementary Table 3 and Methods). These circRNAs were successfully validated by PCR and Sanger sequencing. In addition, a high correlation (Pearson correlation coefficient = 0.75) between the cycle threshold (CT) values of circRNAs obtained by quantitative RT–PCR and predicted expression levels was observed (Fig. 3h). Moreover, we obtained 40 circRNAs

as a benchmark set from two previous studies[28,29] and validated the quantification consistency between different sequencing protocols (Supplementary Fig. 10). Taken together, these results demonstrate the high reliability of CIRI-long for detection and quantification of circRNAs of the nanopore datasets.

**CIRI-long uncovers the complexity of circular transcripts and splicing events.** Alternative circularization and alternative splicing of circRNAs can generate multiple circular transcripts from the same gene locus, which are considered the main contributors to diversity of circRNAs[17,18]. To accurately identify these events using the nanopore sequencing data, a sequence-based strategy inspired by a previous study was modified to characterize various full-length isoforms of circRNAs[19]. In brief, circular sequences from the same circRNA were aligned to the reference genome and clustered based on pair-wise sequence similarity to distinguish various isoforms (Fig. 4a). Then the consensus sequence of each isoform was generated, and the canonical GT/AG splice signal was used for curation of exon boundaries. Finally, a total of 115,755 alternative circularization events of 15,905 genes were identified in the nanopore sequencing libraries (Supplementary Fig. 11). In contrast, only 25,159 alternative circularization events of 6,928 genes were determined using the Illumina RNA-seq data, which indicates higher sensitivity of circular isoform identification using nanopore sequencing. To investigate the alternative splicing (AS) pattern of circRNAs, four types of AS events in our data were classified, including exon skipping (ES), alternative 5′ or 3′ splicing sites (A5SS and A3SS) and intron retention. For comparison, CIRI-AS was used to identify circRNA exons (cirexons) and AS events in Illumina RNA-seq libraries. A total of 90,759 and 49,646 cirexons were identified in the nanopore and Illumina data, respectively, and 93.6% of cirexons detected in the Illumina data were also detected in the nanopore libraries (Supplementary Fig. 12). A total of 6,714 AS events were detected in the nanopore data, and only 3,243 AS events were identified in the Illumina data (Supplementary Fig. 13). Notably, ES events constitute approximately 50% of AS events identified in the nanopore libraries, and only 33% of AS events in Illumina libraries were considered ES, suggesting that prediction of AS events using short-read sequencing technology might lead to biased detection of circRNA sequences and might fail to detect skipping events of internal cirexons. For instance, 65 full-length isoforms of high-confidence circRNAs (≥10 supporting reads) were detected in the case of the *Rims2* gene based on the nanopore sequencing data (Fig. 4b). Only ten isoforms of these circRNAs were fully reconstructed by CIRI-AS using the Illumina RNA-seq data. Moreover, 17 of 31 isoforms detected only in the nanopore data were validated using circAtlas, demonstrating better sensitivity of detection of complex circRNA isoforms and splicing events using nanopore sequencing.

In previous studies, back-spliced events in mitochondrial genomes were often considered false-positive alignments[4,30,31] because mitochondria lack canonical spliceosome machinery required for biogenesis of most exonic circRNAs. However, recent studies confirmed the presence of mitochondria-derived circRNAs and demonstrated the biological functions of these circRNAs[12–14]. Using nanopore sequencing technology, 156 circRNAs with direct evidence of full-length circular structures were detected based on the mitochondrial genome. Consistent with a previous study[12], most of these mitochondrial circRNAs were encoded by the light chain and were located in the antisense region of the protein-coding genes in mitochondrial DNA (Fig. 4c). The array of circRNA sequences and mitochondrial genes suggests that these circRNAs might be trans-acting factors and can play certain roles in the post-transcriptional regulation of mitochondrial gene expression. Moreover, 44 circRNAs were observed to originate from locations spanning different gene loci; most of these circRNAs originate from a combination of two protein-coding genes (Fig. 4d). Further examination demonstrated that most of these circRNAs were derived from transcriptional read-through events of adjacent genes rather than trans-spliced transcripts, which is consistent with the relatively low incidence of trans-splicing events in vertebrates[32] (Supplementary Fig. 14). For experimental validation, outward primers were designed to amplify the targeting BSJ region, and Sanger sequencing was used for sequence validation of 21 randomly selected circRNAs (Supplementary Table 4). In addition, two transcriptional read-through and two intronic self-ligated circRNAs (mentioned below) were validated with northern blotting (Methods and Supplementary Fig. 15). Overall, these findings demonstrate high performance of our methods in the investigation of various circRNAs.

**A new type of intronic self-ligated circRNAs.** To explore the diversity of circRNA biogenesis, all circRNAs were divided into two groups according to the start and end coordinates of back-splicing events. CircRNAs were defined as annotated if their start and end coordinates matched the annotated splice sites in the GENCODE vM20 annotation; all other circRNAs were categorized as novel. In agreement with previous studies[17,18], 90% of exonic circRNAs were derived from annotated splice signals, whereas most intronic and intergenic circRNAs were classified as novel circRNAs (Fig. 5a). Notably, a small proportion of intronic circRNAs matched the annotated splice sites. To determine potential mechanisms of biogenesis of these circRNAs, intronic circRNAs were classified into three categories: (1) intronic exonic circRNAs (intronic circRNAs with flanking canonical AG/GT splice signals), (2) intronic self-ligated circRNAs and (3) lariat intronic circRNAs (Fig. 5b and Supplementary Fig. 16). Intronic exonic circRNAs use the canonical AG/GT splice signals, and their back-splicing pattern is similar to that of most exonic circRNAs. In contrast, intronic self-ligated circRNAs are generated by direct ligation of the 3′ and 5′ splice sites of the host intron, which was not fully characterized previously. Lariat intronic circRNAs are derived from the 2′–5′ branch of intronic lariats, as described in previous studies[33]. For instance, the sequences of an intronic self-ligated circRNA (circPbrm1) and a lariat intronic circRNA (circCtdsp2) are shown in Fig. 5c. The former is composed of the full-length sequence of intron 27 of *Pbrm1*, and the latter contains only a truncated sequence of intron 3 of *Ctdsp2*. The motif analysis of the flanking sequences adjacent to the BSJ regions demonstrated high similarity between intronic exonic circRNAs and exonic circRNAs because both groups have conserved flanking AG/GT signals (Fig. 5d). In contrast, intronic self-ligated circRNAs have conserved internal GT/AG signals, indicating full-length circularization of the host intron[34]. The PhastCons score[35] was used to assess the conservation of flanking regions in each group of circRNAs. As shown in Fig. 5e, intronic exonic circRNAs have lower conservation scores, and intronic self-ligated and lariat intronic circRNAs are associated with flanking regions with strong conservation scores, indicating distinct biogenesis mechanisms of these circRNAs.

