nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-025-02835-1

# Detecting and quantifying circular RNAs in terabyte-scale RNA-seq datasets with CIRI3

Received: 25 February 2025

Accepted: 29 August 2025

Published online: 01 October 2025

Check for updates

Xin Zheng[1,2,5], Jinyang Zhang[3,5], Lipu Song[1,5], Xiang Jennie Li[1,2], Fangqing Zhao[2,3,4] ✉ & Yuan Gao[1,2] ✉

To address recent challenges in circular RNA (circRNA) analysis, we present CIRI3, a tool for circRNA detection and quantification in terabyte-scale RNA-sequencing datasets. Using dynamic multithreaded task partitioning and a blocking search strategy for junction reads, CIRI3 is an order of magnitude faster than existing tools, while providing increased accuracy. We identified differentially spliced circRNAs across 2,535 cancer-related samples, and constructed a pretraining model and a biomarker network provided as the CIRIonco database.

Circular RNAs (circRNAs) are a unique class of noncoding RNA molecules characterized by their covalent circular structure[1]. Recent studies have revealed their important roles in various cellular processes, including the regulation of signaling pathways, sequestration of miRNA and RNA-binding proteins, initiation of gene transcription, inhibition of mRNA translation and promotion of protein degradation[1–10]. Advances in RNA-sequencing (RNA-seq) technologies have enabled the rapid accumulation of large RNA-seq datasets, providing unprecedented opportunities for circRNA research. For instance, Chinnaiyan et al. generated sequencing data from over 2,000 human cancer samples, identifying circRNAs with potential as cancer biomarkers[11]. These expansive datasets also serve as crucial resources for training large-scale models to predict circRNAs and their regulation mechanisms. In our previous work, we used RNA-seq data from 394 tissue and cell line samples to train a deep learning model capable of predicting circRNA differential splicing from single-cell and spatial transcriptomics data, as well as other low-depth datasets[12]. Given the diverse roles and complex regulation networks of circRNAs, the continued expansion of RNA-seq data holds immense potential for uncovering novel biogenesis and degradation mechanisms, predicting unknown functions and optimizing sequence designs for therapeutic applications.

Current circRNA detection methods can be generally classified into alignment-based and pseudo-reference-based approaches[13,14] (Supplementary Table 1). Alignment-based approaches identify circRNA-specific alignment features distinct from linear RNAs or background noise[15]. For example, we developed CIRI[16] and CIRI2 (ref. 17), a tool series renowned for reliable de novo circRNA detection[18] and superior sensitivity in recent large-scale benchmarking[13]. In contrast, pseudo-reference-based approaches rely on predefined back-splice junction (BSJ) libraries constructed from annotated exon combinations. While effective in reducing false positives, this strategy is limited to well-annotated genomes and cannot detect novel circRNAs with unannotated splice sites[13,14]. Despite a decade of development, current tools still suffer from poor scalability due to exponentially increasing runtime and memory demands on large RNA-seq datasets, and accurate quantification of circRNAs remains challenging due to their low abundance relative to mRNAs[19] and strong batch effects across different RNA-seq cohorts.

To address these challenges, we introduce CIRI3, a tool for large-scale circRNA detection and characterization. Building on its predecessor, CIRI3 is optimized for rapid circRNA detection from multisample alignment results, offering improved quantification accuracy, runtime efficiency and memory usage. Key innovations in CIRI3 include robust identification of intronic self-ligated circRNAs and targeted quantification for user-defined circRNA lists. CIRI3 supports alignments from both BWA[20] and STAR[21], yielding consistent results across aligners (Extended Data Fig. 1a–c). CIRI3 outputs both BSJ and forward-splice junction (FSJ) reads, enabling comprehensive statistical analyses, including differential expression, differential splicing, and differential circularization (Methods).

The CIRI3 workflow consists of two primary alignment-scanning modules—high-confidence BSJ discovery and FSJ/BSJ read recovery (Fig. 1a). To optimize efficiency, CIRI3 adopts a dynamic multithreaded task partitioning approach during the two scanning phases, mitigating

[1]China National Center for Bioinformation, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]Institute of Zoology, Chinese Academy of Sciences, Beijing, China. [4]Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Beijing, China. [5]These authors contributed equally: Xin Zheng, Jinyang Zhang, Lipu Song. ✉e-mail: zhfq@ioz.ac.cn; gaoy@big.ac.cn

https://doi.org/10.1038/s41587-025-02835-1

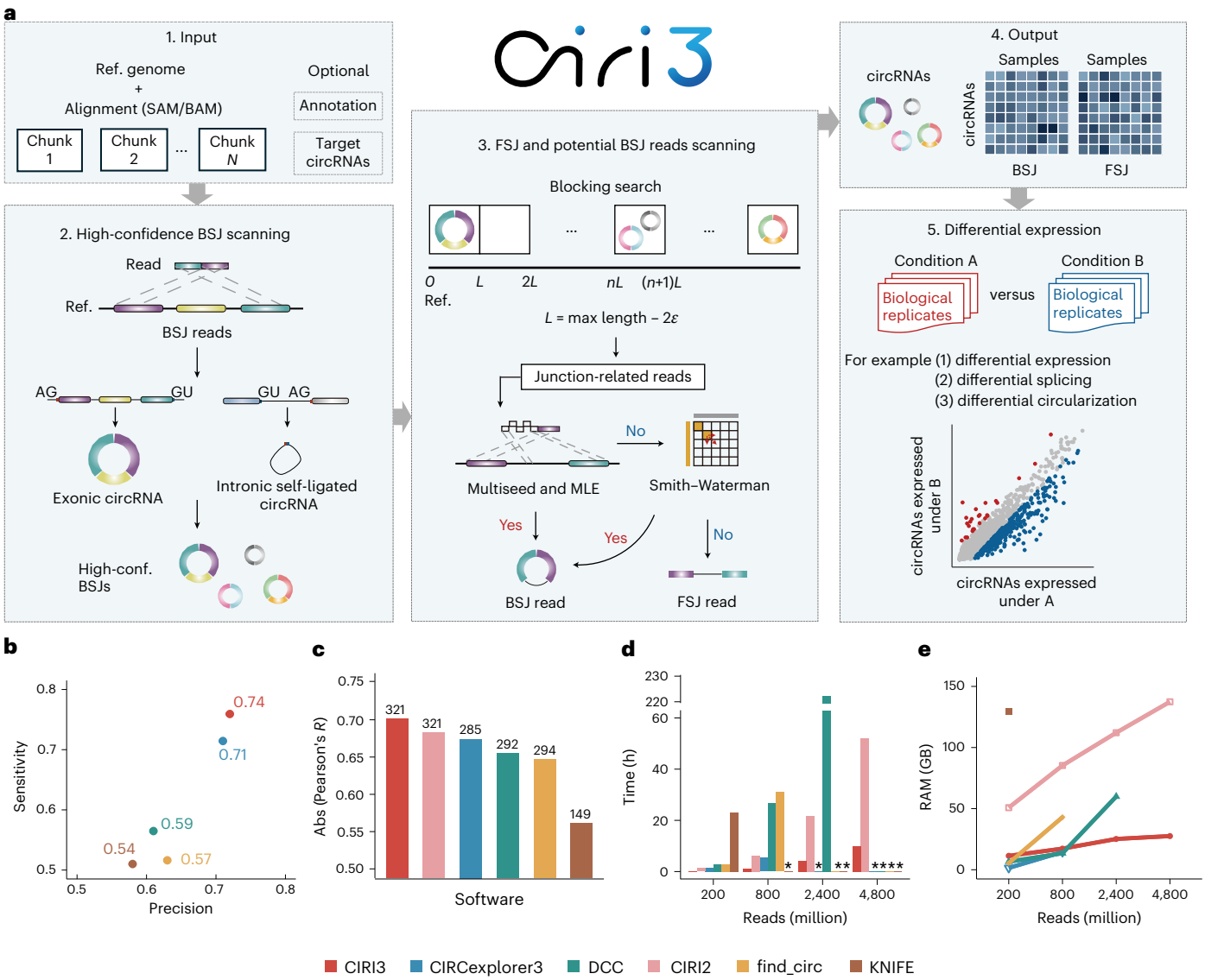



**Fig. 1 | Overview of the CIRI3 pipeline. a**, Step 1: input files include SAM/BAM alignments and a reference (ref.) genome (required), with annotation (including exon coordinates) and circRNA list files optional. Step 2: high-confidence (high-conf.) BSJs are identified by correcting and filtering chiastic clipping signals using splicing signals (for example, GT–AG). Step 3: FSJ and candidate BSJ reads are obtained using blocking search, multiseed matching with MLE and Smith–Waterman alignment. Step 4: CIRI3 generates detailed circRNA annotations and expression profiles for BSJ and FSJ reads. Step 5: CIRI3 provides three types of statistical analysis. **b**, Sensitivity and precision of five circRNA detection tools on the Hs68 cell line, with corresponding *F*1 scores indicated. **c**, The bar plot shows the absolute (Abs) Pearson correlation values between $\log_2$(BSJ read counts) of software-predicted circRNAs and the $C_q$ values of experimentally validated circRNAs from RT–qPCR experiments in the SW480 cell line. The numbers above each bar represent the counts of experimentally validated circRNAs detected by each tool. **d**, Runtime of six tools on datasets with 200, 800, 2,400 and 4,800 million reads. All tools were run using 10 threads, except for KNIFE, which used the default thread setting. The asterisk indicates that the entire circRNA analysis process could not be completed within 14 days on a 256 GB memory server. **e**, The memory required (random-access memory (RAM)) by the tools for the same data as in **d**.

runtime inefficiencies caused by variability in single-thread performance (Methods and Fig. 1a, step 1). In the first scan, CIRI3 identifies potential paired chiastic clipping signals in read alignment (Fig. 1a, step 2), a highly sensitive proxy for BSJ[16,17]. To ensure reliability, CIRI3 refines and filters paired chiastic clipping signals by requiring perfectly matched splicing signals flanking the putative BSJs, including canonical GT–AG dinucleotides or noncanonical splicing motifs. When a GTF/GFF annotation is provided, CIRI3 extracts exon and intron boundaries to further enhance splicing signal filtration. These steps yield a set of high-confidence BSJs.

In the second scan, CIRI3 employs a blocking search approach to recover missed BSJ reads and identify FSJ reads associated with each high-confidence BSJ (Fig. 1a, step 3). The reference genome is segmented into small blocks and each high-confidence BSJ is indexed into a hash table using the blocks corresponding to its splice sites as key values. This enables targeted searches for junction reads within the associated blocks and their neighbors, substantially improving search efficiency. For reads supporting only one splice site of a BSJ in raw alignments, CIRI3 re-evaluates their splice junction positions and orientations through pseudo- or re-alignment with genome sequences. Building on the multiseed pseudo-alignment and maximum likelihood estimation (MLE) adopted in CIRI2, CIRI3 incorporates Smith–Waterman local sequence alignment as an additional criterion to improve classification accuracy. By applying count or ratio thresholds to the BSJs identified in the first scan, CIRI3 generates detailed annotations and expression profiles of circRNAs across multiple samples (Fig. 1a,

step 4). Finally, CIRI3 facilitates downstream differential expression analyses using its integrated statistical algorithms (Fig. 1a, step 5).

To evaluate circRNA detection performance, we compared CIRI3 with five widely used tools (that is, find_circ[8], KNIFE[22], CIRCexplorer3 (ref. 23), DCC[24] and CIRI2) using RNA-seq data from Hs68 cell line samples treated with or without RNase R[25]. Detected circRNAs were classified as enriched (putative positives), depleted (false positives) or unaffected (Methods). Compared with find_circ, KNIFE, CIRCexplorer3 and DCC, CIRI3 demonstrated a higher putative positive and lower false positive rate (Extended Data Fig. 1d–h). While CIRI2 showed substantial overlap with CIRI3, 54 out of 109 circRNAs uniquely detected by CIRI3 were putative positives. Among all these tools, CIRI3 achieved the highest sensitivity and precision ($F1$ score of 0.74) (Fig. 1b). Notably, CIRI3 detected the most putative positives among circRNAs unique to each tool.

