



Consiglio Nazionale
delle Ricerche



ISTITUTO
ITALIANO DI
TECNOLOGIA



NextGenIT
Consolidation of the Italian Infrastructure
for Omics Data and Bioinformatics



MUR
Ministero dell'Istruzione e
della Ricerca



ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

TRAINING COURSE IN Computational Methods for Epitranscriptomics

Bari, 26th-28th April 2023



Overview of transcriptomics and epitranscriptomics

General overview on state-of-the-art approaches in transcriptome research

Francesco Nicassio



CENTER FOR GENOMIC SCIENCE OF IIT@SEMM - Milan

Credits: Tommaso Leonardi & Luca Pandolfini

TRANSCRIPTOMICS IN A NUTSHELL

Transcriptomics is the study of the ‘**transcriptome**,’ a term now widely understood to mean the complete set of all the ribonucleic acid (RNA) molecules (transcripts) expressed in some given entity, such as a cell, tissue, or organism.



Applications for Transcriptomics:

Study gene expression
and regulation (DEGs)

Identify novel transcripts
(de novo)

Investigate
RNA modifications

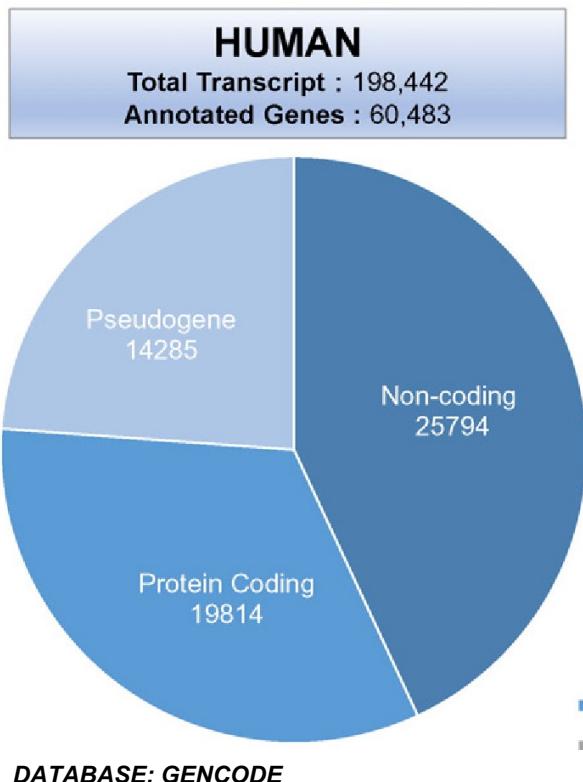
Uncover alternative
splicing events

MULTIMODAL approach

In recent years, advances in sequencing technology, bioinformatics, and computational tools have revolutionized the field of transcriptomics.

DECIPHERING TRANSCRIPTOME COMPLEXITY

Large Scale Consortia and tailored NGS approaches helped highlight functions and regulation of the human genome

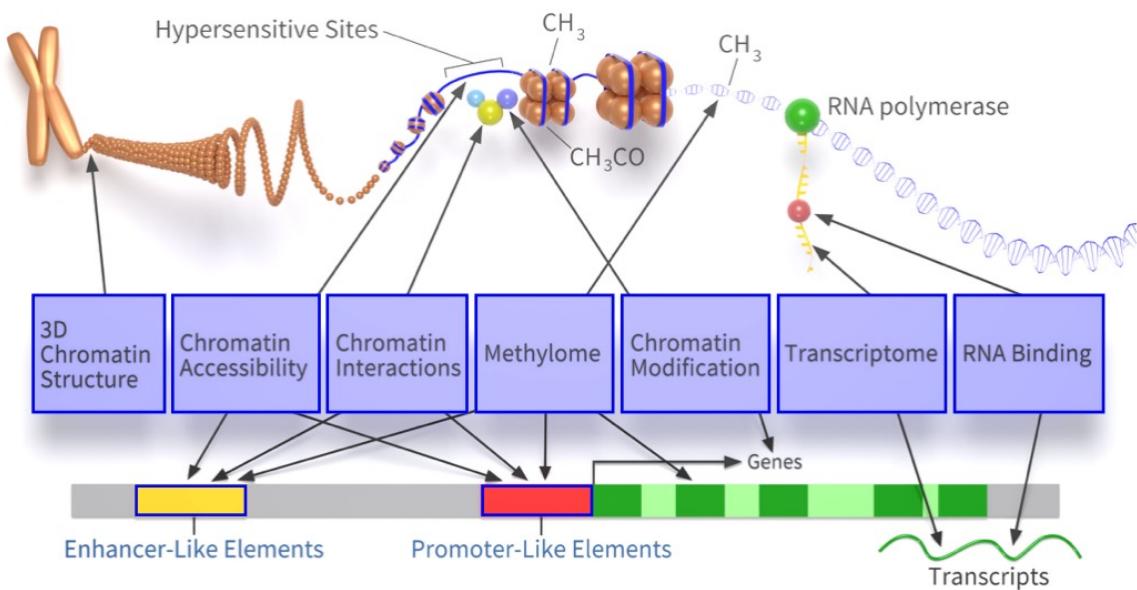


>> **ENCODE** - Encyclopedia of DNA Elements

<https://www.encodeproject.org>

>> **FANTOM** - Functional Annotation of Mammalian Genome

<https://fantom.gsc.riken.jp>

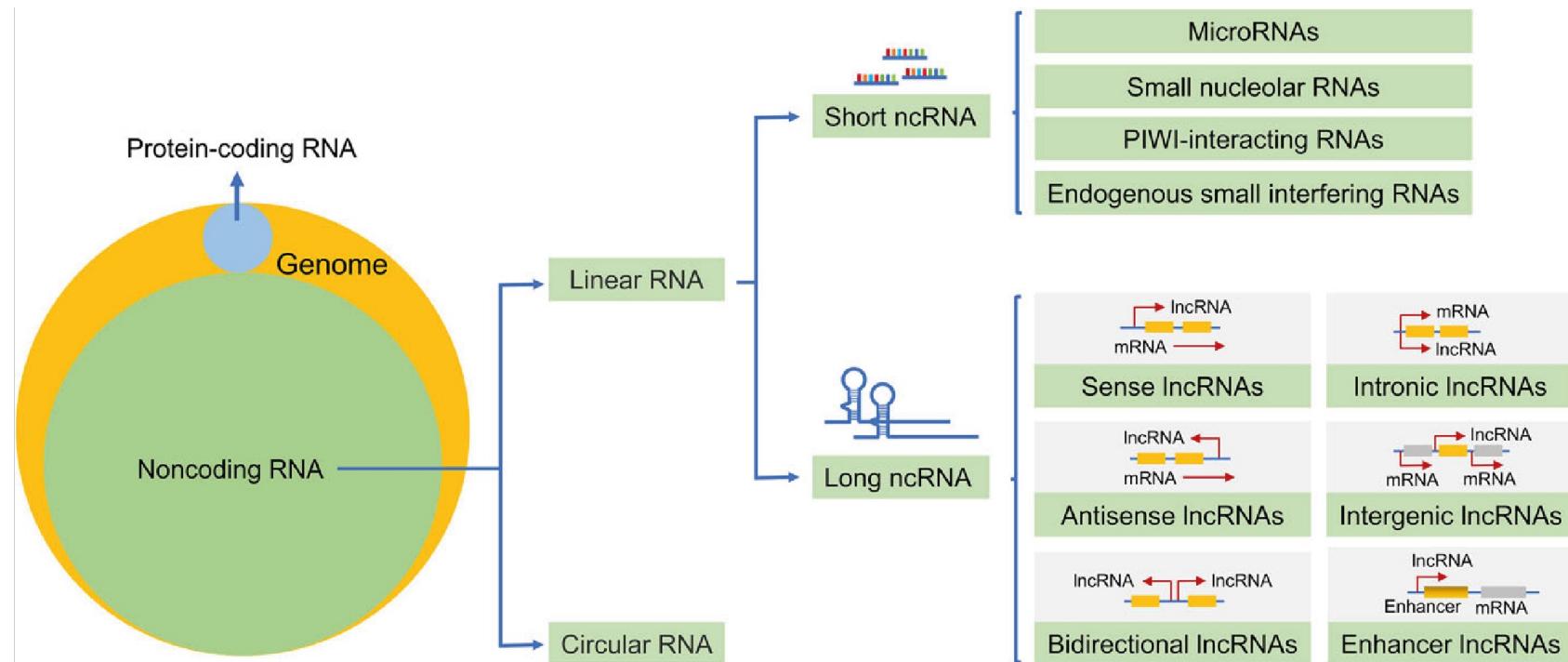


Non-coding DNA elements
promoters, enhancers, insulators

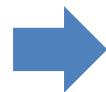
Coding and Non-coding transcripts
small or long by size

DON'T JUST CALL IT «RNA»: MIND THE BIOTYPES!

Technological advances have unravelled the remarkable complexity of RNA species



Many biotypes, many functions



Sophisticated regulatory system for gene expression

- Coordination of expression
- Spatio-temporal regulation
- Precision of gene dosage

TRANSCRIPTOME REPOSITORIES

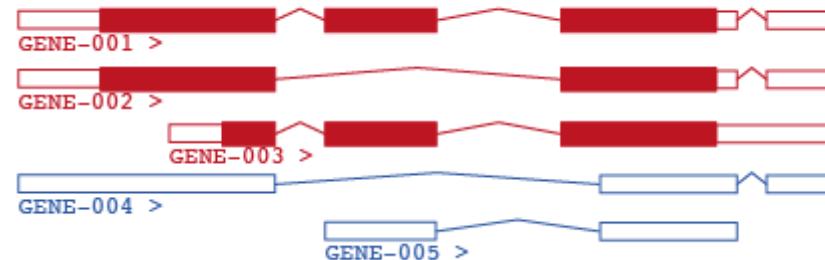
GENCODE

- ENCODE project
- >> **highly curated**
- 58721 human genes (206,694 transcripts)
 - Protein coding 19,940 >> 83,124 RNA
 - Long ncRNAs 16,066 >> 29,566 RNA
 - Small ncRNAs 7,577
- RNA biotype >> transcript origin, processing type, localization or coding potential

ENSEMBL

- part of the ENCODE project
- HAVANA = manually curated annotation (high trust)
- Transcript Support Level (TSL) is a method to highlight the well-supported vs. poorly-supported transcript
- Gene classification:
 - Coding (with >1 ORF)
 - Processed transcript (lncRNA, ncRNA, pseudogenes)
 - IG and TR Genes (variants)
 - TEC (to be experimentally confirmed)
- Transcript classification