The expression pattern of intronic self-ligated circRNAs was investigated by profiling of circRNAs and their associated forward-spliced junctions (FSJs) and BSJs in the total RNA and RNase R-treated Illumina RNA-seq libraries from 15 normal mouse tissues[36]. All libraries were sequenced using 250-bp paired-end reads, and CIRIquant[25] was used to detect circRNAs using circular decoys constructed based on all introns in the protein-coding genes. As shown in Fig. 5f, the FSJ reads across the BSJ site of intronic self-ligated circRNAs are present in all total RNA libraries of various tissues; however, the intronic self-ligated circRNAs are enriched only in the central nervous system (whole brain, cerebellum and spinal cord), which matches the expression preference of most circRNAs[29]. To estimate potential function of intronic self-ligated circRNAs, the conservation of flanking exons and introns was analyzed (Fig. 5g). The flanking exons of intronic self-ligated circRNAs have

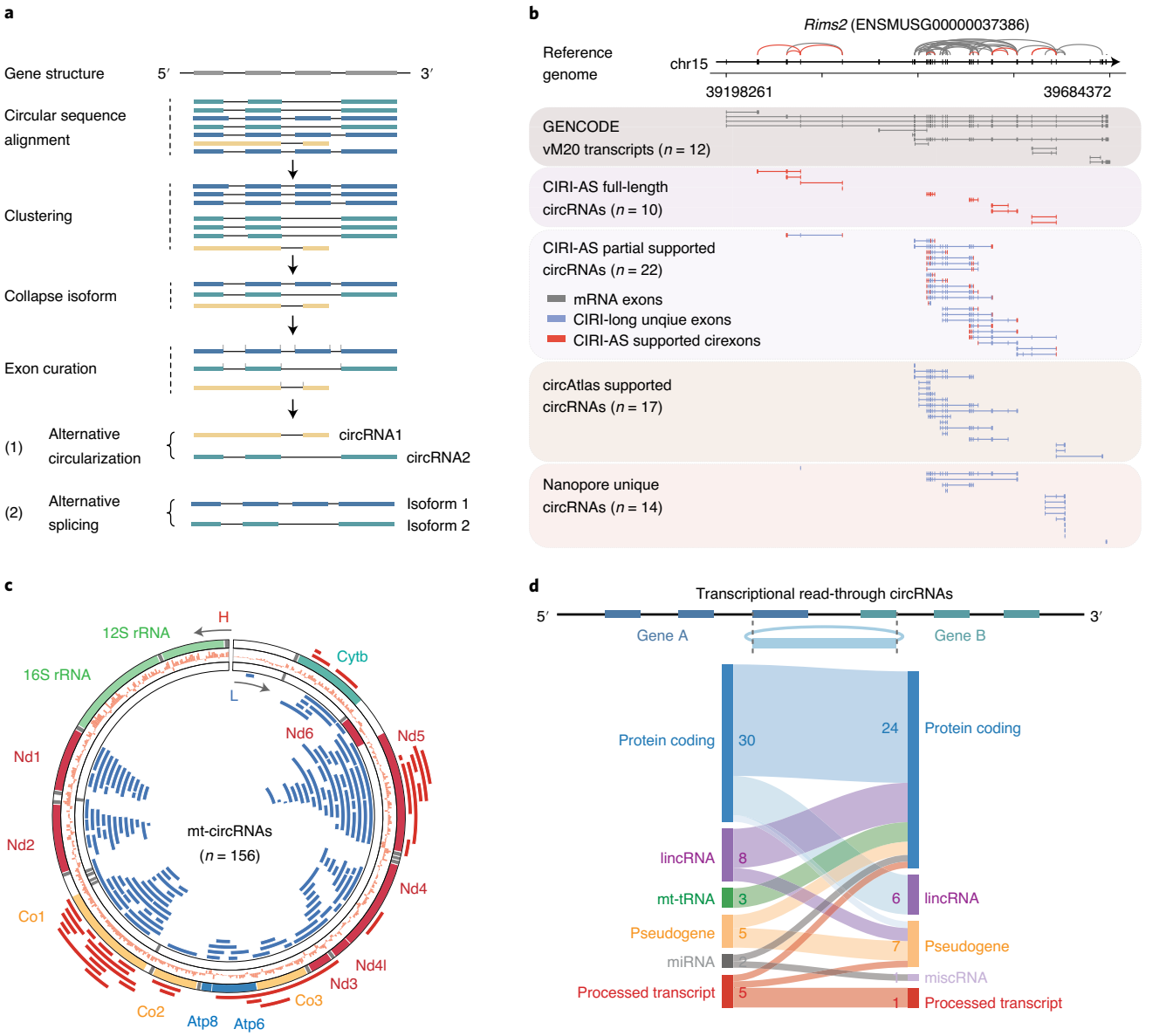


**Fig. 4 | The hidden complexity of circular transcripts and splicing events. a**, Schematic view of detection of alternative circularization and alternative splicing events. Circular reads are aligned to the reference genome and clustered based on pair-wise sequence similarity. A consensus sequence for each cluster is constructed and corrected using annotated splice sites and canonical AG/GT signals. **b**, Example of 65 circRNA isoforms (≥10 supporting reads) detected in the *Rims2* loci. Linear transcripts are extracted from GENCODE annotation and shown in gray. Blue rectangles represent cirexons that are uniquely identified using CIRI-long and nanopore sequencing libraries, and red rectangles are exons supported by CIRI-AS in Illumina RNA-seq data. **c**, Distribution of circRNAs in the mitochondrial genome. A total of 126 circRNAs derived from the light chain (blue) and 30 circRNAs derived from the heavy chain (red) are plotted, and the bars in the middle layer represent the PhastCons scores of each region. **d**, The source and target genes of transcriptional read-through circRNAs. Transcriptional read-through circRNAs were identified from BSJs that spanned multiple genes and annotated using the types of the host genes upstream and downstream of the junction sites. lincRNA, long intervening/intergenic non-coding RNA; miscRNA, miscellaneous RNA; mt-circRNA, mitochondrial circRNA; mt-tRNA, mitochondrial transfer RNA.

a significantly higher conservation compared to the average conservation score of all exons of the host gene (paired-sample *t*-test, $P < 0.001$), and no significant differences in the conservation were observed between circular introns and the corresponding flanking introns.