Moreover, CIRI3 can detect intronic self-ligated circRNAs that were undetectable by other short-read tools[26] (Extended Data Fig. 2). Applied to liver RNA-seq samples from five species[27], CIRI3 identified 59 such events, where all 16 detected in opossum were enriched after RNase R treatment. The RNase R validated intronic self-ligated circRNAs were predominantly ranged from 300 to 800 bp, and 90% originated from protein-coding genes. In a prostate cancer cohort ($n = 181$), CIRI3 detected 2,286 intronic self-ligated circRNAs derived from introns that were significantly shorter than those not involved in circRNA formation (Wilcoxon rank-sum test, $P < 2.2 \times 10^{-16}$), suggesting that shorter introns are more prone to back-splicing and circRNA biogenesis. Together, these results highlight the scalability and efficiency of CIRI3 in identifying diverse circRNA subtypes across different species and disease contexts.

To benchmark the quantification accuracy of CIRI3 against other tools, we analyzed simulated paired-end RNA-seq datasets with 20–100× coverage, calculating the Pearson correlation coefficient (PCC) and root mean squared error (r.m.s.e.) between the BSJ read counts identified by each tool and the simulated BSJ read counts[17] (Extended Data Fig. 3). CIRI3 consistently achieved PCC values above 0.983, with a mean of 0.990, outperforming all others across coverage levels. This improvement over CIRI2 (mean PCC of 0.954) can be attributed to the integration of the Smith–Waterman alignment, which recovers BSJ reads missed by CIRI2. Furthermore, CIRI3 accurately quantified FSJ reads and junction ratios, achieving mean PCC values of 0.977 and 0.980, respectively. In terms of r.m.s.e., CIRI3 consistently exhibited the lowest errors across all coverage levels, further confirming its superior quantification accuracy.

Next, we further evaluated CIRI3's performance on real RNA-seq data from three cell lines (SW480, NCI-H23 and HLF), with 1,479 circRNAs quantified using RT–qPCR[13]. Among the six tools evaluated, CIRI3 and CIRI2 were the most sensitive, detecting 1,172 and 1,174 validated circRNAs, respectively. CIRI3 also demonstrated the highest quantification accuracy for the SW480 and NCI-H23 datasets, with PCC values of −0.701 and −0.728, respectively, when comparing log-transformed BSJ read counts to $C_q$ (quantification cycle) values (Fig. 1c and Extended Data Fig. 4). For the HLF cell line, all tools exhibited lower accuracy, but CIRI3, CIRCexplorer3 and DCC showed comparable correlations ranging from −0.656 to −0.653, outperforming other methods. We also benchmarked computational efficiency (Extended Data Fig. 5a,b and Supplementary Table 1). CIRI3 processed the 295-million-read SW480 dataset in just 0.25 h, while other tools were 8–149 times slower, requiring 2.0–37.1 h with 25 threads. Memory usage was also a notable challenge for most tools. CIRCexplorer3, find_circ, DCC, CIRI2 and KNIFE required 27.7, 34.9, 50.8, 139.2 and 205.1 GB of memory, respectively, substantially exceeding the modest 12.2 GB required by CIRI3.

In large-scale data analysis, a common strategy to reduce computational resource requirements is to process datasets individually before combining results during downstream integration. To assess the impact of this separate-detection mode on circRNA analysis, we divided the SW480 dataset into three subsets and compared the results obtained from processing the entire dataset (joint-detection mode) versus combining results from the subsets. We found that all tools showed compromised performance in circRNA detection and quantification when using the separate-detection mode. Taking find_circ and DCC as examples (Extended Data Fig. 5c–f), the separate-detection mode reduced memory usage by 49.3% and 22.6%, respectively, but detected 22,719 and 8,312 fewer circRNAs, missing 53 and 11 out of 294 and 292 RT–qPCR validated circRNAs, respectively. In addition, quantification accuracy declined, with the absolute PCC dropping from 0.647 and 0.655 to 0.528 and 0.592, respectively. When focusing on low-abundance circRNAs, the correlation between read counts from the separate-detection mode and the joint-detection mode was only moderate, with PCC values of 0.682 and 0.789 for find_circ and DCC, respectively. Similarly, reprocessing the Hs68 cell line samples using this strategy resulted in a substantial reduction in enriched circRNAs detected by find_circ. To further evaluate the impact on cohort-level circRNA analysis, we analyzed data from 181 prostate cancer samples using CIRI3 for a comparison between the joint-detection mode and the separate-detection mode. The results showed that the separate-detection mode identified 24,885 fewer circRNAs, corresponding to a 16.2% decrease (Extended Data Fig. 5g). Furthermore, the separate-detection mode resulted in lower average BSJ read counts (47.5 versus 38.6) and fewer identified samples (14.7 versus 11.3) per circRNA. In particular, the joint-detection mode identified 38,578 highly prevalent circRNAs expressed in at least 15 samples, which is 38.7% more than the separate-detection mode. These findings underscore the limitations of the separate-detection mode and highlight the importance of computational efficiency, which enables simultaneous processing of large or multiple datasets for comprehensive circRNA analysis.

To further evaluate the computational efficiency, we tested all tools on RNA-seq datasets with 200 million, 800 million, 2,400 million and 4,800 million reads. CIRI3 was the only tool capable of processing the terabyte-sized 4,800 million-read datasets in 24 h (Fig. 1d). While all tools could handle 200 million-read dataset, KNIFE failed to process the 800 million-read dataset, and DCC was the only tool besides CIRI2 and CIRI3 that could process the 2,400 million-read dataset within 2 weeks using less than 256 GB of memory (with ten threads). Compared with CIRI2, CIRI3 demonstrated much faster processing speeds and lower memory usage. Notably, while memory usage increased with dataset size for all tools, CIRI3 showed the smallest increase, ranging from 12 GB to 27.5 GB (Fig. 1e). As an ultimate stress test, we attempted to process a collection of 296 deeply sequenced samples from RNAAtlas[28], totaling 39.2 trillion reads. CIRI3 was the only tool capable of completing this task, processing 21 TB of SAM files in 105.31 h with a peak memory usage of 45.85 GB.

To systematically evaluate the robustness of circRNA detection tools, we performed subsampling analyses using RNA-seq data from two cell lines (Methods and Extended Data Fig. 6a–f). In Hs68, the proportion of enriched circRNAs detected by each tool remained stable, while the absolute number of enriched circRNAs increased markedly with sequencing depth. Notably, CIRI3 consistently identified the highest number of enriched circRNAs across all subsampling levels. Further validation was conducted using RNA-seq data from the SW480 cell line, which includes 416 RT–PCR-validated circRNAs. The results demonstrated a steady improvement in detection performance for all tools as sequencing depth increased. Among them, CIRI3 exhibited the best performance at all subsampling levels, identifying the largest number of validated circRNAs. Taken together, these findings indicate that although all tools demonstrate good robustness, CIRI3 shows a clear advantage in detection sensitivity.

Inspired by the Percent Spliced In metric widely used in splicing analysis, the BSJ ratio was designed to quantify the relative abundance

