nature genetics

Review article

https://doi.org/10.1038/s41588-025-02157-7

Circular RNA discovery with emerging sequencing and deep learning technologies

Received: 23 November 2024

Accepted: 7 March 2025

Published online: 17 April 2025

Check for updates

Jinyang Zhang [1] & Fangqing Zhao [1,2,3]

Circular RNA (circRNA) represents a type of RNA molecule characterized by a closed-loop structure that is distinct from linear RNA counterparts. Recent studies have revealed the emerging role of these circular transcripts in gene regulation and disease pathogenesis. However, their low expression levels and high sequence similarity to linear RNAs present substantial challenges for circRNA detection and characterization. Recent advances in long-read and single-cell RNA sequencing technologies, coupled with sophisticated deep learning-based algorithms, have revolutionized the investigation of circRNAs at unprecedented resolution and scale. This Review summarizes recent breakthroughs in circRNA discovery, characterization and functional analysis algorithms. We also discuss the challenges associated with integrating large-scale circRNA sequencing data and explore the potential future development of artificial intelligence (AI)-driven algorithms to unlock the full potential of circRNA research in biomedical applications.

CircRNAs constitute a distinct class of covalently closed RNAs that are widely distributed across various organisms. Recent research has revealed their expanding functions, including sequestration of micro(mi)RNA[1,2] and RNA-binding proteins (RBPs)[3], regulation of mitochondrial reactive oxygen species[4], encoding cryptic peptides[5] and modulation of innate immunity[6]. Notably, their circular structure confers resistance to degradation by endogenous RNA exonucleases, imparting exceptional stability compared to linear RNA counterparts. This stability advantage has been leveraged to engineer circRNAs for various applications, such as vaccines for severe acute respiratory syndrome coronavirus 2 (ref. 7), genome-editing platforms[8], RNA editing[9] and RNA therapeutics[10,11]. With increasing interest in circRNAs, a comprehensive profiling of their molecular composition and spatiotemporal regulation is crucial for understanding their roles in disease and developing circRNA-based therapeutics.

However, profiling circRNA molecular sequences and cellular heterogeneity remains a substantial challenge. CircRNAs are generally expressed at low levels in many tissues, with approximately 10,000 copies per HeLa cell[6], representing an extremely small fraction within transcriptome sequencing data. This scarcity complicates comprehensive circRNA profiling in cells. Studies performed over the last decade have demonstrated the tissue and organism specificity of circRNAs, suggesting that bulk RNA sequencing (RNA-seq) approaches may yield a biased view of circRNA expression profiles that is affected by varying cell proportions and compositions across samples. Therefore, in-depth investigation of the cellular landscape of circRNAs is imperative.

The biological functions of circRNAs largely depend on *cis*-acting elements embedded in their sequences. Similar to linear transcripts, alternative splicing of circRNAs generates extensive isoform diversity, expanding their functional repertoire. Thus, accurate characterization of full-length circRNA isoforms has become pivotal in circRNA research. Traditional algorithms for circRNA identification rely on distinctive features at the back-splicing junction (BSJ) to detect these events based on short-read RNA-seq data. However, the high sequence similarity between circRNAs and their linear counterparts makes it difficult to distinguish circRNAs from their linear counterparts, especially in overlapping exonic regions, thereby hindering the reconstruction of full-length circRNA isoforms.

Recent advances in long-read and single-cell RNA-seq (scRNA-seq) techniques have substantially enhanced our ability to investigate circRNA heterogeneity in depth. In particular, various efforts have been made to achieve comprehensive profiling of full-length circRNA isoforms using long-read sequencing technologies, which overcomes previous limitations in circRNA reconstruction efficiency and accuracy[12–14].

[1]Institute of Zoology, Chinese Academy of Sciences, Beijing, China. [2]Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China. [3]University of Chinese Academy of Sciences, Beijing, China. e-mail: zhangjinyang@ioz.ac.cn; zhfq@ioz.ac.cn

At the same time, the development of single-cell whole-transcriptome sequencing methods has enabled circRNA profiling at single-cell resolution. Recent studies have integrated large-scale scRNA-seq data to elucidate the cellular landscape of circRNAs[15]. In addition, AI-based algorithms have been employed in predicting cell type-specific circRNA expression, providing insights into the spatiotemporal regulation of circRNAs in disease and development[16].

In this Review, we outline recent advances in circRNA identification, quantification and differential expression analysis. We also examine how new sequencing techniques and AI-driven algorithms are advancing the understanding of circRNA molecular and cellular heterogeneity. Additionally, we discuss the emerging role of and challenges associated with integrating large-scale circRNA datasets. Finally, we discuss future prospects for optimizing circRNA characterization approaches and leveraging AI to expedite functional analyses of circRNAs.

Quantitative analysis of circRNAs

CircRNAs are formed through the ligation of the 3′ and 5′ ends of circularized exons via back-splicing. This process is catalyzed by the exon definition complex on long exons[17] and facilitated by flanking intronic complementary sequences[18] and RBPs[19], which bring the splice sites into close proximity. Back-splicing results in unique chimeric sequence features at BSJs that distinguish circRNAs from linear RNA isoforms (Fig. 1a). Standard circRNA analysis workflows begin by identifying these BSJ features from RNA-seq data, followed by quantification and differential expression analysis similar to gene expression studies. In addition, circRNA-specific analyses, such as differential back-splicing and alternative back-splicing analysis, can elucidate intricate changes in circRNA biogenesis. These analyses offer insights into the dynamics of competition between linear and circular splicing as well as switching between different BSJs.

CircRNA identification and quantification

The covalently closed structure of circRNAs produces a unique feature at their BSJs where the splice site order differs from that of linear isoforms (Fig. 1a). Most circRNA identification algorithms employ alignment-based strategies to detect and quantify this back-splicing feature from non-colinear alignment segments, and the abundance of circRNAs is then calculated using the number of BSJ-supporting reads (Fig. 1b). Most tools, such as acfs[20], CIRI2 (ref. 21), find_circ[2], PTESFinder[22] and UROBORUS[23] use standard aligners (for example, BWA[24], Bowtie[25], Bowtie 2 (ref. 26) and TopHat[27]) for de novo circRNA identification. Other tools, such as circRNA_finder[28], CircSplice[29], DCC[30] and CIRCexplorer2 (ref. 31) rely on chimeric-aware aligners such as STAR[32] and TopHat-Fusion[33] to detect BSJs from reported chimeric alignments. Specialized aligners, such as segmehl[34], MapSplice[35] and SPLASH2 (ref. 36), can directly identify back-splicing patterns. Subsequently, many algorithms filter results using canonical GT/AG splice signals. While this enhances accuracy and enables strandedness determination, it may exclude noncanonical circRNAs, including intronic circRNAs derived from a lariat (a looped intermediate)[37], full-length intron circles[14], transfer RNA intronic circular RNAs[38] and certain plant circRNAs[39]. Notably, the intricate nature of BSJ alignment often challenges the distinction between circRNAs and alignment artifacts, leading to substantially long processing times and compromised precision. By contrast, pseudo-reference-based approaches (for example, KNIFE[40] and NCLscan[41]) employ prebuilt candidate BSJ sequences to streamline alignment and reduce false positives. However, these methods require a well-annotated genome and cannot detect circRNAs with novel splice sites.

Combining multiple circRNA identification strategies improves detection sensitivity and quantification accuracy. For instance, pseudo-reference-based quantification approaches (CIRIquant[42] and NCLcomparator[42,43]) realign reads against both the reference genome and a pseudo-circular reference (for example, candidate circRNAs from CIRI2 (ref. 21) and NCLscan[41]) to reduce false chimeric alignment and improve quantification precision. Tools such as CirComPara2 (ref. 44) further enhance reliability by integrating results from multiple prediction tools. Systematic benchmarking studies have shown that, while most circRNA detection algorithms exhibit reliable accuracy, their sensitivity varies widely[45]. Therefore, integrating high-sensitivity tools with pseudo-reference or comparative-based filtration algorithms can offer a more balanced approach for accurate identification and quantification.

Model-based quantification methods, such as Sailfish[46] and Kallisto[47], have been widely used for rapid and accurate linear RNA quantification. These tools rely on matching short sequence fragments (*k*-mers) to estimate transcript abundance. However, efforts to adapt these strategies for circRNA quantification[48] have been limited by the high sequence similarity between circRNA and their linear counterparts. Nonetheless, these models offer valuable insights for future developments in model-based circRNA quantification tools.

Estimation of the circular-to-linear ratio

CircRNA biogenesis involves competition between back-splicing and canonical forward-splicing, which generates linear RNAs[49]. Thus, the ratio of circular-to-linear transcripts serves as a key measure of splice site utilization efficiency in circRNA formation. Various metrics have been proposed to assess this ratio (Fig. 1c). Several tools, including CIRI2 (ref. 21), CIRIquant[42] and CirComPara2 (ref. 44), calculate the back-splicing inclusion ratio by dividing the number of back-splicing reads by the sum of back-splicing and forward-splicing reads at the same splice junction. Similar to the percentage of spliced in (PSI) metric used in the analysis of messenger RNA (mRNA) alternative splicing[50], this measure reflects the relative usage of back-splicing versus forward-splicing and represents the efficiency of specific BSJs.

By contrast, tools such as CircTest[30], CIRCexplorer3-CLEAR[51] and CiLiQuant[52] calculate the BSJ ratio, which measures the abundance of BSJ reads relative to the average linear junction reads within the same gene. This approach provides insights into the overall balance between circular and linear transcripts. To ensure accuracy, ambiguous reads within BSJ regions are typically excluded, as misclassification of internal circRNA junctions as linear junctions can skew BSJ ratio calculations. However, calculation of the BSJ ratio may be less precise for large circRNAs spanning long genomic distances or for loci generating multiple overlapping circRNAs. Similarly, Sailfish-cir[48] calculates the circular read ratio by dividing the expression level of circRNAs by the sum of circular and linear RNA expression. However, limitations in current model-based quantification strategies constrain the precision of this metric. Therefore, accurately estimating the relative expression levels of circular and linear transcripts remains an ongoing challenge.

It is important to note that both metrics reflect the steady-state levels of circRNAs within a given sample, which are influenced by the dynamic regulation of circRNA and mRNA biogenesis and degradation. Due to the greater stability of circRNAs versus mRNAs, a high circular-to-linear ratio may result from circRNA accumulation rather than active biogenesis[53]. To directly estimate the biogenesis and degradation rates, experimental approaches such as metabolic RNA labeling have been employed to track nascent circRNA synthesis[54]. Nevertheless, the interpretation of these metrics may vary depending on the biological context, underscoring the need for careful consideration of their application.

