Trends in Genetics

CellPress
REVIEWS

Review

Computational Strategies for Exploring Circular RNAs

Yuan Gao[1] and Fangqing Zhao[2,*]

Recent studies have demonstrated that circular RNAs (circRNAs) are ubiquitous and have diverse functions and mechanisms of biogenesis. In these studies, computational profiling of circRNAs has been prevalently used as an indispensable method to provide high-throughput approaches to detect and analyze circRNAs. However, without an overall understanding of the underlying strategies, these computational methods may not be appropriately selected or used for a specific research purpose, and some misconceptions may result in biases in the analyses. In this review we attempt to illustrate the key steps and summarize tradeoff of different strategies, covering all popular algorithms for circRNA detection and various downstream analyses. We also clarify some common misconceptions and put emphasis on the fields of application for these computational methods.

Highlights

circRNAs are a large class of RNAs that are diverse in gene source, exon composition, localization, biogenesis, and function.

Annotation-independent detection algorithms can be used in a wide range of organisms, but often need more cautious strategies to ensure reliability.

Most of the detection methods are optimized for their designated aligners, and these can be further divided into splice-aware aligners and versatile read-mappers.

Paired end sequencing provides more information to reduce false positives for detection methods that adopt filtering based on paired end mapping.

Based on detection algorithms, promising computational methods are emerging for downstream analyses of circRNAs.

New computational methods to reconstruct full-length of circRNAs and quantify their expression are urgently needed.

Overview of circRNAs

Circular RNA (circRNA; see Glossary) transcripts are a special class of RNAs that can be produced by a covalent linkage between the 5′ and 3′ ends of an RNA molecule [1]. Such closed circles are distinct from other RNA transcripts such as mRNAs that typically contain a 3′ poly(A) tail and a 5′ cap. The low expression levels of circRNAs have caused misinterpretation in early studies because they were assumed to be artefacts or mis-splicing products due to technical limitations [2]. Thanks to the development of high-throughput sequencing technology and the efforts of many RNA biologists, the presence of circRNAs has been accepted and the study of circRNAs has attracted more and more attention starting about 5 years ago. circRNAs are not only found in classic model mammalian organisms such as human and mouse, but are also present in almost all of the main eukaryotic clades, including insects, plants, and fungi [1,3–6].

In addition to their ubiquity, circRNAs are also more diverse than expected. First, circRNAs can be expressed from a large fraction of genes, as well as from the antisense strand of some genes and even from intergenic regions [7–9]. Second, they have very different lengths, mainly ranging from 100 bp to 4 kb [10], and contain single or multiple exons [9], sometimes with retained introns in their sequences [11]. Third, they are present in different cell lines (such as cancer and non-cancer) and tissues [7,9,12,13], different subcellular compartments (including nucleus and cytosol) [11,14], as well as in extracellular exosomes [15]. Fourth, different mechanisms contribute to their biogenesis; these include intron pairing-driven (Figure 1A) [16–19], lariat-driven [20], and RNA-binding protein-driven circularization [21]. As to their function, the most commonly proposed function of circRNAs is that they act as microRNA (miRNA) sponges that can indirectly regulate the expression of multiple genes [8,22], however, some circRNAs have recently been reported to function as transcription competitors or enhancers for parental gene and translatable transcripts [11,23,24].

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Laboratory of Computational genomics, Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China

*Correspondence: zhfq@biols.ac.cn (F. Zhao).

In studies exploring the nature of these ubiquitous and diverse circRNAs, computational methods play important roles because of their convenience in the analysis of high-throughput RNA-seq data and their advantages in profiling circRNA expression. Currently more than 10 computational methods for circRNA detection have been released and published [6–8,19,25–32]. According to a recent third-party evaluation, CIRI [7,27], CIRCexplorer [19], and KNIFE [31] show more balanced performance than do other methods (Figure 2), although all 11 methods have different advantages and shortcomings regarding sensitivity, precision, and computational cost [33]. Such differences are largely due to the different strategies adopted in these methods (Table 1). Other downstream computational methods are all based on the primary detection results and are mainly related to expression quantification, differential expression analysis, and transcript construction [14,25,34–37].

Genome Reference, Annotation, and GT–AG Splicing Signals in circRNA Detection

A reference genome is necessary for all detection algorithms, but can be used in different ways in the detection workflow. The most common use involves direct alignment of all sequencing reads against the reference genome. Most of the detection algorithms, such as find_circ [8], CIRCexplorer [19], CIRI [7,27], and UROBORUS [30], can be classified into this category. Because circRNAs are different from other RNAs in view of their circularity, an obvious feature that can be captured from an alignment is the circle junction, termed the **back-spliced junction** (BSJ). In contrast to **forward-spliced junctions** (FSJs) in mRNAs that generate sequencing reads that are collinearly aligned on the genome, reads spanning BSJs are split into segments and are aligned to the reference sequence in reverse order (Figure 1B). Therefore, detection algorithms in this category can be termed **split-alignment-based approaches**. For other algorithms, such as KNIFE [31] and NCLscan [26], the reference genome is combined with the corresponding **genome annotation** to build pseudo-sequences around putative BSJs in the first few steps (Figure 1C). Subsequent steps are then focused on the complete alignment of sequencing reads against such pseudo-sequences to recognize BSJ reads. In addition to a BSJ pseudo-sequence database, KNIFE also constructed a FSJ sequence database according to the annotation to remove candidate reads with high-score alignment in both databases [31]. Detection algorithms in this category may be termed **pseudoreference-based approaches**.

However, it should be mentioned that the dependency of a detection algorithm on genome annotation does not completely correlate with the above-stated classification based on a reference genome. In addition to pseudoreference-based approaches such as NCLscan [26] that cannot be used without annotation, several split-alignment-based approaches such as UROBORUS [30] and CIRCexplorer [19] also require annotation to remove false positives during the filtering step. Although KNIFE was developed on the basis of annotated exons, it adds *de novo* detection as a remedy [31], except that such *de novo* detection cannot provide exact breakpoints, according to a recent evaluation [33]. By contrast, find_circ [8] and CIRI [7,27] are wholly annotation-independent, and thus they can be used for circRNA detection in organisms with no complete annotation or with only a draft genome.