https://doi.org/10.1038/s41587-025-02835-1

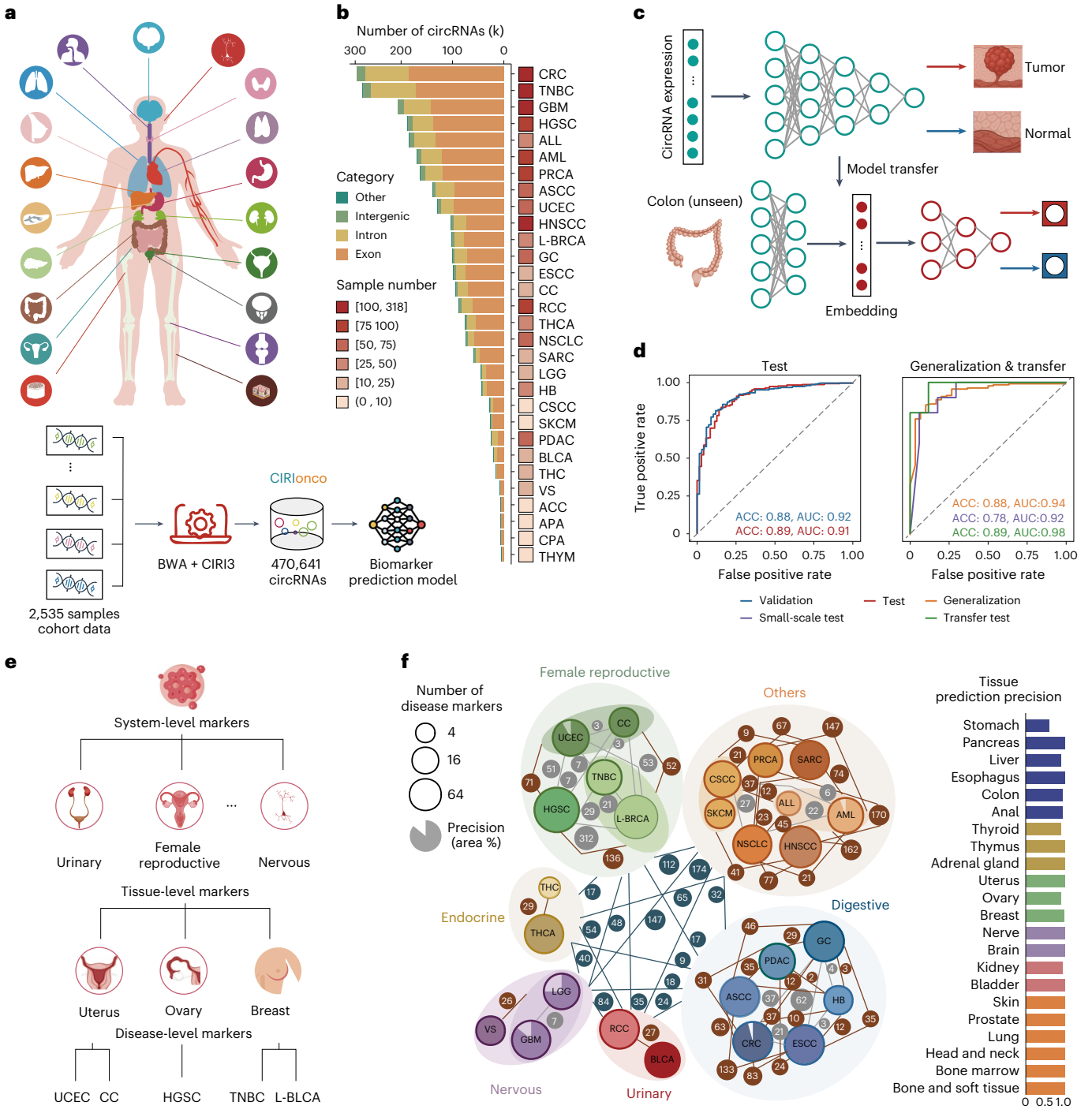



**Fig. 2 | Cohort analysis of human cancer tissue data by CIRI3. a**, An overview of the CIRIonco database dataset: 2,535 total RNA samples from the GEO database were analyzed, encompassing 22 tissue types and 30 cancer types. circRNAs were identified using the BWA and CIRI3 pipeline to construct a tumor-specific circRNAs database. **b**, A stacked bar plot showing the number of circRNAs detected across different cancer types, with colors indicating different circRNA types. The filled squares represent sample number for each disease. CRC, colorectal cancer; TNBC, triple-negative breast cancer; GBM, glioblastoma multiforme; HGSC, high-grade serous carcinoma; ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; PRCA, prostate cancer; ASCC, anal squamous cell carcinoma; UCEC, uterine corpus endometrial carcinoma; HNSCC, head and neck squamous cell carcinoma; L-BRCA, luminal breast cancer; GC, gastric cancer; ESCC, esophageal squamous cell carcinoma; CC, cervical cancer; RCC, renal cell carcinoma; THCA, thyroid carcinoma; NSCLC, non-small cell lung cancer; SARC, sarcoma; LGG, low-grade glioma; HB, hepatoblastoma; CSCC, cutaneous squamous cell carcinoma; SKCM, skin cutaneous melanoma; PDAC, pancreatic ductal adenocarcinoma; BLCA, bladder urothelial carcinoma; THC, thymic carcinoid; VS, vestibular schwannoma; ACC, adrenocortical carcinoma; APA, aldosterone-producing adenoma; CPA, cortisol-producing adenoma; THYM, thymoma. **c**, A schematic of the pretrained model, which leverages differentially spliced circRNAs to classify cancer versus normal samples and the transfer learning model applied to colon tissue. **d**, Left: validation and test AUROC of the pretrained model on non-colon samples. Right: AUROC on colon samples for generalization, small-scale training and transfer learning. ACC, accuracy. **e**, A schematic of the hierarchical tree based on system, tissue and disease levels, containing candidate markers for classification. **f**, Left: the network plot illustrates overlaps in circRNAs markers across hierarchical levels. The node size and pie chart represent marker count and prediction precision. The circle on the edge shows the overlap marker number with different background colors (gray, disease level; brown, tissue level; blue, system level). Right: the bar plot shows the predict precision at the tissue level.

and differential splicing of circRNAs, offering insights into their biogenesis regulation[16,17,24,29]. However, batch effects in RNA-seq data have been reported to impact expression quantification based on read counts[30]. To evaluate this for BSJ ratios calculated by CIRI3, we analyzed RNA-seq data from 62 samples across four tissues (Supplementary Table 2)[27,31–36], and performed principal component analysis on both BSJ read counts and junction ratios (Extended Data Fig. 7a). While BSJ read counts failed to distinguish samples by tissue type, circRNA junction ratios clearly clustered samples according to their tissue of origin, with minimal batch effects. For example, liver tissue samples showed unique circRNA abundance profiles based on BSJ ratios but showed similar expression patterns to spleen and testis samples when using BSJ read counts.

We next investigated whether BSJ ratios could serve as biomarkers in clinical research. Using RNA-seq data from four studies, we analyzed 44 pairs of tumor and adjacent normal tissue samples from patients with hepatocellular carcinoma (HCC)[31,37,38]. Applying the statistical tests in CIRI3, we identified 102 differentially expressed circRNAs based on read counts and 563 differentially spliced circRNAs based on junction ratios. Only 18 circRNAs overlapped between the two analyses, and the host genes of these two circRNA groups showed significant enrichment in distinct biologically relevant pathways, suggesting that BSJ ratios capture a different set of circRNAs with potential clinical relevance (Extended Data Fig. 6g,h). While differentially spliced circRNAs exhibited clear BSJ ratio patterns distinguishing tumor and normal samples (Extended Data Fig. 7b,c), no such patterns were observed in the counts per million (CPM) values of differentially expressed circRNAs. Notably, several experimentally validated HCC-associated circRNAs, including hsa-MET-0001 (ref. 39), hsa-SMARCA5_0005 (ref. 40) and hsa-ZKSCAN1_0001 (ref. 41) (circAtlas ID), showed significant changes in junction ratios (Extended Data Fig. 7d).

To compare the utility of junction ratios and read counts to identify biomarkers, we trained support vector machine models using BSJ ratios and CPM values of 30 representative circRNAs from three studies to classify normal and tumor samples across all four studies. BSJ ratio models consistently outperformed CPM-based models, achieving higher accuracy in the test datasets (mean value 0.924 versus 0.661) (Extended Data Fig. 7e,f), The smaller performance gap between training and test data for BSJ ratio models further underscores their low batch effect and generalization ability in biomarker studies.

To systematically investigate the expression pattern and diagnostic potential of circRNAs in cancer, we collected 2,535 total RNA-seq data from human cancerous and normal tissue samples, covering 30 cancer types (Methods) (Fig. 2a). We identified 470,641 circRNAs across all samples, with an average of 8,245 detected in each sample. Colorectal cancer (CRC), triple-negative breast cancer (TNBC) and glioblastoma multiforme (GBM) samples exhibited the highest numbers, indicating abundant circRNA expression in these cancer types (Fig. 2b). On the basis of this dataset, we constructed CIRIonco (CIRI Oncology, https://ngdc.cncb.ac.cn/ciriionco), a comprehensive circRNA database. A comparison with existing circRNA databases[11,42,43] revealed an overlap of 294,692 circRNAs (62.6%) (Extended Data Fig. 8a). A higher proportion of BSJs recorded exclusively in CIRIonco were located in intronic regions, suggesting the high sensitivity of CIRI3 in identifying these previously underrepresented circRNAs due to its annotation-independent design (Extended Data Fig. 8b).

Next, we used differentially spliced circRNAs between cancer and normal samples as input features to train a five-layer fully connected deep neural network (pretrained model) for sample classification (Fig. 2c). This pretrained model performed well on both the validation and test datasets, achieving overall accuracies and areas under the receiver operating characteristic curve (AUROCs) over 88% and 0.91, respectively (Fig. 2d). A generalization test on colon tissue samples not included in the training set achieved an accuracy of 88% and an AUROC of 0.94, indicating strong performance on unseen tissue types. We further assessed the model's transferability using 172 colon tissue samples from small cohorts in multiple studies, with one study used as the test set and the others used for training in each evaluation. While the newly trained model yielded a test accuracy of only 77.78%, transfer learning by freezing the first two layers of the pretrained model and fine-tuning the remaining layers achieved an accuracy of 88.89% and an area under the curve of 0.98, demonstrating the pretrained model's superior performance in small-sample data.

Building upon this framework, we further used circRNAs as biomarkers to stratify cancer samples at the system, tissue and disease levels (Fig. 2e and Extended Data Fig. 8c). We constructed a system, tissue and disease stratification tree, and used differential spliced circRNAs as candidate markers at each hierarchical level. These markers were then used to train LightGBM classifiers to predict the system, tissue or disease origin of each sample. Our results revealed substantial overlap and connection among marker circRNAs across different systems, tissues and diseases, and the consequent biomarker network highlighted the complexity and diversity of circRNA regulation (Fig. 2f). LightGBM classifiers achieved high classification performance, with mean precision values of 0.959 at both the system and tissue levels, and 0.974 at the disease level, further demonstrating the strong potential of BSJ ratio-based circRNAs as robust biomarkers. The CIRIonco database provides an extensive and scalable resource for cancer-related circRNA research and functional exploration, laying a solid foundation for their application in cancer subtyping and precision diagnostics.

Over the past decade, our understanding of circRNAs has greatly improved, in part due to advancements in circRNA detection methodology that have facilitated the discovery of their biogenesis and functions[44–48]. In this study, we presented CIRI3, which addresses several critical challenges in circRNA detection. The scalable design of CIRI3 makes it highly efficient at processing cohort-scale data, and also capable of discovering underexplored circRNAs lacking canonical GT–AG splice sites, such as intronic self-ligated circRNAs. CIRI3 also facilitates diverse downstream analysis by providing accurate identification and quantification of circRNAs. For example, while CIRI3 was not developed to directly detect the internal structure or full-length isoform of circRNAs, its precise BSJ position output can enhance the detection of these features by leveraging either overlapped paired-end data (for example, CIRI-full[33]) or long-read (for example, CIRI-long[26]) data.

It should be noted that the performance of circRNA detection methods is affected by library preparation strategies. RNase R treatment is a circRNA-specific enrichment strategy, which largely improves the sensitivity of detection. However, samples treated by RNase R only constitute a relatively small proportion of existing RNA-seq data and the efficiency of enrichment varies across samples and protocols, making RNase R treatment unsuitable for quantitative analysis[14]. Nevertheless, CIRI3 enhances the accuracy of BSJ ratio measurement, a valuable metric for filtering candidate circRNAs based on relative changes between RNase R-treated and untreated total RNA-seq data[17,24]. In contrast, total RNA-seq data preserves the features of both circRNAs and linear RNAs, representing the most commonly used data for circRNA analysis. The BSJ ratio calculated from total RNA-seq data reflects the natural proportion of circRNAs compared with other RNAs, and our study further demonstrated its high reliability and low variability across different datasets, supporting its utility in circRNA biomarker identification. However, CIRI3 was not specifically designed for correcting batch effects arising from variations in RNA integrity values[33] or circRNA sequencing protocols[14]. Therefore, careful experimental design and rigorous quality control of circRNA libraries remain essential to minimize technical batch effects.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-025-02835-1.

