Genomics Proteomics Bioinformatics 17 (2019) 511–521

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

APPLICATION NOTE

CIRCexplorer3: A CLEAR Pipeline for Direct Comparison of Circular and Linear RNA Expression

Xu-Kai Ma [1,a], Meng-Ran Wang [1,b], Chu-Xiao Liu [2,c], Rui Dong [1,$,d], Gordon G. Carmichael [3,e], Ling-Ling Chen [2,4,f], Li Yang [1,4,*,g]

[1] *CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[2] *State Key Laboratory of Molecular Biology, CAS Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[3] *Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT 06030, USA*
[4] *School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China*

Received 16 October 2019; accepted 20 November 2019
Available online 3 January 2020

Handled by Yi Xing

KEYWORDS

Circular RNA;
Back-splicing;
Linear RNA;
Pre-mRNA splicing;
Ribo$^-$ RNA-seq

Abstract Sequences of **circular RNAs** (circRNAs) produced from **back-splicing** of exon(s) completely overlap with those from cognate **linear RNAs** transcribed from the same gene loci with the exception of their back-splicing junction (BSJ) sites. Therefore, examination of global circRNA expression from RNA-seq datasets generally relies on the detection of RNA-seq fragments spanning BSJ sites, which is different from the quantification of linear RNA expression by normalized RNA-seq fragments mapped to whole gene bodies. Thus, direct comparison of circular and linear RNA expression from the same gene loci in a genome-wide manner has remained challenging. Here, we update the previously-reported CIRCexplorer pipeline to version 3 for circular and linear RNA expression analysis from ribosomal-RNA depleted RNA-seq (CIRCexplorer3-CLEAR). A new

* Corresponding author.
E-mail: liyang@picb.ac.cn (Yang L).
$ Present address: Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA.
[a] ORCID: 0000-0003-3779-4070.
[b] ORCID: 0000-0002-9281-9501.
[c] ORCID: 0000-0002-6725-8327.
[d] ORCID: 0000-0003-0985-6211.
[e] ORCID: 0000-0002-0379-6580.
[f] ORCID: 0000-0001-9501-0305.
[g] ORCID: 0000-0001-8833-7473.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2019.11.004

quantitation parameter, fragments per billion mapped bases (FPB), is applied to evaluate circular and linear RNA expression individually by fragments mapped to circRNA-specific BSJ sites or to linear RNA-specific splicing junction (SJ) sites. Comparison of circular and linear RNA expression levels is directly achieved by dividing FPB_{circ} by FPB_{linear} to generate a CIRCscore, which indicates the relative circRNA expression level using linear RNA expression level as the background. Highly-expressed circRNAs with low cognate linear RNA expression background can be readily identified by CIRCexplorer3-CLEAR for further investigation. CIRCexplorer3-CLEAR is publically available at https://github.com/YangLab/CLEAR.

Introduction

Eukaryotic pre-mRNA splicing is catalyzed by spliceosomes to join upstream 5′ splice donor sites with downstream 3′ splice acceptor sites to produce linear (m)RNAs. Interestingly, downstream 5′ splice donor sites can also be linked to upstream 3′ splice acceptor sites, referred to as back-splicing, leading to the production of circular RNAs (circRNAs) [1–3]. Unlike most mature linear RNAs (including both coding and long non-coding RNAs), circRNAs are covalently closed and lack 3′-end poly(A) tails, resulting in their depletion in poly(A)$^+$ RNA-seq datasets. By taking advantage of RNA-seq datasets that profile non-polyadenylated transcripts and computational approaches that aim to identify fragments mapped to back-splicing junction (BSJ) sites [4,5], a large number of circRNAs have been successfully profiled as being co-expressed with their cognate linear RNAs from the same gene loci [2,3,6–8]. Recent studies have shown that the biogenesis of circRNAs is catalyzed by canonical spliceosomal machinery and modulated by both *cis*-elements and *trans*-factors [1–3,9,10]. Importantly, increasing lines of evidence have revealed that some circRNAs play important roles under physiological and pathological conditions, such as neurogenesis, cancer metastasis, and innate immune responses, with different modes of action [6,11–14].

Despite these findings, comprehensive characterization of circRNA biogenesis and function has been impeded because the majority of circRNAs are processed from middle exons of genes and their sequences almost completely overlap with those of their cognate linear RNAs except for the BSJ sites [2]. Thus, a direct expression comparison of circular and linear RNAs from the same gene loci in a genome-wide manner has remained challenging. The primary obstacle for direct expression comparison is owing to distinct strategies for circular and linear RNA quantification from mapped RNA-seq fragments. In general, RNA-seq fragments that are solely mapped to BSJ sites are used to represent circRNA expression, such as by raw or normalized fragment counts (fragments per million mapped fragments, FPM) as shown in **Figure 1**A (left). On the other hand, RNA-seq fragments mapped to exon bodies and exon-exon splicing junction (SJ) sites are summed up and normalized for linear RNA quantification, such as by fragments per kilobase of transcript per million mapped fragments (FPKM) [15] as shown in Figure 1A (right). Since FPM is unscaled to FPKM, the relative expression levels of most circRNAs are not comparable to those of their cognate linear RNAs when analyzing RNA-seq datasets.

To solve this problem, we have further updated our previously-reported CIRCexplorer [7] and CIRCexplorer2 [16] pipelines to version 3 for circular and linear RNA expression analysis from ribosomal-RNA depleted RNA-seq (CIRCexplorer3-CLEAR, or CLEAR for simplicity, Figure 1B). With the CLEAR pipeline, RNA-seq fragments mapped to circRNA-specific BSJ sites or linear RNA-specific SJ sites are individually normalized to evaluate circular or linear RNA expression, each in fragments per billion mapped bases (FPB). Unlike using the non-comparable FPM and FPKM values, expression levels of circular and linear RNAs are both quantified by FPB values with the CLEAR pipeline, and thus can be directly compared by dividing FPB_{circ} by FPB_{linear} to generate a CIRCscore. In this scenario, relative circRNA expression can be evaluated by using linear RNA expression as an expression background, and highly-expressed circRNAs with low cognate linear RNA expression background can be identified for further functional studies. Parallel analyses further suggest that CLEAR is more reliable for circular and linear RNA expression comparison than other related methods, with economic memory usage and comparable time consumption.

Method

Direct circular and linear RNA expression comparison by the CLEAR pipeline

CLEAR was developed to achieve direct circular and linear RNA expression comparison. Ribo$^-$ RNA-seq datasets that profile both polyadenylated linear and non-polyadenylated circular RNAs in parallel are used for precise circular and linear RNA expression comparison.

The CLEAR pipeline includes two main steps: alignment and quantification (Figure 1B). For the alignment, ribo$^-$ RNA-seq fragments were first mapped by HISAT2 [17] (version 2.0.5; parameters: hisat2 --no-softclip --score-min L,-16,0 --mp 7,7 --rfg 0,7 --rdg 0,7 --dta -k 1 --max-seeds 20) against the GRCh38/hg38 human reference genome with known gene annotations (Figure S1) for subsequent linear RNA quantification analysis. HISAT2-unmapped fragments were then mapped to the same GRCh38/hg38 reference genome using TopHat-Fusion (version 2.0.12; parameters: tophat2 -fusion-search --keep-fasta-order --bowtie1 --no-coverage-search) for subsequent circRNA quantification.

