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

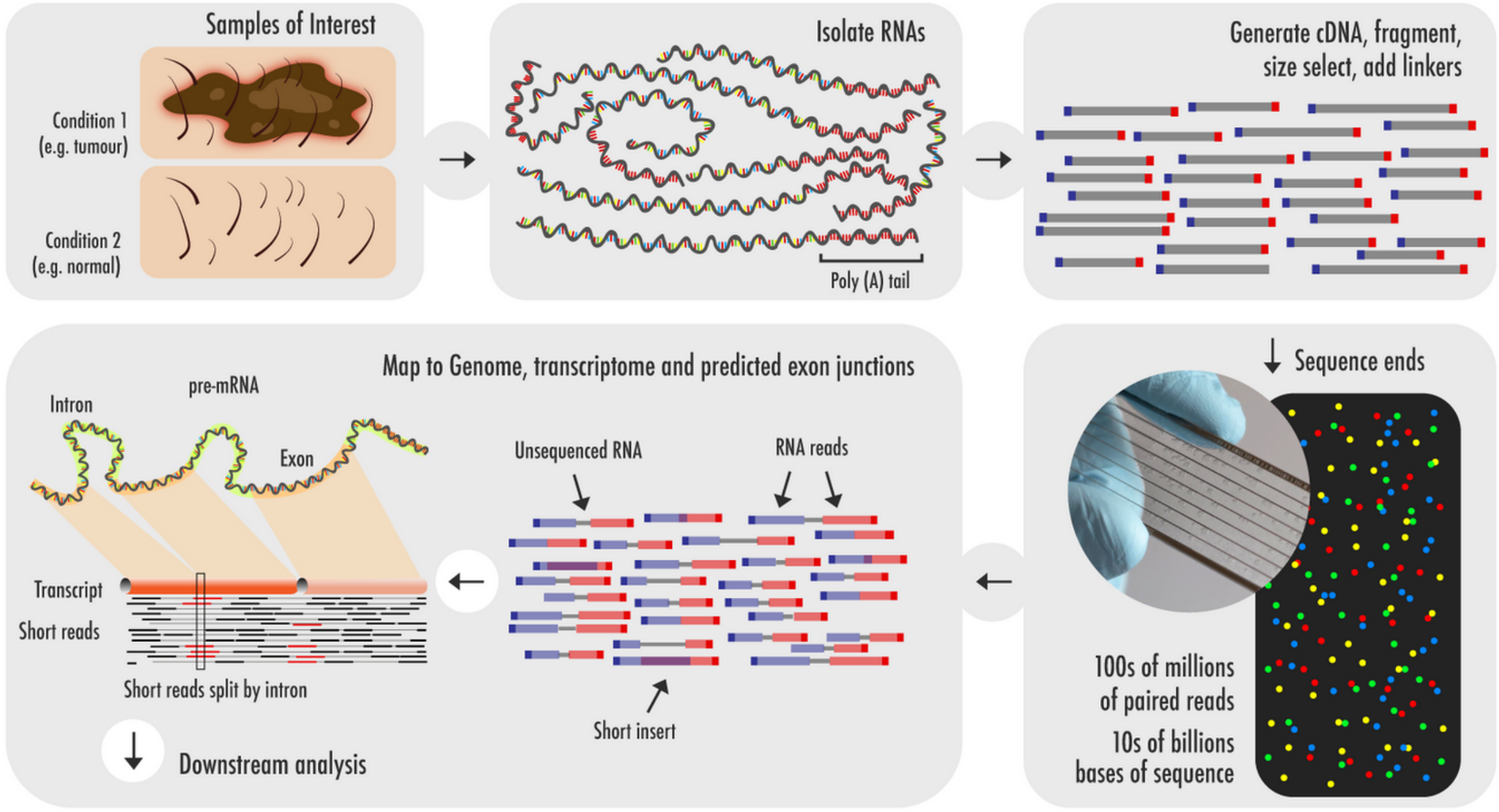
Audrey Beatriz Watanabe do Valle Queiroz
Emilly Regina Ramos
Thuany Giovana Pereira Daniel

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

Technology Networks - Acesso em março de 2024

RNA-SEQ STEPS

NGS-based RNA-Seq analysis

1 - Read alignment

2 - Read summarization

3 - Differencial 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 mesured by the percentage of uniquely aligned genes.
- Efective popular tools are Bowtie, Subread and STAR

Attention point

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

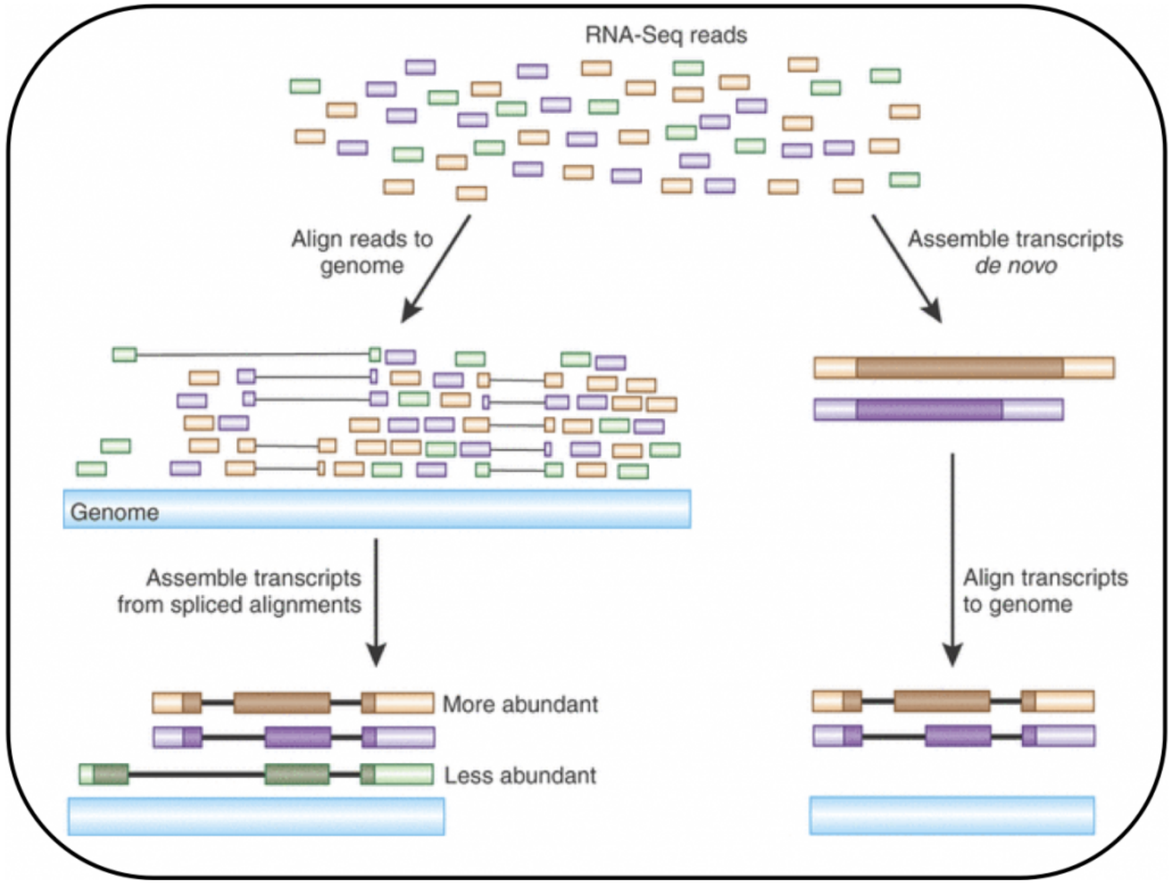
There are two forms to dispose raw reads:

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



Haas BJ and Zody MC. Advancing RNA-seq analysis. 2010

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 approches:

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 defalt choice.