## References


1. Kristensen, L. S. et al. The biogenesis, biology and characterization of circular RNAs. *Nat. Rev. Genet.* **20**, 675–691 (2019).
2. Li, X. et al. Linking circular intronic RNA degradation and function in transcription by RNase H1. *Sci. China Life Sci.* **64**, 1795–1809 (2021).
3. Conn, V. M. et al. A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation. *Nat. Plants* **3**, 17053 (2017).
4. Ashwal-Fluss, R. et al. circRNA biogenesis competes with pre-mRNA splicing. *Mol. Cell* **56**, 55–66 (2014).
5. Xia, P. et al. A circular RNA protects dormant hematopoietic stem cells from DNA sensor cGAS-mediated exhaustion. *Immunity* **48**, 688–701 e687 (2018).
6. Liu, C. X. et al. Structure and degradation of circular RNAs regulate PKR activation in innate immunity. *Cell* **177**, 865–880 e821 (2019).
7. Hansen, T. B. et al. Natural RNA circles function as efficient microRNA sponges. *Nature* **495**, 384–388 (2013).
8. Memczak, S. et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338 (2013).
9. Rossi, F. et al. Circular RNA ZNF609/CKAP5 mRNA interaction regulates microtubule dynamics and tumorigenicity. *Mol. Cell* **82**, 75–89 e79 (2022).
10. Li, S. et al. Screening for functional circular RNAs using the CRISPR–Cas13 system. *Nat. Methods* **18**, 51–59 (2021).
11. Vo, J. N. et al. The landscape of circular RNA in cancer. *Cell* **176**, 869–881 e813 (2019).
12. Zhou, Z. et al. CIRI-Deep enables single-cell and spatial transcriptomic analysis of circular RNAs with deep learning. *Adv. Sci.* **11**, e2308115 (2024).
13. 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).
14. Zhang, J. & Zhao, F. Circular RNA discovery with emerging sequencing and deep learning technologies. *Nat. Genet.* **57**, 1089–1102 (2025).
15. Gao, Y. & Zhao, F. Computational strategies for exploring circular RNAs. *Trends Genet.* **34**, 389–400 (2018).
16. Gao, Y., Wang, J. & Zhao, F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.* **16**, 4 (2015).
17. Gao, Y., Zhang, J. & Zhao, F. Circular RNA identification based on multiple seed matching. *Brief. Bioinform.* **19**, 803–810 (2018).
18. Hansen, T. B. Improved circRNA identification by combining prediction algorithms. *Front. Cell Dev. Biol.* **6**, 20 (2018).
19. Rybak-Wolf, A. et al. Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol. Cell* **58**, 870–885 (2015).
20. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
21. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
22. 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).
23. Ma, X. K. et al. CIRCexplorer3: a CLEAR pipeline for direct comparison of circular and linear RNA expression. *Genomics Proteom. Bioinform.* **17**, 511–521 (2019).
24. Cheng, J., Metge, F. & Dieterich, C. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* **32**, 1094–1096 (2016).
25. Jeck, W. R. et al. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**, 141–157 (2013).
26. Zhang, J. et al. Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long. *Nat. Biotechnol.* **39**, 836–845 (2021).
27. Gruhl, F., Janich, P., Kaessmann, H. & Gatfield, D. Circular RNA repertoires are associated with evolutionarily young transposable elements. *eLife* https://doi.org/10.7554/eLife.67991 (2021).
28. Lorenzi, L. et al. The RNA Atlas expands the catalog of human non-coding RNAs. *Nat. Biotechnol.* **39**, 1453–1465 (2021).
29. Zhang, J., Chen, S., Yang, J. & Zhao, F. Accurate quantification of circular RNAs identifies extensive circular isoform switching events. *Nat. Commun.* **11**, 90 (2020).
30. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
31. Yang, Y. et al. Recurrently deregulated lncRNAs in hepatocellular carcinoma. *Nat. Commun.* **8**, 14421 (2017).
32. Nielsen, M. M. et al. Identification of expressed and conserved human noncoding RNAs. *RNA* **20**, 236–251 (2014).
33. 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).
34. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
35. Ji, P. et al. Expanded expression landscape and prioritization of circular RNAs in mammals. *Cell Rep.* **26**, 3444–3460 e3445 (2019).
36. Zheng, Q. et al. Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs. *Nat. Commun.* **7**, 11215 (2016).
37. Sheng, Z., Wang, X., Xu, G., Shan, G. & Chen, L. Analyses of a panel of transcripts identified from a small sample size and construction of RNA networks in hepatocellular carcinoma. *Front. Genet.* **10**, 431 (2019).
38. Chen, X. et al. circGLS2 inhibits hepatocellular carcinoma recurrence via regulating hsa-miR-222-3p–PTEN–AKT signaling. *Signal Transduct. Target. Ther.* **8**, 67 (2023).
39. Huang, X. Y. et al. Circular RNA circMET drives immunosuppression and anti-PD1 therapy resistance in hepatocellular carcinoma via the miR-30-5p/snail/DPP4 axis. *Mol. Cancer* **19**, 92 (2020).
40. Yu, J. et al. Circular RNA cSMARCA5 inhibits growth and metastasis in hepatocellular carcinoma. *J. Hepatol.* **68**, 1214–1227 (2018).
41. Zhu, Y. J. et al. Circular RNAs negatively regulate cancer stem cells by physically binding FMRP against CCAR1 complex in hepatocellular carcinoma. *Theranostics* **9**, 3526–3540 (2019).
42. 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).
43. 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).
44. Guo, S. K. et al. Therapeutic application of circular RNA aptamers in a mouse model of psoriasis. *Nat. Biotechnol.* **43**, 236–246 (2025).
45. Liang, R. et al. Prime editing using CRISPR–Cas12a and circular RNAs in human cells. *Nat. Biotechnol.* **42**, 1867–1875 (2024).
46. Qu, L. et al. Circular RNA vaccines against SARS-CoV-2 and emerging variants. *Cell* **185**, 1728–1744 e1716 (2022).
47. Zhang, W. et al. Exosomal circEZH2_005, an intestinal injury biomarker, alleviates intestinal ischemia/reperfusion injury by mediating Gprc5a signaling. *Nat. Commun.* **14**, 5437 (2023).

48. Xu, C. et al. A circulating panel of circRNA biomarkers for the noninvasive and early detection of pancreatic ductal adenocarcinoma. *Gastroenterology* **166**, 178–190 e116 (2024).

## Methods

### CIRI3 input and output

CIRI3 requires alignment files (SAM/BAM) generated by mapping FASTQ files to a reference genome (FASTA) using BWA or STAR (Supplementary File 1), as well as the same reference genome (FASTA) as inputs. Optionally, users can provide a gene annotation file (GTF) and a list of circRNAs of interest. This pipeline outputs detailed annotations of circRNAs and their expression profiles for downstream analysis.

### Partition of alignment files

CIRI3 employs dynamic multithreaded task partitioning to optimize computational resource allocation. Alignment files are segmented into chunks, with the number of chunks adjusted based on input size and the preset thread number.

If the input data are less than 200 GB or if dividing the input size by the number of threads results in chunks less than 20 GB, the number of chunks is set to be the number of threads, with each thread processing one chunk. Otherwise, the data are split into 20 GB chunks. As scanning proceeds, CIRI3 assigns available threads to unscanned chunks, preventing overruns for single thread and improving overall efficiency.

### High-confidence BSJ discovery

CIRI3 begins by scanning SAM alignments to identify reads exhibiting paired chiastic clipping patterns, characterized by chiastic soft/hard clipping in their CIGAR strings (for example, $x$S/H$y$M$z$S/H paired with ($x+y$)S/H$z$M and/or $x$M($y+z$)S/H), which indicate candidate back-splice junctions. Putative junctions are further validated by identifying canonical (GT–AG) or noncanonical splicing signals and, if provided, exon–intron boundaries from the annotation file. Reads passing these filters are classified as supporting 'high-confidence BSJs'. Among these, reads with mapping scores above a threshold of 10 for both segments are designated as 'high-confidence BSJ reads'.

### FSJ/BSJ reads recovery

In the second scan, CIRI3 utilizes a blocking search approach. The reference genome is segmented into non-overlapping blocks, with each block's length defined as: block_size = max_read_length − 2 × ε, where max_read_length is the maximum read length and ε is a tolerance parameter. This parameter provides a margin of error for the alignment positions of reads supporting BSJ sites, thereby improving detection sensitivity. High-confidence BSJs are indexed into these intervals based on their coordinates, and targeted searches are conducted within the associated block and neighboring blocks to recover junction-related reads. Reads supporting only one splice site of a high-confidence BSJ are re-evaluated using MLE based on multiple seed matching. If necessary, the Smith–Waterman algorithm is applied for precise alignment and classification.

### circRNA annotation and quantification

CIRI3 quantifies circRNAs using the identified junctions. For BSJs sharing start or end sites, the Smith–Waterman algorithm assesses sequence similarity to determine if they represent distinct junctions. High-confidence BSJ reads from similar junctions are merged toward the junction with higher counts. Finally, CIRI3 outputs detailed circRNA annotations and expression profiles based on supporting BSJ and FSJ reads.

### Identification of the intronic self-ligated circRNAs

CIRI3 identifies intronic self-ligated circRNAs during the first scan. The coordinates of these BSJ reads are corrected using intron boundaries from the annotation file. The Smith–Waterman algorithm aligns the 5 bases from both ends of these reads against the reference genome, filtering out reads with >1 bp insertion, deletion or mismatch.

### Differential expression analysis

CIRI3 provides three levels of differential expression analysis: (1) differential expression and (2) differential splicing of circRNAs and (3) differential circularization (relative abundances of circRNAs from the same gene).

For differential expression analysis of single circRNA, CIRI3 employs algorithms from CIRIquant[29]. For datasets without biological replicates, differential expression scores are calculated using a generalized fold change approach. For datasets with replicates, CIRI3 mitigates systematic batch effects before performing differential expression analysis. First, gene expression profile is obtained using featureCounts (version 2.0.2)[49]. Next, using Trimmed Mean of $M$-value normalization in edgeR package (version 4.2.2)[50], normalization factors based on gene expression are calculated and applied to normalize circRNA expression profile. Differential circRNA expression across conditions is then performed using the quasi-likelihood ratio test in edgeR. For differential splicing and circularization, CIRI3 employs algorithms from rMATS (version 4.1.2)[51].

### Benchmarking circRNA detection

Using the Hs68 cell line dataset, circRNA enrichment is determined by comparing read counts between RNase R-treated and untreated datasets of CIRI3 and five commonly used detection tools: CIRI2, CIRCexplorer3, DCC, and find_circ and KNIFE. The parameters used for each tool are provided in Supplementary File 1. A circRNA is labeled as enriched if its BSJ read count in treated samples is at least twice that in untreated samples, depleted if its count in treated samples is less than in untreated samples and unaffected otherwise, and accordingly, enriched circRNAs in RNase R-treated samples were considered putative positives, while depleted ones were treated as false positives. For each run, the start and end times were recorded to monitor the total runtime, while RAM usage was assessed by executing the 'qstat -f Job_ID' command at 10 s intervals.

### Separate-detection mode and joint-detection mode

We divided the FASTQ files of the SW480 cell line into three equal subsets. Each subset was processed individually, and results from the three subsets were merged by summing the corresponding BSJ read counts (separate-detection mode). In the joint-detection mode, CIRI3 treats the three subsets as a single unified dataset and performs joint analysis in a single run, whereas other tools process the entire SW480 dataset directly without supporting joint analysis across subsets.

### Enrichment analysis

We performed Kyoto Encyclopedia of Genes and Genomes and Gene Ontology enrichment analyses on the host genes of differentially expressed circRNAs and differentially spliced circRNAs using the R package clusterProfiler.

### Robustness evaluation

To evaluate the robustness of circRNA detection tools, RNA-seq data from the Hs68 and SW480 cell lines were used for subsampling. Reads were randomly subsampled from the untreated dataset to generate subsets at 20%, 40%, 60% and 80% of the original sequencing depth. The detection tools were applied to each subset and the number of enriched, unaffected and depleted circRNAs was determined as described above. Each subsampling was repeated three times to calculate the mean and standard deviation.

### Construction of a deep learning model for disease classification based on circRNA features

We constructed a five-layer fully connected deep neural network to classify tumor and normal samples based on circRNA BSJ ratios. All samples except those from colon tissue were randomly split into training, validation and test sets in a 3:1:1 ratio. All colon-derived samples were

held out as an independent test set to evaluate the model's generalizability to unseen tissue types.

The input features consisted of differentially spliced circRNAs that were selected according to the following criteria: (1) expressed in at least 10% of normal samples, with an average BSJ ratio in normal samples at least twofold higher than in tumor samples, (2) expressed in at least 10% of tumor samples but in less than 10% of normal samples or (3) expressed in at least 10% of disease samples with an average BSJ ratio at least 1.5-fold higher than in the corresponding normal samples. A total of 9,631 circRNAs were selected as input features. The model was trained on the training set and evaluated on the validation set, the test set and the independent colon dataset.

During transfer learning, the parameters of the first two layers of the neural network were fixed and only the last three layers were fine-tuned to improve the model's adaptability to new datasets. Specifically, colon samples from multiple studies were divided so that one study served as the test set and the rest served as the training set.

### Construction of LightGBM classifiers based on differentially spliced circRNAs

We used circRNAs as biomarkers to construct a multilevel classification system for stratifying cancer samples according to their system, tissue and disease origin. At each hierarchical level (that is, system, tissue and disease), we performed differential analysis to identify circRNAs with high specificity, which were used as candidate features for subsequent classification. Specifically, a one-versus-rest differential expression strategy was applied at each level: for each system (or tissue or disease), we compared all samples belonging to the target category against samples from all other categories at the same level and identified circRNAs that were significantly differentially spliced in each comparison. The union of all differentially spliced circRNAs was defined as the final feature set for each category. Using these features, we trained three separate LightGBM classification models for predicting system, tissue and disease labels, respectively. The input of the model consisted of the BSJ ratio values for the selected circRNAs and the output was the predicted category label (system, tissue or disease). The data were split into training and test sets, with 80% used for training and 20% for testing.