An intronic self-ligated circRNA derived from the *Tpm1* gene has relatively high supported reads and conservation score; *Tpm1* encodes a highly conserved actin-binding protein involved in the formation of striated and smooth muscle and cytoskeleton in non-muscle cells[37]. circTpm1 is generated from intron 5 of the *Tpm1* gene and has an exceptionally higher conservation level (assessed by phyloP[38] score) compared to introns in the flanking regions (Fig. 5h). In addition, the existence of circTpm1 was further validated using RT–PCR and Sanger sequencing (chr9:67032027–67032793 in Supplementary Table 4). Analysis of circTpm1 indicates a distinct expression pattern of FSJ and BSJ reads. The *Tpm1* gene and the FSJ of circTpm1 are highly expressed in the skeletal muscle and heart, and back-spliced reads are also detected in the cerebellum. It should be noted that a switch between the FSJ and BSJ reads was observed in the whole brain and cerebellum samples, suggesting that circTpm1 is not a byproduct of linear intron splicing. Moreover, the results of de novo profiling using mouse and human tissues

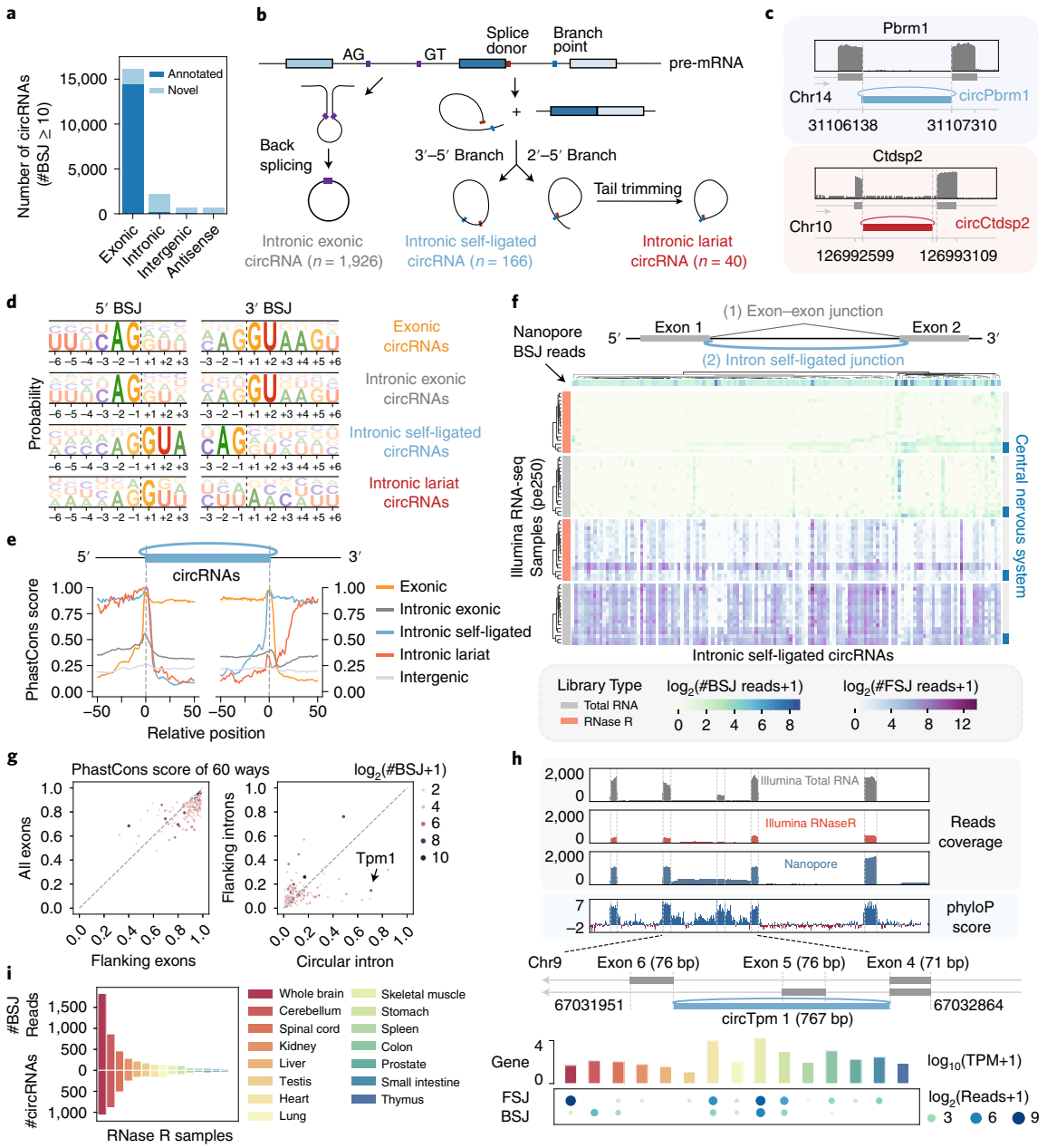


**Fig. 5 | A new type of intronic self-ligated circRNAs. a**, Several high-confidence circRNAs (≥10 supporting reads) with annotated and novel splice sites. The *x* axis represents various types of circRNAs. **b**, Overview of various types of intronic circRNAs. Intronic circRNAs are divided into three categories based on possible biogenesis mechanisms: (1) intronic exonic circRNAs, (2) intronic self-ligated circRNAs and (3) lariat intronic circRNAs. Intronic exonic circRNAs are derived from canonical back-splicing events using flanking AG/GT splice signals. Intronic self-ligated circRNAs are generated through direct ligation of the 3′ and 5′ ends of the host intron, and lariat intronic circRNAs are derived from excised lariat introns. **c**, Examples of an intronic self-ligated circRNA and a lariat intronic circRNA. Upper, intronic self-ligated circPbrm1 consists of the full-length sequence of intron 28 in *Pbrm1*. Bottom, lariat intronic circCtdsp2 consists of the truncated intron sequence of intron 3 in *Ctdsp2*. **d**, Nucleotide frequency of flanking regions of 5′ and 3′ BSJ. The intronic exonic circRNAs and exonic circRNAs have similar external AG/GT signals, and intronic self-ligated circRNAs have internal GT/AG signals. **e**, Average PhastCons score of the flanking regions of the BSJ sites. **f**, Landscape of 166 intronic self-ligated circRNAs in 15 mouse tissues. Intronic self-ligated circRNAs were identified using CIRIquant with all intron coordinates as decoy, and the numbers of the BSJ reads (green) and FSJ reads (purple) were extracted from the CIRIquant output. The red and gray colors on the left of each row represent RNase R-treated and total RNA libraries, and blue colors in the right bar represent tissues related to the central nervous system (whole brain, cerebellum and spinal cord). **g**, Conservation of flanking exons and introns of intronic self-ligated circRNAs. For each exon or intron, the average PhastCons score of all bases was calculated. Left, the *x* axis and *y* axis represent the average score of two flanking exons and all exons in the host gene, respectively. Right, the *x* axis and *y* axis represent the score of the circular intron and the average score of the nearest upstream and downstream introns, respectively. **h**, An example of an intronic self-ligated circRNA, circTpm1. The line plot shows the sequence coverage of circTpm1 in the Illumina and nanopore sequencing datasets and the phyloP score of the flanking regions. The bar plot shows the log-scaled expression levels (measured by TPM) of the host gene and the number of the BSJ/FSJ reads of circTpm1 in various mouse tissues. **i**, De novo profiling of intronic self-ligated circRNAs in mouse RNase R-treated samples from 15 tissues. The bars above and below zero show the total number of the BSJ reads and intronic self-ligated circRNAs detected in each tissue, respectively.

revealed that most intronic self-ligated circRNAs are enriched in the whole brain and cerebellum samples (Fig. 5i and Supplementary Figs. 17 and 19). Thus, a unique expression pattern suggests a role of circTpm1 in the regulation of biological processes. Overall, these findings demonstrate that nanopore sequencing and CIRI-long systematically characterized a new type of circRNA and are helpful in the studies of circRNA functions and biogenesis.

