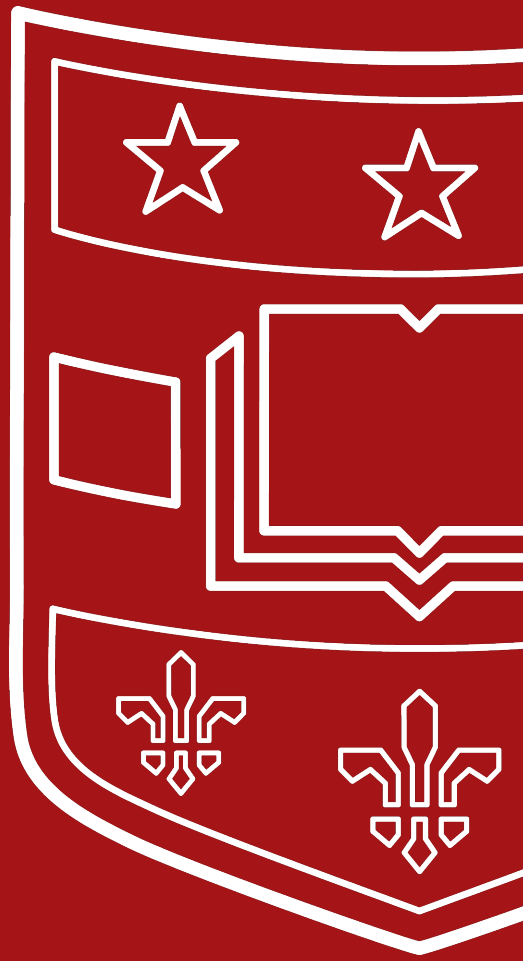
Noncoding RNAs

Christopher Maher

Associate Professor, Internal Medicine and Biomedical Engineering

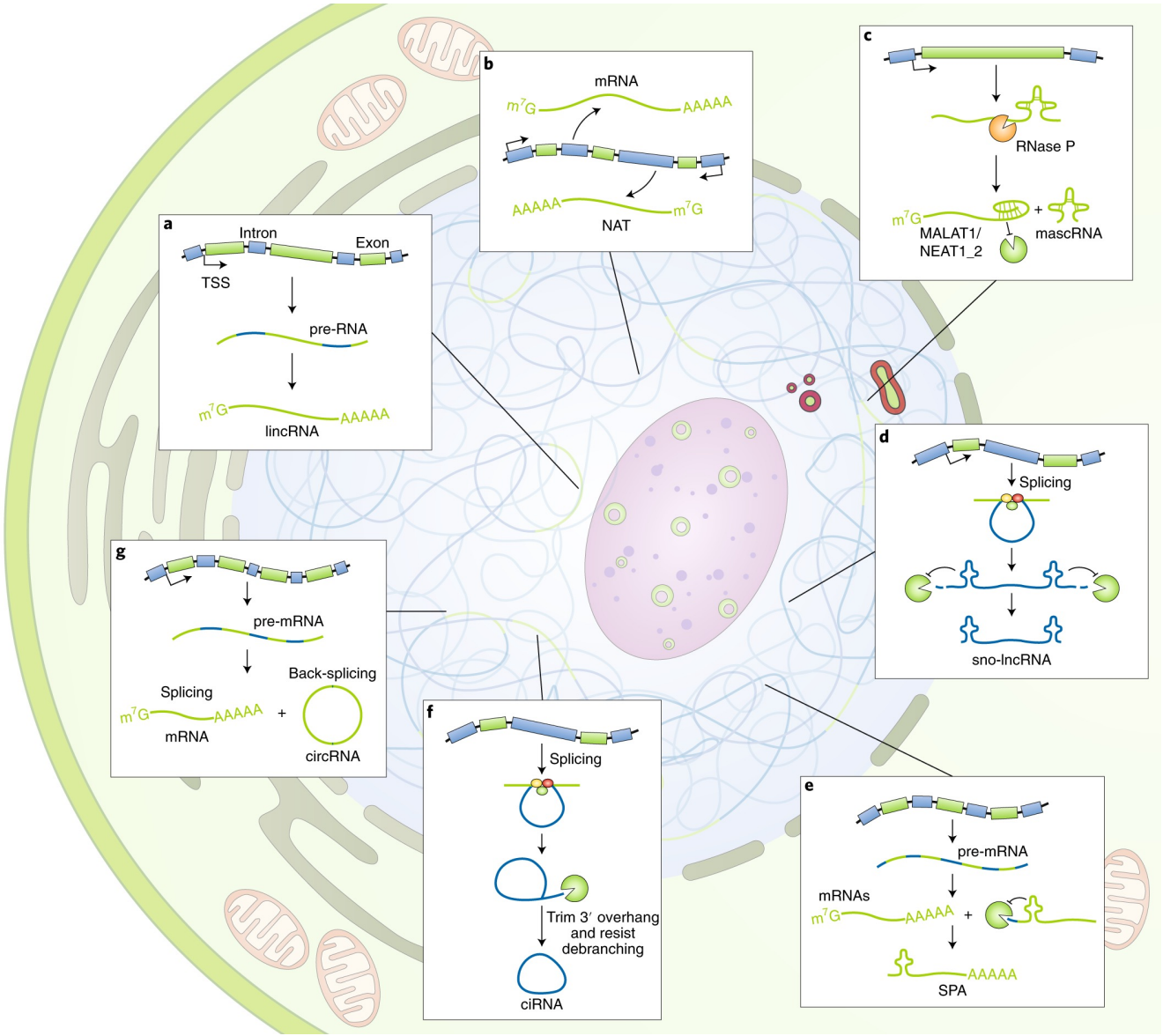
Washington University School of Medicine



November 10th, 2022

CSHL Advanced Sequencing Technologies & Bioinformatics Analysis Course

Diversity of noncoding RNAs



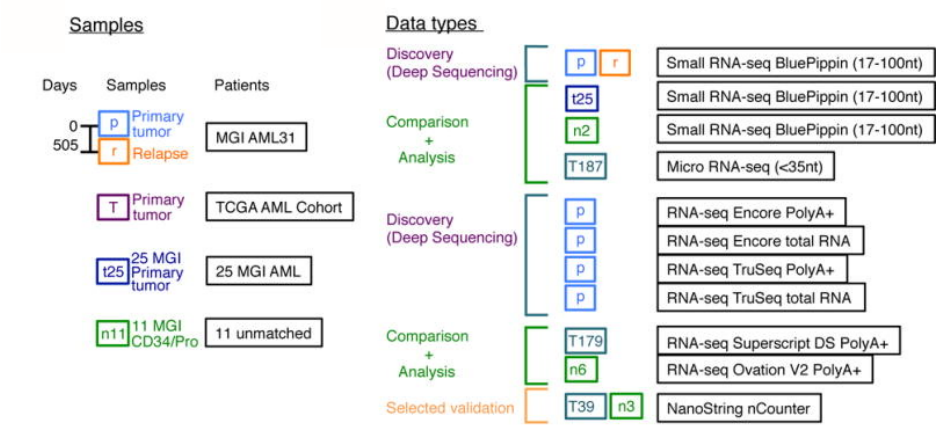
Classes of non-coding RNAs

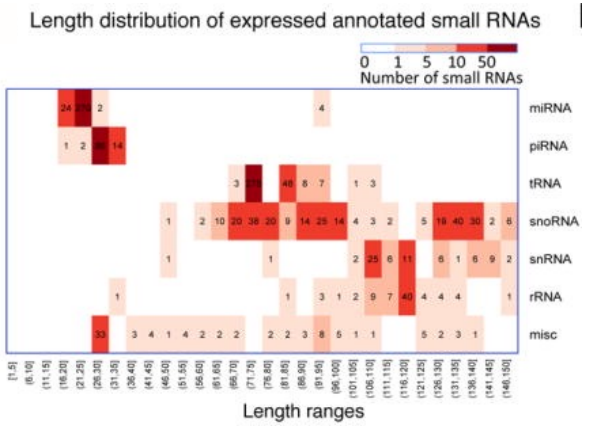
Category	Name	Quality of supporting data	Specific role in carcinogenesis	Aberration in cancer
Housekeeping RNAs	Transfer RNAs	High	No	No
	Ribosomal RNAs	High	No	No
	Small nucleolar RNAs	High	No	No
	Small nuclear RNAs	High	No	No
Small ncRNAs (<200 bp in size)	MicroRNAs	High	Yes	Amplification, deletion, methylation, gene expression
	Tiny transcription initiation RNAs	High	Not known	Not known
	Repeat-associated small interfering RNAs	High	Not known	Not known
	Promoter-associated short RNAs	High	Not known	Not known
	Termini-associated short RNAs	High	Not known	Not known
	Antisense termini-associated short RNAs	High	Not known	Not known
	Transcription start site antisense RNAs	Moderate	Not known	Not known
	Retrotransposon-derived RNAs	High	Not known	Not known
	3′UTR-derived RNAs	Moderate	Not known	Not known
	Splice-site RNAs	Poor	Not known	Not known
Long ncRNAs (> 200 bp in size)	Long or large intergenic ncRNAs	High	Yes	Gene expression, translocation
	Transcribed ultraconserved regions	High	Yes	Gene expression
	Pseudogenes	High	Yes	Gene expression, deletion
	Enhancer RNAs	High	Yes	Not known
	Repeat-associated ncRNAs	High	Not known	Not known
	Long intronic ncRNAs	Moderate	Not known	Not known
	Antisense RNAs	High	Yes	Gene expression
	Promoter-associated long RNAs	Moderate	Not known	Not known
	Long stress-induced noncoding transcripts	Moderate	Yes	Gene expression

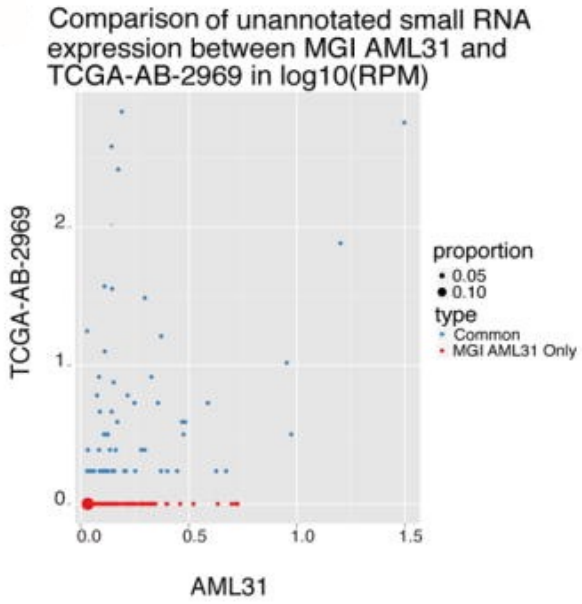
- Existing small noncoding RNA analysis tools are optimized for processing short sequencing reads (17-35 nucleotides) to monitor microRNA expression.
- These strategies under-represent many biologically relevant classes of small noncoding RNAs in the 36-200 nucleotides length range (tRNAs, snoRNAs, etc.)

(Cancer Discovery - Prensner et al., 2011)

Discovered previously unannotated small RNAs using deep sequencing of a libraries with broader insert size selection







Existing small RNA analysis tools were not intended to analyze sequence reads of varying lengths, handle larger quantities of sequence reads, or support for diverse small RNA species

(Zhang et al., 2017)

DANSR: A Tool for the Detection of Annotated and Novel Small RNAs

Raw instrument data (FASTQ)

I. Remove adapter sequences (Cutadapt)

Reads with various lengths (FASTQ)

II. Align to reference genome (BWA)

III. Cluster overlapping read (BEDtools)

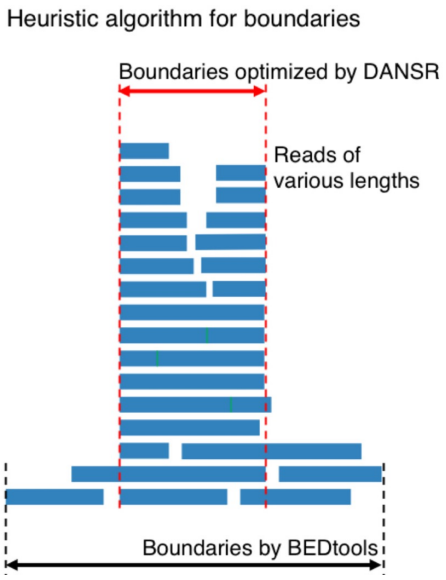
Small RNA read clusters (BED file)

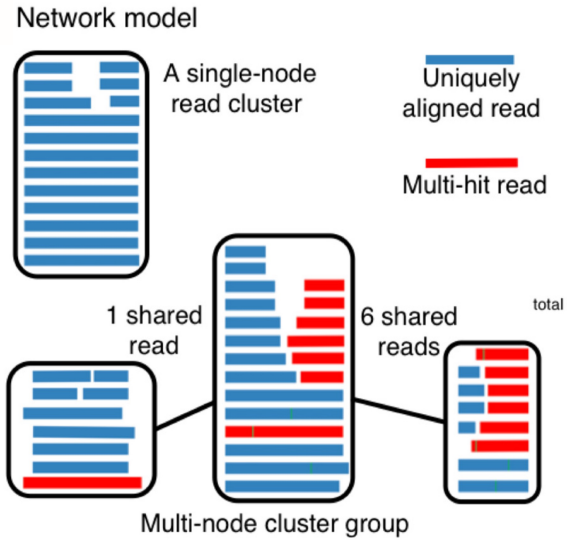
IV. Optimize small RNA boundaries (Heuristic algorithm; Panel B)

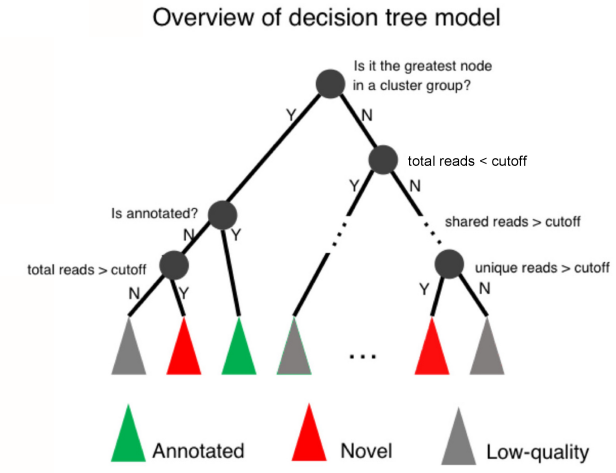
V. Identify single- and multi-node clusters (Network model; Panel C)

VI. Identify annotated/ unannotated small RNAs (Decision tree model; Panel D)

Annotated/ unannotated small RNAs (BED file)

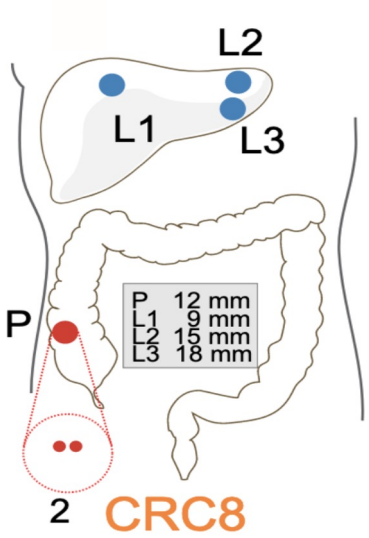


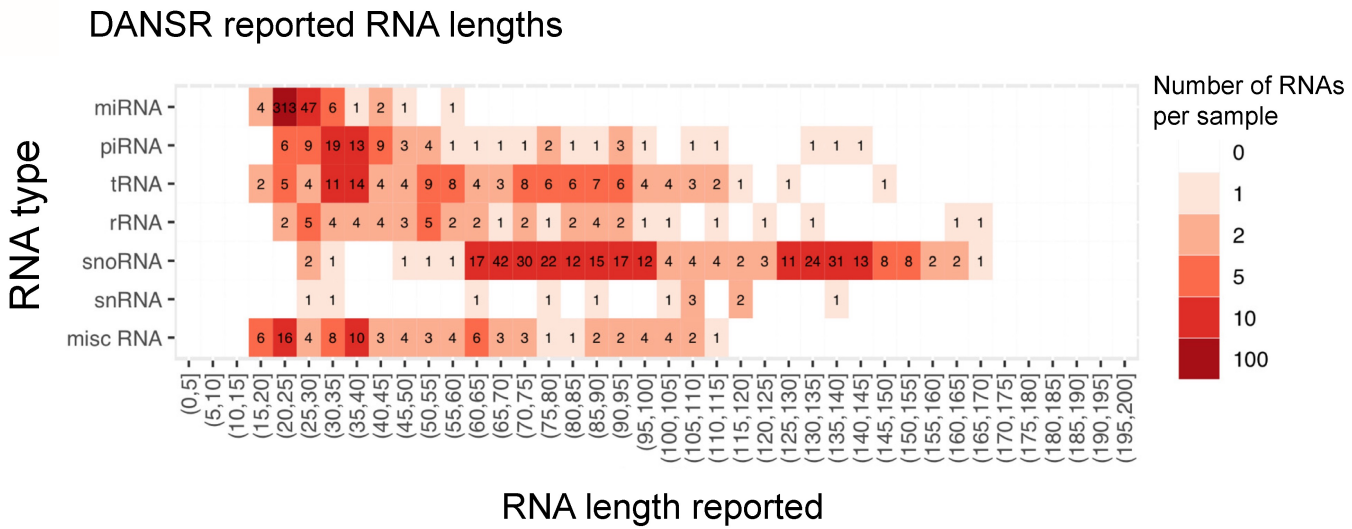




https://github.com/ChrisMaherLab/DANSR
(Eteleeb et al., 2022)

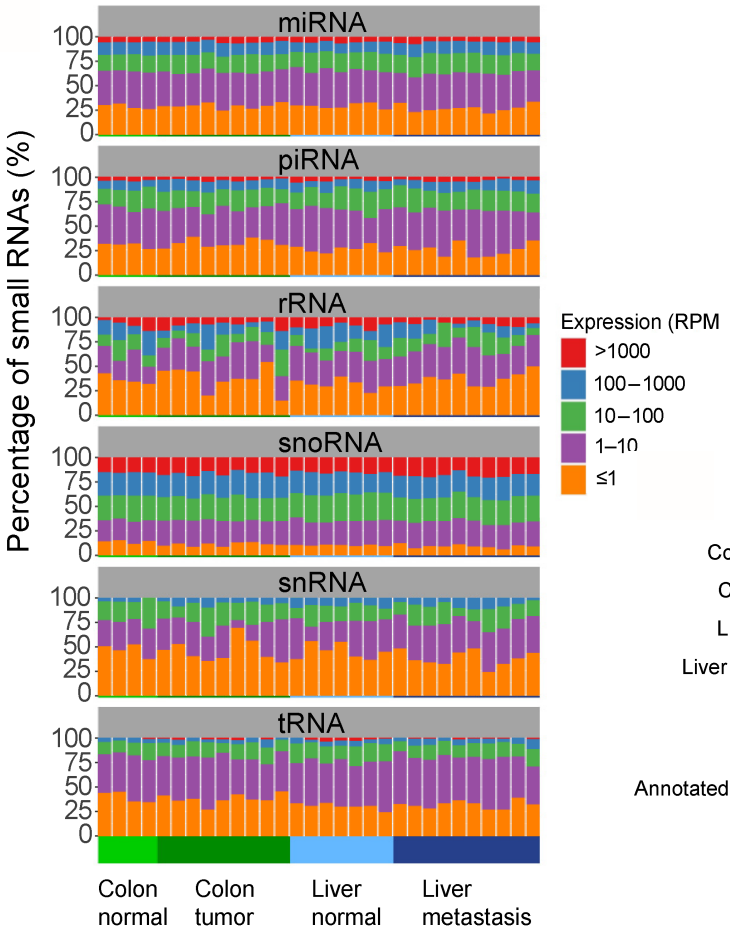
Accurate categorization of annotated small RNAs in metastatic colorectal cancer (mCRC) patients

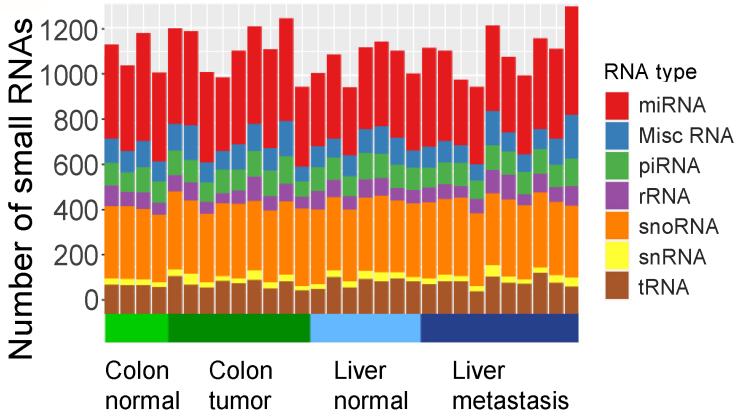


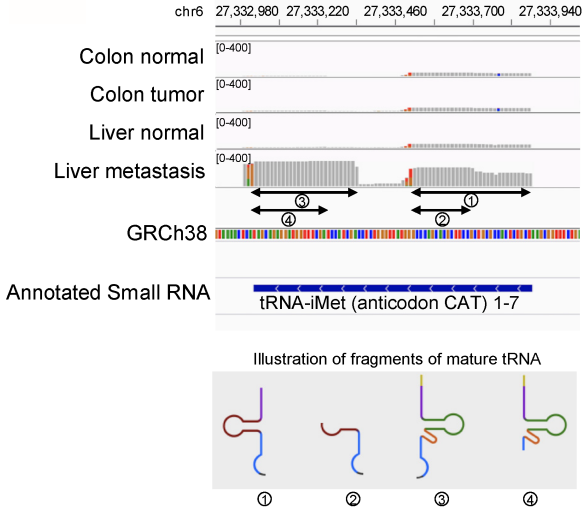


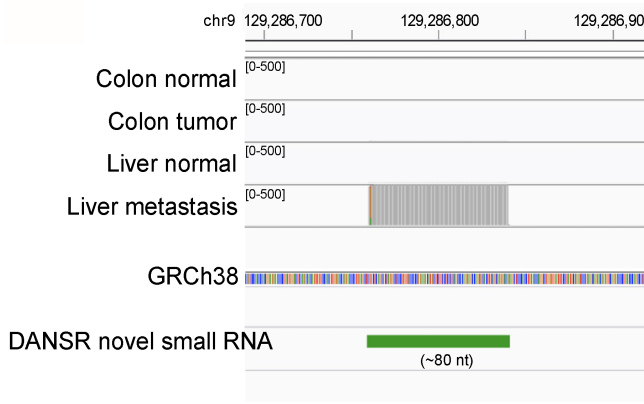
https://github.com/ChrisMaherLab/DANSR
(Eteleeb et al., 2022)

Discovery of altered small RNAs in metastatic colon cancer progression





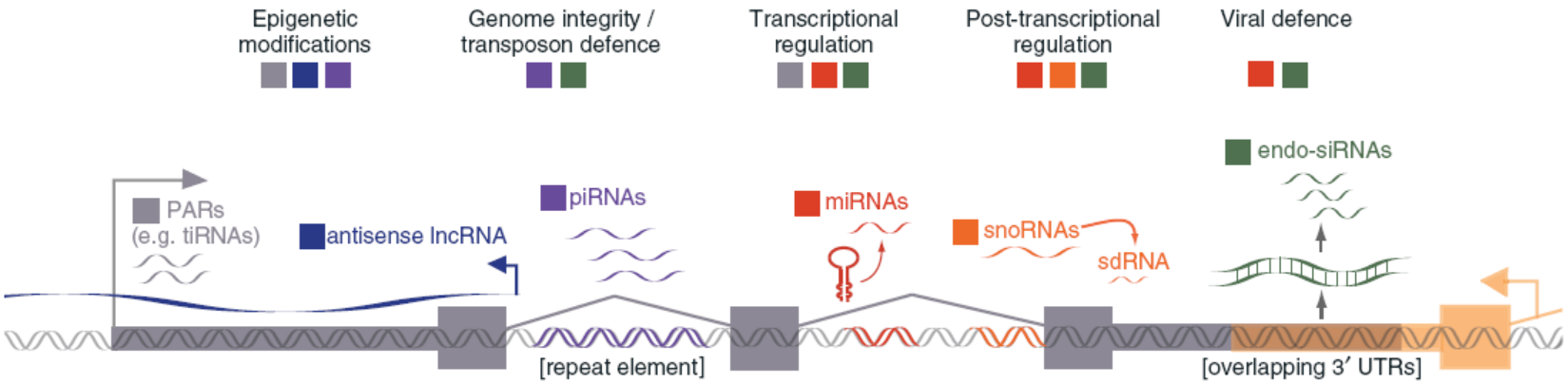




https://github.com/ChrisMaherLab/DANSR
(Eteleeb et al., 2022)

