

# Characterization of the small RNA transcriptome

Lorena Pantano

@lopantano [lpantano@hsph.harvard.edu](mailto:lpantano@hsph.harvard.edu)  
Harvard TH Chan School of Public Health

<https://goo.gl/uZWG0E>

2016-12-02

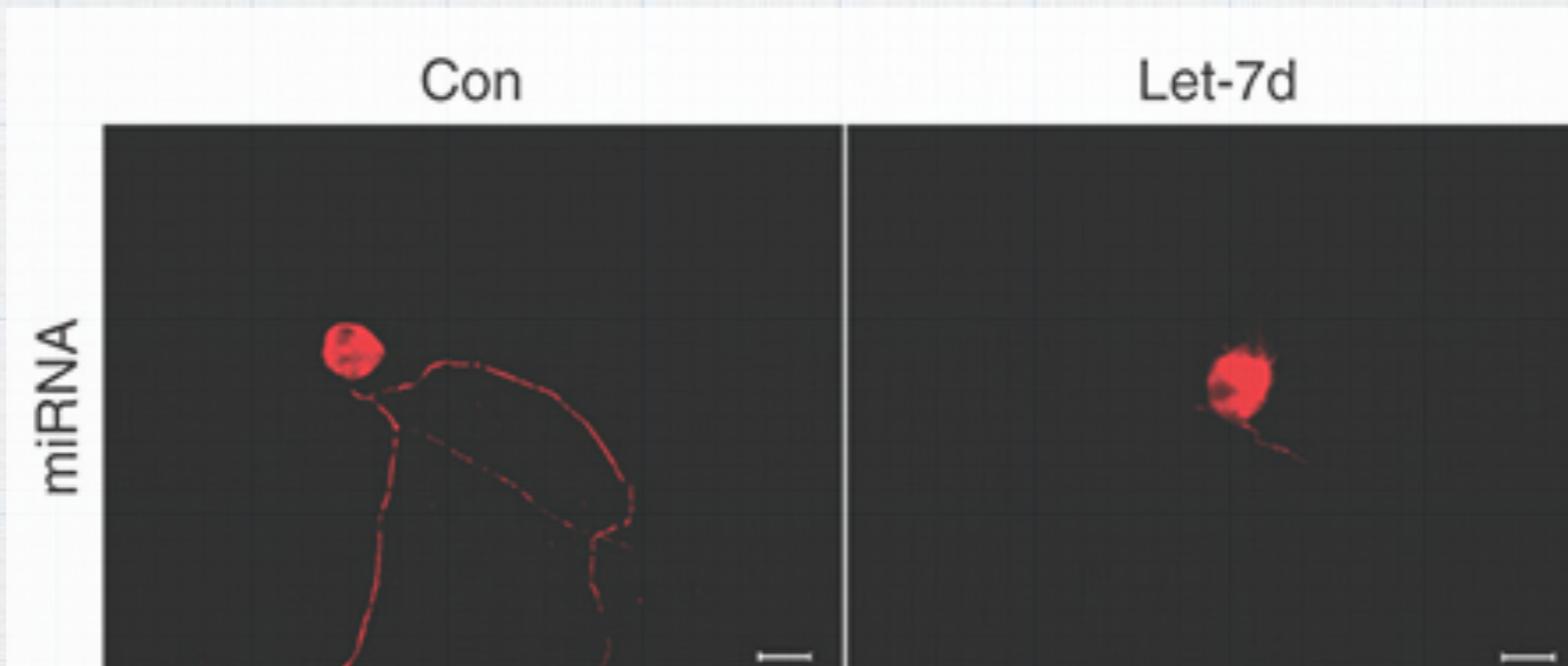
# small RNA

RNA molecules of 18-36 nts  
long with regulation  
function



# 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](#)

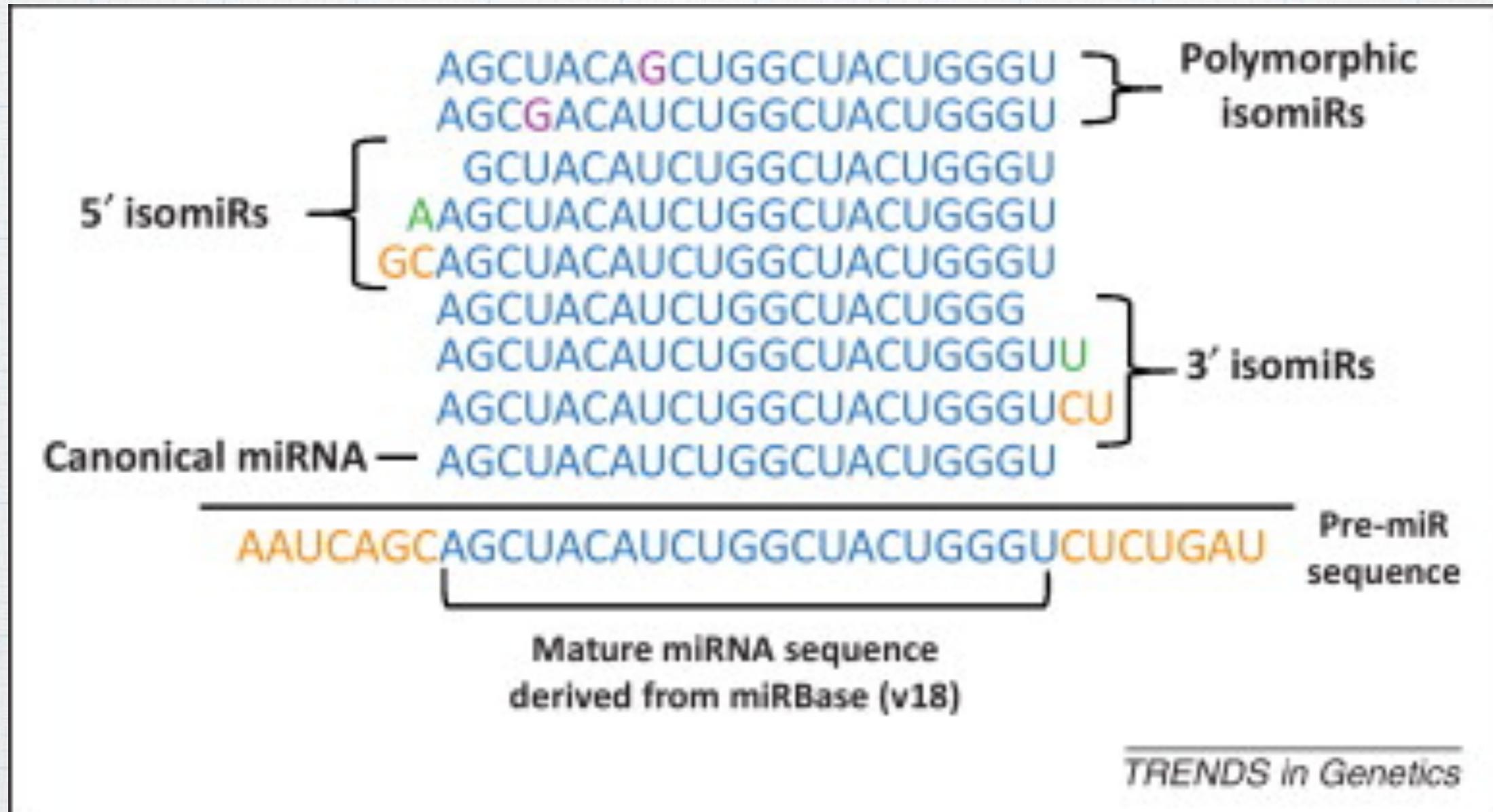
GGUGCCUACUGAGCUGAUUAUC  
GUGCCUACUGAGCUGAUUAUCAGU  
GUGCCUACUGAGCUGAUUAUCAG  
GUGCCUACUGAGCUGAUUA  
UGCCUACUGAGCUGAUUAUCA  
UGCCUACUGAGCUGAUUAUCAGU  
UGCCUACUGAGCUGAUUAUC  
UGCCUACUGAGCUGAUUA  
CCUACUGAGCUGAUUAUCA  
CCUACUGAGCUGAUUAUCAGU  
CUACUGAGCUGAUUAUCA  
CUACUGAGCUGAUUAUC

## hsa-miR-24-3p

GGUGCCUACUGAGCUGAUUAUC . . . . .  
.. GUGCCUACUGAGCUGAUUAUCAGU . . . . .  
.. GUGCCUACUGAGCUGAUUAUCAG . . . . .  
.. GUGCCUACUGAGCUGAUUA . . . . .  
.. UGCCUACUGAGCUGAUUAUCA . . . . .  
.. UGCCUACUGAGCUGAUUAUCAGU . . . . .  
.. UGCCUACUGAGCUGAUUAUC . . . . .  
.. UGCCUACUGAGCUGAUUA . . . . .  
.. . CCUACUGAGCUGAUUAUCA . . . . .  
.. . CCUACUGAGCUGAUUAUCAGU . . . . .  
.. . CUACUGAGCUGAUUAUCA . . . . .  
.. . CUACUGAGCUGAUUAUC . . . . .

