Published in final edited form as:

Nat Rev Genet. 2016 October 14; 17(11): 679–692. doi:10.1038/nrg.2016.114.

Detecting circular RNAs: bioinformatic and experimental challenges

Linda Szabo[1] and **Julia Salzman**[2]

[1]Stanford Biomedical Informatics Training Program, Stanford, California 94305, USA

[2]Stanford Department of Biochemistry, Department of Biomedical Data Science, Stanford Cancer Institute, Stanford, California 94305, USA

Abstract

The pervasive expression of circular RNAs (circRNAs) is a recently discovered feature of gene expression in highly diverged eukaryotes. Numerous algorithms that are used to detect genome-wide circRNA expression from RNA sequencing (RNA-seq) data have been developed in the past few years, but there is little overlap in their predictions and no clear gold-standard method to assess the accuracy of these algorithms. We review sources of experimental and bioinformatic biases that complicate the accurate discovery of circRNAs and discuss statistical approaches to address these biases. We conclude with a discussion of the current experimental progress on the topic.

In 2012, a surprising feature of gene expression programmes that had been overlooked for decades was discovered: the pervasive expression of circular RNAs (circRNAs) in eukaryotic genes, with circRNAs constituting the dominant isoform in hundreds of human genes[1]. circRNAs are RNA molecules in which a covalent and canonical linkage termed a “backsplice” has formed between a downstream 3′ splice site and an upstream 5′ splice site in a linear pre-messenger RNA (FIG 1). Most backsplices reported so far occur at annotated exon boundaries or at locations that contain canonical splice signals that are recognized by the spliceosome. The size of a spliced circRNA molecule can range from smaller than 100 nt to larger than 4 kb (REF 2), although the most common size in human cells seems to be a few hundred nucleotides[3–5]. They can be formed from spliced introns or from one or more exons (they are most commonly formed from two or three exons in humans[4]), sometimes with retained introns (the characterization of circRNA types is reviewed elsewhere[2]). Most genes with circular isoforms produce only one or two distinct circRNAs, although some produce tens of distinct circular products[4–9]. In addition to their topology, circRNAs are distinguished from mRNAs in that they lack poly(A) tails and 5′ caps. Although not

Correspondence to: J.S. julia.salzman@stanford.edu.

Competing interests statement

The authors declare no competing interests.

FURTHER INFORMATION

Rateratio test: https://cran.r-project.org/web/packages/rateratio.test/

R code: https://github.com/lindaszabo/NRG

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

essential, long flanking introns that contain inverted repeat sequences seem to promote exon circularization[3,4,9,10].

Genome-wide statistical analysis of splicing led to the discovery, in 2012, of transcripts with exons arranged in a scrambled order (compared with the reference genome) for approximately 10% of the genes expressed in human leukocytes. In hundreds of these events the scrambled exons were expressed at levels that were comparable to those of the linear isoforms of the gene[1]. Further statistical and biochemical tests revealed that these splicing events were contained in topologically circular RNA molecules[1], a finding that was subsequently confirmed by other groups[3,11]. Prior to this work, the expression of circRNAs was an almost completely uncharacterized component of eukaryotic gene expression, owing to the depletion of circRNAs by a poly(A) selection step in most RNA sequencing (RNA-seq) library preparations at the time, as well as bioinformatics filters imposed by the most widely used algorithms that detect unannotated splicing events. Initial rare observations of circular RNA were serendipitous in the context of the indepth study of a particular gene and, with the exception of *SRY*[12] and *CDR1* antisense RNA (*CDR1as*)[13], were generally dismissed as anomalies owing to their low abundance (*DLL*[14], *ETS-1* (REF 15), *MLL*[16], sodium/calcium exchanger 1 (*NCX1*)[17], *dystrophin*[18], *Mbnl*[19] and *ANRIL*[20]). Current estimates suggest that the abundance of circRNA is approximately 2–4% of the total mRNA in cells and can be much higher in some cell types, such as platelets[3,6,21,22].

During the past few years there has been a considerable increase in interest in circRNA expression, where it has been most extensively studied in metazoans: from humans to mice, flies and worms[3–7,11,21,23–29]. The expression of circRNAs is an ancient genomic feature that has either been conserved over billions of years of evolution or independently evolved multiple times. circRNAs are expressed from the genomes of very simple organisms such as *Saccharomyces cerevisiae* and other fungi, amoeba and *Plasmodium falciparum*[30]. circRNAs are also expressed in plants, which may not be surprising given that they have extensive alternative splicing programmes (reviewed in REFS 31–33). The discovery of circRNAs provides additional evidence that ‘protein-coding genes’ and their post-transcriptional regulation and processing may have functions that are completely independent of protein coding. Although an engineered circRNA mini-gene that contains an internal ribosome entry site (IRES) can be efficiently translated[34], all current evidence points to a non-coding function for naturally occurring circRNA. If true, this could have a substantial impact on our view of the evolution and function of genes and genomes. However, with the exception of *CDR1as* and *SRY*, which are abundant circRNAs that are now known to function as microRNA sponges[11,23], the functions of most circRNAs remain unknown.

The pervasive expression of circRNAs that comprise annotated exons from protein-coding genes challenges a broad range of assumptions: first, the sufficiency of algorithms to analyse RNA-seq data — although several algorithms have since been ‘patched’ to allow the discovery of circRNA, no current algorithm provides estimates of false-positive and false-negative rates for the isoforms that are detected; second, the perspective that the field has an adequate mechanistic model of splice site selection and RNA processing; and third, the characterization of long non-coding RNA (lncRNA) as mostly distinct from protein-coding genes.

In this Review, we focus on computational approaches used to characterize the potential functions of circRNAs, which hinge on experimental, bioinformatic and statistical approaches to identify them. We begin with a brief discussion of the current state of RNA-seq algorithms for linear splice detection, which predate circRNA algorithms and are more mature. We then highlight sources of error and bias in both RNA-seq experiments and downstream analysis that complicate the genome-wide discovery of circRNA, discussing the approaches used by published circRNA algorithms, and underline the need for improved standards to evaluate their accuracy. Finally, we consider the evidence for circRNA functionality on the basis of the published results of existing algorithms. In contrast to benchmarking these algorithms, this Review provides insight into the sources of false-positive circRNAs, both experimental and bioinformatic, which will inform the development of improved algorithms and will guide users of current algorithms in the interpretation of their data.

Challenges in detecting splicing

The advent of RNA-seq initially suggested that a complete and precise reconstruction of transcriptomes, with low false-positive splice site identification, would be feasible and straightforward. A plethora of algorithms capable of detecting spliced (linear) alignments were developed, and were quickly followed by algorithms that use these spliced alignments to identify and quantify full-length transcripts. Although much progress has been made on both of these tasks, objective benchmarks that are based on simulated and experimental data by numerous groups demonstrate that substantial conceptual and computational improvements are needed to improve accuracy[35–40]. The field now has a much better appreciation of the fact that the accurate detection of spliced alignments is an important unsolved problem, even before the additional challenges of detecting circRNAs or other novel classes of RNA are considered.

Even in the ideal scenario in which the sequenced genome is essentially identical to the reference genome, as is the case for the mouse strain C57BL/6NJ, algorithms do not agree on the expressed isoforms and have striking differences in recall and false-discovery rates[35]. One explanation for this discrepancy is that each algorithm implements a distinct set of heuristics designed to minimize a particular known source of false positives or false negatives and each uses distinct methods to achieve reasonable run-time or memory usage to deal with the ever-increasing size of RNA-seq data sets. Another factor that distinguishes algorithms is whether they carry out splice-junction discovery before final alignment or directly assign final alignments for each read, which also influences accuracy, as evidenced by comparing the output of algorithms that can run in either mode[41].

Sequence homology and degenerate sequences at exon boundaries can complicate the assignment of a read to the correct splice junction. Distinct approaches to handling mismatches and indels (insertions and deletions) lead to an even greater variation in accuracy, with different aligners preferentially reporting indels with specific features and preferentially assigning indels either to the middle or to the ends of reads[35].

Spliced-alignment algorithms can be either annotation dependent, thus identifying splice junctions that occur between annotated exons, or annotation independent, thus identifying splice junctions from read alignments to a reference genome independently of gene annotations. For published algorithms that provide the option to run in either mode, using annotations improves accuracy[35], but there is a need for improved annotation-independent algorithms to enable the genome-wide discovery of novel types of RNA and the comprehensive characterization of all of the transcripts that are expressed in a cell.

This need is exemplified by circRNAs, which had until recently ‘flown under the radar’ of all published algorithms[1]. Like other classes of RNA that lack poly(A) tails, circRNAs are generally depleted by the poly(A) selection step that is commonly used in library preparation. The development of biochemical protocols for RNA purification during library preparation, such as ribosomal RNA (rRNA) depletion and poly(A) depletion, resulted in RNA-seq libraries in which circRNAs were more prevalent, and such libraries are now regularly used in the study of non-coding RNA. However, some non-coding RNAs, including circRNAs, are present, usually at low levels, in poly(A)$^+$ (poly(A)-enriched) libraries because the selection step is not completely efficient and because some abundant circular RNAs contain A-rich sequences. Thus, circRNAs should have been detected in poly(A)$^+$ RNA-seq libraries by algorithms along with lowly expressed mRNA isoforms. Because existing algorithms had previously failed to detect circRNAs, several groups have developed algorithms specifically for the detection of circRNAs since their initial discovery (TABLE 1). Although there has been extensive work in this area, how to achieve highly sensitive and specific genome-wide detection of circRNAs remains an unsolved problem. Like algorithms for detecting linear splicing, circRNA detection algorithms implement distinct alignment methodologies and heuristics, leading to highly divergent results[41].