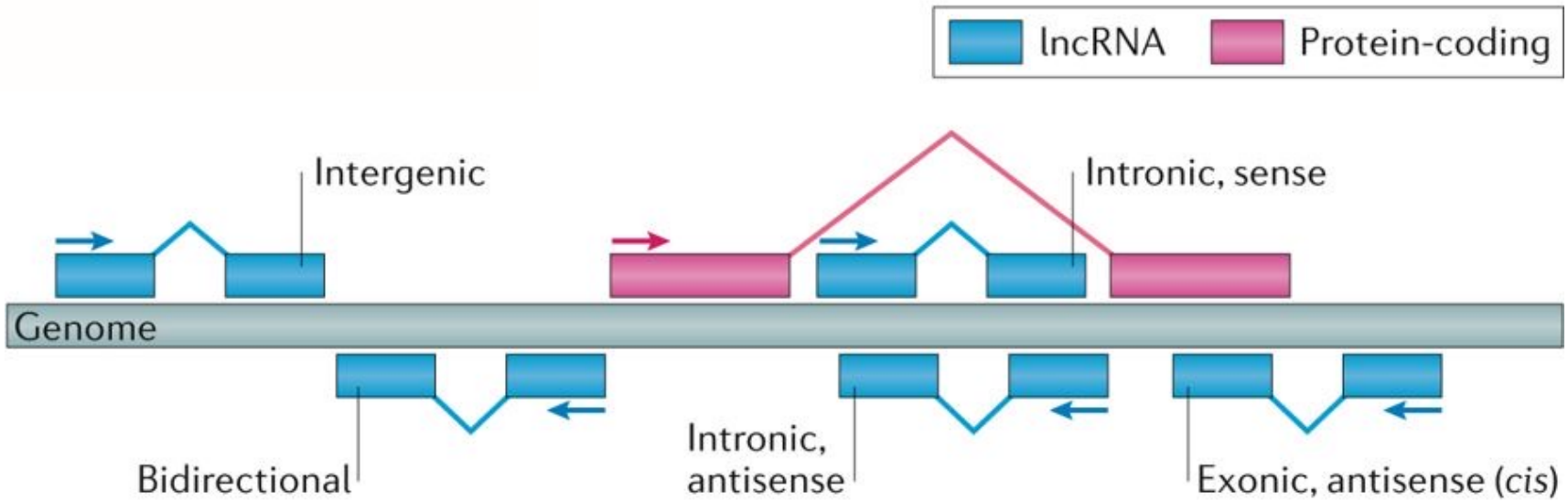
Long noncoding RNAs (lncRNAs)

Characteristics of long non-coding RNAs (lncRNAs)

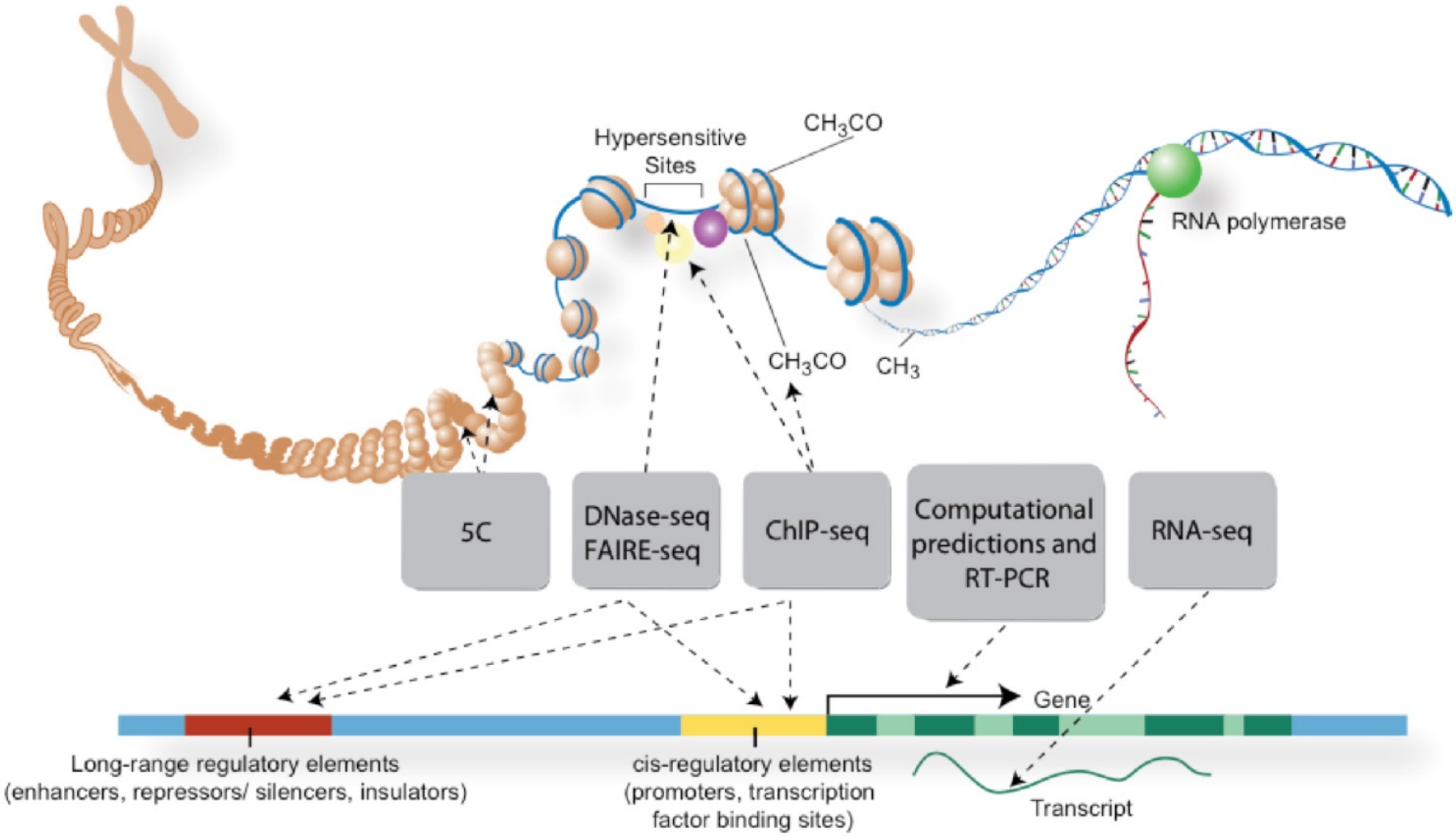


- Transcription via RNA polymerase II
- Polyadenylation
- Frequent splicing of multiple exons via canonical genomic splice site motifs
- Regulation by well-established transcription factors
- Epigenetic marks consistent with a transcribed gene (H3K4me3 at the gene promoter, H3K36me3 throughout the gene body)
- Frequent expression in a tissue-specific manner

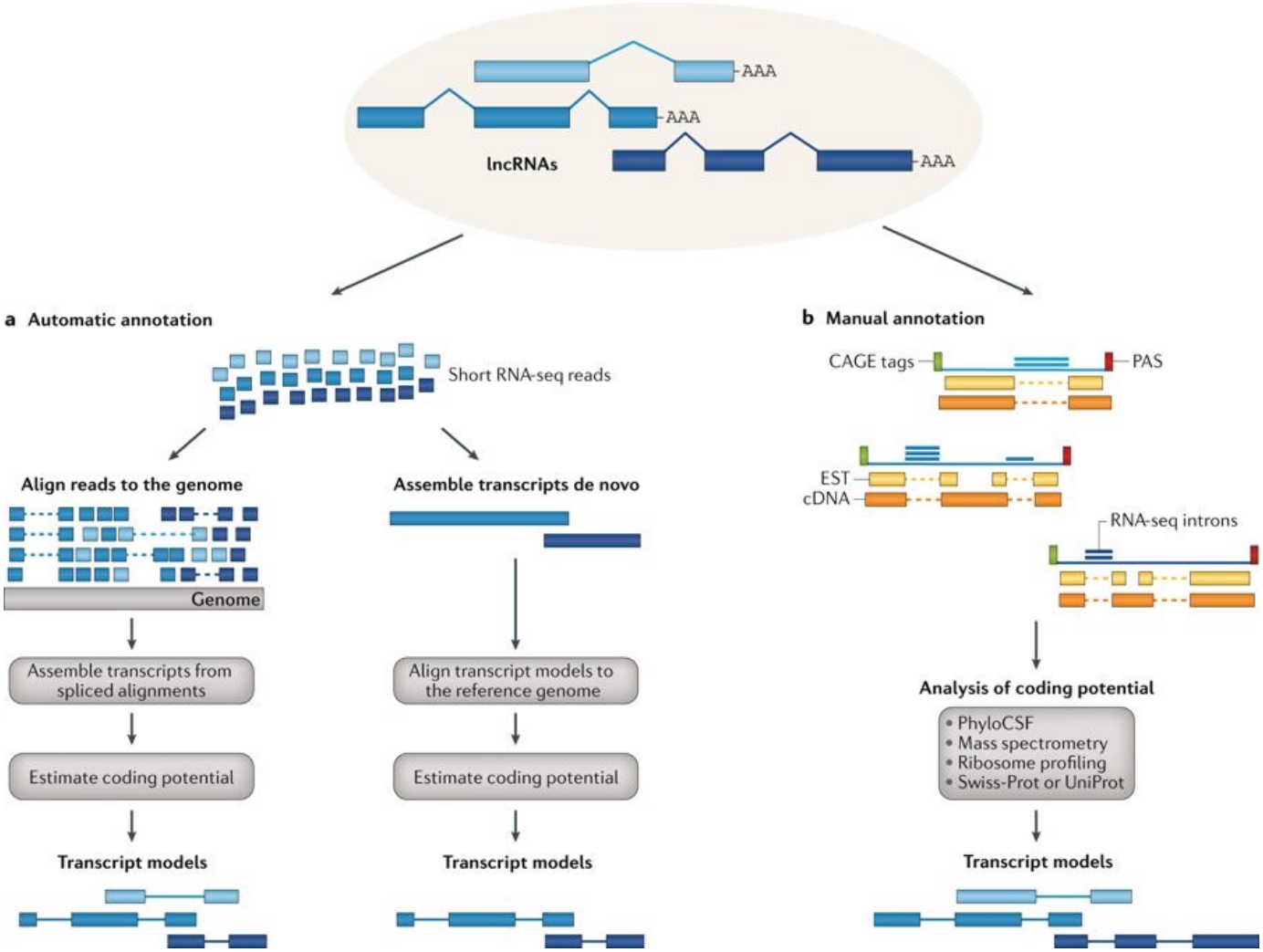
Positional classification of lncRNAs



Integrative methods for discovering lncRNAs



RNA-Seq focused strategies for lncRNA discovery



(Cell -- Bartel et al., 2013)
(Uszczynska-Ratajczak et a., 2018)

How many lncRNAs have been annotated

Table 1 | **lncRNA annotations**

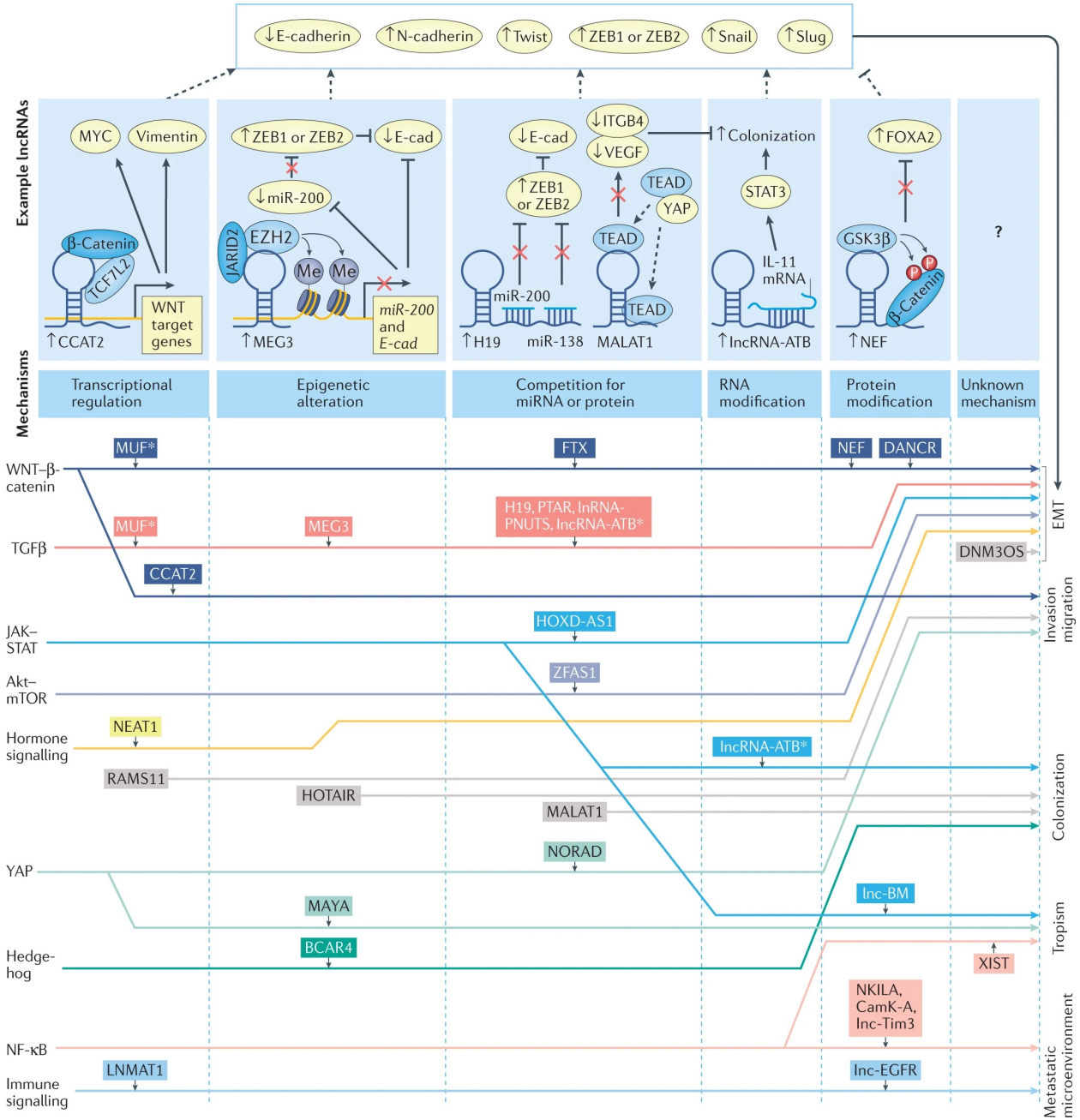
Name (version)	Reported size (gene loci)	Methods[a]	Comments	Completeness	Comprehensiveness[b]	Exhaustiveness[c]
NONCODE (v5)	96,308	Integration of other databases	The most comprehensive resource	8.9%	67,276	2.3
MiTranscriptome (v2)	63,615	Assembly from short reads	Mainly cancer samples	4.4%	45,088	4.4
FANTOM CAT (v1)	27,919	Assembly, other annotations and CAGE evidence	Mapped 5′ ends using CAGE tags	15.8%	27,278	3.3
RefSeq (GCF_000001405.37_GRCh38.p11)	15,791	Manual (based on cDNA) and automated annotation (based on RNA-seq data)	The oldest annotation	11.0%	14,889	1.9
GENCODE (v27)	15,778	Manual annotation based on cDNA, ESTs and high-quality long-read data	Used by most consortia and integrated with Ensembl	13.5%	15,063	1.9
BIGTranscriptome (v1)	14,158	Assembly, with CAGE and 3 P-seq evidence	Full-length transcripts	27.7%	12,632	2.1
GENCODE+	13,434	Union of GENCODE (v20) and CLS lncRNAs with anchor-merged CLS transcript models	Extension of GENCODE by CLS	24.0%	13,434	3.3
CLS FL	807	lncRNAs from GENCODE+ with CAGE and poly(A) evidence	Full-length transcripts	71.7%	807	5.5
Protein-coding[d]	19,502	GENCODE confident protein-coding transcripts	Not tagged *mRNA_end_NF* nor *mRNA_start_NF* in the original GENCODE v27 GTF file	53.8%	18,995	2.9

Comprehensiveness
The fraction of all gene loci that are included;

Exhaustiveness
The fraction of all transcripts from each locus that are known;

Completeness
The fraction of transcript models that cover the entire length, from start to end, of the physical
RNA molecule

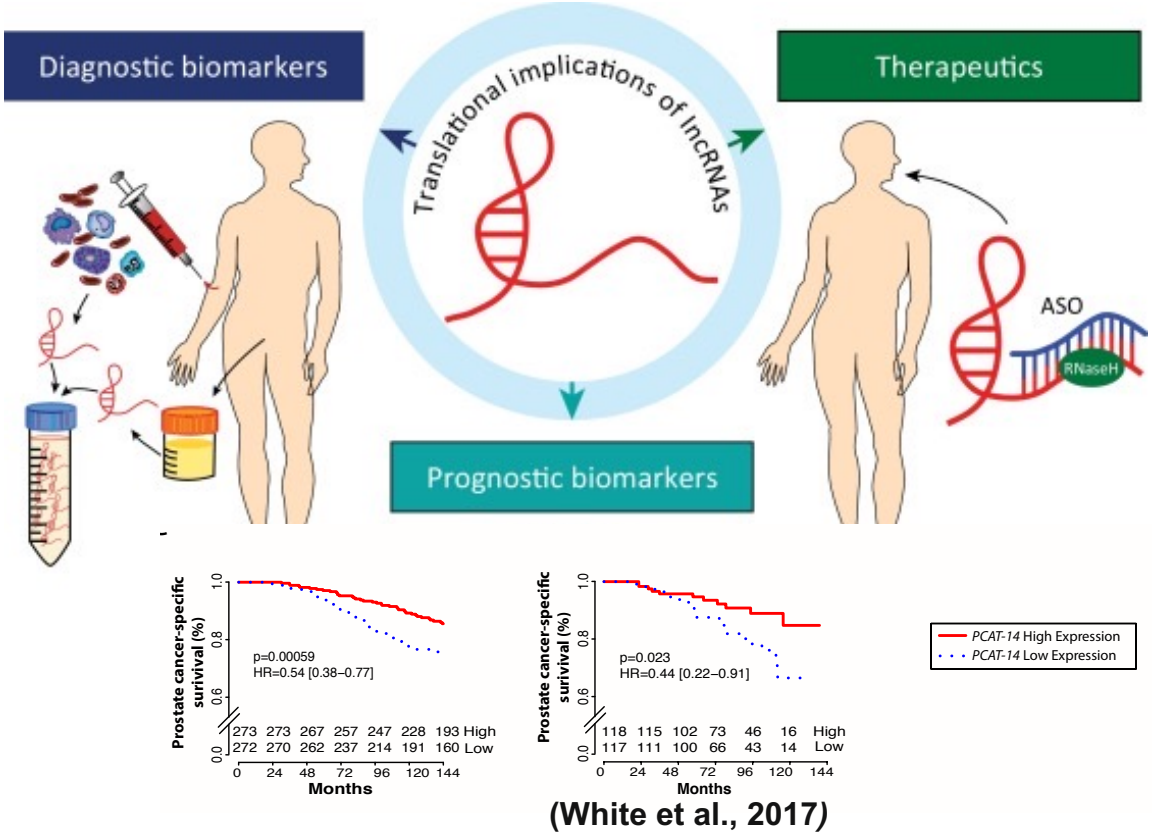
(Uszczynska-Ratajczak et al., 2018)



Long noncoding RNAs regulate metastasis via various pathways using diverse mechanisms

(Nature Reviews Cancer
Liu et al., 2021)

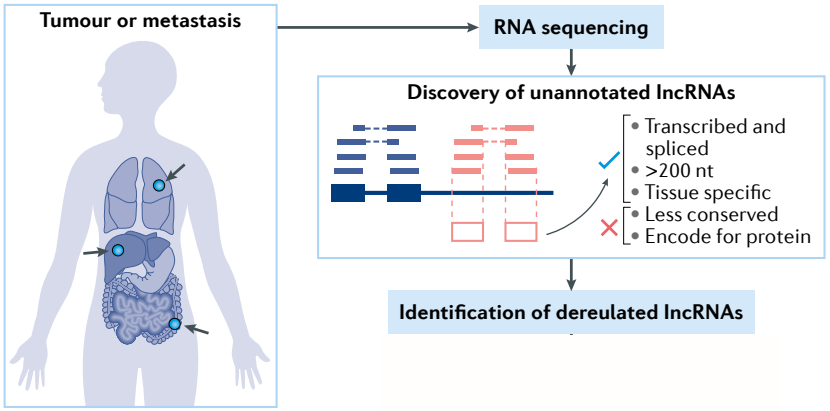
Clinical applications of lncRNAs



(White et al., 2017)

- LncRNAs are emerging as diagnostic/prognostic biomarkers in tissue, serum, and urine
- Antisense oligonucleotides (ASOs) can be used to directly target lncRNAs and are a promising therapeutic strategy in cancer

Despite discovering thousands of lncRNAs, only a minor subset have been well characterized

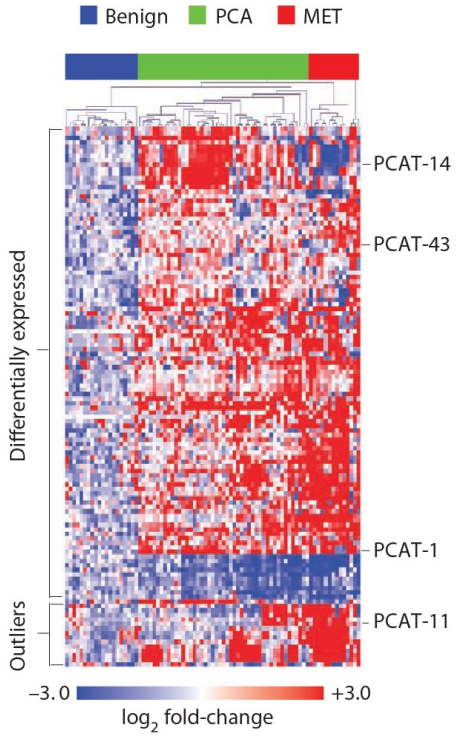


Challenges:

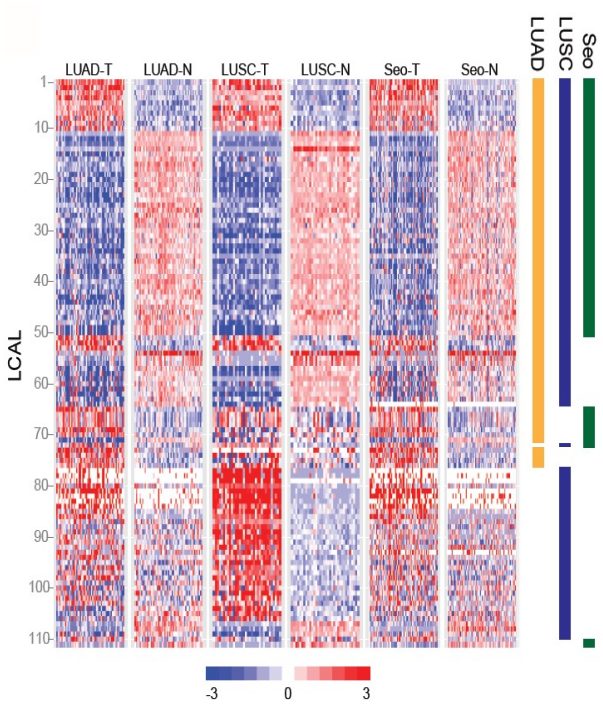
- Prioritizing biologically and clinically relevant lncRNAs
- Lack of “domains” is a barrier for predicting function
- Molecular interrogation is labor intensive

(Nature Reviews Cancer Liu et al., 2021)

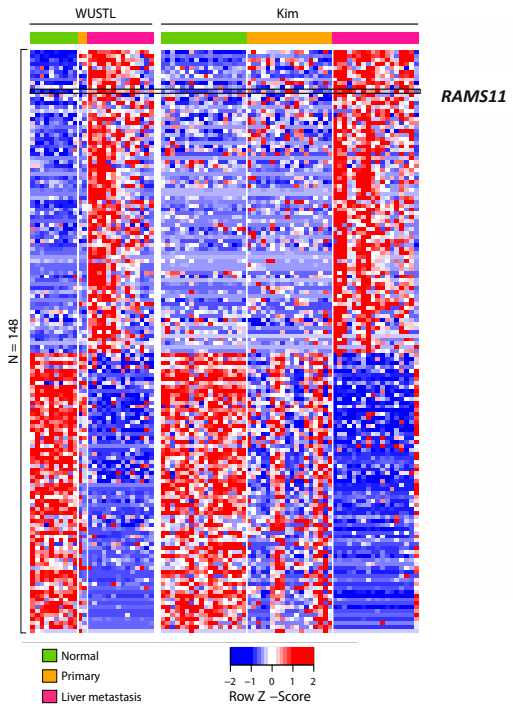
Only a fraction of lncRNAs are altered in a given cancer type