For the quantification, we applied a new FPB value to quantitate linear RNA expression by HISAT2-mapped fragments to SJ sites of the maximally-expressed transcript annotation (Figure S2). The maximally-expressed transcript of a given gene is selected with the highest FPKM value, which is calculated by StringTie (version 1.3.3; parameters: stringtie -e -G) from HISAT2 aligned BAM file [18]. Fragments mapped to BSJs were retrieved from TopHat-Fusion as previously reported (version 2.3.6; parameters: CIRCexplorer2

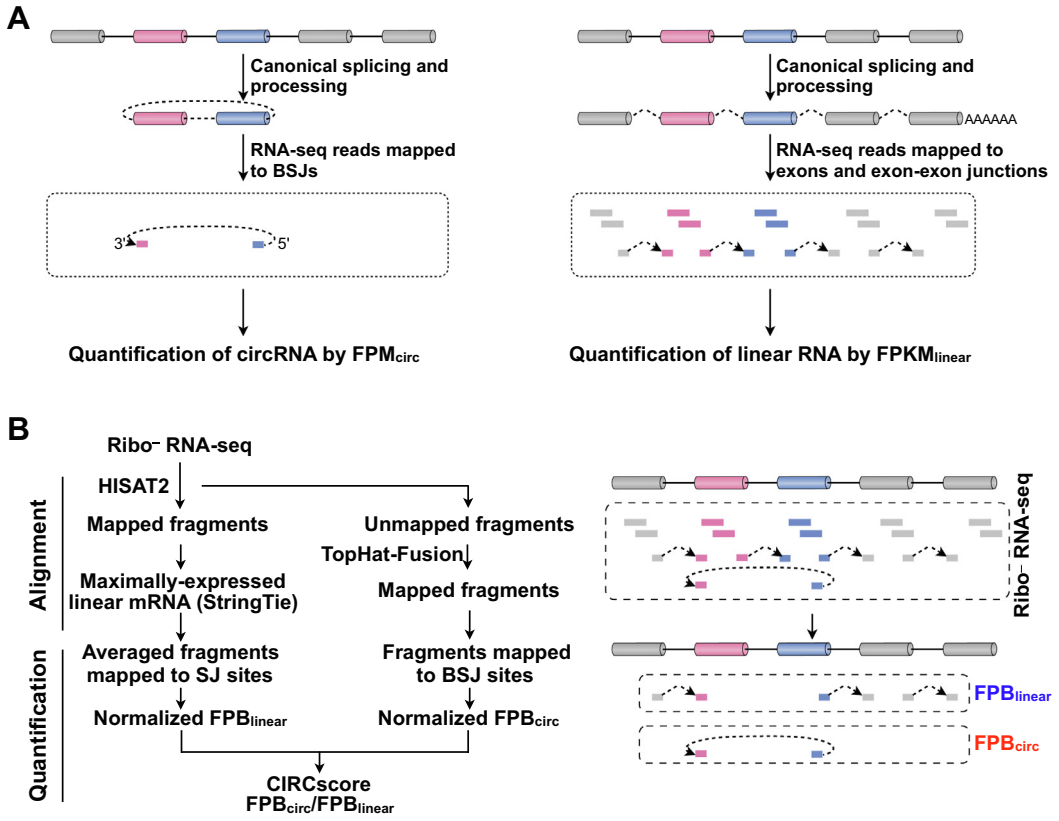

Figure 1 A computational pipeline for direct circular and linear RNA expression comparison

A. Schematic diagram to show different quantification strategies for circRNAs (left) and linear RNAs (right). FPM for circRNA quantification and FPKM for linear RNA quantification are unscaled and incomparable. **B.** Development of the CIRCexplorer3-CLEAR pipeline for circular and linear RNA expression analysis from ribosomal-RNA depleted (ribo$^-$) RNA-seq. Schematic diagram of the CIRCexplorer3-CLEAR pipeline (left) and the same strategy for circular or linear RNA quantification by SJ or BSJ site mapped FPB (right). The CIRCscore derived by dividing FPB_{circ} by FPB_{linear} allows direct comparison of circular and linear RNA expression. FPM, fragments per million mapped fragments; FPKM, fragments per kilobase of transcript per million mapped fragments; FPB, fragments per billion mapped base; SJ, splicing junction; BSJ, back-splicing junction.

parse -f -t TopHat-Fusion) [16,19] and normalized by totally-mapped bases to obtain FPB values for circRNA quantification.

Direct comparison of circular and linear RNA expression is achieved using the CIRCscore value that divides FPB_{circ} by FPB_{linear}, which represents relative circRNA expression using linear RNA expression as the background.

Flexibility of the CLEAR pipeline

Other aligners, including TopHat2 (version 2.0.12; parameter: tophat2 -a 6 --microexon-search -m 2 -g 1) with known gene annotations (Figure S3) or MapSplice (version 2.1.8 with default parameters) with gene annotations (ensGene_v89.txt updated at 2017/05/08) can also be used in the CLEAR pipeline with similar outputs.

In the CLEAR pipeline, comparable circular or linear RNA expression by FPBs and their direct comparison by the CIRCscore can be obtained directly from raw RNA-seq FASTQ files or processed RNA-seq results, such as CIRCexplorer2 output files [16]. Please see https://github.com/YangLab/CLEAR for details.

Cell culture

PA1 cells were purchased from the American Type Culture Collection (ATCC; http://www.atcc.org), and maintained in MEMα supplemented with 10% FBS, 1% glutamine and 0.1% penicillin/streptomycin at 37 °C in a 5% CO_2 cell culture incubator. PA1 cells were routinely tested to exclude mycoplasma contamination.

Comparison of FPB with qPCR quantification

Total RNAs from cultured PA1 cells were extracted with Trizol (Thermo Fisher Scientific; Cat No. 15596018, Waltham, USA) according to the manufacturer's protocol. Extracted RNAs were treated with DNase I (DNA-free™ kit; Thermo Fisher Scientific; Catalog No. AM1907, Waltham, USA), and reversely transcribed with SuperScript III (Thermo Fisher Scientific; Catalog No. 18080044) to produce cDNA and then applied for qPCR analysis. Expression of *ACTB*, which encodes β-actin, was examined as an internal control for normalization. Expression of examined linear and circular RNAs

was determined from three independent experiments. The primers used in this study are listed in Table S1.

Mapping efficiencies of circRNAs by different pipelines

Three different mapping strategies, including CLEAR-embedded CIRCexplorer2, MapSplice, and circtools with embedded tool (detect circRNAs from chimeric reads, DCC) [20,21], were applied to fetch fragments mapped to BSJ and/or SJ sites in PA1 ribo$^-$ RNA-seq dataset [16,22]. Efficiencies of BSJ-mapped fragments were compared by all three pipelines. Normalized circRNA expression was compared between CLEAR with CIRCscores and circtools with circular over linear ratios (CLRs).

