Accurate Assembly of Circular RNAs with TERRACE

Tasfia Zahin[1] Qian Shi[1,*] Xiaofei Carl Zang[2,*] Mingfu Shao[1,2]

[1]Department of Computer Science and Engineering, School of Electrical Engineering and Computer Science, The Pennsylvania State University

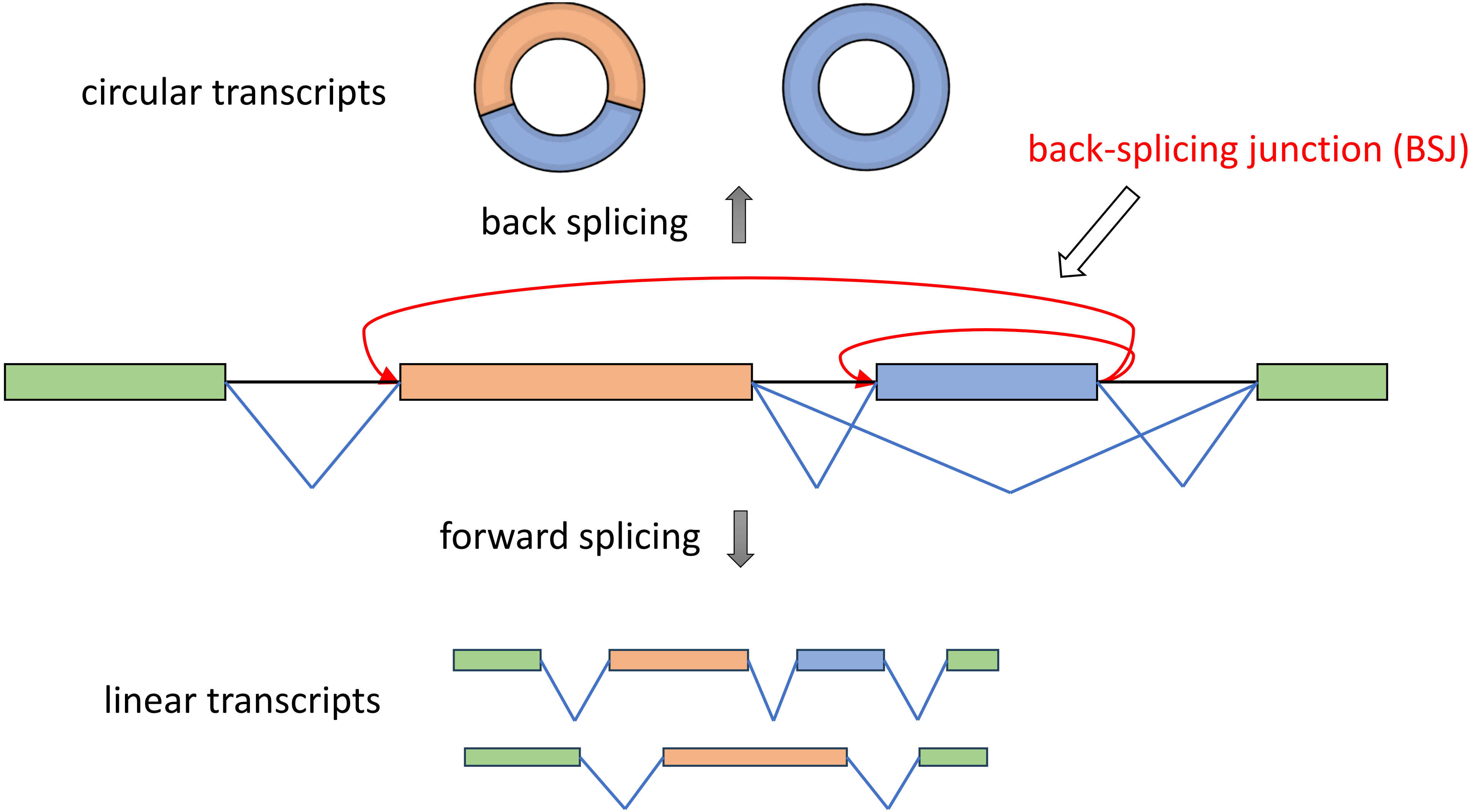
[2]Huck Institutes of the Life Sciences, The Pennsylvania State University

Cambridge | 2024
RECOMB

PennState College of Engineering | ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

PennState Huck Institutes of the Life Sciences

Circular RNA



circular transcripts

back splicing

back-splicing junction (BSJ)