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

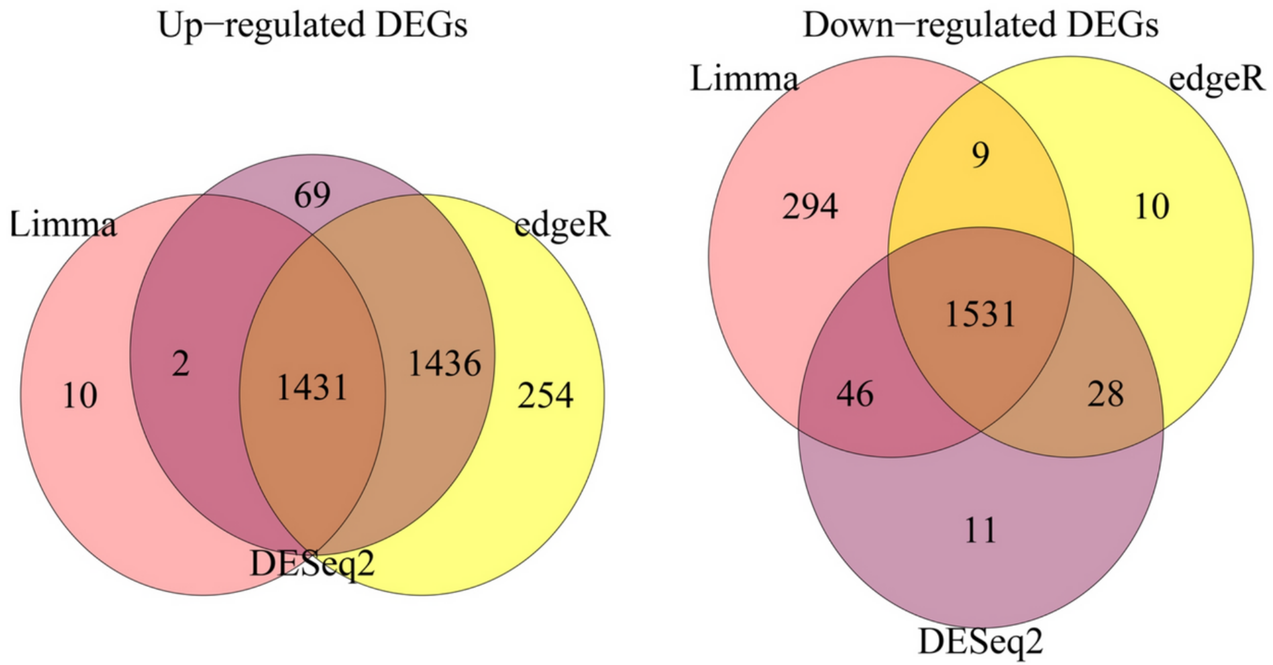
Challenges:

- 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.



Liu, S., Wang, Z., Zhu, R., Wang, F. et Al (2021)

STEP 4 – GENE SET ANALYSIS

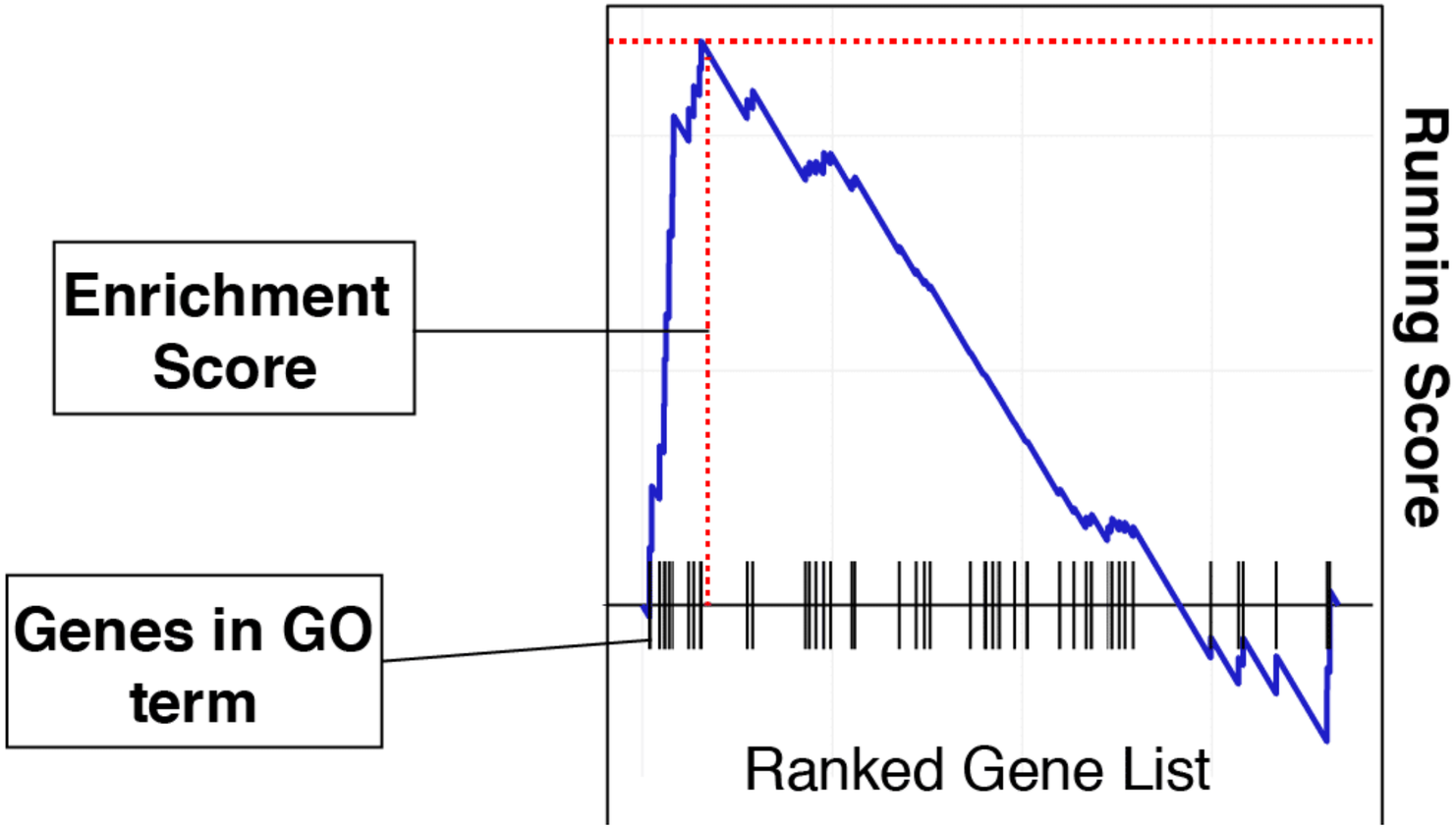
Gene sets are groups of genes related by functionality, participation in signaling pathways, or similar expression patterns.
Ex: Gene Ontology (GO) and KEGG pathways.

