

Small RNA transcriptomics

Trine B Rounge



Learning outcomes – small RNAseq

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. How to analyse small RNA-sequencing data

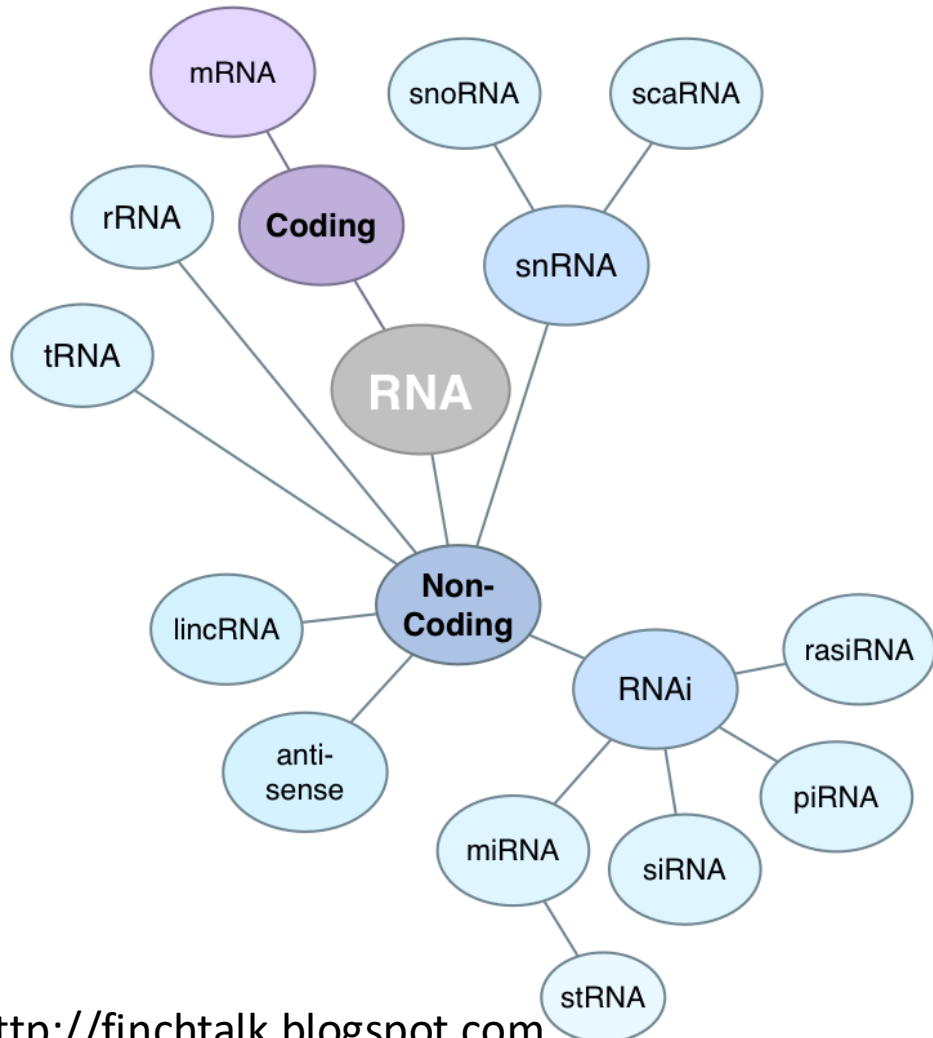
4. Research examples

1. Small RNAs

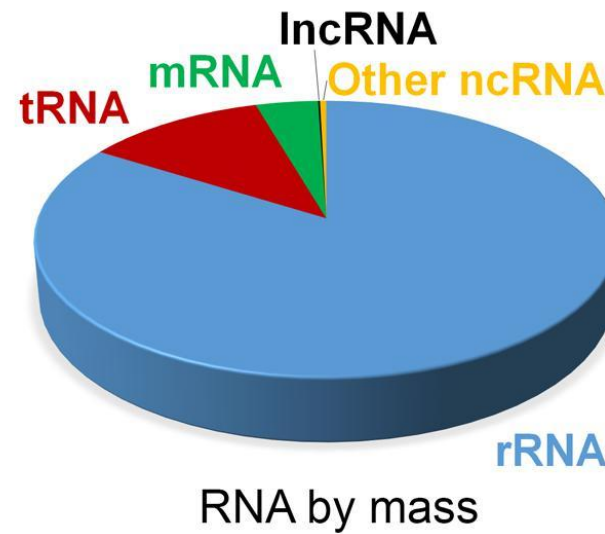


What are small RNAs

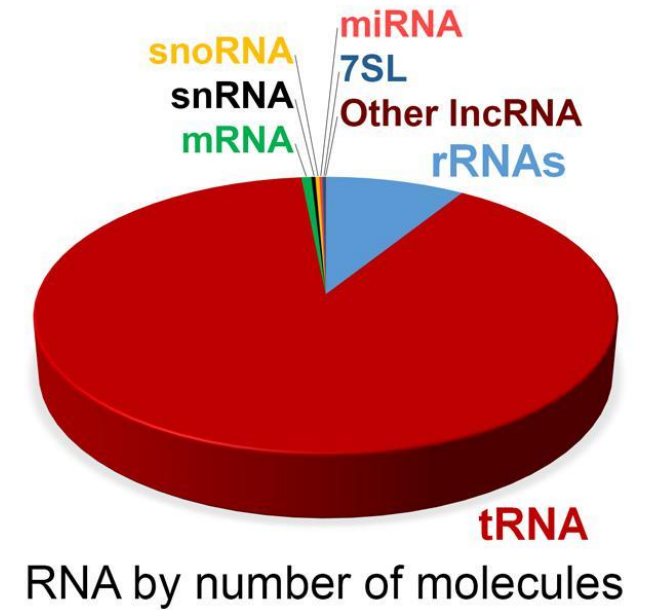
RNA World



A



B

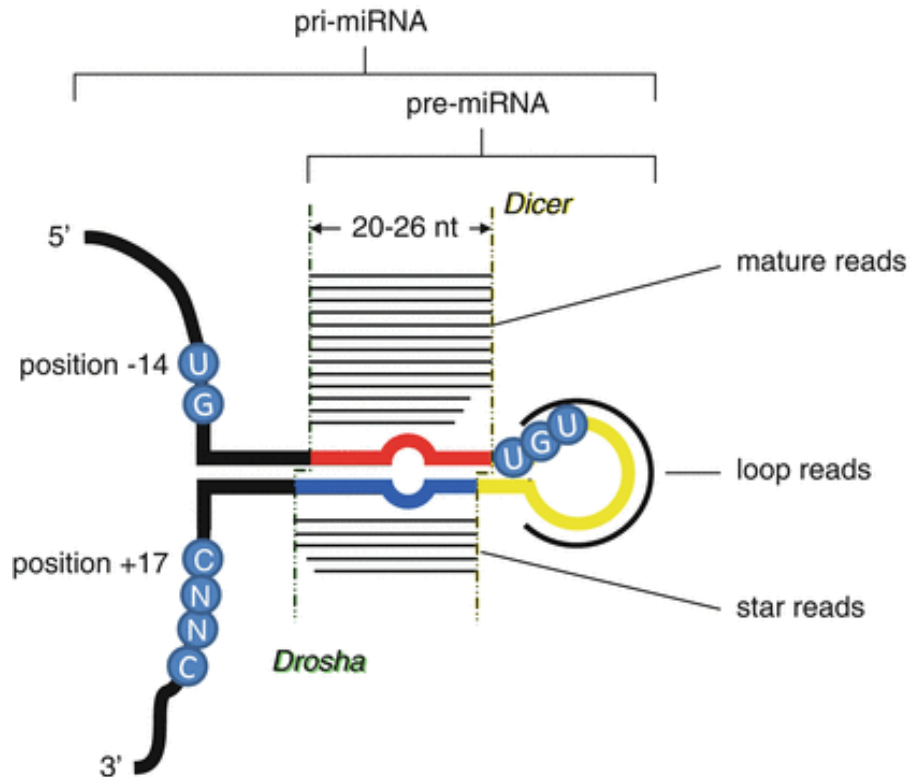


IN-BIOS 5000/9000

Palazzo & Lee, Front Genet, 2015
Small RNAseq

miRNAs

~ 22 nucleotide in length



miRNA - Lee et al, Cell, 1993

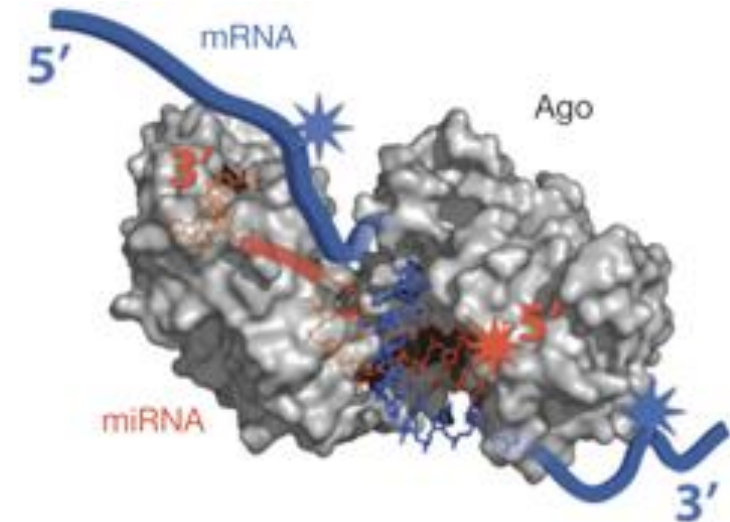
Dicer - Hutvagner et al, Science, 2001

Drosha - Lee et al, Nature, 2003



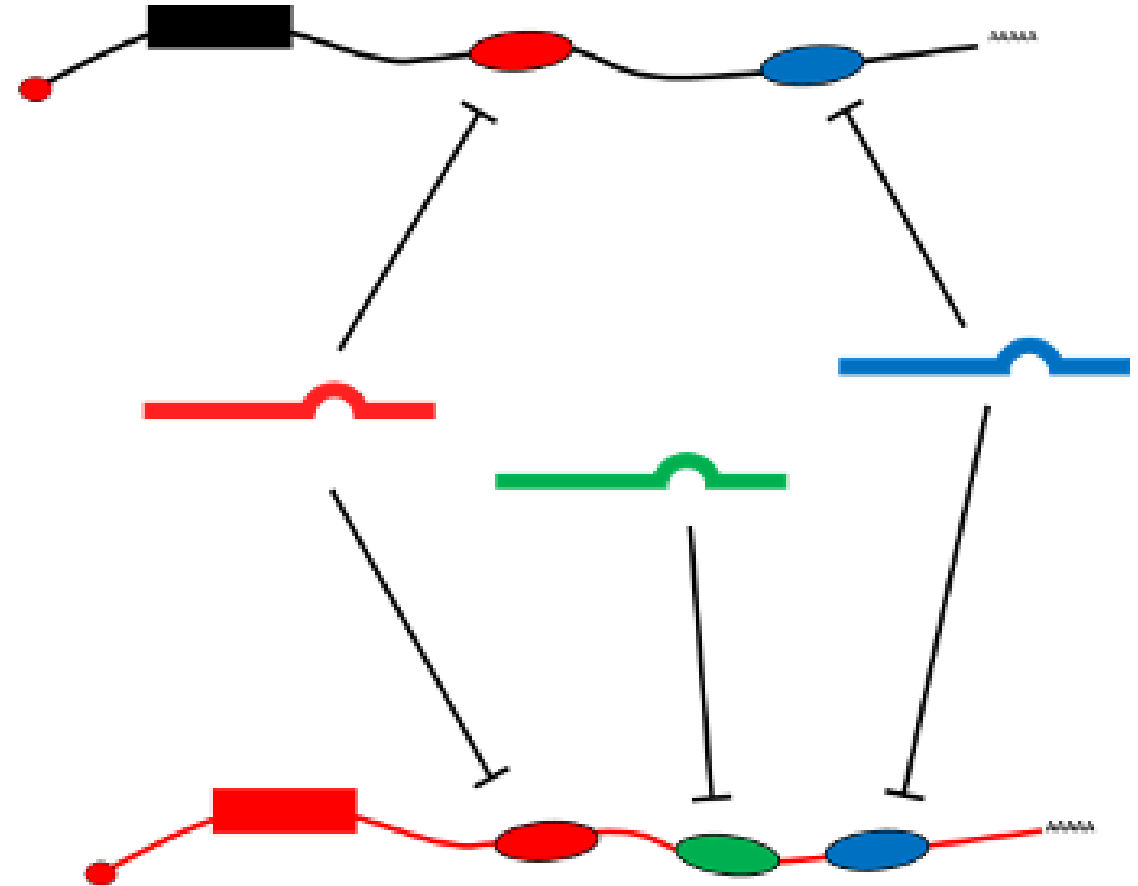
UiO : University of Oslo

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.



Medicine Nobel awarded for gene-regulating 'microRNAs'

Victor Ambros and Gary Ruvkun identified a class of tiny molecules that have a crucial role in controlling gene expression.

By [Ewen Callaway](#) & [Katharine Sanderson](#)



The researchers found that the *lin-4* RNA strand, later called a microRNA, attaches to a stretch of the *lin-14* messenger RNA, preventing the protein from being made through a process known as translation.