forward splicing

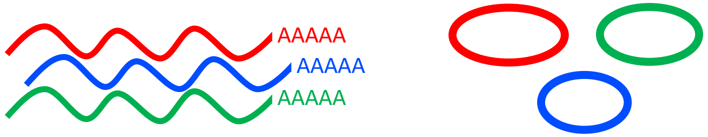
linear transcripts

Significance of Circular RNAs

widely expressed in human tissues

More than 60% of human genes express at least one circular RNA

very stable in nature

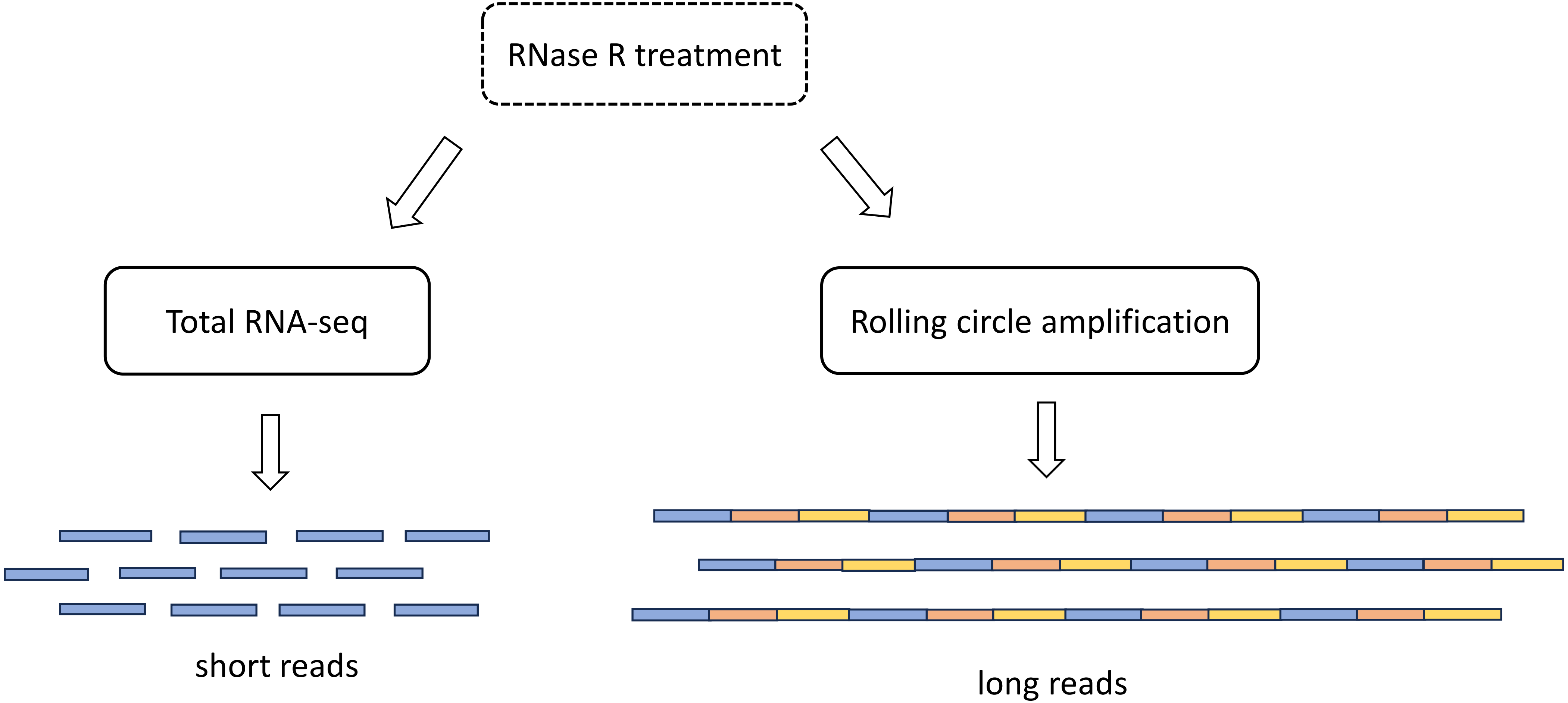


Circular RNAs have at least 2.5 times longer half-life than linear RNAs in mammary cells

reliable disease biomarkers
diagnostic and therapeutic targets

circ-Foxo3 interacts with certain proteins to form a complex and inhibits cell cycle progression in cancer

Existing Experimental Protocols



RNase R treatment

Total RNA-seq

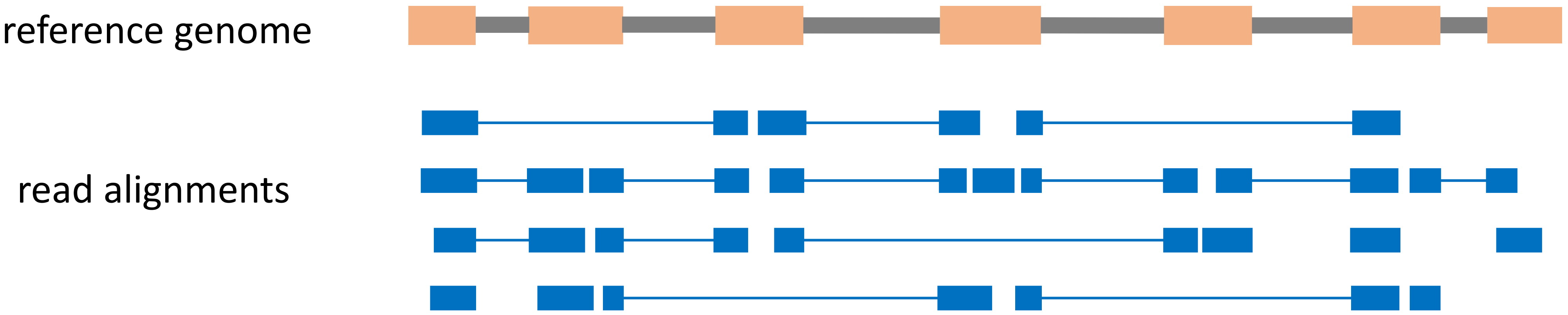
Rolling circle amplification

short reads

long reads

Circular Transcript Assembly

Input: a set of paired-end total RNA-seq reads aligned to a reference genome



Optional: reference gene annotation

Goal: reconstruct a set of full-length circular transcripts

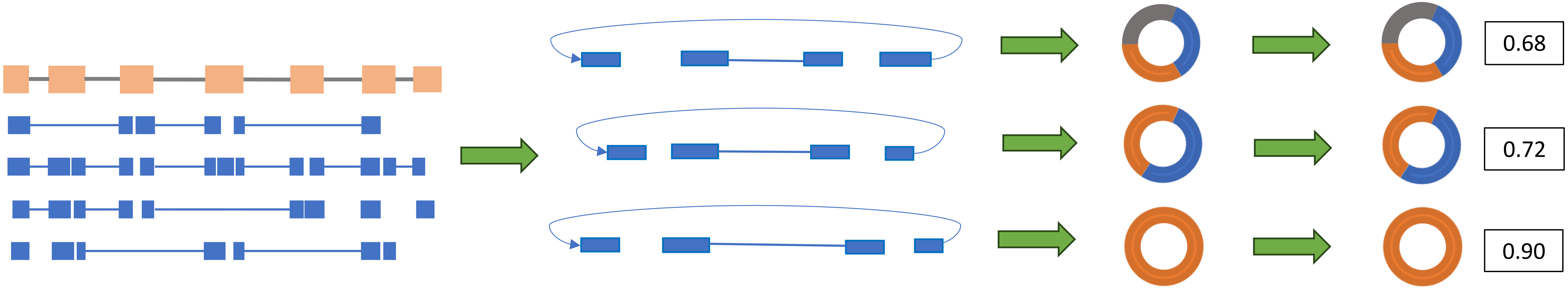
Challenges in Computational Methods

- Several computational methods exist for circular RNA detection, however:
- Majority lack full-length detection capabilities (CIRI2, Circall, CircMiner, CircMarker).
- Other full-length assemblers require a reference gene annotation (CIRCexplorer2, CircAST, cycler, psirc).
- Detection accuracy is often not sufficiently high.

Develop a new circular RNA assembler

- **does full length detection**
- **does not rely on reference annotation**
- **achieves high assembly accuracy**

Methodology



reads alignment

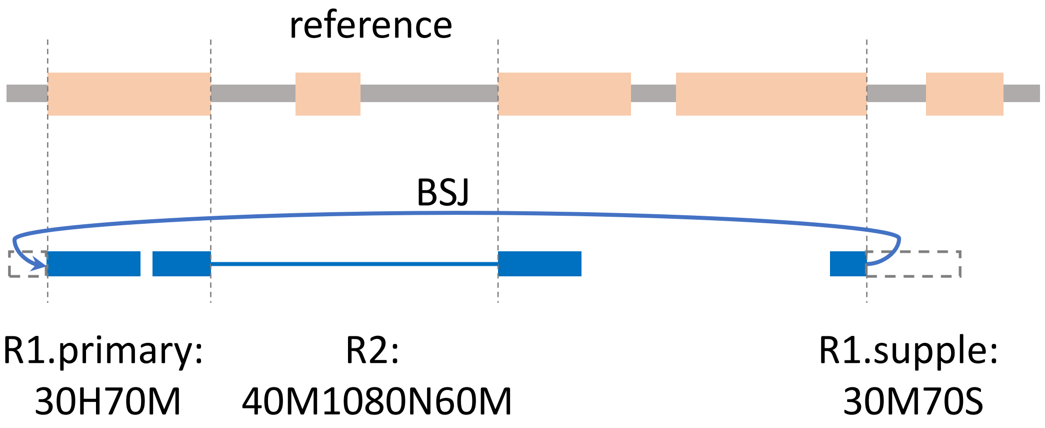
identify reads with back-splicing junctions

