CYCLeR– a novel tool for the full isoform assembly and quantification of circRNAs

Stefan R. Stefanov & Irmtraud M. Meyer

27. September 2022

1 Simulation of RNA-seq data

1.1 Simulated transcript generation

At the moment thse is still no comprehensive gold-standard dataset of *circRNA* isoforms containing complete alternative splicing (*AS*) information available. This forces us to artificially design a set of *circRNA* sequences. One key design goal is to capture key features of real data. To this end, we use publicly available *D.* melanogaster head data with available RNase R treatment to select real BSJ sites and *circRNA*-specific genomic features. This allows us to detect novel genomic features (exons, junctions, retained introns) and to incorporate them into our dataset. The necessary features that a simulation set should have are shown in Figure S1.
One key goal is to design a dataset with the highest possible complexity in terms of transcript reconstruction in order to be able to test the capabilities of the different tools to discover unannotated genomic features. To specifically test our tool, we also require multiple overlapping linear isoforms as additional challenge. For this, we not only need to simulate multiple overlapping circular isoforms with comparable quantities. Table S1 summarises the design goals of our simulated transcript set.
We map *D.* melanogaster adult head RNA-seq data with STAR (1) for novel junction detection and process the junction boundaries into exons with the *SGSeq* R (2) package (based on the commonly used *GenomicAlignments* (3) package). We select exons within BSJ loci, based on a union of the BSJs predicted by *CIRCexplorer2* (4) and *CIRI2* (5). We filter the exons to keep only features enriched after RNase R treatment. We also filter the BSJ of monoexonic or diexonic circRNAs as they pose no problem for assembly or quantification,as they inflate the benchmark statistics with true positives without much challenge to the algorithms. For every set of edge exons, internal exons are selected by Bernoulli trial with success 0.75. Every selection is repeated (number of internal exons)/3 times. Only unique transcripts are selected for simulations. We select all annotated linear transcripts originating from the same genes as the circRNAs. Many analyses require normalization of the feature abundances to facilitate statistical testing. To assess the normalization of feature abundances, we need to simulate full RNA-seq libraries. We pad the libraries with additional 10000 randomly selected protein coding transcripts and 5000 randomly selected non-coding transcripts. The protein coding transcripts serve as a placeholder for all the transcripts that are depleted by circRNA enrichment procedures, while the non-coding transcripts represent the linear transcripts increased as a byproduct of circRNA enrichment. Our library size is based on a common sequencing library depth of 25 million reads.
We select transcript quantities by random sampling within set ranges. Transcripts containing unannotated exons and retained introns are simulated at lower quantities. We selected the ranges empirically to accommodate a requirement of 0.1% BSJ spanning reads from all the reads of the simulated library. The number of simulated reads per transcript is calculated based on the length of the transcript multiplied by a fixed factor. For linear transcripts, this factor value is randomly sampled in the range of 10-40, while for circular in the range of 8-20. For nascent RNA

simulations, the factor is in range of 1-1.5.

$$\text{Number of simulated reads per transcript} = \frac{(\text{length of transcript}) \cdot \text{factor}}{50}$$

In the simulation of total ribo-depleted RNA-seq, the simulated circular transcript reads account for ∼0.1% of the simulated linear reads.

Linear transcript derived reads are simulated with polyester (6), while circular transcript derived reads are simulated with polyestercirc (see polyestercirc section).

1.2 Reference sets

Recent studies (7; 8) have shed light on the complexity of the circRNA *AS*. We design the complexity of our datasets based on the findings of those studies. Therefore, our most common *AS* events is alternative circulazation (alternative BSJ in the same locus), followed by exon skipping and alternative 5'/3'-splicing, and, at even lower rate, intron retention. We separated the simulated transcripts into two reference data sets, Figure S2. One that involves extreme cases hindering reconstruction constituting our *high complexity* reference set. The second set involves simpler cases - with less than one *AS* event coming from a BSJ locus, constituting our reference set. This allows us to test how linear RNA *AS* events affect the detection of circular *AS* events. We provide a table summarsing the characteristics of the two reference sets in Table S2.

1.3 Polyestercirc

As a benchmark, an RNA-seq library containing linear and circular transcripts with multiple isoforms with overlapping sequences needs to be simulated. The available RNA-seq simulation tools, used in previous studies (9; 10), are not realistic representations of the data. The common RNA-seq library simulation tools are not designed to handle circular transcripts (6; 11; 12). This leads to their inability to generate fragments spanning the BSJ. Some RNA-seq simulation programs were designed with the purpose of simulating circular reads (9), but the algorithms do not provide a proper representation of the RNA-seq data, due to unrealistic edge effect, sequencing errors and GC-bias. For that simulation, the "sequence bias" is induced by a random selection of read start sites, which does not represent a realistic RNA-seq scenario.
The simulation done in (10) utilises Polyester by providing pseudo-linear transcripts extended by the length of a read minus one nucleotide (L-1). This approach, however, does not account for circular fragments that span the BSJ with start coordinates with a distance longer than a read length from the splice site.

Polyester is an RNA-seq simulator that fits the input transcript expression levels into negative binomial model and is able to generate realistic replicates (6). We have modified the original code to make Polyester capable of generating reads from circular transcripts. The simulation parameters correspond to the library parameters discussed in the previous section.
The modification of the code of Polyester is made with the assumption that during RNA fragmentation, *circRNA*s will hydrolyse (i.e. break) at least once. The start site of the fragmentation is chosen from a uniform distribution along the sequence of the circle. We disabled the edge effect model for the simulation of circular reads, because it does not correspond to the topology of a circle. The suggested pipeline for using polyester is applying GC-bias by increasing the amount of transcripts being simulated based on their GC-content. For a more accurate GC-bias modelling, we have switched to sampling transcripts as Bernoulli trials with probability based on GC-content. The circular and linear transcripts are simulated separately and are subsequently merged into one library. The linear transcript sequences are based on known annotation. To simulate intronic noise, we added the full sequence of the gene as part of the reference for the linear transcript simulation. CircRNA enrichment effect is simulated by five-fold decrease in the linear transcript abundance and a 4.5 times increase in circular transcript abundance. The resulting simulated library successfully mirrors the characteristic features of real RNA-seq data, see Figure S1.

2 *CircRNA* transcript reconstruction

As a prerequisite of the transcript assembly (discussed in the main part of the manuscript), splice graph creation is required. Figure S3 shows the splice graph construction process of the 5-HT2A gene, based on simulated data. After the STAR (1) mapping, the reads mapped to *circRNA*s are extracted with SAMtools(13) to eliminate unnecessary reads slowing down the computation. Note: that the data still contains reads outside the boundaries of the *circRNA*s. Feature selection is performed based with the SGSeq R package(2) and the feature information is stored in SGRanges object part of the same package. SGRanges functions provide easy fix of the artefacts of the feature selection and allow for a smooth transition to the next stesp of the process.
After the artefact correction, a recount of the reads per features is required. This is performed with RSubread package (14). The featured depleted after circRNA-enrichment are identified using the statistical test of the DEXSeq R package (15). The standard workflow of DEXSeq is modified to circumvent the normalization step. RPKM is used in all calculations.
The final step of the *CYCLeR*workflow is a step-wise reconstruction of possible transcripts using a tailor-made graph algorithm (see main Figure 2).

3 *CircRNA* transcript quantification

The available reads of a library need to be collapsed into a single value per transcript (11).The EM based methods have shown significantly better accuracy than the alternative (16). An EM algorithm can be performed using short sequences of length k (k-mers), fitting into an *a priori* created de Bruijn graph (16). The latest k-mer based quantification tools employ pseudo-alignments to speed up and increase the accuracy of the counting (16; 17). In order to adjust the algorithm to work properly for circRNAs, we increase the length of the reference transcripts to cover the extra alignment possibilities around the BSJ, see Figure S5. That adjustment extends the effective length of the transcript by the size of a k-mer minus one base. The resulting pseudo-linear sequence then needs to be filled with pseudo-random (not similar to the reference sequences) nucleotides to account for the increase in the effective length of a circular as opposed to a linear RNA.

References

1. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics,* **29**(1), 15–21.

2. Goldstein, L. D., Cao, Y., Pau, G., Lawrence, M., Wu, T. D., Seshagiri, S. and Gentleman, R. (2016) Prediction and quantification of splice events from RNA-seq data. *PLoS ONE,* **11**(5), 1–18.

3. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T. and Carey, V. J. (2013) Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.,* **9**(8), 1–10.

4. Zhang, X.-o., Dong, R., Zhang, Y., Zhang, J.-l., Luo, Z., Zhang, J., Chen, L.-l. and Yang, L. (2016) Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.,* **26**(9), 1277–1287.

5. Gao, Y., Zhang, J. and Zhao, F. (2018) Circular RNA identification based on multiple seed matching. *Brief. Bioinform.,* **19**(5), 803–810.

6. Frazee, A. C., Jaffe, A. E., Langmead, B. and Leek, J. T. (2015) Polyester: Simulating RNA-seq datasets with differential transcript expression. *Bioinformatics,* **31**(17), 2778–2784.