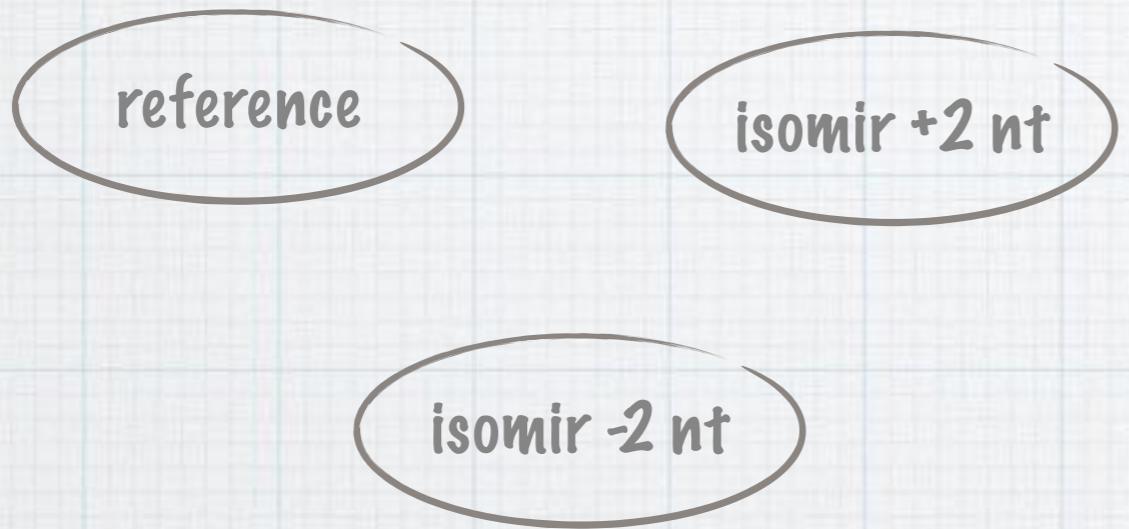
# precursor

# types of isomiRs



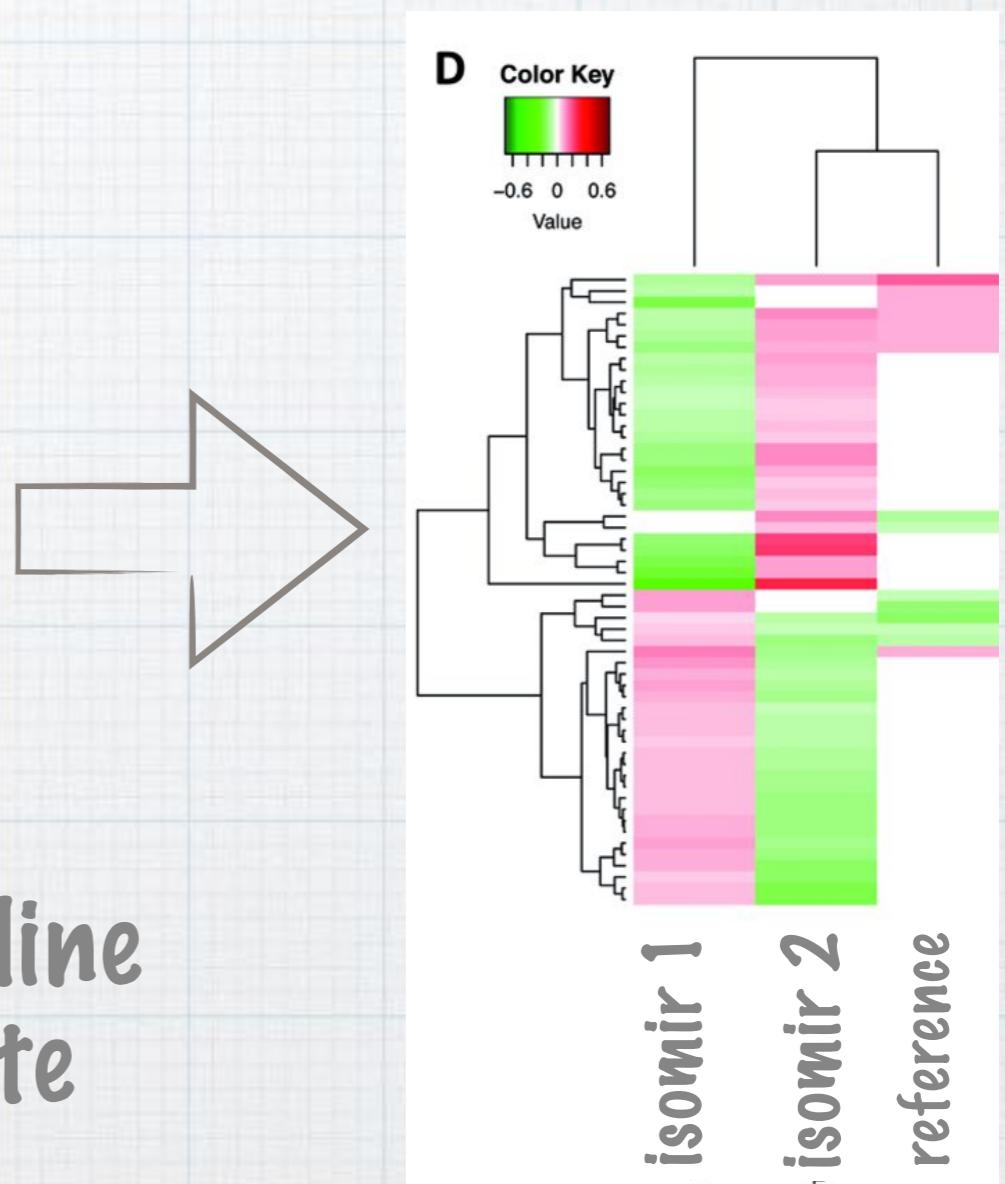
