

Characterization of the small RNA transcriptome

Lorena Pantano

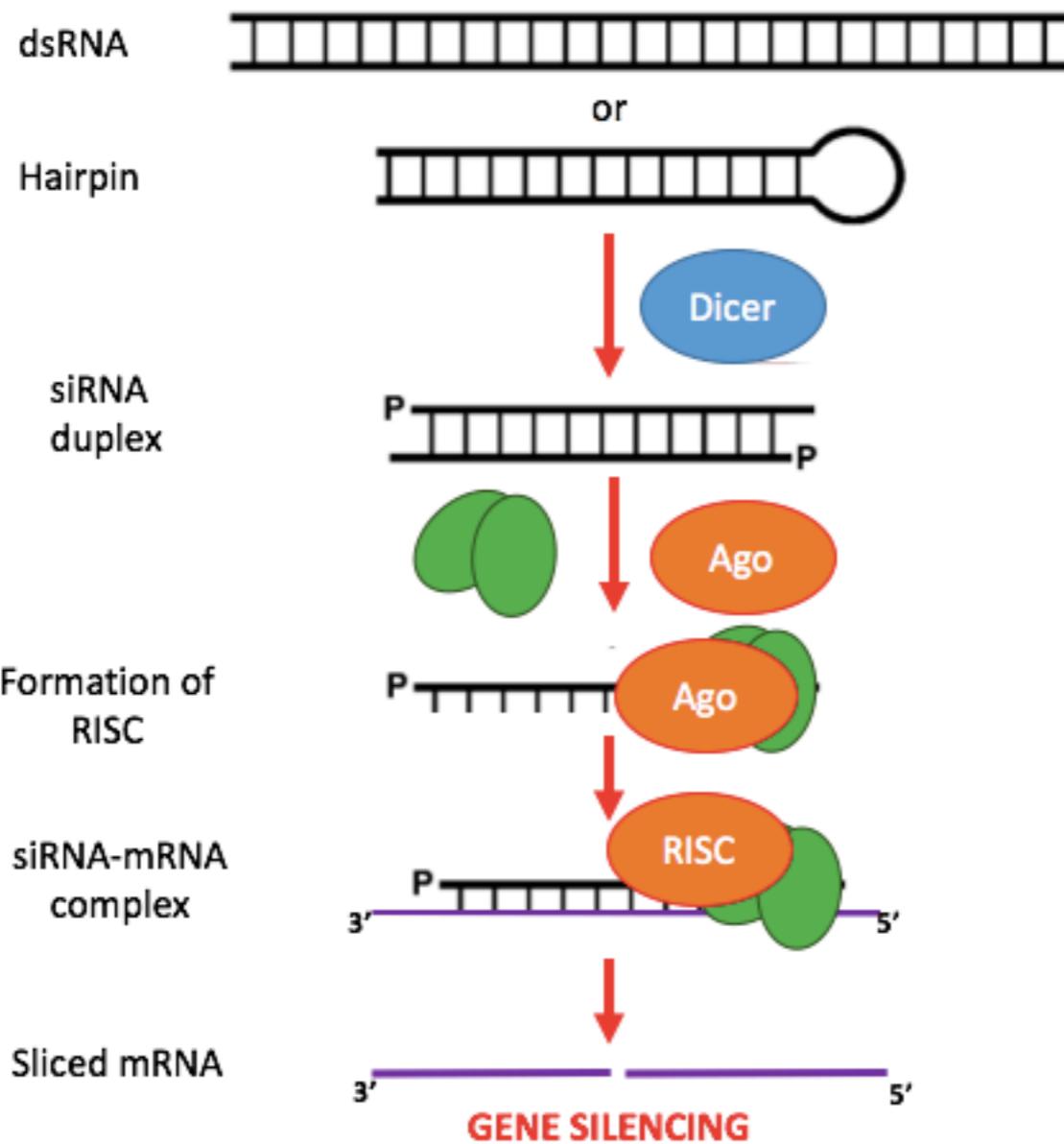
@lopantano lpantano@hsph.harvard.edu

Harvard TH Chan School of Public Health

http://bit.ly/hbc_course_2018_srna

2018-9-26

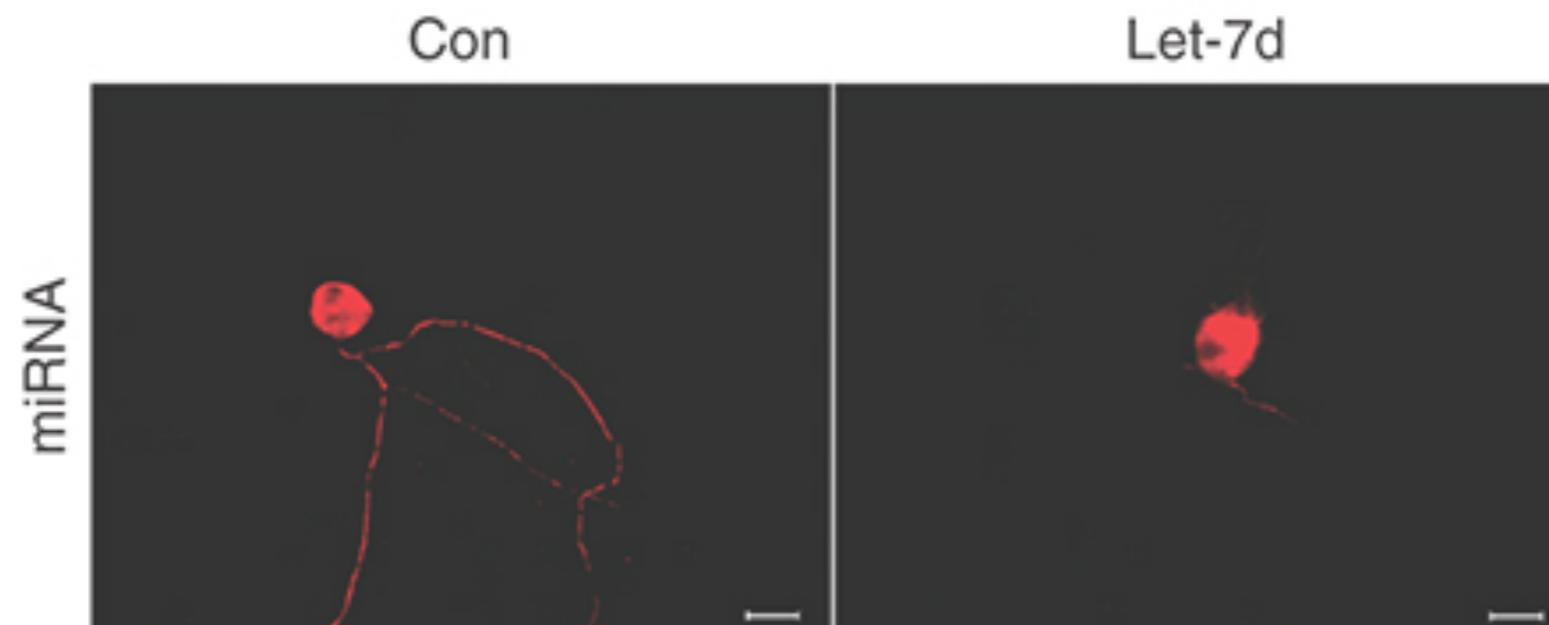
small interference RNA



- miRNA (18-25nt)
- endo-siRNA (20-25nt)
- piRNA (25-33nt)

miRNA

axon outgrowth



Let-7 microRNAs Regenerate Peripheral Nerve Regeneration by Targeting Nerve Growth Factor
Shiying Li, Xinghui Wang, Yun Gu, Chu Chen, Yaxian Wang, Jie Liu, Wen Hu, Bin Yu, Yongjun Wang, Fei Ding, Yan Liu and
Xiaosong Gu

isomiRs

hsa-miR-24-1-5p

....GGUGCCUACUGAGCUAGAUUAUC...

GUGCCUACUGAGCUAGAUUCAGU

GUGCCUACUGAGCUGAUUAUCAG

GUGCCUACUGAGCUGAUA

... UGCCUACUGAGCUGAUUAUCA ...

... UGCCUACUGAGCUGAUUAUCAGU

... UGCCUACUGAGCUGAUUAUC .

... UGCCUACUGAGCUGAUA

.....CCUACUGAGCUAUAUCA.....

.....CCUACUGAGCUGAUAUCAC

.....CUACUGAGCUCAUUAUCA.....

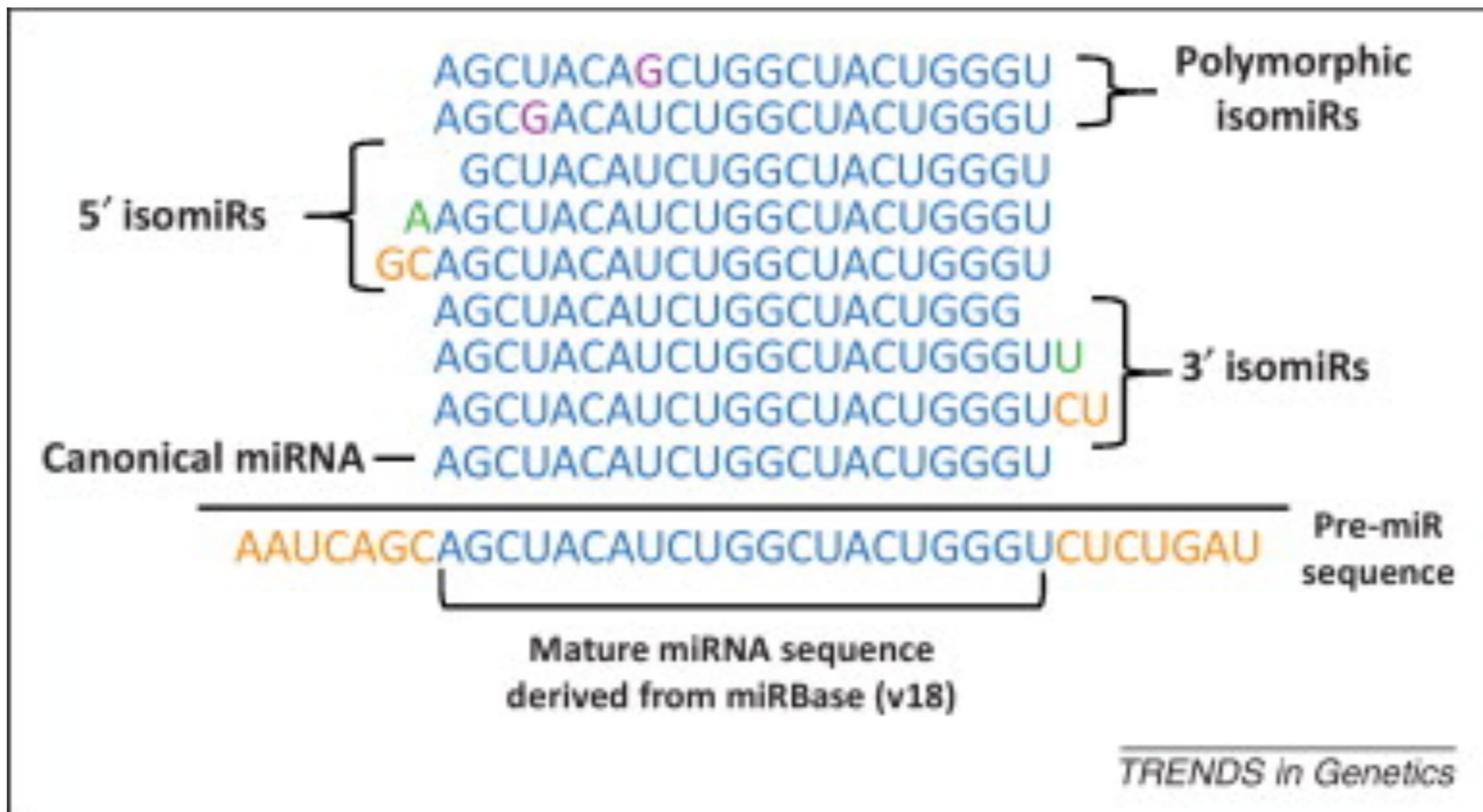
.....CUACUGAGCUCAUUAUC.....

