Small RNA transcriptomics

IN-BIOS5000/9000

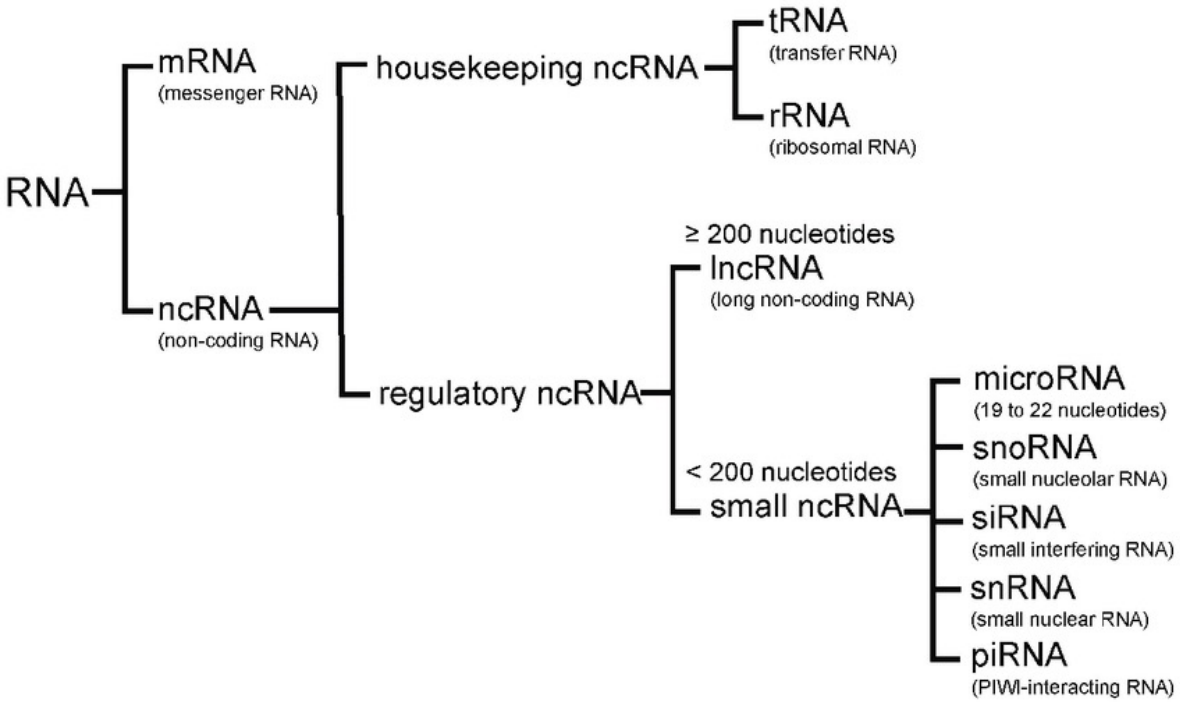
Trine B Rounge

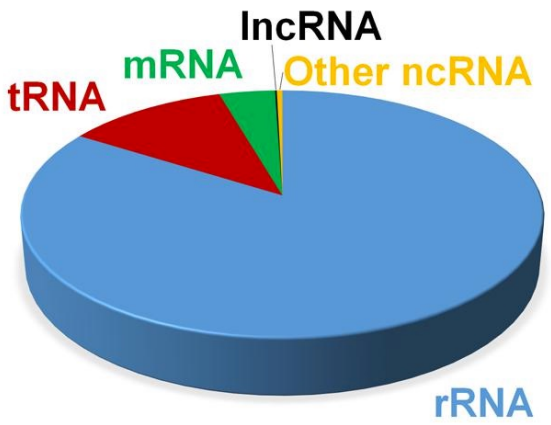
Learning outcomes

1. What are small RNAs
    - Their role in the cell and in bodily fluids
2. What is small RNAs transcriptomics
    - Methods/technologies
    - Experimental design
3. Analyse small RNA-sequencing data
    - Read Counts
    - How to identify differences
4. Research examples (if time)
5. Practical tasks

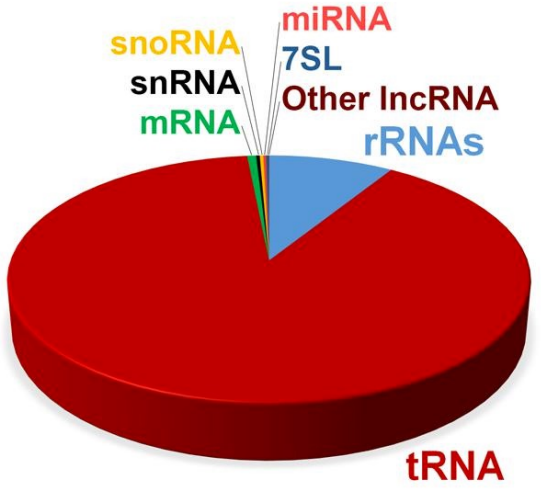
1. Small RNAs

What are small RNAs





RNA by mass

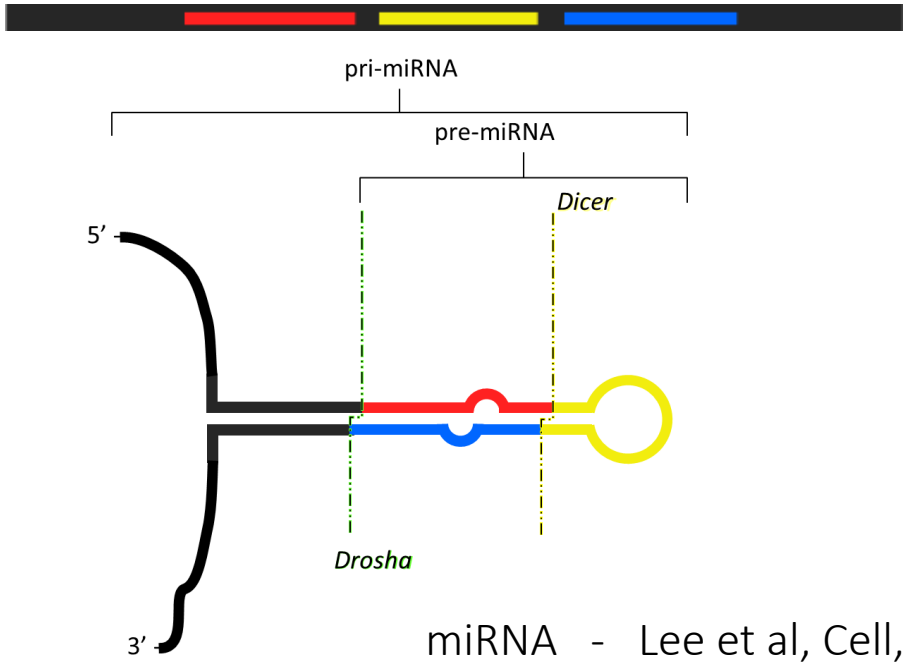


RNA by number of molecules

Palazzo & Lee, Front Genet, 2015

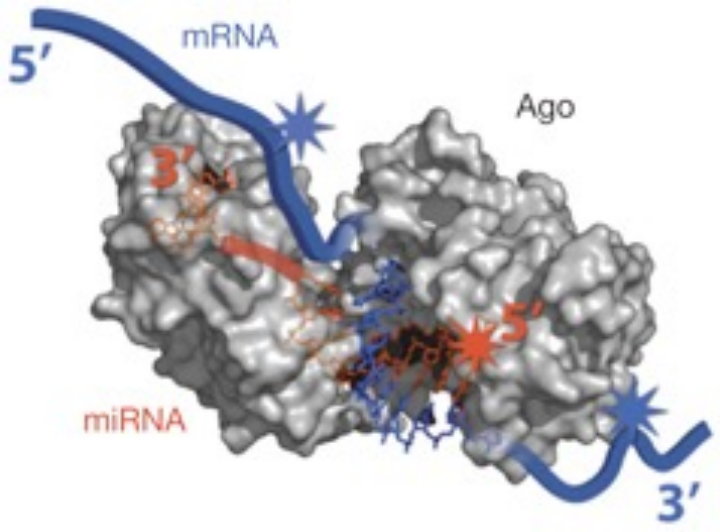
miRNAs

~ 22 nucleotide in length



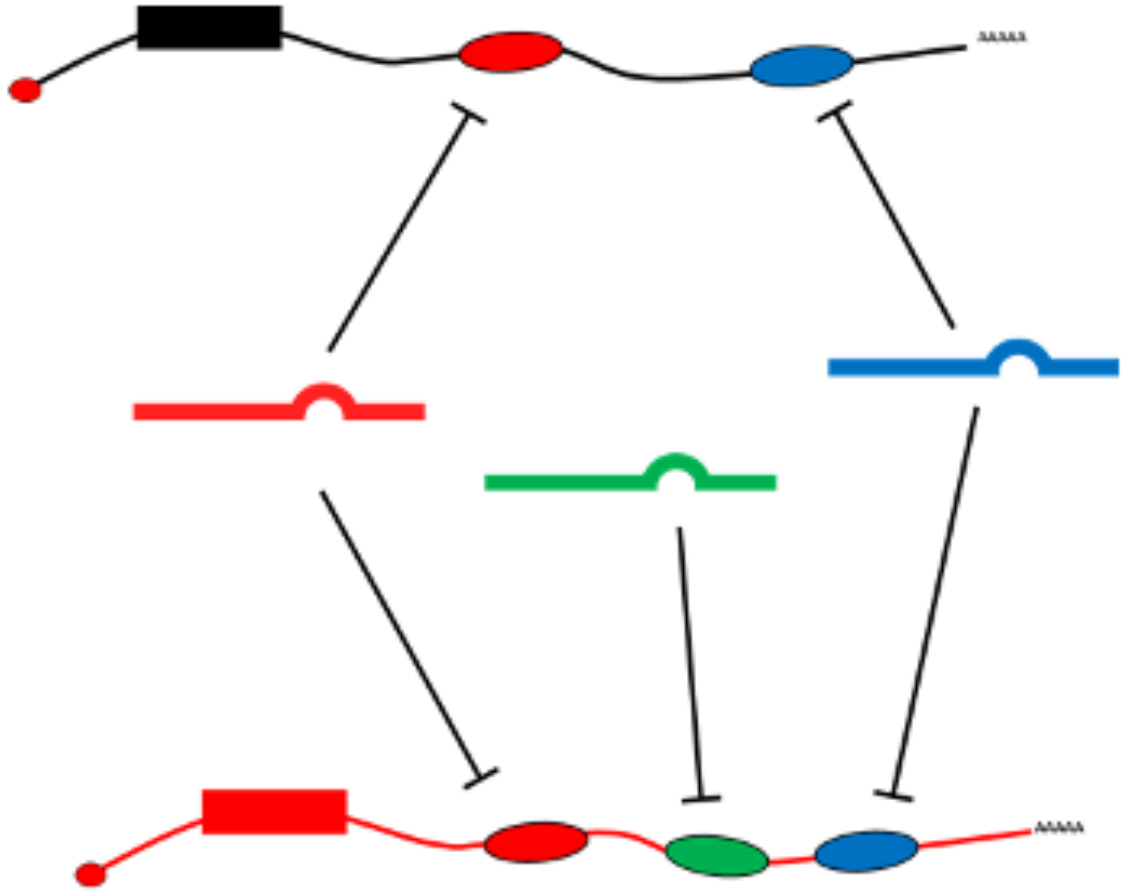
miRNA - Lee et al, Cell, 1993
Dicer - Hutvagner et al, Science, 2001
Drosha - Lee et al, Nature, 2003

RNA-induced silencing complex (RISC) including the Argonaute protein



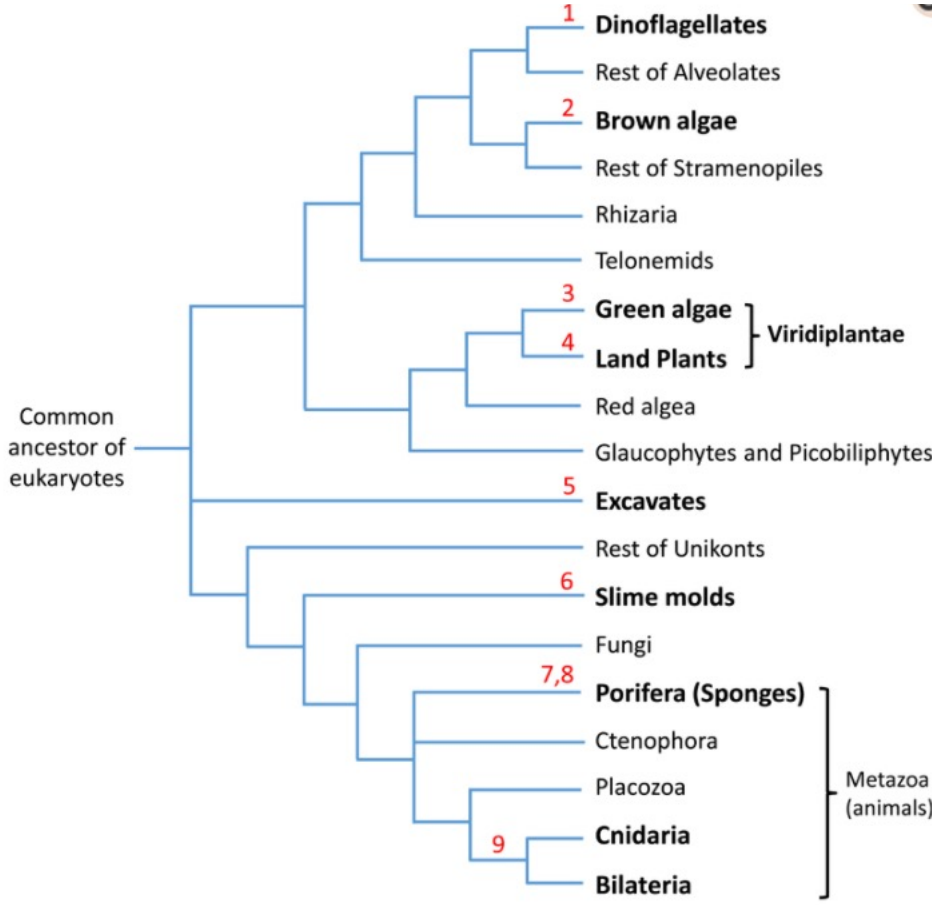
miRNA function

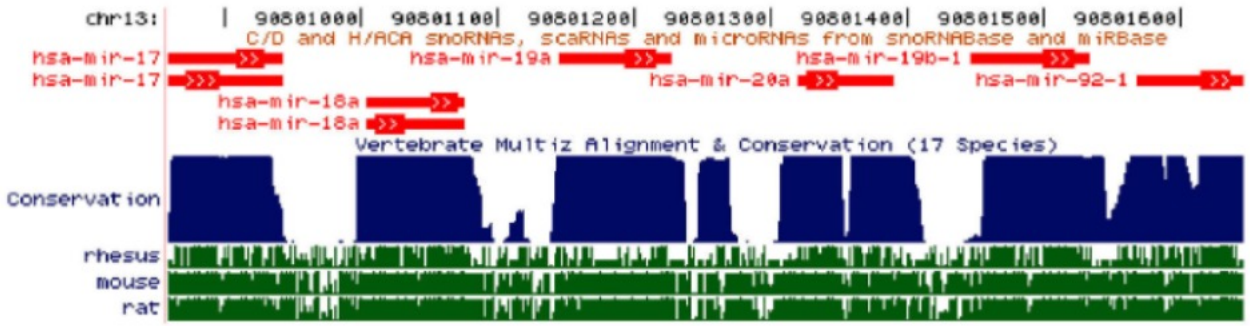
- The miRNA-RISC complex targets mRNA for silencing.
- target multiple mRNAs and multiple miRNAs can target the same transcript.
- the fine tuning of most protein products within the cell.