7. Xin,R., Gao,Y., Gao,Y., Wang,R., Kadash-Edmondson,K.E., Liu,B., Wang,Y., Lin,L. and Xing,Y. (2021) isoCirc catalogs full-length circular RNA isoforms in human transcriptomes. *Nat. Commun.,* **12**, 1–11.

8. Zhang,J., Hou,L., Zuo,Z., Ji,P., Zhang,X., Xue,Y. and Zhao,F. (2021) Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long. *Nat. Biotechnol.,* **39**(7), 836–845.

9. Gao, Y., Wang, J., Zheng, Y., Zhang, J., Chen, S. and Zhao, F. (2016) Comprehensive identification of internal structure and alternative splicing events in circular RNAs. *Nat. Commun.,* **7**(May), 1–13.

10. Li, M., Xie, X., Zhou, J., Sheng, M., Yin, X., Ko, E.-a., Zhou, T. and Gu, W. (March, 2017) Quantifying circular RNA expression from RNA-seq data using model-based framework. *Bioinformatics,***33**(14), 2131–2139.

11. Li, B. and Dewey, C. N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics,* **12**(1), 323.

12. Griebel,T., Zache,B., Ribeca,P., Raineri,E., Lacroix,V., Guigó,R. and Sammeth,M. (2012) Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.,* **40**(20), 10073–10083.

13. Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. and Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience,* **10**(2), 1–4.

14. Liao, Y., Smyth, G. K. and Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.,* **47**(8), e47

15. Anders, S., Reyes, A. and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data.*Genome res.,*, **22**(10), 2008–2017

16. 000 Bray, N. L., Pimentel, H., Melsted, P. and Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.,* **34**(5), 525–527.

17. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. and Kingsford, C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods,* **14**(4), 417–419.

18. Patro, R., Mount, S. M. and Kingsford, C. (2013) Sailfish: Alignment-free Isoform Quantification from RNA-seq Reads using Lightweight Algorithms. *Nat. Biotechnol.,* **32**(5), 462–464.

19. Zhang,J. and Zhao,F. (2021) Reconstruction of circular RNAs using Illumina and Nanopore RNA-seq datasets *Methods,* **196**, 17–22.

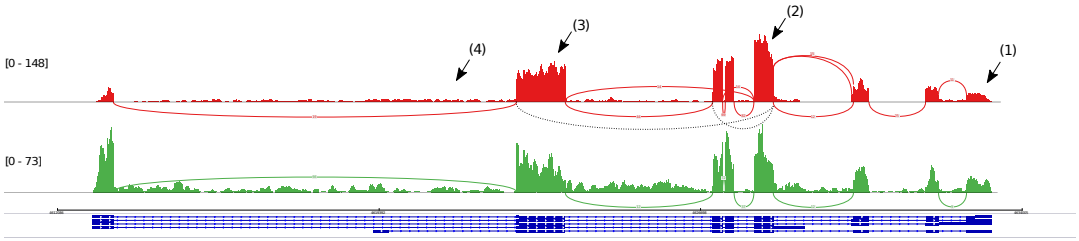

Figure S1: **Simulation of RNA-seq data-locus of the 5-HT2A gene.** Sashimi plot comparing simulated (red) versus real (green) RNA-seq data. BSJs are marked with dotted line. The necessary behaviour to consider a simulated sample "realistic": (1) Decrease in coverage around the start/end of linear transcripts, (2) relative decrease in coverage around back-splicing sites, (3) CG-bias in coverage and (4) intronic "noise" caused by unspliced transcripts.

Reconstruction problem	Dataset design
Identifying *circRNA* exons	Selected exons after *circRNA* enrichment
Identifying un-annotated exons	Integrated novel SJ from STAR output
Overlapping linear *AS*	Included overlapping linear transcripts
Overlapping *circRNA AS*	Included overlapping circular transcripts
Nascent RNA noise	Included full gene sequence

Table S1: **Benchmarking set design goals.** The benchmarking dataset is designed to test the capability of different *circRNA* transcript reconstruction tools to deal with common problems for *circRNA* reconstruction-summarized in the table

Dataset	Reference set	High complexity set
Unannotated exons	Yes	Yes
Retained introns	Yes	Yes
Overlapping linear RNA *AS* events	Yes	Yes
# of overlapping *circRNA AS* events	≤ 1	All

Table S2: **Reference vs High complexity sets** This table specifies the differences between the *High complexity set* and the *Reference set*. Both datasets allow us to assess the evaluation of the effect of linear splicing on the *circRNA* assembly. The high complexity set also enables us to evaluate the effect of multiple overlapping *circRNA* s on *circRNA* assembly.

	Rep1 control	Rep2 control	Rep1 treated	Rep2 treated
2x75	26,964,485	26,937,052	25,760,332	25,752,251
2x250	27,011,990	27,005,653	25,863,703	25,855,256

Table S3: **Information on simulated libraries type and depth.** To accommodate the requirements of all tools, libraries were simulated in: 1) replicates; 2) pair-end; 3) two types of read length 4) 5 time decrease in the linear transcript abundance and 4.5 times increase in circular transcript abundance (based on empirical observations)

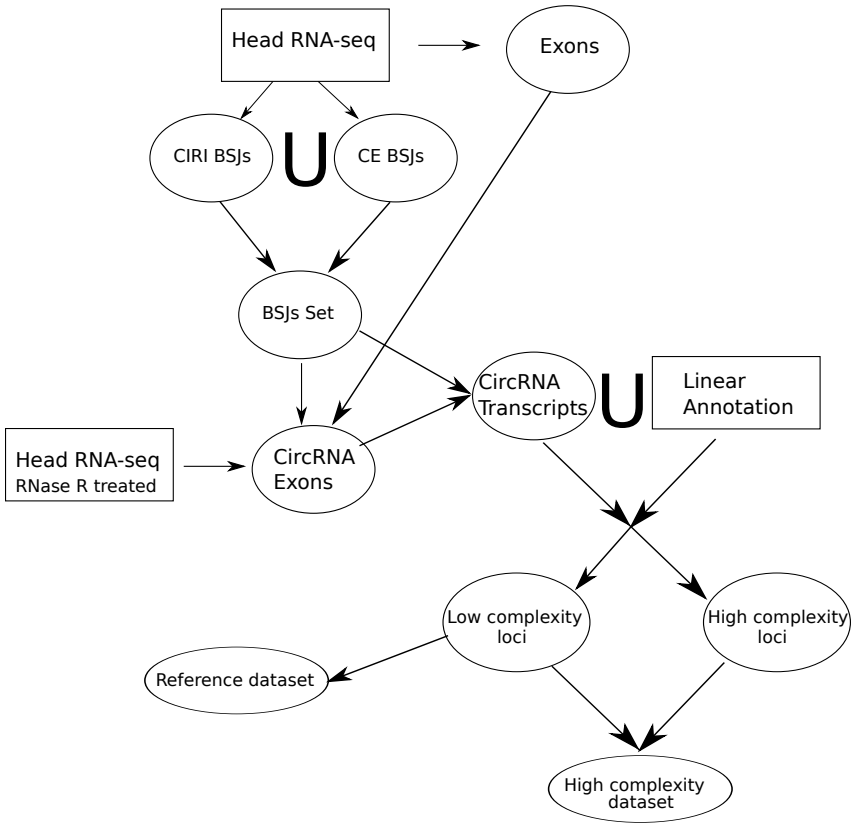

Figure S2: **Selection of reference sets.** The simulated dataset contains cases that present common challenges for circRNA assembly that we know to be present in real data (see section "Reference sets") as well as even more challenging cases. We select a set of lower complexity cases to better highlight the differences in performance of the tools and employ the higher complexity set to investigate the limits of transcriptome assembly based on RNA-seq data. Our *reference set* contains loci with a low number of overlapping *AS* events. Our *high complexity set* expands the *reference set* by combining it with *circRNA* transcripts with multiple overlapping *AS* events.

Organism	Reference genome	Annotation	Source
Fruit fly	BDGP6 (dm6)	BDGP6.87	Ensembl Top level Assembly
Mouse	GRCm39 (mm39)	GRCm39.104	Ensembl Primary Assembly
Human	GRCh38 (hs38)	GRCh38.101	Ensembl Top level Assembly

Table S4: **Summary of reference genome and annotation versions.** The source of all files is the Ensembl FTP server. *Note:* for *CYCLeR* feature annotation, the corresponding R TxDb package was used.