hsa-miR-24-3p

CUCCGGUGCCUACUGAGCUGAUUAUCAGUUCUCAUUUACACACUGGCUCAGUUCAGCAGGAACAGGAG
(((((((((.((((((((((.,((((((.....))))))),.)))))))))).))))..))))))) (-26.32)

precursor

types of isomiRs



Search results

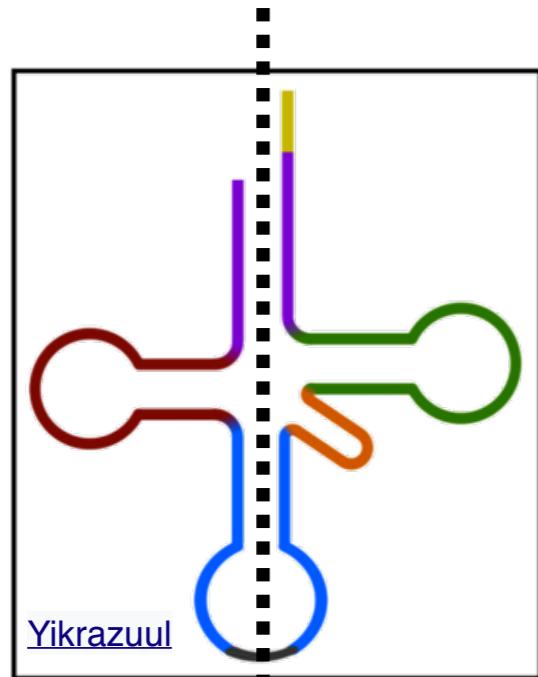
Items: 1 to 20 of 186

<< First < Prev Page of 10 Next > Last >>

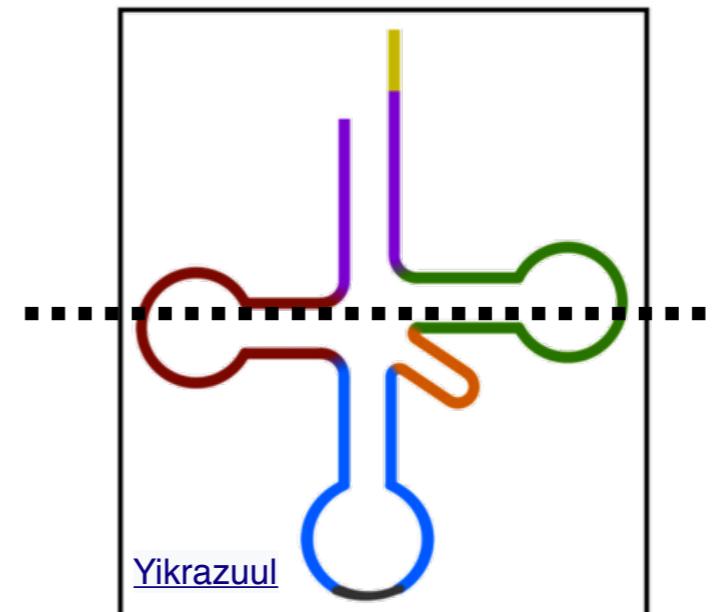
- [NGS analysis of total small non coding RNAs from low input RNA from dried blood sampling.](#)
 1. Pirritano M, Fehlmann T, Laufer T, Ludwig N, Gasparoni G, Li Y, Meese E, Keller A, Simon M. Anal Chem. 2018 Sep 10. doi: 10.1021/acs.analchem.8b03557. [Epub ahead of print]
PMID: 30198258
[Similar articles](#)
- [miRge 2.0 for comprehensive analysis of microRNA sequencing data.](#)
 2. Lu Y, Baras AS, Halushka MK. BMC Bioinformatics. 2018 Jul 23;19(1):275. doi: 10.1186/s12859-018-2287-y.
PMID: 30153801 [Free PMC Article](#)
[Similar articles](#)
- [Analysis of the expression, function, and evolution of miR-27 isoforms and their responses in metabolic processes.](#)
 3. Ma M, Yin Z, Zhong H, Liang T, Guo L. Genomics. 2018 Aug 23. pii: S0888-7543(18)30297-0. doi: 10.1016/j.ygeno.2018.08.004. [Epub ahead of print]
PMID: 30145283

tRNA derived fragments

tRNAs function as carriers that transport amino acids to the growing polypeptide chain during the translation of mRNA.



tRNA-halves (30-33nt)



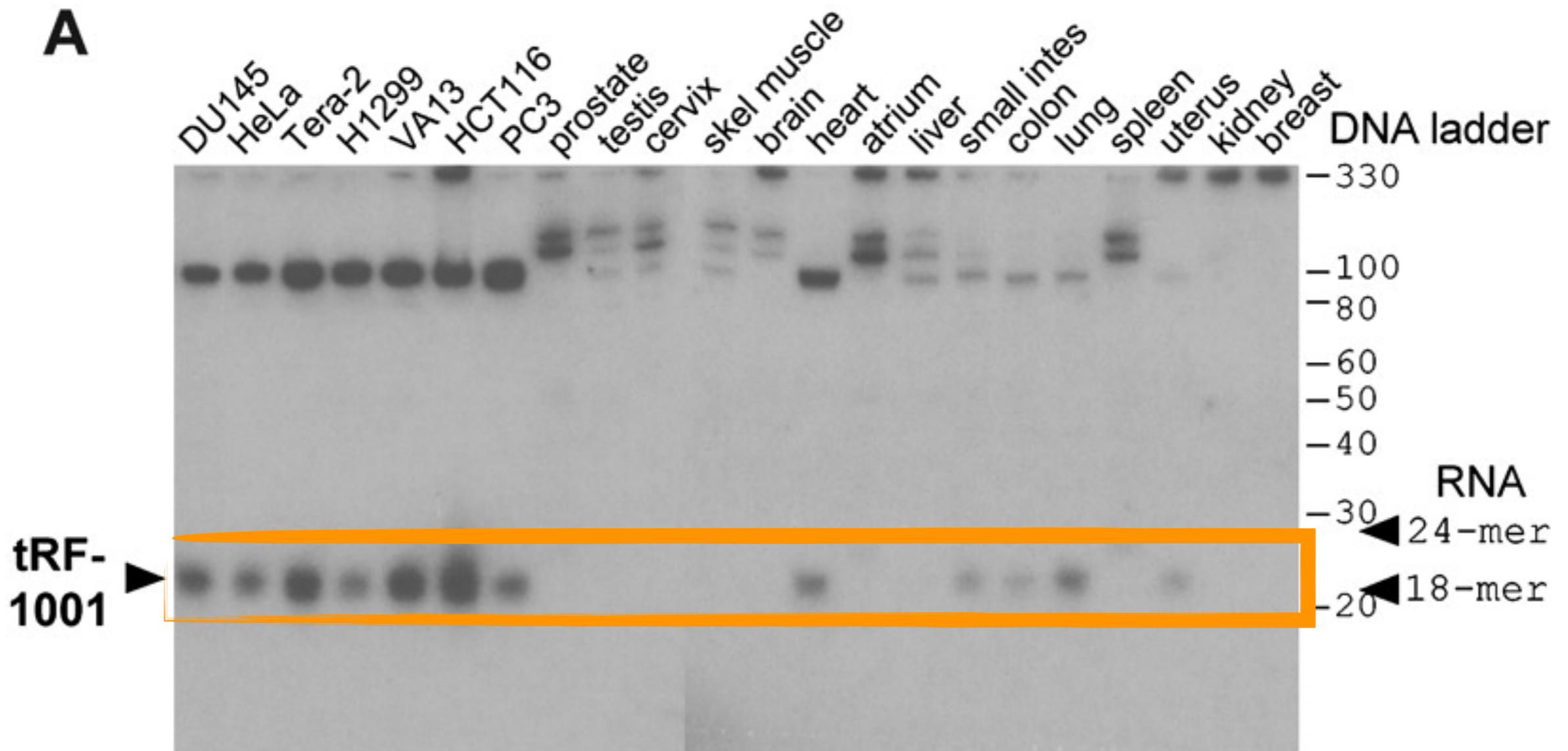
5'-tRNAs (18-22) 3'-tRNAs (18-22)

associated with **genetic disorders** and malignancies such as **prostate, liver, lung (tRF-Leu-CAG)** or **breast cancer**, and related processes like **aging, oxidative stress**, and embryonic development

In Arabidopsis, they are miRNA-like sequences, targeting transposable elements.

They have been found in extracellular samples like: plasma, saliva and urine.

small tRNAs



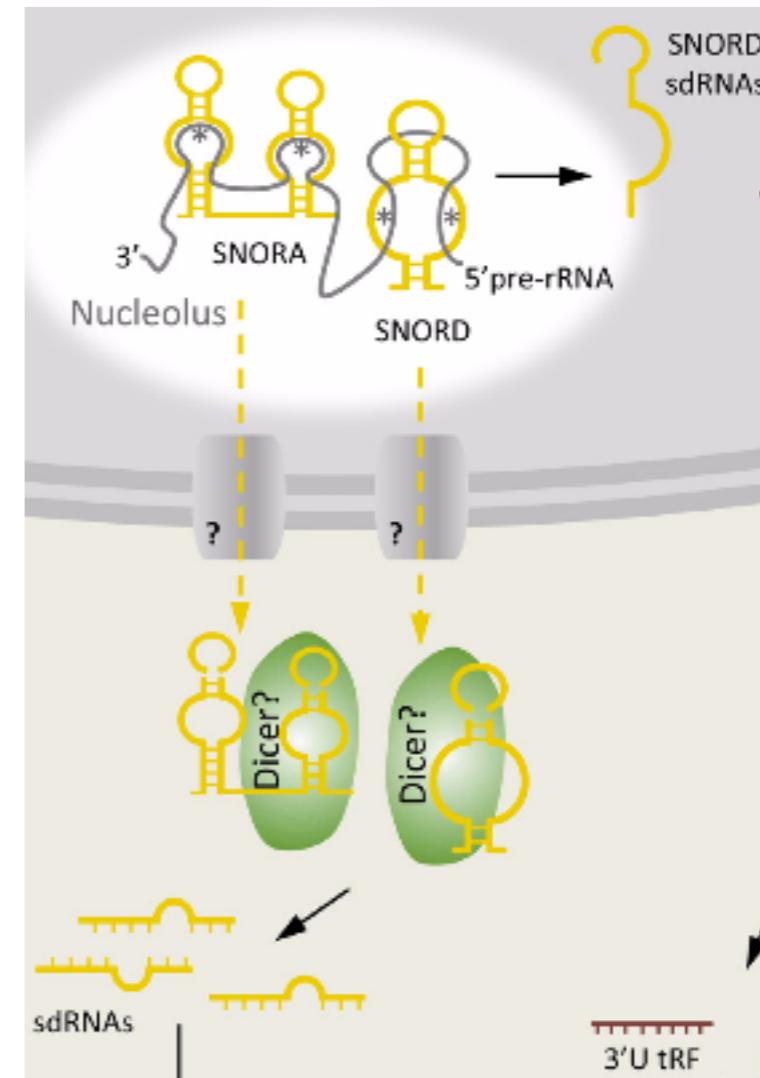
sncRNA fragments

Small nucleolar RNAs (snoRNA) are well-conserved, abundant, short non-coding RNA molecules, 60–300 nucleotides (nt) in length, which localize to a specific compartment of the cell nucleus – the nucleolus

In HEK293, SCARNA15 **miRNA-like** sequence targeting CDK11B (22nt)

SNORD88C-sdRNAs can regulate **alternative splicing** of fibroblast growth factor receptor 3. (FGFR3)

SNORD44/78 up-regulated in **prostate cancer**.