Specifically, for the CLEAR pipeline, fragments mapped to BSJ or SJ sites and CIRCscores were directly obtained by one single command line: -g hg38.fa -i hisat2_index -j bowtie1_index -G gene.gtf -o out_dir -p 10. For MapSplice pipeline, fragments mapped to BSJ sites were obtained by: -p 10 --fusion --min-fusion-distance 200 --gene-gtf gene.gtf -o out_dir -c chromosomes -x bowtie1_index -1 PA1. For circtools, PA1 ribo$^-$ RNA-seq dataset [16,22] were mapped by circtools pipeline with tool (spliced transcripts alignment to a reference, STAR) as suggested at https://docs.circ.tools/en/latest/Detect.html. After a series of reformatting steps, fragments mapped to BSJ or SJ sites were obtained by circtools with parameters: detect @samplesheet -T 10 -N -D -an gene.gtf -F -Nr 1 1 -fg -G -A hg38.fa -B @bam_files.txt. CLRs of circRNAs were finally calculated by dividing BSJ fragments with the mean of SJ-mapped fragments with customized scripts according to circtools.

For the comparison of consumed memories and elapsed time by CLEAR or circtools, ribo$^-$ RNA-seq datasets in PA1 [16,22] or cortex [23] were used for the analysis with parameters described above. Consumed memories were recorded by linux command *ps* every 20 s.

RNA-seq datasets used in this study

Datasets used for this study include publicly available ribo$^-$, poly(A)$^+$, poly(A)$^-$/ribo$^-$, and RNase R RNA-seq datasets from PA1 cell line [16,22], ribo$^-$ RNA-seq datasets of 12 tissues from ENCODE [23] (Table S2), as well as ribo$^-$ RNA-seq datasets of 20 human hepatocellular carcinoma (HCC) samples and their paired normal samples from Gene Expression Omnibus (GEO: GSE77509) [24].

Results

Development of the CLEAR pipeline

The CLEAR pipeline was set up for direct circular and linear RNA expression comparison on a genome-wide scale (Figure 1B). Two characteristic features for circular and linear RNA quantification are applied in the CLEAR pipeline. Similar to circRNA quantification by RNA-seq fragments solely mapped to BSJ sites, fragments that only map to canonical SJ sites by HISAT2 are used for linear RNA quantification (Figure 1B, right). Different from commonly-used FPKM that counts fragments mapped to both exon bodies and SJ sites, linear RNA quantification by fragments only mapped to canonical SJ sites is comparable to circRNA quantification by those mapped to BSJ sites (Figure 1B). In addition, fragments mapped to SJ or BSJ sites are normalized by totally mapped bases, rather than by totally mapped fragments, to get FPB for linear or circular RNA quantification (Figure1B, left, Quantification). Direct circular and linear RNA expression comparison can then be achieved with the CIRCscore that divides FPB_{circ} by FPB_{linear} (Figure 1B, left).

Comparison of FPB with FPKM for linear RNA quantification

To evaluate the accuracy of FPB for RNA quantification, commonly-used FPKM values are obtained from the same HISAT2-mapped results. Basically, HISAT2-mapped results are first converted to BAM format by SAMtools [25]. StringTie [18] is then used to calculate transcript expression by FPKM. Since multiple linear RNAs can be produced from a given gene locus, the average FPB value of fragments mapped to all SJ sites in the maximally-expressed linear transcript is used to represent the expression of this gene in the current study (Figure 1B and Figure S2A, S2B).

With the requirement of $FPB_{linear} > 0$ and $FPKM_{linear} > 0$, linear RNA expression, when quantitated by FPB_{linear}, is highly correlated with that quantitated by $FPKM_{linear}$ in the PA1 cell line [22] (**Figure 2**A). Indeed, the value of FPB_{linear} is theoretically equivalent to that of $FPKM_{linear}$ (Figure S2C). Furthermore, FPB_{linear} is highly correlated with the relative expression of 13 linear RNAs as measured by RT-qPCR in PA1 cells (Figure 2B, Table S3). We observe a high correlation between FPB_{linear} and $FPKM_{linear}$ when using different aligners, such as TopHat2 [26] and MapSplice [27], to analyze the ribo$^-$ RNA-seq dataset of PA1 (Figure S3). Finally, FPB_{linear} is also highly correlated with $FPKM_{linear}$ in ENCODE RNA-seq datasets from the 12 human tissues examined (Figure S4 and Table S2). Collectively, these findings reveal that FPB_{linear} is applicable for linear RNA quantification.

Comparison of FPB with FPM for circRNA quantification

As expected, circRNA expression, when quantitated by FPB_{circ}, is highly correlated with that by FPM_{circ} (Figure 2C). Experimentally, FPB_{circ} is also highly correlated with the relative expression of 13 examined circRNAs as measured by RT-qPCR in PA1 cells (Figure 2D, Table S3). The expression of these 13 circRNAs ranges from ~1 to 10 FPB (Figure2D), and their cognate linear RNAs are evaluated above (Figure 2B).

Importantly, compared to commonly-used FPM, FPB is resistant to differences in sequencing lengths and strategies, such as 1×50 *vs.* 1×100 or single-end *vs.* paired-end RNA-seq datasets (Figure 2E). These results are in reasonable agreement with the definitions of FPB and FPM. For example, 1 FPB is equivalent to 0.1 FPM for 1×100 bp single-end RNA-seq datasets (Figure S5A) and to 0.2 FPM for 2×100 bp paired-end RNA-seq datasets (Figure S5B). Importantly, in this scenario, FPB can be used directly for cross-sample comparison regardless of different sequencing lengths and strategies employed.