Challenges for circRNA detection

Experimental challenges

Many variations in RNA-seq library preparations exist, significantly affecting the abundance of circRNAs in the resulting RNA-seq data sets. Biochemical steps in library preparation that have the most important influence on circRNA detection are, first, RNA purification; second, size selections at the RNA or the cDNA level; and third, RNA fragmentation, method of priming and/or adaptor ligation (FIG 2Aa–Ac). Currently, RNA-seq libraries from eukaryotic cellular RNA are typically either poly(A)-selected or depleted of rRNA before library preparation. For studies that aim to identify protein-bound or ribosome-bound RNA (small RNA libraries), RNA may be biochemically purified using affinity purifications or a sucrose gradient. The only purification that is predicted to significantly deplete a sample of circRNAs is a poly(A) enrichment step, as circRNAs lack a poly(A) tail[1]. By contrast, circRNAs are retained in rRNA-depleted libraries and are enriched in libraries treated with RNase R to digest linear RNA. Because the size selections that are used in RNA-seq typically exclude molecules that are under 200 nt (REF 6), they are likely to influence circRNA detection only by excluding very small circRNAs, if they exist. Random priming, unlike oligo (dT) priming, does not require a poly(A) stretch, such as a poly(A) tail, and will thus result in an RNA-seq library that is not biased against circRNAs. Finally, small RNA libraries will be biased against circRNAs only if the RNA is not fragmented before either adaptor ligation or priming, as circRNAs do not have free ends unless they are nicked *in vivo* or *in vitro*.

Common RNA-seq protocols also introduce technical artefacts that can result in spurious identification of circRNA isoforms. As has been appreciated for some time, technical artefacts can be introduced during the ligation and reverse transcription steps of RNA-seq library preparation (FIG 2Ba–Bc). Reverse transcriptase (RT) can introduce substantial template-switching artefacts, in which two distinct RNA molecules are joined by RT (FIG 2Ba), which can confound RNA-seq analyses that attempt to discover novel RNA isoforms[42–46]. Template switching can mimic linear splicing or backsplicing.

Long homologous sequences promote template switching[45], so this is particularly problematic for genes that produce multiple isoforms that share identical constitutive exons. In fact, such artefacts can dominate the results for some genes, with specific examples suggesting that these artefacts can account for 34–55% of the isoforms detected, sometimes with specific artefactual isoforms detected at higher levels than any of the truly expressed isoforms[44]. As such artefacts can be generated at high levels, filtering out the circRNAs that are supported by only a few reads is not sufficient to eliminate the false positives generated due to template switching, and can result in true isoforms being overlooked. Additionally, RNA-seq library preparation that involves ligation steps can produce chimeric cDNAs and, therefore, can generate artefactual circRNAs at a low rate[47] (FIG. 2Bb). It is important for algorithms that aim to detect circular and linear RNA to test and account for these artefacts, as they can lead to false positives in algorithmic prediction; examples of approaches to account for these artefacts can be found in REF. 46.

In addition to introducing template-switching artefacts, RT is known to have the potential to strand displace[48], thus artificially inflating quantitative estimates of abundance (FIG. 2Bc). This can influence both circular and linear isoform detection, but it may be especially relevant for circRNA detection in which multiple cDNA copies of a single small circRNA could be generated through rolling circle amplification if RNA is not fragmented before cDNA synthesis or if cDNA molecules are not tagged at the 3′ or 5′ ends (as in the ScriptSeq protocol[49]). Because RT has limited processivity[50], this consideration is likely to have little effect on large circRNAs.

Statistical approaches can be used to determine whether the above biases significantly influence circRNA ascertainment. Methods that have previously been used include estimating the read count depth inside and outside exonic boundaries defined by a circular junction, and testing whether circRNA counts are enriched or depleted in libraries that should be depleted of the linear isoform (for example, RNase R^+ libraries). Each of these methods provides a computational test of whether detected circRNAs can be explained by biochemical artefacts. However, a recent study has highlighted the inaccuracy of exon-density-based estimates of isoform expression[51]. Some methods have been developed to address this issue, but more work is needed[52].

Bioinformatic challenges

circRNAs constitute a small fraction of reads in common cell lines, approximately 1–3% of the level of mRNA[6], and although this level is higher in some primary tissues such as platelets[22], most circRNAs are expressed at low levels[1,5,11,26]. In single-end RNA-seq data, circRNAs can only be identified by reads aligned to the backsplice junction, as all other reads may have been generated by either a linear or a circular isoform. In addition to reducing the sample size compared with linear isoforms, relying on junctional reads is problematic because read density in any fixed window can have significant biases that are not yet understood[51]. Furthermore, the degenerate sequence motifs at exon boundaries mean that a convolution of homology and sequencing errors can lead to false-positive alignments (FIG. 2C). Given the large number of publicly available transcriptome high-throughput sequencing experiments, even low technical error rates can generate false-positive alignments to backsplice junctions at sufficient levels that they seem to represent truly expressed circRNAs.

Biases in the algorithms that have been designed to minimize common sources of false positives can cause systematic ‘blind spots’ that lead to incorrect conclusions about the production and regulation of circRNAs[5]. Common strategies to reduce false positives include the use of gene annotations or a requirement for canonical U2 (major) splice signals. However, such restrictions reduce sensitivity, as gene annotations are incomplete and a small fraction of genes are known to use non-U2 splice signals. As most algorithms do not control false positives that occur at canonical sequence motifs, many algorithms additionally apply (high) thresholds on absolute read count[3,4,7,27,28] or on counts relative to linear reads from the host gene[26]. However, many essential genes are expressed at a low level (for example, *SMO* (Smoothened)) and read count is not highly predictive of whether a junction is truly expressed.

An important study[53] of the accuracy of splice junction classification into true-positive and false-positive junctions for a training set using several scoring metrics that did not include read count compared these metrics with the read count. Importantly, this study found that read count was the least reliable metric and yielded the worst performance in terms of junction classification. For this reason, other algorithms take statistical approaches to reduce reliance on read count thresholds[1,5,6,53]. In a comprehensive benchmark[35], the statistical algorithm for linear splice detection[53] was ranked among the top performers across several evaluation metrics, including highest accuracy for novel splice detection. The benefit of statistical approaches for circRNA detection was highlighted in a recent paper[5] in which several apparently highly expressed circRNA junctions that contained exons from different but homologous genes, discarded post-facto as probable RT or alignment artefacts by other algorithms, were flagged as false positives by their statistical score.

Additionally, a statistical approach to discover unannotated splice sites used in circRNAs allowed the authors to relax the requirement of U2 splice signals typically imposed to minimize false-positive rates and to discover the first examples of circRNAs that are spliced by the U12 (minor) spliceosome[5]. One such example occurs in the gene *RANBP17*, the mRNA of which codes for a GTPase that is thought to be a nuclear transport receptor[54]. However, further development of statistical techniques for the *de novo* discovery of circRNAs is necessary, as multiple distinct splice sites within close proximity are still discarded as probable artefacts, underestimating splicing diversity, as exemplified by the splicing of *RANBP17* (REF. 12) (FIG. 3). ‘Hotspot’ genes that produce multiple circRNA isoforms using U2 splice signals were initially identified using a statistical strategy[6] and have also recently been reported[8,9] using find circ[11], although only one group[8] has experimentally validated some of these predictions. Notably, before selecting hotspot circRNA candidates for validation, additional stringent filters were imposed because the authors believed that 10 of the 50 most highly expressed circRNAs reported were likely to be errors due to homology. Improved statistical approaches that eliminate the need for such stringent post-filtering of results are thus needed to further explore the genome-wide prevalence of circRNA hotspots and their functional implications.

The use of RNase R, a highly processive 3′ to 5′ exonuclease that digests nearly all linear RNA that contains at least seven unstructured nucleotides at the 3′ end[55], to enrich for circRNAs before sequencing[3] is becoming more common. However, as has already been noted with respect to linear splice detection, increased coverage improves sensitivity but actually results in a higher false-positive rate[36,53,56], and longer reads do not sufficiently address this issue[56]. Thus, improved enrichment strategies and sequencing technologies cannot be expected to eliminate the need for further algorithm development to increase specificity. Moreover, many known circRNAs are sensitive to RNase R under some regimes[3,4,7,26,57].

Comparison of circRNA algorithms

Some circRNA algorithms are specifically limited to single-end (SE)[11,25] or to paired-end (PE) data[1,6,7], but most algorithms provide options to use either. In all cases, PE data increase the sensitivity, and in some cases also the specificity, reported by algorithm developers. Higher read coverage also improves sensitivity[58]. With the exception of segemehl[25], all pipelines use an external aligner, with Bowtie and STAR being common choices, and begin by filtering out reads that contiguously align to the genome and/or to the transcriptome. Subsequent processing of the unaligned reads identifies reads that align to a backsplice junction. As for linear spliced alignment algorithms, some carry out splice-junction discovery before final alignment, whereas others directly assign final alignments for each read. Algorithm-specific criteria for the types of back-splice junctions considered and for what constitutes a junction-aligned read or a true-positive junction based on the features of aligned reads are applied to limit false positives. Read counts, detection in multiple samples, RNase R resistance, lack of good linear explanation and statistical scores have been used by various algorithm developers to reduce false positives. However, these filters all inevitably result in the inability of the algorithms to detect some circRNA isoforms (blind spots) (TABLE 1).

