



Universidade de São Paulo - USP  
Faculdade de Medicina de Ribeirão Preto  
Curso de Ciências Biomédicas - Turma IX  
RCB0300 – Tópicos em Biotecnologia III



# THE HITCHHIKERS' GUIDE TO RNA SEQUENCING AND FUNCTIONAL ANALYSIS

**O guia dos mochileiros para sequenciamento de  
RNA e análise funcional**

Audrey Beatriz Watanabe do Valle Queiroz

Emilly Regina Ramos

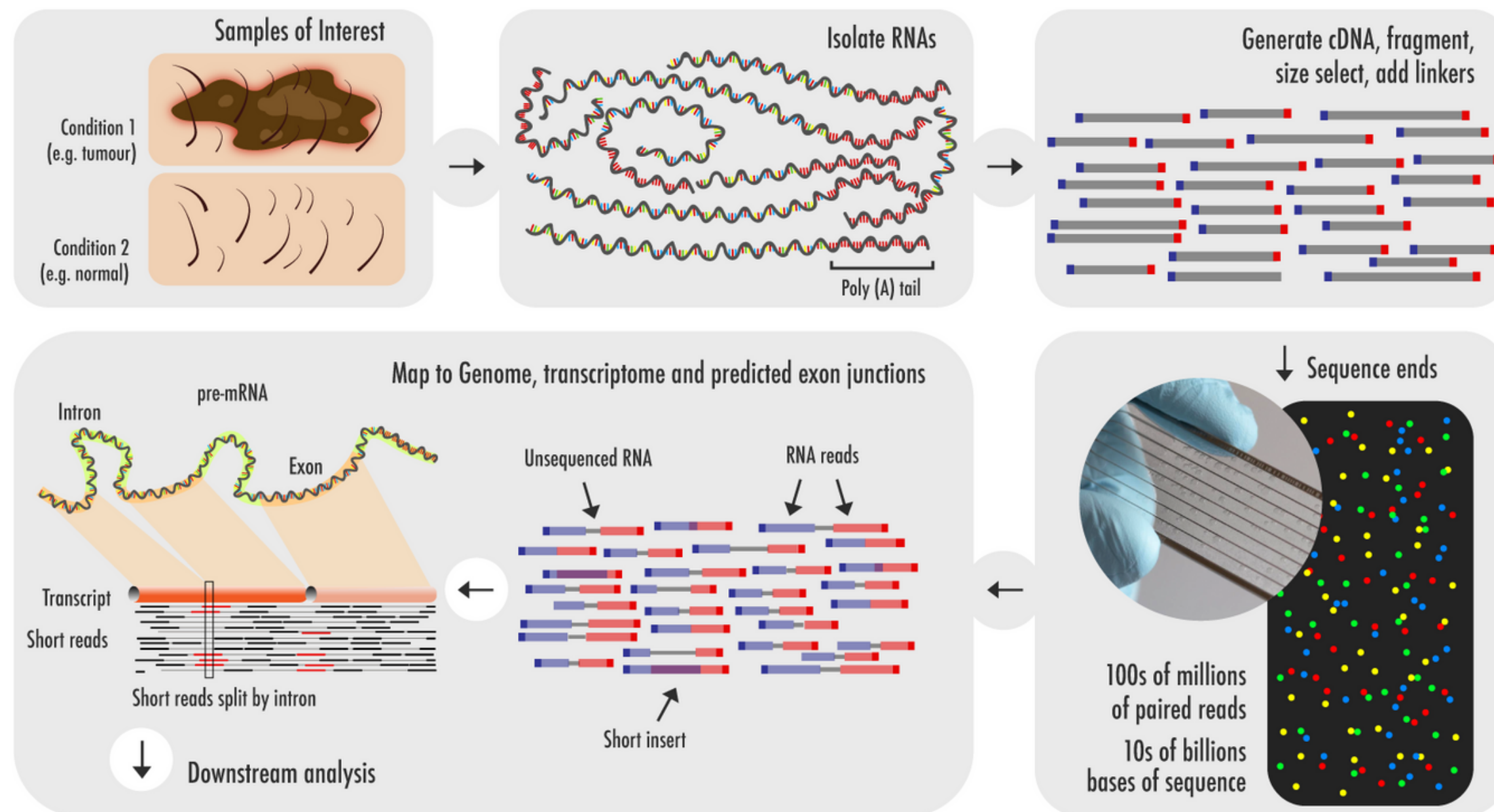
Thuany Giovana Pereira Daniel

# SUMMARY

- Introduction
- RNA-Seq steps
- Sequencing and analysis of non-coding RNAs
- Recent developments
- Evaluation
- Conclusions
- Key Points

# INTRODUCTION

The paper is a guide to properly chose the right tools and analyse the results of a RNA sequencing (RNA-Seq) read



Different methods can lead to different results and conclusions

It's important to conduct comprehensive comparative analyses and justify specific study choices

# RNA-SEQ STEPS

NGS-based RNA-Seq analysis

**1 - Read alignment**

**2 - Read summarization**

**3 - Differential expression analysis**

**4 - Gene set analysis**

**5 - Functional enrichment analysis**

# STEP 1 – READ ALIGNMENT

Alignment is the process of matching reads to specific regions of the genome or transcriptome

- The quality of the read will be measured by the percentage of uniquely aligned genes.
- Effective popular tools are Bowtie, Subread and STAR

## Attention point

- Reads can originate from exons, introns or exon-intron junctions

## There are two forms to dispose raw reads:

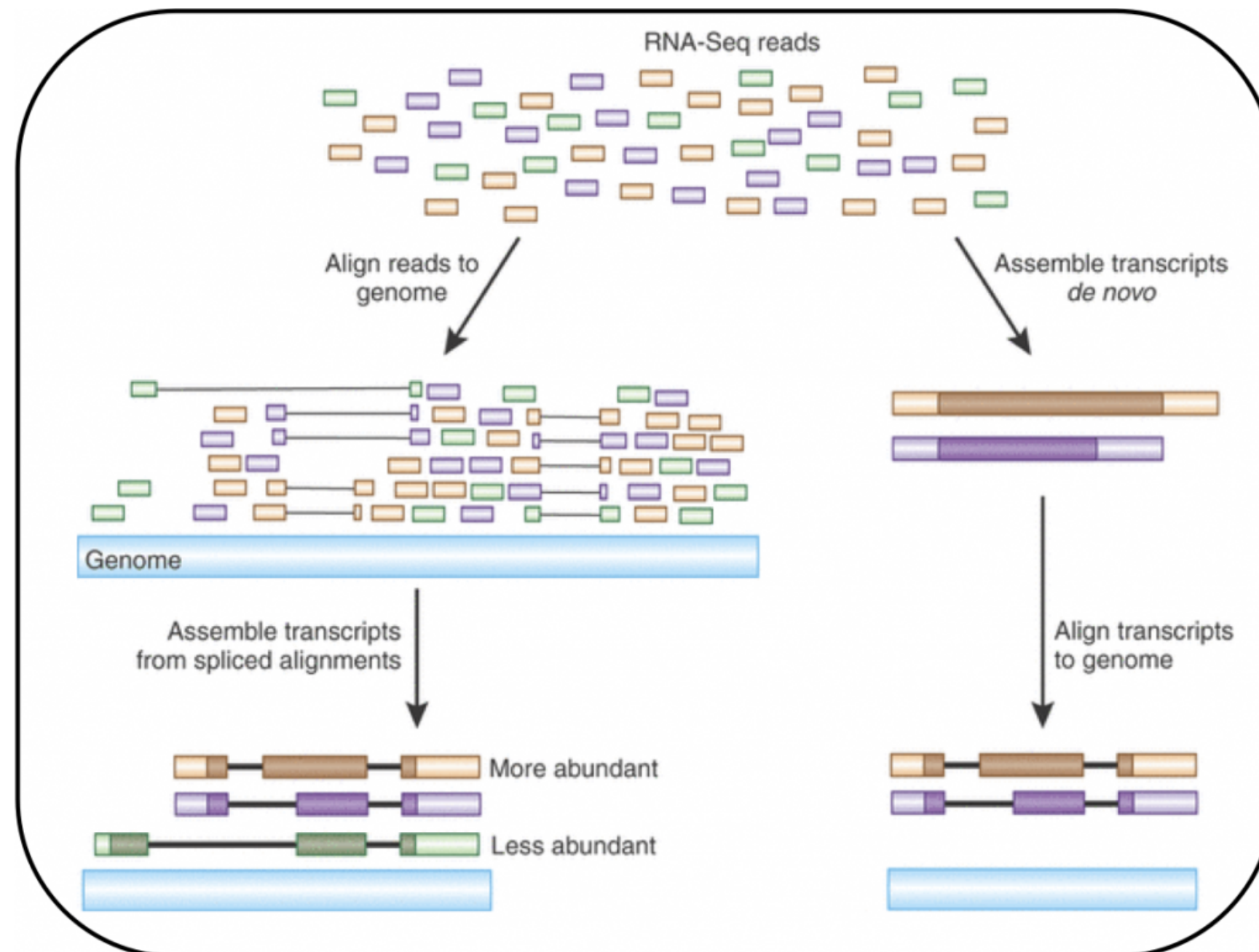
1. Sequence alignment/map (SAM) file
2. Binary alignment/map (BAM) file

The next step in the pipeline is to know which genes or exons they were matched to.