Gene and transcript variants



* five transcripts, some **coding (red)** and **non-coding (blue)**

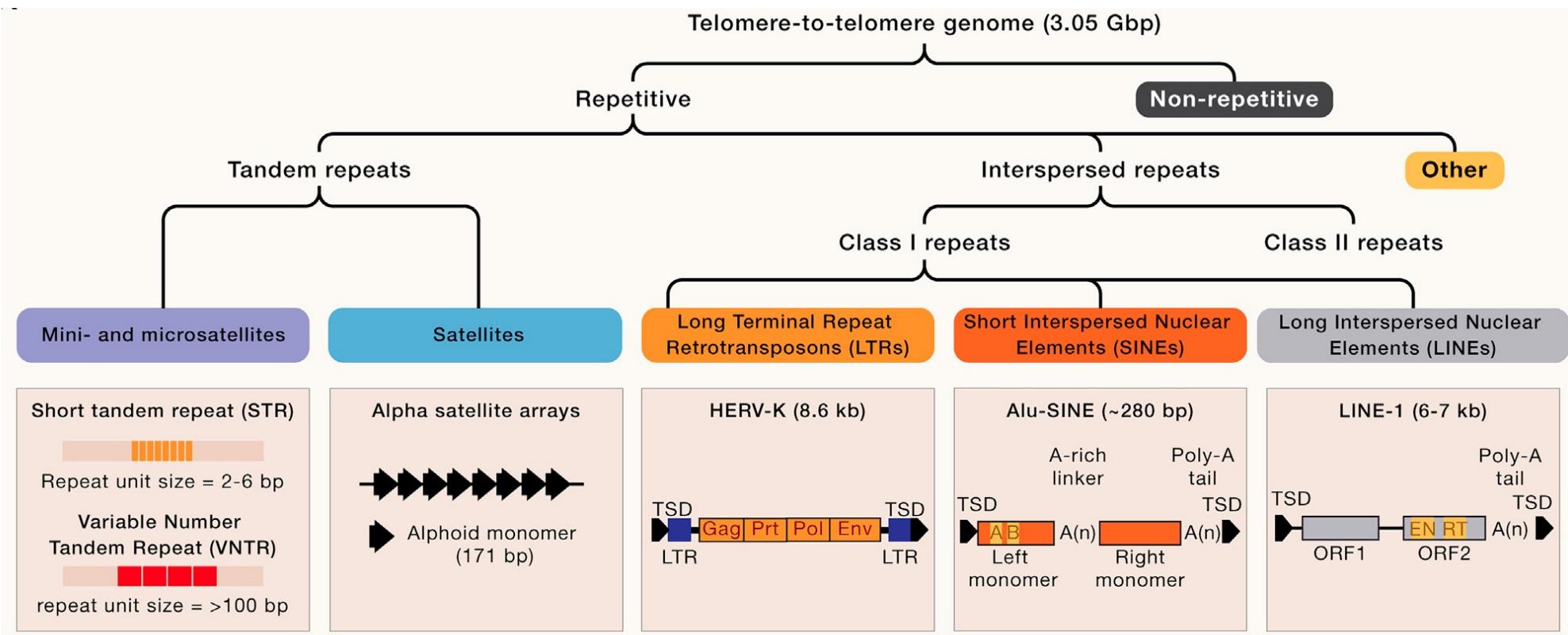
Types of ncRNA

Abbreviation Definition

tRNA	transfer RNA
Mt-tRNA	transfer RNA located in the mitochondrial g
rRNA	ribosomal RNA
scRNA	small cytoplasmic RNA
snRNA	small nuclear RNA
snoRNA	small nucleolar RNA
miRNA	microRNA precursors
misc_RNA	miscellaneous other RNA
lincRNA	Long intergenic non-coding RNAs

COMPLEXITY OF TRANSCRIPTOME: REPETITIVE ELEMENTS

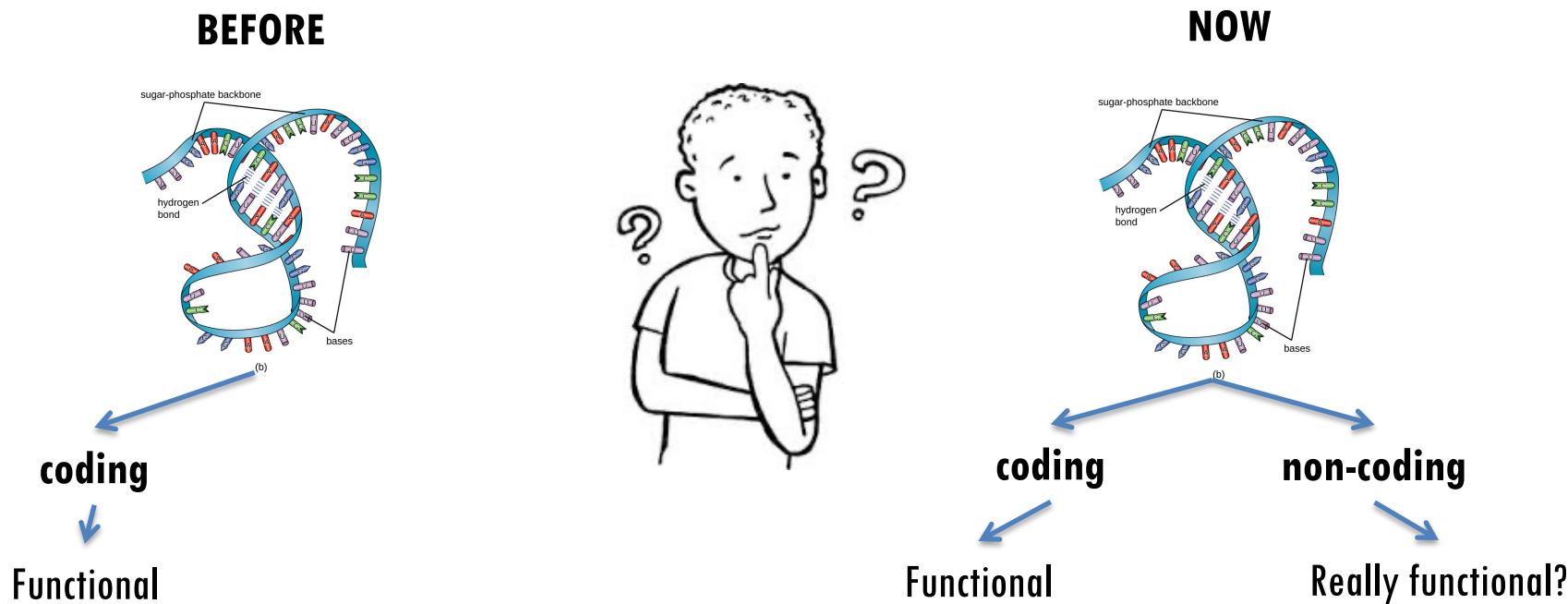
The non-coding landscape is further expanded by multiple DNA repetitive elements that are transcribed into non-coding RNAs.



Haws SA., Cell. 2022

From a computational perspective, repeats create ambiguities in alignment and assembly, which, in turn, can produce biases and errors when interpreting results.

TO CODE OR NOT TO CODE?



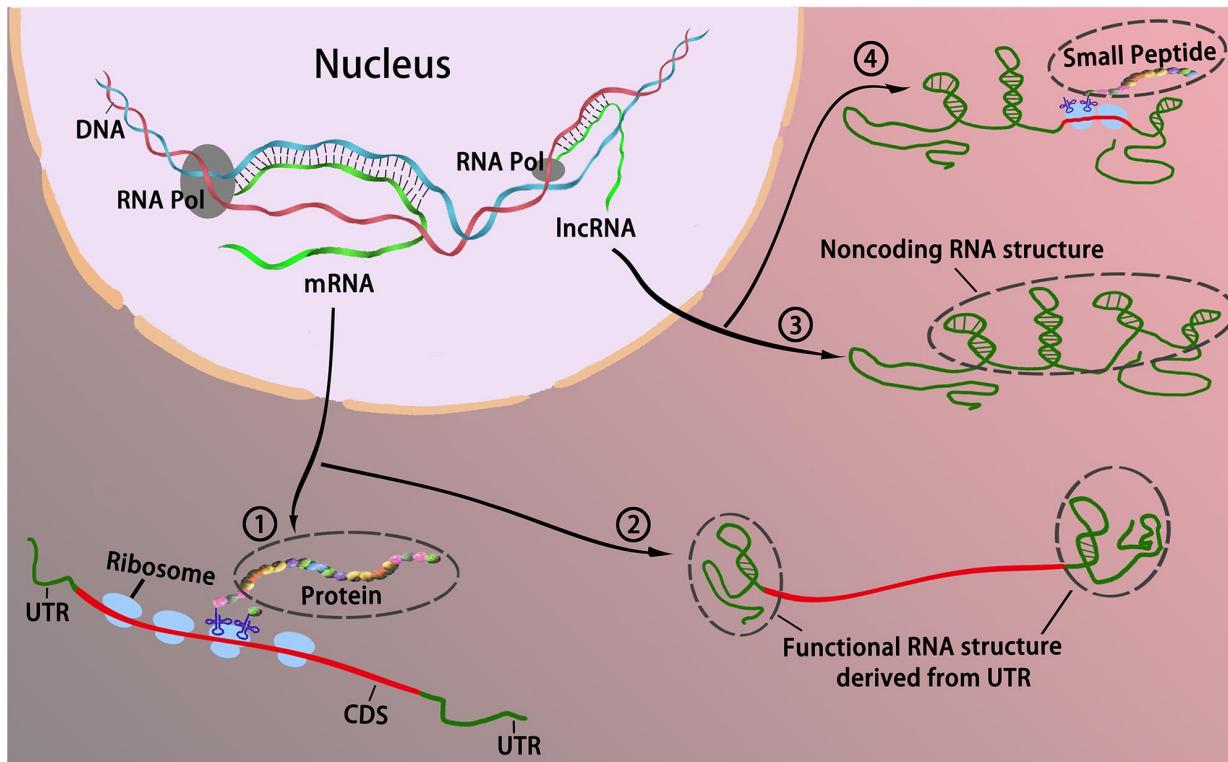
Coding vs Non-coding: biology doesn't care !!

DON'T RELY TOO MUCH ON DEFINITIONS!

We should care on what is functional and on what has a role
in a phenotype/process control (it depends on the context, too)

TO CODE OR NOT TO CODE? BIFUNCTIONAL RNAs

RNA molecules may possess both coding and non-coding functions, evolved to diversify phenotypes or produce convergent effects.



Li et al. *Frontiers in Genetics* 2019;
doi.org/10.3389/fgene.2019.00496

ARTICLE
DOI: [10.1038/s41467-018-05182-9](https://doi.org/10.1038/s41467-018-05182-9) OPEN

Endogenous transcripts control miRNA levels and activity in mammalian cells by target-directed miRNA degradation

Francesco Ghini¹, Carmela Rubolino¹, Montserrat Climent¹, Ines Simeone¹, Matteo J. Marzì¹ & Francesco Nicassio¹

Small ORFs into ncRNAs
Functional peptides



ARTICLE
<https://doi.org/10.1038/s41467-019-09754-1> OPEN

LncRNA EPR controls epithelial proliferation by coordinating *Cdkn1a* transcription and mRNA decay response to TGF-β

Martina Rossi^{1,2}, Gabriele Buccì³, Dario Rizzotto⁴, Domenico Bordo⁵, Matteo J. Marzì⁵, Margherita Puppo^{1,2}, Arielle Filhois⁶, Domenica Spadaro⁶, Sandra Citi⁶, Laura Emriente⁷, Michele Cilli⁷, Michele Nicassio^{1,5}, Alberto Inga⁴, Paola Briata¹ & Roberto Gherzi¹

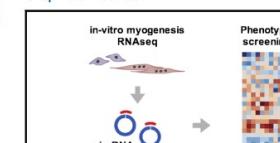
Non-coding functions into mRNAs
5' UTRs or 3'UTRs
circRNAs