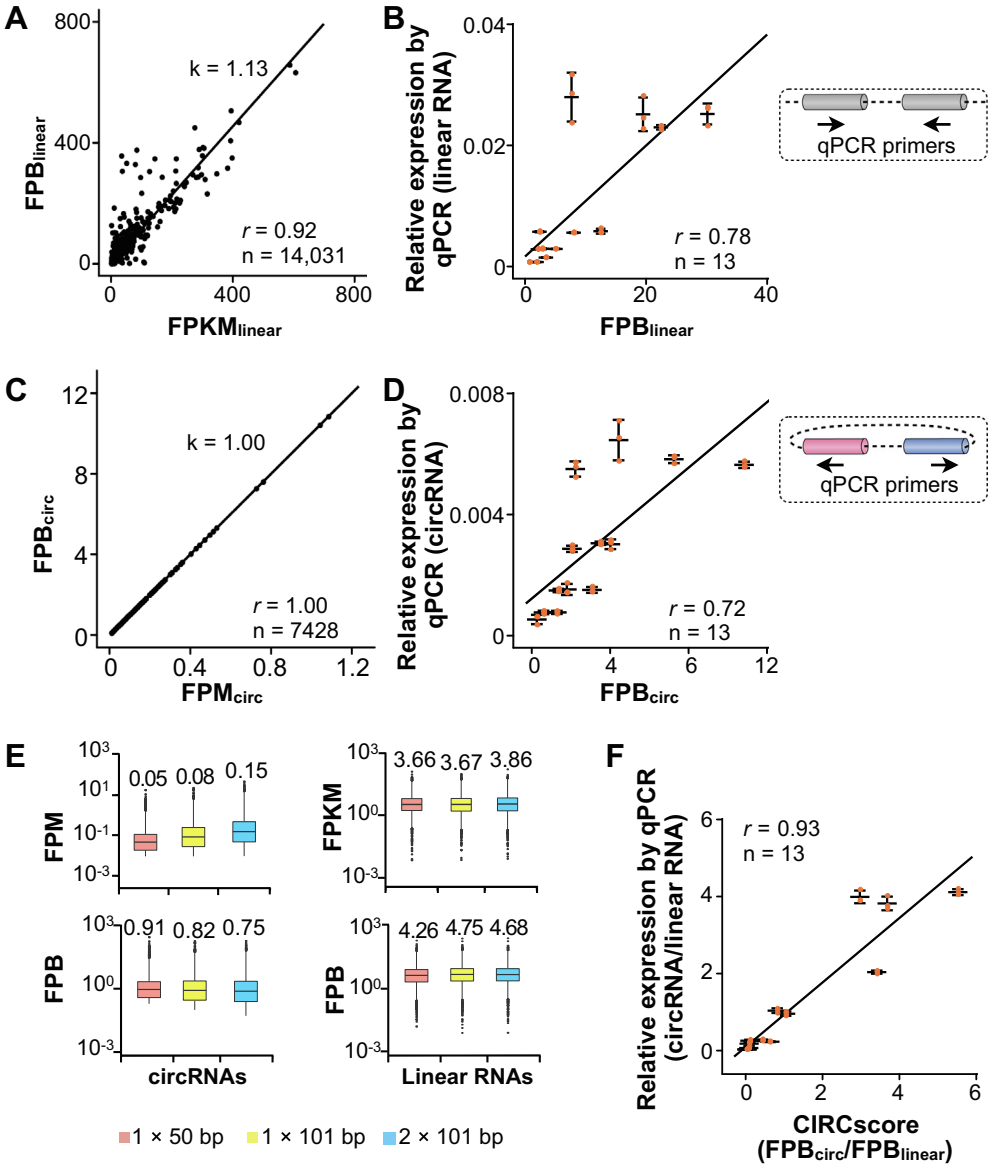

Figure 2 Comparison of FPB with other quantification statistics

A. Comparison of FPB_{linear} and $FPKM_{linear}$. FPB_{linear} is highly correlated with $FPKM_{linear}$ (r = 0.92; PCC) in PA1 cells, under conditions where both FPB_{linear} and $FPKM_{linear}$ > 0. The slope (k) was calculated from linear regression by the lm function in R. **B.** FPB_{linear} is highly correlated with linear RNA relative expression by RT-qPCR. Relative expression of 13 linear RNAs (Table S3) was measured by RT-qPCR, and highly correlated with FPB_{linear} values obtained from PA1 ribo$^-$ RNA-seq (r = 0.78; PCC). Right, qPCR primers convergently spanned linear RNA exons. Relative expression of linear RNA was obtained from three independent experiments using RT-qPCR and presented as 2^{-Ct} relative to those of *ACTB* and shown as mean ± STD. **C.** Comparison of FPB_{circ} and FPM_{circ}. FPB_{circ} is highly correlated with FPM_{circ} (r = 1.00; PCC) in PA1 cells. The slope (k) was calculated from linear regression by the lm function in R. **D.** FPB_{circ} is highly correlated with circRNA expression measured by RT-qPCR. Relative expression of 13 circRNAs (Table S3) was measured by RT-qPCR, and highly correlated with FPB_{circ} values obtained from PA1 ribo$^-$ RNA-seq data (r = 0.72; PCC). Right, qPCR primers divergently spanned circRNA exons. Relative expression of circRNA was obtained from three independent experiments using RT-qPCR and presented as 2^{-Ct} relative to those of *ACTB* and shown as mean ± STD. **E.** FPB is resistant to changes in sequencing lengths or strategies. Two virtual RNA-seq datasets were constructed from original 2 × 101 bp cortex ribo$^-$ RNA-seq dataset (GEO: GSM2072380) to mimic different sequencing lengths and strategies, including 1 × 101 bp (extracting the read1 of the fragment from the paired-end dataset) and 1 × 50 bp (extracting first 50 bp sequence from read1 of the fragment from the paired-end dataset). All three RNA-seq datasets were used for circular and linear RNA quantification to obtain related FPM, FPB, and/or FPKM values. Unlike FPM, FPB largely remained unchanged with different sequencing lengths and strategies. Note, FPKM was set as a control that was also not greatly altered by the changes of sequencing lengths and strategies. The numbers at the top of boxes represent the median values of datas. **F.** CIRCscore is highly correlated with the relative expression of circular *vs.* linear RNA measured by RT-qPCR (r = 0.93; PCC). Relative expression of circRNA or linear RNA was obtained from three independent experiments using RT-qPCR and presented as 2^{-Ct} relative to those of *ACTB* and shown as mean ± STD. PCC, Pearson correlation coefficient.

Evaluation of relative circRNA expression by CIRCscore

Different from unscaled and non-comparable values of $FPKM_{linear}$ (for linear RNA expression) and FPM_{circ} (for circRNA expression), FPB_{linear} for linear RNA measurement is comparable to FPB_{circ} for circRNA measurement. We divide FPB_{circ} by FPB_{linear} to obtain CIRCscore values, by which expression levels of circular and linear RNAs are directly compared in a genome-wide manner. Importantly, the CIRCscore was highly correlated with the experimental comparison of circular *vs.* linear RNA relative expression as measured by RT-qPCR in 13 gene loci from PA1 cells examined in this study (Figure 2F and Table S3), confirming that CIRCscore provides an additional parameter to evaluate circRNA expression normalized by their cognate linear RNA expression background.

We further compared CIRCscore by CLEAR with CLR by another previously reported circRNA quantification toolkit, circtools [20]. Different from CIRCscore that can be achieved by the CLEAR pipeline directly with a simple command, multiple steps are required to obtain CLR with circtools [20]. Importantly, CIRCscore by CLEAR is more accurate than CLR by circtools. For example, CIRCscore of a highly-expressed circRNA, *circCAMSAP-1*, is shown as ~3.65 (**Figure 3**A, blue), which is similar to RT-qPCR validation with relative expression (circRNA/linear RNA) of ~3.82 (Figure 3A, gray). However, CLR calculated by circtools with a customized script is about ~0.80 (Figure 3A, light blue), which is very different from the value derived from RT-qPCR validation. To find out what causes the difference, we performed mapping analysis. It shows that about 126 fragments at the BSJ site of *circCAMSAP*-1 can be identified by the CLEAR-embedded CIRCexplorer2 pipeline, while only 36 fragments are identified by the circtools-embedded DCC pipeline (Figure 3A), suggesting that DCC could be less efficient for BSJ-mapped fragment calling. The comparison of CIRCexplorer2, MapSplice, and DCC confirms that DCC is less efficient for circRNA identification (Figure 3). In the ribo$^-$ RNA-seq dataset from PA1 cells, 356 overlapping circRNAs (with fragments mapped to BSJ ≥ 3) identified by both CIRCexplorer2 and MapSplice failed to be detected by DCC, while only 107 or 99 overlapping circRNAs identified by both CIRCexplorer2 and DCC or MapSplice and DCC were undetected by MapSplice or CIRCexplorer2 (Figure 3B), respectively. Among 787 overlapping circRNAs identified by all three pipelines, DCC was also shown to inefficiently call fragments mapped to BSJs in general (Figure 3C). Of note, CIRCexplorer2 and MapSplice are two reliable pipelines for circRNA profiling [4,28]. Taken together, CIRCscore from the CLEAR pipeline is reliable for circRNA normalization using cognate linear RNA expression as background.