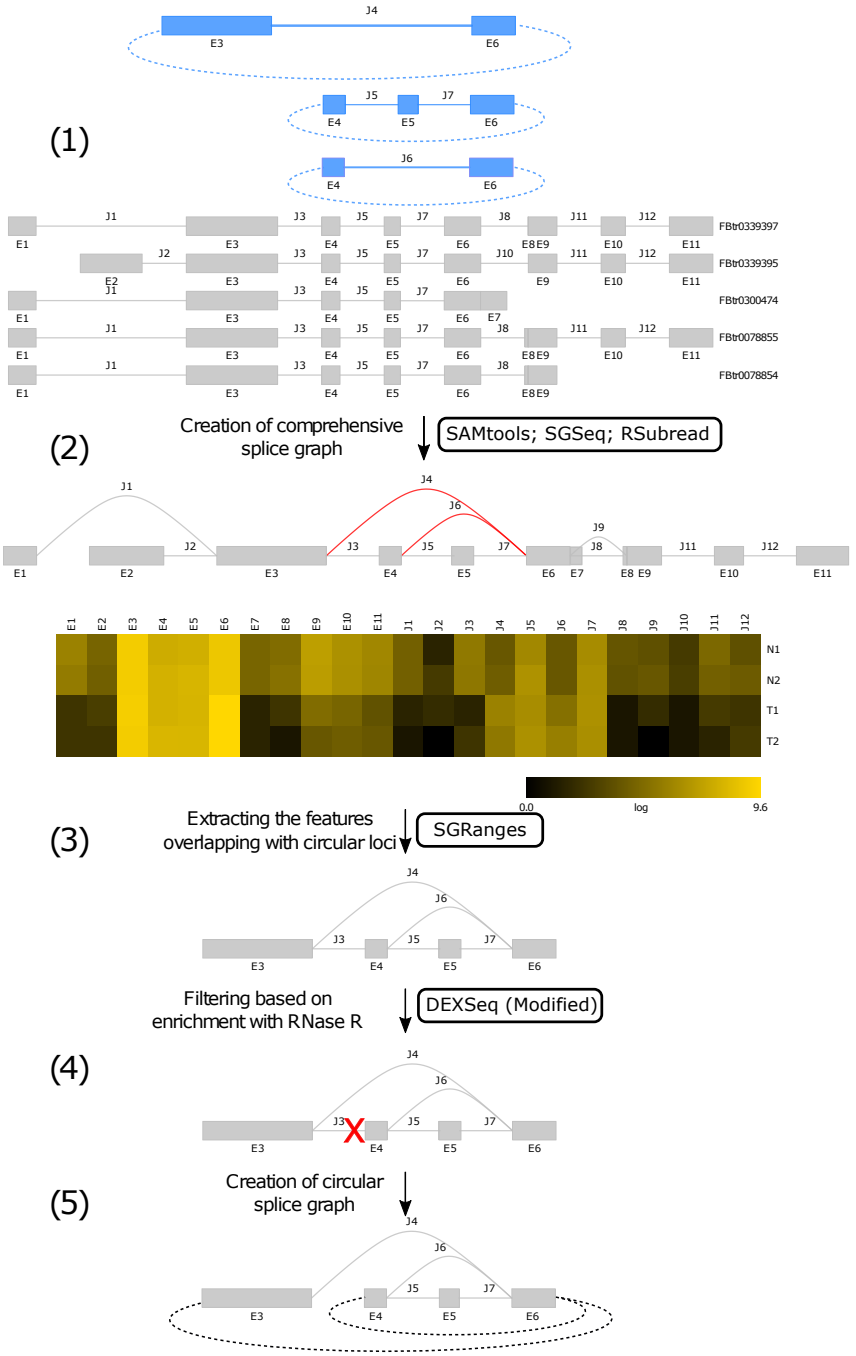

Figure S3: **Circle-specific splice graph construction of the 5-HT2A gene.** (1) Reference linear and circular transcripts are used for sample simulation, (2) comprehensive splice graph is created using SGSeq, the features for 4 samples are quantified and can be represented in a heatmap(N (non-treated), T (treated); the features that have no previous annotation are marked red), (3) a subgraph with only features located between the BSJ start/end boundaries is extracted, (4) filtering based on depletion of features in circRNA enriched samples (T) (5) Final splice graph creation.

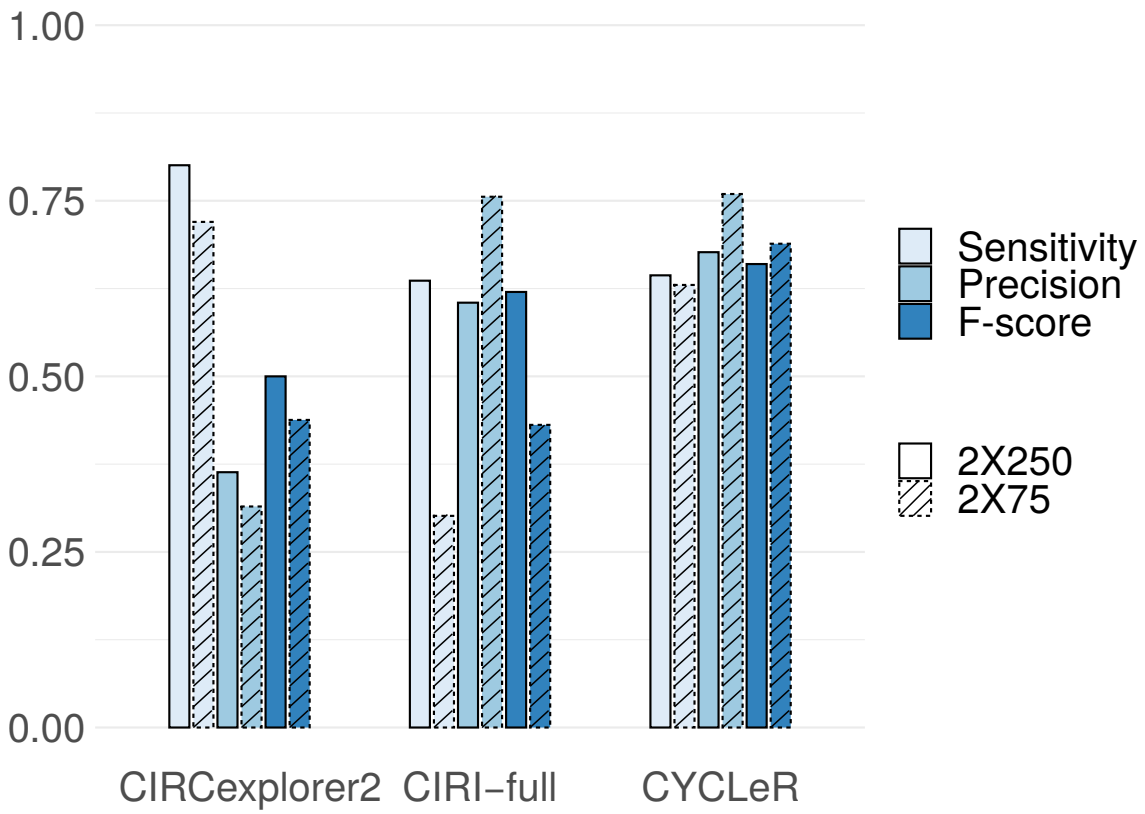

Figure S4: **Benchmark with the *High complexity* dataset.** This barplot depicts the *sensitivity, precision and F-score* of the assembled transcripts by *CYCLeR* and comparable tools in the benchmark for reference set of simulated data. The advantages of *CYCLeR* compared to other tools are apparent. The superior F-score of *CYCLeR* shows a good balance between sensitivity and precision. *CYCLeR* outperforms *CIRI-full* on all metrics. *CIRCexplorer2* has an output with higher sensitivity than *CYCLeR*, but the number of false positive assemblies shown by the precision measure makes *CIRCexplorer2* an unreliable choice. Note that *CYCLeR* is only minimally affected by the library difference in library read length.

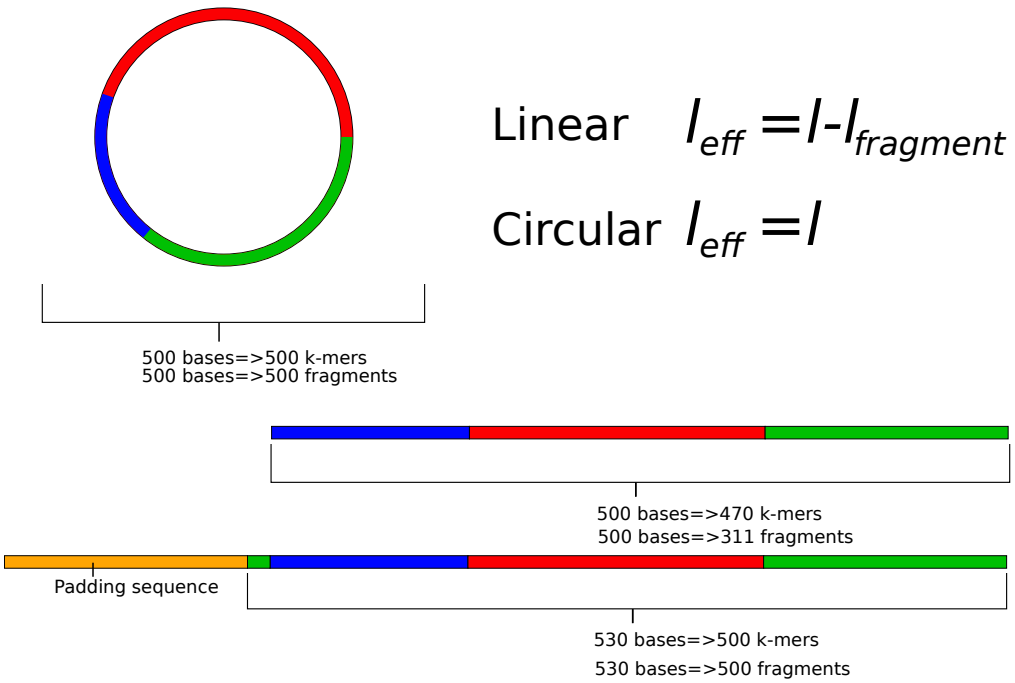

Figure S5: **Creation of a pseudo-linear reference sequence.** Comparison between an example circular and a linear transcript of the same length (l)=500 bp. Considering a k-mer size of 31 bases, the *circRNA* produces 30 more unique k-mers compared with the linear transcript of the same size. Therefore, an extra of 30 bases need to be added to the pseudo-linear transcript when an isoform quantification is used that is based on pseudo-alignments. Additional padding sequence of pseudo-random nucleotides is added for adjustment of effective length(l_{eff}) equal to one insert(*aka* fragment) length(l_{fragment}).

