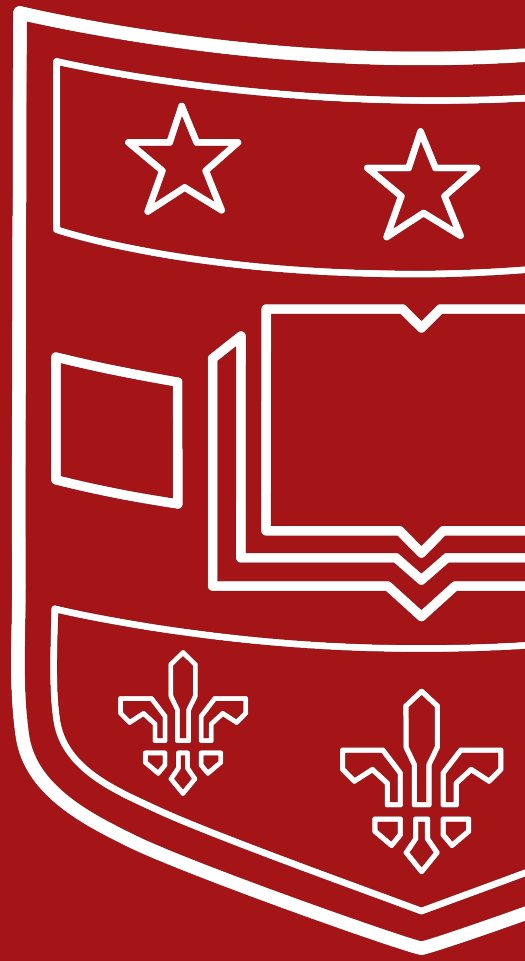
The long and short of noncoding RNAs

Christopher Maher

Professor, Internal Medicine and Biomedical Engineering

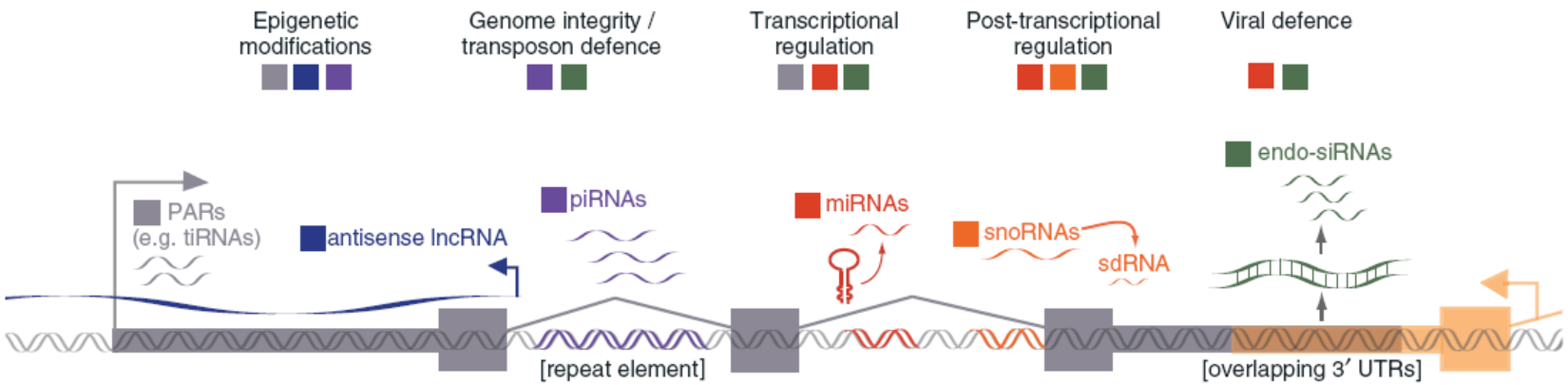
Washington University School of Medicine



November 9th, 2023

CSHL Advanced Sequencing Technologies & Bioinformatics Analysis Course

Diverse classes of non-coding RNAs (ncRNAs)



Classes of non-coding RNAs (ncRNAs)

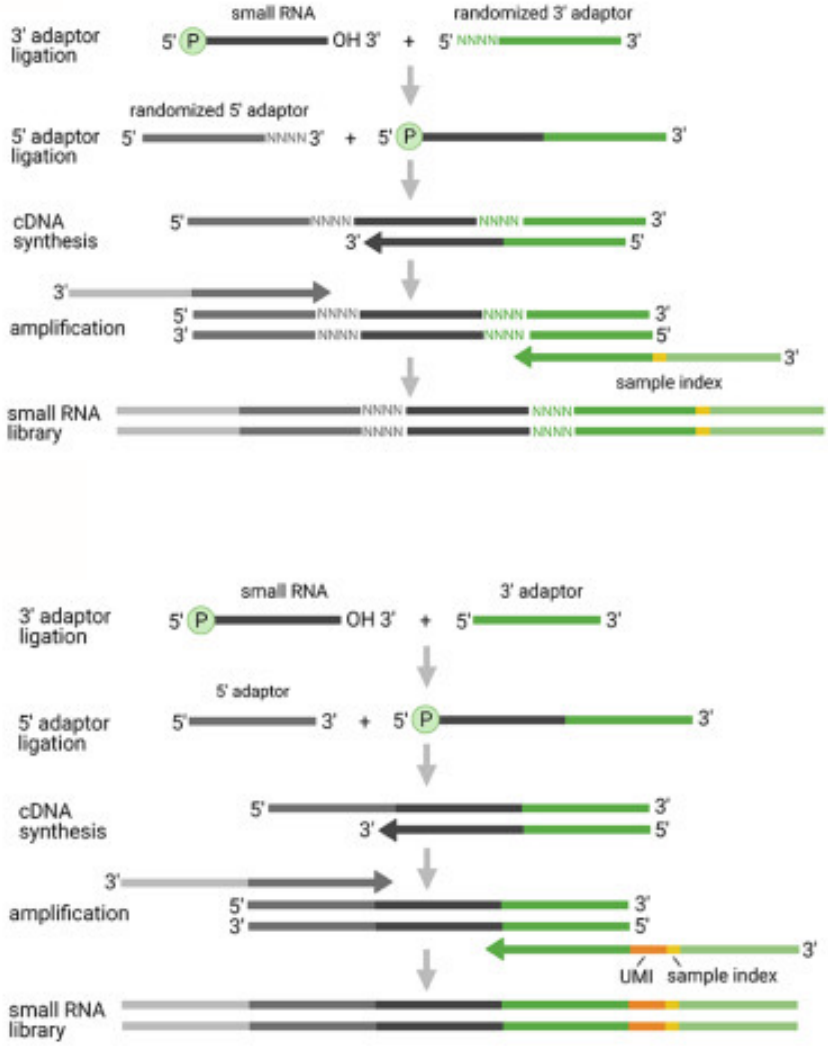
Category	Name	Quality of supporting data	Specific role in carcinogenesis	Aberration in cancer
Housekeeping RNAs	Transfer RNAs	High	No	No
	Ribosomal RNAs	High	No	No
	Small nucleolar RNAs	High	No	No
	Small nuclear RNAs	High	No	No
Small ncRNAs (<200 bp in size)	MicroRNAs	High	Yes	Amplification, deletion, 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)

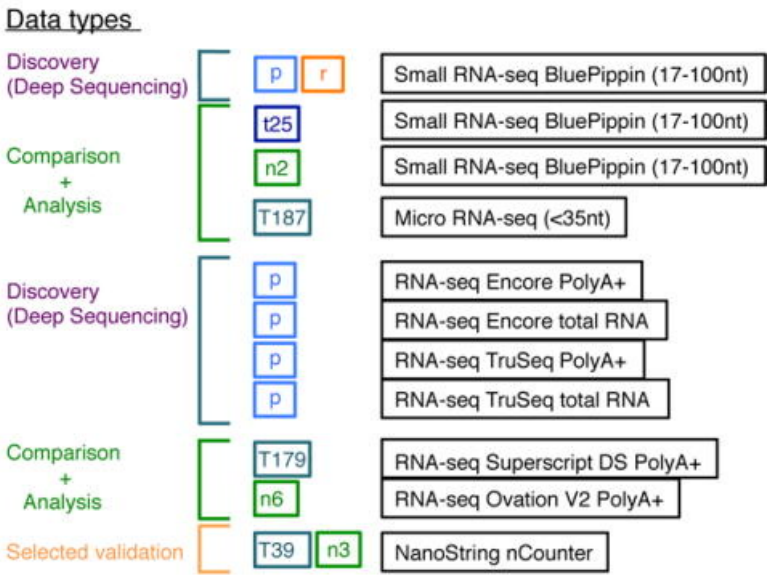
Small RNA sequencing

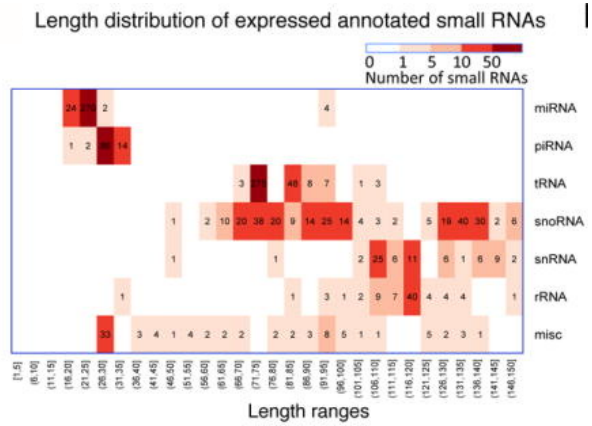
- Small RNA-seq workflow typically involves
  - Isolation of RNA
  - Adaptor ligation
  - cDNA library construction
  - Sequencing

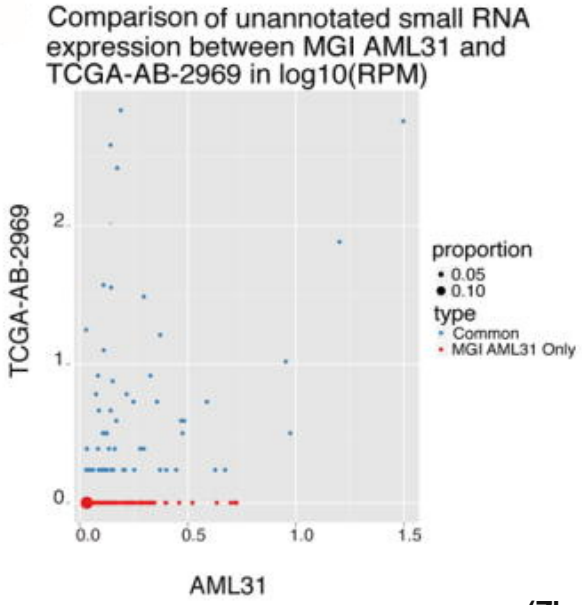


(Benesova et al., 2021)

Increasing insert size selection capture unannotated sncRNAs



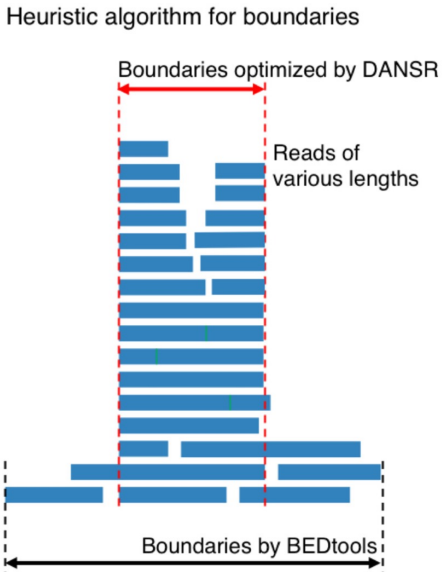


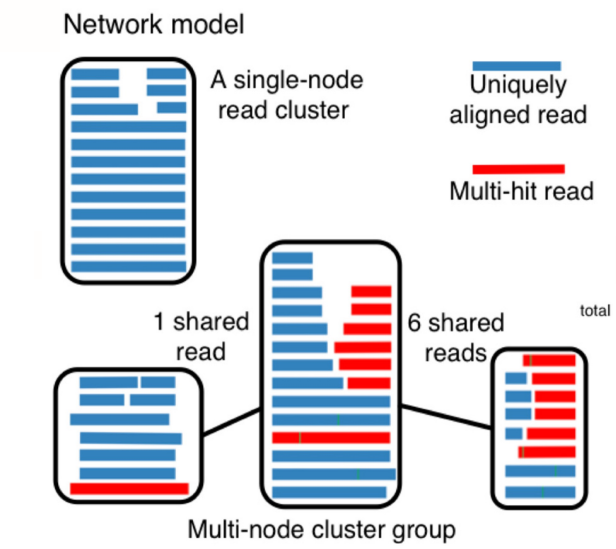


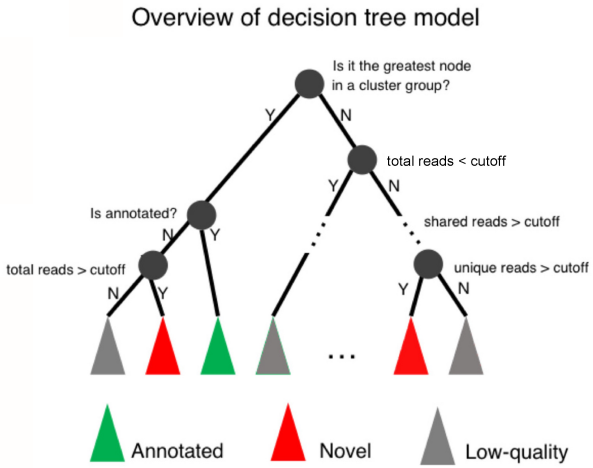
(Zhang et al., 2017)

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

- **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**

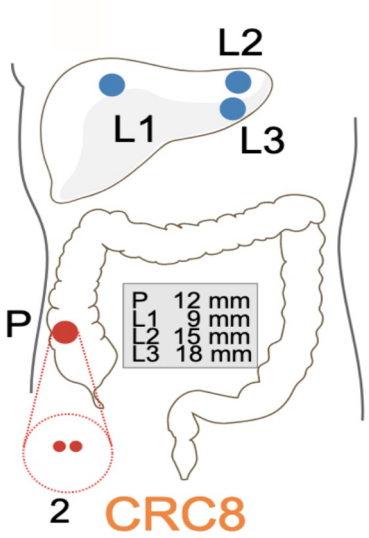


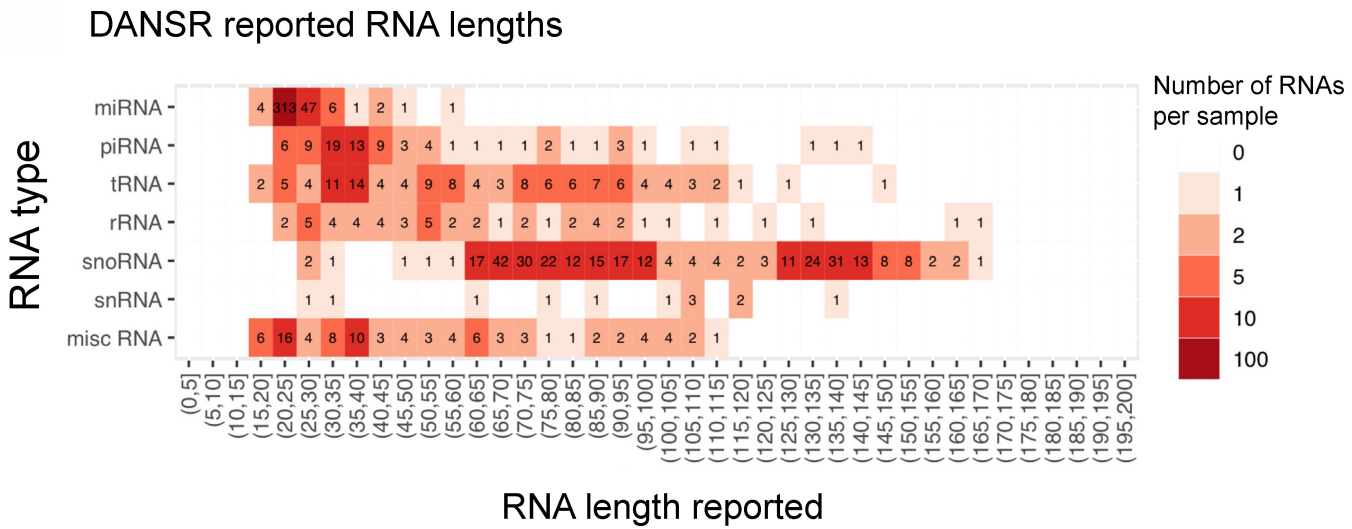




https://github.com/ChrisMaherLab/DANSR
(Eteleeb et al., 2022)

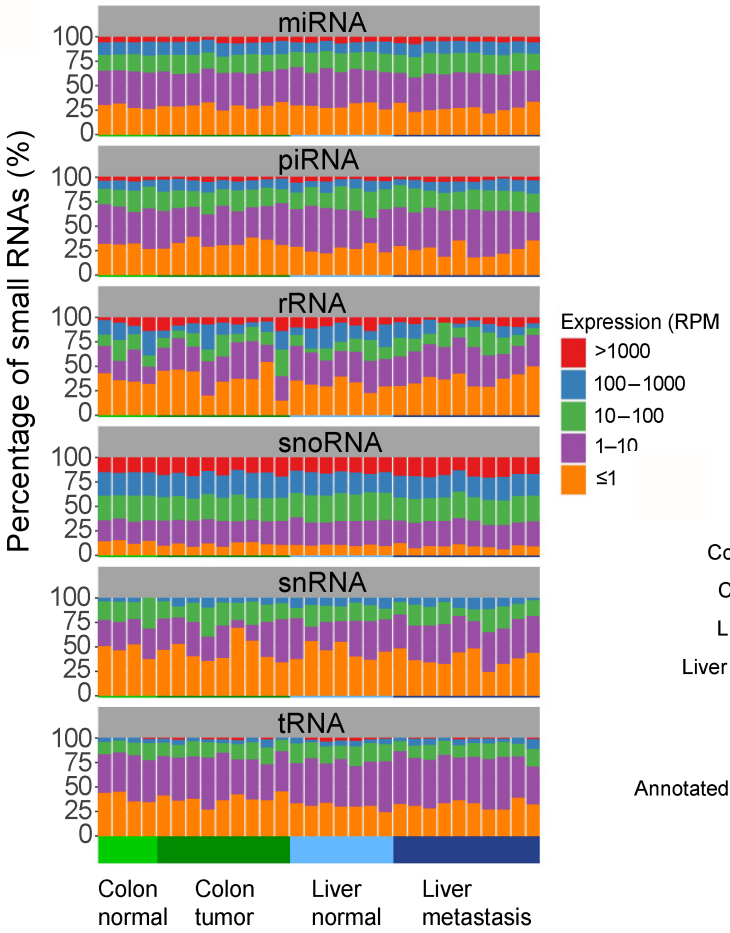
Accurate categorization of annotated small RNAs in metastatic colorectal cancer (mCRC) patients

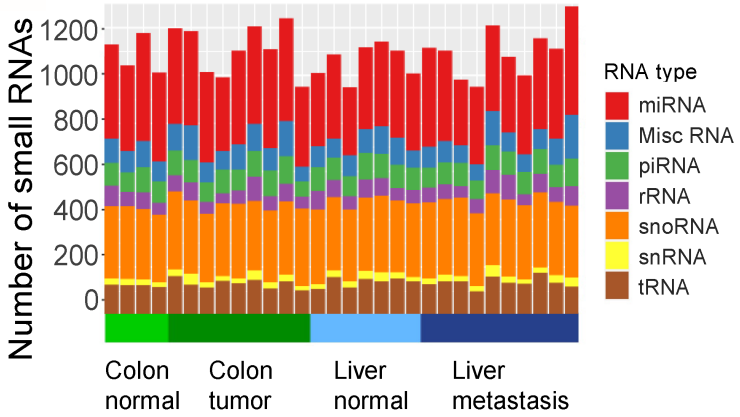


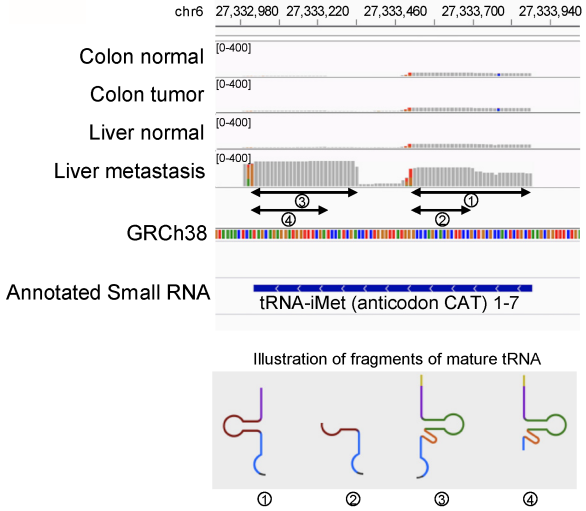


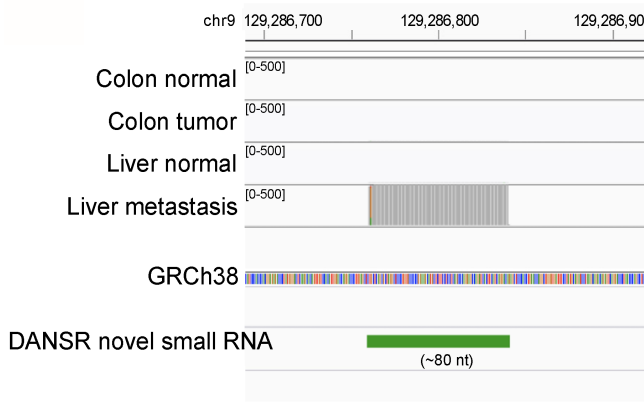
https://github.com/ChrisMaherLab/DANSR
(Eteleeb et al., 2022)

Discovery of altered small RNAs in metastatic colon cancer progression





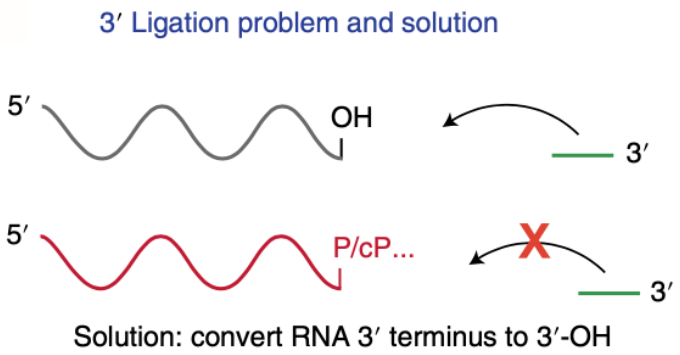




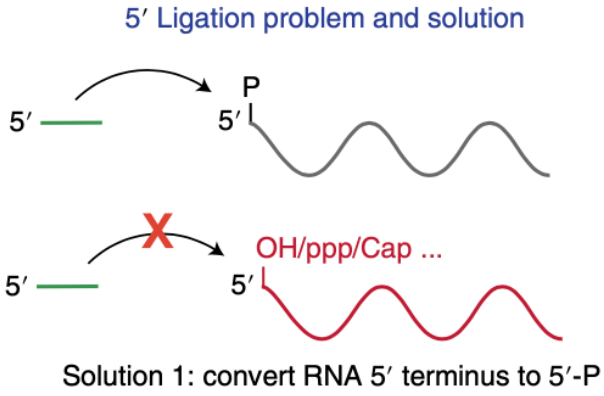
https://github.com/ChrisMaherLab/DANSR
(Eteleeb et al., 2022)

Adaptor-ligation process as a source of sequencing bias in sncRNA discovery

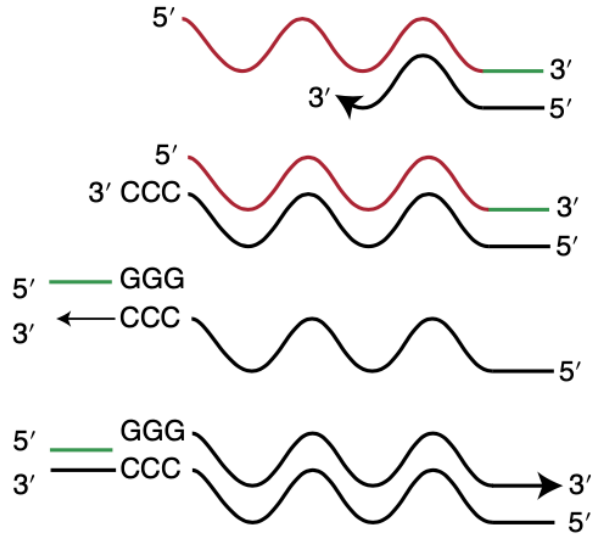
- Ligation process adds adaptor sequences to the termini of all sncRNAs in the pool
- Different sncRNAs harbor distinct termini generated by different enzymes and thus canno be uniformly ligated
- Most widely used sncRNA-sequencing protocol i optimized for 5′-P and 3′-OH termini



Solution: convert RNA 3′ terminus to 3′-OH



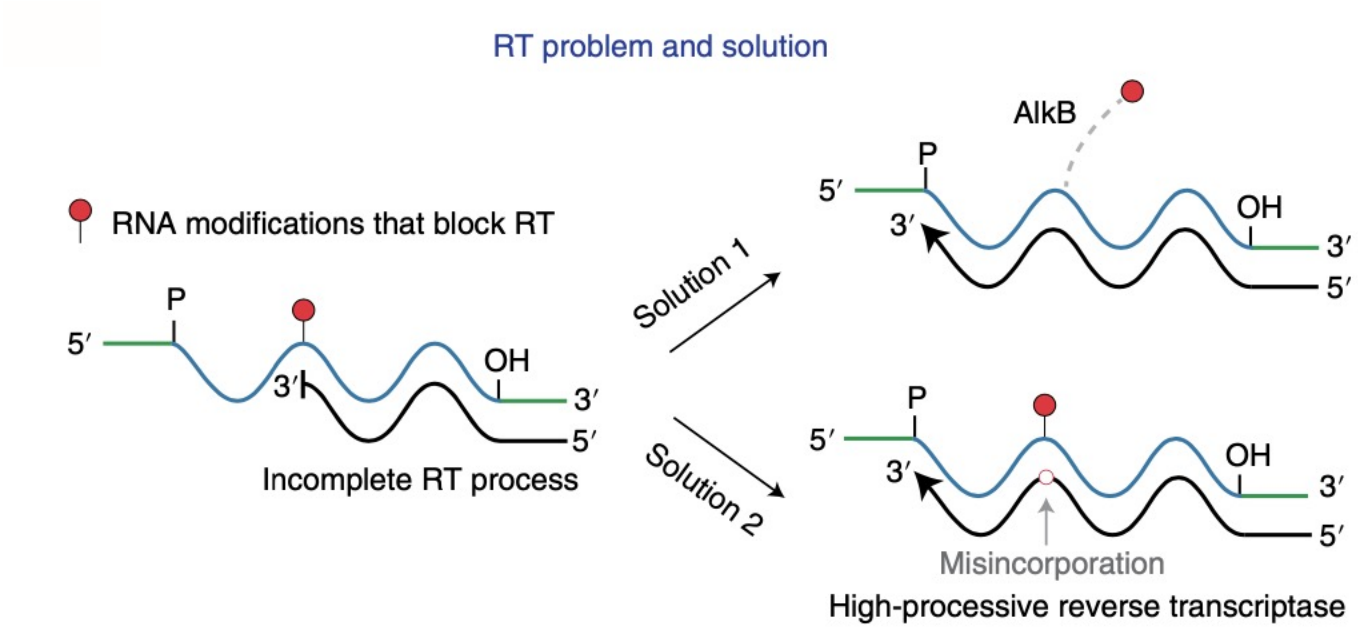
Solution 1: convert RNA 5′ terminus to 5′-P