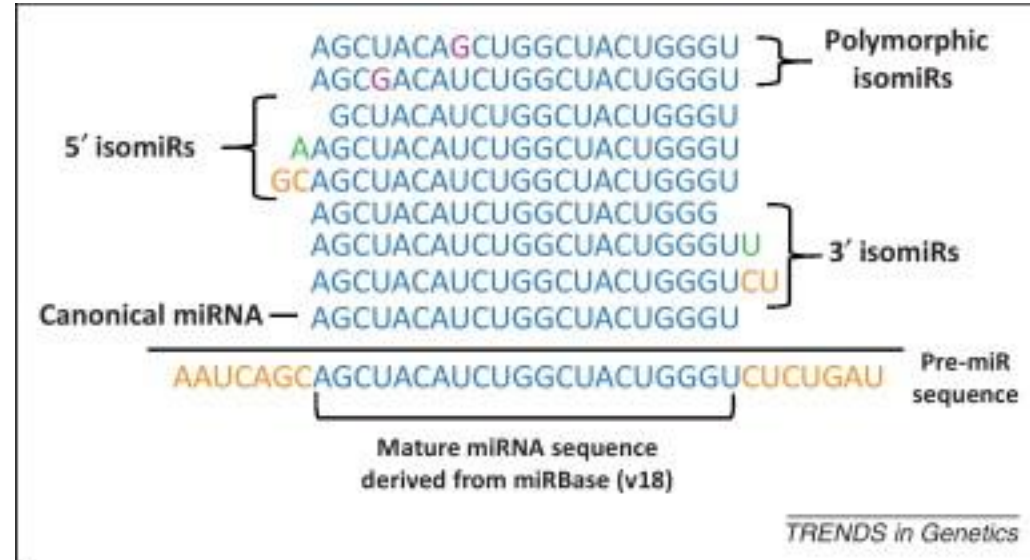
For years, the discovery was viewed as a quirk unique to roundworms, without much relevance to other organisms. That view was shattered in 2000, when Ruvkun's team found that miRNAs was shared by humans, mice and most of the rest of the animal kingdom.

<https://www.nature.com/articles/d41586-024-03212-9>

Small RNAseq

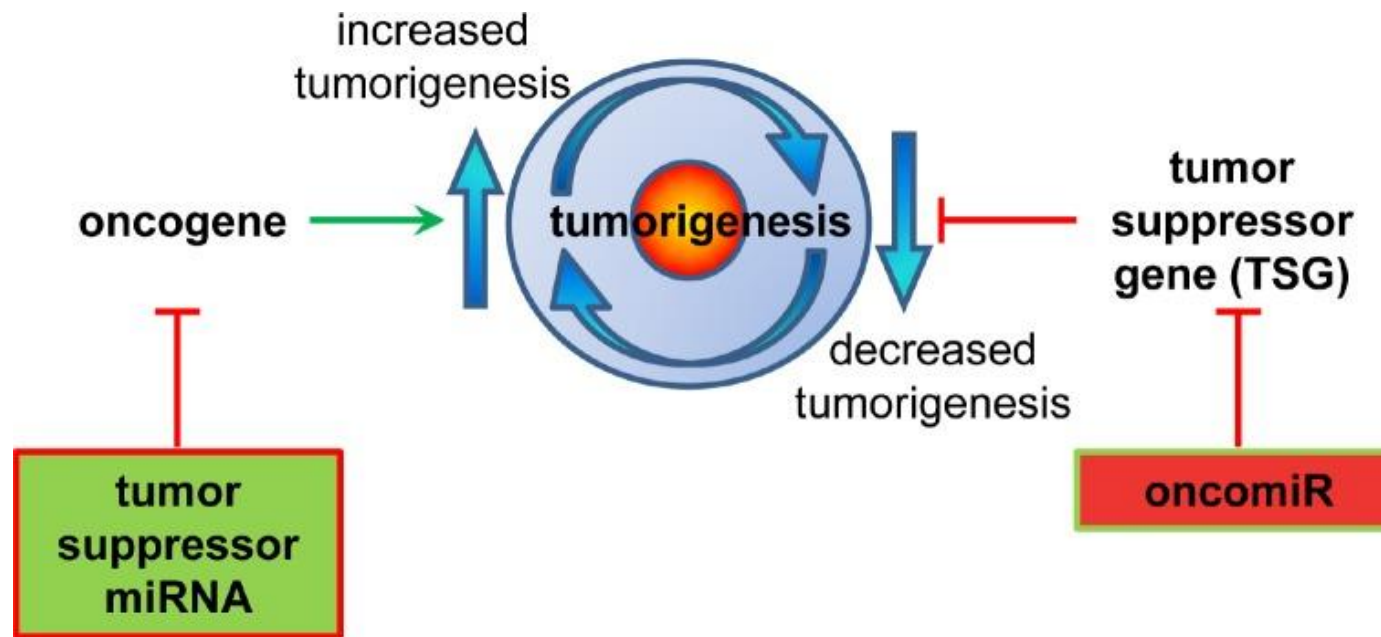
Isoforms

IsomiRs



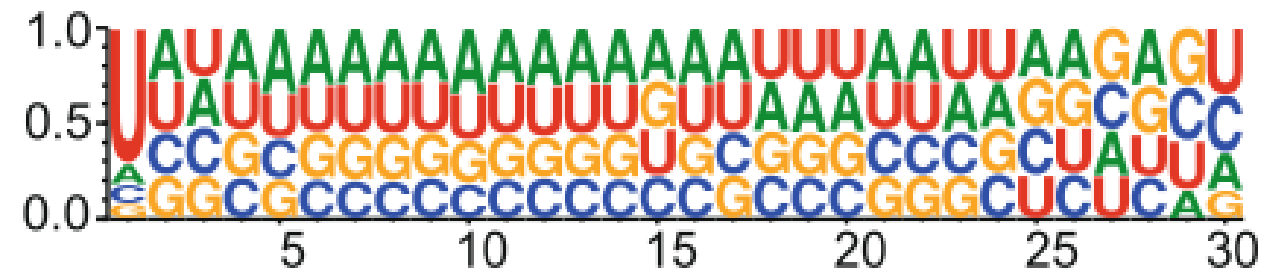
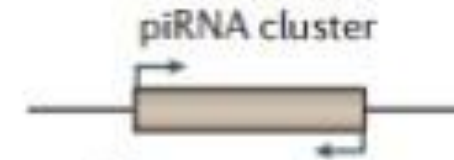
Variability in Dicer and Drosha processing

miRNA expression in cancer



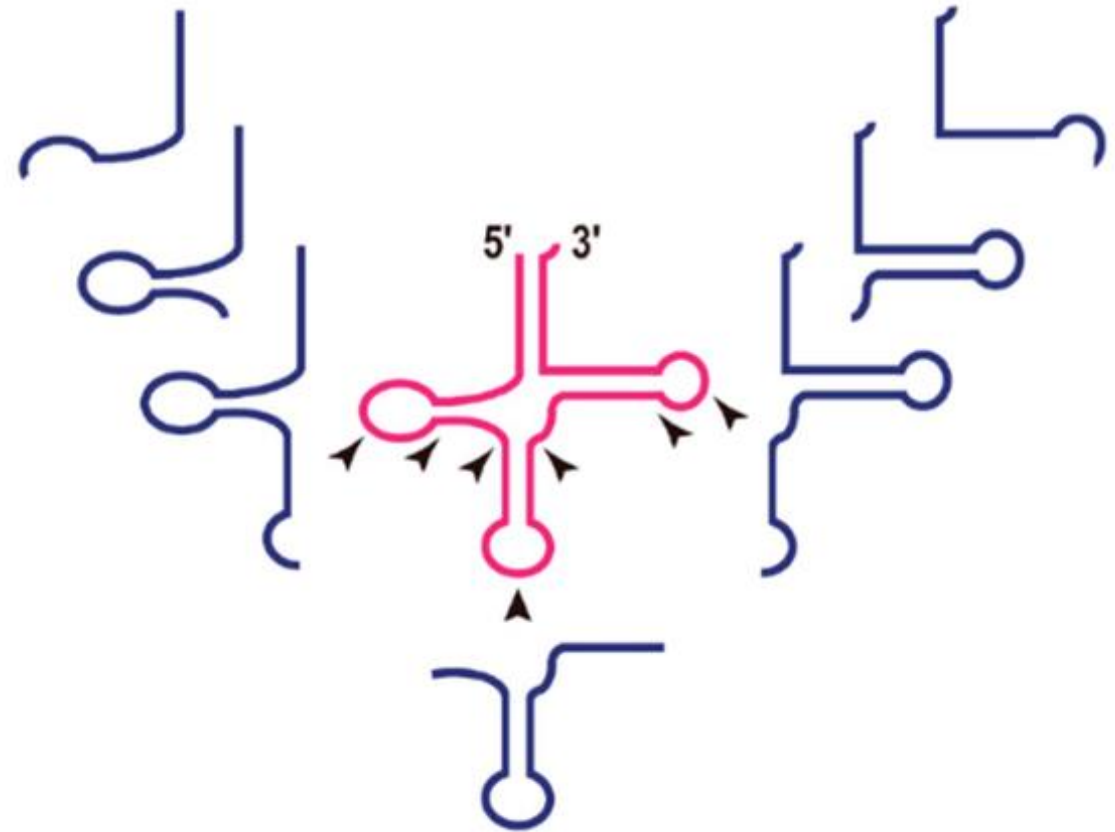
piRNA

- 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



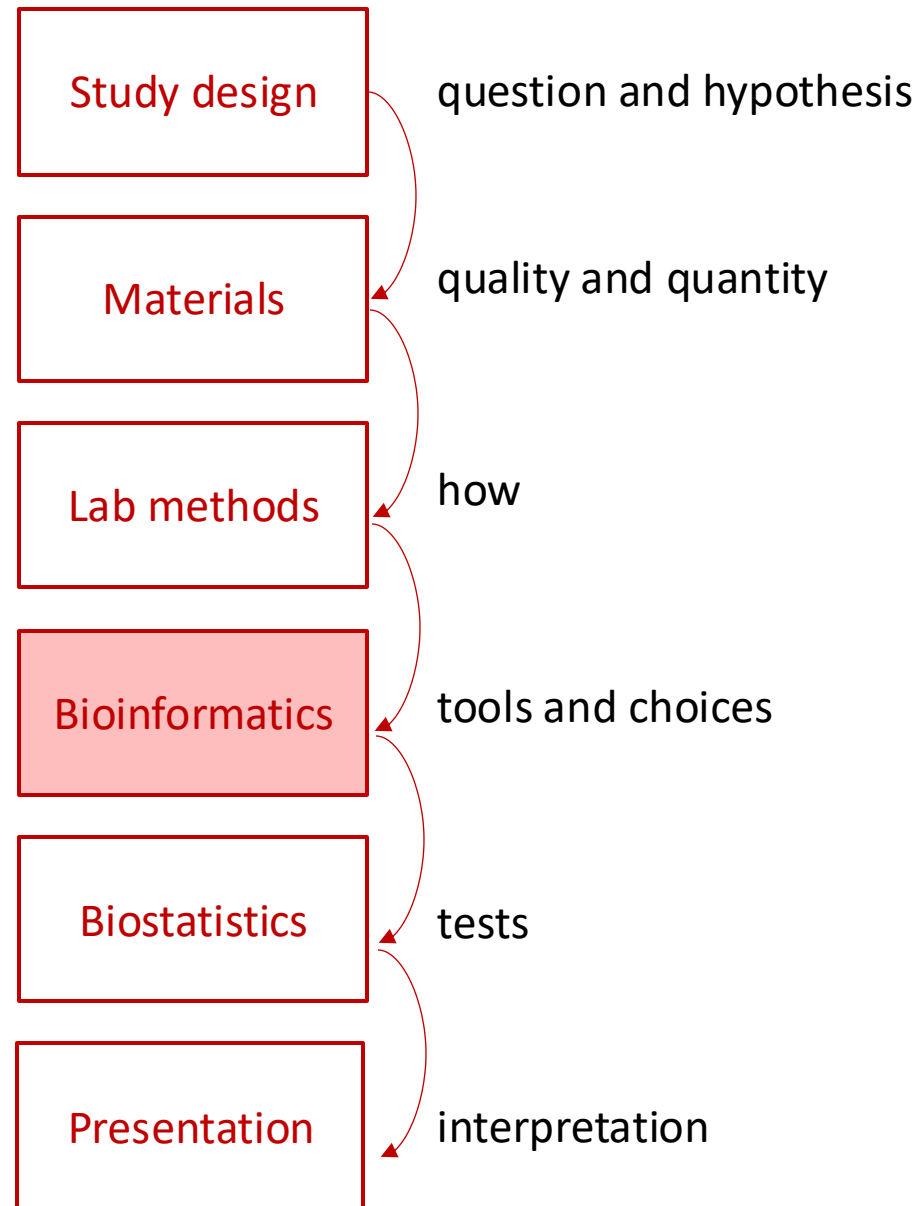
tRNA fragments

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



2. Small RNAs transcriptomics





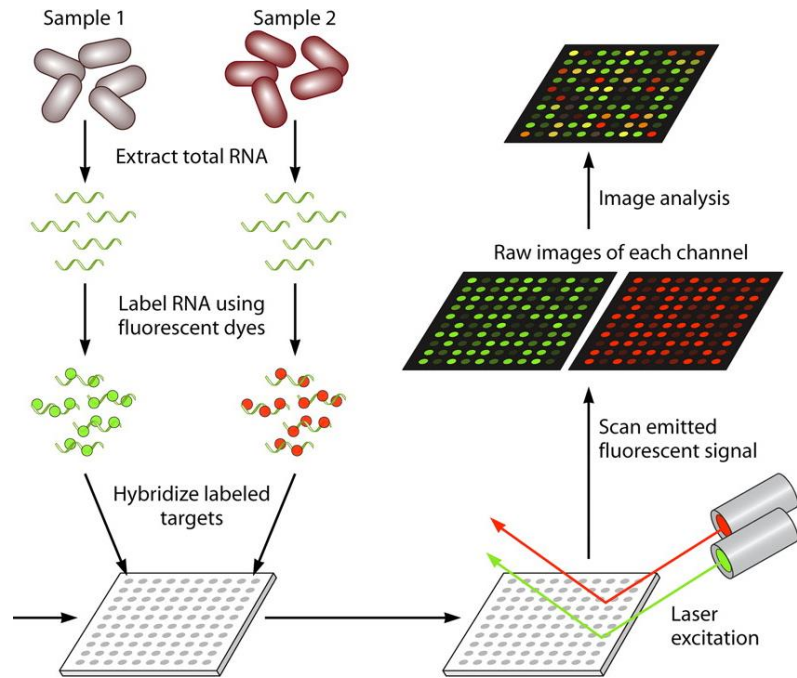
Small RNA transcriptomics

- Study of the the complete set of small RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell—using high-throughput methods
- Transcribed as small RNAs
- Processed to small RNA
- Degraded to small RNA