The use of annotation is often beneficial. For example, a comprehensive evaluation of RNA-seq aligners recently demonstrated that annotation can help to increase the sensitivity for junction identification versus *de novo* detection [38]. For this reason some *de novo* algorithms such as CIRI [7,27] can also accept annotation as complementary information to facilitate comprehensive detection. Another potential advantage of using annotation is that the **false discovery rate** (FDR) can be better controlled because there are fewer candidates. This also brings about

Glossary

Back-spliced junction (BSJ): a junction between two related exons in the opposite order relative to their positions on the reference sequence. A circular junction is a typical type of BSJ.
BSJ read pair: a BSJ read and its mate in paired end sequencing, the former of which indicates a sequencing read spanning a BSJ.
Circular RNA (circRNA): a type of RNA molecule that forms a covalently closed loop.
circRNA exon (cirexon): an exon retained in a circRNA transcript, which may or may not be included in any mature linear transcripts.
False discovery rate (FDR): the proportion of false positives in all detected candidates.
Forward-spliced junction (FSJ): a junction between two related exons in the same order relative to their positions on the reference sequence. Splice junctions in mRNA transcripts are a typical type of FSJ.
Genome annotation: a note on the reference genome that records boundaries of known or predicted gene structures such as exons and protein-coding regions.
Split-alignment-based approach: circRNA detection based on split alignment of sequencing reads against the reference genome.
Paired end mapping (PEM): mapping information of a read and its mate in paired end sequencing; can be more informative than single-end mapping because of the inherent link between the two reads.
Pseudoreference-based approach: circRNA detection based on sequencing read alignment against the pseudo-sequence by combining annotated exons.
Unbalanced BSJ read: a BSJ read where one segment flanking the BSJ is much shorter than the other segment.

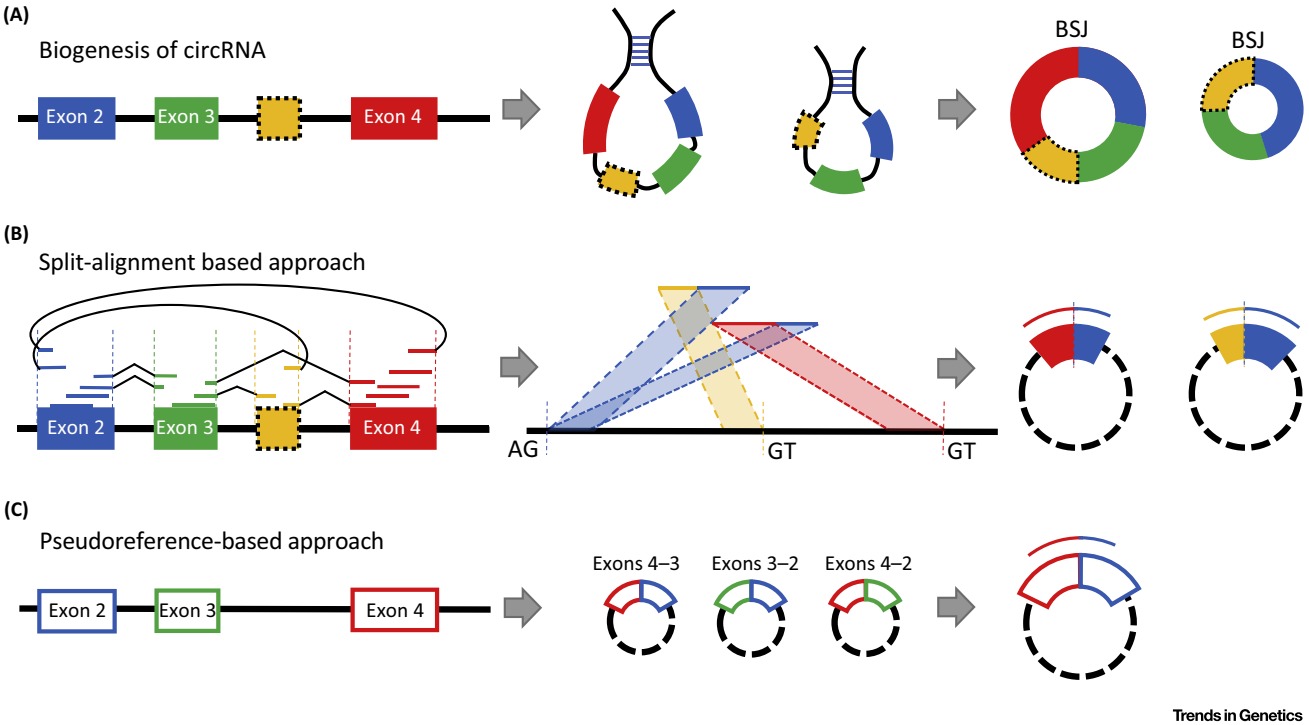

Figure 1. Biogenesis of Circular RNA (circRNA), and Two Types of circRNA Detection Algorithms. (A) Biogenesis of two circRNAs from the same gene (intron pairing-driven mechanism). Three mRNA exons and one circRNA-specific exon (yellow) are involved. (B) Split-alignment-based approach for circRNA detection. (C) Pseudoreference-based approach for circRNA detection. Because annotation is necessary for construction of pseudo-circRNA sequence, a circRNA-specific exon that has not been recorded in the annotation cannot be detected. Abbreviation: BSJ, back-spliced junction.

less restrictions on the detection of non-canonical junctions. Because non-canonical (e.g., AT–AC) sites only comprise a very small proportion of all splicing junctions, most *de novo* algorithms such as find_circ [8] choose to use GT–AG splicing signals as one of the default filters for the control of false positives and the accurate localization of splice junctions, and for this reason they are unable to detect non-canonical sites. Such a dilemma can be mostly resolved by a combination of *de novo* strategy with flexible annotation usage, where the vast majority of canonical junctions can be detected with FDR control by virtue of GT–AG splicing signals. Other non-canonical junctions can then be detected when the annotation of the reference genome is provided [7].