assembled circular transcripts

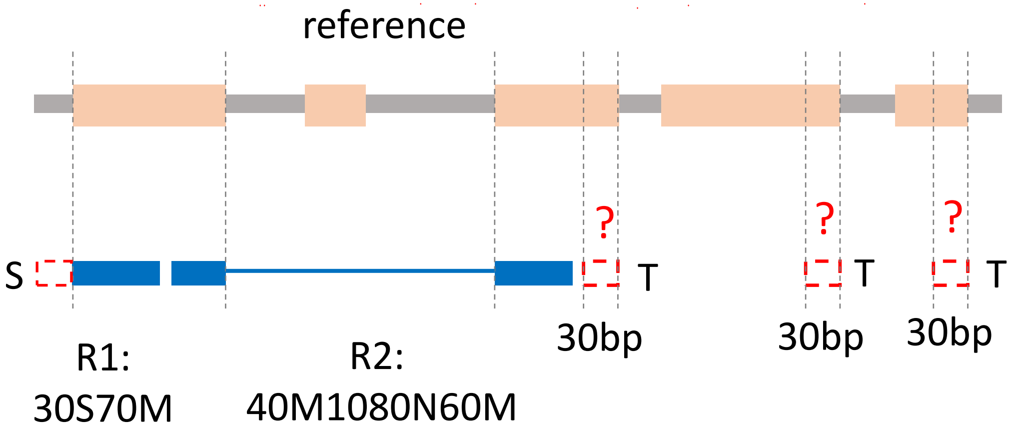
scored circular transcripts

Identifying Back-spliced Reads

From chimeric alignments



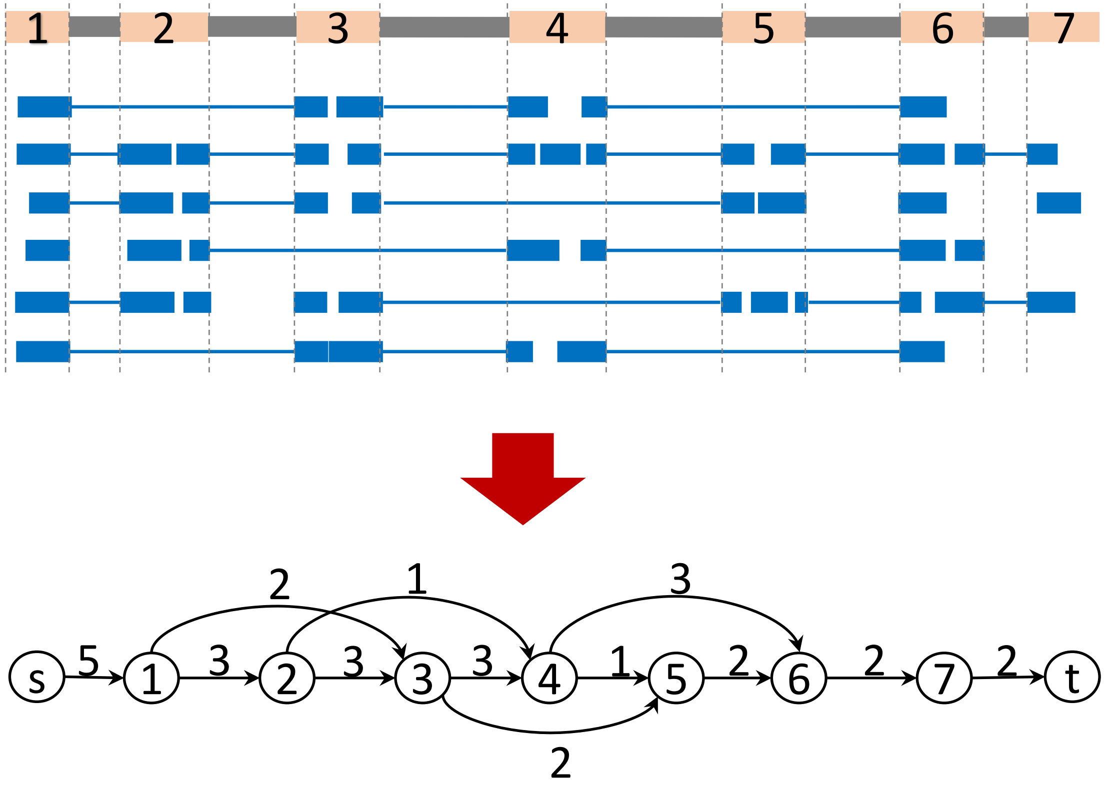
From new junction mapping algorithm



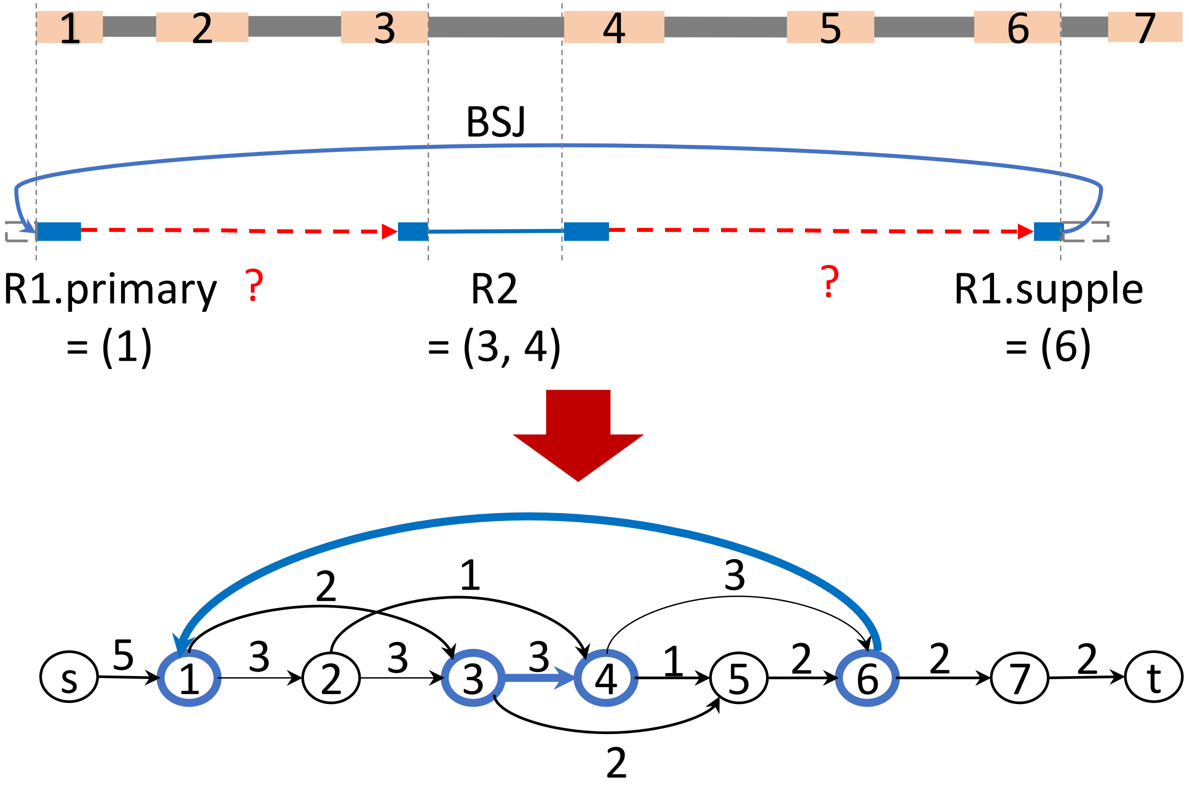
Jaccard index (S, T) > 0.9

Transforming Assembly to Bridging

splice graph formation



back-spliced read represented by 3 paths in splice graph



Formulation and Algorithm for Bridging

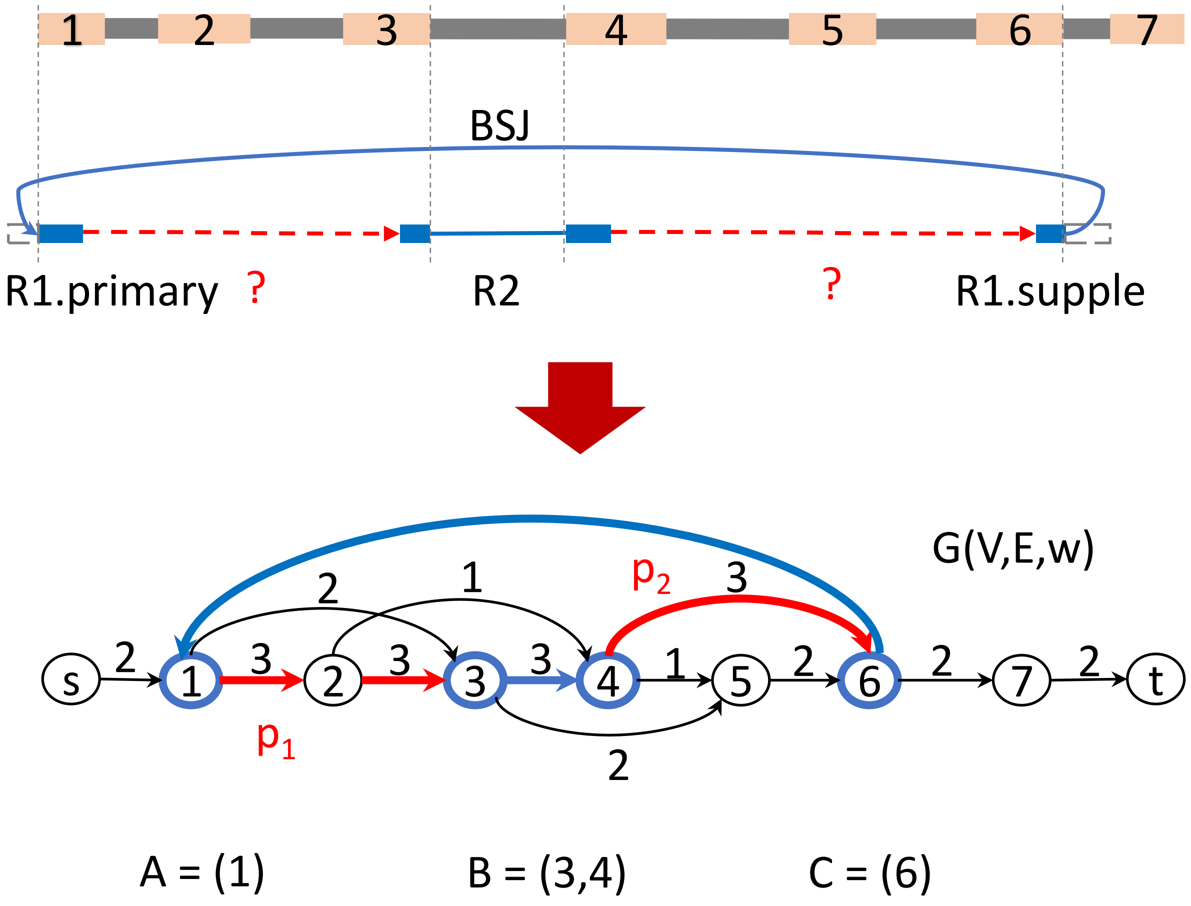