Experimental design

- **Biological question**
- Species specific information
 - Genome and annotation available
- Sample variation and quality
- **Technology**
 - Technical variation
 - Technical bias
- Depth vs sample size/replicates
- Data analyses

Small RNA Technologies



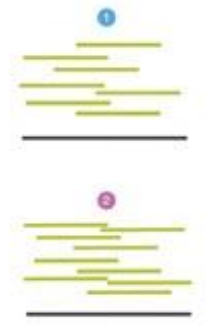
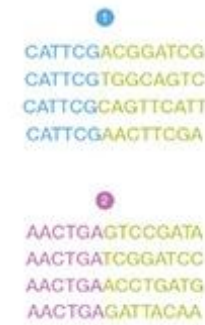
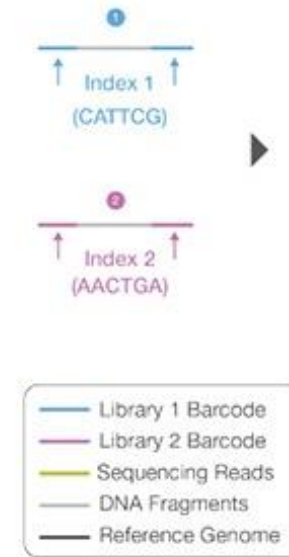
Library Preparation