TRENDS in Genetics

# isomiRs

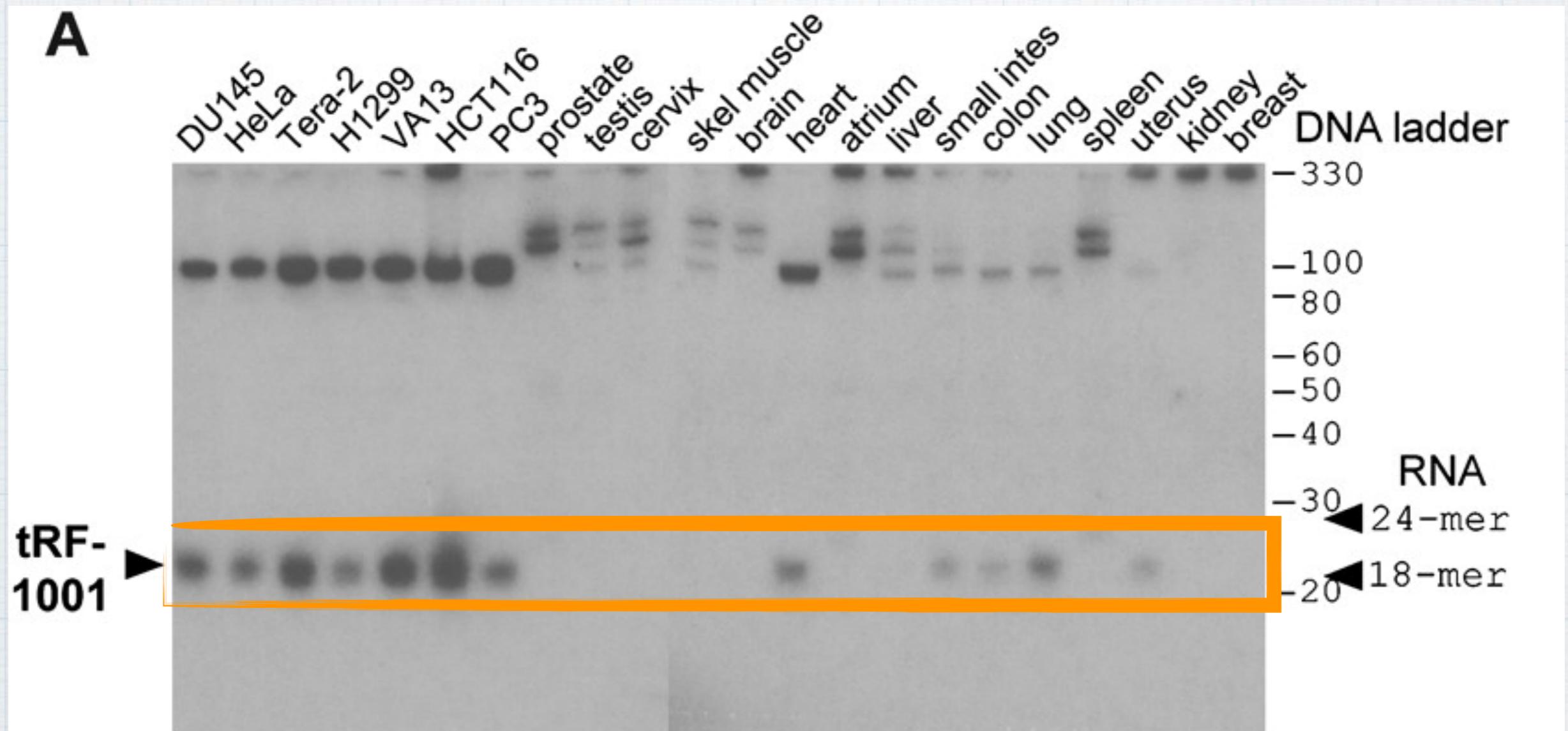


transfected mammary cells line  
derived from metastatic site