# STEP 1 - READ ALIGNMENT

Preexisting reference genome vs. alignment free techniques



Those algorithms work mostly for long RNAs, but typically fail to accurately quantify expression in low-expressed genes and small RNAs

# STEP 2 - READ SUMMARIZATION

Process of mapping reads to genes, exons, or transcripts and quantifying them into a count matrix

**Computational tools:** TopHat, featureCounts and HTSeq-count

**Annotation databases:** RefSeq, UCSC, Ensembl, and GENCODE

## **Challenges to be a program:**

MUST handle both DNA and RNA sequences, single and pair-ended reads, and accommodate splice variants.

## **Common approaches:**

1. Count reads that match annotated exons - tests splice variants
2. Count at the gene level - any exon within a gene

# STEP 2 - READ SUMMARIZATION

The output is a count matrix indicating the number of aligned reads to each feature in each sample

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

Lafzi, A., Moutinho, C., Picelli, S. et al. (2018)

It will be input data for further analysis such as DE analysis



# STEP 3 - DE ANALYSIS

Differential expression is the process of comparing gene expression levels between two different conditions, treatments, or phenotypes.

## Gene Expression Measurement

**Counts:** Number of reads for each transcript. It's the default choice.

**Other measures:** Such as RPKMs, FPKMs, CPM, and TPM normalize according to gene length or millions of base pairs.

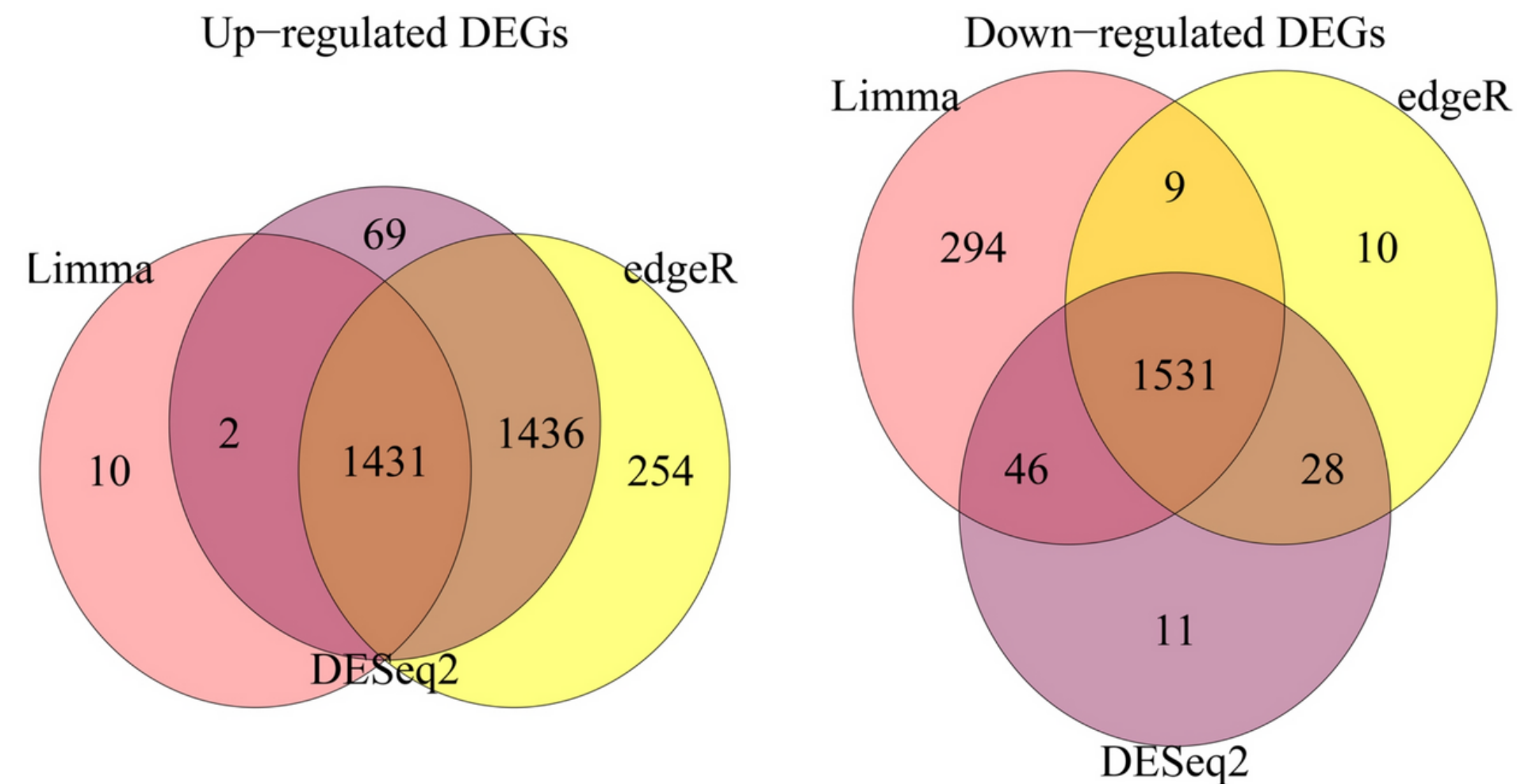
## Challenges:

- Biological variation and statistical testing methods.
- Multiple testing leads to false discoveries, addressed by adjusting for false discovery rate (FDR).
- Gene length bias affects count variance and test power.

# STEP 3 - DE ANALYSIS

## Statistical Approaches:

- Parametric tests (e.g., t-test) and non-parametric tests (e.g., Mann-Whitney) perform poorly with RNA-Seq data.
- Specialized methods like DESeq2 and edgeR are developed for RNA-Seq DE analysis.