Versatile Read-Mapper or Specified Splice-Aware Aligner

Alignment of transcriptomic reads is involved in the identification of BSJ reads for both split-alignment-based and pseudoreference-based approaches. Some algorithms, such as find_circ [8], CIRI [7,27], and KNIFE [31], choose read-mappers that are widely used in most reference-based DNA/RNA sequence analyses, such as Bowtie [39,40] and BWA [41,42]. Although the choice is explicit for KNIFE [31] for the end-to-end alignment of BSJ reads against a constructed pseudo-template, it will often necessitate more exquisite designs in split-alignment-based approaches such as find_circ [8] and CIRI [7,27] to tackle local alignment boundary ambiguity, spurious alignments, and low-complexity sequences. An easier way is to use splice-aware aligners which were developed specifically for RNA-seq reads across intron-sized gaps on genome references, such as TopHat [43,44], STAR [45], and Novoalign. Detection algorithms, such as circRNA_finder [6], CIRCexplorer [19], and DCC [25], depend

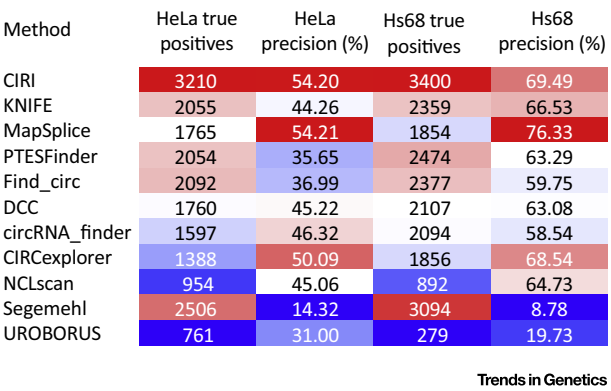

Figure 2. Performance Comparison Among Eleven Circular RNA (circRNA) Detection Methods by Third-Party Evaluation.

Table 1. Summary of Eleven circRNA Detection Methods

Method	Category	Annotation-dependent	Mapper type	Mapper(s)	Other characteristics
CIRCexplorer	Split-alignment-based	Yes	Splice-aware	TopHat/STAR	TopHat-fusion/STAR are required for non-collinearity detection
circRNA_finder	Split-alignment-based	No	Splice-aware	STAR	GT–AG splice sites PEM filtering
CIRI	Split-alignment-based	No	Versatile	BWA-MEM	GT–AG splice sites combined with flexible annotation usage Stringent PEM filtering Recovery of unbalanced BSJ reads Multiple seed matching
DCC	Split-alignment-based	Yes	Splice-aware	STAR	GT–AG splice sites PEM filtering
find_circ	Split-alignment-based	No	Versatile	Bowtie2	GT–AG splice sites No PEM filtering Two 20 bp anchors for non-collinearity detection
KNIFE	Pseudoreference-based	Yes	Versatile	Bowtie, Bowtie2	Binary logistic regression model based on PEM *De novo* detection as remedy (exact breakpoint not detectable)
MapSplice	Split-alignment-based	Yes	Versatile[a]	Bowtie[a]	CircRNA detection is embedded in its alignment algorithm
NCLscan	Pseudoreference-based	Yes	Mixed	BWA, BLAT, Novoalign	Fusion and *trans*-spliced transcript detection in addition to circRNA detection
PTESFinder	Pseudoreference-based	Yes	Versatile	Bowtie, Bowtie2	No PEM filtering Two 20 bp anchors for non-collinearity detection
UROBORUS	Split-alignment-based	Yes	Mixed	Bowtie/Bowtie2, TopHat	PEM filtering Two 20 bp anchors for non-collinearity detection Recovery of unbalanced BSJ reads
segemehl	Split-alignment-based	No	Versatile	*Per se*	No PEM filtering Few filters adopted

[a]MapSplice is a splice-aware aligner but invokes Bowtie as the underlying read-mapper.

on this type of aligner. An obvious advantage is that these aligners are all optimized according to current knowledge of eukaryotic transcription, and are thus usually much more convenient than the versatile read-mappers mentioned above. However, splice-aware aligners are often less flexible than versatile mappers which have developed both end-to-end and local alignment, as well as more diverse reporting modes and arguments [39–42]. In addition, most of the splice-aware aligners are also dependent on versatile read-mappers.

Some detection algorithms may require multiple aligners. For example, NCLscan [26] requires BWA [42], BLAT [46], and Novoalign, while UROBORUS [30] takes advantage of Bowtie [39,40] and TopHat [43,44]. A possible benefit is that such integrated analyses may combine the advantages of different aligners, but they also unavoidably increase dependencies for installation and usage [33]. Two detection algorithms can align reads by themselves: segemehl [28] is itself an algorithm for split mapping, and thus its circRNA detection has no dependency on any other aligner; MapSplice [32] is a splice-aware aligner which invokes Bowtie [40] as the underlying read-mapper.

Two misconceptions about aligners need to be clarified. It is a common misunderstanding that the required aligner can be replaced by another for a given detection algorithm. Almost all detection algorithms are optimized for their designated aligners. Although the SAM formatted alignments generated by different aligners are similar, and are compatible with some analyses such as sequence depth calculation, circRNA detection algorithms usually require very detailed information such as alternative mapping location, and mapping length, quality, and score, which are often provided in different formats or rules. For example, the CIGAR (concise idiosyncratic gapped alignment report) strings in the SAM files generated by TopHat [43,44] and BWA [42] have different patterns, and the maximum mapping quality is 42 in Bowtie2 [39] but 60 in BWA-MEM [41]. The above unique characteristics of circRNA detection compared to other analyses also help to clarify another misconception – that the aligner is believed to take over the major work of circRNA detection, and detection performance is mostly influenced by the aligner rather than by the detection algorithm itself. Although the aligner can facilitate detection by providing accurate and comprehensive mapping information, a detection algorithm aims to consider systematic strategies to deal with challenges unique to circRNAs, including low abundance, possible biased detection, and multiple sources of false positives [47], and such strategies are highly likely to influence the performance of the entire pipeline. For example, although DCC [25] and circRNA_finder [6] are both designed based on the alignment records generated by STAR [45], they showed very different detection performance in regard to sensitivity and precision [27,33].