Given: $G(V,E,w)$, $A = (a_1, a_2, \dots, a_i)$, $B = (b_1, b_2, \dots, b_j)$, and $C = (c_1, c_2, \dots, c_k)$

Goal: Find a path p_1 from a_i to b_1 and a path p_2 from b_j to c_1 such that the **score** of p_1 and that of p_2 are maximized.

score(p): smallest weight over all the edges in p.

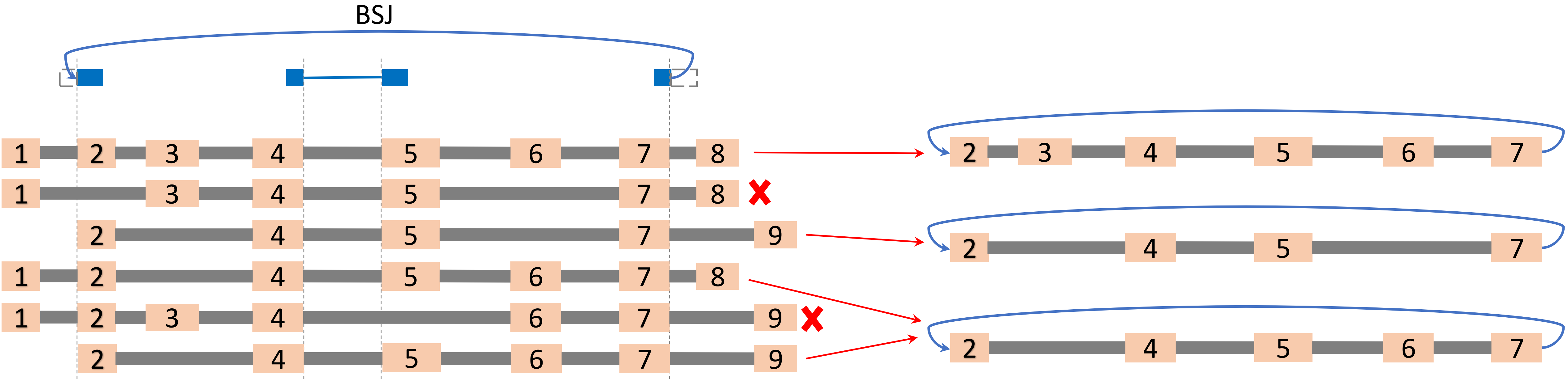
Adopt an efficient dynamic programming algorithm ($O(|V|^2 \cdot |E|)$) to find optimal p_1 and p_2, previously proved to effective for linear transcript assembly.

Collect the top 10 optimal bridging paths (P).