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.

www.bioinformaticsbreakdown.com - acesso em 13/03/2023

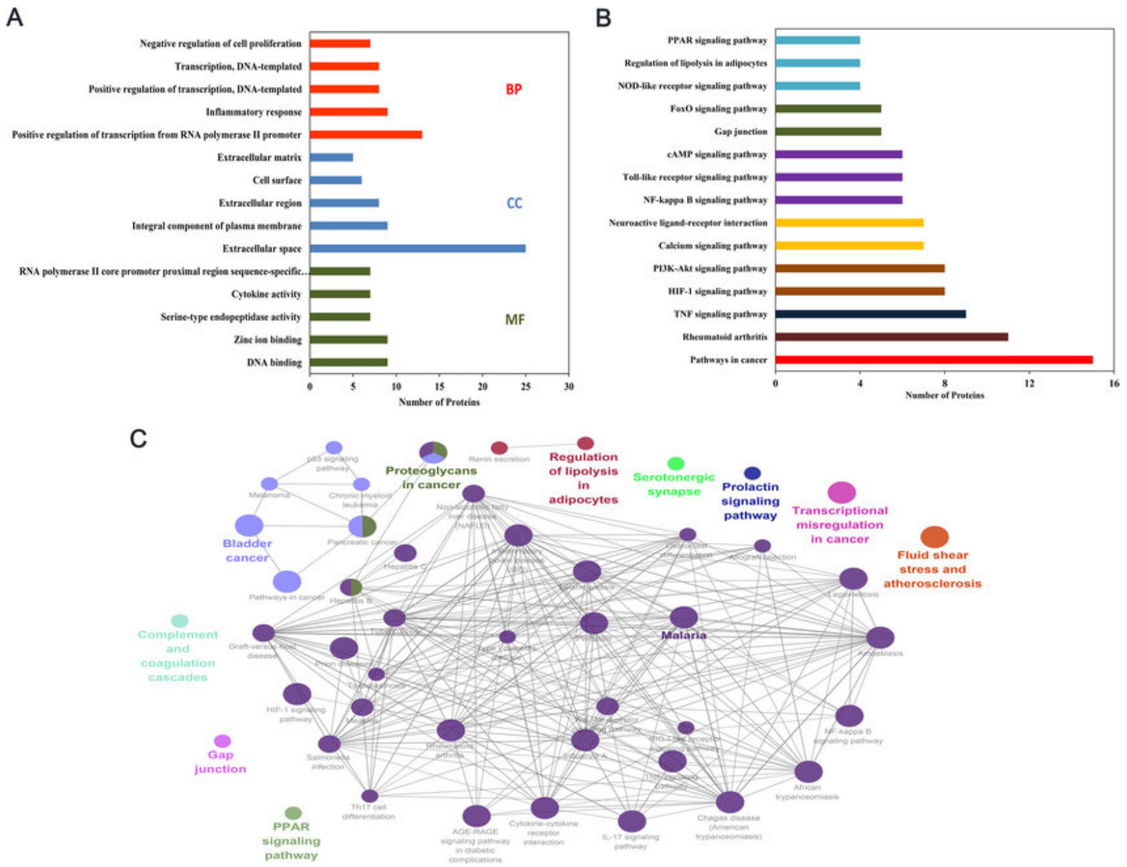
STEP 5 – FUNCTIONAL ENRICHMENT ANALYSIS

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.



Wenjun Wang, Tianlong Liu, Liudi Yang et al. (2019)

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)

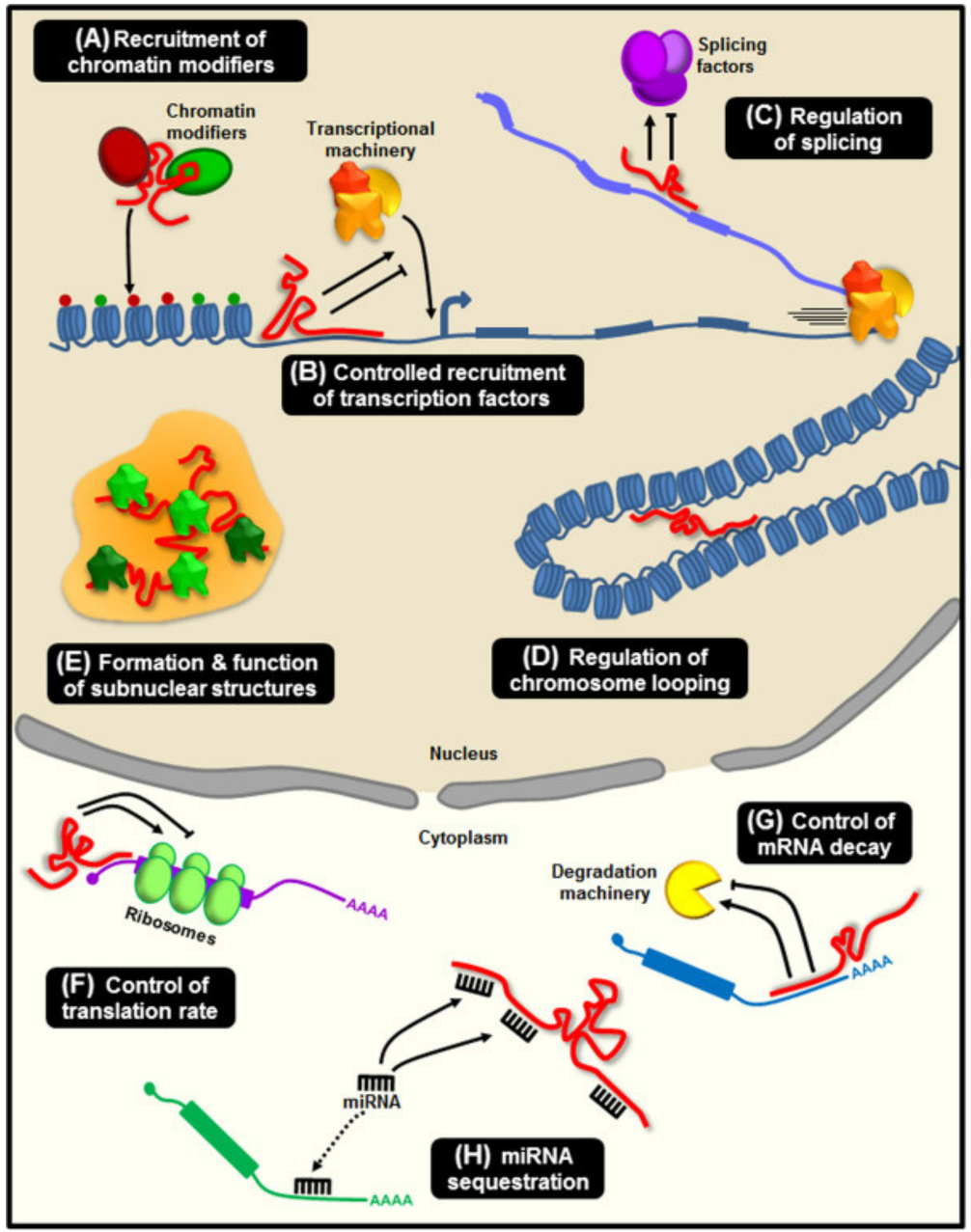
CIRCULAR RNAs (circRNAs)