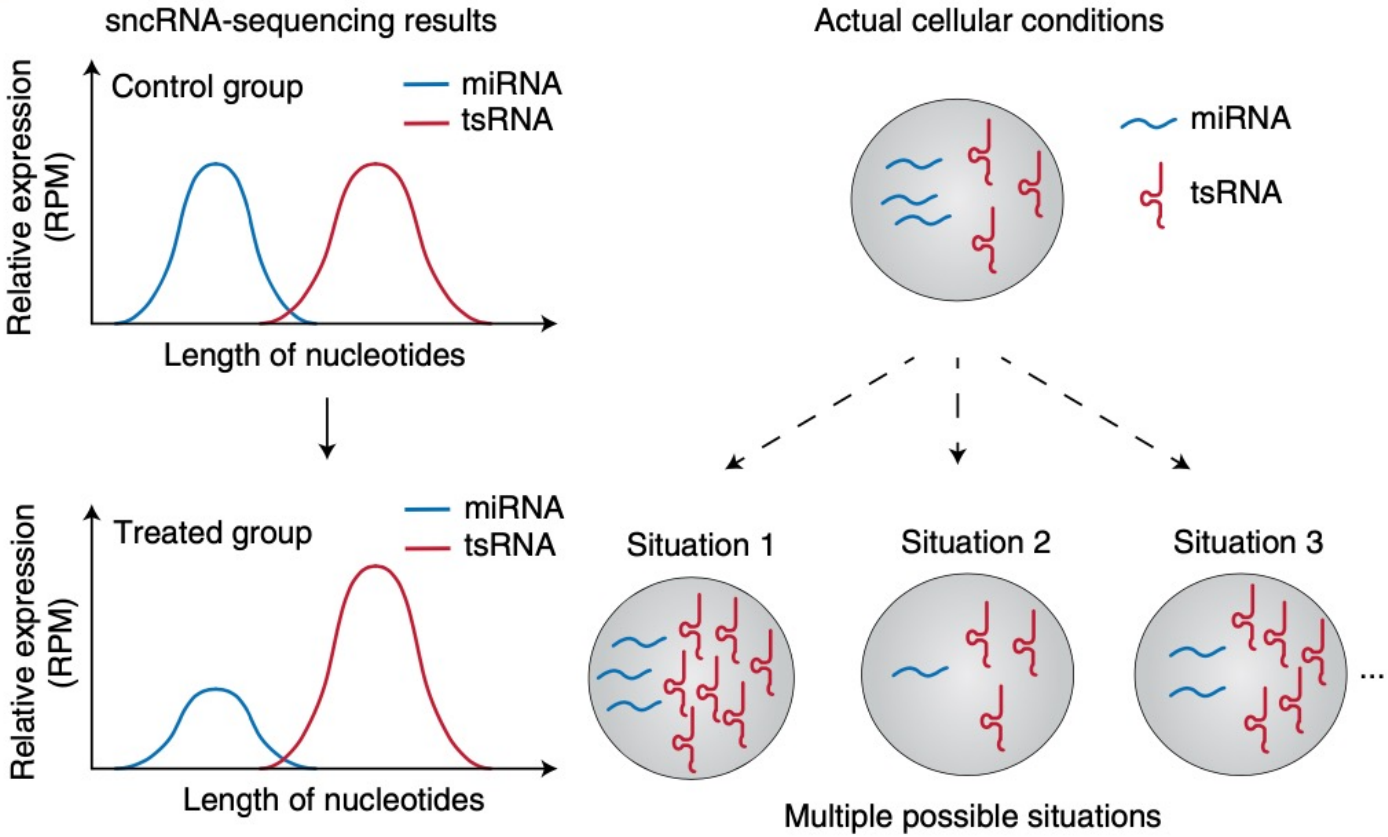
Solution 2: add 5′ adaptor to cDNA after template switching

Additional source of bias

- RT process converts the adaptor-ligated RNA into cDNA
- If RT process is interrupted before reaching the 5′ terminus, truncated cDNA will not be further amplified from the 5′ end and will not be detected

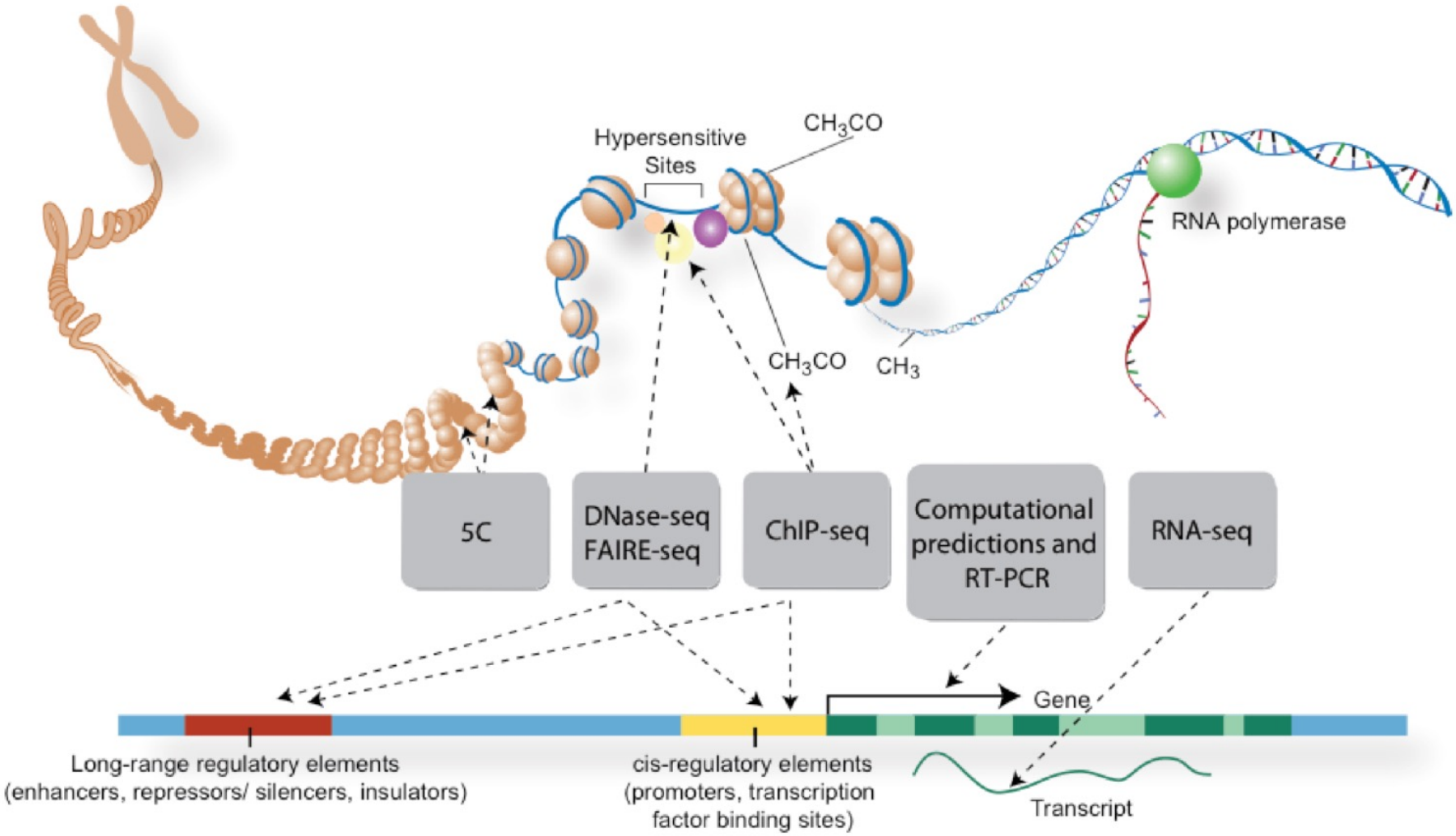


Caveats to analyzing sncRNA sequencing

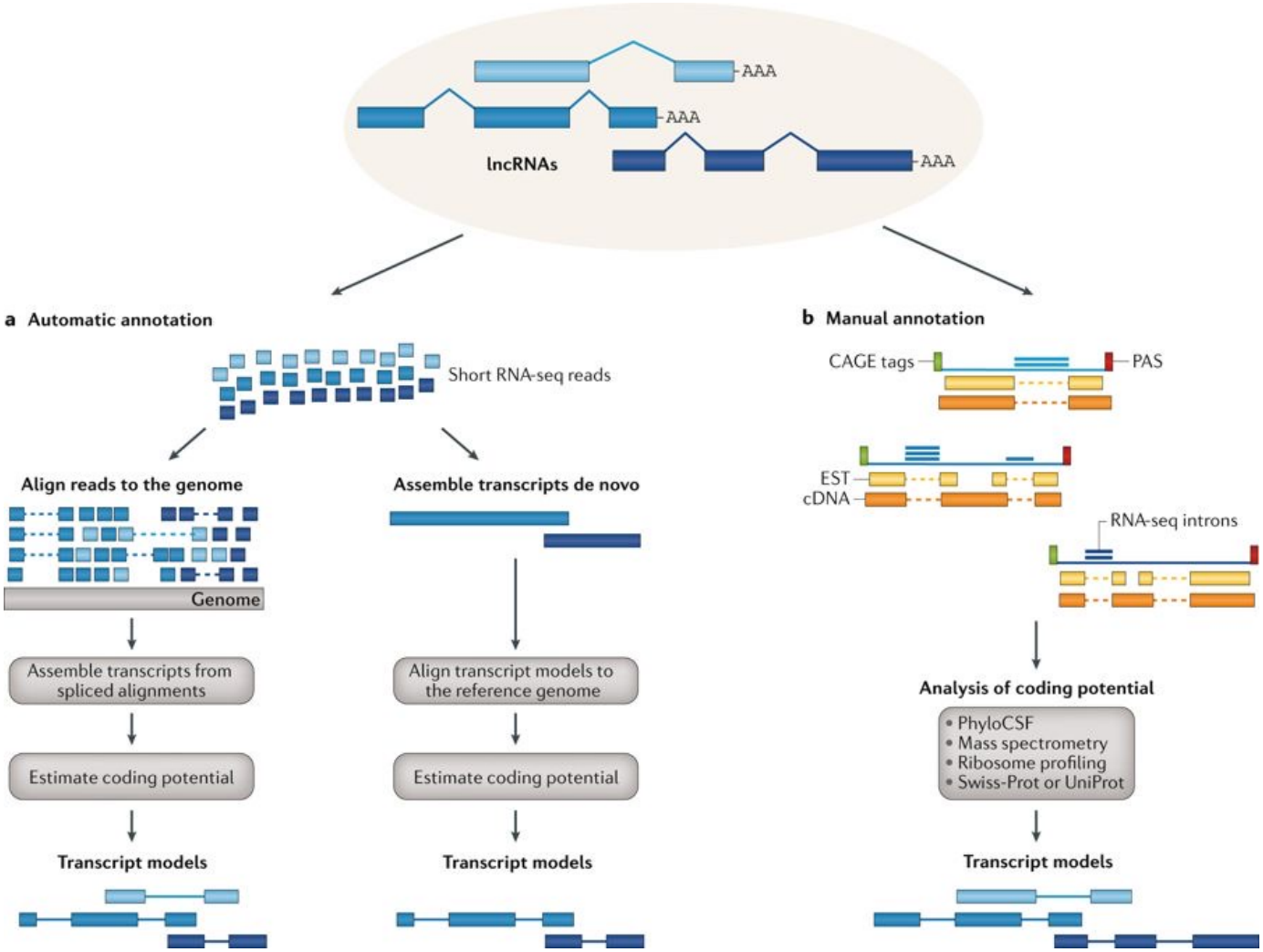


Long noncoding RNAs (lncRNAs)

Integrative methods for discovering lncRNAs

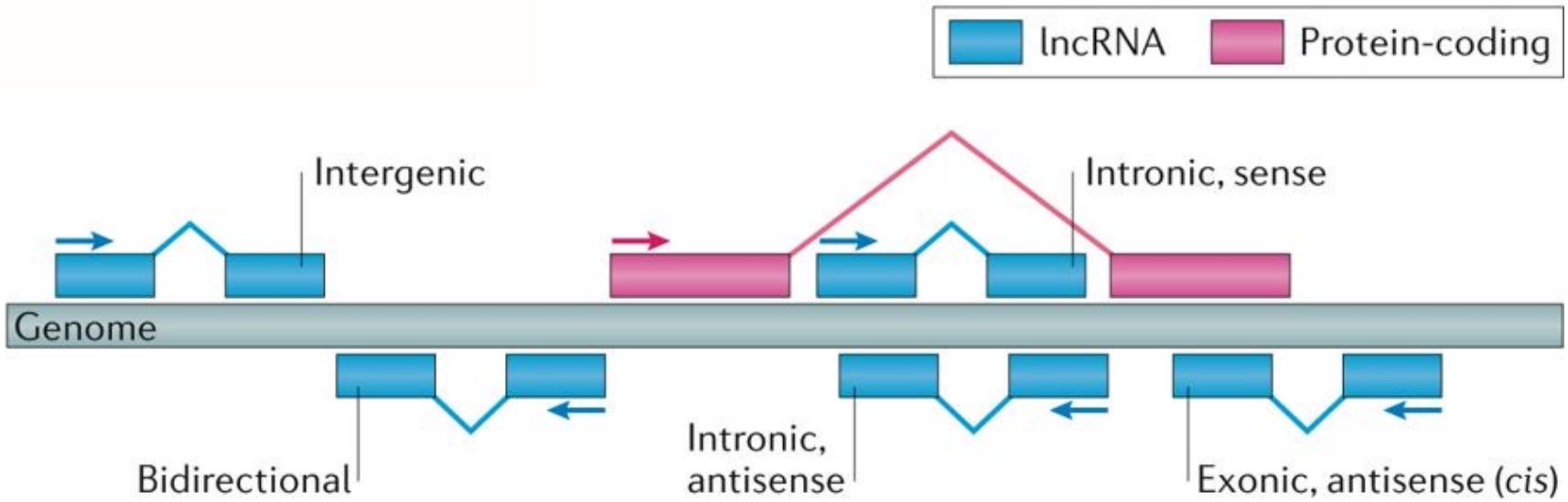


RNA-Seq focused strategies for lncRNA discovery



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

Positional classification of lncRNAs



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;

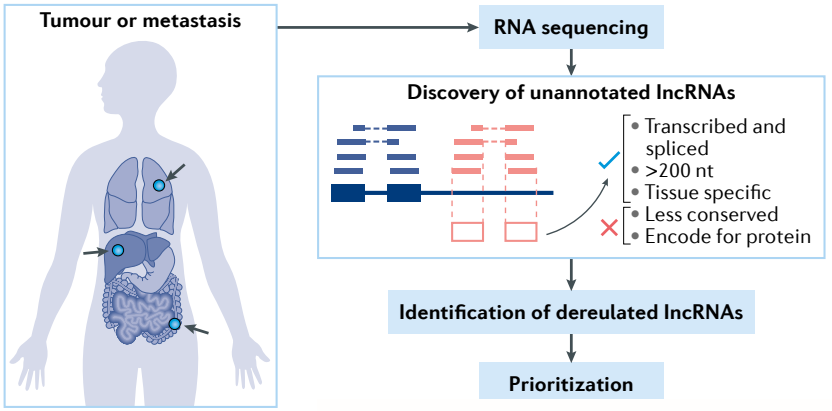
Completeness
The fraction of transcript models that cover the entire length, from start to end, of the physical
RNA molecule

(Uszczynska-Ratajczak et al., 2018)

Characteristics of long noncoding RNAs

- **Expression**
  - Restricted expression in different cells / stages of development and generally low copy number (owing to their regulatory nature) accounts for their sparse representation in bulk- tissue RNA sequencing datasets
- **Regulation**
  - Transcription via RNA polymerases (I, II, III)
  - Polyadenylation
  - Frequent splicing of multiple exons via canonical genomic splice site motifs (a feature that was not obvious before the advent of high-depth sequencing)
  - Regulation by well-established transcription factors (promoters exhibit levels of conservation comparable to protein-coding genes)
- **Conservation**
  - LncRNAs evolve rapidly due to more relaxed structure–function constraints
  - LncRNAs also have conserved exon structures, splice junctions and sequence patches
  - Retain orthologous functions despite rapid sequence evolution Indeed, low sequence conservation can be misleading
  - At least 18% of the human genome is conserved among mammals at the level of predicted RNA structure

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



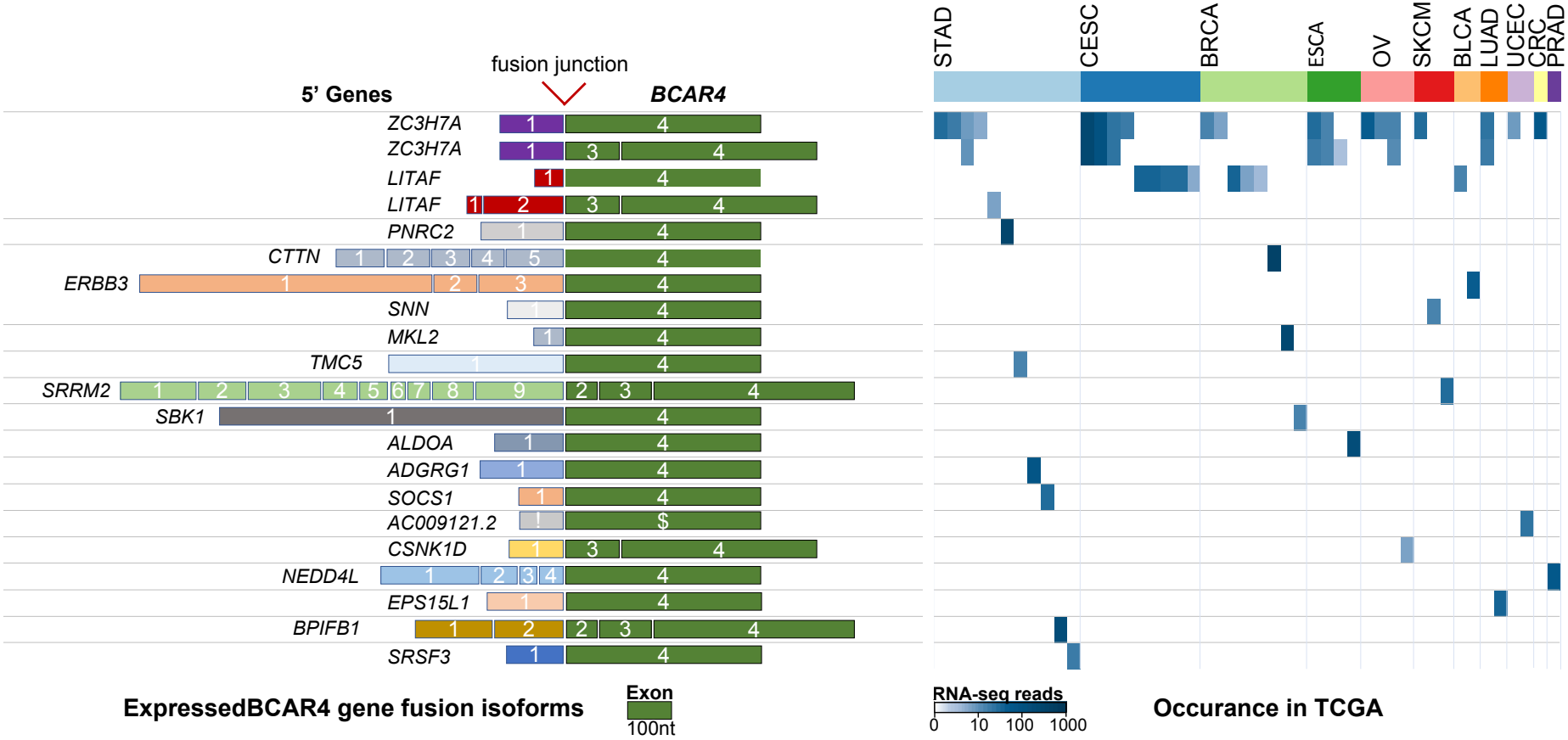
Challenges:

- 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)

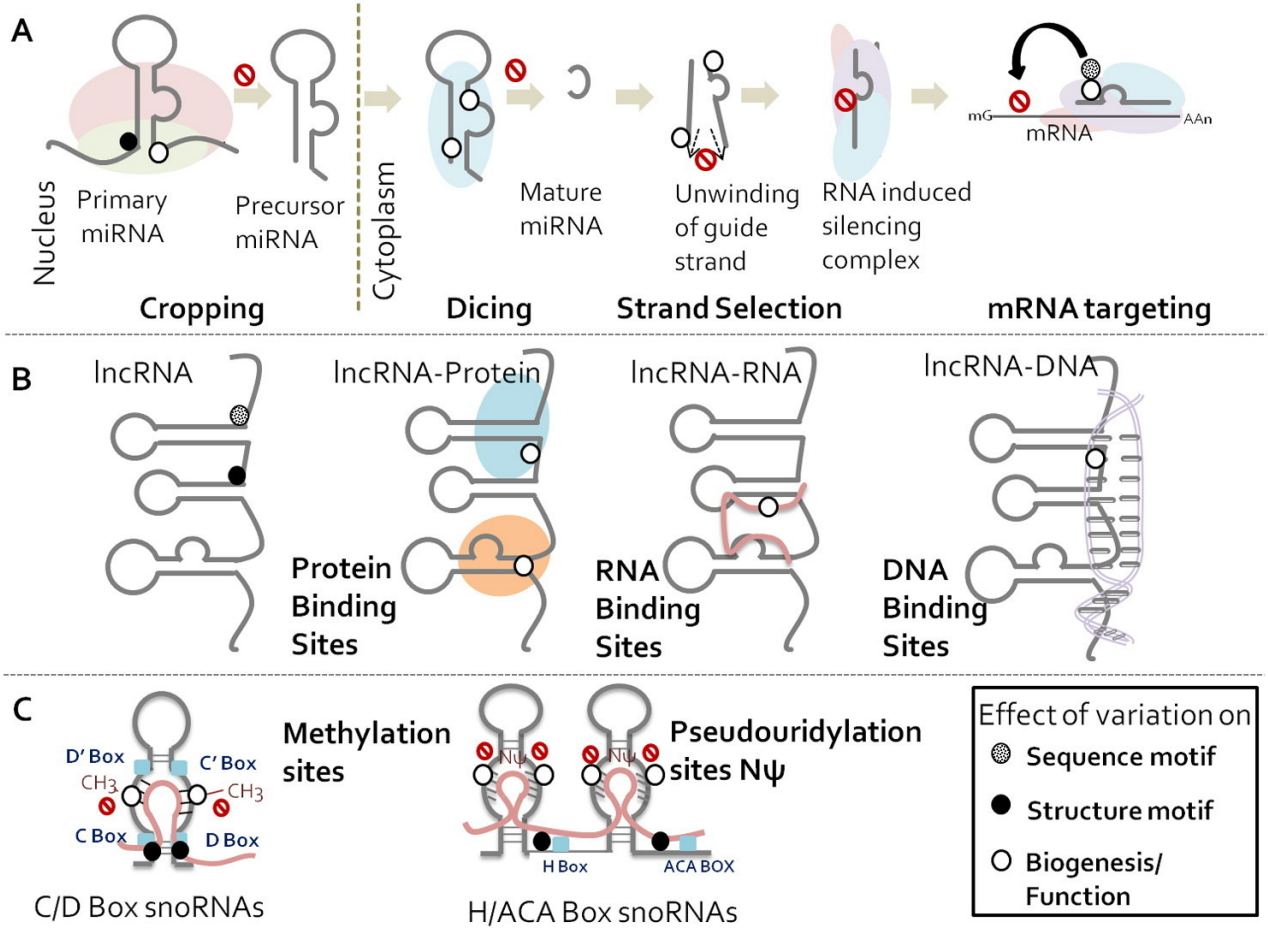
Recurrent somatic mutations suggest importance in cancer

Common gene fusions activating 3’ end of an annotated lncRNAs



(Nickless et al., 2022)

Genetic variations can altert A) microRNAs; B) long non-coding RNAs; C) small nucleolar RNAs.