- Analysis 121 prostate cancer patients (normal, primary, and metastatic samples)
- In total, we identified 121 prostate cancer associated transcripts (PCATs)



- Analysis of ~600 LUAD and LUSC cancer patients
- 111 novel transcripts were differentially expressed in at least one histology
- Referred to as **l**ung **c**ancer **a**ssociated **l**ncRNAs (LCALs)



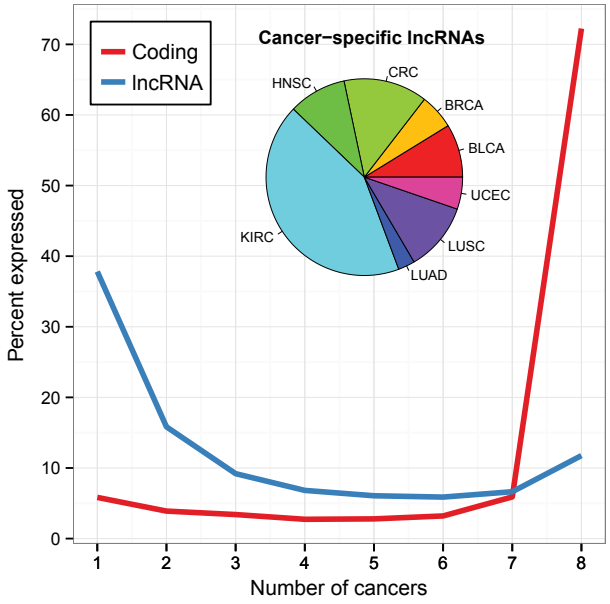
- 148 lncRNAs that performed as well as known biomarkers in differentiating benign, primary, and metastatic tissues
- 51 lncRNAs differentially expressed in metastatic tumors compared to non-metastatic (primary and adjacent normal)
  - 17 Unannotated
- Referred to as **R**NAs **A**ssociated with **M**etastasi**s** (**RAMS**)

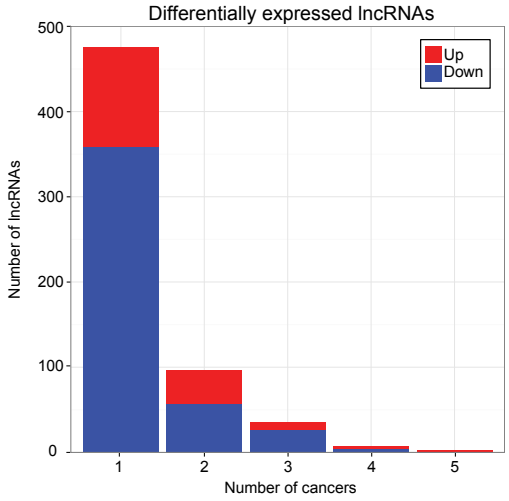
(Nature Biotechnology-- Prensner et al., 2011)

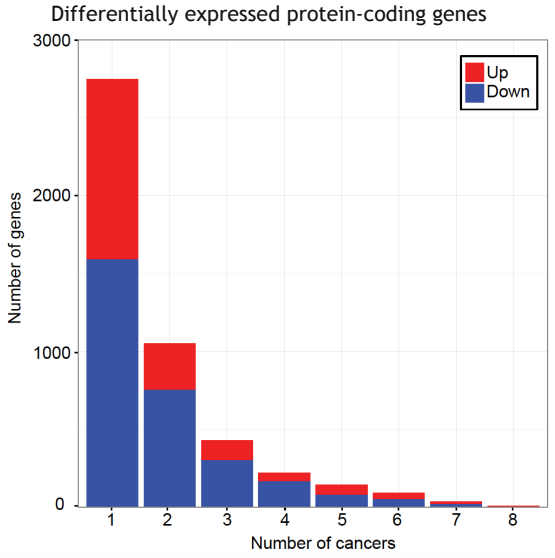
(Genome Biology -- White et al., 2014)

(Nature Communications –Silva et al., 2021)

LncRNAs have greater tissue-specificity in pan-cancer analysis across ~3,000 patients



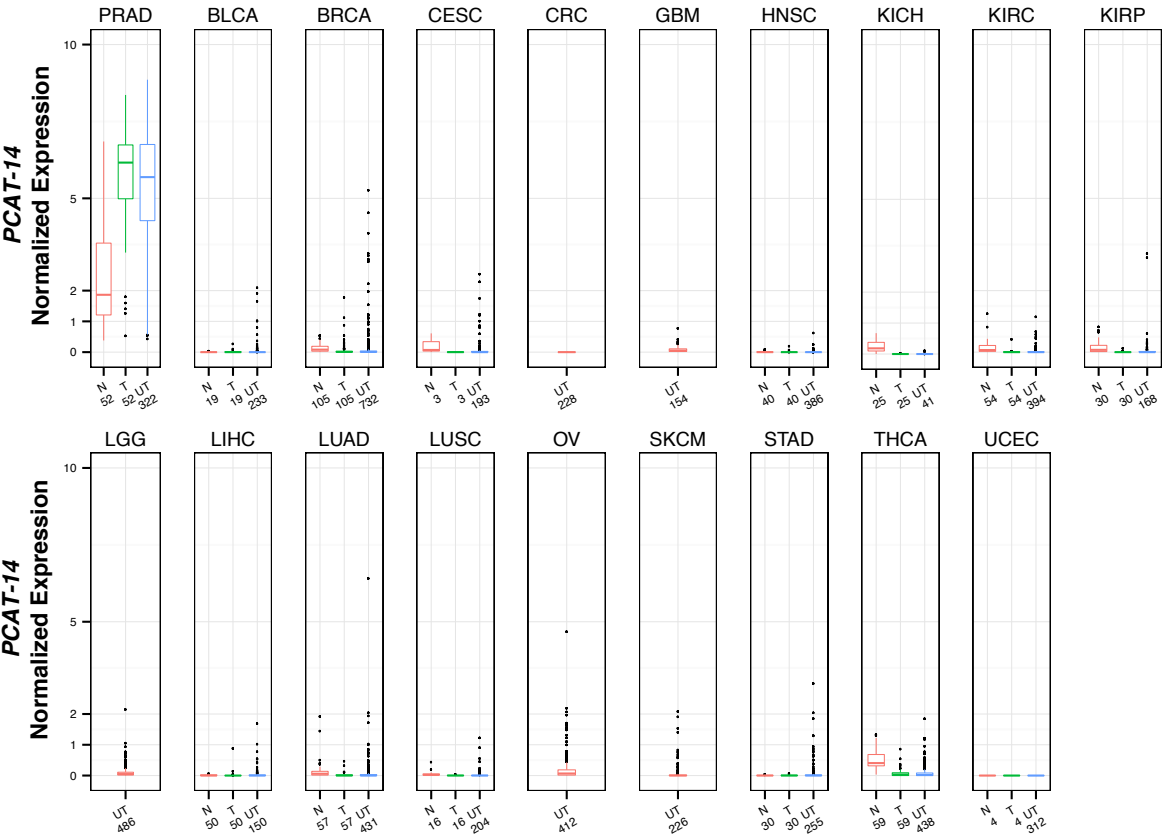


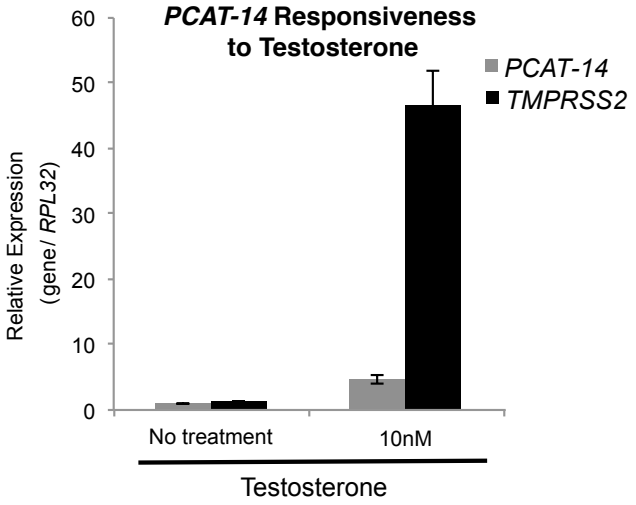


- ~10% of protein-coding genes are altered across 2 or more cancer types
- ~2% of lncRNAs are altered across 2 or more cancer types

(Cabanski et al., 2015)

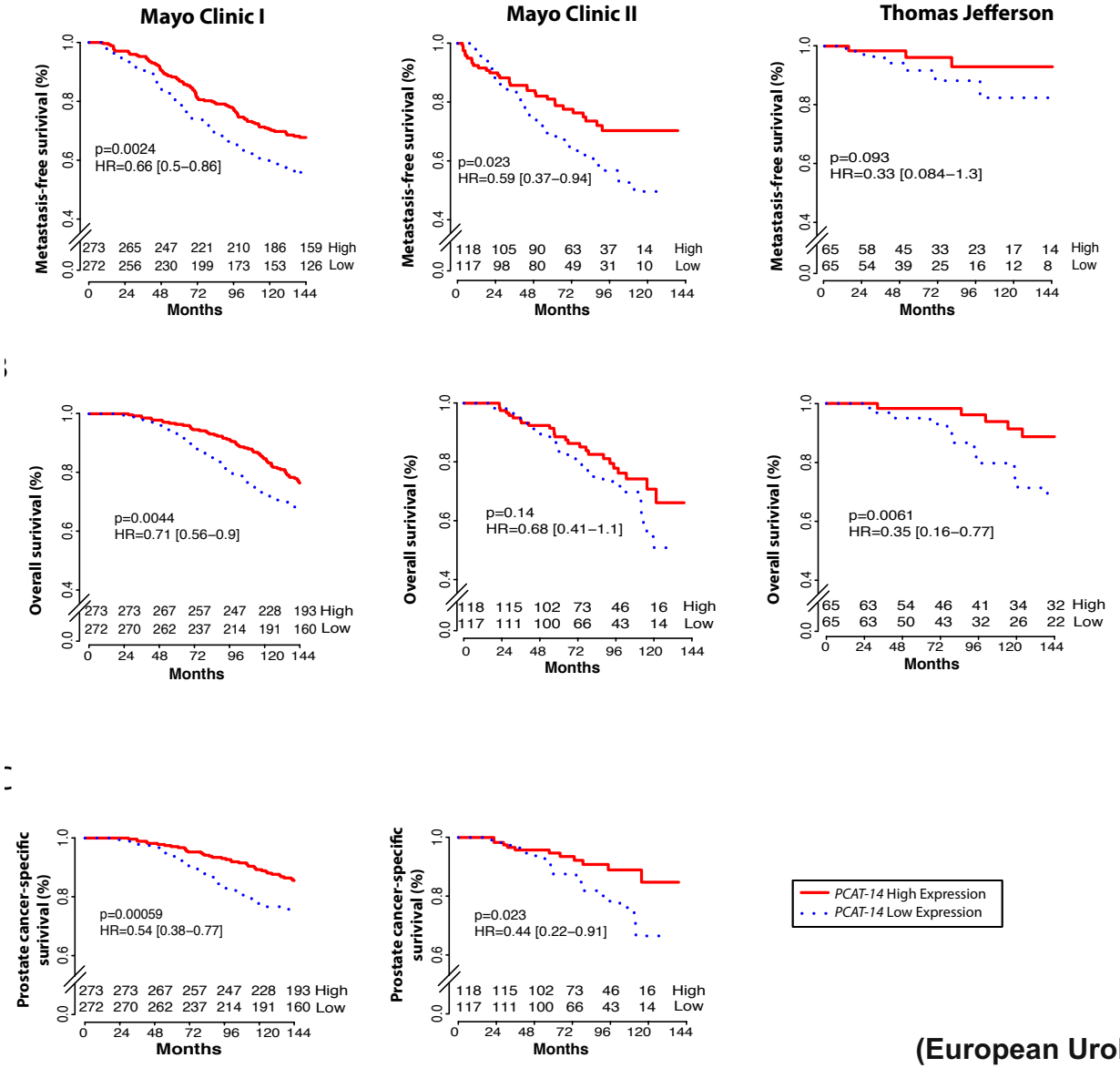
PCAT-14 expression is enriched in prostate cancer





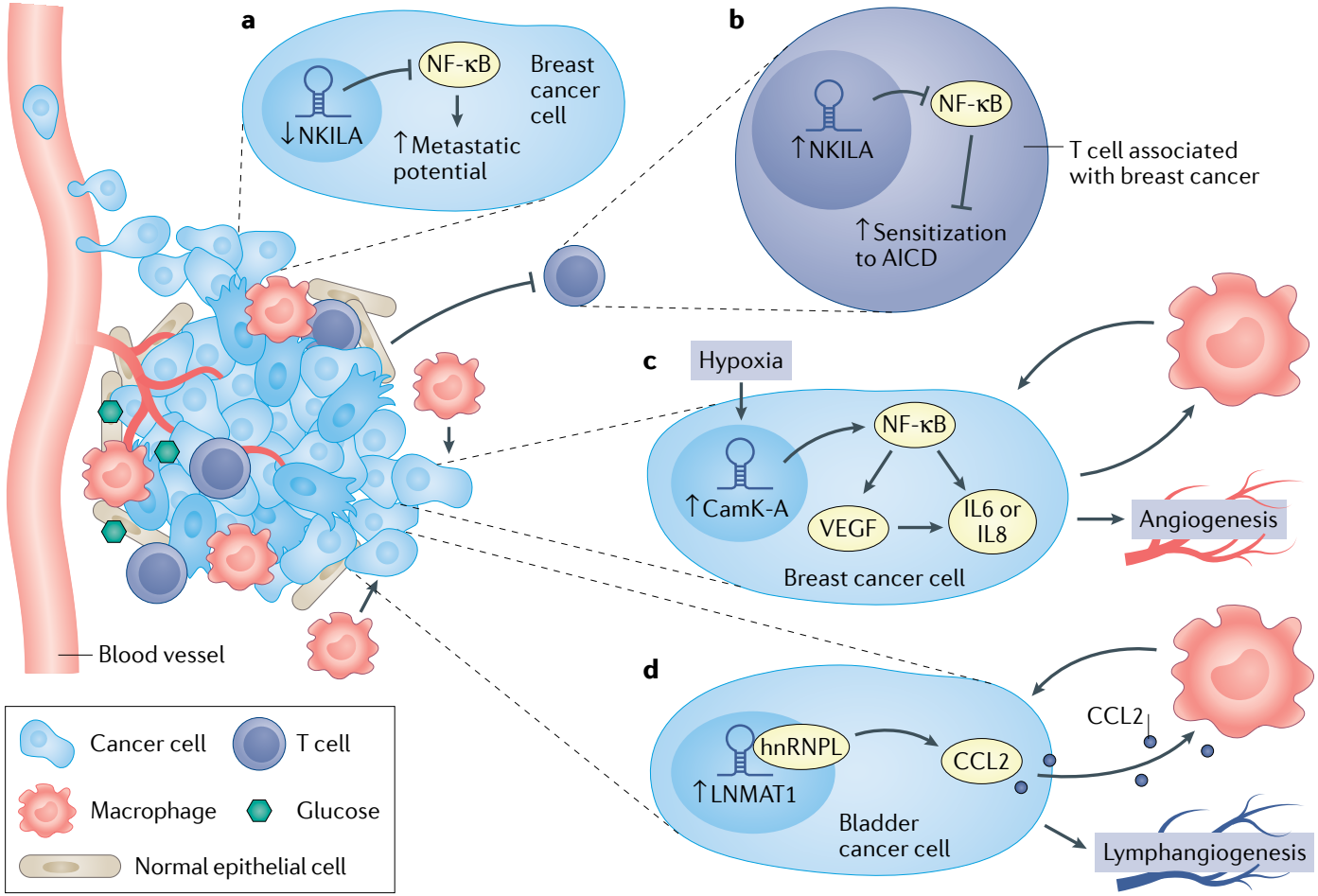
(European Urology -- White et al., 2017)

PCAT-14 as a single gene predictor of aggressive disease



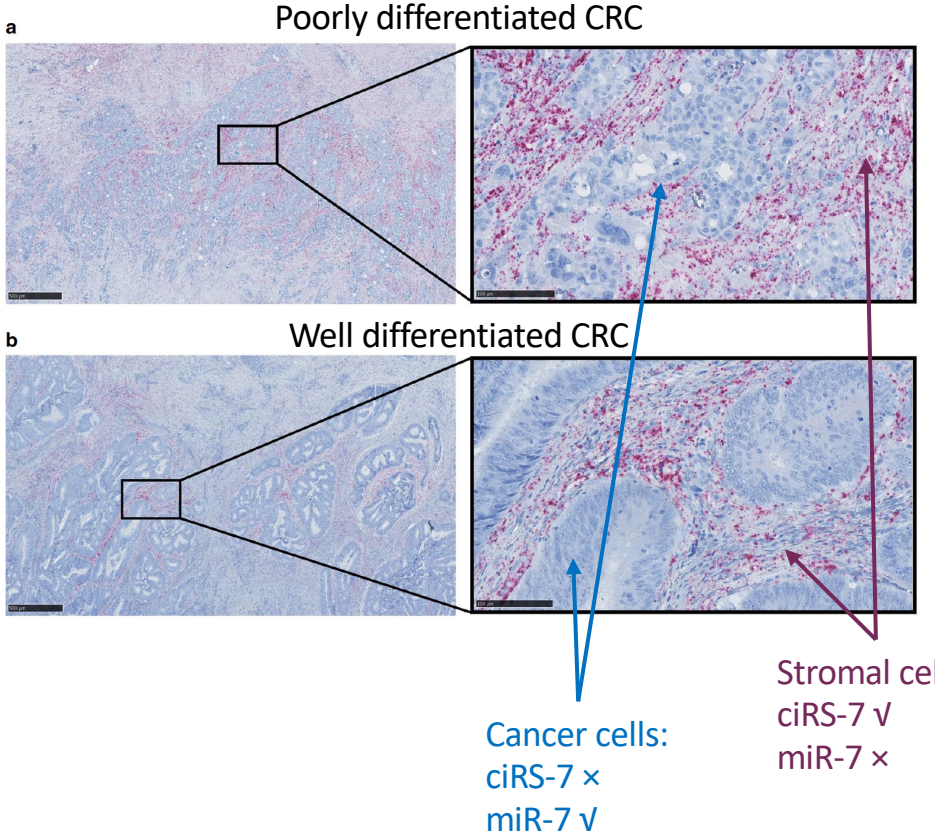
(European Urology -- White et al., 2017)

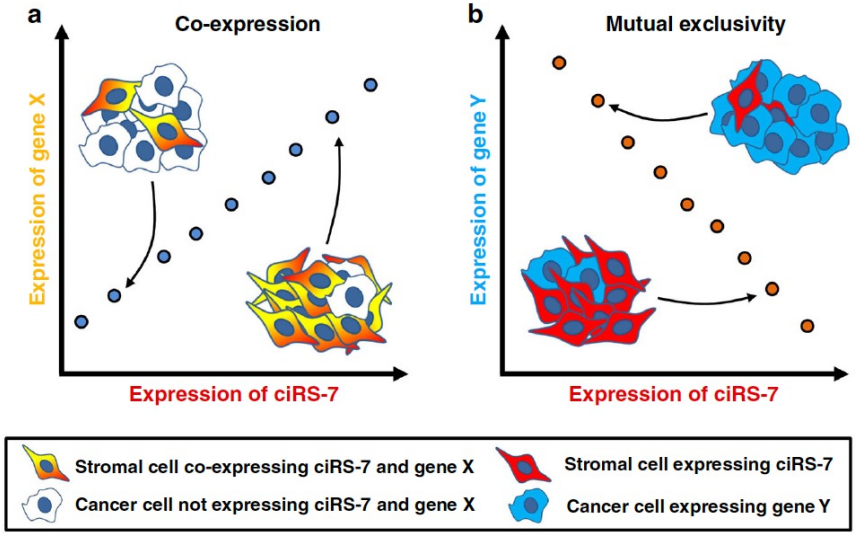
Long noncoding RNAs and the tumor microenvironment



(Nature Reviews Cancer Liu et al., 2021)

Challenge to the miRNA sponge mechanism

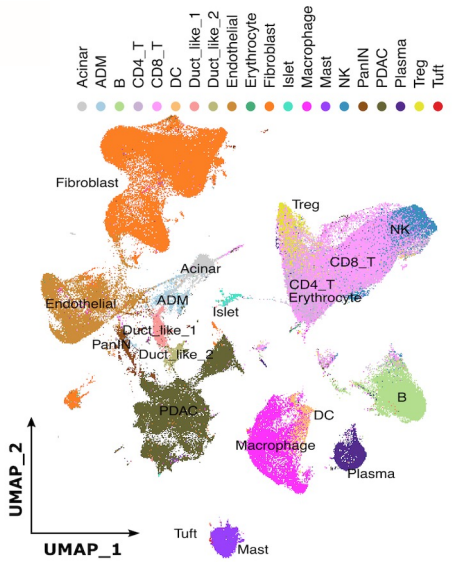




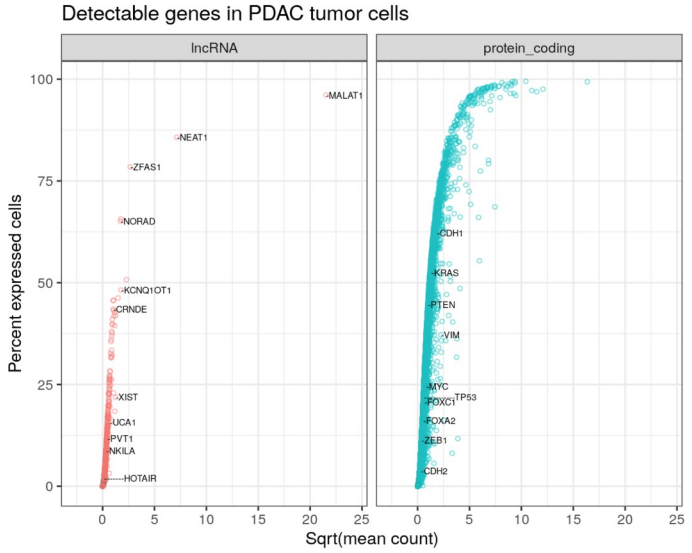
(Kristensen, L.S., et al., 2020)

Analysis of lncRNAs in single cell RNA-Seq data from pancreatic cancer patients

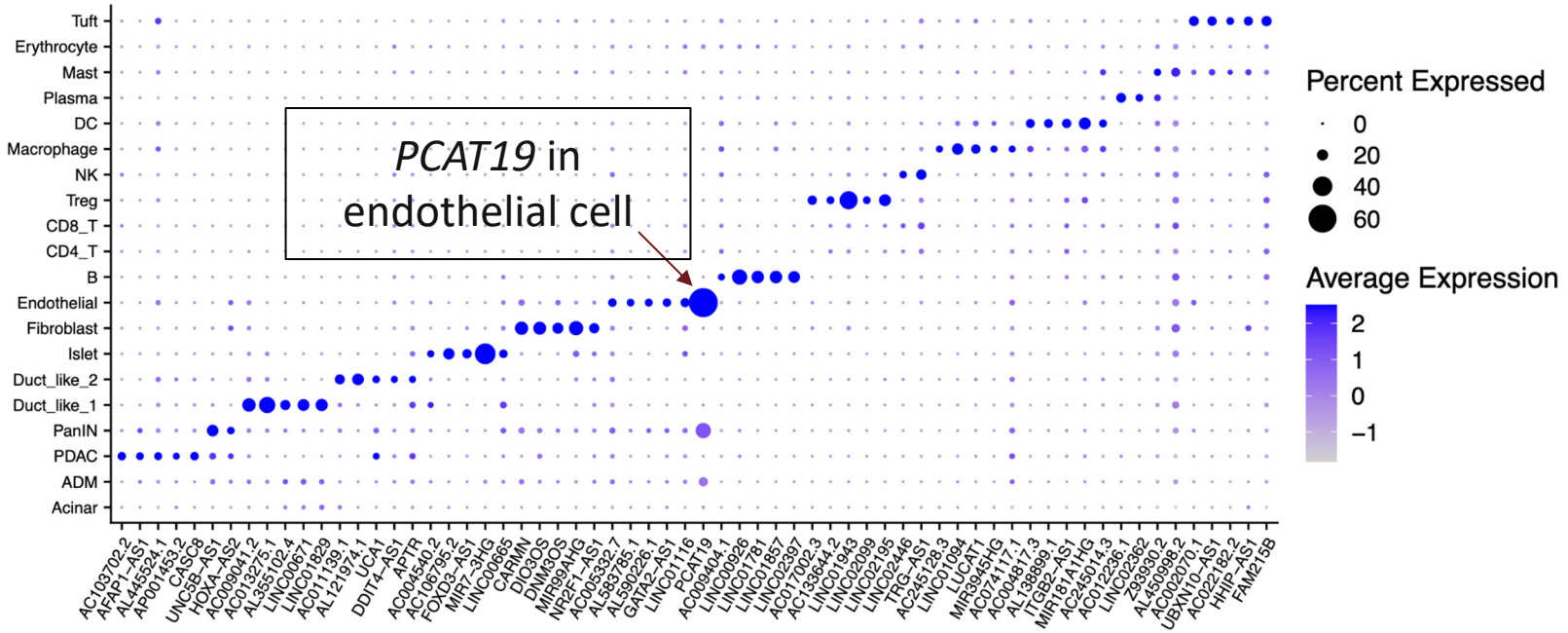
- 73 samples from 21 patients with PDAC
  - 10x Genomics scRNA-Seq data (~50K reads per cell)
  - Various cell types identified in TME including PDAC tumors, immune cells, and stromal cells