Hypotheses about genome-wide circRNA regulation and function must be based on the accurate quantification of circular and linear RNA to avoid the propagation of these errors in downstream bioinformatic analyses. Before testing a genome-wide hypothesis, many authors define criteria to select a subset of high-confidence circRNAs reported by an algorithm (TABLE 2). A combination of the algorithm-specific filtering criteria used to identify circRNAs and the criteria for selecting a high-confidence subset can lead to very different conclusions about circRNA regulation. For example, the length of single-exon circRNAs has been examined using three algorithms[3–5]. CircRNAseq[3] reported an average length of 690 nt for circRNAs that comprise a single exon in human fibroblasts, which is three times longer than the average expressed exon, suggesting that longer exons are more easily circularized. By contrast, CircExplorer[4] reported a median length of 353 nt in H9 cells, and KNIFE[5] reported a median length of 260 nt from the same data. The inference for the conservation of circRNA expression is also algorithm-dependent, even when based on an analysis of the same RNA-seq data sets[5].

An obvious limitation of the initial method[1] that demonstrated the prevalence of circRNAs was its reliance on gene annotations, so most subsequent algorithms have focused on annotation-independent discovery. In an effort to reduce false positives, these algorithms only count uniquely mapped reads and require canonical splice signals — a filter that excludes some known circRNA isoforms. For example, current third-party evaluations have reported that the most commonly used algorithm, find circ[11], is less sensitive than other algorithms and can report many false positives[5,8,27,41,58].

Although there are limitations (discussed in detail below) to the methodologies that were used in two recent benchmarking studies, their findings provide insights into the current state of circRNA algorithms[41,48]. Both studies reported little overlap between predictions from different circRNA detection algorithms, with as many as 40% of circRNAs reported by only a single algorithm. Analyses based on the expected enrichment of circRNAs in rRNA$^-$ libraries that have been treated with RNase R or that have been poly(A)-depleted, revealed a high level of predicted false positives; 31–76% of circRNAs detected in an rRNA$^-$ library were not detected by the same algorithm in either of the enrichment libraries[58], and 12–28% of detected circRNAs were depleted by RNase R[41], indicating that they were in fact false-positive circRNAs. Using simulated data, all algorithms demonstrate improved sensitivity with increased read count as expected, and specificity improved by increased read count to a much lesser degree, but the algorithms with the highest specificity had the lowest sensitivity and vice versa[58]. The trade-off between sensitivity and specificity was also observed in the analysis of real data, in which RNase R sensitivity was used to measure false positives[41]

Finally, a distinct biochemical species, known as a lariat, can be detected by many algorithms that identify circRNAs. Lariats are circular by-products of linear splicing that form through a $5' - 2'$ linkage as opposed to the $5' - 3'$ linkage of circRNAs. Similar to circRNAs, lariats can be stable[59] and are resistant to RNase R[3]. They can be distinguished from circRNAs in RNA-seq data by a characteristic decreased coverage of the backsplice owing to inefficient RT traversal of the $5' - 2'$ junction and the insertion of a T residue at this junction[3]. Fewer than 0.17% of the circRNAs reported by all algorithms so far seem to be lariats[41].

Despite the poor agreement between circRNA algorithms reported by these benchmarks, some genome-wide observations relating to circRNAs have been consistently made regardless of the algorithm used, reflecting a signal robust to variation in current algorithms. These observations include the ubiquity of circ-RNA expression, a lack of correlation between linear and circular RNA levels from the same gene, an enrichment of longer flanking introns and no global enrichment of microRNA binding sites in circRNAs.

Benchmarking circRNA detection

Targeted validation of the accuracy of a circRNA algorithm is accomplished by PCR using outward-facing primers and Sanger sequencing for a semi-random selection of tens of predicted circRNAs and testing for RNase R resistance. However, there is currently no gold standard to assess genome-wide sensitivity and specificity of circRNA algorithms. Developers have used a variety of methods to benchmark new algorithms, some that are specific to circRNAs and others that are more broadly applicable to algorithms that detect novel linear or circular splicing. We discuss below the benefits of each method, as well as important biochemical and computational limitations that must be taken into account when interpreting results. TABLE 3 summarizes this discussion.

Method 1: RNase R treatment

RNase R treatment followed by RT-quantitative PCR (qPCR) is the most widely used experimental approach to validate the circRNAs identified from rRNA-depleted samples, and is a method for targeted confirmation of true positives. Extending this methodology, the false-positive rate of some algorithms has been estimated by the fraction of circRNAs detected in a control sample that are not detected after RNase R treatment[3,4,27,41].

However, this probably provides an inaccurate estimate of the genome-wide false-positive rate due to the biochemical variability of RNase R. qPCR validation has shown that some experimentally validated circRNAs are depleted by RNase R, including human *CDR1as*, a *CAMSAP1* isoform with intron retention, *MAN1A2*, *NCX1* and *Drosophila melanogaster Pangolin* and *Ank2* (REFS 3,4,7,26,57). Some circRNAs may be prone to being nicked during library preparation, allowing them to be degraded by RNase R, although it is unclear whether there are specific features of some circRNAs that systematically result in RNase R sensitivity. In addition, there can be a high variability in results between RNase R-treated replicates, with fewer than 50% of the circRNAs that are resistant in one replicate also resistant in the second replicate prepared by the same laboratory[3,27]. Therefore, the list of true circRNAs is typically presented as the union of all the circRNAs that are resistant in any replicate when multiple replicates are carried out. This obscures the statistical variation that would reduce confidence in ‘true-positive’ circRNAs defined by RNase R resistance in any RNA-seq experiment.

Read count fold change for a candidate circRNA between RNase R^- and RNase R^+ samples to determine genome-wide true-positive and false-positive rates, as used in REF. 41, has known statistical issues[60]. Simple cut-offs will inevitably lead to widely discrepant results for the same algorithm. This is due to inherent variability when sampling read counts for a given circle (a Poisson variable): observing fewer reads for this circle in an RNase R^+ sample does not imply that the species was depleted. Basic statistics show that the 95% confidence interval (CI) for the ratio of RNase R^+/RNase R^- reads is larger for genes that are sampled at a lower depth, which is illustrated using simulated read counts from circular junctions representing two libraries sequenced at an equal depth in FIG. 4a. For example, if

100 reads are observed for a given circle in each of the libraries, the 95% CI for the true RNase R^+/RNase $R^−$ ratio assuming Poisson counts is [0.75–1.33], whereas the 95% CI is [0.23–4.34] when five reads are observed in each library.

Other quantitative considerations are also required when comparing RNase R-treated and mock-treated control libraries. Naive comparison of read count fold change without controlling for different sequencing depths in the libraries can clearly lead to inaccurate inferences, either an overestimation or an underestimation of RNase R enrichment. As a quantitative example, in simulated data in which a circle is known to be fivefold enriched in the RNase R library compared with the control, the expected value for the ratio of observed reads in the two libraries is 5/1 when both libraries are sequenced at equal depth. But the expected ratio of observed counts drops if the control library is more deeply sequenced than the RNase R library, with the true ratio not even included in the 95% CI if the control library contains twice as many reads as the RNase R library (FIG. 4b). Therefore, normalization of the counts in the two libraries is needed to make a meaningful comparison.

Controlling for sequencing depths is not sufficient, and statistical methods are required to use RNase R resistance to assess the genome-wide false-positive rate of circRNA algorithms. Intuitively, because RNase R libraries contain fewer distinct RNAs, even if the two libraries contain the same number of reads, circRNAs remaining after RNase R treatment will be more deeply sampled than in the mock treatment, despite having no absolute increase in abundance. The fact that normalization procedures are essential for drawing genome-wide conclusions from RNA-seq data is well documented[61–65], but such procedures have not been applied to matched RNase R^+ and RNase $R^−$ libraries.

Finally, circRNAs are also expected to be enriched in libraries that are both $rRNA^−$ and $poly(A)^−$ relative to $rRNA^−$ libraries, although to a much lesser degree than after RNase R treatment. On this basis, it has been proposed that circRNAs only detected in $rRNA^−$ and not in matched $rRNA^−$ libraries that are poly(A)-depleted are likely to be false positives[58]. The statistical concerns discussed here with respect to RNase R enrichment are relevant to any circRNA detection methodology that uses expected enrichment profiles between libraries.

Method 2: depletion in $poly(A)^+$ libraries

circRNAs are not expected to be found in $poly(A)^+$ libraries because they lack poly(A) tails. Therefore, depletion in $poly(A)^+$ libraries has been used as evidence that the circRNAs identified in matched rRNA-depleted or $poly(A)^−$ libraries are truly circular[5,7,26], and, conversely, the number of circRNAs detected after poly(A) selection has been used as a proxy for false-positive rates.

As with all RNA isolation protocols, poly(A) selection is not perfect, and some circRNAs remain in $poly(A)^+$ libraries, usually at low levels[5,7,27]. Furthermore, some circRNAs that are expressed at low levels may be absent from a $poly(A)^−$ library but present in a matched $poly(A)^+$ library by chance, just as there are differences between technical replicates. Care must be taken while interpreting results, as the simple presence or absence of a putative circRNA in a $poly(A)^+$ library does not necessarily reflect its status as a false positive or a true positive.

Although relative depletion or enrichment by poly(A) selection, as opposed to absence or presence in poly(A) libraries, may be a meaningful proxy for the genome-wide accuracy of circRNA algorithms, the statistical and quantitative limitations discussed with respect to RNase R resistance also apply here. Normalization procedures for matched poly(A)$^+$ and poly(A)$^-$ libraries are needed, and CIs must be evaluated when assessing whether the read counts observed for a given circle in the two libraries support depletion by poly(A) selection.