Differential expression analysis

Differential circRNA expression analyses evaluate changes in both expression levels and the circular-to-linear ratio (Fig. 1d). The number of reads spanning the BSJ site is conceptionally similar to gene read counts normalized by transcript length, as in RPKM (reads per kilobase of transcript per million reads mapped) or its improved

https://doi.org/10.1038/s41588-025-02157-7

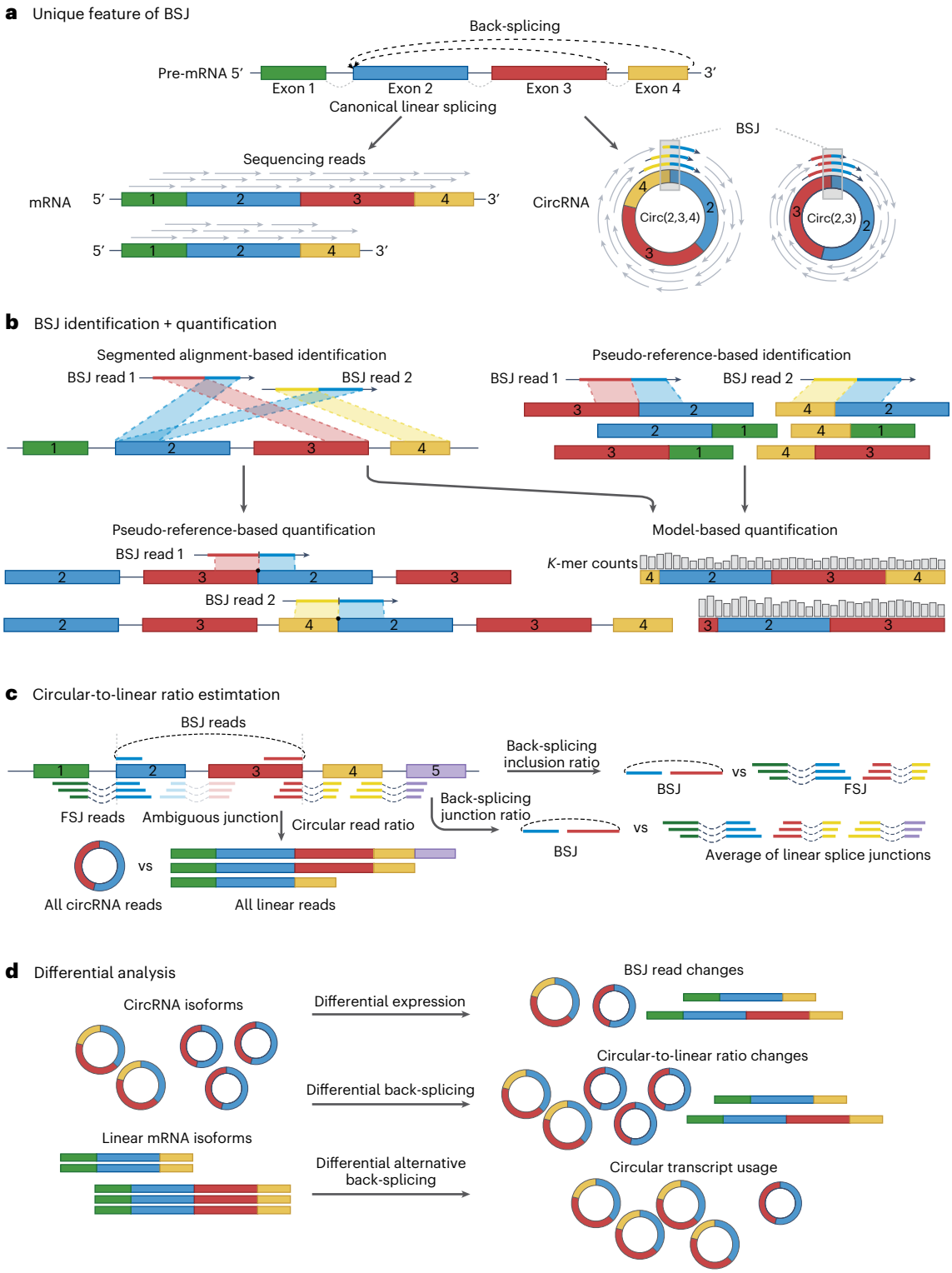

Fig. 1 | Identification, quantification and differential expression analysis of circRNAs. a, The BSJ of circRNA provides a unique non-colinear alignment feature essential for circRNA identification. **b**, BSJs are identified from segmented alignment against a reference genome or by direct alignment against a pseudo-reference that mimics BSJ sequences. For circRNA quantification, the pseudo-reference-based strategy uses the candidate circRNAs from other tools for false positive filtering based on BSJ read realignment. In addition, model-based strategies can estimate circRNA abundance through iterative *k*-mer allocation. **c**, The circular-to-linear ratio indicates the proportion of circRNAs generated from pre-mRNAs; it can be measured by the number of BSJ reads relative to the sum of BSJ and forward-splicing junction (FSJ) reads (back-splicing inclusion ratio) or to the sum of all linear splicing junction reads (BSJ ratio). Ambiguous junction reads within BSJ regions are often discarded during calculations, as they cannot be definitively assigned to circRNAs or mRNAs. This metric can also be derived from relative circRNA expression levels estimated using *k*-mer model-based quantification strategies. Shown here is a simplified transcript model to illustrate the concept, but more complex overlaps between circRNAs and linear transcripts are often observed. **d**, Changes in circRNA expression between different conditions can be measured by three methods: differential circRNA expression measured by changes in BSJ reads, differential back-splicing measured by the linear/circular ratio and differential alternative back-splicing, which reflects the shifts between different circRNAs originating from the same gene locus.

successor TPM (transcripts per million)[55]. CircRNA expression levels can therefore be estimated by dividing BSJ reads by the total number of mapped reads. However, unlike gene expression analysis, in which the null hypothesis assumes that most genes remain unchanged across conditions, circRNA expression is influenced by factors, such as circRNA accumulation, degradation or detection biases introduced by circRNA-enriched sequencing protocols. This makes normalization a critical step in circRNA differential expression analysis.

One precise normalization method involves quantitative PCR with reverse transcription (RT–qPCR)[56] or spike-in RNAs[57] to estimate normalization factors linking BSJ reads to RPKM or TPM[56]. While effective, this approach depends on the choice of RT–qPCR targets or synthetic spike-in RNA molecules, limiting its scalability for large-scale integrative studies. Alternatively, bioinformatic pipelines employ various normalization strategies. For example, normalization factors can be derived from all circRNA reads, capturing relative changes between circRNAs, but they potentially introduce bias by ignoring host gene expression. Other approaches apply canonical gene differential expression models to estimate normalization factors from mRNA expression levels, which are then used to normalize circRNA expression[42,58]. While these methods improve cross-sample comparability, they are less suitable for RNase R-treated samples due to variations in RNase R efficiency across different samples and protocols. Thus, integrating these complementary normalization approaches may better capture the complex changes in circRNA expression under different experimental conditions.

To quantify changes in the circular-to-linear ratio, CircTest[30] uses a β-binomial model to measure relative changes between linear and circular isoforms. CIRIquant[42] uses the exact rate ratio test[59] to assess significant changes in back-splicing inclusion ratios. For studies lacking biological replicates, β-distribution and generalized fold change[60] methods estimate expression and back-splicing inclusion ratio changes, providing a robust approach for preliminary experiments.

CircRNA expression analysis also requires multifaceted exploration. Differential alternative back-splicing analysis examines shifts in the usage of distinct BSJs, quantified by calculating the ratio of specific BSJ to total BSJs within the same gene. For instance, the fly *mbl* gene has been shown to express context-specific circRNA isoforms, such as circMbl(2), which dominates in fly brain cells, while alternative isoforms prevail in eye cells[61]. The switching of different circular transcripts also contributes to the regulation of MBL-C, MBL-O and MBL-P protein isoforms[61]. Together, these metrics offer valuable insights into the dynamic regulation of circRNA biogenesis, accumulation and degradation.

Reconstruction of internal structure using short- and long-read sequencing

Examining full-length circRNA sequences offers valuable insights into the biological functions of circRNAs. Recent advancements in short-read-based algorithms have enabled effective reconstruction of short circRNA isoforms, while long-read sequencing strategies have further enabled the direct reconstruction of full-length circRNAs across a broader size range.

Identification of alternative splicing from short-read RNA-seq

The study of circRNAs has revealed their intricate and unique alternative splicing patterns[31,62], highlighting that using annotated linear exons to represent circRNA structures can lead to biased and inaccurate conclusions[63]. Consequently, various methods have been developed to profile the internal structures of circRNAs using short-read RNA-seq data, categorized as either indirect or direct approaches (Fig. 2a).

Direct methods identify circRNA-specific splicing using paired-end reads spanning BSJs. Tools such as CIRI-AS[62], CircSplice[29] and FUCHS[64] detect internal splice sites through the alignment of BSJ read pairs, providing strong evidence for cryptic circular exons (cirexons)[62]. However, their resolution is constrained by RNA-seq fragment lengths, often leading to missed internal splicing events in large circRNAs[12,14]. Indirect methods compare exon coverage between circRNA-enriched samples (often treated with RNase R) and untreated samples. For example, CIRCexplorer2 (ref. 31) uses poly(A)-depleted and/or RNase R-treated and poly(A)+ RNA-seq to map circRNA and linear splicing, respectively. This circumvents fragment length limitation but still faces challenges, as circRNA-enriched samples still typically contain >90% linear reads[13,14] and RNase R inefficiency against structured 3′ ends or G-quadruplexes[65] can bias results. Additionally, indirect approaches lack direct BSJ read evidence to support internal splicing structures and require paired treated and/or untreated datasets, which are often unavailable in clinical studies.

Circular isoform reconstruction from short-read RNA-seq

Accurately determining full-length circRNA sequences is critical for predicting their biological functions[66]. Multiple methods now aim to assemble circRNA isoforms from short-read RNA-seq, enabling insights into isoform-level changes during various biological processes (Fig. 2b).

Similar to paired-end mapping approaches for identification of circRNA alternative splicing, CIRI-full[67] merges overlapping BSJ read pairs from Illumina PE250 and PE300 platforms to reconstruct circRNAs at single-molecule resolution, achieving high-fidelity assemblies of circRNAs under 500 bp. However, its ability to assemble longer circRNA isoforms is limited. circseq_cup[39] extends this by assembling all BSJ reads from the same circRNA loci using CAP3 (ref. 68); this expands the representation of circRNA isoforms but requires high circRNA coverage and remains limited by fragment length. CIRI-full also incorporates a Monte Carlo-based algorithm to estimate isoform structure and abundance using the splice graphs from CIRI-AS[62] but faces similar limitation of detection range[67].