Pool

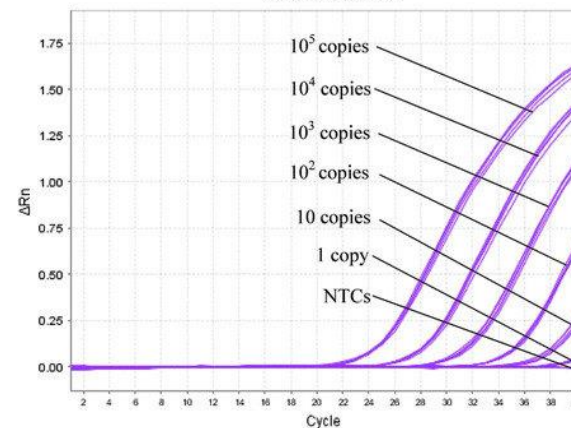
Sequence

Demultiplex

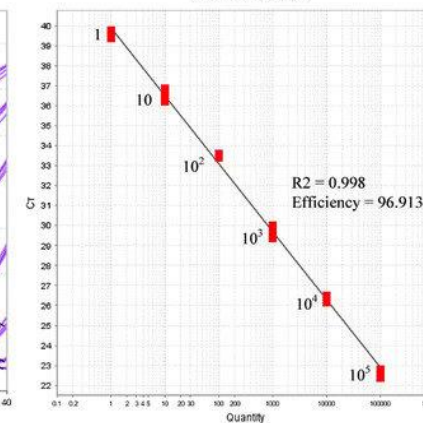
Align



Amplification Plot



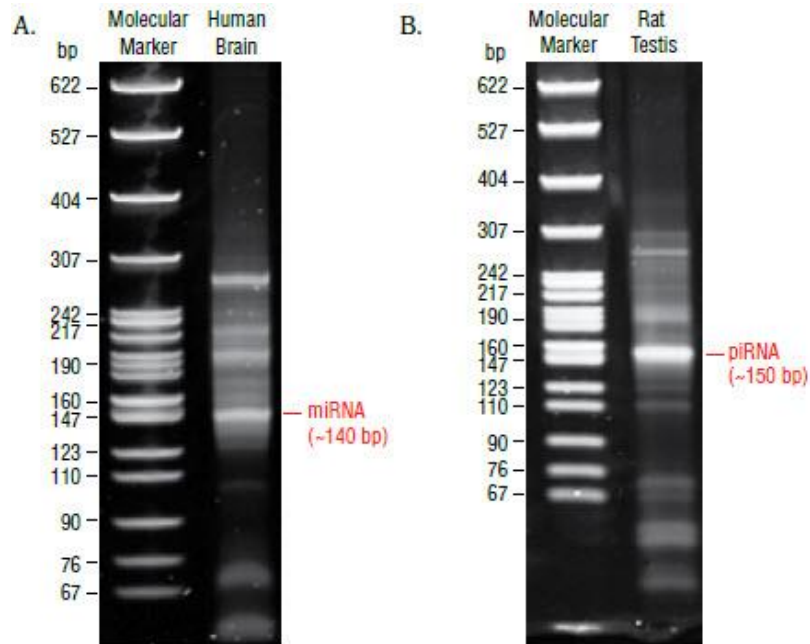
Standard Curve



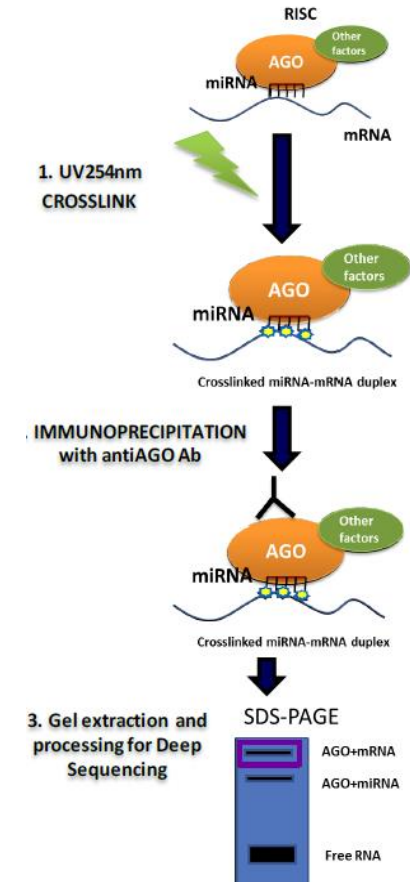
Small RNAseq

Small RNA transcriptomics

Size selection

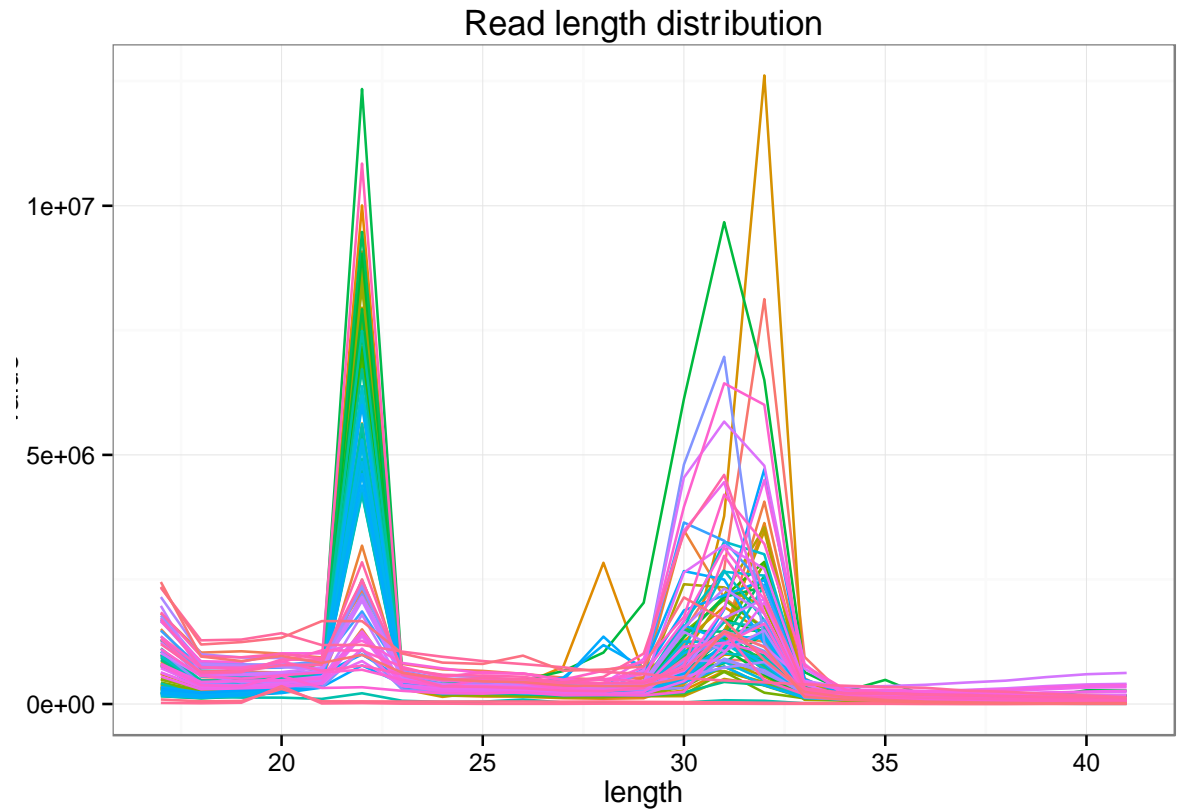


AGO immunoprecipitation

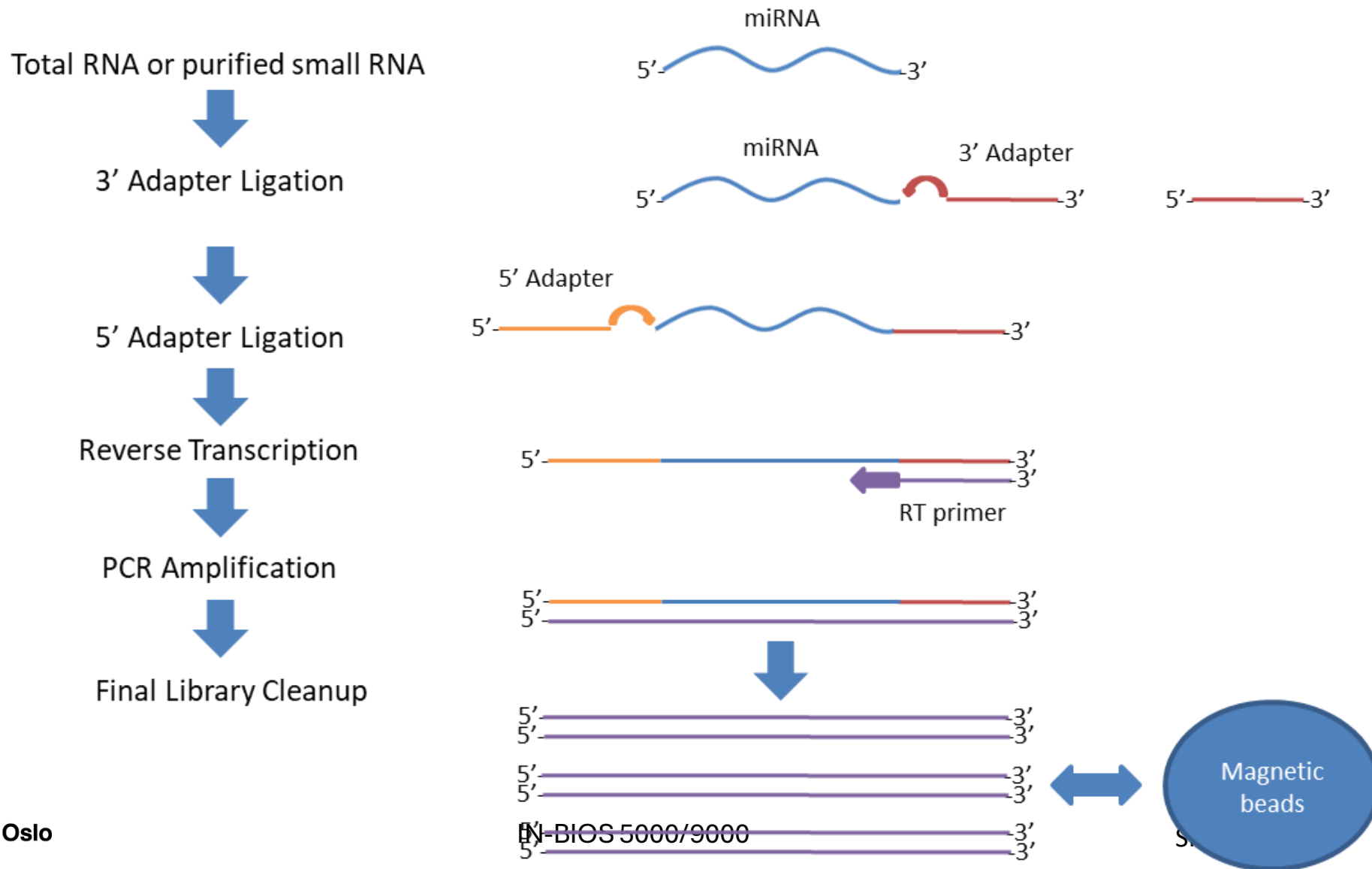


Small RNA transcriptomics

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



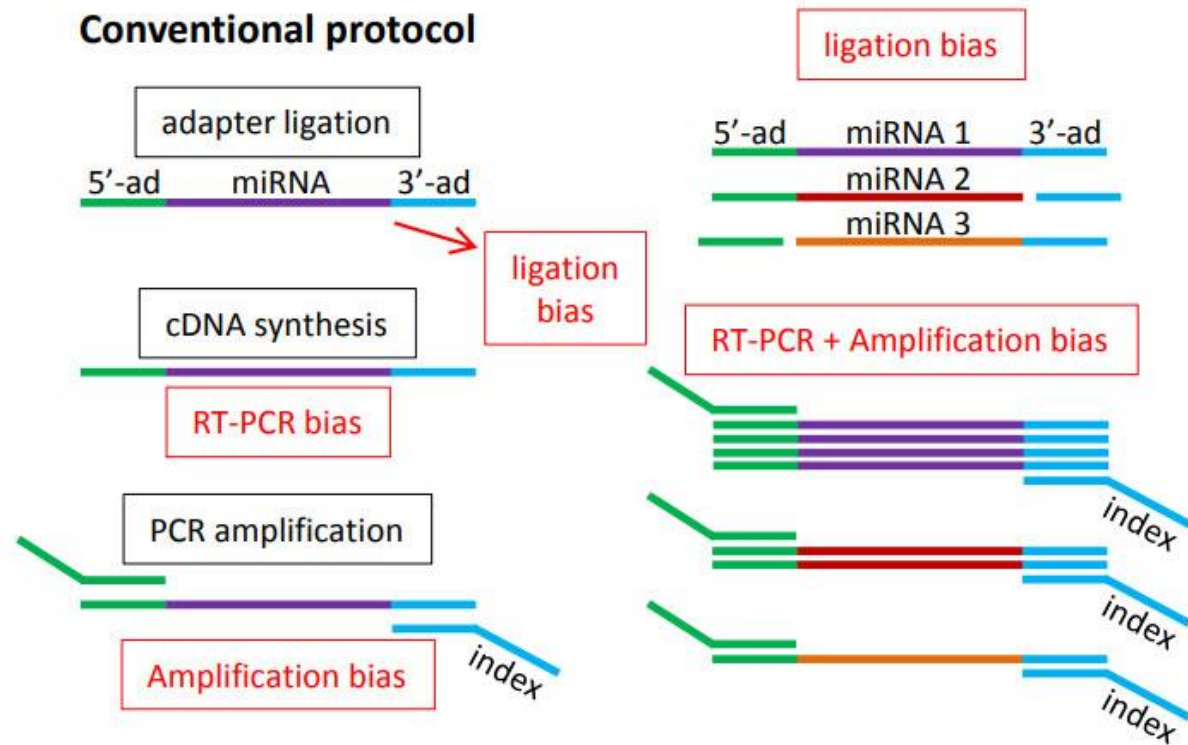
Small RNA sequencing – library preparation