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 (≥30 million)

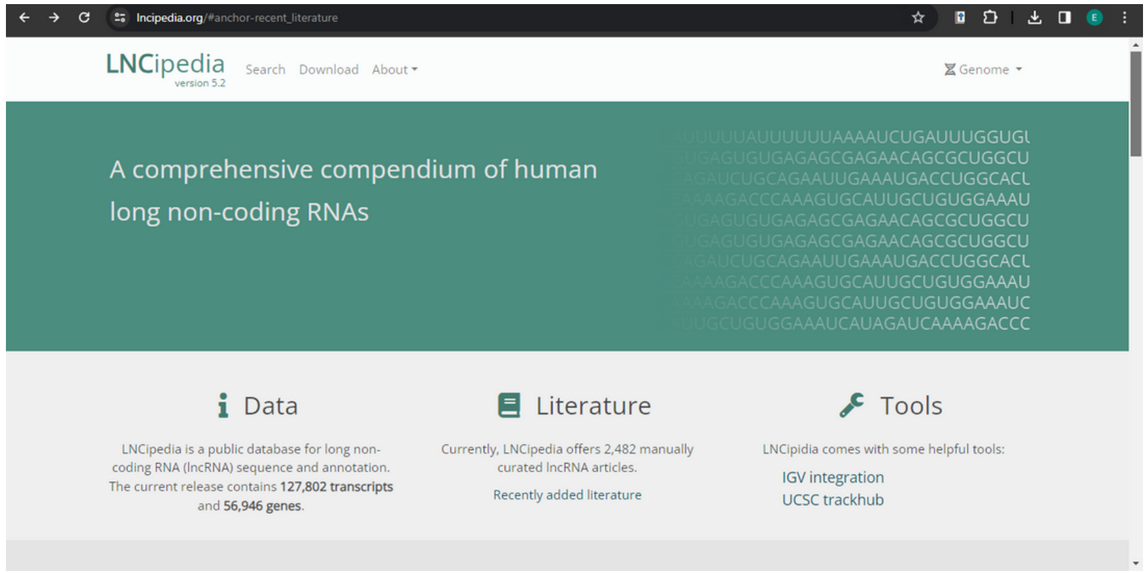


(MARIA VICTORIA NEGUEMBOR; MATHIVANAN JOTHI; DAVIDE GABELLINI, 2014)

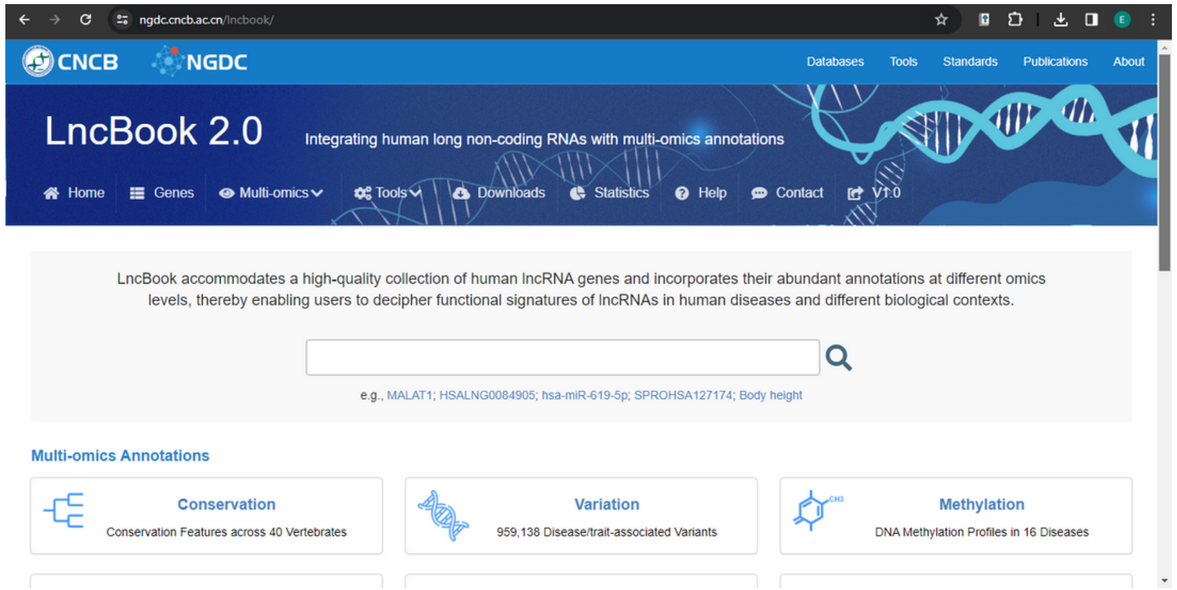
LONG NON-CODING RNAs (lncRNAs)