Gene expression



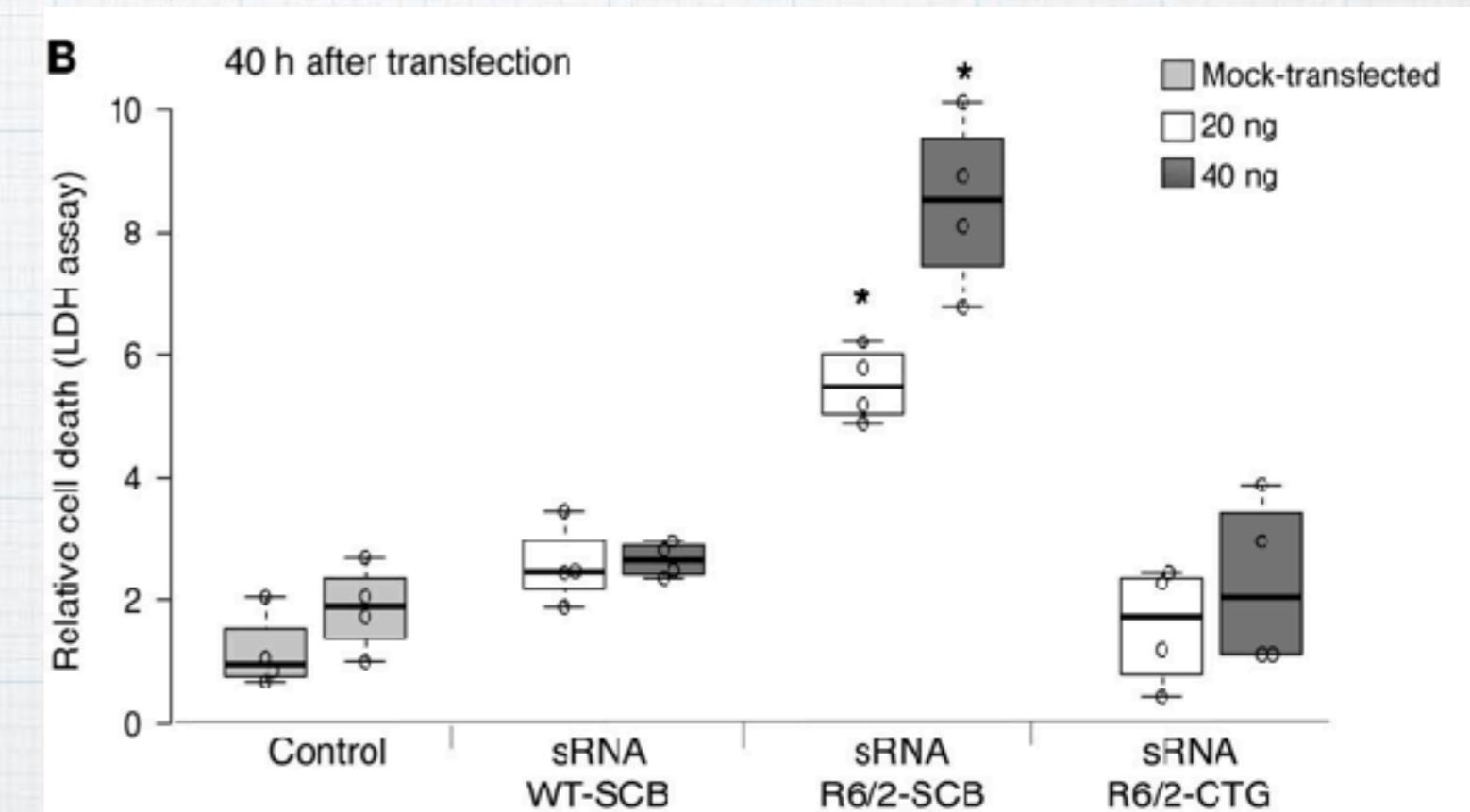
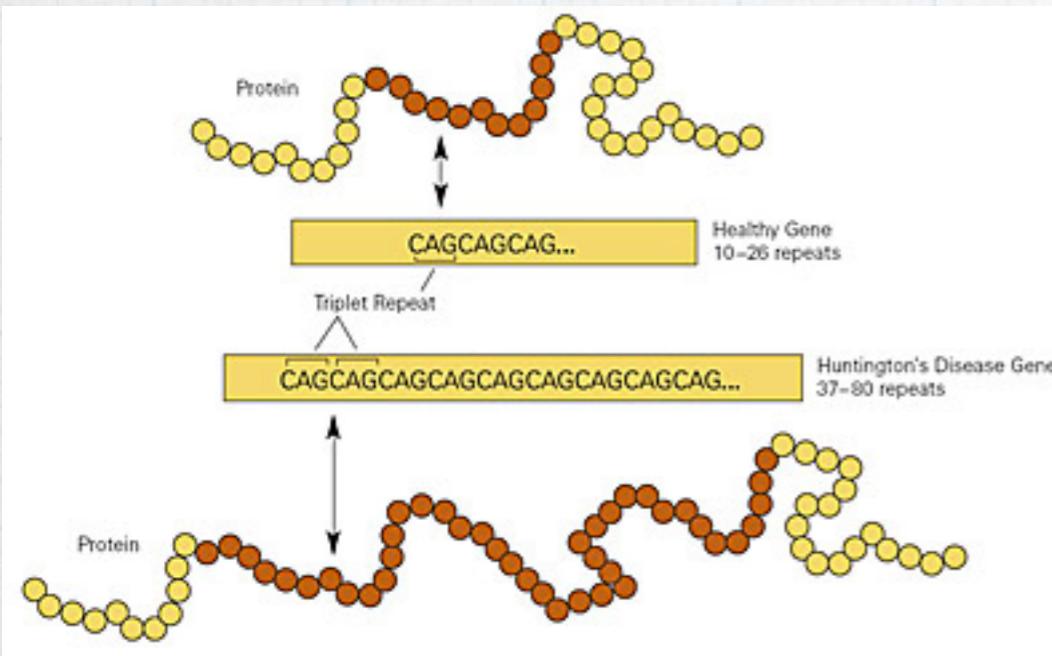
# small tRNAs