## Discussion

We present an experimental and computational protocol for direct detection of full-length circRNA isoforms using nanopore sequencing. circRNAs are reverse transcribed in multiple rounds of circular amplification of the same circRNA molecule to provide direct evidence of the circular structure of the transcripts. Comprehensive evaluation of CIRI-long using simulated datasets, Illumina RNA-seq data and experimental validation demonstrates the reliability of our approach in decoding the complex array of the circular transcripts and splicing events that have not been investigated previously.

Comparison of various experimental conditions demonstrated that selection of the length of the fragments is the major factor influencing the efficiency of circRNA detection. Other experimental treatments have a slight effect on the length of sequencing reads (Supplementary Fig. 4); however, an increase in the fragment size significantly enhances circRNA detection. Selection of appropriate length (~1 kb) and optimization of other treatment conditions resulted in a 20-fold increase in the percentage of circular reads compared to that obtained by traditional Illumina-based sequencing (Supplementary Figs. 20 and 21). However, considering a limited number of circRNAs, selection of a longer fragment size might result in a lower load of starting cDNA, which leads to a reduction in the sequencing quality. Evaluation using Illumina RNA-seq libraries and the circAtlas database demonstrated that our method can efficiently detect circRNAs with higher expression levels and reliably detect several low-abundance circRNAs. However, a simulation study suggests that the overselection of the fragment size above 1 kb might result in an enrichment in longer circRNAs, which can cause bias in the detection of circRNAs.

Considering that the fraction of circRNAs is usually <1% of total RNA after rRNA depletion, a large sequencing dataset is required for optimal detection of circRNAs using Illumina RNA-seq. Recent studies detected an overwhelming number of circRNAs across several vertebrate species[27]. For instance, circAtlas collected a total of 252,811 circRNAs detected in mouse, whereas only 138,835 linear transcripts were annotated in the GENCODE project[39]. Vast diversity of circRNAs indicates that de novo detection of circular transcripts in various transcriptomes remains challenging. Most currently available approaches to circRNA identification rely on the detection of reads spanning BSJs, which are unable to distinguish linear and circular reads aligned to the internal region of circRNAs. In our method, each transcribed cDNA molecule contains multiple copies of the corresponding full-length circRNA sequence, and each long read provides direct evidence of the presence and sequence of a circRNA. CIRI-long can accurately determine additional alternative circularization and alternative splicing events and provides over a five-fold increase in alternative circularization events using similar data size compared to the results obtained using Illumina RNA-seq. In addition, nanopore sequencing libraries detected a two-fold higher number of cirexons and AS events compared to those detected using the Illumina data. Moreover, 65 full-length circRNA isoforms with more than ten supporting reads are detected within the *Rim2* gene locus, whereas only ten circRNA isoforms can be fully reconstructed using the Illumina RNA-seq data. Furthermore, quantitative RT–PCR of randomly selected circRNAs validated high reliability of circRNA quantification using nanopore sequencing data and demonstrated enhanced applicability of this approach to circRNA studies.

In most circRNA studies, circRNA candidates without the canonical AG/GT splice sites and candidates mapped to the mitochondrial genome are removed to increase the accuracy of circRNA identification[3,17,40,41]. However, these methods fail to detect circRNAs with non-canonical splice signals or circRNAs derived from the mitochondrial genome. We applied our method to investigate 156 mitochondria-derived circRNAs with AG/GT signals. In agreement with a previous study, >80% of these circRNAs are derived from the light chain of mitochondrial genome and might serve as trans-factors for the regulation of mitochondrial gene expression. Moreover, we identified a new type of intronic self-ligated circRNA with a distinct, previously incompletely characterized internal GT/AG splice signal rather than the flanking AG/GT signal characteristic of most exonic and intronic exonic circRNAs.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-00842-6.