Identification and Recovery of BSJ Reads

How to identify BSJ reads from an alignment is a prominent challenge in a detection pipeline. Pseudoreference-based approaches usually put emphasis on the upstream construction of putative circRNA sequences, as described above, to simplify this challenging process. In split-alignment-based approaches, however, alignment continuity and collinearity obtained from raw alignment information are used to determine potential BSJ reads, and this may involve multiple steps. In brief, most algorithms contain preliminary filtering step in which all reads are first aligned against the reference genome. For example, find_circ [8], UROBORUS [30], and CIRCexplorer [19] collect reads that cannot be continuously or collinearly aligned to the reference so as to generate BSJ read candidates (Figure 3A). To further validate the non-collinearity of these reads, find_circ [8] and UROBORUS [30] extract 20 bp anchors from both ends, and then align these anchors individually against the reference. Such alignments will be extended if they are found to be on the same strand, but in reverse order, to obtain a specific

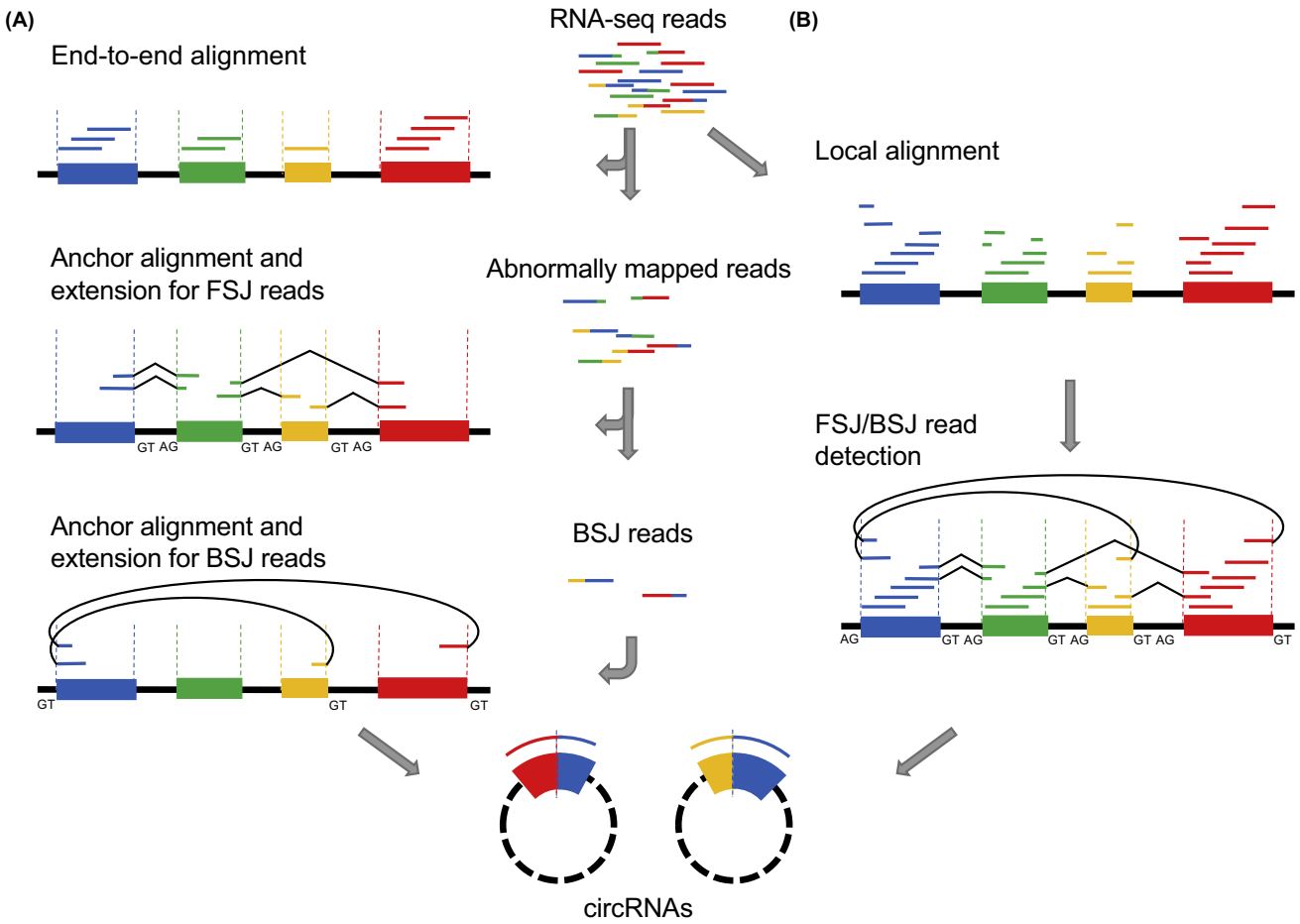

Figure 3. Strategies for Back-Spliced Junction (BSJ) Read Identification. (A) A typical pipeline with preliminary filtering for continuity and collinearity. RNA-seq reads are filtered after each alignment step for the final detection of BSJ reads. (B) A typical pipeline with only one alignment step. Continuity and collinearity are determined based on the local alignments of all sequencing reads. Abbreviations: BSJ, back-spliced junction; FSJ, forward-spliced junction.

location for a potential BSJ. As another solution, CIRCexplorer takes advantage of TopHat-fusion [48] to determine the collinearity of these unmapped reads [19], and this largely simplifies its pipeline. In contrast to the above stepwise determination, BSJ reads are identified in CIRI based on local alignment records, including all reads or read segments continuously aligned against the reference [7] (Figure 3B). Such a strategy makes preliminary filtering unnecessary, and also facilitates the detection of BSJ reads with more than two segments.