Article

Molecular Cell

Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis

Graphical Abstract



Authors

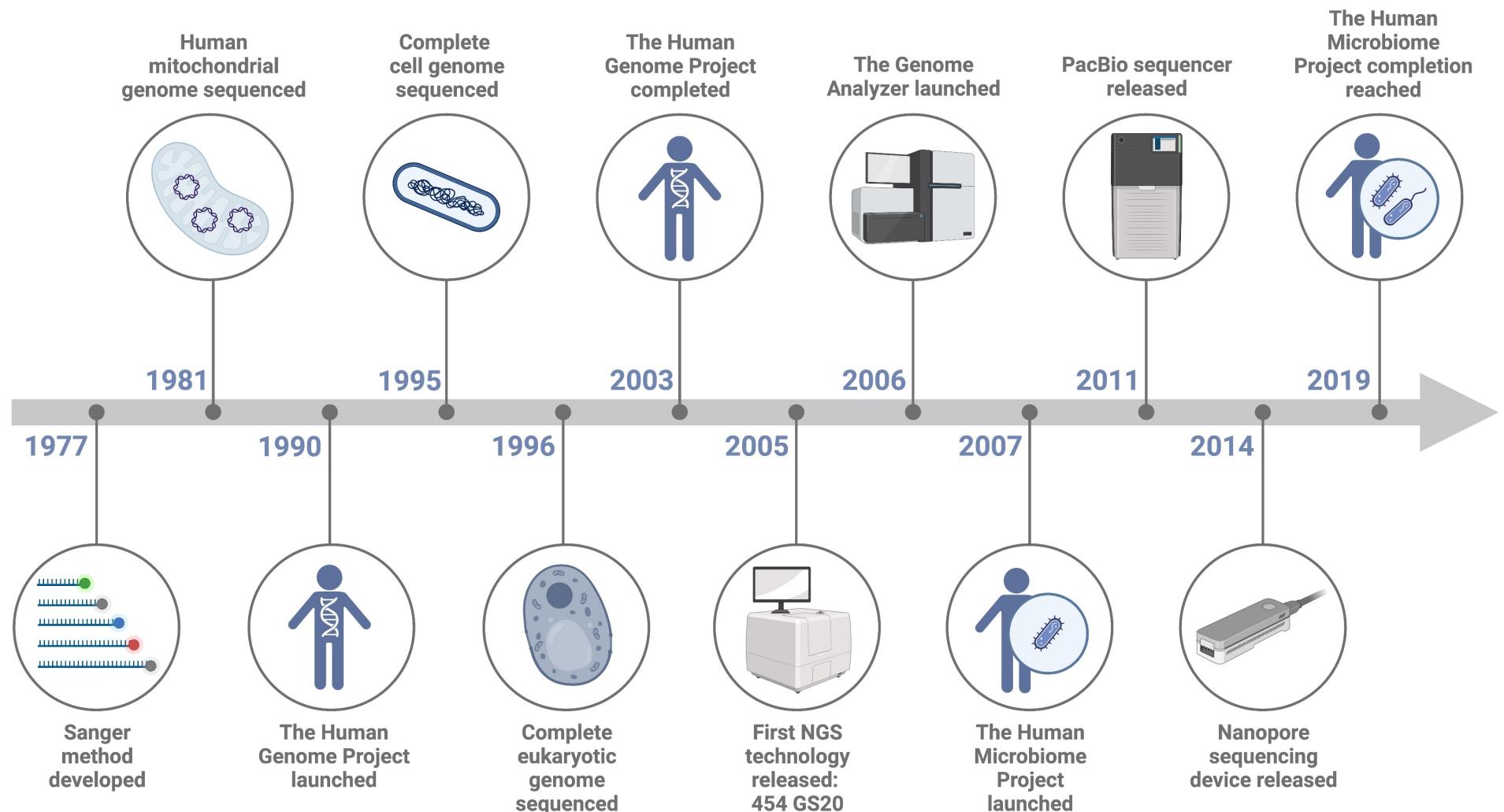
Ivano Legnini, Gaia Di Timoteo, Francesca Rossi, ..., Pietro Laneve, Niklaus Rajewsky, Irene Bozzoni

Correspondence
irene.bozzoni@uniroma1.it

In Brief
Legnini et al. identified circ-ZNF609, a

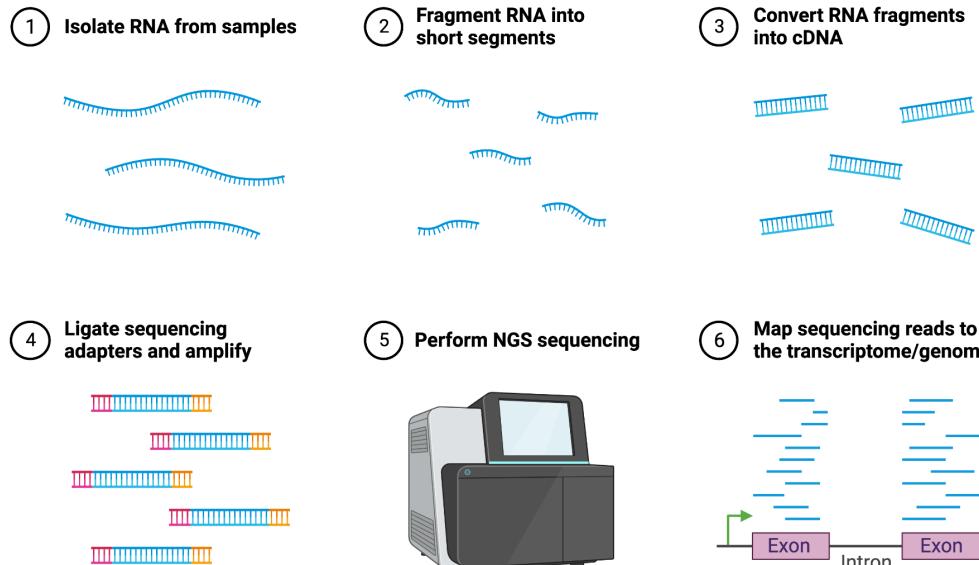
THE TECHNOLOGY – THE STORY SO FAR

In the past decades, several sequencing technologies have entered the research practice, generating new opportunities and challenges

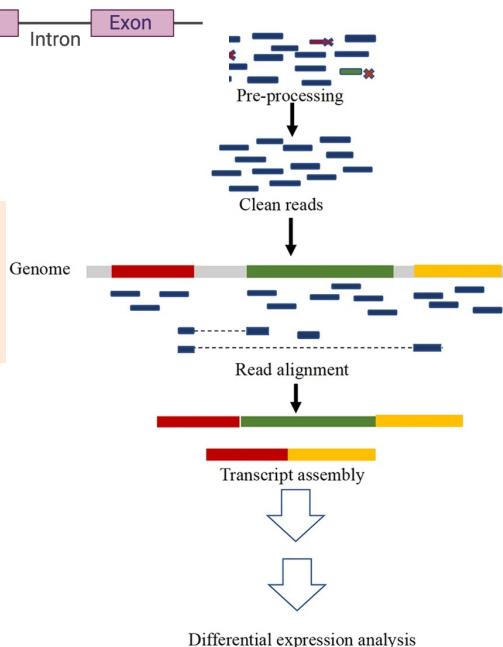


ILLUMINA **SHORT-READ** SEQUENCING

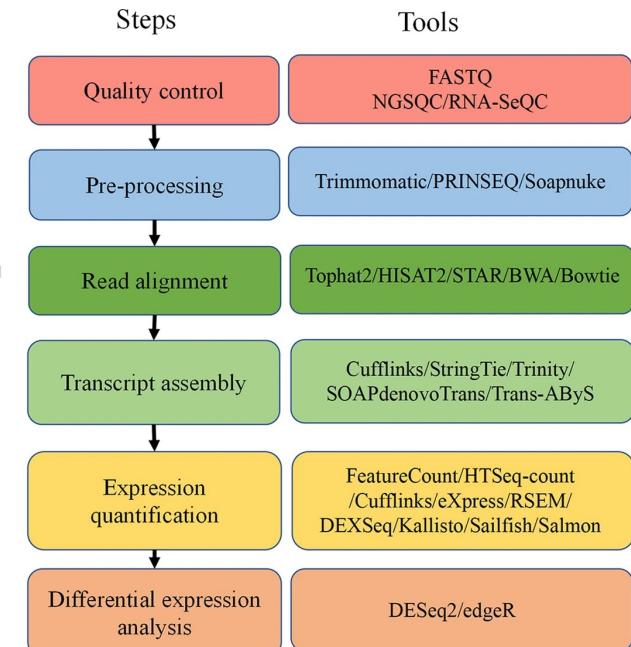
RNA Sequencing



Experimental Step: Library preparation and Sequencing

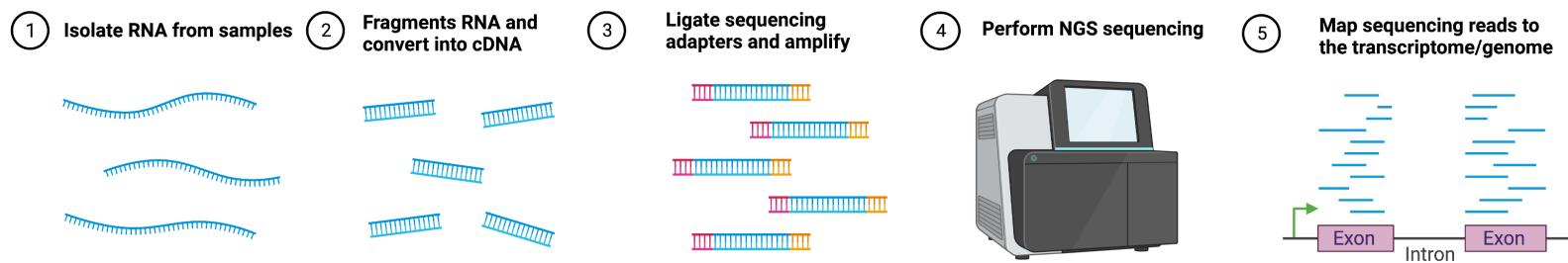


Analytical Step: Assembly/Alignment and Quantification



ILLUMINA **SHORT-READ** SEQUENCING

RNA Sequencing



PROs

- The first “unbiased” characterization of RNA species
- Quantitative and very sensitive (ultra low input protocols up to <1ng RNA)
- Coding and Non-coding species
- polyA-based or Total RNA
- Strand-specific reactions
- Pair-end mode helps isoform alignments

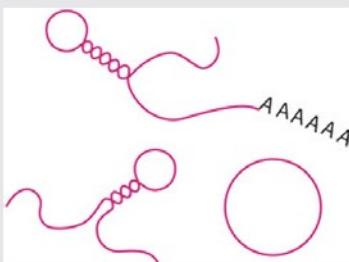
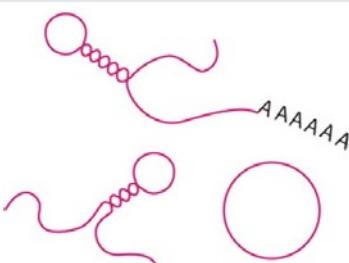
CONs

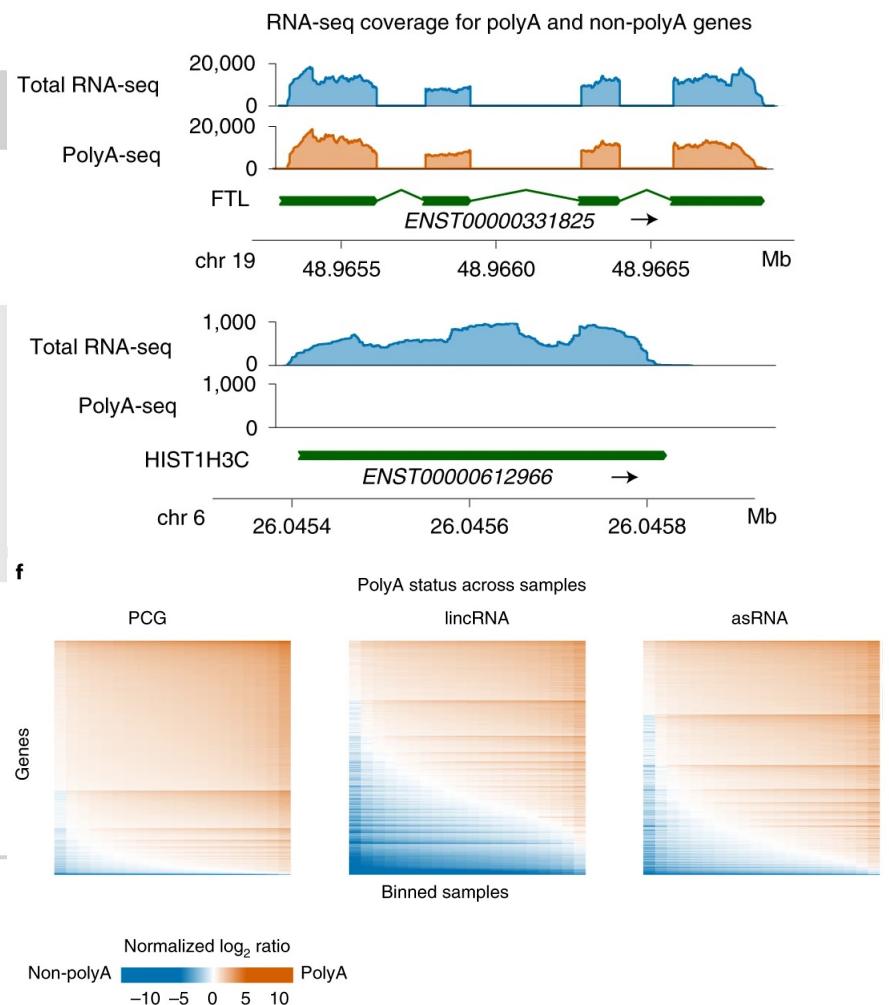
- **Small RNA species:** some small RNA species, (tRNAs or microRNAs), requires a different ad hoc protocol
- Repetitive elements are hard to analyse (**multi-mapping**)
- **cDNA** based (Artifacts may be introduced upon cDNA and amplification steps)