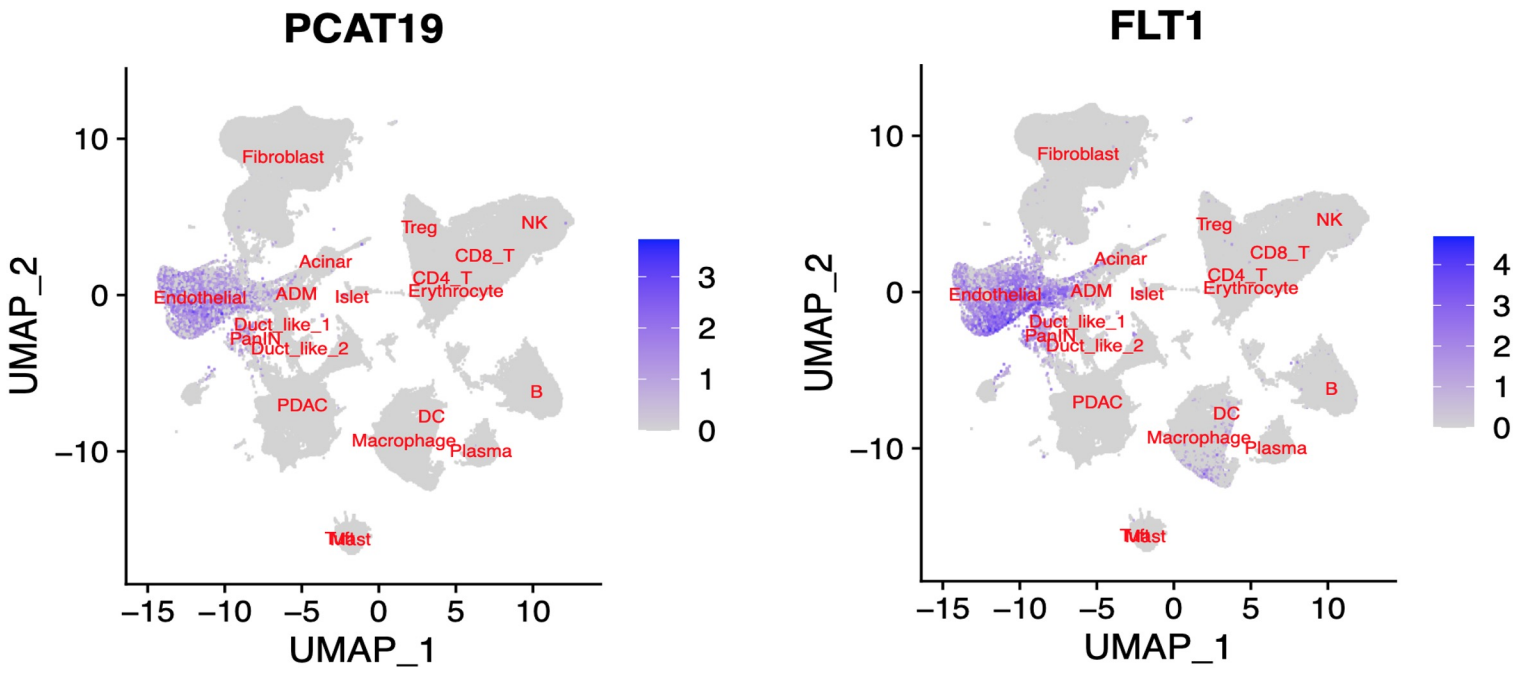
- Most genes are only detected in a small fraction of cells
- LncRNAs are more likely to be missed at individual cell level due to their lower expression level
- Significant number of lncRNAs are detected in as many cells as protein coding cancer genes



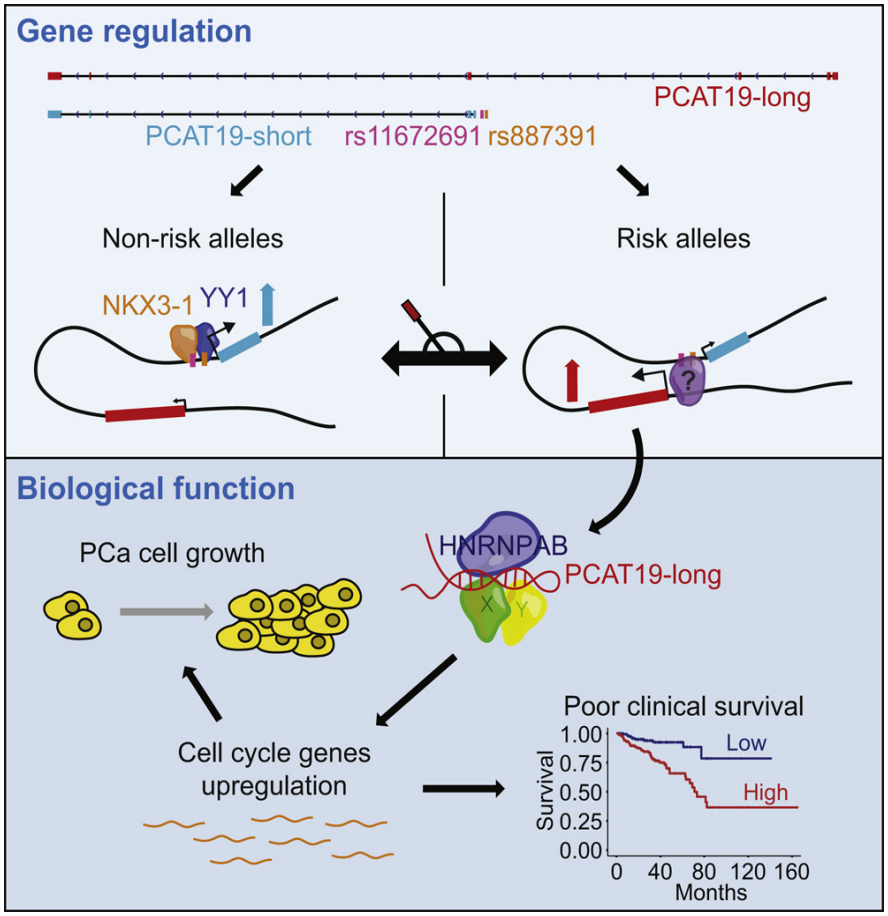
LncRNAs as markers of PDAC TME cell types

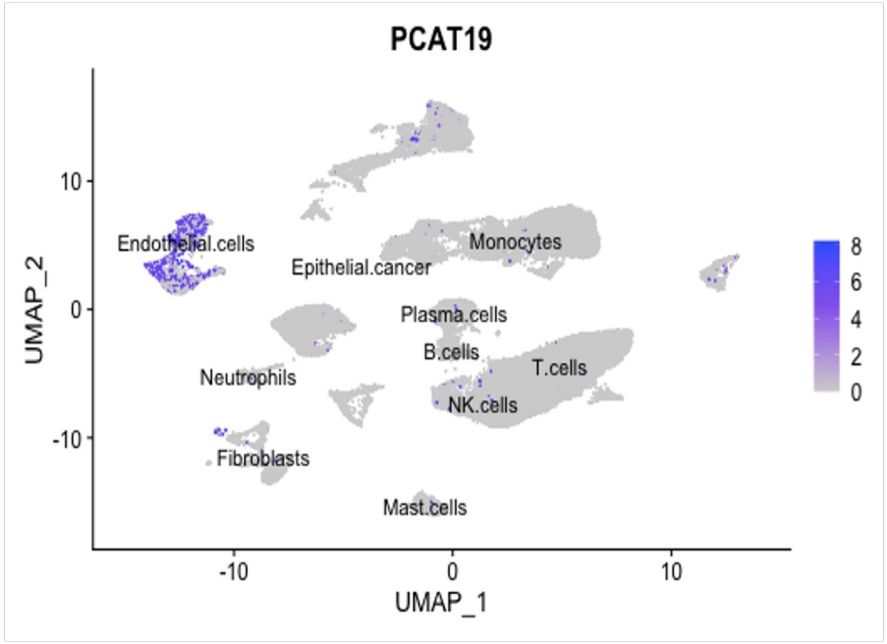


PCAT19: a strong marker of endothelial cells



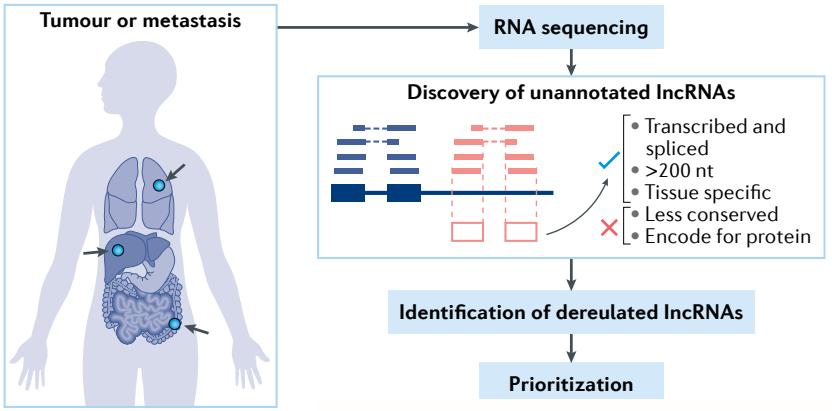
PCAT19 activates a subset of cell-cycle genes associated with PCa progression, thereby promoting PCa tumor growth and metastasis





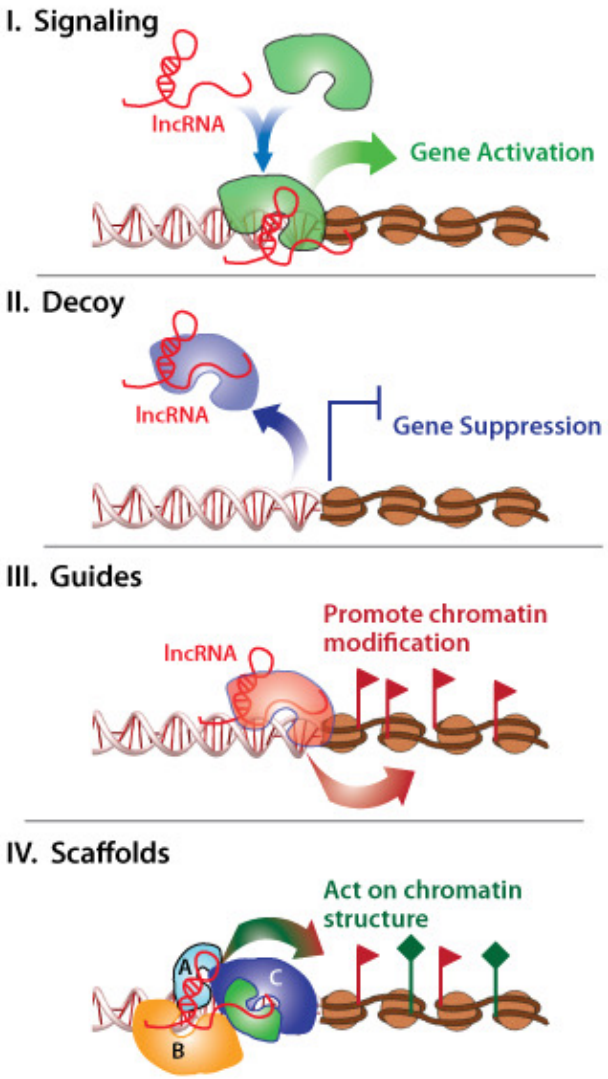
(Cell -- Hua et al., 2018)

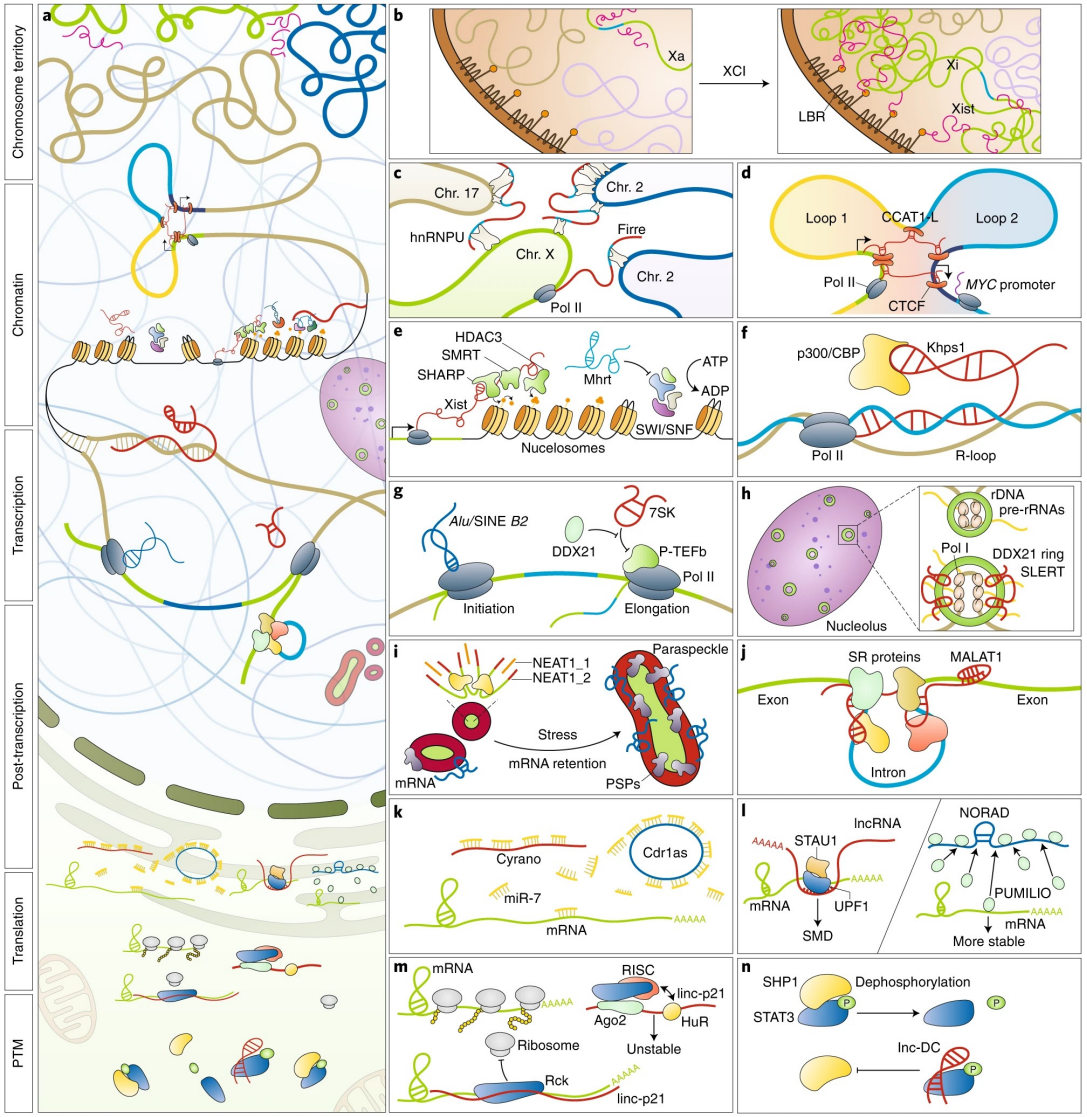
Despite discovering thousands of lncRNAs, only a minor subset have been well characterized



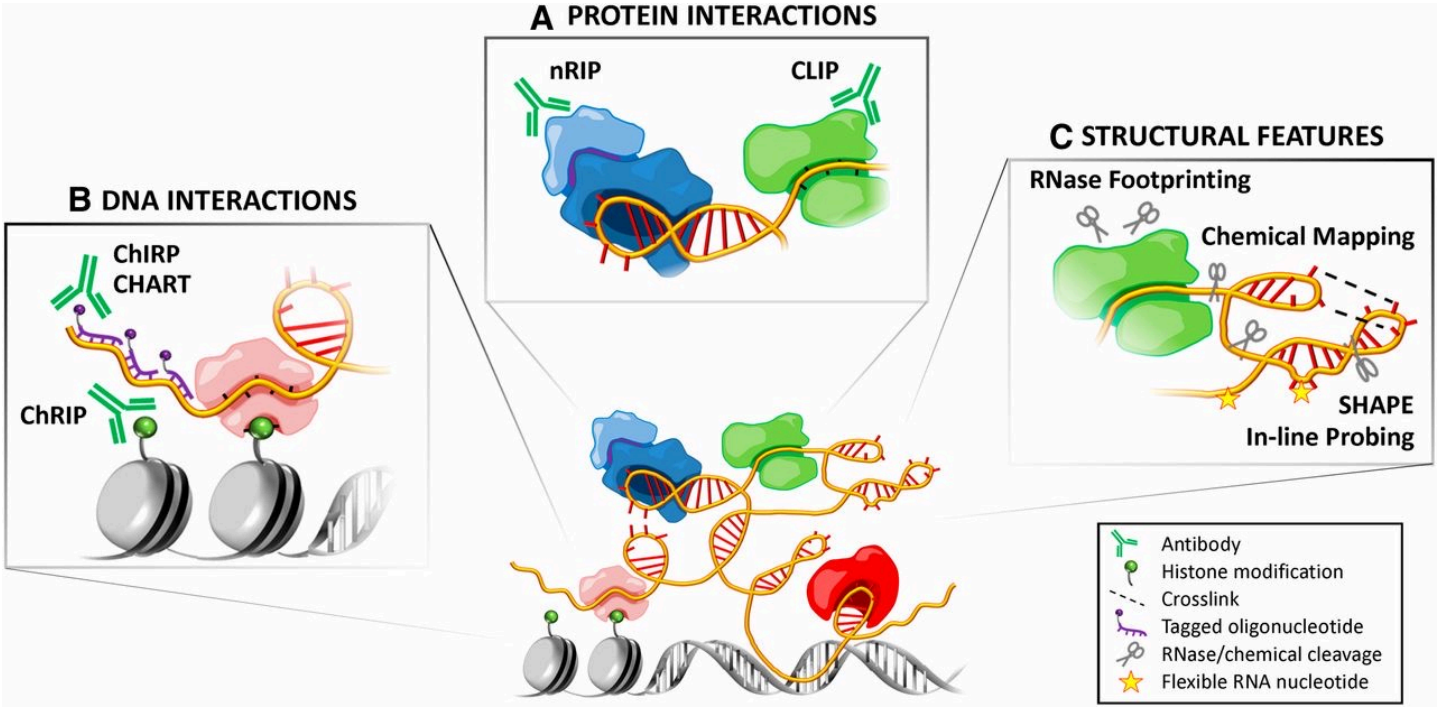
(Nature Reviews Cancer Liu et al., 2021)

Putative lncRNA regulatory mechanisms

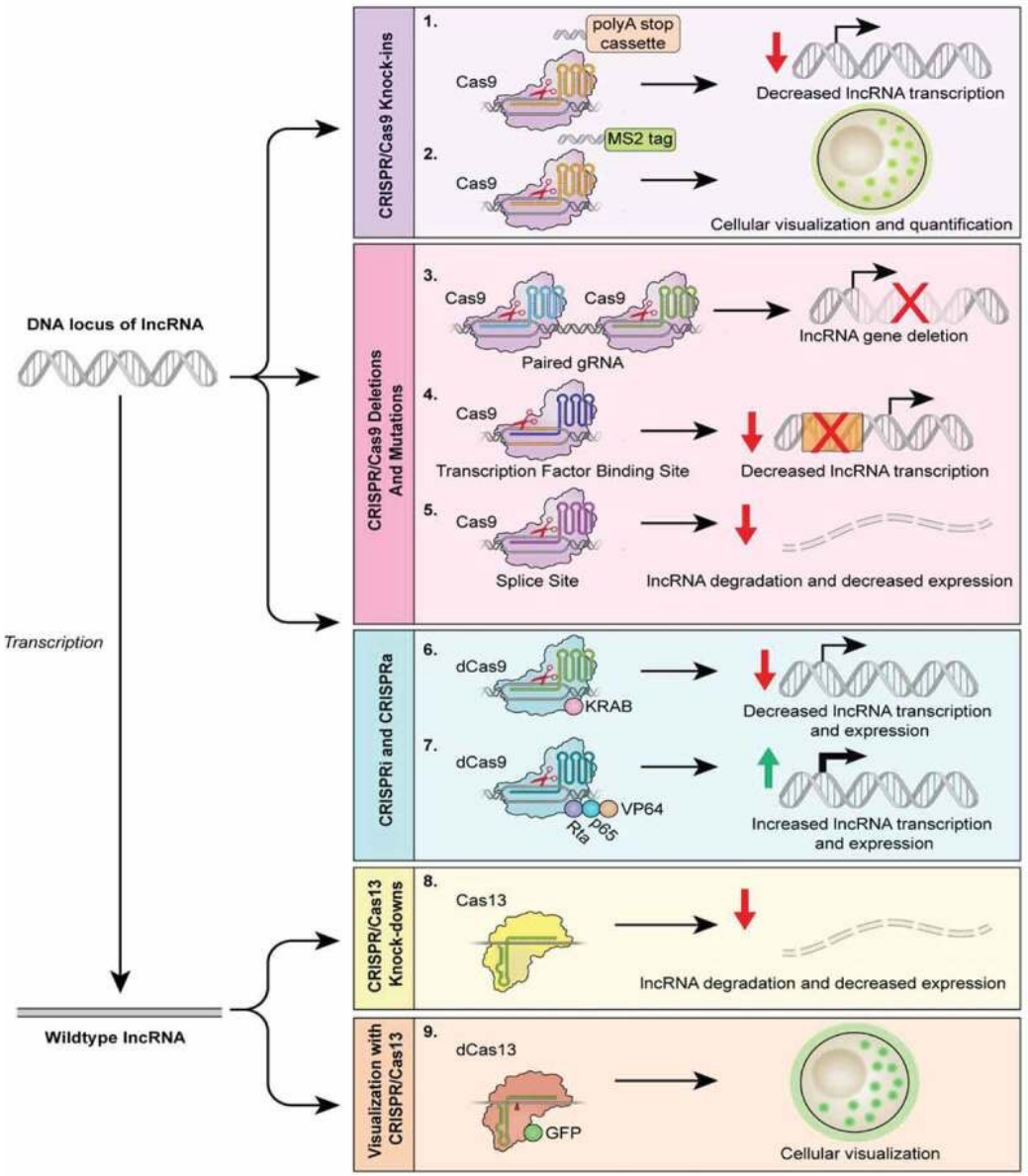




Our ability to understand how a lncRNA functions requires knowledge of its interacting partners: DNA, RNA, and/or protein

