Large-scale benchmarking of circRNA detection tools reveals large differences in sensitivity but not in precision

Marieke Vromman[1], Jasper Anckaert[1], Stefania Bortoluzzi[2], Alessia Buratin[2], Chia-Ying Chen[3], Qinjie Chu[4], Trees-Juen Chuang[3], Roozbeh Dehghannasiri[5], Christoph Dieterich[6], Xin Dong[7], Paul Flicek[8], Enrico Gaffo[2], Wanjun Gu[9], Chunjiang He[7], Steve Hoffmann[10], Osagie Izuogu[8], Michael S. Jackson[11], Tobias Jakobi[12], Eric C. Lai[13], Justine Nuytens[1], Julia Salzman[5], Mauro Santibanez-Koref[11], Peter Stadler[14], Olivier Thas[15], Eveline Vanden Eynde[1], Kimberly Verniers[1], Guoxia Wen[16], Jakub Westholm[17], Li Yang[18], Chu-Yu Ye[4], Nurten Yigit[1], Guo-Hua Yuan[19], Jinyang Zhang[20], Fangqing Zhao[20], Jo Vandesompele[1]*, Pieter-Jan Volders[1]*

* Both authors contributed equally to this work.

All authors have been ranked alphabetically according to their surname, except the first author and the two shared last authors.

[1] OncoRNALab, Cancer Research Institute Ghent (CRIG), Department of Biomolecular Medicine, Ghent University, Belgium

[2] Department of Molecular Medicine, University of Padova, Italy

[3] Genomics Research Center, Academia Sinica, Taiwan

[4] Institute of Crop Science & Institute of Bioinformatics, Zhejiang University, China

[5] Department of Biomedical Data Science and of Biochemistry, Stanford University, USA

[6] Klaus Tschira Institute for Integrative Computational Cardiology, Department of Internal Medicine III, University Hospital Heidelberg, German Center for Cardiovascular Research (DZHK), Germany

[7] School of Basic Medical Science, Department of Medical Genetics, Wuhan University, China

[8] EMBL-EBI, UK

[9] Collaborative Innovation Center of Jiangsu Province of Cancer Prevention and Treatment of Chinese Medicine, School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, China

[10] Computational Biology Group, Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), Jena, Germany

[11] Biosciences Institute, Faculty of Medical Sciences, Newcastle University, UK

[12] Translational Cardiovascular Research Center, University of Arizon - College of Medicine Phoenix, USA

[13] Developmental Biology Program, Sloan Kettering Institute, New York, USA

[14] Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Germany

[15] Data Science Institute, I-Biostat, Hasselt University, Belgium

[16] State Key Laboratory of Bioelectronics, School of Biological Sciences and Medical Engineering, Southeast University, China

[17] Dept of Biochemistry and Biophysics, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University, Sweden

[18] Center for Molecular Medicine, Children's Hospital, Fudan University and Shanghai Key Laboratory of Medical Epigenetics, International Laboratory of Medical Epigenetics and Metabolism, Ministry of Science and Technology, Institutes of Biomedical Sciences, Fudan University, China

[19] CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, China

[20] Beijing Institutes of Life Science, Chinese Academy of Sciences, China

Abstract

The detection of circular RNA molecules (circRNAs) is typically based on short-read RNA sequencing data processed by computational detection tools. During the last decade, a plethora of such tools have been developed, but a systematic comparison is missing. Here, we set up a circRNA detection tool benchmarking study, in which 16 tools were used and detected over 315,000 unique circRNAs in three deeply sequenced human cell types. Next, 1,516 predicted circRNAs were empirically validated using three orthogonal methods. Generally, tool-specific precision values are high and similar (median of 98.8%, 96.3%, and 95.5% for qPCR, RNase R, and amplicon sequencing, respectively) whereas the number of predicted circRNAs is the largest tool differentiator (ranging from 1,372 to 58,032). Furthermore, we demonstrate the complementarity of tools through the increase in detection sensitivity by considering the union of highly-precise tool combinations while keeping the number of false discoveries low. Finally, based on the benchmarking results, recommendations are put forward for circRNA detection and validation.

Introduction

Circular RNAs (circRNAs) are a class of non-coding RNA molecules numerously present in humans and other eukaryotic species. For a long time, circRNAs were regarded as unimportant byproducts of splicing. However, since the advancement of RNA sequencing technologies and the development of circRNA detection bioinformatics pipelines, there has been a significant increase in circRNA research, with a compound annual growth rate of scientific publications of 181% over the last five years (Figure 1A) (1).

Although an *in vivo* function for most circRNAs remains unknown and functional analyses are typically restricted to *in vitro* experiments, some circRNAs have been linked to specific diseases, including cancer. CircRNAs have also been reported to be more stable than linear transcripts due to the absence of a free 5' or 3' end to be recognized by exonucleases (1). In line with this, a higher fraction of circRNA relative to linear RNA has been observed in a wide range of human biofluids, which makes them interesting biomarker candidates, with the potential to be used for minimally-invasive tests for diagnosis or response monitoring (2). Wang *et al.* reviewed 112 differentially expressed circRNAs in various biofluids from patients with different cancer types (3). Furthermore, 15 clinical trials incorporating circRNAs as disease biomarkers have been initiated (ClinicalTrials.gov).

Eukaryotic circRNAs are formed through a process called back-splicing, where the 5' end of an RNA molecule forms a covalent bond with its own 3' end, forming a circular molecule with a characteristic back-spliced junction (BSJ) sequence (Figure 1B) (1). CircRNAs comprise of one or multiple exons, and analogous to linear RNA, there is alternative splicing of circRNAs, where circRNAs with the same BSJ sequence may have a different exon (and/or intron) composition (1).

In a targeted manner, circRNAs can be quantified using RT-qPCR (reverse transcription quantitative polymerase chain reaction) using BSJ-spanning primer pairs to amplify the region flanking the BSJ (Figure 1C). These primer pairs are divergent (facing away from each other) when hybridizing to the linear host transcripts and can therefore only amplify the circRNA (4). However, false positives resulting from alignment ambiguity, repeat sequences, trans-splicing, or RT template-switching artifacts have been described (5, 6). In all these cases, a linear RNA molecule is formed with the same exon orientation and, therefore, the same sequence as the circRNA BSJ. To prevent false-positive circRNA identifications, linear RNA is often digested with the exonuclease ribonuclease R (RNase R) followed by RT-qPCR. RNase R typically degrades linear RNA, whereas circRNAs are generally not affected. Of note, it has been suggested that long circRNAs may be somewhat sensitive to RNase R degradation and various challenges of validating circRNAs have been recognized (7, 8).

In general, high-throughput or exploratory circRNA detection is performed using bioinformatics approaches that analyze total RNA sequencing data. For this, the RNA sequencing reads are first mapped against a reference genome. The unmapped reads are subsequently used to identify BSJ-spanning reads that map divergently (in reverse order) on the linear genome (Figure 1D).

Over the last decade, numerous computational circRNA detection tools have been developed and tested. Whereas multiple sets of circRNA detection tools using a bioinformatics approach have been compared (9–11), a systematic and comprehensive evaluation of many circRNA detection tools using an orthogonal validation method is still missing. In our benchmarking study, we aimed to evaluate all currently available circRNA detection tools with an orthogonal approach using RT-qPCR, RNase R, and amplicon sequencing (Figure 2A). Our study highlights that although the precision of the tools is generally excellent, their sensitivities are highly variable.

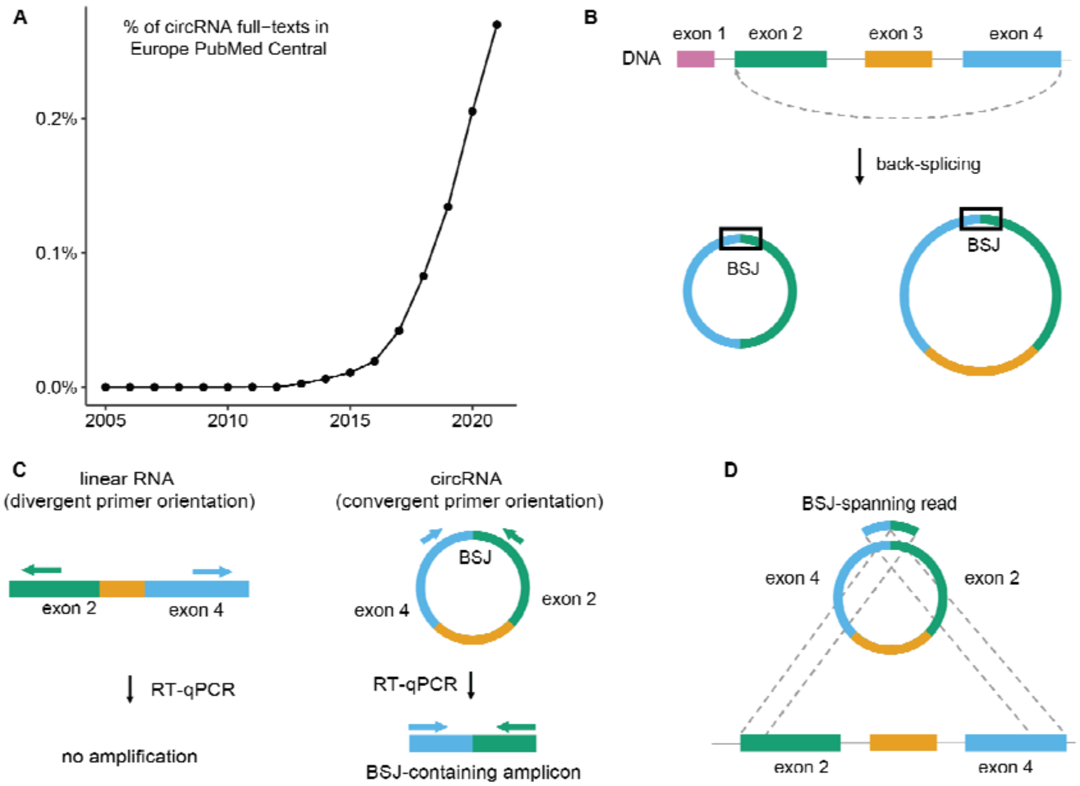


Figure 1 CircRNA scientific relevance, structure, and detection. **A.** Over the last decade, circRNA research has increased rapidly, as illustrated by the proportional growth of publications mentioning circRNA in Europe PubMed Central. **B.** CircRNAs are formed through back-splicing, which results in a circular molecule with a back-spliced junction (BSJ). Black boxes highlight the BSJ in the circRNA isoforms. **C.** CircRNAs can be detected with RT-qPCR using a BSJ-specific primer pair. The primer pair can only bind in a divergent manner (facing away from each other) to linear RNA, where no amplification will be possible, yet binds the circRNA in a convergent manner (facing towards each other), amplifying the BSJ sequence. **D.** Large-scale circRNA detection is typically performed using total RNA sequencing datasets and specialized computational tools. These tools identify BSJ-spanning reads, which map divergently (in reverse order) on the linear reference genome.

Results

CircRNA detection tools predict a wide variety of circRNAs

CircRNA detection tools differ in detection strategies and filtering

For this study, 16 different circRNA detection tools were included (Table 1, Supplementary Table 1) (11–25). CircRNA detection tools differ in their circRNA detection approach (including strand assignment), reliance on linear annotation, and filtering methods. CircRNAs can be detected from RNA sequencing data using the pseudo-reference-based approach (also called the candidate-based approach) or the fragmented-based approach (also called the segmented-read-based approach) (26). The former approach uses a reference list of potential BSJ sequences, often based on all possible combinations of known annotated exons within a gene. This approach is therefore limited to species with annotated genomes and to previously annotated genes and will only detect circRNAs that use the same splicing sites as the linear RNAs. The latter approach splits unmapped sequencing reads into shorter sequences and remaps these against the reference genome. Lastly, integrative tools, such as CirComPara2 and ecircscreen, combine the results of multiple tools.