RNA EXPRESSION BY SHORT-READ SEQUENCING

polyA vs non-polyA

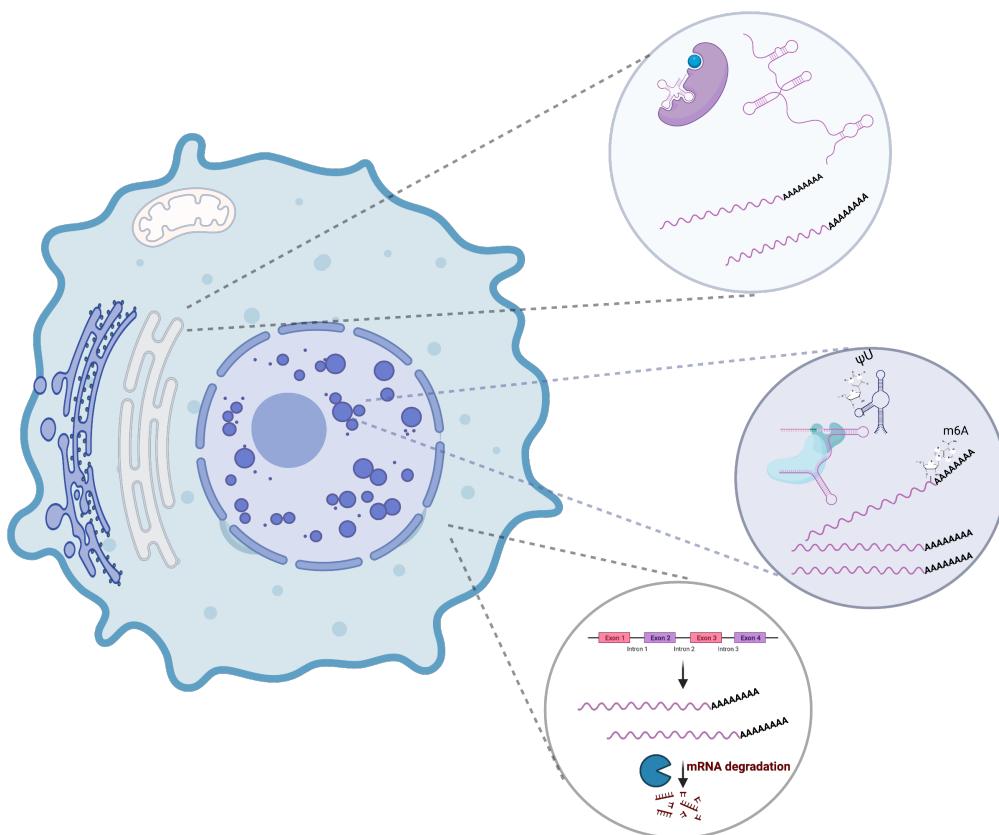
TABLE 2 RNA-seq technologies

RNA-seq technology	lncRNA	Strength/weakness
polyA+		+ High coverage for polyadenylated transcripts + Limited sequencing depth required - Misses out on non-polyadenylated/circular transcripts
Total RNA		+ Any lncRNA +/- High percentage of intronic reads - High sequencing depth required
Capture seq		+ Any lncRNA + Focus reads on genes of interest + Useful for the profiling of low-abundant RNAs in circulation - Custom probe design and cost



SHORT-READ SEQUENCING METHODOLOGIES

short read sequencing is extremely versatile and allowed the implementation of specific applications to characterize the transcriptome in its details



- **RNA localization** (*subcellular localization*)

- Fractionation protocols >> sequencing
- Proximity ligation >> APEX-sequencing

- **RNA modifications** and **RNA structure**

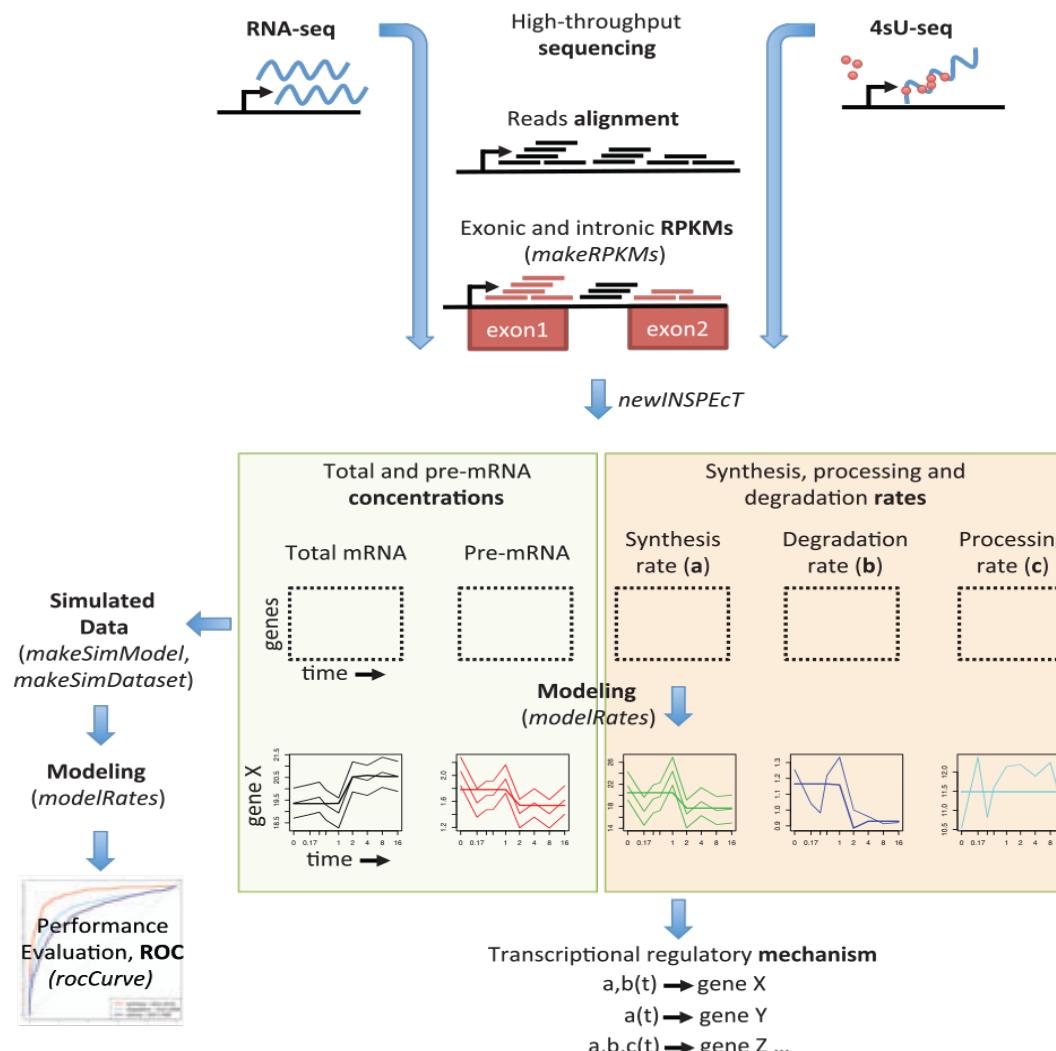
- Biochemical protocols
 - AB-based (*miCLIP2*)
 - Chemical modification (*GLORI, SHAPE/C-SHAPE*)

- **RNA dynamics** (*synthesis, maturation, decay*)

- 5'UTR and TSS (*CAGE, NET-CAGE*)
- 3'UTR (*polyA-seq, FLAM-seq*)
- Metabolic labelling (*4SU or IAA – SLAM-SEQ*)
- Global Run-On sequencing (*GRO-seq*)
- Computational: intron/exon ratios

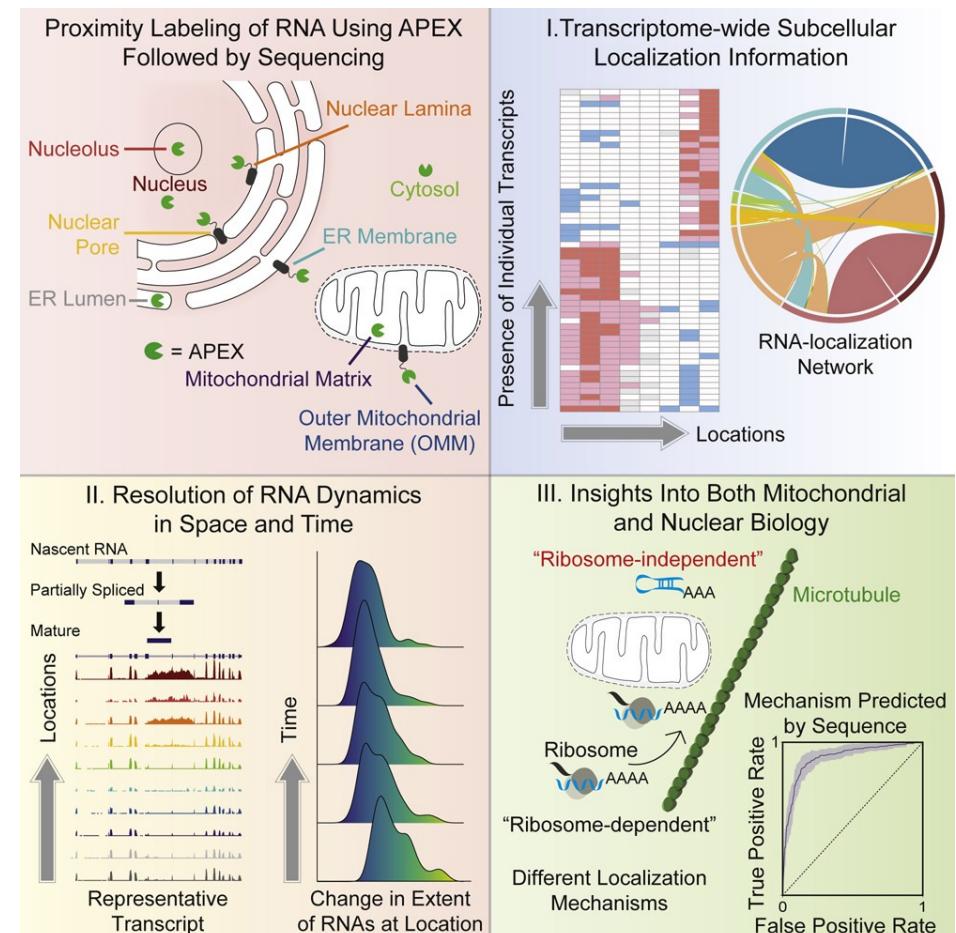
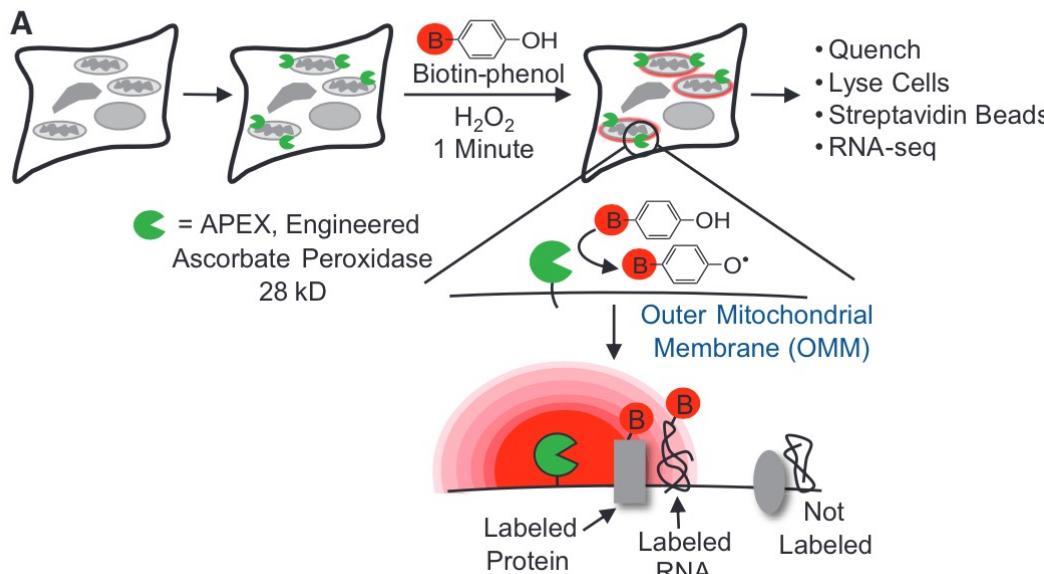
SHORT-READ SEQUENCING METHODOLOGIES

INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments



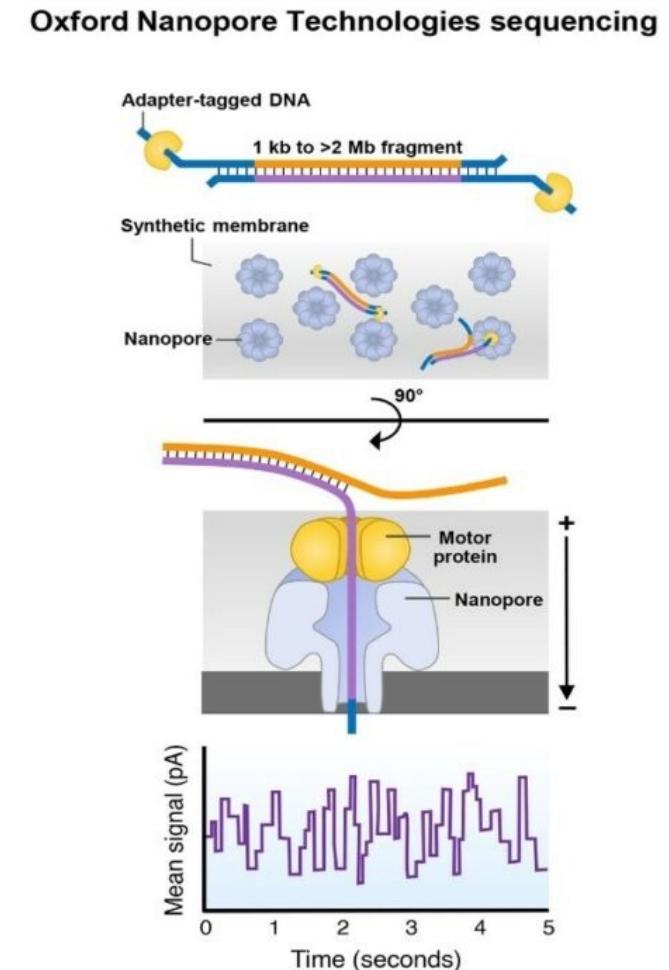
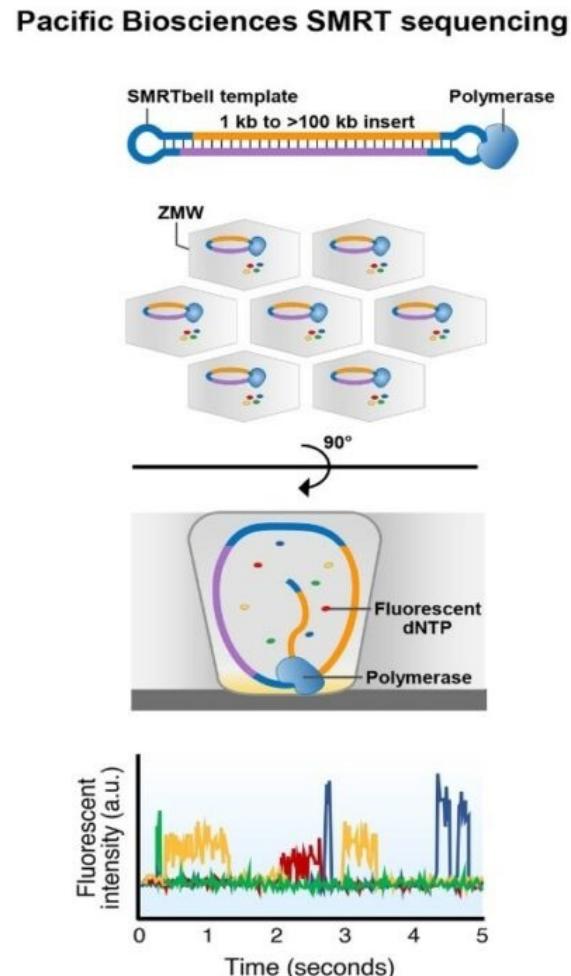
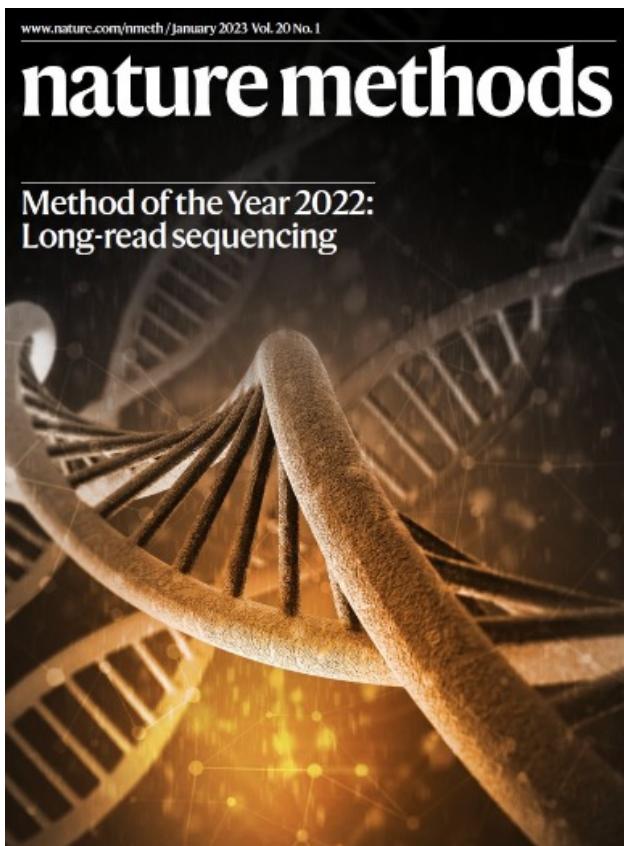
SHORT-READ SEQUENCING METHODOLOGIES

APEX-seq: a method for RNA sequencing based on direct proximity labelling of RNA using the peroxidase enzyme APEX2.



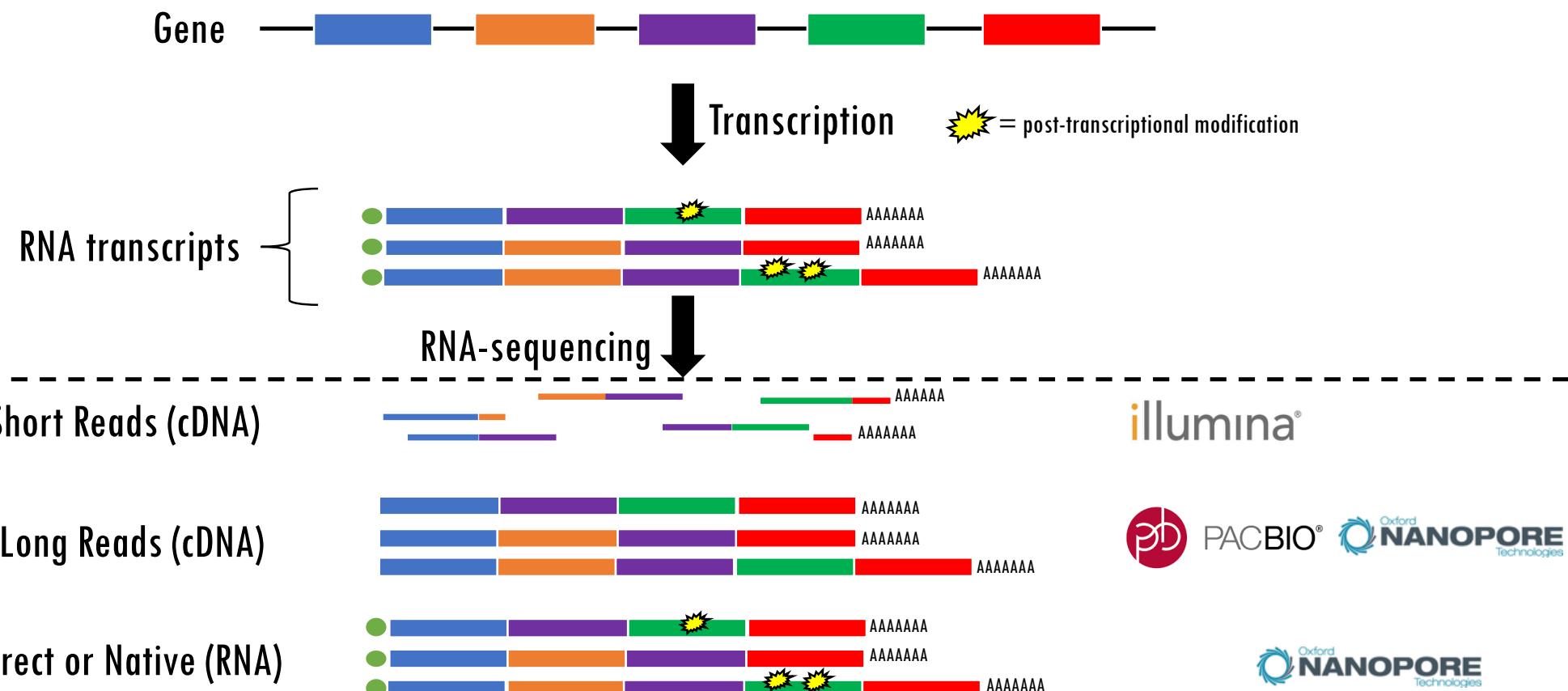
THIRD-GENERATION - LONG-READ SEQUENCING

Advent of **Long Read sequencing technologies** shifted the paradigm of sequencing, with single-molecule applications