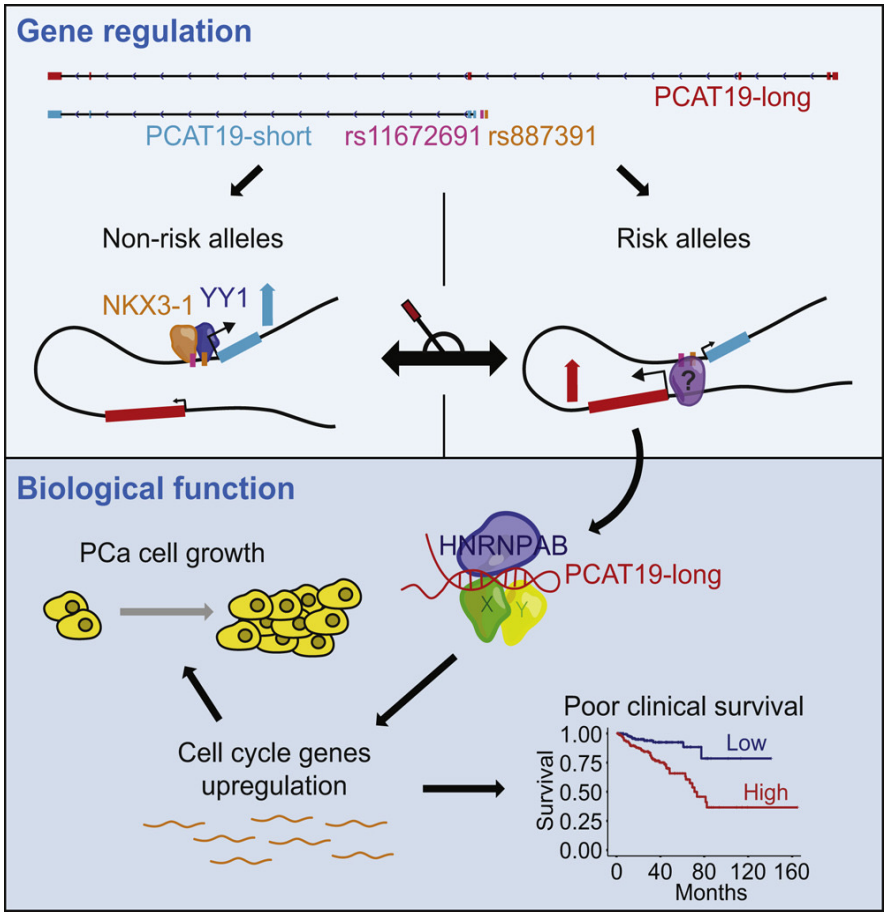


(Bhartiya et al., 2016)

Databases curating genomic variations in non-coding RNAs

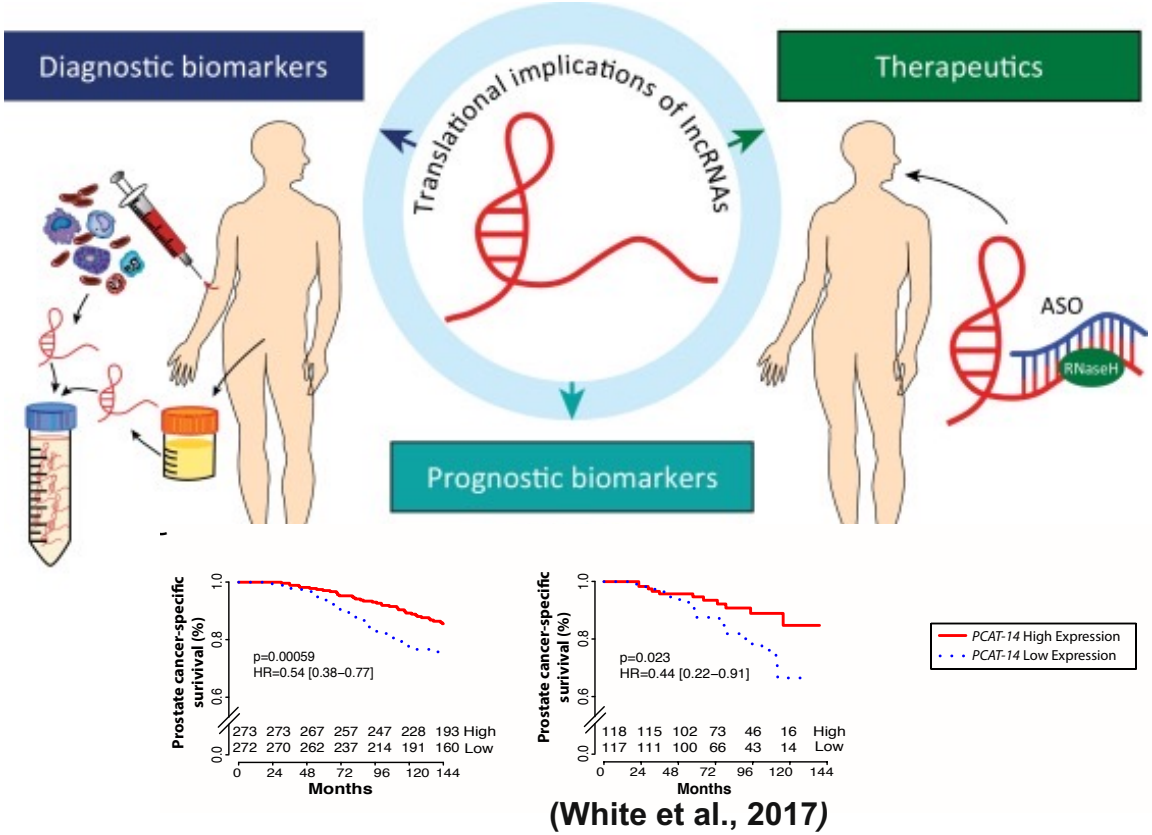
Database	Features
PolymiRTS	DNA variations in miRNA seed sequences
lncRNASNP	SNPs in lncRNAs, effects on lncRNA structure and lncRNA:miRNAbinding; functional SNP selection
rSNPbase	Database of curated regulatory SNPs
lincSNP	Disease-associated SNPs in human lincRNAs
SNP@lincTFBS	SNPs in potential TFBSs of human lincRNAs
lncRNAdisease	Experimentally supported lncRNA-disease association data
lncRNome	SNPs in human lncRNAs
snolovd	SNPs in human snoRNA loci
miRNASNP	SNPs in human pre-miRNAs and miRNA flanks
miRNASNiPer	Polymorphisms within miRNA genes in vertebrates
miRvar	Variations in human miRNA loci

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



(Cell -- Hua et al., 2018)

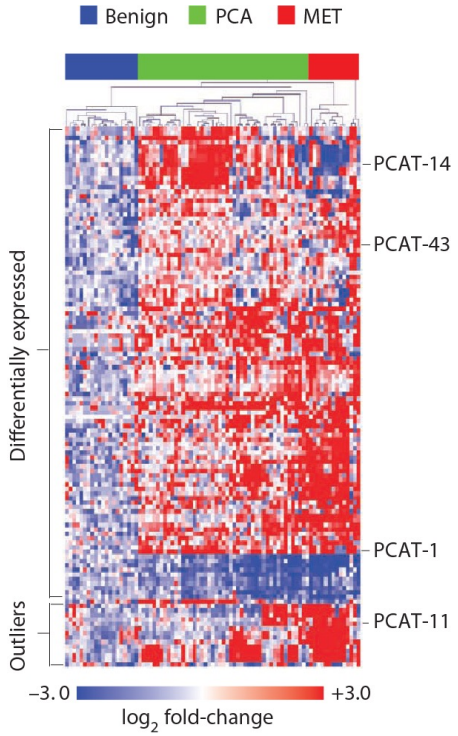
Clinical applications of lncRNAs



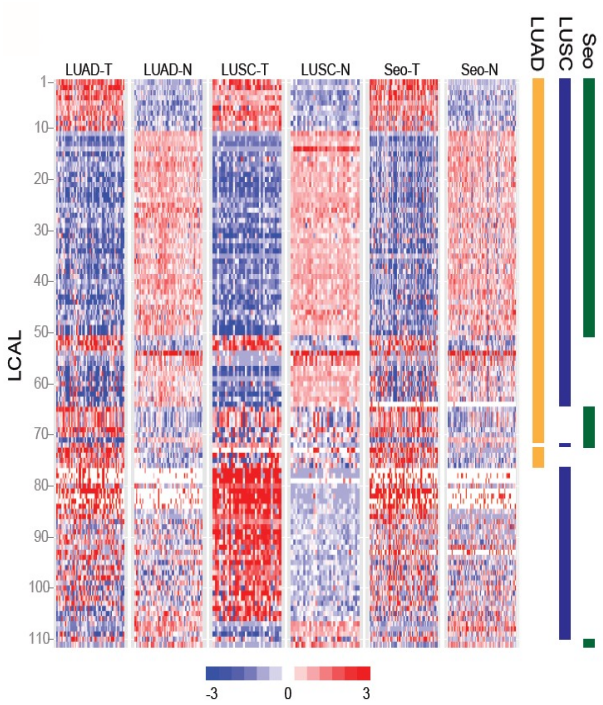
(White et al., 2017)

- 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

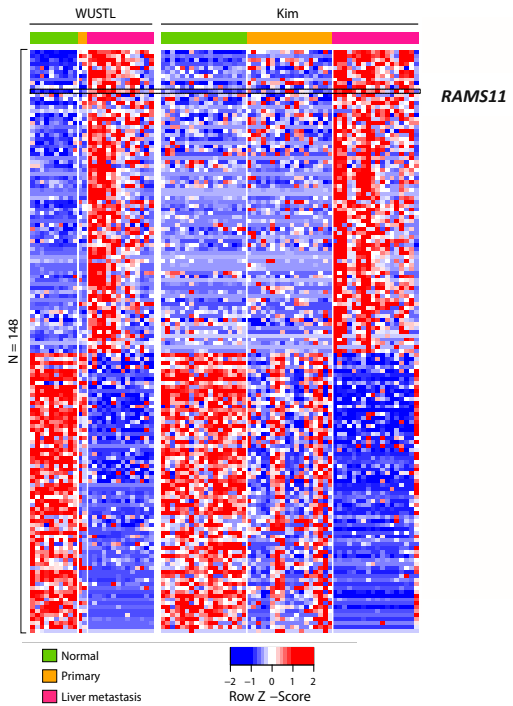
Only a fraction of lncRNAs are altered in a given cancer type



- 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)



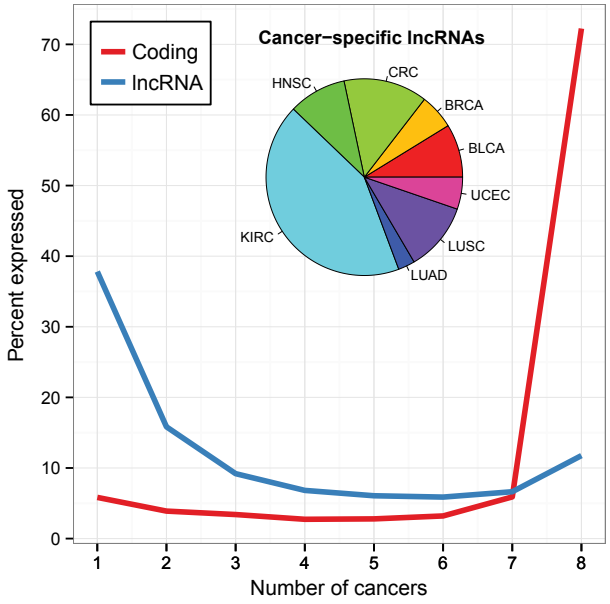
- 148 lncRNAs that performed as well as known biomarkers in differentiating benign, primary, and metastatic tissues
- 51 lncRNAs differentially expressed in metastatic tumors compared to non-metastatic (primary and adjacent normal)
  - 17 Unannotated
- Referred to as **R**NAs **A**ssociated with **M**etastasi**s**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

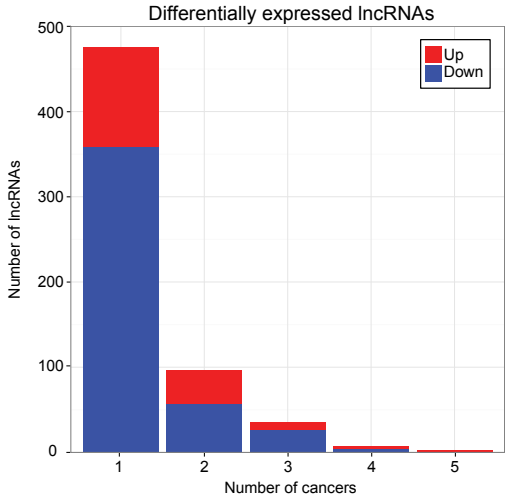
(Nature Biotechnology-- Prensner et al., 2011)

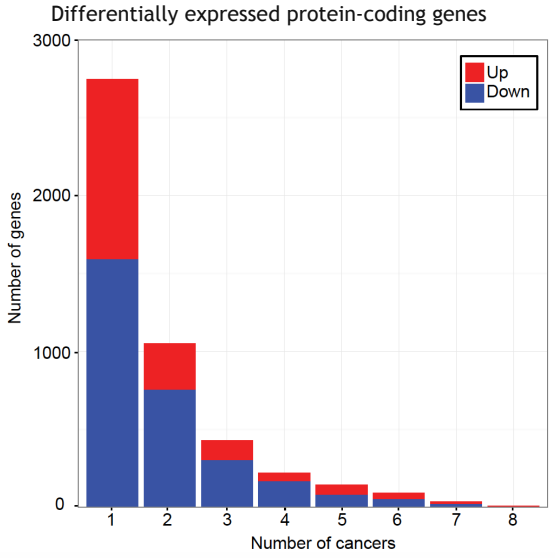
(Genome Biology -- White et al., 2014)

(Nature Communications –Silva et al., 2021)

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



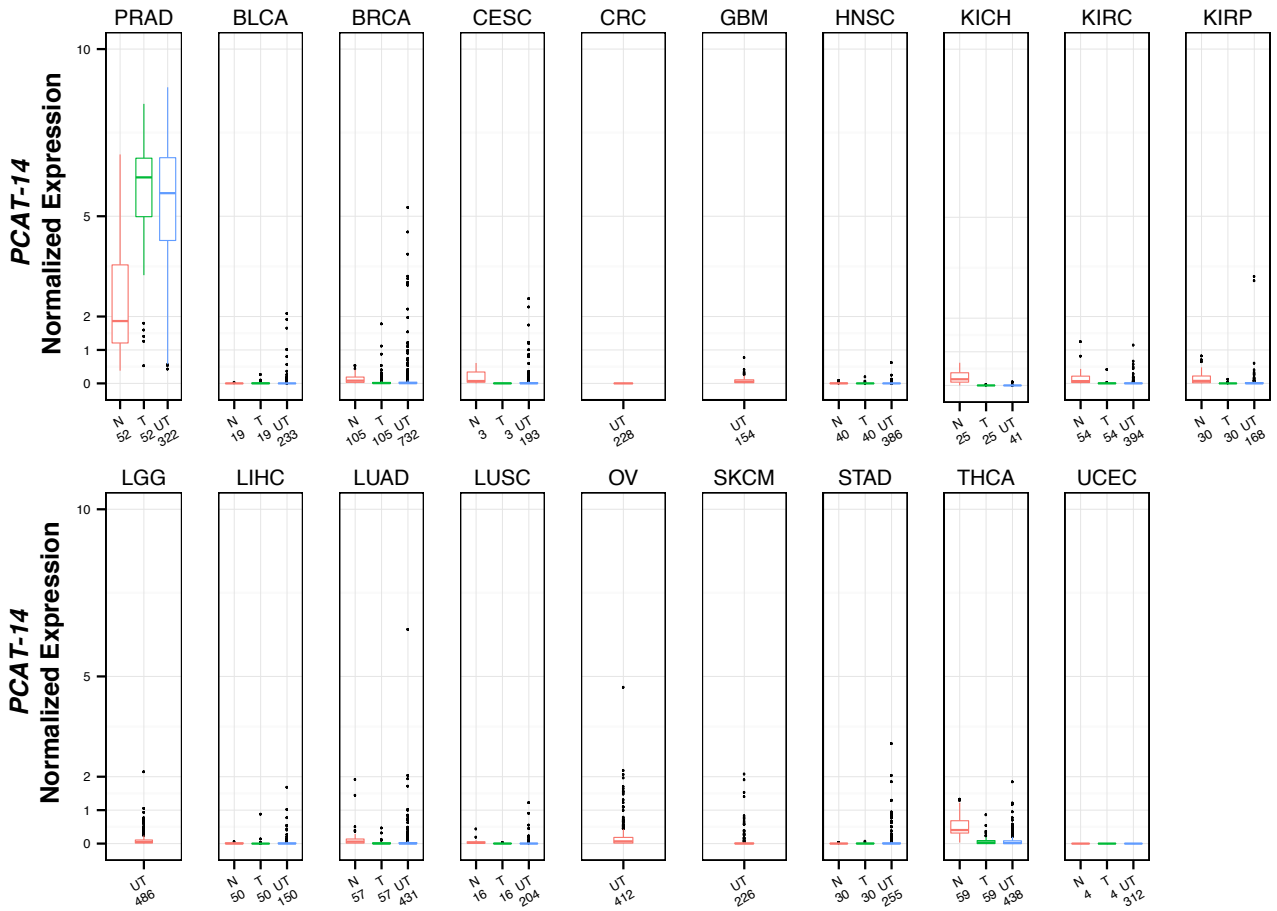




- ~10% of protein-coding genes are altered across 2 or more cancer types
- ~2% of lncRNAs are altered across 2 or more cancer types

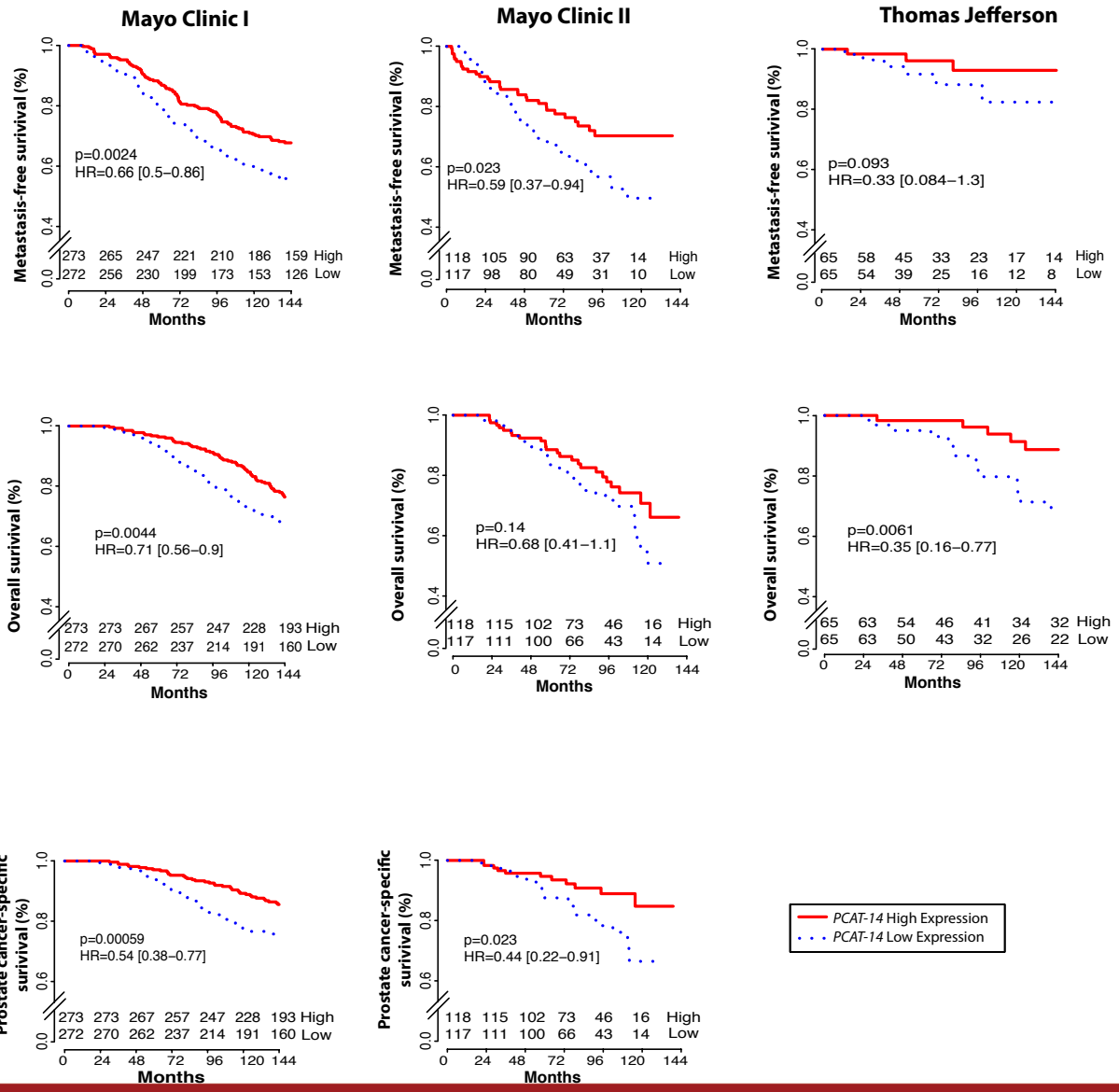
(Cabanski et al., 2015)

PCAT-14 expression is enriched in prostate cancer



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

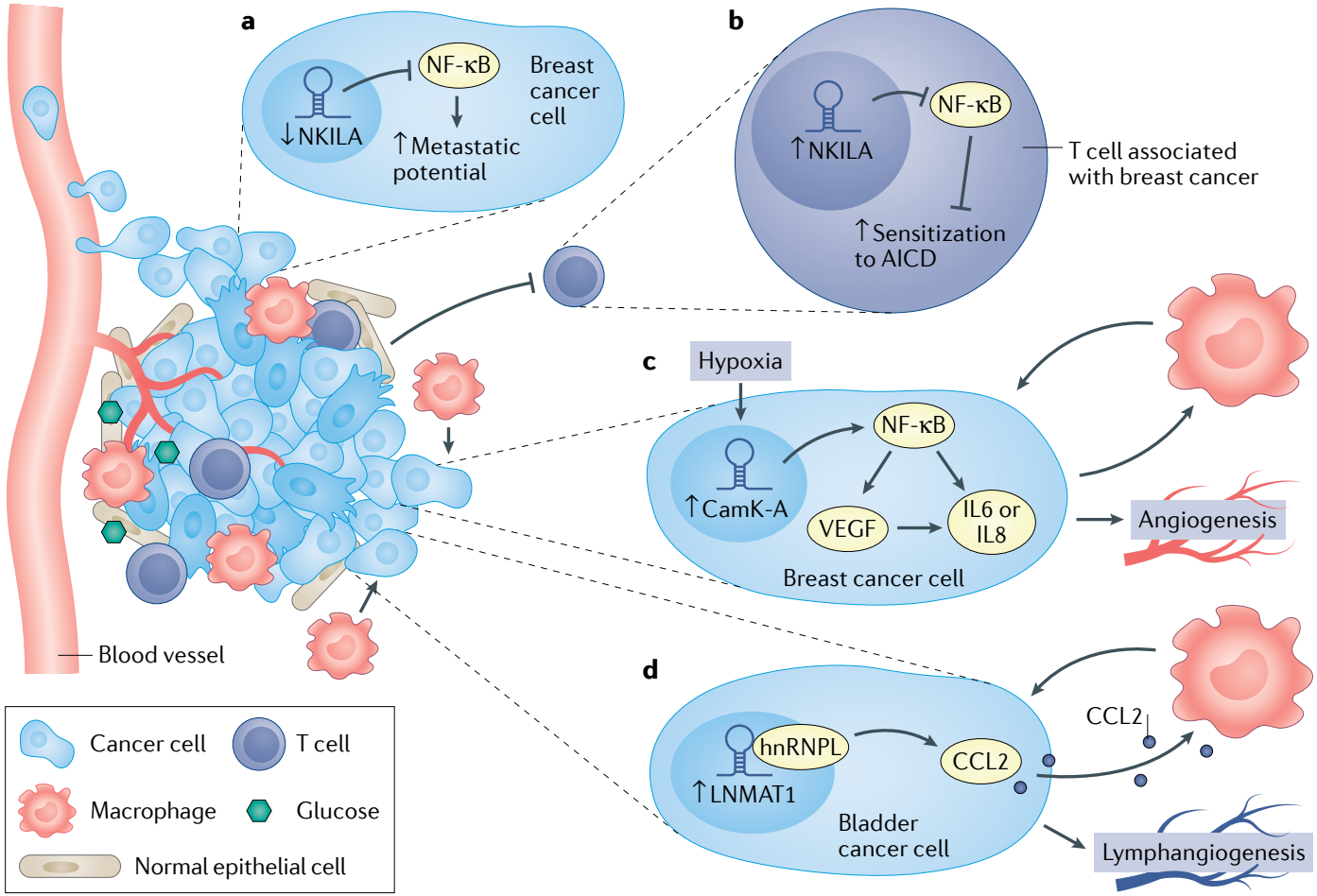
PCAT-14 as a single gene predictor of aggressive disease



How does *PCAT14* promote aggressive phenotypes in prostate cancer?