The number of detected circRNAs differs greatly between tools

A total of 315,312 unique circRNA predictions (corresponding to 1,137,099 unique circRNA/strand/tool/sample tuples) were detected using 16 different tools based on deeply sequenced total RNA from three human cancer cell lines (Supplementary Table 2, because of large size, only available on https://github.com/OncoRNALab/circRNA_benchmarking). There is a striking almost 40-fold difference between the tool with the highest number of predicted circRNAs (circseq_cup with 58,032 circRNAs) and the tool with the lowest number of predicted circRNAs (segemehl with 1,372 circRNAs) for one of the cell lines (Figure 2B shows results for HLF cells, similar results for the other cell lines are shown in Supplementary Figure 1). Tools were further compared based on different metrics, including predicted circRNA length, strand information, correspondence to linear annotation, and predicted exon composition (Supplementary Data 1-4, Supplementary Figure 2-5). No notable differences were observed among tools, except for CIRI2 and PFv2 having a higher number of circRNAs for which no canonical linear annotation match was found compared to the other circRNA detection tools. Across all tools, 53.7% of circRNAs uniquely match one canonical linear transcript, 10.3% match more than one canonical transcript, and 35.9% do not match any canonical transcript. CircRNAs were found for 17,461 different human genes, demonstrating the pervasive nature of back-splicing.

Most circRNAs are characterized by low BSJ counts

CircRNA abundance is reflected by the BSJ count, which is the number of reads uniquely assigned to a given circRNA. The majority of circRNAs (86.6%) are detected with a BSJ count below 5 (Figure 2B), with only 46.1% of the detected circRNAs being observed with at least 2 BSJ counts (detailed distribution in Supplementary Figure 6). To increase confidence, circRNA_finder and segemehl filtered their results to report only circRNAs with a BSJ count of at least 5, and CirComPara2 and KNIFE filtered for circRNAs with a BSJ count of at least 2. Circtools filtered circRNAs with at least 2 counts in at least 2 samples. Of note, Sailfish-cir does not report raw BSJ counts, but transcripts per million (TPM) instead. The similarity of circRNA BSJ counts between tool pairs is reasonable, according to correlation analysis (median $r^2 = 0.71$, median slope = 0.75, p-value < 0.001, Supplementary Figure 7).

Half of the circRNAs are reported by only one circRNA detection tool

Half of all circRNAs in this study (49.9%) are reported only by one tool, which is largely due to circseq_cup's high number of uniquely predicted circRNAs (Figure 2C for HLF cells, similar results for the other cell lines are shown in Supplementary Figure 8). The overlap of circRNA predictions among different tools is visualized in a heat map for each cell line in Supplementary Figure 9.

tool	approach	circRNAs detected in ...	strand assignment[+]	splicing	BSJ count filter*	min circRNA length*	max circRNA length*
CIRCexplorer3	segmented read-based	entire genome	based on linear annotation	AGNGT	none	none	none
CirComPara2	integrative[x]	entire genome	no strand reported	AGNGT, ACNCT	≥ 2	299	2,304,996
circRNA_finder	segmented read-based	entire genome	based on mapping to genome	AGNGT	≥ 5	200	100,000
circseq_cup	based on segemehl, with full-length circRNA assembly	entire genome	no strand reported	non-canonical	none	none	5,000
CircSplice	segmented read-based	known splice sites	based on linear annotation	AGNGT, ACNCT	none	78	none
circtools	segmented read-based	entire genome	based on mapping to genome	AGNGT	≥ 2 in ≥ 2 samples	31	1,000,000
CIRI2	segmented read-based	entire genome	based on GT-AG splice sites	AGNGT	none	135	200,000
CIRIquant	based on CIRI2, with improved quantification	entire genome	based on GT-AG splice sites	AGNGT	none	135	200,000
ecircscreen	integrative[x]	entire genome	based on consensus from tools	AGNGT	none	none	none
find_circ	segmented read-based	entire genome[-]	based on mapping to genome[-]	AGNGT	none	none	100,000
KNIFE	candidate-based	entire genome	based on linear annotation	non-canonical	≥ 2	none	1,000,000
NCLscan	candidate-based	known splice sites	based on linear annotation	non-canonical	none	100	none
NCLcomparator	filtered results of NCLscan	known splice sites	based on linear annotation	non-canonical	none	100	none
PFv2	segmented read-based	entire genome	based on mapping to genome	AGNGT, ACNCT	none	50	1,000,000
Sailfish-cir	based on CIRI2 v2.0.6	entire genome	no strand reported	AGNGT, ACNCT	no BSJ counts reported	135	200,000
segemehl	segmented read-based	entire genome	no strand reported	non-canonical	≥ 5	none	200,000

Table 1 CircRNA detection tools with their circRNA detection approach, strand assignment approach, reliance on linear annotation, and filtering approach. [x] Integrative tools combine the results of multiple circRNA detection tools. This includes CirComPara2 (combining CIRCexplorer2 v2.3.8 (on any of BWA v0.7.15, TopHat2 v2.1.0, STAR v2.6.1e, and Segemehl v0.3.4), CIRI2 v2.0.6, DCC v0.4.8, and find_circ v1.2, and the filtering all circRNAs detected by at least two methods) and ecircscreen (combining CIRI2 v2.0.6, circRNA_finder v1.2, PFv2 v2.0.0, find_circ v1.2, and CIRCexplorer v1.1.10, and then filtering all circRNA detected by at least three methods). [+] Some tools did not report strand information for this study, but the (updated) circRNA tool might report circRNA strand information. * The BSJ count, and minimum and maximum circRNA length filters, are the filters used for this specific study. The user can choose these parameters freely. Of note, the minimum and maximum length filters are based on the estimated circRNA length with introns, calculated by subtracting the start position from the end position of the BSJ. [-] Inferred based on publication and available code.

Canonical splice sites characterize the majority of predicted circRNAs

Out of 16 circRNA detection tools, 8 exclusively report circRNAs flanked by canonical splice sites (with an AGNGT pattern, where AG is the splice acceptor, N represents the circRNA sequence in between, and GT is the splice donor) (Figure 2D). CirComPara2, circseq_cup, Sailfish-cir, and segemehl do not report circRNA strand orientation, explaining most of the ACNCT patterns. Lastly, KNIFE, NCLscan, and NCLcomparator also report a substantial number of circRNAs with a GGNGG splicing pattern.

Two-thirds of predicted circRNAs are novel

Two-thirds of all predicted circRNAs in this study (68.5%) are novel compared to a set of previously reported circRNAs extracted from 13 published circRNA databases (Circ2Disease, circad, CircAtlas, circbank, circBase, CIRCpediav2, CircR2disease, CircRiC, circRNADb, CSCD, exoRBase, MiOncoCirc, and TSCD) (Supplementary Figure 10) (27). Of note, approximately half of these novel circRNA candidates originate solely from circseq_cup. Looking at the tools individually, circseq_cup, KNIFE, NCLscan, and NCLcomparator report a higher number of novel circRNAs (87.8%, 53.9%, 53.4%, 53.3%, respectively) compared to the other tools (median 19.7%, interquartile range (IQR) 4.9-34.8%).

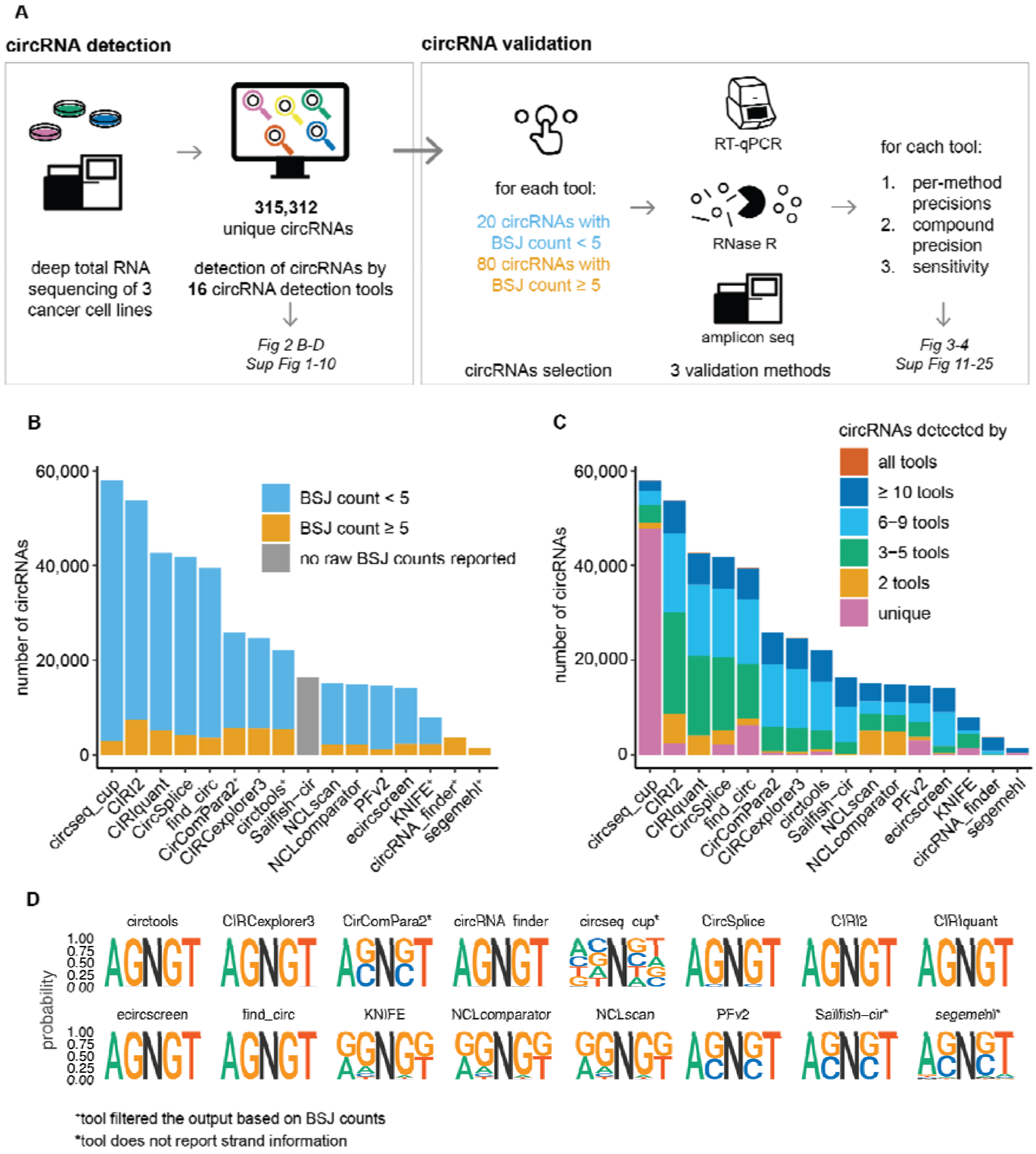


Figure 2 CircRNA detection tools predict a wide variety of circRNAs **A.** This study consists of a circRNA detection phase and a circRNA validation phase. For the former, 16 circRNA detection tools were used to predict circRNAs in three deeply sequenced cancer cell lines. For the latter, a set of circRNAs was selected per tool and validated using three orthogonal methods, generating tool-specific precision values for each method. These values were also used to compute a compound precision value. **B.** The number of reported circRNAs differs greatly between tools (shown for HLF cells, similar results for the other cell lines are shown in Supplementary Figure 1). The tools are ordered according to the total number of predicted circRNAs. The vast majority of circRNAs are predicted with a BSJ count below 5 (in blue). Two tools, circRNA_finder and segemehl, filtered their results to report only circRNAs with a BSJ count of at least 5 (in orange). C**.** The majority of circRNAs (49.9%) are detected by only one tool. Circseq_cup reports the largest set of unique circRNAs (shown for HLF cells, similar results for the other cell lines in Supplementary Figure 8). A small set of 55 circRNAs is detected by all tools (column n_db in Supplementary Table 2). **D.** CircRNA splice sites

differ among circRNA detection tools. Most commonly, the canonical AGNGT pattern is observed, with AG being the splice acceptor, N the circRNA, and GT the splice donor. Circseq_cup, CirComPara2, Sailfish-cir, and segemehl do not report strand information. To be able to retrieve a splicing sequence for the circRNAs from these tools, it was assumed the circRNA originated from the positive strand. This led to the ACNCT pattern (reverse complement of AGNGT), most probably from circRNAs that were assigned to the positive strand incorrectly. Lastly, there are some tools displaying the GGNGG pattern.