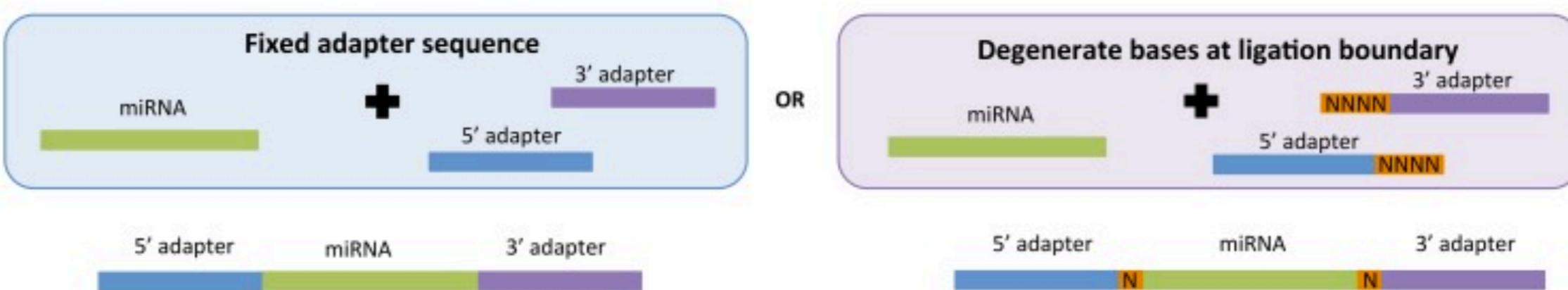
snoRNA-derived RNAs (sdRNAs)

Protocols

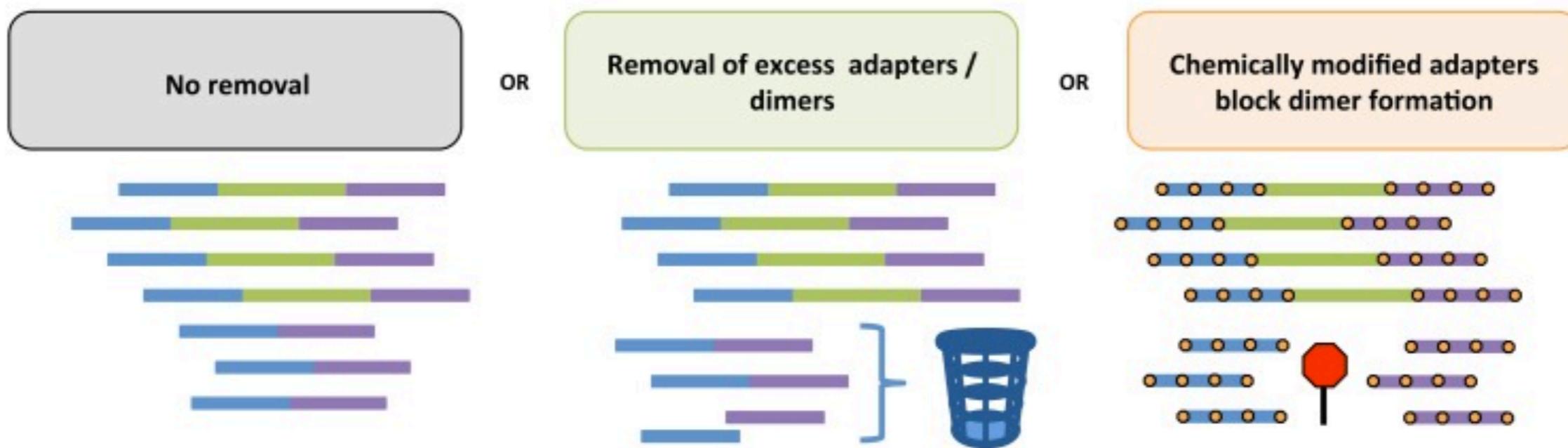
7

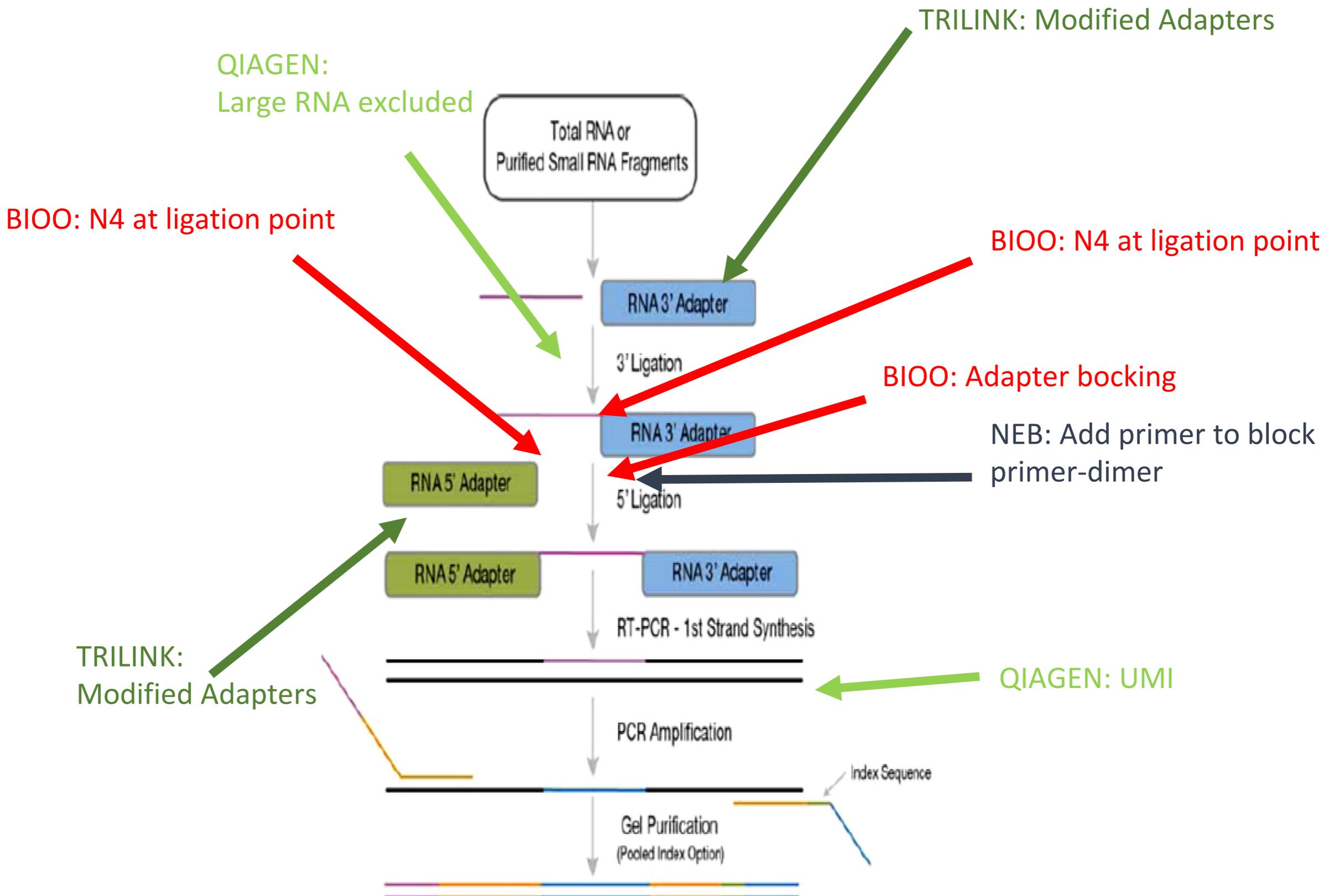
Critical differences in small RNA library preparation protocols

Issue 1: Adapter ligation introduces bias



Issue 2: Adapter dimers compete with small RNAs, reducing effective sequencing depth





Systematic comparison of small RNA library preparation protocols for next-generation sequencing

Cloelia Dard-Dascot¹, Delphine Naquin¹, Yves d'Aubenton-Carafa¹, Karine Alix², Claude Thermes¹ and Erwin van Dijk^{1*} 

Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling

Maria D Giraldez , Ryan M Spengler [...] Muneesh Tewari 

Nature Biotechnology **36**, 746–757 (2018) | Download Citation 

Caveats

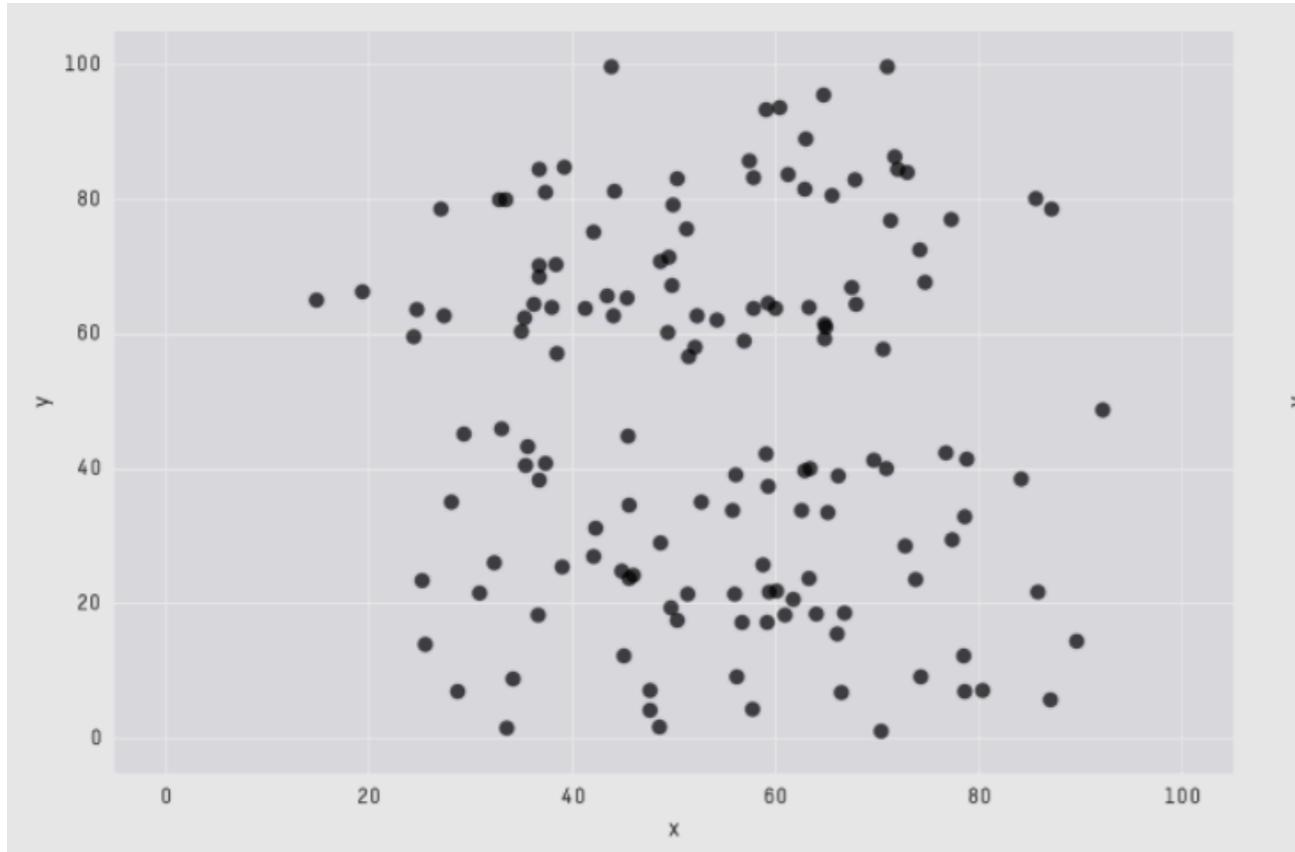
- TrueSeq Illumina: ligation bias
- NEBNext: ligation bias (potentially less)
- 4N + (NextFlex): generation of random sequences.
We lose the accuracy to detect isomiRs. Partial
ligation bias.
- 20m small RNA like piRNA, or plant miRNA needs
protocol modifications to be detected.