Yong Sun Lee et al. Genes Dev. 2009;23:2639-2649

# small RNA

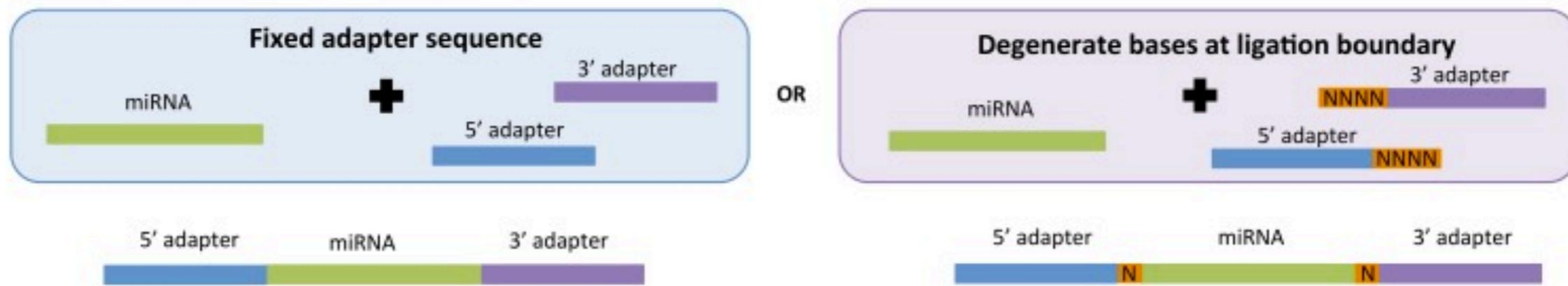
## Huntington disease therapy



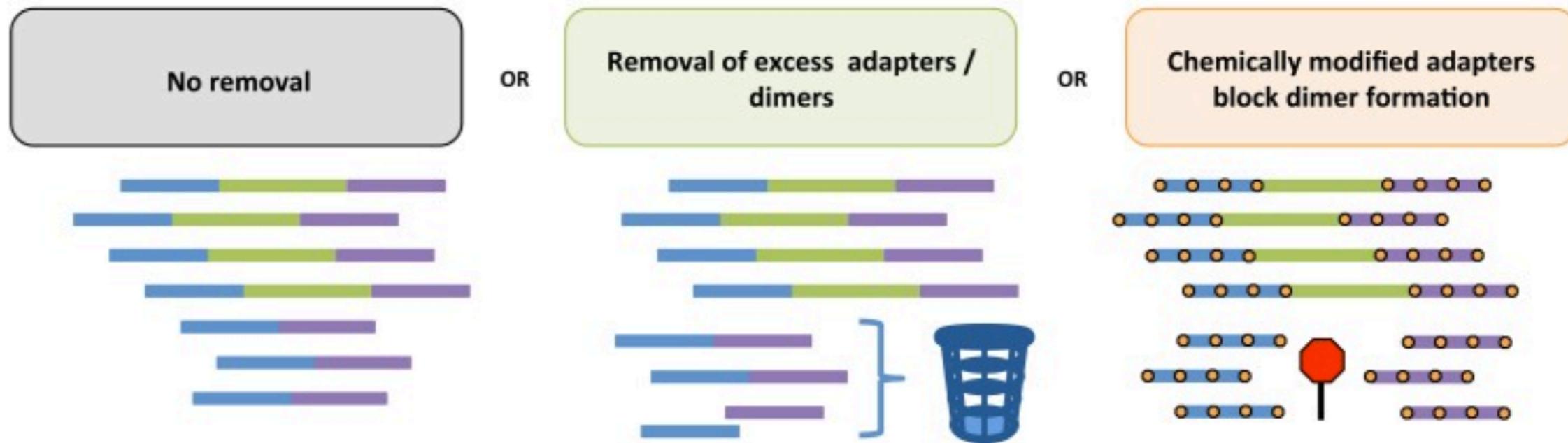
