Benchmarking circRNA Detection Tools from Long-Read Sequencing Using Data-Driven and Flexible Simulation Framework

Anastasia RUSAKOVICH[1], Sebastien CORRE[1], Edouard CADIEU[1], Rose-Marie FRABOULET[1], Marie-Dominique GALIBERT [1,2], Thomas DERRIEN[1]* and Yuna BLUM[1]*

[1] Univ Rennes, CNRS, INSERM, IGDR (Institut de Génétique et Développement de Rennes) - UMR 6290, Av. du Professeur Léon Bernard Bâtiment 4, F-35000 Rennes, France.

[2] Department of Molecular Genetics and Genomics, Hospital University of Rennes (CHU Rennes), 2 Rue Henri le Guilloux, F-35000 Rennes, France.

*Corresponding author, co-last authors

Correspondence: anastasia.rusakovich@univ-rennes.fr, thomas.derrien@univ-rennes.fr, yuna.blum@univ-rennes.fr

ABSTRACT

Circular RNAs (circRNAs) are unique non-coding RNAs with covalently closed loop structures formed through backsplicing events. Their stability, tissue-specific expression patterns, and potential as disease biomarkers have garnered increasing attention. However, their circular structure and diverse size range pose challenges for conventional sequencing technologies. Long-read Oxford Nanopore (ONT) sequencing offers promising capabilities for capturing entire circRNA molecules without fragmentation, yet the effectiveness of bioinformatic tools for analyzing this data remains understudied.

This study presents the first comprehensive benchmark comparison of three specialized tools for circRNA detection from ONT long-read data: CIRI-long (Zhang et al., 2021), IsoCIRC (Xin et al., 2021), and circNICK-lrs (Rahimi et al., 2021). To address the lack of standardized evaluation frameworks, we developed a novel computational pipeline, open-source and freely available, to generate realistic simulated circRNA ONT long-read datasets. Our pipeline integrates several molecular features of circRNAs extracted from established databases and real datasets into NanoSim tool (Hafezqorani et al., 2020) and outputs FASTQ reads reflecting therefore biological diversity and technical properties.

We systematically assessed tool performance across key metrics, including precision, recall, specificity, accuracy, and F1 score. Our analysis revealed distinct

performance profiles: while all tools exhibited high specificity, they varied in precision and their ability to detect different circRNA subtypes, often showing limited sensitivity and precision. Notably, the overlap in detected circRNAs among tools was relatively low. Additionally, computational efficiency varied significantly across the tools. This suggests that relying on a single tool might not be ideal, and combining tools or improving algorithms could be necessary for more accurate circRNA detection from ONT data.
This benchmark provides valuable insights for researchers selecting appropriate tools for circRNA studies using ONT sequencing. Furthermore, our customizable simulation framework, offering a resource to optimize detection approaches and advance bioinformatic tool development for circRNA research is freely available at: https://gitlab.com/bioinfog/circall/nano-circ.

Keywords: circRNA, long-read nanopore sequencing, benchmark comparison, simulation framework

Introduction

Non-coding RNAs (ncRNAs), which make up a significant portion of the transcriptome, have emerged as critical players in various cellular processes, ranging from gene regulation to disease pathogenesis (Esteller, 2011). Among them, circular RNAs (circRNAs) represent a unique class of ncRNAs, formed through non-canonical backsplicing events, resulting in a covalently closed loop structure. Since their discovery, circRNAs have garnered increasing attention due to their stability, tissue-specific expression, and potential roles as biomarkers in diseases, including cancer (Anastasiadou et al., 2018; Verduci et al., 2021). However, their circular structure and diverse size range—spanning from less than 100 to almost 100000 nucleotides—pose significant challenges to conventional sequencing technologies, particularly second-generation sequencing methods, which often rely on read fragmentation.

Recent advancements in third-generation sequencing technologies, such as Oxford Nanopore sequencing, have provided novel opportunities to explore circRNAs in greater detail (Wang et al., 2021). Unlike second-generation methods, Nanopore sequencing offers long-read capabilities that can capture entire circRNA molecules without the need for fragmentation, making it a promising approach for the comprehensive characterization of circRNAs. However, the full potential of long-read sequencing for circRNA discovery and annotation depends on bioinformatics tools that can accurately detect and quantify circRNAs from these datasets.

Despite the growing number of bioinformatics tools developed for circRNA detection (Drula et al., 2024), a comprehensive benchmark comparing their performance on long-read Nanopore sequencing data remains absent. To address this gap, we performed a systematic comparison of 3 circRNA detection tools — CIRI-long (Zhang et al., 2021), IsoCIRC (Xin et al., 2021) and circNICK-lrs (Rahimi et al., 2021).

The complexity of circRNA detection requires a robust and reproducible evaluation framework that goes beyond traditional wet-lab datasets. Experimental data inherently suffers from limitations such as biological variability, sequencing biases, and the challenge of definitively establishing ground truth. Simulated datasets offer a solution to these challenges, providing a controlled environment with precisely known circRNA and linear RNA annotations. By generating *in silico reads* that mimic the molecular and sequencing characteristics of Nanopore technologies, we can create a comprehensive ground truth dataset that allows for rigorous assessment of performance by each of circRNA detection tools.

In this study, we provide a first framework for the generation of simulation data of long-read sequencing of circRNA and present the first benchmark analysis of these 3 tools using simulated datasets. By evaluating their performance across different metrics, we aim to provide valuable insights into the strengths and limitations of each tool, guiding researchers in selecting the most appropriate software for their circRNA studies (Fig.1).