Visualization

“Visualization gives you answers to questions you didn’t know you had.” – [Ben Schneiderman](#)

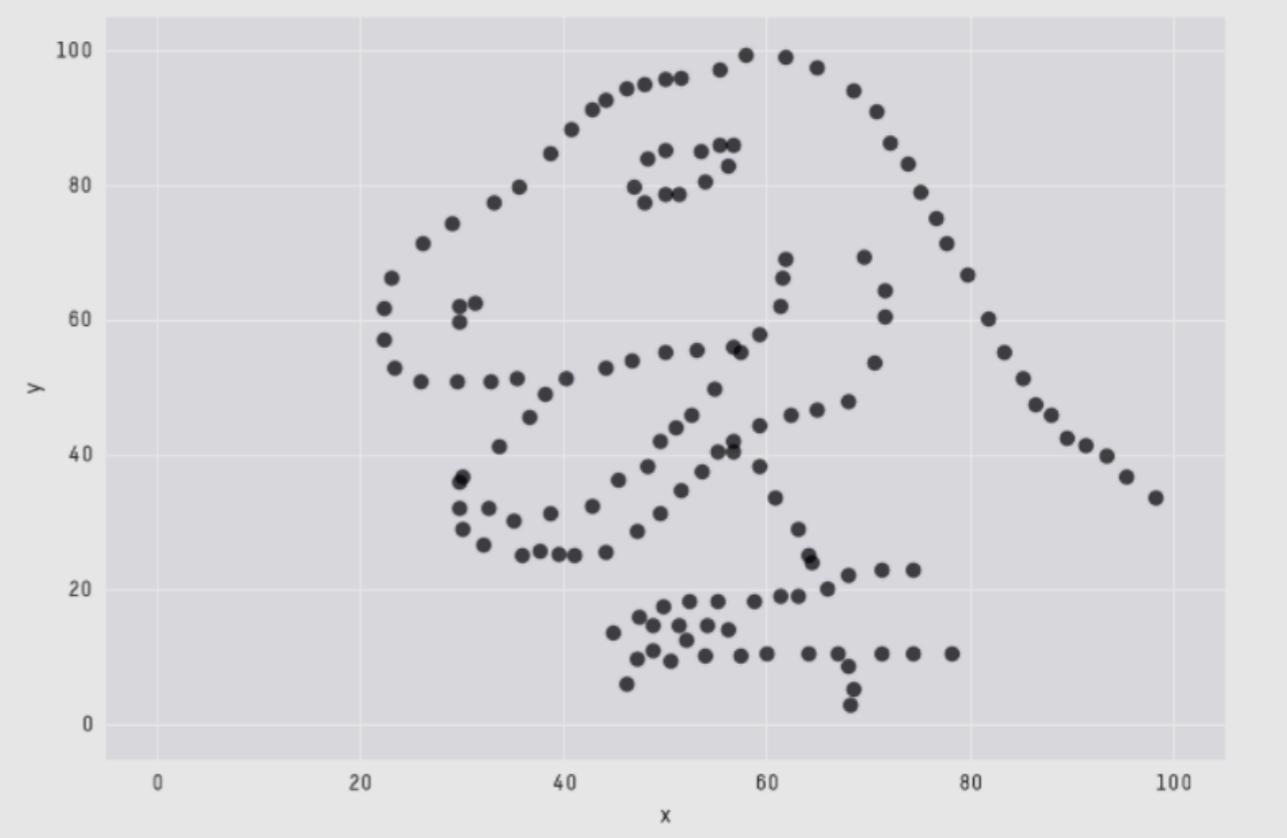
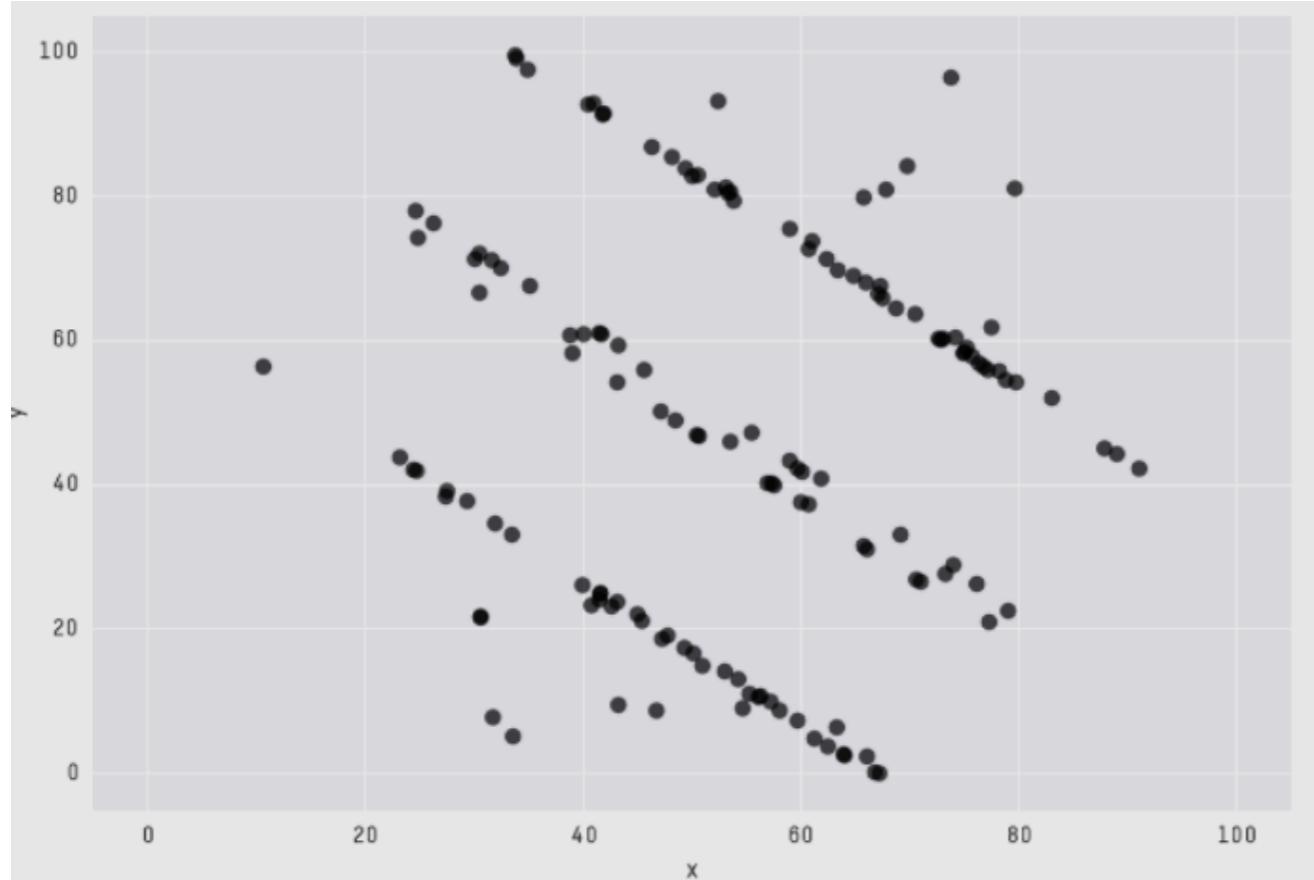
“There is no such thing as information overload. There is only bad design.” – [Edward Tufte](#)

Same Stats, Different graphs



X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

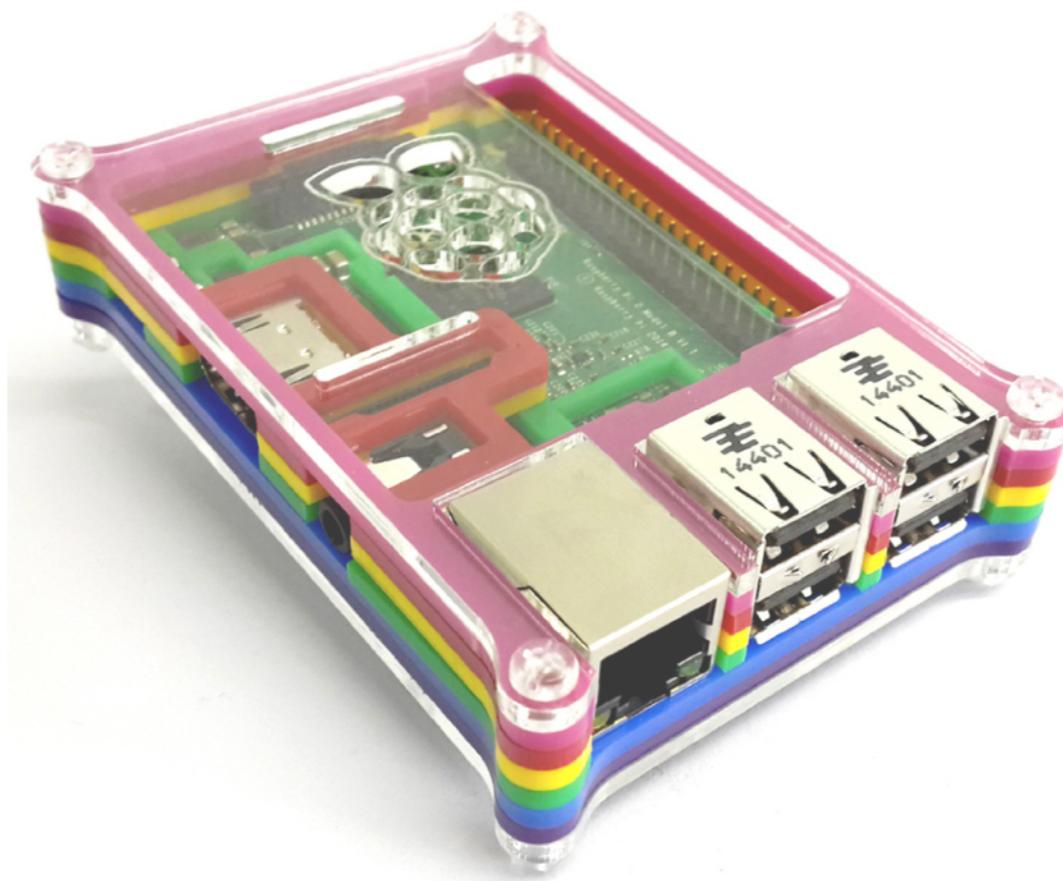
A TRex in your data



challenges

- isomiRs detection
- small RNAs coming from multiple precursors over the genome (multi-mapped reads can be 40% of the data.)
- differentiate degradation and functional molecules
- non-model organism

bcbio-nextgen



Variant calling, RNA-seq, small RNA-seq
over 200 peer reviewed tools **BIOCONDA**[®]

small RNA-seq analysis

processing & QC

cutadapt
fastqc
qualimap
multiqc

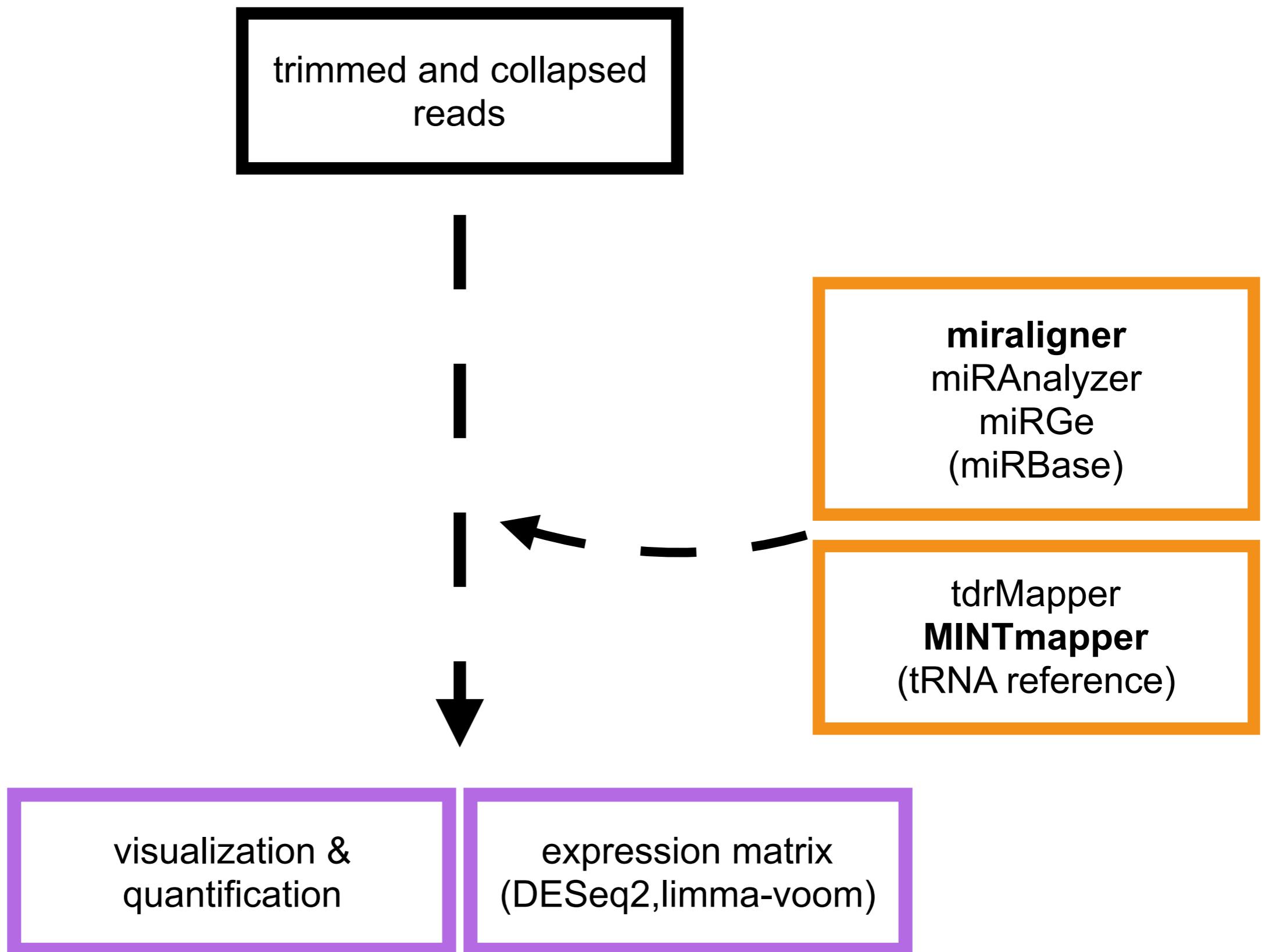
de-novo

seqcluster
mirdeep2 for miRNA
protac for piRNA

detection & annotation

mimaligner
miRAnalyzer
miRGe
tdrmapper
MINTmapper

Detection & Annotation



mirGFF3

A proxy for [miRNA/isomiR](#) data analysis where all tools meet with the idea to create an ecosystem of data analysis promoting community collaboration.