By contrast, strategies that are not confined to BSJ reads offer broader reconstruction coverage. CIRIT[69] uses de novo transcriptome assembly (for example, IDBA-tran[70]) to identify circRNAs via head-to-tail overlap in assembled transcripts. However, most transcript assemblers may not optimize for circRNA assembly performance. Similarly, CircAST[71] and TERRACE[71] constructs splice graphs from aligned fragments within each BSJ or gene locus and apply path-finding algorithms (for example, extended minimum coverage paths, dynamic programming) to infer isoforms covering all BSJ read-supported structures. While these methods bypass circRNA size limits, they lack direct BSJ read support for certain internal structures.

Despite progress, all current short-read-based methods face trade-offs of reconstruction length and reliability, underscoring the urgent need to improve the accuracy and comprehensiveness of circRNA studies.

Full-length circRNA detection using long-read sequencing

With the advent of long-read sequencing technology[72], several methods have emerged for the direct identification of full-length circRNA structures (Fig. 2c). These methods primarily use the strand-displacement activity[73] of reverse transcriptase to perform rolling circle reverse transcription (RCRT), generating concatemers of multiple complementary DNA (cDNA) copies of a single circRNA. CIRI-long[14] and circFL-seq[13] use template switching and poly(A) tailing to enable second-strand synthesis of RCRT cDNAs. CIRI-long optimizes RNase R treatment conditions[65] to digest linear transcripts into small fragments, followed by size selection to enrich longer RCRT products over linear cDNAs. The enriched cDNAs are subsequently amplified and sequenced using the Oxford Nanopore Technology platform. By contrast, isoCirc[12] uses exonuclease treatment to remove strand-displacement overhangs, followed by ligation of cDNAs from the circRNA into a full-length circle, which is amplified using rolling circle amplification. All strategies produce long concatemer reads, enabling identification of individual

https://doi.org/10.1038/s41588-025-02157-7

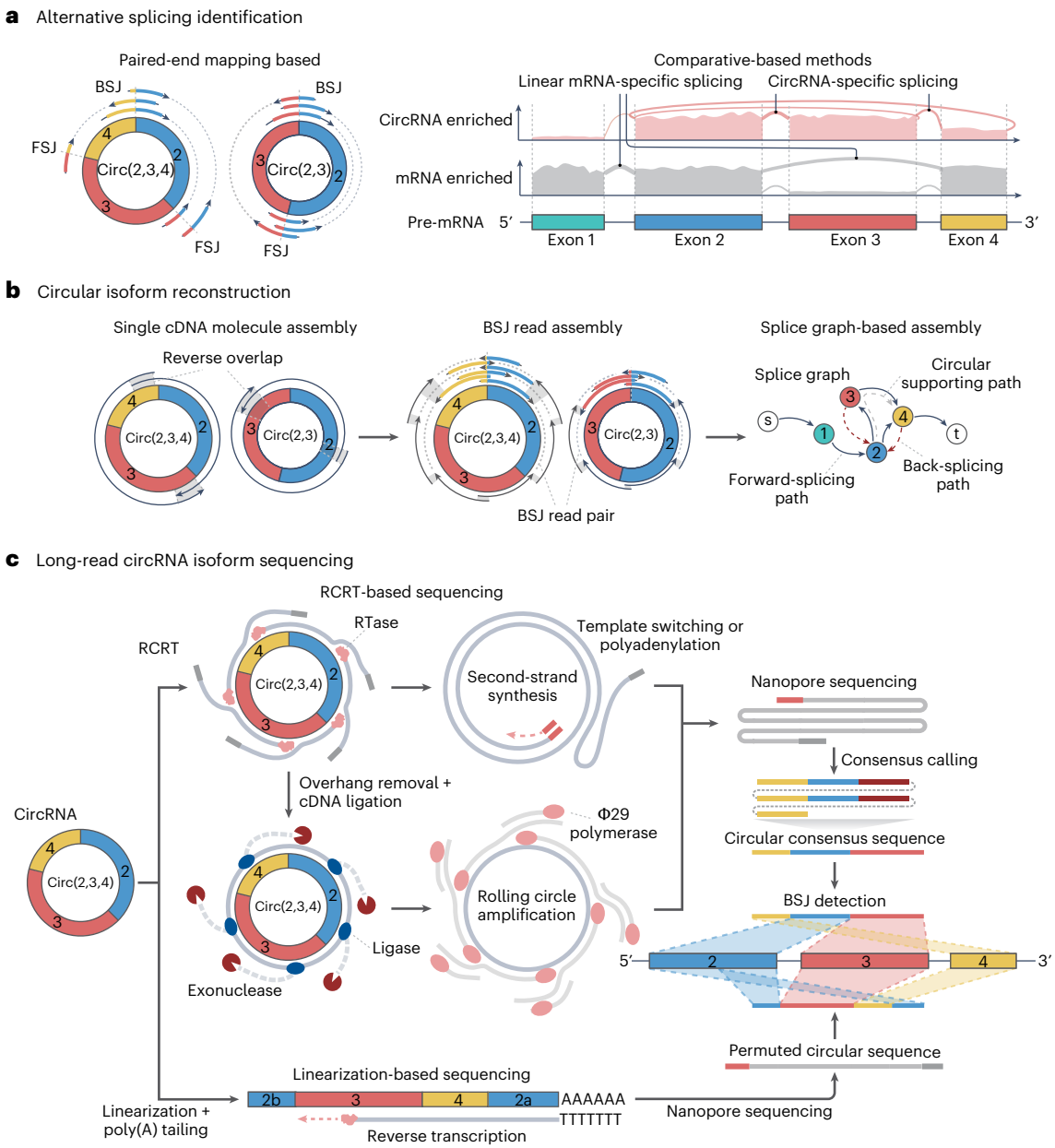

Fig. 2 | Characterization of circRNA isoform structures. a, Circular alternative splicing events are identified either using read pairs spanning BSJ sites or by comparing coverage in circRNA-enriched (red) and mRNA-enriched (gray) libraries. The top track indicates coverage of circRNA-specific splicing events that correspond to the two circRNA isoforms originating from exon (2, 3) and exon (2, 3, 4) of the example mRNA. **b**, Full-length circRNA isoforms are reconstructed by merging reverse overlapping read pairs, assembling all BSJ read pairs from the same junction site or predicting from the splice graph of the entire gene locus. The BSJ read pair assembly method provides strong evidence for the reconstructed isoform, whereas the splice graph-based approach is effective for constructing long circRNAs but lacks direct evidence of internal structures. **c**, Long-read sequencing strategies can be used for single-molecule sequencing of full-length circRNA isoforms. In linearization-based strategies, circRNAs are fragmented and polyadenylated to generate linearized RNA with poly(A) tails, which are further sequenced using the standard Oxford Nanotechnologies cDNA sequencing protocol. CircRNA sequences are identified using a split alignment strategy, similar to short-read based analysis. Alternatively, RCRT-based strategies employ template switching or polyadenylation to capture RCRT products or use exonuclease and single-stranded DNA ligase to generate circular cDNAs that replicate the circRNA sequences. These ligated products undergo rolling circle amplification with Φ29 polymerase for library construction and nanopore sequencing. Sequencing reads from RCRT-based strategies consist of concatemers of multiple copies of full-length circRNA sequences. The full-length circRNA sequences are then identified through consensus calling and downstream BSJ detection. RTase, reverse transcriptase.

copies of full-length circRNAs. Consensus sequences are calculated using trf[74] or partial order alignment algorithms[75,76] and aligned to the reference genome to identify BSJs and full-length isoforms. Apart from RCRT-based strategies, circNick-LRS[77] combines different fragmentation conditions to linearize circRNAs, followed by polyadenylation and nanopore sequencing. Sequenced molecules exhibit permuted circular sequences similar to those in RCRT methods but risk omitting internal circRNA structures if linearization occurs at multiple sites in one circRNA.

These long-read-based strategies resolve length limitations, enabling accurate full-length isoform reconstruction and novel circRNA discovery. However, circRNA length distributions differ between the methods that are based on RCRT (~500 nucleotides) or linearization (~800 nucleotides), likely due to biases in RCRT and fragmentation

https://doi.org/10.1038/s41588-025-02157-7

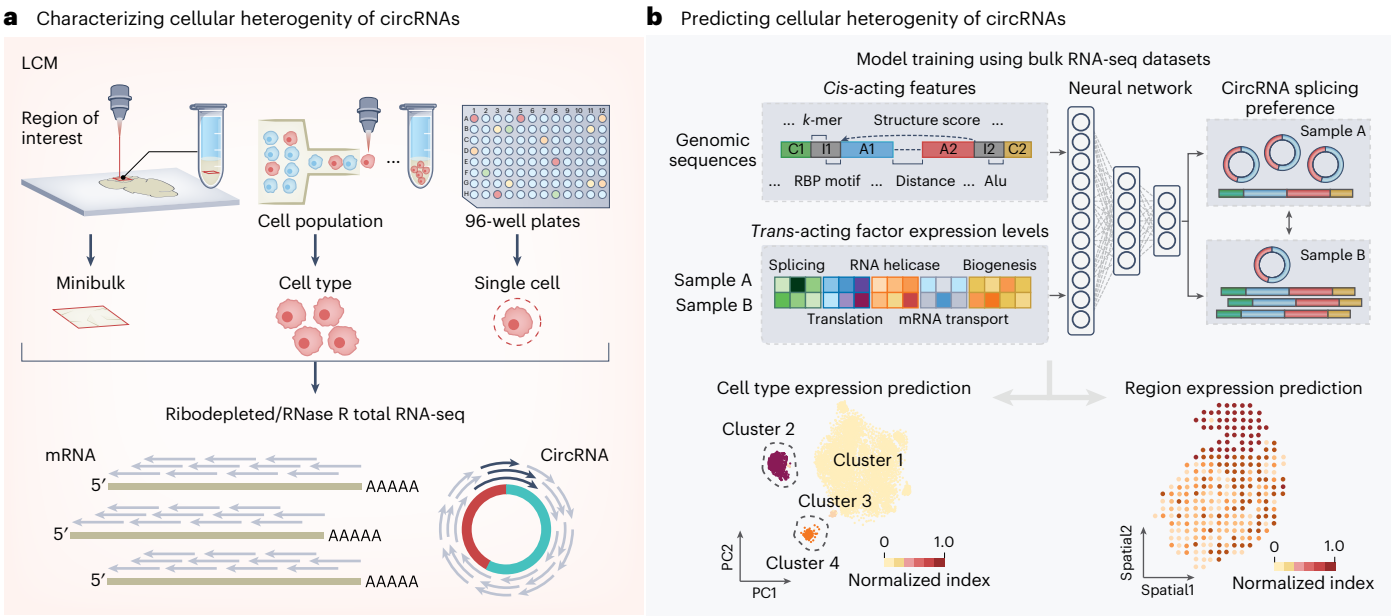

Fig. 3 | Profiling the cellular heterogeneity of circRNAs. a, Single-cell circRNA sequencing strategies use LCM, flow cytometry sorting or single-cell separation techniques to isolate minibulk samples, target cell populations or individual cells for RNA extraction. The extracted RNA is then sequenced using ribosomal RNA-depleted total RNA-seq with an optional RNase R treatment to enrich circRNAs. **b**, Deep learning-based models are used to predict the outcomes of circRNA splicing in specific cell types or spatial domains. These models incorporate *cis*-acting genomic features and the expression level of *trans*-acting factors across different samples. A deep neural network is trained to determine the circRNA splicing preference for each circRNA in a pair of samples. This trained model can then predict cell type-specific circRNA expression patterns or estimate region-specific circRNA expression from spatial transcriptomic data. Color bars represent the normalized circRNA index, which indicates the relative splicing preference of circRNAs across cell types or regions. PC, principal component.

efficiency toward circRNAs of different lengths. Notably, recent advances using group II intron reverse transcriptase have generated RCRT products >10 kb, revealing that RNase R-treated samples may exhibit biased circRNA distributions due to nonspecific nicking during RNase R digestion[78,79]. These findings suggest that the true circRNA length spectrum requires further experimental validation. Additionally, variations in RCRT efficiency across circRNAs of different lengths may also impact quantitative analysis[80], necessitating further evaluation of quantification outcomes.