Pichler & Calin, Br J Cancer 2015

Evolutionary conserved



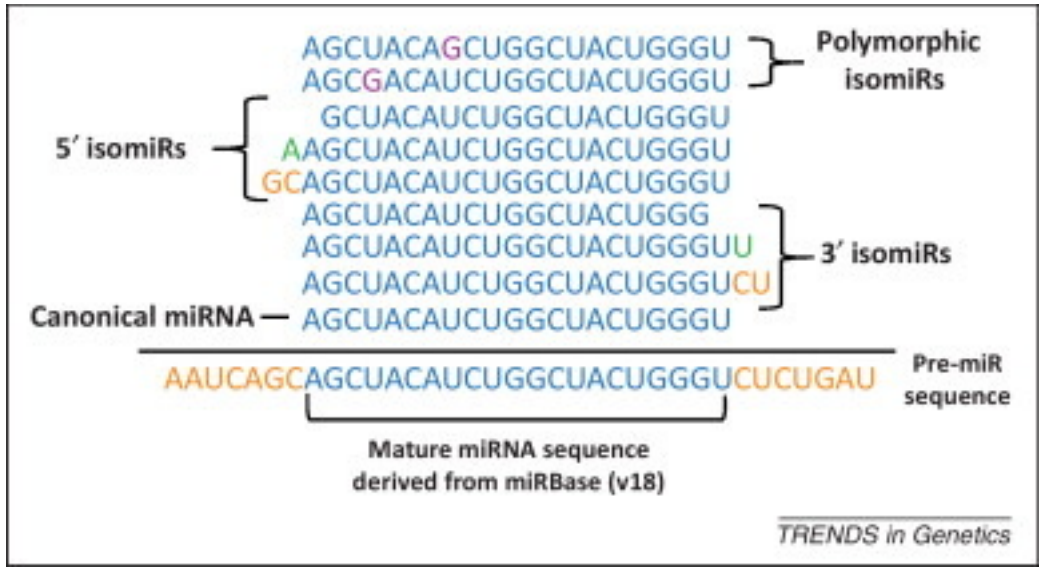


Moran, Nat Eco Evol, 2017

Yue, BMC Genomics, 2008

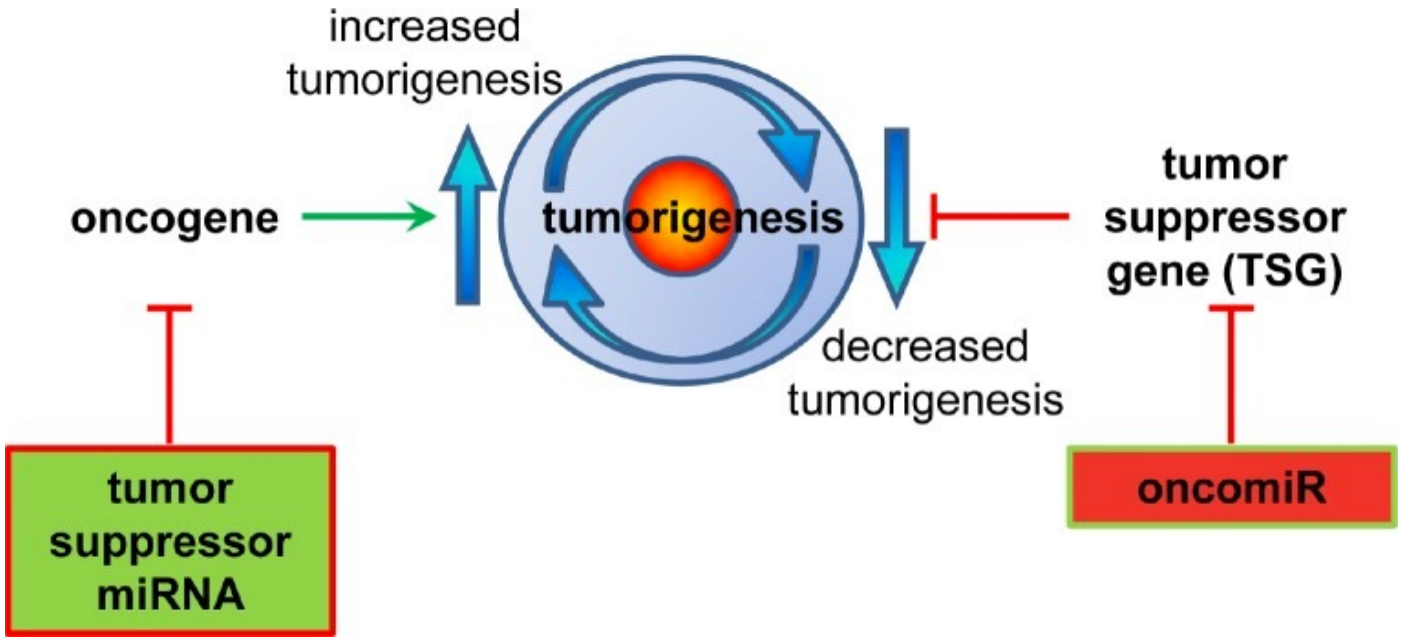
Isoforms

IsomiRs



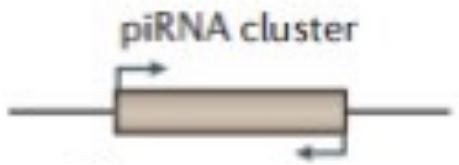
Variability in Dicer and Drosha processing

miRNA expression in cancer

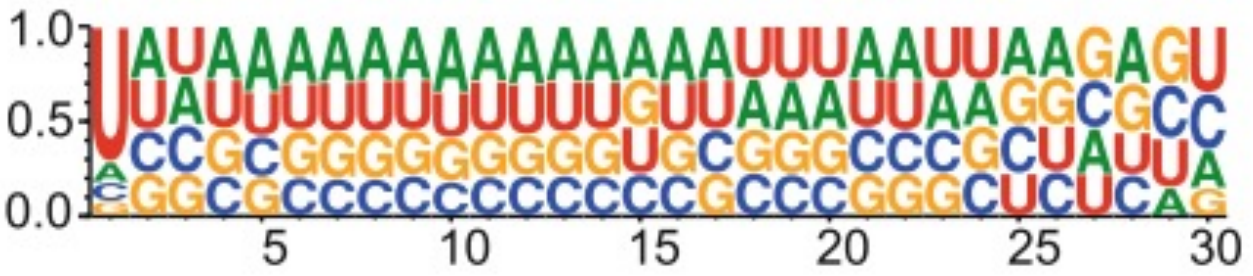


Pichler & Calin, Br J Cancer 2015

piRNA

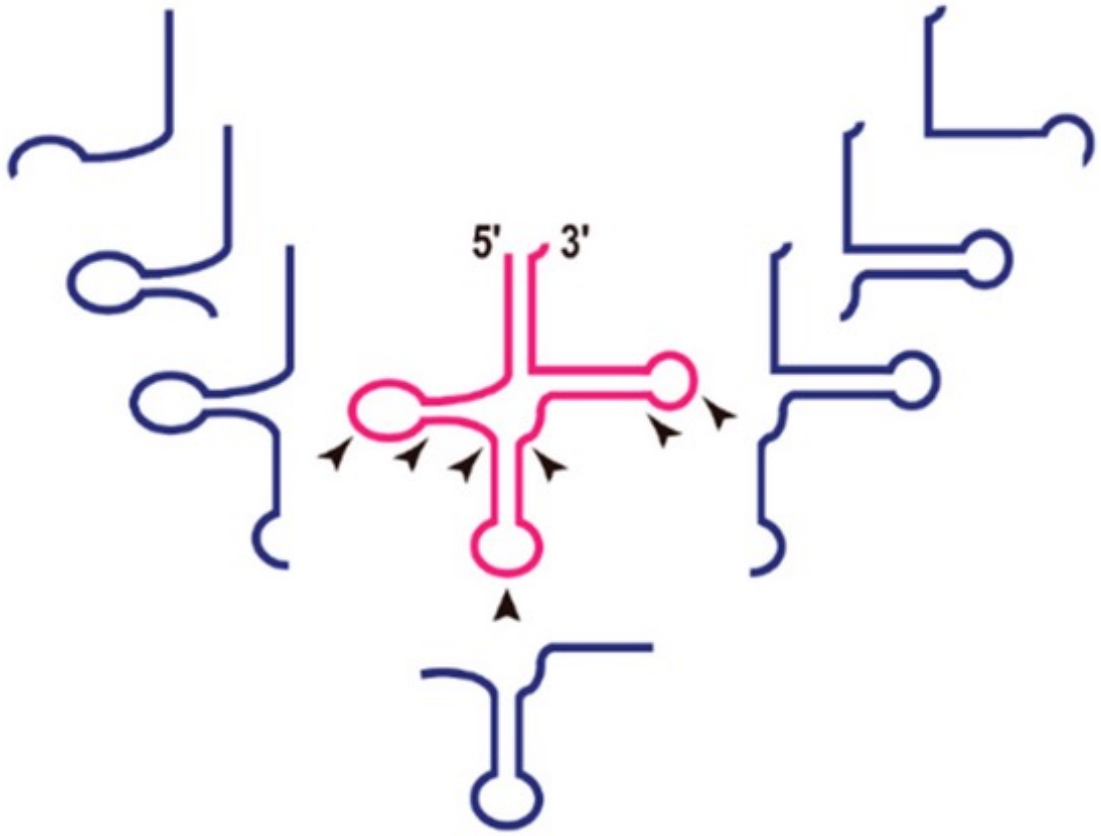


- 26-31 nucleotides in length
- transcribed in clusters
- 5’ uridine
- a role in RNA silencing via PIWI
- active in the testes of mammals
- silencing of transposons
- silencing via RISC
- amplified by a ping-ping mechanism



tRNA fragments

- 76 - 90 nucleotides in length
- carrying an amino acid to the ribosome
- Fragmented in:
  - halfs
    - stress induced
  - fragments (1/4)
    - RISC regulation,
    - epigenetic control,
    - metabolism,
    - immune activity
    - stem cell fate commitment



Kanai, Life, 2015

snRNA and snoRNA

- snRNA – small nuclear RNA
  - Role in splicing
  - ~ 150 nucleotides.
    - Bohnsack et al, Biol Chem, 2018

- snoRNA – small nucleolar RNAs
  - role in the modification, maturation, and stabilization of rRNA
  - regulation of gene expression and alternative splicing
  - stress response
    - Liang, Frontiers in Oncol, 2019

2. Small RNAs transcriptomics

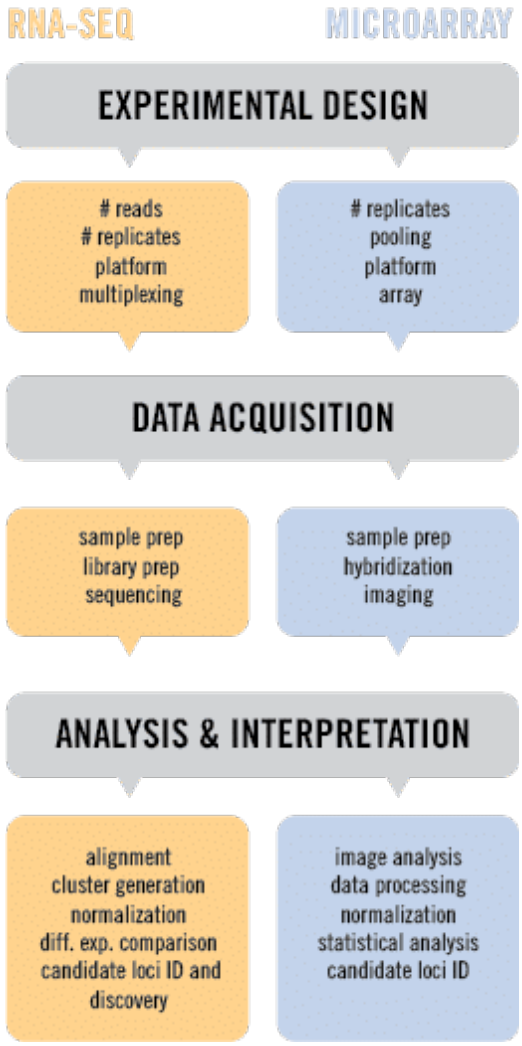
Small RNA transcriptomics

- Study of the the **complete set of small RNA transcripts** that are produced by the genome, under specific circumstances or in a specific cell or body fluids—using high-throughput methods