Received: 28 September 2020; Accepted: 2 February 2021;
Published online: 11 March 2021

## References

1. Kristensen, L. S. et al. The biogenesis, biology and characterization of circular RNAs. *Nat. Rev. Genet.* **20**, 675–691 (2019).
2. Li, X., Yang, L. & Chen, L.-L. The biogenesis, functions, and challenges of circular RNAs. *Mol. Cell* **71**, 428–442 (2018).
3. Memczak, S. et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338 (2013).
4. Westholm, J. O. et al. Genome-wide analysis of *Drosophila* circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep.* **9**, 1966–1980 (2014).
5. Barrett, S. P., Parker, K. R., Horn, C., Mata, M. & Salzman, J. ciRS-7 exonic sequence is embedded in a long non-coding RNA locus. *PLoS Genet.* **13**, e1007114 (2017).
6. Ruan, H. et al. Comprehensive characterization of circular RNAs in ~1000 human cancer cell lines. *Genome Med.* **11**, 1–14 (2019).
7. Du, W. W. et al. Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2. *Nucleic Acids Res.* **44**, 2846–2858 (2016).
8. Yang, W., Du, W. W., Li, X., Yee, A. J. & Yang, B. B. Foxo3 activity promoted by non-coding effects of circular RNA and Foxo3 pseudogene in the inhibition of tumor growth and angiogenesis. *Oncogene* **35**, 3919–3931 (2016).
9. Legnini, I. et al. Circ-ZNF609 is a circular RNA that can be translated and functions in myogenesis. *Mol. Cell* **66**, 22–37 (2017).
10. Liu, C.-X. et al. Structure and degradation of circular RNAs regulate PKR activation in innate immunity. *Cell* **177**, 865–880 (2019).
11. Xu, X. et al. CircRNA inhibits DNA damage repair by interacting with host gene. *Mol. Cancer* **19**, 128 (2020).
12. Liu, X. et al. Identification of mecciRNAs and their roles in the mitochondrial entry of proteins. *Sci. China Life Sci.* **63**, 1429–1449 (2020).
13. Wu, Z. et al. Mitochondrial genome-derived circRNA mc-COX2 functions as an oncogene in chronic lymphocytic leukemia. *Mol. Ther. Nucleic Acids* **20**, 801–811 (2020).
14. Zhao, Q. et al. Targeting mitochondria-located circRNA SCAR alleviates NASH via reducing mROS output. *Cell* **183**, 76–93 (2020).
15. Wu, J. et al. CircAST: full-length assembly and quantification of alternatively spliced isoforms in circular RNAs. *Genomics Proteomics Bioinformatics* **17**, 522–534 (2019).
16. Zheng, Y., Ji, P., Chen, S., Hou, L. & Zhao, F. Reconstruction of full-length circular RNAs enables isoform-level quantification. *Genome Med.* **11**, 2 (2019).
17. Zhang, X.-O. et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.* **26**, 1277–1287 (2016).
18. Gao, Y. et al. Comprehensive identification of internal structure and alternative splicing events in circular RNAs. *Nat. Commun.* **7**, 12060 (2016).
19. Tang, A. D. et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.* **11**, 1438 (2020).

20. You, X. et al. Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nat. Neurosci.* **18**, 603–610 (2015).
21. Rahimi, K., Venø, M. T., Dupont, D. M. & Kjems, J. Nanopore sequencing of full-length circRNAs in human and mouse brains reveals circRNA-specific exon usage and intron retention. Preprint at *bioRxiv* https://doi.org/10.1101/567164 (2019).
22. Xiao, M.-S. & Wilusz, J. E. An improved method for circular RNA purification using RNase R that efficiently removes linear RNAs containing G-quadruplexes or structured 3′ ends. *Nucleic Acids Res.* **47**, 8755–8769 (2019).
23. Lee, C. Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics* **19**, 999–1008 (2003).
24. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
25. Zhang, J., Chen, S., Yang, J. & Zhao, F. Accurate quantification of circular RNAs identifies extensive circular isoform switching events. *Nat. Commun.* **11**, 90 (2020).
26. Yang, C., Chu, J., Warren, R. L. & Birol, I. NanoSim: nanopore sequence read simulator based on statistical characterization. *Gigascience* **6**, 1–6 (2017).
27. Wu, W., Ji, P. & Zhao, F. CircAtlas: an integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. *Genome Biol.* **21**, 101 (2020).
28. Gruner, H., Cortés-López, M., Cooper, D. A., Bauer, M. & Miura, P. CircRNA accumulation in the aging mouse brain. *Sci. Rep.* **6**, 38907–38907 (2016).
29. Rybak-Wolf, A. et al. Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol. Cell* **58**, 870–885 (2015).
30. Akers, N. K., Schadt, E. E. & Losic, B. STAR chimeric post for rapid detection of circular RNA and fusion transcripts. *Bioinformatics* **34**, 2364–2370 (2018).
31. Ragan, C., Goodall, G. J., Shirokikh, N. E. & Preiss, T. Insights into the biogenesis and potential functions of exonic circular RNA. *Sci. Rep.* **9**, 2048 (2019).
32. Lei, Q. et al. Evolutionary insights into RNA trans-splicing in vertebrates. *Genome Biol. Evol.* **8**, 562–577 (2016).
33. Talhouarne, G. J. S. & Gall, J. G. Lariat intronic RNAs in the cytoplasm of vertebrate cells. *Proc. Natl Acad. Sci. USA* **115**, E7970–E7977 (2018).
34. Taggart, A. J. et al. Large-scale analysis of branchpoint usage across species and cell lines. *Genome Res.* **27**, 639–649 (2017).
35. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
36. Ji, P. et al. Expanded expression landscape and prioritization of circular RNAs in mammals. *Cell Rep.* **26**, 3444–3460 (2019).
37. Takenaga, K., Nakamura, Y., Tokunaga, K., Kageyama, H. & Sakiyama, S. Isolation and characterization of a cDNA that encodes mouse fibroblast tropomyosin isoform 2. *Mol. Cell. Biol.* **8**, 5561–5565 (1988).
38. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
39. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
40. Gao, Y., Zhang, J. & Zhao, F. Circular RNA identification based on multiple seed matching. *Brief. Bioinform.* **19**, 803–810 (2018).
41. Cheng, J., Metge, F. & Dieterich, C. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* **32**, 1094–1096 (2016).
42. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
43. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).


**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021, corrected publication 2021

## Methods

**Nanopore library preparation.** Total RNA from two healthy adult mice was isolated using TRIzol (Invitrogen), and the quality of RNA was assessed with an Agilent fragment analyzer system. A RiboErase kit (human/mouse/rat, Kapa Biosystems) was used to remove rRNA from 1 μg of total RNA. Then total RNA was divided into two groups for construction of RNase R and A-tailing/RNase R libraries. For normal RNase R libraries, RNase R was purchased from Epicentre, and rRNA-depleted RNA was incubated at 37 °C with 20 U $\mu l^{-1}$ of RNase R in a 0.5-μl reaction for 15 min. In the latter group, 5 U $\mu l^{-1}$ of poly(A) polymerase (NEB) was used to add an additional poly(A) tail to the linear transcripts and to increase the digestion efficiency of RNase R. Ribosomal-depleted total RNA was incubated at 37 °C with 1 μl of poly(A) polymerase for 30 min and subjected to RNase R treatment as described above.