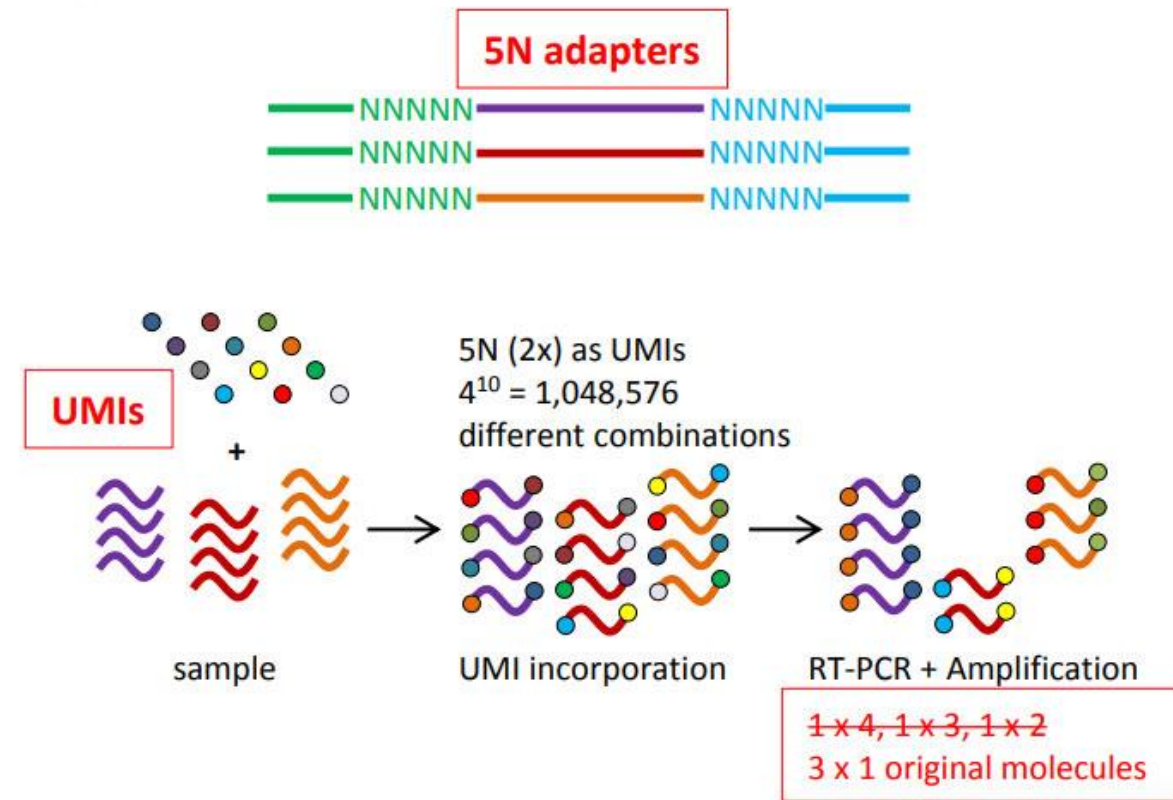
UMIs in small RNA seq

A

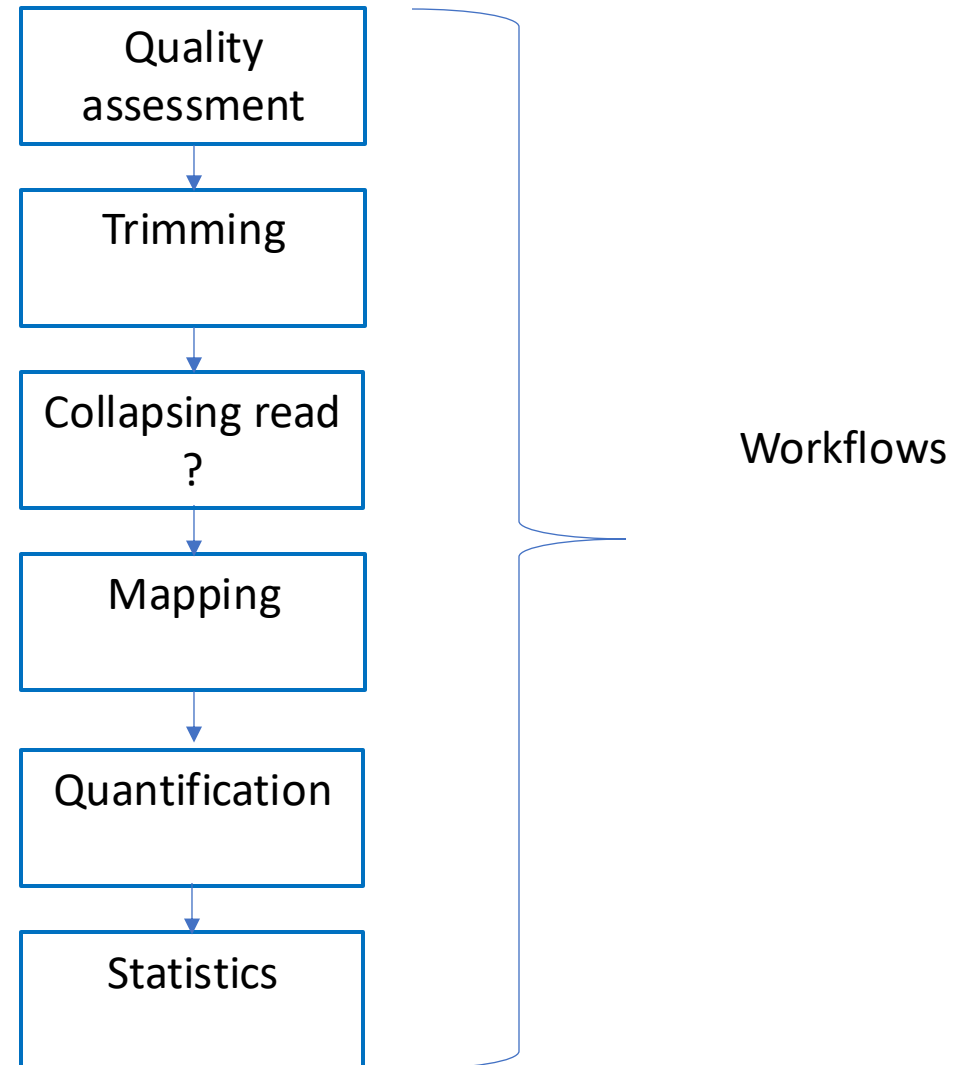
Conventional protocol



B



Analyse small RNA sequencing data



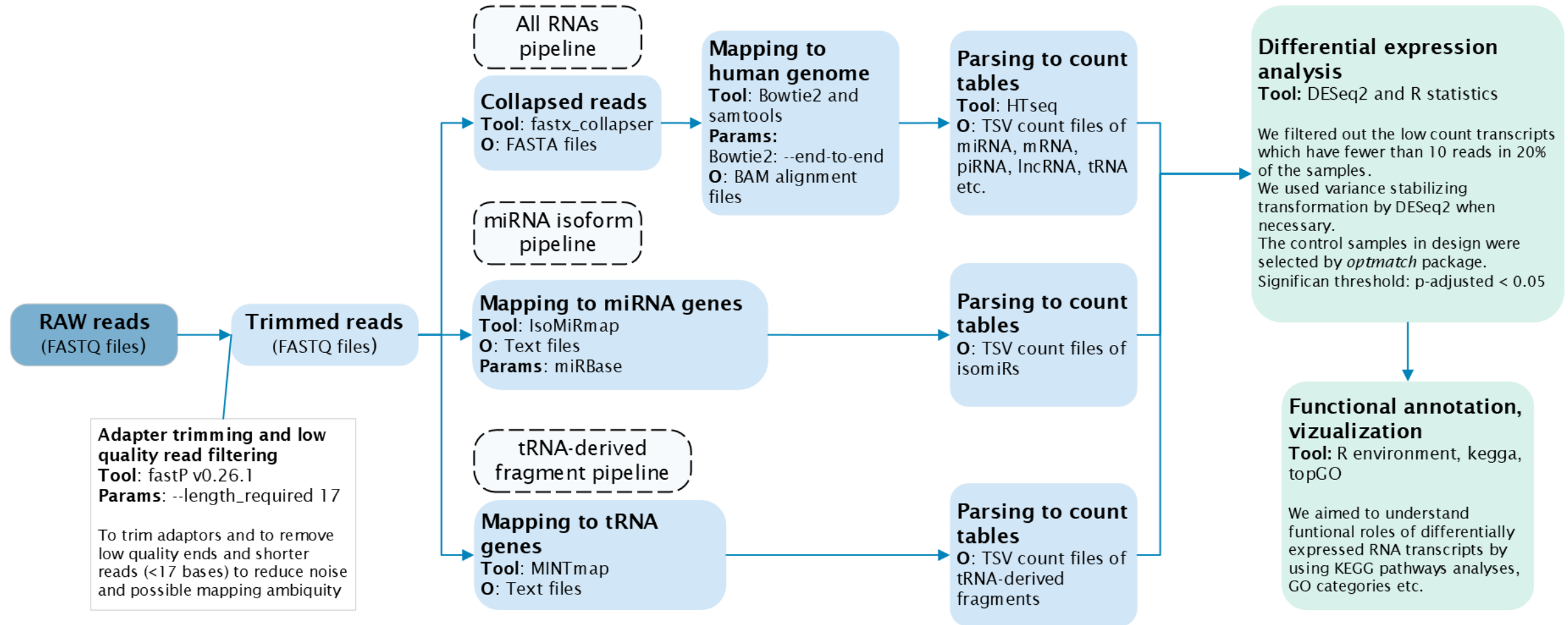
IN-BIOS 5000/9000

Small RNAseq

Small RNAseq workflow

From Fastq files to small RNA read counts

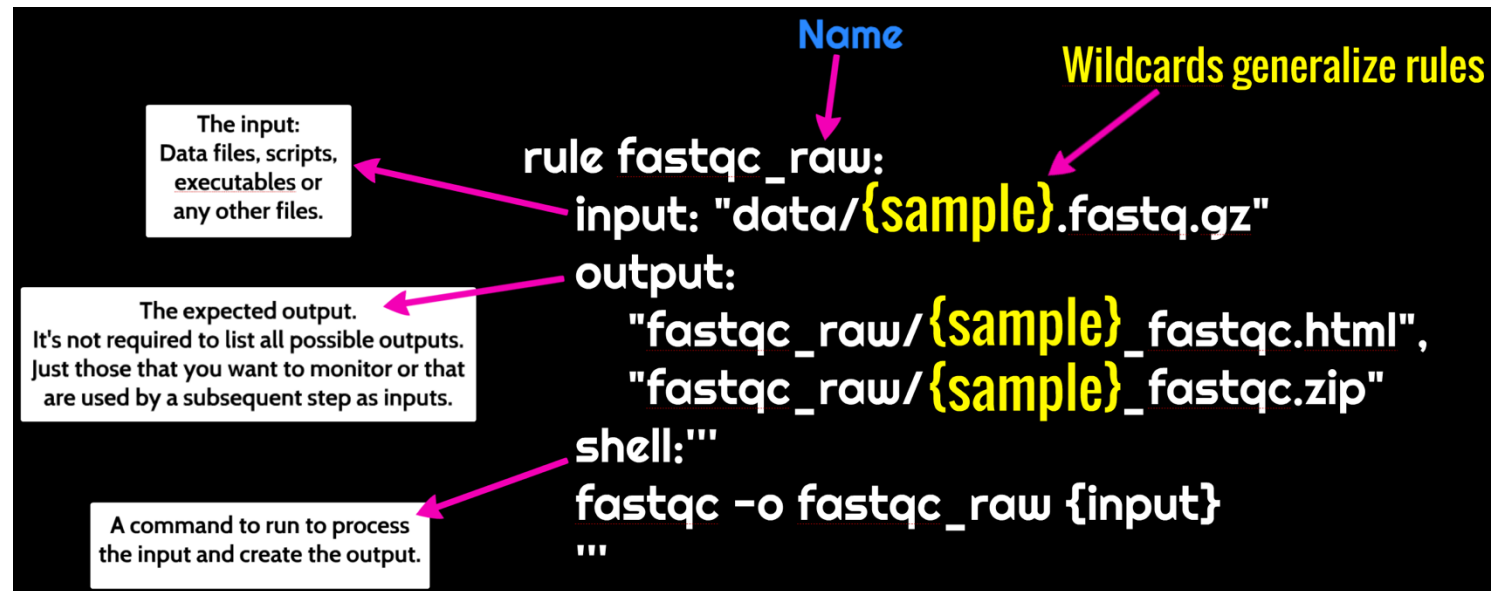
- Reproducible
- Detect isoforms
- Detect most small RNA classes
- Efficient and can use multiple cores



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

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/>

Nextflow is built around the idea that Linux is the *lingua franca* of data science.



Fast prototyping

Nextflow allows you to write a computational pipeline by making it simpler to put together many different tasks.

You may reuse your existing scripts and tools and you don't need to learn a new language or API to start using it.

Reproducibility

Nextflow supports [Docker](#) and [Singularity](#) containers technology.

This, along with the integration of the [GitHub](#) code sharing platform, allows you to write self-contained pipelines, manage versions and to rapidly reproduce any former configuration.

Portable

Nextflow provides an abstraction layer between your pipeline's logic and the execution layer, so that it can be executed on multiple platforms without it changing.

It provides out of the box executors for GridEngine, SLURM, LSF, PBS, Moab and HTCondor batch schedulers and for [Kubernetes](#), [Amazon AWS](#), [Google Cloud](#) and [Microsoft Azure](#) platforms.

Unified parallelism

Nextflow is based on the *dataflow* programming model which greatly simplifies writing complex distributed pipelines.

Parallelisation is implicitly defined by the processes input and output declarations. The resulting applications are inherently parallel and can scale-up or scale-out, transparently, without having to adapt to a specific platform architecture.

Continuous checkpoints

All the intermediate results produced during the pipeline execution are automatically tracked.

This allows you to resume its execution, from the last successfully executed step, no matter what the reason was for it stopping.

Stream oriented

Nextflow extends the Unix pipes model with a fluent DSL, allowing you to handle complex stream interactions easily.

It promotes a programming approach, based on functional composition, that results in resilient and easily reproducible pipelines.

<https://github.com/nf-core/smrnaseq>

<https://www.nextflow.io/>

Preprocessing FASTQ files

- Quality assessment and trimming is very similar to RNA seq
 - Check read lengths after trimming to assess if the library is matching your target RNA classes
- Collapsing reads
 - Many identical sequences – RNAs
 - Saves compute time to collapse these
 - Add copy number to sequence header

Bioinformatics

Quality
assessment

Trimming

Collapsing read
?

Mapping and counting

- Mapping
 - Mapper and setting must be adapted to short sequences
 - Bowtie2 – in end-to-end mode
 - BWA
 - Bowtie1
- Reference for mapping
 - Whole genome
 - RNA database
 - MirBase
 - Mirgenedb
- Counting
 - Featurecount
 - HTSeq

Bioinformatics

Mapping

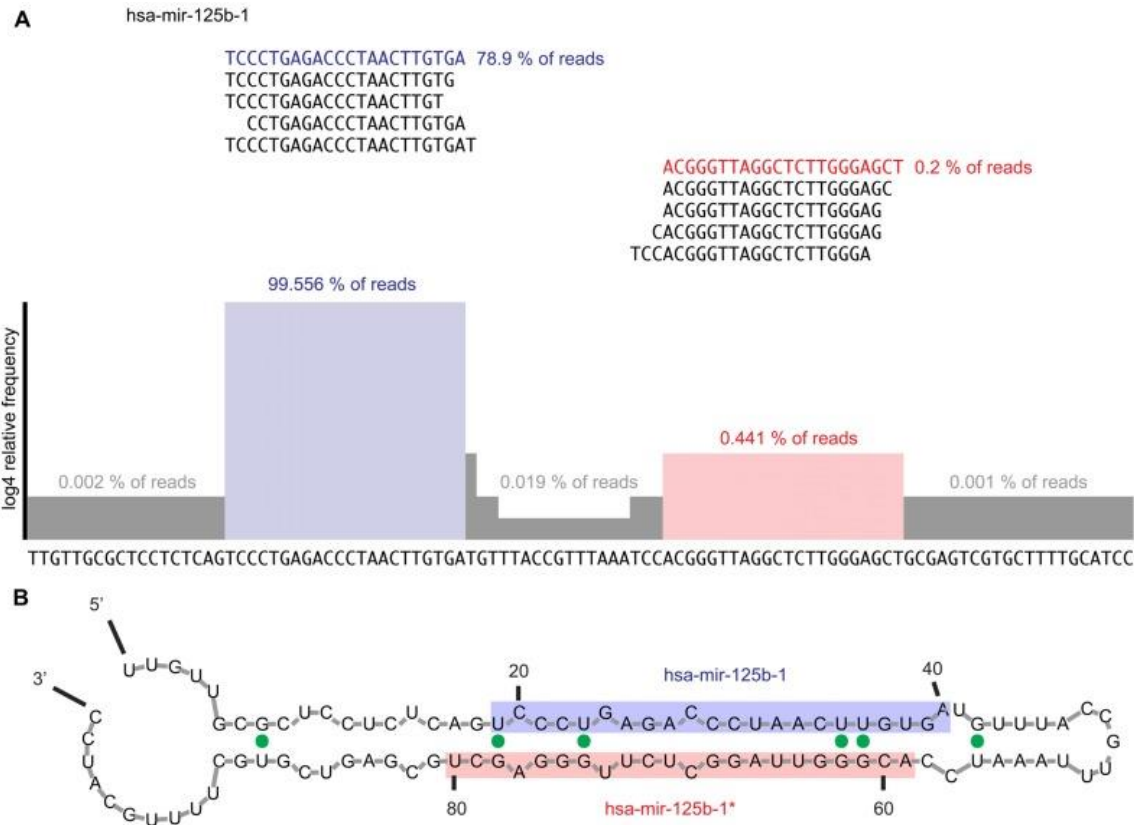


Quantification

Small RNA count data

- Similar to mRNA count data
- Each class analysed separately
 - 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

IN-BIOS 5000/9000

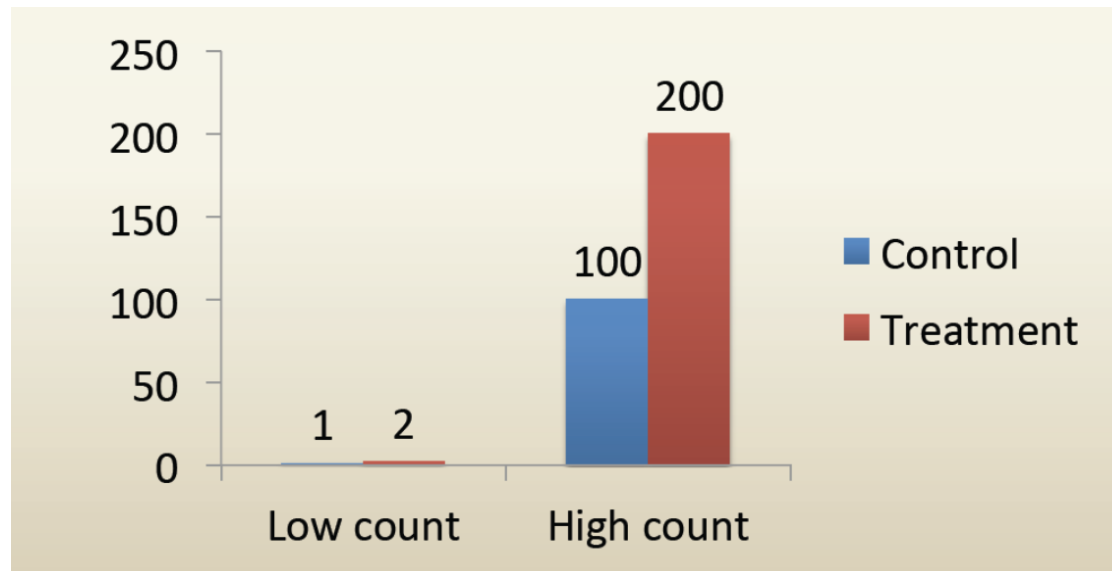
Small RNAseq

Normalization

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

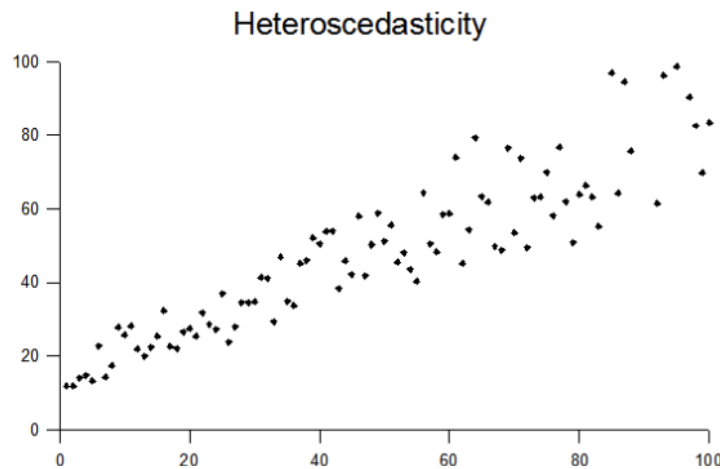
Noise for low counts

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

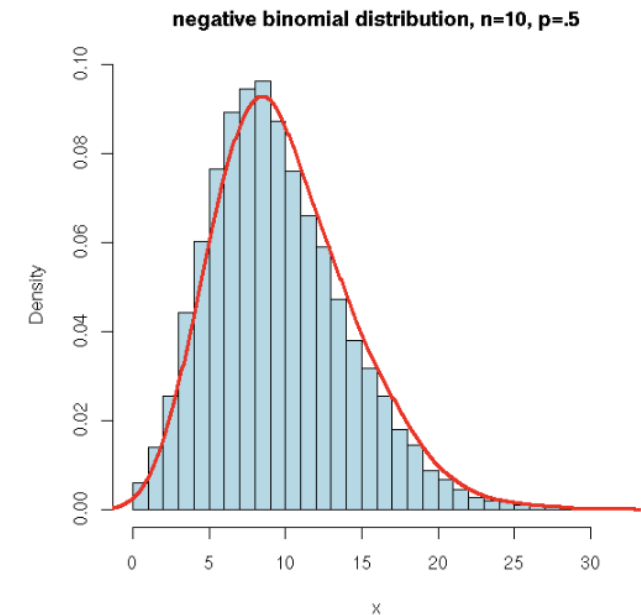


Challenges with using count data

- Large dynamic range – from zero to millions
- Variance is not equal across range of counts
- Positive integers and non-symmetric distribution – not normally distributed