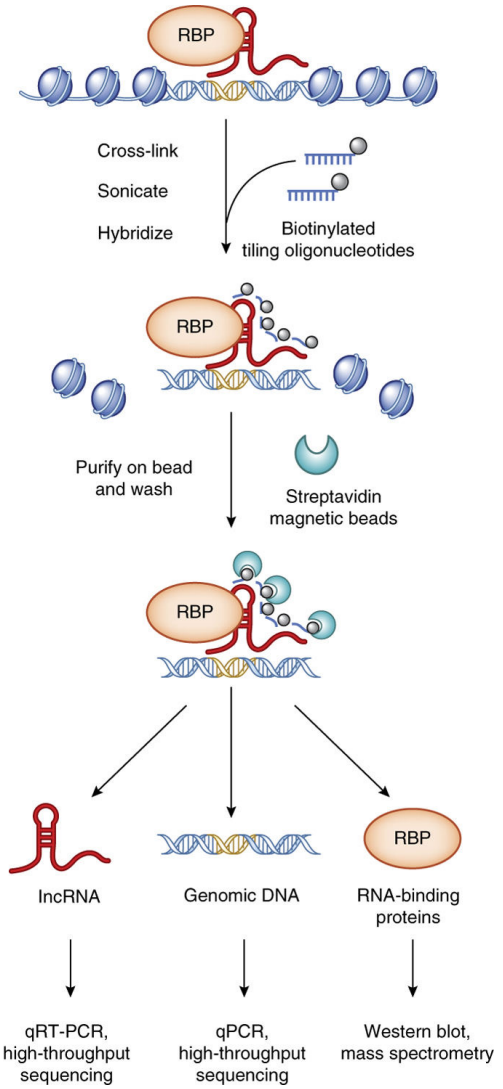


CRISPR based strategies are dependent on the proposed functional mechanism



A LncRNA-Centric Approach

Elucidating lncRNA interactions with proteins, RNA, and DNA

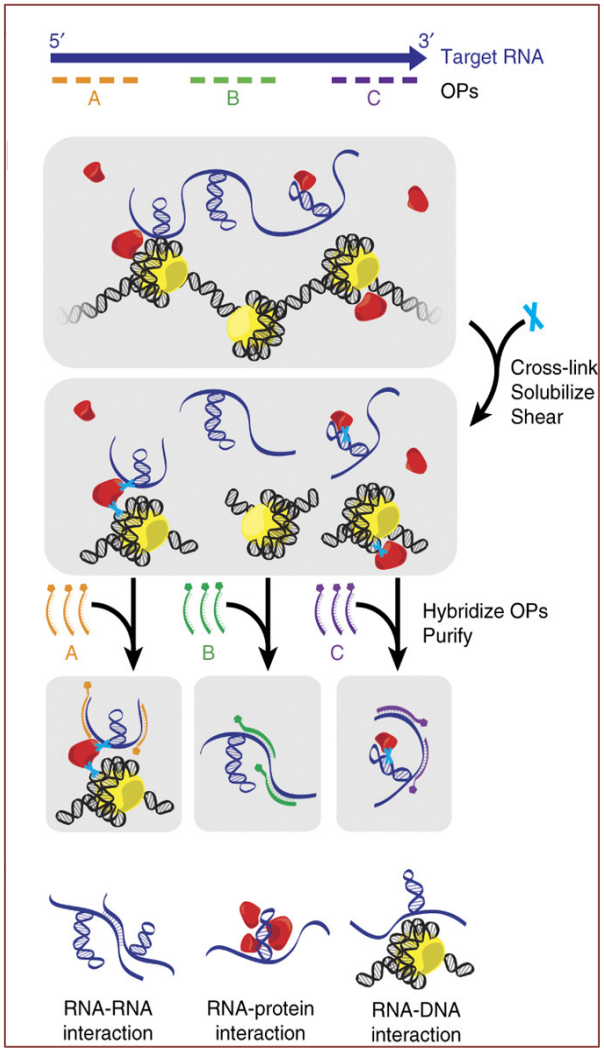


- **ChIRP (chromatin isolation by RNA purification)**
  - RNA–protein–DNA complexes are cross-linked *in vivo* and solubilized by sonication
  - Biotinylated tiling oligonucleotides are hybridized to target lncRNAs
  - Oligonucleotide-bound RNA and associated complexes are efficiently pulled down with streptavidin magnetic beads
  - Enriched RNA, protein and DNA can be isolated and subjected to downstream analysis

(Chu et al., 2014)

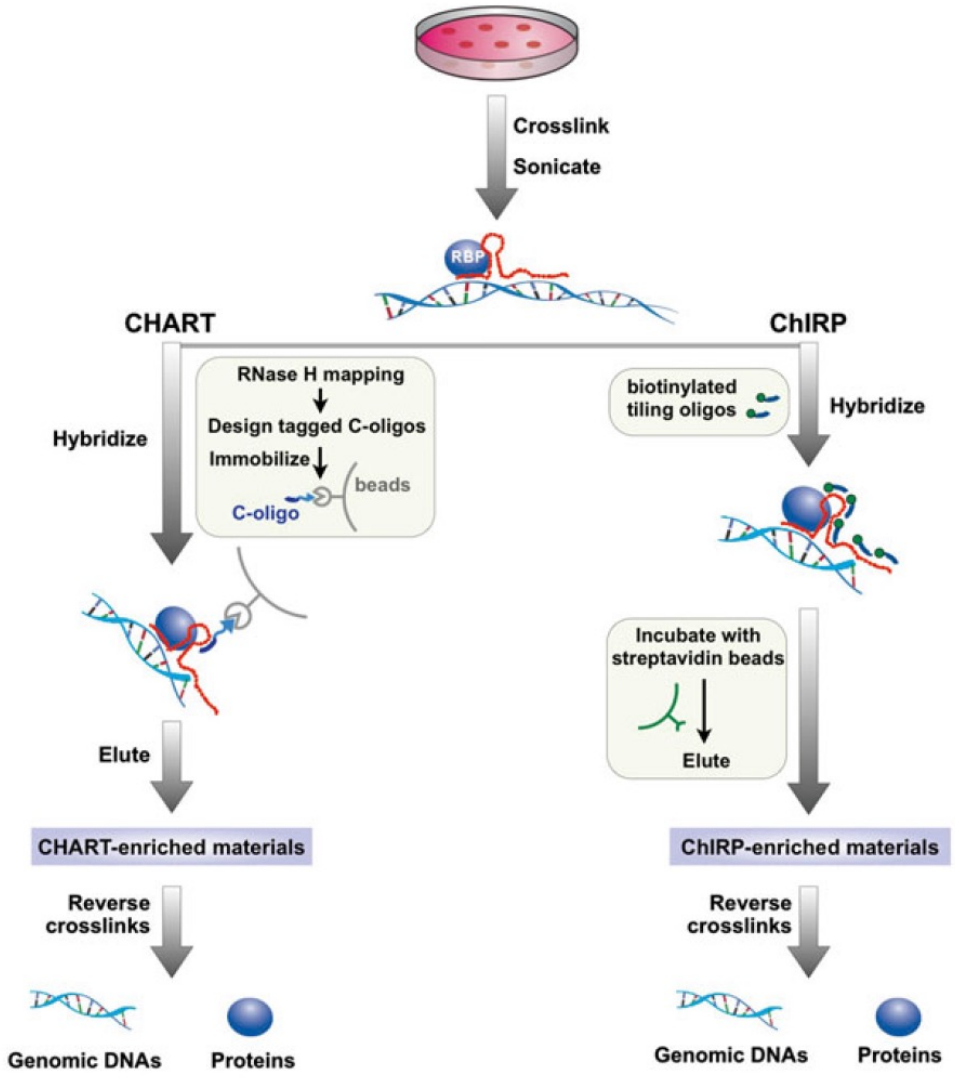
Mapping functional domains within lncRNAs

Domain Specific ChIRP (dChIRP)



(Chu et al., 2011; Quinn et al., 2014)

Additional methods to interrogate lncRNA interactions



Methods are more similar than different

Approach	Probe	Pros	Cons
ChIRP	20-nt; unbiased	Probes are cheap and have minimal off-target effect	Irrelevant probes increase noise
dChIRP	20-nt; unbiased	Improve signal-to-noise by reducing probes	Requires more independent experiments
RAP	120-nt; unbiased	High specificity of longer probes	Probes cost more to synthesize
CHART	Rnase H assay to narrow search space	Background signal reduced due to relatively few probes used	RNase Assay only indicates a probe can bind, not chromatin interaction; time consuming

What have these methods revealed?

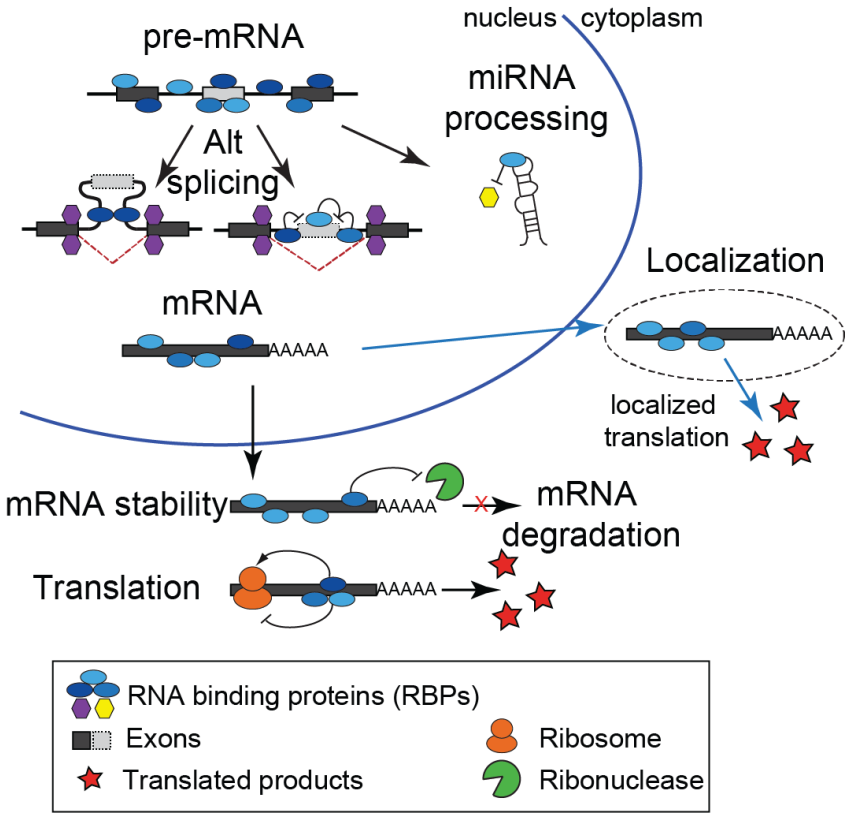
Table 1 Summary of RNAs analyzed by ChIRP, Chart and RAP

RNA	Method of purification	Biological function	Genomic occupancy	References
roX2	ChIRP or Chart	Dosage compensation in male *Drosophila* (one of two lncRNAs required)	With MSL-complex gene bodies, co-occupies active X-linked genes, with strong bias toward 3′ end	4,5
TERC	ChIRP	Acting as scaffold for telomerase complex and template for telomere DNA synthesis	Binds telomeric ends of chromosomes and *Wnt* genes	4
HOTAIR	ChIRP	Recruitment of PRC2 to silence target-gene expression	Exhibits focal binding at loci that overlap with PRC2	4
7SK	ChIRP	Regulation of transcription by controlling the positive transcription elongation factor P-TEFb	JMJD6 and Brd4 co-bind distal enhancers	15
FOXC1 enhancer RNA	ChIRP	Induction of enhancer-promoter looping and enhancement of ligand-dependent induction of target coding gene	Binds its own transcribed enhancer locus (does not appear to act *in trans*)	16
Pan	ChIRP	Interaction with viral and cellular proteins to affect host gene expression (abundantly made during lytic cycle of KSHV)	Binds *ORF50* promoter	17
FMR1 mRNA	ChIRP	Silencing by the 5′ untranslated region of the promoter of its own gene locus	CGG-repeat portion of *FMR1* mRNA binds to the gene promoter through DNA-RNA hybridization to result in silencing	22
116HG	ChIRP	Regulation of diurnal energy expenditure of the brain	Occupies genes enriched for brain expression and protein transport, including *Mtor*, *Crebbp* and *Igf2r*	18
RMST	ChIRP	Regulation of transcription of *SOX2* and modulation of neurogenesis in humans	Together with SOX2, interacts with promoter region of neurogenic genes to co-regulate their expression	19
THRIL	ChIRP	Expression of many immune-response genes	Binds TNF-α promoter	20
roX1	dChIRP	Dosage compensation in male *Drosophila* (one of two lncRNAs required)	Colocalizes with roX2 and MSL on genomic DNA through the three fingers of minimal chromatin-binding domains	14
Xist	Chart	Dosage compensation in mice	Binds gene-rich islands and then the gene-poor domain on to-be-silenced X chromosomes in two steps during *de novo* inactivation	13
Paupar	Chart	Cell-cycle profile of neuroblastoma cells (with loss inducing neural differentiation)	Has ~3,000 binding sites across genome; enriched on X chromosomes, preferentially within promoters and 5′ untranslated regions	21
Xist	RAP	Dosage compensation in mice	Binds spatially close loci and then spreads to the entire chromosome; requires A-repeat region to spread to active genes	6
FIRRE	RAP	Modulation of nuclear architecture across chromosomes	Has 34 global binding sites, some of which are in spatial proximity; five overlap with mRNAs	12

- LncRNAs do not follow any single paradigm in their regulation
- General characteristics
  - Focal or broad binding
  - *Cis* or *trans* regulation
  - Relatively few to thousands of binding sites
  - Activation and Repression
  - 3D conformation dependent

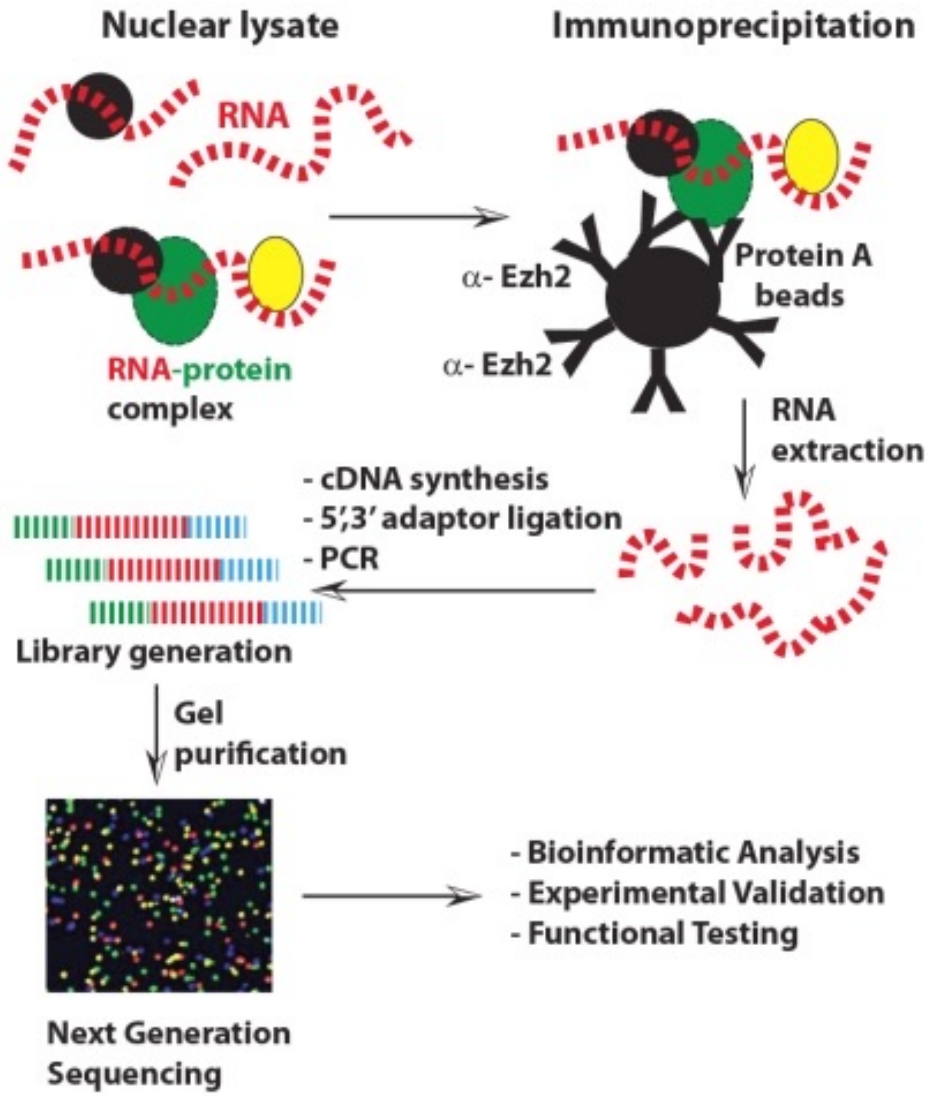
A Protein-Centric Approach

RNA Binding Proteins (RBPs)



- Estimated >1000 RBPs in human
- Have diverse roles in post-transcriptional gene expression, including regulation of alternative splicing, RNA export and localization, RNA stability and translation
- Functionality in gene regulation is naturally dependent on their ability to selectively recognize and bind target RNAs within the cell
- Mutation or alteration of RNA binding proteins plays critical roles in disease

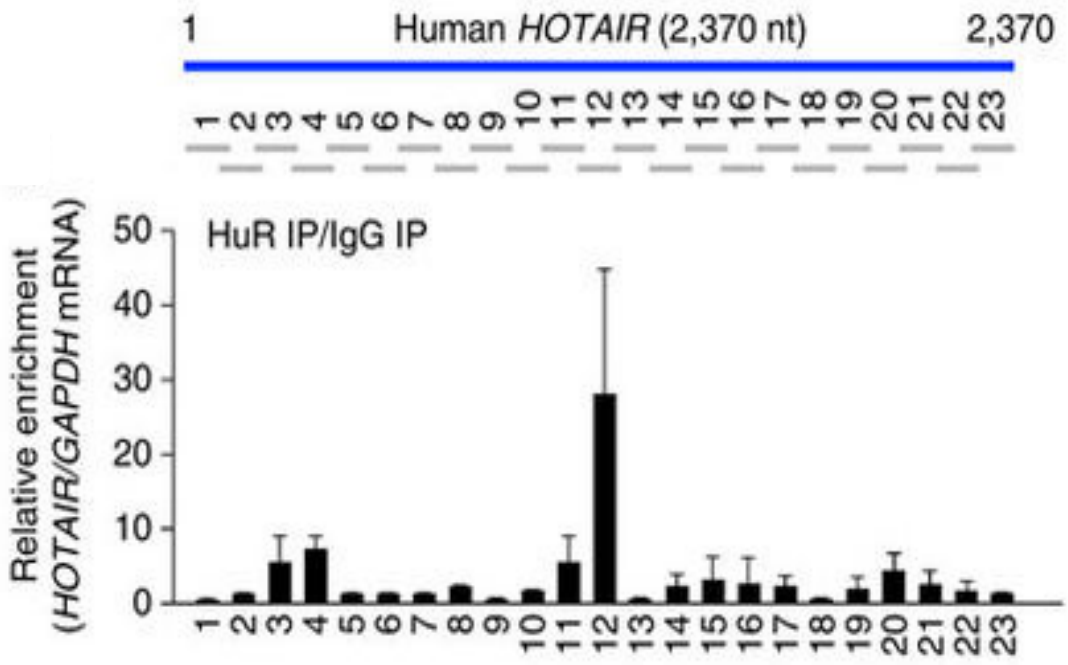
RNA Immunoprecipitation coupled with NGS (RIP-Seq)



- RIP allows identification of the target RNA molecules binding to an RBP
- Limitations
  - Data may include indirectly bound sequences
  - High variability
  - Requires high quality antibody
  - Precise locations of the binding site on the target mRNA may be difficult to determine
- RIP conditions must be calibrated to minimize reassociation of RBPs with mRNA in vitro after cell lysis

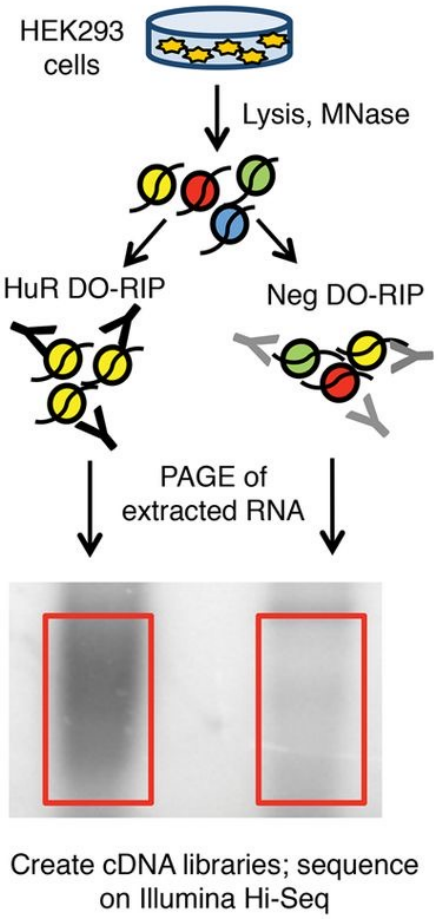
Locating specific interaction sites

- RNAse protection assay can help localize the potential interaction site



(Yoon et al., 2013)

DO-RIP-Seq Overview

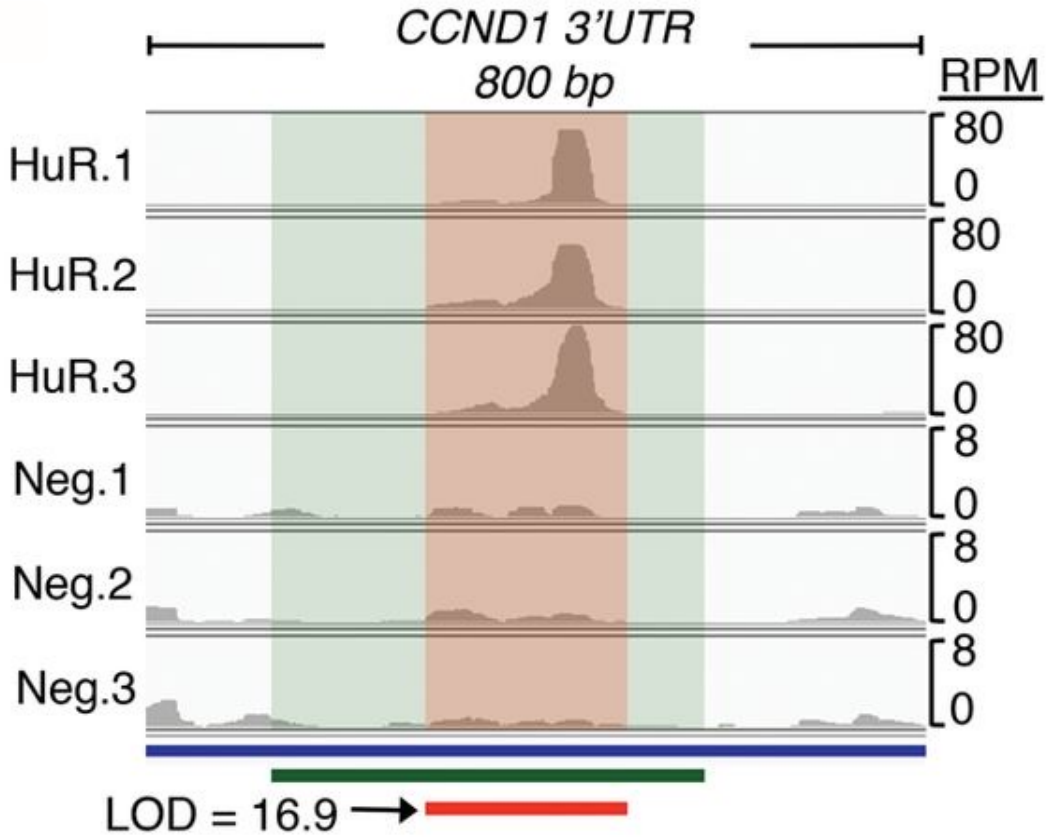


(RNA -- Nicholson et al. 2017)

- Cell lysates treated with micrococcal nuclease (MNase) under optimized conditions to partially digest RNA to fragments bound by the RBP.
- RNAs from parallel immunoprecipitations using a nonspecific control antibody or similar negative sample were extracted for normalization of the positive sample

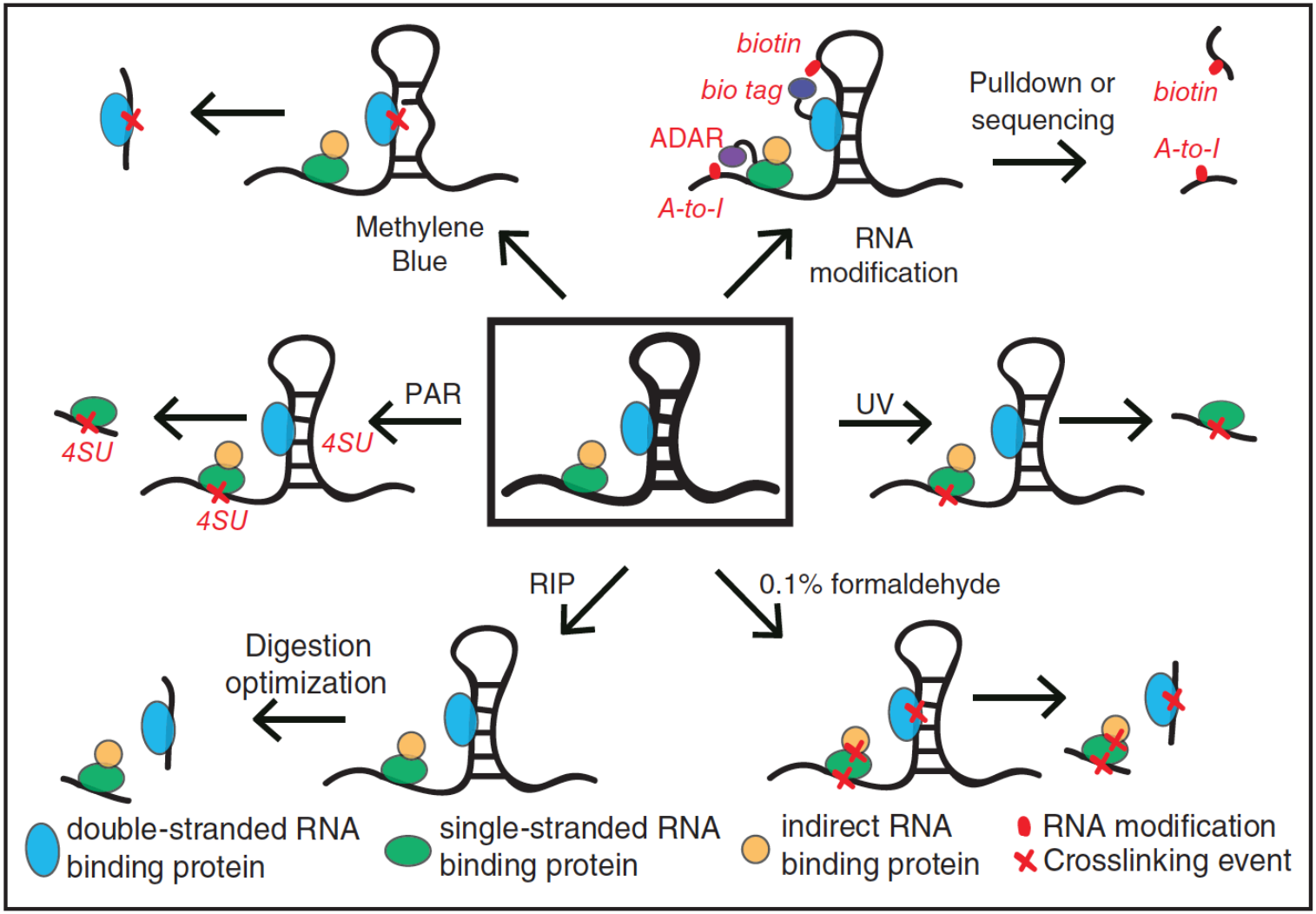
DO-RIP-Seq detects validated interaction sites

HuR DO-RIP-seq binding site (red bar and shading) in the CCND1 mRNA 3′UTR in comparison to the binding site deduced by a previous study (green bar and shading) using deletion analysis (Lal et al. 2004).