Despite the higher cost associated with nanopore sequencing, which has restricted the application of these circRNA sequencing strategies in clinical research, ongoing efficiency improvements may mitigate these limitations[78,81,82]. With continued enhancements, long-read circRNA sequencing could become the standard approach in circRNA studies, offering accurate and comprehensive insights into circRNA structures and functions.

Profiling circRNA cellular and spatial heterogeneity

CircRNAs exhibit high tissue and cell type specificity[56,83–86], making bulk analyses prone to bias arising from variations in cell type composition[87,88]. For example, the ciRS-7 (CDR1as), initially proposed as an oncogene due to its miR-7 sponge activity and overexpression in tumors[89], was later found to originate predominantly from stromal cells rather than cancer cells in colon cancer[88]. Similarly, the correlation between circRNA and mRNA expression levels is more reflective of varying cellular compositions than the competitive endogenous RNA roles traditionally attributed to circRNAs[87]. These findings emphasize the urgent need for single-cell-resolution methods to dissect circRNA heterogeneity. However, current single-cell sequencing platforms such as the widely used 10x Chromium system, which primarily capture 3′ or 5′ sequences of linear transcripts[90], are inadequate for detecting circRNAs (Fig. 3a). As a result, profiling circRNAs at single-cell resolution remains a formidable challenge, and systematic characterization of cellular expression patterns in this context has yet to be achieved.

Characterization of cellular heterogeneity of circRNAs with single-cell sequencing

To explore the cellular heterogeneity of circRNAs, researchers employ laser capture microdissection (LCM) to dissect regions of interest and sequence circRNAs in minibulk samples of target cell types[91]. While LCM improves cell type resolution over bulk-level analyses, it is labor intensive and relies on the purity of the isolated cell population. Flow cytometry cell sorting offers higher-throughput sequencing of circRNAs in specific cell types[92]. However, both LCM and flow cytometry approaches rely on prior knowledge of the cell types under investigation, limiting their utility for discovering novel cell types.

Single-cell sequencing approaches such as SUPeR-seq[93] combine single-cell sorting with random primer reverse transcription; these enable circRNA characterization in mouse and human embryos and have revealed stage-specific circRNA dynamics during embryonic development[93,94]. Other random reverse transcription-based scRNA-seq strategies, including SMARTer single-cell total RNA-seq[95], MATQ-seq[96], VASA-seq[97] and snRandom-seq[98], have also been able to capture these circular transcripts. At the same time, polyadenylation-based protocols, such as Smart-seq-total[99], can also detect degraded circRNA fragments containing BSJ sequences.

While other single-cell full-length scRNA-seq methods primarily rely on poly(A) selection, which inadequately captures circRNAs that lack poly(A) tails, a considerable number of circRNAs have been detected in these poly(A)-enriched datasets[61,100]. For example, internal poly(A) tracts in circRNAs enable oligo(dT) primer binding, which facilitated the detection of its cell type-specific expression of different circRNAs from the *mbl* locus in the fly brain and eye using poly(A)-enriched scRNA-seq datasets[61]. Moreover, circSC[15] aggregates data from 171 full-length scRNA-seq studies to map circRNAs in human and mouse cells, revealing highly cell type-specific expression patterns

in brain samples, developing embryos and breast tumors. However, the lack of circRNA enrichment and low sequencing depth limit detection to a few circRNAs per cell, where many circRNAs are supported by one or two BSJ reads, increasing the risk of false positives. These limitations highlight the urgent need for high-throughput single-cell circRNA sequencing strategies.

Prediction of circRNA cellular heterogeneity with deep learning models

Deep learning algorithms have revolutionized circRNA detection, enabling analysis at single-cell or spatial resolutions. CircRNA biogenesis is intricately regulated by factors, such as the spliceosome[17], RBPs[19,49] and flanking intronic complementary sequences[101,102]. This regulatory complexity implies that circRNA expression may be predictable based on these *cis*-acting sequence features and *trans*-acting regulator expression levels.

On this basis, CIRI-deep introduces a deep neural network to predict changes in the back-splicing inclusion ratio between paired samples[16] (Fig. 3b). Trained on 25 million circRNA splicing events from bulk RNA-seq samples, this model is adapted to predict circRNA splicing preferences in single-cell and spatial transcriptomic data and incorporates an adapted integrated gradient strategy to assess the contribution of various *cis* and *trans* regulatory features, enhancing exploration of circRNA regulation across different sequencing methodologies. However, the model's training on samples from normal tissues limits its utility in tumors or other disease contexts. Bulk RNA-seq training sets also conflate RBP–circRNA regulatory relationships with cell type composition heterogeneity and technical batch effects, which can lead to false positive predictions.

Recent spliceosome perturbation studies[103] highlight opportunities to refine models with RBP knockdown datasets[104] or genome-wide CRISPR screening data[105] to better model the regulatory mechanisms underlying circRNA back-splicing. Such approaches would provide more robust evidence for predicting circRNA expression across diverse biological contexts.

Functional characterization of circRNAs

The regulatory functions of circRNAs can be explored through standard differential expression and splicing analyses across various experimental conditions. In large cohorts, correlating circRNA levels with gene and/or miRNA expression can reveal potential regulatory networks such as circRNA–miRNA–mRNA axes. For large-scale analysis, deep learning models integrate experimentally validated circRNA–disease associations to predict novel associations.

Canonical expression-based analysis

Standard differential expression and splicing analysis measures circRNA changes across different conditions, such as disease states or experimental treatments (Fig. 4a). The regulatory functions of differentially expressed circRNAs are then annotated using public databases[106,107] or predicted de novo based on sequence features, including RBP-binding motifs[108] and miRNA-responsive elements[109]. While many studies use host gene functions of circRNAs for Gene Ontology and Kyoto Encyclopedia of Genes and Genomes enrichment analysis, this approach can be misleading, as circRNA expression often diverges from that of their host genes.

Recent efforts instead prioritize functional circRNAs using network-based algorithms. Such mRNA–circRNA coexpression networks are often constructed from large-scale circRNA and mRNA expression profiles to reveal functional relationships (Fig. 4b). Strong positive correlations between circRNA and mRNA expression suggest coexpression or cofunction, whereas negative correlations indicate potential negative regulation. To prioritize disease-related circRNAs, a random walk algorithm could be employed to quantify the proximity of candidate circRNAs to known disease-related genes[83]. Furthermore, the conservation of genomic sequences could further be combined to provide an effective strategy for ranking disease-associated circRNAs.

Deep learning architectures for functional circRNA prediction

Recent advances in deep learning models have spurred the development of numerous deep learning-based algorithms for predicting disease-related circRNAs[110]. These models typically leverage the circRNA–disease network (Fig. 4c). First, they gather experimentally validated circRNA–disease associations from curated circRNA databases[111]. Relatedness between circRNAs and diseases is then measured using different approaches: circRNA–circRNA similarity (via sequence matching), disease–disease similarity (via semantic approaches, such as shared clinical features and molecular mechanisms) and circRNA–disease interaction similarity (via entropy, topology, functional characteristics[112] and Gaussian interaction profile kernels[113]). To predict novel associations, these algorithms employ deep learning models, such as convolution neural networks[114], autoencoders[115] and graph neural networks[116,117]. Notably, recent models, such as CLCDA[118] and CircDA[116], incorporate RBP-binding sites and miRNA-responsive elements to provide additional insight into circRNA regulatory mechanisms.

However, many of these models remain largely proof of concept, relying on interaction profile similarities or data from RBP and miRNA interaction databases, which restricts their ability to identify novel disease-related circRNAs independently. Consequently, there is a pressing need for more generalized models that are capable of systematically prioritizing newly detected circRNAs. In addition, efforts should focus on developing interpretable and versatile deep learning models that integrate gene expression profiles, rather than relying solely on semantic disease–disease similarities, to enhance the interpretability of disease–disease similarities and improve predictions of target genes.

Although functional annotation algorithms for circRNAs have advanced substantially, not all circRNAs are functionally relevant. Recent studies suggest that most circRNAs may arise as nonadaptive by-products of eukaryotic splicing rather than as functional entities[119]. Therefore, incorporating changes in host gene expression or shifts in cell type composition during prioritization could improve the identification of functional circRNAs while distinguishing them from splicing by-products. Notably, the suggested role of circRNAs in sequestering miRNAs and RBPs[120] may hold true for circRNAs with a high density of binding sites[1,2,49] but should not be considered as a universal property of all circRNAs, especially when the number of miRNA- and/or RBP-binding sites per circRNA is low[121]. Thus, prediction of any miRNA and RBP regulatory axis should only be interpreted as a guide to prioritize candidates for experimental validation. Overall, advancing the characterization of functional circRNAs requires refining algorithms to integrate multidimensional evidence to distinguish functional and nonfunctional circRNA. These computational efforts must also be coupled with systematic experimental validation to confirm the true regulatory role of candidate circRNAs.

Challenges in large-scale integration of circRNA sequencing data

Extensive studies over the last decade have generated a vast amount of circRNA sequencing datasets, offering a valuable resource for investigating circRNA biogenesis and biological functions (Supplementary Table 1). However, integrating these large datasets presents several challenges related to sample preparation, library construction and data analysis pipelines (Fig. 5a).