CircRNA validation with empirical methods

CircRNA primer design inherently introduces a selection bias

Based on previous experiments (Supplementary Data 5, Supplementary Figure 11), for each tool, we aimed to select 80 random *high-abundance* circRNAs with a BSJ count of at least 5 and 20 random *low-abundance* circRNAs with a BSJ count below 5. Of note, circRNA primer design inherently introduces a bias (Supplementary Data 6, Supplementary Figure 12). A selection of 1,560 circRNAs was obtained (Supplementary Table 3, BSJ count distribution in Supplementary Figure 13). As some circRNAs were selected more than once (by chance, for different tools or in different cell lines), the total number of unique circRNAs/sample pairs was 1,516, from here on termed *selected circRNAs* (detailed description in Supplementary Figure 14). An overview of the circRNA validation results and (compound) precision values described below are represented in Supplementary Figure 15 and Supplementary Table 4, respectively. Reproducibility evaluations were performed and are described in Supplementary Data 7 (Supplementary Figure 16 and 17) and Supplementary Data 8 (Supplementary Figure 18).

High circRNA detection precision using RT-qPCR validation

Of the 1,516 selected circRNAs, 1,479 (97.6%) could be validated with RT-qPCR, i.e., the primer pair flanking the BSJ site resulted in a detectable amplicon. For the *low-abundance* circRNAs there is some variation in the tool-specific precision values (median 95.0%, range 80.0-100%), which is expected (Figure 3A). *High-abundance* circRNAs have high RT-qPCR precision values for most tools (median 98.8%, range 90.0-100%). It is important to note that RT-qPCR-based validation is the net result of a successful primer pair and the actual presence of a sufficiently abundant circRNA in the amount of RNA tested.

One in sixteen predicted circRNAs fail validation upon RNase R treatment

RNase R was used as a second, more stringent validation approach. RNase R selectively degrades linear transcripts, ensuring the RT-qPCR primers amplify a circular molecule. For 112 out of 1,516 selected circRNAs (7.4%), RNase R treatment could not be evaluated, as their abundance in the untreated sample was too low, leaving no room to confirm RNase R degradation in the event of a false-positive circRNA (hence labeled as NAs). In the remaining set of 1404 predicted circRNAs, 1319 circRNAs (93.9%) could be successfully validated using RT-qPCR on RNase R treated RNA. For most tools, high RNase R precision values were observed for *high-abundance* circRNAs (median 96.3%, range 74.0-100%) (Figure 3A). PFv2 displays the lowest precision (74.0%). For *low-abundance* circRNAs, reduced precision values were observed (median 86.7%, range 50.0-100%). Of note, the number of circRNAs per tool in this bin is lower than the original 20 that were selected, as more circRNAs were excluded because of too low abundance (resulting in only 10-18 circRNAs per tool, with a median of 14 circRNAs).

Amplicon sequencing is the most stringent orthogonal validation method

The RT-qPCR amplicons of the untreated RNA were sequenced for further validation of the circRNAs (Figure 3A). A random subset of circRNAs (337/1,516, 22.2%) was not included in the amplicon sequencing experiment (hence labeled as NAs). For the remaining 1,179 circRNAs, 1,014 circRNAs (86.0%) could be readily validated with amplicon sequencing, i.e., the majority of reads aligned to the expected BSJ sequence. A cumulative plot of the amplicon sequencing precision values can be found in Supplementary Figure 19. Most tools have similar amplicon sequencing precision values for *high-abundance* circRNAs (median 95.5%, range 30.0-100%), with PFv2 displaying a very low (30.0%) amplicon sequencing precision value. Of note, as PFv2 was developed to retain repeat sequences, it is expected to result in more false positives. The most obvious are caused by linear read-through between exons in neighboring tandemly repeated gene clusters/interspersed repeats, and these tend to be abundant. For *low-abundance* circRNAs, performance is more diverse, with generally lower amplicon sequencing precision (median 73.3%, range 17.6-94.1%).

Different validation methods should be concurrently used to compensate for their inherent limitations

Although the three validation strategies were used independently, it is interesting to evaluate to what extent they support each other (Figure 3B). Considering 1,103 circRNAs for which all three validation results are available 957 circRNAs (86.8%) pass all validation methods, 128 circRNAs (11.6%) fail one or two of the validation methods, and 18 circRNAs (1.6%) fail all three validation methods. These observations show that orthogonal validation with more than one empirical approach is important. It is beyond the scope of this study to investigate why there are some discrepancies among the validation results (some hypotheses are discussed in Supplementary Discussion 1). First, they are rare (for most circRNAs, the different methods completely agree). Second, the same methods are used to compare the tools, whereby no tool should be favored over the other.

The theoretical number of true positive circRNAs can be estimated based on the compound precision value

All three validation methods can be combined into one compound precision value per tool (Supplementary Figure 20). The theoretical number of true positive circRNAs for each tool can then be computed and used as an estimate for sensitivity (Figure 3C for HLF cells, similar results for the other cell lines are shown in Supplementary Figure 21). Of note, sensitivity can also be estimated based on the total number of true-positive circRNAs (n = 957) (Supplementary Figure 22, Supplementary Table 4). There is a significant correlation between the theoretical number of true-positive circRNAs and the estimated sensitivity (Spearman correlation of 0.92 with a p-value < 0.001 for *low-abundant* circRNAs, and a Spearman correlation of 0.59 with a p-value = 0.0211 for *high-abundant* circRNAs).

Comparing precision values in function of circRNA annotation

To compare precision values in function of circRNA annotation, we restrict the analyses to *high-abundance* circRNAs with information for all validation techniques. Furthermore, a strict validation definition was used, where all circRNAs failing for at least one technique were classified as unvalidated. CircRNAs previously described in databases have higher chances of getting validated (Chi-squared test with p-value < 0.001, odds ratio (OR) = 13.6). Nevertheless, false-positive circRNAs according to our data are still present in multiple published databases (Supplementary Figure 23). For example, false-positive circRNA chr6:47526627-47554766 (hg38, 0-based) is present in CircAtlas (as hsa-CD2AP_0048) and in exoRBase (as exo_circ_65199). A difference in validation rate in function of the splicing pattern was observed, with better validation of circRNAs surrounded by canonical splice sites (Chi-squared test with p-value < 0.001, OR = 5.0). Similarly, circRNAs that originate from a region with an annotated linear transcript have higher validation rates (Chi-squared test with p-value < 0.001, OR = 20.9). Surprisingly, single-exon circRNAs displayed significantly lower validation rates than multi-exon circRNAs (Chi-squared test with p-value < 0.001, OR = 3.8). Lastly, while tools with a 'candidate-based' approach seem more precise than tools using the 'segmented read-based' approach (Chi-squared test with p-value =

0.0087, OR = 2.4), we cannot be sure that these results are not confounded by other algorithmic differences.

Evaluation of tool detection sensitivity in function of circRNA annotation

There is a significantly higher sensitivity for tools reporting circRNAs surrounded by canonical splice sites, resulting in a median difference in sensitivity of 38.5% (Mann-Whitney U test with p-value < 0.001 and large effect size of 0.78, 95% CI [0.56 - 0.85]). No link could be found between estimated sensitivity and tool approach, use of linear annotation, or strand annotation status.

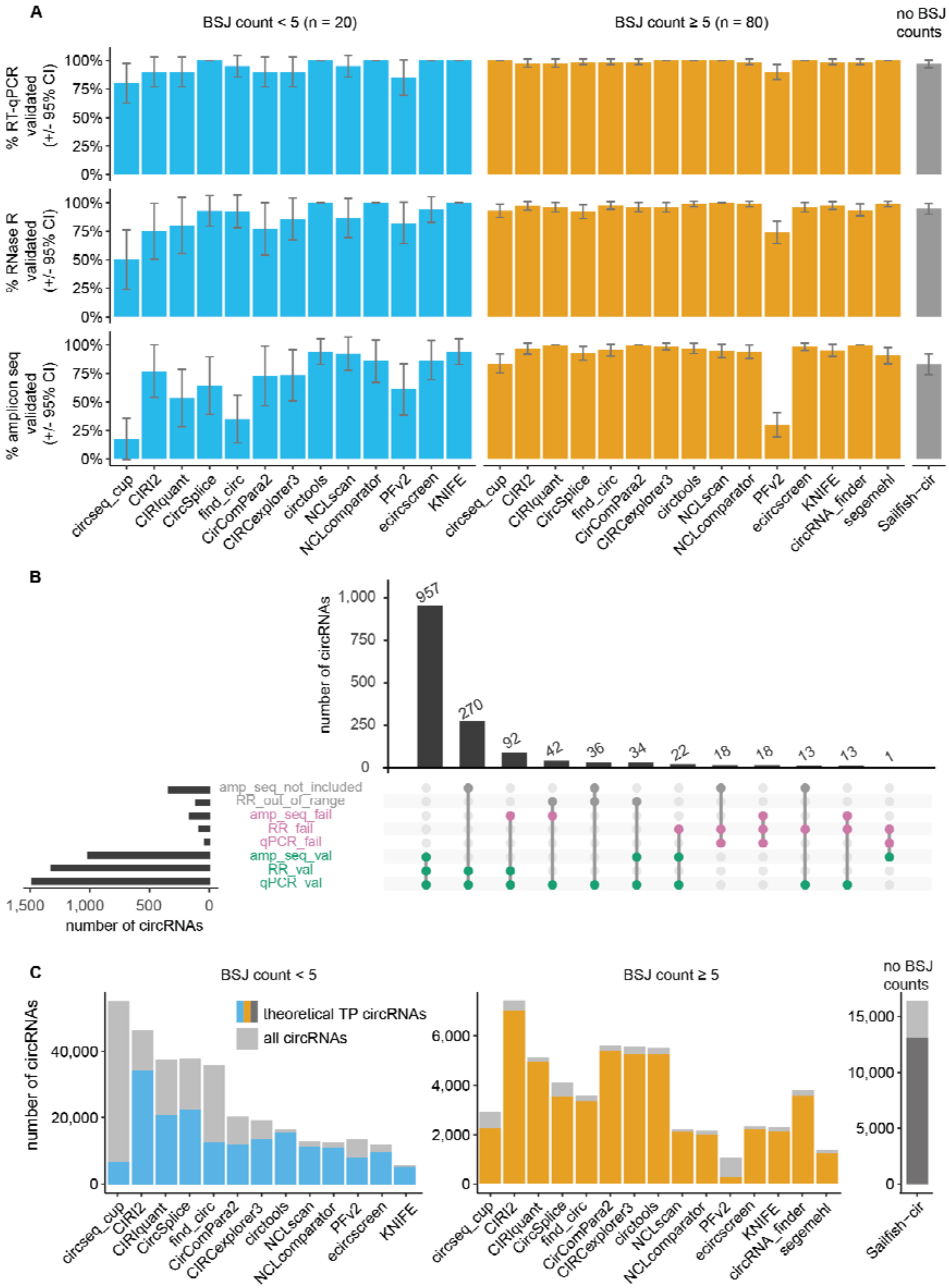


Figure 3 The precision of circRNA detection tools is generally high and similar, whereas tools largely differ with respect to the number of predicted circRNAs. The plots are separated based on circRNA BSJ count below 5 (*low-abundance*, in blue) or a BSJ count of at least 5 (*high-abundance*, in orange). Sailfish-cir reports TPM (transcripts per million) values, instead of BSJ counts and is therefore depicted separately. **A.** CircRNAs were validated using three different techniques: RT-qPCR detection, resistance to degradation by RNase R, and