### circRNA database construction (CIRIonco)

We systematically searched and collected RNA-seq datasets from the GEO database using keywords such as 'total RNA' and 'ribo-zero', focusing on studies that employed total RNA library preparation strategies. circRNAs were identified from these datasets using the BWA aligner and the CIRI3 algorithm. Differentially spliced circRNAs were subsequently used to train a deep neural network and to construct a hierarchical stratification tree. The resulting circRNA resource was organized into CIRIonco (https://ngdc.cncb.ac.cn/cirionco), an online database system implemented using the Django framework. CIRIonco provides a global overview and visualization of circRNA profiles and supports precise querying of biomarkers that distinguish different systems, tissues and diseases. In addition, it presents detailed information on the BSJ junction ratios of each circRNA across disease and normal samples in different tissues. Metadata and accession numbers are provided in Supplementary Table 3. The database is freely accessible and no login is required.

### Simulated data

Simulated RNA-seq datasets were generated using CIRI simulator[17]. The inputs are human reference genome and gene annotations from GENCODE Release 43 (GRCh38.p13). To enable a fair comparison across methods, read length was fixed at 100 bp, while insert sizes were drawn from a mixture of two normal distributions ($N(320, 70)$ and $N(550, 70)$). Sequencing coverage for both linear transcripts and circRNAs was varied across 20, 40, 60, 80 and 100 in separate datasets. All other parameters were kept at their default settings.

### Real data

Publicly available RNA-seq datasets were downloaded from the SRA database (SRR444975, SRR445016, GSE162152 and GSE138734). These datasets contain RNase R-treated and untreated libraries. Data of Hs68 cell line (SRR444975 and SRR445016) were used to assess accuracy and sensitivity of typical circRNAs identification by each software. The accuracy of intronic self-ligated circRNAs identification by CIRI3 was evaluated using the GSE162152 dataset, which encompassed testis samples from five different species, including human, mouse, rat, rhesus and opossum. The reference genome for human (GRCh38.p13) and mouse (GRCm38.p6) were downloaded from GENCODE database. The reference genomes for rat (Rnor_6.0) and rhesus (Mmul_10) were downloaded from Ensembl. The reference genome for opossum (mMonDom1.pri) was downloaded from the NCBI Genome database. Assessment of circRNA quantification used the ribosomal R-treated SW480 cell line (SRR17235468), NCI-H23 cell line (SRR17235469) and HLF cell line (SRR17235470) RNA-seq data.

In addition, several RNA-seq datasets of human brain, testis, liver and spleen tissues were used for batch effect analysis (for accession numbers, see Supplementary Table 2) and RNA-seq datasets of tumor and adjacent normal liver samples from 44 HCC patients were used for DE analysis (GSE128274, GSE169289, GSE216613 and GSE77276). Further datasets from multiple cancer types and adjacent normal tissues were used for training the deep neural network and constructing the stratification tree (for accession numbers, see Supplementary Table 3).

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The data supporting the findings of this study, which were generated in previous studies, include the RNAAtlas dataset (GSE138734)[28], Hs68 cell line data (SRR444975, SRR445016)[25], SW480 cell line data (SRR17235468)[13], NCI-H23 cell line data (SRR17235469)[13], HLF cell line data (SRR17235470)[13], Hepatocellular Carcinoma data (GSE128274, GSE169289, GSE216613 and GSE77276)[31,37,38], and testis data from five different species (GSE162152)[27]. Additionally, several RNA-seq datasets from human brain, testis, liver and spleen tissues, along with other datasets from multiple cancer types and adjacent normal tissues, are also included (for accession numbers, see Supplementary Tables 2 and 3)[27,31–36]. All data are available from the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) and the Genome Sequence Archive in the National Genomics Data Center, China National Center for Bioinformation (https://ngdc.cncb.ac.cn/gsa/). Source data are provided with this paper.

## Code availability

CIRI3 is implemented in Java and is freely available via GitHub at https://github.com/gyjames/CIRI3. Our package includes CIRI3 tool and the example dataset, which has been extensively tested on Linux and OS X. We also provide a web interface https://ngdc.cncb.ac.cn/bit/ciri3 for users to run CIRI3 for circRNA analysis.

## References

49. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
50. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
51. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-seq data. *Proc. Natl Acad. Sci. USA* **111**, E5593–E5601 (2014).

## Acknowledgements
This work was supported by National Key R&D Program of China (grant no. 2021YFF0704500), Strategic Priority Research Program of the Chinese Academy of Sciences (XDA0460302), National Natural Science Foundation of China (grant nos. 32025009, 32270108, 32130020 and 32422020), Beijing Natural Science Foundation (grant no. JQ23025 to Y.G. and Z230007 to F.Z.) and Chinese Academy of Sciences Hundred Talents Program (to Y.G.). The authors thank Z. Zhou for the valuable suggestion and discussion.

## Author contributions
F.Z. and Y.G. conceived the project. X.Z., J.Z., L.S. and Y.G. designed the method. X.Z., J.Z., L.S. and X.L. conducted the analysis. X.Z. and X.L. wrote the manuscript with contributions from all authors.

## Competing interests
The authors declare no competing interests.

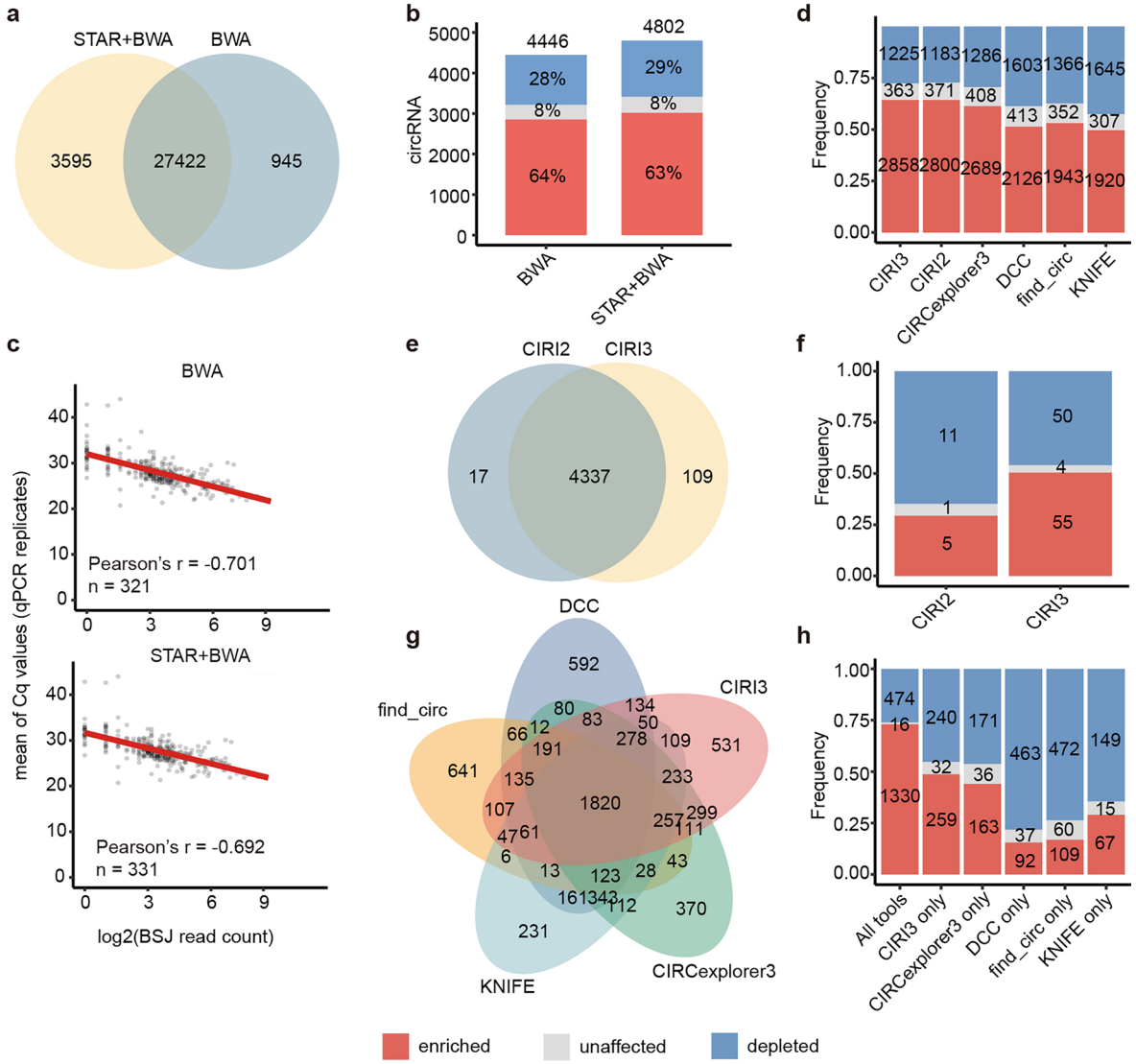
## Additional information
**Extended data** is available for this paper at https://doi.org/10.1038/s41587-025-02835-1.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41587-025-02835-1.

**Correspondence and requests for materials** should be addressed to Fangqing Zhao or Yuan Gao.

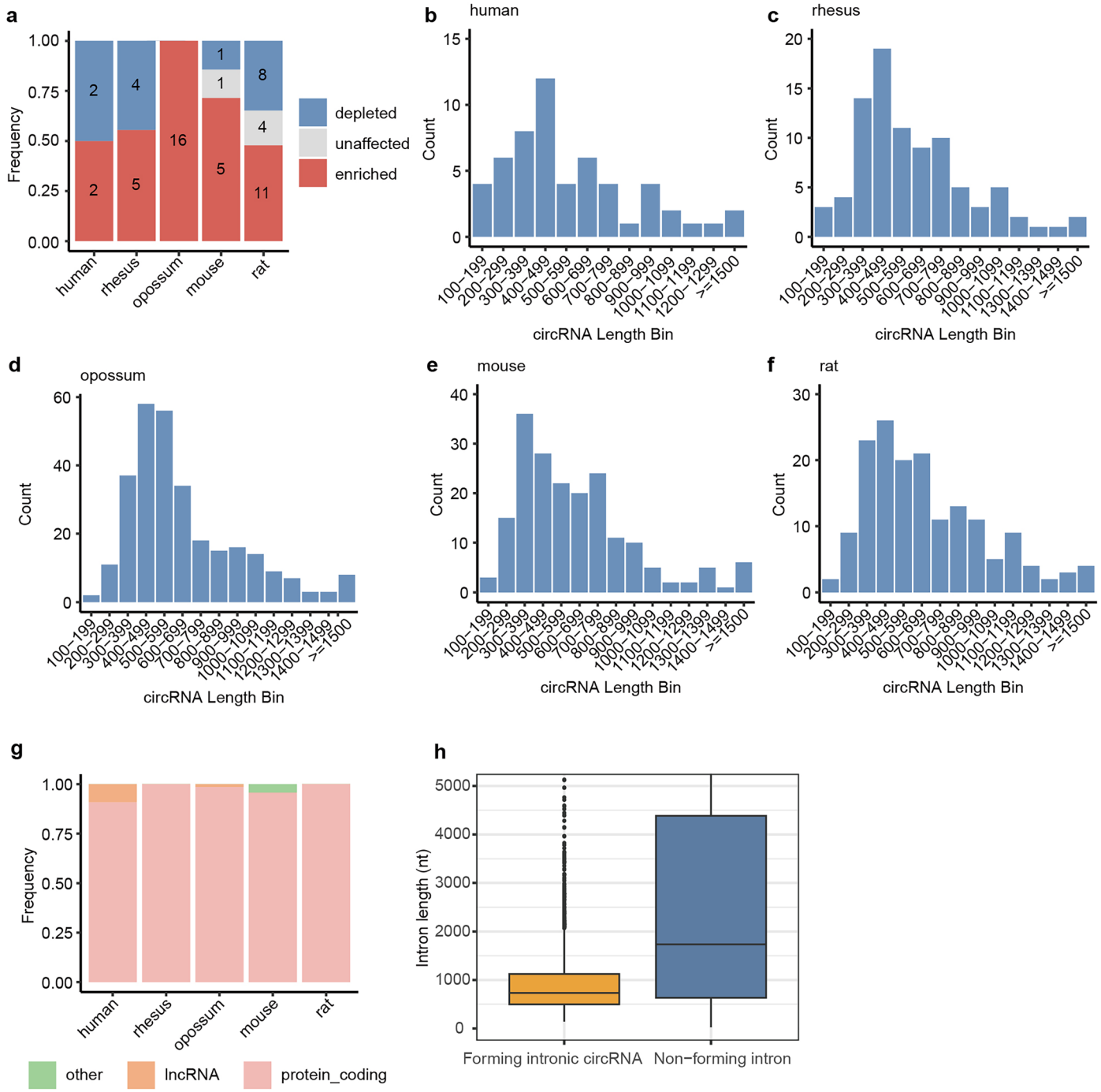
**Peer review information** *Nature Biotechnology* thanks Christoph Dieterich, Leng Han, Markus List, Xuerui Yang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41587-025-02835-1