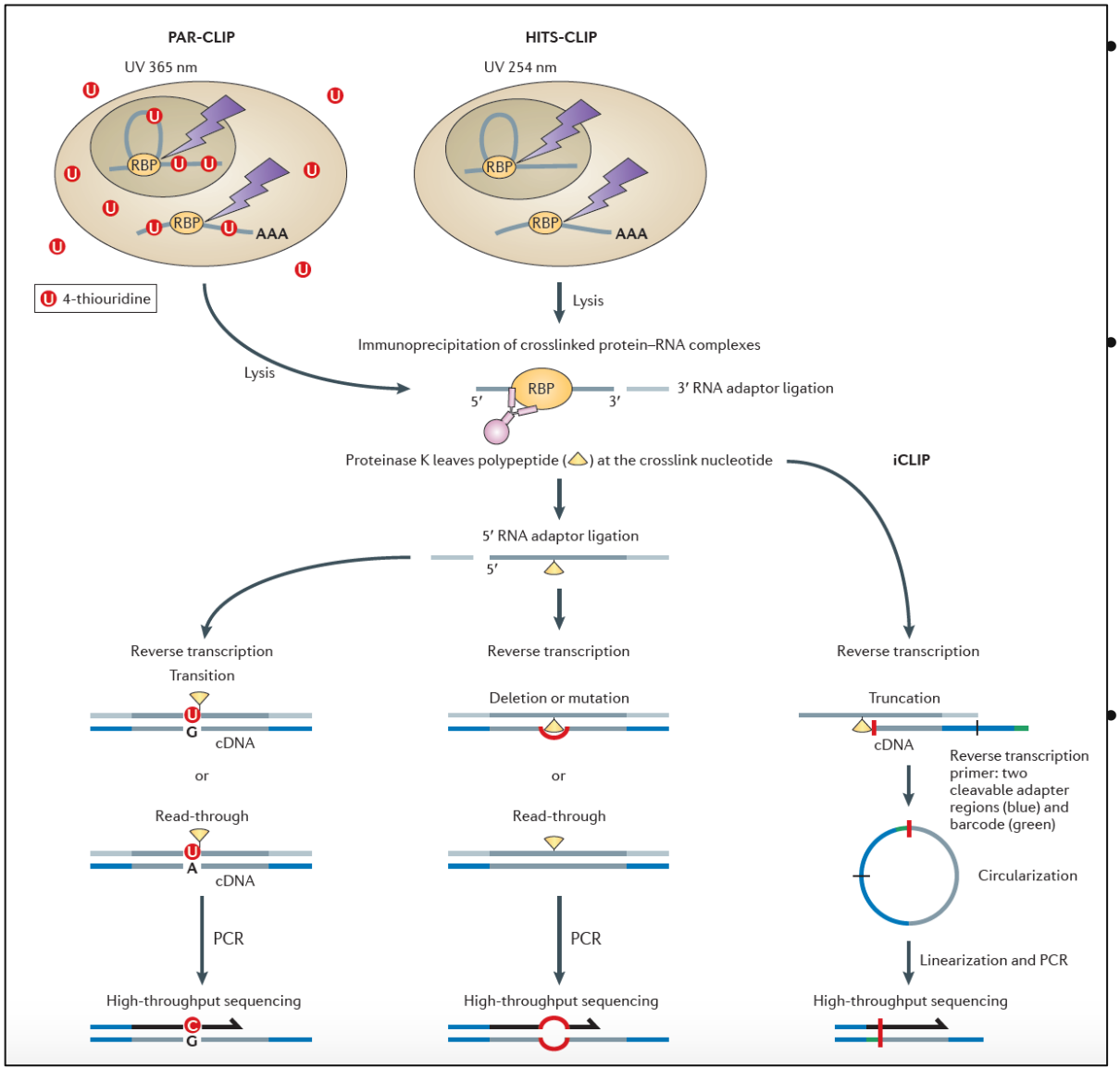
Cindo O. Nicholson et al. RNA 2017;23:32-46

Methods to capture protein-RNA interactions



(Wheeler et al., 2017)

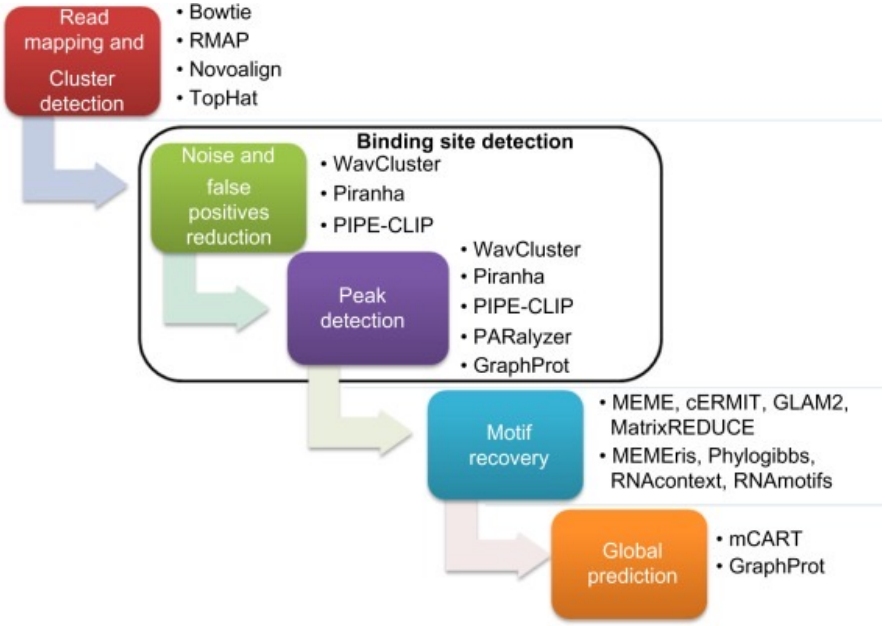
Common variations of crosslinking immunoprecipitation (CLIP)

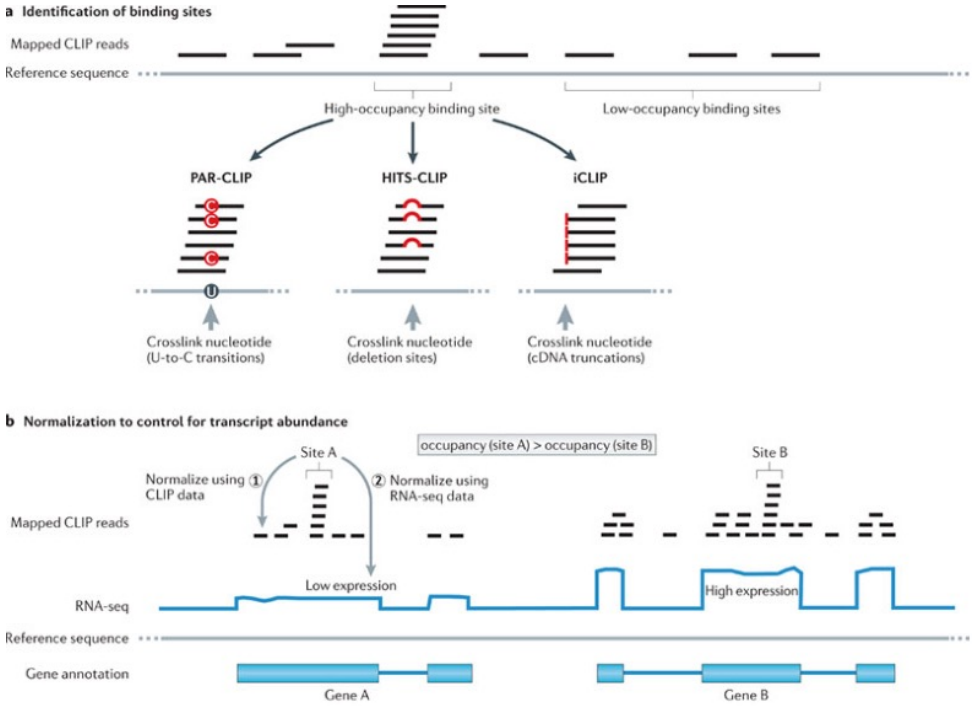


- **HITS-CLIP** 254 nm ultraviolet UV cross-linking and immunoprecipitation allows more stringent washing and RNase treatment of bound RNAs
- **PAR-CLIP** is another modification of CLIP-seq that first treats the cell with a modified nucleoside (4SU or 6SG), which is incorporated into transcribed RNA. The modified nucleotide can be cross-linked using longer wavelength UV radiation
- **iCLIP** identifies binding sites more precisely by taking advantage of the fact that the amino acid tag left by proteinase K treatment terminates reverse transcription The truncated cDNA molecules can be marked with cleavable adaptor and barcode allowing for self circularization

(Nature Reviews Genetics – Konig et al., 2012)

CLIP-Seq data analysis workflow



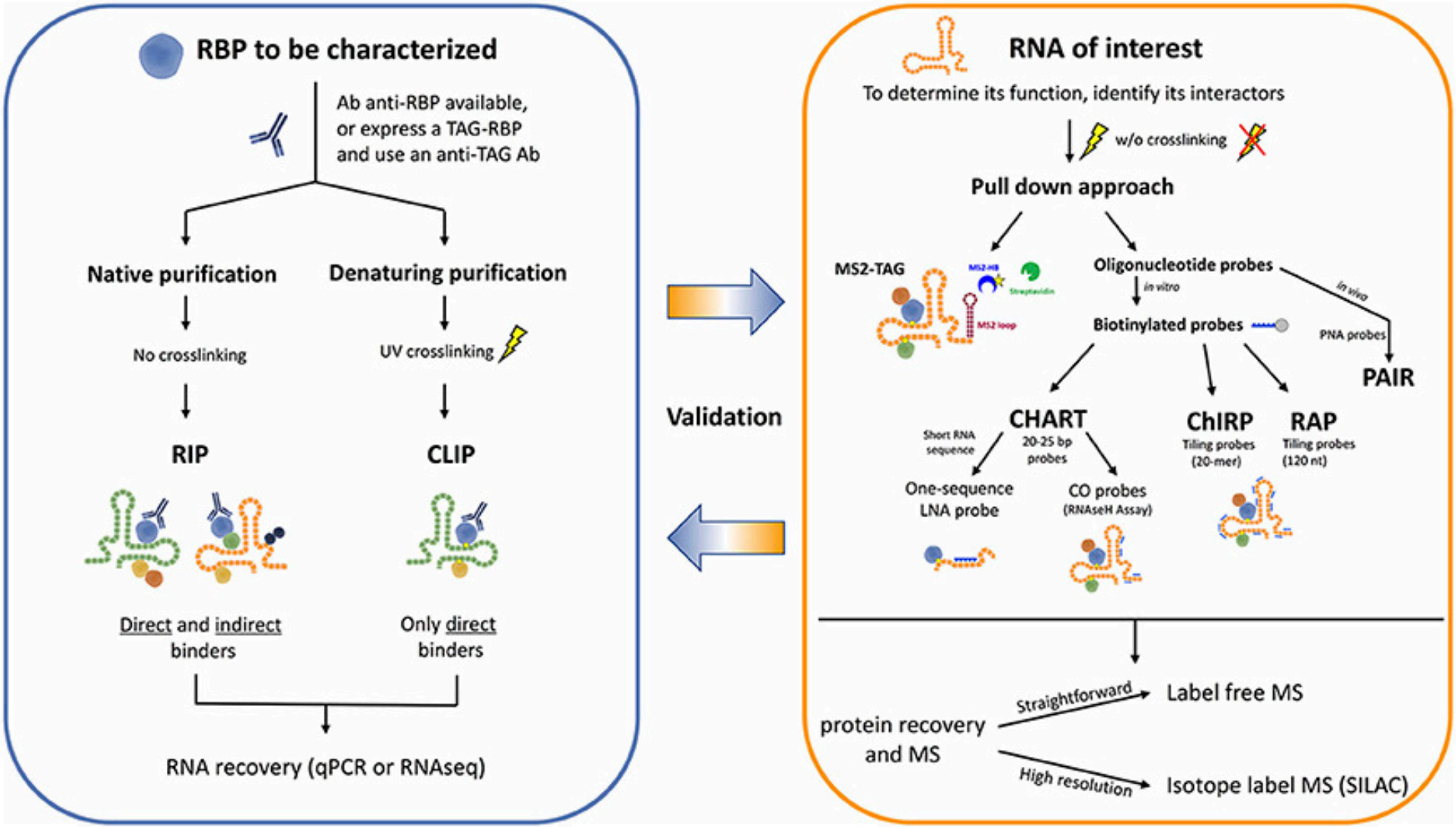


How do you choose?

Methods	PROS	CONS
RIP	• Performed under physiological conditions to preserve the native complexes • Requires little specialized equipment and/or reagents	• Relies on the availability of good antibodies, or the use of tagged RBPs • Lacks high-stringency washes and crosslinking of RBPs to RNAs, which leads to low signal to noise ratio and frequent misinterpretations in the data analysis • Additional control conditions may be required to distinguish true interactions from non-specific ones • Does not determine the exact location of the binding site of RBPs
CLIP	• Application of strong washing steps allows to get rid of non-specific binders	• UV radiation can alter the RNP infrastructure, and crosslinking is not homogeneously efficient • Low efficiency of UVC (254 nm) RNA-protein crosslinking • Difficult identification of the exact site of crosslink within the sequenced fragment
HITS-CLIP	• Genome-wide tool	• The eluted RNA must be de-crosslinked, cDNAs are truncated at the crosslink site and get lost during the standard library preparation protocol
PAR-CLIP	• Single nucleotide resolution to identify the exact site of binding of the RBP on the RNA (the nucleotide analogs are converted into cytosine (C) for 4-SU, or adenine (A) for 6-SG, and can be used to specifically mark the exact binding site)	• The eluted RNA must be de-crosslinked, cDNAs are truncated at the crosslink site and are lost during the standard library preparation protocol • Nucleotide analogs can be toxic for cells and animal models • More expensive than the classic CLIP approach
iCLIP	• Single nucleotide resolution to identify the exact site of binding of the RBP on the RNA	• Needs special adaptors to allow the circularization step, not always highly efficient • Input material required: high

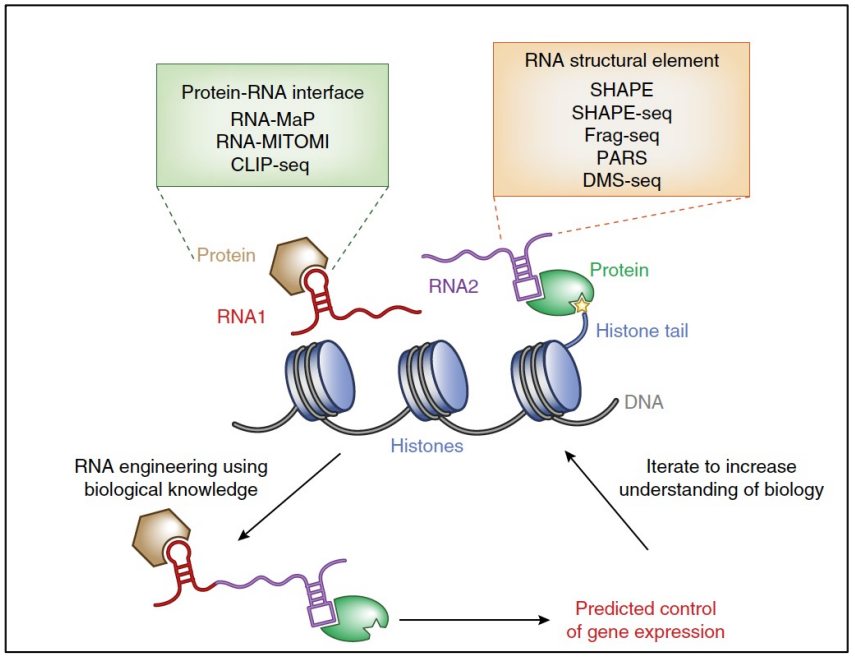
(Barra et al., 2017)

Integration of various strategies



LncRNA summary

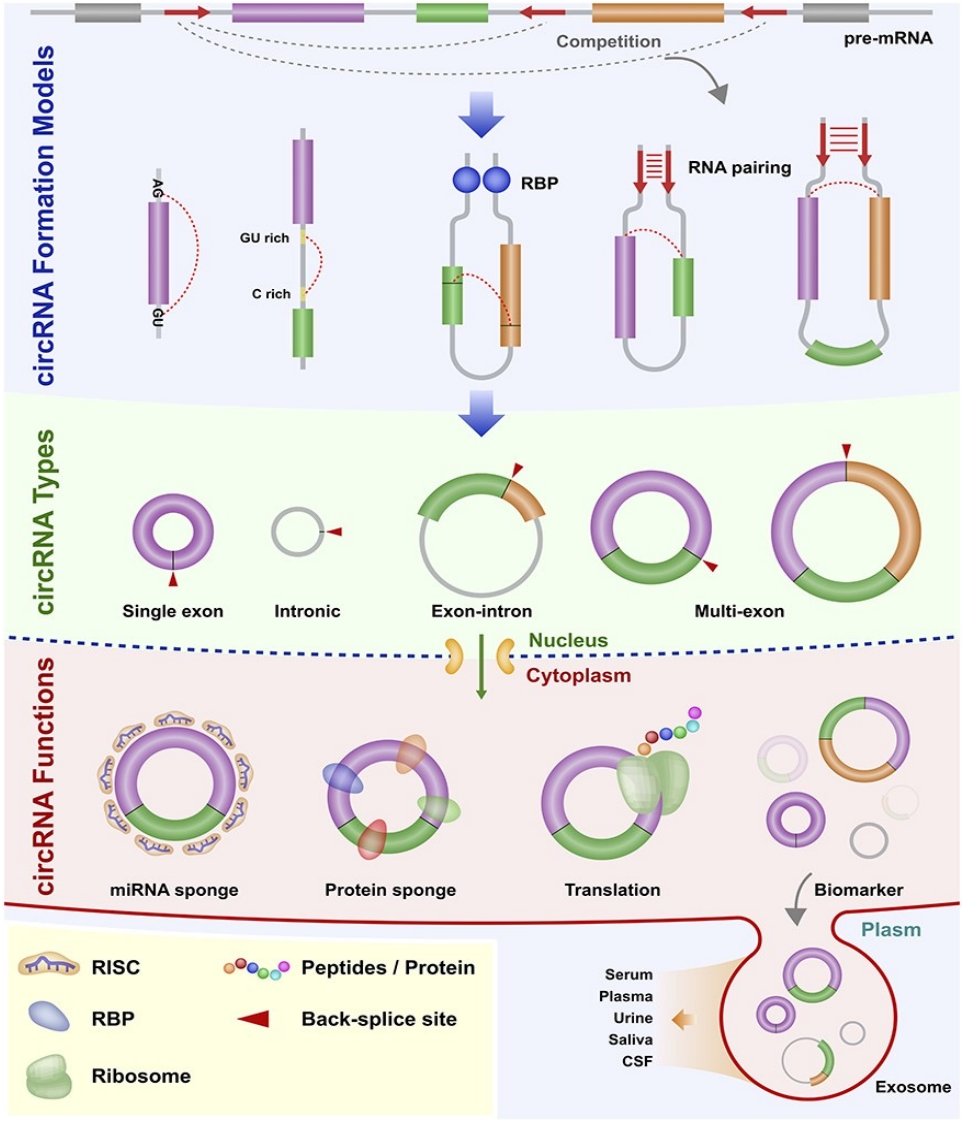
- LncRNAs are an abundant class of biologically and clinically relevant class of genes with a broad range of functionality
- Despite the rapid emergence of lncRNAs, the methods to interrogating their regulatory mechanisms are still evolving
- Ongoing development is still necessary to fully understand the limitations and biases of existing NGS applications and the corresponding computational tools for analysis and interpretation
- Integration of orthogonal strategies will increase the likely of uncovering real lncRNA regulatory mechanisms



(Chu et al., 2015)

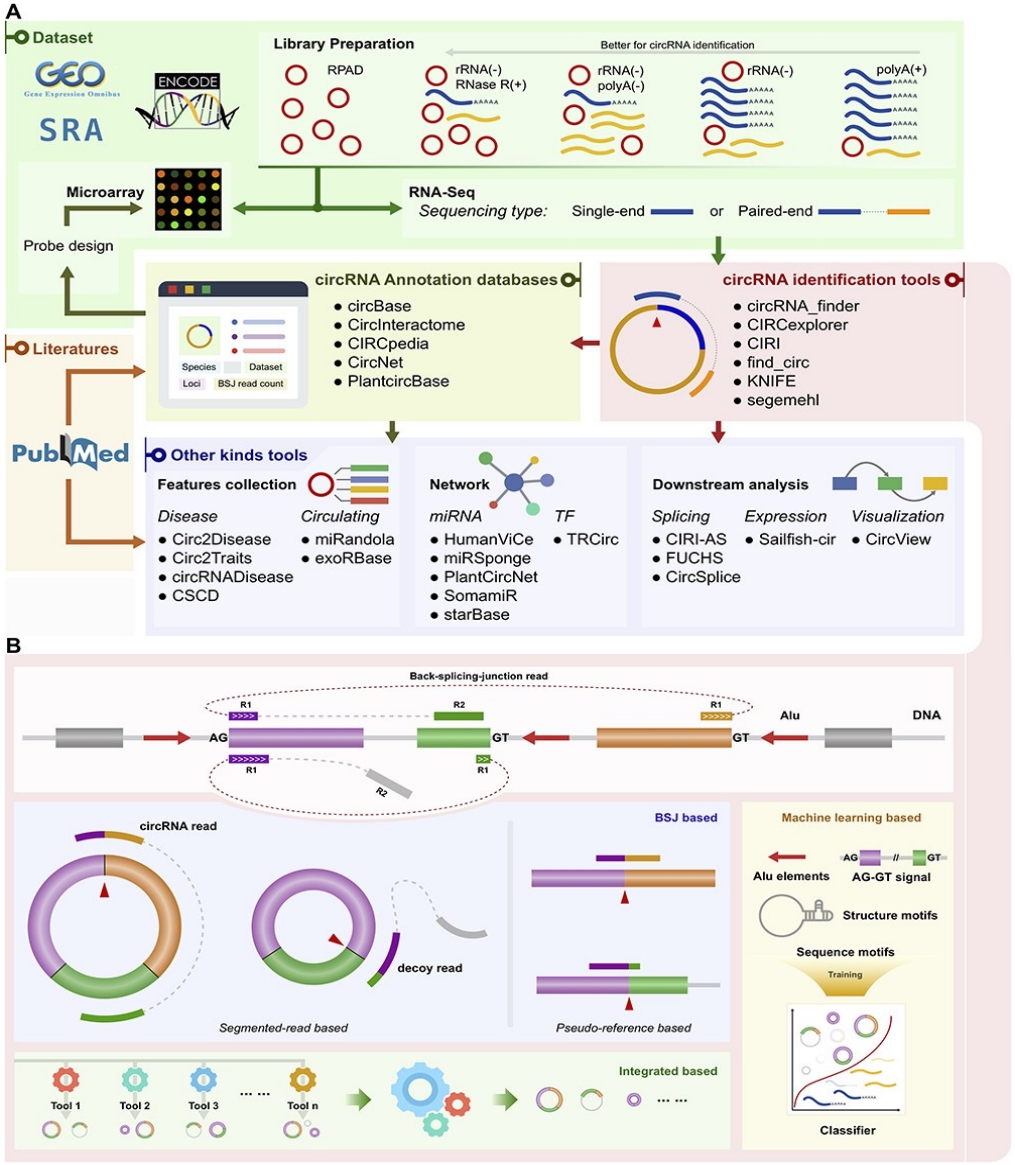
Circular RNAs

Circular RNAs (circRNAs)