Use of Reference Annotation

- Set of circular paths from reference annotation, if provided (P_2).



annotated linear transcripts

compatible full-length circular paths

Selection of Candidate Paths

Back-spliced read

Set of bridging paths P

Filter paths in P:
- Discard paths with uncovered partial exon regions
- Discard paths with low scores
- Discard paths with indication of intron retentions

Reference annotation

Unique compatible paths from annotated transcripts: P_2

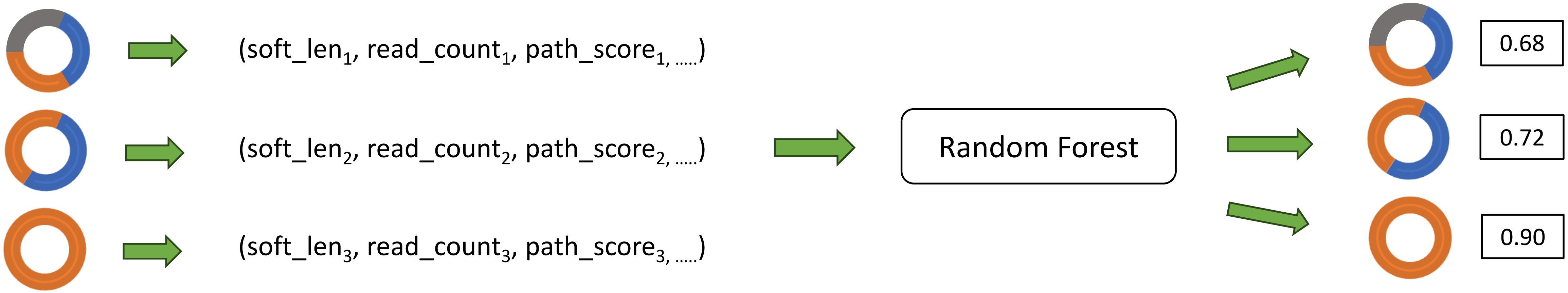
+

Set of survived bridging paths: P_1

Selection Criteria:
- If $P_1 \cap P_2 \neq \emptyset$, pick the path in $P_1 \cap P_2$ with maximum score.
- If $P_1 \cap P_2 = \emptyset$ and $P_1 \neq \emptyset$, pick the path in P_1 with maximum score.
- If $P_1 = \emptyset$ and P_2 contains a single path, pick that path.
- If $P_1 = \emptyset$ and P_2 contains multiple path, discard the back-spliced read.
- If $P_1 \cup P_2 = \emptyset$, discard the back-spliced read.

Assembled circular RNA

Scoring Assembled Transcripts



(soft_len$_1$, read_count$_1$, path_score$_{1,}$ .....)

(soft_len$_2$, read_count$_2$, path_score$_{2,}$ .....)

(soft_len$_3$, read_count$_3$, path_score$_{3,}$ .....)

Random Forest

assembled circular transcripts from one sample

extract list of 13 features for each circRNA

train a machine learning model and test all samples

assign confidence scores to circRNAs

Experimental Setup

Datasets: Short-read total RNA-seq datasets of 8 human tissues (accession number PRJCA000751 from BIGD)

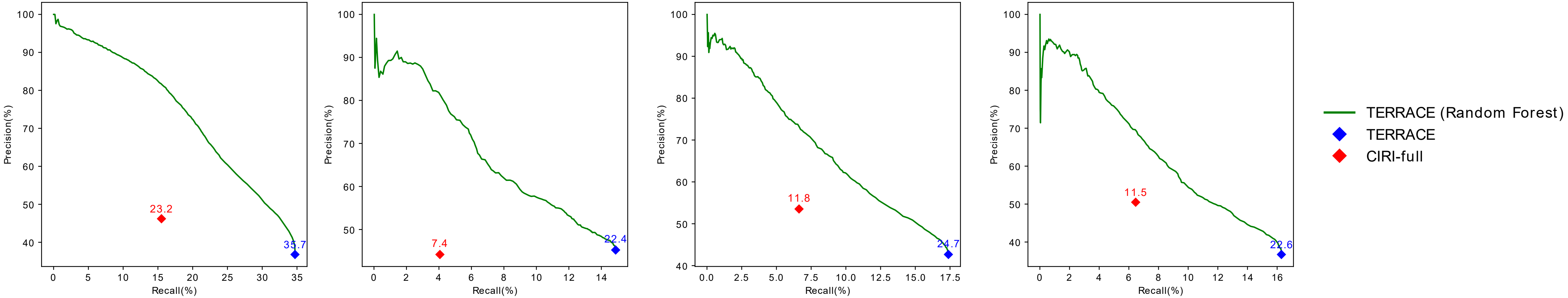
Ground truth: Full length circular rRNAs assembled by isoCirc using long-reads and reference gene annotation

Methods Compared: CIRI-full, CIRCexplorer2, CircAST

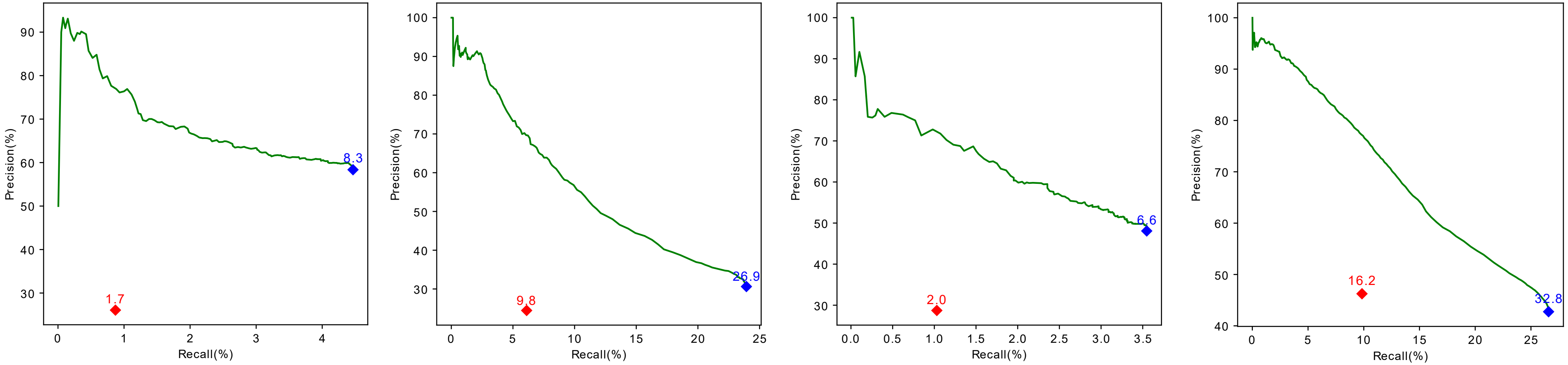
Evaluation: number of matching transcripts with ground truth