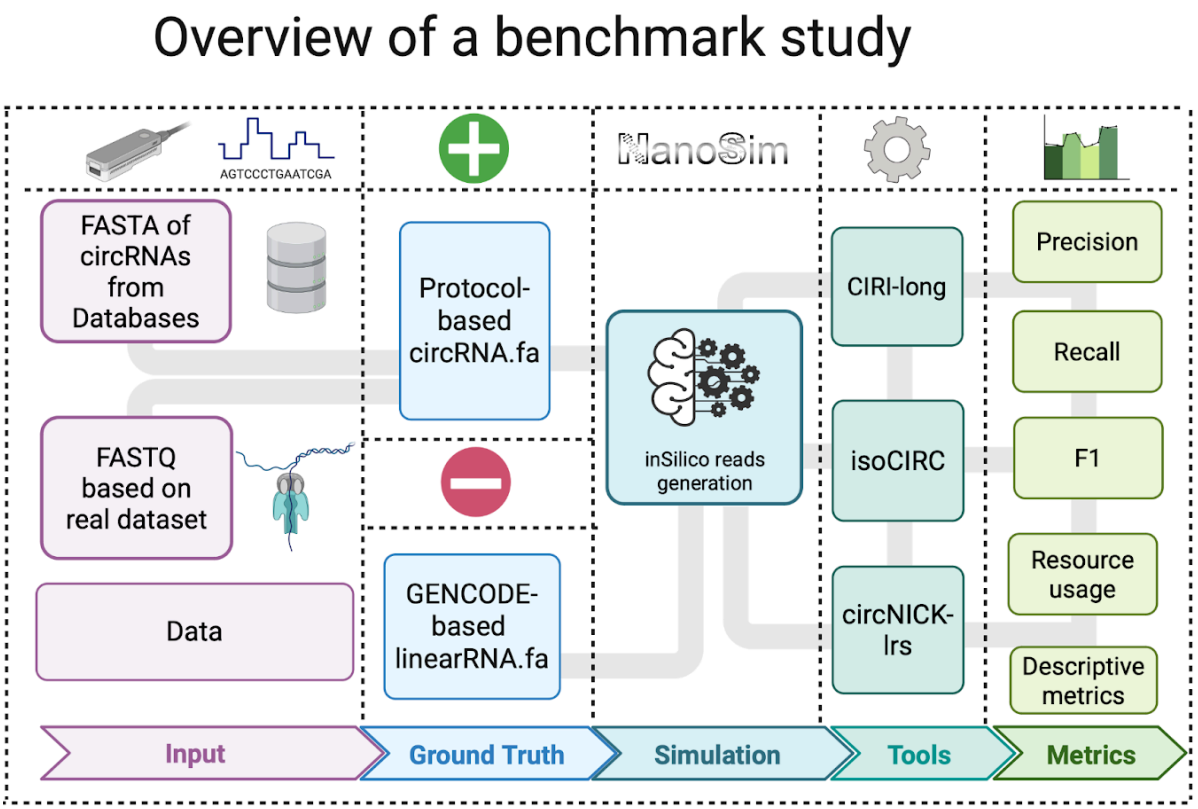

Figure 1 - Overview of the benchmarking study. Schematic representation of the benchmarking framework for circular RNA (circRNA) detection tools, using an available real dataset of circRNA long read sequencing from Nanopore from (mouse brain cell lines) as training model. The workflow progresses through five stages: (I) input FASTQ sequences from circRNA databases (see Methods section) and real wet lab protocol as training model ; (II) ground truth preparation; (III) in silico read simulation with NanoSim version 3.1.0; (IV) computational analysis using CIRI-long, isoCirc, and circNICK-lrs tools; (V) comprehensive performance evaluation through multiple metrics including sensitivity, precision, specificity, accuracy, and F1 score, calculated across different overlap thresholds. Figure created in https://BioRender.com.

Methods

Tools for circular RNA identification using long-read sequencing data

CIRI-long (Zhang et al., 2021): is a computational method designed for profiling circRNAs using nanopore long-read sequencing data based on a rolling circle reverse transcription approach. The algorithm reconstructs full-length circRNA sequences through a multi-step process, beginning with demultiplexing, quality control, and adapter removal using tools such as guppy (v3.3.0), Porechop (v2.0.4), and pycoQC (v2.5.0.14). Next, repetitive pattern identification is performed using a k-mer approach, followed by mapping with minimap2. Back-splice junction (BSJ) detection is then conducted using SPOA, mappy (v2.17), and bwapy (v0.1.4). The method employs filtering strategies, including an 80% sequence similarity threshold and splice signal validation, to ensure high-confidence circRNA detection and full-length isoform reconstruction. More details can be found at [https://github.com/bioinfo-biols/CIRI-long].

isoCirc (Xin et al., 2021): is a method for sequencing and characterizing full-length circular RNA isoforms. The approach relies on rolling circle amplification followed by nanopore long-read sequencing. The computational pipeline begins with demultiplexing, quality control, and adapter removal using guppy (v2.1.3). Repetitive pattern identification is then performed with Tandem Repeat Finder (v4.0.9), and mapping is conducted using minimap2 (v2.17). BSJ detection also utilizes minimap2 (v2.17). The pipeline incorporates filtering methods such as tandem repeat detection, multi-tiered alignment scoring, and stringent validation of back-spliced junctions (BSJs) and forward-spliced junctions (FSJs). isoCirc can identify circRNA isoforms with precise exonic compositions, including those ≥500 nucleotides in length, and characterizes alternative splicing events within circRNAs. More information can be found at [https://github.com/Xinglab/isoCirc/tree/main].

circNick-LRS (Rahimi et al., 2021): is a computational workflow for profiling circRNAs from nanopore long-read sequencing data. The pipeline begins with demultiplexing, quality control, and adapter removal using Guppy (v3.4.5) and NanoFilt (v2.6.0). Reads are filtered for quality (≥Q7) and length (≥250 bp). Mapping is performed using pblat (v35) against the human (hg19) or mouse (mm10) genome. BSJ detection utilizes pblat (v35) and bedtools (v2.29.2), requiring a minimum Blat score of 30 on both sides of the BSJ. Reads are filtered based on genomic criteria, retaining only those aligning to the same strand, within 1 Mb, and in reverse order relative to the read sequence. Full-length circRNA isoform identification involves aligning reads to the closest refSeq exon using bedtools. Reads overlapping annotated circRNAs (≥95%) are corrected and retained. Unique circRNA regions are annotated with gene, circBase ID, circAtlas ID, CIRCpedia ID, read count, and mean annotation values. Further details can be found at [https://github.com/omiics-dk/long_read_circRNA].