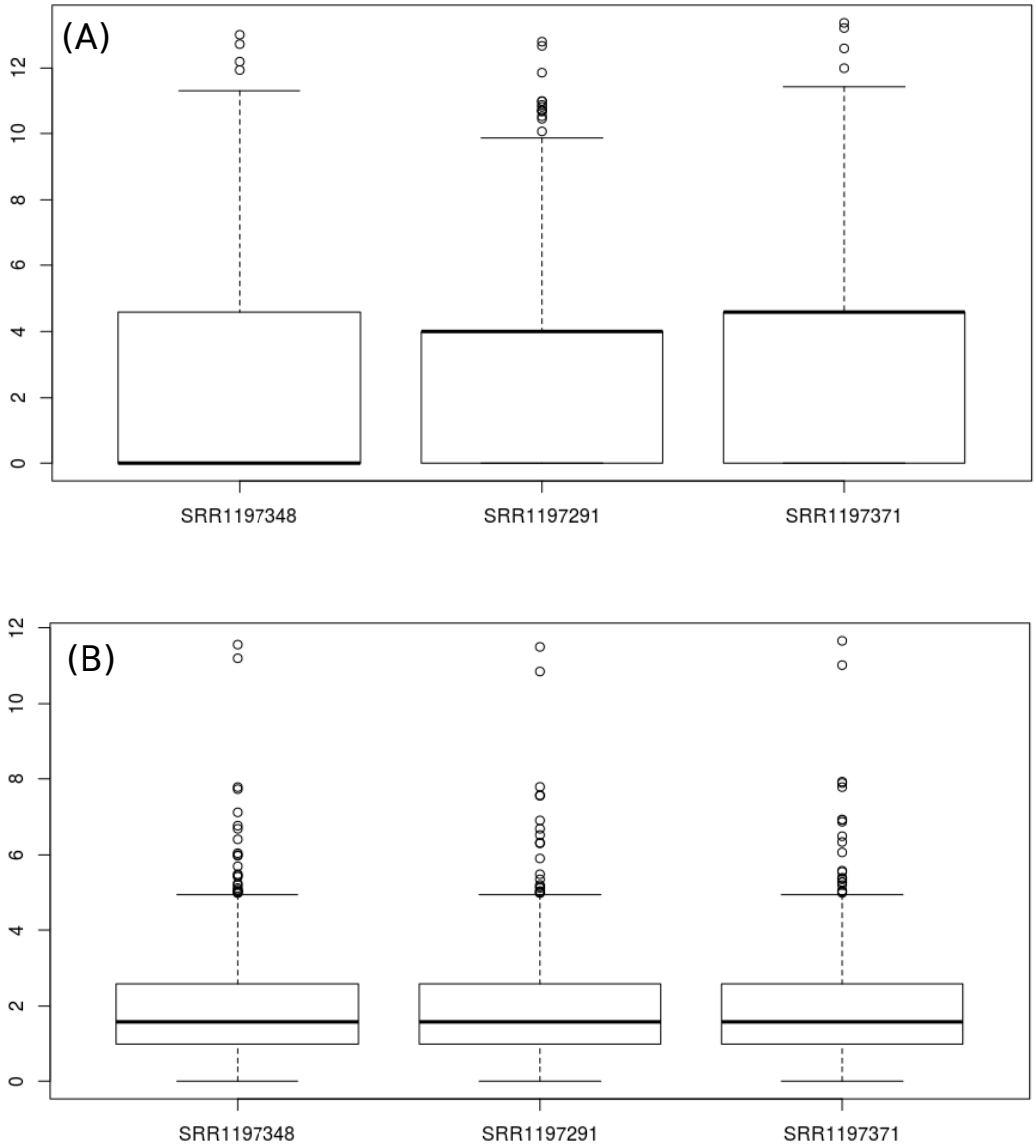

Figure S6: **Variance stability in third instar larvae, wandering stage, CNS samples.** The median of the box plots of the normalized *CYCLeR* (B) counts is more stable than the normalized *CIRCexplorer2* BSJ counts(A). This shows the higher variance stability between replicates, when quantifying with *CYCLeR*. The outputs are filtered to contain results from BSJ sites shared between the tools.

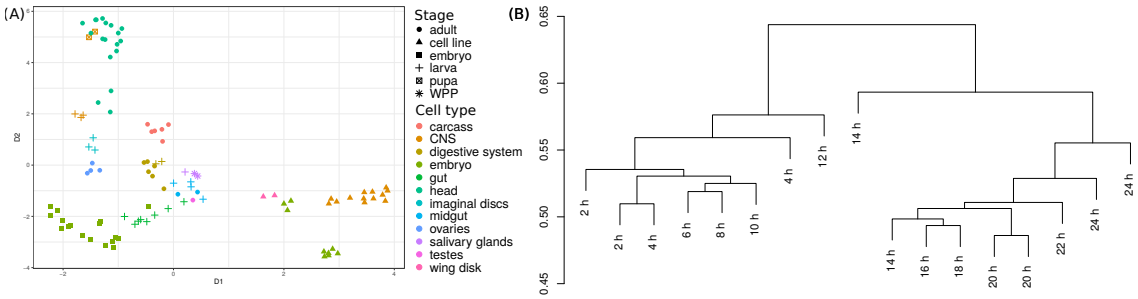

Figure S7: **Sailfish-cir applied to Lai lab 2014 dataset.** Extension from the corresponding figure in the manuscript. In part (A), is shown the UMAP-dimensional scaling of the abundances generated by *sailfish-cir* of all 103 samples from the dataset. In part (B), we use the abundances computed by sailfish-cir to plot a dendrogram of the embryo stages subset based on between-sample distance calculations. *sailfish-cir* has the same quantification strategy as *CYCLeR*, but lacks a proper assembly algorithm. The results indicate that the quantification strategy itself is not sufficient for improved clustering and the correct sequences of the isoforms are imperative for correct quantification.