amplicon sequencing. *Low-abundance* circRNAs are in general more difficult to validate. Of note, the precision values for *low-abundance* circRNAs are based on a limited set of circRNAs. *High-abundance* circRNAs have good precision values for most tools and most validation methods. The error bars represent the 95% confidence intervals (CI). **B.** The vast majority of circRNAs obtain the same verdict based on the three different validation methods. However, some circRNAs have conflicting results. For example, there are 13 circRNAs that are detectable by RT-qPCR but also degraded upon RNase R treatment and seem to amplify the wrong product. **C.** The compound precision value is used to compute the theoretical number of true positive circRNAs by multiplying it with the original number of circRNAs detected by that tool (shown for HLF, similar results for the other cell lines are shown in Supplementary Figure 21).

Evaluation of tool combinations to improve performance

For the combination of two or more tools, both the intersection and the union have been proposed (11) (Supplementary Tables 5 and 6). While not evaluated here, the increased time and resource consumption should also be taken into account when considering the use of multiple tools.

A circRNA predicted by two detection tools is not necessarily a true positive result

Figure 4A shows that circRNAs uniquely detected by a single tool generally have lower precision values. In line with this, circRNAs detected by at least two tools have a higher chance of getting validated (Chi-squared test with p-value < 0.001, OR = 57.3). On the other hand, out of 1,380 unique circRNAs detected by at least two tools, 7 circRNAs (0.5%) failed all three validation methods, and 137 (9.9%) failed at least one of the validation methods, illustrating that the practice of using the intersection is not a guarantee to avoid false positive results.

The union of high-precision circRNA detection tools substantially increases the number of true positive circRNAs

To maximize detection sensitivity and maintain precision, we evaluated the union of pairs or triples of circRNA detection tools. Generating all possible combinations of the better tools with individual compound precision ≥ 90% for *high-abundance* circRNAs (n = 11 tools) consistently results in higher detection sensitivity while maintaining a high weighted precision value. The median increase in the number of detected circRNAs (and interquartile range) is 36% (18.4-108.2%) and 69.4% (33.1 - 196.1%) for combinations of 2 or 3 tools, respectively. In other words, when combining very precise tools, the number of false positives does not counteract the gain in additional true positives. A subset of tool combinations with a ≥ 7.5% (~ 1,000 circRNAs) increase in the number of circRNAs due to the combination is shown in Figure 4B (shown for HLF cells, similar results for the other cell lines are shown in Supplementary Figure 24; the combo of three tools shown in Supplementary Figure 25). One obvious consideration when selecting two different tools is their circRNA detection approach, their reliance on linear annotation, and their filtering methods.

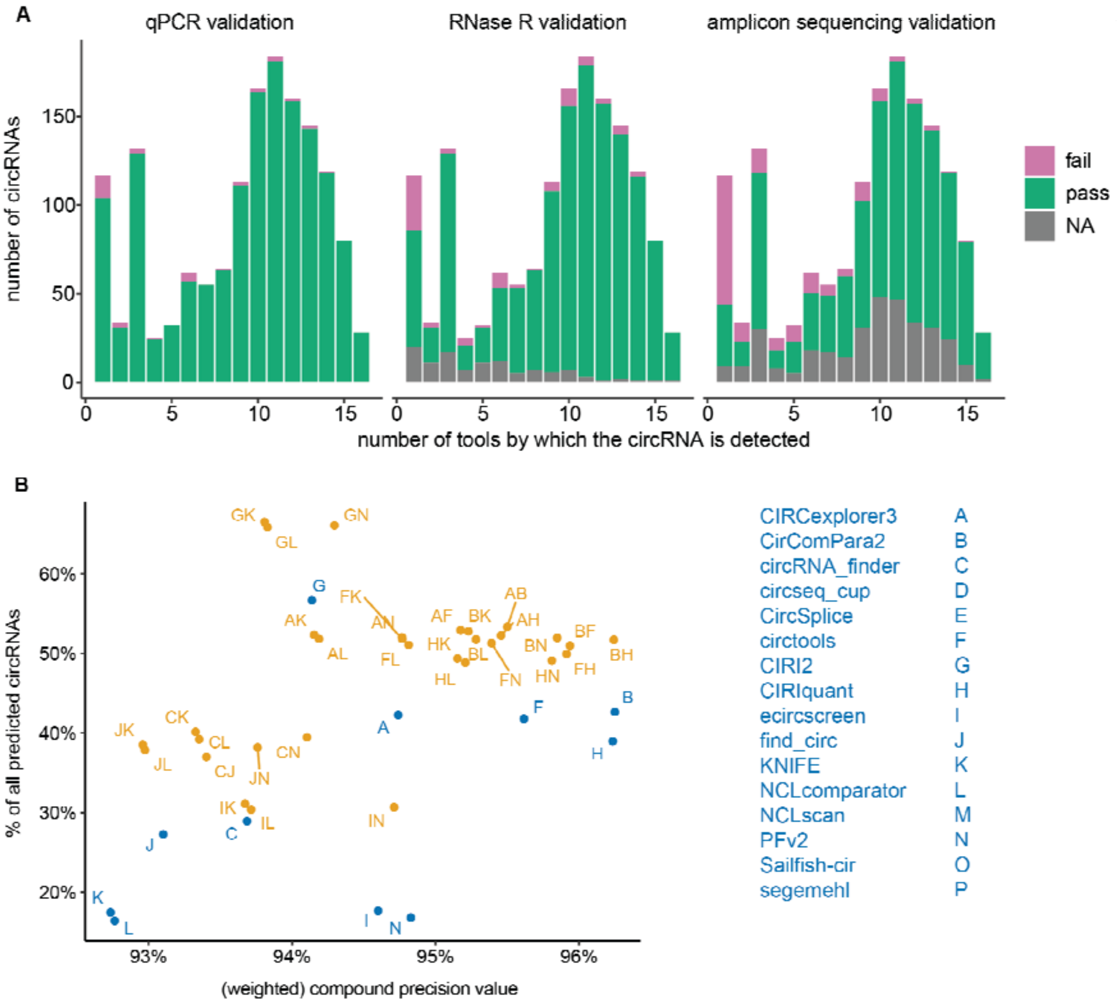


Figure 4 The intersection or union of two circRNA detection tools decreases the number of false positives, or increases the overall number of detected circRNAs, respectively. **A.** CircRNAs detected by multiple tools generally have higher precision values. However, the often-used practice of using the intersection of two tools is not necessarily a guarantee for avoiding false positive results. **B.** By considering the union of two circRNA detection tools, the number of circRNAs can be significantly increased, whilst keeping the number of false positive predictions low (shown for the HLF cell line, similar results for the other two cell lines are shown in Supplementary Figure 24). For the y-axis, the percentage of detected circRNAs is calculated by dividing the number of circRNA detected by that tool combination by the total number of predicted circRNAs for that sample taking the union of all tools (13,087 for the HLF sample). For this analysis, the compound precision value of *high-abundance* circRNAs was used due to its higher precision. Some circRNA detection tools are integrative and combine the results of multiple other tools. It is therefore assumed an integrative tool would have large similarities with its integrated tools. However, a difference in tool version and filtering can still give a different set of circRNAs. For example, CirComPara2 is an integrative tool that combines CIRCexplorer2, CIRI2, DCC, and find_circ, nevertheless, the combination of CirComPara2 and CIRCexplorer3 still gives a significant increase (10%) in circRNA predictions.

Discussion

Multimodal orthogonal validation of bioinformatics tools that predict circular RNAs from total RNA sequencing data is currently lacking. Hence, their precision and sensitivities are unknown and scientific data is confounded with false positive and false negative predictions. To accommodate this lacune, we set up a large-scale international collaborative circRNA detection tool benchmarking study. First, a deeply sequenced total RNA sequencing dataset was processed by the developers of 16 different circRNA detection tools. Next, 3 empirical validation strategies were used to evaluate a random selection of 1,560 circRNAs representing each tool: 1) RT-qPCR to determine if the circRNA was detectable; 2) RNase R treatment to confirm that the detected RNA was most likely circular and not linear; and 3) amplicon sequencing to confirm the BSJ sequence. Of note, both circRNA RT-qPCR and RNAse R validation protocols were extensively validated (4, 28).

The precision values are similarly high among tools (Figure 3A), especially when considering the subset of *high-abundance* circRNAs (with a back-spliced junction (BSJ) count ≥ 5). In contrast, the number of predicted circRNAs varies greatly among tools, in line with previous studies (11). The striking differences in sensitivity are in part dependent on the operator applying BSJ count filters. Since the greater majority of predicted circRNAs are low abundant, setting a threshold for abundance may have a large negative effect on the detection sensitivity.

The three validation methods each have their own strengths and biases, with conflicting results for several circRNAs (Figure 3B, Supplementary Discussion 1). This underscores the importance of using different validation methods to compensate for their intrinsic limitations and to increase the validation status confidence (as previously suggested in (7)). Although long-read sequencing has been implemented to study full-length circRNAs (29–32), the bulk of currently available data remains short-read sequencing. Therefore, this benchmarking study evaluated circRNA detection tools for short-read sequencing data, which typically report circRNAs by their BSJ position (chr, start, end, strand). However, it remains unknown if the detected BSJ corresponds to one circRNA, or multiple alternatively spliced circRNAs with different exon/intron compositions. Henceforth, the prediction precision values reported here might be influenced by more than one circRNA with the same BSJ. As this study is focused on circRNA detection in short-read sequencing data, the internal circRNA composition was not evaluated. Furthermore, no distinction can be made between circRNAs on the positive strand or negative strand using RT-qPCR and amplicon sequencing (9.4% of circRNAs were reported to originate from different strands according to different tools).

Based on a pilot study (Supplementary Data 5, Supplementary Figure 11), a cut-off was set at BSJ count 5, as circRNAs under this cut-off approached the qPCR limit to reliably detect RNase R based degradation of falsely predicted circRNAs. While very deep sequencing of a large RNA input amount was performed, it is beyond the scope of this study to evaluate if the BSJ count should be reconsidered in function of sequencing depth. However, as the majority of predicted circRNAs have a BSJ count below 5, we decided to include at least a subset of these *low-abundance* circRNAs to calculate the corresponding prediction precision. It is no surprise that the precision values for *low-abundance* circRNAs are significantly lower compared to *high-abundance* circRNAs (Chi-squared test with p-value < 0.001, OR = 2.8). This difference is likely due to the detection limits of the applied validation strategies in conjunction with the sampling bias of *low-abundance* analytes, and not due to inherently more false positive predictions for circRNAs with a lower count. Of note, it can be presumed that weakly expressed circRNAs are less relevant for both functional studies and biomarker research.