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

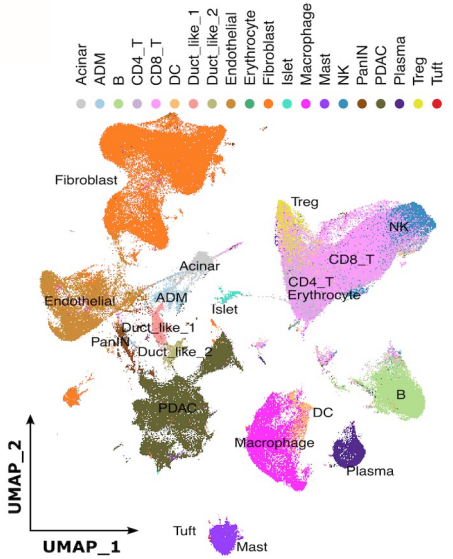
lncRNAs and the tumor microenvironment



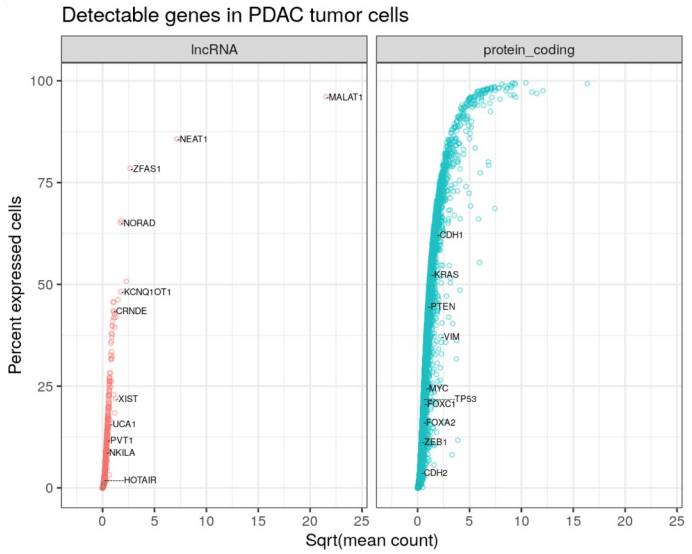
(Nature Reviews Cancer Liu et al., 2021)

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

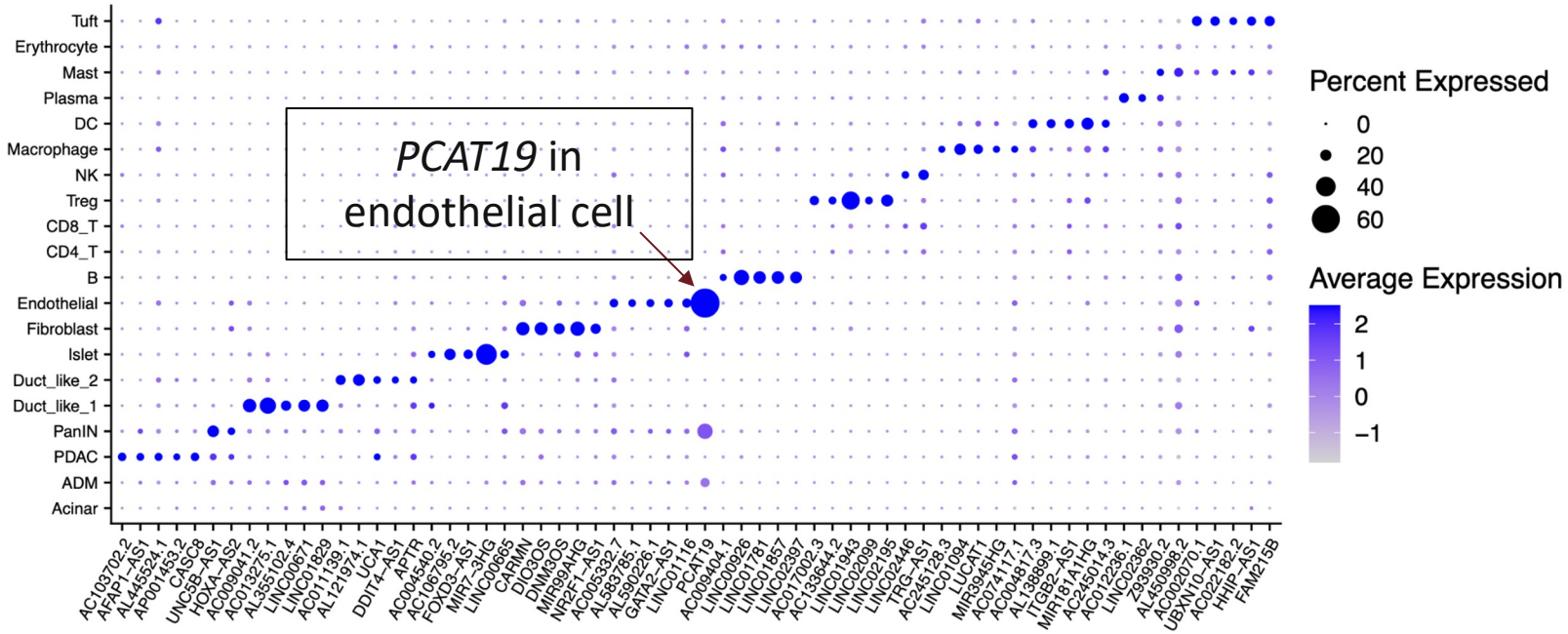


- Most genes are only detected in a small fraction of cells
- LncRNAs are more likely to be missed at individual cell level due to their lower expression level
- Significant number of lncRNAs are detected in as many cells as protein coding cancer genes



(*Accepted* NAR Cancer – Saha et al., 2023)

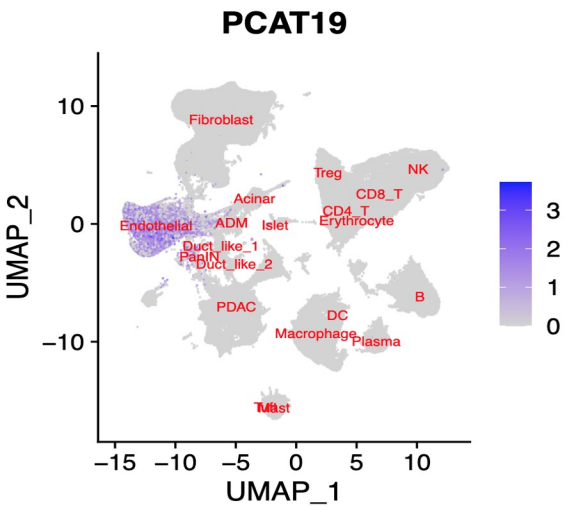
LncRNAs as markers of PDAC TME cell types

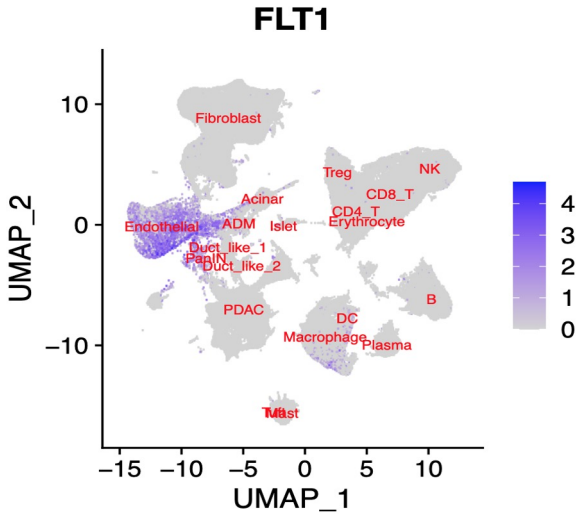


(*Accepted* NAR Cancer – Saha et al., 2023)

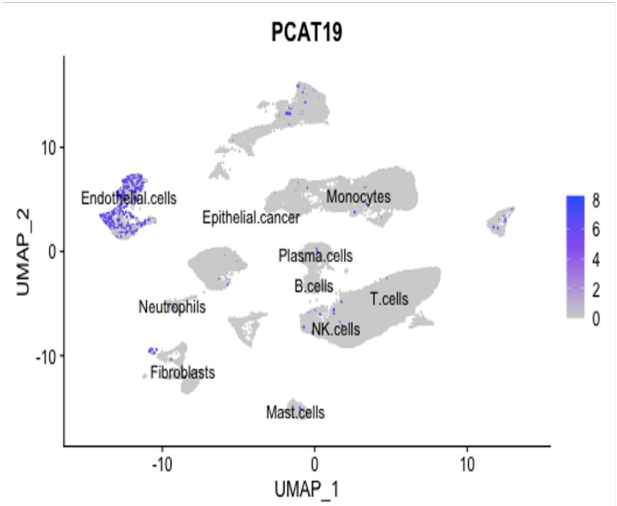
PCAT19: a strong marker of endothelial cells in pancreatic and prostate cancer

Pancreatic cancer





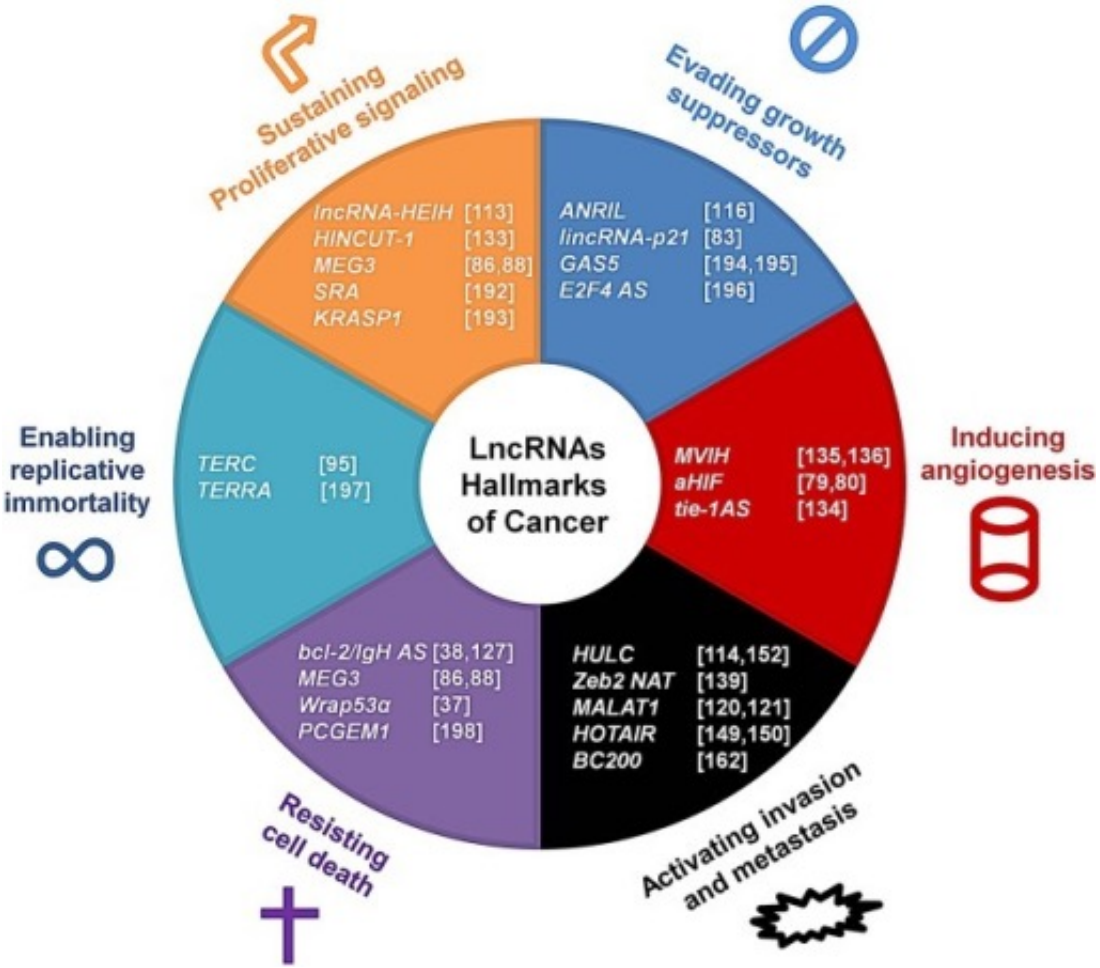
Prostate cancer



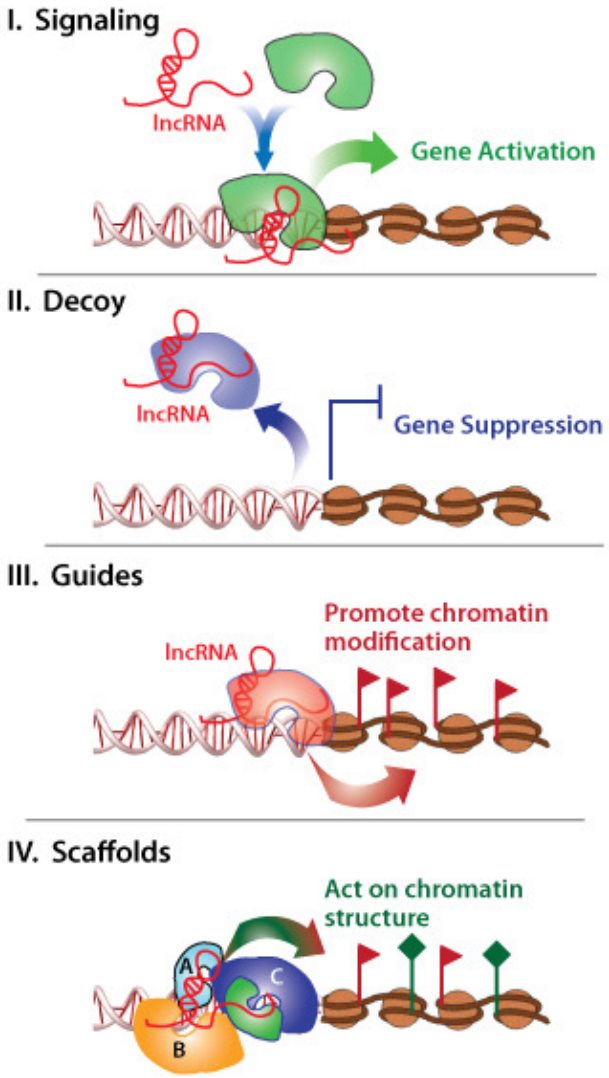
(*Accepted* NAR Cancer – Saha et al., 2023)

You've found an interesting lncRNA, now what?

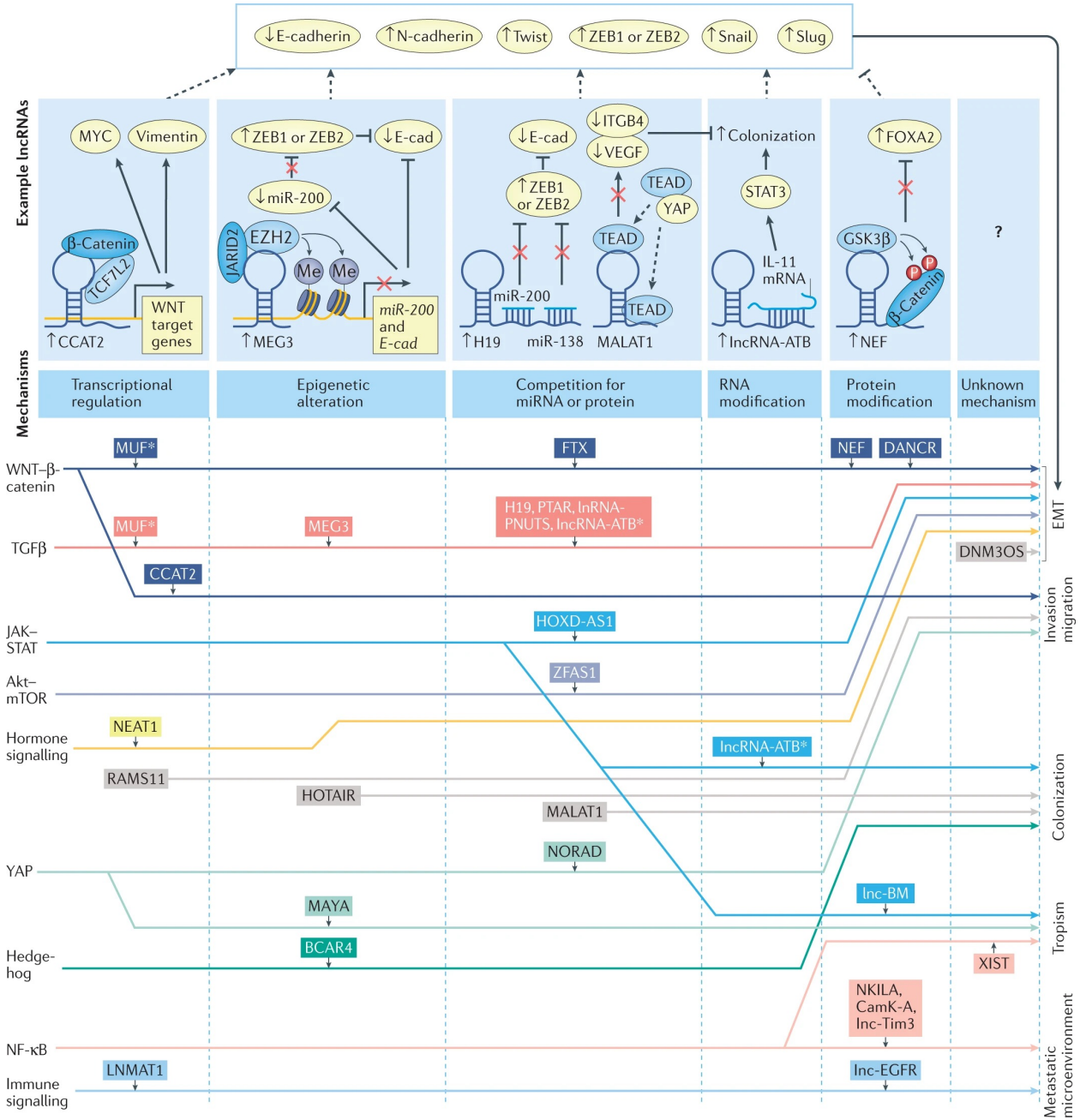
LncRNAs impact the hallmarks of cancer



Putative lncRNA regulatory mechanisms



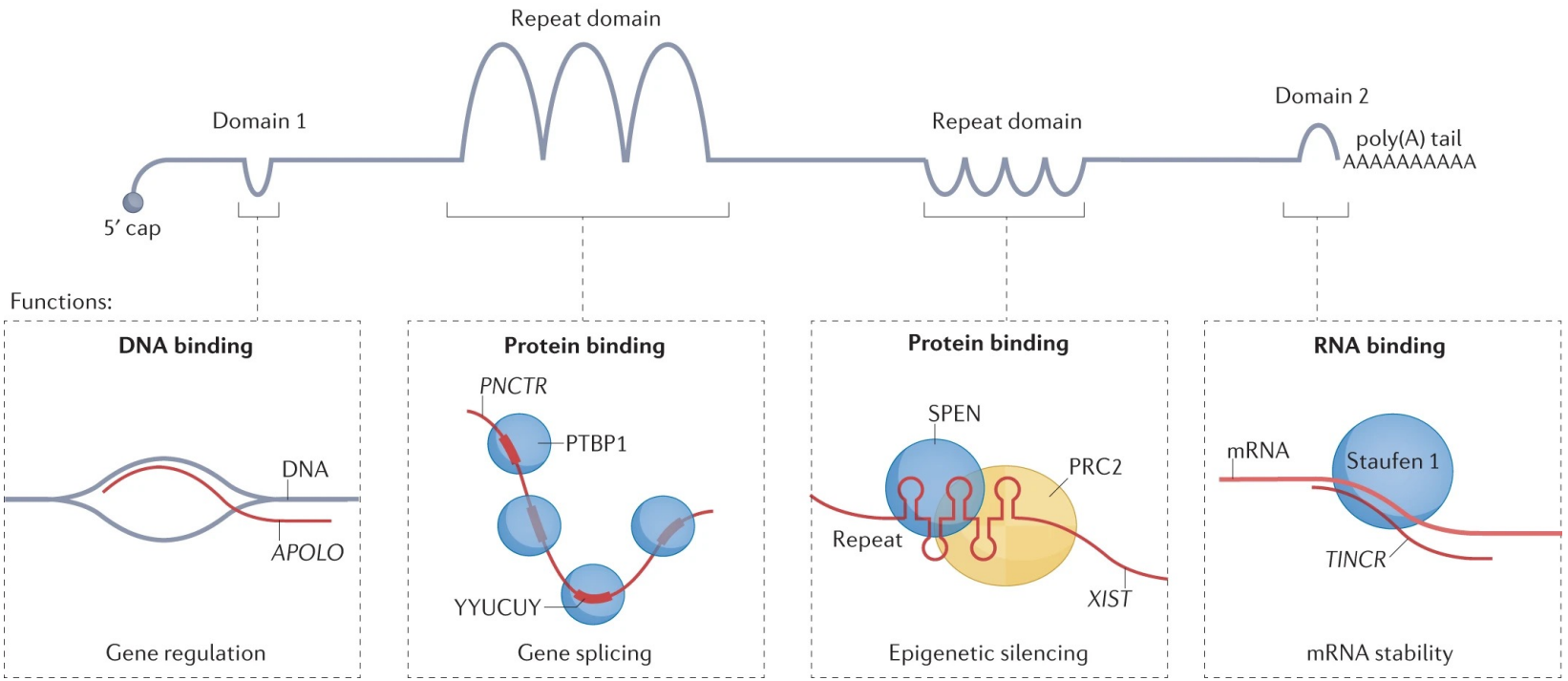
(Nature Cell Biology – Yao et al., 2019)



Long noncoding RNAs regulate metastasis via various pathways using diverse mechanisms

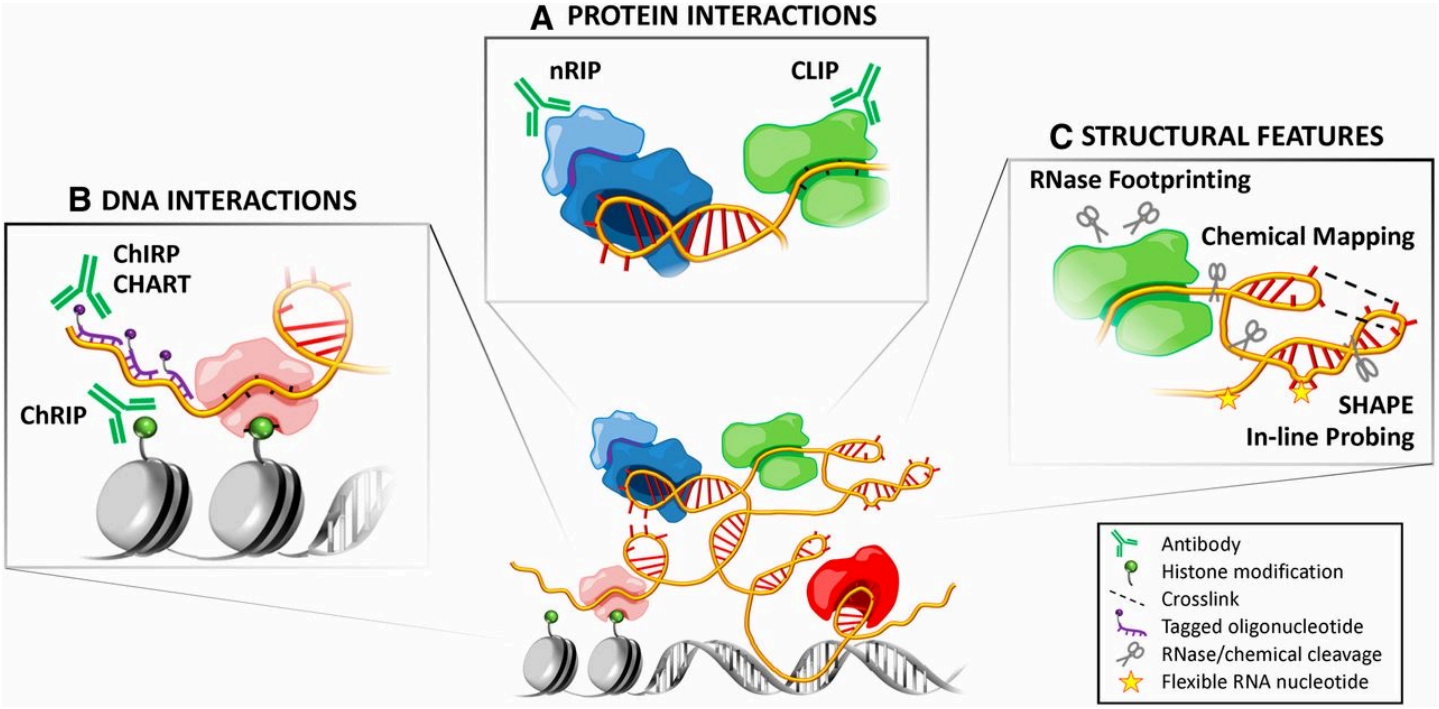
(Nature Reviews Cancer
Liu et al., 2021)

Domain architecture of lncRNAs



(Nature Reviews Molecular Cell Biology – Mattick et al., 2023)

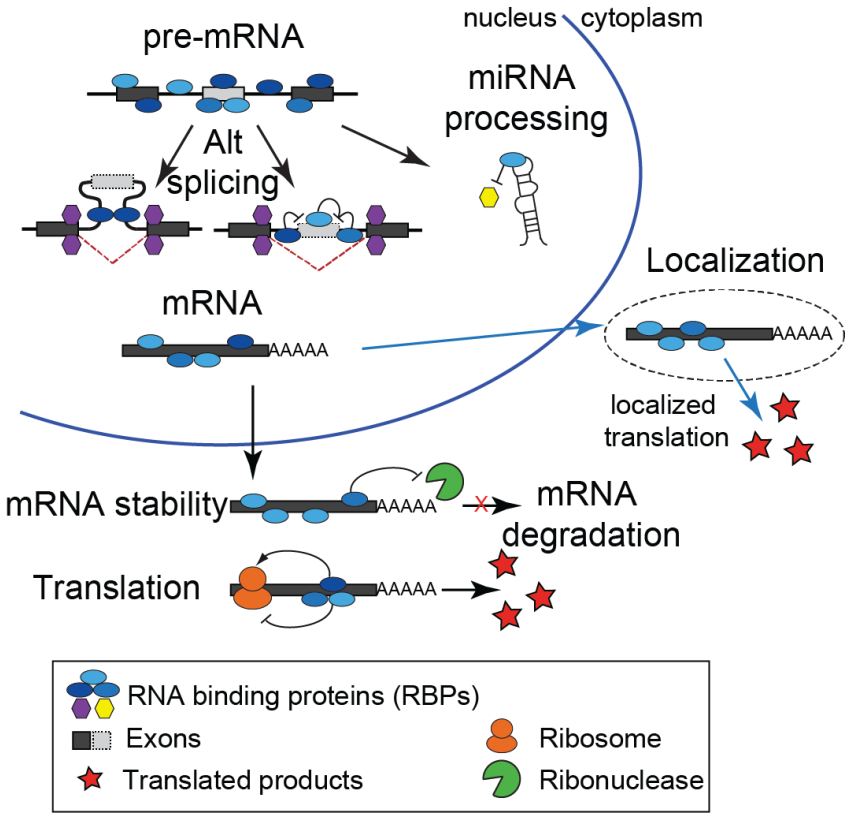
Our ability to understand how a lncRNA functions requires knowledge of its interacting partners: DNA, RNA, and/or protein



LncRNA-Centric Approaches

Methods start with a lncRNA of interest and characterize its interaction with DNA/RNA/proteins

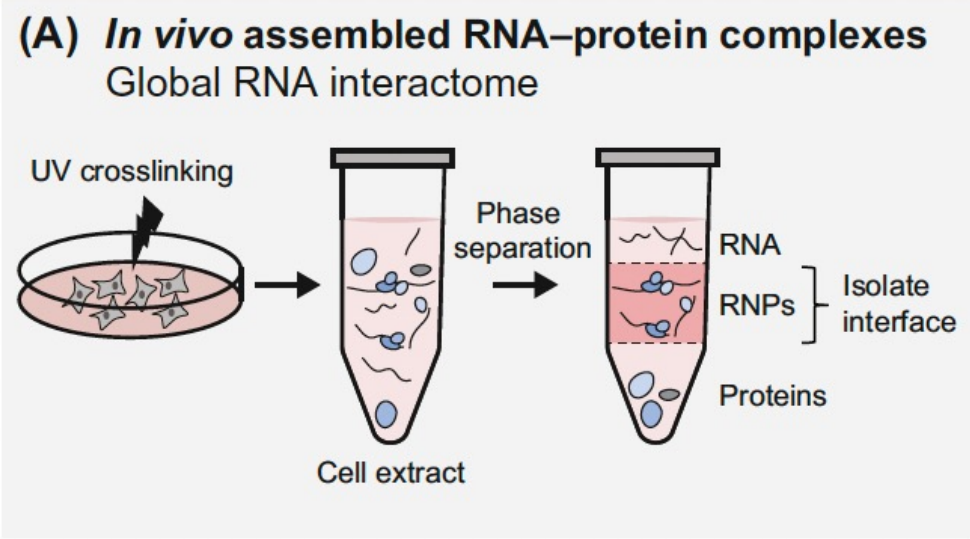
RNA Binding Proteins (RBPs)



- 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
- Recent efforts have identified novel RBPs with no annotated RNA-binding domains
- Mutation or alteration of RNA binding proteins plays critical roles in disease