Tool for ONT in-silico generation of long-read sequencing data

NanoSim (Hafezqorani et al., 2020): is a fast and scalable read simulator that captures the technology-specific features of Oxford Nanopore Technologies (ONT) data. The tool analyzes ONT reads from experimental data to model read features such as length, error profiles, and k-mer biases. In its latest version (v3.0), NanoSim supports the simulation of genomic, transcriptomic (cDNA and direct RNA), and metagenomic reads, accommodating features like intron retention events and chimeric reads. The simulation process involves characterizing the input data to learn these features and then generating synthetic reads that mimic the observed characteristics. NanoSim is implemented in Python and utilizes tools such as minimap2 for alignment and HTSeq for efficient reading of SAM alignment files. Pre-trained models for organisms like E. coli and S. cerevisiae are available, and users can also train NanoSim on their own datasets to tailor the simulation to specific applications. More details and access to the software can be found at [https://github.com/bcgsc/NanoSim].

Wet-lab dataset

Mouse Brain Dataset: The FASTQ file from the Zhang et al. (2021) study was obtained from the National Genomics Data Center (China National Center for

Bioinformation) under the accession number CRA003317 [https://ngdc.cncb.ac.cn/gsa/browse/CRA003317].

Ground truth construction

circRNA Databases: We utilized two circRNA databases for our analysis: CircAtlas v.3 (Wu et al., 2020) and CircBase (Glažar et al., 2014) as they provided both human and mouse circRNAs with their sequences in FASTA format that allowed us to extract features from them.

Genome assembly: We used the GRCm38.p4 mouse genome assembly, as our reference genome as circNICK-lrs tool is compatible only with this version of the reference genome. Reference genome and genomic annotations were obtained from the GENCODE consortium (Mudge et al., 2025), specifically the Gencode version M10 mouse gene annotation file.

Data simulation

Feature extraction from databases: To understand the composition of circular RNAs, we have analysed circRNAs found at an intersection between circRNA databases. We developed a computational pipeline to extract comprehensive features from circRNA annotations using Python-based bioinformatics tools (pybedtools, pysam, pandas). The analysis integrated genomic coordinates, sequence information, and transcriptomic context from input BED files, reference genomes (FA), and gene annotation (GTF) files. Our methodology extracted key circRNA characteristics, including genomic location (chromosome, start, end, strand), mature RNA length, and splice site details. We classified splice sites based on canonical motifs (GT-AG, GC-AG, AT-AC) and performed intersectional analysis to annotate gene and transcript types.

Feature extraction from nanopore sequences: To understand the effect of different wet lab protocols on data we have developed a comprehensive computational pipeline to extract and visualize detailed sequencing characteristics from FASTQ files (Fig 2A). The analysis systematically quantified read metrics, encompassing comprehensive length distribution and repeat pattern present in wet lab data.

in-Silico circRNA generation: Building upon our database and wet-lab protocol feature extraction pipelines, we developed an approach to simulate circular RNA types (Fig. 2B). Leveraging genomic features extracted from exonic, intronic, and intergenic regions, we generated four circRNA types: exonic circRNAs (ecircRNAs), circular intronic RNAs (ciRNAs), exon-intron circRNAs (EIciRNAs) and intergenic circRNAs. Key generation features included random sequence extraction, length-constrained generation, splice site preference and rolling circle amplification. We implemented sequence quality filtering to exclude sequences with more than 10% unidentified nucleotides (N bases). The generation process dynamically selected sequence start positions, controlled rolling circle replication defined by user, and captured detailed metadata including transcript identifiers, gene coordinates, and strand information. To validate the simulated circRNAs, we employed BLAT to identify and verify backsplice

junctions. The BLAT analysis screened generated sequences, filtering reads with characteristic circular RNA mapping patterns to ensure biological plausibility by detecting potential backsplice junction signatures.

in-silico circRNA long-reads data simulation: We utilized NanoSim version 3.1.0 to simulate Nanopore sequencing reads from the circRNA sequences generated in our previous simulation step to use as a positive ground truth and from linear RNA sequences to use as a negative ground truth (Fig.2C). The simulation uses multiple input files: a control FASTQ file (CRR194180.fq), reference genome (GRCm38.p4), and transcriptome annotation (gencode.vM10).

The NanoSim simulation process involved four stages:

- Read Analysis: Characterizing sequencing properties using the control FASTQ file with minimap2 alignment
- CircRNA Read Simulation: Generating 200,000 reads from our previously created circRNA.fa file
- Transcriptome Quantification: Analyzing expression levels
- Linear Read Simulation: Generating 200,000 linear RNA reads from transcriptome annotation

Key NanoSim parameters included:

- Sequencing technology: Guppy basecaller
- Read type: cDNA_1D
- Output format: FASTQ
- Intron retention model: Disabled

The simulation generated a comprehensive dataset that captured the molecular and sequencing characteristics of circular and linear RNAs, providing a computationally derived representation of Nanopore sequencing data.

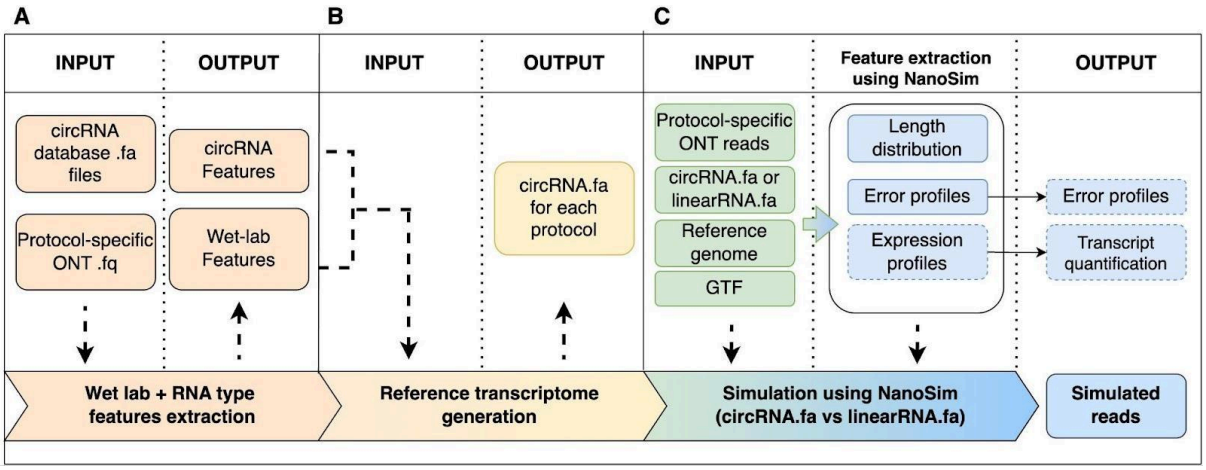

Figure 2 - Computational Workflow for Nanopore Sequencing circRNA Simulation. The workflow integrates wet-lab Oxford Nanopore Technologies (ONT) read

characteristics with computational simulation techniques. Panels A-C depict sequential stages of data processing and simulation: (A) initial feature extraction from circRNA databases and experimental reads, (B) generation of protocol-specific reference circRNA sequences, and (C) NanoSim-based simulation generating detailed computational read models.