Focusing on *high-abundance* circRNAs, interesting links were found between circRNA annotation and validation rates. As such, circRNAs had higher validation rates when they were detected by multiple tools, when they were previously reported in a circRNA database, when they were surrounded by canonical splice sites, and when they originate from a region with an annotated linear transcript. CircRNA detection tools with a ‘candidate-based’ approach are more precise than tools using the ‘segmented read-based' approach, which is in line with the higher validation likelihood of circRNAs originating from known linear genes and surrounded by canonical splice sites.

Based on our study, we compiled a list of recommendations for circRNA detection and validation, and for the future development of circRNA detection tools and their performance evaluation (Table 2). Ideally, publicly available (spike-in) reference material (consisting of known synthetic circRNAs) should be used to benchmark existing and novel circRNA detection tools. However, such reference material is currently not available. As the main goal of this study was to perform a neutral assessment of circRNA detection tool sensitivity and precision, the developers of the tools were asked to run the tools themselves. Therefore, execution time, memory usage, and ease of use could not be compared and were not assessed here.

Furthermore, this study resulted in a circRNA resource containing > 315,000 circRNAs detected in three human cancer cell lines from different tissue origins and provides validation results for 1500 circRNAs that can be used as a reference for the development of new or improved circRNA detection tools. Finally, our study can also serve as an example framework for empirical validation of benchmarking results from other bioinformatics tools in the future.

circRNA detection	1. An orthogonal validation method must be used to validate a predicted circRNA; qPCR validation on its own is not sufficient, at least qPCR + RNase R treatment or preferably qPCR + amplicon sequencing should be used. 2. Filtering based on a minimum BSJ count is recommended to increase the likelihood of successful empirical validation.
circRNA validation	3. For a precision-focused approach, the *intersection* of two tools with a high individual precision (for example ≥ 90%) should be used. 4. For a sensitivity-focused approach, the *union* of two tools with a high individual precision (for example ≥ 90%) should be used. 5. The choice of tools to be combined may be informed based on the tools' underlying principles (circRNA detection approach, reliance on linear annotation, and filtering).
circRNA tool development	6. Tools should report the originating strand information, the BSJ count evidence, and the chromosomal start and end position of the BSJ.
circRNA tool validation	7. For evaluation of sensitivity, novel and updated tools are encouraged to use the empirically validated set of 957 true-positive circRNAs. 8. For evaluation of precision, a random set of 100 predicted circRNAs should be validated with empirical methods.

Table 2 CircRNA research recommendations.

Online methods

Study set-up

As this study includes executing and evaluating circRNA detection tools, the co-authors can be divided into two groups 1) a first independent group (with no circRNA detection tool of their own) which initiated and designed the study and performed all the wet-lab work and data analysis (the validation co-author group), and 2) a second group of tool developer co-authors, which detected circRNAs using their own circRNA detection tools according to their expertise (the circRNA prediction co-author group) (details in Author Contribution section). During the study, meetings and emails were used to share the results (first anonymously) and discuss the final manuscript with the circRNA prediction co-authors.

Cell culture

Three cancer cell lines were randomly chosen as biological replicates. SW480 cells were cultured at 37 °C, 0% CO_2 in Leibovitz's L-15 medium (#31415-029, ThermoFisher). HLF cells and NCI-H23 cells were cultured at 37 °C, 5% CO_2 in DMEM, low glucose, GlutaMAX Supplement, pyruvate (#21,885,025, ThermoFisher) and RPMI 1640 Medium, HEPES (#52,400,041, ThermoFisher), respectively. 10% fetal bovine serum (FBS) (#F7524, Sigma) and 1% penicillin-streptomycin (10,000 U/mL) (#15,140,122, ThermoFisher) were added to all three media.

RNA isolation

RNA was isolated from the cells using the miRNeasy Mini kit (#217004, Qiagen) according to the manufacturer's instructions, including the optional on-column DNase treatment (#79254, Qiagen). For each cell line, a sufficient number of cells was cultured to be able to harvest a minimum of 330 µg RNA per cell line. The RNA concentration was measured spectrophotometrically using a NanoDrop instrument and the RNA integrity was evaluated using a Fragment Analyzer. For each cell line, the RNA was pooled and aliquoted (1000 ng RNA in 100 µL nuclease-free water per aliquot) and stored at –80 °C, making a uniform RNA collection to use for all downstream experiments.

Library preparation and sequencing

For each cell line, 1000 ng input RNA was used. First, ribosomal RNA (rRNA) was removed with the NEBNext rRNA Depletion Kit (#E6350X, New England Biolabs), following the manufacturer's instructions. The NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (#E7760L, New England Biolabs) was used in combination with the NEBNext Multiplex Oligos for Illumina (#E7600S, New England Biolabs) to index and prepare the samples for sequencing. The library preparation protocol was adjusted to obtain relatively long insert sizes (average size of 636 nucleotides measured using a Fragment Analyzer): RNA fragmentation of 7.5 minutes; first-strand cDNA synthesis elongation step of 50 minutes instead of 15 minutes. The last bead clean-up step was performed twice to completely remove all indexes from the samples. Finally, the samples were pooled equimolarly and sequenced on a NovaSeq 6000 instrument using a NovaSeq 6000 S1 Reagent Kit v1.5 (300 cycles) (#20028317, Illumina), resulting in approximately 300 million paired-end 150-nucleotides reads per sample. Raw FASTQ files are stored in the Sequence Read Archive (SRX13414572 (HLF), SRX13414573 (NCI-H23), SRX13414574 (SW480)).

CircRNA detection

In November 2020, a comprehensive list of all published circRNA detection tools was compiled, and all developers were invited to collaborate. Upon consent, they were asked to detect circRNAs using their own circRNA detection tool as they seemed fit for the data that was provided. The circRNA detection steps for each tool are detailed in the Supplementary Notes. We were unable to get into contact with the authors of find_circ (23) and decided to run this tool ourselves, as it is one of the most frequently cited and broadly used circRNA detection tools. After collecting all circRNA detection results, a uniform list of circRNAs defined by their BSJ position (chr, start, end, strand) and the BSJ count for each tool was compiled (Hg38, 0-based).

CircRNA selection and primer design

Guided by a pilot experiment assessing circRNA RT-qPCR detectability in function of abundance and RNA input amount (Supplementary Data 5, Supplementary Figure 11), for each tool, 80 *high-abundance* circRNAs (with a BSJ count of at least 5), and 20 *low-abundance* circRNAs (with a BSJ count below 5) were selected (as two separate count bins). Primer pairs were designed using our primer design tool CIRCprimerXL (28). All primer sequences are available in Supplementary Table 3. If no primer pair could be designed for a given circRNA, a substitution was randomly selected from the complete dataset, considering the BSJ count bin. In total, 1,560 circRNA/tool/cell line tuples were selected. As some circRNAs were selected more than once (for different tools) the total number of unique circRNA/cell line pairs is 1,516, and the number of unique circRNAs (not taking into account the strand) is 1,457 (Supplementary Figure 14). Additionally, most of the selected circRNAs are detected by multiple tools (for which they were not selected). For this study, only the 20 + 80 selected circRNAs for a specific tool were used to evaluate that tool to keep the number of observations equal for each tool, even though more of its predicted circRNAs might have been validated.

RNase R and RT-qPCR

The RNA aliquots derived from the three cell lines were used for the circRNA RT-qPCR validation. A total of 1,080, 900, and 780 µl RNA (100 ng/µL) was required to validate 579, 500, and 437 circRNAs in HLF, NCI-H23, and SW480 cells, respectively. RNase R treatment was performed according to our previously reported protocol (4), adapted for this large-scale experiment. In summary, one RNA aliquot of a given cell line was treated with RNase R (#RNR07250 (250 U), Lucigen) and another was treated as a buffer control, for a total of 92 RNase R treated replicates and 92 buffer control replicates (2 * 36 for HLF, 2 * 30 for NCI-H23, and 2 * 26 for SW480 RNA). All volumes were doubled during the buffer and RNase R reaction (total reaction volume of 20 µL). This was followed by a clean-up step using Vivacon 500, 10,000 MWCO Hydrosart columns (#VN01H02, Sartorius). Next, the RT reaction was performed on the 184 separate replicates using the iScript Advanced cDNA Synthesis Kit (#172-5038, Bio-Rad), according to the manufacturer's instructions. After RT, the cDNA was diluted 1:2 and an aliquot (2.5 µL) was further diluted 1:4 to evaluate the success of the RNase R reaction for each individual replicate. For this, ACTB and a known circRNA (chr1:117402185-117420649) previously described (4) (primer sequences available in Supplementary Table 7) were measured with qPCR using 2.5 µL 2x SsoAdvanced Universal SYBR Green Supermix (#172-5274, Bio-Rad), 0.5 µL forward and reverse primer (5 nM), and 2 µl cDNA per well, with qPCR duplicates. Once the RNase R treatment was successfully validated, all cDNA replicates were pooled per cell line and treatment condition. The cDNA was diluted 1:5 in 2× SsoAdvanced Universal SYBR Green Supermix (#172-5274, Bio-Rad). All 1,560 circRNA primer pairs were ordered from IDT in 96-well plates at a concentration of 100 µM in RNase-free water. All primers were diluted 1:160 to obtain a 0.625 µM concentration. In each well of a qPCR plate, 2 µl diluted primers and 3 µl cDNA-master mix combination were added, resulting in an equivalent of 25 ng input RNA per qPCR reaction. Each assay (circRNA) was measured 4 times to include qPCR duplicates and to measure the abundance in both an RNase R untreated and treated sample, resulting in a total of more than 6000 qPCR reactions. A pipetting robot (EVO100, TECAN L) was used to dilute the primers and fill the qPCR plates. The qPCR reactions were run on a CFX384 instrument (Bio-Rad). The plates were stored at -20 °C for amplicon sequencing.

Amplicon sequencing

After RT-qPCR, ~80% of the circRNAs were randomly selected for amplicon sequencing. To make the sequencing library, the amplicons were pooled by combining 2 µL of the PCR reaction from one of the untreated qPCR duplicates, per cell line. Next, the 3 samples were cleaned using Vivacon 500, 10,000 MWCO Hydrosart columns (#VN01H02, Sartorius). The PCR product pools were analyzed using a TapeStation 4150 (Agilent) and the concentration was measured using a Qubit fluorometer (ThermoFisher). Next, the three pools were diluted in RNase-free water to obtain 50 µl samples with a concentration of 20 ng/µl. Finally, the samples were prepared for sequencing using the NEBNext Ultra II DNA Library Prep Kit for Illumina (#E7645S, New England Biolabs) and NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1) (#E7600S, New England Biolabs). To retain all amplicons, no size selection was performed after adaptor ligation, and 1.0x AMPure XP beads (#A63881, Beckman Coulter) in a 1:1 sample:beads ratio was used instead. After library preparation, the samples were pooled equimolarly. The pool was sequenced on a NextSeq 500 instrument using a Mid Output Kit v2.5 (150 cycles) (#20024904, Illumina), resulting in approximately 25-30 million paired-end 75-nucleotides reads per library.

Data analysis

Data analysis was mostly done using R (33) (version 4.2.1) in RStudio (34) (version 2022.07.1, build 554). The following R packages were used: tidyverse (version 1.3.2), conflicted (version 1.1.0), ggrepel (version 0.9.1), ggseqlogo (version 0.1), europepmc (version 0.4.1), gplots (version 3.1.3), ggpubr (version 0.4.0), quantreg (version 5.94) and UpSetR (version 1.4.0). For sequencing data analyses, including circRNA detection and amplicon sequencing analysis, the Ghent University high-performance cluster was used. For this, Python3 (version 3.6.8) (35), Bowtie2 (version 2.3.4.1) (36), fastahack (version 1.0.0), SAMtools (version 1.11) (37), and BEDTools (version 2.30.0) (38) were used. The human reference transcriptome was downloaded as a GTF file from Ensembl (39). All data analysis scripts are available at https://github.com/OncoRNALab/circRNA_benchmarking.