Method 3: decoy reads

If a read is truly generated from a splice junction, for paired-end RNA-seq the mate read should align such that mates are consistent with being generated from the ends of a single RNA fragment. Inconsistent 'decoy' reads are typically discarded as experimental or alignment artefacts. For circRNA, decoys include reads for which one mate mapped to a backsplice junction and the other mapped outside the genomic region defined by the backsplice. These decoy reads have been hypothesized to be due to experimental artefacts or genomic rearrangements[7] and a convolution of sequencing errors and exon homology[5]. The proportion of decoy reads aligning to a putative circRNA has been used along with other evidence to assess the quality of the prediction of an algorithm[7].

It is well appreciated that experimental and alignment artefacts can produce RNA-seq reads that are consistent with circRNA. With this in in mind, decoy reads have also been used to fit models that provide a statistical score to filter false-positive circRNAs that have many mapped consistent reads[1,5,6]. A limitation to this methodology is that 'decoy' reads under one model (in this case circRNA versus artefact) may be consistent with a different model that has not been considered, such as exon duplication. Using reads truly generated from a splice junction between duplicated exons as examples of experimental or alignment artefacts would produce an inaccurate statistical model. Even so, models that incorporate decoy reads can detect known false-positive circRNAs that are detected in RNA-seq owing to exon homology. Additional development of statistical models that extend upon this methodology may be useful for independent unbiased benchmarking of the genome-wide false-positive rates of circRNA algorithms.

Method 4: RT specificity

Apparent circRNA reads can be generated from template-switching artefacts (FIG. 2Ba–Bc). These artefacts are often reproducible in independent libraries created using the same RT[66], so cannot be ruled out on the basis of replicability. One group[46] suggested that circRNAs amplified only by either AMV (avian myeloblastosis virus) or MMLV (Moloney murine leukaemia virus) RT (that is, those displaying RT specificity) are false positives, reporting that only six of the 13 candidates amplified by both RTs could be validated. However, another group[22] indicated that four of the remaining seven candidates are also true circRNAs; all four are enriched in platelets and one was RNase R resistant and amplified by both RTs in their hands. On the basis of these reports, although lack of RT specificity can provide an additional line of evidence to support true-positive circRNA, it does not seem to be a reliable experimental method to distinguish circRNAs from artefacts and may result in a high false-negative rate, although additional work is required for confirmation.

Method 5: simulated data

Simulated data are also commonly used to assess sensitivity and specificity on the basis of a known ground truth, and are valuable for identifying systematic limitations of particular algorithms. Several tools for simulating data exist, with a common choice being BEERS[67], which simulates human or mouse paired-end RNA-seq from the Illumina platform with varying levels of gene expression, splicing, sequencing error, single-nucleotide variants (SNVs) and indels. Simulation provides a means to examine the trade-off between sensitivity and specificity for an algorithm, enabling algorithm developers to better understand the types of splicing that are identified as false positives and false negatives, and enabling users to select the tool that best meets their needs. However, it is important to note that experimentally generated RNA-seq data are more complex than simulated data, owing to biochemical events that are not fully understood. It remains unclear how closely performance on simulated data reflects performance on real data. Although one group[35] observed general agreement between results on simulated and real data sets, another group[36] found that the linear splice algorithms that performed best on simulated data were not the same as those that did best in the *in vitro* transcription (IVT) data.

Suggested statistical practice

In light of the convolution of experimental and informatic biases and errors that we have discussed, further statistical method development is required before a gold-standard method to assess the genome-wide accuracy of circRNA algorithms in real data can be achieved. Therefore, evaluation on the basis of a combination of benchmark methods, including simulated data, is necessary.

Two groups[27,58] have provided simulated circRNA data sets that will be valuable for benchmarking algorithms when used in combination with simulated negative data sets that contain only linear reads, such as those provided by the RNA-seq Genome Annotation Assessment Project (RGASP)[35], or a mixture of linear and circular isoforms, to more accurately reflect the circRNA detection task.

In order to conclude that circRNAs have been enriched or depleted in matched libraries, CIs must be computed rather than basing inference on a simple comparison of read counts. As normalization methods remain to be developed for tests of enrichment or depletion in matched libraries, an alternative approach is to compare the circular-to-linear splicing ratio for a given exon in the two libraries, as this method does not rely on normalization. For identifying genome-wide false-positive circRNAs in real data, we consider the depletion of a circRNA in poly(A)$^+$ libraries to be a more appropriate metric than failure to be enriched by RNase R for two reasons: first, only a few validated circRNAs have been detected with more than a few reads in poly(A)$^+$ libraries, whereas some validated circRNAs, such as *CDR1as*, have been reported to be depleted by RNase R[3]; and, second, circRNAs naively identified by some algorithms with high read counts in poly(A)$^+$ libraries are often known common false positives due to sequence homology.

General statistical principles hold for the analysis of circRNA expression: when multiple replicates are carried out for any test for enrichment or depletion in matched libraries, each replicate must be separately analysed and the standard error reported as for any experiment in which multiple replicates are performed. Statistical principles and empirical findings support using PE RNA-seq data to discover and quantify circRNA expression.

Functionality: circumstantial evidence

One of the most convincing and classic arguments for the biological functionality of gene expression is genetic conservation: when a feature of gene expression is conserved, this suggests that an evolutionary fitness pressure has maintained it. Early work showed that the ubiquitous expression of circRNAs is conserved from humans to mice, which also suggested that it was a shared feature of more distantly related metazoans. This was later confirmed in worms and flies by several groups[6,7,11,25], and subsequently extended to a much more divergent group of eukaryotes, including yeast, plants and parasites, the last common ancestor of which existed more than a billion years ago[30].

High circRNA expression from many genes is also conserved across evolution, including the microRNA sponge *CDR1as* and hundreds of other examples[21]. However, whether there is significant genome-wide enrichment of conserved circRNAs is a hypothesis that is still debated in the literature. Some studies have found significant conservation in genes that host circRNA isoforms between humans and mice but other studies have not[11,26]. Tests of genome-wide conservation are always subject to confounding by Simpson’s paradox[68]: it could be the case that genes that host circRNAs expressed in the brain are statistically conserved between humans and mice, whereas, after collapsing over all organs, conservation of gene sets is lost; the converse is also possible.

Another computational test for evolutionary selection on circRNAs is whether wobble bases in exons that are circularized have greater conservation than control exons. This would be expected if circRNAs served as microRNA sponges, RNA-binding protein sponges or had other functions, including those that depend on structure. This point has also been debated in the literature. Some studies found negative evidence of this regulation[26], whereas other studies have reported significant evolutionary pressure on wobble bases[11,26]. Future studies will clarify this issue, and will be dependent on appropriate statistical controls. Of course, evolutionary conservation is not a requirement for genetic function, so whatever consensus is achieved will not conclusively determine whether circRNAs have a function in the cell.

Multiple groups have studied the expression profiles of circRNA in cell culture and in primary cells to find circumstantial evidence of function. circRNAs have been found to be enriched in the ageing fly brain[7], and multiple groups have found that circRNAs are regulated in fetal development, including in humans[58]. circRNAs also exhibit cell type-specific and tissue-specific expression patterns that are independent of linear isoform levels, which also suggests that they are actively regulated[5,8,69], although a competing explanation is that linear RNA abundance is actively regulated, whereas circRNA abundance is not, as has been suggested to be the case for platelets[22].

Conclusions

With little agreement between the predictions of the different circRNA detection algorithms[41,58], and in the absence of a gold standard that can be used to assess the accuracy of the predictions from these algorithms, current hypotheses on the regulation and function of circRNA are based on a subset of the identified circRNAs considered to be high confidence. Both the choice of algorithm and the thresholds used to select the high-confidence set of circRNAs for further analysis can greatly alter these conclusions. As read count is a poor predictor of whether a junction (linear or circular) is truly expressed[5,8,53], an important first step will be the continued development of methods that provide a statistical test that can be used to estimate the false discovery rate and to select appropriate thresholds for high-confidence circRNA detection. Normalization procedures for assessing enrichment or depletion in matched libraries are necessary in order to allow researchers to accurately interpret genome-wide results from these experiments. Additional work on methods to test for biochemical artefacts in common RNA-seq protocols will also be essential to reduce the need for adhoc filtering of bioinformatic results. Importantly, such improvements will also be beneficial to linear splice detection algorithms that are also negatively influenced by such artefacts.

Although many tools exist for simulating linear reads with a variety of error and splicing profiles, these tools will need to be extended to enable the simulation of different ratios of circular and linear isoforms using both annotated and unannotated exons. Appropriately designed *in vitro* transcribed circRNA libraries providing a ground truth would be instrumental both for informing the development of statistical methods and for presenting a more realistic alternative to simulated data for evaluating the trade-off between sensitivity and specificity. Methods for *in vitro* circularization are reviewed in REF. 70.

Finally, the very recent revelation of the widespread existence of circRNAs raises the question: are there other classes of linear or circular RNA isoforms that are still being overlooked? This question is difficult to answer because novel isoform detection is confounded by many factors, and perhaps no single biochemical approach may be comprehensive. Completely reference-free approaches to determine and quantify expressed RNAs are needed and present challenges and opportunities for biological insight into the full repertoire of RNA expressed by cells.

Acknowledgments

The authors thank G. Hsieh and P. Wang for useful discussion and comments. This work was supported by NCI grant R00 CA168987-03, NIGMS grant R01 GM116847, a JIMB seed grant and an NSF CAREER award to J.S. and McCormick-Gabilan and a Baxter Family Fellowship. The authors would also like to acknowledge the support of the Stanford Center for Computational, Evolutionary and Human Genomics. J.S. is an Alfred P. Sloan fellow in Computational & Evolutionary Molecular Biology.

Glossary

Splice signals	Conserved sequences delineating introns in pre-mRNA and recognized by the spliceosome. Nearly all introns contain a GU at the 5′ end of the intron and an AG at the 3′ end

(canonical U2 splice signal); the U12 splice signal is (A|G)TATCCT(C|T), and is present in a minority of exons.