Standardisation of tool output

To ensure comprehensive and standardized evaluation, we converted all tool prediction results to BED12 format, which provides a consistent representation of genomic features including chromosomal location, exon structure, and strand information. This normalization allowed for precise comparative analysis across different circRNA detection tools despite different output formats.

Performance metrics

The bedtools intersect approach utilized three critical parameters to ensure precise genomic feature comparison. The -f (fraction overlap) parameter specifies the minimum fraction of overlap required between predicted and ground truth circRNA annotations, controlling the stringency of feature matching. We tested three different overlap values: 0.25, 0.5, and 0.75. At 0.25, at least 25% of a feature must overlap, representing a lenient matching criterion. The 0.5 threshold requires 50% overlap, providing moderate stringency, while the 0.75 value demands at least 75% overlap, representing a more stringent matching approach.

The -split option was used for long-read circRNA detection, enabling exon-aware intersection by respecting exon boundaries during comparison. This parameter ensures that circRNA predictions are evaluated considering their specific exonic composition, allowing for a more nuanced performance assessment that accounts for multi-exonic transcript structures. By implementing this option, we can more accurately capture the complexity of circRNA annotations.

Additionally, the -r (reciprocal overlap) parameter ensures that the overlap is symmetric between prediction and ground truth. This approach prevents asymmetric matching where only one feature partially overlaps, providing a stringent criterion for considering a prediction as a true positive. The reciprocal requirement adds an extra layer of precision to our computational analysis.

We utilized bedtools version v2.31.1 to systematically compare BED12 format annotations of in silico positive and negative ground truth reads against tool outputs. By utilizing three distinct overlap fractions—0.25, 0.5, and 0.75—we captured a comprehensive range of detection scenarios, from lenient criteria capturing broader genomic matches to more stringent approaches demanding closer genomic alignment.

We utilized a set of performance metrics to assess the accuracy of circRNA detection. True Positives (TP) were defined as circRNAs correctly identified by the tool, fitting the reciprocal fraction overlap criteria between predicted and ground truth circRNA annotations. False Positives (FP) represented computational artifacts or

misannotated RNA predictions that did not meet our reciprocal fraction overlap criteria despite being predicted as circRNAs by a tool. False Negatives (FN) comprised circRNAs present in the simulated dataset but undetected by the tools, indicating potential limitations in circRNA identification methodologies.

We calculated three primary performance metrics to provide a comprehensive evaluation of tool performance. Precision, calculated as TP / (TP + FP), quantifies the proportion of predicted circRNAs that are genuinely correct, effectively measuring the tools' specificity in avoiding spurious predictions. Recall (or Sensitivity), calculated as TP / (TP + FN), assesses the tools' ability to detect existing circRNAs, revealing their capacity to comprehensively identify circRNA events within the dataset. The F1 Score, calculated as the harmonic mean of precision and recall, provides a balanced performance metric that simultaneously considers both false positives and false negatives, offering a nuanced view of each tool's overall detection capabilities.

Results

Feature extraction from intersected databases and real data for the generation of realistic simulated datasets results

Our circRNA simulation pipeline integrates Nanopore sequencing data from Zhang et al., database annotations, and computational modeling. (Fig. 3).

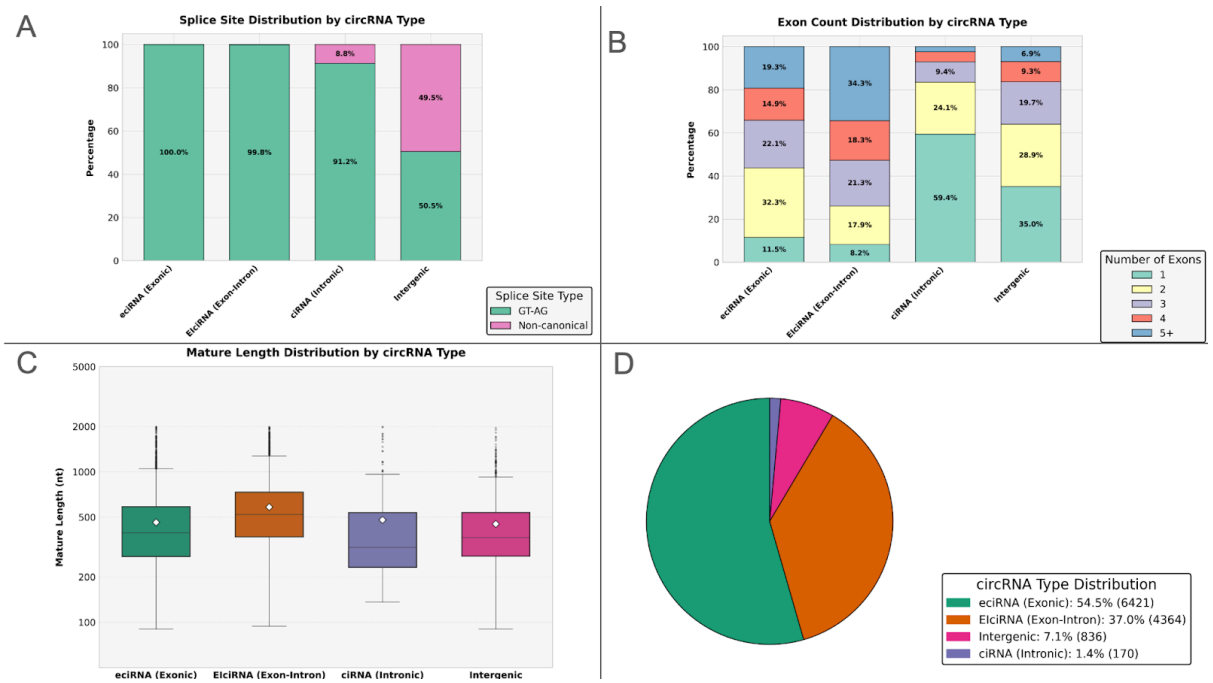

Figure 3 - Characterization of Circular RNA molecular features across circRNA types from CircAtlas and CircBase databases. Panels A-D depict: (A) Splice site composition across circRNA types, (B) Exon count distribution, (C) mature length (exonic concatenation) distribution, and (D) circRNA type distribution.