Main Tools

LNCipedia



LNCbook

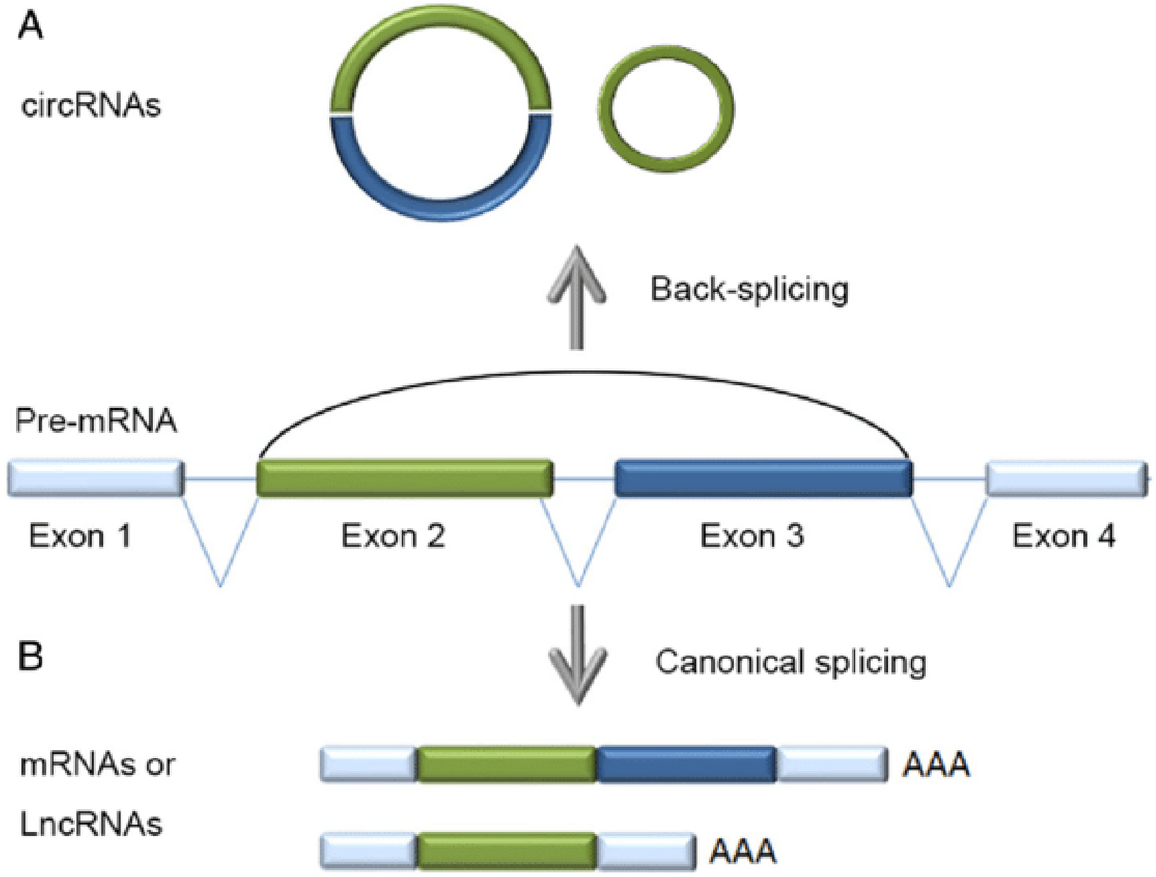


Other Tools

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

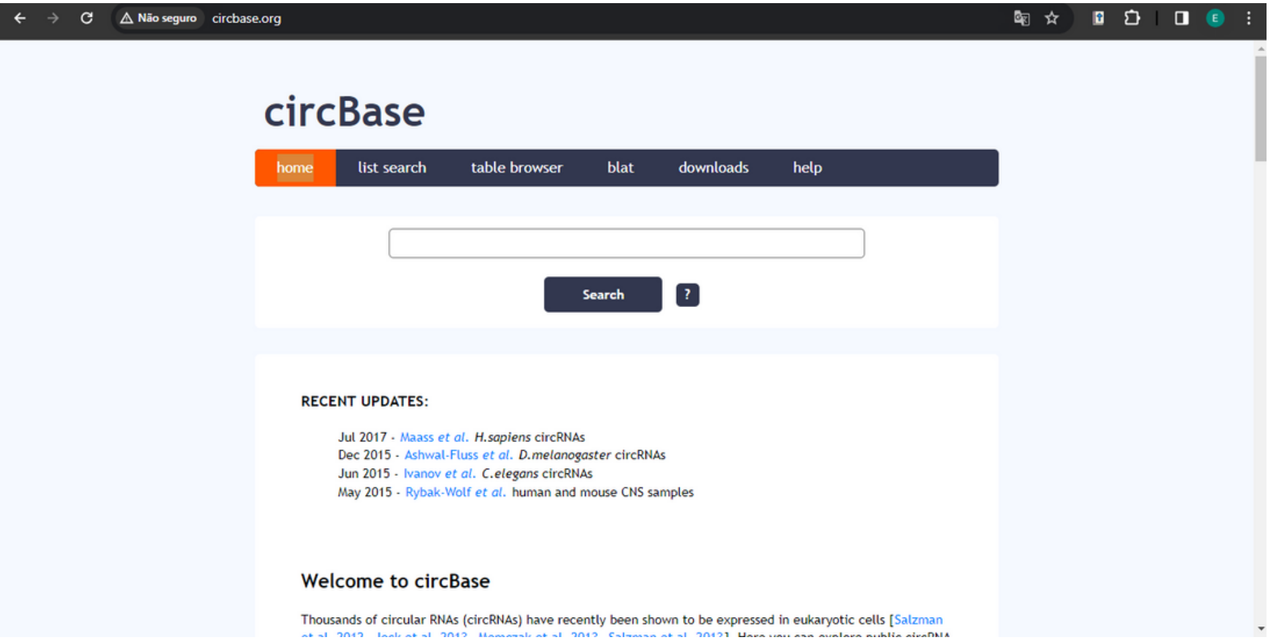


(CHUN YING YU; KUO, 2019)

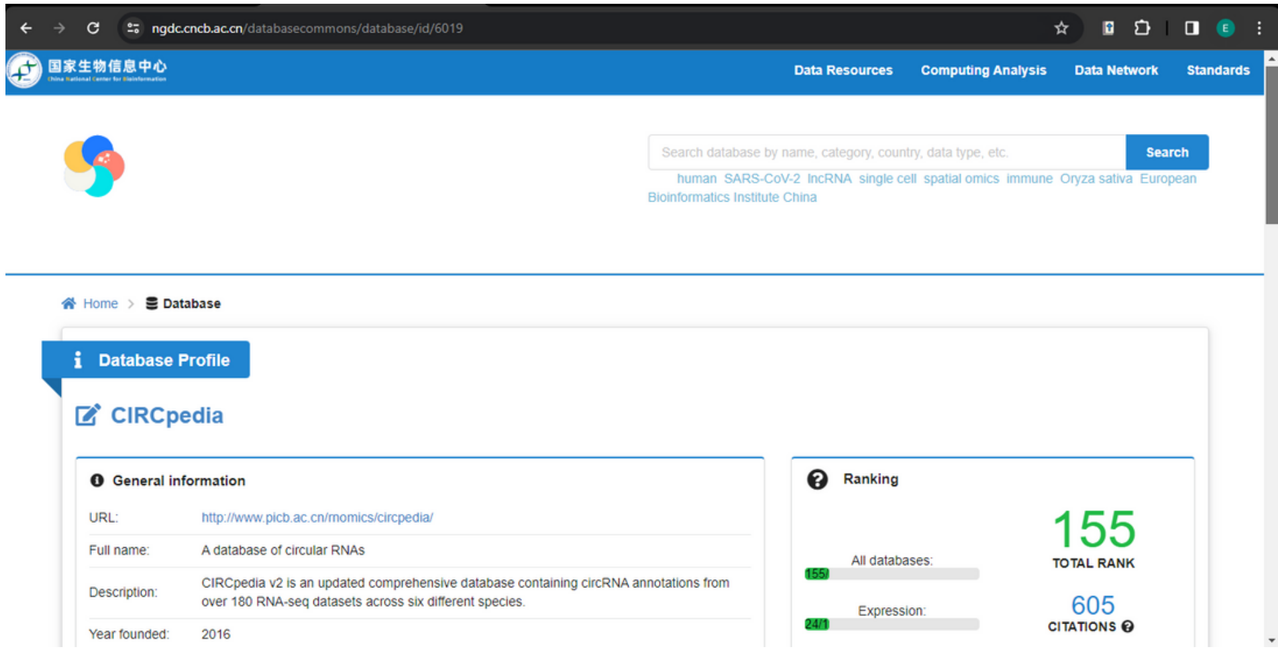
CIRCULAR RNAs (circRNAs)