# STEP 4 - GENE SET ANALYSIS

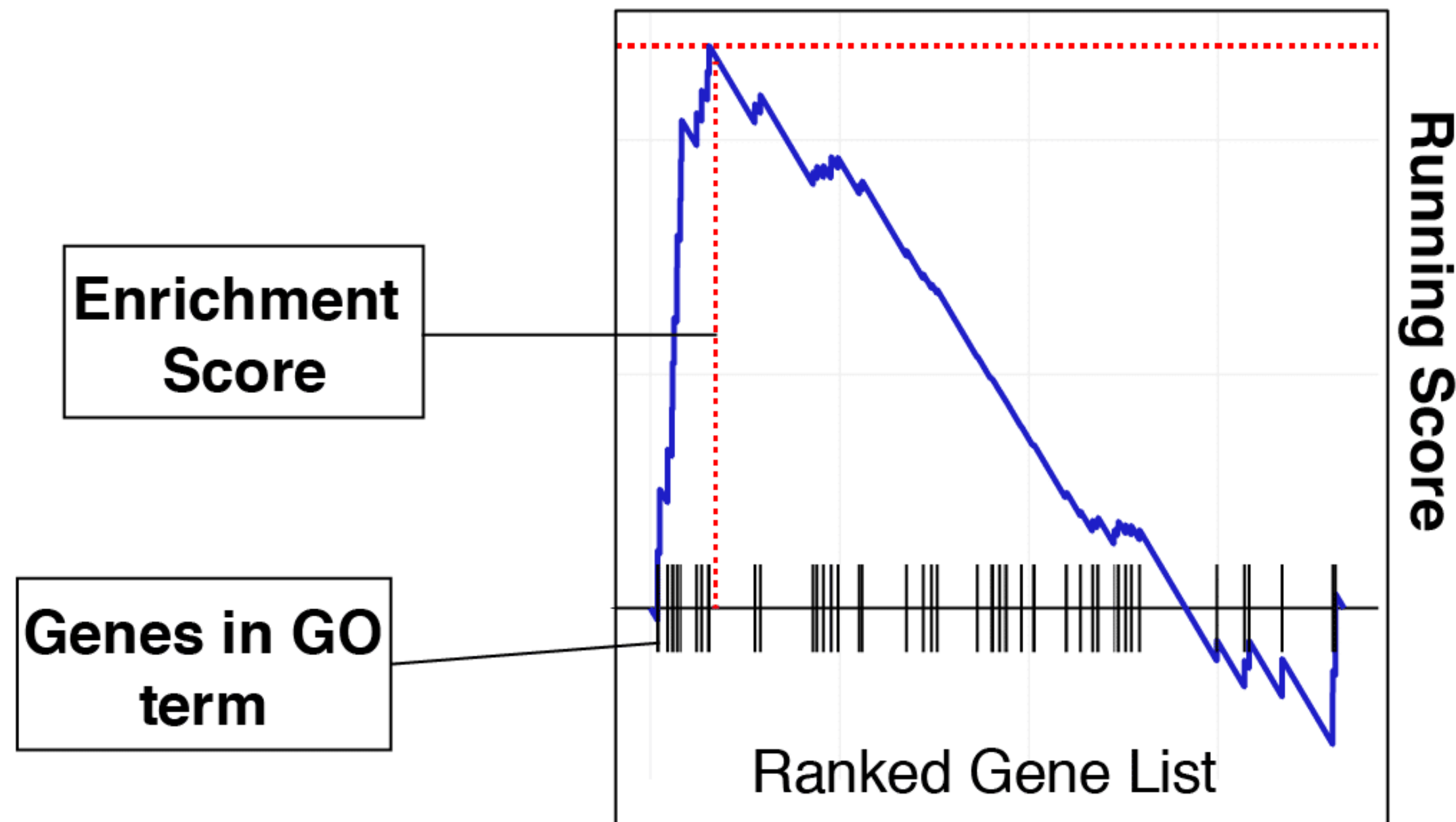
Gene sets are groups of genes related by functionality, participation in signaling pathways, or similar expression patterns.  
Ex: Gene Ontology (GO) and KEGG pathways.

**Hypergeometric test:** investigates the over-representation of gene sets in a group of differentially expressed genes (DEGs).

**GOseq and SeqGSA:** address bias by weighting gene statistics according to length.

# STEP 4 - GENE SET ANALYSIS

Gene Set Enrichment Analysis (GSEA):



GSEA determines if a predefined set of genes shows significant consistent differences between study groups.

It creates a ranked list of genes based on DE levels and tests for overrepresentation of gene sets at the top or bottom of the list.

# STEP 5 - FUNCTIONAL ENRICHMENT ANALYSIS

Used to determine whether a particular set of genes or proteins shows statistically significant enrichment for specific biological functions, pathways, or processes.

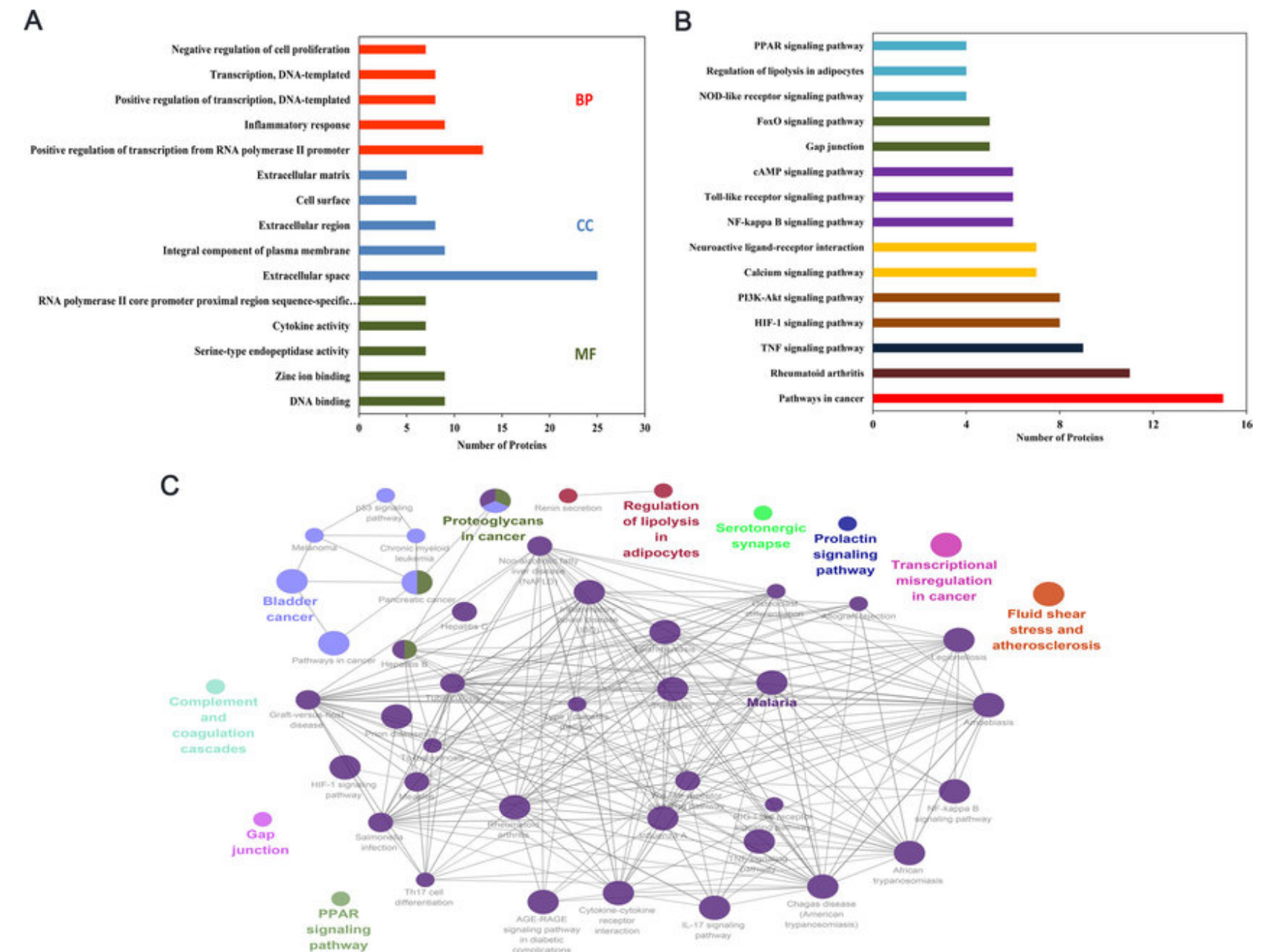
## DAVID - Functional Classification Tool:

(Database for Annotation, Visualization, and Integrated Discovery)

- identifies overlaps
- Organizes genes into biological modules
- Utilizes Fisher's exact test for significance assessment and clustering algorithm

## Ingenuity Pathway Analysis (IPA):

Commercial software for canonical pathway analysis against a manually curated pathway database.