RNA sequencing (RNA-seq). A technique to obtain the sequence of the transcriptome (all expressed RNA) in a sample. It enables the identification and quantification of alternative splicing, as well as gene-level expression.

MicroRNA sponges An RNA molecule containing microRNA-binding sites that sequesters the microRNA away from its target in a sequence-specific manner.

Indels Insertions and deletions in the sequenced genome compared with a reference genome.

Oligo(dT) priming Priming with a primer that hybridizes to the poly(A) tail of mRNA.

Wobble bases The third position in a 3 nt codon in which more than one nucleotide in this position codes for the same amino acid.

References

1. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. PloS One. 2012; 7:e30733. This article provided the first demonstration that circRNA was a ubiquitous and overlooked feature of eukaryotic gene expression. [PubMed: 22319583]
2. Lasda E, Parker R. Circular RNAs: diversity of form and function. RNA. 2014; 20:1829–1842. [PubMed: 25404635]
3. Jeck WR, et al. Circular RNAs are abundant, conserved, and associated with *ALU* repeats. RNA. 2013; 19:141–157. [PubMed: 23249747]
4. Zhang XO, et al. Complementary sequence-mediated exon circularization. Cell. 2014; 159:134–147. [PubMed: 25242744]
5. Szabo L, et al. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. Genome Biol. 2015; 16:126. The first published circRNA algorithm to develop a statistical score independent of read count for identifying true and false positives. [PubMed: 26076956]
6. Salzman J, Chen RE, Olsen MN, Wang PL, Brown PO. Cell-type specific features of circular RNA expression. PLoS Genet. 2013; 9:e1003777. [PubMed: 24039610]
7. Westholm JO, et al. Genome-wide analysis of *Drosophila* circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. Cell Rep. 2014; 9:1966–1980. [PubMed: 25544350]
8. Veno MT, et al. Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development. Genome Biol. 2015; 16:245. [PubMed: 26541409]
9. Ivanov A, et al. Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. Cell Rep. 2015; 10:170–177. [PubMed: 25558066]
10. Liang D, Wilusz JE. Short intronic repeat sequences facilitate circular RNA production. Genes Dev. 2014; 28:2233–2247. [PubMed: 25281217]
11. Memczak S, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature. 2013; 495:333–338. [PubMed: 23446348]
12. Capel B, et al. Circular transcripts of the testis-determining gene Sry in adult mouse testis. Cell. 1993; 73:1019–1030. [PubMed: 7684656]

13. Hansen TB, et al. miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. EMBO J. 2011; 30:4414–4422. [PubMed: 21964070]

14. Nigro JM, et al. Scrambled exons. Cell. 1991; 64:607–613. [PubMed: 1991322]

15. Cocquerelle C, Daubersies P, Majerus MA, Kerckaert JP, Bailleul B. Splicing with inverted order of exons occurs proximal to large introns. EMBO J. 1992; 11:1095–1098. [PubMed: 1339341]

16. Caldas C, et al. Exon scrambling of *MLL* transcripts occur commonly and mimic partial genomic duplication of the gene. Gene. 1998; 208:167–176. [PubMed: 9540777]

17. Li XF, Lytton J. A circularized sodium-calcium exchanger exon 2 transcript. J Biol Chem. 1999; 274:8153–8160. [PubMed: 10075718]

18. Surono A, et al. Circular dystrophin RNAs consisting of exons that were skipped by alternative splicing. Hum Mol Genet. 1999; 8:493–500. [PubMed: 9949208]

19. Houseley JM, et al. Noncanonical RNAs from transcripts of the *Drosophila muscleblind* gene. J Hered. 2006; 97:253–260. This study reports the first evidence of a highly enriched circRNA from the fly. [PubMed: 16714427]

20. Burd CE, et al. Expression of linear and novel circular forms of an *INK4*/*ARF*-associated non-coding RNA correlates with atherosclerosis risk. PLoS Genet. 2010; 6:e1001233. [PubMed: 21151960]

21. Rybak-Wolf A, et al. Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. Mol Cell. 2015; 58:870–885. [PubMed: 25921068]

22. Alhasan AA, et al. Circular RNA enrichment in platelets is a signature of transcriptome degradation. Blood. 2015; 127:e1–e11. [PubMed: 26660425]

23. Hansen TB, et al. Natural RNA circRNAs function as efficient microRNA sponges. Nature. 2013; 495:384–388. [PubMed: 23446346]

24. Ashwal-Fluss R, et al. circRNA biogenesis competes with pre-mRNA splicing. Mol Cell. 2014; 56:55–66. [PubMed: 25242144]

25. Hoffmann S, et al. A multi-split mapping algorithm for circular RNA, splicing. *trans*-splicing and fusion detection. Genome Biol. 2014; 15:R34. [PubMed: 24512684]

26. Guo JU, Agarwal V, Guo H, Bartel DP. Expanded identification and characterization of mammalian circular RNAs. Genome Biol. 2014; 15:409. This paper provides a comprehensive controlled analysis of the enrichment in circRNAs from microRNA binding sites. [PubMed: 25070500]

27. Gao Y, Wang J, Zhao F. CIRI: an efficient and unbiased algorithm for *de novo* circular RNA identification. Genome Biol. 2015; 16:4. [PubMed: 25583365]

28. Cheng J, Metge F, Dieterich C. Specific identification and quantification of circular RNAs from sequencing data. Bioinformatics. 2016; 32:1094–1096. [PubMed: 26556385]

29. Kramer MC, et al. Combinatorial control of *Drosophila* circular RNA expression by intronic repeats, hnRNPs, and SR proteins. Genes Dev. 2015; 29:2168–2182. [PubMed: 26450910]

30. Wang PL, et al. Circular RNA is expressed across the eukaryotic tree of life. PLoS ONE. 2014; 9:e90859. [PubMed: 24609083]

31. Yang S, Tang F, Zhu H. Alternative splicing in plant immunity. Int J Mol Sci. 2014; 15:10424–10445. [PubMed: 24918296]

32. Filichkin S, Priest HD, Megraw M, Mockler TC. Alternative splicing in plants: directing traffic at the crossroads of adaptation and environmental stress. Curr Opin Plant Biol. 2015; 24:125–135. [PubMed: 25835141]

33. Meyer K, Koester T, Staiger D. Pre-mRNA splicing in plants: *in vivo* functions of RNA-binding proteins implicated in the splicing process. Biomolecules. 2015; 5:1717–1740. [PubMed: 26213982]

34. Wang Y, Wang Z. Efficient backsplicing produces translatable circular mRNAs. RNA. 2015; 21:172–179. [PubMed: 25449546]

35. Engstrom PG, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods. 2013; 10:1185–1191. Competition-style independent evaluation of linear spliced alignment algorithms identifying systematic discrepancies and blind spots in all algorithms. [PubMed: 24185836]

36. Hayer KE, Pizarro A, Lahens NF, Hogenesch JB, Grant GR. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. Bioinformatics. 2015; 31:3938–3945. [PubMed: 26338770]

37. Carrara M, et al. Alternative splicing detection workflow needs a careful combination of sample prep and bioinformatics analysis. BMC Bioinformatics. 2015; 16:S2.

38. Liu R, Loraine AE, Dickerson JA. Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. BMC Bioinformatics. 2014; 15:364. [PubMed: 25511303]

39. Chandramohan R, Wu PY, Phan JH, Wang MD. Benchmarking RNA-seq quantification tools. Conf Proc IEEE Eng Med Biol Soc. 2013; 2013:647–650. [PubMed: 24109770]

40. Hatem A, Bozdag D, Toland AE, Catalyurek UV. Benchmarking short sequence mapping tools. BMC Bioinformatics. 2013; 14:184. [PubMed: 23758764]

41. Hansen TB, Veno MT, Damgaard CK, Kjems J. Comparison of circular RNA prediction tools. Nucleic Acids Res. 2015; 44:e58. [PubMed: 26657634]

42. Luo GX, Taylor J. Template switching by reverse transcriptase during DNA synthesis. J Virol. 1990; 64:4321–4328. [PubMed: 1696639]

43. Houseley J, Tollervey D. Apparent non-canonical *trans*-splicing is generated by reverse transcriptase *in vitro*. PLoS ONE. 2010; 5:e12271. [PubMed: 20805885]

44. Roy CK, Olson S, Graveley BR, Zamore PD, Moore MJ. Assessing long-distance RNA sequence connectivity via RNA-templated DNA–DNA ligation. eLife. 2015; 4:e03700. This study provided important biochemical evidence for artefactual splicing from RNA-seq and technological solution.