Metric: F-score**,** precision-recall curve

Assembly Accuracy on Human Tissue Samples



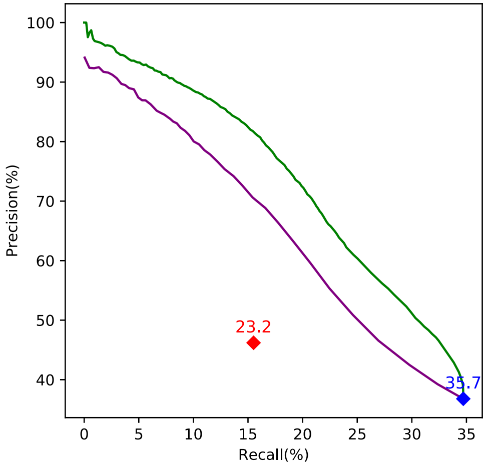
brain heart kidney liver



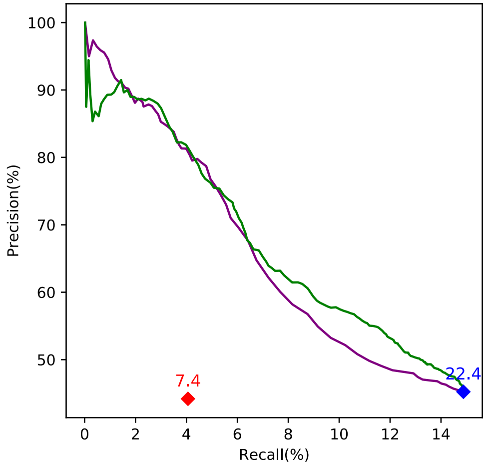
lung prostate skeletal muscle testis

without annotation

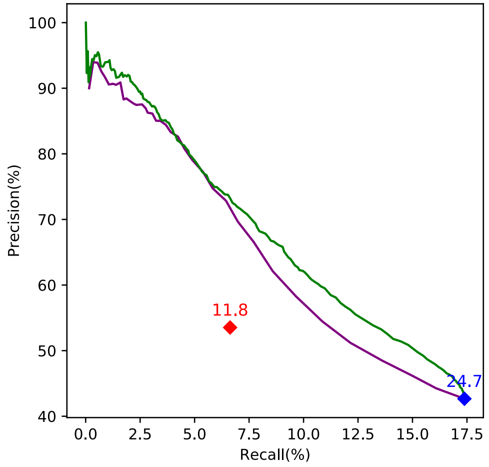
Abundance vs Random Forest



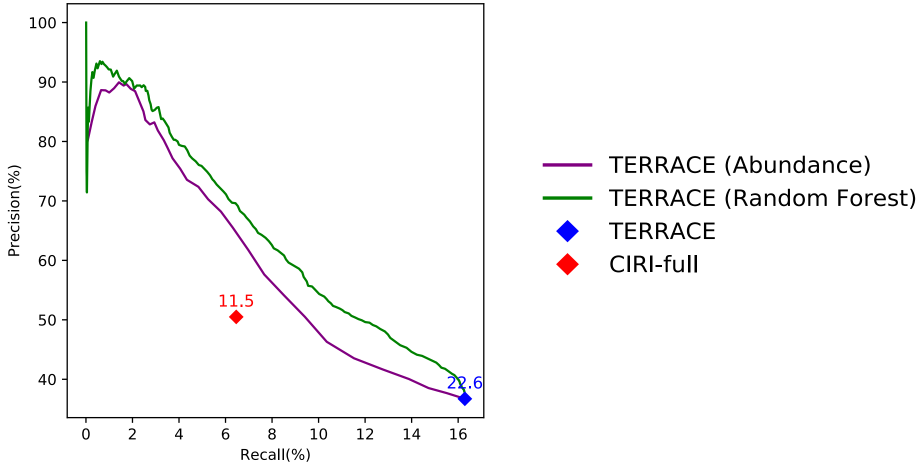
brain



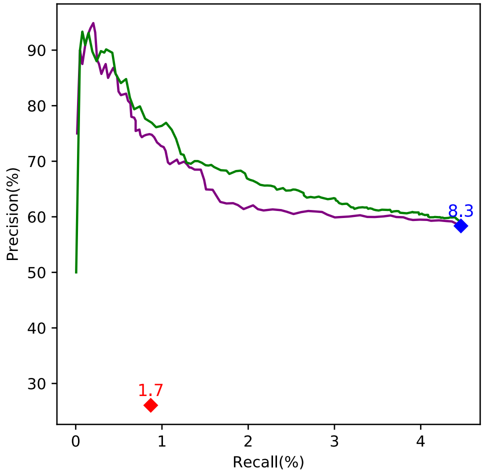
heart



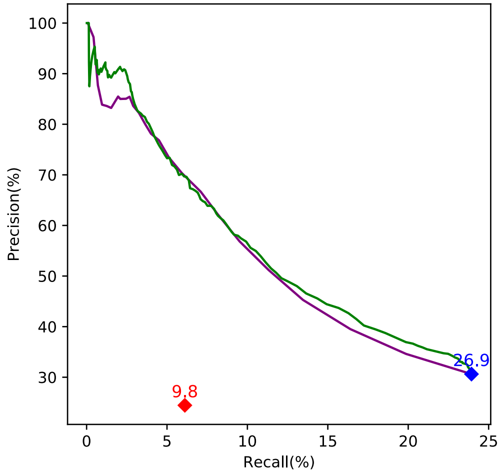
kidney



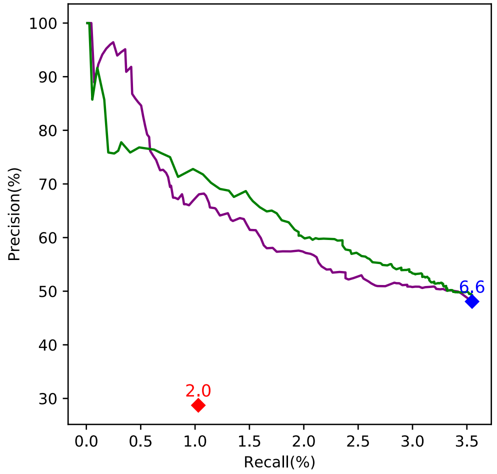
liver



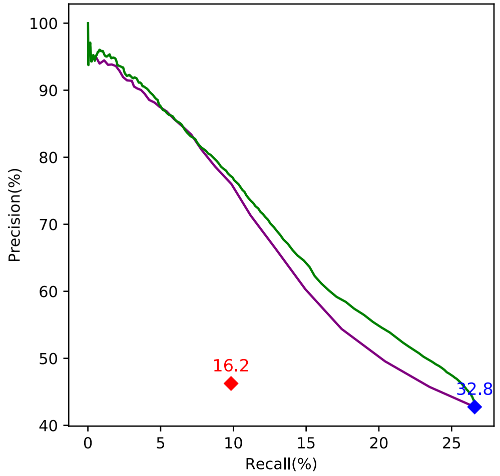
lung



prostate



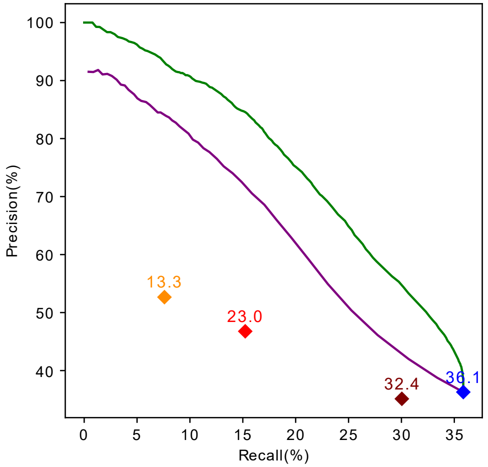
skeletal muscle



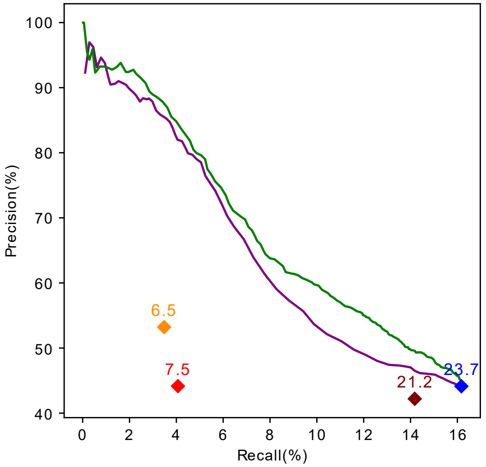
testis

without annotation

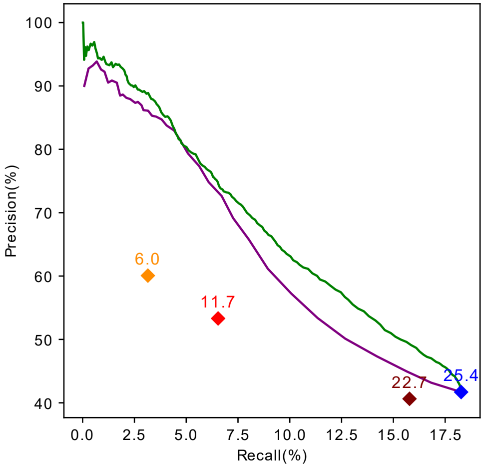
Assembly Accuracy on Human Tissue Samples