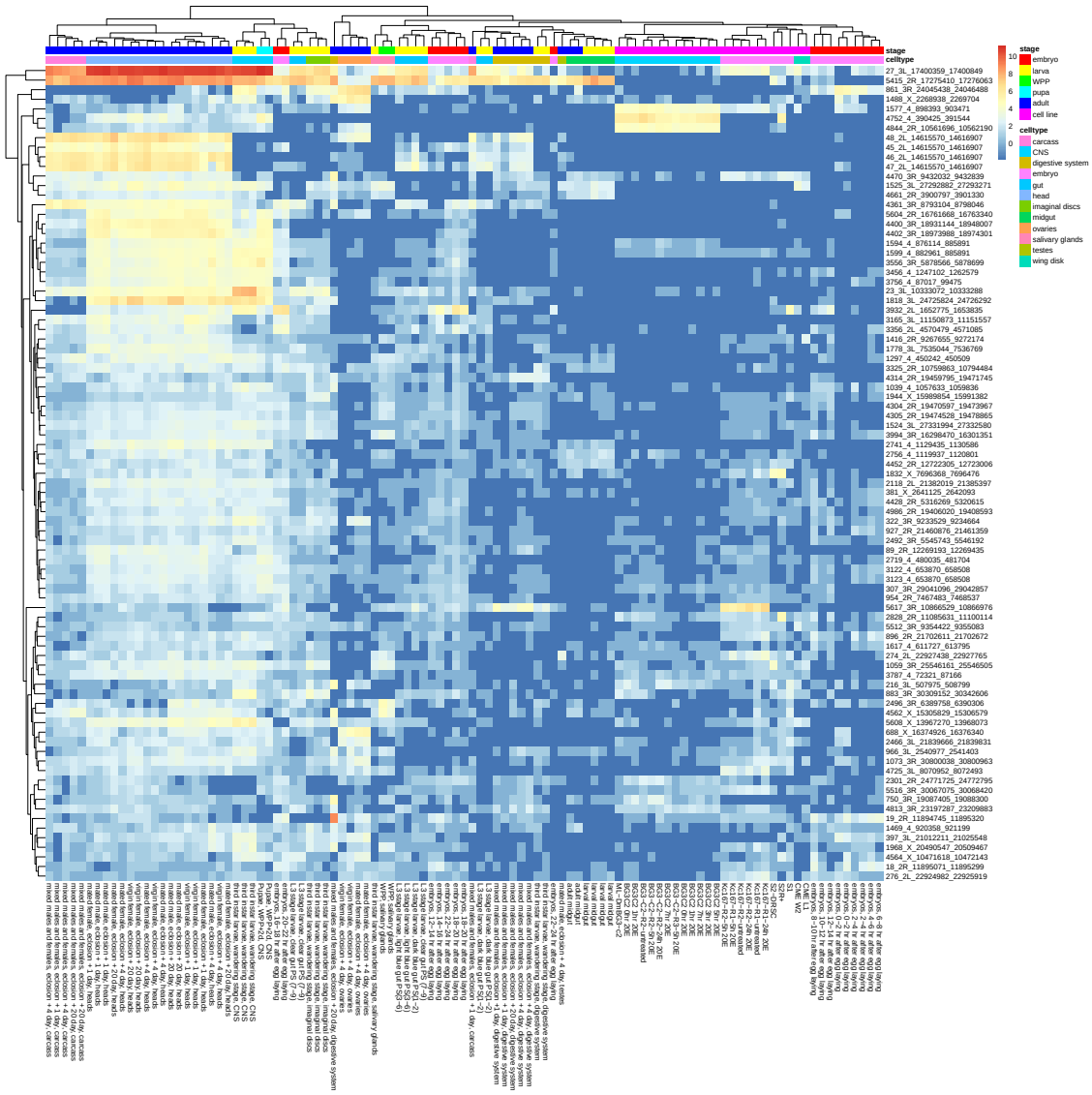

Figure S8: **Heatmap of most variable *circRNA*s** for the D. Lai dataset. *CYCLeR* can identify *circRNA*s that are specific for a particular stage of development or a particular cell type. Furthermore, it can differentiate *circRNA*s specific for the CNS or *circRNA*s specific for adult flies. We can also conclude that embryo derived cell lines show a similar pattern as early stage embryo samples.

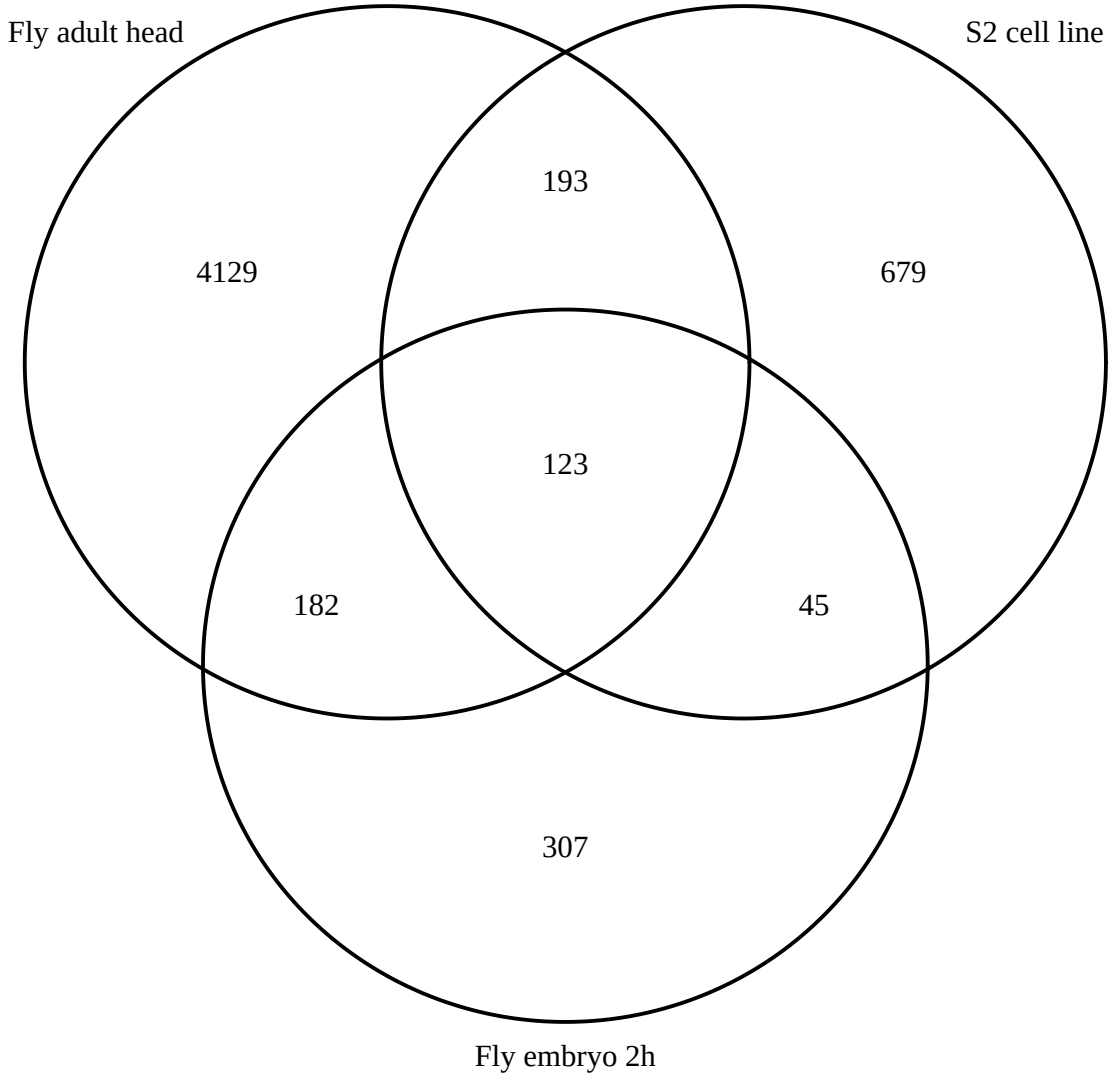

Figure S9: **Fly transcripts overlap** Overlap between the sets of transcripts that are assembled by *CYCLeR*. The data from replicates is merged into one set.

Gene	Locus of circRNA	qPCR (circRNA)	CLEAR	CYCLeR
CORO1C	Chr12:108652271-108654410	0.0008	1.28	Iso1 - 1.048 Iso2 - 0.047
FKBP8	Chr19:18539370-18539720	0.0015	1.79	1.37
KIAA0368	Chr9:111386376-111391824	0.0015	1.36	2.325
SMO	Chr7:129205202-129206587	0.0015	3.07	3.772
ARHGAP12	Chr10:31908171-31910563	0.0030	4.01	5.422
HIPK3	Chr11:33286412-33287511	0.0058	7.25	9.149
CAMSAP1	Chr9:135881632-135883078	0.0057	10.83	Iso1 - 3.456 Iso2 – 0
ZBTB46	Chr20:63775677-63790790	0.0030	3.5	4.404
CAPRIN1	Chr11:34071725-34076642	0.0008	0.6	1.564
CDK8	Chr13:26400452-26401624	0.0005	0.26	0.083
MGA	Chr15:41668827-41669958	0.0065	4.43	3.983
FAM13B	Chr5:137985256-137988315	0.0055	2.22	3.409
PLEKHM3	Chr2:207976650-207977586	0.0029	2.05	1.186
		Correlation	0.75	0.67

Figure S10: **Results for PA1 RNA-seq data.** All values in the table correspond to averages between replicates. The yellow rows indicate the filtered items from the benchmark due to known multiple isoforms. The correlation cells show the Pearson product correlation of the filtered values of qPCR results and estimated abundances. Note that the CAMSAP1 Iso2 has zero value due to the fact that with default parameters, *CYCLeR* fails to recover the more abundant isoform. This is due to multiple overlapping *circRNA* isoforms, whose reconstruction leads to a premature depletion in the reconstruction algorithm.

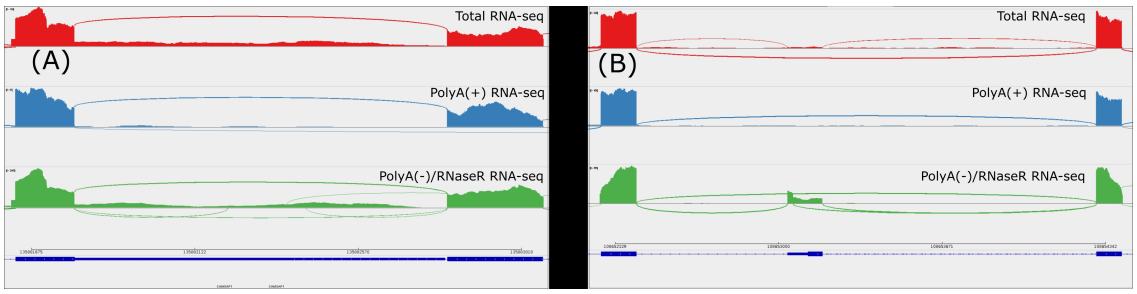

Figure S11: **Sashimi plots for PA1 RNA-seq data.** The sashimi plots show the BSJ loci of CAMSAP1 (A) and COROC1C (B). After circRNA enrichment, additional exons are present compared to the results for polyA(+) data. In the case of CAMSAP1, there is an alternative isoform with a retained intron between the two exons. In the COROC1C locus, there is an exon with an alternative 3'-splice site. Splice site junctions with fewer than three reads are not shown.