RNA extraction and preparation

RNA quality (measured by RNA integrity number (RIN)) substantially impacts circRNA detection. Low RIN values are associated with fewer detected circRNAs, as degraded circRNAs are susceptible to RNase R digestion during circRNA enrichment[18,67], reducing the number of BSJ reads required for circRNA detection algorithms.

https://doi.org/10.1038/s41588-025-02157-7

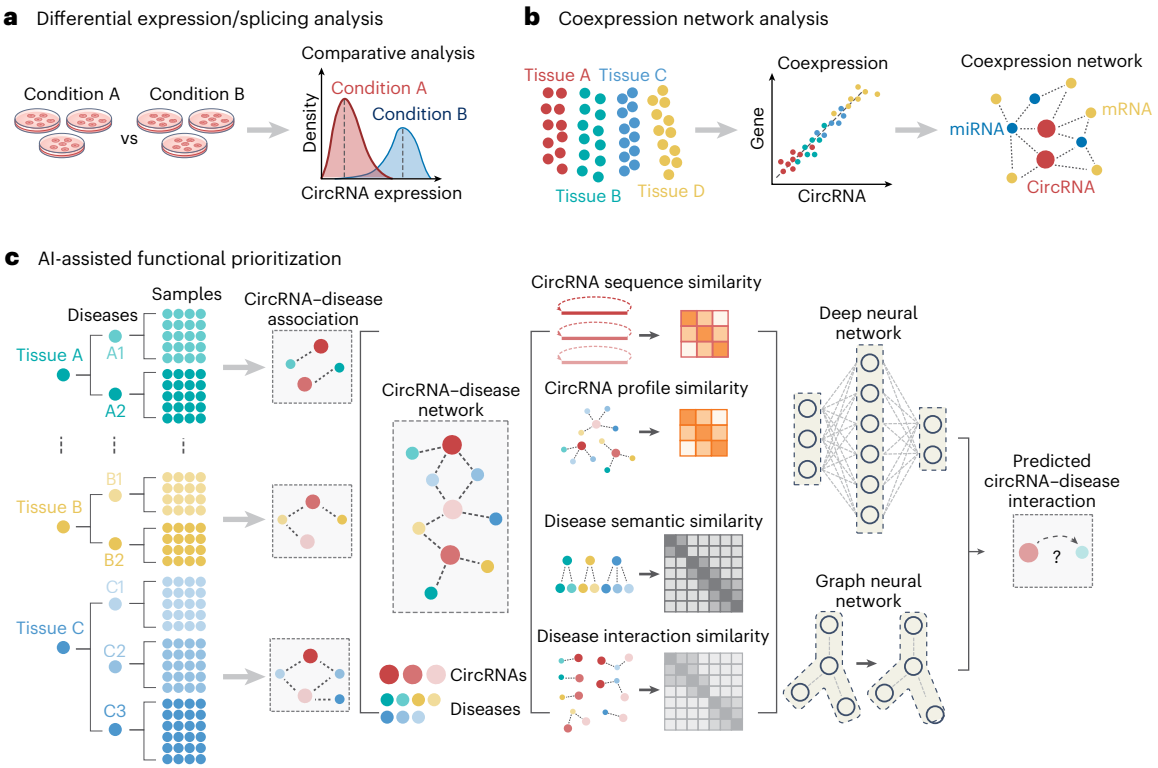

Fig. 4 | Functional characterization of circRNAs. a, Analysis of differential expression and splicing can evaluate changes in circRNA expression and splicing under different conditions. **b**, In a cohort-level analysis, the circRNA regulatory network is inferred by coexpression analysis of circRNA, miRNA and mRNAs. This network helps to reveal potential circRNA–miRNA–mRNA regulatory axes. **c**, For large-scale characterization of functional circRNAs, any circRNA–disease associations are calculated for individual diseases and integrated to construct a circRNA–disease association network. Parameters, such as circRNA sequence similarity, degree of disease semantic similarity (for example, shared clinical features and molecular mechanisms) and interaction profile similarity between circRNAs and diseases, are calculated. Deep learning-based models, including deep neural networks or graph neural networks, are then used to model these associations. These trained models facilitate the prediction of novel circRNA–disease interactions.

The choice of enrichment strategy is another critical factor affecting the efficiency of circRNA detection. Most studies use ribosomal RNA-depleted total RNA-seq for unbiased circRNA quantification, but this approach yields a low proportion of back-splicing reads. By contrast, RNase R-treated RNA-seq improves circRNA detection sensitivity but introduces challenges such as variable digestion efficiency due to RNA secondary structure[65,122], high variability between experimental replicates[42] and nonspecific circRNA nicking during RNase R treatment[78,123]. Improved RNase R protocols employ polyadenylation-based linear RNA removal or G-quadruplex unfolding buffers to optimize linear RNA elimination. Optimizing RNase R concentration and treatment duration further enhances circRNA detection[124], underscoring the need for standardized protocols to ensure reproducible experimental results. Some studies, such as circSC[15] and CircRiC[100], integrate poly(A)-enriched datasets for circRNA analysis. Platforms, such as MiOncoCirc[125] and the Human Biofluid RNA Atlas[57], use exome capture RNA-seq to profile the circRNA transcriptome in cancer and biofluids; this enhances circRNA enrichment while preserving accurate estimation of circular-to-linear ratios.

The variability in circRNA-detecting protocols introduces strong batch effects, which poses substantial challenges for large-scale integrative analysis. Tools such as CIRIquant[42] have implemented a Gaussian mixture model to characterize enrichment efficiency and perform circRNA expression-level correction using paired RNase R-treated and untreated samples. A similar framework could be expanded to characterize and mitigate batch effects between total RNA and circRNA-enriched samples from similar tissue sources, thereby enhancing the robustness of integrative circRNA analyses.

Sequencing technologies and depth

Illumina and nanopore circRNA sequencing strategies each offer distinct trade-offs in circRNA detection. Despite its lower efficiency, Illumina sequencing is favored in most circRNA studies due to its cost-effectiveness. By contrast, long-read sequencing improves circRNA detection by capturing full-length isoforms. Oxford Nanopore Technology's high sequencing speed facilitates rapid experiments within days[126], yet biases in rolling circle amplification and RCRT processes may impact quantitative analysis[80]. The FL-circAS[127] database contains full-length circRNA isoforms detected from several long-read sequencing studies[12–14,77], while circAtlas version 3.0 (ref. 66) integrates both short-read and long-read sequencing datasets[81,128]. Both resources provide functional annotation, but integration of expression data from short-read and long-read platforms remains challenging. While long-read sequencing costs remain higher than those of current short-read sequencing platforms, ongoing improvements in detection efficiency and decreasing cost of the Oxford Nanopore and PacBio sequencing systems suggest that it may soon become standard for circRNA analysis.

Sequencing depth substantially impacts circRNA detection. In short-read data, only <1% of reads typically represent BSJs, necessitating large datasets for comprehensive circRNA detection. Long-read sequencing captures more circRNAs per gigabase of data, yet both Oxford Nanopore and PacBio technologies yield limited molecules per flow cell, making it difficult to achieve saturated detection of all expressed circRNAs[14]. Empirically, approximately 12 GB of total RNA-seq or 3–4 GB of RNase R-treated RNA-seq can detect around 10,000 circRNAs in brain samples, whereas 5 GB of long-read sequencing data can detect >50,000 brain circRNAs[13,14]. Variations in circRNA