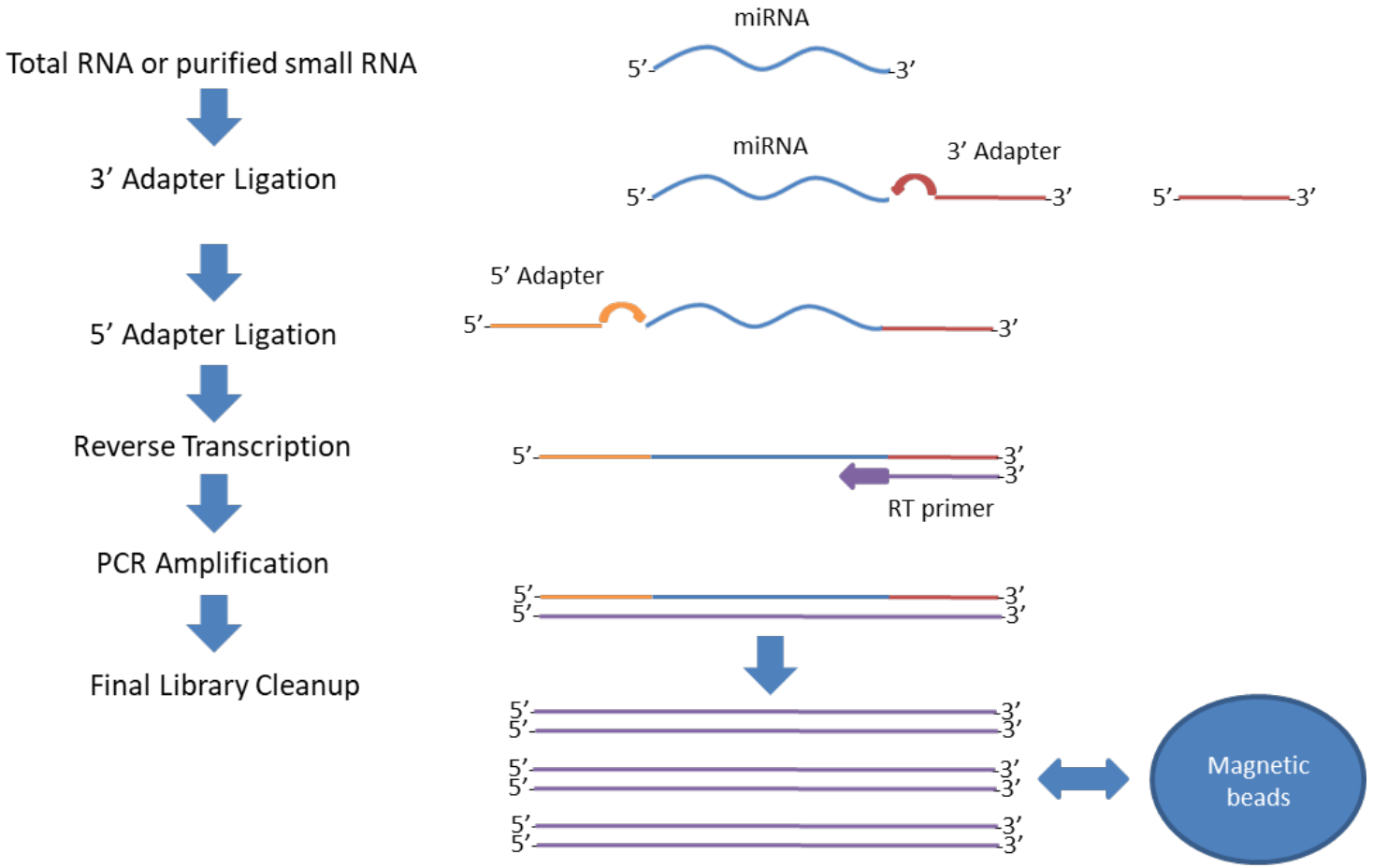
- <100 nucleotides or <200 nucleotides
  - Transcribed as small RNAs
  - Processed to small RNA
  - Degraded to small RNA

Small RNA Technologies

- RT-qPCR panels
  - limited part of the transcriptome
  - good reproducibility
  - good for validation
- Microarray
  - limited to known genomes and known transcripts
  - In some cases poor sensitivity and a limited dynamic range
  - non-specific hybridization or cross-hybridization
  - well defined RNAs
  - low cost
  - low noise
  - MirNome
- Small RNA sequencing
  - potential ambiguity in the mapping
  - protocol specific biases
  - discovery
  - large range
  - Many kits – not comparable results

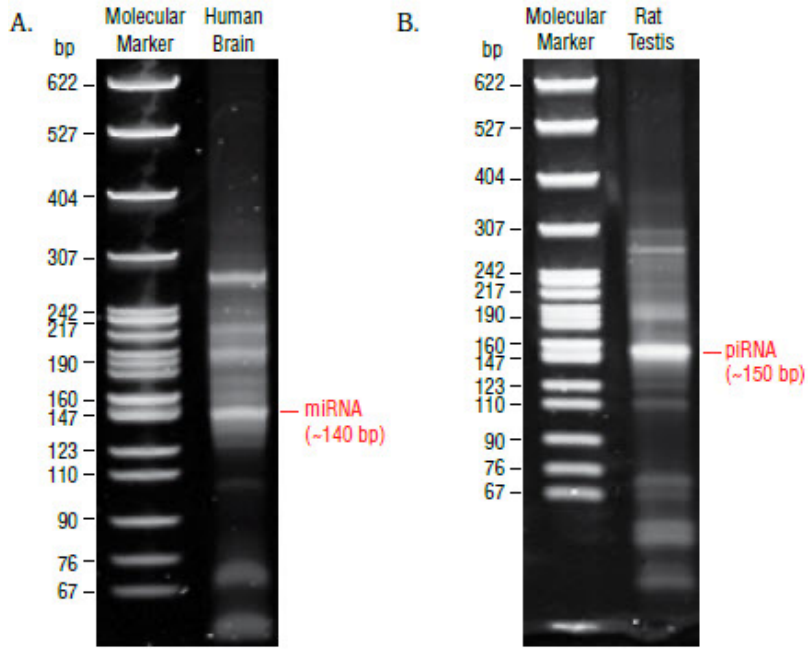


Small RNA sequencing – library preparation

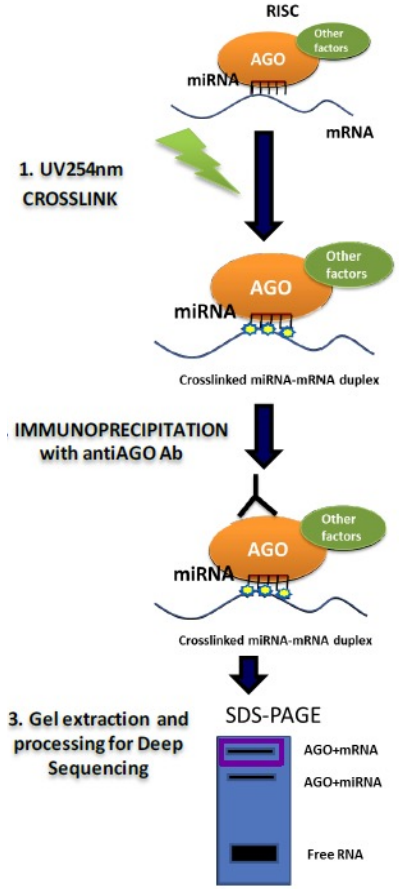


Small RNA transcriptomics

Size selection

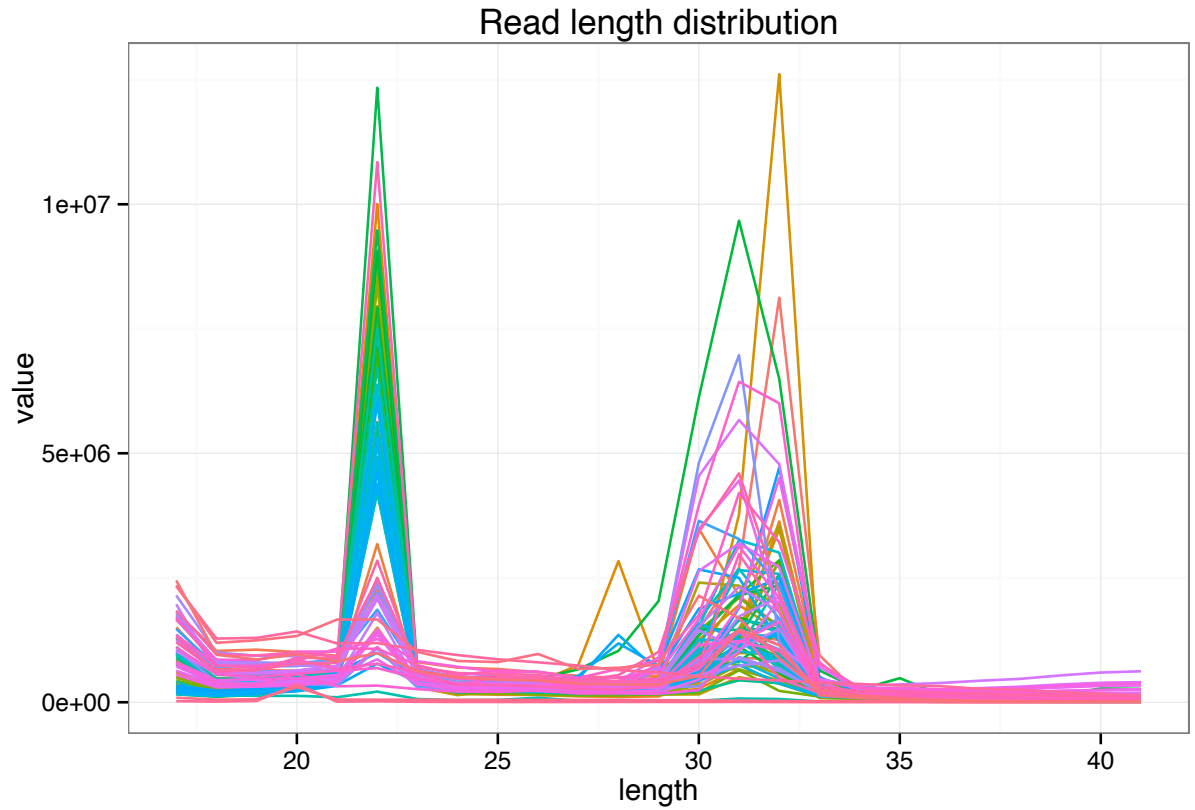


AGO immunoprecipitation



Small RNA transcriptomics

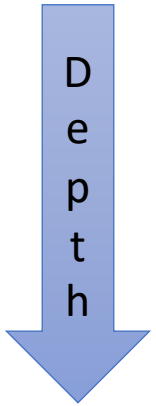
- Size selection is **not** perfect
- Bias can be introduced
- Size distribution need to be checked



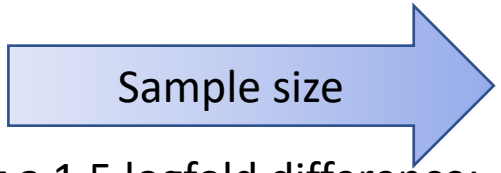
Sequencing

- The most common technology for small RNA-seq
- Reconstruction of isoforms
- Detection of novel transcripts
- Suited for **expression analysis**
- High number of samples

Sample size vs depth



- RNA sequencing
  1. Highly expressed known transcripts
  2. Novel isoforms
  3. Low expressed / rare transcript



To detect a 1.5 logfold difference:

- 3 samples/group –> 43% statistical power
- 10 samples/group –> 91% statistical power

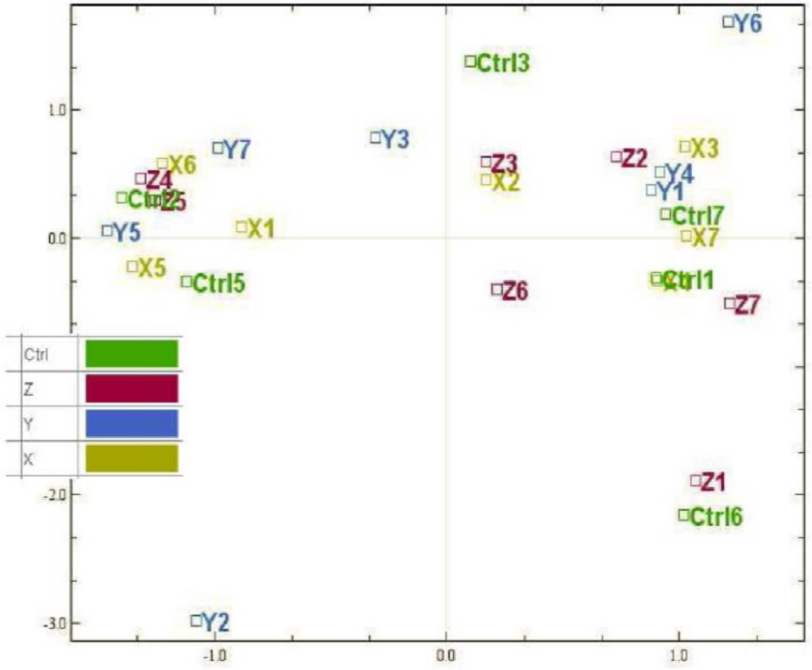
With 10 sample/group:

- 3 mill read/sample –> 52% statistical power
- 10 mill read/sample –> 80% statistical power

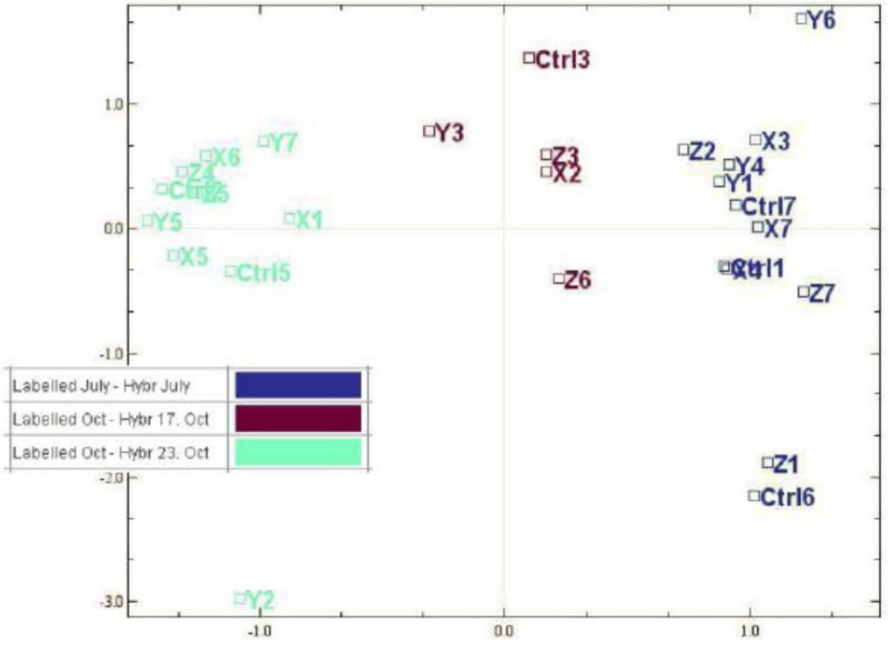
Talk to a statistichian

A survey of best practices for RNA-seq data analysis
https://doi.org/10.1186/s13059-016-0881-8