Main Tools

circBase



CIRCpedia



Other Tools

CIRCExplorer2
find_circ

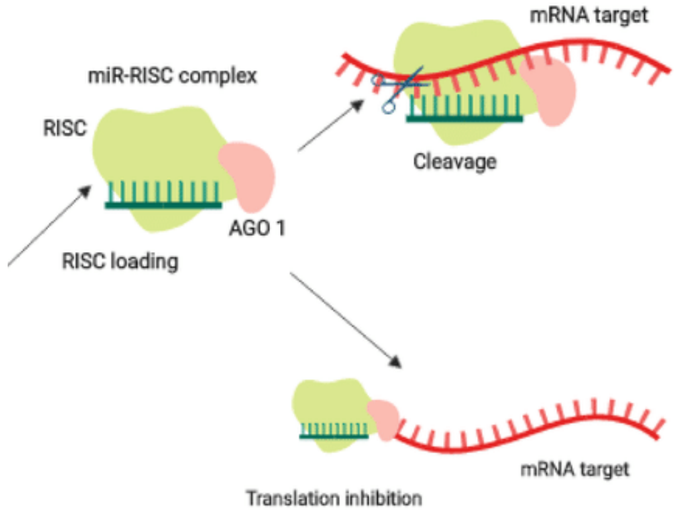
CIRCInteractome

Circ2Disease
circRNADisease
Circ2Traits

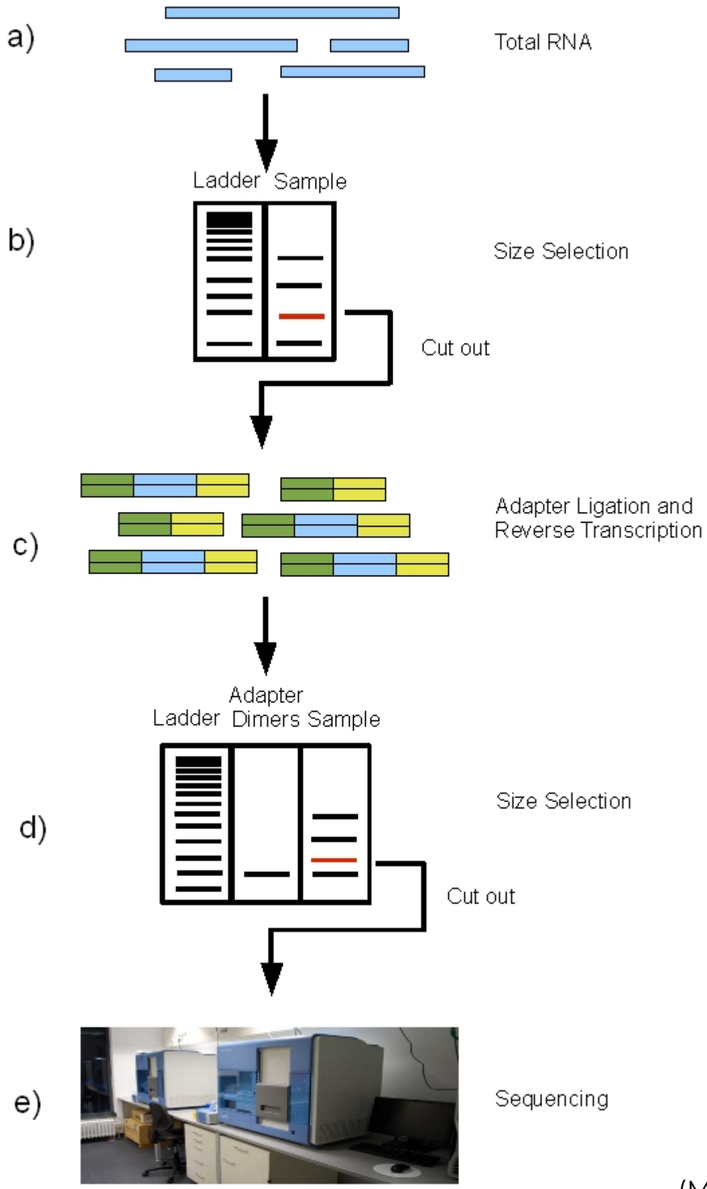
MICRO RNAs (miRNAs)

- 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)

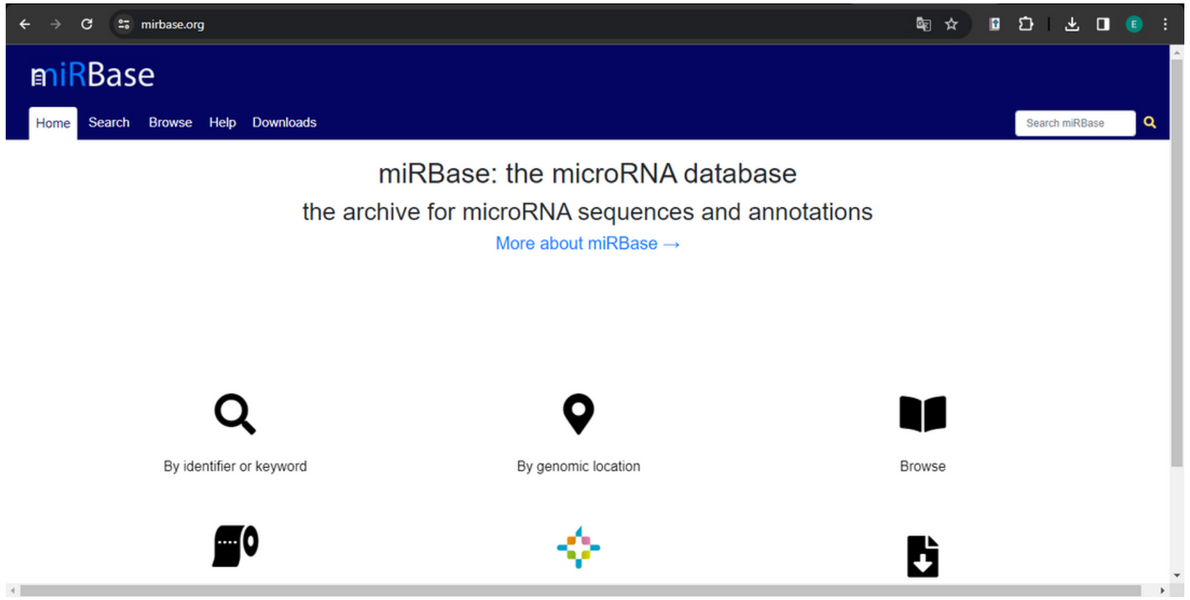


(MOTAMENY et al., 2010)