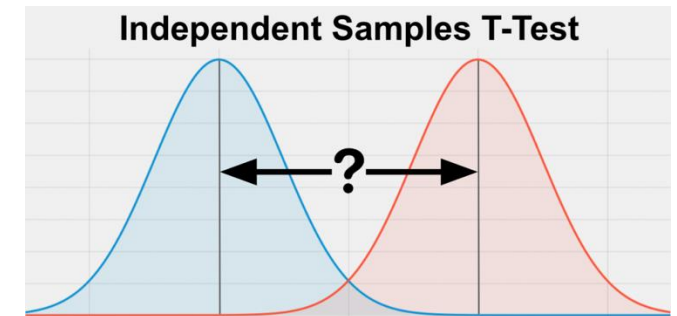


IN-BIOS 5000/9000



Differential gene expression (DE) analyses

- Statistically testing if gene expression differ
 - Between conditions
 - With continuous variable
- Wald test – DESeq2 – similar to t-test
- Assumptions: most genes are not DE
- Null hypothesis: the same abundances across groups/conditions



Gene enrichment analyses

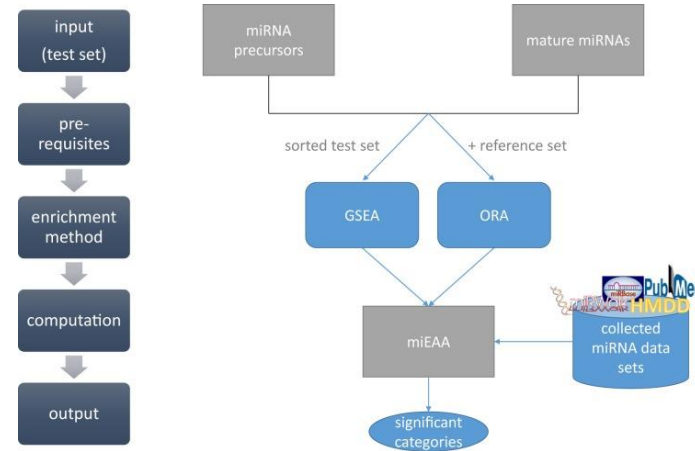
Bioinformatics

Statistics

DE miRNA

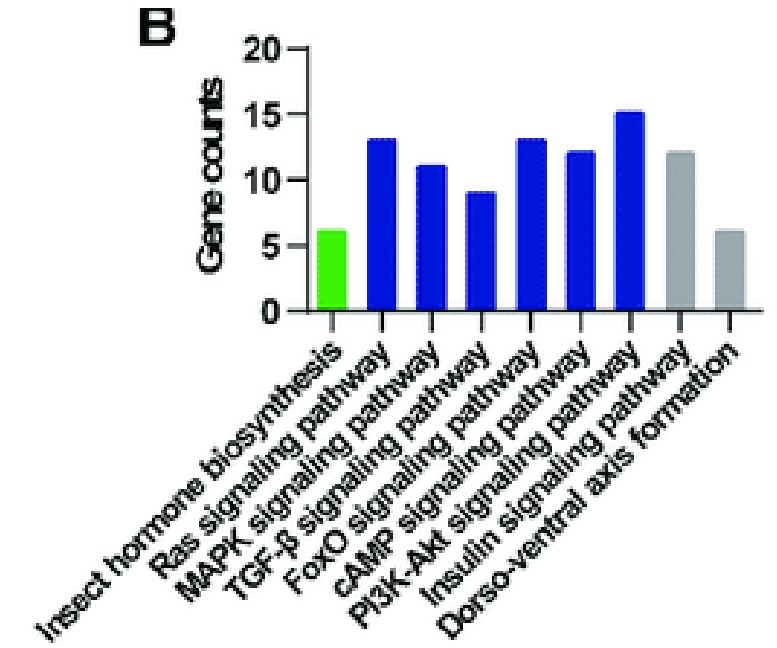
Target mRNA

Enrich
pathways



<https://ccb-compute2.cs.uni-saarland.de/mieaa/>

B



4. Research examples



Prediagnostic serum RNA dynamics in lung cancer

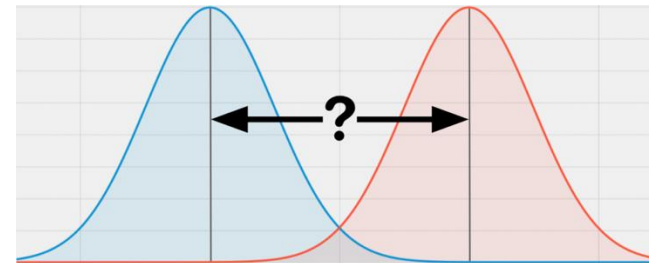


Aims



Identify circulating ncRNAs as early detection biomarkers of cancer

RNAs that discriminate healthy and those who will develop lung cancer



Pre-diagnostic biomarker → pre-diagnostic samples

- Collected from 1972-2003
- Norwegian population
 - 90 % health surveys
 - 10% red cross blood donors
- > 300 000 donors
- Linkage to the cancer registry
- Questionnaires
 - Smoking
 - BMI
 - Physical activity ++

pre-diagnostic samples → old samples → sample quality

- > low volume
- > low ncRNA yields
- > storage effects
- > sampling over time
- > differing protocols



JANUS serumbank

Langseth et al, Int J Epidemiol, 2016, Hjerkind et al, Int J Epidemiol, 2017



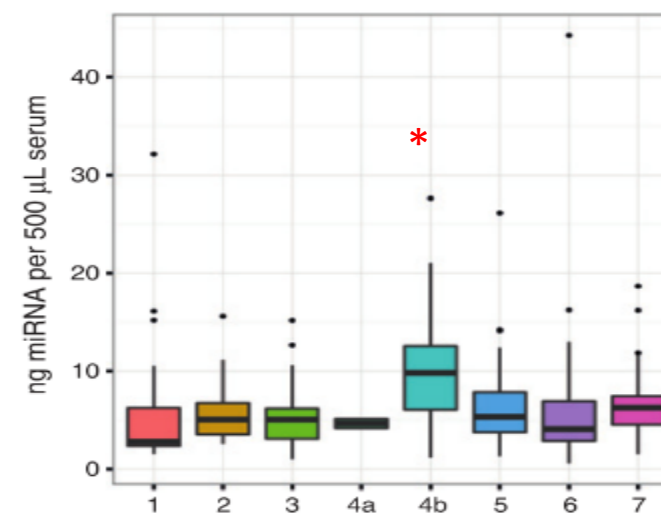
Sample quality



JANUS serumbank

Group	Sample collection period	Sample source	Serum processing
1	1972-1978	HE	Iodoacetate added
2	1979-1986	HE	No additives
3	1987-2004	HE	Separating gel tubes
4A	1973-1979	RCBD	No additives
4B	1973-1979	RCBD	Lyophilization
5	1980-1990	RCBD	No additives
6	1997-2004	RCBD	No additives
7	2013-2014	Fresh	No additives

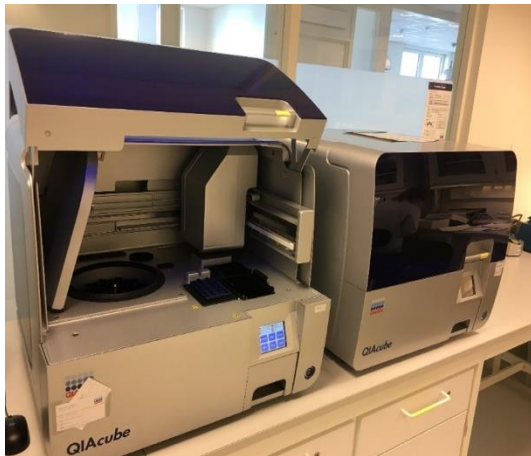
A miRNA yields in janus serum samples groups





ncRNA detection in serum

STEP 1: RNA extraction



- 2x200 µl serum
- miRNeasy and Trizol
- Glycogen carrier
- RNA internal control
- QIAcube

STEP 2: sncRNA library prep.



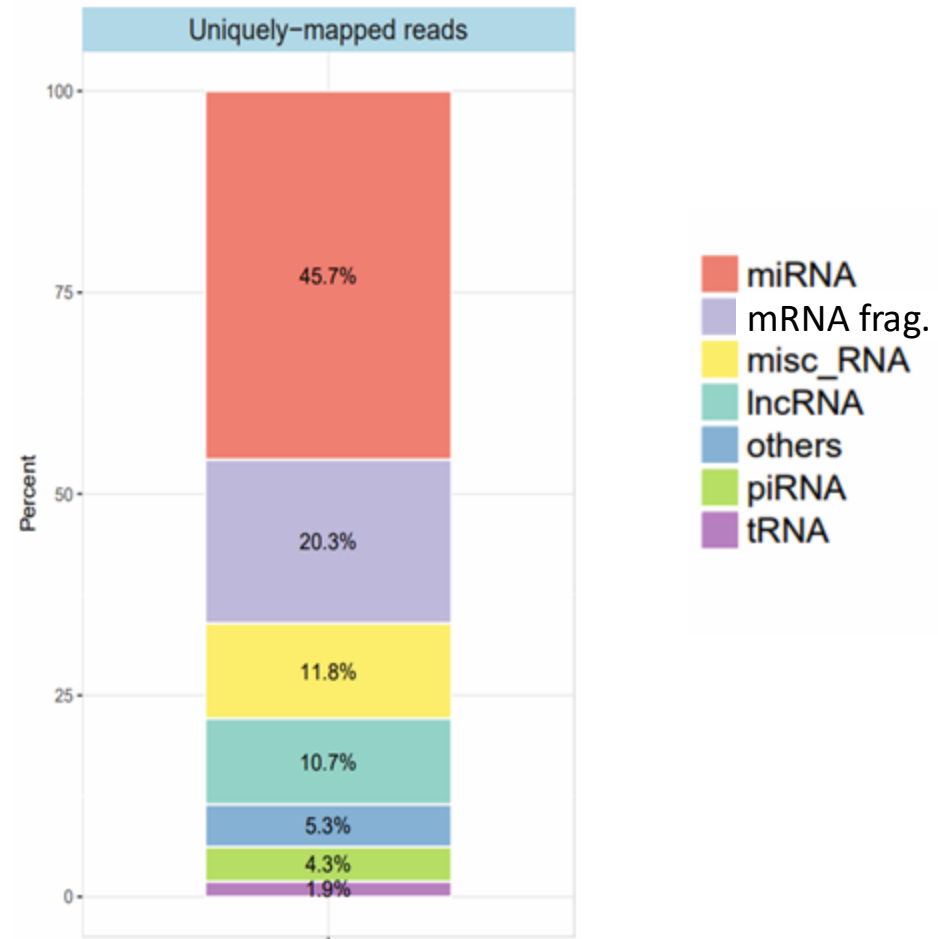
- NEBNext small RNA Libraries.
- **17-47 nt RNA size**

STEP 3: Sequencing



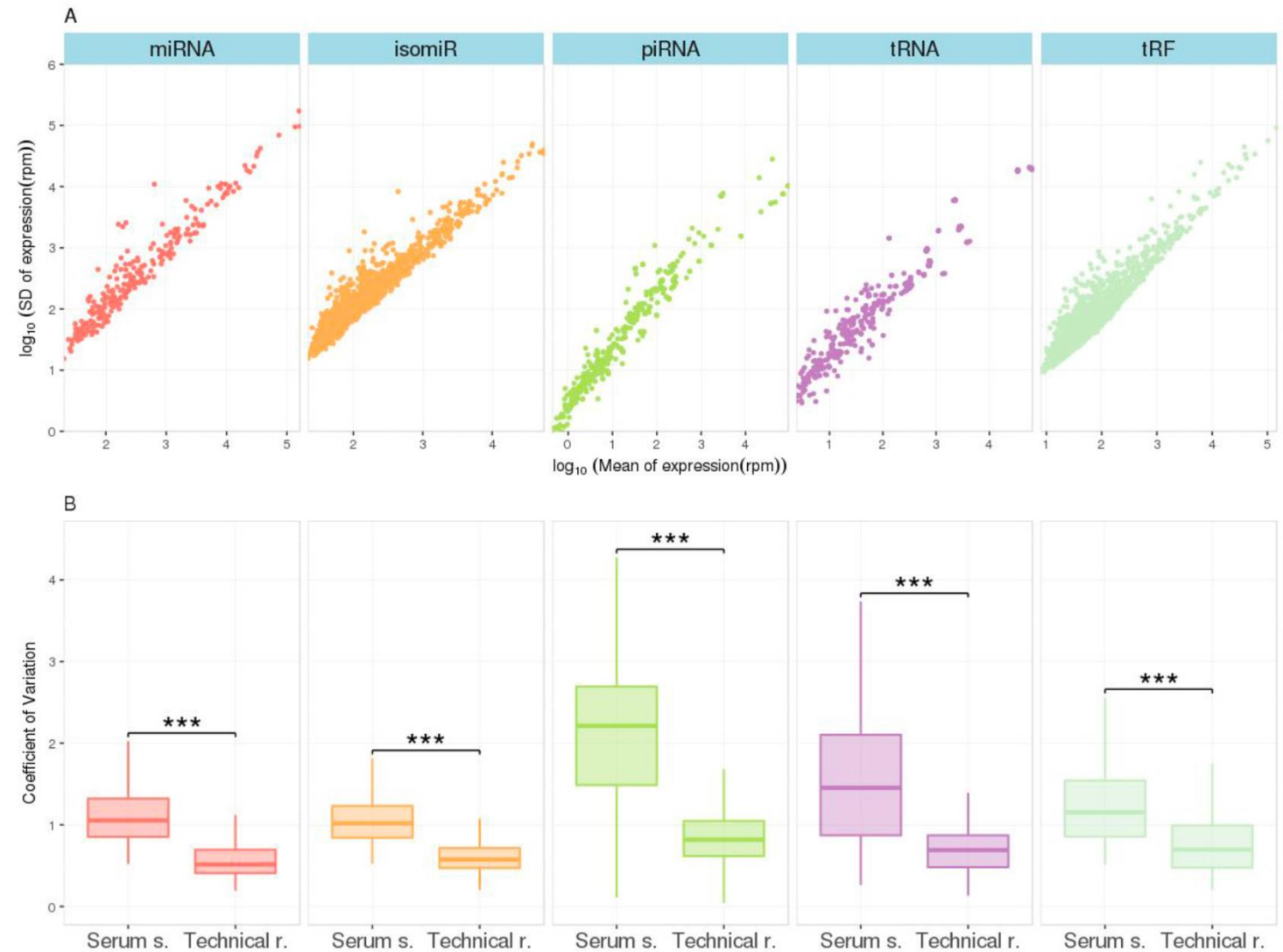
- Sequencing on the Illumina HiSeq 2500
- 12 libraries per lane
- **~ 18 mill sequences per sample**