Among 11,791 circRNAs common to the two databases, canonical GT-AG splice sites dominated the junction boundaries of exonic, exon-intronic, and intronic circRNAs (100%,

99.8%, and 91.2%, respectively), while intergenic circRNAs showed significantly higher usage of non-canonical splice sites (49.5%) (Fig. 3A).

The exon composition analysis demonstrated substantial variation in complexity across circRNA types (Fig. 3B). Notably, intronic circRNAs (ciRNAs) were predominantly single-exon structures (59.4%), whereas exonic circRNAs (ecircRNAs) showed more diverse exon count distributions with significant proportions containing 2 exons (32.3%) and 3 exons (22.1%). Exon-intron circRNAs (EIciRNAs) exhibited the highest complexity, with 34.3% containing 5 or more exons. Intergenic circRNAs showed an intermediate distribution pattern, with substantial representation across various exon count categories.

Mature length distribution analysis revealed that EIciRNAs possessed the highest median length, followed by ecircRNAs, intergenic, and ciRNAs (Fig. 3C). This pattern reflects the fundamental structural differences between these circRNA subtypes, with EIciRNAs incorporating both exonic and intronic sequences, contributing to their increased overall length.

When examining the overall distribution of circRNA types in the reference databases (Fig. 3D), ecircRNAs constituted the majority (54.5%, n=6421), followed by EIciRNAs (37.0%, n=4364), intergenic circRNAs (7.1%, n=836), and ciRNAs (1.4%, n=170). This distribution highlights the predominance of exon-derived circular RNAs in the current reference datasets and reflects potential biases in detection methodologies favoring exonic circRNA identification.

These different features were incorporated in our simulation framework based on NanoSim to generate circRNA molecules reflecting biological variations. We verified that the simulated data accurately reflected the characteristics of the input parameters. Specifically, the distributions of circRNA types, splicing patterns, and repeat number variability observed in real Nanopore data and database annotations were preserved in the simulated output

Descriptive comparison of circRNA outputs across tools

We performed a comparison of three circRNA detection tools (CIRI-long, isoCirc, and circNick-lrs) run on a simulated dataset ($n_{circRNAs}$ = 7503, n_{reads} = 400000), to evaluate their performance characteristics and detection biases (Fig. 4).

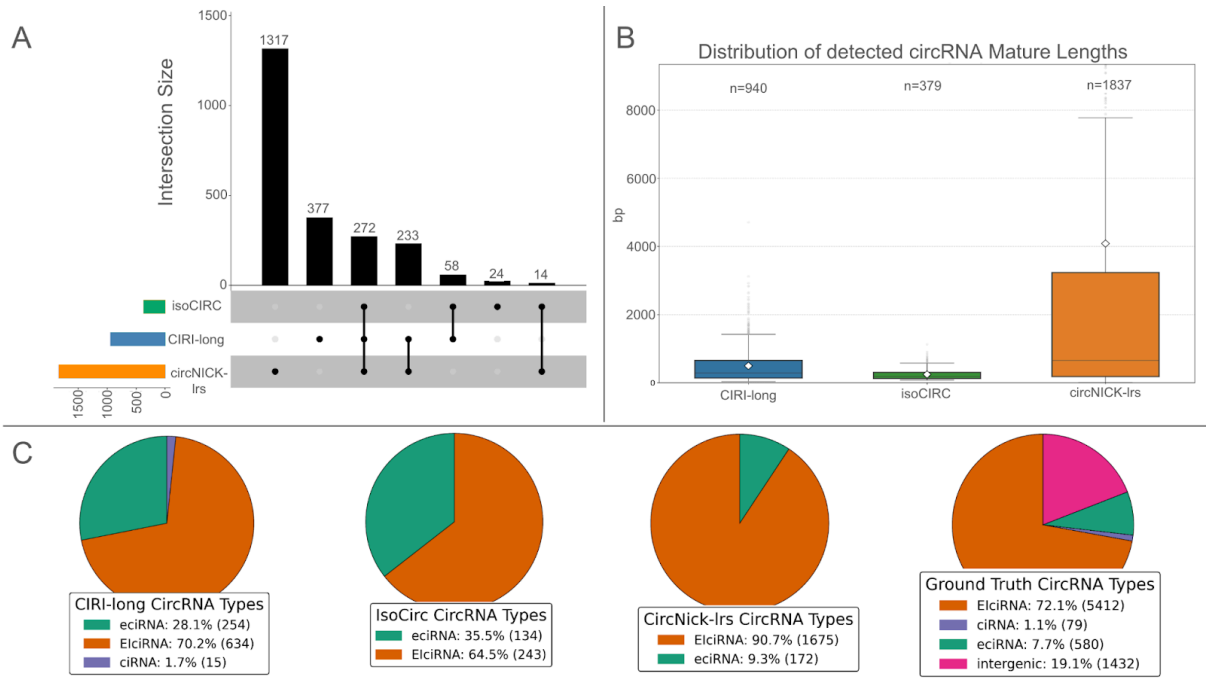

Figure 4 - Descriptive comparison of circRNA outputs across tools. Panels A-C depict: (A) intersection analysis showing shared and unique isoform detection across tools, (B) boxplots of circRNAs mature length distribution found by each tool, and (C) proportional distribution of circRNA subtypes identified by each tool compared to ground truth.

The intersection analysis revealed significant differences in detection capability among the three tools (Fig. 4A). Notably, circNick-lrs demonstrated the highest unique detection capability, identifying 1317 circRNAs not found by other methods. CIRI-long uniquely detected 377 circRNAs, while isoCirc showed the lowest unique detection with only 24 circRNAs. The consensus between all three tools was limited to 272 circRNAs (11.85% of total detections), highlighting the complementary nature of these detection approaches. Additional intersection patterns included 233 circRNAs detected by both circNick-lrs and CIRI-long but missed by isoCirc, and 58 circRNAs identified by both CIRI-long and isoCirc but not by circNick-lrs.