Another important strategy that influences the sensitivity of circRNA detection and the accuracy of downstream quantification analysis is the recovery of **unbalanced BSJ reads**. Such BSJ reads have a very short segment that cannot be captured by a reasonable anchor length (e.g., 20 bp set in UROBORUS [30] and 19 bp in BWA-MEM by default [41]). The detection of unbalanced BSJ reads is challenging because the short segment is usually not sufficiently informative to accurately locate on the reference. As a solution, CIRI was designed to detect unbalanced BSJ reads on the basis of balanced BSJ reads by dynamic programming alignments to control the FDR [7]. However, it has been demonstrated that multiple circRNA

transcripts may originate from the same BSJ [14], and thus a detected balanced BSJ read sometimes may not be a perfect reference for an unbalanced BSJ read. This step has been improved by maximum likelihood estimation based on multiple seeds from the unbalanced BSJ read [27]. In brief, two putative genomic regions of the short segment corresponding to BSJ and FSJ are first determined according to the location of the long segment, and this restricts the search range compared to the whole genome. The short segment is then divided into multiple seeds that are used to count matches in the two regions. By comparing the matching counts and relative positions of seeds, the most likely location of the short segment can be determined. This recovery of unbalanced BSJs was found to further improve precision while retaining high sensitivity for circRNAs with low abundance [27].

Paired End Mapping and Other Filtering Strategies

Paired end reads and their mapping information proved to be useful with the development of RNA-seq alignment and assembly algorithms [38,44,49,50]. In circRNA detection, the use of paired end reads exhibited a similar trend. Two relatively early algorithms, find_circ [8] and segemehl [28], did not take **paired end mapping** (PEM) information into account. Owing to the prevalence of paired end sequencing, most of the algorithms developed later have utilized such information to reduce false positives from spurious mapping. However, the specific patterns and extents of PEM information used in these tools are different. For example, UROBORUS sets a loose requirement that **BSJ read pairs** should be aligned to the same chromosome but in the opposite orientation [30]. In DCC [25], BSJ read pairs are considered to be valid only if they are consistent with putative templates of the corresponding circRNAs (Figure 4A).

To achieve a low FDR for circRNA detection, the above filtering strategies are necessary but not sufficient. In addition to the false positive BSJ reads that can be filtered out as described above, other candidate BSJ reads can still originate from linear RNAs if both orientation and library insert length by chance coincide with the genomic region of a putative circRNA (Figure 4B). If these false BSJ reads corresponding to putative circRNA regions accumulate from highly abundant transcripts, false detections may be generated. A more stringent strategy has been adopted by algorithms such as CIRI [27] to evaluate the percentage of consistent BSJ read pairs in all reads associated with the same candidate BSJ, because approximately one half of false BSJ read pairs should violate the putative circRNA region given the two strand sequencing of RNA-seq libraries (Figure 4C). As another solution, KNIFE collects these inconsistent BSJ read pairs as the negative training set, and uses forward-spliced reads of canonical linear isoforms as the positive training set [31]. Three features – alignment score, mapping quality, and mapping length across the junctions – from both sets are extracted to train a binary logistic regression model, which is finally used to distinguish true and false BSJ reads from candidate read pairs within putative circRNA regions.

The above three features used in KNIFE [31] are often adopted as filters in other algorithms. For example, find_circ [8] requires at most two mismatches during the extension of one anchor as a control for alignment score, and CIRI [27] has a parameter (–U) to set threshold of mapping quality for one of the segments. As to the mapping length across BSJ, split-alignment-based approaches often set a minimum for each segment. When one segment is shorter than the anchor length, this can lead to unbalanced BSJ read detection. In addition to these three filtering features, other filters can be employed for circRNA detection. One interesting filter specific to split-alignment-based approaches is the maximum genomic range of a putative circRNA. UROBORUS [30] has the largest cutoff of 3 million bp, whereas in some other algorithms such as find_circ [8] this parameter is set as short as 100 kb. A large cutoff may result in a relatively high FDR, whereas a small cutoff may be insufficient for the detection of