https://doi.org/10.1038/s41588-025-02157-7

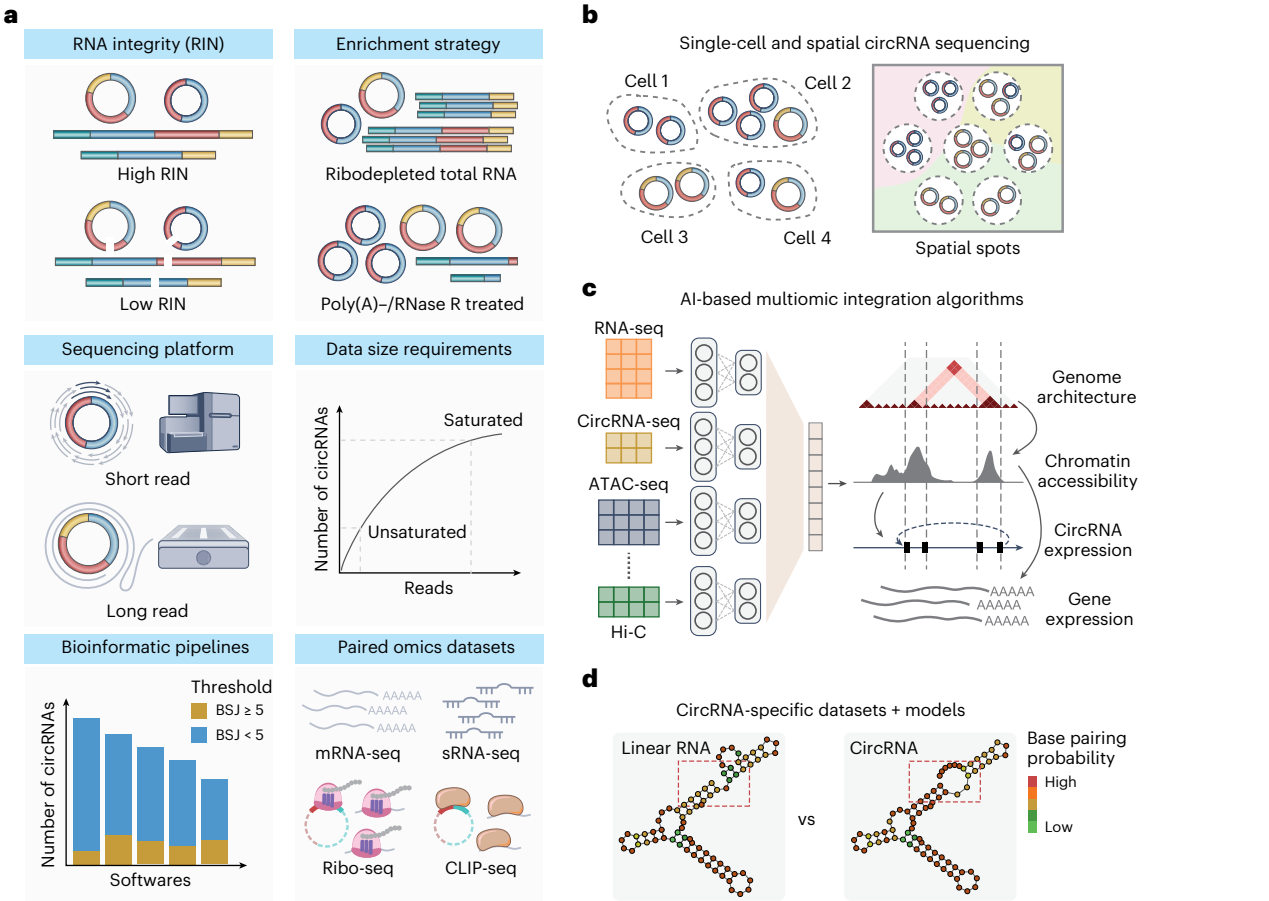

Fig. 5 | Challenges and opportunities in large-scale analysis of circRNA data. **a**, Challenges in integrating large-scale circRNA sequencing data. First, RIN is critical, as degraded circRNAs are lost during downstream processing, impeding the detection of BSJ reads. Diverse circRNA-enrichment strategies also lead to variations in detection efficiency and give rise to substantial batch effects, affecting the accuracy of circRNA quantification. Furthermore, short-read sequencing typically provides results at the BSJ level, while long-read methods allow for more complicated isoform-level quantification. Sequencing depth affects the saturation of circRNA detection, and variation in the sensitivity and filter threshold of different bioinformatic pipelines also increases the difficulty of integration. In addition, although paired sequencing datasets can provide evidence to infer circRNA regulation and functions, heterogeneity between different studies poses further challenges in the integration process. CLIP–seq, cross-linking immunoprecipitation followed by high-throughput sequencing. sRNA-seq, small RNA sequencing. **b**, Development of single-cell and spatial circRNA sequencing technologies provides an opportunity to characterize circRNA expression patterns with improved resolution. Here, optimized circRNA sequencing strategies are essential to detect circRNAs from limited material from single cells and spatial spots. **c**, Development of AI-based multiomic algorithms could integrate omic-specific features, providing critical insights into how genome architecture, chromatin accessibility and other epigenetic features affect circRNA biogenesis. ATAC–seq, assay for transposase-accessible chromatin with sequencing. **d**, A circRNA and its cognate linear RNA can exhibit distinct structures and functionalities. Therefore, the development of circRNA-specific AI models requires the accumulation of circRNA training sets and the development of tailored algorithms to incorporate circRNA-specific features, such as exon scrambling patterns and unique BSJ sequences.

expression across tissues further complicate data size estimation. For example, circRNAs are highly abundant in tissues with elevated back-splicing activity (for example, brain, spinal cord, testis and heart[83,84]) or in those with high circRNA accumulation (for example, biofluids such as blood plasma[57,129,130]), which require less sequencing depth. By contrast, circRNA expression is generally low in other tissues, necessitating a substantially greater sequencing depth[131]. Importantly, many circRNAs are rare and may not be functionally relevant[119]; so pursuing the detection of low-abundance circRNAs should not be the highest priority.

CircRNA identification and multiomic integration algorithms

All circRNA analysis algorithms rely on identifying BSJ sites and employ various strategies to filter low-confidence circRNAs. Filtering based on the canonical GU/AG splice signal improves identification accuracy but may exclude novel yet unidentified circRNAs that arise from noncanonical back-splicing. In addition, tools vary in their default stringency for minimum supporting reads[45]. While most tools do not filter based on BSJ counts or offer options to report all circRNAs, CirComPara2 (ref. 44), circtools[132] and KNIFE[40] require a minimum of two supporting reads, whereas circRNA_finder[28] and segemehl[34] require a more stringent filter of five supporting reads. These discrepancies can substantially affect the sensitivity of circRNA detection. Furthermore, most tools require sophisticated processing of alignment results, leading to long computation times and high demands on resources. Therefore, highly efficient circRNA detection and quantification algorithms, such as the *k*-mer counting-based SPLASH2 tool[36], are more suitable for scalable analysis of large datasets. However, advancements in this area remain limited and require further optimization.

The high sequence similarity between nuclear mitochondrial pseudogenes and the mitochondrial genome presents specific challenges for identifying mitochondrial-derived circRNAs. Initially, circRNAs detected from the mitochondrial genome were often excluded as artifacts, but recent studies have identified genuine mitochondrial-derived circRNAs with regulatory roles[4,14,133]. However, the biogenesis mechanism of mitochondrial circRNA remains

unclear, and the applicability of the canonical GU/AG splice site filter requires further investigation. Therefore, special efforts should also be made to improve algorithms for accurate identification of these mitochondrial circRNAs.

Integrating multiomics datasets could enhance the prioritization of functional circRNAs. Databases such as POSTAR2 (ref. 106) and starBase[134] integrate cross-linking immunoprecipitation followed by high-throughput sequencing data to predict RBP- and miRNA-binding sites, while other studies include mass spectrometry[135–137], ribosome and/or polysome profiling[138–141] and m^{6}A or methylated RNA immunoprecipitation sequencing[142] for assessing circRNA coding potential. However, mass spectrometry analysis relies on a stringent estimation of false positive rates and may yield false discoveries if not carefully controlled[143]. Integrating evidence from diverse omic approaches could provide more robust functional predictions[144]. Moreover, these datasets often originate from independent studies, introducing sample heterogeneity that may bias predictions. Additionally, the lack of large-scale multiomic circRNA studies hinders the development of circRNA-specific integration algorithms.

Future development of circRNA sequencing techniques and AI-based algorithms

Current strategies for circRNA sequencing have substantially improved the sensitivity and accuracy of circRNA identification and reconstruction. However, neither of these strategies is able to achieve highly efficient circRNA detection when the starting materials are limited, posing challenges for further characterization of circRNAs at single-cell and spatial resolution. Additionally, existing single-cell and spatial barcoding methods primarily capture mRNA poly(A) tails[145], which do not align well with current circRNA sequencing approaches. Therefore, advancing high-throughput techniques for single-cell and spatial RNA-seq is crucial for understanding circRNA functions across diverse microenvironments and biological processes (Fig. 5b). Promising solutions include integrating efficient long-read circRNA sequencing with single-cell barcoding technologies to achieve full-length circRNA profiling at cellular resolution and developing deep learning-based imputation algorithms to enable a comprehensive understanding of the cellular circRNA expression landscape[146].

Numerous deep learning-based computational methods have emerged for integrating single-cell and multiomics data. These models facilitate accurate integration of diverse omic modalities based on cell identities[147] or spatial information[147] and hold promise for regulatory inference[148]. However, the lack of paired circRNA sequencing datasets has impeded the widespread integration of multiomics circRNA analysis. Because many circRNAs are by-products of host gene expression, future efforts could incorporate their coexpression as anchors to align circRNA expression with gene expression and chromatin accessibility datasets. Moreover, the development of single-cell and spatial circRNA sequencing strategies would also enable adaptation of these integrative models for circRNA research. The availability of multiomic resources and deep learning models thus could expedite exploration of circRNA biogenesis and regulatory mechanisms (Fig. 5c).

Simultaneously, AI-driven RNA language models have demonstrated promising capabilities in predicting RNA structures and functions. For example, an mRNA 5′ untranslated region language model accurately predicts ribosome loading, translation efficiency and mRNA expression levels[149]. Additionally, generative diffusion models have successfully designed novel proteins de novo[150]. Therefore, the development of circRNA-specific language models could facilitate rational design of circRNA-based vaccines and therapeutics with desired properties[151]. However, given the distinct structures and biological functions of circRNAs compared to mRNAs, most current mRNA models are not directly applicable to circRNA analysis. Thus, establishing circRNA-specific AI models requires the accumulation of circRNA-specific training data. However, the limited number of experimentally validated circRNA structures and functions presents a challenge to progress in this field. Although efforts have been made to profile the functionality of circRNAs, many studies even relied on bioinformatic prediction to perform this task, the reliability of these predictive results remains unexamined and only circRNAs validated by sophisticated means should be taken as the basis for subsequent works. In addition, the lack of free ends in circRNA poses specific computational challenges[152], necessitating models that can accurately incorporate circRNA-specific features to capture the unique characteristics of these molecules (Fig. 5d).

Conclusions

Research on circRNAs has underscored their potential as therapeutic agents and RNA drug platforms. Advances in long-read circRNA sequencing have overcome challenges posed by high sequence similarity between circRNAs and linear transcripts, enabling in-depth profiling of circRNA diversity. The field now prioritizes developing single-cell and spatial methods to map cellular expression patterns and spatiotemporal regulation of circRNAs and to enable robust statistical analysis across cells and spatial spots. Combining these approaches with AI-powered multiomic algorithms holds promise for cluster- and spatial domain-guided integration of transcriptomic and epigenetic data. This integration could connect circRNA expression to gene activity and epigenetic modifications, providing insights into the biogenesis mechanisms and regulatory functions of endogenous circRNAs.

However, the relatively low abundance of circRNA presents challenges in detecting these molecules from limited material in single-cell or spatial contexts. While long-read sequencing strategies have improved the sensitivity and accuracy of circRNA characterization, their effectiveness is still limited. Optimizing circRNA sequencing techniques, such as enhancing reverse transcriptase for RCRT, is crucial for refining methods suitable for single-cell and spatial characterization.

Additionally, recent advances in AI models have demonstrated remarkable accuracy in protein and RNA modeling[64,149], setting the stage for AI-based functional prediction and rational design of circRNAs. However, it should be noted that the biogenesis and regulation mechanisms of circRNAs are distinct from those of mRNA transcripts, necessitating the development of circRNA-specific AI models. In addition, canonical deep learning RNA models often overlook circularization constraints and may not be directly applicable to circRNA modeling. Consequently, developing AI algorithms tailored for effectively mining the vast circRNA datasets accumulated thus far is imperative. Current circRNA studies provide rich resources but also pose challenges in integrating data from diverse sources, sequencing protocols and analysis pipelines, a task in which AI algorithms excel. We believe that the establishment of circRNA-specific AI models holds immense potential to advance circRNA characterization and applications in the near future.

References

1. Hansen, T. B. et al. Natural RNA circles function as efficient microRNA sponges. *Nature* **495**, 384–388 (2013).
2. Memczak, S. et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338 (2013).
3. Abdelmohsen, K. et al. Identification of HuR target circular RNAs uncovers suppression of PABPN1 translation by CircPABPN1. *RNA Biol.* **14**, 361–369 (2017).
4. Zhao, Q. et al. Targeting mitochondria-located circRNA SCAR alleviates NASH via reducing mROS output. *Cell* **183**, 76–93 (2020).
5. Huang, D. et al. Tumour circular RNAs elicit anti-tumour immunity by encoding cryptic peptides. *Nature* **625**, 593–602 (2024).
6. Liu, C. X. et al. Structure and degradation of circular RNAs regulate PKR activation in innate immunity. *Cell* **177**, 865–880 (2019).