(Chen et al., 2021)

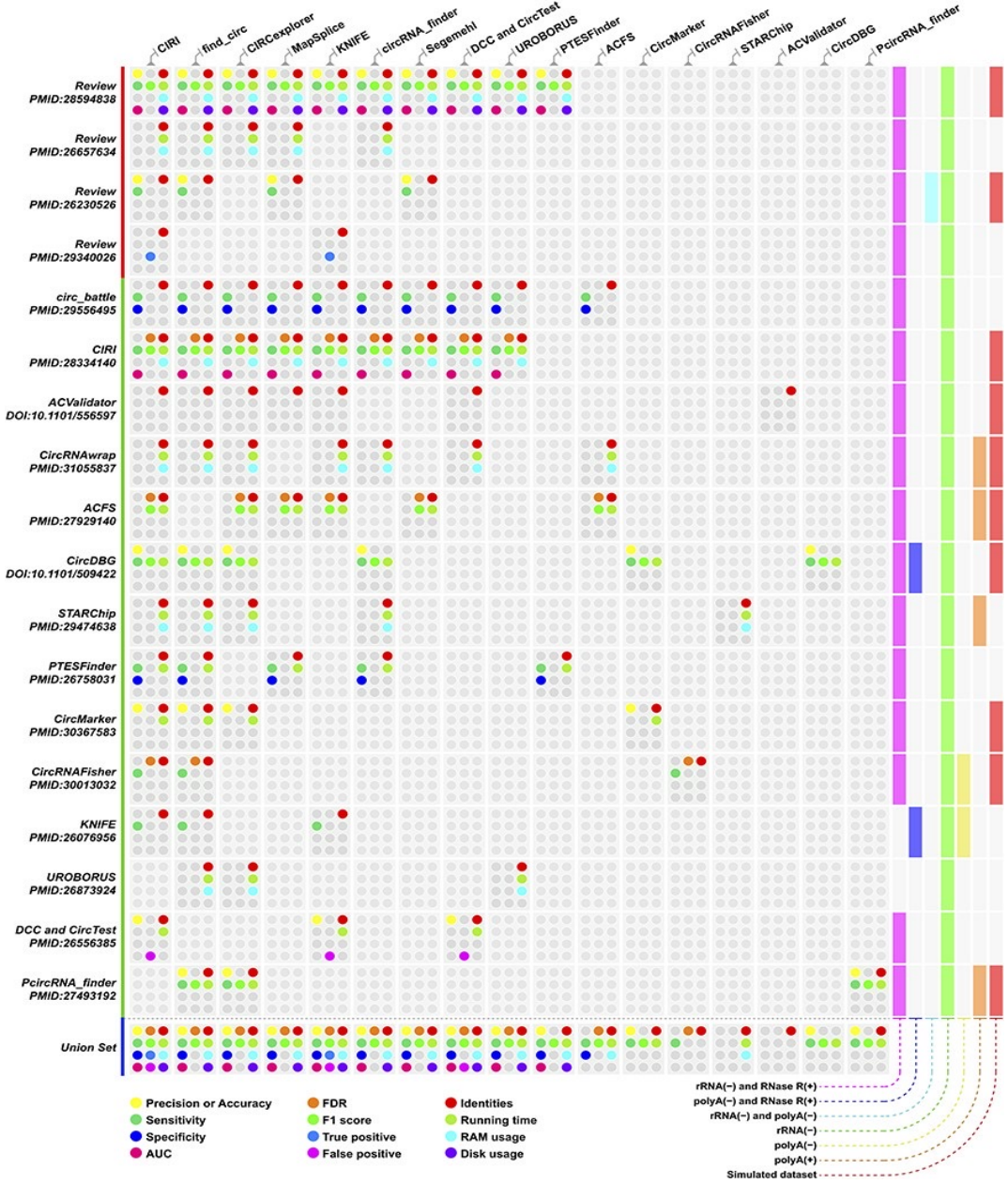
Overview of existing circRNA resources and tools



(Chen et al., 2021)

Evaluation of tools

- Extensive quantity of tools available
- Most existing tools are designed for short read sequencing

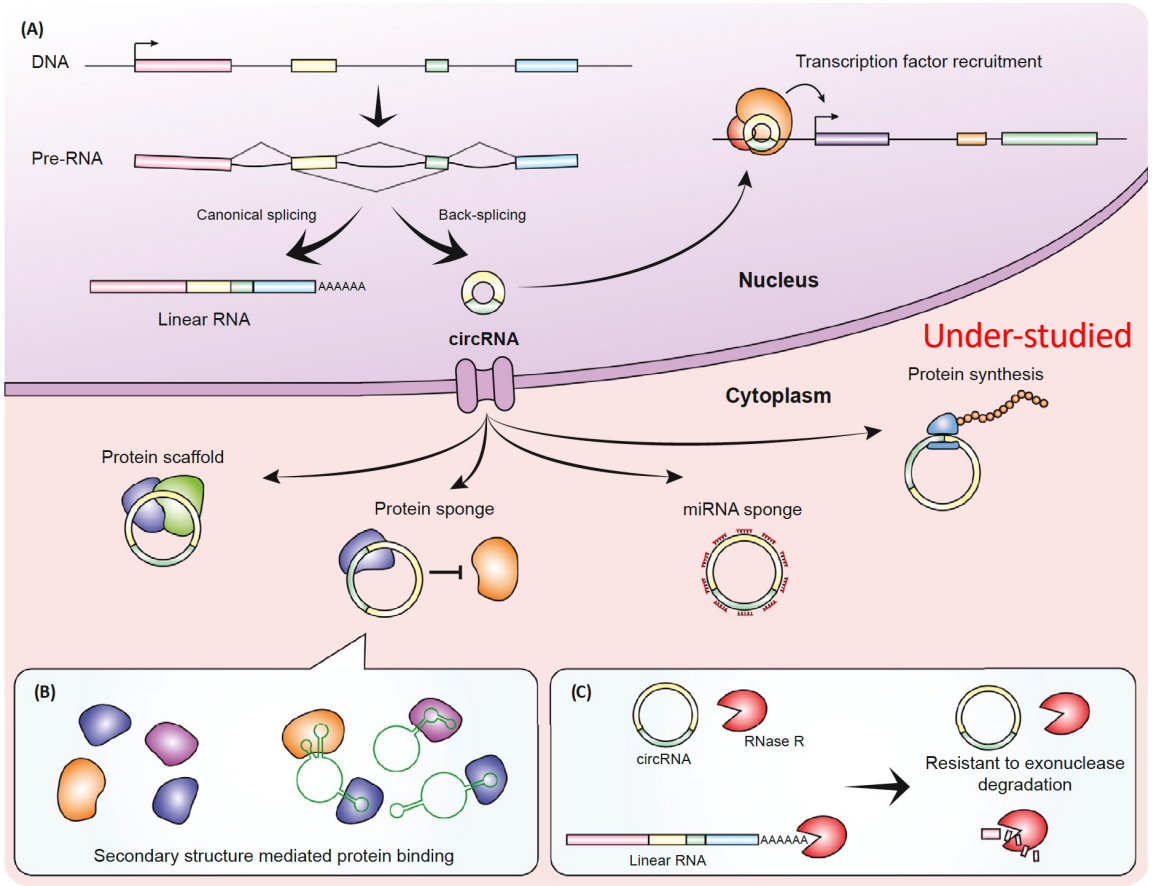


(Chen et al., 2021)

Limited understanding of circRNAs contributing to metastatic colon cancer progression (mCRC)

- Large scale studies (i.e., TCGA) mostly used poly-A selection
- No standard RNA quantification method
- Cell lines or limited patient cohorts
- Lack of genome-wide systematic analysis
- Existing databases lack inclusion of CRC (and particular matched patients throughout progression)
  - MiOncoCirc, a cancer focused database, contains only 14 CRC out of 880 patients

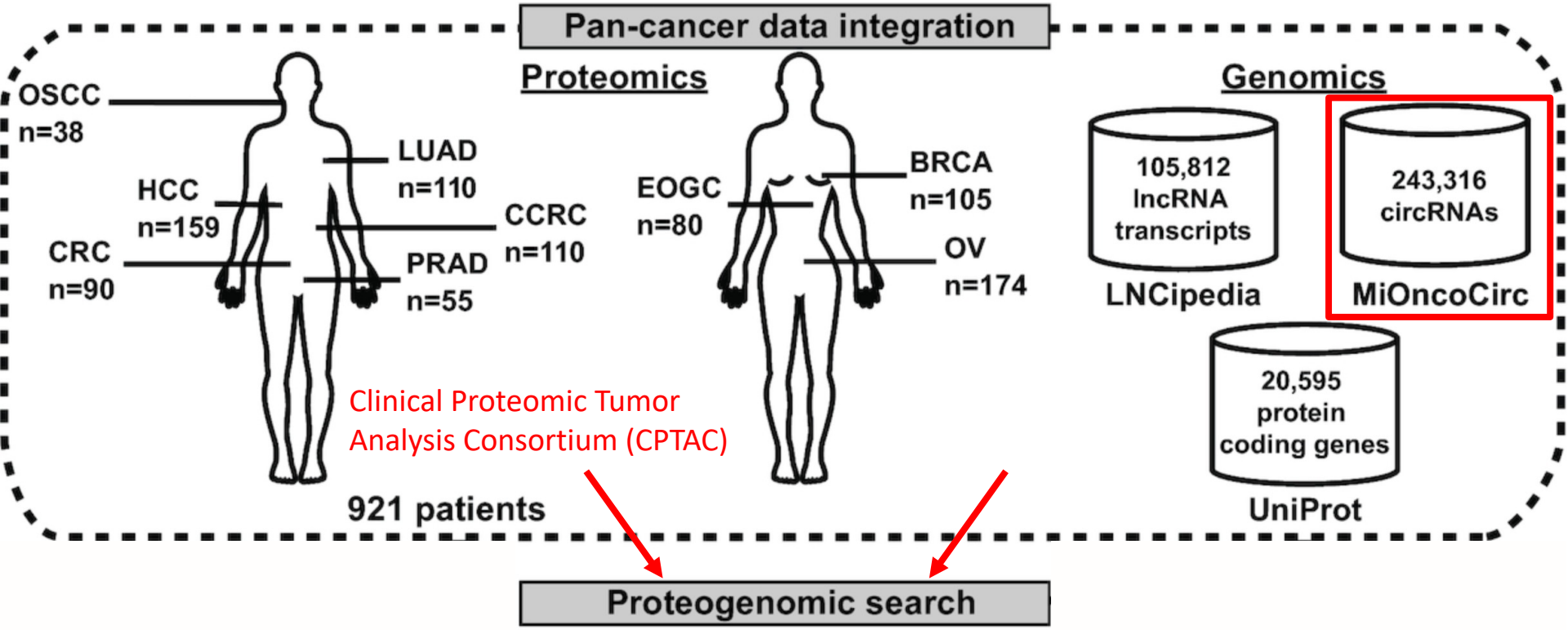
Putative functions of circRNAs remain under-studied in cancer



Hua, J.T., S. Chen, and H.H. He, *Landscape of Noncoding RNA in Prostate Cancer.* Trends Genet, 2019.
Othoum, G., et al., *Pan-cancer proteogenomic analysis reveals long and circular noncoding RNAs encoding peptides.* NAR Cancer, 2020.

- Limitations of circRNA translation studies
  - Ribo-Seq only shows initiation of translation ≠ peptide products
  - Proteomics study typically discard noncoding RNAs

Pan-cancer proteogenomic integration of circRNAs: PepTransDB

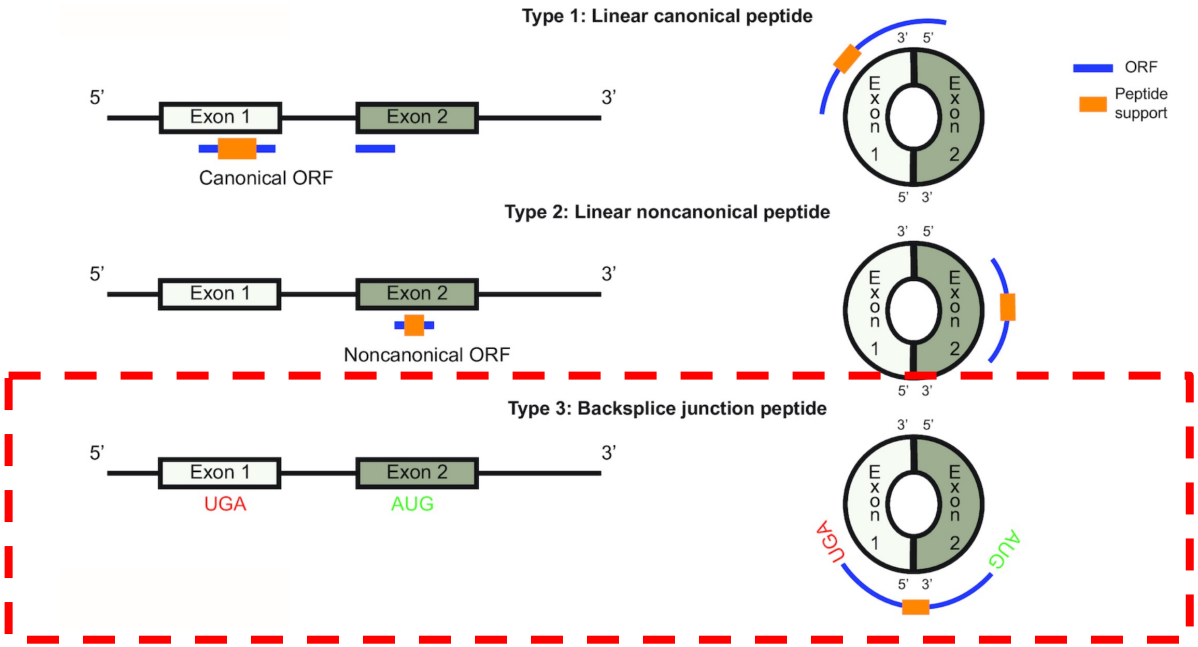


https://www.maherlab.com/peptransdb

(Othoum et al., 2020)

Possible types of peptides encoded by circRNAs

PepTransDB:
Total: 3238 peptides
from 2834 circRNAs



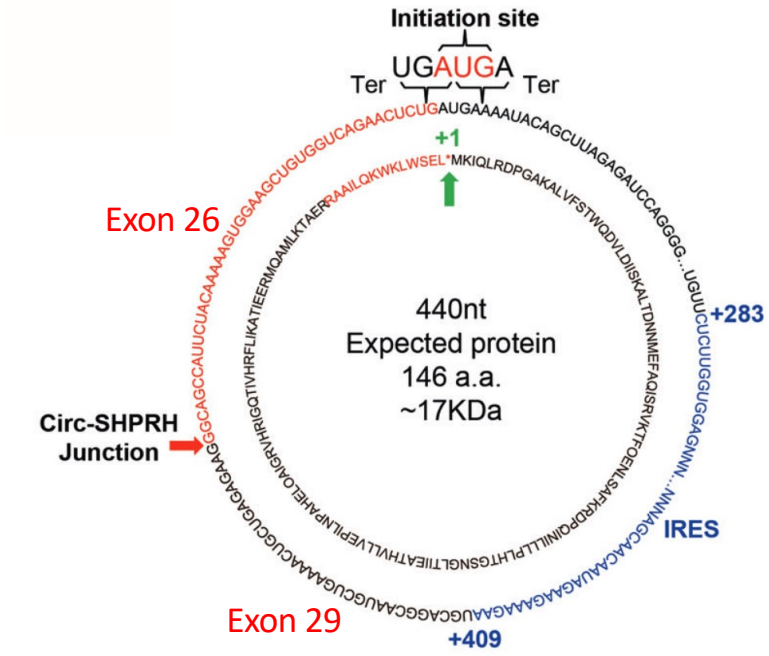
Junction: 2010 peptides
from 1964 circRNAs

(Othoum et al., 2021)

Short read strategies are limited to terminal exons

A novel protein encoded by the circular form of the *SHPRH* gene suppresses glioma tumorigenesis

Maolei Zhang[1,2] · Nunu Huang[1,2] · Xuesong Yang[1,2] · Jingyan Luo[3] · Sheng Yan[1,2] · Feizhe Xiao[4] · Wenping Chen[1,2] · Xinya Gao[1,2] · Kun Zhao[1,2] · Huangkai Zhou[1,2] · Ziqiang Li[5] · Liu Ming[5] · Bo Xie[6] · Nu Zhang[1,2]

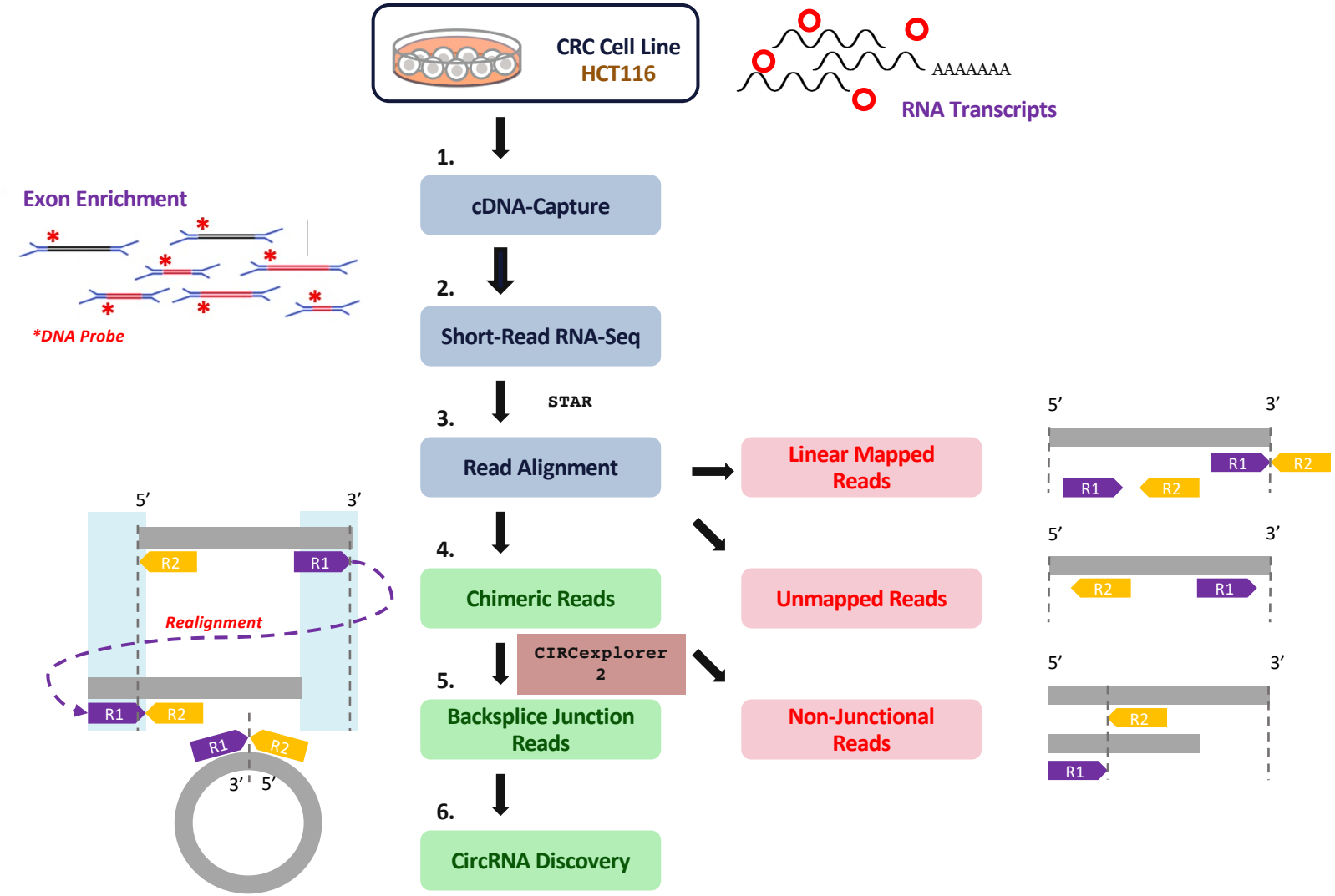


E.g. CircSHPRH was missed in PepTransDB

- 4 exons involved (exon 26, 27, 28, 29)
- Translated open reading frame (ORF) spans beyond backsplice junction (exon 29-26)

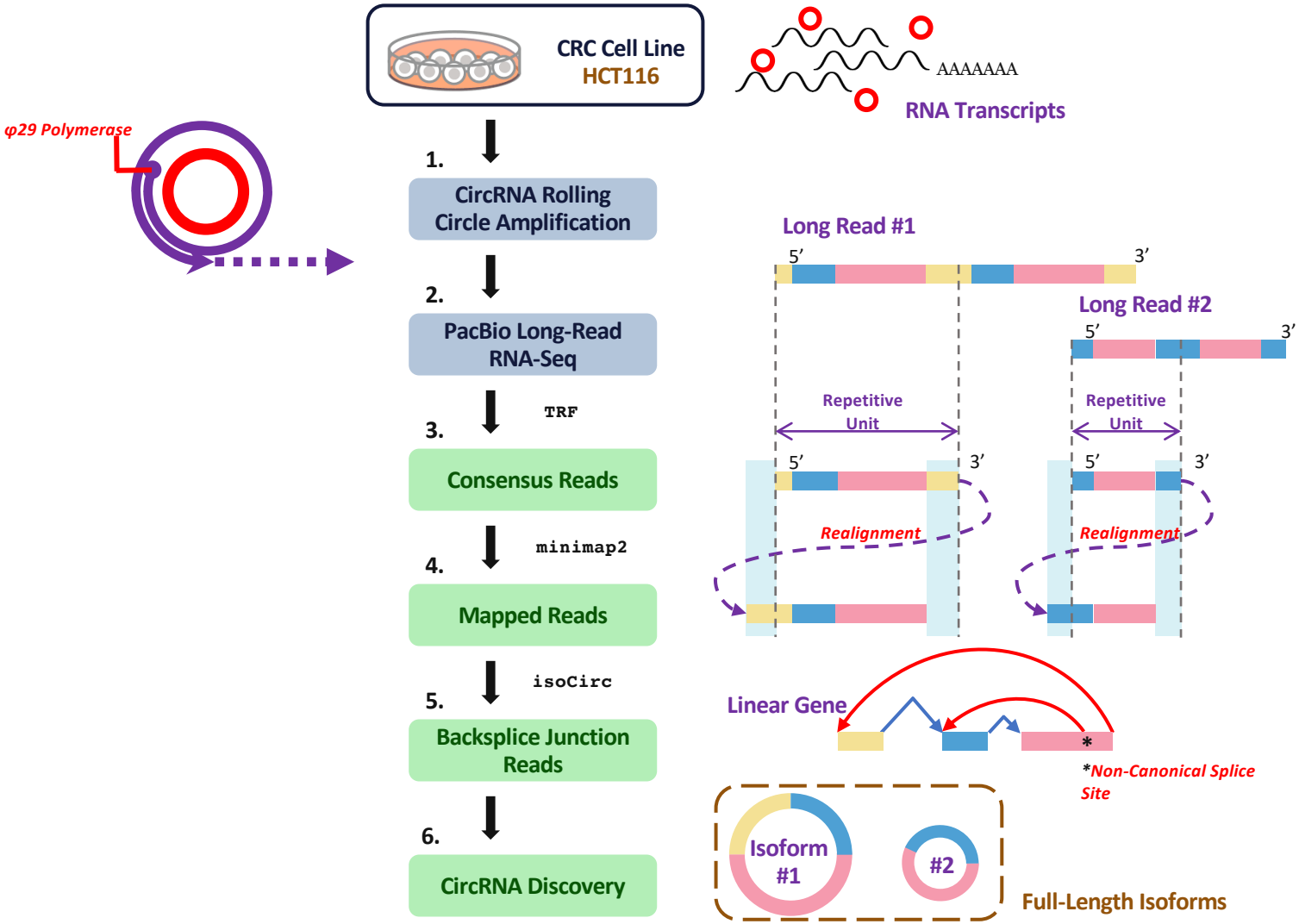
➢ To comprehensively capture ORFs of full circular transcripts: PacBio long-read sequencing