# SEQUENCING AND ANALYSIS OF NON-CODING RNAs (ncRNAs)

Sequencing helped uncover **the significance of ncRNAs** in various physiological mechanisms and regulations during disease pathogenesis

**LONG NON-CODING RNAs (lncRNAs)**

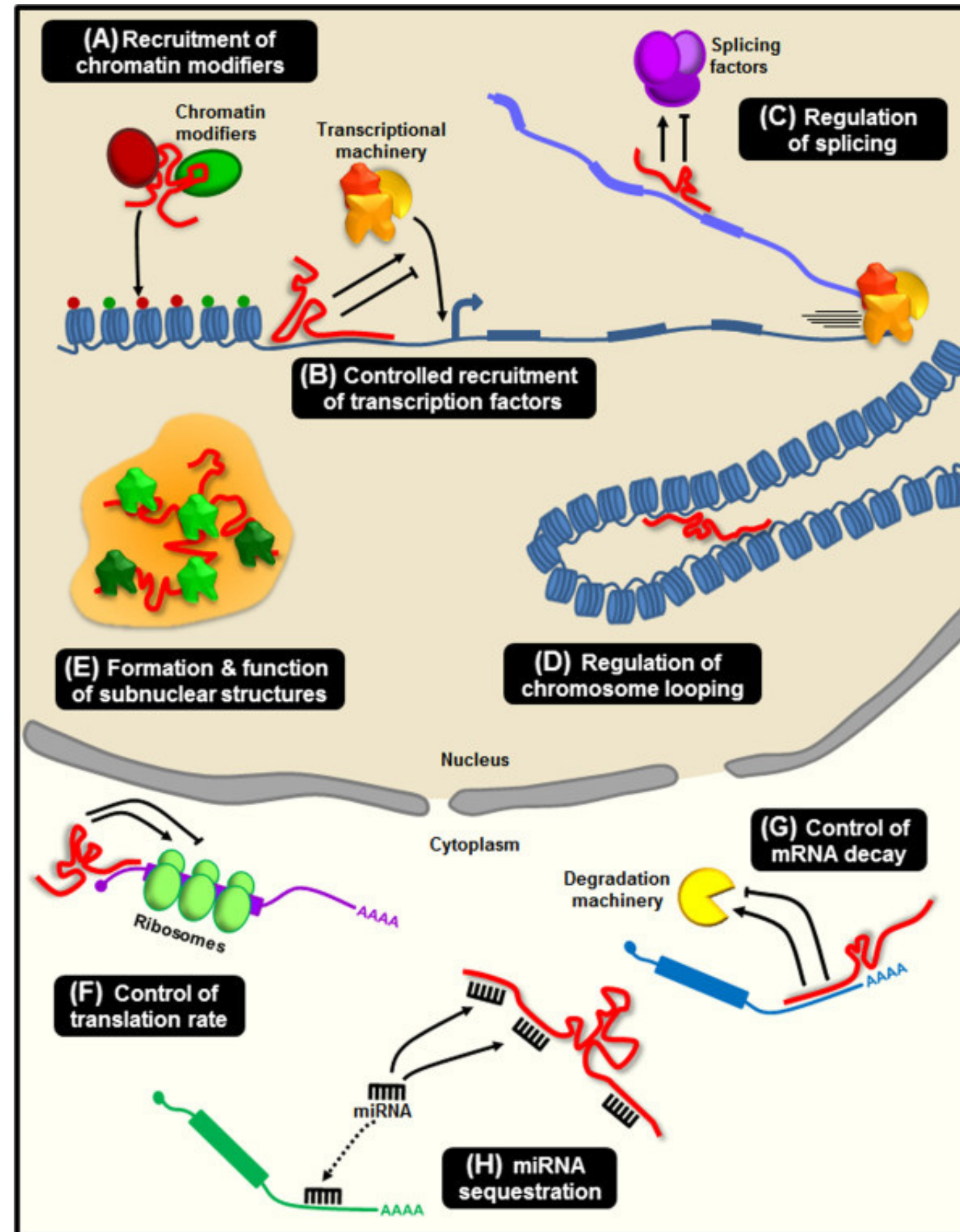
**CIRCULAR RNAs (circRNAs)**

**MICRO RNAs (miRNAs)**



# LONG NON-CODING RNAs (lncRNAs)

- lncRNAs are larger than 200 nucleotides
- Sequencing Methods
  - PolyA-selected or
  - Stranded ribosomal RNA (rRNA)-depleted libraries from total RNA samples with high sequencing depth ( $\geq 30$  million)



# LONG NON-CODING RNAs (lncRNAs)

## Main Tools

### LNCipedia

The screenshot shows the LNCipedia website homepage. The header includes the LNCipedia logo (version 5.2), navigation links for Search, Download, and About, and a Genome dropdown menu. The main content area features a green banner with the text "A comprehensive compendium of human long non-coding RNAs" and a background image of a DNA double helix. Below the banner, there are three columns of information: "Data" (LNCipedia is a public database for long non-coding RNA (lncRNA) sequence and annotation. The current release contains 127,802 transcripts and 56,946 genes.), "Literature" (Currently, LNCipedia offers 2,482 manually curated lncRNA articles. Recently added literature), and "Tools" (LNCipedia comes with some helpful tools: IGV integration, UCSC trackhub).

### LNCbook

The screenshot shows the LncBook 2.0 website homepage. The header includes the CNCB and NGDC logos, navigation links for Databases, Tools, Standards, Publications, and About, and a search bar. The main content area features a blue banner with the text "LncBook 2.0 Integrating human long non-coding RNAs with multi-omics annotations" and a background image of a DNA double helix. Below the banner, there is a search bar with a magnifying glass icon and a search button. Below the search bar, there is a text box stating "LncBook accommodates a high-quality collection of human lncRNA genes and incorporates their abundant annotations at different omics levels, thereby enabling users to decipher functional signatures of lncRNAs in human diseases and different biological contexts." Below this text box, there are three columns of information: "Multi-omics Annotations" (Conservation: Conservation Features across 40 Vertebrates; Variation: 959,138 Disease/trait-associated Variants; Methylation: DNA Methylation Profiles in 16 Diseases).