Analysis of mature length distribution revealed striking differences between the tools (Fig. 4B). CircNick-lrs demonstrated substantially greater capability in detecting longer circRNAs, with a much wider interquartile range extending beyond 3000 bp, and a significant proportion of detections in the higher ranges (25.97% between 1-5 kb, 8.44% between 5-10 kb, and 9.04% above 10 kb). In contrast, both CIRI-long and isoCirc predominantly detected shorter circRNAs. CIRI-long showed an intermediate range capability with 66.60% of detections under 500 bp and 14.47% in the 1-5 kb range, but none above 5 kb. IsoCirc exhibited the most restricted length distribution, with 89.71% of detections under 500 bp, nearly no representation (0.26%) in the 1-5 kb range, and none above 5 kb. This severe limitation in isoCirc can be attributed to its built-in length cutoff at 4000 nucleotides, making it incapable of detecting longer circRNAs without modifying the original code in the TRF (Tandem Repeat Finder) section.

When examining the proportional distribution of circRNA types (Fig. 4C), we observed substantial variation in type bias across tools compared to ground truth. Ground truth composition revealed a distribution of 72.1% EIciRNAs, 7.7% ecircRNAs, 1.1% ciRNAs, and 19.1% intergenic circRNAs. However, all three tools showed significant detection biases. CircNick-lrs demonstrated extreme bias toward EIciRNAs (90.7%) with limited ecircRNA detection (9.3%). CIRI-long showed better balance between EIciRNAs (70.2%) and ecircRNAs (28.1%), with minimal ciRNA detection (1.7%). IsoCirc exhibited a more balanced detection of EIciRNAs (64.5%) and ecircRNAs (35.5%) but completely missed ciRNAs and intergenic circRNAs. Notably, intergenic circRNAs were missed by all three tools despite representing 19.1% of the ground truth, while ciRNAs were detected only by CIRI-long, suggesting limitations in current circRNA detection methodologies. These findings highlight the importance of tool selection based on the specific circRNA types of interest and suggest that combining multiple tools may provide the most comprehensive circRNA detection.

Performance evaluation of circRNA detection Tools

To evaluate the three circRNA detection tools, we assessed their performance across multiple dimensions including detection accuracy, expression profiling, computational efficiency, and resource requirements (Fig. 5).

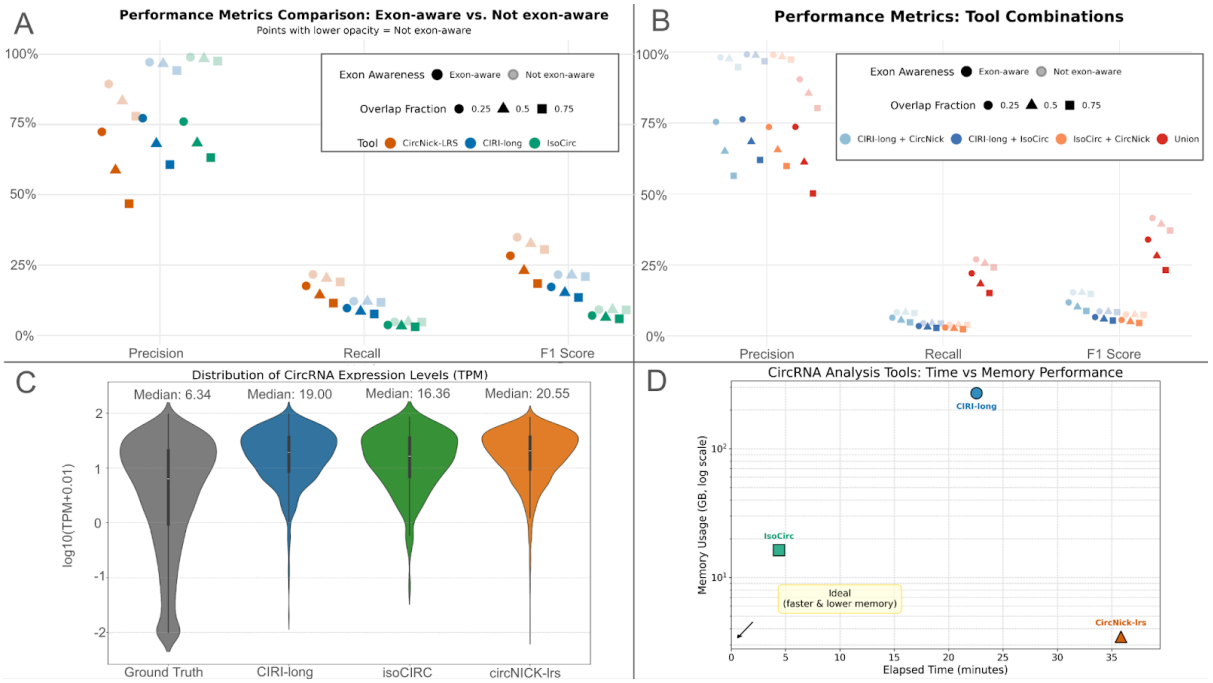

Figure 5 - Performance Evaluation of circRNA Detection Tools on Simulated Datasets. Panels A-D depict: (A) quantitative comparison of performance metrics, (B) quantitative comparison of performance metrics on combinations and union of tools, (C) violin plot distribution of circRNA expression levels (TPM) detected by each tool compared to ground truth, (D) dot plot showing peak memory usage requirements for each computational approach and displaying processing time per read when analyzing a standardized dataset of 400,000 reads.

Performance metrics analysis revealed varying detection capabilities across different tools and overlap fractions (Fig. 5A). At the 0.25 fraction, IsoCirc demonstrated the

highest precision (99%), followed by CIRI-long (0.97) and CircNick-LRS (0.89). Notably, all tools exhibited low recall, with CircNick-LRS performing significantly better (0.22) compared to IsoCirc (0.05) and CIRI-long (0.12). As the overlap fraction increased to 0.5, a consistent decline in precision was observed: IsoCirc dropped to 0.98, CIRI-long to 0.97, and CircNick-LRS to 0.83. Recall remained persistently low across all tools: CircNick-LRS at 0.20, CIRI-long at 0.12, and IsoCirc at 0.05. At the 0.75 fraction, the precision decline continued: IsoCirc decreased to 0.98, CIRI-long to 0.94, and CircNick-LRS further dropped to 0.78, with recall values remaining consistently low across all tools: CircNick-LRS at 0.19, CIRI-long at 0.12, and IsoCirc at 0.05. Based on this, we can see that each tool exhibits unique performance profiles. CircNick-LRS shows the most balanced approach with relatively higher recall, CIRI-long maintains consistent performance, while IsoCirc prioritizes extreme precision at the cost of sensitivity.