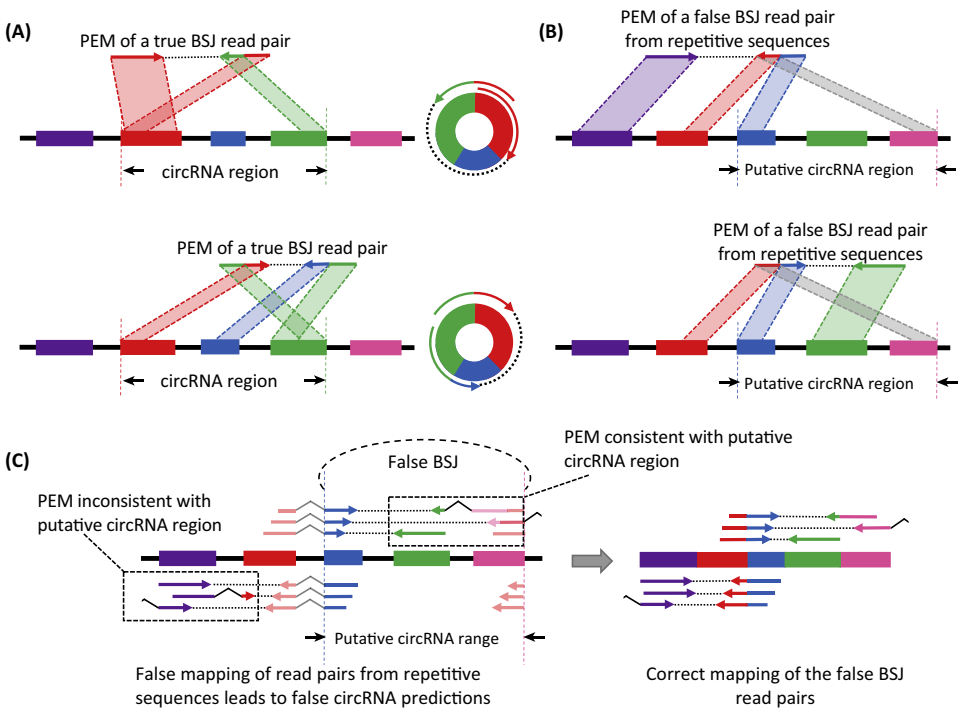

Figure 4. Paired End Mapping (PEM) Information Can Be Used To Filter False Positives in Circular RNA (circRNA) Detection. (A) All alignments of read pairs associated with a true Back-Spliced Junction (BSJ) should be located within the genomic region of putative circRNA on the reference genome. (B) Although the putative circRNA region can be used to remove some false BSJ reads, it is not sufficiently efficient to remove them all. When both orientation and library insert length coincide with the genomic region of a putative circRNA by chance, paired end mapping (PEM) of false BSJ reads can be consistent with putative circRNA region. The red and pink exons share similar sequences in this example, which results in spurious alignment and the false BSJ reads. (C) False detections may be generated if PEM filtering is individually set for each candidate BSJ read. A more stringent method should consider the percentage of consistent BSJ read pairs for a given BSJ. When a putative BSJ has a fraction of BSJ reads that violate the putative circRNA region, it should be discarded.

long-range circRNAs. Another important filtering feature is the existence of duplicated sequences associated with putative circRNAs. Duplicated sequences that are often contained in homologous regions or genes may confuse algorithms designed to detect FSJs and BSJs, and are one of the main sources of false detection of circRNAs [7]. Indeed, the majority of false positives that the logistic regression model in KNIFE aims to remove are those originating from sequencing errors combined with homologous exons [31]. An alternative solution to this problem is to perform targeted analysis of candidate circRNAs associated with duplicated sequences. By applying maximum likelihood estimation based on matching seeds in two adjacent regions, interference by duplicated sequences on circRNA detection can be minimized. Such a strategy was demonstrated to effectively remove the vast majority of false positives from duplicated regions in human chromosomes and unplaced contigs [27].

Computational Methods for circRNA Downstream Analyses

In addition to the circRNA detection algorithms discussed above, other computational approaches have been designed based on detecting circRNAs for diverse analyses, including differential expression analysis, accurate quantification, alternative splicing detection, full-length assembly, and visualization (Table 2).

Table 2. Summary of Computational Methods for Downstream Analyses of circRNAs

Method	Function	Language	Input requirement
CircPro	Protein-coding potential estimation	Perl	FASTA formatted reference; FASTQ formatted sequencing reads; GTF/GFF3 formatted annotation
circTest	Differential expression test	R	circRNA list with read-count; parental gene list with read-count
CircView	Visualization	Java	circRNA list (optional: MRE/RBP binding site list)
CIRI-AS	Internal structure and alternative splicing detection	Perl	circRNA list; FASTA formatted reference; SAM formatted alignment (optional: GTF formatted annotation)
FUCHS	Internal exon skipping detection, motif enrichment, and miRNA seed analysis	Python	circRNA list; BED formatted annotation; BAM/SAM formatted alignment
Sailfish-cir	Precise quantification	Python	circRNA list; FASTA formatted reference; FASTA/FASTQ formatted sequencing reads; GTF formatted annotation

Previous studies quantified the expression of different circRNAs according to their detected BSJ read-counts [6,13,51]. Although this is a simple and direct way, it may suffer from low precision owing to the rareness of BSJ reads and the resulting high variation. An effective method used for linear RNA quantification in RNA-seq data involves an expectation-maximization (EM) procedure to iterate the assignment of reads and abundance estimation for transcripts [50,52,53]. By transforming circRNAs to pseudo-linear transcripts, sailfish-cir [37] was developed to quantify circRNA expression based on an improved EM procedure [54]. Application of sailfish-cir to 59 circRNAs in 11 rRNA-depleted datasets showed a stronger correlation to their qRT-PCR threshold cycle values compared to direct quantification based on BSJ read-counts [37]. This quantification method mainly benefits from employing all sequencing reads from circRNAs rather than only BSJ reads. They also found that three factors – expression level, length, and relative abundance of circRNAs – may affect the performance of sailfish-cir [37].

In addition to quantification, relative abundance comparisons for circRNAs constitute another difficult task owing to the rareness of BSJ reads. As a solution, a statistical method, circTest, was proposed in the same study of DCC [25]. This method is based on the assumption that all junction reads can be attributed to circRNAs or linear RNAs. When comparing the relative abundance of a circRNA in different samples or treatments, circTest assumes that BSJ read-counts in replicates of the same treatment follow the same binomial distribution, whose parameter of probability θ (the relative abundance of the circRNA in all related transcripts) follows a β distribution with a mean of $\alpha/(\alpha + \beta)$. By selecting between hypotheses that multiple treatments have different means or not according to a likelihood ratio test, the potential differential expression of the circRNA between treatments can be determined. Such a statistical test was applied with DCC to rRNA-depleted RNA-seq data of fruit fly. It was demonstrated that 72 of 116 circRNAs showed significant differential abundances across three developmental stages of brain [25]. The above two methods, circTest [25] and sailfish-cir [37], represent two typical applications for quantification analysis of circRNAs. circTest [25] places emphasis on BSJ read-counts and their distribution in samples, whereas sailfish-cir [37] takes advantage of all related reads, including non-BSJ reads, for estimating circRNA expression levels.

Key Figure

Detection of Circular RNA (circRNA) Internal Structure and Alternative Splicing Relies on Back-Spliced Junction (BSJ) Read Pairs