Batch effects



Samples color coded according to biology



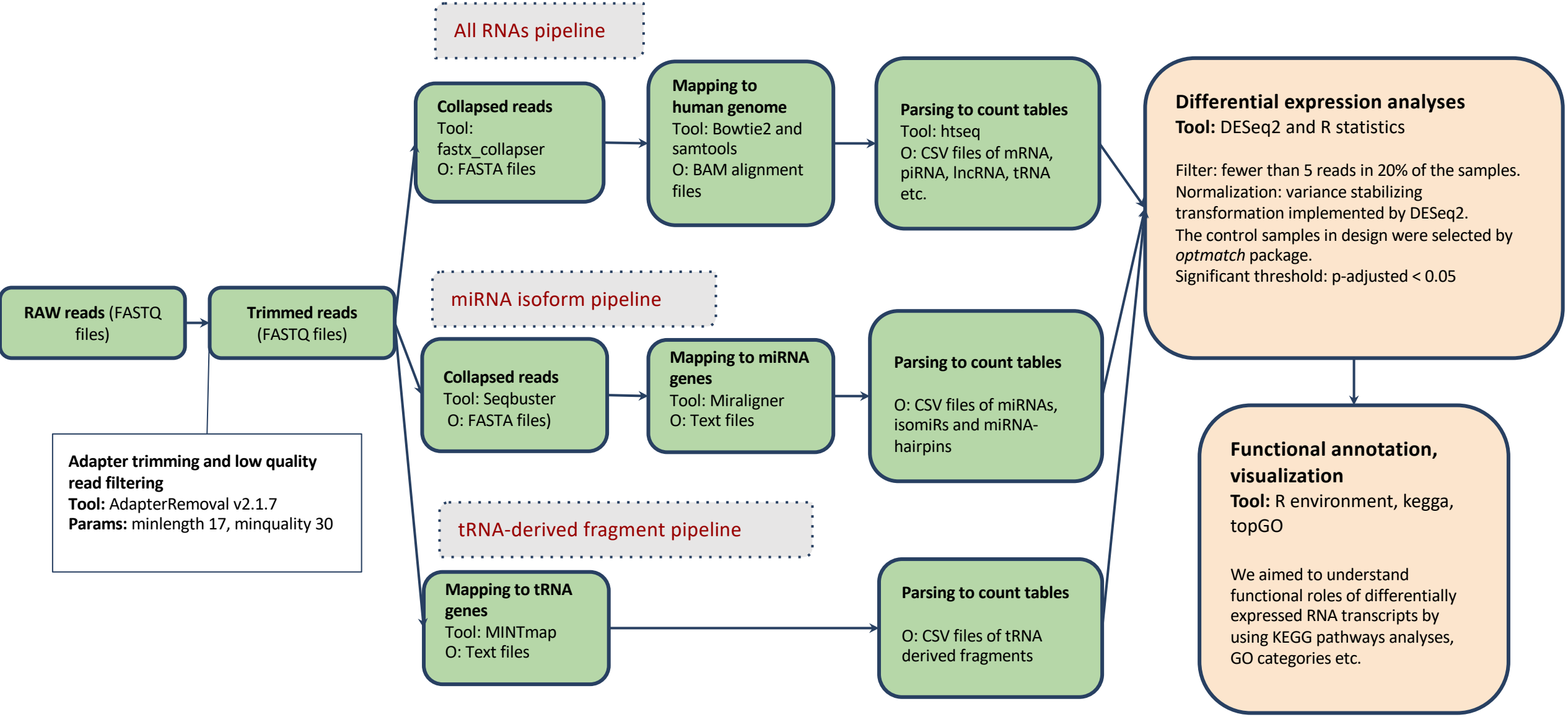
Samples color coded according to labeling date

3. Analyse small RNA sequencing data

Small RNAseq workflow

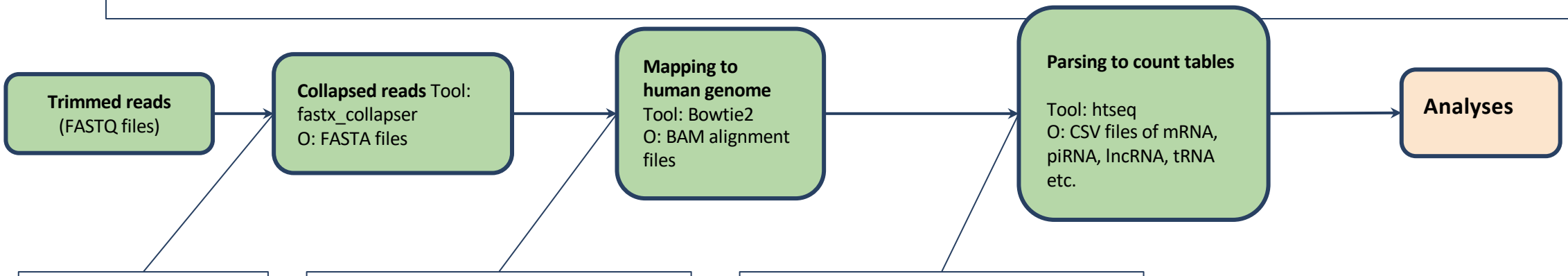
From Fastq files to small RNA read counts

- Reproducible
- Detect isoforms
- Detect most small RNA classes
- Efficient



https://github.com/sinanugur/sncRNA-workflow

All RNAs pipeline



Tool: fastx_collapser

We collapsed overrepresented reads which is essential to reduce mapping time and disk space.

Genome: hg38
Tool: bowtie2
Params: --sensitive-local -k 10

To allow 10 alignments and to use local alignment method, no mismatch in seed.

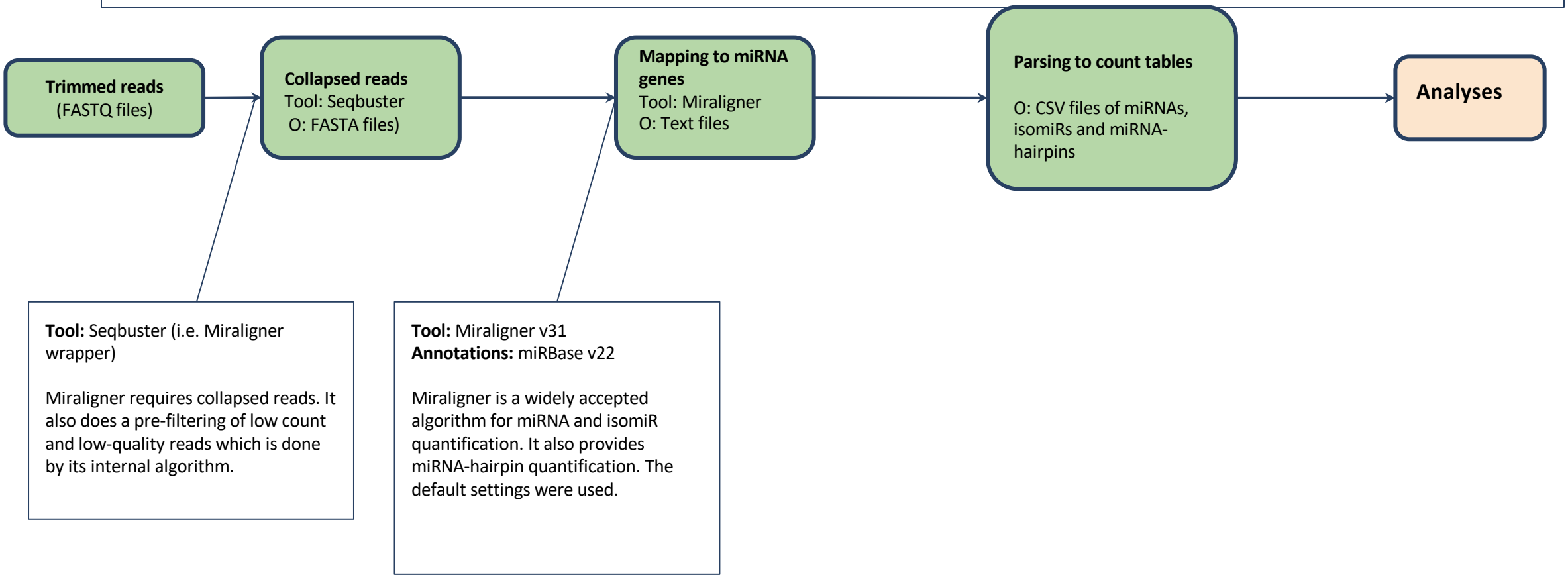
Sensitive-local is usually recommended for small RNA read mapping. We aimed to get a high sensitivity for all transcripts.

Tool: htseq v0.9.1 (in a python script)
Annotations: GENCODE v26, pirbase v1, MirGeneDB v1.1

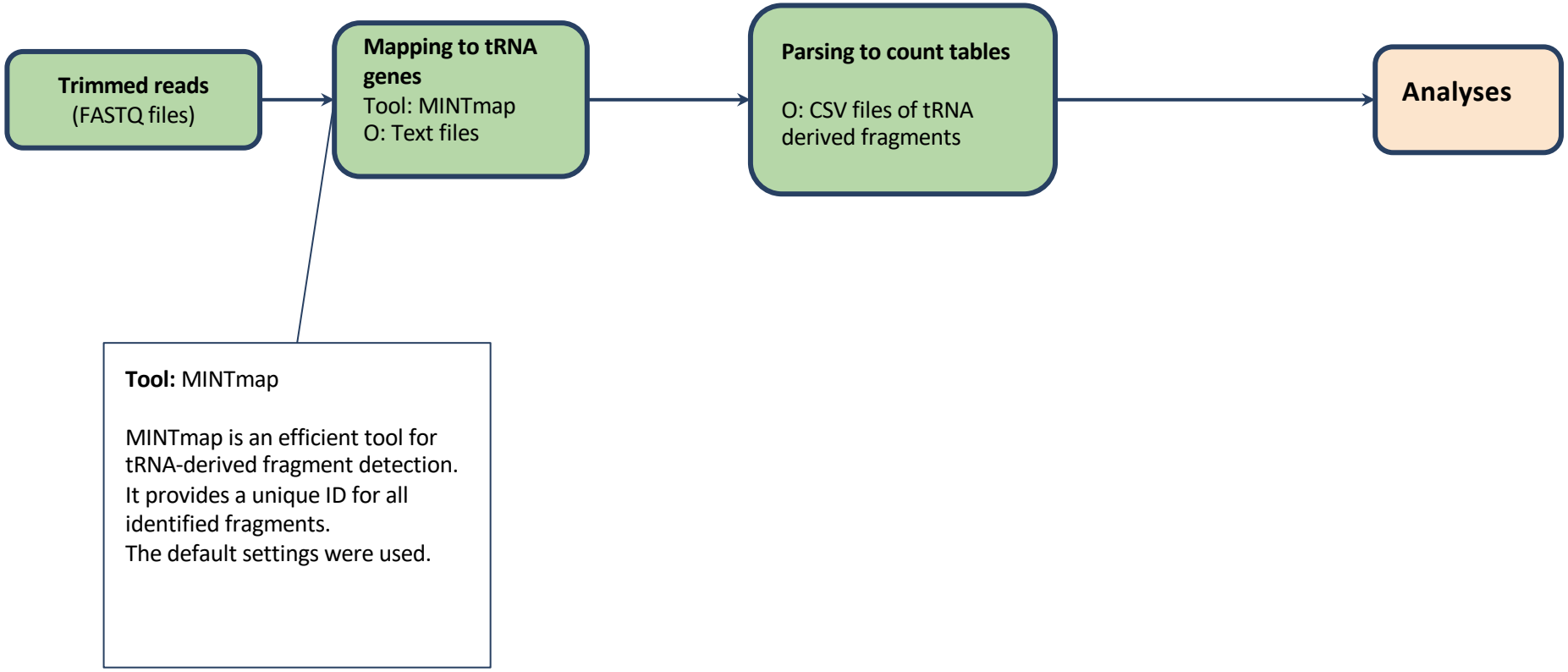
We selected the highest quality alignment for multi-mapped reads to increase specificity while decreasing noise of ambiguous mapping.

The GENCODE database contains almost all transcript classes, including mRNAs, lncRNAs, tRNAs, miscRNAs etc. We aimed to recover all possible transcript types.

miRNA isoform pipeline

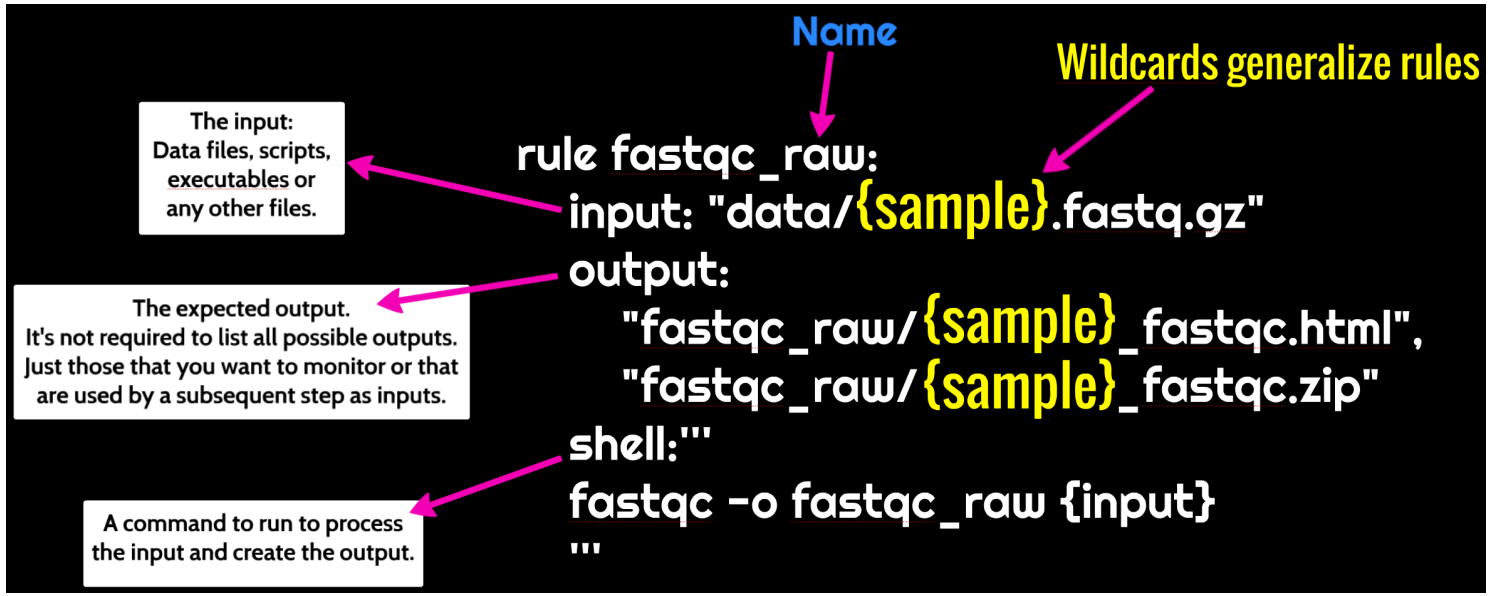


tRNA-derived fragment pipeline



Snakemake

- tool to create **reproducible and scalable** data analyses.
- python based language.
- Can be scaled to server, cluster, grid and cloud environments.