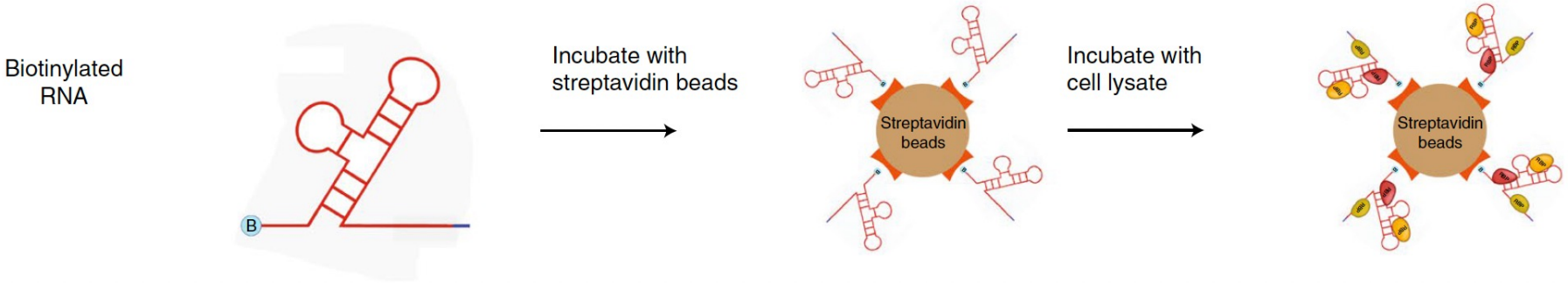
Identify the global RNA-bound proteome

Methods preserve cellular RNA–protein interactions by crosslinking, which is then followed by cell lysis and purification of the RNA–protein complexes



(Trends in Biotechnology – Grawe et al., 2021)

In vitro methods commonly used to identify proteins interacting with specific RNA

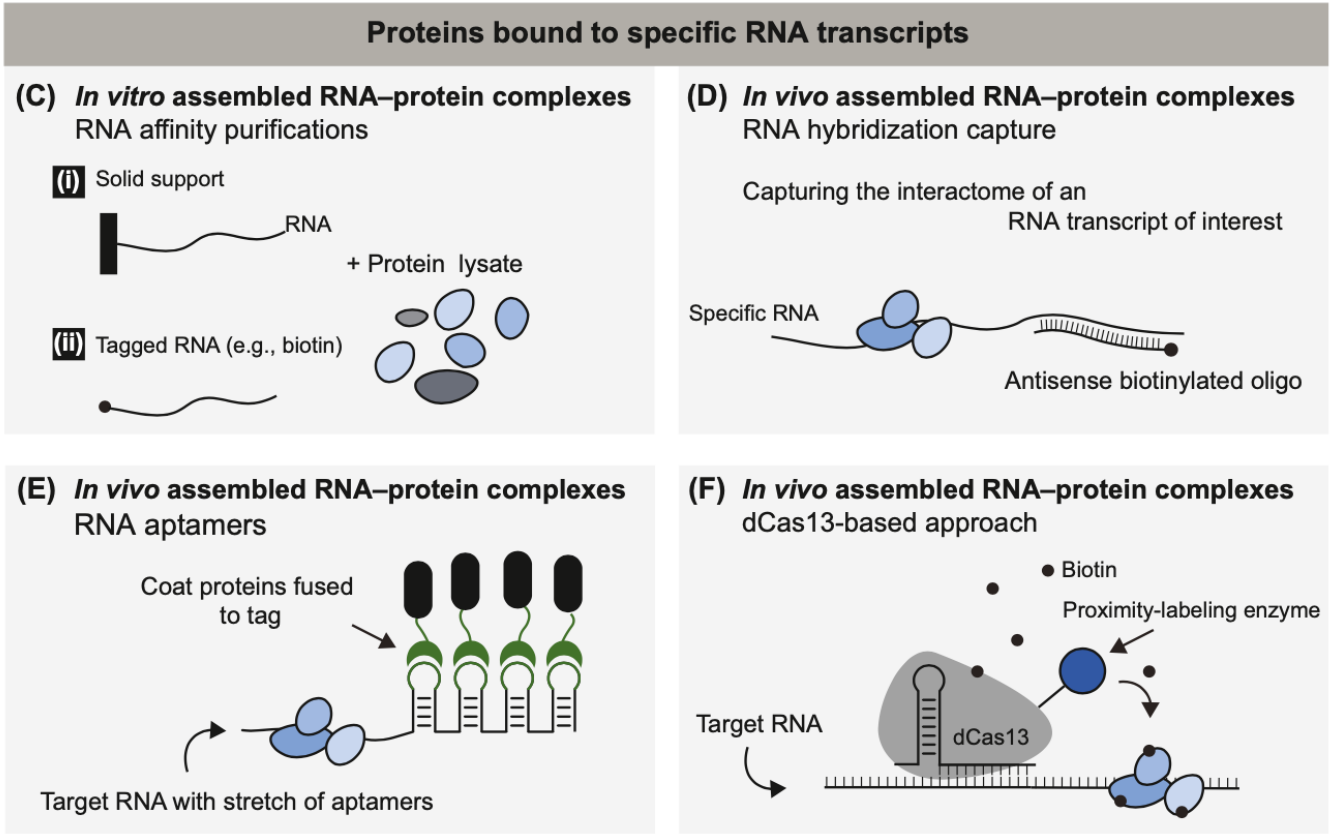


- RNA bait is immobilized on beads and incubated with proteins extracted from tissues or cells
- Unbound proteins are washed away and bound proteins are purified and identified by western blot or mass spectrometry

(Nature Methods – Ramanathan et al., 2019)

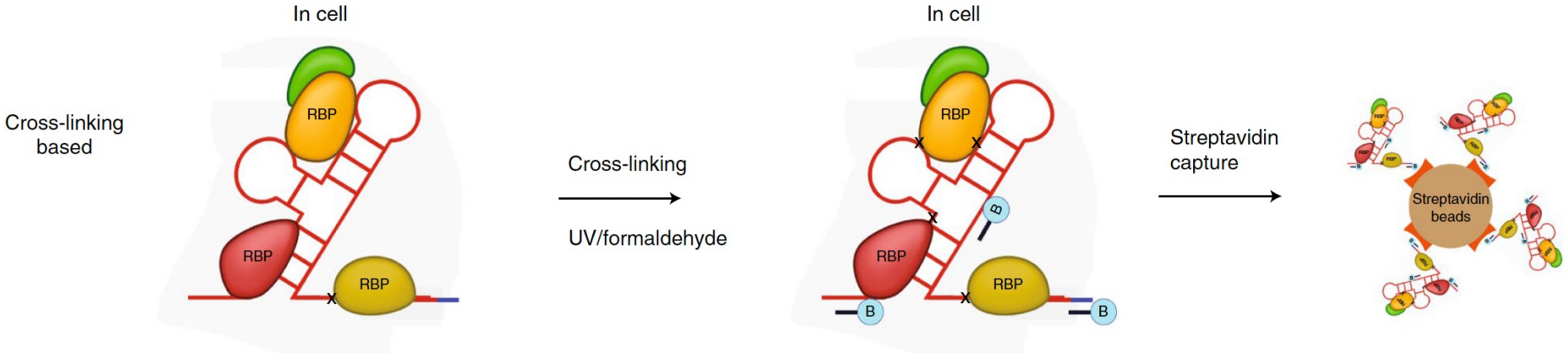
In vivo approaches to investigate RNA–protein interactions in the cellular environment

- These methods can be divided into three categories: RNA hybridization; RNA-tags; or CRISPR-based RNA targeting



(Trends in Biotechnology – Grawe et al., 2021)

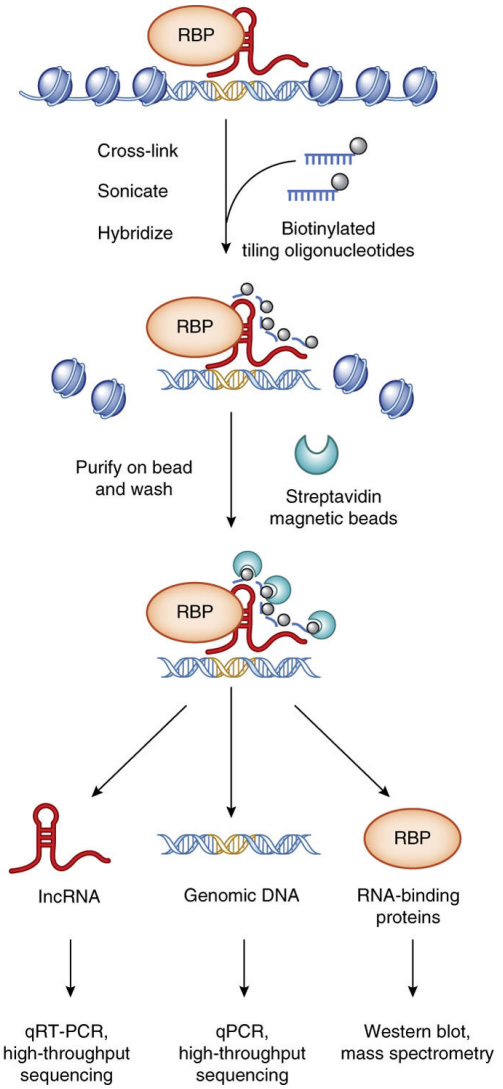
In vivo RNA hybridization approaches using cross-linking to investigate RNA–protein interactions



- Subdivided according to crosslinking method:
  - Ultraviolet (UV) crosslinking methods include:
    - RNA affinity purification (RAP), peptide-nucleic-acid-assisted identification of RBPs (PAIR), MS2 in vivo biotin-tagged RAP (MS2-BioTRAP), and tandem RNA isolation procedure (TRIP)
  - Formaldehyde crosslinking methods include:
    - Chromatin isolation by RNA purification (ChIRP)
    - Capture hybridization analysis of RNA targets (CHART)

(Nature Methods – Ramanathan et al., 2019)

Elucidating lncRNA interactions with proteins, RNA, and DNA

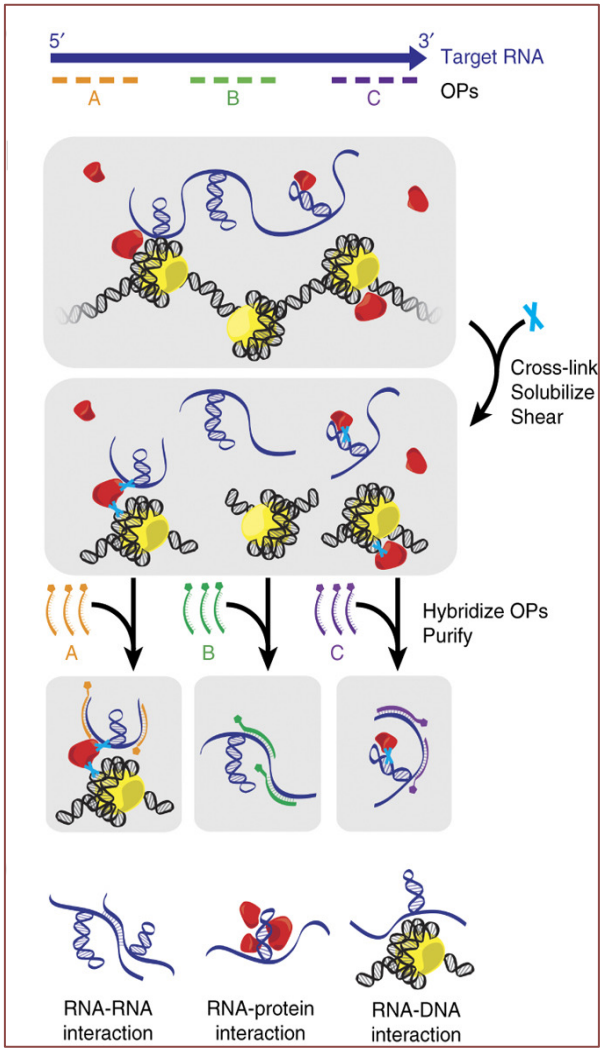


- **ChIRP (chromatin isolation by RNA purification)**
  - RNA–protein–DNA complexes are fixed by cross-linked *in vivo* and solubilized by sonication
  - Capture RNA of interest using specific biotin tagged antisense DNA oligonucleotides
  - Oligonucleotide-bound RNA and associated complexes are efficiently 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

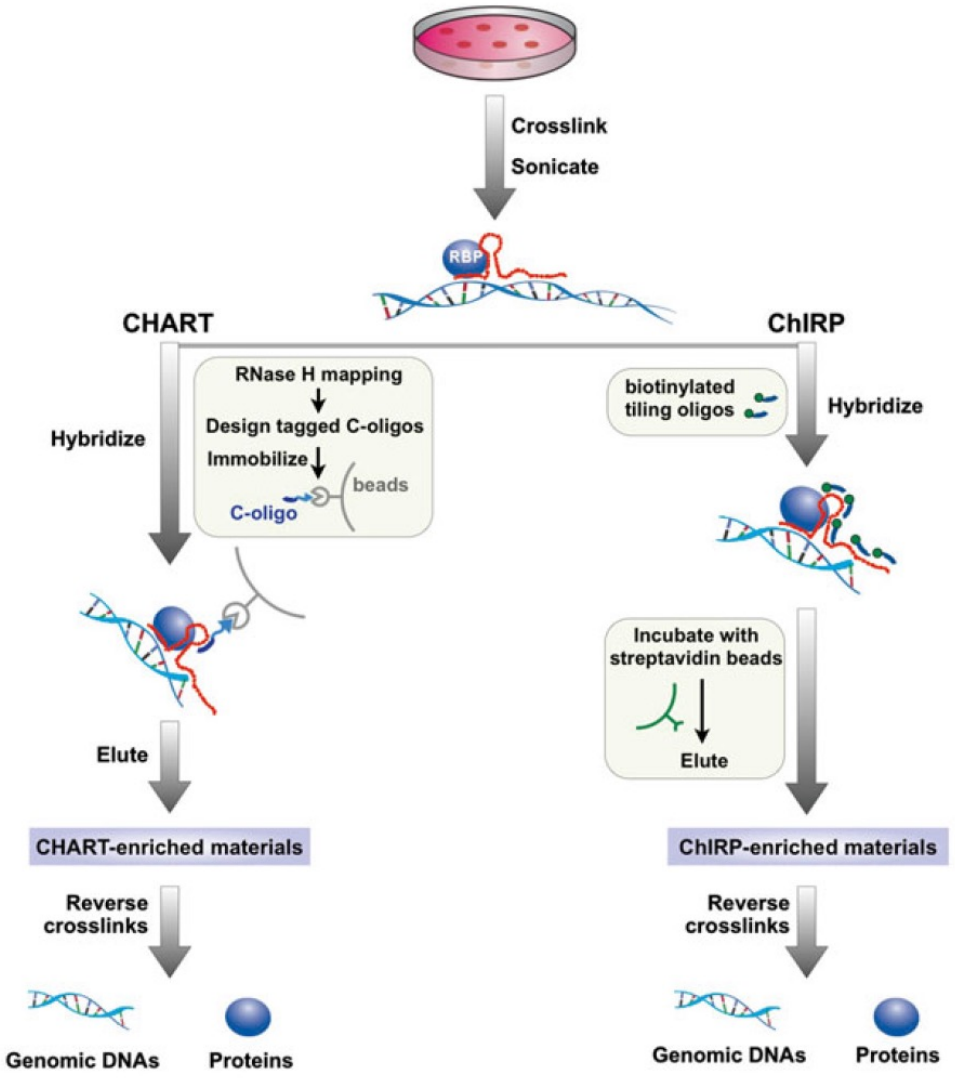
Domain Specific ChIRP (dChIRP)



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

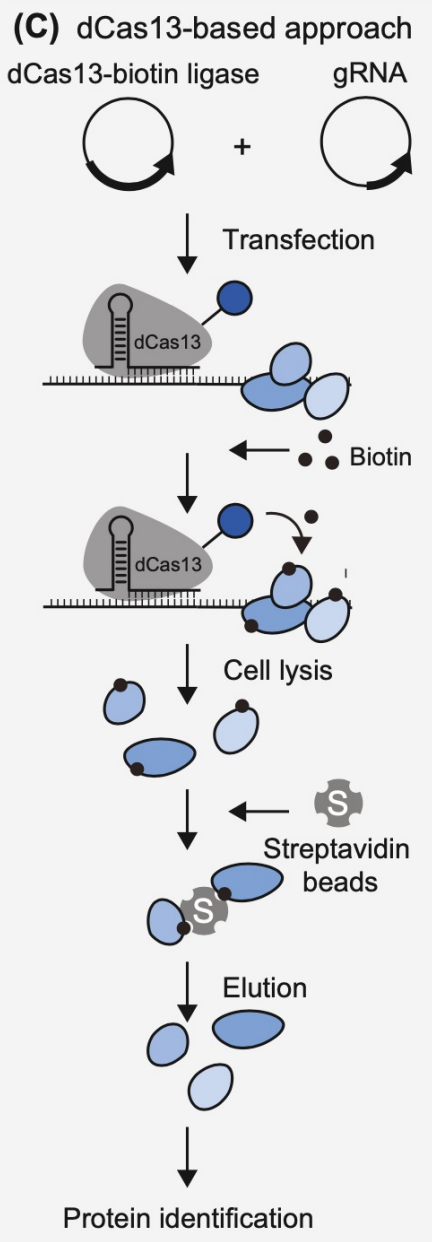
Additional methods to interrogate lncRNA interactions

- CHART utilizes an additional RNase H assay to identify accessible sites for probes



CRISPR-Cas13 RNA targeting system recruits proximity-labeling enzymes to specific RNAs

- Cells are transfected with dCas13 and guide RNAs (gRNAs)
- gRNA recruits dCas13 fused to a proximity-labeling enzyme (biotinylates accessible lysines or tyrosines of proteins within a 10–20-nm radius) to an RNA of interest
- Upon addition of biotin, the proximity-labeling enzyme biotinylates proteins in its proximity
- These proteins are isolated with streptavidin beads and identified by mass spectrometry.



(Trends in Biotechnology – Grawe et al., 2021)

Overview of RNA-centric methods

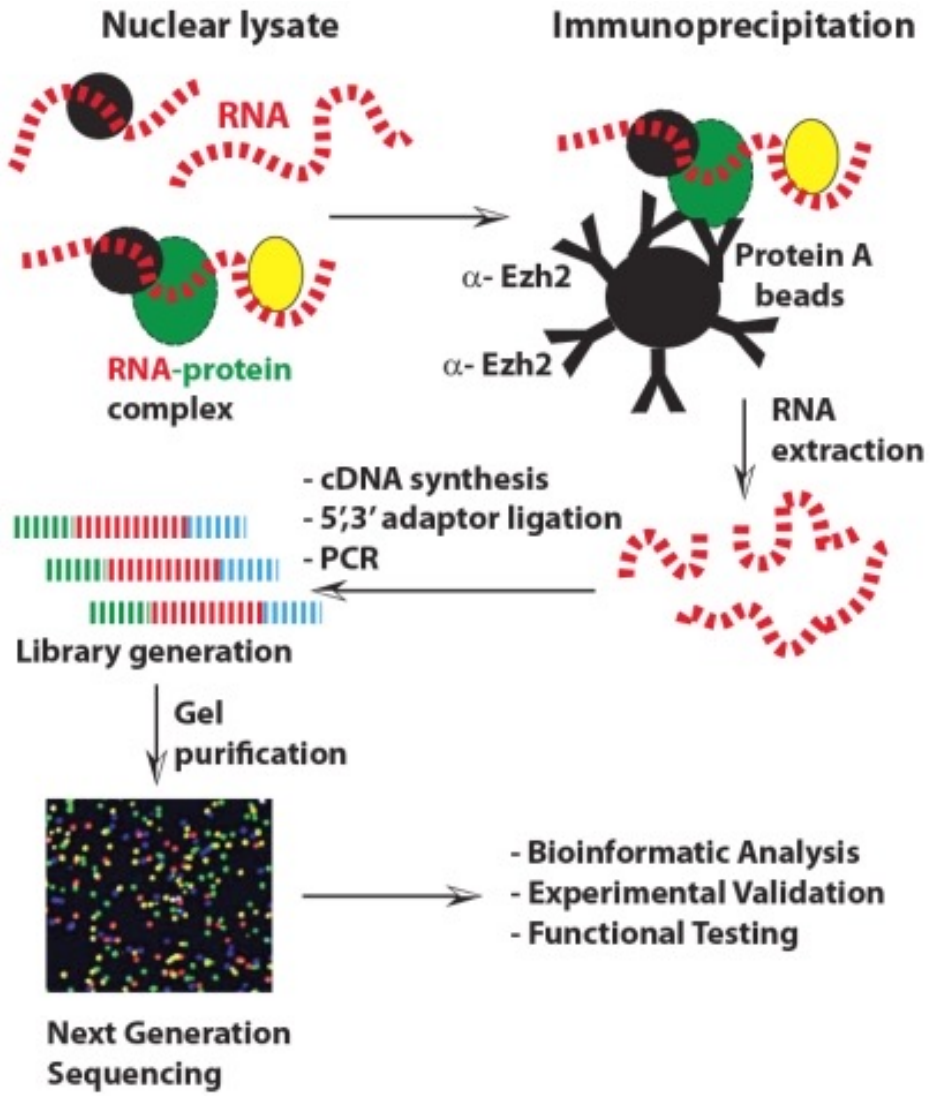
Method	Application	Advantages	Disadvantages
Biotinylated RNA	*SMN* mRNA	Strong binding of biotin-end-labeled RNA with streptavidin beads	In vitro; potentially biased toward abundant proteins in cell extracts
S1 aptamer	ARE motif	Easy elution of RBP complex from streptavidin beads with biotin	In vitro; potentially biased toward abundant proteins
Cys4	Pre-miRNA	Elution of RBP complex with imidazole	In vitro; potentially biased toward abundant proteins
Protein microarray	TINCR, SNORD50	No cellular extract required; no MS required	In vitro; limited to direct interactions with proteins spotted on microarray
RAP	Xist, FIRRE noncoding RNA	In vivo; high specificity with UV cross-linking and long oligonucleotide probes (120 nt)	High input cell numbers
TRIP	*p27* mRNA, *CEP-1* mRNA	In vivo; high specificity with UV cross-linking	Two capture steps with poly(A) and biotinylated ASO capture decrease efficiency
PAIR	*ANK* mRNA	In vivo; high specificity with UV cross-linking	Cost and effort for production of peptide nucleic acid
MS2-BioTRAP	IRES	In vivo; high specificity with UV cross-linking	Requires MS2 conjugation to RNA, transfection/infection of RNA and labeler protein, and high input cell numbers
CHART	Xist, MALAT1, NEAT1	In vivo	Additional RNase H step to identify accessible sites for probes; high input cell numbers
ChIRP	TERC, Xist	In vivo; no prior knowledge of RNA accessibility required for probe design	Short probes may pull down similar sequence fragments; high input cell numbers
RaPID	ZIKV-host protein interactome; 3′ untranslated region motifs	In vivo; low number of cells required; direct labeling of protein	Requires BoxB link to RNA; short sequence limits; transfection/infection of RNA and labeler protein

(Nature Methods – Ramanathan et al., 2019)

Protein-Centric Approaches

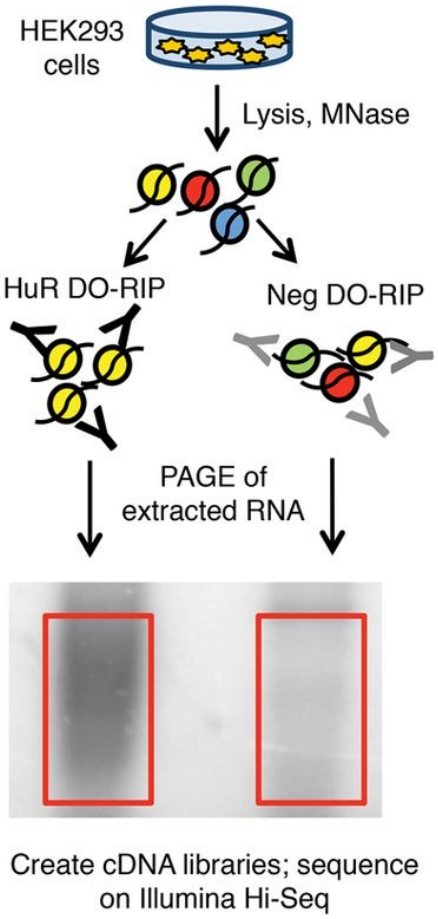
Methods start with a protein of interest and characterize its interaction with RNA

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

DO-RIP-Seq Overview

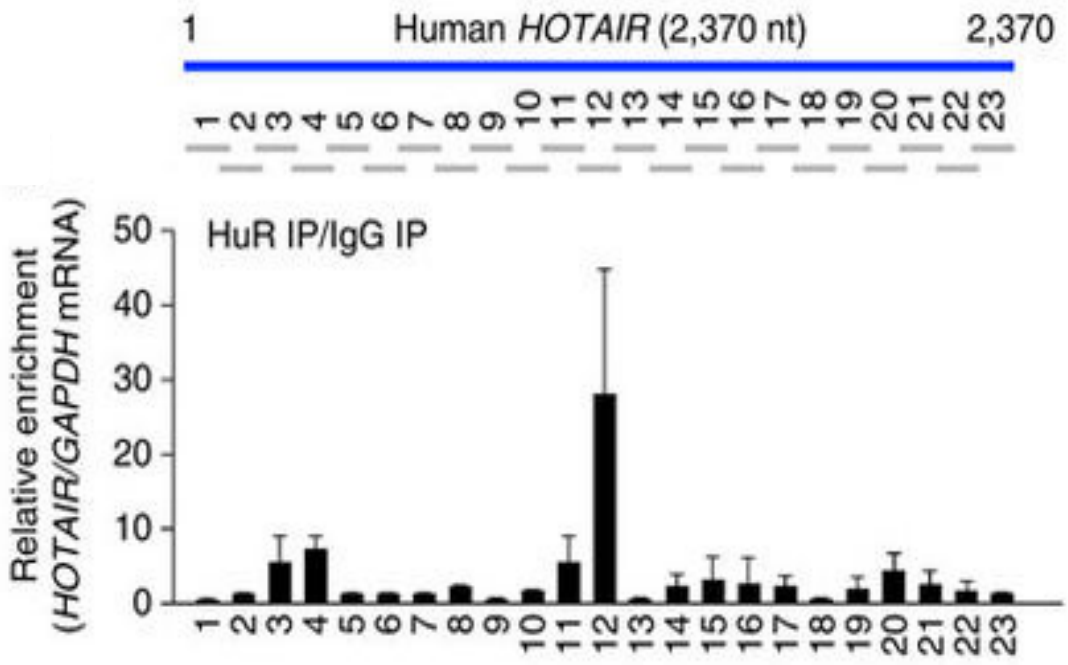


(RNA -- Nicholson et al. 2017)