Comparison of FPB and CIRCscore in circRNA analysis

circRNAs are generally co-expressed with their cognate linear RNAs and that sequences of circRNAs largely overlap with those of linear RNAs. Therefore, the advantage of using CIRCscore to quantitate circRNA expression is that it normalizes circRNA expression to the linear RNA expression background. As shown in the PA1 cell line, among those with $FPB_{circ} \geq 1$, some circRNAs with high FPB values have low CIRCscore values (**Figure 4**A, blue), possibly due to the high expression of their cognate linear RNAs (Figure 4B). However, other circRNAs with comparable FPB values have relatively high CIRCscores (Figure 4A, red), as their cognate linear RNAs are expressed at low levels (Figure 4C). This observation suggests variable expression patterns of circular and their cognate linear RNAs from different genomic loci.

We further applied CLEAR to evaluate circRNAs in 12 additional human tissues with both FPB and CIRCscore values (**Figure 5**A and Table S4). Consistent with previous findings [29], circRNAs are more abundant in brain samples than in non-brain tissues. Among all six brain samples examined, circRNAs are more enriched in the cortex, occipital, and diencephalon, but less in the cerebellum, when evaluated by both FPB (Figure 5A, left) and CIRCscore (Figure 5A, right) values. In the six non-brain tissues, circRNAs are enriched in the heart and thyroid at a comparable level to that in the cerebellum. About 10%–20% of circRNAs with $FPB_{circ} \geq 1$ are expressed at a comparable or even higher level than their cognate linear RNAs, as indicated by CIRCscore ≥ 1 (Figure 5A, right), such as in gene loci for *circTPTE2P5* and *circPHF7* (Figure 5B). Taken together, the identification of highly-expressed circRNAs with high FPB_{circ} and CIRCscore values reveals that some gene loci are particularly favorable for circRNA production (Table S4), and such circRNAs warrant subsequent functional studies.

CIRCscore reduces individual differences

Different from FPB, using CIRCscore to evaluate circRNA expression can reduce individual differences that are caused by RNA-seq samples themselves. For example, compared to paired normal samples, circRNA expression evaluated by the FPB_{circ} value is inconsistent in a batch of 20 human HCC samples (GEO: GSE77509) [24]. Some HCC samples appear to have generally low circRNA expression; while others, such as samples #11 and #16, appear to have significantly high circRNA expression (Figure S6A). Consequently, it is hard to distinguish circRNA expression differences between HCC and their paired normal samples using FPB_{circ} in these 20 HCC samples (**Figure 6**A, $P = 0.99$). Strikingly, however, circRNAs are generally lowly expressed in almost all HCC samples when CIRCscore is used to normalize circRNA expression with cognate linear RNA background (Figure 6B, $P = 3.59 \times 10^{-5}$ and Figure S6B). These results suggest that it is important to take cognate linear RNA expression into consideration for circRNA quantification, which can be achieved by the CLEAR pipeline in a genome-wide manner. Taken together, quantification of circRNA expression by CIRCscore helps to eliminate individual differences among paired comparisons, and can therefore be used to decipher the trend of circRNA expression changes under different conditions and for different diseases across RNA-seq datasets.

Discussion

Recently, circRNAs have been widely detected in cell lines and tissues examined by deep sequencing of non-polyadenylated RNAs and using specific computational pipelines for detecting RNA-seq reads/fragments mapped to BSJ sites [16,29,30]. Due to distinct strategies for circular or linear RNA quantification (Figure 1A), computational pipelines for direct circular and