CircRNA detection using short-read analysis pipeline



(Cabanski et al., 2014)

CircRNA discovery with long reads



Unpublished

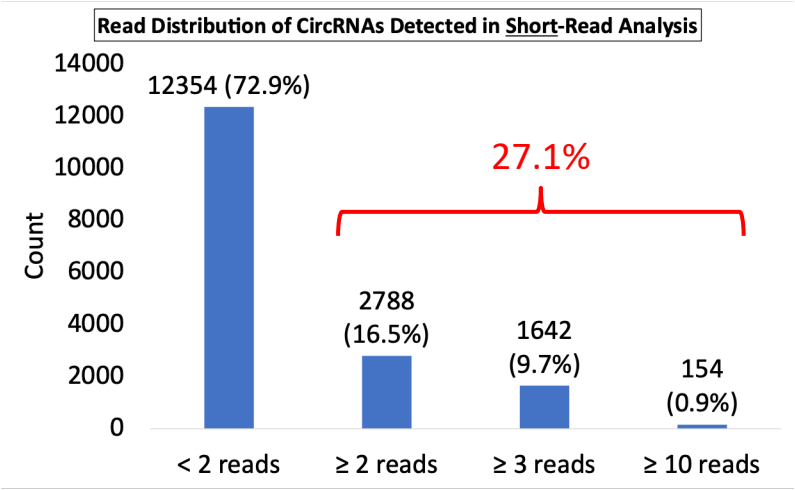
Long-read sequencing produces 35x circRNA enrichment and a larger proportion of high expression circRNAs

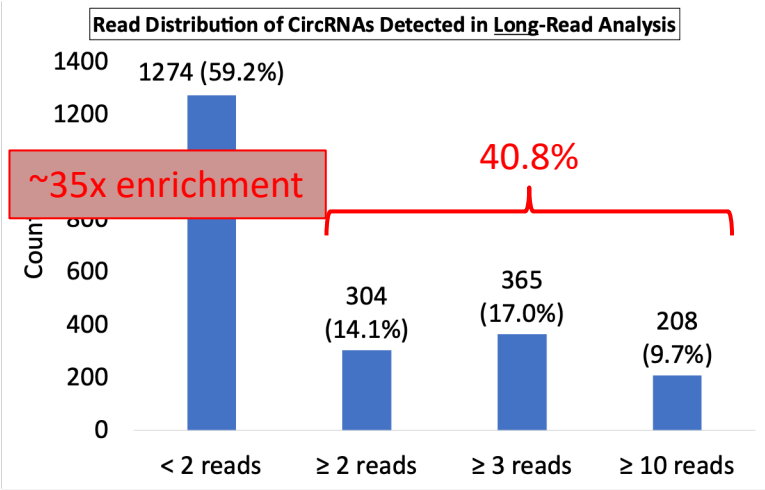
Short-read sequencing summary

Type of reads	Illumina	
	No. reads	% of total reads
Total reads	30,980,769	100.00%
Reads with candidate circRNAs	16,685	0.054%

Long-read sequencing summary

Type of reads	PacBio	
	No. reads	% of total reads
Total reads	1,637,091	100.00%
Reads with candidate circRNAs	34,616	2.11%





Unpublished

Identified novel peptide encoded circRNAs via PepTransDB

- 9 different peptides, including 2 junctional peptides
- 10 different cancer types

Junctional peptide example:

chr	start	end	no. exons	strand	gene
15	80120327	80122800	3	+	ZFAND6
15	80120327	80122800	2	+	ZFAND6

Peptide

3’…GGT…5’

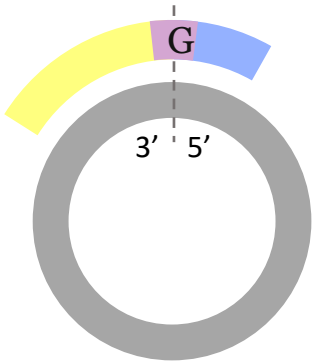
AVPETEDVQGVQLR

Amino Acids from 5' Exons

VQLRNMAQETNHSQVPMLCSTGCGFYGNPRTNGMCSVCYKEHLQRQNSSNGRISPP

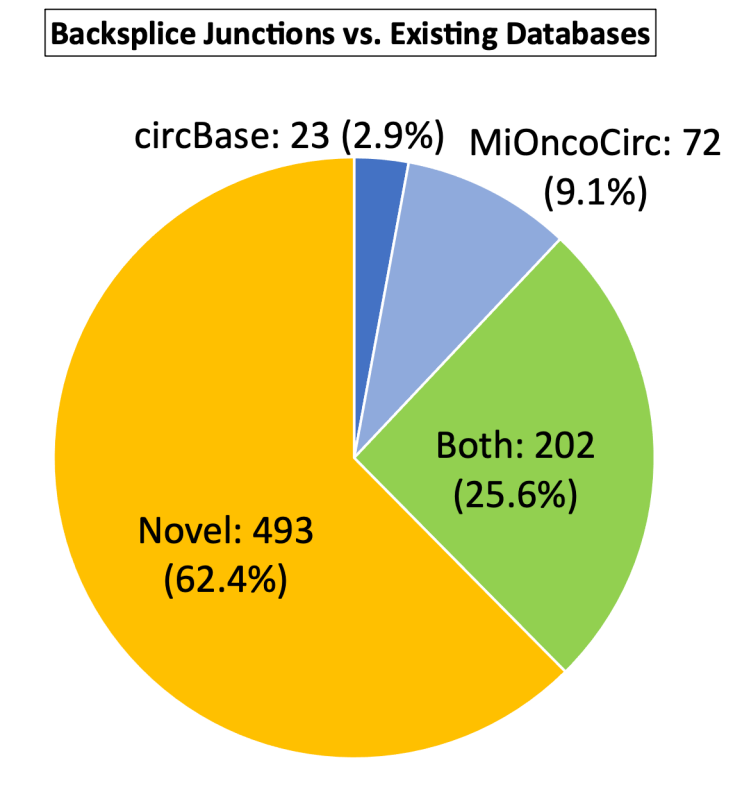
Amino Acids from 3' Exons

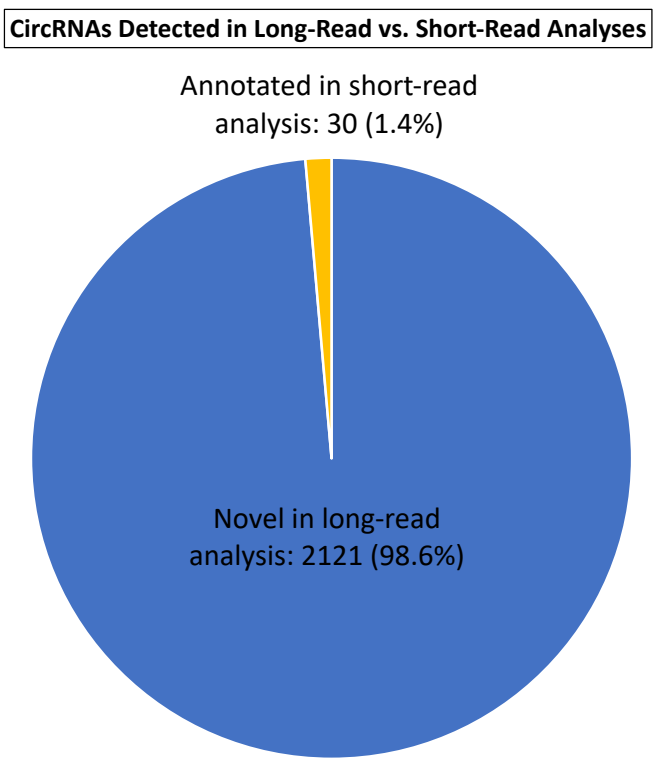
PVSNQSLLSESVASSQLDSTSVDKAVPETEDVQ



https://www.maherlab.com/peptransdb-circrna

Higher percentage of novel circRNAs were detected via long-read sequencing





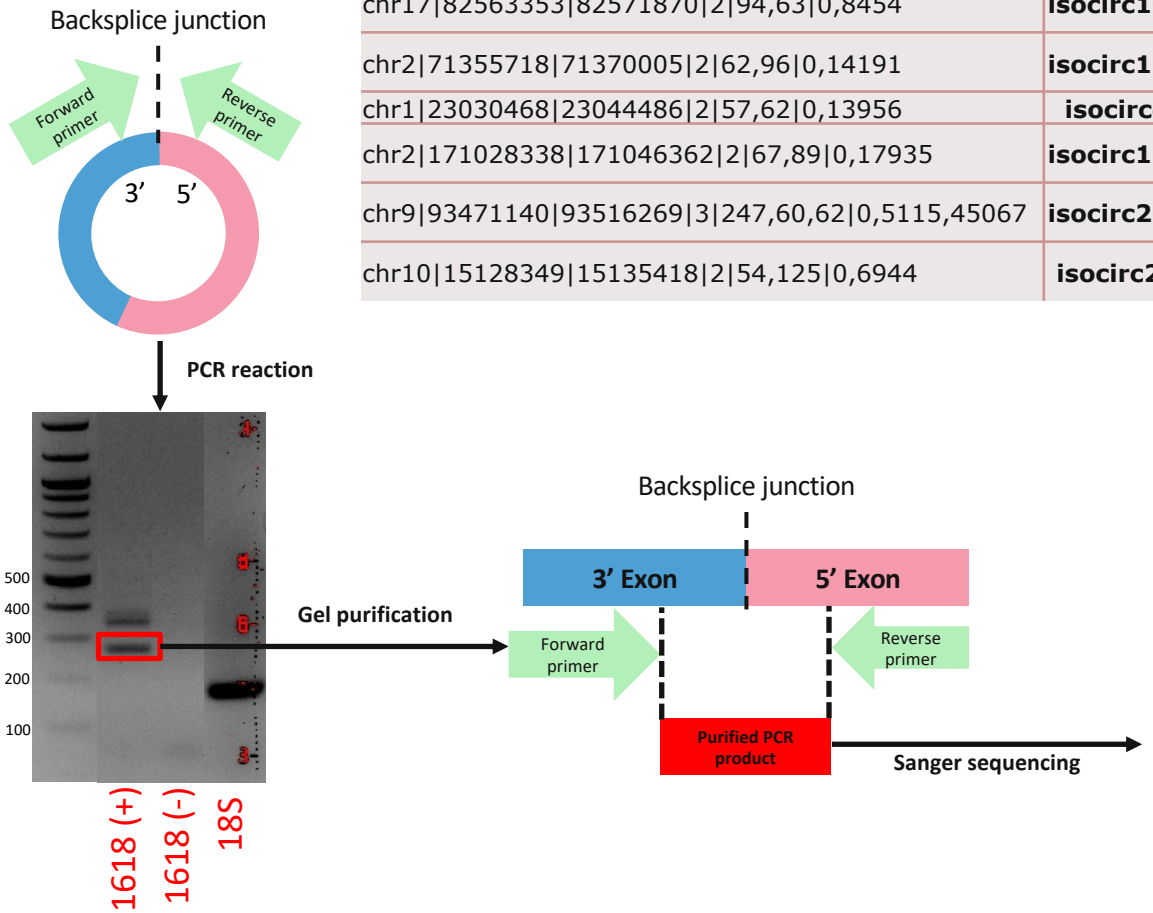
- ***What was missing in short-read?***
- ***How can we leverage long-read data to improve short-read results?***

Vo, J.N., et al., *The Landscape of Circular RNA in Cancer.* Cell, 2019.
Glažar, P., Papavasileiou, P., Rajewsky, N., *circBase: a database for circular RNAs.* RNA, 2014.

Unpublished

Validation of rescued circRNAs

CircRNA Unique Identifier	isoCirc ID	Validated in Experiment	Mean Rescued Read Number
chr4|48369848|48383784|2|149,147|0,13789	**isocirc1618**	Yes	5
chr17|82563353|82571870|2|94,63|0,8454	**isocirc1022**	Yes	3
chr2|71355718|71370005|2|62,96|0,14191	**isocirc1214**	Yes	2.5
chr1|23030468|23044486|2|57,62|0,13956	**isocirc47**	Yes	1.5
chr2|171028338|171046362|2|67,89|0,17935	**isocirc1259**	Yes	1
chr9|93471140|93516269|3|247,60,62|0,5115,45067	**isocirc2052**	Yes	0.5
chr10|15128349|15135418|2|54,125|0,6944	**isocirc299**	No	0.5

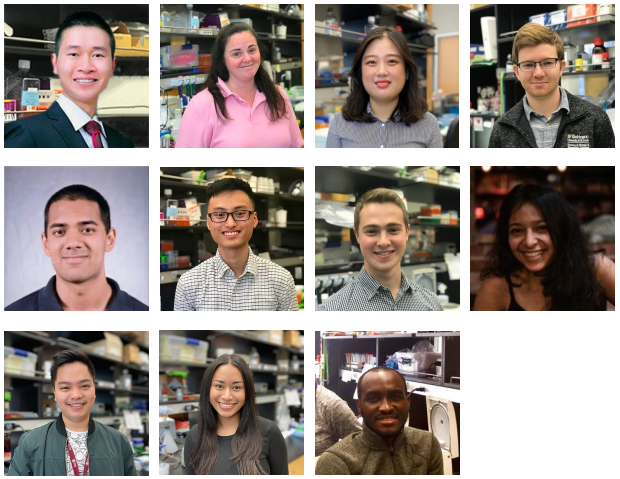


Unpublished

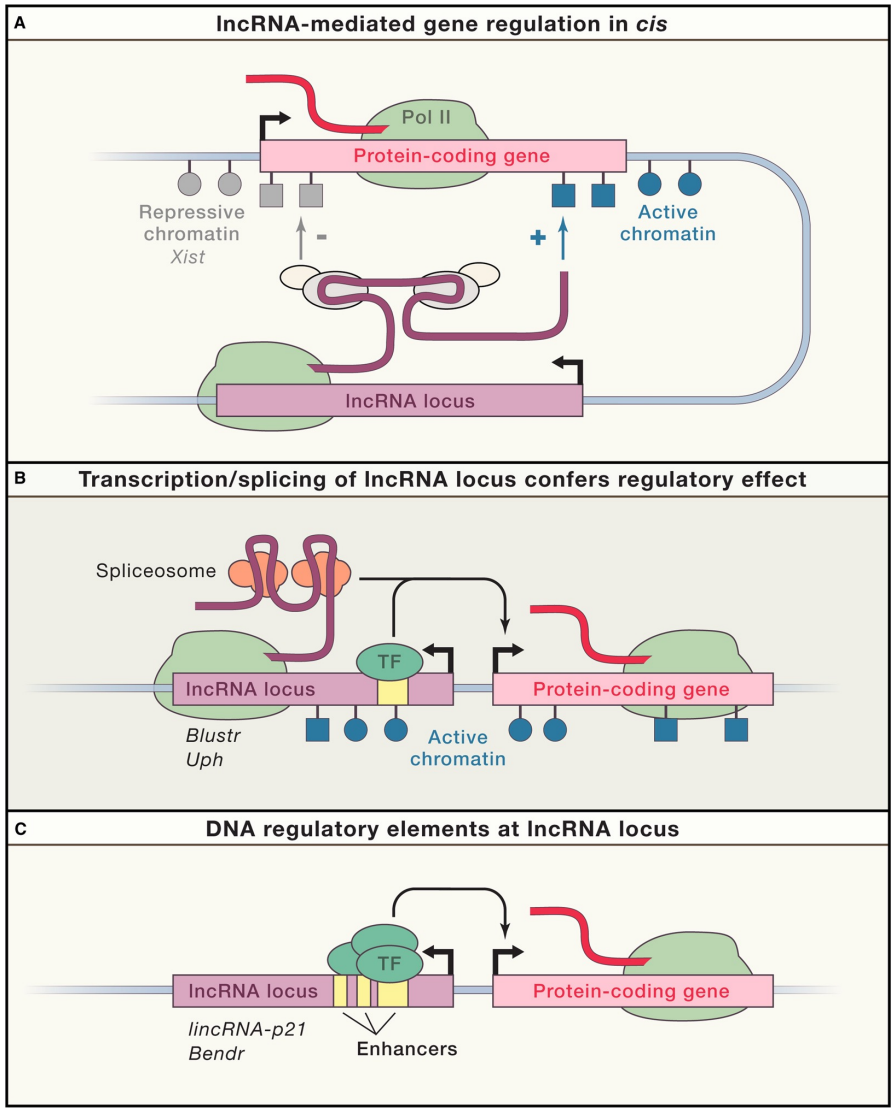
CircRNA conclusions

- Novel, integrated long-read approach discovers beyond annotated circRNAs
  - Eventual improvement to rely on a single strategy
- Improved bioinformatic workflow for comprehensive full-length circRNA characterization
- Aid in future mechanistic studies exploring their function in cancer, such as evaluating the coding potential of circRNAs
- **More cell line and matched patient long-read sequencing data will help to discover circRNAs and encoded peptides in matched cancer patients**

Acknowledgements



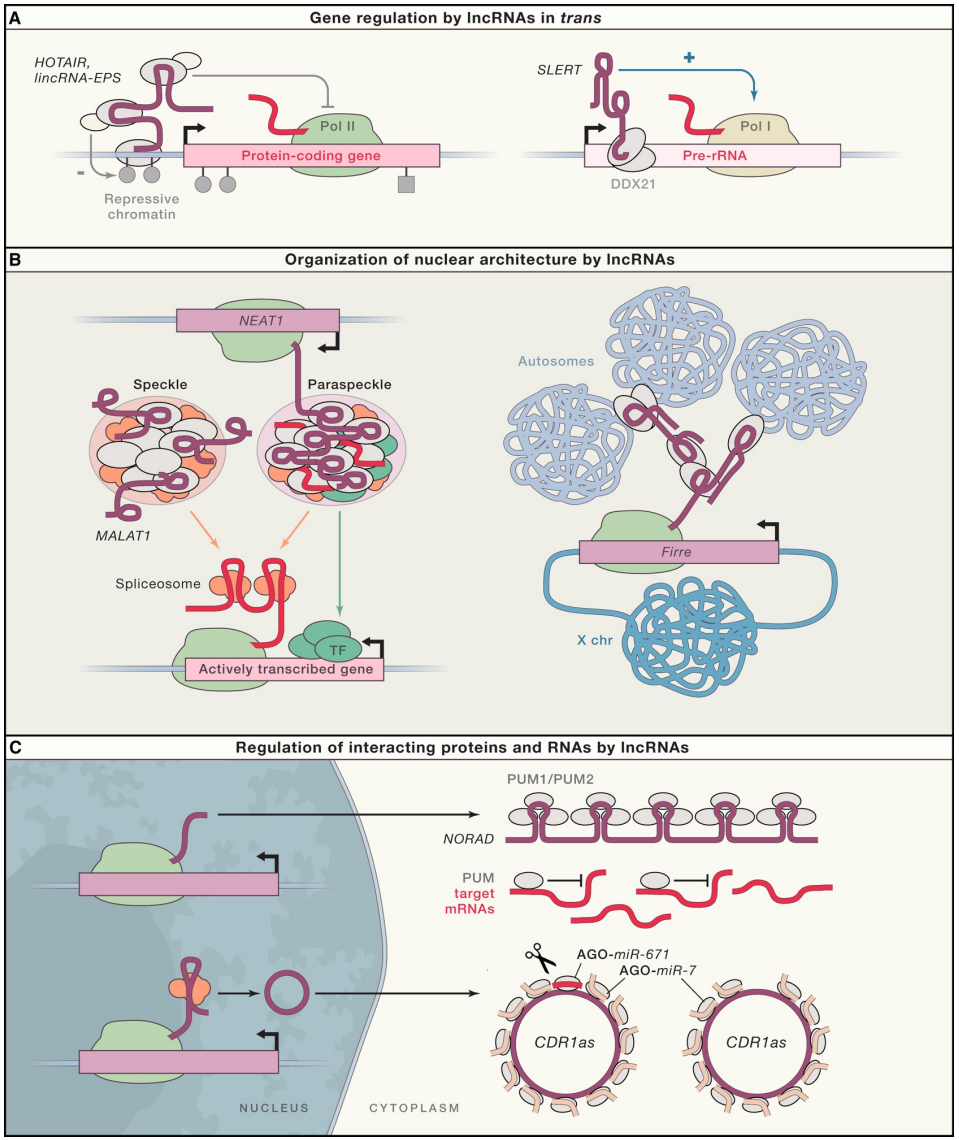
Functional roles of lncRNAs in *cis*



- The lncRNA transcript itself regulates the expression of neighboring genes through its ability to recruit regulatory factors to the locus and/or modulate their function
- The process of transcription and/or splicing of the lncRNA confers a gene-regulation functionality that is independent of the sequence of the RNA transcript
- Regulation in cis depends solely on DNA elements within the lncRNA promoter or gene locus and is completely independent of the encoded RNA or its production

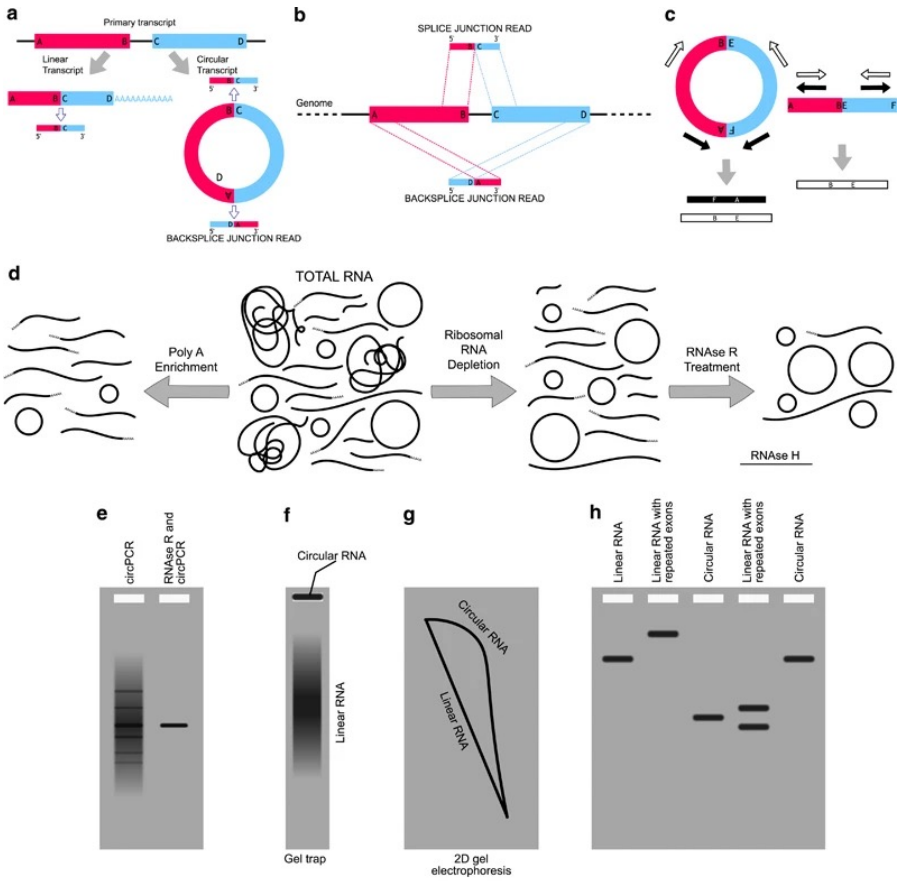
(Cell – Kopp et al., 2018)

Functional roles of lncRNAs in *trans*



(Cell – Kopp et al., 2018)

Enrichment strategies for circRNAs



Sequence reconstruction

CircRNAs length estimation can be obtained by Northern blot or PCR-based methods.

Selective amplification and direct sequencing provide the actual circRNA structure.

Expression quantification

- Only RNA-seq reads derived from the backsplice junction are private of the circRNA.

- CircRNA expression estimation is based on the number of detected backsplice junction reads, since most reads fall in sequences shared with linear transcripts.

(Bonizzato et al., 2016)