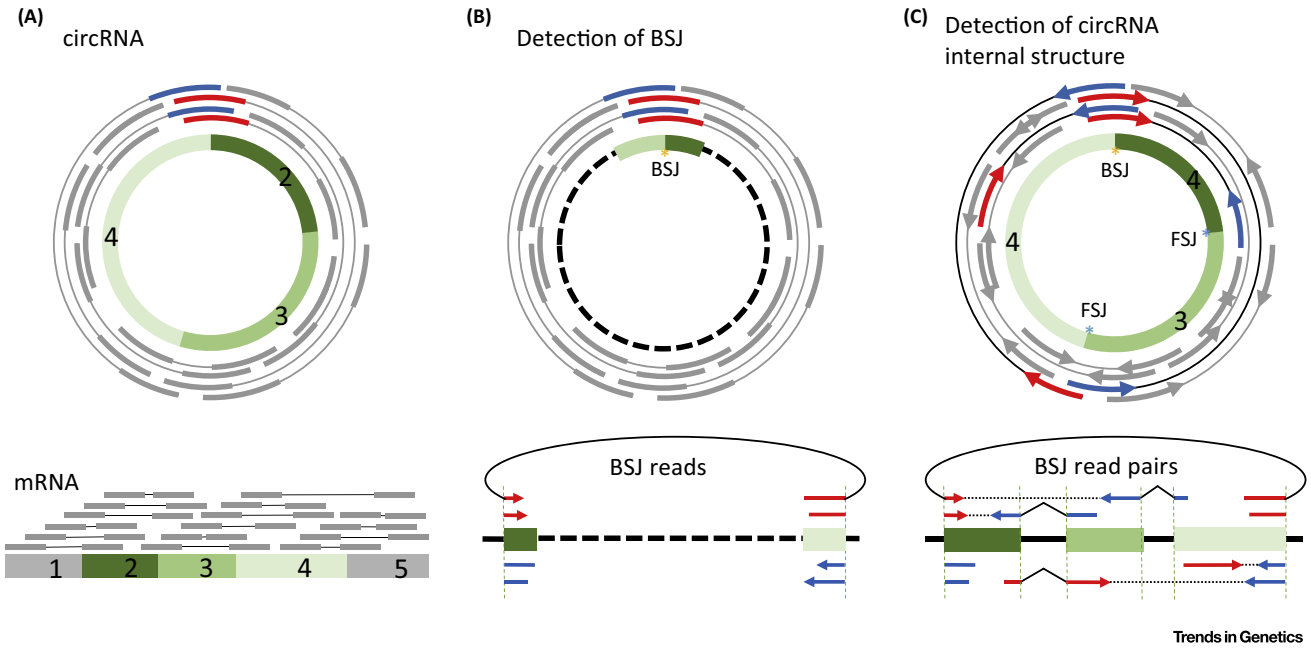

Figure 5. (A) BSJ reads are specific for circRNAs, which can avoid interference from mRNAs. (B) BSJ reads can be used to recognize circRNAs, but often cannot cover forward-spliced junctions (FSJ) within circRNAs. (C) BSJ read pairs with an adequate library insert length can be used for circRNA internal structure and alternative splicing detection.

In contrast to quantification analysis, the determination of circRNA internal structure requires more concrete data (Figure 5, Key Figure). Previous studies unexpectedly found non-exonic circRNAs which contain both intronic and intergenic circRNA fragments (ICFs) [7,8], and two exon–intron circRNAs were demonstrated to be functional in a later study [11]. For a comprehensive investigation of such internal structure, CIRI-AS was designed and applied to multiple datasets of both human and fruit fly samples [14]. In consideration of the significant overlap between circRNAs and their linear counterparts, this tool places emphasis on BSJ reads which are specific to circRNA and thus can avoid interference from linear RNAs. By contrast, it makes flexible use of PEM information, which is an effective filter for circRNA detection as described above, to circumvent short BSJ reads (Figure 5). By analyzing local alignment information for both reads, such as the alignment locations of neighboring segments, FSJs within circRNAs can be identified and used as direct evidence in the detection of **circRNA exons** (cirexons). CIRI-AS applied to real datasets combined with experimental validation demonstrated that >10% of circRNAs contain ICFs which are not recorded in the annotation of the human genome. Furthermore, by reconstructing exon routes based on both detected FSJs and circRNA exons, Gao *et al*. [14] revealed the prevalence of four types of alternative splicing events within circRNAs: exon skipping, alternative 5′ or 3′ splice sites, and intron retention.

Interestingly, these alternative splicing events are distinct from the corresponding mRNAs with regard to their relative abundance as well as their preference for specific splicing factors [14].

Most recently, three additional computational methods have been developed to characterize or visualize the internal structure of circRNAs. FUCHS [36] is able to detect known exon-skipping events within circRNAs by analyzing the alignment of BSJ reads in virtue of BED formatted annotation. CircPro [35] and CircView [55] directly utilize all linear exons recorded in the annotation as the putative full length of the detected circRNAs. Furthermore, these three methods can perform additional analyses. For example, FUCHS [36] can cluster circRNAs according to their sequencing coverage normalized by length, with the purpose of removing false positives with obvious uneven coverage around the BSJ. CircPro [35] focuses on the assessment of protein-coding potential of detected circRNAs, and can output predicted protein sequences with the corresponding potential values provided by a support vector machine (SVM)-based algorithm, CPC [56]. CircView [55] can be used to visualize regulatory elements within circRNAs, including both microRNA response elements and RNA-binding protein binding sites.

Concluding Remarks

In summary, more than 10 circRNA detection methods have been published since 2013. Each of these methods adopts a unique combination of different strategies that vary according to genome reference and annotation usage, read-aligner choice, identification and recovery of BSJ reads, as well as PEM information and other filters. Further efforts will continue to be made to develop detection methods with more comprehensive performance, and new methods based on novel sequencing technology such as single-molecule sequencing will probably be widely used in the future (see Outstanding Questions). On the basis of these circRNA detection methods, current downstream tools focus on differential expression analysis, accurate quantification, alternative splicing detection, full-length assembly, and visualization, and even more diverse downstream analysis methods will be developed to facilitate intense studies on circRNAs.

Outstanding Questions

Is there an appropriate way to combine accurate quantification with a stringent statistical test for differential expression of circRNAs?

How can we quantify the expression of circRNAs at the isoform level?

How can we accurately determine the junction ratio, a measure of the relative expression of circRNA against its linear counterpart, across the BSJ?

The use of BSJ read pairs in the detection of circRNA internal structure is sometimes limited by circRNA length and by the insert length distribution of the library. For example, an exon in a circRNA located >1 kb away from the BSJ in both orientations is unlikely to be captured in a sequencing dataset with an average insert length of 200 bp. How can we reconstruct full-length circRNAs?