45. Cocquet J, Chong A, Zhang G, Veitia RA. Reverse transcriptase template switching and false alternative transcripts. Genomics. 2006; 88:127–131. [PubMed: 16457984]

46. Yu CY, Liu HJ, Hung LY, Kuo HC, Chuang TJ. Is an observed non-co-linear RNA product spliced in trans, in *cis* or just *in vitro*? Nucleic Acids Res. 2014; 42:9410–9423. [PubMed: 25053845]

47. Quail MA, et al. A large genome center's improvements to the Illumina sequencing system. Nat Methods. 2008; 5:1005–1010. [PubMed: 19034268]

48. Kelleher CD, Champoux JJ. Characterization of RNA strand displacement synthesis by Moloney murine leukemia virus reverse transcriptase. J Biol Chem. 1998; 273:9976–9986. [PubMed: 9545343]

49. Pease J, Sooknanan R. A rapid, directional RNA-seq library preparation workflow for Illumina® sequencing. Nat Methods. 2012; 9

50. Mohr S, et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. RNA. 2013; 19:958–970. [PubMed: 23697550]

51. Lahens NF, et al. IVT-seq reveals extreme bias in RNA sequencing. Genome Biol. 2014; 15:R86. [PubMed: 24981968]

52. Jiang H, Salzman J. A penalized likelihood approach for robust estimation of isoform expression. Stat Interface. 2015; 8:437–445. [PubMed: 27239250]

53. Wang K, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010; 38:e178. [PubMed: 20802226]

54. Koch P, et al. Identification of a novel putative Ran-binding protein and its close homologue. Biochem Biophys Res Commun. 2000; 278:241–249. [PubMed: 11071879]

55. Vincent HA, Deutscher MP. Substrate recognition and catalysis by the exoribonuclease RNase, R. J Biol Chem. 2006; 281:29769–29775. [PubMed: 16893880]

56. Stephan-Otto Attolini C, Pena V, Rossell D. Designing alternative splicing RNA-seq studies. Beyond generic guidelines. Bioinformatics. 2015; 31:3631–3637. [PubMed: 26220961]

57. Jeck WR, Sharpless NE. Detecting and characterizing circular RNAs. Nat Biotechnol. 2014; 32:453–461. [PubMed: 24811520]

58. Chen I, Chen CY, Chuang TJ. Biogenesis, identification, and function of exonic circular RNAs. Wiley Interdiscip Rev RNA. 2015; 6:563–579. [PubMed: 26230526]

59. Hesselberth JR. Lives that introns lead after splicing. Wiley Interdiscip Rev RNA. 2013; 4:677–691. [PubMed: 23881603]

60. Witten, D., Tibshirani, R. Tech Report. Stanford Univ; 2007. A comparison of fold-change and the t-statistic for microarray data analysis.

61. Trapnell C, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7:562–578. [PubMed: 22383036]

62. Salzman J, Klass DM, Brown PO. Improved discovery of molecular interactions in genome-scale data with adaptive model-based normalization. PLoS ONE. 2013; 8:e53930. [PubMed: 23349766]

63. Li P, Piao Y, Shon HS, Ryu KH. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-seq data. BMC Bioinformatics. 2015; 16:347. [PubMed: 26511205]

64. Zyprych-Walczak J, et al. The impact of normalization methods on RNA-seq data analysis. Biomed Res Int. 2015; 2015:621690. [PubMed: 26176014]

65. Erhard F, Zimmer R. Count ratio model reveals bias affecting NGS fold changes. Nucleic Acids Res. 2015; 43:e136. [PubMed: 26160885]

66. Wu CS, et al. Integrative transcriptome sequencing identifies *trans*-splicing events with important roles in human embryonic stem cell pluripotency. Genome Res. 2014; 24:25–36. [PubMed: 24131564]

67. Grant GR, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-seq unified mapper (RUM). Bioinformatics. 2011; 27:2518–2528. [PubMed: 21775302]

68. Simpson EH. The interpretation of interaction in contingency tables. J R Statist Soc. 1951; 13:238–241.

69. Boeckel JN, et al. Identification and characterization of hypoxia-regulated endothelial circular RNA. Circ Res. 2015; 117:884–890. [PubMed: 26377962]

70. Petkovic S, Muller S. RNA circularization strategies *in vivo* and *in vitro*. Nucleic Acids Res. 2015; 43:2454–2465. [PubMed: 25662225]

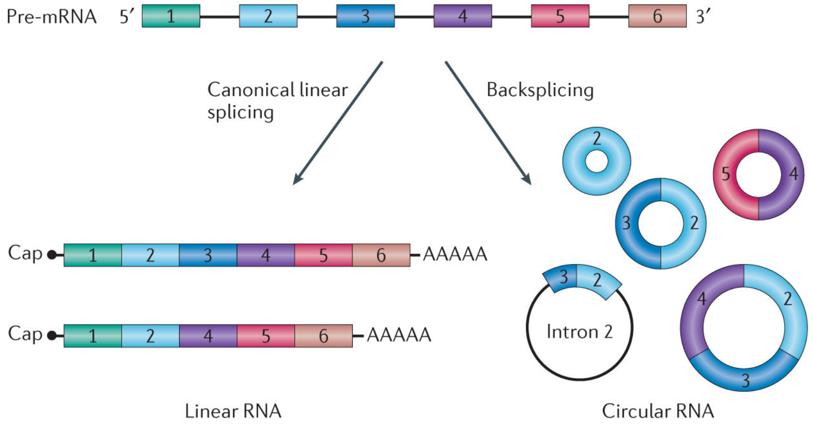

Figure 1. Circular RNA

Circular RNA (circRNA) is produced from both protein-coding genes and non-coding regions of the genome. Linear RNAs are formed by a covalent linkage between an upstream 3′ splice site and a downstream 5′ splice site of pre-messenger RNA (pre-mRNA), whereas circRNA is characterized by a covalent and canonical linkage between a downstream 3′ splice site and an upstream 5′ splice site in a process known as backsplicing. circRNAs lack poly(A) tails and can contain a single exon or multiple exons, as well as introns. Exons are numbered. Adapted from REF. 6.

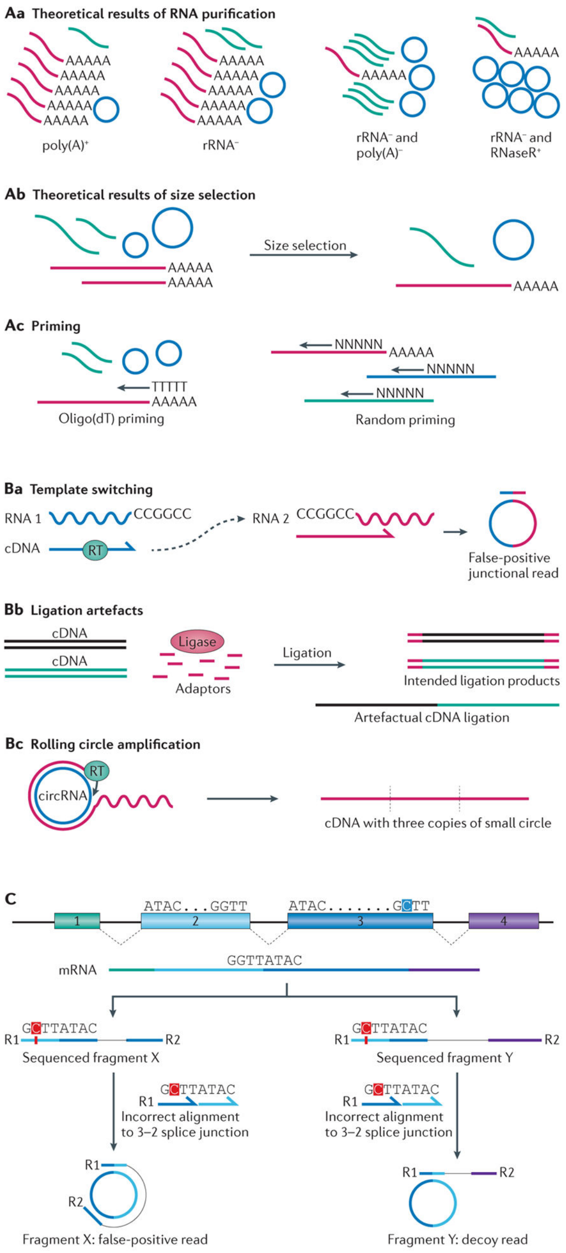

Figure 2. Challenges for circRNA detection in RNA-seq

Aa–**Ac** | Variations in preparation protocols alter the amount of circular RNA (circRNA) in a library. Poly(A) RNA is shown in pink, non-poly(A) RNA is shown in green and circular RNA is shown in blue. **Aa** | Common RNA purification methods, in order of increasing relative amounts of circRNA. circRNAs are depleted by poly(A) selection and retained in ribosomal RNA (rRNA)$^-$ libraries. They constitute a large proportion of reads in an rRNA$^-$ library that has also been depleted of poly(A) RNA, and are the primary RNA in RNase R-treated libraries. **Ab** | Size selection excludes very small circular and linear RNA. **Ac** | Oligo(dT) priming biases against circRNA. **Ba**–**Bc** | Known sources of artefacts from common RNA-seq protocols. **Ba** | Reverse transcriptase (RT) can join two distinct RNA molecules in a non-canonical order, particularly when the two RNAs contain a common sequence. **Bb** | Two distinct cDNAs may be ligated together in non-canonical order during adaptor ligation. **Bc** | RT can displace cDNA from the template, generating a single cDNA that contains multiple copies of a circRNA. **C** | A convolution of homology and sequencing errors can lead to false alignments to a backsplice junction. In this case two fragments generated from a linear exon 2–exon 3 splice junction are sequenced with an error and incorrectly aligned to an exon 3–exon 2 backsplice. If the mate aligns outside the genomic region defined by the backsplice junction it is correctly discarded as a false positive, but if the mate aligns within the presumed circle it is incorrectly considered evidence of circRNA. For clarity, the mRNA sequence shown is the DNA equivalent.