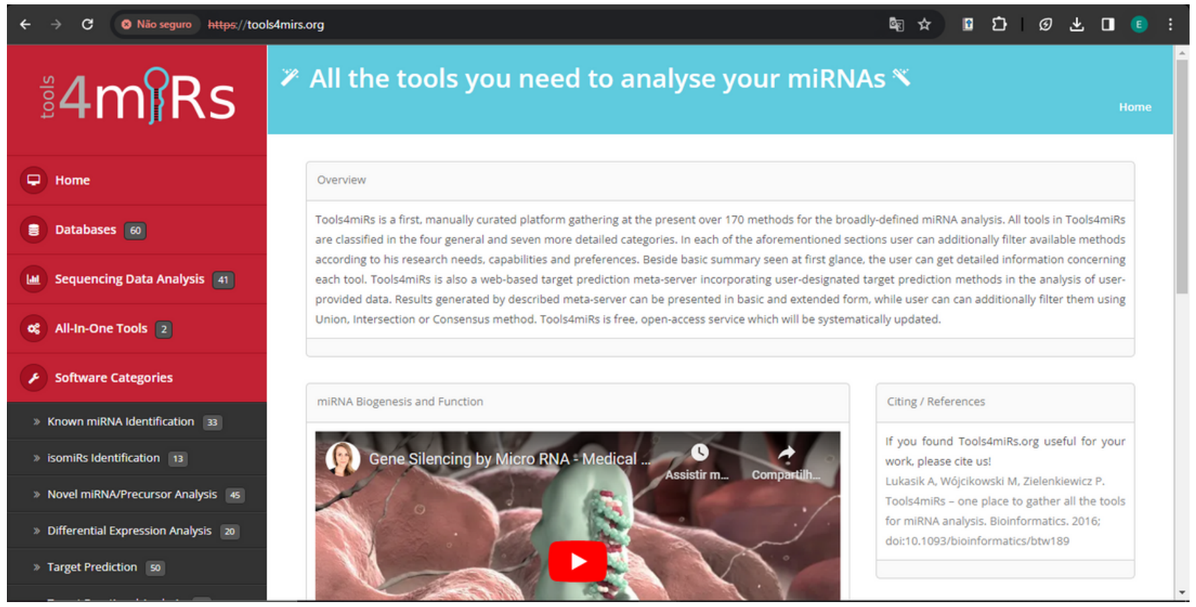
MICRO RNAs (miRNAs)

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
  - Trancript 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 biologial 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	https://www.niehs.nih.gov/research/resources/software/biostatistics/gaknn	[129]
GA/kNN, gradient boosting	Identification of differentially expressed genes	NA	[130]
E-M algorithm	De novo assembly of meta-transcriptomics data	https://sourceforge.net/projects/dnpipe	[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	https://github.com/armbrustlab/trophic-mode-ml	[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	https://github.com/chenlabgccri/CancerTypePrediction	[136]
Deep NN	Alternative splicing analysis	https://github.com/Xinglab/DARTS	[137]
CNN and DeepLIFT	Regulatory mechanisms identification	https://github.com/stasaki/DEcode	[138]
Mixing observation	Data augmentation	NA	[139]
Autoencoder	Cell content inference from bulk RNA-Seq data (which is typically done w/ scRNA-Seq data)	https://github.com/xindd/DCNet	[140]
ICA	Identification of novel regulons	https://github.com/avsastry/modulome-workflow	[141]

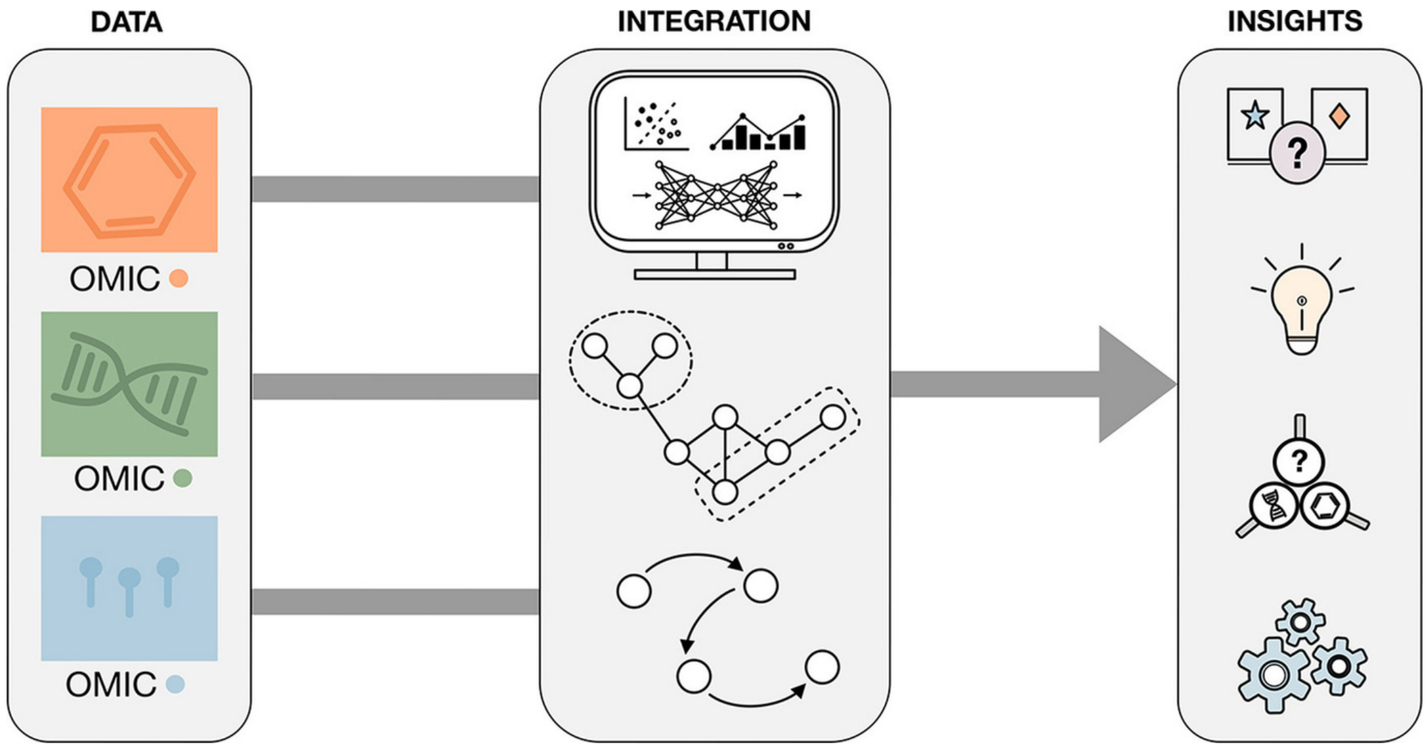
CNN: Convolutional neural network; E-M: expectation–maximization; GA: genetic algorithm; ICA: independent component analysis; kNN: k nearest neighbors; NN: neural network.