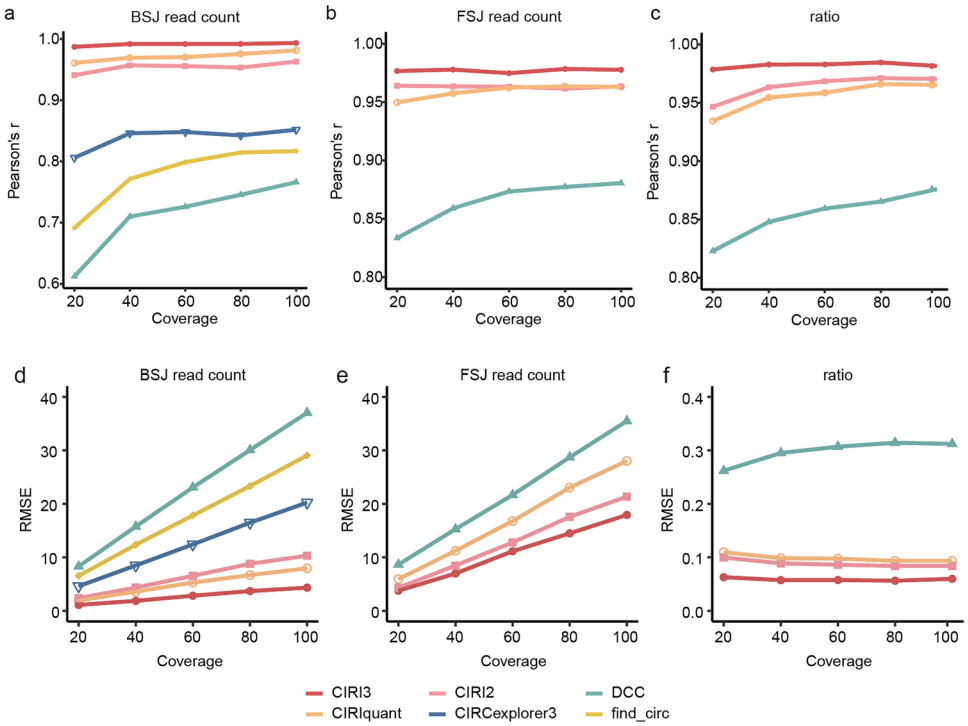

**Extended Data Fig. 1 | The performance of circRNA detection. a**, Venn diagram showing the overlap of circRNAs identified using BWA alone or the combined STAR + BWA alignment strategy. **b**, Stacked bar plot showing RNase R resistance of circRNAs detected by CIRI3 under different alignment strategies. **c**, Linear regression and Pearson correlation between log2(BSJ read count) (x-axis) and mean Cq values (y-axis) in qRT-PCR experiments for circRNAs detected using BWA (top) and STAR + BWA (bottom) strategies. **d**, Stacked bar plot showing RNase R resistance of circRNAs detected by CIRI3 and five commonly used tools. **e**, Venn diagram showing overlap of circRNAs detected by CIRI3 and CIRI2. **f**, Stacked bar plot showing RNase R resistance of circRNAs uniquely detected by CIRI3 or CIRI2. **g**, Venn diagram showing overlap of circRNAs detected by CIRI3 and four other tools. **h**, Stacked bar plot showing RNase R resistance of circRNAs detected by all tools, along with circRNAs uniquely detected by each tool.

https://doi.org/10.1038/s41587-025-02835-1



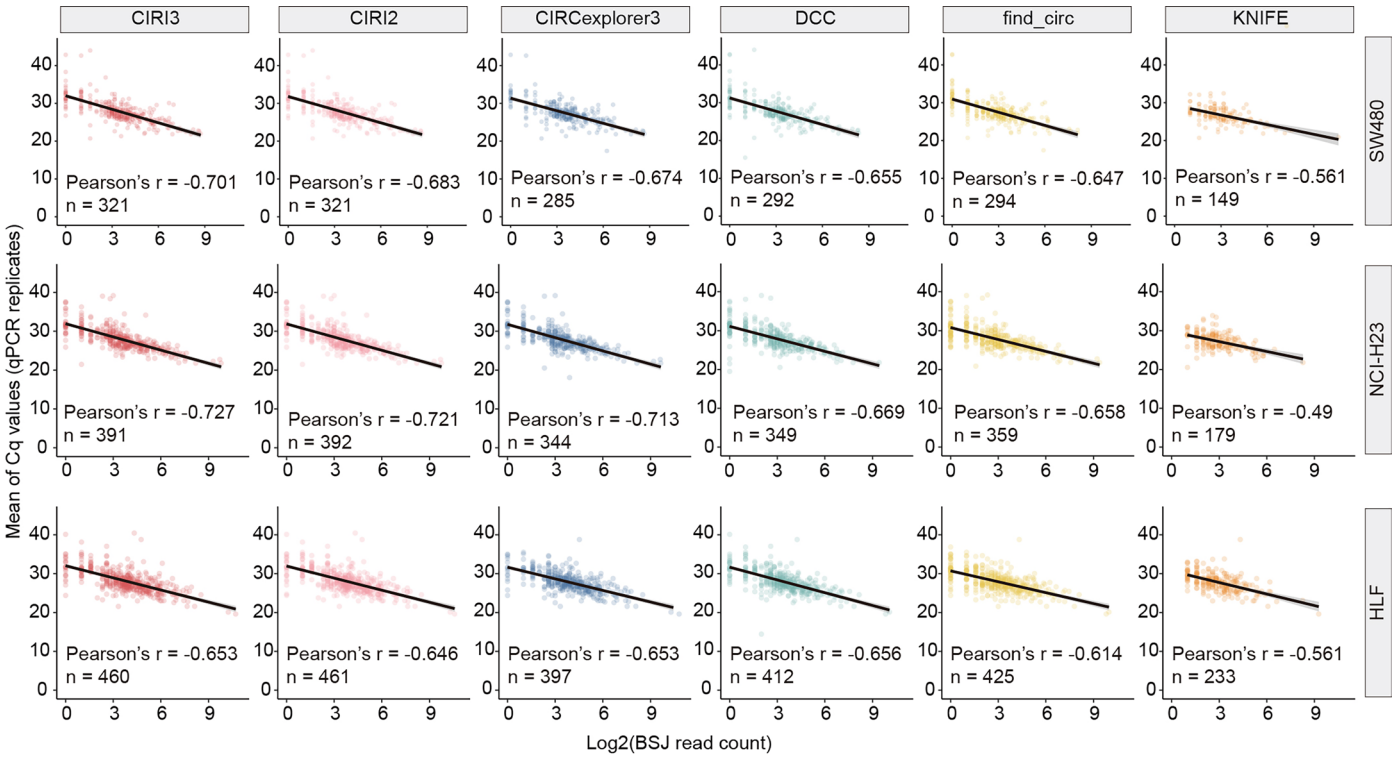

**Extended Data Fig. 2 | Performance of CIRI3 for detection of intronic self-ligated circRNAs. a**, Stacked bar plot showing RNase R resistance of intronic self-ligated circRNAs detected by CIRI3. b–f, Length distributions of self-ligated circRNAs identified from RNase R-treated RNA-seq data in liver tissues of human (**b**), rhesus (**c**), opossum (**d**), mouse (**e**), and rat (**f**). **g**, Functional classification of host genes for self-ligated circRNAs across the five species. **h**, Comparison of intron lengths between introns forming intronic self-ligated circRNAs (n = 2,286) and introns that do not (n = 35,004) within the same genes. Box plots show the median (center line), quartiles (box limits), and 1.5× IQR (whiskers). Values > 5,000 bp are excluded from the plot for visualization, but all data were included in the analysis. Statistical significance was assessed using a two-sided Wilcoxon rank-sum test ($P < 2.2 \times 10^{-16}$).