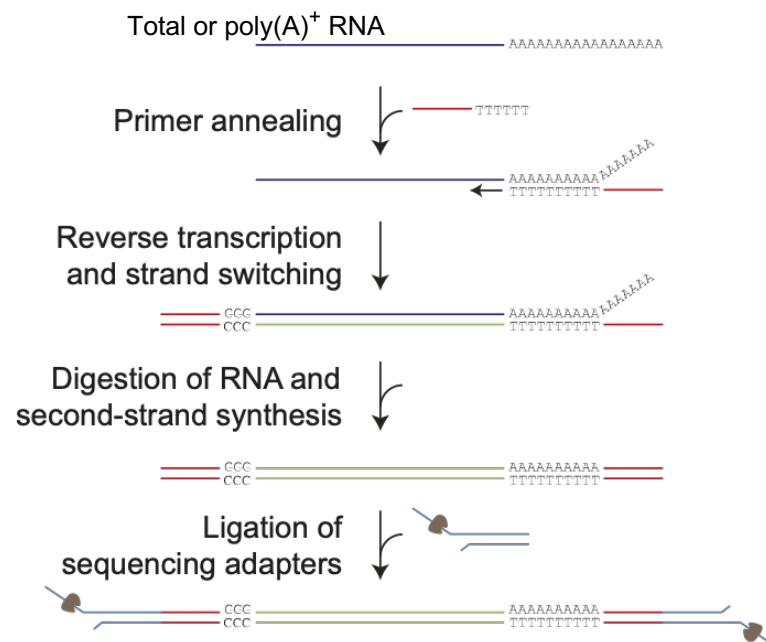


THE LONG READ TECHNOLOGIES IN RNA SEQUENCING

Advent of **Long Read sequencing technologies** improve mRNA analysis
longer read, single molecule information and native sequencing

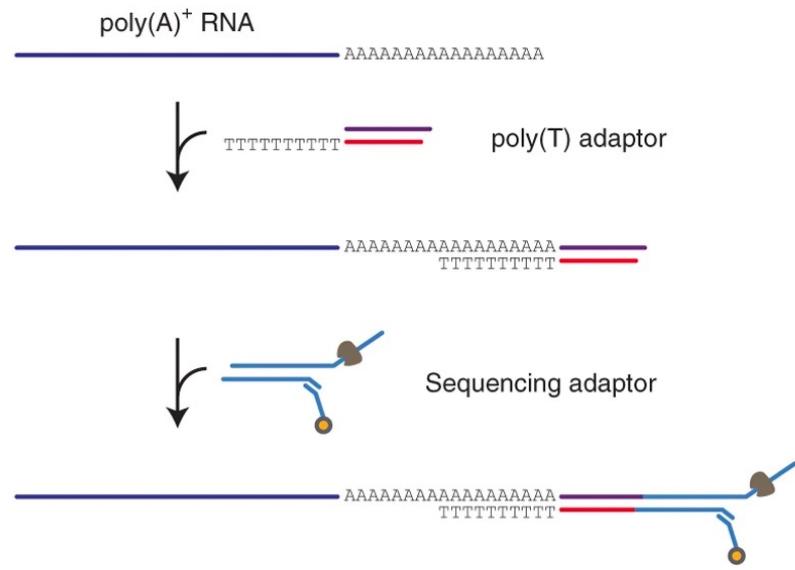


THE LONG READ TECHNOLOGIES IN RNA SEQUENCING



cDNA sequencing

- Higher coverage
- Full-length Sequencing
- Sensitivity can be increased with few cycles of cDNA amplification
- RT-bias can be introduced
- Direct identification of modifications is not possible

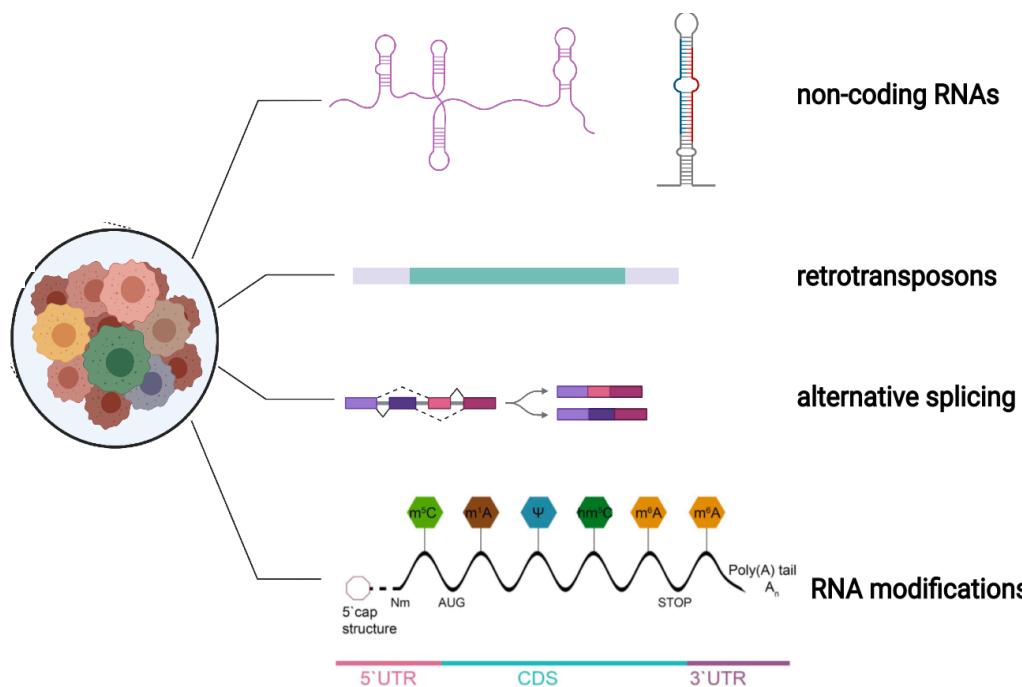


Direct RNA sequencing

- Lower coverage
- Full-length Sequencing is not standard
- Low Sensitivity and high amount of material
- No RT-bias
- Direct identification of modifications

RNA APPLICATIONS OF LONG-READ SEQUENCING

High-resolution and multi-level transcriptional and epitranscriptomics characterisation



Long read RNA Seq applications:

- Viral RNA characterization (es. SARS-CoV2)
- Transposable Element (TE) Characterization
- Full-length RNAs, isoform characterization
- polyA-length analysis
- Multi-level transcriptional analysis upon perturbation
 - Splicing : transcript isoforms
 - Epitranscriptomics: RNA mods
 - PolyA-length
 - Quantification of Repetitive Elements

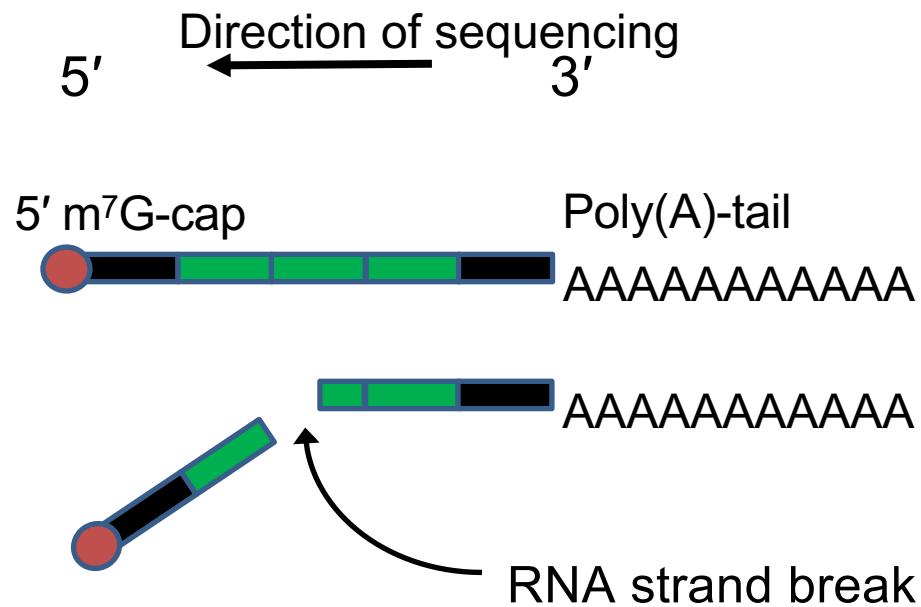
Direct RNA Seq applications:

- Detection of modified bases: epitranscriptomics

NOT ALL *LONG RNA READS* ARE FULL-LENGTH

strand breaks can terminate reads.

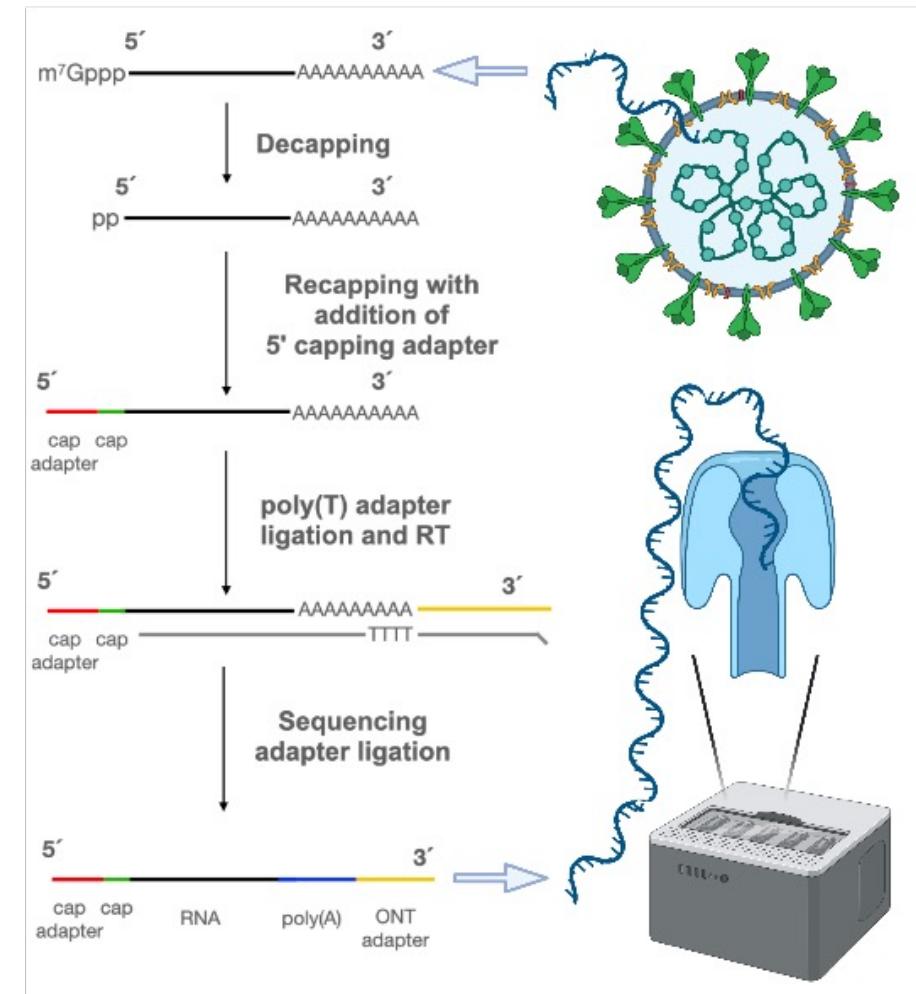
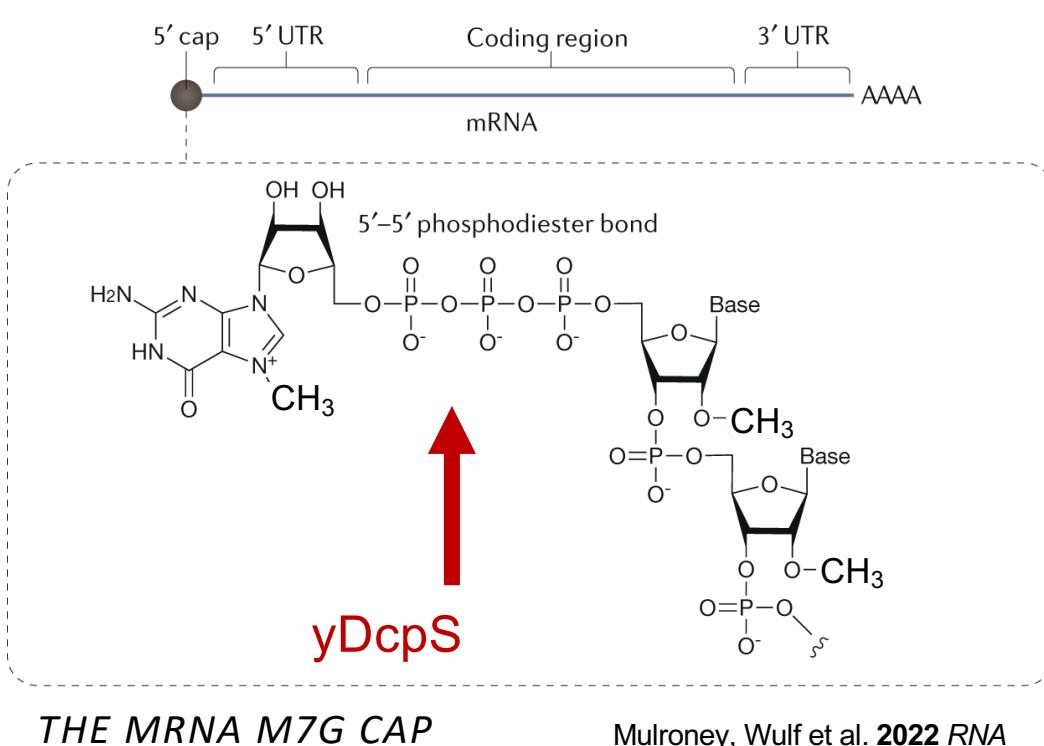
How to discriminate for TRUE 5' ends?