Amplicon sequencing data analysis

First, a custom Python script matches the primer sequences with the first 16-mer of each read (forward and reversed) and generates a separate FASTQ file per primer pair, containing all reads starting with that primer sequence. The FASTQ reads are then clipped to remove the primer sequences. Next, all FASTQ files are mapped against the reference genome (Ensembl version GRCh38.101) supplemented with the theoretical BSJ amplicon sequences using Bowtie2 with default settings. Lastly, the Bowtie2 BAM files are converted to counts using another custom Python script and the percentage on-target amplification was calculated for each primer pair.

Filtering and determination of orthogonal precision values

Several strategies to filter the data prior to precision value calculation were explored. For RT-qPCR, a circRNA was considered validated when at least one of the untreated RNA samples had a Cq value above 10. Multiple variations of this threshold and a potential upper Cq threshold were evaluated. For RNase R validation, a subset of circRNAs with at least one untreated replicate with a Cq value below 32 was selected to ensure that the enzymatic degradation of a false-positive circRNA could be measured. A circRNA was considered validated upon RNase R treatment if the difference in Cq between the untreated and treated RNA sample was equal to or less than 3 cycles, based on a previous study (4). As there were two qPCR replicates available for each (un)treated sample, the ‘best-case scenario’ was used to calculate the difference in Cq by subtracting the maximum untreated Cq replicate from the minimum treated Cq replicate. A circRNA with both untreated replicates having a Cq value above 32 was labeled as NA. For amplicon sequencing, a circRNA was considered validated if the primer pair was found in at least 1000 reads and if at least 50% of these reads matched the expected amplicon upon mapping with Bowtie2. For a random subset of circRNAs, no amplicon sequencing was performed; these were labeled as NA.

To calculate precision values per tool, BSJ count bin, and validation method, the number of circRNAs that passed the validation was divided by the total number of circRNAs that were not NA for that validation method. We also determined a compound precision value by considering both qPCR, RNase R treatment, and amplicon sequencing. For this, the RNase

R and amplicon sequencing precision values were modified by excluding the circRNAs for which qPCR failed (treated as NAs), so that all three precision values could be multiplied to generate a compound precision value per tool, without amplifying the effect of lowly abundant circRNAs that could not easily be measured with qPCR. Lastly, the number of theoretically true positive circRNAs was calculated by multiplying the total number of circRNAs predicted by that tool for that sample with the compound precision value. Sensitivity was also estimated as the percentage of circRNAs each tool detected from the validated set of true-positive circRNAs. For this, no distinction was made between the BSJ count groups.

Annotation of circRNAs

To obtain the circRNA splice site information, the BSJ-flanking nucleotides were extracted from the reference genome using fastahack (Ensembl version GRCh38.104). To compare BSJ positions with known linear annotation, BEDtools intersect was used with a list of canonical transcripts from Ensembl (‘canonical tsv’) with their positions based on the corresponding Ensembl GTF file (Ensembl version GRCh38.103). When a circRNA mapped to multiple isoforms, the annotation was labeled as ‘ambiguous’ and the circRNA was not taken into account for further annotation-based calculations and figures. The annotation was used to compute the length of each circRNA excluding introns, and the number of exons per circRNA. CircRNAs smaller than their host gene exon were labeled ‘single-exon’ circRNAs. For the length of each circRNA including introns, the BSJ start position was simply subtracted from the BSJ end position. Furthermore, for each circRNA, annotation was added to indicate if the BSJ start and end positions match known exon boundaries. When comparing predicted circRNAs to circRNAs previously described in databases, strand information was discarded.

Combination of tools

To compare the circRNA tools, the union and intersection of all circRNAs predicted by each tool pair and triple were calculated. A weighted precision value was calculated for each combination of tools as follows: ((perc_compound_val_1 * total_n_1) + (perc_ compound _val_2 * total_n_2)) / (total_n_1 + total_n_2). For this, strand information was discarded, as 4 out of 16 tools did not report circRNA strands and would therefore have been excluded. These calculations were performed for each cell line separately. To determine the correspondence among tools, the Jaccard distance was calculated and heatmap clusters were generated. The tools were compared based on the mere presence or absence of a circRNA. Also, for the calculation of how many tools detected a given circRNA, circRNA strand information was discarded.

Availability

We anticipate this study will serve as a future resource for the circRNA community. The information on all predicted circRNAs (n = 315,312), including the large extensively validated circRNA set (n = 1,516), along with the validation results, and all the scripts used to compute the metrics and make the figures are available at https://github.com/OncoRNALab/circRNA_benchmarking.

Acknowledgments

We thank Steve Lefever for his contribution to primer design in the early stage of this project. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO, and the Flemish Government – department EWI. Stefania Bortoluzzi would like to acknowledge support of Fondazione AIRC per la Ricerca sul Cancro, Milan, Italy (Investigator Grant 2017 20052), Italian Ministry of Education, Universities, and Research (PRIN 2017 2017PPS2X4_003), EU within the MUR PNRR “National Center for Gene

Therapy and Drugs based on RNA Technology" (Project no. CN00000041 CN3 RNA) and "HPC, big data and quantum computing" (CN1 HPC) and the Department of Molecular Medicine of the University of Padova. Trees-Juen Chuang would like to acknowledge the National Health Research Institutes, Taiwan (NHRI-EX110-11011B1). Christoph Dieterich would like to acknowledge the German Science Foundation (DI 1501/13-1). Paul Flicek would like to acknowledge Wellcome Trust (WT108749/Z/15/Z). Enrico Gaffo would like to acknowledge the Fondazione Umberto Veronesi, Milan, Italy (Fellowship 2020). Steve Hoffmann would like to acknowledge the German Federal Ministry of Education and Research (BMBF 031L0106D). Eric C. Lai would like to acknowledge the NIH (R01-NS083833) and MSK (Core Grant P30-CA008748). Peter Stadler would like to acknowledge the German Federal Ministry of Education and Research (BMBF 031L0164C). Jakub Westholm would like to acknowledge the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab. Li Yang would like to acknowledge the National Natural Science Foundation of China (NSFC, 31925011) and the Ministry of Science and Technology of China (MoST, 2021YFA1300503). Chu-Yu Ye would like to acknowledge the National Science Foundation of China (31871589).

Author contributions

Validation co-author group: Marieke Vromman: methodology, software, validation, formal analysis, investigation, data curation, writing - original draft, visualization, funding acquisition; Jasper Anckaert: software; Justine Nuytens: investigation; Olivier Thas: methodology; Eveline Vanden Eynde: investigation; Kimberly Verniers: investigation; Nurten Yigit: investigation; Jo Vandesompele: conceptualization, methodology, writing - review & editing, supervision, project administration, funding acquisition; Pieter-Jan Volders: conceptualization, methodology, writing - review & editing, supervision, project administration, funding acquisition.

CircRNA prediction co-author group (contribution: formal analysis and writing - review & editing): Stefania Bortoluzzi, Alessia Buratin, Chia-Ying Chen, Qinjie Chu, Trees-Juen Chuang, Roozbeh Dehghannasiri, Christoph Dieterich, Xin Dong, Paul Flicek, Enrico Gaffo, Wanjun Gu, Chunjiang He, Steve Hoffmann, Osagie Izuogu, Michael S. Jackson, Tobias Jakobi, Eric C. Lai, Julia Salzman, Mauro Santibanez-Koref, Peter Stadler, Guoxia Wen, Jakub Westholm, Li Yang, Chu-Yu Ye, Guo-Hua Yuan, Jinyang Zhang, Fangqing Zhao

Funding

This work was supported by a Foundation Against Cancer grant (STK F/2018/1,267), a Standup Against Cancer grant (STIVLK2018000601), and the Concerted Research Action of Ghent University (BOF16/GOA/023, BOF/24J/2021/244) and the Research Foundation - Flanders (FWO 1253321N).