- 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

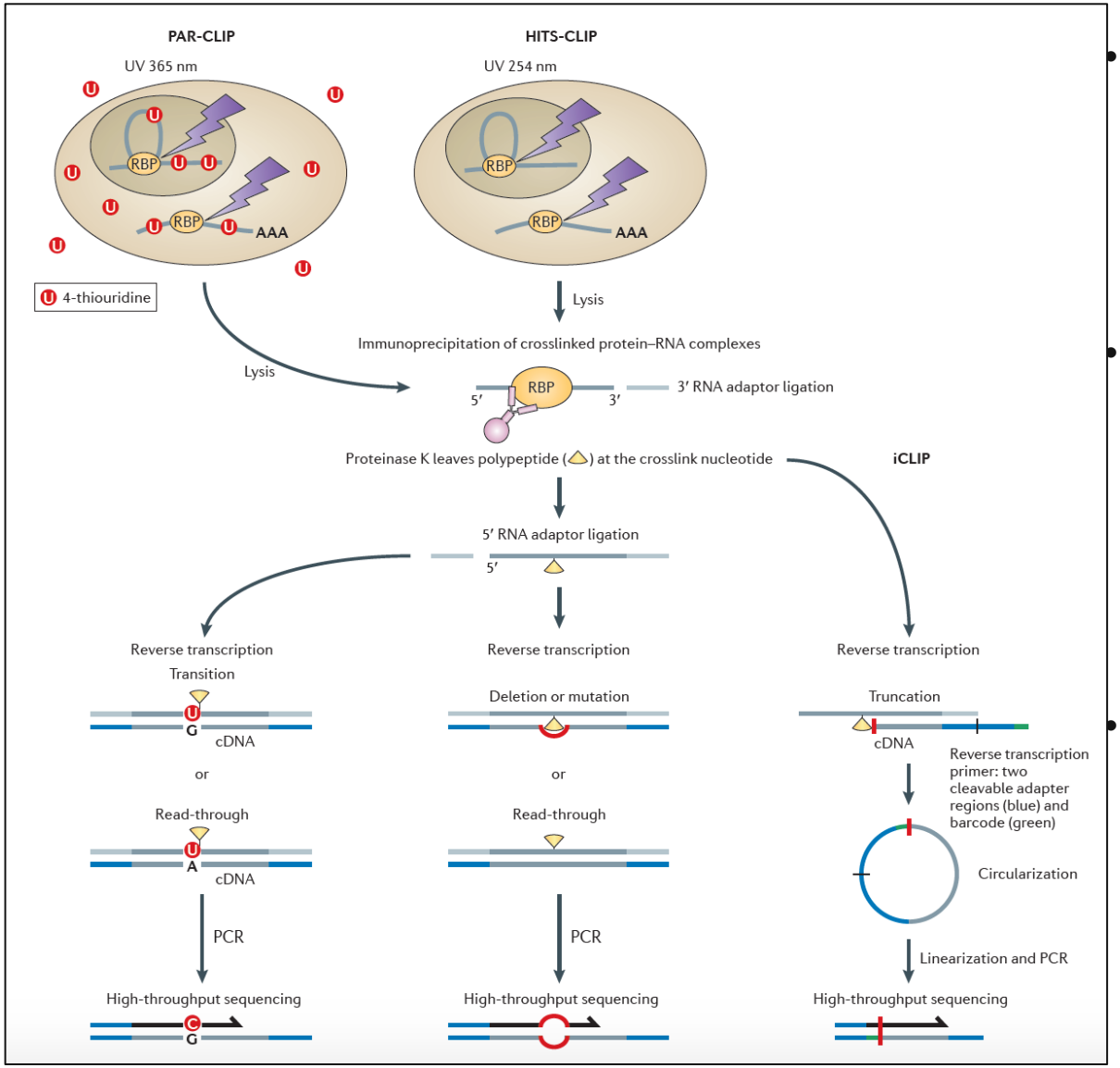
Locating specific interaction sites

- RNAse protection assay can help localize the potential interaction site



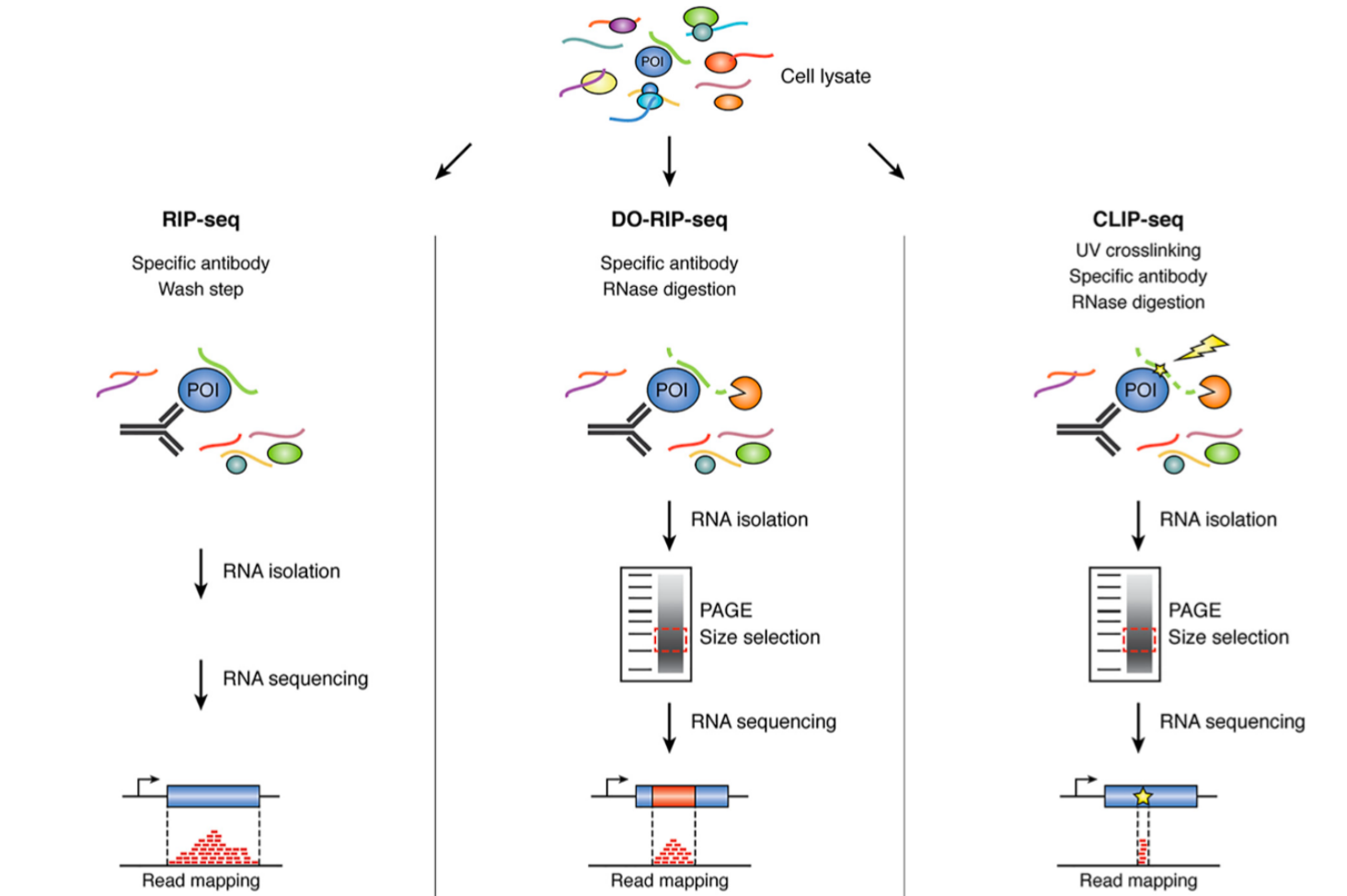
(Yoon et al., 2013)

Common variations of crosslinking immunoprecipitation (CLIP)



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

- **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



RIP-seq

Pros:
- Preserves native-binding conditions
- Well-suited to RBPs with low transcript affinity

Cons:
- Wash conditions are difficult to optimize
- Does not identify binding site within transcript

DO-RIP-seq

Pros:
- Combines advantages of both RIP-seq and CLIP-seq protocols
- Identifies binding region of RBP within transcript

Cons:
- Rarely used technique

CLIP-seq

Pros:
- Allows stringent wash buffers
- Individual nucleotide resolution of binding site

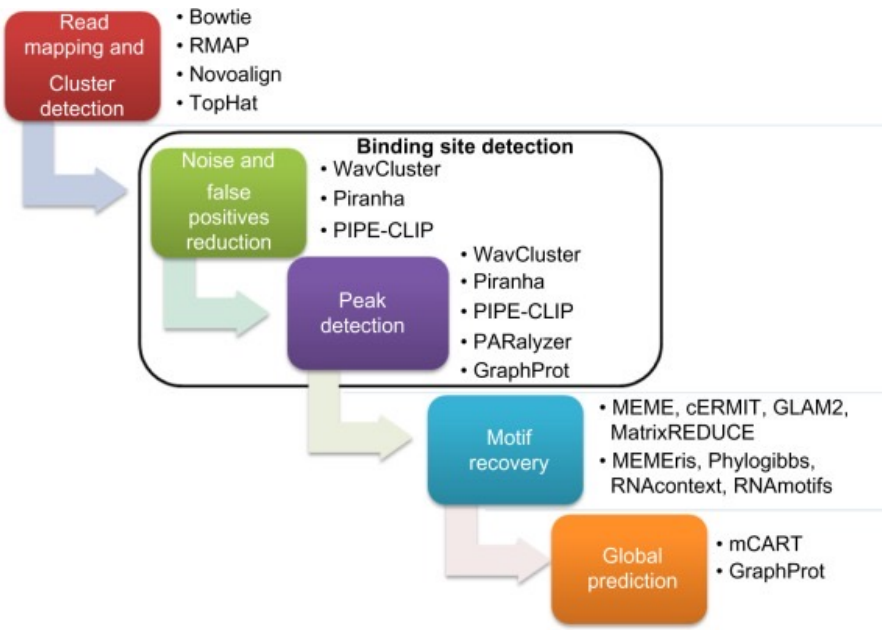
Cons:
- Stabilizes non-specific interactions
- Prone to certain biases

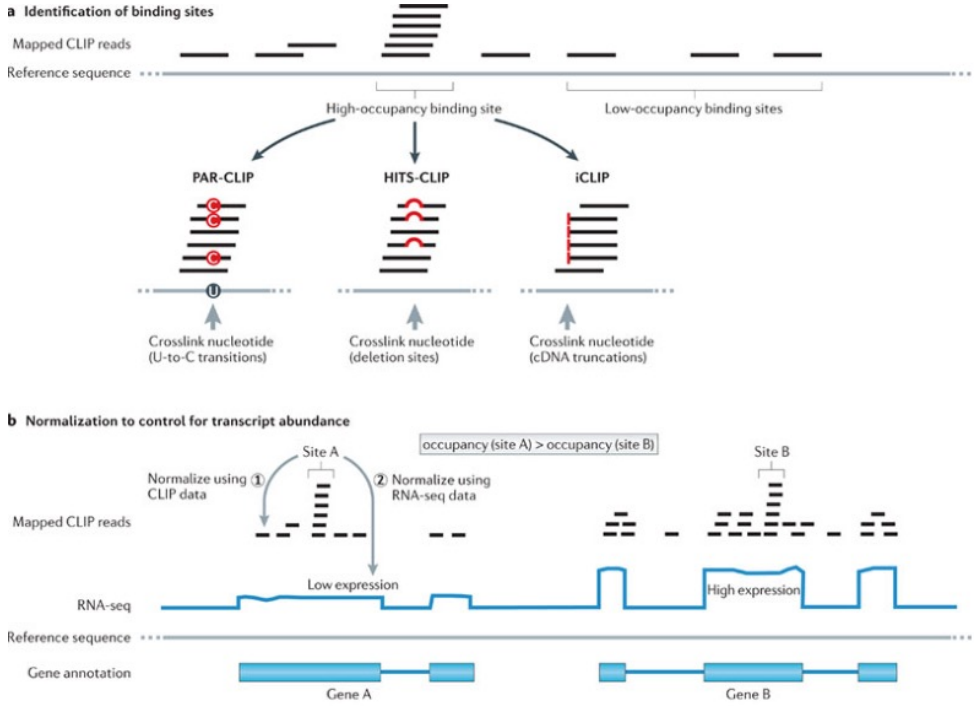
(Moore et al., 2018)

Tools used to analyze CLIP-Seq data

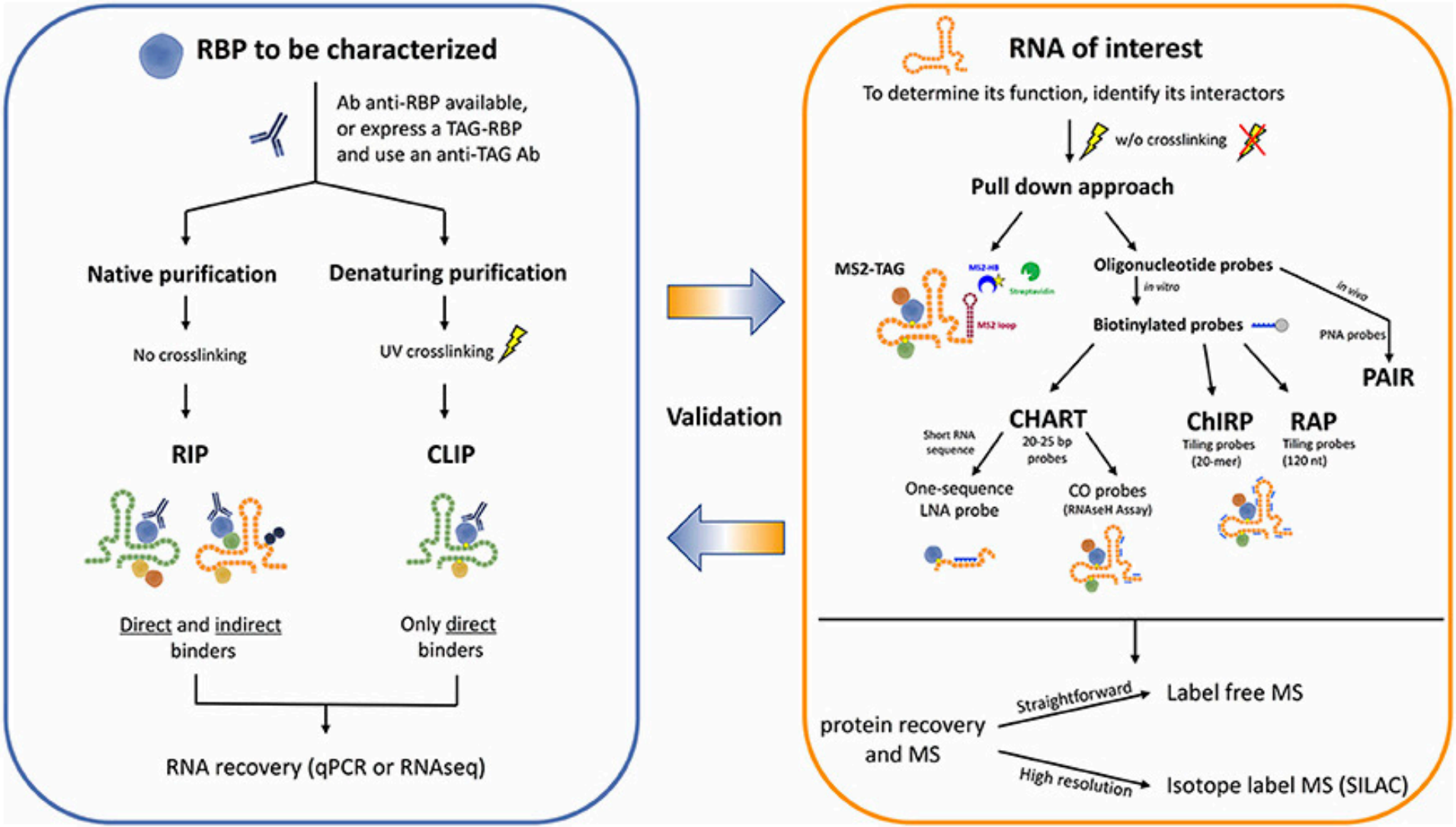
Study	Approach
Nova	CLIP data are evaluated on the basis of enrichment relative to a control immunopurification.
Ago	Calculates the expected number of reads per RNA from RNA abundance and randomizes their positions to determine the odds of a peak as high as that observed.
eCLIP	CLIP data are evaluated on the basis of enrichment relative to RNA from the input lysate run on an SDS–PAGE gel and transferred to a nitrocellulose membrane.
FOX2	Reads are scrambled randomly in transcript to calculate the odds of the observed peak height. This differs from the method used for Ago above[94] in that the reads are drawn from the CLIP data, not RNA abundance. As a result, this approach identifies a locational bias within a given RNA, not affinity for the RNA molecule of interest, although the latter is indirectly tested by the requirement for having enough reads to identify a locational bias.
Puf2p	Raw peak height cutoff, enrichment relative to local CLIP signal, and enrichment over expectation from RNA abundance. This method is a combination of a local signal pileup analysis like that of FOX2[95], enrichment relative to RNA abundance like that of Ago[94], and a third raw peak height requirement.
FBF	Compares each of the different approaches available at the time of its development: local pileup analysis (similar to FOX2 above[95]), enrichment over RNA abundance (similar to Ago[94]), enrichment over a control immunopurification (similar to Nova[62]), CIMS[96], and CITS[80]. Enrichment over a control immunopurification was found to most accurately reflect known biology.
Nova, Ago	Frequency of cross-link-induced mutation pileup versus random permutations. Termed CIMS. The notes on FOX2 apply to this approach as well.
hnRNPC, TIA1, Rbfox	Frequency of reverse-transcription termination positions versus random permutations. Termed CITS[80]. The notes on FOX2 apply to this approach as well.
Many	Frequency of PAR-CLIP-specific mutations. We refer the reader to a more specialized review of CLIP-seq analysis tools[100].

CLIP-Seq data analysis workflow



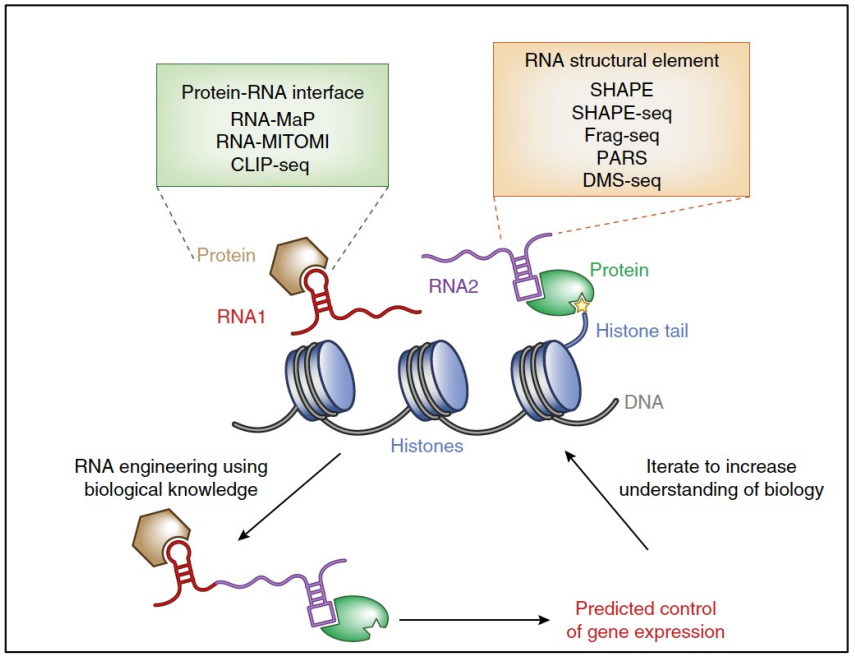


Integration of various strategies



LncRNA summary

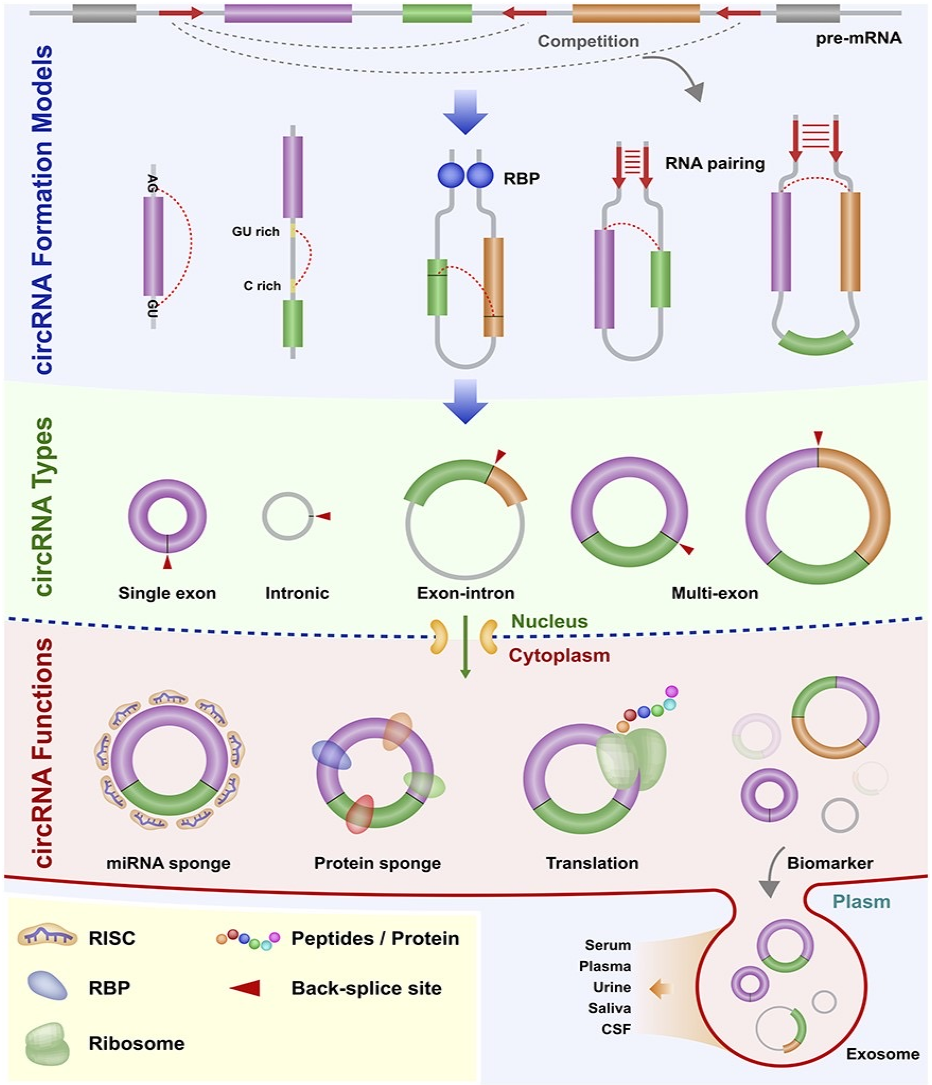
- LncRNAs are an abundant class of biologically and clinically relevant class of genes with a broad range of functionality
- Despite the rapid emergence of lncRNAs, the methods to interrogating their regulatory mechanisms are still evolving
- Ongoing development is still necessary to fully understand 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

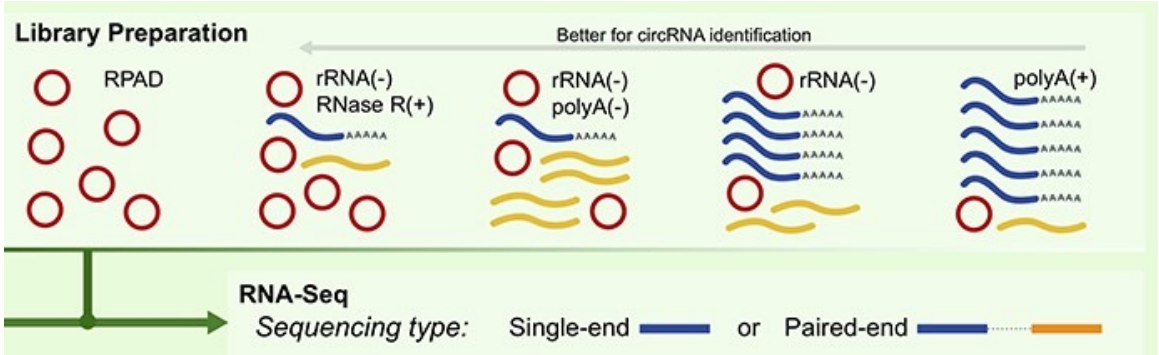
Circular RNA (circRNA) biogenesis



(Chen et al., 2021)

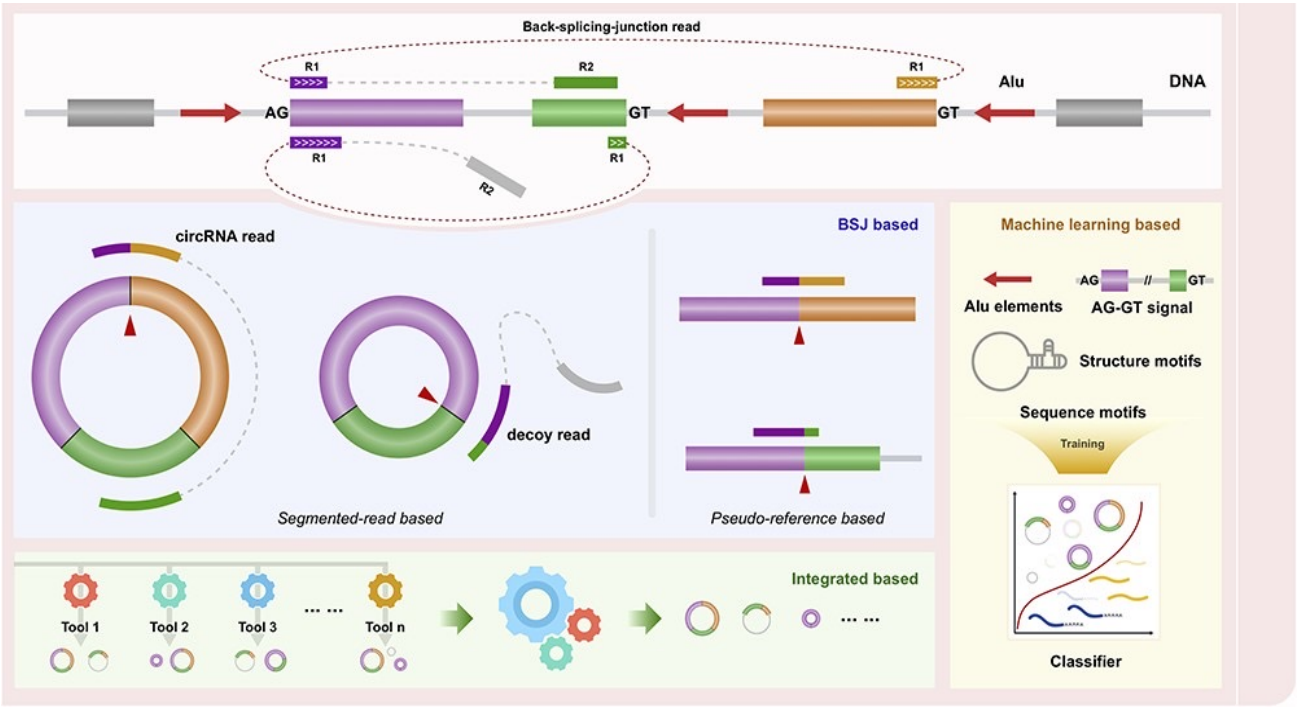
- Single-stranded circular molecules averaging 500 nucleotides in length
- Covalent 3'-5' bond formed in the process of backsplicing; a downstream splice donor on a pre-mRNA pairs with an unspliced upstream splice acceptor
- Exonic circRNAs make up the majority of annotated circRNAs
- Lack of open 3' end makes circRNAs resistant to exonucleolytic decay by the cellular exosome ribonuclease complex
- Higher stability and longer half-life of allow them to accumulate in exosomes and blood