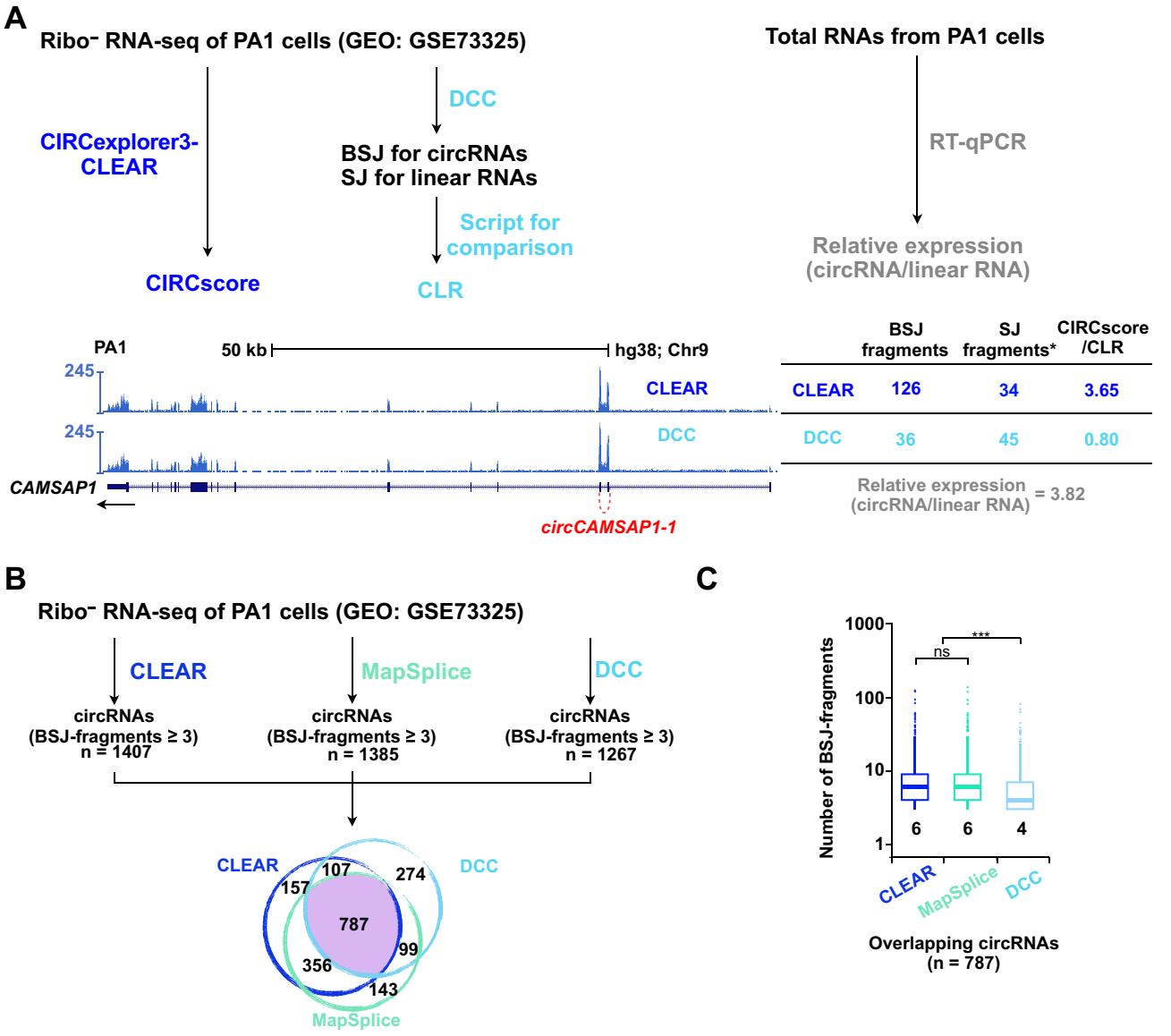

Figure 3 Comparison of circRNA quantification by CIRCexplorer3-CLEAR and other tools

A. Pipelines for calculating CIRCscore by CIRCexplorer3-CLEAR, circular over linear ratios (CLR) by circtools with DCC and a customized script, and validation by RT-qPCR. SJ-mapped fragments, SJ-mapped fragments, CIRCscore, CLR, and RT-qPCR validation are listed. SJ fragments* indicate average number of fragments mapped to all SJ sites in the linear *CAMSAP1* transcript only by CIRCexplorer3-CLEAR or average number of fragments mapped to two SJ sites flanking *circCAMSAP1-1* by DCC. Relative expression of *circCAMSAP1* or linear RNA of *CAMSAP1* was obtained from three independent experiments using RT-qPCR and presented as 2^{-Ct} relative to those of *ACTB* and shown as mean ± STD. **B.** Comparison of circRNAs identified by CIRCexplorer3-CLEAR, MapSplice, and DCC in PA1 cells. The circRNAs with BSJ-fragments ≥ 3 were selected and overlapped among three tools. **C.** CLEAR and MapSplice identify more BSJ-fragments than DCC in 787 overlapping circRNAs with BSJ-fragments ≥ 3 in B. DCC, detect circRNAs from chimeric reads; CLR, circular linear ratio.

linear RNA expression comparison from RNA-seq datasets have remained challenging. In this study, we have developed CLEAR by applying normalized RNA-seq fragments solely mapped to BSJ or canonical SJ sites individually for circular (FPB_{circ}) or cognate linear (FPB_{linear}) RNA quantification (Figure 1B).

The CLEAR pipeline has at least two advantages in circRNA studies. First, the FPB values are highly correlated with canonical FPKMs for linear RNAs and FPMs for circRNAs (Figure 2), which are unlikely affected by RNA-seq strategies, making cross-sample comparisons feasible. Second, direct comparison of circular and cognate linear RNAs with the CIRCscore not only precisely quantitates circRNA expression relative to normalized linear RNA expression background (Figures 4 and 5), but also eliminates possible errors/fluctuations caused by sample preparation/sequencing differences (Figure 6). This reduces inaccuracies for circRNA quantification and subsequent cross-sample comparison. Compared to