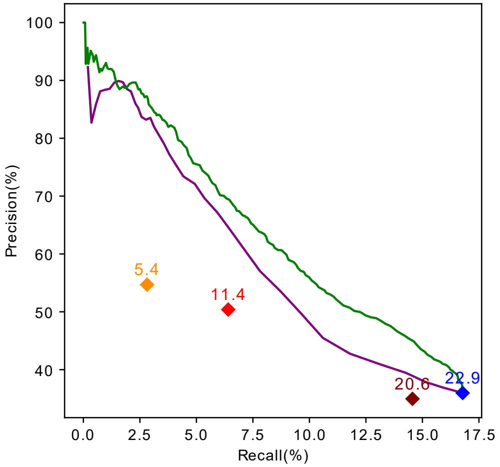
brain



heart

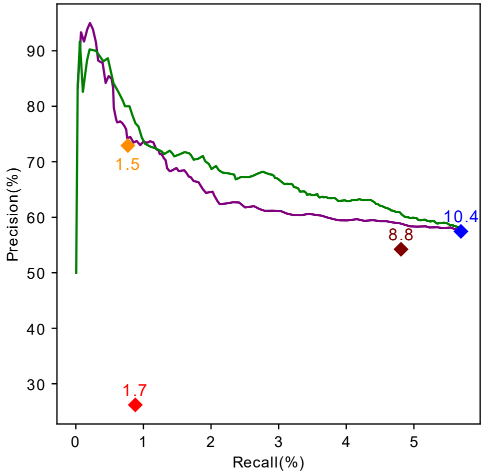


kidney

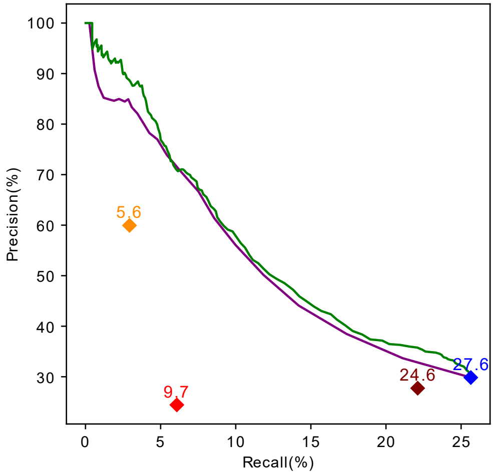


liver

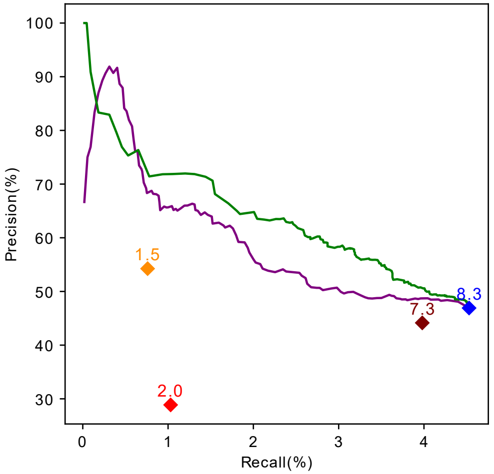
TERRACE (Abundance)
TERRACE (Random Forest)
◆ TERRACE
◆ CIRI-full
◆ CIRCexplorer2
◆ CircAST