Next, RNase R-treated RNA was reverse transcribed using random hexamers and SMARTer or Maxima reverse transcriptase according to the manufacturer's instructions. In the first group, a SMARTer cDNA synthesis kit (Takara Bio) was used, and the 3′ SMART CDS primer II A 5′-AAGCAGTGGTATCAACGCAGAGTACT(30)N-1N-3′ was replaced with 5′-AAGCAGTGGTATCAACGCAGAGTACNNNNNN-3′ to amplify circular RNAs without the poly(A) sequence. Then 3.5 μl of RNA and 1 μl of SMARTer CDS random primer (12 μM) were mixed and incubated at 72 °C for 3 min and 25 °C for 10 min and held at 42 °C. A mixture containing 2 μl of 5× first-strand buffer, 0.25 μl of DTT (100 mM), 1 μl of dNTP (10 mM), 1 μl of SMARTer II A oligo (12 μM), 0.25 μl of RNase inhibitor and 1 μl of SMARTer reverse transcriptase (100 U) was added to the samples followed by incubation at 42 °C for 70 min. During this step, the cDNA was divided into two aliquots; one of the parts was treated with 1 μl of RNase H (5 U $\mu l^{-1}$, NEB) and 0.5 μl of SMARTer reverse transcriptase. Then both aliquots were incubated at 42 °C for another 20 min and then denatured at 70 °C for 10 min.

In the Maxima group, 3.5 μl of RNA was mixed with 7.5 μl of deionized water, 1 μl of random primer (10 μM) and 1 μl of dNTP, incubated at 65 °C for 5 min and cooled on ice. Then a mixture containing 4 μl of 5× RT buffer, 1 μl of RNaseOUT, 1 μl of SMARTer II A oligo and 1 μl of Maxima H Minus reverse transcriptase (2,000 units at 200 U $\mu l^{-1}$ concentration, Thermo Fisher Scientific) was added to the sample and incubated at 25 °C for 10 min and at 50 °C for 70 min. Similarly to the SMARTer group, cDNA was divided into two parts, and one of the parts was treated with 1 μl of RNase H and 0.5 μl of Maxima RT. Both parts were incubated at 50 °C for 20 min and 85 °C for 5 min and finally held at 4 °C.

To obtain sufficient cDNA products for sequencing, PCR amplification was performed using 2 μL of cDNA with NEBNext LongAmp Taq DNA Polymerase and SMARTer primers under the following conditions: 95 °C for 30 s; 19 or 20 cycles of 95 °C for 15 s, 62 °C for 15 s, and 65 °C for 120 s; 65 °C for 6 min and hold at 4 °C. Finally, Agencourt AMPure XP magnetic beads (Beckman) were used for size selection of the cDNA fragments. An increase in the DNA-to-bead ratio (1:1, 1:0.6 and 1:0.5) resulted in capture of longer fragments, and the length of fragments increased accordingly to approximately 400 bp, 600 bp and 1 kbp, respectively. Steps for library construction are detailed in the CIRI-long manual pages at https://ciri-cookbook.readthedocs.io/en/latest/CIRI-long_sequencing.html.

**Nanopore sequencing and base-calling.** cDNA libraries were prepared according to the ONT protocol SQK-LSK109 and barcoded with EXP-NBD104 and EXP-NBD114 kits, and nanopore sequencing was performed using the MinION (MN26543) platform with a FLO-MIN106 flow cell. The MinKNOW (v3.2.6) interface was used with the base-calling option disabled, and real-time GPU-based base-calling and demultiplexing were performed using the GPU version of ont-guppy (v3.3.0) with custom scripts from GitHub (https://github.com/rrwick/MinION-desktop) and the R9.4 high-accuracy model.

**Adapter trimming and alignment.** Quality control of base-called raw reads was performed using pycoQC[44] (v2.5.0.14). To remove barcodes and adapters, reads were trimmed using Porechop v2.0.4 (https://github.com/rrwick/Porechop) with manually specified sequences of SMARTer adapters (AAGCAGTGGTATCAACGCAGAGTAC and GTACTCTGCGTTGATACCACTGCTT). Cleaned reads were aligned to the mouse genome (GRCm38) version M20 (Ensembl 95) using minimap2 (ref. [45]) with the '--x splice' option.

**Overview of CIRI-long.** The identification of circRNA sequences from nanopore reads has two main steps, including candidate circRNA identification and isoform collapsing. Clean nanopore reads were scanned using a set of *k*-mers and homopolymer-compressed *k*-mers to detect repetitive patterns, and a partial order alignment algorithm was employed to generate consensus circular sequences of each read. Then candidate sequences were mapped to the genome for detection of the BSJ, and splice signals were used to determine circRNA boundaries. Finally, all candidate circRNA reads were aggregated and clustered based on pair-wise sequence similarity to identify isoforms within each BSJ site.

**Detection of candidate circRNAs from nanopore sequencing reads.** CIRI-long takes a set of ordinary *k*-mers and homopolymer-compressed *k*-mers as seeds to identify repetitive patterns of raw nanopore reads. By default, reads were scanned using $k=8$ and $k=11$, and the occurrence coordinates of *k*-mer *i* were stored. For each *k*-mer, the coordinates were sorted, and the distance between two adjacent coordinates was defined as a repeat hit. Then the distance with maximum supported hits was assigned as the common distance *d*, which represents the expected length of circular repetitive elements in the candidate read.

Considering the high error rate and prevalence of insertion/deletion in nanopore sequencing reads, the distance between two matching *k*-mers might deviate within certain ranges. According to a random walk model[46], the distance between *k*-mer hits can range between $d \pm 2.3\sqrt{p_{\text{indel}} \cdot d}$ 95% of the time, where $p_{\text{indel}}$ is the probability of insertion and deletion sequencing error and is set to 0.1 by default. To identify circular repetitive patterns, reads were scanned for the second time, and the hits of identical *k*-mers with distance between $d \pm 2.3\sqrt{p_{\text{indel}} \cdot d}$ were chained together.

To distinguish circRNA reads and reads containing tandem repeats, the start and end positions of the *k*-mer chains should be located within the first and last 100 bp of the sequencing reads. The read was then split into repetitive segments according to the position of the first *k*-mer in the final chain. Then an SIMD accelerated partial order alignment library (SPOA, https://github.com/rvaser/spoa.git) was employed with a customized scoring function (-l 2 -m 10 -n -4 -g -8 -e -2 -q -24 -c -1) to generate a CCS from the repetitive segments, which represents the sequence of a putative circRNA. The Levenstein distance between the segments and CCS was calculated, and a threshold of 80% similarity between repetitive segments and CCS was used to screen high-confidence candidates for circRNAs.