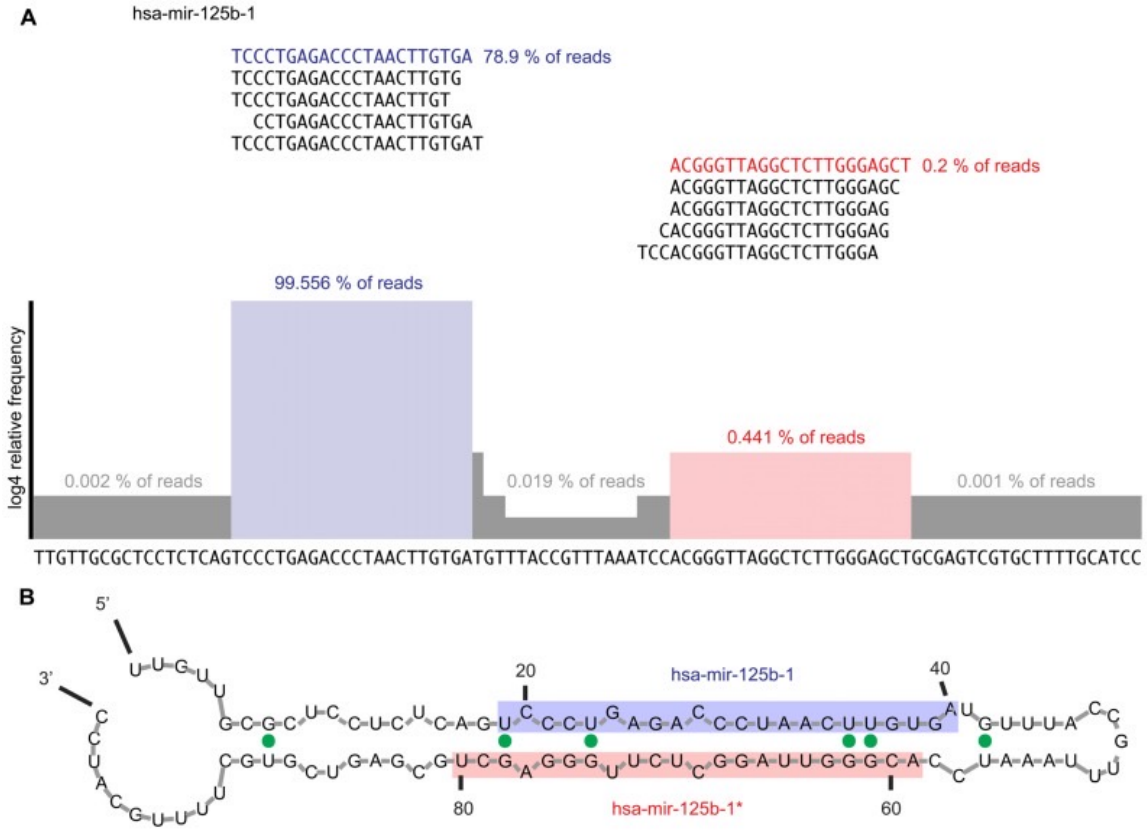
https://snakemake.readthedocs.io/en/stable/

- Break

Small RNA count data

- Similar to mRNA count data
- Each class analysed separatly
  - Detection/sequencing bias due to length differences

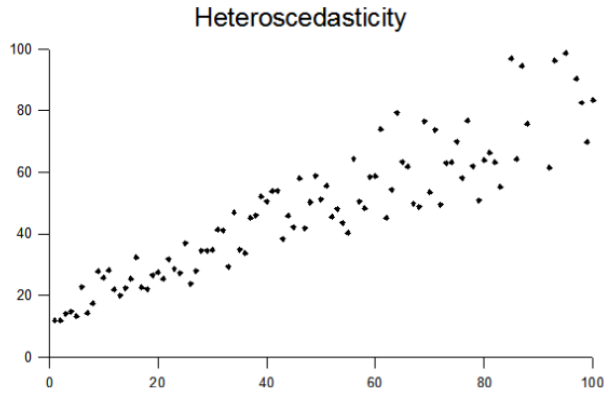
Read counts



Gene/Sample	Sample 1	Sample 2	Sample 3 repA	Sample 3 repB
Gene A	5	0	45	101
Gene B	17	500	32	67
Gene C	752	16432	20020	45078
Total	350250	278090	400890	799009

Normalization

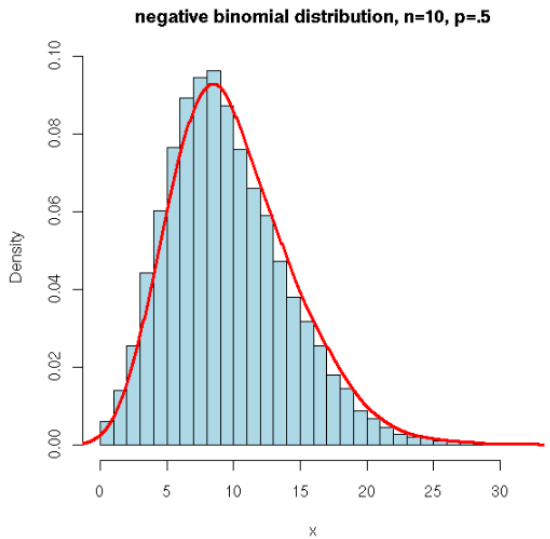
Challenges of using count data



- The unit of measurement is number of reads (count) – not light intensities (microarray), or the shape of a curve (qRT-PCR).
- Large **dynamic range** – from zero up to millions.
- **Heteroscedastic** – variance is not equal across the range of counts
- Positive integers and non-symmetric

distribution – i.e. normal distribution doesn't fit.

- Systematic biases (**normalization**)



Normalization

- Sequencing depth – library size
- Sequencing length
- Variance across means

Normalization for library size

- If sample A has been sampled deeper than sample B, we expect counts to be higher.

- Naive approach: Divide by the total number of reads per sample
- Problem: Genes that are strongly and differentially expressed may distort the ratio of total reads.

Normalization for library size

- To compare more than two samples:
- Form a “virtual reference sample” by taking, for each gene, the geometric mean of counts over all samples

- **DESeq2:** Normalize each sample to this reference, to get one scaling factor (“size factor”) per sample.

Anders and Huber, 2010
similar approach: Robinson and Oshlack 2010

Research question - analyses

You now have a normalised small RNA count table

- Now you must design your analyses according to your research question

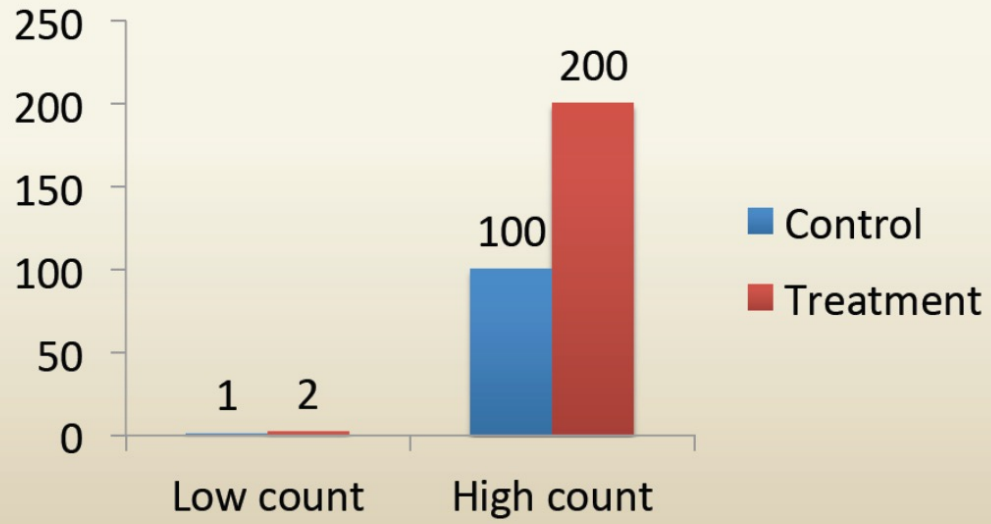
- For example:
  - Identify small RNAs that differes in expression between two groups
    -> Differential expression analyses
  - Construct a model based on RNA counts that classifies two groups
    -> machine learning

Differential gene expression analysis

- DESeq2 implements a **Wald test**
- Conceptually similar to a t-test
- Goal is to identify genes that are differentially abundant (DE) between two conditions.
- Assumption: most genes are not DE
- Null hypothesis: each gene has the same abundance across conditions.

Strong poisson noise for low count values

1) Poisson counting error
   - Uncertainty in count-based measurements
   - Disproportionately large for low-count data



DE testing– adjusted p-values

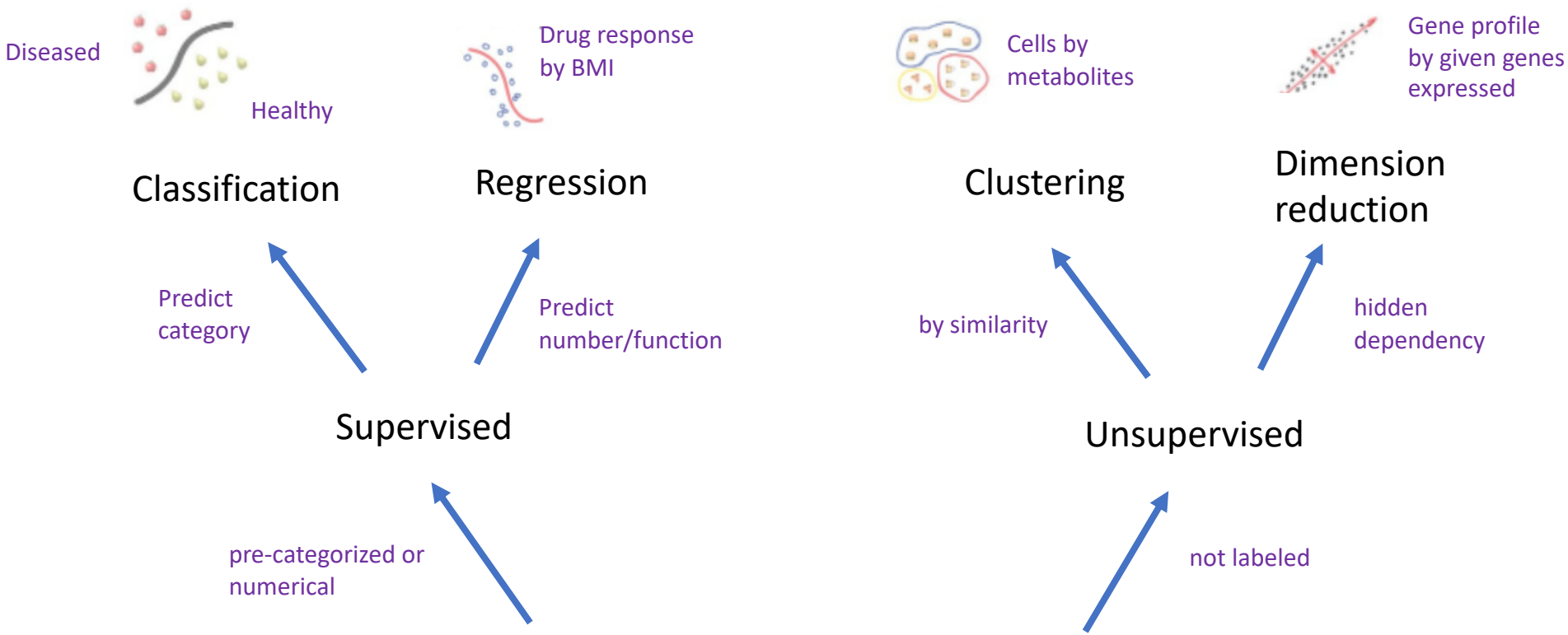
Multiple hypothesis testing

- Thousands of genes = thousands of hypothesis tests (simultaneously)
- Increased chance of false positives! (Type I error)
  - e.g. you test for differential expression in 1000 genes that are not differentially expressed
  - You would expect 1000 x 0.05 = 50 of them to have a *P*-value < 0.05
- Individual *P*-values not useful
  - Need multiple testing statistic instead

Differensial Expression

- **DESeq2** and **edgeR** – two of the most common pacakges for RNA-seq analysis (differential expression).
- **DESeq2** and **edgeR** – based on “raw counts” such as from HTSeq

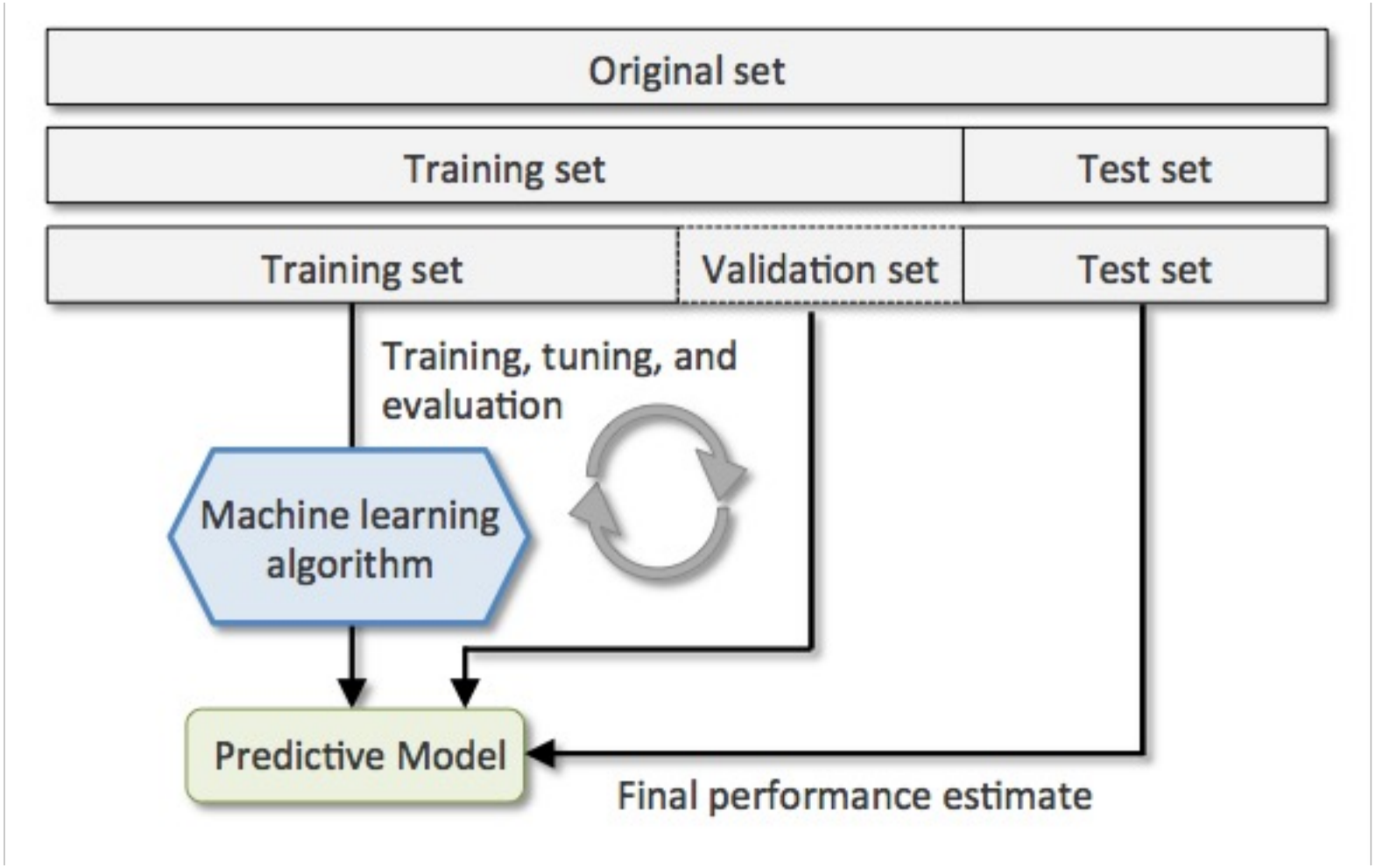
Differential expression vs machine learning