RNAs in serum samples (n=477)



Expression variation

- Technical vs biological variation
- Biological variation
 - Too much
 - Too little



ncRNA variation in 477 serum samples from healthy donors

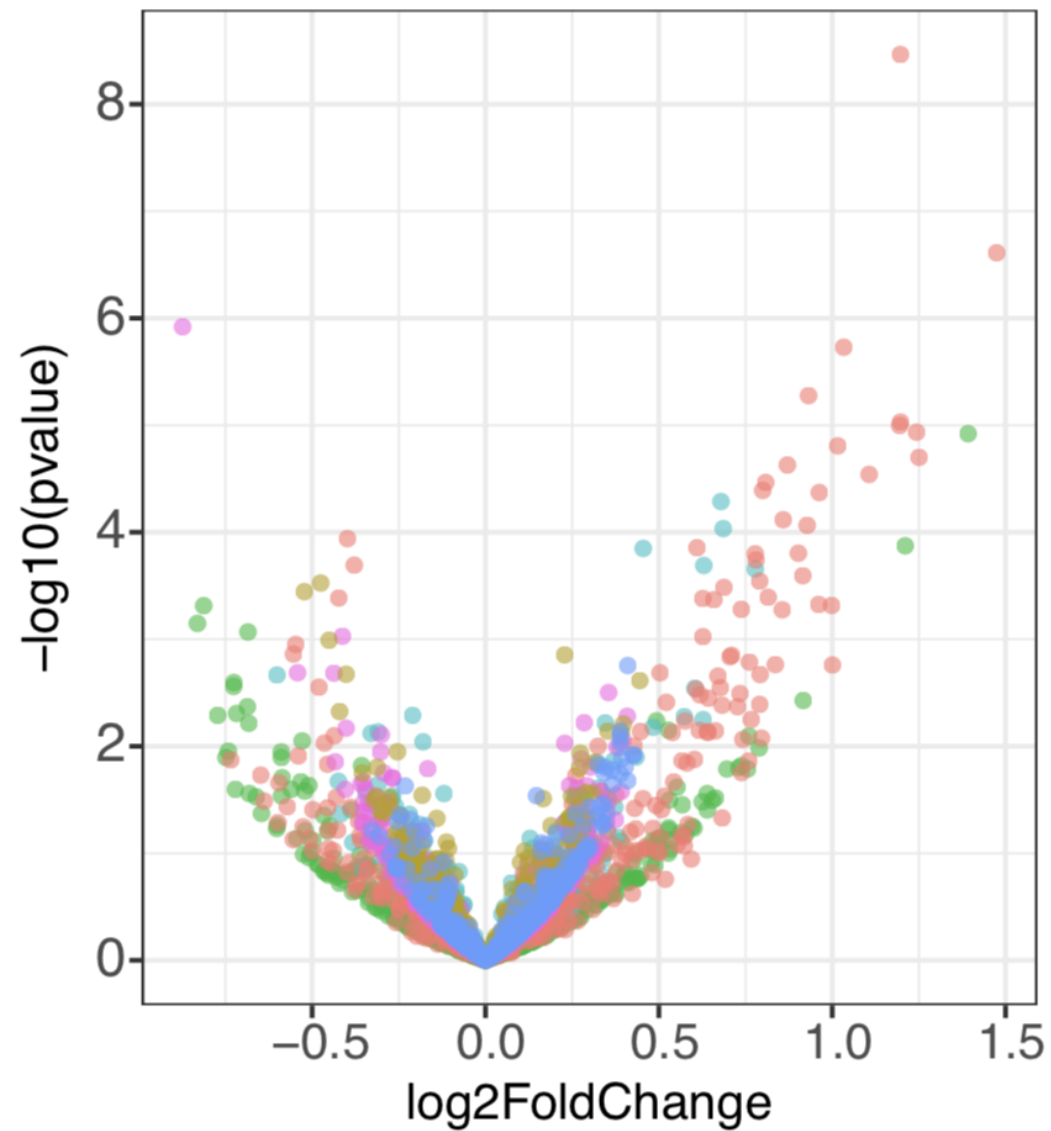
Sinan U Umu

Umu et al, RNA Biol, 2018

Small RNAseq



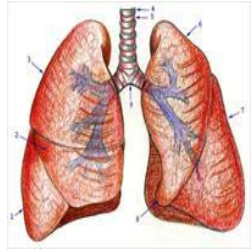
miRNA



Trait

- age
- body mass
- blood donor group
- sex
- physical activity
- smoking

Study design and participants



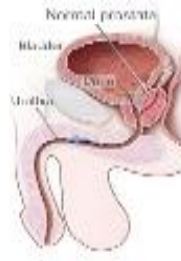
Lung
(404/559)



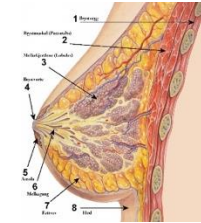
Colorectal
(488/520)



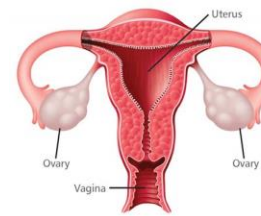
Testicular
(84/84)



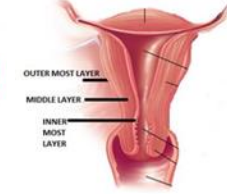
Prostate
(332/450)



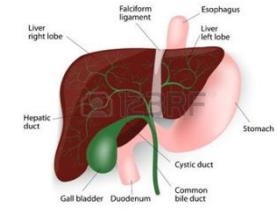
Breast
(206/395)



Ovaries
(88/100)



Endometrium
(320/320)



Gall
bladder
(27/27)

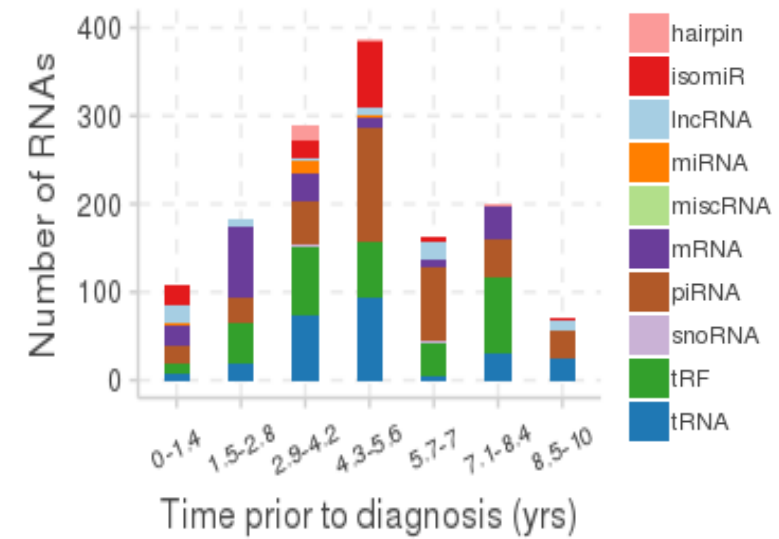
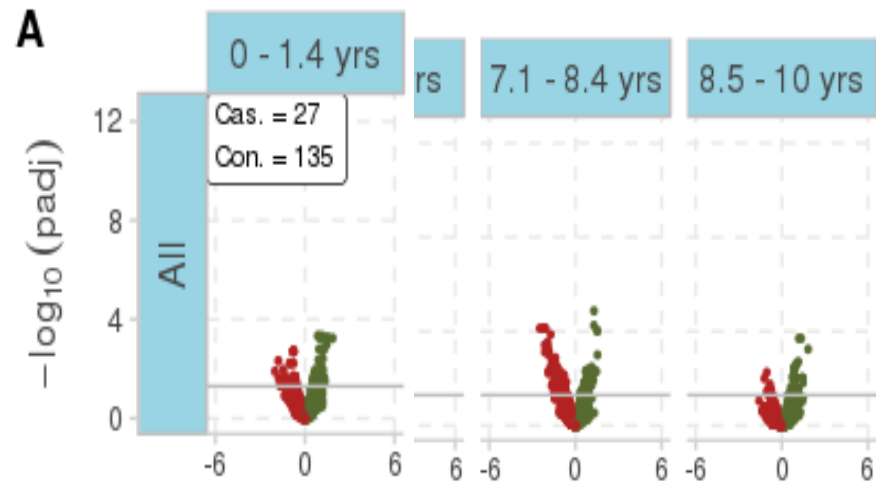
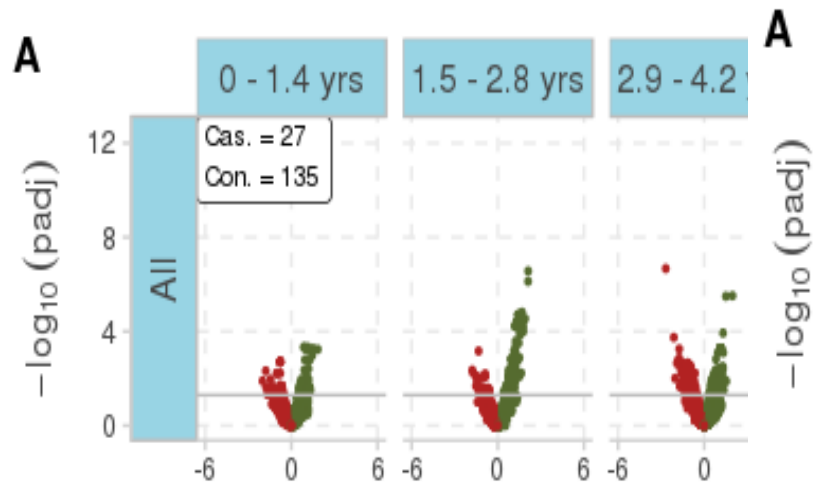


Controls
(N=673)

Pre-diagnostiske Janusprøver - Illumina HiSeq 2500, 3500 og 6000




- Janus serum samples donated up to 10 years prior to diagnosis
- Controls :
 - Free from cancer at least 10 years after blood collection
 - Frequency matched on age, sex, storage time
- Serial samples in approximately 20% of the patients

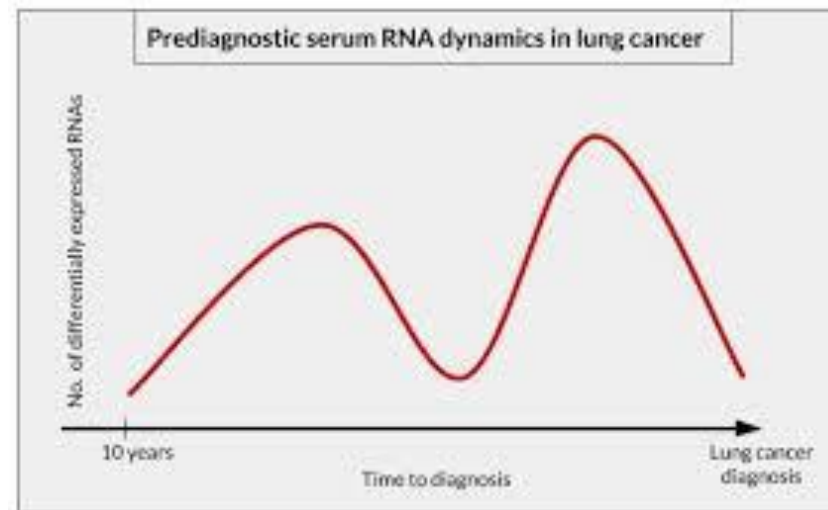


Summary

- Large-scale serum ncRNA analyses for biomarker research require
 - optimized methods
 - knowledge of technical and biological variation
 - RNA classes have similar variation
 - control for traits and exposome effects,
 - Age
 - Smoking
- ncRNA 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