[repo status](#) [Active](#) [gff3 definition](#) [gff3 example](#) [fairsharing accepted](#) [edam accepted](#)

Introduction

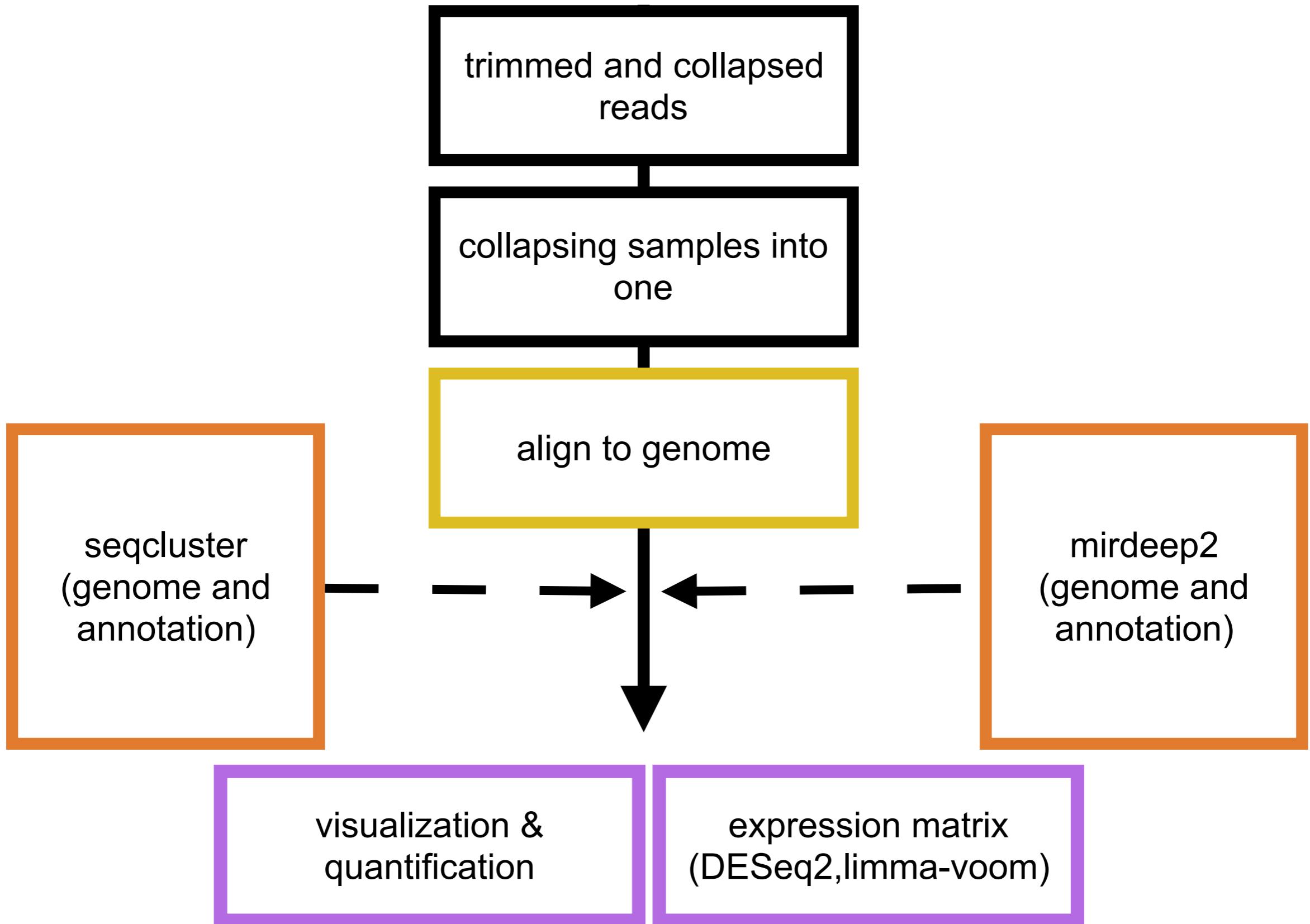
As discussed here: <https://github.com/miRTop/incubator/issues/10> we've defined a GFF3 format for output of small RNA pipelines focused on miRNA data currently. We is an open community project ([read more](#)) and joint us!

This output is based on the current GFF3 definition: [https://github.com/The-Sequence-Ontology/Specifications
/blob/master/gff3.md](https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md)

Note: Keep in mind this is for the output of a pipeline, so we know there will be bias toward methodology, but the idea is to put enough information to be able to re-analyze or filter sequences using information described here. As well, it would be a proxy for downstream analysis or packages.

[Advantages](#)

De-novo detection



miRDeep2 output

```

Provisional ID      : chr12_16160
Score total        : 1869.4
Score for star read(s) : 3.9
Score for read counts : 1866.6
Score for mfe       : -1
Score for randfold  :
Score for cons. seed:
Total read count   : 3673
Mature read count  : 3670
Loop read count     : 0
Star read count     : 3

```



Mature	Star			
5' - gaaaugccugguccaaugguaugggguauucagaacuuauuaguggcaaauacguugcuacuaccccc	cucugauaaaauuugacuggcuu	-3'	obs	
gaaaugccugguccaaugguaugggguauucagaacuuauuaguggcaaauacguugcuacuaccccc	cucugauaaaauuugacuggcuu		exp	
.....(((((((.....(((((.....((((((.....)))))).....)))))).....))))..		reads	nn	sample
....cccugguccGauugguaugggguauucagaacuu.....		38	1	seq
....cccugguccGauugguaugggguauucagaacuu.....		3	1	seq
....cccugguccGauugguaugggguauucagaacuuuu.....		3	1	seq
....cugguccGauugguaugggguauucagaacuuu.....		3	1	seq
....cugguccGauugguaugggguauucagaacuu.....		3	1	seq
....cugguccGauugguaugggguauucagaacuuuu.....		3	1	seq
....guccGauugguaugggguauucagaacuu.....		16	1	seq

seqcluster visualization

The screenshot shows a web browser window titled "clusters information". The address bar displays "file:///Users/lpantano/repos/seqclusterViz/reader.html". The toolbar includes standard icons for back, forward, search, and file operations. Below the toolbar, there are several tabs and links: "BioC 3.3: BUILD/CHECK", "Files - OneDrive", "Home - Dropbox", "Timesheet - HBC - Ha", "Niner Cogalicious - Mo", "Build your own images", "Home - PubMed - NC", and "Google Scholar". A blue button labeled "Browse..." is visible. On the left, there are two input fields: "Clusters Filter:" and "Clusters Id:". In the center, there are two tables: "Table with clusters" and "Table with Locus". The "Table with clusters" has columns "Sel.", "I.D.", and "Description:". The "Table with Locus" has columns "I.D.", "Index", and "Locus:". At the bottom, there are two sections: "Abundance profile along precursor" and "Secondary structure".

<https://github.com/lpantano/seqclusterViz>

Bcbio-nextgen done,
Now what?

bcbioSmallRna: R package

- Load all the data from bcbio pipeline
 - `miRNA`
 - `isomiRs`
 - `clusters`
- QC figures

<https://lpantano.github.io/bcbioSmallRna>

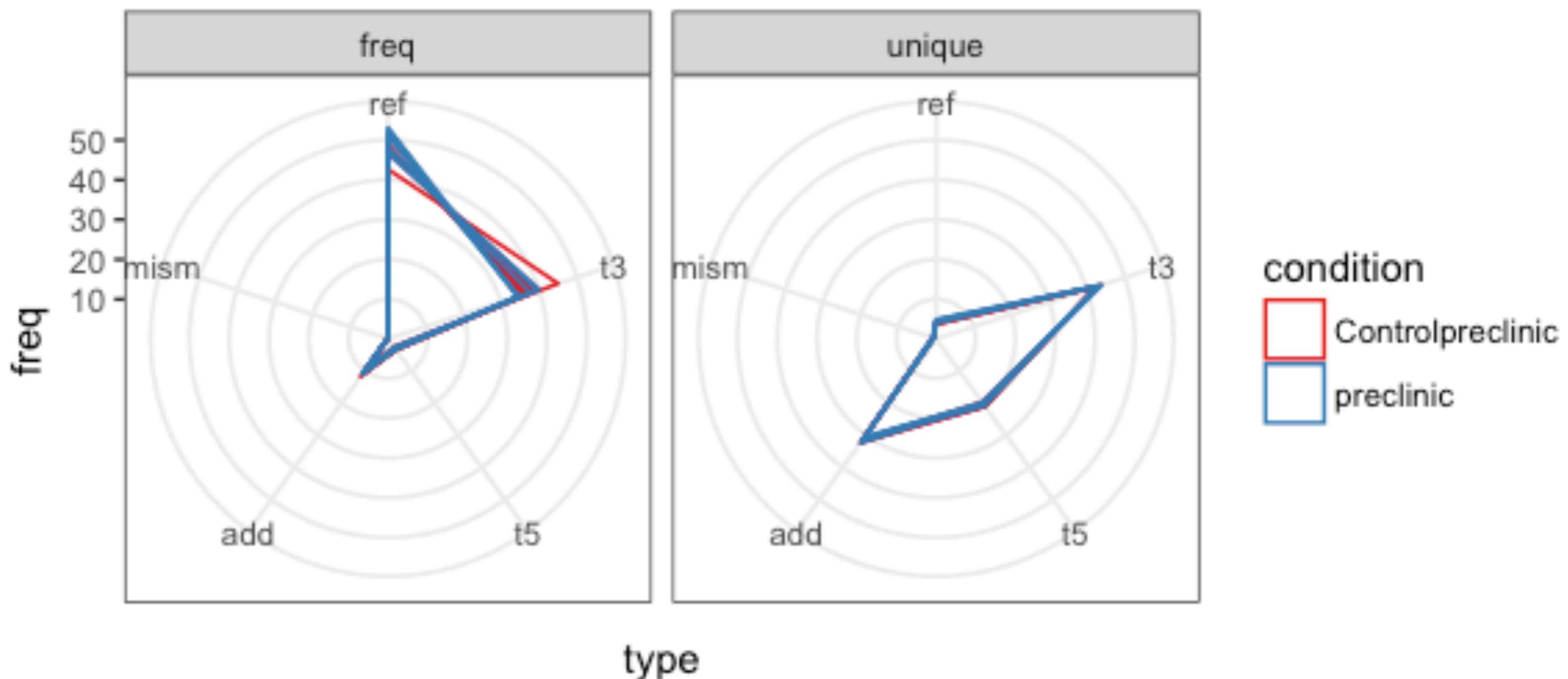
isomiRs: R package

- General Characterization of isomiRs.
`biocLite("isomiRs")`
- Collapsing isomiRs in different ways
- Supervised clustering analysis to detect important miRNAs (PLS-DA)
- RNAseq and miRNA time serie data
- Help with DE analysis