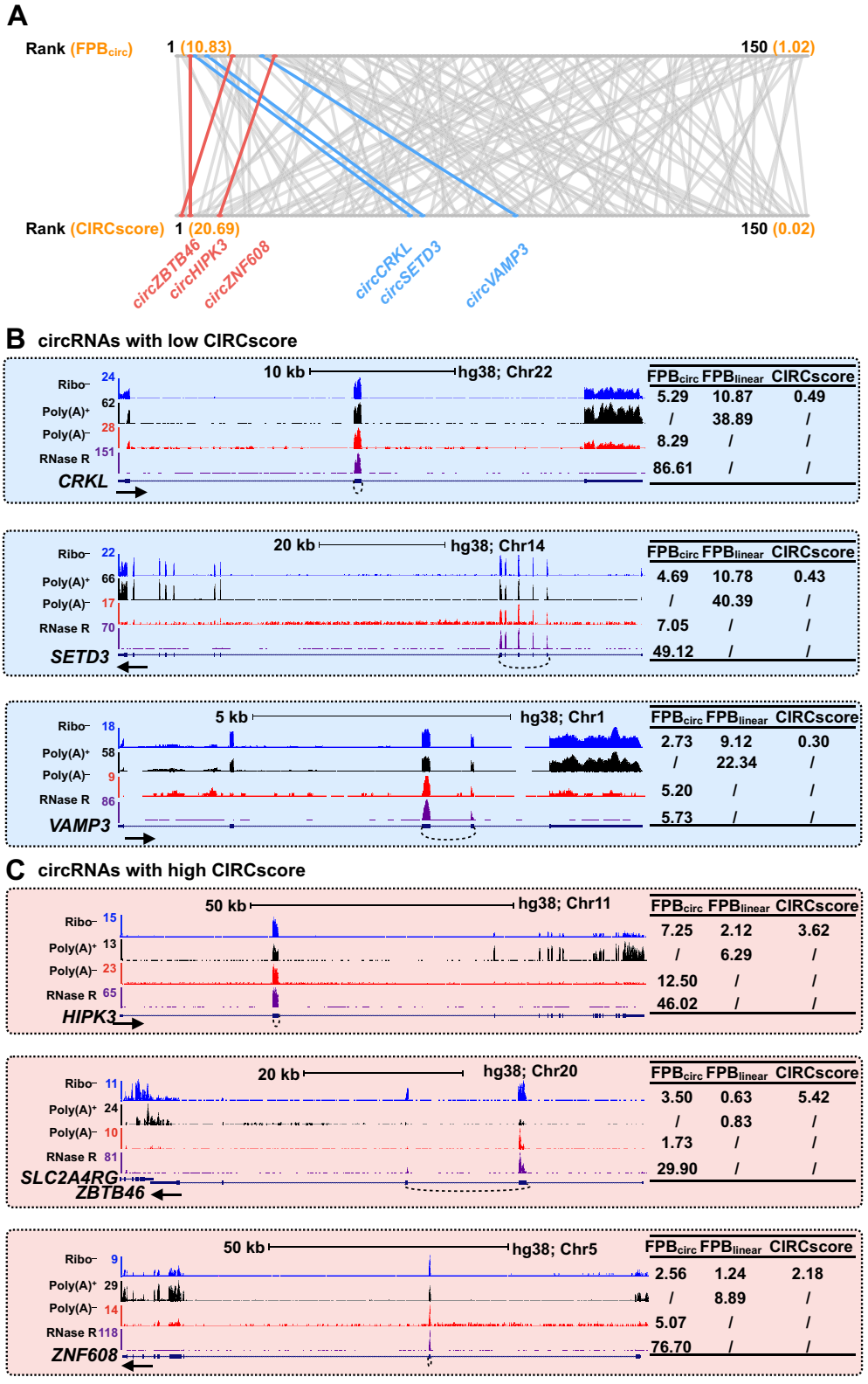

Figure 4 Difference of circRNA quantification by FPB and CIRCscore

A. Difference in circRNA quantitation by FPB_{circ} or CIRCscore in PA1 cells. About 150 circRNAs with $FPB_{circ} \geq 1$ were identified in PA1 cells from published ribo$^-$ RNA-seq (GEO: GSE73325 and GEO: GSE75733) [16,22]. Some circRNAs with high FPB_{circ} values have high CIRCscore values due to the low background of cognate linear RNA expression (in red), while some others have low CIRCscore values due to the high background of cognate linear RNA expression (in blue). **B.** Three highly-expressed circRNAs, *circCRKL*, *circSETD3*, and *circVAMP3*, are co-expressed with their cognate linear RNAs at high levels, indicated by relatively low CIRCscores. The arcs represent the positions of circRNAs. **C.** Three highly-expressed circRNAs, *circHIPK3*, *circZBTB46*, and *circZNF608* are co-expressed with their cognate linear RNAs at low levels, indicated by relatively high CIRCscores. The arcs represent the positions of circRNAs. Gene *SLC2A4RG* is presented in the figure, since it partially overlaps with the circRNA host gene *ZBTB46*.

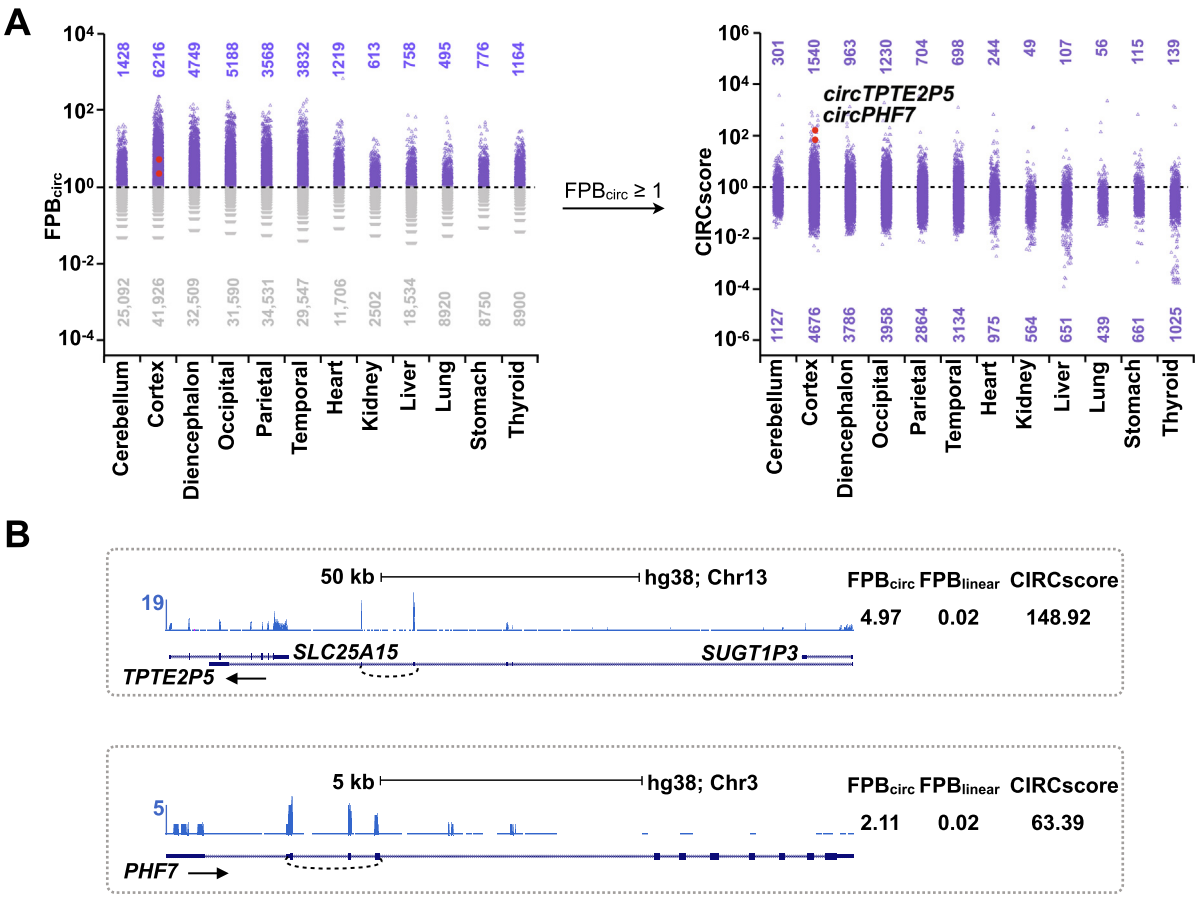

Figure 5 Application of CIRCexplorer3-CLEAR among 12 human tissue samples

A. Quantification of circRNAs by the CIRCexplorer3-CLEAR pipeline. The circRNAs in 12 ENCODE human tissues are quantitated by FPB_{circ} (left), and those with $FPB_{circ} \geq 1$ (blue) are further evaluated by CIRCscore (right). Two representative circRNAs in cortex tissue with medium FPB but high CIRCscore values are highlighted in red. The numbers flanking points represent the numbers of circRNAs. **B.** Visualization of *circTPTE2P5* and *circPHF7* in ENCODE human cortex sample. Of note, *circTPTE2P5* and *circPHF7* are co-expressed with their cognate linear RNAs at low levels, indicated by high CIRCscores. The numbers of circRNAs in different tissues are indicated in the plots. The arcs represent the positions of circRNAs. Genes *SLC25A15* and *SUGT1P3* are presented in the figure due to their overlap with the circRNA host gene *TPTE2P5*.

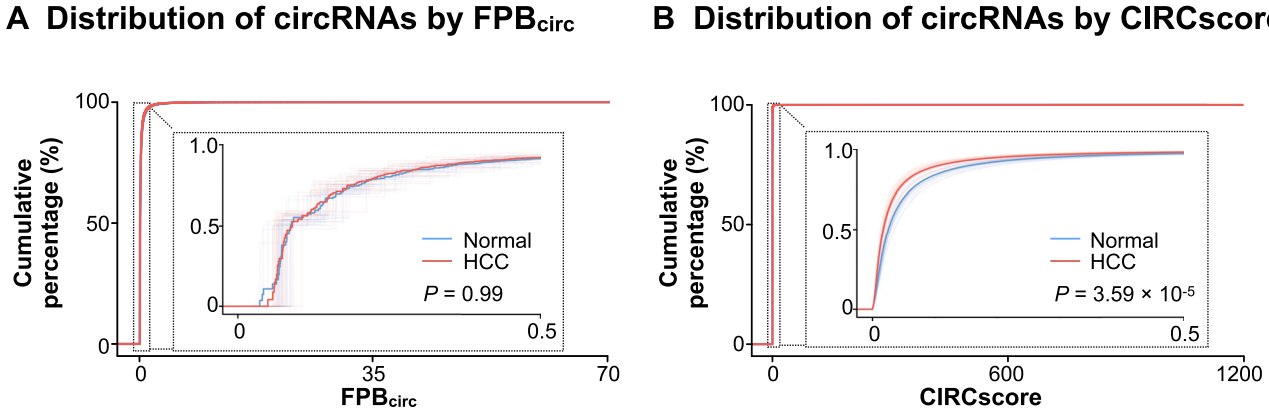

Figure 6 Removal of possible errors/fluctuations and individual differences using CIRCscore quantification