## Other Tools

lncRScan-SVM

lncFinder

iSeeRNA

linc-SF

lncRNADisease 2.0

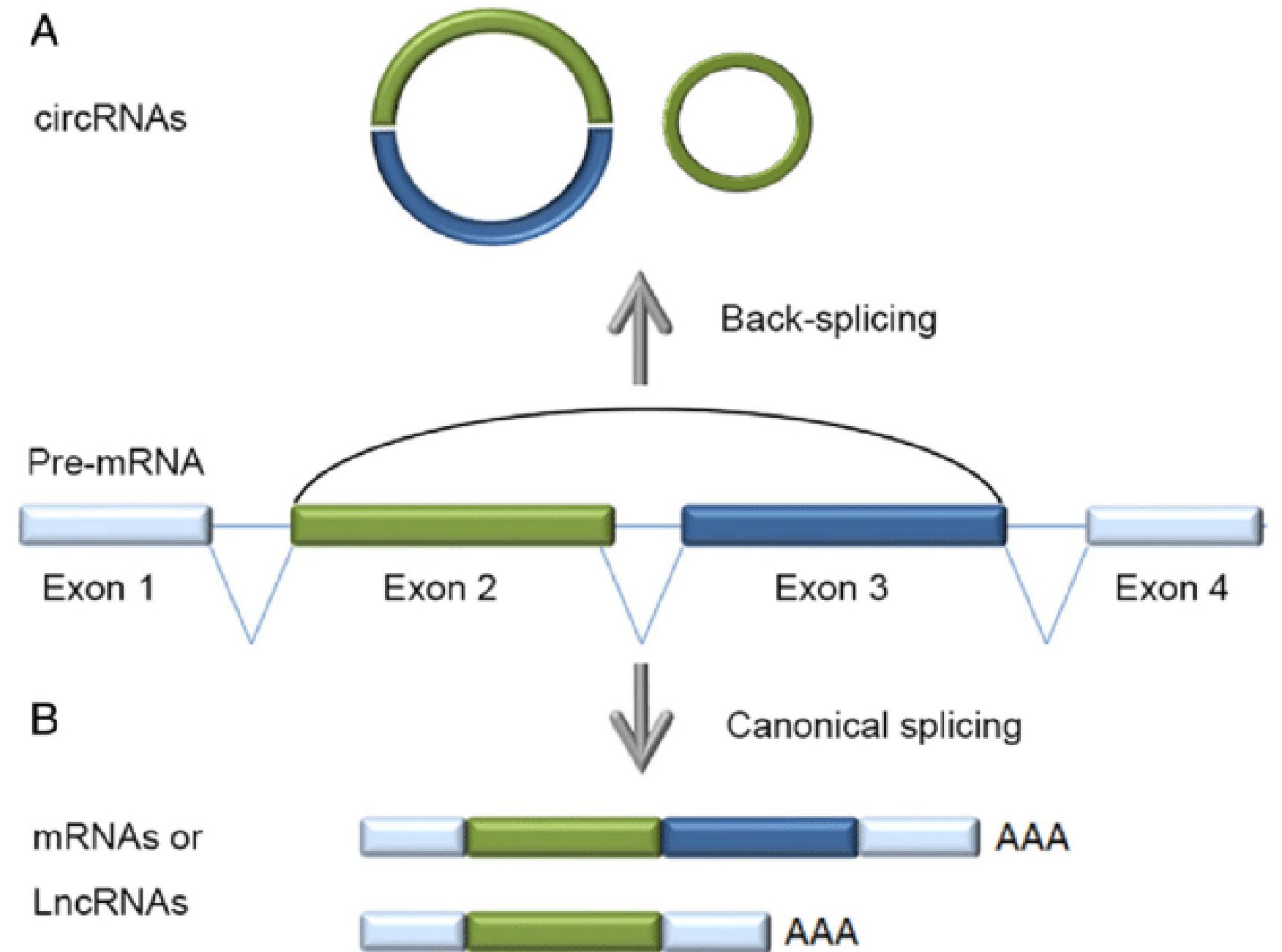
lnc2Cancer 3.0

NNLDA

IDLDA

# CIRCULAR RNAs (circRNAs)

- Covalently closed and non-polyadenylated circular transcripts
- Formed by either exon or protein driven back-splicing mechanisms
- Sequencing Methods
  - Extra step: Treat the samples with RNase R for circRNA enrichment or
  - Nanopore long-read sequencing with a modified RNA-Seq sample preparation protocol



(CHUN YING YU; KUO, 2019)



# CIRCULAR RNAs (circRNAs)

## Main Tools

### circBase

The screenshot shows the homepage of circBase (circbase.org). The page features a navigation menu with options: home, list search, table browser, blat, downloads, and help. A search bar is prominently displayed in the center. Below the search bar, there is a 'RECENT UPDATES' section listing several publications from July 2017 to May 2015, including works by Maass et al., Ashwal-Fluss et al., Ivanov et al., and Rybak-Wolf et al. A 'Welcome to circBase' message is also present, mentioning that thousands of circular RNAs have been shown to be expressed in eukaryotic cells.

### CIRCpedia

The screenshot shows the database profile page for CIRCpedia on the ngdc.cnpc.ac.cn website. The page includes a search bar with a 'Search' button and a list of search filters: human, SARS-CoV-2, lncRNA, single cell, spatial omics, immune, Oryza sativa, and European Bioinformatics Institute China. The main content area is titled 'Database Profile' and 'CIRCpedia'. It provides 'General information' such as the URL (http://www.picb.ac.cn/momics/circpedia/), full name (A database of circular RNAs), description (CIRCpedia v2 is an updated comprehensive database containing circRNA annotations from over 180 RNA-seq datasets across six different species), and year founded (2016). A 'Ranking' section shows a 'TOTAL RANK' of 155 and '605 CITATIONS'.

## Other Tools

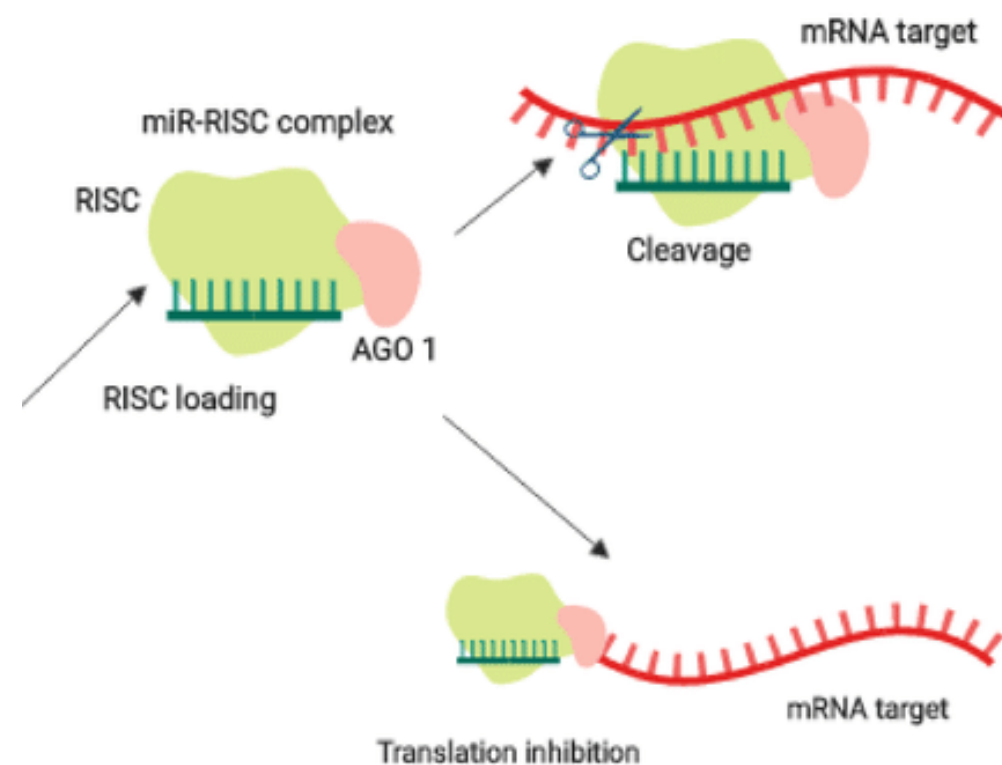
CIRCExplorer2  
find\_circ

CIRCInteractome

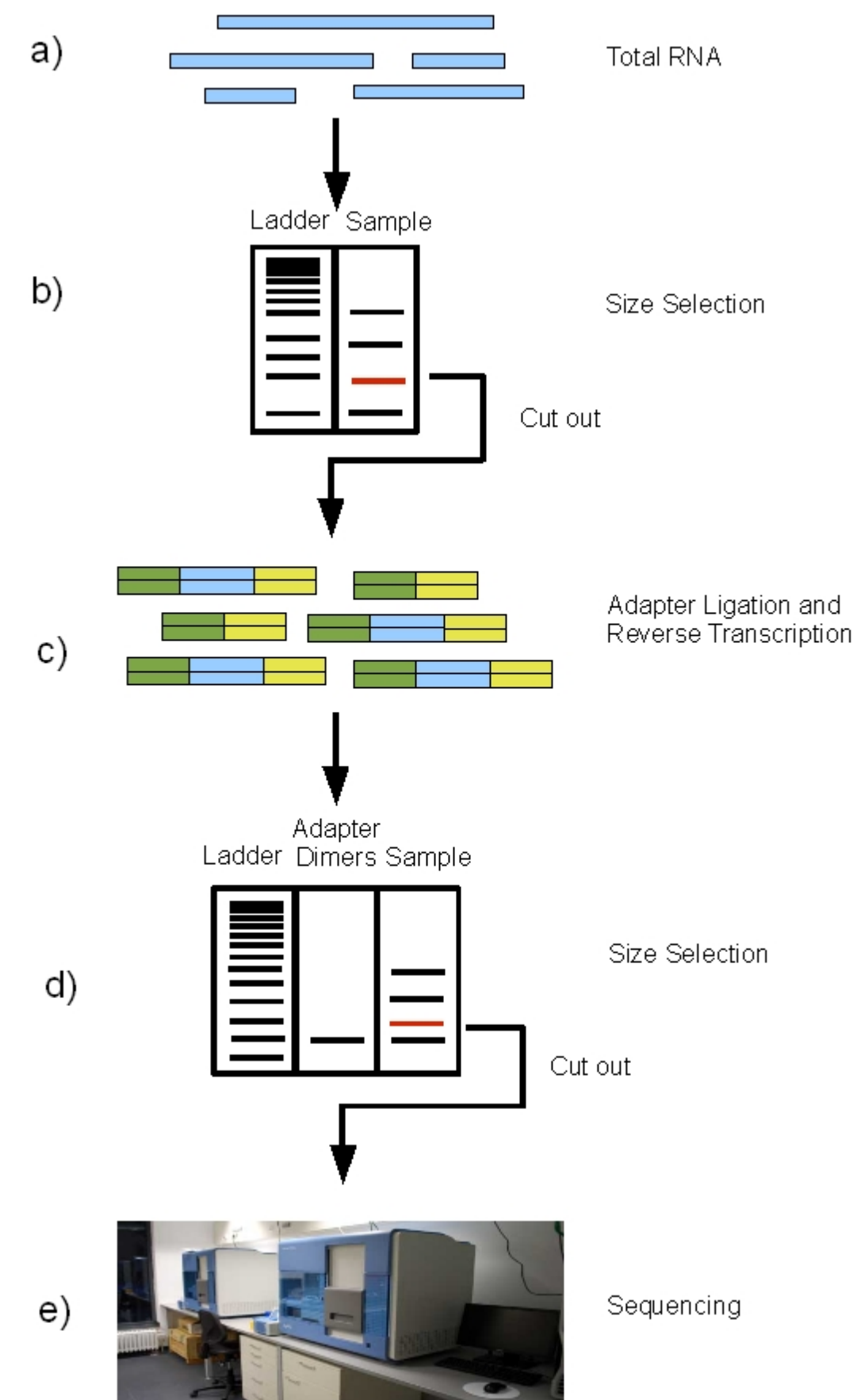
Circ2Disease  
circRNADisease  
Circ2Traits