7. Qu, L. et al. Circular RNA vaccines against SARS-CoV-2 and emerging variants. *Cell* **185**, 1728–1744 (2022).
8. Liang, R. et al. Prime editing using CRISPR–Cas12a and circular RNAs in human cells. *Nat. Biotechnol.* **42**, 1867–1875 (2024).
9. Yi, Z. et al. Engineered circular ADAR-recruiting RNAs increase the efficiency and fidelity of RNA editing in vitro and in vivo. *Nat. Biotechnol.* **40**, 946–955 (2022).
10. Guo, S. K. et al. Therapeutic application of circular RNA aptamers in a mouse model of psoriasis. *Nat. Biotechnol.* **43**, 236–246 (2025).
11. Feng, Z. et al. An in vitro-transcribed circular RNA targets the mitochondrial inner membrane cardiolipin to ablate EIF4G2^{+}PTBP1^{+} pan-adenocarcinoma. *Nat. Cancer* **5**, 30–46 (2024).
12. Xin, R. et al. isoCirc catalogs full-length circular RNA isoforms in human transcriptomes. *Nat. Commun.* **12**, 266 (2021).
13. Liu, Z. et al. circFL-seq reveals full-length circular RNAs with rolling circular reverse transcription and nanopore sequencing. *eLife* **10**, e69457 (2021).
14. Zhang, J. et al. Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long. *Nat. Biotechnol.* **39**, 836–845 (2021).
15. Wu, W., Zhang, J., Cao, X., Cai, Z. & Zhao, F. Exploring the cellular landscape of circular RNAs using full-length single-cell RNA sequencing. *Nat. Commun.* **13**, 3242 (2022).
16. Zhou, Z. et al. CIRI-deep enables single-cell and spatial transcriptomic analysis of circular RNAs with deep learning. *Adv. Sci.* **11**, e2308115 (2024).
17. Li, X. et al. A unified mechanism for intron and exon definition and back-splicing. *Nature* **573**, 375–380 (2019).
18. Jeck, W. R. et al. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**, 141–157 (2013).
19. Conn, S. J. et al. The RNA binding protein Quaking regulates formation of circRNAs. *Cell* **160**, 1125–1134 (2015).
20. You, X. & Conrad, T. O. Acfs: accurate circRNA identification and quantification from RNA-seq data. *Sci. Rep.* **6**, 38820 (2016).
21. Gao, Y., Zhang, J. & Zhao, F. Circular RNA identification based on multiple seed matching. *Brief. Bioinform.* **19**, 803–810 (2018).
22. Izuogu, O. G. et al. PTESFinder: a computational method to identify post-transcriptional exon shuffling (PTES) events. *BMC Bioinformatics* **17**, 31 (2016).
23. Song, X. et al. Circular RNA profile in gliomas revealed by identification tool UROBORUS. *Nucleic Acids Res.* **44**, e87 (2016).
24. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).
25. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
26. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
27. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
28. Westholm, J. O. et al. Genome-wide analysis of *Drosophila* circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep.* **9**, 1966–1980 (2014).
29. Feng, J. et al. Genome-wide identification of cancer-specific alternative splicing in circRNA. *Mol. Cancer* **18**, 35 (2019).
30. Cheng, J., Metge, F. & Dieterich, C. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* **32**, 1094–1096 (2016).
31. Zhang, X. O. et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.* **26**, 1277–1287 (2016).
32. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
33. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12**, R72 (2011).
34. Hoffmann, S. et al. A multi-split mapping algorithm for circular RNA, splicing, *trans*-splicing and fusion detection. *Genome Biol.* **15**, R34 (2014).
35. Wang, K. et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, e178 (2010).
36. Kokot, M., Dehghannasiri, R., Baharav, T., Salzman, J. & Deorowicz, S. Scalable and unsupervised discovery from raw sequencing reads using SPLASH2. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-024-02381-2 (2024).
37. Talhouarne, G. J. S. & Gall, J. G. Lariat intronic RNAs in the cytoplasm of vertebrate cells. *Proc. Natl Acad. Sci. USA* **115**, E7970–E7977 (2018).
38. Lu, Z. et al. Metazoan tRNA introns generate stable circular RNAs in vivo. *RNA* **21**, 1554–1565 (2015).
39. Ye, C. Y. et al. Full-length sequence assembly reveals circular RNAs with diverse non-GT/AG splicing signals in rice. *RNA Biol.* **14**, 1055–1063 (2017).
40. Szabo, L. et al. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol.* **16**, 126 (2015).
41. Chuang, T. J. et al. NCLscan: accurate identification of non-co-linear transcripts (fusion, *trans*-splicing and circular RNA) with a good balance between sensitivity and precision. *Nucleic Acids Res.* **44**, e29 (2016).
42. Zhang, J., Chen, S., Yang, J. & Zhao, F. Accurate quantification of circular RNAs identifies extensive circular isoform switching events. *Nat. Commun.* **11**, 90 (2020).
43. Chen, C. Y. & Chuang, T. J. NCLcomparator: systematically post-screening non-co-linear transcripts (circular, *trans*-spliced, or fusion RNAs) identified from various detectors. *BMC Bioinformatics* **20**, 3 (2019).
44. Gaffo, E., Buratin, A., Dal Molin, A. & Bortoluzzi, S. Sensitive, reliable and robust circRNA detection from RNA-seq with CirComPara2. *Brief. Bioinform.* **23**, bbab418 (2022).
45. Vromman, M. et al. Large-scale benchmarking of circRNA detection tools reveals large differences in sensitivity but not in precision. *Nat. Methods* **20**, 1159–1169 (2023).
46. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32**, 462–464 (2014).
47. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
48. Li, M. et al. Quantifying circular RNA expression from RNA-seq data using model-based framework. *Bioinformatics* **33**, 2131–2139 (2017).
49. Ashwal-Fluss, R. et al. circRNA biogenesis competes with pre-mRNA splicing. *Mol. Cell* **56**, 55–66 (2014).
50. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
51. Ma, X. K. et al. CIRCexplorer3: a CLEAR pipeline for direct comparison of circular and linear RNA expression. *Genomics Proteomics Bioinformatics* **17**, 511–521 (2019).
52. Morlion, A. et al. CiLiQuant: quantification of RNA junction reads based on their circular or linear transcript origin. *Front. Bioinform.* **2**, 834034 (2022).
53. Enuka, Y. et al. Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor. *Nucleic Acids Res.* **44**, 1370–1383 (2016).
54. Zhang, Y. et al. The biogenesis of nascent circular RNAs. *Cell Rep.* **15**, 611–624 (2016).

55. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
56. Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L. & Brown, P. O. Cell-type specific features of circular RNA expression. *PLoS Genet.* **9**, e1003777 (2013).
57. Hulstaert, E. et al. Charting extracellular transcriptomes in the Human Biofluid RNA Atlas. *Cell Rep.* **33**, 108552 (2020).
58. Liu, Z. et al. Detection of circular RNA expression and related quantitative trait loci in the human dorsolateral prefrontal cortex. *Genome Biol.* **20**, 99 (2019).
59. Fay, M. P. Two-sided exact tests and matching confidence intervals for discrete data. *R. J.* **2**, 53–58 (2010).
60. Feng, J. et al. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics* **28**, 2782–2788 (2012).
61. Pamudurti, N. R. et al. circMbl functions in *cis* and in *trans* to regulate gene expression and physiology in a tissue-specific fashion. *Cell Rep.* **39**, 110740 (2022).
62. Gao, Y. et al. Comprehensive identification of internal structure and alternative splicing events in circular RNAs. *Nat. Commun.* **7**, 12060 (2016).
63. Hossain, M. T., Peng, Y., Feng, S. & Wei, Y. FcircSEC: an R package for full length circRNA sequence extraction and classification. *Int. J. Genomics* **2020**, 9084901 (2020).
64. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
65. Xiao, M. S. & Wilusz, J. E. An improved method for circular RNA purification using RNase R that efficiently removes linear RNAs containing G-quadruplexes or structured 3′ ends. *Nucleic Acids Res.* **47**, 8755–8769 (2019).
66. Wu, W., Zhao, F. & Zhang, J. circAtlas 3.0: a gateway to 3 million curated vertebrate circular RNAs based on a standardized nomenclature scheme. *Nucleic Acids Res.* **52**, D52–D60 (2024).
67. Zheng, Y., Ji, P., Chen, S., Hou, L. & Zhao, F. Reconstruction of full-length circular RNAs enables isoform-level quantification. *Genome Med.* **11**, 2 (2019).
68. Huang, X. & Madan, A. CAP3: a DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).
69. Qin, Y. et al. Reference-free and de novo identification of circular RNAs. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.21.050617 (2020).
70. Peng, Y. et al. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics* **29**, i326–i334 (2013).
71. Wu, J. et al. CircAST: full-length assembly and quantification of alternatively spliced isoforms in circular RNAs. *Genomics Proteomics Bioinformatics* **17**, 522–534 (2019).
72. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).
73. Kelleher, C. D. & Champoux, J. J. Characterization of RNA strand displacement synthesis by Moloney murine leukemia virus reverse transcriptase. *J. Biol. Chem.* **273**, 9976–9986 (1998).
74. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
75. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
76. Gao, Y., Liu, B., Wang, Y. & Xing, Y. TideHunter: efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain. *Bioinformatics* **35**, i200–i207 (2019).
77. Rahimi, K., Veno, M. T., Dupont, D. M. & Kjems, J. Nanopore sequencing of brain-derived full-length circRNAs reveals circRNA-specific exon usage, intron retention and microexons. *Nat. Commun.* **12**, 4825 (2021).
78. Unlu, I., Maguire, S., Guan, S. & Sun, Z. Induro-RT mediated circRNA-sequencing (IMCR-seq) enables comprehensive profiling of full-length and long circular RNAs from low input total RNA. *Nucleic Acids Res.* **52**, e55 (2024).
79. Vincent, H. A. & Deutscher, M. P. Insights into how RNase R degrades structured RNA: analysis of the nuclease domain. *J. Mol. Biol.* **387**, 570–583 (2009).
80. Szabo, L. & Salzman, J. Detecting circular RNAs: bioinformatic and experimental challenges. *Nat. Rev. Genet.* **17**, 679–692 (2016).
81. Fuchs, S. et al. Generation of full-length circular RNA libraries for Oxford Nanopore long-read sequencing. *PLoS ONE* **17**, e0273253 (2022).
82. Zhang, J. et al. Real-time and programmable transcriptome sequencing with PROFIT-seq. *Nat. Cell Biol.* **26**, 2183–2194 (2024).
83. Ji, P. et al. Expanded expression landscape and prioritization of circular RNAs in mammals. *Cell Rep.* **26**, 3444–3460 (2019).
84. Xia, S. et al. Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. *Brief. Bioinform.* **18**, 984–992 (2017).
85. Nicolet, B. P. et al. Circular RNA expression in human hematopoietic cells is widespread and cell-type specific. *Nucleic Acids Res.* **46**, 8168–8180 (2018).
86. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
87. Garcia-Rodriguez, J. L. et al. Spatial profiling of circular RNAs in cancer reveals high expression in muscle and stromal cells. *Cancer Res.* **83**, 3340–3353 (2023).
88. Kristensen, L. S. et al. Spatial expression analyses of the putative oncogene ciRS-7 in cancer reshape the microRNA sponge theory. *Nat. Commun.* **11**, 4551 (2020).
89. Weng, W. et al. Circular RNA ciRS-7—a promising prognostic biomarker and a potential therapeutic target in colorectal cancer. *Clin. Cancer Res.* **23**, 3918–3928 (2017).
90. He, R., Zhu, J., Ji, P. & Zhao, F. SEVtras delineates small extracellular vesicles at droplet resolution from single-cell transcriptomes. *Nat. Methods* **21**, 259–266 (2024).
91. Dong, X. et al. Circular RNAs in the human brain are tailored to neuron identity and neuropsychiatric disease. *Nat. Commun.* **14**, 5327 (2023).
92. Gaffo, E. et al. Circular RNA differential expression in blood cell populations and exploration of circRNA deregulation in pediatric acute lymphoblastic leukemia. *Sci. Rep.* **9**, 14670 (2019).
93. Fan, X. et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.* **16**, 148 (2015).
94. Dang, Y. et al. Tracing the expression of circular RNAs in human pre-implantation embryos. *Genome Biol.* **17**, 130 (2016).
95. Verboom, K. et al. SMARTer single cell total RNA sequencing. *Nucleic Acids Res.* **47**, e93 (2019).
96. Sheng, K., Cao, W., Niu, Y., Deng, Q. & Zong, C. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods* **14**, 267–270 (2017).
97. Salmen, F. et al. High-throughput total RNA sequencing in single cells using VASA-seq. *Nat. Biotechnol.* **40**, 1780–1793 (2022).
98. Xu, Z. et al. High-throughput single nucleus total RNA sequencing of formalin-fixed paraffin-embedded tissues by snRandom-seq. *Nat. Commun.* **14**, 2734 (2023).