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



Wörheide MA, *et al* (2021).

Multi-omics workflow

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

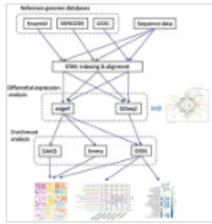
	xMWAS	PaintOmics 3	TIMEOR	Mergeomics 2	OmicsAnalyst	BIOMEX	miodin	3Omics	multiGSEA
Implementation	R, online	Online	Online, command line	Online	Online	R	R	Online	R
Functionality									
Pre-processing			x	x	x	x	x		
Data integration	x		x		x	x	x	x	
Network analysis	x	x	x	x	x			x	
Enrichment analysis		x	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[a]		x	x						
Regulatory omics[b]	x	x	x		x				
Reference	[177]	[178]	[179]	[180]	[181]	[182]	[183]	[184]	[185]

[a]ChIP-seq, ATAC-seq or Methyl-seq. [b]miRNAs or other transcription factors.

Thuany

EVALUATION

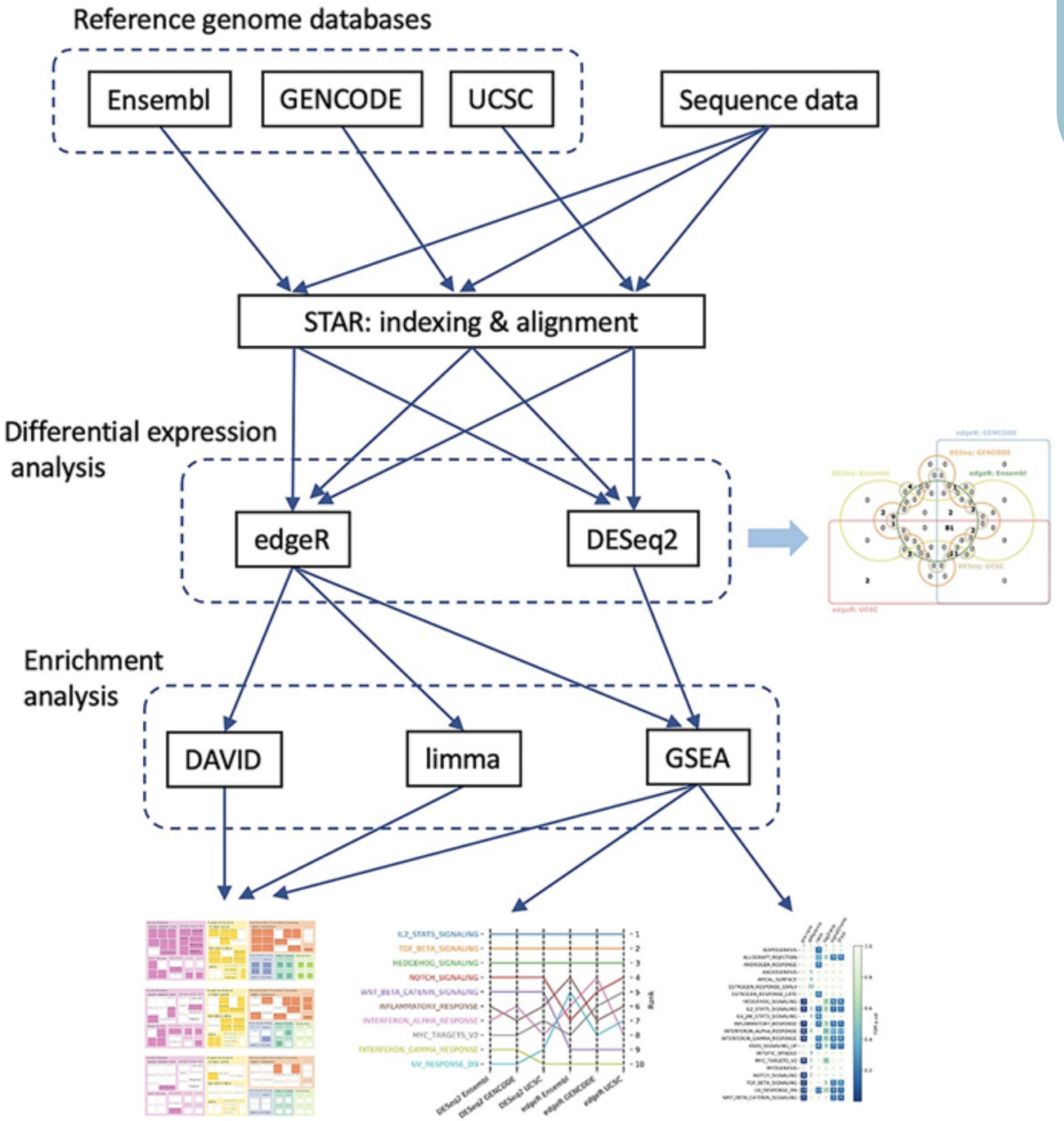
Typical RNA-Seq analysis process



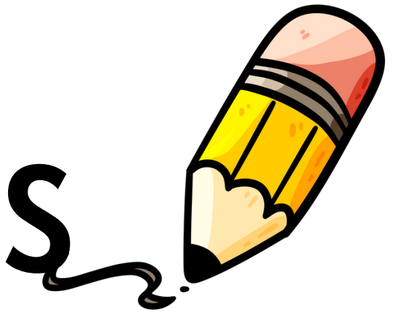
The hitchhikers' guide to RNA sequencing and functional analysis

Abstract. DNA and RNA sequencing technologies have revolutionized biology and biomedical sciences,...

OUP Academic

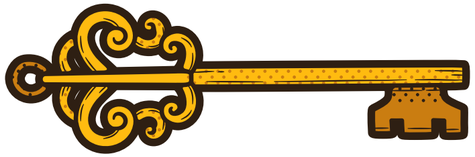


CONCLUSIONS



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

Thuany

- RNA-Seq analysis and software options.
  - Steps: Read alignment, read summarizaiton, differencial 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.

THANK YOU