# MICRO RNAs (miRNAs)

- miRNAs are 20-23 nucleotides
- Sequencing Methods
  - Library construction with reverse transcription → size exclusion gel or size selection magnetic beads → cDNA → PCR amplification



Adaptado: (CHAUDHARY; GROVER; PRAKASH CHAND SHARMA, 2021)

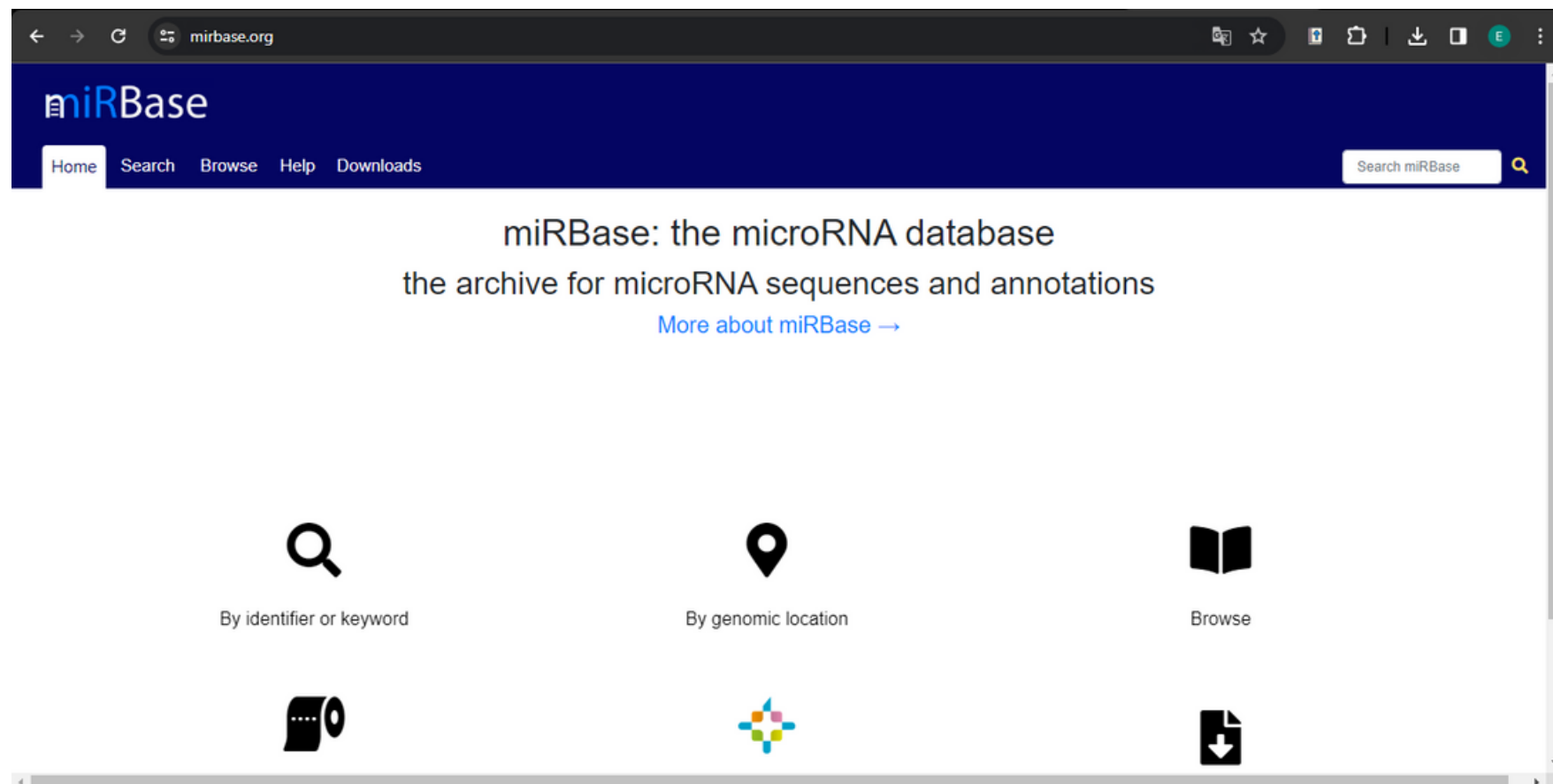


(MOTAMENY et al., 2010)

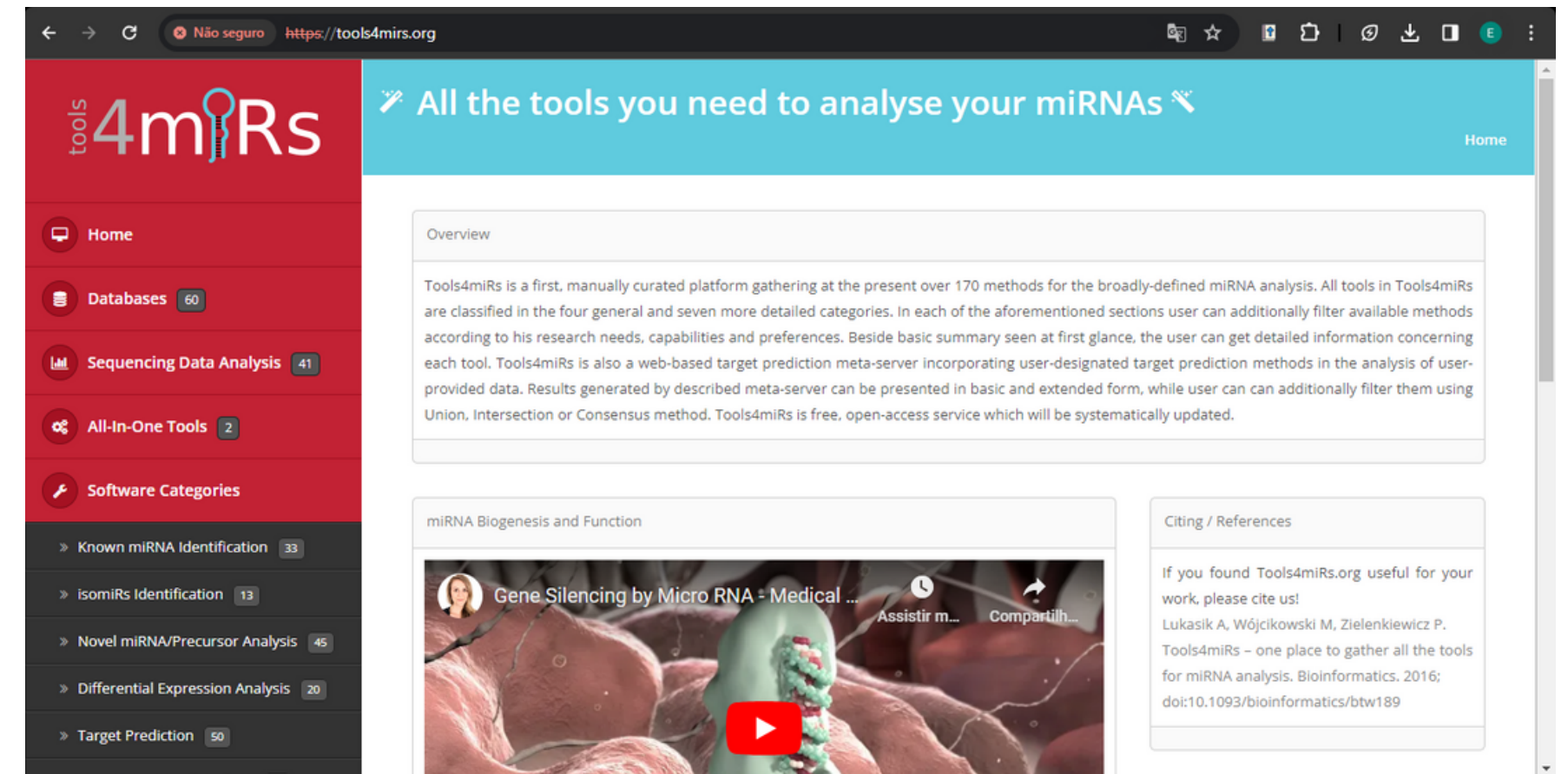
# MICRO RNAs (miRNAs)