NOT ALL LONG RNA READS ARE FULL-LENGTH

strand breaks can terminate reads.

How to discriminate for TRUE 5' ends?

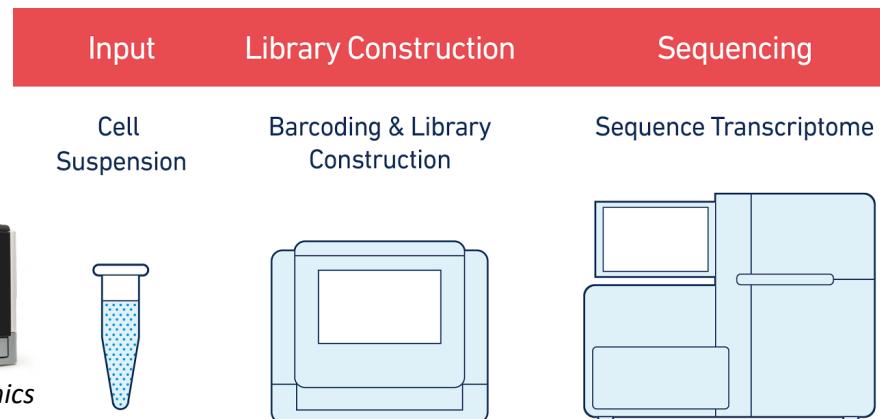


Ugolini et al. 2022 Nucleic Acid Research

SINGLE-CELL TECHNOLOGIES IN RNA SEQUENCING

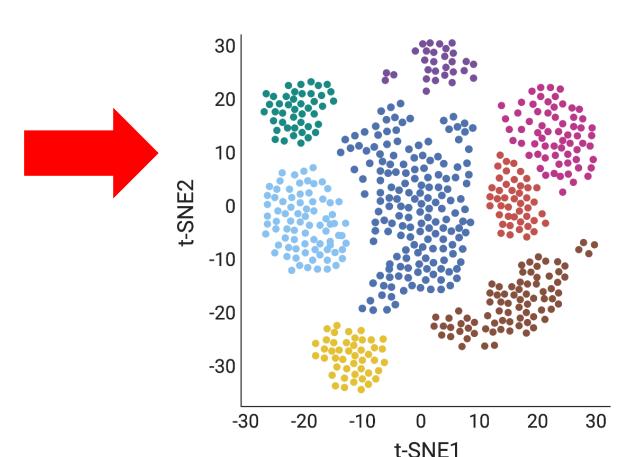
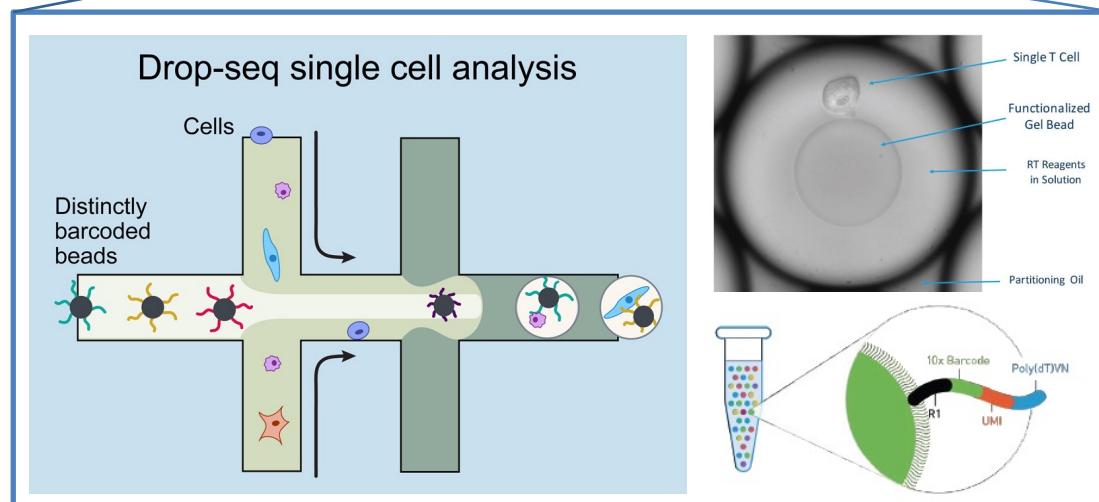
RNA sequencing has been evolved at single cell resolution (different from BULK)

Dropseq(i.e.10XGenomics) vs **Single-well** approach (icell8)



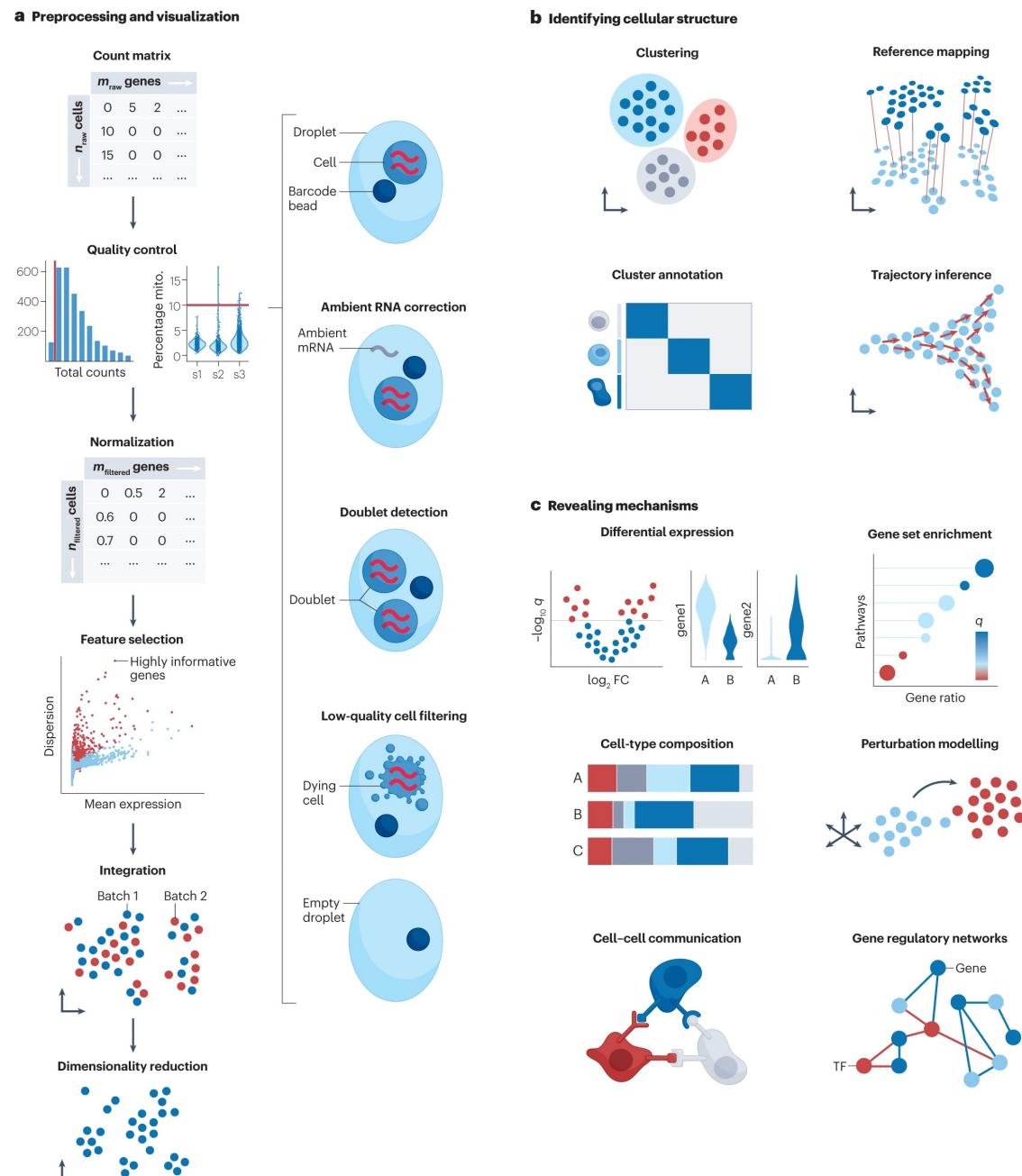
Deconvolution of **Cellular** Heterogeneity

- Diversification of cell identity
- Different cell states within the same population
- Transcriptional markers
- Trajectories (developmental)



SINGLE-CELL TECHNOLOGIES IN RNA SEQUENCING

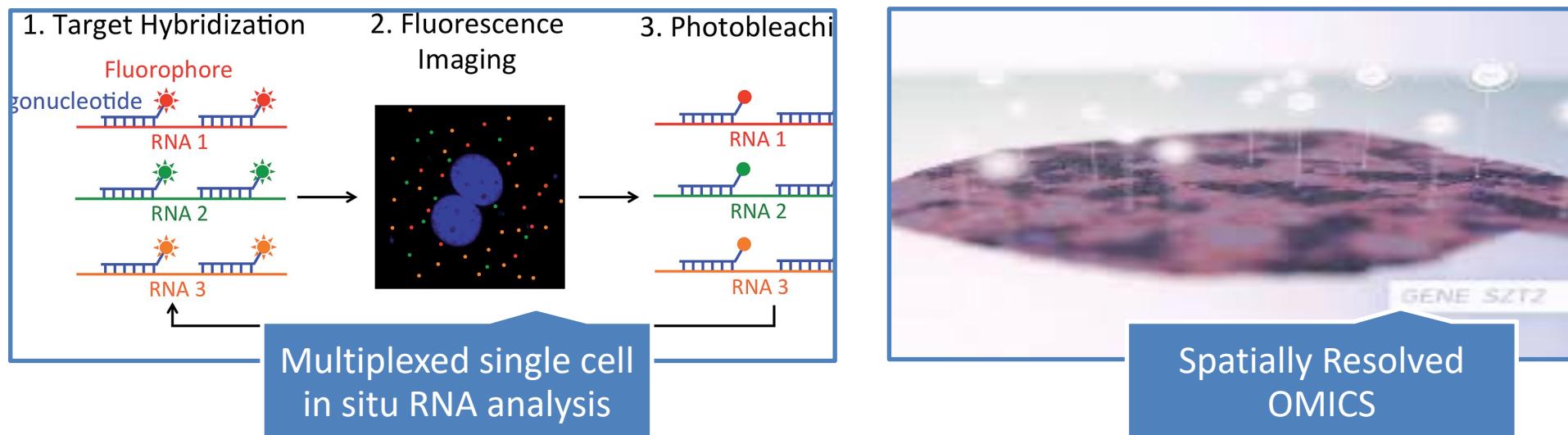
Overview of unimodal analysis steps for scRNA-seq