Classical machine learning

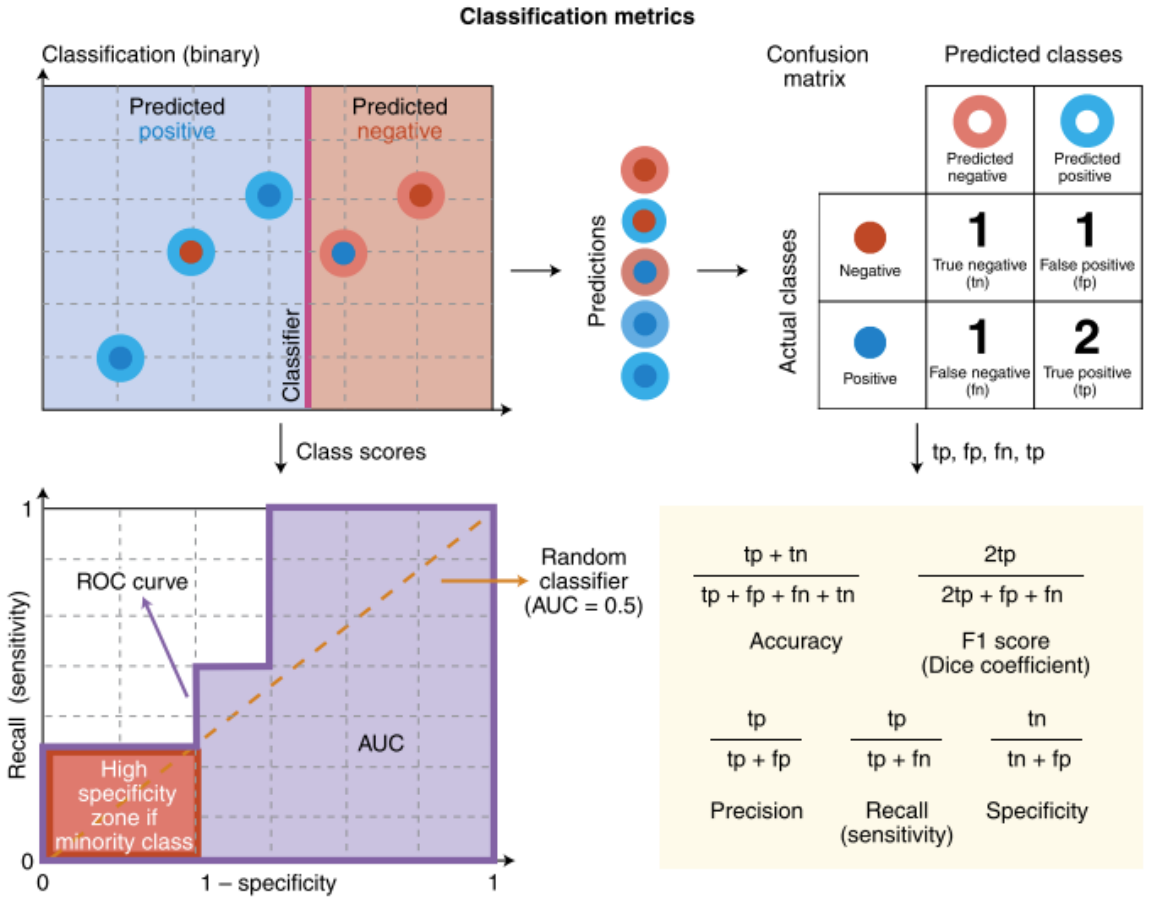
Credit: https://vas3k.com/blog/machine_learning/

Training – test – cross validation



Evaluation

- Evaluation scenarios in life sciences
  - experimental validation of the predictions
  - computational assessment
- Quantifiable indicators of a model's ability to solve the given task
- Confidence intervals

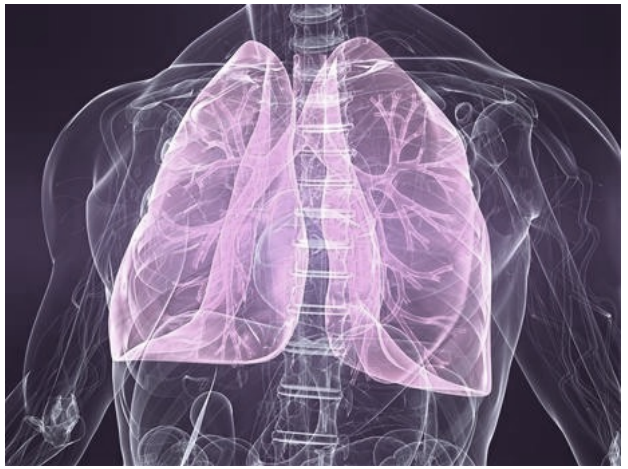


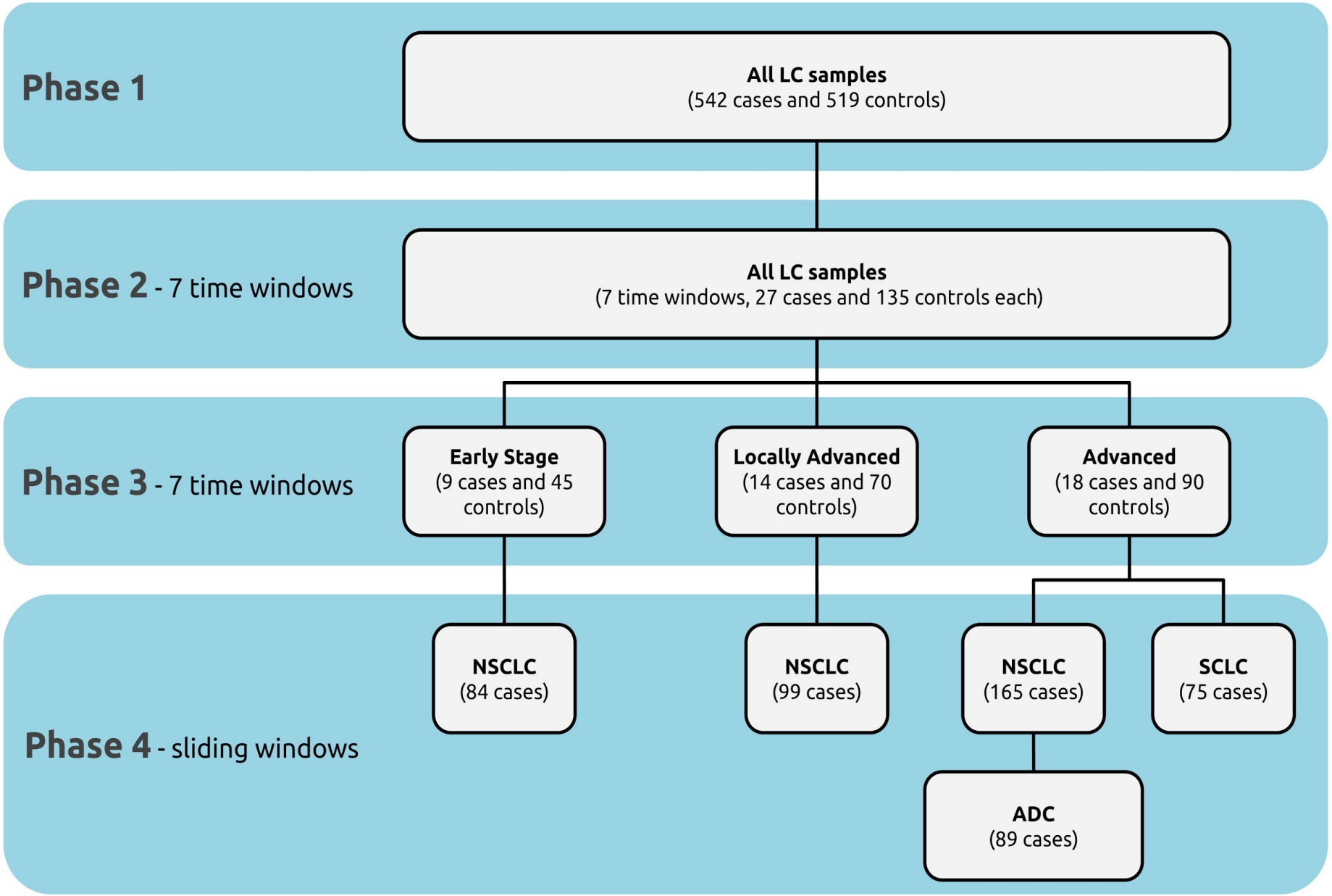
Walsh et al, Nature Method, 2021

Mentimeter - quiz

4. Research example

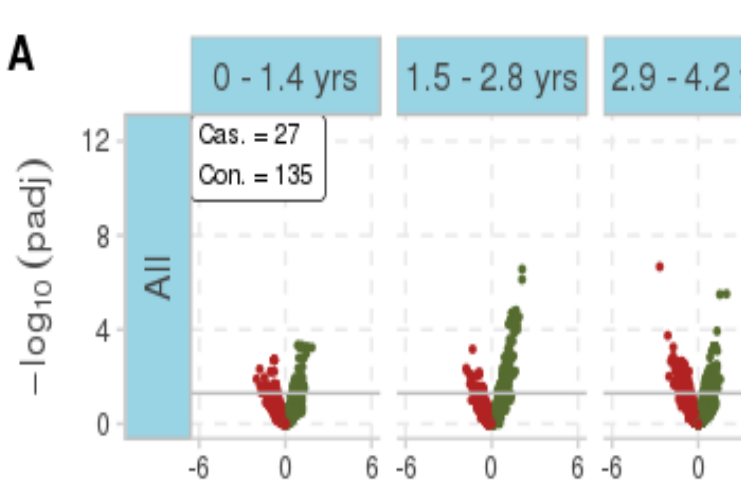
Prediagnostic serum RNA dynamics in lung cancer

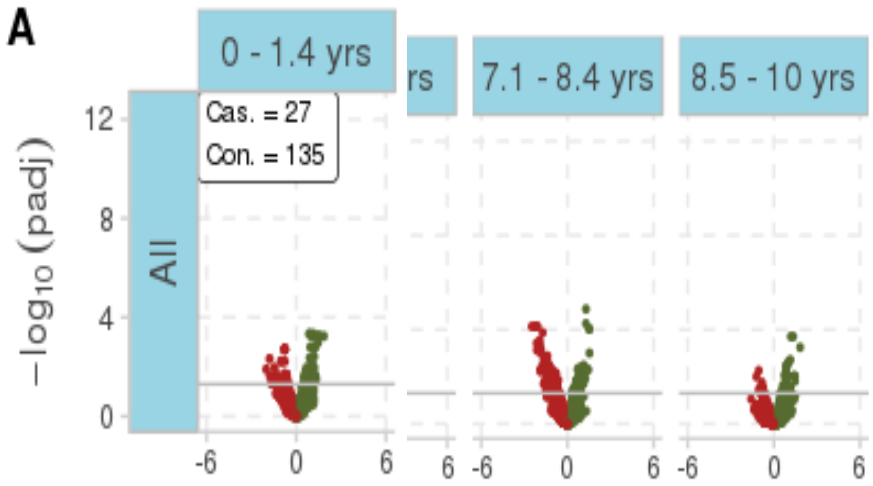


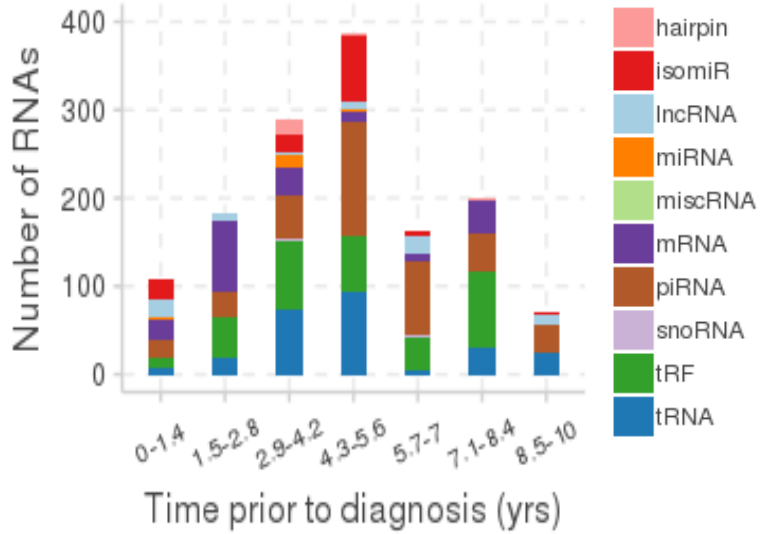

Phase 1
All LC samples
(542 cases and 519 controls)
Phase 2 - 7 time windows
All LC samples
(7 time windows, 27 cases and 135 controls each)
Phase 3 - 7 time windows
Early Stage
(9 cases and 45 controls)
Locally Advanced
(14 cases and 70 controls)
Advanced
(18 cases and 90 controls)
Phase 4 - sliding windows
NSCLC
(84 cases)
NSCLC
(99 cases)
NSCLC
(165 cases)
SCLC
(75 cases)
ADC
(89 cases)

Sample and patient characteristics

		Stage			
	Control	Early (Localized)	Locally Advanced (Regional)	Advanced (Distant)	Unknown
Histology					
NSCLC	-	84	101	171	11
SCLC	-	9	35	76	4
Others	-	10	5	32	4