A. Cumulative distribution and comparison of circRNAs in 20 paired human HCC and normal control samples by FPB_{circ}. Thick blue and red lines represent the mixture distribution of FPB_{circ} from 20 normal or HCC samples, respectively. *P* value for statistical significance of difference between two distributions (normal *vs.* HCC) was calculated using two-tailed unpaired Student's *t*-test. **B.** Cumulative distribution and comparison of circRNAs in 20 paired HCC and normal control samples by CIRCscore. Thick blue and red lines represent the mixture distribution of CIRCscores from 20 normal or HCC samples, respectively. *P* value for statistical significance of difference between two distributions (normal *vs.* HCC) was calculated using two-tailed unpaired Student's *t*-test. HCC, hepatocellular carcinoma.

other multi-step methods for circular and linear RNA comparison, such as DCC/circtools, the CLEAR pipeline is efficient (Figure 3), memory-economical (Figure S7), easily performed with a single command (Figure S7), and user-friendly due to the application of reliable CIRCexplorer2 [4,28]. By using cognate linear RNAs as background, CLEAR has the potential to allow users to identify highly expressed circRNAs in different biological settings for subsequent functional studies. This is important, because so far it has often been difficult to identify the circRNAs with the highest expression levels in contexts of interest, or those more highly expressed than their cognate linear RNAs, for functional studies.

It is worth noting that different RNA sequencing strategies have been applied to profile circRNAs, including ribo$^-$, poly(A)$^-$/ribo$^-$, and RNase R-treated RNA-seq datasets (Figure 4). Different from poly(A)$^+$ RNA-seq datasets that are used to detect polyadenylated cognate linear RNAs, all three types of non-polyadenylated RNA-seq can be used to determine circRNA expression by FPB. However, only ribo$^-$ RNA-seq datasets that profile both polyadenylated linear and non-polyadenylated circular RNAs in parallel are suitable for direct circular and linear RNA expression comparison by CIRCscore (Figure 4). In contrast, in poly(A)$^-$/ribo$^-$, and RNase R-treated RNA-seq datasets, polyadenylated linear RNAs are largely depleted, which is unsuitable for accurate linear RNA quantification and subsequent CIRCscore evaluation.

Taken together, the CLEAR pipeline provides a comprehensive way to quantitatively evaluate circRNA expression across samples and to identify highly expressed circRNAs with low linear RNA expression background.

Availability

The CIRCexplorer3-CLEAR pipeline and its application can be downloaded from https://github.com/YangLab/CLEAR.

Authors' contributions

LY conceived and designed the project. XKM, MRW, and RD performed computational analyses under the supervision by LY. CXL performed experiments under the supervision by LLC. LY, LLC, and GGC wrote the paper with input from XKM and MRW. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences, China (Grant No. XDB19020104), the National Natural Science Foundation of China (Grant Nos. 31730111, 31925011, and 91940306), and the Howard Hughes Medical Institute International Program, the United States (Grant No. 55008728).

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2019.11.004.

References

[1] Chen LL. The biogenesis and emerging roles of circular RNAs. Nat Rev Mol Cell Biol 2016;17:205–11.
[2] Li X, Yang L, Chen LL. The biogenesis, functions, and challenges of circular RNAs. Mol Cell 2018;71:428–42.
[3] Wilusz JE. A 360 degree view of circular RNAs: from biogenesis to functions. Wiley Interdiscip Rev RNA 2018;9:e1478.
[4] Hansen TB, Veno MT, Damgaard CK, Kjems J. Comparison of circular RNA prediction tools. Nucleic Acids Res 2016;44:e58.
[5] Gao Y, Zhao F. Computational strategies for exploring circular RNAs. Trends Genet 2018;34:389–400.
[6] Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature 2013;495:333–8.
[7] Zhang XO, Wang HB, Zhang Y, Lu X, Chen LL, Yang L. Complementary sequence-mediated exon circularization. Cell 2014;159:134–47.
[8] Guo JU, Agarwal V, Guo H, Bartel DP. Expanded identification and characterization of mammalian circular RNAs. Genome Biol 2014;15:409.
[9] Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, et al. Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA 2013;19:141–57.
[10] Liang D, Wilusz JE. Short intronic repeat sequences facilitate circular RNA production. Genes Dev 2014;28:2233–47.
[11] Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, et al. Natural RNA circles function as efficient microRNA sponges. Nature 2013;495:384–8.
[12] Piwecka M, Glazar P, Hernandez-Miranda LR, Memczak S, Wolf SA, Rybak-Wolf A, et al. Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. Science 2017;357:eaam8526.
[13] Kleaveland B, Shi CY, Stefano J, Bartel DP. A network of noncoding regulatory RNAs acts in the mammalian brain. Cell 2018;174:350–62.e17.
[14] Liu CX, Li X, Nan F, Jiang S, Gao X, Guo SK, et al. Structure and degradation of circular RNAs regulate PKR activation in innate immunity. Cell 2019;177:865–80.e21.
[15] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010;28:511–5.
[16] Zhang XO, Dong R, Zhang Y, Zhang JL, Luo Z, Zhang J, et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res 2016;26:1277–87.
[17] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 2015;12:357–60.
[18] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;33:290–5.
[19] Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 2011;12:R72.
[20] Jakobi T, Uvarovskii A, Dieterich C. circtools-a one-stop software solution for circular RNA research. Bioinformatics 2019;35:2326–8.
[21] Cheng J, Metge F, Dieterich C. Specific identification and quantification of circular RNAs from sequencing data. Bioinformatics 2016;32:1094–6.

[22] Zhang Y, Xue W, Li X, Zhang J, Chen S, Zhang JL, et al. The biogenesis of nascent circular RNAs. Cell Rep 2016;15:611–24.
[23] Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, et al. The 1000 Genomes Project: data management and community access. Nat Methods 2012;9:459–62.
[24] Yang Y, Chen L, Gu J, Zhang H, Yuan J, Lian Q, et al. Recurrently deregulated lncRNAs in hepatocellular carcinoma. Nat Commun 2017;8:14421.
[25] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.
[26] Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009;25:1105–11.
[27] Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 2010;38:e178.
[28] Hansen TB. Improved circRNA identification by combining prediction algorithms. Front Cell Dev Biol 2018;6:20.
[29] Rybak-Wolf A, Stottmeister C, Glazar P, Jens M, Pino N, Giusti S, et al. Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. Mol Cell 2015;58:870–85.
[30] Salzman J, Chen RE, Olsen MN, Wang PL, Brown PO. Cell-type specific features of circular RNA expression. PLoS Genet 2013;9:e1003777.