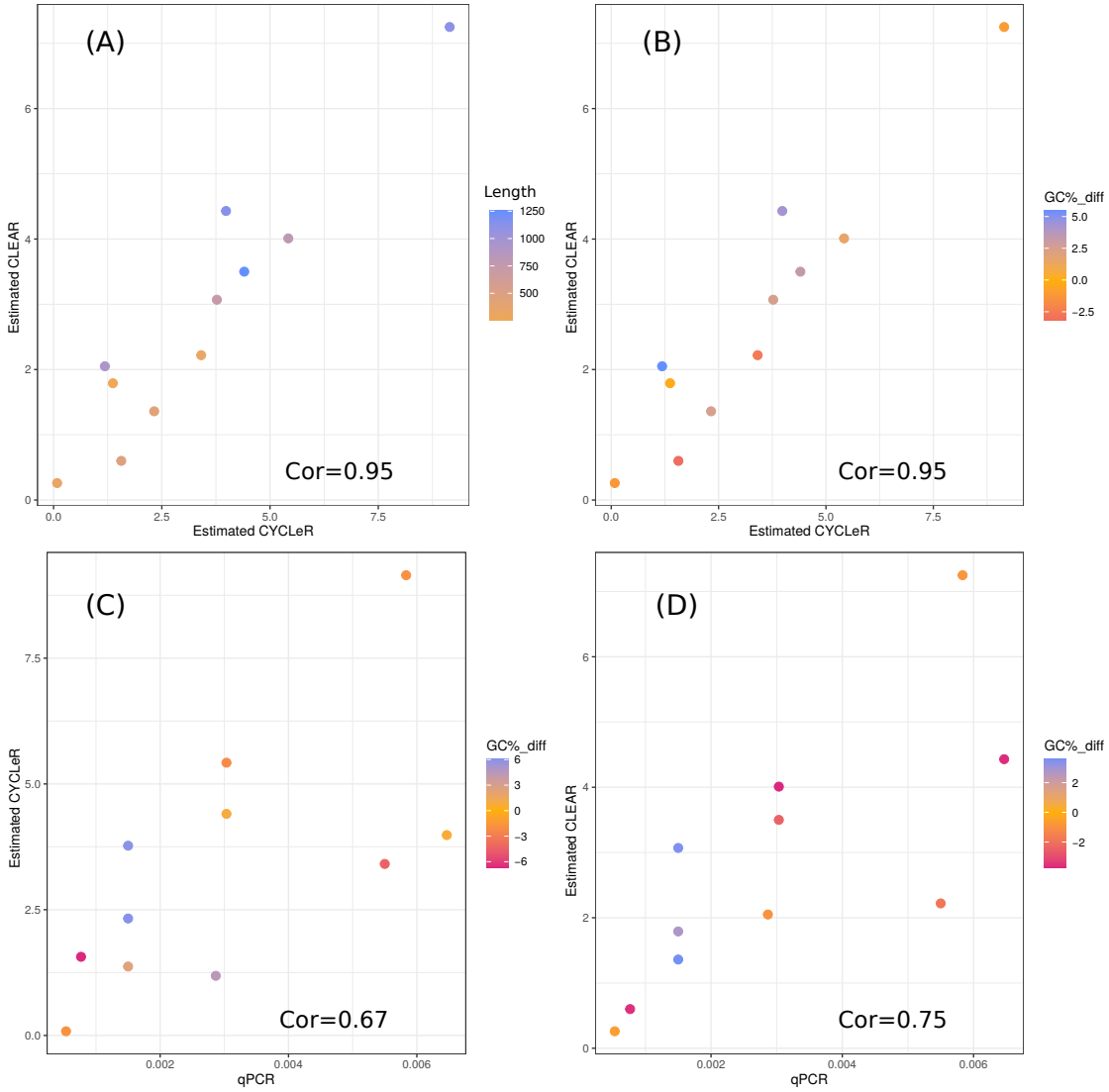

Figure S12: **Evaluation of the difference between CLEAR and *CYCLeR*.** The output of *CYCLeR* and CLEAR are in very good agreement as shown in (A) and (B). The most likely sources of difference is the length of transcript (A) and GC-content (B). The comparison between the off-diagonal points in (A) and (B) indicates that the source of the difference is most likely the difference in GC content. The fact that CLEAR focuses only on the region around the BSJ makes the GC-content affecting CLEAR output closer to the GC-content affecting qPCR results. This is supported by (C) and (D) showing a comparison between the GC-content between the evaluated locus of the qPCR product to the GC-content of the *CYCLeR* transcript and 200 nt region around the BSJ respectively. The difference in GC-content is higher for the locus evaluated by *CYCLeR* and the off-diagonal points account for the difference between *CYCLeR* abundance estimation and qPCR results. Naturally, the differences between qPCR results and abundance estimation cannot be explained by those plots, as the difference between experimental procedures is influenced by a numerous biases, yet these plots at least manage to explain the better agreement between CLEAR and qPCR data.

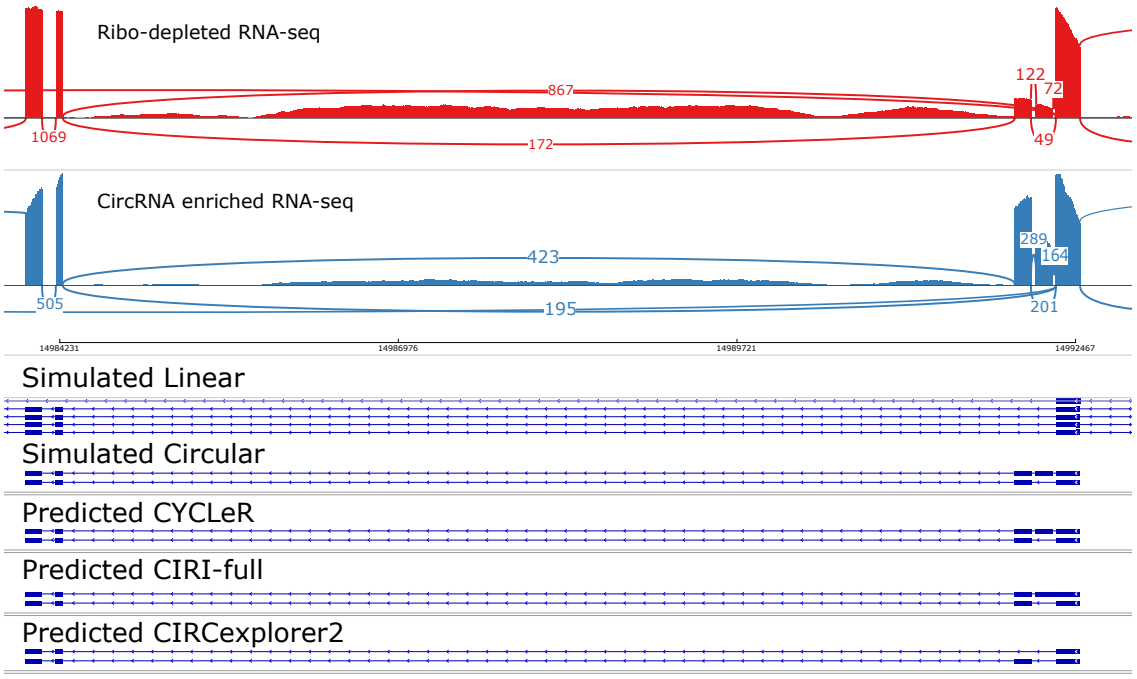

Figure S13: **Comparison of the assembly of BSJ locus chr2L:14,983,950-14,992,506 in *D.* melanogaster.** The Sashimi plots show the STAR mapping of simualted data for the exons and FSJ encompassed by the BSJ sites of the circle in chr2L:14,983,950-14,992,506. The plot shows comparison between total ribo-depleted RNA-seq simulation and circRNA enriched RNA-seq. The sequence mode used is 2 x 250 (leading to less observable biases that in Figure S1). This example was selected to show the advantage of *CYCLeR* in handling unannotated exons. While all tools can handle identification of one of the unannotated exons, the second exon is accounted for properly only by *CYCLeR*. We observe that the *CIRCexplorer2* assembly is biased by the provided linear annotation. The *CIRI-full* assembly disregards one of the FSJs and assembles an erroneous full exon. *CYCLeR* not only correctly identifies all exons and splice junctions, bus also manages to properly manage the AS event and reconstructs the correct isoforms.