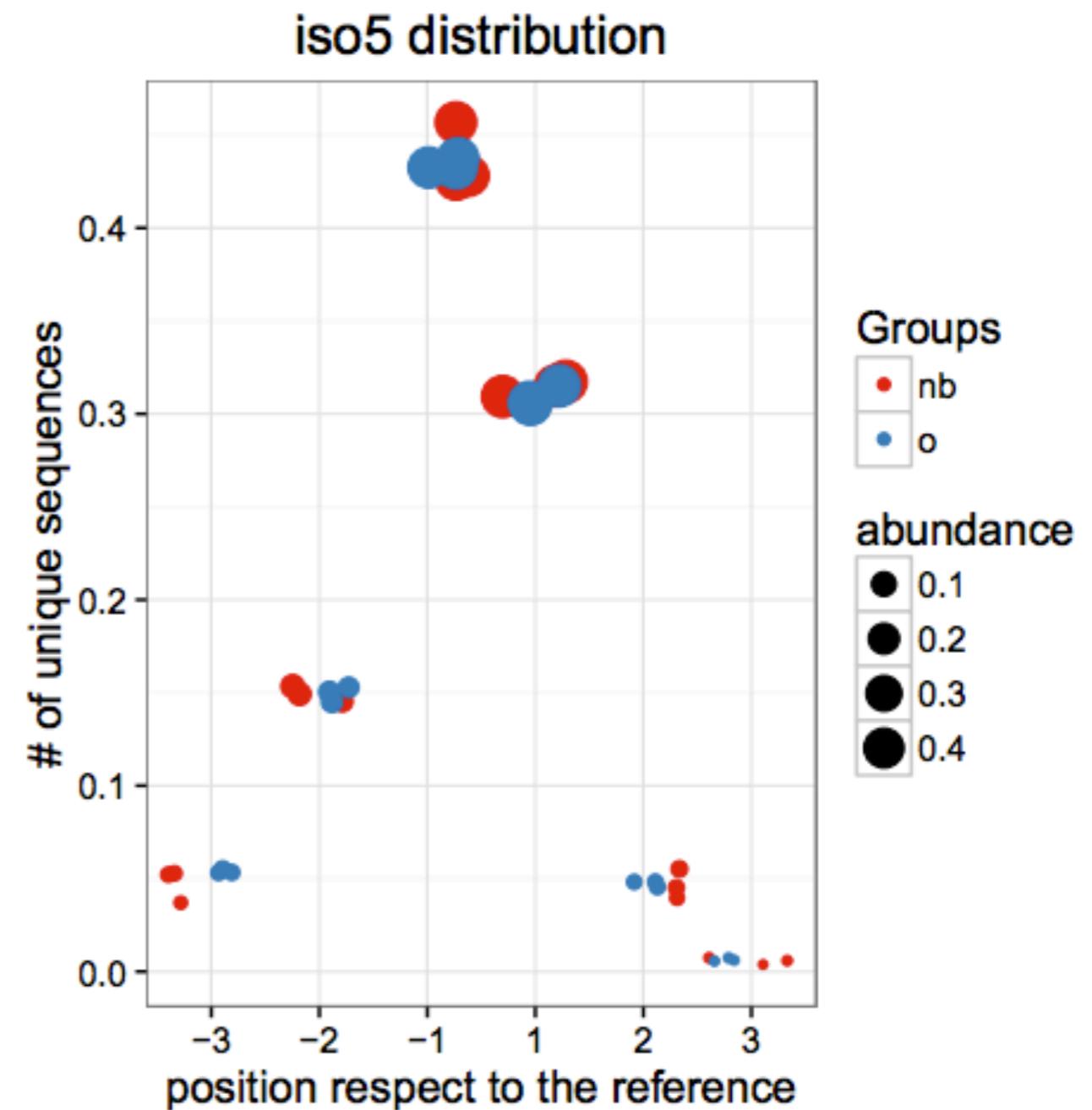
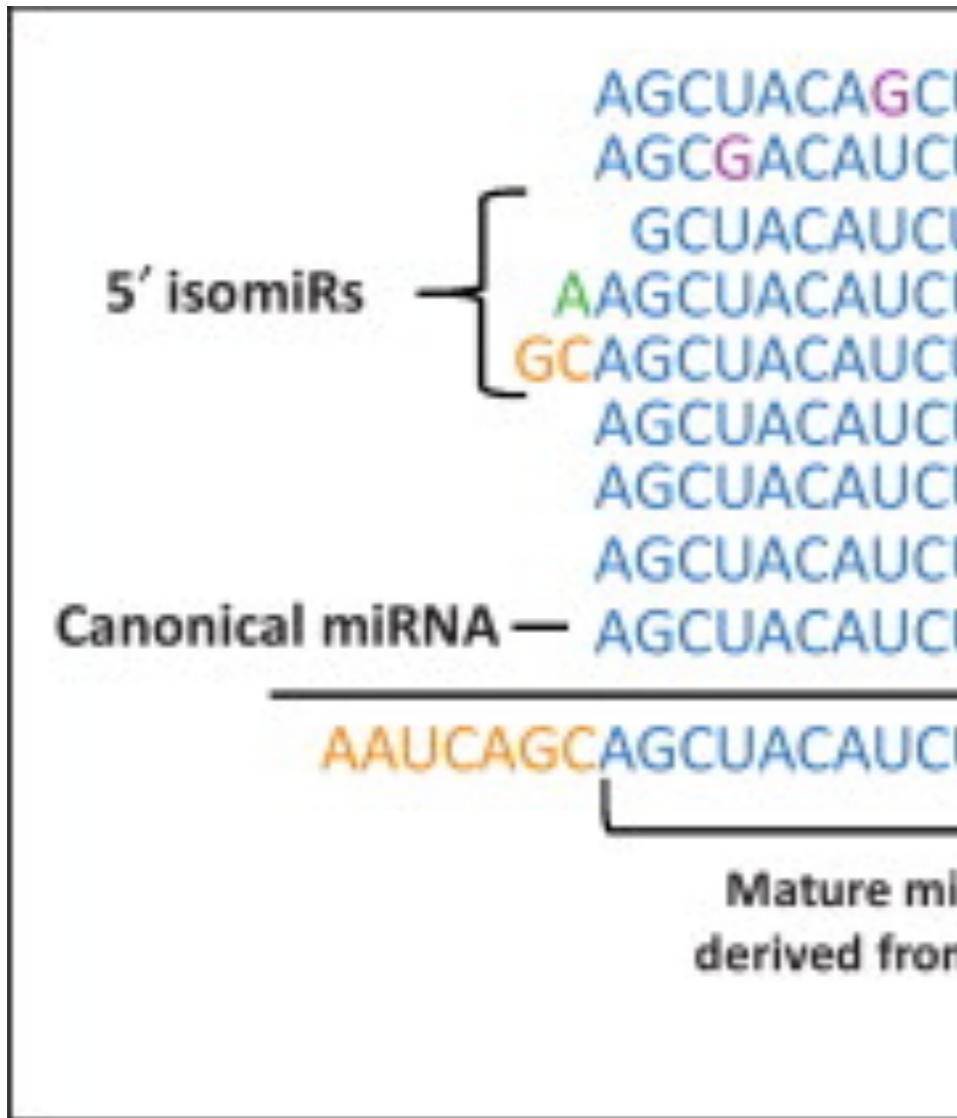
[http://bioconductor.org/packages/release/bioc/html/
isomiRs.html](http://bioconductor.org/packages/release/bioc/html/isomiRs.html)

isomiRs: R package

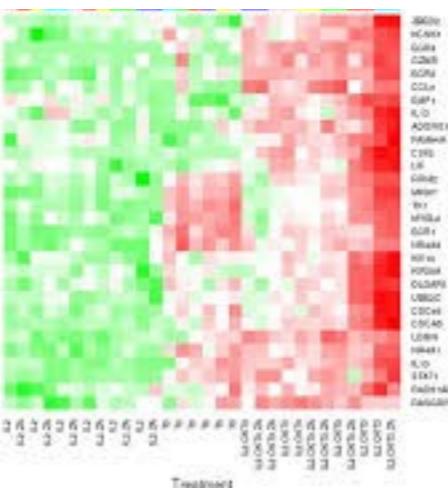
Spider plots



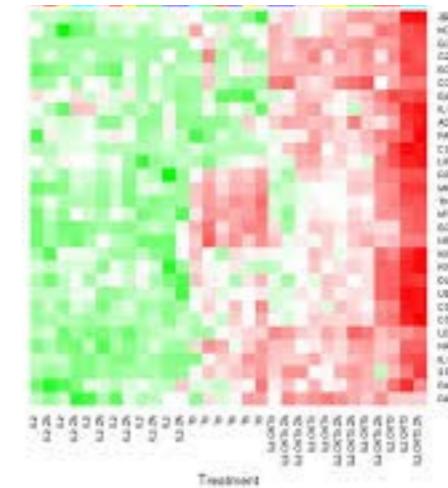
isomiRs: R package



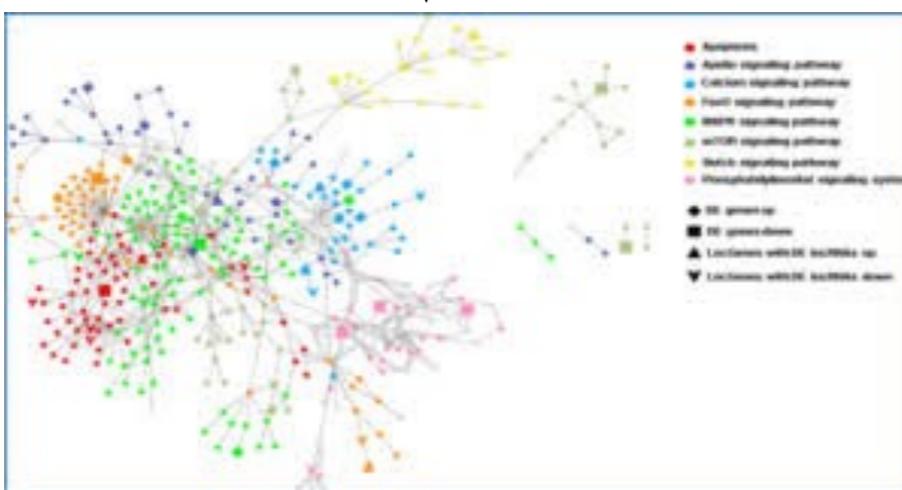
Omics integration



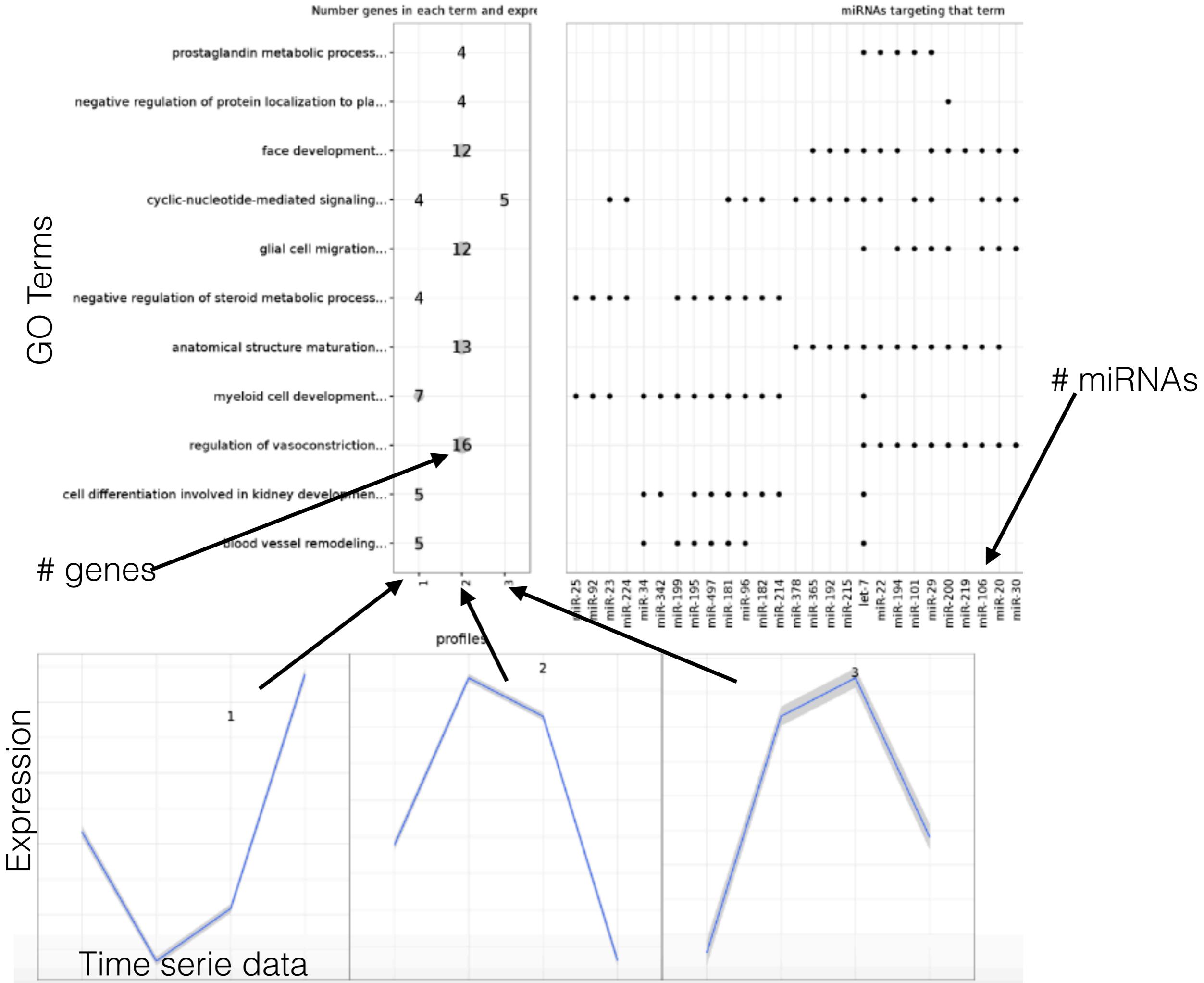
Gene expression



miRNA expression



Functional enrichment



Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study

Pieter Mestdagh, Nicole Hartmann, Lukas Baeriswyl, Ditte Andreasen, Nathalie Bernard, Caifu Chen, David Cheo, Petula D'Andrade, Mike DeMayo, Lucas Dennis, Stefaan Derveaux, Yun Feng, Stephanie Fulmer-Smentek, Bernhard Gerstmayer, Julia Gouffon, Chris Grimley, Eric Lader, Kathy Y Lee, Shujun Luo, Peter Mouritzen, Aishwarya Narayanan, Sunali Patel, Sabine Peiffer, Silvia Rüberg, Gary Schroth  et al.