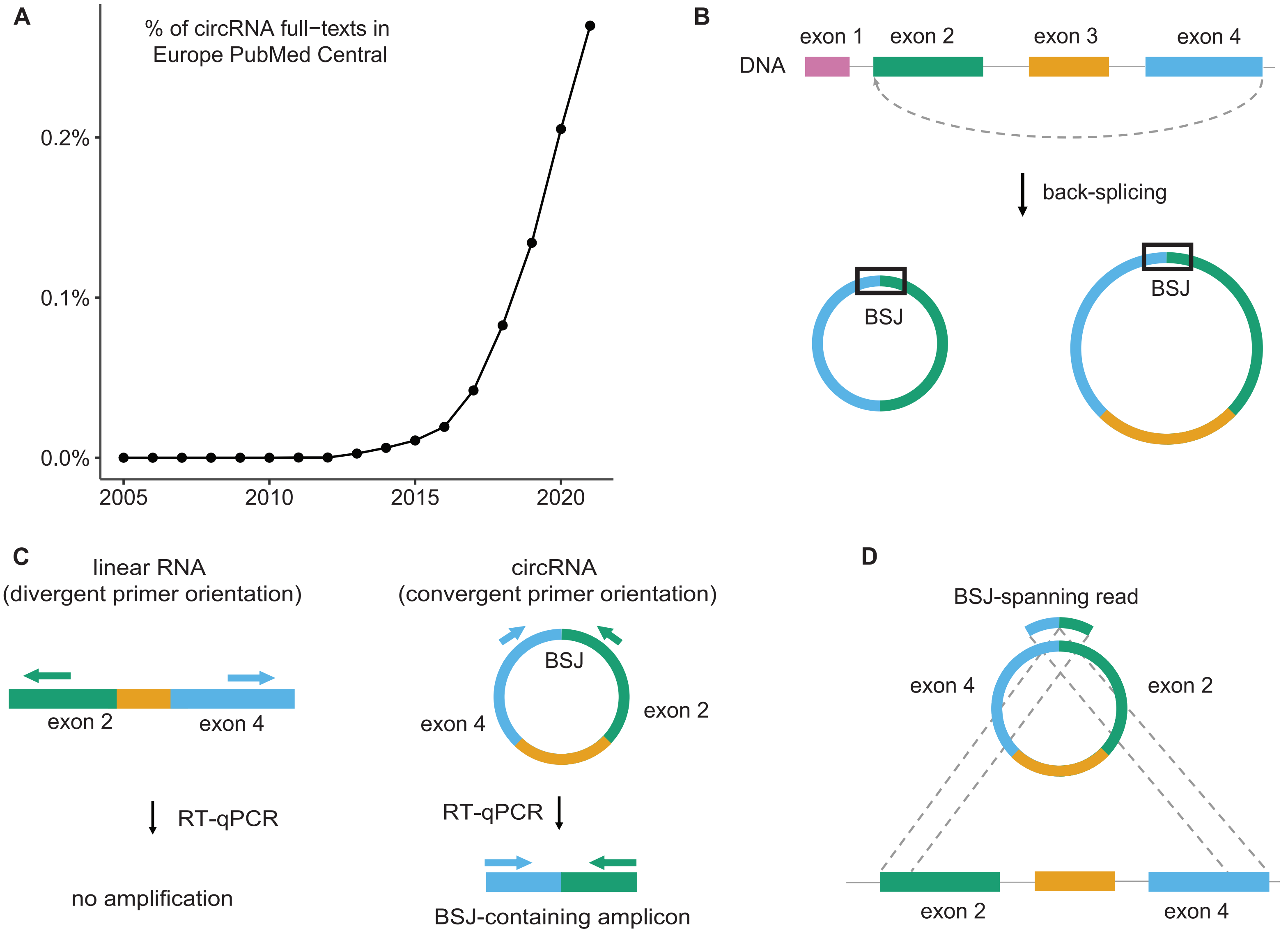
Conflict of interest

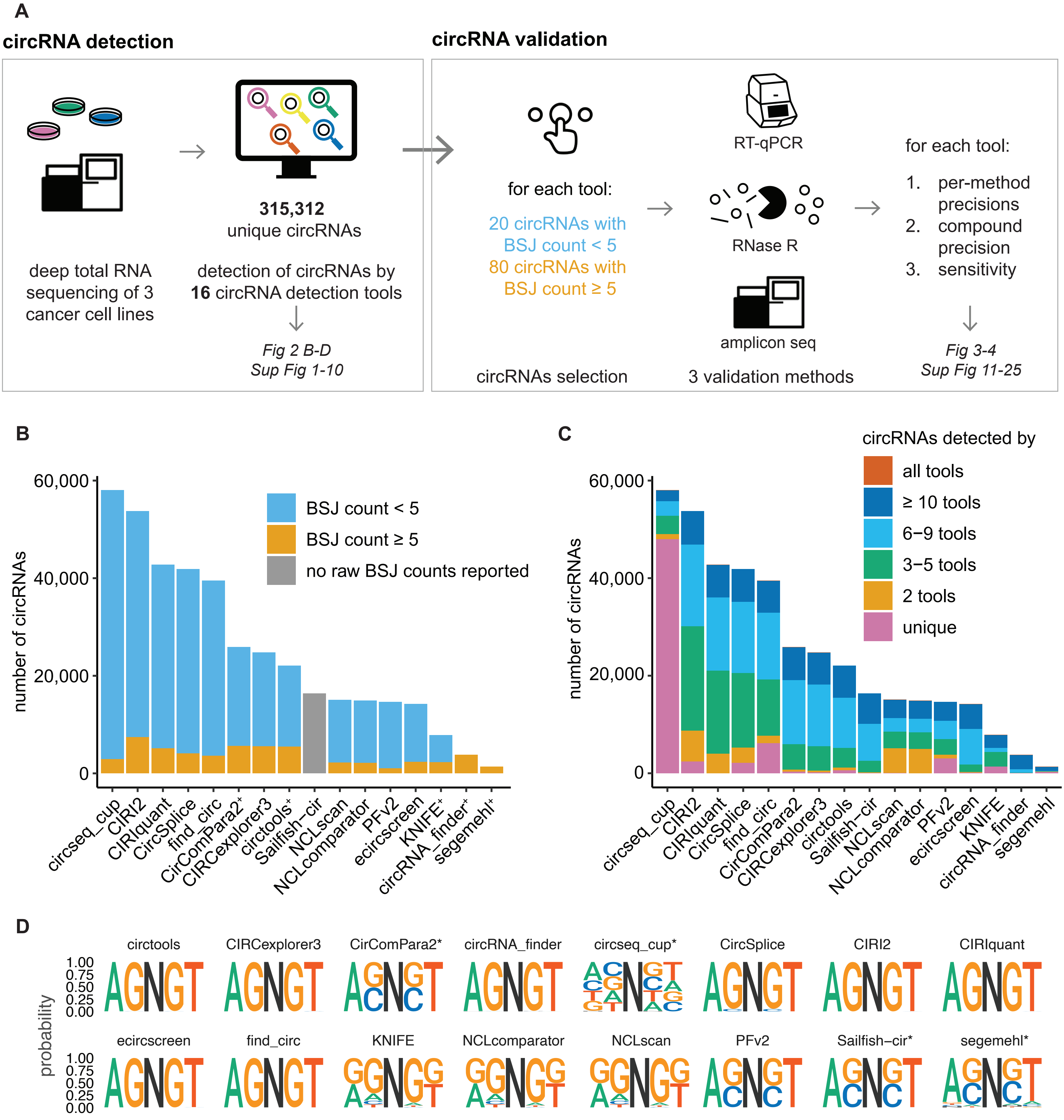
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Kristensen,L.S., Andersen,M.S., Stagsted,L.V.W., Ebbesen,K.K., Hansen,T.B. and Kjems,J. (2019) The biogenesis, biology and characterization of circular RNAs. *Nat Rev Genet*, 10.1038/s41576-019-0158-7.
2. Hulstaert,E., Morlion,A., Avila Cobos,F., Verniers,K., Nuytens,J., vanden Eynde,E., Yigit,N., Anckaert,J., Geerts,A., Hindryckx,P., *et al.* (2020) Charting Extracellular Transcriptomes in The Human Biofluid RNA Atlas. *Cell Rep*, **33**.
3. Wang,S., Zhang,K., Tan,S., Xin,J., Yuan,Q., Xu,H., Xu,X., Liang,Q., Christiani,D.C., Wang,M., *et al.* (2021) Circular RNAs in body fluids as cancer biomarkers: the new frontier of liquid biopsies. *Mol Cancer*, **20**, 1–10.
4. Vromman,M., Yigit,N., Verniers,K., Lefever,S., Vandesompele,J. and Volders,P. (2021) Validation of Circular RNAs Using RT-qPCR After Effective Removal of Linear RNAs by Ribonuclease R. *Curr Protoc*, **1**, 1–16.
5. Yu,C.Y., Liu,H.J., Hung,L.Y., Kuo,H.C. and Chuang,T.J. (2014) Is an observed non-co-linear RNA product spliced in trans, in cis or just in vitro? *Nucleic Acids Res*, **42**, 9410.
6. Szabo,L. and Salzman,J. (2016) Detecting circular RNAs: Bioinformatic and experimental challenges. *Nat Rev Genet*, **17**, 679–692.
7. Dodbele,S., Mutlu,N. and Wilusz,J.E. (2021) Best practices to ensure robust investigation of circular RNAs: pitfalls and tips. *EMBO Rep*, **22**, 1–12.
8. Nielsen,A.F., Bindereif,A., Bozzoni,I., Hanan,M., Hansen,T.B., Irimia,M., Kadener,S., Kristensen,L.S., Legnini,I., Morlando,M., *et al.* (2022) Best practice standards for circular RNA research. *Nat Methods*, 10.1038/s41592-022-01487-2.
9. Jakobi,T. and Dieterich,C. (2019) Computational approaches for circular RNA analysis. *Wiley Interdiscip Rev RNA*, **2019**, e1528.
10. Hansen,T.B., Venø,M.T., Damgaard,C.K. and Kjems,J. (2015) Comparison of circular RNA prediction tools. *Nucleic Acids Res*, **44**, e58.
11. Gaffo,E., Buratin,A., Dal Molin,A. and Bortoluzzi,S. (2022) Sensitive, reliable and robust circRNA detection from RNA-seq with CirComPara2. *Brief Bioinform*, **23**, 1–12.
12. Guo,L., Zhu,Q.-H., Zhang,X., Ye,C.-Y., Liu,C., Yu,Y., Chu,Q., Jiang,W. and Fan,L. (2016) Full-length sequence assembly reveals circular RNAs with diverse non-GT/AG splicing signals in rice. *RNA Biol*, **14**, 1055–1063.
13. Gao,Y., Zhang,J. and Zhao,F. (2018) Circular RNA identification based on multiple seed matching. *Brief Bioinform*, **19**, 803–810.
14. Zhang,J., Chen,S., Yang,J. and Zhao,F. (2020) Accurate quantification of circular RNAs identifies extensive circular isoform switching events. *Nat Commun*, **11**, 90.
15. Feng,J., Chen,K., Dong,X., Xu,X., Jin,Y., Zhang,X., Chen,W., Han,Y., Shao,L., Gao,Y., *et al.* (2019) Genome-wide identification of cancer-specific alternative splicing in circRNA. *Mol Cancer*, **18**, 1–5.
16. Jakobi,T., Uvarovskii,A. and Dieterich,C. (2019) Circtools—a one-stop software solution for circular RNA research. *Bioinformatics*, **35**, 2326–2328.
17. Li,M., Xie,X., Zhou,J., Sheng,M., Yin,X., Ko,E.A., Zhou,T. and Gu,W. (2017) Quantifying circular RNA expression from RNA-seq data using model-based framework. In *Bioinformatics*.Vol. 33, pp. 2131–2139.
18. Chuang,T.J., Wu,C.S., Chen,C.Y., Hung,L.Y., Chiang,T.W. and Yang,M.Y. (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**.
19. Szabo,L., Morey,R., Palpant,N.J., Wang,P.L., Afari,N., Jiang,C., Parast,M.M., Murry,C.E., Laurent,L.C. and Salzman,J. (2015) Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol*, **16**.
20. Westholm,J.O., Miura,P., Olson,S., Shenker,S., Joseph,B., Sanfilippo,P., Celniker,S.E., Graveley,B.R. and Lai,E.C. (2014) Genome-wide Analysis of Drosophila Circular RNAs