Performance metrics analysis revealed subtle changes in detection capabilities when implementing the -split option (see Methods for more details) across different tools and overlap fractions (Fig. 5A). The observed performance decline with the -split option reflects the critical challenge of precisely mapping circRNA isoform exon structures. This stringent criterion dramatically reduces detection sensitivity by enforcing a more rigorous matching of exon structures, essentially demanding that the computational tools precisely align with the ground truth of circRNA splicing events. The substantial drops in precision and recall across all tools—most notably for CircNick-LRS—highlight the fundamental difficulty in computationally reconstructing exact circRNA isoform boundaries.

When we merged the circRNA sets of the three tools, significant performance variations emerged (Fig. 5B). The union of all three tools produced the highest recall values across all fractions, substantially outperforming any individual tool or pairwise combination. However, this comprehensive approach came at the cost of reduced precision (0.90) compared to any individual tool or two-tool combination. This trade-off highlights the complementary nature of these detection methods, where combining all tools captures more true positives but introduces additional false positives.

The expression level analysis (Fig. 5C) revealed that all three tools demonstrated significant bias toward detecting highly expressed circRNAs. While the ground truth dataset had a median expression of 6.34 TPM, the detected circRNAs showed substantially higher median expression levels: CircNick-lrs (20.55 TPM), CIRI-long (19.0 TPM), and IsoCirc (16.36 TPM). This observation aligns with detection capabilities, as highly expressed circRNAs generate more supporting reads, making them easier to identify. Minimum expression values were similar across tools, with CIRI-long, CircNick-lrs, and the ground truth dataset showing a minimum expression of 1.0×10^{-2} TPM, while IsoCirc had a slightly higher minimum TPM of approximately 6.0×10^{-2} TPM.

Resource utilization varied dramatically between tools (Fig. 5D). CIRI-long exhibited exceptionally high memory consumption (269.95 GB), representing a substantial computational burden. In contrast, IsoCirc (16.28 GB) and especially CircNick-lrs (3.46 GB) demonstrated significantly more efficient memory utilization, making them more

accessible for researchers with limited computational resources. Processing speed analysis showed IsoCirc as the clear performance leader, requiring only 0.66 milliseconds per read. CIRI-long demonstrated intermediate efficiency (3.39 ms/read), while CircNick-lrs required substantially more processing time (5.38 ms/read), making it the slowest of the three tools despite its memory efficiency.

Discussion

Our benchmarking study reveals the current state of circRNA detection tools for long-read Nanopore sequencing, highlighting both the promising capabilities and significant challenges in accurately identifying circular RNAs. The analysis underscores the complementary strengths and limitations of current computational approaches, providing critical insights for researchers navigating circRNA detection. Each tool demonstrated distinct performance characteristics that warrant careful consideration.

IsoCirc emerged as the computational efficiency leader, requiring only 0.66 milliseconds per read, making it an attractive option for researchers with limited computational resources. However, this efficiency came at the cost of significant limitations, including a built-in length cutoff at 4000 nucleotides that severely restricts its ability to detect longer circRNAs.

CIRI-long offered a more balanced approach to circRNA detection, with intermediate performance across metrics. However, it imposed a substantial computational burden, consuming an exceptional 269.95 GB of memory—a significant constraint for many research environments.

CircNick-lrs distinguished itself by demonstrating the most comprehensive circRNA detection, particularly for longer circRNAs. It uniquely identified 1,317 circRNAs not found by other methods, with a notable capability to detect circRNAs across diverse length ranges—25.97% between 1-5 kb, 8.44% between 5-10 kb, and 9.04% above 10 kb. However, the tool came with significant limitations, most notably its restriction to only built-in mouse and human reference genomes (mm10 and hg19) with predefined annotation files, preventing its use with custom or alternative genome annotations. Despite its extensive detection capabilities, the tool also required the most processing time at 5.38 ms/read, presenting an additional performance challenge for researchers with large datasets and struggled the most with predicting correct exon structure.

Notably, the extremely low intersection between tools—with only 272 circRNAs (11.85% of total detections) identified by all three methods—strongly suggests that relying on a single detection tool is suboptimal. This stark divergence in detection capabilities demonstrates the methodological challenges in circRNA identification. Each tool essentially acted as a unique lens, capturing distinct aspects of the circRNA landscape that other tools missed. The tools also exhibited significant bias towards detecting highly expressed circRNAs, with median expression levels substantially higher than the ground truth dataset, further complicating comprehensive circRNA characterization.

Our analysis suggests that the choice of tool combination should be guided by specific research priorities. For maximum sensitivity (recall), the union of all tools provides the highest likelihood of detecting true circRNAs, though with more false positives than other approaches. For highest precision, the combination of CIRI-long and isoCIRC maintains the highest precision values across all settings, making it ideal for applications requiring high confidence in detected circRNAs. For balanced performance,

while no combination achieves ideal balance, the union approach offers the best compromise between precision and recall, as reflected in its superior F1 scores compared to any two-tool combination.

The study also revealed notable limitations in current detection methodologies when tools can not rely on database validation or standard circRNA composition. All three tools completely missed intergenic circRNAs, which represented 19.1% of the ground truth dataset. Similarly, ciRNAs were detected only by CIRI-long, suggesting significant gaps in comprehensive circRNA identification across types.

Performance metrics consistently demonstrated low sensitivity across all tools. This highlights the formidable challenge of accurately identifying exact exon structures in circRNA detection.