Overview of existing circRNA resources and tools



(Chen et al., 2021)

circRNA identification tools can be divided into three categories: (1) Backsplice Junction (BSJ)-based, (2) integrated-based and (3) machine learning-based



BSJ-based:

- Most algorithms are based on splitting the reads (called segmented-read-based), while several other tools are based on a pre-defined BSJ and flanking sequence of a circRNA

Integrated tools:

- These ensemble tools merge the results of multiple stable tools

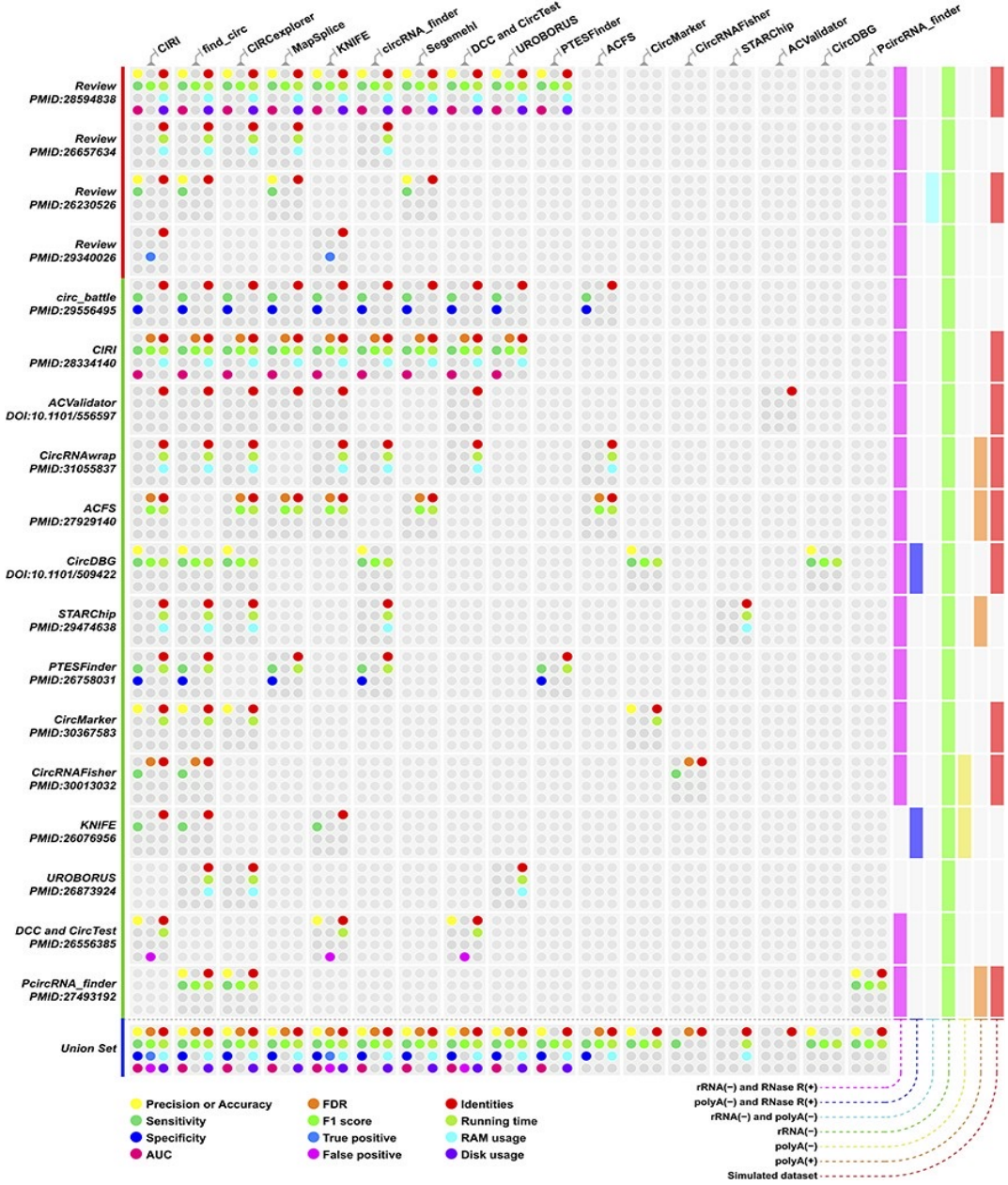
Machine learning:

- Most circRNA identification tools require RNA-Seq data as input, but machine learning techniques predict circRNAs using existing features (i.e., ALU repeats, structure motifs and sequence motifs) of known circRNA to train a classification model

(Chen et al., 2021)

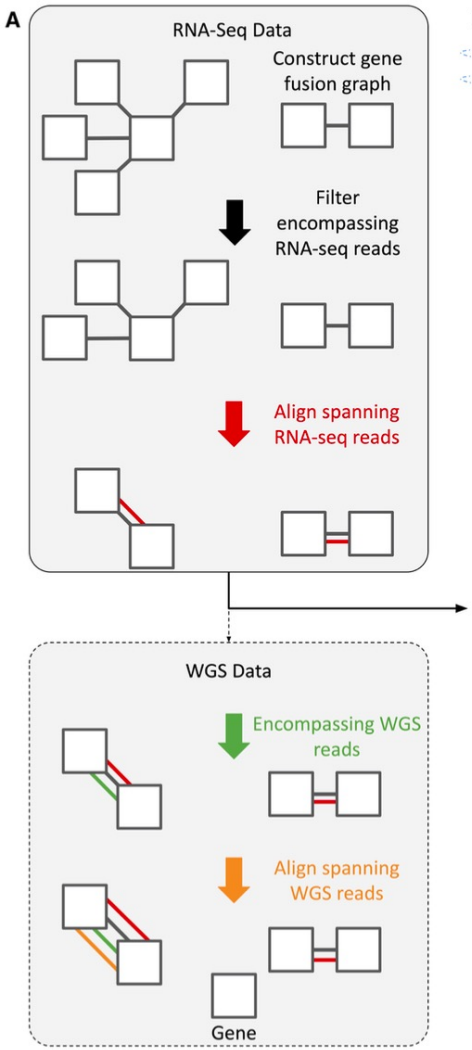
Evaluation of tools

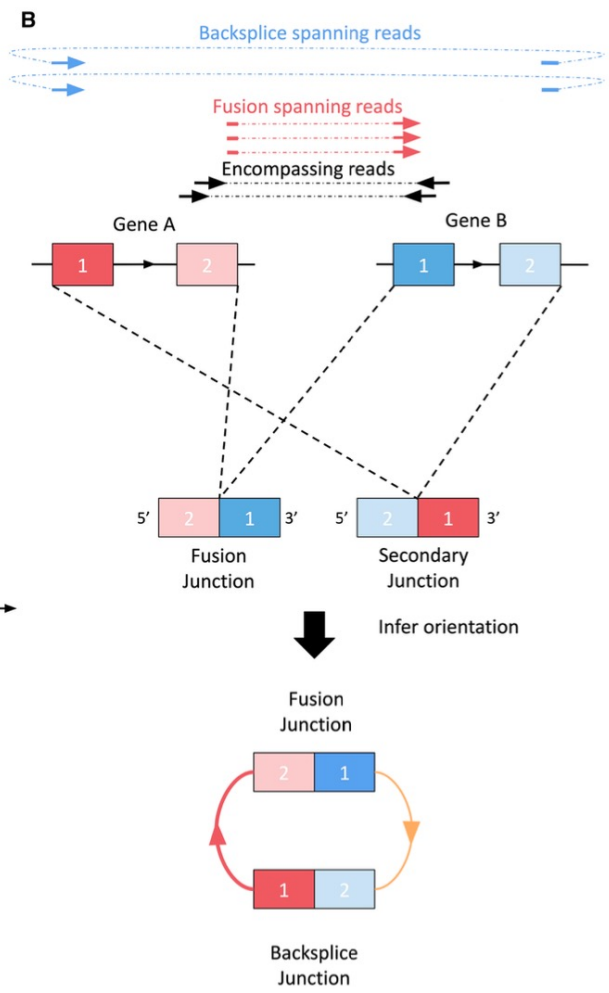
- Tested using different real and simulated datasets with variable criteria
- Extensive quantity of tools available (>40)
- Most existing tools are designed for short read sequencing

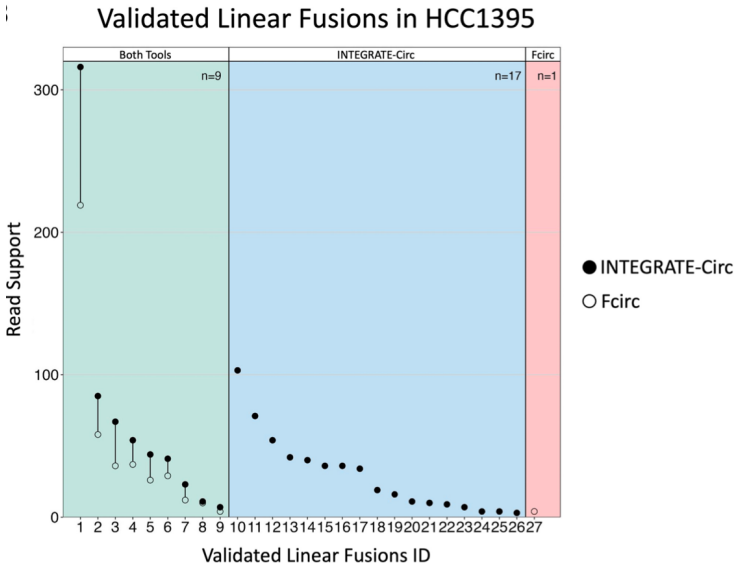


(Chen et al., 2021)

INTEGRATE-Circ and INTEGRATE-Vis: unbiased detection and visualization of fusion-derived circular RNA





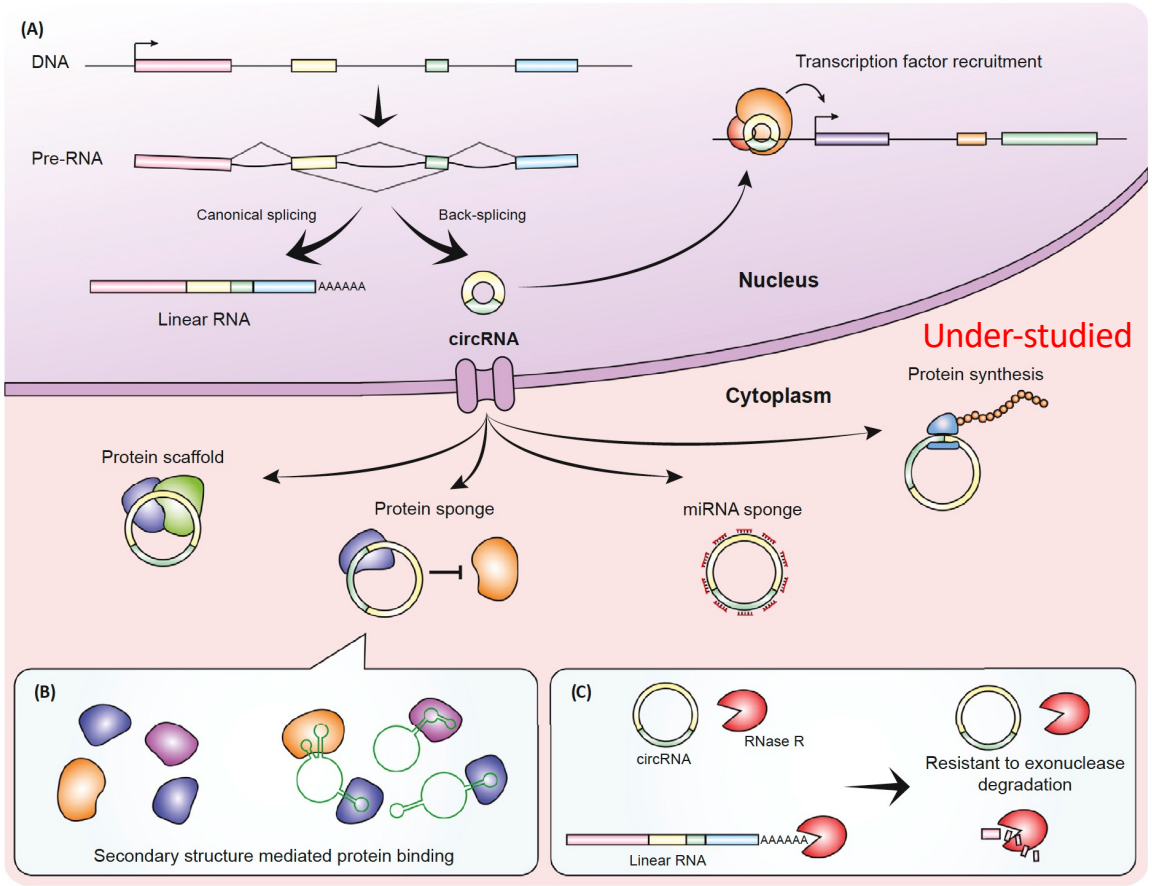


(Bioinformatics - Webster et al., 2023)

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

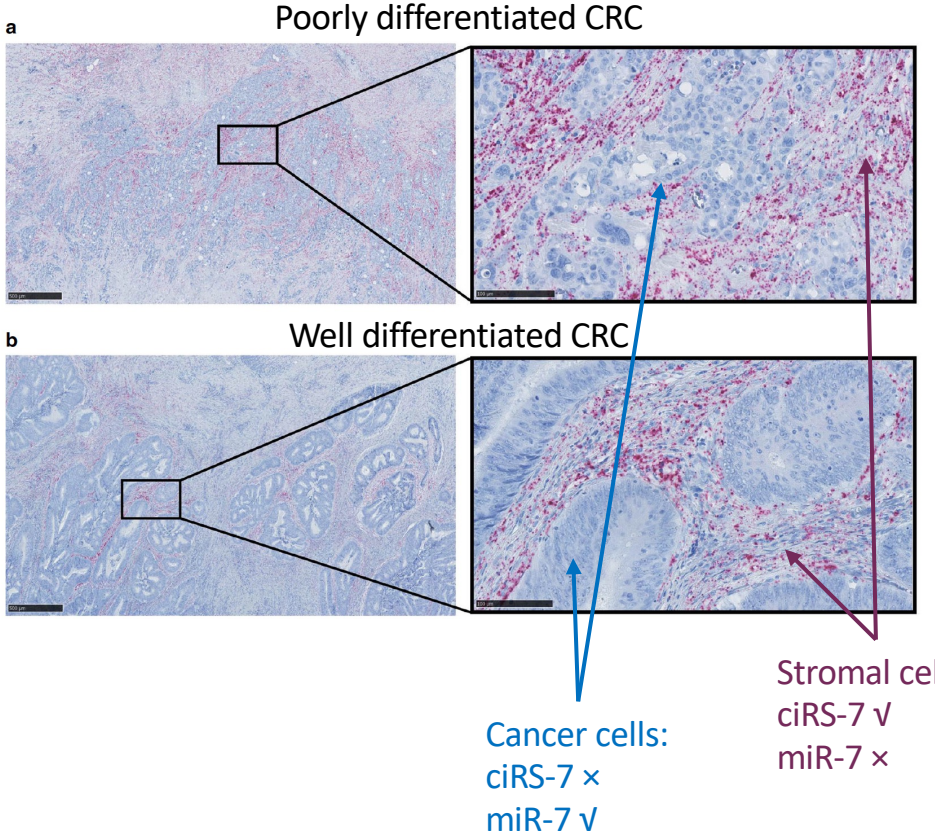
- Large scale studies (i.e., TCGA) mostly used poly-A selection
- No standard RNA quantification method
- Cell lines or limited patient cohorts
- Lack of genome-wide systematic analysis
- Existing databases lack inclusion of CRC (and particular matched patients throughout progression)
  - MiOncoCirc,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

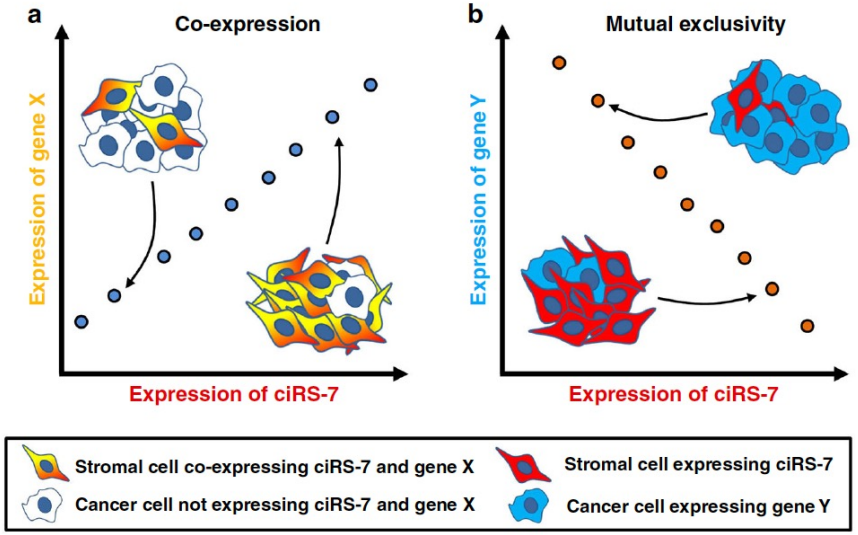
Putative functions of circRNAs remain under-studied in cancer



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

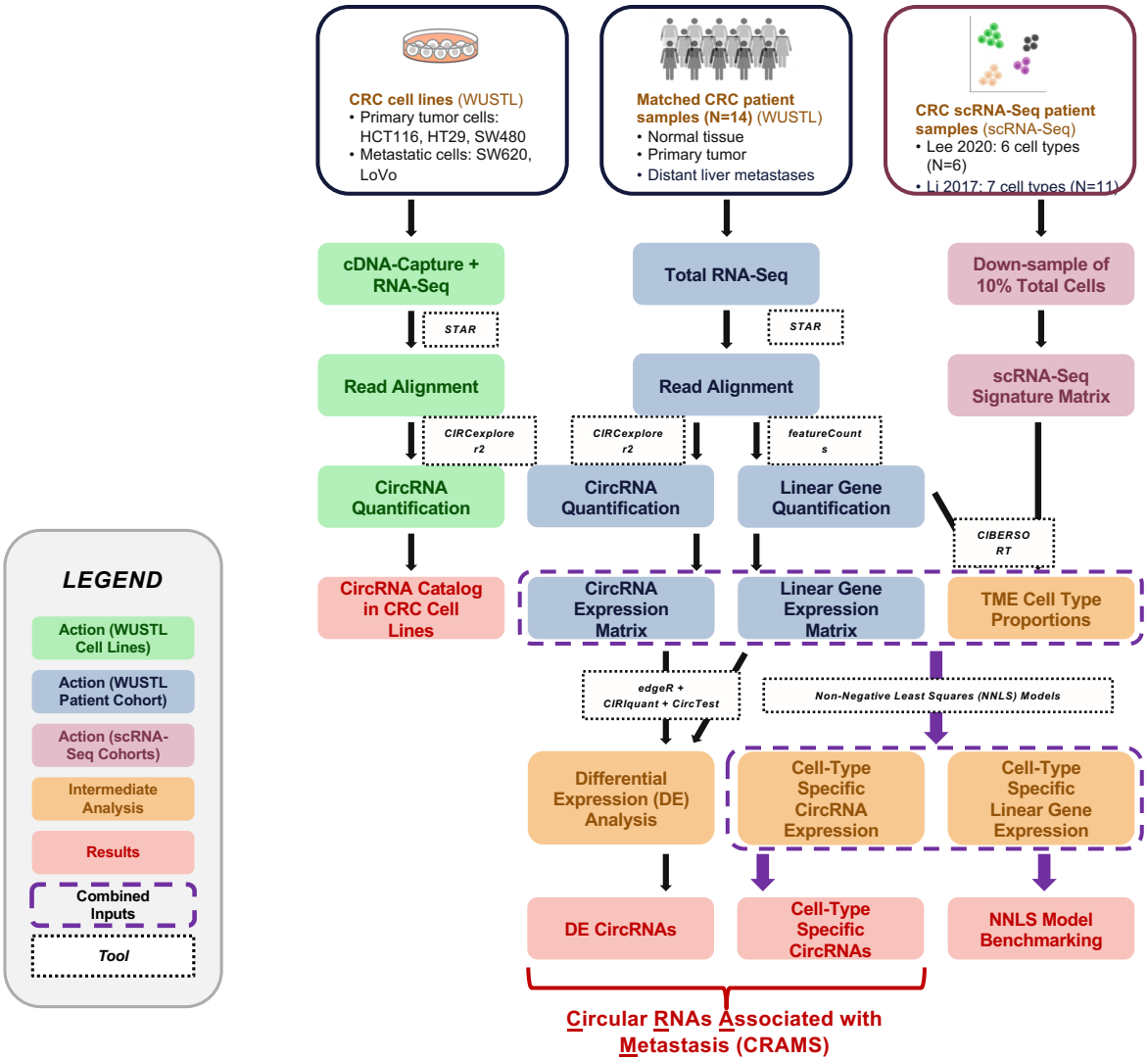
Challenge to the miRNA sponge mechanism





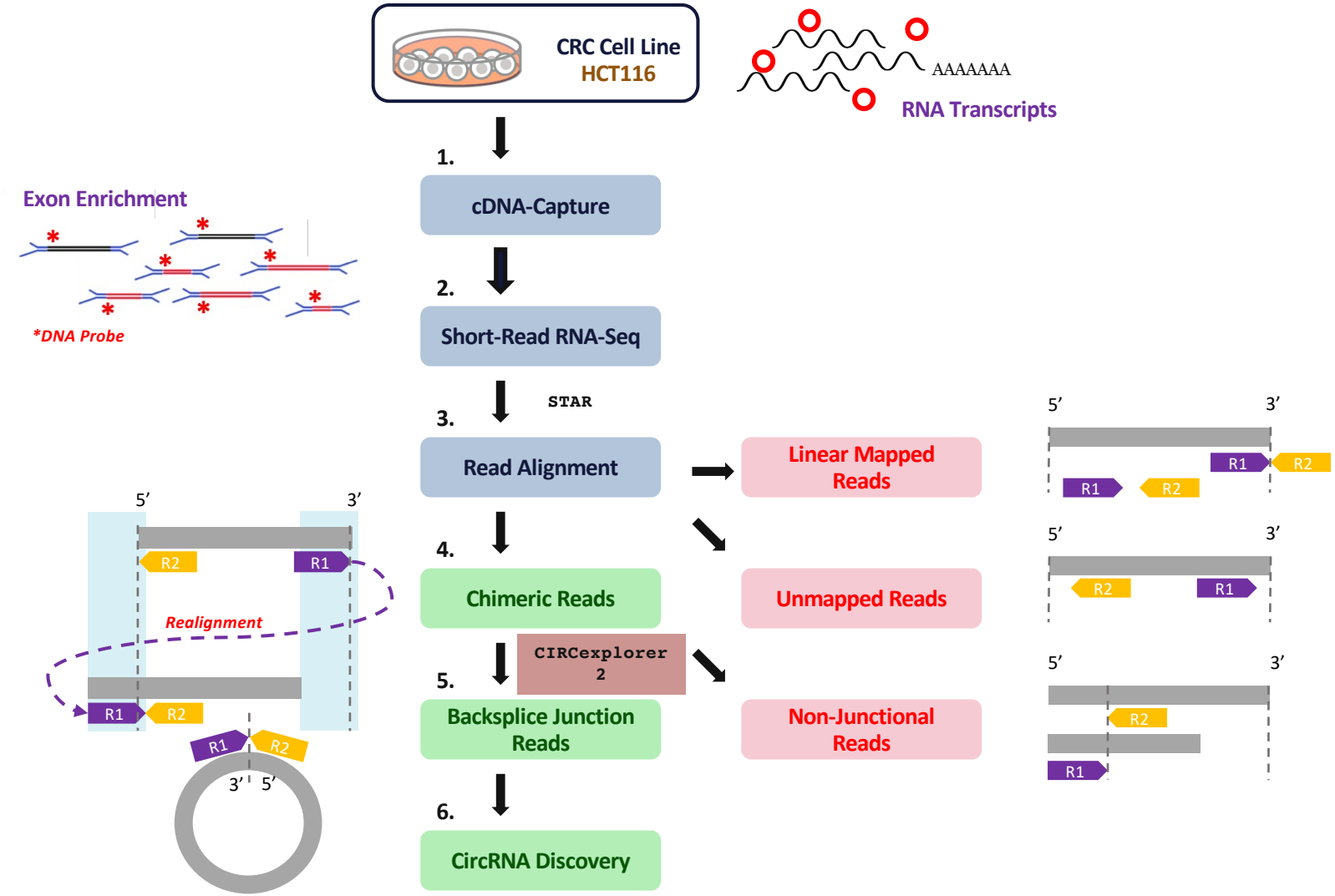
(Kristensen, L.S., et al., 2020)

Comprehensive workflow for characterization of cell-type specific and altered circRNAs in mCRC