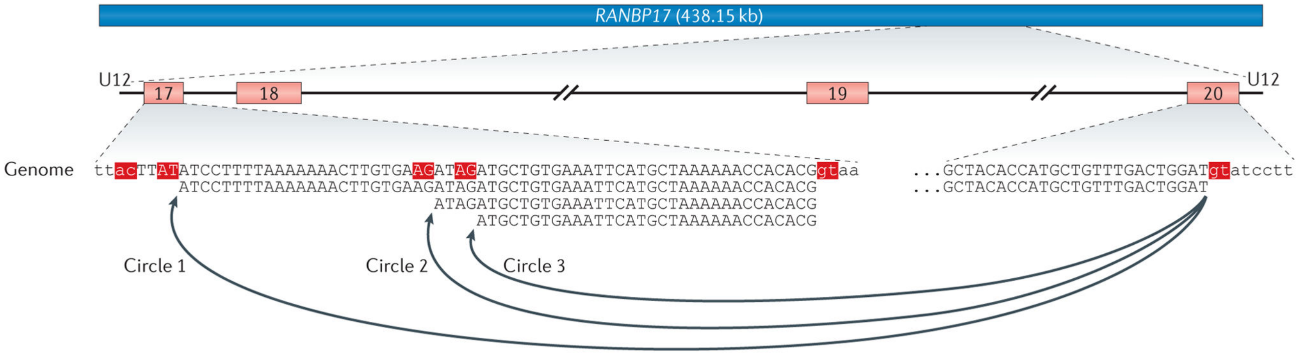
RANBP17 (438.15 kb)
U12
17
18
19
20
U12
Genome
ttacTTATATCCTTTTAAAAAAACTTGTGAAGATAGATGCTGTGAAATTCATGCTAAAAAACCACACGgtaa
ATCCTTTTAAAAAAACTTGTGAAGATAGATGCTGTGAAATTCATGCTAAAAAACCACACG
ATAGATGCTGTGAAATTCATGCTAAAAAACCACACG
ATGCTGTGAAATTCATGCTAAAAAACCACACG
...GCTACACCATGCTGTTTGACTGGATgtatcctt
...GCTACACCATGCTGTTTGACTGGAT
Circle 1
Circle 2
Circle 3

Figure 3. Multiple circRNAs can be generated from a single locus

The *RANBP17* locus is shown at the top, with the circularized region expanded below. The boxes represent annotated exons, with the location of the U12-type splice signal labelled. Three circular isoforms of *RANBP17*, formed by splicing of the 5′ end of exon 20 into three distinct locations within exon 17, were validated by PCR and clone sequencing; only circle 1 and circle 2 were algorithmically predicted[5].

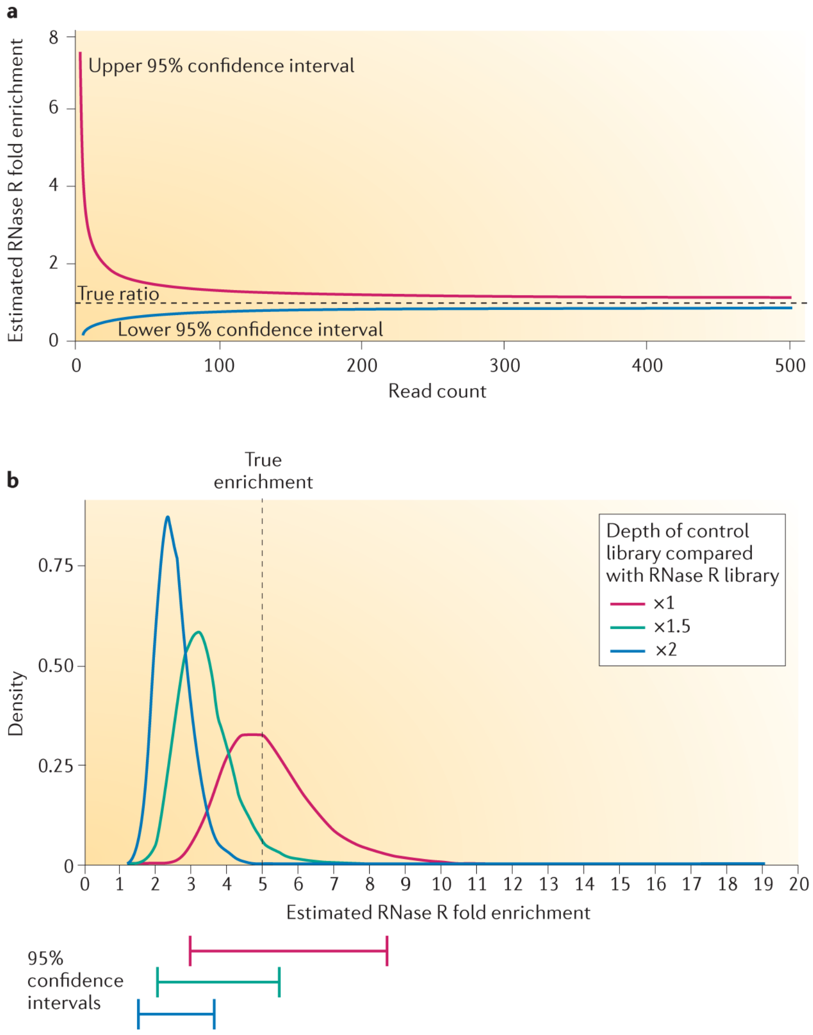

Figure 4. Statistical considerations when using RNase R enrichment to assess genome-wide accuracy

Read counts were simulated in R (code available at https://github.com/lindaszabo/NRG) and confidence intervals were computed using rateratio.test (https://cran.r-project.org/web/packages/rateratio.test). **a** | Upper and lower bounds of the 95% confidence interval for the estimated RNase R fold enrichment when the same number of reads is observed for a given circular RNA (circRNA) in RNase R-treated and mock-treated control libraries sequenced at the same depth. **b** | Density distributions for the ratio of observed read counts for a given circRNA in RNase R^+/control (that is, fold enrichment by RNase R) when the underlying true ratio is 5/1. When the two libraries have equal number of reads (red line), the expected value is 5. If the control library sequenced more deeply, then the expected observed fold enrichment decreases although the underlying rate parameter has not changed.

Table 1

circRNA detection algorithms*

Algorithm	Improvements	Read type	Aligner	Junctions considered	Filtering rules: per read (R) or per junction (J)	Blind spots	Reported validation rate	Refs
MapSplice‡	• Genome-wide independent identification of splicing, including exon scrambling and fusion events • Logistic regression to classify TPs and FPs independently of read count	PE and SE	Bowtie	No restrictions on splice sequence or intron length	• R: assign single best alignment for each read using composite score based on mismatches, base call quality and junction score (based on overlap and uniform read distribution)	Junctions where sampling or alignment properties differ from training set used to estimate regression model parameters	• 20/20 exon skipping events validated by qPCR • 96.3% TP and 8% FP in synthetic data based on ASTD database • 98% specificity and 96% sensitivity in simulated data after parameter tuning	53
Salzman 2012	Genome-wide analysis of circRNA and local rearrangements in RNA-seq data	PE	Bowtie	RefSeq annotated exons within single gene	• R: R2 overlaps junction by 10 nt and R1 within same gene; <4 mismatches	circRNAs using unannotated exons or comprising multiple genes	• 13/13 most highly expressed circRNAs validated by PCR • 17/17 by PCR and Sanger sequencing • 9/9 RNase R resistant	1
CircRNAseq	• Increased sensitivity by using RNase R to enrich for circRNAs prior to genome-wide identification of circRNA	PE and SE	MapSplice (Bowtie)	GT–AG, donor or acceptor within 2MB	• J: enrichment of junctional reads in RNase R library	circRNAs sensitive to RNase R	• 31% FP based on lack of enrichment by RNase R • 7/7 from low-stringency set	3

Algorithm	Improvements	Read type	Aligner	Junctions considered	Filtering rules: per read (R) or per junction (J)	Blind spots	Reported validation rate	Refs
	• Annotation independent						validated by PCR	
find_circ	• Source code available • Reports circular and linear splicing	SE	Bowtie2	GT–AG, anchors within 100kb	• R: unique anchor alignment; anchor extension completely aligns read with <3 mismatches; 22 nt junction overlap • J: unambiguous breakpoint	circRNAs comprising small exons or using non-canonical splice signals	• 75% sensitivity and 0.2% FDR in simulated data • 37/46 (80%) validated by PCR and Sanger sequencing • 37/37 RNase R resistant	11
Salzman 2013	• Reduced FP rate by estimating FDR based on alignment scores instead of hard thresholding • Increased sensitivity for small circRNAs • Per-sample modeling instead of model based on training data	PE	Bowtie2	UCSC KnownGene annotated exons within single gene	• R: R1 overlaps junction by 10 nt • J: FDR < 0.025	circRNAs using unannotated exons or comprising multiple genes	• 8/8 RNase R resistant • 6/6 validated by qPCR confirming predicted variation in levels across cell lines	6
Segemehl	• Improved sensitivity in simulated data • Identification of circular, *trans*-spliced and fusions	PE and SE	Segemehl	No restrictions on splice sequence or intron length	• R: only highest ranking chain; chain must cover ≥80% of read; single alignment picked per seed to minimize genomic distance of chain	circRNAs lacking high-quality seeds or not included in alignment obtained by greedy chaining, such as circRNAs in genes with homologous exons	• 85% recall and 98% precision in simulated data • Segemehl detected 19/19 circRNAs	25