# Protocols

## Critical differences in small RNA library preparation protocols

### Issue 1: Adapter ligation introduces bias

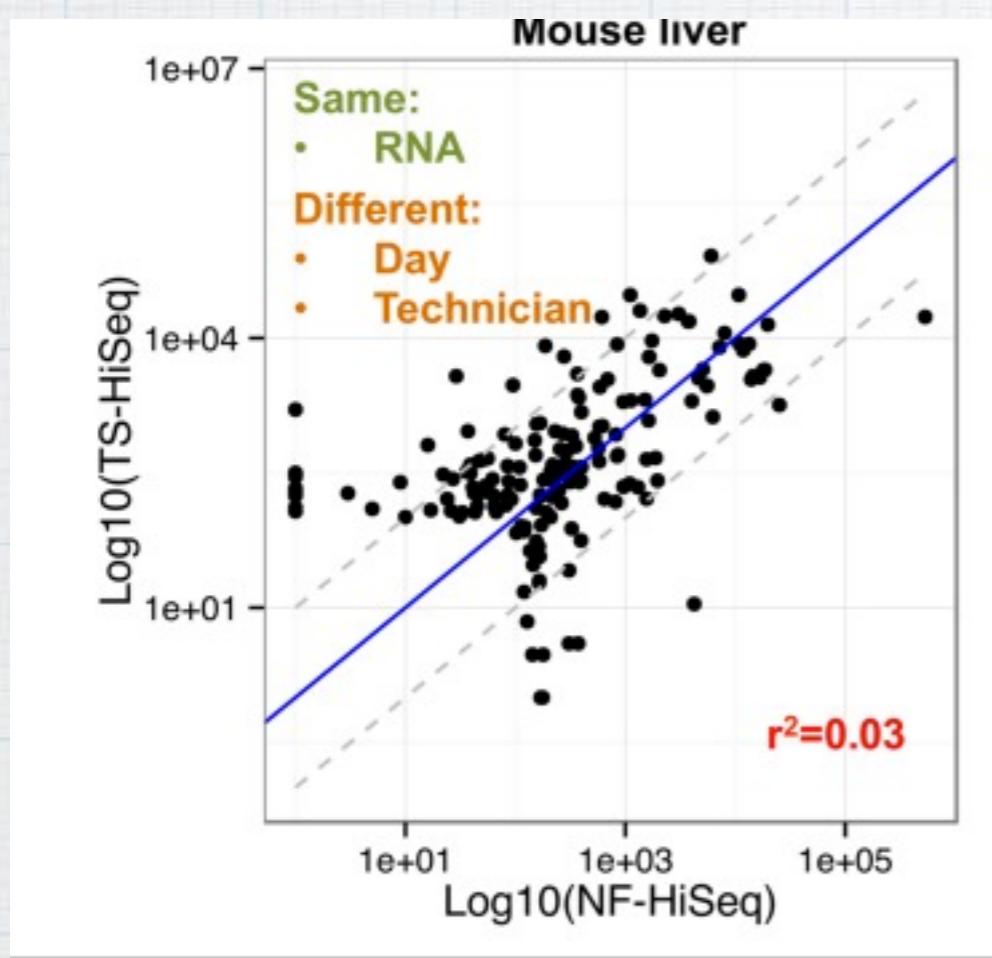


### Issue 2: Adapter dimers compete with small