https://doi.org/10.1038/s41587-025-02835-1



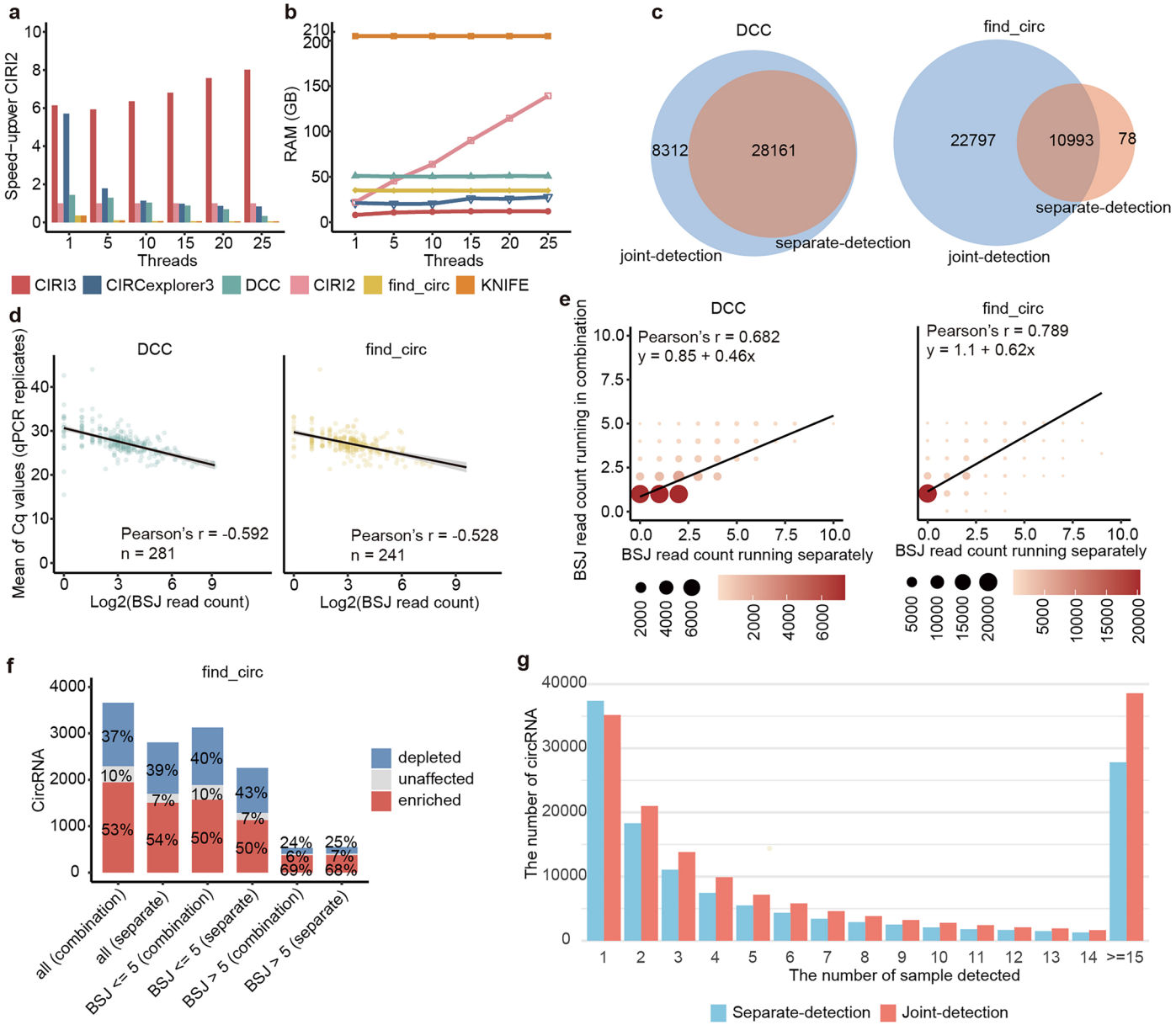

**Extended Data Fig. 3 | Performance of circRNA quantification for CIRI3 and other tools on simulated data.** Pearson correlation coefficients between BSJ read counts, FSJ read counts, or junction ratios of circRNAs identified by CIRI3 and other tools and the ground truth across simulated datasets with varying coverage. **a**–**c**, Pearson correlations for BSJ read counts (**a**), FSJ read counts (**b**), and junction ratios (**c**). r.m.s.e. between BSJ read counts, FSJ read counts, or junction ratios of circRNAs identified by CIRI3 and other tools versus the ground truth. **d**–**f**, The r.m.s.e. for BSJ read counts (**d**), FSJ read counts (**e**), and junction ratios (**f**).

https://doi.org/10.1038/s41587-025-02835-1



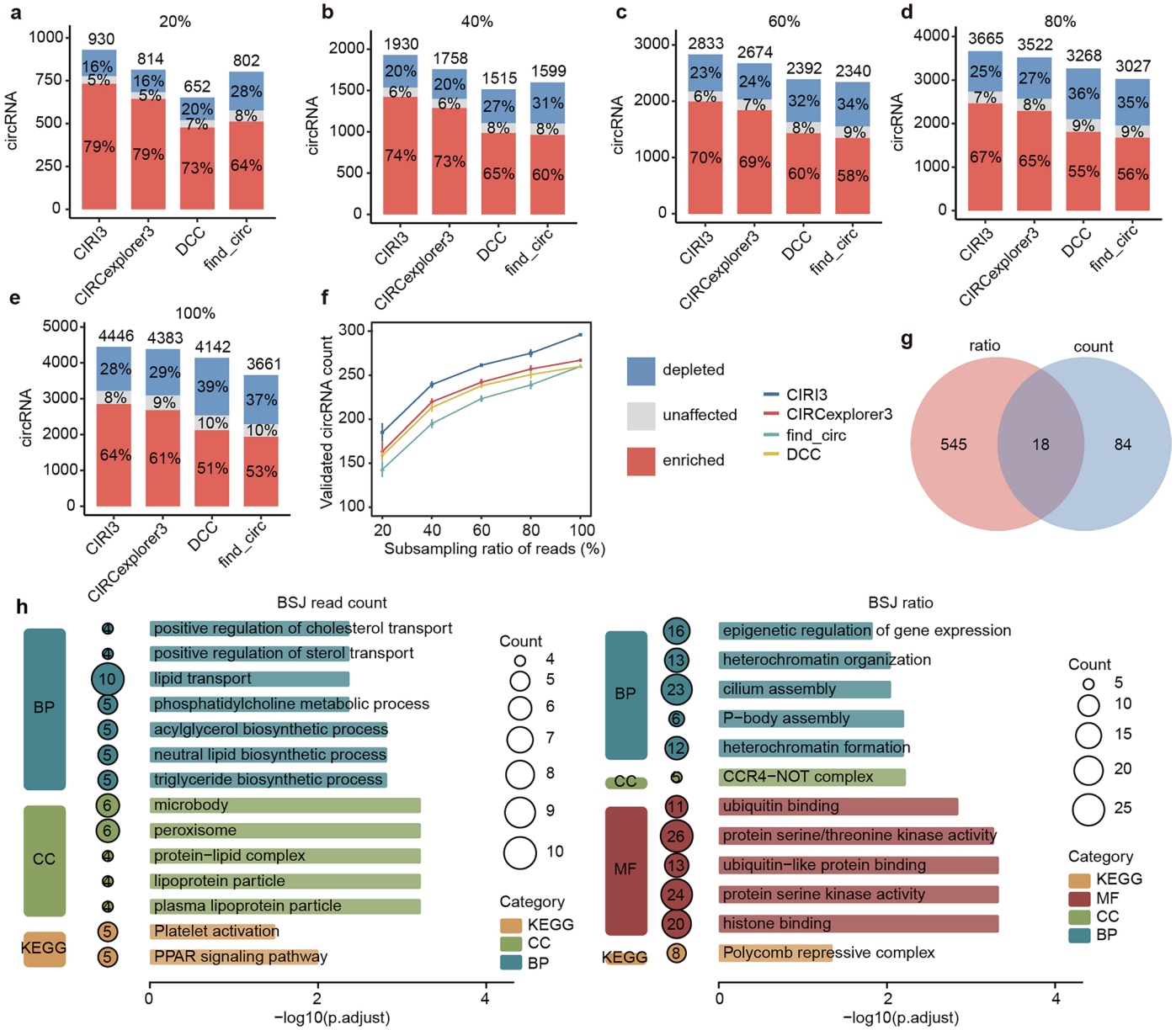

**Extended Data Fig. 4 | Evaluation of circRNA quantification accuracy.** Linear regression and Pearson correlation between log2(BSJ read counts) (x-axis) and mean Cq values (y-axis) from qRT-PCR experiments in SW480, NCI-H23, and HLF cell lines. n indicates the number of circRNAs identified by each tool and validated in the corresponding cell line.

https://doi.org/10.1038/s41587-025-02835-1



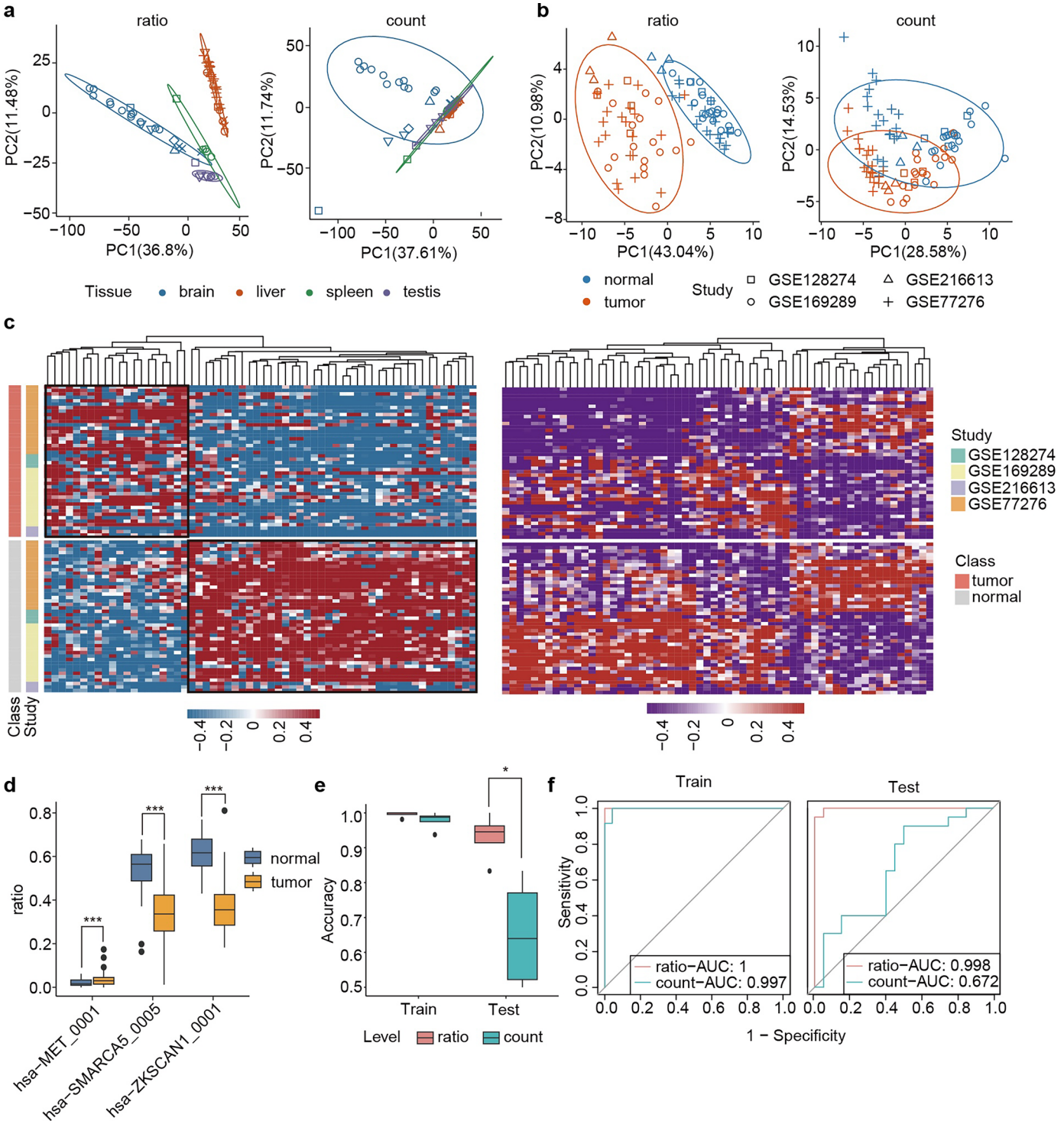

**Extended Data Fig. 5 | Benchmarking and comparative analysis of circRNA detection tools. a**, Runtime of six tools on the SW480 dataset using 1, 5, 10, 15, 20, and 25 threads. KNIFE used the default number of threads. The y-axis shows the speedup of each tool relative to CIRI2 in circRNA identification. **b**, Memory usage of the same tools for the corresponding thread settings in a. **c**, Venn diagram showing circRNAs identified by DCC (find_circ) in separate-detection and joint-detection modes. **d**, Linear regression and Pearson correlation between log2(BSJ read counts) (x-axis) and Cq values (y-axis) from qRT-PCR experiments on the SW480 dataset using DCC (find_circ) in separate-detection mode. **e**, Scatter plot of BSJ read counts ( ≤ 5) for circRNAs identified by DCC (find_circ) in joint-detection mode versus the same circRNAs in separate-detection mode. **f**, Stacked bar plot showing RNase R resistance of circRNAs detected by find_circ in separate-detection and joint-detection modes. **g**, Distribution of circRNAs identified by CIRI3 in both separate-detection and joint-detection modes.