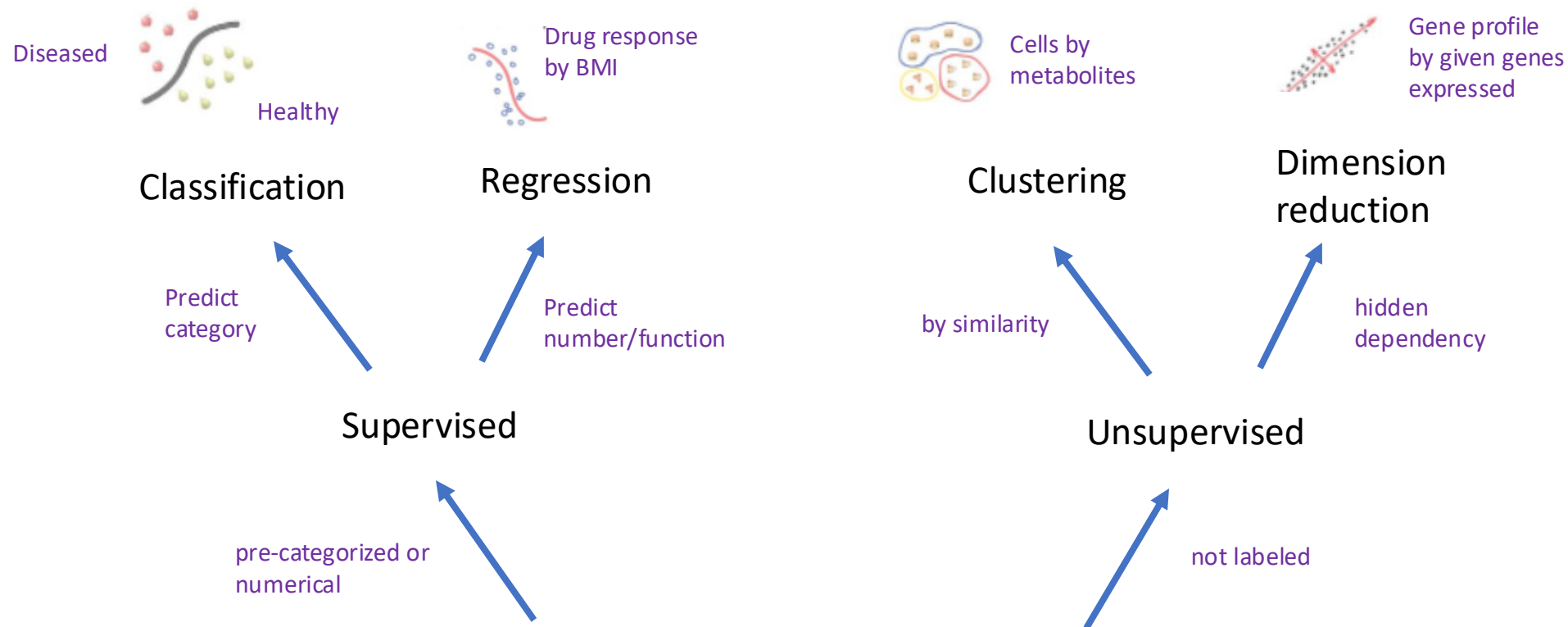
A 10-year prediagnostic follow-up study shows that serum RNA signals are highly dynamic in lung carcinogenesis

Sinan Uğur Umu¹, Hilde Langseth¹, Andreas Keller^{2,3}, Eckart Meese⁴, Åslaug Helland^{5,6,7}, Robert Lyle^{8,9} and Trine B. Rounge^{1,10} 

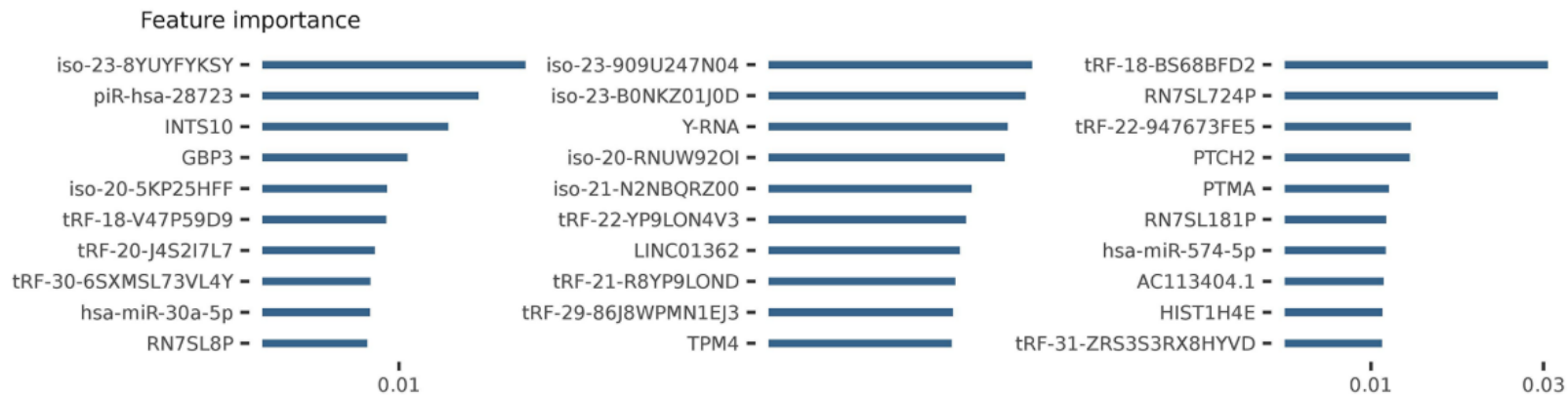
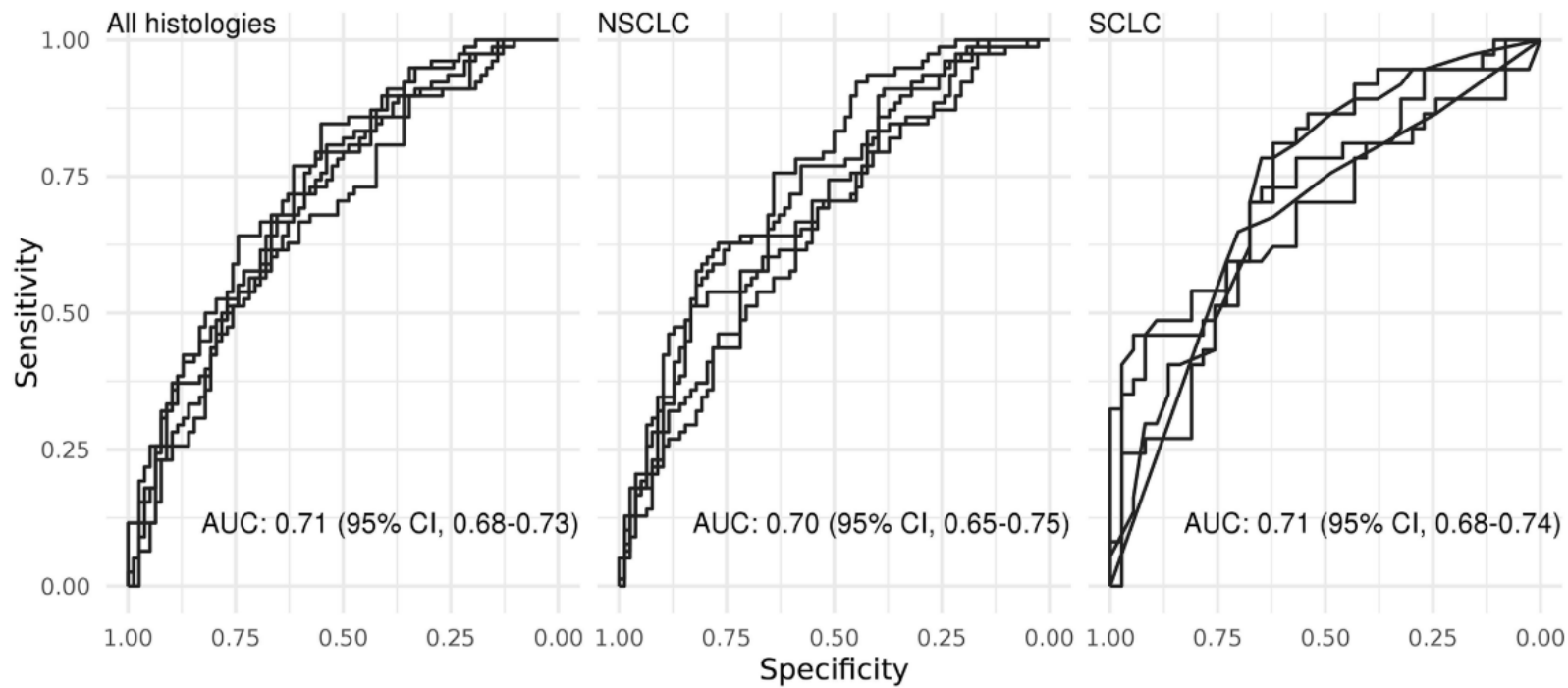


Small RNAs used in classification

- Machine learning (ML) algorithms can be used for classification
- Differential expression analyses that analyse one RNA at the time
- ML can identify “patterns” in the data that can be used to classify a sample
 - Disease or healthy



Classical machine learning



RESEARCH ARTICLE



CC

Serum RNAs can predict lung cancer up to 10 years prior to diagnosis

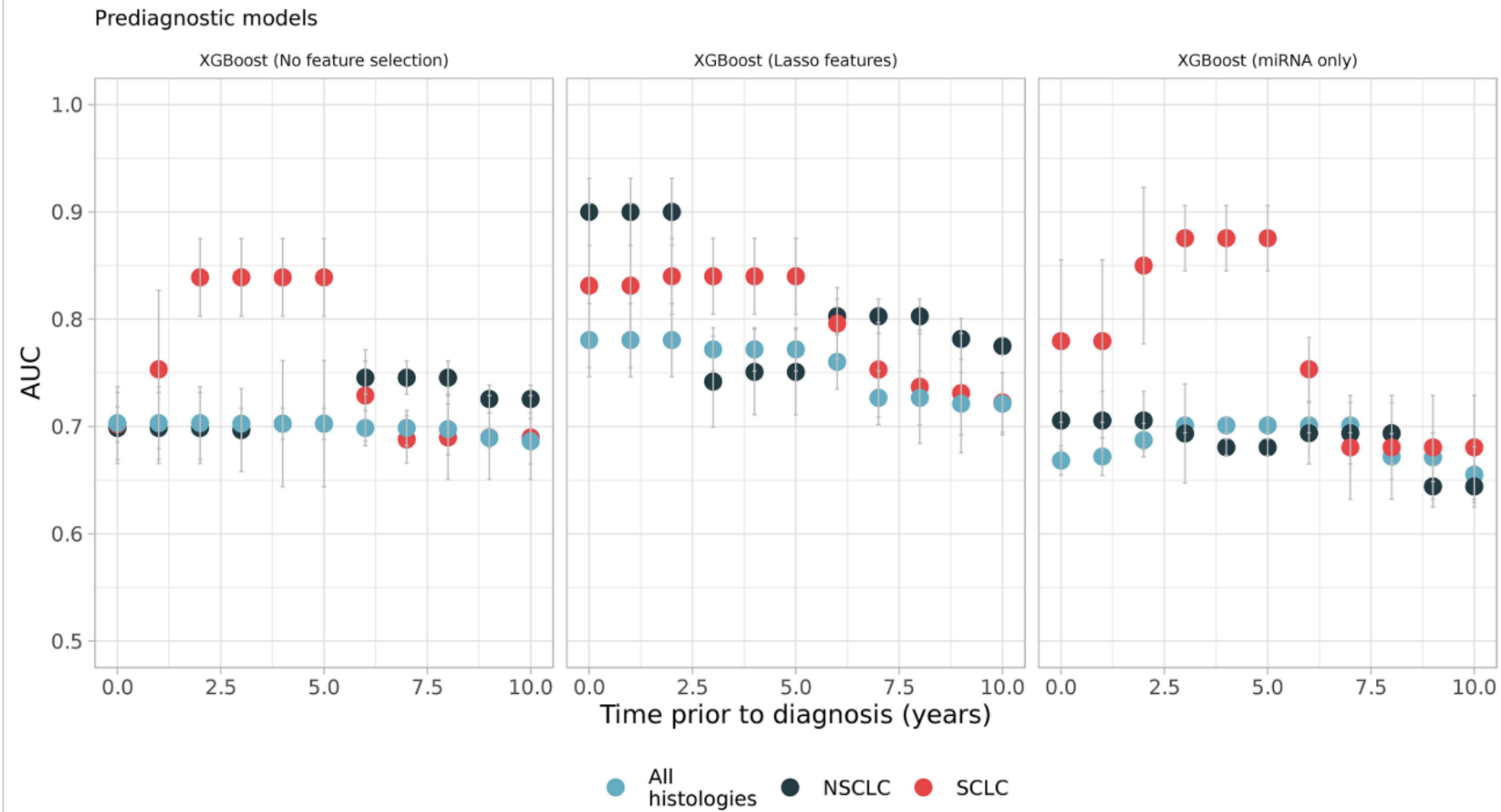
Sinan U Umu^{1*}, Hilde Langseth^{1,2}, Verena Zuber², Åslaug Helland^{3,4,5}, Robert Lyle^{6,7}, Trine B Rounge^{1,8*}



UiO : University of Oslo

IN-BIOS 5000/9000

Small RNAseq



RESEARCH ARTICLE



Serum RNAs can predict lung cancer up to 10 years prior to diagnosis

Sinan U Umu^{1*}, Hilde Langseth^{1,2}, Verena Zuber², Åslaug Helland^{3,4,5}, Robert Lyle^{6,7}, Trine B Rounge^{1,8*}



UiO : University of Oslo

IN-BIOS 5000/9000

Small RNAseq

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