To identify the genome position of back-splicing events, CCSs were aligned to the reference genome, where CCSs longer than 150 bp are initially aligned using mappy (the Python interface of minimap2) with splice preset, and shorter sequences are aligned using bwapy (the Python bindings of bwa mem[47]) with '-x ont2d -T 19' options. To accurately identify BSJ sites from the CCS, an iterative alignment strategy was used, whereby the CCS is initially mapped to the genome, and the unmapped segment on the head or tail region is appended to the opposite end. Then the reordered CCS is aligned again to determine if the alignment has a better score than the previous alignment. The iterative alignments converge after the CCS is transformed to the right order on the genome. For converged CCSs that have a remaining soft-clipped segment, a stripped Smith–Waterman algorithm is used to map the clipped segments to the flanking upstream and downstream regions as described in a previous study[40]. Then annotated splice sites are extracted from annotation GTF, and de novo canonical GT/AG splice signals are used to correct the junction sites using dynamic programming. If no annotated splice signals are detected, the flanking region is scanned for non-canonical U2- (GC/AG), and the combination of U11/U12-type (AT/AC, GT/AC and AT/AG) signals and reads containing no splicing signals are filtered out.

**Isoform collapsing.** To accurately determine the sequence of each isoform, candidate circRNA reads are clustered based on their coordinates in the reference genome, and each cluster represents a putative circRNA. Then all sequences in the same cluster are hierarchically clustered based on pair-wise sequence similarity, and consensus sequences of each cluster are generated as described above. Each consensus sequence indicates an isoform of circRNA. Finally, consensus sequences are aligned to the reference genome, and annotated splice site and de novo canonical splice signals are used for correction of the junctions of circRNA and cirexon boundaries. Finally, the supported reads are counted and assigned, and the expression matrices of high-confidence circRNAs are generated as the final output.

**Simulation dataset.** We used NanoSim[26] to generate simulated nanopore sequencing datasets. To model real sequencing error in nanopore sequencing, total RNA of the control mouse liver attached in SMARTe PCR cDNA synthesis kit was initially sequenced using the same MinNION instrument. Nanopore reads are aligned to the mm10 (GENCODE, vM20) genome using transcriptome mode with -a minimap2 and --no_intron_retention options, and features of the sequencing errors are characterized in simulation. Transcript expression patterns are quantified against the reference transcriptome sequences using the quantification mode of NanoSim. Then circRNA coordinates from exonic regions are randomly generated, and the full-length sequences of simulated circRNAs are joined multiple times (>10×) to simulate circular transcripts. Finally, 200,000 reads from the reference transcriptome and 200,000 reads from the simulated circular transcripts are simulated using the error profiles generated as described above and pooled together as the simulation dataset for the downstream analysis. To evaluate the performance of CIRI-long, the precision and recall rates of circRNA identification were calculated, and the overall performance was assessed by the $F_1$ score using the equation

$$F_1\ \text{score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

**Quantitative PCR validation.** To evaluate the reliability of circRNA quantification using nanopore sequencing data, outward divergent primer sets (Supplementary Table 3) were designed to quantify the expression of 16 randomly selected

circRNAs with variable numbers of supported reads ranging from 200 to 30,000. To validate various types of detected circRNAs, specific primers were designed targeting 21 randomly selected circRNAs, including one exonic, five mitochondria-derived, five transcriptional read-through, five intronic self-ligated and five lariat intronic circRNAs. For quantitative real-time PCR, cDNAs were prepared using RNase R and the reverse transcription kit described above. Sequences of the PCR products were determined using Sanger sequencing (Supplementary Table 4). Dual peaks or noises conducted by homopolymers have been observed in the BSJ regions for five circRNAs (marked as 'ambiguous boundaries'), but the back-splicing structure can still be inferred from the alignment results of adjacent sequences.

**Northern blotting.** RNA was isolated from adult mouse brain samples using TRIzol and incubated with or without RNase R. Probes targeting BSJ sequences of four circRNAs and 5.8S rRNA as control for northern blotting were synthesized with 3′ Biotin label (Supplementary Table 5). RNA samples were electrophoresed with 8% denaturing urea polyacrylamide gel and transferred to Hybond-N+ nylon membranes (GE), which were then incubated with the hydration buffer containing the probes and hybridized overnight. Finally, the RNA signal was detected using the Chemiluminescent Nucleic Acid Detection Module (Thermo Fisher Scientific).

**Preparation and analysis of Illumina RNA-seq library.** To construct the Illumina RNA-seq libraries, rRNA-depleted total RNA was extracted as described above. Total RNA was divided into two groups after depletion of rRNA for construction of riboMinus and riboMinus/RNase R libraries. To effectively enrich circRNAs, RNase R from Epicentre was used, and rRNA-depleted RNA was incubated at 37 °C with 20 U $\mu l^{-1}$ of RNase R in a 0.5-μl reaction for 15 min. Then both riboMinus and riboMinus/RNase R libraries were reverse transcribed using a SMARTer PCR cDNA synthesis kit (Takara Bio). The 3′ SMART CDS Primer II A 5′-AAGCAGTGGTATCAACGCAGAGTACT(30)N-1N-3′ was replaced with 5′-AAGCAGTGGTATCAACGCAGAGTACNNNNNN-3′ to amplify circular RNAs without the poly(A) sequence. All four cDNA libraries were shipped to Annoroad Gene Technology and sequenced using an Illumina HiSeqX10 sequencer according to the manufacturer's instructions. In summary, a total of 758,741,512 paired-end 150-bp reads were generated with an average size of 26.5 Gb of data for each library.

For analysis of Illumina RNA-seq data, raw sequencing reads were assessed using FastQC (v0.11.9). Cleaned reads were mapped to the GRCm38 mouse genome using HISAT version 2.1.0 with default parameters. Eight tools, including CIRI2 (ref. [40]) (v2.0.6), CIRCexplorer2 (ref. [17]) (v2.3.5), DCC[41] (v0.4.7), find_circ[48] (v1.2), KNIFE[49] (v1.4), Sailfish-cir[50] (v0.11a), MapSplice[51] (v2.2.1), UROBORUS[52] (v2.0.0), circRNA_finder[4] (v1.1) and segemehl[53] (v0.3.4), were used for circRNA predictions in all libraries following the instructions of the software documentation; circRNAs with at least two reads detected by two tools were retained for downstream analysis. In addition, CIRI-AS (v1.2) was used to detect the sequence of circRNAs in the output of CIRI2.