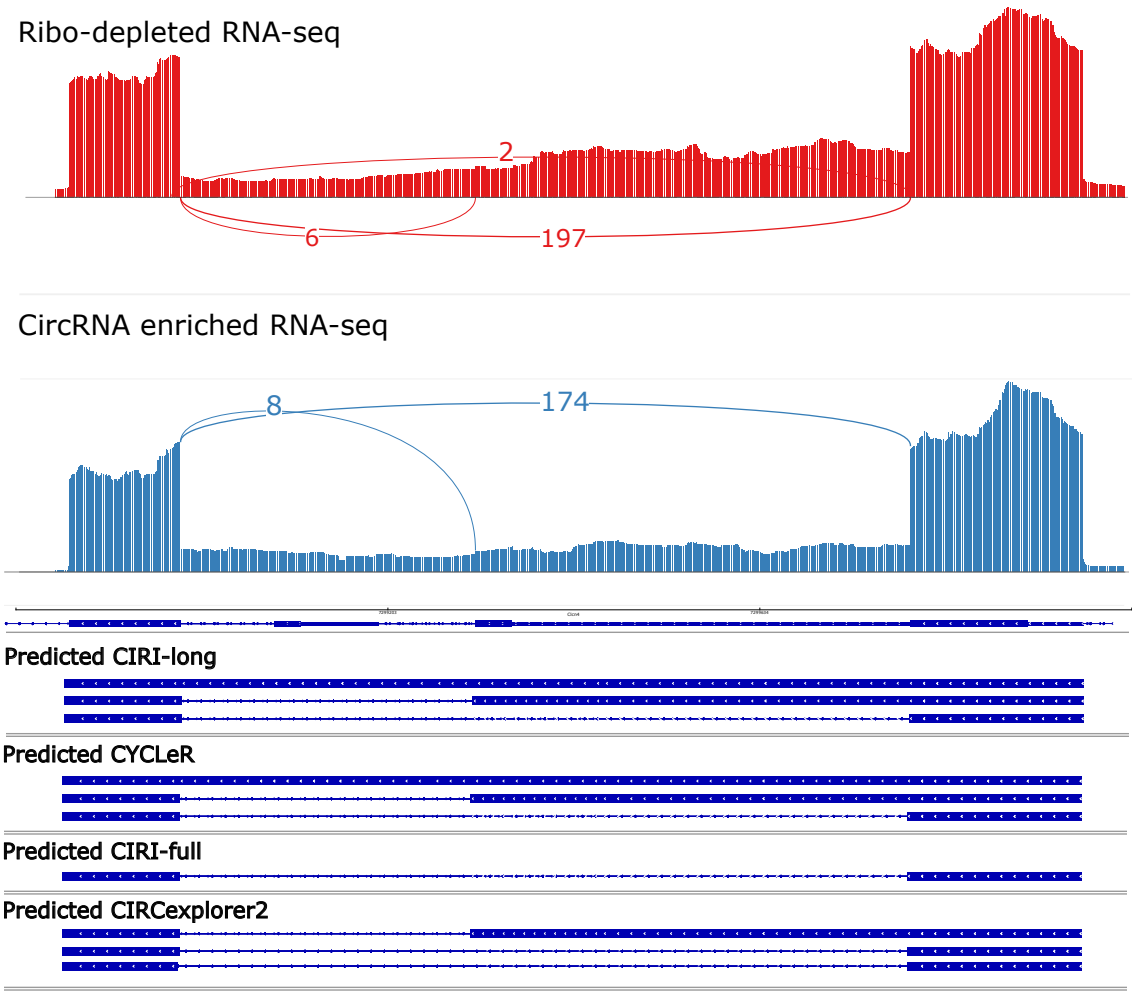

Figure S14: **Comparison of the assembly of BSJ locus chr7:7,298,969-7,299,877 in *M. musculus*.** The Sashimi plots show STAR mapping of the exons and FSJ encompassed by the BSJ sites of the chr7:7,298,969-7,299,877 circles. The FSJs further away not participating in circRNAs have been removed for visibility. The plot shows a comparison between total ribo-depleted RNA-seq and circRNA enriched RNA-seq. The CIRI-LONG output serves as a true positive reference. This example was selected to show the advantage of *CYCLeR* to handle retained introns. As shown, *CIRI-full* does not account for an alternative 5'-splicing as well as the retained intron. *CIRCexplorer2* makes an assembly error as the tool attempts to match the assembly to the given linear annotaion. The superior feature selection of *CYCLeR* compared to *CIRCexplorer2* is the reason for avoiding an exclusively linear transcript FSJ, thereby preventing an incorrect isoform assembly. *CYCLeR* is the only tool that manages to assemble the isoform containing the retained intron.

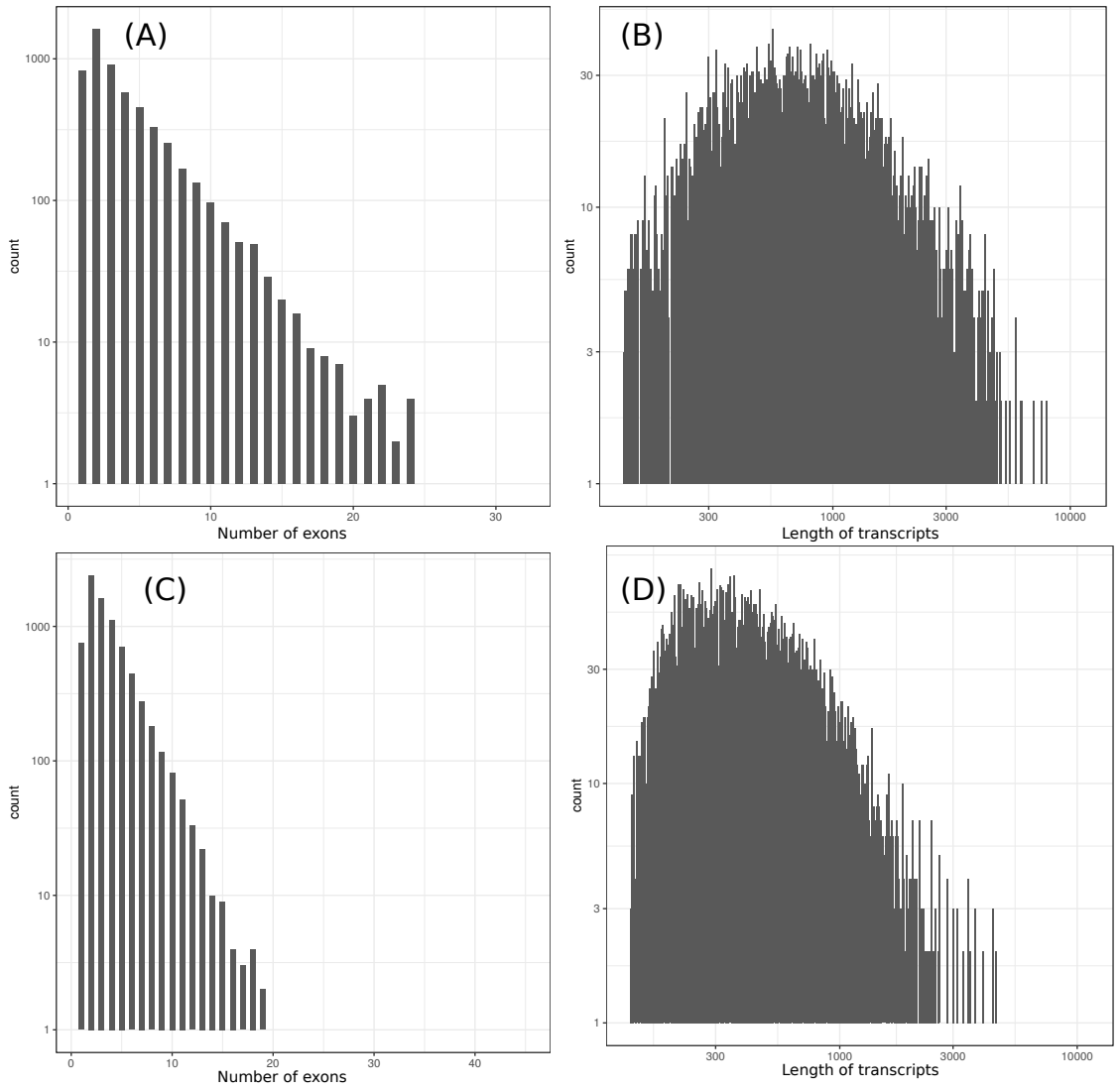

Figure S15: **Statistics of real data assembly.** The results show the reconstruction of the PA1 dataset (A) and (B) and the accumulated data of the fruit fly dataset (C) and (D). (A) and (C) show the number of exons per transcript and (B) and (D) show the length of transcript on a logarithmic scale.