Algorithm	Improvements	Read type	Aligner	Junctions considered	Filtering rules: per read (R) or per junction (J)	Blind spots	Reported validation rate	Refs
							validated by REF. 11	
Guo 2014	Quantifies and enables filtering by circular/linear ratio	PE and SE	Bowtie	GT–AG, donor or acceptor within 100kb	• J: ratio of circular to linear reads (circular fraction ≥10%)	circRNAs expressed at low levels relative to linear host (authors report 2/3 of their candidates filtered out by 10% rule, 50% filtered out in data from REF. 11, and 90% in data from REF. 3 and REF. 6)	NA	26
circExplorer	More sensitive than MapSplice, as sensitive as Segemehl but requires ~23-times less memory	PE and SE	TopHat Fusion	UCSC KnownGene annotated exons within single gene	• R: aligns uniquely	circRNAs using unannotated exons or comprising multiple genes	• 7/7 validated by combination of RT-PCR, northern blot on denaturing PAGE and RNase R resistance	4
circRNA_finder	STAR aligner is optimized for speed enabling analysis of >100 rRNA-depleted libraries across multiple tissues and developmental time points	PE and SE	STAR	Annotated exons within gene or intergenic GT–AG; donor or acceptor within 100kb	• R: <4 mismatches; 15–20nt overlap depending on read length; unique alignment	circRNAs expressed at moderate to low levels	• 4/4 validated by northern blot, depletion in poly(A)$^+$ RNA or RNase R resistance • 8/10 resistant to RNase R as measured by qPCR	7
CIRI	• *De novo* detection without 2-segment alignment allows detection of circRNAs using small exons	PE and SE	BWA-MEM	GT–AG	• J: filter circRNAs in homologous genes or repeat regions and those lacking PCC signal	circRNAs using non-canonical splice signals	• 24/33 (73%) circRNAs selected from circRNAs with >5 reads validated by PCR	27

Algorithm	Improvements	Read type	Aligner	Junctions considered	Filtering rules: per read (R) or per junction (J)	Blind spots	Reported validation rate	Refs
	• Simulator to generate linear and circular reads						• 5/5 using exons <70 nt validated by PCR	
KNIFE	• Improved statistical score per circle • Combined annotation - dependent and - independent approach increased sensitivity without increasing FPs	PE and SE	Bowtie and Bowtie2	UCSC KnownGene annotated exons within 1Mb	• R: <4% of read bases are mismatches; circular junctional read does not align to linear junction; user-specified junction overlap; R2 maps within circle; unique alignment with ≤2 mismatches for *de novo* • J: high posterior probability	• circRNAs in regions of genomic variation • circRNAs with SNVs • circRNAs between exons >100kb apart	• 13/13 validated by RNase R resistance • 14/14 by qPCR • 5/5 *de novo* circRNAs by PCR and Sanger sequencing	5
DCC	Uses circRNA versus host gene expression levels to test host gene independence	PE and SE	STAR	GT–AG	• R: R2 maps within circle • J: remove candidates in repetitive or homologous genes	• circRNAs sensitive to RNaseR (for example, *CDR1as*) • Lowly expressed circRNAs • Non-canonical splice signals	97% precision (TP defined as those detected in RNase R and rRNA-depleted; FP defined as those detected in rRNA-depleted only)	28

ASTD, Alternative Splicing and Transcript Diversity; BWA-MEM, Burrow–Wheeler aligner; *CDR1as*, *CDR1* antisense RNA; circRNA, circular RNA; FDR, false discovery rate; FP, false positive; NA, not applicable; PAGE, polyacrylamide gel electrophoresis; PCC, paired chiastic clipping; PE, paired end; qPCR, quantitative PCR; RNA-seq, RNA sequencing; rRNA, ribosomal RNA; RT-PCR, reverse transcription PCR; SE, single end; SNV, single-nucleotide variant; TP, true positive; UCSC, University of California, Santa Cruz.

[*] Algorithm-specific criteria for which junctions are considered and which reads or junctions are filtered as false positives before reporting results are listed, along with the computational improvements over prior algorithms and blind spots for each algorithm.

[‡] Later versions modified to improve reporting of circRNA.

Table 2

Filtering criteria for selection of high-confidence circRNAs

Algorithm	Criteria[*]	Novel findings[‡]	Refs
Salzman 2012	≥1 junctional reads	• circRNAs from 481 genes expressed at similar level to linear transcripts and additional 399 genes with circRNAs constituting >10% of transcripts in leukocytes • circRNAs in 10% of expressed genes in leukocytes • circRNAs are not polyadenylated • circRNAs are enriched for exon 2 and longer flanking introns • Cytoplasmic localization	1
CircRNAseq	• Enriched by RNase R in 2/2 samples • Three stringency levels from ≥1 read to ≥10 SRBPM[§] in RNase R$^-$ libraries	• Long flanking introns and complementary *Alu* elements are highly correlated with circularization • Single circle exons are ~3× longer than average • >25,000 circRNAs detected in fibroblasts • circRNAs are more stable than linear isoforms • Introns are spliced out of most circRNAs • circRNAs in 14% of expressed genes in fibroblasts • Mouse and human circRNAs in homologous genes use the same exons more often than expected by chance • No evidence of circRNA translation • siRNA can target circRNAs • circRNA-producing genes are enriched for protein kinases	3
find_circ	≥2 junctional reads	• *CDR1as* acts as mir-7 sponge in brain • Tissue- and development-specific expression • circRNA enriched in conserved nucleotides • Most circRNAs from CDS exons and contain 1–5 exons	11
Salzman 2013	FDR ≤0.025 and ≥1 junctional reads for ENCODE or ≥2 junctional reads for *Drosophila melanogaster*	• Cell type-specific circRNAs and ratios of circular to linear isoform expression • circRNA in humans ~1% of mRNA level, with most circRNAs ~5–10% of linear isoform from host gene • ~47,000 isoforms from ~8,500 genes • Most circRNAs transcribed from same strand as linear isoforms • Linear and circular isoforms not correlated	6
Guo 2014	Circular isoform ≥10% of linear RNA from gene in ≥2 samples	• ~20% of circRNAs have intron retention in CD34$^+$ cells • Identification of 57 circRNAs that are ≥50% of total circular and linear transcripts across most cell types • Highly expressed circRNAs are not more cell type specific than mRNA • Catalogue of 7,112 human circRNAs with expression ≥10% of linear	26
circExplorer	≥5 reads in either poly(A)$^-$ or poly(A)$^-$ and RNase R$^+$ libraries	• Non-repetitive complementary sequences promote circularization, >50% of genes with circRNA produce multiple circular isoforms (alternative circularization)	4

Algorithm	Criteria[*]	Novel findings[‡]	Refs
		• Exons of single-exon circRNAs are larger than exons in multi-exon circRNAs, most circRNAs contain 2–3 exons and usually not the first or last exon of a gene	
circRNA_Finder	≥10 reads; ≥2 reads for gene-level conservation analysis	• circRNA accumulates in the ageing *D. melanogaster* brain • *D. melanogaster* circRNA is enriched for conserved microRNA seeds • 2,500 high-confidence *D. melanogaster* circRNAs • Gene-level conservation of circRNA in heads of three *Drosophila* species • Last exon in circRNA biased to 5′ end of gene • circRNAs are flanked by long introns but circularization in *D. melanogaster* is not driven by flanking intronic sequences	7
CIRI	≥5 junctional reads	• Prevalence of intronic or intergenic circRNAs (estimated as 20% and 5% of circRNAs from ENCODE data, respectively) • Greater variation in circRNAs than in linear isoforms in cancer cell lines • More highly expressed circRNAs detected in more cell lines	27
KNIFE	Annotation -dependent: ≥1 or ≥2 junctional reads depending on sequencing depth and statistical score	• Global and tissue-specific circRNA induction detected in human fetal development • circRNA is spliced by minor (U12) spliceosome • *NCX1* induction in hES cells recapitulated	5
DCC	≥5 reads, detected in at least 6/18 samples	• Catalogue of 72 circRNAs in mouse brain with temporal expression during development independent of host gene expression	28

CDR1as, *CDR1* antisense RNA; CDS, coding DNA sequence; circRNA, circular RNA; ENCODE, Encyclopedia of DNA Elements; FDR, false discovery rate; hES cells, human embryonic stem cells; *NCX1*, sodium/calcium exchanger 1; siRNA, small interfering RNA; SRPBM, spliced reads per billion mapped.

[*] Criteria used to select the subset of high-confidence circRNA from all circRNAs reported based on the criteria listed in TABLE 1.

[‡] Genome-wide novel findings reported based on these circRNAs in the original publication for each algorithm.

[§] Calculated as (spliced reads/total mapped reads) × 10^9.

Table 3

Methods used to assess the genome-wide accuracy of algorithms

Method	circRNA specific	Benefits	Experimental limitations	Bioinformatic limitations	Refs
RNase R Resistance	Yes	Enriches for circRNA by degrading linear RNA, making it easier to detect lowly expressed circRNAs	• Requires matched RNase R- and mock-treated libraries • Some validated circRNA sensitive to RNase R • Variability between RNase R-treated replicates	• Inaccurate conclusions from read count fold change without considering confidence intervals • Appropriate normalization procedures need to be developed	3,4, 27,28, 41
Depletion in poly(A)$^+$ libraries	Yes	Uses expected depletion profile to assess results	• Requires matched poly(A)$^+$ and poly(A)$^-$ libraries • Variability in detection of lowly expressed circRNAs	• Inaccurate conclusions from read count fold change without considering confidence intervals • Appropriate normalization procedures need to be developed	5,7,26
Decoy reads [*]	No	• Rules out experimental and alignment artefacts • Can be used to identify artefacts within non-decoy reads	Experimental and alignment artefacts can generate reads consistent with circRNA so FP rate can be underestimated	• ‘Decoy’ reads under one model may be consistent with an alternative model not evaluated • Only applicable to PE data	5,7
RT specificity	No	Rules out FP from template switching	• Varying results reported by different groups • High FN rate[22]	Cannot rule out FP from sequencing and alignment errors	46
Simulated data	No	Known truth for evaluating sensitivity and specificity based on known sources of error		Unclear how this translates to real sequencing data where there are additional unmodelled biases or unknown sources of error	11,25, 27,53, 58

circRNA, circular RNA; FN, false negative; FP, false positive; PE, paired end; RT, reverse transcriptase.

[*] Mate alignments inconsistent with isoform inferred by junctional alignment.