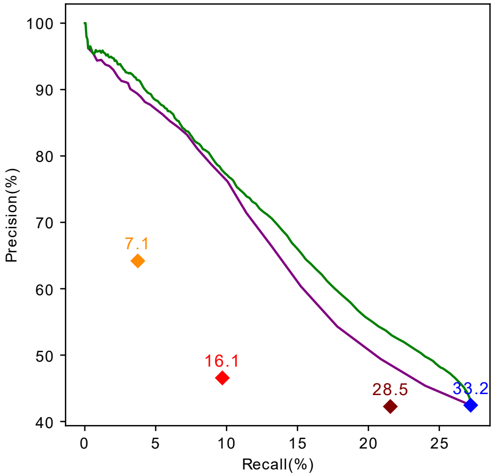
lung



prostate



skeletal muscle

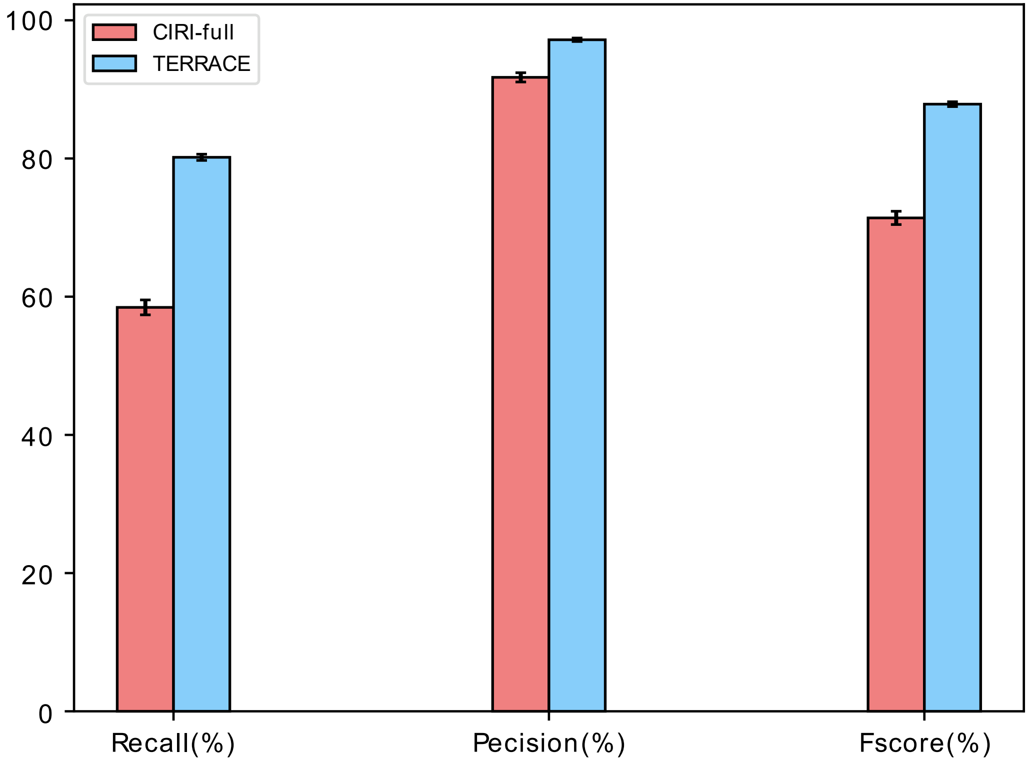


testis

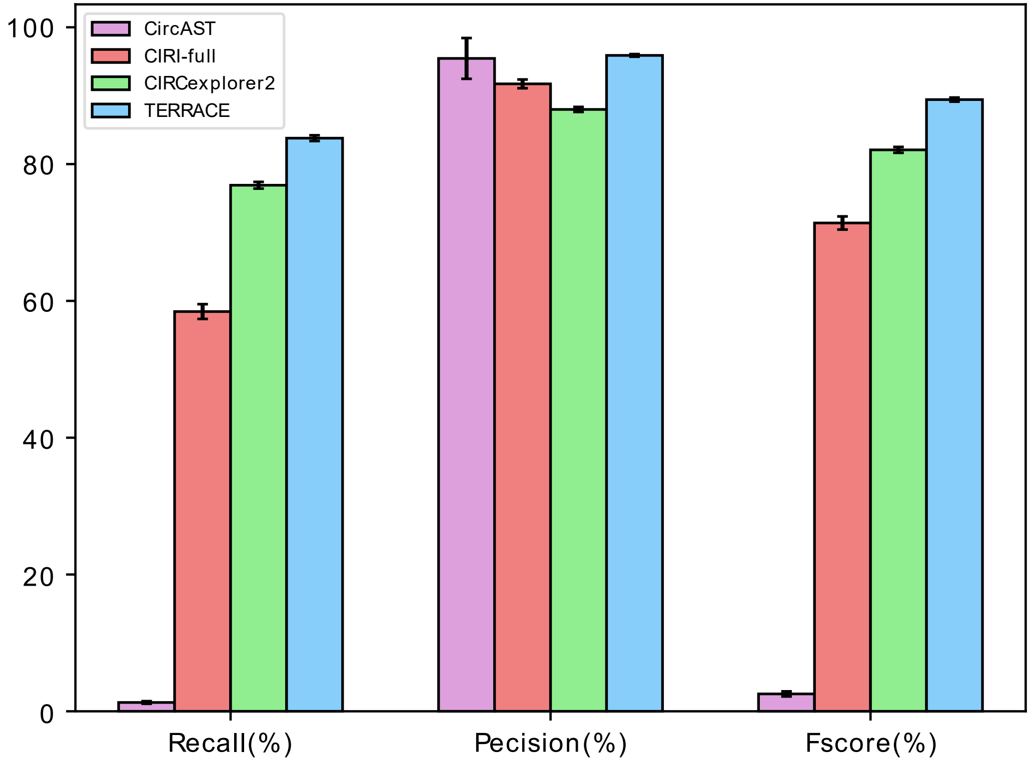
with annotation

Assembly Accuracy on Simulated Data

without annotation



with annotation



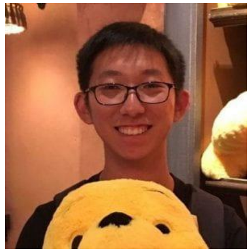
Summary

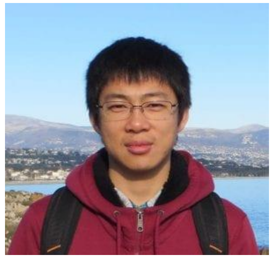
- We present TERRACE, a full-length circular RNA assembler.
- Four algorithmic innovations: identifying accurate back-spliced reads, formulating assembly into bridging, designing new heuristics for path selection, assigning confidence scores to circRNAs.
- TERRACE allows accurate detection of circular RNAs without requiring reference gene annotation, particularly useful for species lacking well-annotated transcriptomes.
- Tool availability: https://github.com/Shao-Group/TERRACE.

Acknowledgements

- Co-authors


Qian Shi


Carl Zang


Mingfu Shao

- Funding support

DBI-2019797
and 214517

R01HG011065

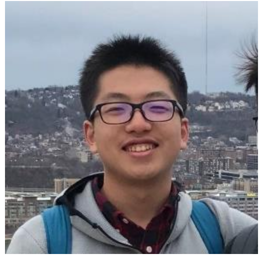
- Members of Shao Group


Ke Chen


Xin Yuan


Qimin Zhang


Xiang Li


Zhezheng Song


Abhishek Talesara


Saadya Rao

Thank You!

Run Time and Memory Usage

CPU time (minutes) for different tools on human tissues

sample	**methods w/o annotation**		**methods with annotation**			
	CIRI-full	TERRACE	CIRCexplorer2	CircAST	CIRI-full	TERRACE
Lung	265	29	0.14	249	266	26
Brain	471	42	0.21	1234	489	45
Skeletal	334	40	0.1	260	337	50
Heart	380	32	0.1	488	389	35
Testis	427	39	0.14	922	472	40
Liver	330	31	0.08	423	345	31
Kidney	352	31	0.1	554	370	32
Prostate	253	52	0.24	229	307	57
Average	**352**	**37**	**0.14**	**545**	**372**	**40**

Peak memory usage (GB) for different tools on human tissues

sample	**methods w/o annotation**		**methods with annotation**			
	CIRI-full	TERRACE	CIRCexplorer2	CircAST	CIRI-full	TERRACE
Lung	10.1	16.9	0.14	0.07	11	16.7
Brain	80.4	11.2	0.16	0.11	81.1	11.4
Skeletal	22.2	19.8	0.14	0.1	22.8	20.2
Heart	30.8	10.9	0.14	0.61	31.4	11.1
Testis	74.8	5.5	0.15	0.09	76.2	5.6
Liver	271.9	17.7	0.14	0.08	31.7	17.9
Kidney	32	10.3	0.14	0.08	33.1	10.1
Prostate	21.7	23.4	0.15	0.08	22.4	23.6
Average	68	14.5	0.15	0.15	38.7	14.6

pAUC (without annotation)

sample	**TERRACE vs CIRI-full**			**TERRACE vs. CIRI-full**		
	constrained by recall			*constrained by precision*		
	TERRACE	CIRI-full	Δ%	TERRACE	CIRI-full	Δ%
Lung	65.2	52.3	24.6	66.8	14.3	365.9
Brain	1380.1	1047.8	31.7	969.7	339	185.9
Skeletal	75.7	62.5	21	59.9	17.3	246
Heart	347.6	294.4	18	334.2	113	195.5
Testis	851.9	645.4	31.9	618.4	196.4	214.9
Liver	527.1	450.4	17	255.9	136.8	87
Kidney	553.9	483.3	14.6	267.9	133.8	100.1
Prostate	495.7	364.1	36.1	598	161.4	270.3

Statistics

sample	**# reads**	**#circRNAs**	**w/o annotation**				**With annotation**			
			TERRACE		CIRI-full		TERRACE		CIRCexplorer2	
			# detected	# correct	# detected	# correct	# detected	# correct	# detected	# correct
Lung	87M	18136	1388	810	606	158	1798	1033	1608	872
Brain	82M	35801	33785	12428	12024	5553	35365	12835	30611	10754
Skeletal	93M	10908	805	387	390	112	1053	494	983	434
Heart	79M	11223	3692	1670	1032	456	4113	1815	3770	1591
Testis	90M	42633	26509	11333	9070	4195	27329	11603	21740	9188
Liver	87M	11978	5314	1951	1533	774	5588	2010	4989	1744
Kidney	93M	22521	9176	3915	2791	1494	9869	4115	8747	3554
Prostate	83M	8114	6342	1942	2029	496	6973	2081	6469	1794