99. Isakova, A., Neff, N. & Quake, S. R. Single-cell quantification of a broad RNA spectrum reveals unique noncoding patterns associated with cell types and states. *Proc. Natl Acad. Sci. USA* **118**, e2113568118 (2021).
100. Ruan, H. et al. Comprehensive characterization of circular RNAs in ~1000 human cancer cell lines. *Genome Med.* **11**, 55 (2019).
101. Zhang, X. O. et al. Complementary sequence-mediated exon circularization. *Cell* **159**, 134–147 (2014).
102. Liang, D. & Wilusz, J. E. Short intronic repeat sequences facilitate circular RNA production. *Genes Dev.* **28**, 2233–2247 (2014).
103. Rogalska, M. E. et al. Transcriptome-wide splicing network reveals specialized regulatory functions of the core spliceosome. *Science* **386**, 551–560 (2024).
104. Van Nostrand, E. L. et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
105. Gonatopoulos-Pournatzis, T. et al. Genome-wide CRISPR–Cas9 interrogation of splicing networks reveals a mechanism for recognition of autism-misregulated neuronal microexons. *Mol. Cell* **72**, 510–524 (2018).
106. Zhu, Y. et al. POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Res.* **47**, D203–D211 (2019).
107. Chen, Y. et al. CircNet 2.0: an updated database for exploring circular RNA regulatory networks in cancers. *Nucleic Acids Res.* **50**, D93–D101 (2022).
108. Paz, I., Argoetti, A., Cohen, N., Even, N. & Mandel-Gutfreund, Y. RBPmap: a tool for mapping and predicting the binding sites of RNA-binding proteins considering the motif environment. *Methods Mol. Biol.* **2404**, 53–65 (2022).
109. Riffo-Campos, A. L., Riquelme, I. & Brebi-Mieville, P. Tools for sequence-based miRNA target prediction: what to choose? *Int. J. Mol. Sci.* **17**, 1987 (2016).
110. Chen, Y., Wang, J., Wang, C., Liu, M. & Zou, Q. Deep learning models for disease-associated circRNA prediction: a review. *Brief. Bioinform.* **23**, bbac364 (2022).
111. Vromman, M., Vandesompele, J. & Volders, P. J. Closing the circle: current state and perspectives of circular RNA databases. *Brief. Bioinform.* **22**, 288–297 (2021).
112. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C. F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274–1281 (2007).
113. van Laarhoven, T., Nabuurs, S. B. & Marchiori, E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* **27**, 3036–3043 (2011).
114. Fan, C., Lei, X. & Pan, Y. Prioritizing circRNA–disease associations with convolutional neural network based on multiple similarity feature fusion. *Front. Genet.* **11**, 540751 (2020).
115. Deepthi, K. & Jereesh, A. S. An ensemble approach for circRNA–disease association prediction based on autoencoder and deep neural network. *Gene* **762**, 145040 (2020).
116. Niu, M., Wang, C., Zhang, Z. & Zou, Q. A computational model of circRNA-associated diseases based on a graph neural network: prediction and case studies for follow-up experimental validation. *BMC Biol.* **22**, 24 (2024).
117. Niu, M., Zou, Q. & Wang, C. GMNN2CD: identification of circRNA–disease associations based on variational inference and graph Markov neural networks. *Bioinformatics* **38**, 2246–2253 (2022).
118. Wang, Y. et al. Collaborative deep learning improves disease-related circRNA prediction based on multi-source functional information. *Brief. Bioinform.* **24**, bbad069 (2023).
119. Xu, C. & Zhang, J. Mammalian circular RNAs result largely from splicing errors. *Cell Rep.* **36**, 109439 (2021).
120. Liu, C. X. & Chen, L. L. Circular RNAs: characterization, cellular roles, and applications. *Cell* **185**, 2016–2034 (2022).
121. Jarlstad Olesen, M. T. & L, S. K. Circular RNAs as microRNA sponges: evidence and controversies. *Essays Biochem.* **65**, 685–696 (2021).
122. Panda, A. C. et al. High-purity circular RNA isolation method (RPAD) reveals vast collection of intronic circRNAs. *Nucleic Acids Res.* **45**, e116 (2017).
123. Abe, B. T., Wesselhoeft, R. A., Chen, R., Anderson, D. G. & Chang, H. Y. Circular RNA migration in agarose gel electrophoresis. *Mol. Cell* **82**, 1768–1777 (2022).
124. Vromman, M. et al. Validation of circular RNAs using RT–qPCR after effective removal of linear RNAs by ribonuclease R. *Curr. Protoc.* **1**, e181 (2021).
125. Vo, J. N. et al. The landscape of circular RNA in cancer. *Cell* **176**, 869–881 (2019).
126. Hou, L., Zhang, J. & Zhao, F. Full-length circular RNA profiling by nanopore sequencing with CIRI-long. *Nat. Protoc.* **18**, 1795–1813 (2023).
127. Chiang, T. W. et al. FL-circAS: an integrative resource and analysis for full-length sequences and alternative splicing of circular RNAs with nanopore sequencing. *Nucleic Acids Res.* **52**, D115–D123 (2024).
128. Fuchs, S. et al. Defining the landscape of circular RNAs in neuroblastoma unveils a global suppressive function of MYCN. *Nat. Commun.* **14**, 3936 (2023).
129. Lai, H. et al. exoRBase 2.0: an atlas of mRNA, lncRNA and circRNA in extracellular vesicles from human biofluids. *Nucleic Acids Res.* **50**, D118–D128 (2022).
130. Zhao, J. et al. Circular RNA landscape in extracellular vesicles from human biofluids. *Genome Med.* **16**, 126 (2024).
131. Rybak-Wolf, A. et al. Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol. Cell* **58**, 870–885 (2015).
132. Jakobi, T., Uvarovskii, A. & Dieterich, C. circtools—a one-stop software solution for circular RNA research. *Bioinformatics* **35**, 2326–2328 (2019).
133. Liu, X. et al. Identification of mecciRNAs and their roles in the mitochondrial entry of proteins. *Sci. China Life Sci.* **63**, 1429–1449 (2020).
134. Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. starBase v2.0: decoding miRNA–ceRNA, miRNA–ncRNA and protein–RNA interaction networks from large-scale CLIP-seq data. *Nucleic Acids Res.* **42**, D92–D97 (2014).
135. Kim, M. S. et al. A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
136. Yang, Q. et al. dbDEPC 3.0: the database of differentially expressed proteins in human cancer with multi-level annotation and drug indication. *Database* **2018**, bay015 (2018).
137. Perez-Riverol, Y. et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552 (2022).
138. Yang, Y. et al. Extensive translation of circular RNAs driven by N^6-methyladenosine. *Cell Res.* **27**, 626–641 (2017).
139. Zhang, M. et al. A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nat. Commun.* **9**, 4475 (2018).
140. Ragan, C., Goodall, G. J., Shirokikh, N. E. & Preiss, T. Insights into the biogenesis and potential functions of exonic circular RNA. *Sci. Rep.* **9**, 2048 (2019).
141. Floor, S. N. & Doudna, J. A. Tunable protein synthesis by transcript isoforms in human cells. *eLife* **5**, e10921 (2016).
142. Liu, S., Zhu, A., He, C. & Chen, M. REPIC: a database for exploring the N^6-methyladenosine methylome. *Genome Biol.* **21**, 100 (2020).
143. Hansen, T. B. Signal and noise in circRNA translation. *Methods* **196**, 68–73 (2021).
144. Huang, W. et al. TransCirc: an interactive database for translatable circular RNAs based on multi-omics evidence. *Nucleic Acids Res.* **49**, D236–D242 (2021).
145. Zhu, J. et al. Custom microfluidic chip design enables cost-effective three-dimensional spatiotemporal transcriptomics with a wide field of view. *Nat. Genet.* **56**, 2259–2270 (2024).

146. Hu, B. et al. High-resolution spatially resolved proteomics of complex tissues based on microfluidics and transfer learning. *Cell* **188**, 734–748 (2025).
147. Long, Y. et al. Deciphering spatial domains from spatial multi-omics with SpatialGlue. *Nat. Methods* **21**, 1658–1667 (2024).
148. Cao, Z. J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).
149. Chu, Y. et al. A 5′ UTR language model for decoding untranslated regions of mRNA and function predictions. *Nat. Mach. Intell.* **6**, 449–460 (2024).
150. Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
151. Cao, X., Cai, Z., Zhang, J. & Zhao, F. Engineering circular RNA medicines. *Nat. Rev. Bioeng.* https://doi.org/10.1038/s44222-024-00259-1 (2024).
152. Rettie, S. A. et al. Cyclic peptide structure prediction and design using AlphaFold. Preprint at *bioRxiv* https://doi.org/10.1101/2023.02.25.529956 (2023).

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (32130020 and 32025009 to F.Z. and 32200530 and 32422020 to J.Z.) and the National Key R&D Project (2021YFA1300500 and 2021YFA1302000 to J.Z.).

Author contributions

F.Z. conceived the project. J.Z. conducted the initial literature research. All authors contributed to writing and edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02157-7.

Correspondence and requests for materials should be addressed to Jinyang Zhang or Fangqing Zhao.

Peer review information *Nature Genetics* thanks Jo Vandesompele and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.