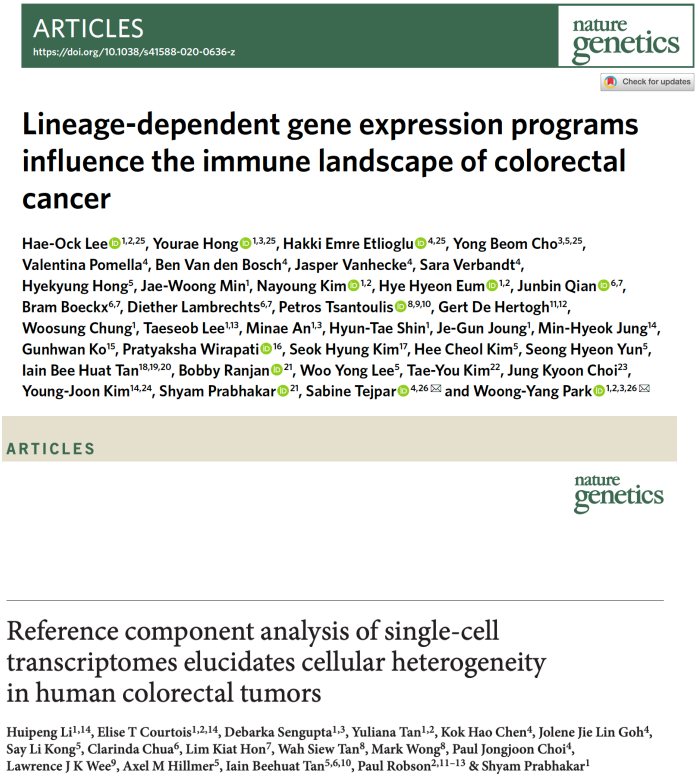
(NAR Cancer – Zhou et al., 2023)

CircRNA detection using short-read analysis pipeline



(Cabanski et al., 2014)

Leverage existing scRNA-Seq cohorts to infer circRNA expression



ARTICLES
https://doi.org/10.1038/s41588-020-0636-z
nature genetics
Check for updates

Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer

Hae-Ock Lee[1,2,25], Yourae Hong[1,3,25], Hakki Emre Etlioglu[4,25], Yong Beom Cho[3,5,25], Valentina Pomella[4], Ben Van den Bosch[4], Jasper Vanhecke[4], Sara Verbandt[4], Hyekyung Hong[5], Jae-Woong Min[1], Nayoung Kim[1,2], Hye Hyeon Eum[1,2], Junbin Qian[6,7], Bram Boeckx[6,7], Diether Lambrechts[6,7], Petros Tsantoulis[8,9,10], Gert De Hertogh[11,12], Woosung Chung[1], Taeseob Lee[1,13], Minae An[1,3], Hyun-Tae Shin[1], Je-Gun Joung[1], Min-Hyeok Jung[14], Gunhwan Ko[15], Pratyaksha Wirapati[16], Seok Hyung Kim[17], Hee Cheol Kim[5], Seong Hyeon Yun[5], Iain Bee Huat Tan[18,19,20], Bobby Ranjan[21], Woo Yong Lee[5], Tae-You Kim[22], Jung Kyoon Choi[23], Young-Joon Kim[14,24], Shyam Prabhakar[21], Sabine Tejpar[4,26] ✉ and Woong-Yang Park[1,2,3,26] ✉

ARTICLES
nature genetics

Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors

Huipeng Li[1,14], Elise T Courtois[1,2,14], Debarka Sengupta[1,3], Yuliana Tan[1,2], Kok Hao Chen[4], Jolene Jie Lin Goh[4], Say Li Kong[5], Clarinda Chua[6], Lim Kiat Hon[7], Wah Siew Tan[8], Mark Wong[8], Paul Jongjoon Choi[4], Lawrence J K Wee[9], Axel M Hillmer[5], Iain Beehuat Tan[5,6,10], Paul Robson[2,11–13] & Shyam Prabhakar[1]

Lee 2020:
6 Belgian patients,
6 cell types:

1. T cells
2. B cells
3. Mast cells
4. Stromal cells
5. Myeloid cells
6. Epithelial cells

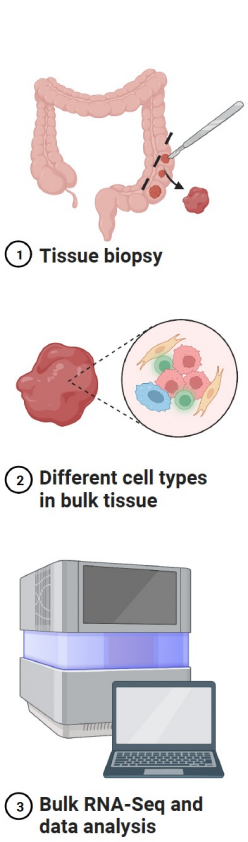
Li 2017:
11 Singaporean patients,
7 cell types:

1. T cells
2. B cells
3. Mast cells
4. Endothelial cells
5. Fibroblasts
6. Macrophages
7. Epithelial cells

CIBERSORT signature matrices

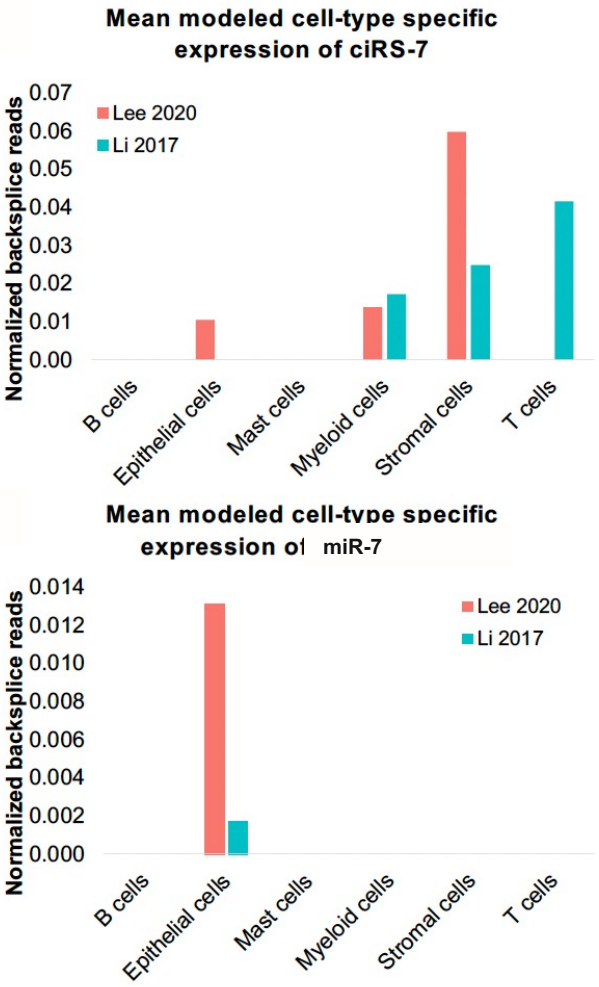
Lee et al (Nat Genetics, 2020); Li et al. (Nat Genetics, 2017); Zhao et al. (NAR Cancer, 2023)

Custom NNLS statistical model to extrapolate cell-type specific circRNA expression



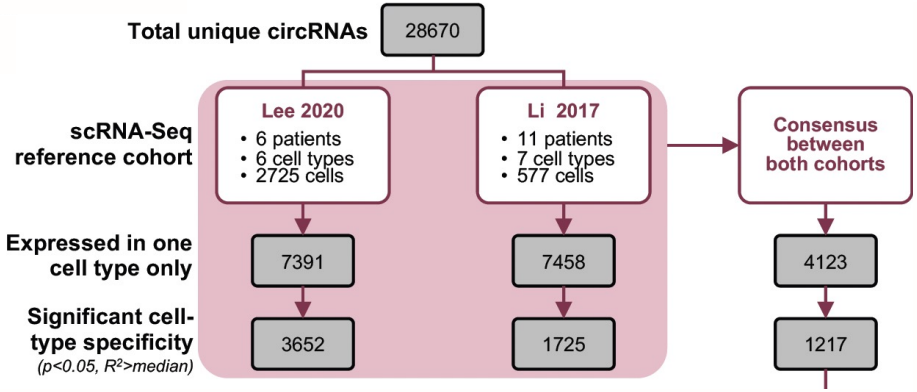
(NAR Cancer – Zhou et al., 2023)

Validation of ciRS-7 and miR-7 cell type enriched expression



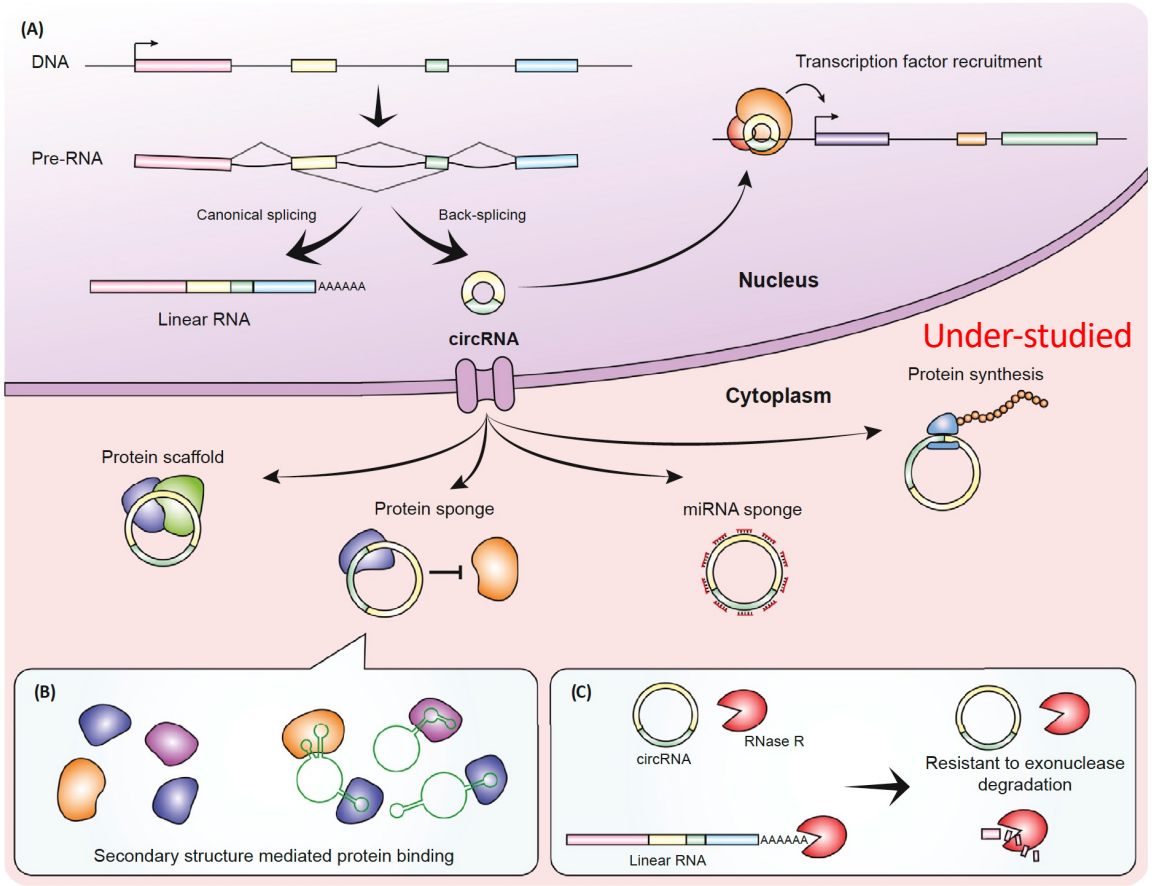
(NAR Cancer – Zhou et al., 2023)

NNLS model results reveal cell-type specific circRNAs in CRC TME



(NAR Cancer – Zhou et al., 2023)

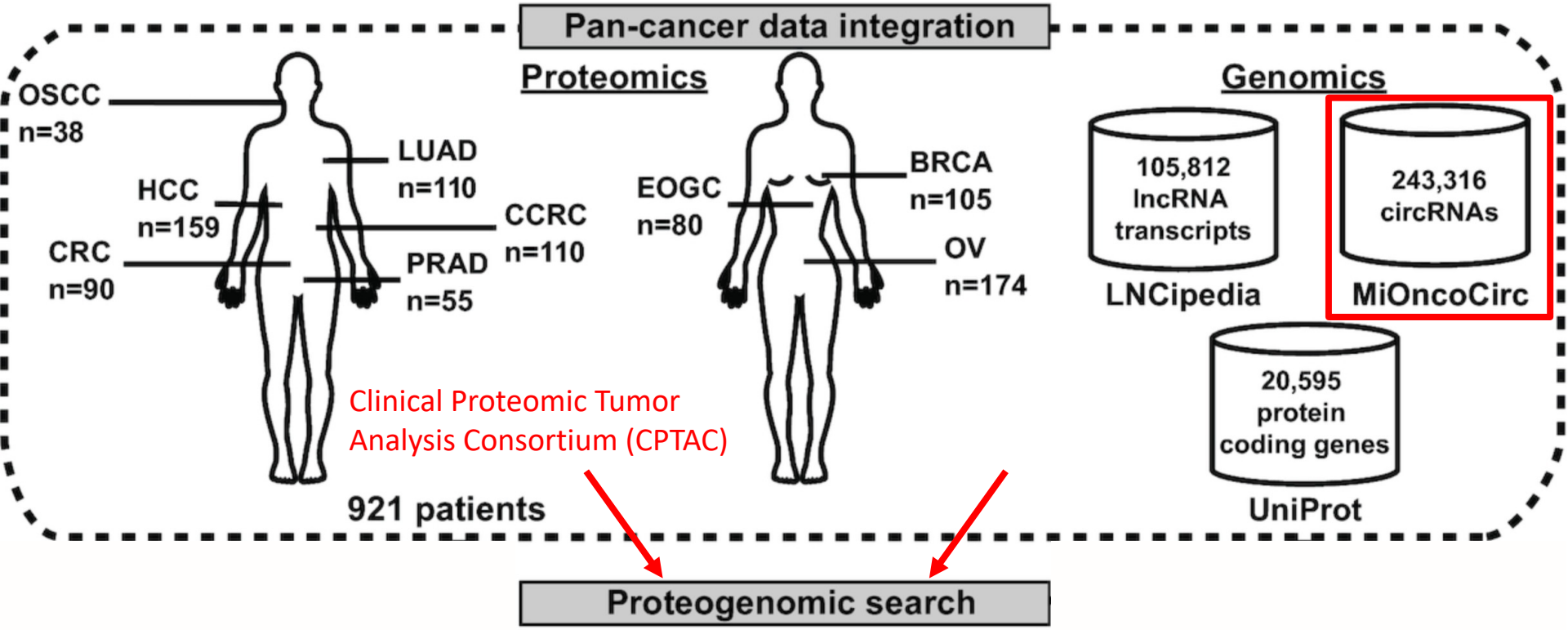
Putative functions of circRNAs remain under-studied in cancer



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

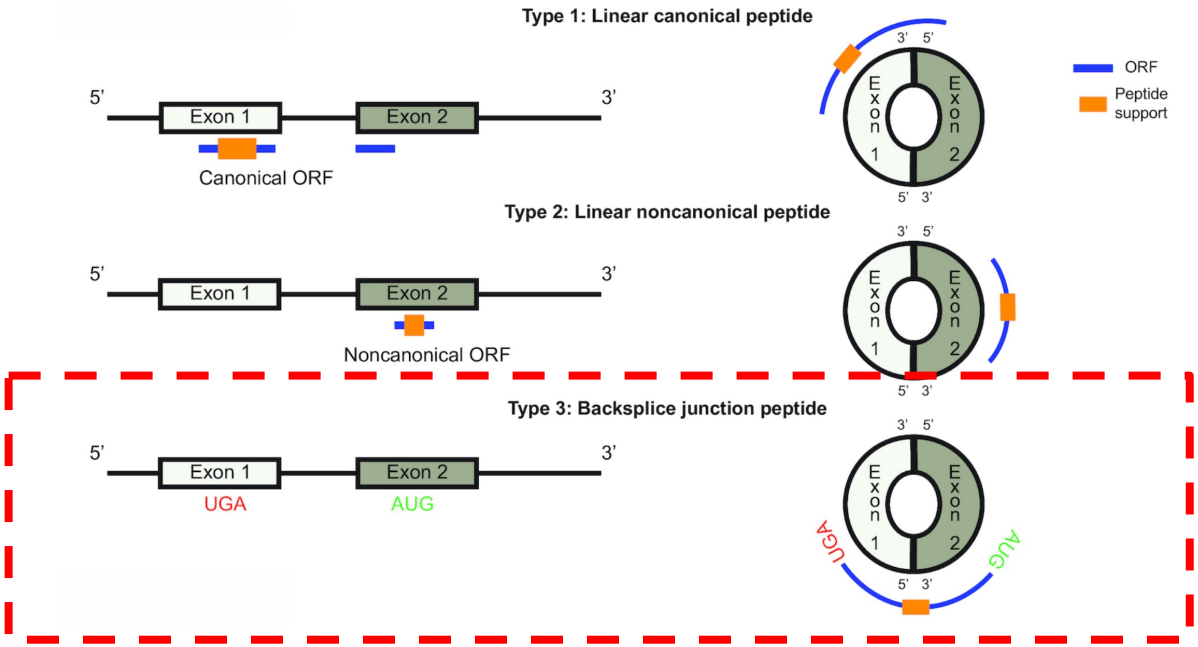


https://www.maherlab.com/peptransdb

(Othoum et al., 2020)

Possible types of peptides encoded by circRNAs

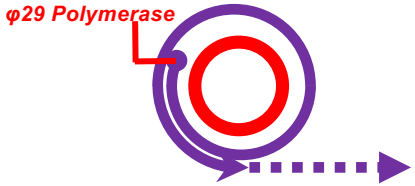
PepTransDB:
Total: 3238 peptides
from 2834 circRNAs



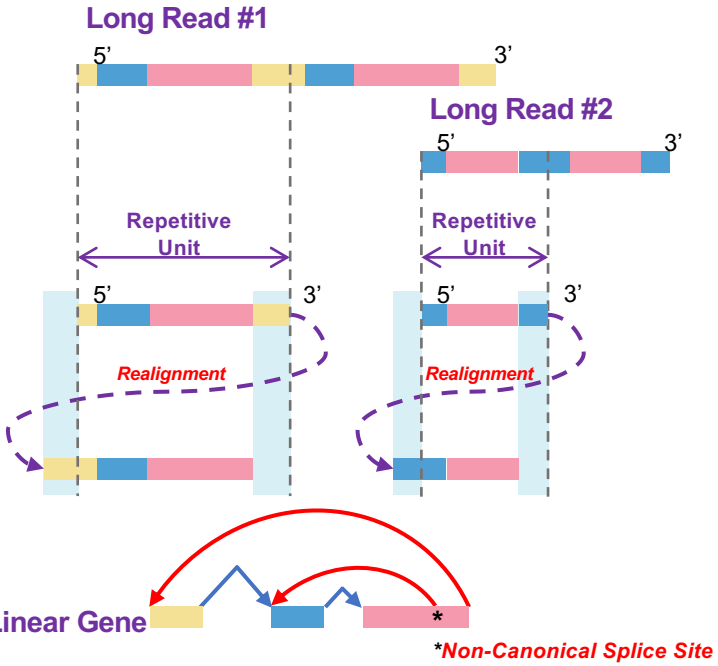
Junction: 2010 peptides
from 1964 circRNAs

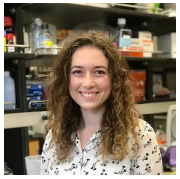
(Othoum et al., 2021)

Improved circRNA detection pipeline using long-read RNA-Seq



1. CircRNA Rolling Circle Amplification
2. PacBio Long-Read RNA-Seq
3. TRF → Consensus Reads
4. minimap2 → Mapped Reads
5. isoCirc/CIRI-long → Backsplice Junction Reads
6. CircRNA Discovery

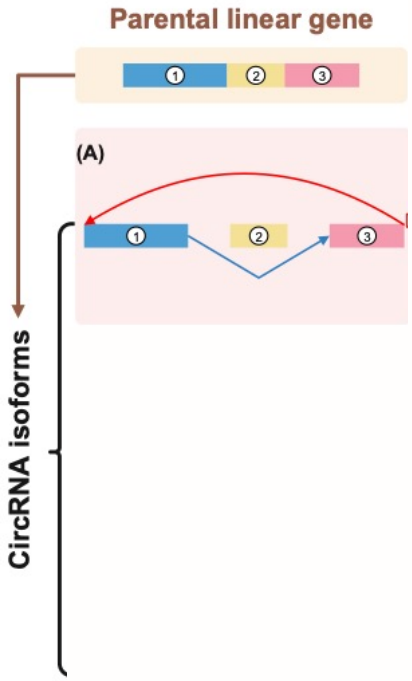




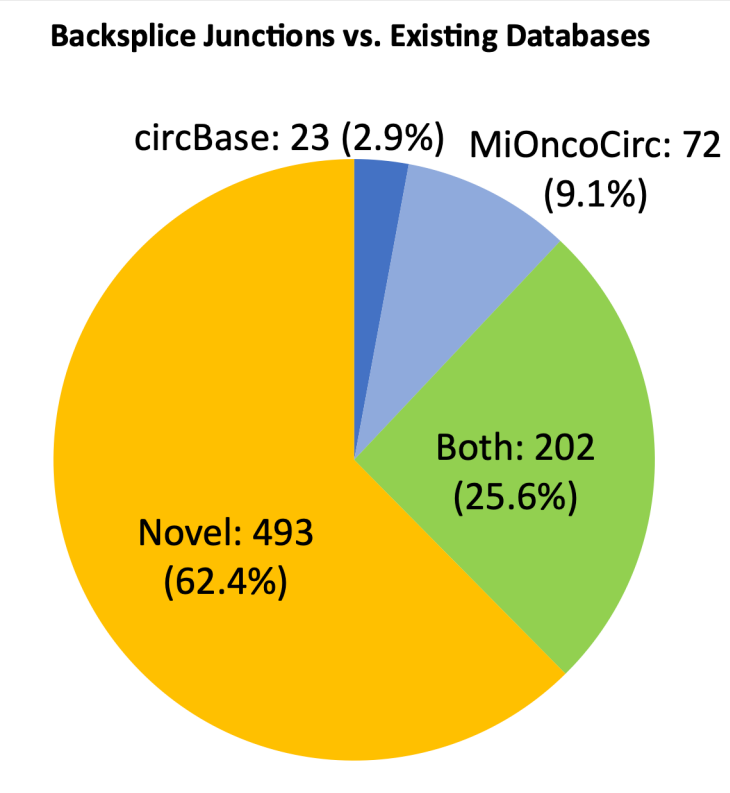
Long-read sequencing produces 48x enrichment of circRNA-containing reads

	Short-read	Long-read				
		isoCirc		CIRI-long		
	Percentage of short reads with circRNAs	Percentage of short reads with circRNAs	Fold read enrichment vs. short-read	Percentage of short reads with circRNAs	Fold read enrichment vs. short-read	Average fold read enrichment vs. short-read
Average among all samples	**0.04%**	**0.57%**	**15.88x**	**2.23%**	**80.24x**	**48.06x**

Integrative proteogenomic search revealed novel peptides undetected by short-read sequencing alone



Long-read sequencing revealed novel circRNAs

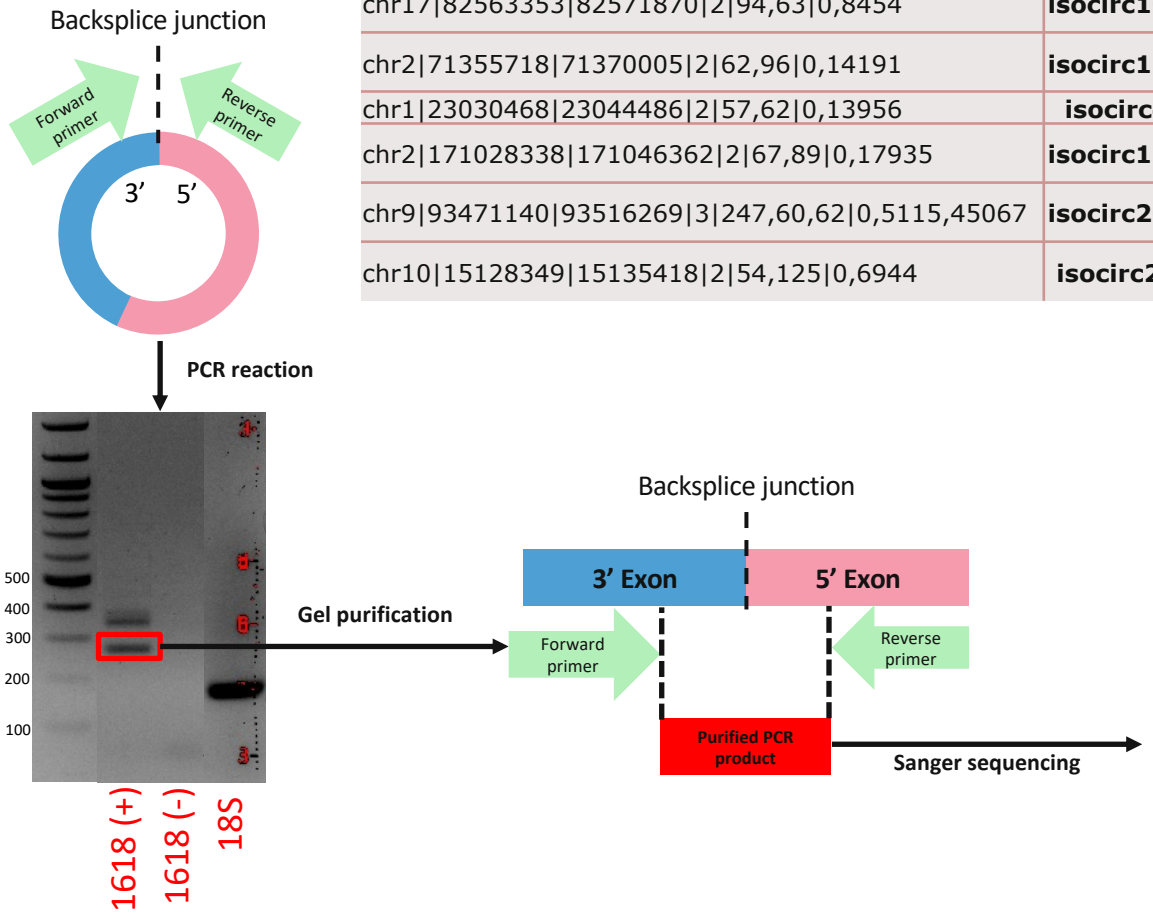


- ***What was missing in short-read?***
- ***How can we leverage long-read data to improve short-read results?***

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

- The biological and clinical significance of circRNAs is still emerging
- Integrated long-read approach discovers novel circRNAs
- Improved bioinformatic workflow for comprehensive full-length circRNA characterization
- More accurate circRNA transcript reconstruction will aid future mechanistic studies, such as evaluating the coding potential of circRNAs