Limitations of the current study include the exclusion of the circFL-seq tool (Liu et al., 2021) from the benchmarking analysis. This tool was initially considered for inclusion but was ultimately omitted from this version of the paper due to unresolved dependency issues that prevented its execution on our computational cluster. The dataset for this study was generated using CIRI-long dataset parameters, which may introduce some methodological bias. In future versions of this study, we plan to expand our approach by using FASTQ files from different wet-lab protocols as a base for in-silico dataset generation, as well as incorporating various biological sources (mouse brain tissue and HEK293 cell cultures) for a more comprehensive performance evaluation across diverse experimental conditions. This expansion will provide a more robust assessment of tool performance in varied biological contexts and sequencing approaches, further enhancing the utility of our benchmarking framework for the circRNA research community.

Notably, other circRNA detection tools in the study also presented significant installation challenges, lacking container environments and experiencing varying difficulties during cluster deployments. These installation difficulties can pose substantial barriers to researchers without bioinformatics expertise, potentially limiting the broader adoption and comparative evaluation of emerging circRNA detection methodologies. This highlights a broader challenge in bioinformatics tool comparison: the practical difficulties of integrating different computational approaches, especially those with complex dependencies or limited ongoing maintenance. Future iterations of this benchmarking study should continue efforts to incorporate additional circRNA detection tools, potentially developing more standardized approaches to tool integration, documentation, and user accessibility.

In conclusion, this study provides a comprehensive evaluation of circRNA detection tools, emphasizing the need for continued methodological refinement. Researchers must carefully consider the specific requirements of their studies—computational resources, circRNA length, and detection sensitivity—when selecting circRNA detection tools and are advised, when possible, to use their combination.

By integrating wet-lab data, database annotations, and computational modeling, our framework captures circRNA biogenesis complexity and provides a valuable resource for studying circular RNA function and regulation. Our study comes with a customizable framework that allows researchers to analyse database features given mature circRNA sequences from any database as an input or protocol-specific features given .fastq files corresponding to a protocol and tailor circRNA simulation parameters to their specific needs.

Acknowledgements

XXXXinsert here the sentence of the form Preprint version XXXX of this article has been peer-reviewed and recommended by Peer Community In XXXName of the PCI (https://doi.org/10.24072/pci.xxx [replace by the right doi of the recommendation], *Citation of the recommendation* eg Kalinkat, 2023).

The authors would like to thank the Gene Expression and Oncogenesis team, the Dog Genetics team and Biology in Silico 2.0 group (IGDR CNRS UMR6290) for helpful discussions. We also acknowledge the GenOuest bioinformatics core facility (https://www.genouest.org/) for providing the computing infrastructure.

Data, scripts, code, and supplementary information availability

Data are available online: https://doi.org/10.1038/s41587-021-00842-6 (Zhang et al., 2021);

Scripts and code are available online: https://doi.org/10.5281/zenodo.15229596 (Rusakovich et al, 2025).

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Funding

This study received financial support from the Ligue National Contre le Cancer (LNCC) Départements du Grand-Ouest. AR is a recipient of a doctoral fellowship from the French Ministry of Higher Education and Research.

References

Anastasiadou, E., Jacob, L. S., & Slack, F. J. (2018). Non-coding RNA networks in cancer. Nature Reviews Cancer, 18(1), 5–18. https://doi.org/10.1038/nrc.2017.99

Drula, R., Braicu, C., & Neagoe, I.-B. (2024). Current advances in circular RNA detection and investigation methods: Are we running in circles? Wiley Interdisciplinary Reviews. RNA, 15(3), e1850. https://doi.org/10.1002/wrna.1850

Esteller, M. (2011). Non-coding RNAs in human disease. Nature Reviews Genetics, 12(12), 861–874. https://doi.org/10.1038/nrg3074

Glažar, P., Papavasileiou, P., & Rajewsky, N. (2014). circBase: A database for circular RNAs. RNA, 20(11), 1666–1670. https://doi.org/10.1261/rna.043687.113

Hafezqorani, S., Yang, C., Lo, T., Nip, K. M., Warren, R. L., & Birol, I. (2020). Trans-NanoSim characterizes and simulates nanopore RNA-sequencing data. GigaScience, 9(6), giaa061. https://doi.org/10.1093/gigascience/giaa061

Liu, Z., Tao, C., Li, S., Du, M., Bai, Y., Hu, X., Li, Y., Chen, J., & Yang, E. (2021). circFL-seq reveals full-length circular RNAs with rolling circular reverse transcription and nanopore sequencing. eLife, 10, e69457. https://doi.org/10.7554/eLife.69457

Mudge, J. M., Carbonell-Sala, S., Diekhans, M., Martinez, J. G., Hunt, T., Jungreis, I., Loveland, J. E., Arnan, C., Barnes, I., Bennett, R., Berry, A., Bignell, A., Cerdán-Vélez, D., Cochran, K., Cortés, L. T., Davidson, C., Donaldson, S., Dursun, C., Fatima, R., … Frankish, A. (2025). GENCODE 2025: Reference gene annotation for human and mouse. Nucleic Acids Research, 53(D1), D966–D975. https://doi.org/10.1093/nar/gkae1078

Rahimi, K., Venø, M. T., Dupont, D. M., & Kjems, J. (2021). Nanopore sequencing of brain-derived full-length circRNAs reveals circRNA-specific exon usage, intron retention and microexons. Nature Communications, 12(1), 4825. https://doi.org/10.1038/s41467-021-24975-z

Verduci, L., Tarcitano, E., Strano, S., Yarden, Y., & Blandino, G. (2021). CircRNAs: Role in human diseases and potential use as biomarkers. Cell Death & Disease, 12(5), 468. https://doi.org/10.1038/s41419-021-03743-3

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. Nature Biotechnology, 39(11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Wu, W., Ji, P., & Zhao, F. (2020). CircAtlas: An integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. Genome Biology, 21(1), 101. https://doi.org/10.1186/s13059-020-02018-y

Xin, R., Gao, Y., Gao, Y., Wang, R., Kadash-Edmondson, K. E., Liu, B., Wang, Y., Lin, L., & Xing, Y. (2021). isoCirc catalogs full-length circular RNA isoforms in human transcriptomes. Nature Communications, 12(1), 266. https://doi.org/10.1038/s41467-020-20459-8

Zhang, J., Hou, L., Zuo, Z., Ji, P., Zhang, X., Xue, Y., & Zhao, F. (2021). Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long. Nature Biotechnology, 39(7), 836–845. https://doi.org/10.1038/s41587-021-00842-6