https://doi.org/10.1038/s41587-025-02835-1



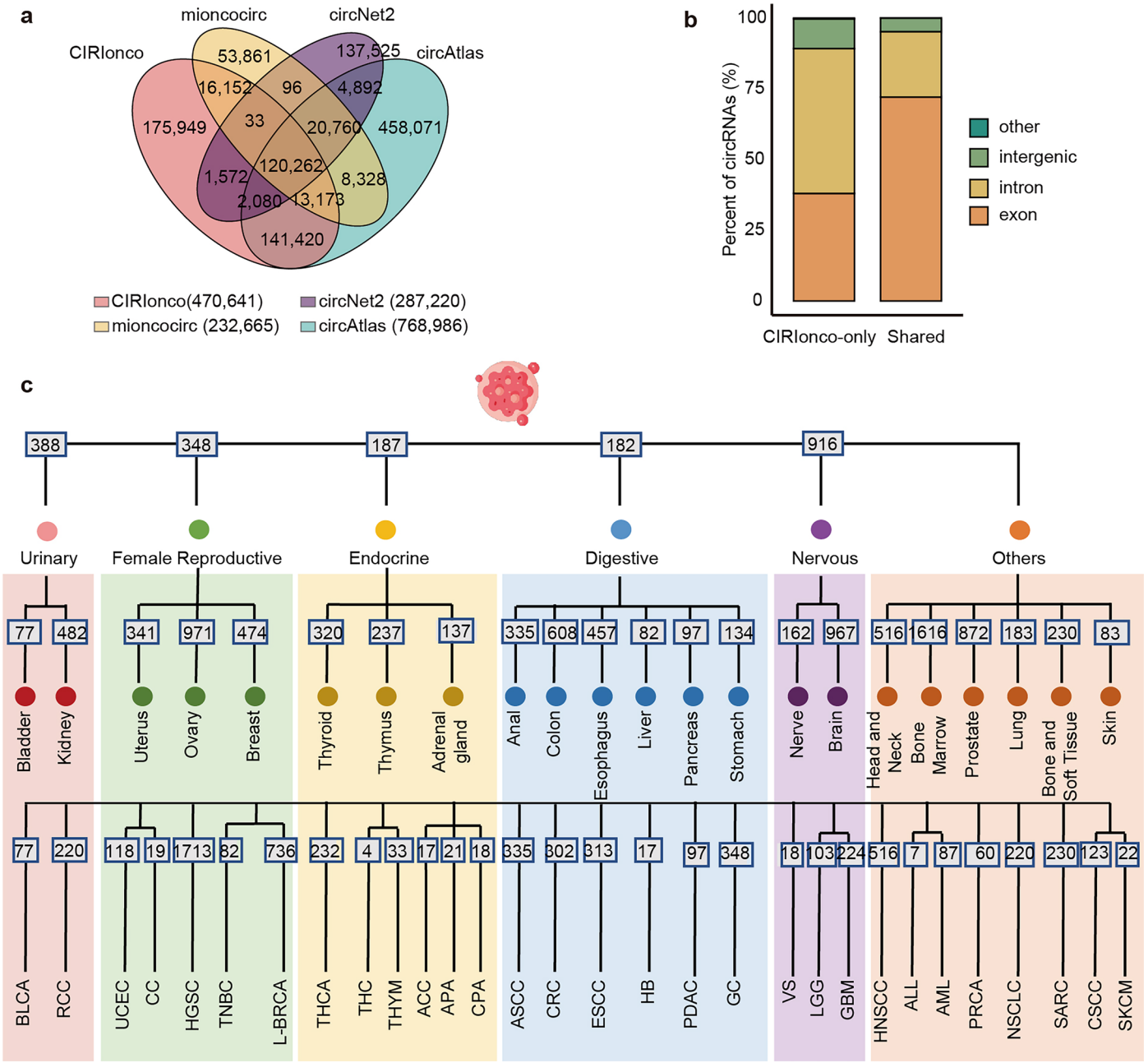

**Extended Data Fig. 6 | Subsampling analysis and functional characterization of circRNAs. a**–**e**, Stacked bar plots showing RNase R resistance of circRNAs detected by each tool in subsampled datasets at different sampling levels: 20% (a), 40% (b), 60% (c), 80% (d), and 100% (e). **f**, Number of RT-PCR-validated circRNAs detected by each tool in subsampled SW480 datasets of varying sizes. Points indicate mean ± SD of three independent replicates per subsampling level. **g**, Venn diagram showing overlap between circRNAs with differential BSJ read counts and differential junction ratios. **h**, KEGG and GO enrichment analyses of host genes associated with circRNAs. Left: circRNAs identified by differential BSJ read counts; right: circRNAs identified by differential junction ratios.

https://doi.org/10.1038/s41587-025-02835-1




**Extended Data Fig. 7 | Evaluation of batch effects on BSJ read counts and junction ratios of circRNAs. a**, Principal component analysis (PCA) based on circRNA junction ratios (left) and BSJ read counts (right) across four tissue types from multiple projects. CircRNAs expressed in ≥50% of samples were retained. Point shapes indicate project origin. **b**, PCA of differential junction ratios (left) and differential BSJ read counts (right) of circRNAs. **c**, Heatmaps of the top 20 upregulated and top 40 downregulated circRNAs based on junction ratios (left) and BSJ read counts (right). **d**, BSJ ratios for hsa-MET_0001, hsa-SMARCA5_0005, and hsa-ZKSCAN1_0001 in paired normal (n = 44) and tumor (n = 44) samples. Box plots show the median (center line), quartiles (box limits), and 1.5× IQR (whiskers). Statistical significance was assessed using a two-sided test implemented in rMATS, with "***" indicating $P < 1 \times 10^{-16}$. **e**, Accuracy of SVM models trained on differential circRNAs, based on BSJ ratios or read counts, across four datasets (GSE128274, GSE169289, GSE216613, and GSE77276). In each round, three datasets were used for training and one for testing. Box plots show the median (center line), quartiles (Q1 and Q3), and 1.5× IQR (whiskers). Each point represents the accuracy in a given training/testing round (n = 4). Statistical significance was assessed using a two-sided Wilcoxon rank-sum test, with "*" indicating P = 0.042. **f**, ROC curves of SVM models trained on BSJ ratios or read counts, using GSE77276 as the test set.

https://doi.org/10.1038/s41587-025-02835-1




**Extended Data Fig. 8 | Characterization of circRNAs in the CIRIonco database.** **a**, Venn diagram showing overlap of circRNAs between CIRIonco and other human cancer circRNA databases. **b**, Stacked bar plot showing the composition of circRNA types in CIRIonco-specific circRNAs and those shared with other databases. **c**, Hierarchical classification tree based on system, tissue, and disease levels, highlighting the number of circRNA markers.

nature portfolio

Corresponding author(s): Fangqing Zhao, Yuan Gao

Last updated by author(s): Jul 17, 2025

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ A description of all covariates tested |
| ☒ | ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | Use sratoolkit (version 3.0.1) to download RNA-seq data from the SRA database (https://www.ncbi.nlm.nih.gov/sra/?term=) and use Edge Turbo to download RNA-seq data from GSA (https://ngdc.cncb.ac.cn/gsa/). |
| Data analysis | RNA-seq data is converted from SRA format to fastq format using sratoolkit (version 3.0.1). The following tools are then used to identify circRNA from the RNA-seq data: CIRI3 (https://github.com/gyjames/CIRI3), CIRI2 (version 2.0.6) (https://ciri-cookbook.readthedocs.io/en/latest/CIRI2.html), find_circ (version 1.2) (https://github.com/marvin-jens/find_circ), CIRIquant (version 1.1) (https://github.com/bioinfo-biols/CIRIquant), DCC (version 0.5.0) (https://github.com/dieterich-lab/DCC), KNIFE (version 1.5) (https://github.com/lindaszabo/KNIFE), and CIRCexplorer3 (version 1.0.1) (https://github.com/YangLab/CLEAR). Gene expression profile is obtained using featureCounts (version 2.0.2). Differential expression analysis of identified circRNAs is performed using the built-in differential expression script of CIRI3, which calls parts of the rMATS (version 4.1.2) (https://github.com/Xinglab/rmats-turbo) and edgeR package (version 4.2.2) code. Data analysis was performed using R (version 4.4.0), utilizing the stats package (version 4.4.0) for principal component analysis (PCA), the e1071 package (version 1.7.16) for support vector machine (SVM) analysis, the clusterProfiler package (version 4.12.6) for enrichment analysis, and the ggplot2 (version 3.5.1) and pheatmap (version 1.0.12) packages for figure visualization. In addition, Python's (version 3.11.5) torch package (version 2.1.0) was used to build fully connected neural networks, and the lightgbm package (version 4.6.0) was employed to construct LightGBM models. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The RNA-seq data analyzed in this study are publicly available in the SRA database under the accession numbers SRR444975, SRR445016, SRR17235468, SRR17235469, SRR17235470, SRR1723546, SRP294141, SRP069760, SRP013565, SRP150447, SRP019807, PRJNA787399, PRJNA1066280, PRJNA1018507, PRJNA1180454, PRJNA875895, PRJNA268523, PRJNA528269, PRJNA806896, PRJNA1149753, PRJNA268523, PRJNA528269, PRJNA683653, PRJNA757576, PRJNA510124, PRJNA613939, PRJNA689295, PRJNA806896, PRJNA1095632, PRJNA486023, PRJNA875895, PRJNA1047096, PRJNA1089512, PRJNA689313, PRJNA891435, PRJNA913584, PRJNA533799, PRJNA727315, PRJNA844356, PRJNA704974, PRJNA757576, PRJNA635121, PRJNA821888, PRJNA875895, PRJNA913584, PRJNA996979, PRJEB29932, PRJNA786565, PRJNA875895, PRJNA996979, PRJNA875895, PRJNA1150102, PRJNA551007, PRJNA396544, PRJNA551007, PRJNA789867, PRJNA850175, PRJNA996979, PRJNA751379, PRJNA552058, PRJNA997353, PRJNA551007, PRJNA875895, PRJNA996979, PRJNA520916, PRJNA579999, PRJNA438844, PRJNA789867, PRJNA792194, PRJNA850175, PRJNA996979, PRJNA737121, PRJNA523137, PRJNA695543, PRJNA857849, PRJNA704389, PRJNA723838, PRJNA754677, PRJNA849660, and PRJNA895900. The data can be accessed at https://www.ncbi.nlm.nih.gov/sra.The RNA-seq data are also publicly available in the GEO database under the accession numbers GSE138734, GSE128274, GSE169289, GSE216613, GSE77276, and GSE162152. The data can be accessed at https://www.ncbi.nlm.nih.gov/gds/?term=.Additionally, the RNA-seq data are publicly available in the GSA database under the accession number PRJCA000751. The data can be accessed at https://ngdc.cncb.ac.cn/gsa/.

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | NA |
| Reporting on race, ethnicity, or other socially relevant groupings | NA |
| Population characteristics | NA |
| Recruitment | NA |
| Ethics oversight | NA |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | No additional methods were used to pre-determine the sample size; the sample size depends on the availability of public data. For each analysis, the sample size was sufficient to obtain statistically significant results. |
| Data exclusions | In the analysis of tissue samples, we only selected data from total RNA library preparation and did not use RNase R-enriched tissue data. |
| Replication | We used differentially expressed circRNAs from Hepatocellular Carcinoma (HCC) samples (four datasets) to train a support vector machine (SVM) model to distinguish tumor samples from normal samples. Three datasets were used as the training set, and each dataset was used once as the test set in a rotating manner. This process was repeated four times to evaluate the performance of models constructed based on BSJ read counts and BSJ junction ratios. To assess the robustness of circRNA detection tools, we performed subsampling analyses using RNA-seq data from the Hs68 and SW480 cell lines. Reads were randomly sampled from the datasets to generate subsets representing 20%, 40%, 60%, and 80% of the original sequencing depth. Each subsampling was repeated three times to calculate the mean and standard deviation. |
| Randomization | When the data, environment, and parameters are consistent, the results from circRNA identification and differential expression analysis are consistent. |

| | |
|---|---|
| Blinding | Since this study uses publicly available datasets, the researchers were unable to implement blinding for group allocation. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Plants

| | |
|---|---|
| Seed stocks | NA |
| Novel plant genotypes | NA |
| Authentication | NA |