Novel sequencing technology such as single-molecule sequencing has a significant advantage regarding read length. To further improve the efficiency of detecting and reconstructing circRNAs, how can such sequencing technology be combined with regular high-throughput sequencing?

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (31722031, 91640117, 31671364, and 91531306).

References

1. Jeck, W.R. *et al.* (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19, 141–157
2. Cocquerelle, C. *et al.* (1993) Mis-splicing yields circular RNA molecules. *FASEB J.* 7, 155–160
3. Lu, T.T. *et al.* (2015) Transcriptome-wide investigation of circular RNAs in rice. *RNA* 21, 2076–2087
4. Salzman, J. *et al.* (2012) Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* 7, e30733
5. Wang, P.L. *et al.* (2014) Circular RNA is expressed across the eukaryotic tree of life. *PLoS One* 9, e90859
6. Westholm, J.O. *et al.* (2014) Genome-wide analysis of *Drosophila* circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep.* 9, 1966–1980
7. Gao, Y. *et al.* (2015) CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.* 16, 4
8. Memczak, S. *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338
9. Salzman, J. *et al.* (2013) Cell-type specific features of circular RNA expression. *PLoS Genet.* 9, e1003777
10. Lasda, E. and Parker, R. (2014) Circular RNAs: diversity of form and function. *RNA* 20, 1829–1842
11. Li, Z.Y. *et al.* (2015) Exon–intron circular RNAs regulate transcription in the nucleus. *Nat. Struct. Mol. Biol.* 22, 256–264
12. Guo, J.U. *et al.* (2014) Expanded identification and characterization of mammalian circular RNAs. *Genome Biol.* 15, 409
13. Veno, M.T. *et al.* (2015) Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development. *Genome Biol.* 16, 245
14. Gao, Y. *et al.* (2016) Comprehensive identification of internal structure and alternative splicing events in circular RNAs. *Nat. Commun.* 7, 12060
15. Li, Y. *et al.* (2015) Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res.* 25, 981–984
16. Ashwal-Fluss, R. *et al.* (2014) circRNA biogenesis competes with pre-mRNA splicing. *Mol. Cell* 56, 55–66

17. Kramer, M.C. *et al.* (2015) Combinatorial control of *Drosophila* circular RNA expression by intronic repeats, hnRNPs, and SR proteins. *Genes Dev.* 29, 2168–2182
18. Liang, D. and Wilusz, J.E. (2014) Short intronic repeat sequences facilitate circular RNA production. *Genes Dev.* 28, 2233–2247
19. Zhang, X.O. *et al.* (2014) Complementary sequence-mediated exon circularization. *Cell* 159, 134–147
20. Barrett, S.P. *et al.* (2015) Circular RNA biogenesis can proceed through an exon-containing lariat precursor. *Elife* 4
21. Conn, S.J. *et al.* (2015) The RNA binding protein quaking regulates formation of circRNAs. *Cell* 160, 1125–1134
22. Hansen, T.B. *et al.* (2013) Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388
23. Legnini, I. *et al.* (2017) Circ-ZNF609 is a circular RNA that can be translated and functions in myogenesis. *Mol. Cell* 66, 22–37
24. Yang, W. *et al.* (2015) 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
25. Cheng, J. *et al.* (2015) Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* 32, 1094–1096
26. Chuang, T.J. *et al.* (2016) NCLscan: accurate identification of non-co-linear transcripts (fusion, trans-splicing and circular RNA) with a good balance between sensitivity and precision. *Nucleic Acids Res.* 44, e29
27. Gao, Y. *et al.* (2017) Circular RNA identification based on multiple seed matching. *Brief. Bioinform.* Published online February 28, 2017. http://dx.doi.org/10.1093/bib/bbx014
28. Hoffmann, S. *et al.* (2014) A multi-split mapping algorithm for circular RNA, splicing, trans-splicing, and fusion detection. *Genome Biol.* 15, R34
29. Izuogu, O.G. *et al.* (2016) PTESFinder: a computational method to identify post-transcriptional exon shuffling (PTES) events. *BMC Bioinform.* 17, 31
30. Song, X. *et al.* (2016) Circular RNA profile in gliomas revealed by identification tool UROBORUS. *Nucleic Acids Res.* 44, e87
31. Szabo, L. *et al.* (2015) Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol.* 16, 126
32. Wang, K. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* 38, e178
33. Zeng, X. *et al.* (2017) A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput. Biol.* 13, e1005420
34. Glazar, P. *et al.* (2014) circBase: a database for circular RNAs. *RNA* 20, 1666–1670
35. Chen, X.M.Q.C.P.Z.M. (2017) CircPro: an integrated tool for the identification of circRNAs with protein-coding potential. *Bioinformatics* 33, 3314–3316
36. Metge, F. *et al.* (2017) FUCHS-towards full circular RNA characterization using RNAseq. *PeerJ.* 5, e2934
37. Li, M.S. *et al.* (2017) Quantifying circular RNA expression from RNA-seq data using model-based framework. *Bioinformatics* 33, 2131–2139
38. Baruzzo, G. *et al.* (2017) Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods* 14, 135–139
39. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359
40. Langmead, B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25
41. Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*
42. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760
43. Kim, D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14, R36
44. Trapnell, C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111
45. Dobin, A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21
46. Kent, W.J. (2002) BLAT – the BLAST-like alignment tool. *Genome Res.* 12, 656–664
47. Szabo, L. and Salzman, J. (2016) Detecting circular RNAs: bioinformatic and experimental challenges. *Nat. Rev. Genet.* 17, 679–692
48. Kim, D. and Salzberg, S.L. (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* 12, R72
49. Wang, Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63
50. Trapnell, C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511
51. Rybak-Wolf, A. *et al.* (2015) Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol. Cell* 58, 870–885
52. Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12, 323
53. Xing, Y. *et al.* (2006) An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res.* 34, 3150–3160
54. Patro, R. *et al.* (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* 32, 462–464
55. Feng, J. *et al.* (2017) CircView: a visualization and exploration tool for circular RNAs. *Brief. Bioinform.* Published online June 30, 2017. http://dx.doi.org/10.1093/bib/bbx070
56. Kong, L. *et al.* (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, W345–W349