[Affiliations](#) | [Contributions](#) | [Corresponding author](#)

Nature Methods 11, 809–815 (2014) | doi:10.1038/nmeth.3014

Received 27 February 2014 | Accepted 22 May 2014 | Published online 29 June 2014

| Corrected online **30 July 2014**

Analyze Public Dataset

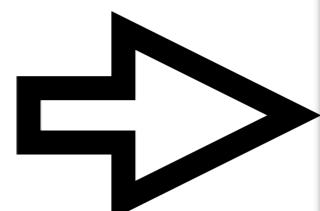
Samples (20)
≡ Less...

GSM1207643 miRQC A
GSM1207644 miRQC A repeat
GSM1207645 miRQC B
GSM1207646 miRQC B repeat
GSM1207647 miRQC C
GSM1207648 miRQC C repeat
GSM1207649 miRQC D
GSM1207650 miRQC D repeat

```
samplenames,description,group
GSM1207643,miRQCA,A
GSM1207644,miRQCArepeat,A
GSM1207645,miRQCB,B
GSM1207646,miRQCBrepeat,B
GSM1207647,miRQCC,B
GSM1207648,miRQCCrepeat,B
GSM1207649,miRQCD,B
GSM1207650,miRQCDrepeat,B
```

```
lp113@loge:~$ bcbio_prepare_samples.py --csv test.csv --out fastq
```

test-merged.csv



bcbio_nextgen.py -w template

bcbio_nextgen.py config.yaml

fastq/*fastq.gz

Analyze Public Dataset

NCBI Site map All databases Search

 Sequence Read Archive

Main Browse Search Download Submit Software Trace Archive Trace Assembly Trace BLAST

Studies Samples Analyses Run Browser Run Selector Provisional SRA

Evaluation of quantitative miRNA gene expression platforms i

Identifiers:

SRA: SRP028738
BioProject: [PRJNA214981](#)
GEO: [GSE49816](#)

Study Type:

Other

Abstract:

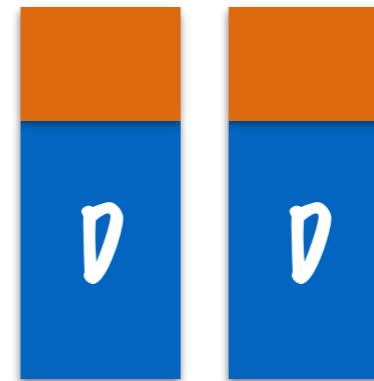
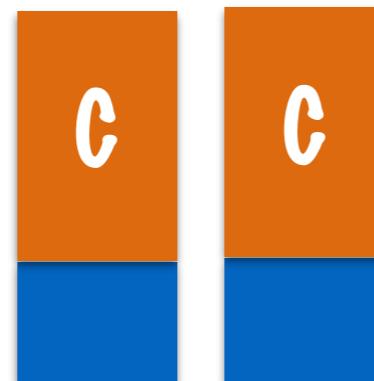
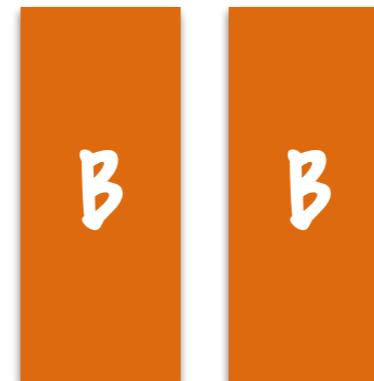
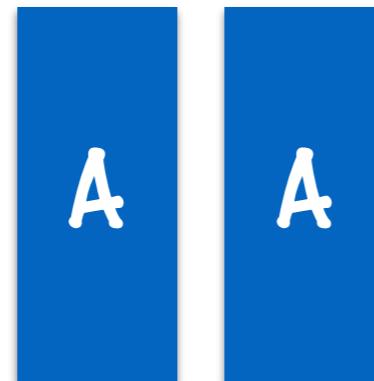
MiRNAs are important negative regulators of gene expression that have been studied intensively over the last few years. To determine their relative RNA abundance, we used four different platforms to measure miRNA levels. In this study, we systematically compared the performance of four platforms by measuring an identical set of 20 standard miRNAs, human universal reference RNA, human spike-in control RNA, and synthetic spikes from miRNA factories. The four platforms were: (1) miRCURY LNA™ qPCR Array (Exiqon), (2) miRCURY LNA™ qPCR Array (Exiqon), (3) miRCURY LNA™ qPCR Array (Exiqon), and (4) miRCURY LNA™ qPCR Array (Exiqon). The results show that all four platforms can be used to evaluate miRNA expression levels.

samplename , description
SRR950876, miRQC_A
SRR950877, miRQC_A_repeat
SRR950878, miRQC_B
SRR950879, miRQC_B_repeat
SRR950880, miRQC_C
SRR950881, miRQC_C_repeat
SRR950882, miRQC_D
SRR950883, miRQC_D_repeat
SRR950884, liver_miR_302a
SRR950885, liver_miR_302b
SRR950886, liver_miR_302c
SRR950887, liver_miR_302d
SRR950888, MS2_let_7a_5p
SRR950890, MS2_let_7c
SRR950889, MS2_let_7b_5p
SRR950891, MS2_let_7d_5p
SRR950892, serum_miRs_constant
SRR950893, serum_miRs_constant_repeat
SRR950894, serum_miRs_variable
SRR950895, serum_miRs_variable_repeat

Support of remote files

```
details:
- algorithm:
  adapters:
  - AGATCGGAAGAG
aligner: star
expression_caller:
- trna
- seqcluster
species: mmu
spikein_fasta: /home/lp113/scratch/charest_egfr_srna/spikeins/all.fa
analysis: smallRNA-seq
description: sampleone
files:
- ftp://ftp.sra.ebi.ac.uk/vol1/ERA169/ERA169754/fastq/NA07000.1.MI_120104_3_1.fastq.gz
genome_build: mm10
metadata: {}
fc_date: '2017-06-21'
fc_name: sample
```

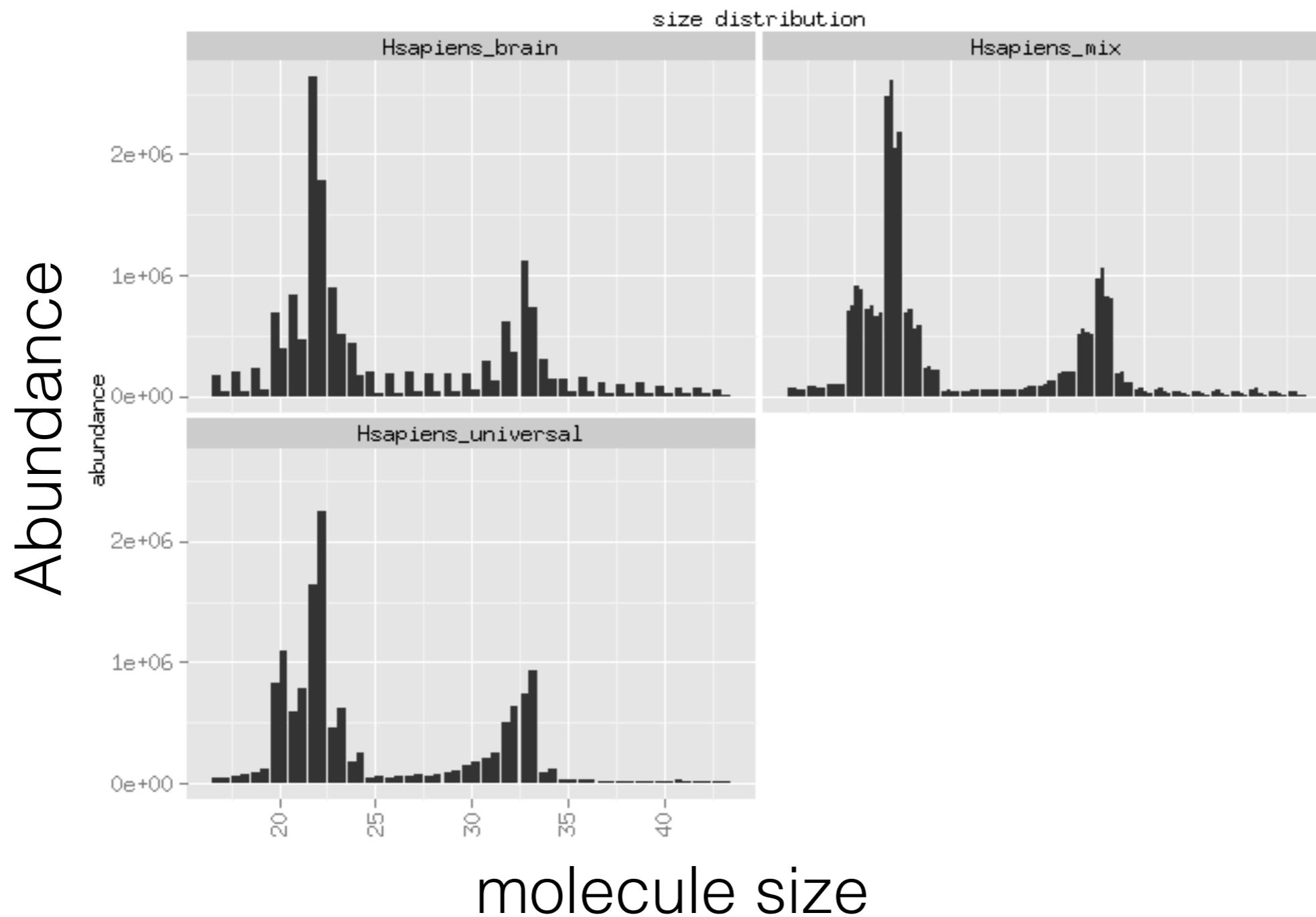
Quality Control samples



For each molecule:

- * If $A > B$ then $A > D > C > B$
- * If $B > A$ then $A < D < C < B$

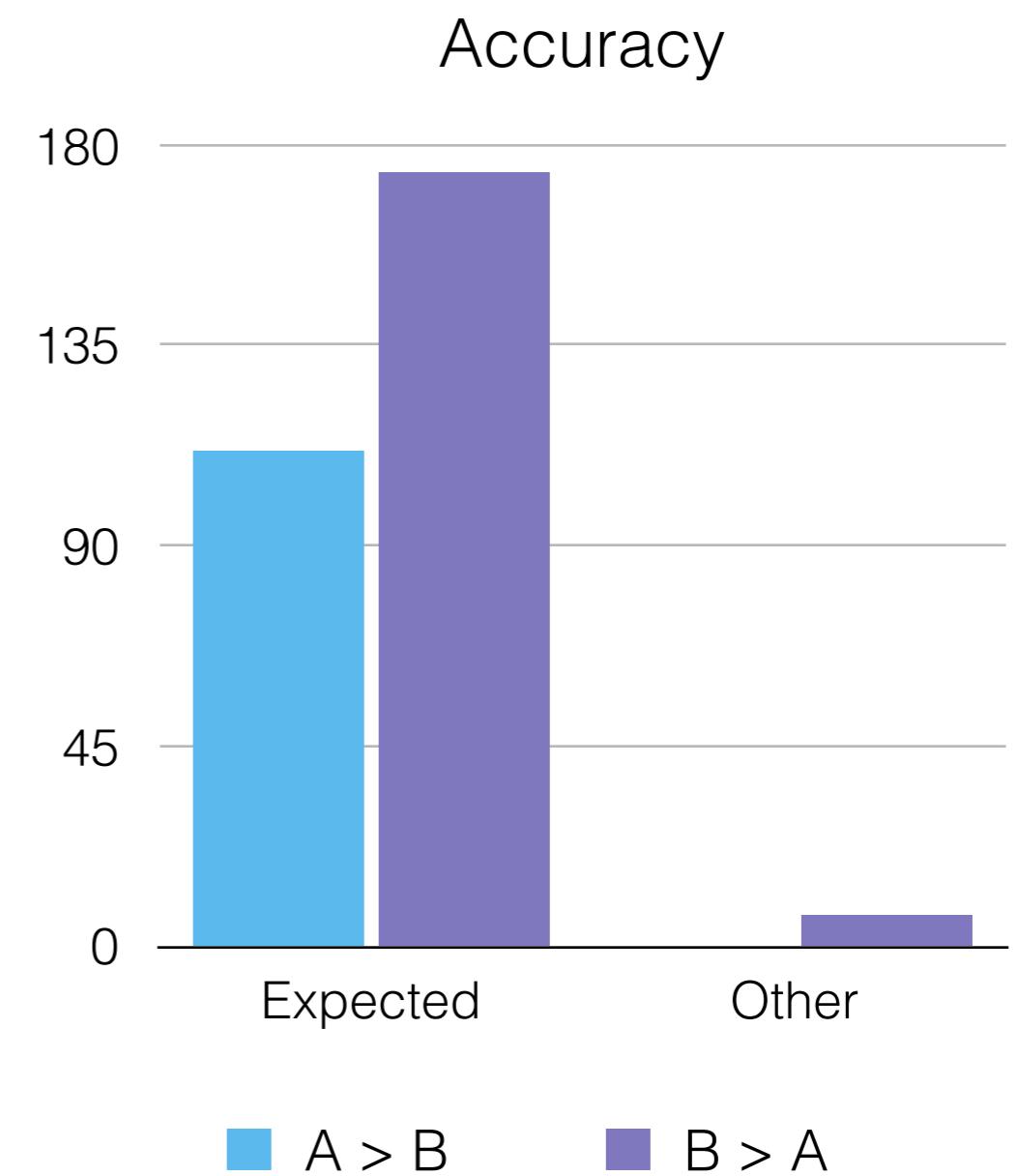
Good samples



miRNA quantification

miRNAs > 5 counts in average

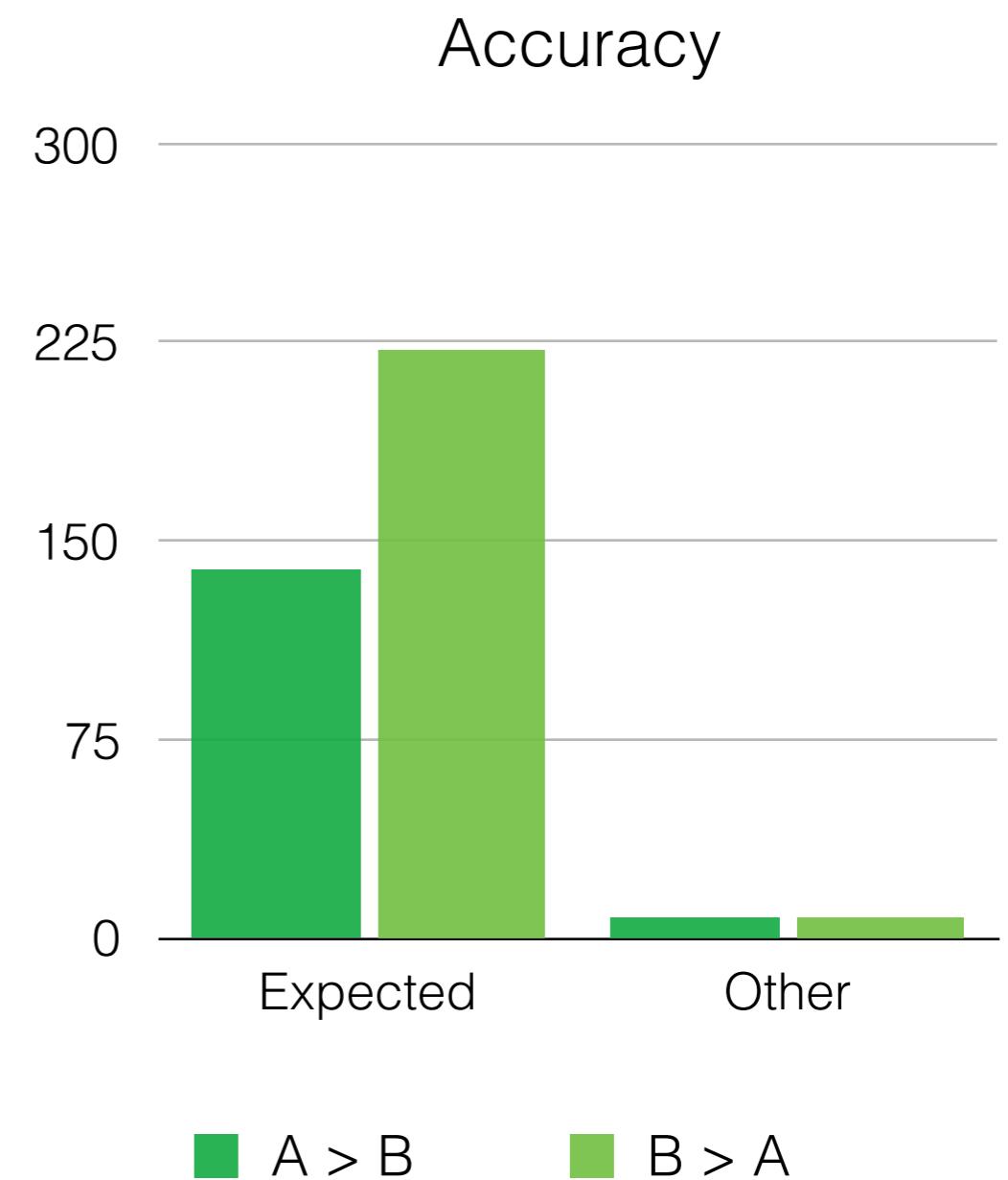
upper quantile normalization



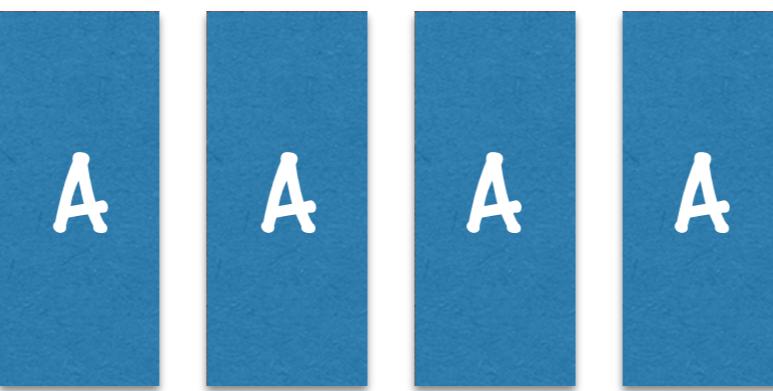
clusters quantification

expression > 5 counts in average

upper quantile normalization



Positive controls



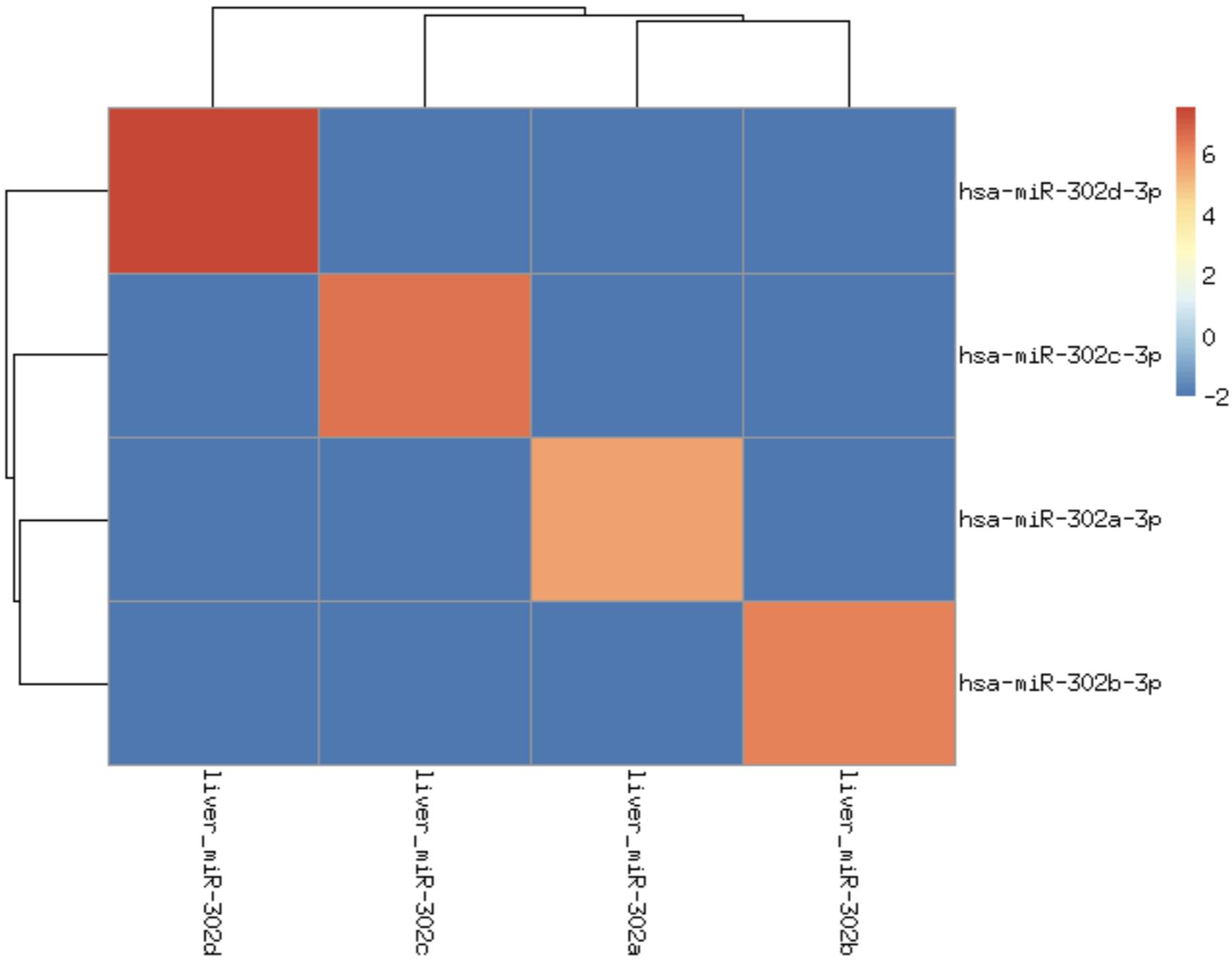
miR-302a

miR-302b

miR-302c

miR-302d

Specificity



bcbio template

```
upload:  
    dir: ../final  
  
details:  
    - analysis: smallRNA-seq  
  
        algorithm:  
            aligner: star  
  
            # change adapter according project  
            adapters: ["TGGAATTCTCGGGTGC"]  
            expression_caller: [trna, seqcluster, mirdeep2]  
            species: hsa  
  
            genome_build: hg19
```

<https://github.com/chapmanb/bcbio-nextgen/blob/master/config/templates/illumina-srnaseq.yaml>

Resources

	Time (h)
organize	0:01
adapter	0:27
alignment	0:26
annotation	3:43
cluster + mirdeep2	4:15
qc	0:04

The time for 8 samples with 6 millions reads each was 8 hours and 57 minutes.



Boston-area Women's Bioinformatics Meetup

★★★★★ 35 ratings

📍 Cambridge, MA

👤 500 members · Private group ?



Organized by
Lorena Pantano and 3 others

Share: [f](#) [t](#) [in](#) [↗](#)

About

Meetups

Members

Photos

Discussions

More

Manage group ▼

Plan a Meetup ▼

Next Meetup

See all

2
OCT

Tuesday, October 2, 2018, 6:00 PM

Develop your Bioinformatics Shiny App - day

2 out of 4

A shiny app - day

OCT
2

6:00 PM, October 2, 2018



thanks

Harvard T.H. Chan School of Public Health



Research Computing at Harvard Medical School: Chris Botka,
Director of Research Computing and all the people in the
team.



miRToP consortium



Teaching Team