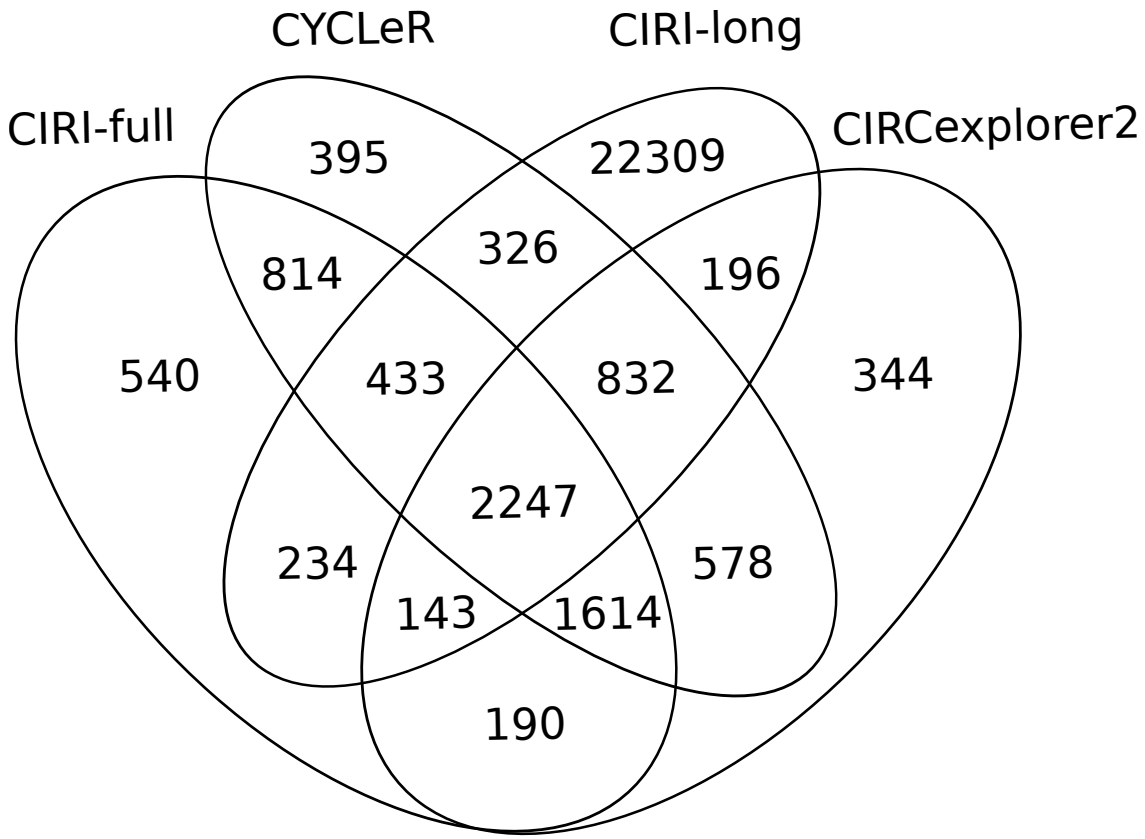

Figure S16: **Venn diagram of unique BSJ per tool in the benchmark versus CIRI-long Nanopore data.** The assembly of each tool is dependent of the set of input BSJs. This plot complements the transcript assembly (shown in Figure S17 and Figure 6 in the manuscript)and sheds light on the differences in prediction. The *CIRI-full* and *CIRCexplorer2* BSJ that are not part of the *CYCLeR* output derive from loci that failed the BSJ enrichement requirement. The BSJs that are unique for *CYCLeR* are BSJs identified by CIRI2, which belong to circRNAs longer than the *CIRI-full* detection limit. *CIRCexplorer2* has unique set of BSJs disproportional to the amount of unique transcripts assembled (see also Figure S17).

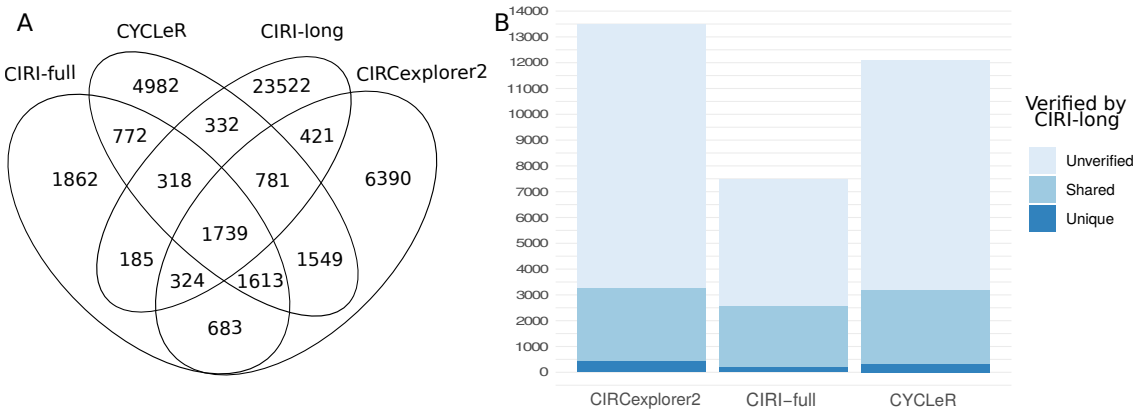

Figure S17: **CIRI-long Nanopore benchmark.** These figures show the results of the comparison between Illumina-based methods and an Oxford Nanopore-based method. (A) shows a Venn diagram of the full set of assembled transcripts for each tool. (B) is a bar graph representation of the same data, but with emphasis on overlapping regions from the Venn diagram. On (B), the assembled transcripts for each Illumina-based tool are divided into *verified* by CIRI-long or *unverified*. The latter are further subdivided into *unique* - the transcripts that are shared only by one Illumina-based tool and CIRI-long, and *shared* - the transcripts that are shared by two or more Illumina-based tools and CIRI-long. *CIRI-full* has the lowest transcript count in every category. This is due to the length limit of assembly based on the library insert size. When comparing *CIRCexplorer2* and *CYCLeR* , we notice that *CIRCexplorer2* has only ∼100 more *verified* transcripts, while simultaneously having ∼1400 more unverified transcripts. Based on the information provided by the simulated benchmark, it is safe to conclude that the extra isoforms produced by *CIRCexplorer2* are primarily erroneous assemblies.

Command logs:

Mapping:

STAR --runThreadN 8 --chimSegmentMin 15 --chimScoreMin 1 --alignIntronMax 100000 --outFilterMismatchNmax 4 --alignTranscriptsPerReadNmax 10000 --outFilterMultimapNmax 500 --limitOutSAMoneReadBytes 300000

bwa mem -T 19

tophat -o <.> -p 4 -G <gtf> <fasta>

BSJ identification:

CIRI_v2.0.3.pl -I <sam> -O ./ciri_$i -T 4 -F <fasta> -A <gtf>

CIRCExplorer2 parse -t STAR Chimeric.out.junction
CIRCExplorer2 annotate -r <annot_flat> -g <fasta> ./circ_out

CircRNA characterization/assembly:

CIRI_AS_v1.2.pl -S ./$i.sam -C ./ciri_$i -O ./ciri_as_$i -F <fasta> -A <gtf> -D yes
java -jar ./CIRI-full_v2.0/CIRI-full.jar RO1 -1 ./${i}_1.fasta -2 ./${i}_2.fasta -o ./ciri_ro_$i
bwa mem -T 19 /scratch/AG_Meyer/fly_data/dm6/genome_ensembl/bwa_index/bwa_index ./ciri_ro_${i}_ro1.fq> ./${i}_ro1.sam
java -jar ./CIRI-full_v2.0/CIRI-full.jar RO2 -r /<fasta> -s ./${i}_ro1.sam -l 250 -o ./${i}_
java -jar ./CIRI-full_v2.0/CIRI-full.jar Merge -c ./ciri_$i -as ciri_as_${i}_jav.list -ro ./${i}__ro2_info.list -a <gtf> -r <fasta> -o ./${i}_
java -jar ./CIRI-full_v2.0/CIRI-vis.jar -i ./${i}__merge_circRNA_detail.anno -l ciri_as_${i}_library_length.list -r <fasta> -min 2 -o vis_out_${i} -d stdir_${i}

CIRCexplorer2 assemble -r <annot_flat> -m <path> -o assemble
CIRCExplorer2 denovo --as --rpkm --tophat-dir ./sample1 -a ./sample2 -r <annot_flat> -g <fasta> ./circ_out

CircRNA quantification:

CIRCExplorer2 results across all samples are accumulated into a single bed file - for_sailfish.bed
python ./sailfish-cir/sailfish_cir.py -g <fasta> -a <gtf> -1 ${i}_1.fastq -2 ${i}_2.fastq --bed for_sailfish.bed -o ./${i}

CYCLeR:

CYCLeR runs are performed as described in the manual: http://www.e-rna.org/cycler/

CIRI-long:

CIRI-long is run as shown in Methods Zhang et al .2021 (16).

Figure S18: **Overview of commands used.** Information about the annotation and parameters used for the tools. (4; 19).