(F. Theis lab) Heumos, L., Schaar, A.C., Lance, C. *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* (2023). <https://doi.org/10.1038/s41576-023-00586-w>

SPATIALLY-RESOLVED OMICS IN RNA SEQUENCING

RNA sequencing with spatial information provides new challenges and opportunities for Transcriptome Analysis



- Imaging
 - Few RNA species but specific (with probes)
 - Single Molecule + Single Cell Resolution
 - 2D cell culture
- Sequencing
 - Many RNA species ($>10,000$, with probes)
 - Few cell resolution
 - Clinical oriented - (frozen samples or FFPE) Brain- or Cancer-



COSTS OF THE TECHNOLOGY IS HIGH

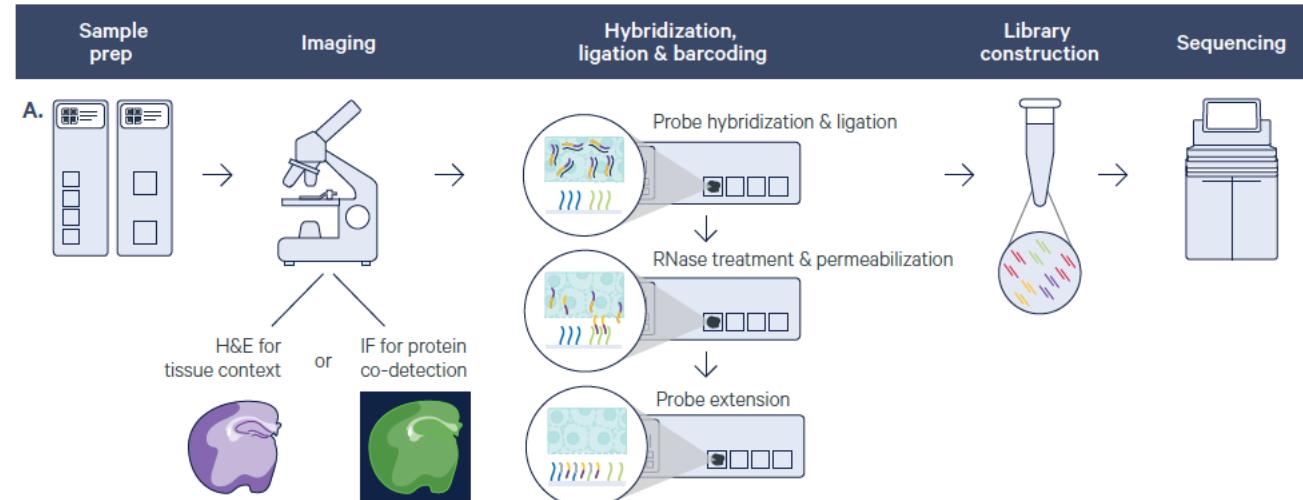
a preliminary assessment of the technology before acquiring the infrastructure is advisable



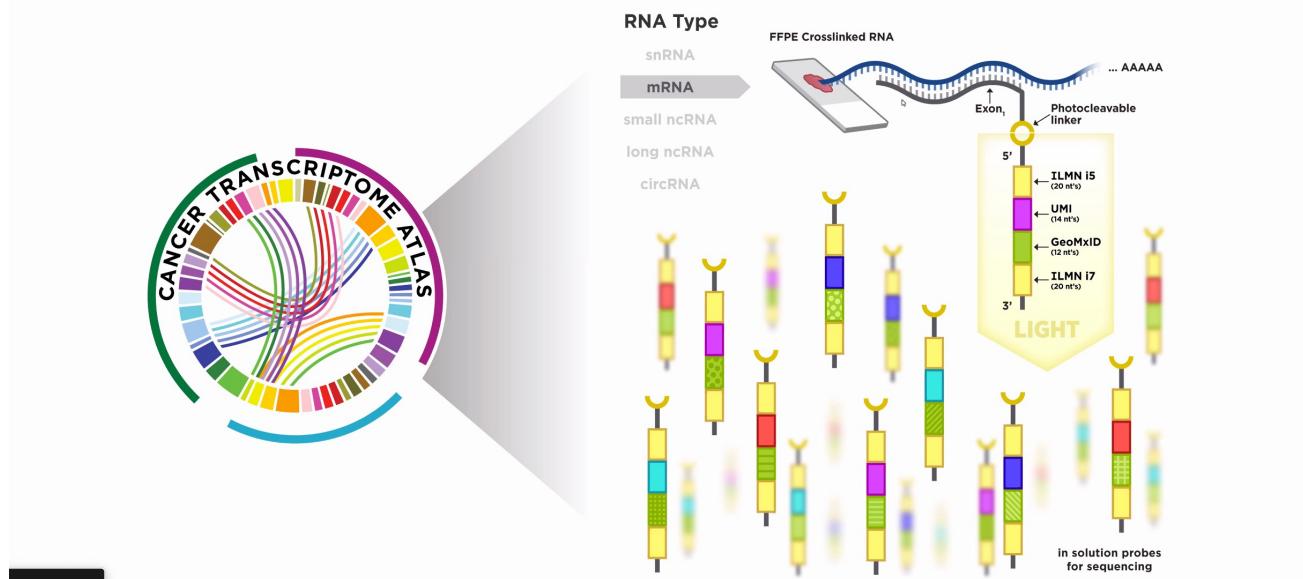
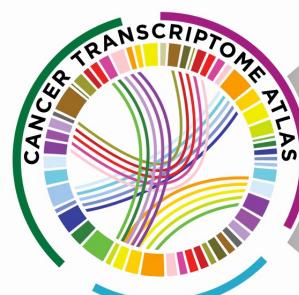
AD HOC ANALYTICAL TOOLS REQUIRED

exploitation of a technology depends mostly on the development of proper analytical tools

SPATIALLY-RESOLVED OMICS IN RNA SEQUENCING

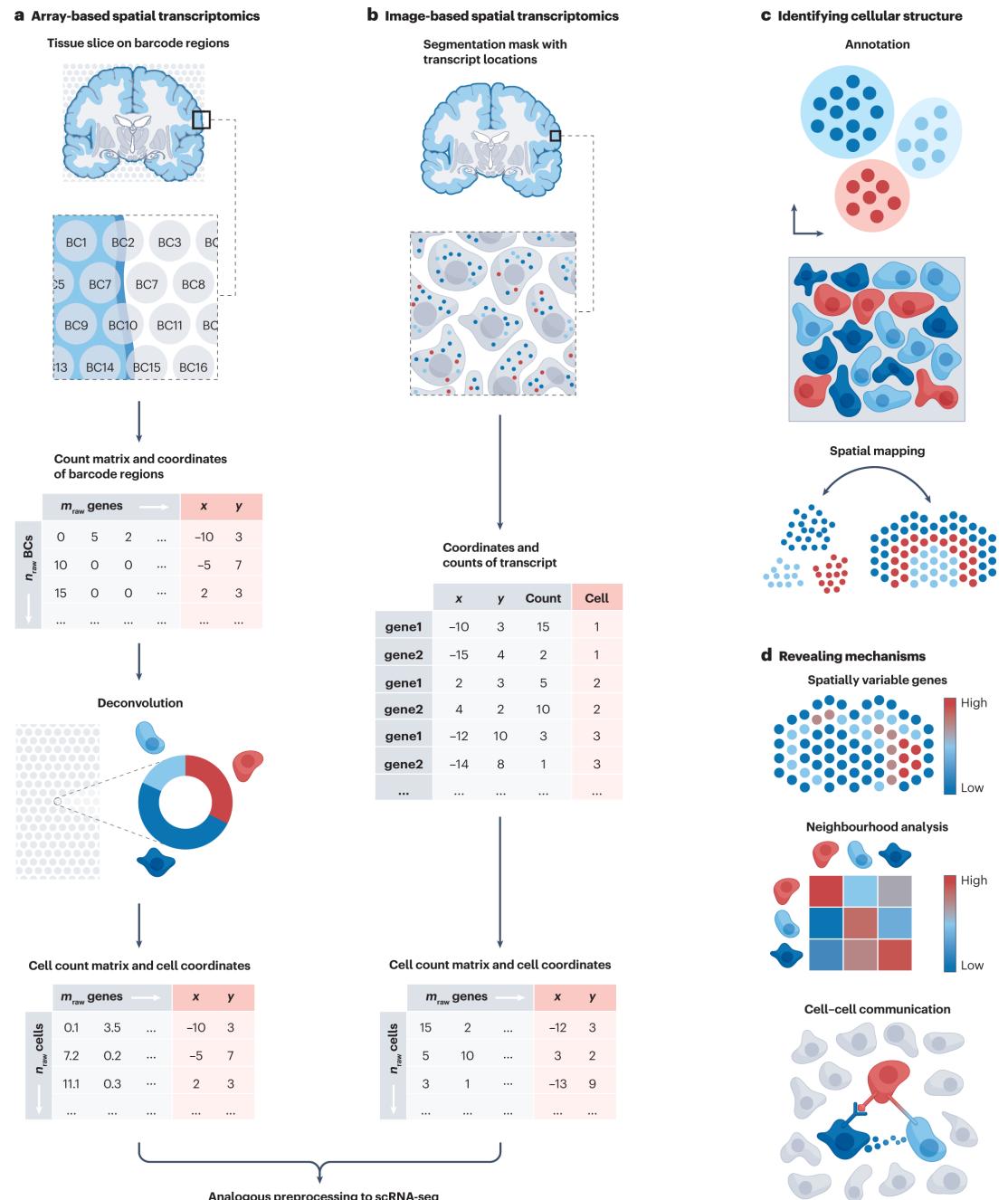


GeoMx DSP High-plex RNA Chemistry: Sequencing of Photocleaved Barcodes



SPATIALLY-RESOLVED OMICS IN RNA SEQUENCING

Overview of spatial transcriptomics pre-processing and downstream analysis steps



(F. Theis lab) Heumos, L., Schaar, A.C., Lance, C. *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* (2023). <https://doi.org/10.1038/s41576-023-00586-w>

CONCLUDING REMARKS



- **It's time to move from a GENE-centered to an RNA-centered approach**
 - the functions of the cells, as well as cell identity and properties are specified by regulating the number and the type of transcripts which are generated (a.k.a. the transcriptome)
-
- **The structure and functions of RNA are remarkably complex**
 - Several RNA biotypes exist with specific functions
-
- **Technology can help decipher the complexity**
 - Each approach has pros and cons >> nothing is perfect !
 - Biological question can be answered by a set of orthogonal approaches, including omics
-
- **Multidisciplinary approach is mandatory**
 - Computational Biology is discipline. Bioinformatics is not just a support for research: it actively participates in all the steps of a research project