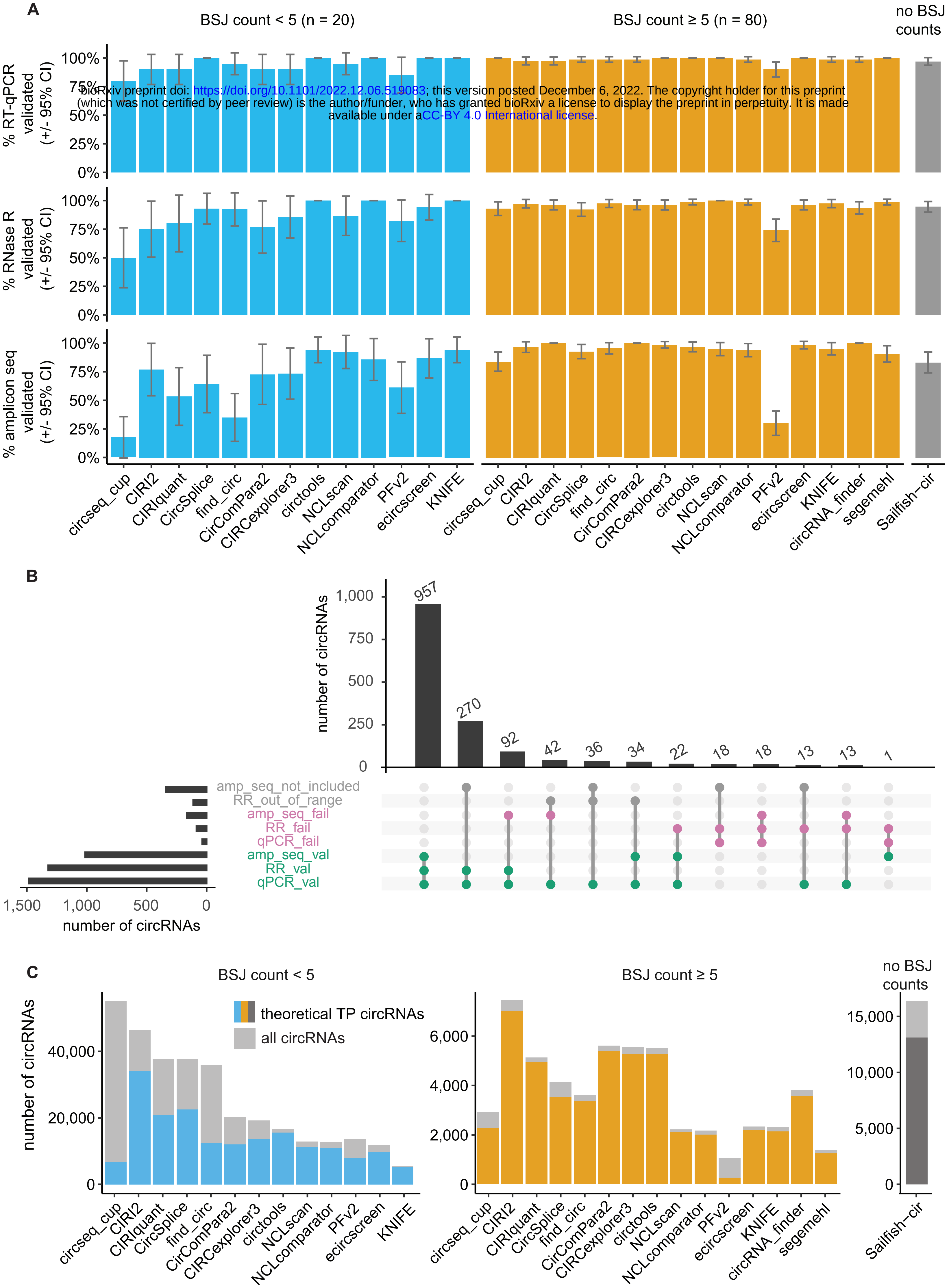
Reveals Their Structural and Sequence Properties and Age-Dependent Neural Accumulation. *Cell Rep*, **9**, 1966–1980.
21. Hoffmann,S., Otto,C., Doose,G., Tanzer,A., Langenberger,D., Christ,S., Kunz,M., Holdt,L.M., Teupser,D., Hackermüller,J., *et al.* (2014) A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol*, **15**.
22. Ma,X.K., Wang,M.R., Liu,C.X., Dong,R., Carmichael,G.G., Chen,L.L. and Yang,L. (2020) CIRCexplorer3: A CLEAR Pipeline for Direct Comparison of Circular and Linear RNA Expression. *Genomics Proteomics Bioinformatics*, **17**, 511–521.
23. Memczak,S., Jens,M., Elefsinioti,A., Torti,F., Krueger,J., Rybak,A., Maier,L., Mackowiak,S.D., Gregersen,L.H., Munschauer,M., *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, **495**, 333–338.
24. Izuogu,O.G., Alhasan,A.A., Mellough,C., Collin,J., Gallon,R., Hyslop,J., Mastrorosa,F.K., Ehrmann,I., Lako,M., Elliott,D.J., *et al.* (2018) Analysis of human ES cell differentiation establishes that the dominant isoforms of the lncRNAs RMST and FIRRE are circular. *BMC Genomics*, **19**, 1–18.
25. Chen,C.Y. and Chuang,T.J. (2019) NCLcomparator: Systematically post-screening non-co-linear transcripts (circular, trans-spliced, or fusion RNAs) identified from various detectors. *BMC Bioinformatics*, **20**, 1–11.
26. Zeng,X., Lin,W., Guo,M. and Zou,Q. (2017) A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput Biol*, **13**, e1005420.
27. Vromman,M., Vandesompele,J. and Volders,P.-J. (2021) Closing the circle: current state and perspectives of circular RNA databases. *Brief Bioinform*, **22**, 288–297.
28. Vromman,M., Anckaert,J., Vandesompele,J. and Volders,P.-J. (2022) CIRCprimerXL: Convenient and High-Throughput PCR Primer Design for Circular RNA Quantification. *Frontiers in Bioinformatics*, **0**, 20.
29. Zhang,J., Hou,L., Zuo,Z., Ji,P., Zhang,X., Xue,Y. and Zhao,F. (2021) Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long. *Nat Biotechnol*, **39**, 836–845.
30. Rahimi,K., Venø,M.T., Dupont,D.M. and Kjems,J. (2021) Nanopore sequencing of brain-derived full-length circRNAs reveals circRNA-specific exon usage, intron retention and microexons. *Nature Communications 2021 12:1*, **12**, 1–15.
31. Xin,R., Gao,Y., Gao,Y., Wang,R., Kadash-Edmondson,K.E., Liu,B., Wang,Y., Lin,L. and Xing,Y. (2021) isoCirc catalogs full-length circular RNA isoforms in human transcriptomes. *Nature Communications 2021 12:1*, **12**, 1–11.
32. Liu,Z., Tao,C., Li,S., Du,M., Bai,Y., Hu,X., Chen,J. and Yang,E. circFL-seq reveals full-length circular RNAs with rolling circular reverse transcription and nanopore sequencing.
33. R Core Team (2019) R: A language and environment for statistical computing.
34. RStudio Team (2020) RStudio: Integrated Development for R. RStudio.
35. van Rossum,G. and Drake,F.L. (2009) Python 3 Reference Manual.
36. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods 2012 9:4*, **9**, 357–359.
37. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A., Davies,R.M., *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**.
38. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
39. Cunningham,F., Allen,J.E., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Bennett,R., *et al.* (2022) Ensembl 2022. *Nucleic Acids Res*, **50**, D988–D995.

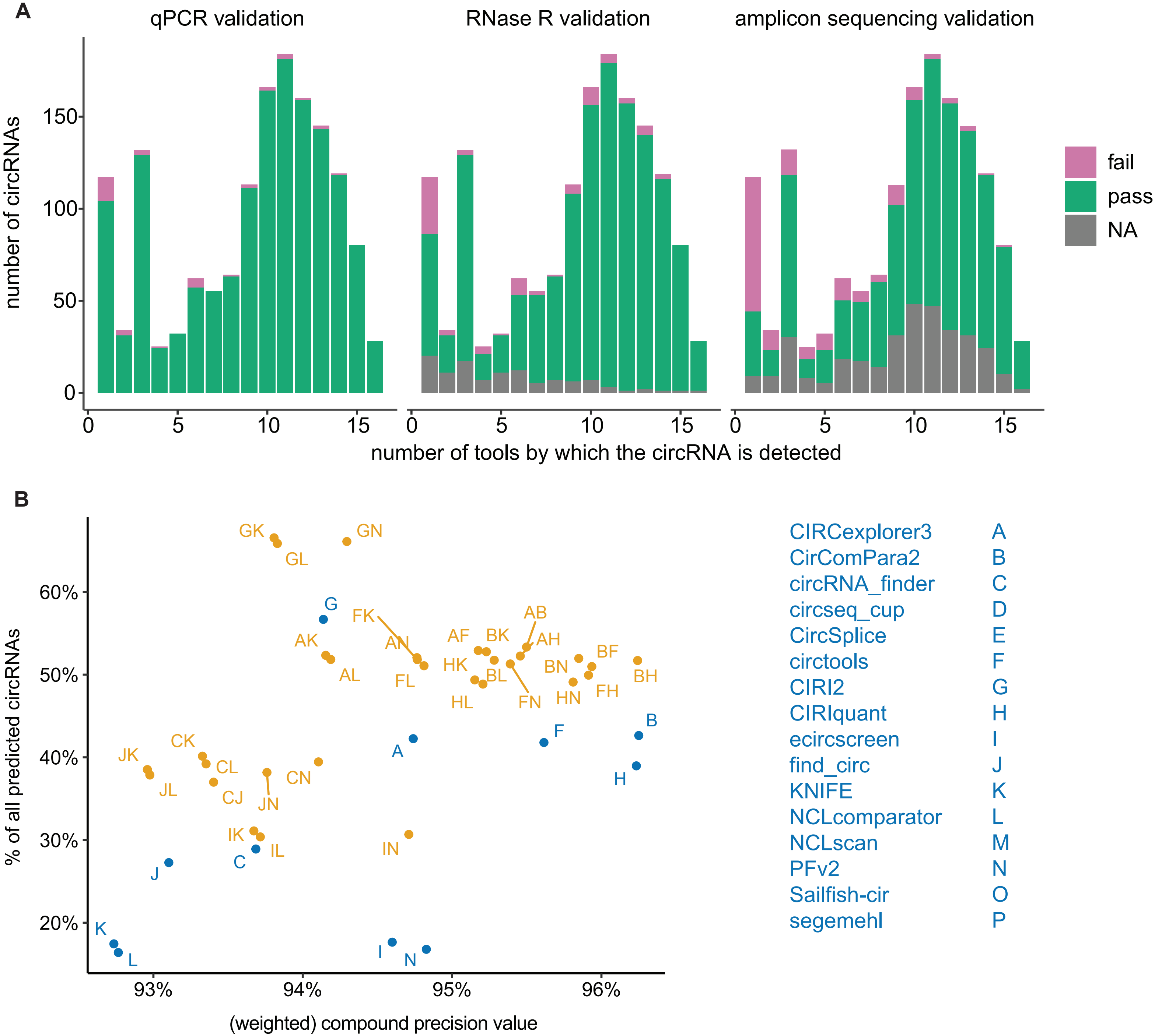

A
% of circRNA full−texts in Europe PubMed Central
0.2%
0.1%
0.0%
2005
2010
2015
2020
B
exon 1
exon 2
exon 3
exon 4
DNA
back-splicing
BSJ
BSJ
C
linear RNA
(divergent primer orientation)
exon 2
exon 4
RT-qPCR
no amplification
circRNA
(convergent primer orientation)
BSJ
exon 4
exon 2
RT-qPCR
BSJ-containing amplicon
D
BSJ-spanning read
exon 4
exon 2
exon 2
exon 4



⁺tool filtered the output based on BSJ counts

*tool does not report strand information


A
BSJ count < 5 (n = 20)
BSJ count ≥ 5 (n = 80)
no BSJ counts
% RT-qPCR validated (+/- 95% CI)
% RNase R validated (+/- 95% CI)
% amplicon seq validated (+/- 95% CI)
0%
25%
50%
75%
100%
circseq_cup
CIRI2
CIRIquant
CircSplice
find_circ
CirComPara2
CIRCexplorer3
circtools
NCLscan
NCLcomparator
PFv2
ecircscreen
KNIFE
circRNA_finder
segemehl
Sailfish−cir
B
number of circRNAs
1,000
750
500
250
0
957
270
92
42
36
34
22
18
18
13
13
1
amp_seq_not_included
RR_out_of_range
amp_seq_fail
RR_fail
qPCR_fail
amp_seq_val
RR_val
qPCR_val
1,500
1,000
500
0
number of circRNAs
C
BSJ count < 5
BSJ count ≥ 5
no BSJ counts
theoretical TP circRNAs
all circRNAs
number of circRNAs
40,000
20,000
0
6,000
4,000
2,000
0
15,000
10,000
5,000
0
circseq_cup
CIRI2
CIRIquant
CircSplice
find_circ
CirComPara2
CIRCexplorer3
circtools
NCLscan
NCLcomparator
PFv2
ecircscreen
KNIFE
circRNA_finder
segemehl
Sailfish−cir


A
qPCR validation
RNase R validation
amplicon sequencing validation
number of circRNAs
0
50
100
150
0
5
10
15
number of tools by which the circRNA is detected
fail
pass
NA
B
% of all predicted circRNAs
20%
30%
40%
50%
60%
93%
94%
95%
96%
(weighted) compound precision value
GK
GL
GN
G
FK
AB
AH
AF
BK
AK
AN
AL
FL
HK
BL
HL
FN
BN
HN
BF
FH
BH
B
F
A
H
CK
JK
CL
JL
CJ
JN
CN
IK
IL
IN
C
J
K
L
I
N
CIRCexplorer3 A
CirComPara2 B
circRNA_finder C
circseq_cup D
CircSplice E
circtools F
CIRI2 G
CIRIquant H
ecircscreen I
find_circ J
KNIFE K
NCLcomparator L
NCLscan M
PFv2 N
Sailfish-cir O
segemehl P