## Main Tools

### miRBase



### Tools4miRs



### Other Tools

miRMaster 2.0  
CAP-miRSeq

miRDeep2

isomiR2Function  
miRge

isomiRs - variations  
with respect to a  
reference sequence



# RECENT DEVELOPMENTS

**METATRANSCRIPTOMICS**

**MACHINE LEARNING APPROACHES**

**MULTI-OMICS WORKFLOW**

# METATRANSCRIPTOMICS

Microbiome is a characteristic microbial community occupying a reasonable well-defined habitat which has distinct physio-chemical properties”

- Microbial metabolism can contribute to host health
- Metatranscriptomics capturing the transcripts of a whole microbial community
- Bioinformatics Workflow is necessary
- Same RNA-seq steps
  - Requires more computational resources
  - Transcript abundance - models have been proposed
    - Data normalization, for example
  - Data analysis tools
    - SqueezeMeta, IMP, MUFFIN, MetaPro...

# MACHINE LEARNING APPROACHES

- RNASeq problem: large  $p$ , small  $N$
- The power of ML in metatranscriptomics is still an under-researched topic
- Traditional approaches has been used for DEGs discovery
- Build a predictive model based on metatranscriptomic data is desirable
- It is a black box model without understanding of biological processes

# MACHINE LEARNING APPROACHES

**Table 1.** Applications of ML/DL in (meta-)transcriptomics

Method	Application	Source code	References
GA/kNN	Identification of differentially expressed genes	<a href="https://www.niehs.nih.gov/research/resources/software/biostatistics/gaknn">https://www.niehs.nih.gov/research/resources/software/biostatistics/gaknn</a>	[129]
GA/kNN, gradient boosting E-M algorithm	Identification of differentially expressed genes  De novo assembly of meta-transcriptomics data	NA  <a href="https://sourceforge.net/projects/dnpipe">https://sourceforge.net/projects/dnpipe</a>	[130]  [131]
Logistic regression w/ L2 regularization	Identification of predictive microbial taxa and KOs from meta-transcriptomics	NA	[132]
Random forest, gradient boosting	Construction of predictive models from meta-transcriptomics	<a href="https://github.com/armbrustlab/trophic-mode-ml">https://github.com/armbrustlab/trophic-mode-ml</a>	[133]
CNN/Grad-Cam	Identification of marker genes and classification of cancer types	NA	[134]
CNN/Grad-Cam	Identification of marker genes and classification of oral cancer types	NA	[135]
CNN/saliency maps	Identification of marker genes and classification of cancer types	<a href="https://github.com/chenlabgcrci/CancerTypePrediction">https://github.com/chenlabgcrci/CancerTypePrediction</a>	[136]
Deep NN	Alternative splicing analysis	<a href="https://github.com/Xinglab/DARTS">https://github.com/Xinglab/DARTS</a>	[137]
CNN and DeepLIFT	Regulatory mechanisms identification	<a href="https://github.com/stasaki/DEcode">https://github.com/stasaki/DEcode</a>	[138]
Mixing observation Autoencoder	Data augmentation Cell content inference from bulk RNA-Seq data (which is typically done w/ scRNA-Seq data)	NA <a href="https://github.com/xindd/DCNet">https://github.com/xindd/DCNet</a>	[139] [140]
ICA	Identification of novel regulons	<a href="https://github.com/avsastory/modulome-workflow">https://github.com/avsastory/modulome-workflow</a>	[141]

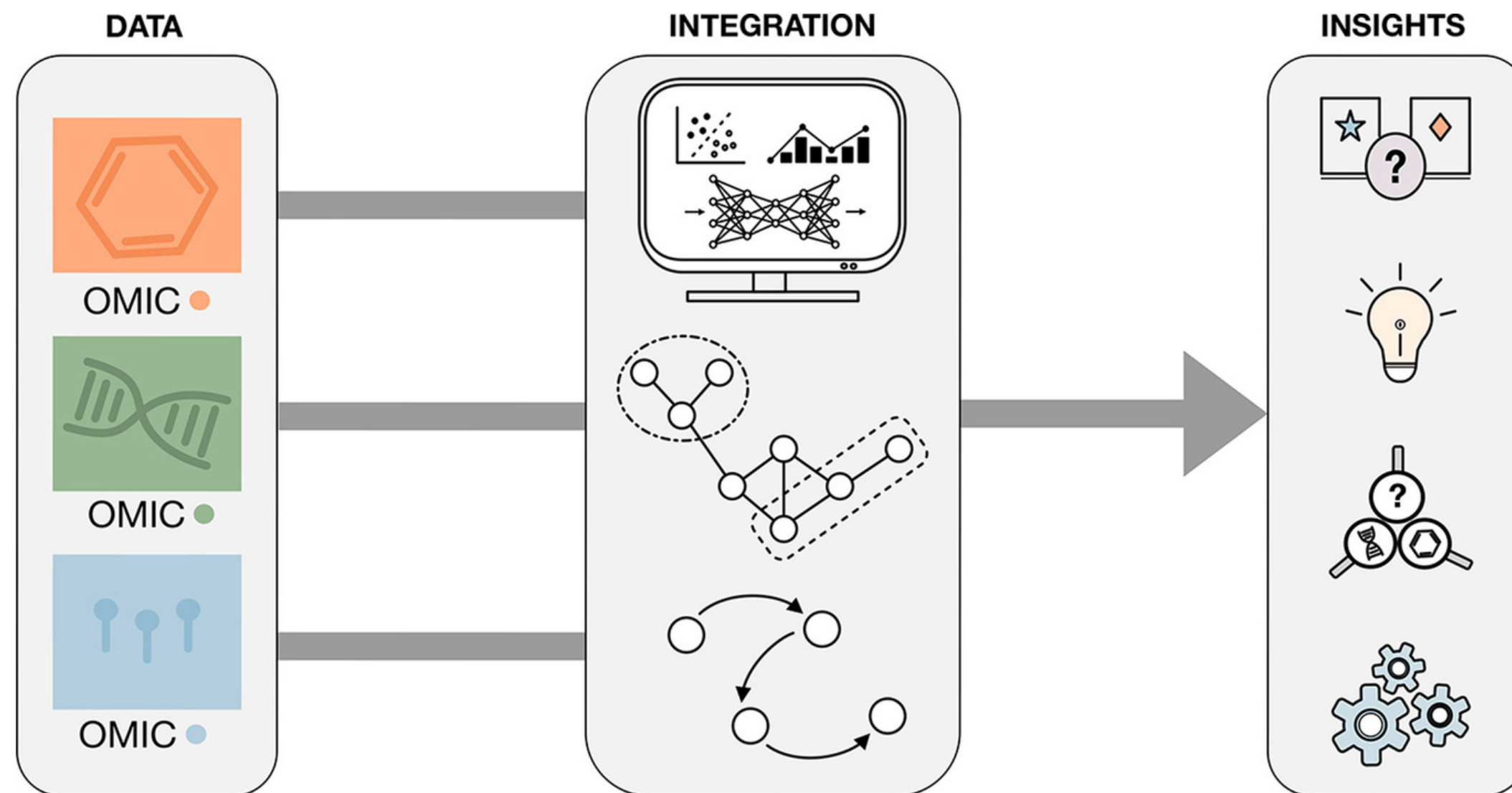
CNN: Convolutional neural network; E-M: expectation-maximization; GA: genetic algorithm; ICA: independent component analysis; kNN: k nearest neighbors; NN: neural network.