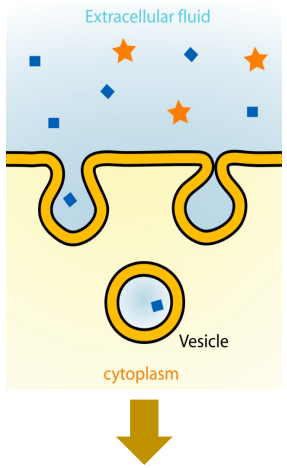


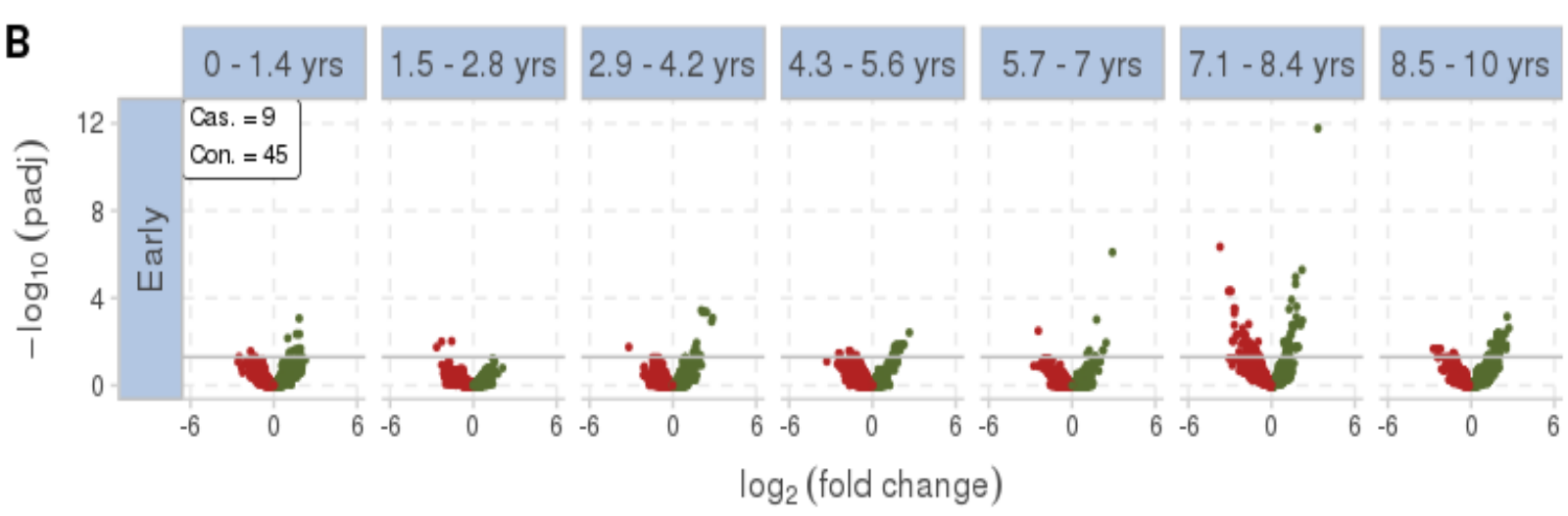


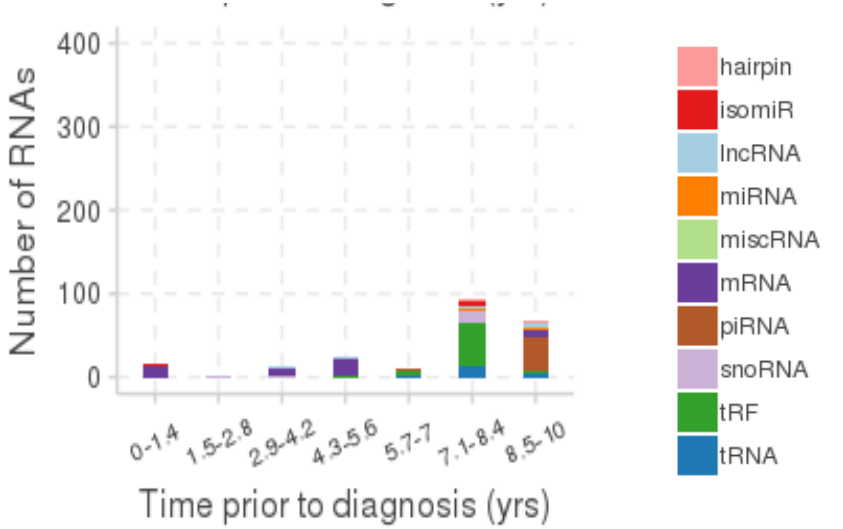


Umu et al. in prep

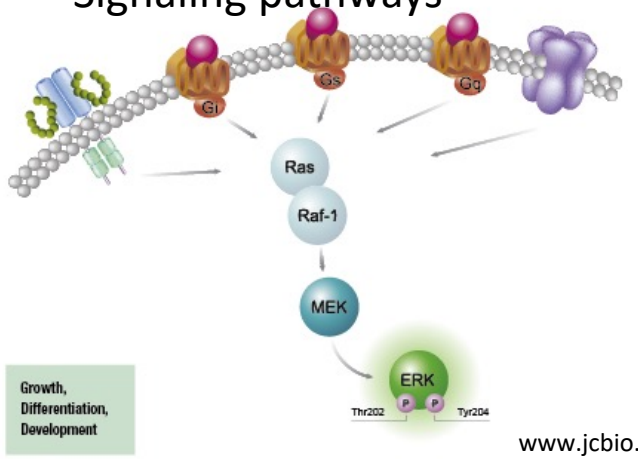
Endocytosis pathways





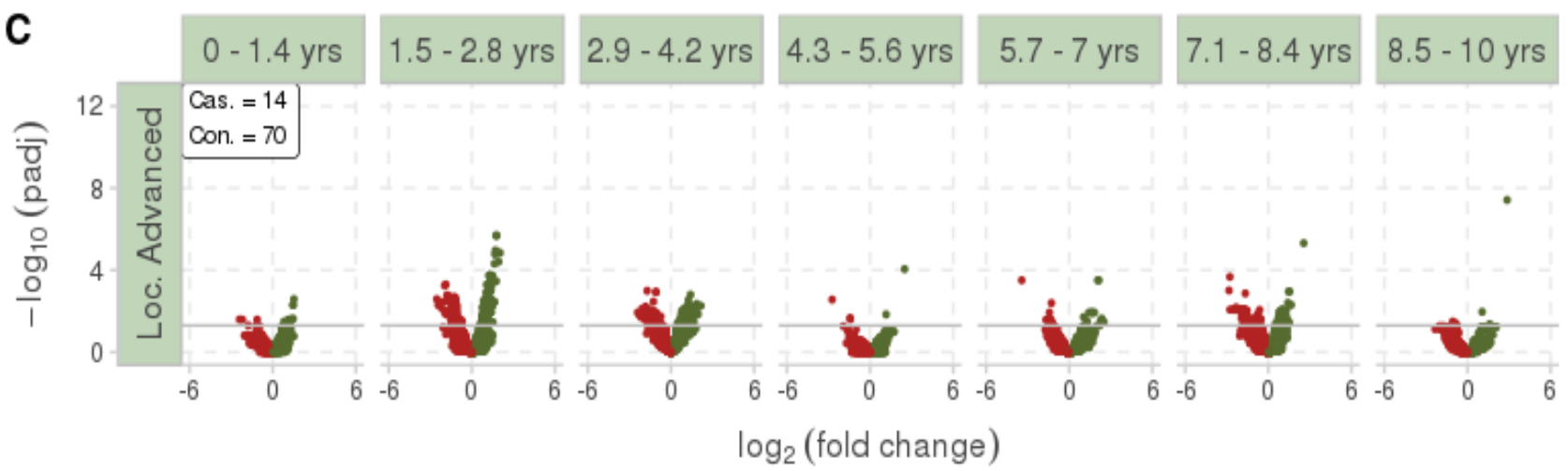


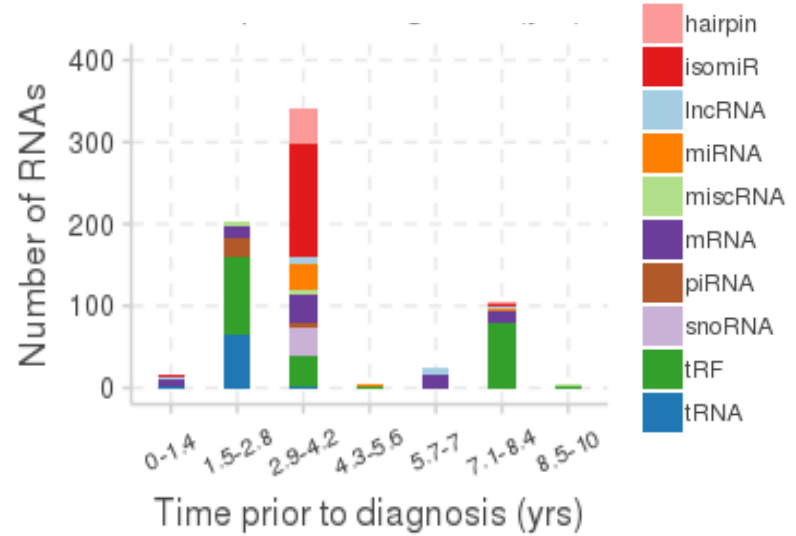
Signaling pathways



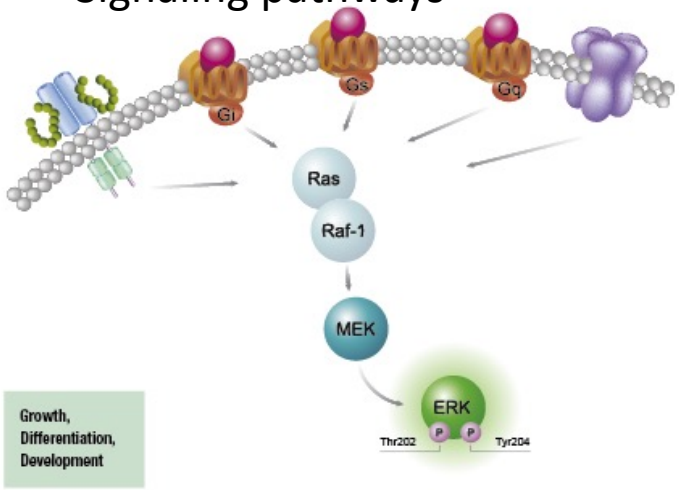
www.jcbio.co.kr/

MAPK, mTOR, ErbB, Ras





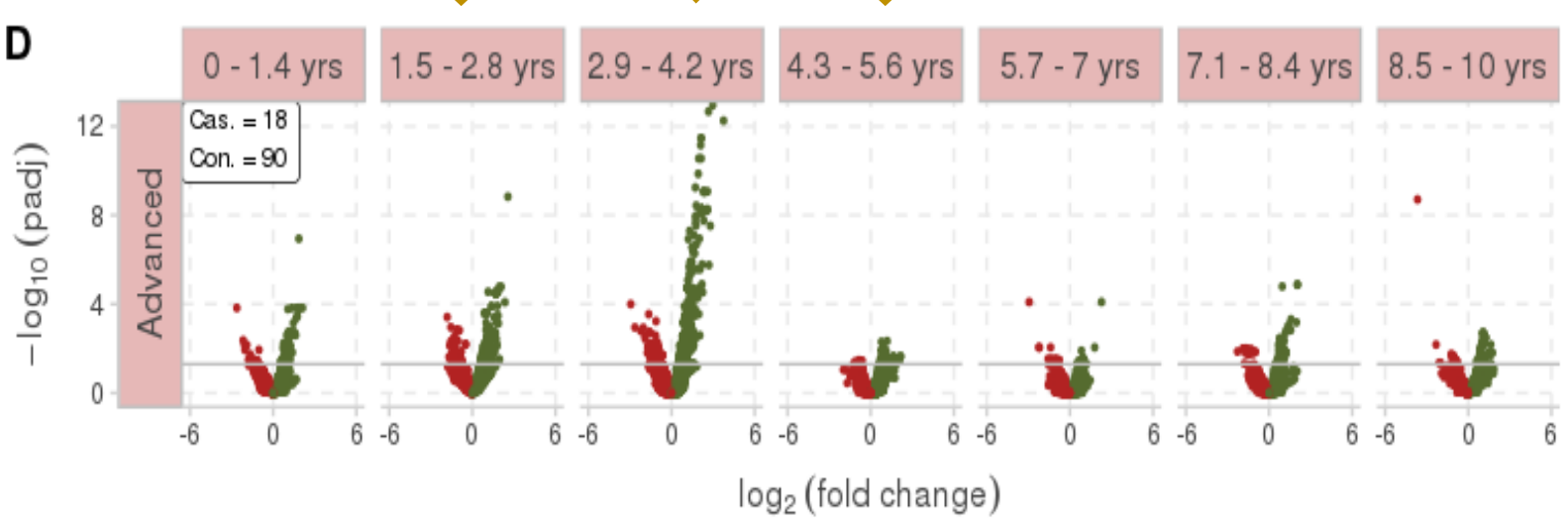
Signaling pathways

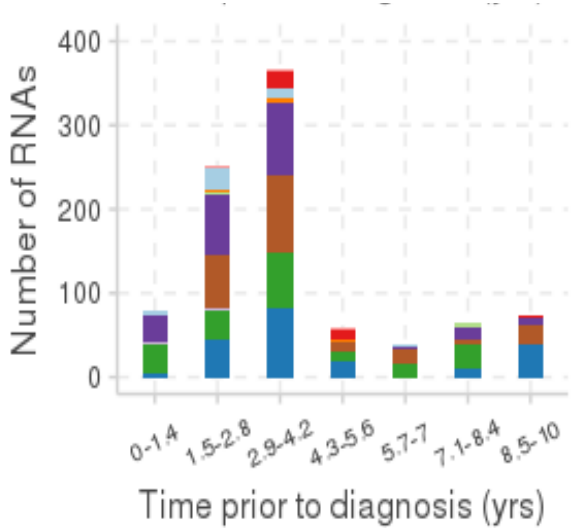


MAPK, ErbB, Ras, mTOR

MAPK, ErbB, Ras

ErbB



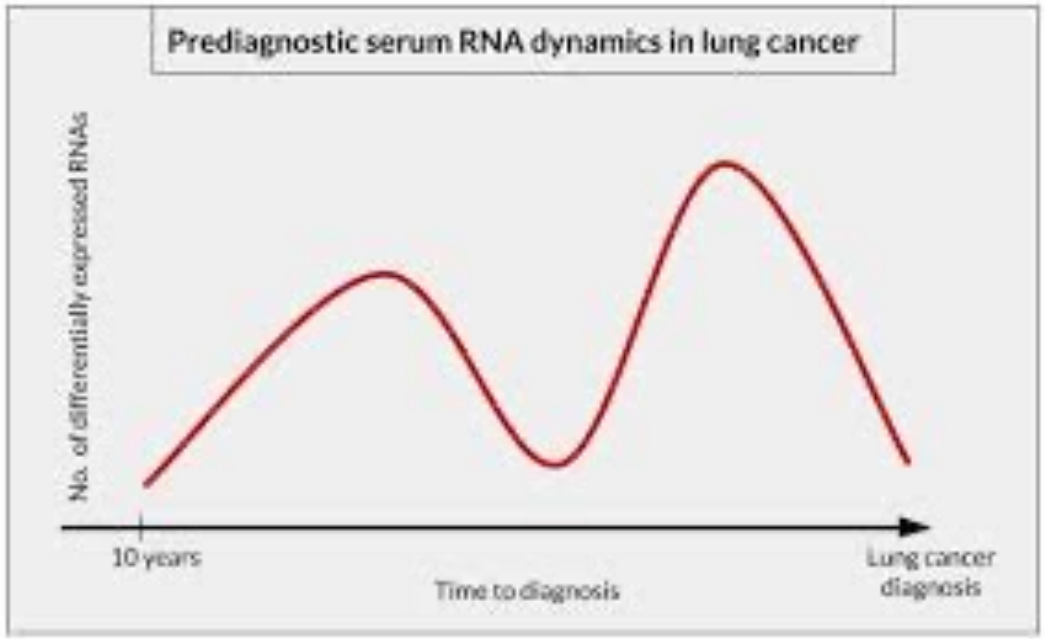


Molecular Oncology

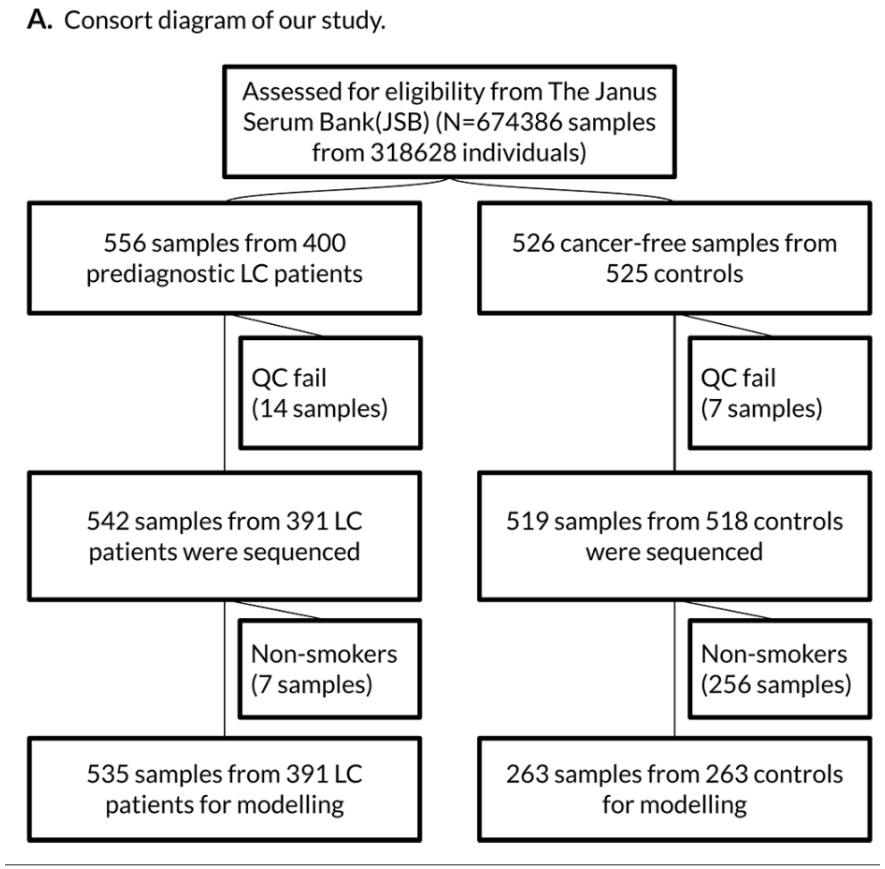
FEBS PRESS science publishing by scientists

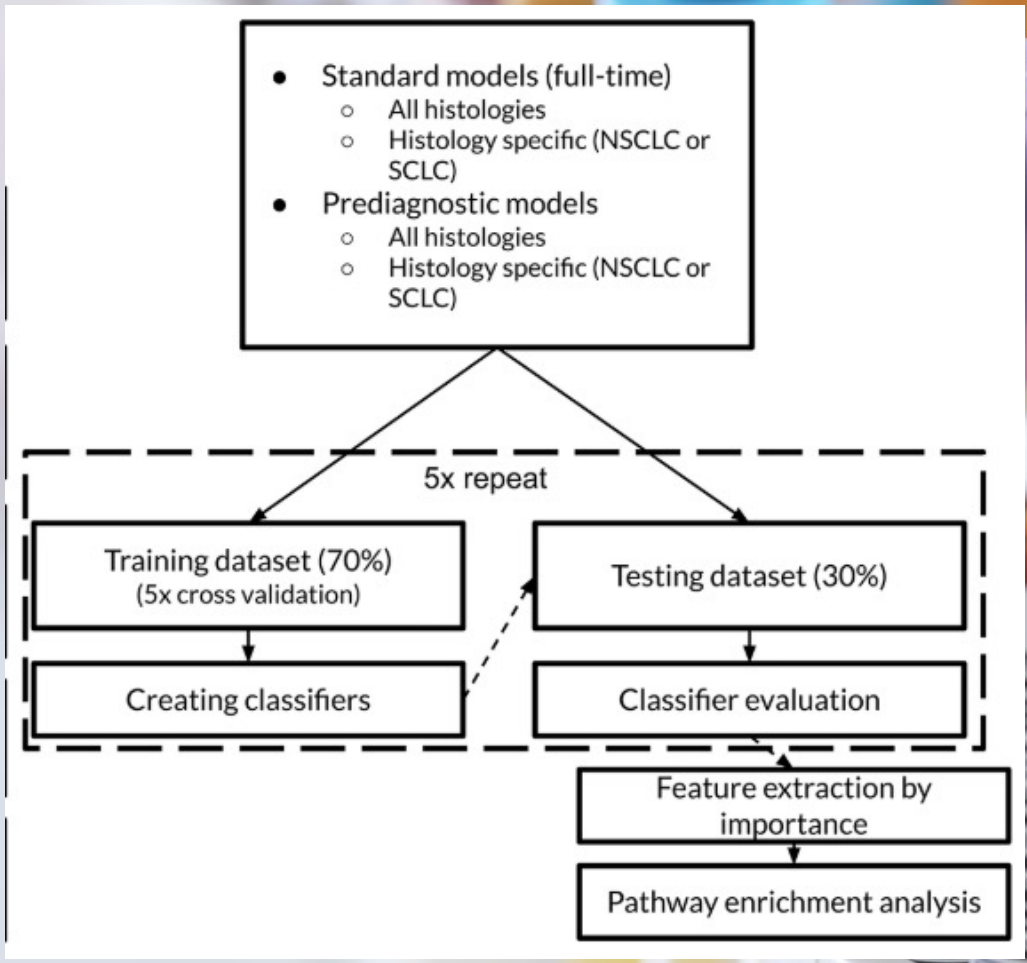
A 10-year prediagnostic follow-up study shows that serum RNA signals are highly dynamic in lung carcinogenesis

Sinan Uğur Umu[1], Hilde Langseth[1], Andreas Keller[2,3], Eckart Meese[4], Åslaug Helland[5,6,7], Robert Lyle[8,9] and Trine B. Rounge[1,10]



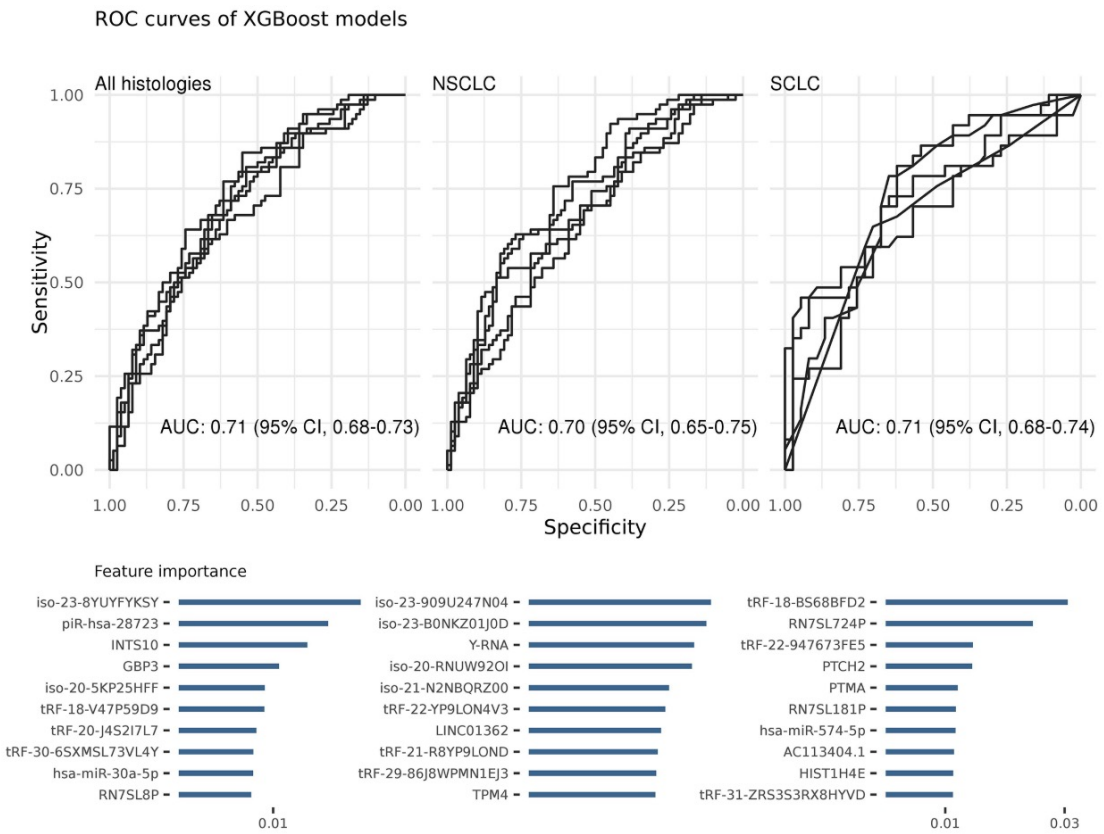
Lung cancer early detection biomarkers

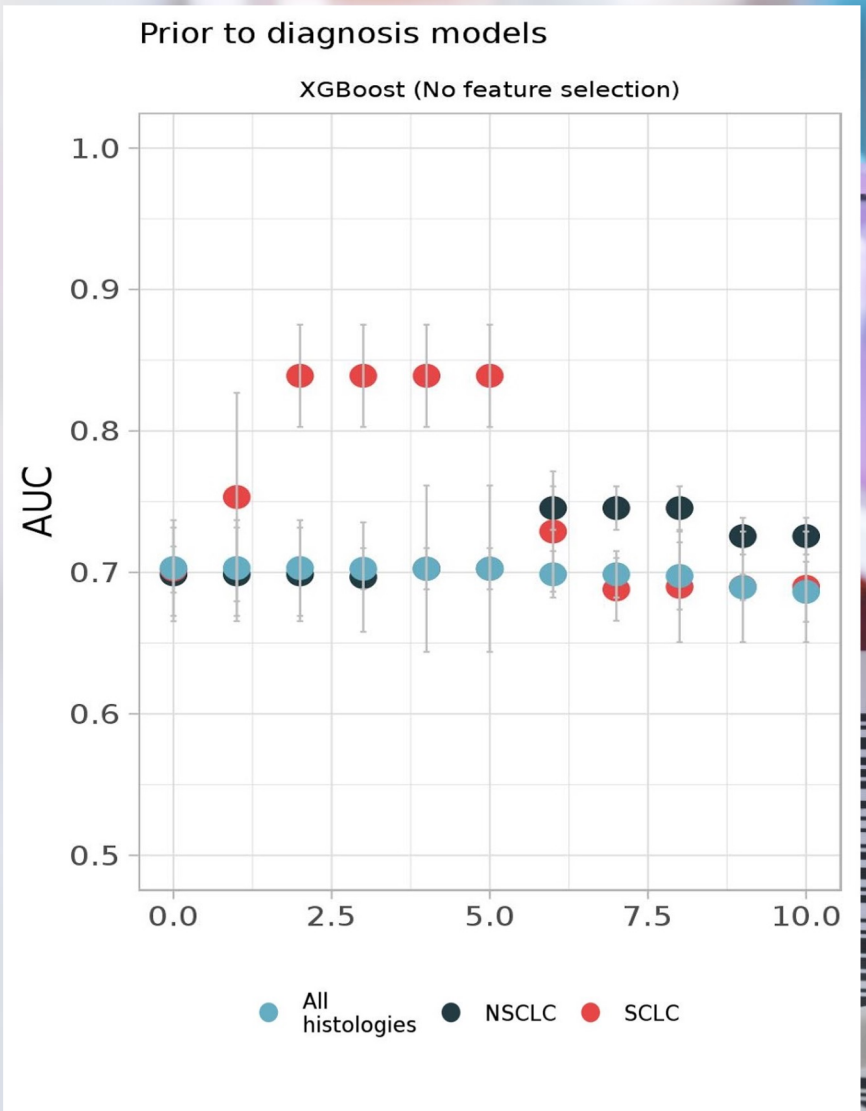