**Detection of intronic self-ligated circRNAs in public RNA-seq data.** Public RNA-seq data from the human and mouse transcriptomes were downloaded from the National Genomics Data Center (China National Center for Bioinformation)[54] under accession number PRJCA000751. The reference genomes of human (release 34, GRCh38.p13) and mouse (release M20, GRCm38.p13) were downloaded from GENCODE. For detection of intronic self-ligated circRNAs in Illumina RNA-seq libraries, coordinates of all introns of the protein-coding genes were initially extracted in the BED format. Then CIRIquant (v1.1.1) was used for detection using the BED file as input with the --bed option. FSJ reads were detected by alignment of the raw reads to the reference genome using HISAT2, where gapped aligned (with 'N' in CIGAR strings) segments[55] spanning circularized introns were extracted as FSJ reads of these intronic self-ligated circRNAs.

**Conservation analysis.** For conservation analysis, conservation scores by PhastCons and phyloP (phylogenetic *P* values) for multiple alignments of 59 vertebrate genomes to the mouse genome were downloaded from the UCSC website (http://hgdownload.cse.ucsc.edu/goldenPath/mm10). For each type of circRNA, the average scores of the flanking 50-bp region of the 5′ and 3′ BSJ sites were calculated for comparison.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The sequence data generated in this study have been deposited to the National Genomics Data Center[54] (China National Center for Bioinformation: https://bigd.big.ac.cn/gsa) with accession number CRA003317. Details of these datasets are included in Supplementary Table 1 and the Methods section.

## Code availability

CIRI-long is implemented in Python and can be freely accessed at https://github.com/Kevinzjy/CIRI-long. The software is packaged with sample datasets and has been extensively tested on Linux.

## References


44. Leger, A. & Leonardi, T. pycoQC, interactive quality control for Oxford Nanopore Sequencing. *J. Open Source Softw.* **4**, 1236 (2019).
45. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
46. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
47. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
48. Hansen, T. B. et al. Natural RNA circles function as efficient microRNA sponges. *Nature* **495**, 384–388 (2013).
49. Szabo, L. et al. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol.* **16**, 126 (2015).
50. Li, M. et al. Quantifying circular RNA expression from RNA-seq data using model-based framework. *Bioinformatics* **33**, 2131–2139 (2017).
51. Wang, K. et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, e178 (2010).
52. Song, X. et al. Circular RNA profile in gliomas revealed by identification tool UROBORUS. *Nucleic Acids Res.* **44**, e87 (2016).
53. Hoffmann, S. et al. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol.* **15**, R34 (2014).
54. Wang, Y. et al. GSA: genome sequence archive. *Genomics Proteomics Bioinformatics* **15**, 14–18 (2017).
55. Gao, Y. et al. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.* **16**, 4 (2015).


## Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (32025009, 91940306, 31722031, 32071463, 91951209 and 91640117) and the National Key R&D Program (2018YFC0910400).

## Author contributions

F.Z. conceived the project. J.Z. implemented the algorithm and performed data analysis. L.H., Z.Z., Y.X. and X.Z. performed the experiments and generated sequencing data. J.Z., J.P. and F.Z. wrote the manuscript with the contribution of all authors.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-00842-6.

**Correspondence and requests for materials** should be addressed to F.Z.

**Peer review information** Nature Biotechnology thanks the anonymous reviewers for their contribution to the peer review of this work.

**Reprints and permissions information** is available at www.nature.com/reprints.

nature research

Corresponding author(s): Fangqing Zhao

Last updated by author(s): Feb 1, 2021

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | Public data were downloaded from the Sequence Read Archive using sratoolkit (2.9.4). |
| Data analysis | Our proposed method CIRI-long is available at https://github.com/Kevinzjy/CIRI-long. Simulation data was generated using NanoSim (v2.6.0). Nanopore reads were sequenced using MinKNOW(v.3.2.6) interface, basecalled with ont-guppy (v3.3.0). Quality control was performed using pycoQC (v2.5.0.14). Porechop (v2.0.4) were used to trim sequencing adapters, and minimap2 (v2.17) were employed for reads alignment. For RNA-seq analysis, CIRI2 (v2.0.6), CIRCexplorer2 (v.2.3.5), DCC (v0.4.7), find_circ (v1.2), KNIFE (v1.4), Mapsplice (v.2.2.1), UROBORUS (v2.0.0) , circRNA_finder (v1.1) and CIRIquant (v1.1.1) were used for circRNA detection, and bwa (v0.7.17) were used for sequence alignment. CIRI-AS (v1.2) was performed to detect the internal structure of circRNAs. All data were analysed using python 3.6.8 with scikit-learn (0.21.3) and scipy (1.3.1), and visualized using matplotlib (3.2.1) |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The sequence data generated in this study have been deposited to the National Genomics Data Center63 (China Nation Center for Bioinformation, https://bigd.big.ac.cn/gsa) with accession numbers CRA003317. Details of these data sets are included in Supplementary Table 1 and the Methods section.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | We generated 32 Nanopore sequencing libraries and 10 Illumina RNA-seq libraries of adult mouse brain from two biological replicates. For circRNA expression analysis, we collected 70 public RNA-seq samples from different human and mouse tissues from previous study, including total RNA and RNase R treated samples for each tissues. We generated two biological replicates for each group in our study, which should be sufficient as in most circRNA studies. |
| Data exclusions | No data were excluded from the analyses. |
| Replication | We generated two biological replicates for all conditions explored in our study. All attempts were confirmed to be successful. |
| Randomization | Randomization was not relevant to our study, as we used public datasets from previous studies and no group assignment was needed. |
| Blinding | Blinding was not necessary for the same reason as no objective scoring was applied in our study. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☐ | ☒ Animals and other organisms |
| ☒ | ☐ Human research participants |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

# Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

| | |
|---|---|
| Laboratory animals | Total RNA extracted from two adult E14.5 male C57 mice were used in our study. |
| Wild animals | The study did not involve wild animals. |
| Field-collected samples | The study did not involve field-collected samples. |
| Ethics oversight | No ethical approval was required , as only extracted total RNA was used in our study and no animal experiments were performed. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.