# MULTI-OMICS INTEGRATION

High performance technology for biological systems

“The multi-omics data, considered collectively, reveals complex biological connections that were not deemed significant in the individual omics analysis”

DR. THOMAS HARTUNG,  
UNIVERSIDADE JOHNS HOPKINS



## Multi-omics workflow

- data collection
- data pre-processing
- data integration
- data analysis

	xMWAS	PaintOmics 3	TIMEOR	Mergeomics 2	OmicsAnalyst	BIOMEX	miodin	3Omics	multiGSEA
Implementation	R, online	Online	Online, command line	Online	Online	R	R	Online	R
Functionality									
Pre-processing			X	X	X	X	X		
Data integration	X		X		X	X	X	X	
Network analysis	X	X	X	X	X			X	
Enrichment analysis		X	X	X	X	X		X	X
Pathway analysis		X	X	X		X		X	X
Time series analysis			X				X		
Visualization	X	X	X	X	X	X	X	X	
Accepted-omics data (besides transcriptomics)									
Metabolomics	X	X		X	X	X		X	X
Proteomics	X	X		X	X	X	X	X	X
Genomics	X	X		X			X		
Epigenomics				X			X		
Region-based omics <sup>a</sup>		X	X						
Regulatory omics <sup>b</sup>	X	X	X		X				
Reference	[177]	[178]	[179]	[180]	[181]	[182]	[183]	[184]	[185]

<sup>a</sup>ChIP-seq, ATAC-seq or Methyl-seq. <sup>b</sup>miRNAs or other transcription factors.



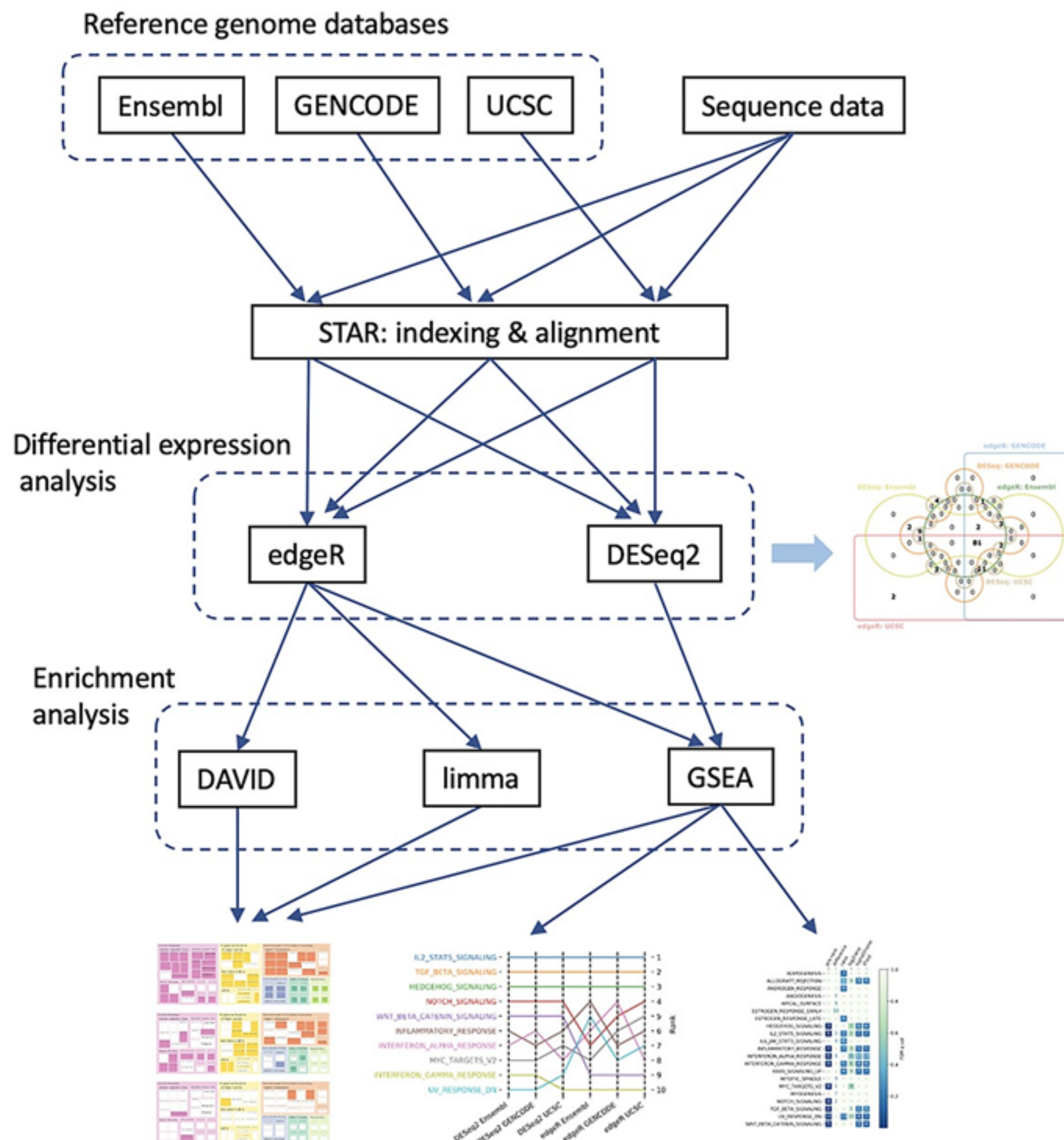
# EVALUATION

Typical RNA-Seq analysis process



**The hitchhikers' guide to RNA sequencing and functional analysis**  
 Abstract. DNA and RNA sequencing technologies have revolutionized biology and biomedical sciences,...

OUP Academic

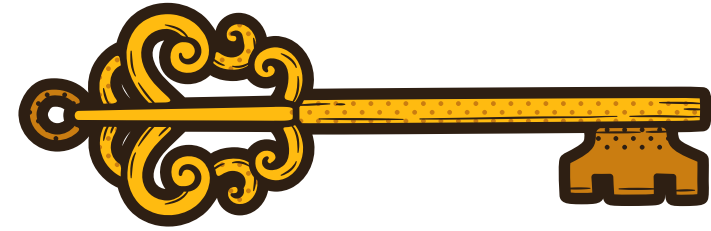


# CONCLUSIONS



Thuany

- The article described the most popular RNA-Seq analysis options instead of providing a gold standard or best practice for RNA-Seq analysis.
- There is no consensus about the best practice of enrichment analysis for a given RNA-Seq experiment.
- The results can be unintentionally impacted by the choice of methods.
- Researchers thus need to cautiously interpret the clinical or biological relevance of the statistically significant features derived from choice of analysis methods.



# KEY POINTS

30

Thuany

- RNA-Seq analysis and software options.
  - Steps: Read alignment, read summarization, differential expression analysis, gene set analysis and functional enrichment analysis.
- RNA-Seq applications, including non-coding RNA analysis and interaction with other technologies.
- Different RNA-Seq results can be obtained depending on the computational method selected.
- The results need to be interpreted and validated with caution.

The background features three vertical stripes on the left side: a wide light red stripe, a narrower teal stripe, and a medium-width light beige stripe. On the right side, there are two rectangular areas filled with a grid of small, light red dots, one in the top right and one in the bottom right.

**THANK YOU**