Umu, et al, 2022, elife

Lung cancer early detection biomarkers





Umu, et al, 2022, elife

Summary

- Large-scale serum ncRNA analyses for biomarker research require
  - optimized methods
  - knowledge of technical and biological variation
- smallRNA signals in pre-diagnostic lung cancer serum samples are highly dynamic
  - Signals appear up to 10 years prior to diagnosis
  - Stage and histology specific
  - Disrupts proliferation related signalling pathways
- smallRNAs can predict lung cancer up to 10 years prior to diagnosis
  - The prediction AUC is dependent on time and subtype

Resources

- https://edu.t-bio.info/course/transcriptomics-1/
- https://edu.t-bio.info/course/transcriptomics-2/
- https://edu.t-bio.info/course/transcriptomics-3/
- https://edu.t-bio.info/course/transcriptomics-4/
- https://www.youtube.com/watch?v=WbJ9OA2vevk&feature=youtu.be&ab_channel=PineBiotech
- https://www.youtube.com/watch?v=UFB993xufUU&ab_channel=StatQuestwithJoshStarmer
- https://www.youtube.com/watch?v=Gi0JdrxRq5s&ab_channel=StatQuestwithJoshStarmer
- https://www.youtube.com/watch?v=tlf6wYJrwKY&ab_channel=StatQuestwithJoshStarmer

5. Practical