RNAs, reducing effective sequencing depth

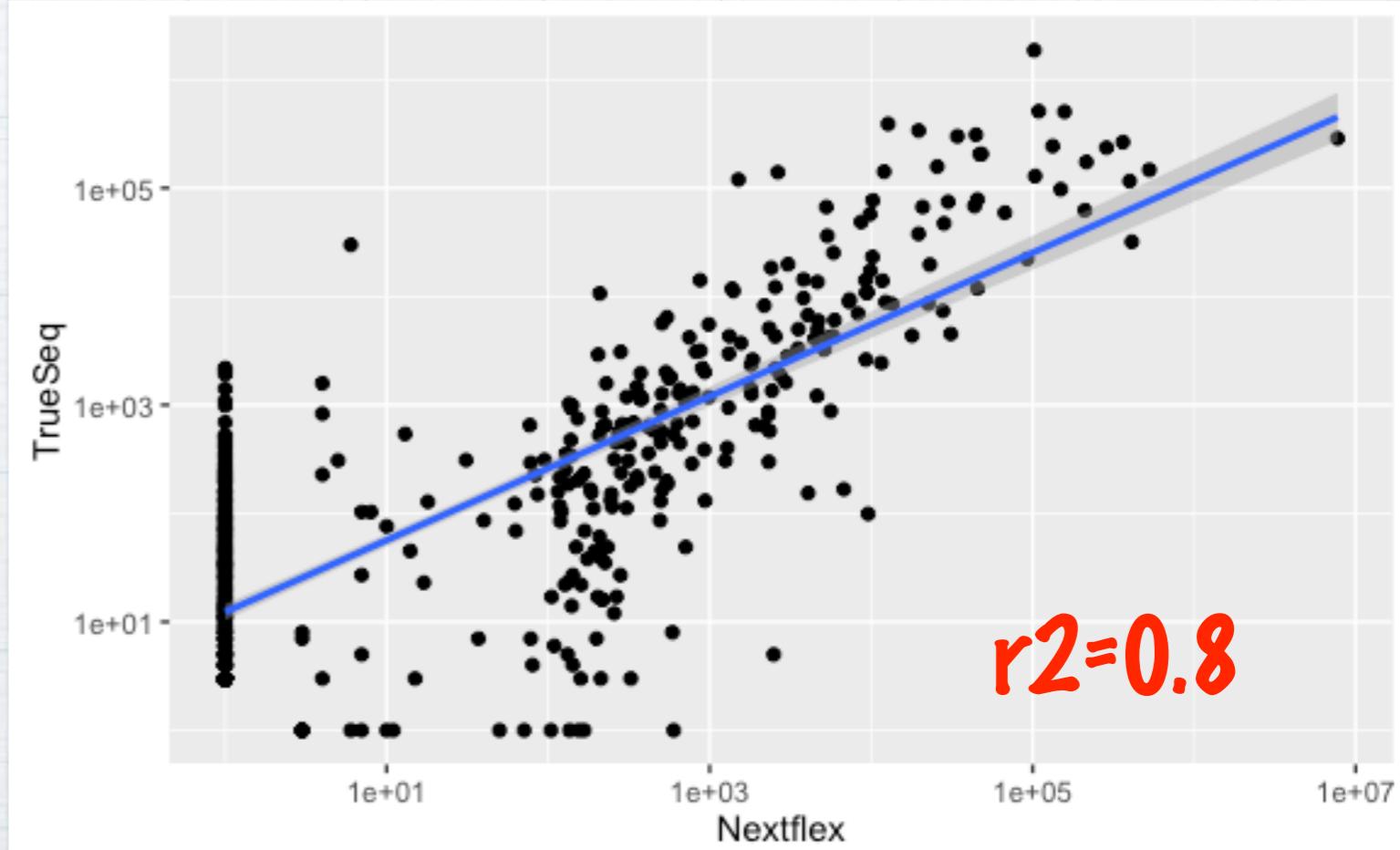


# Protocol correlation

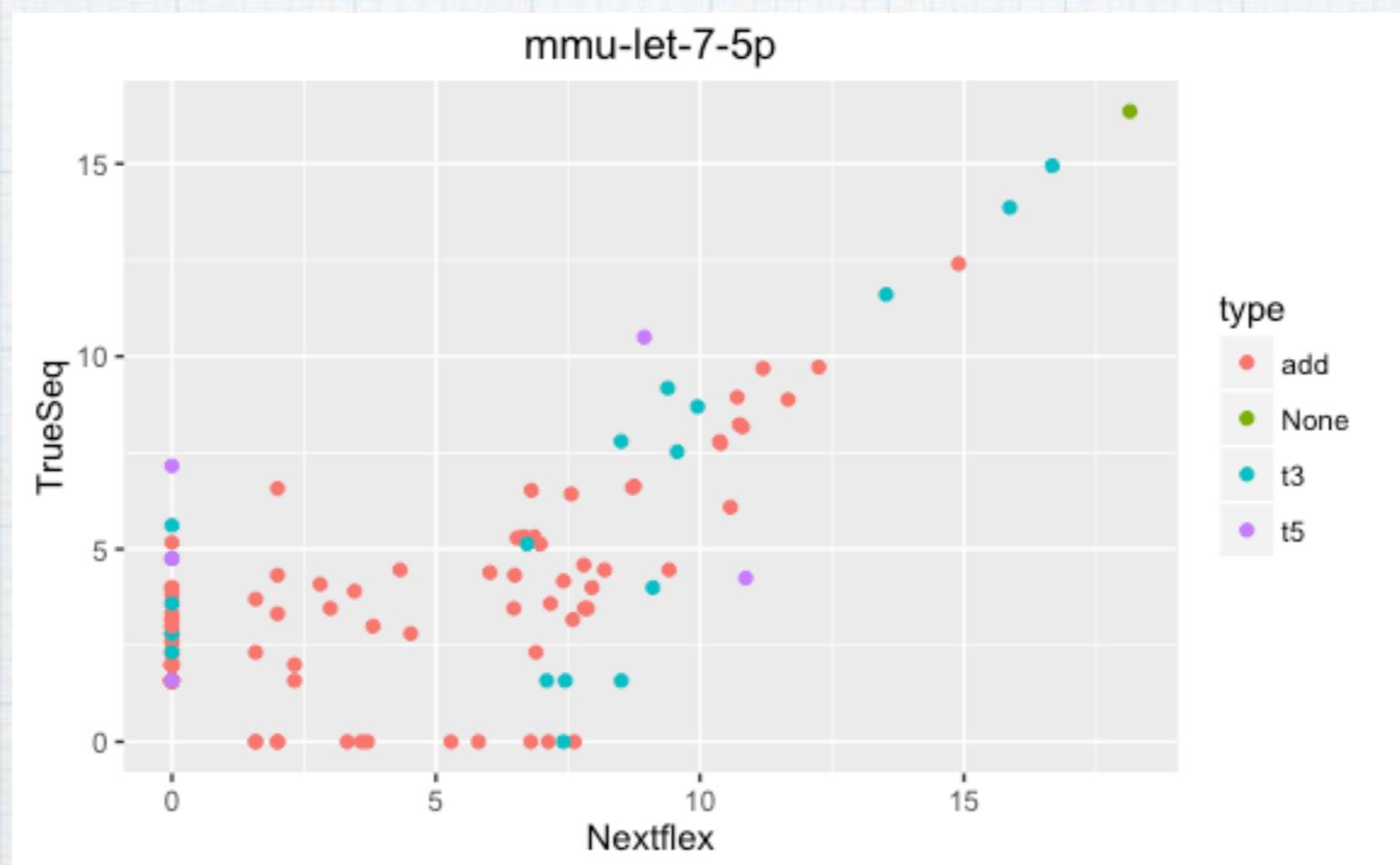
Paper figure



bcbio pipeline



# let7-a-5p miRNA



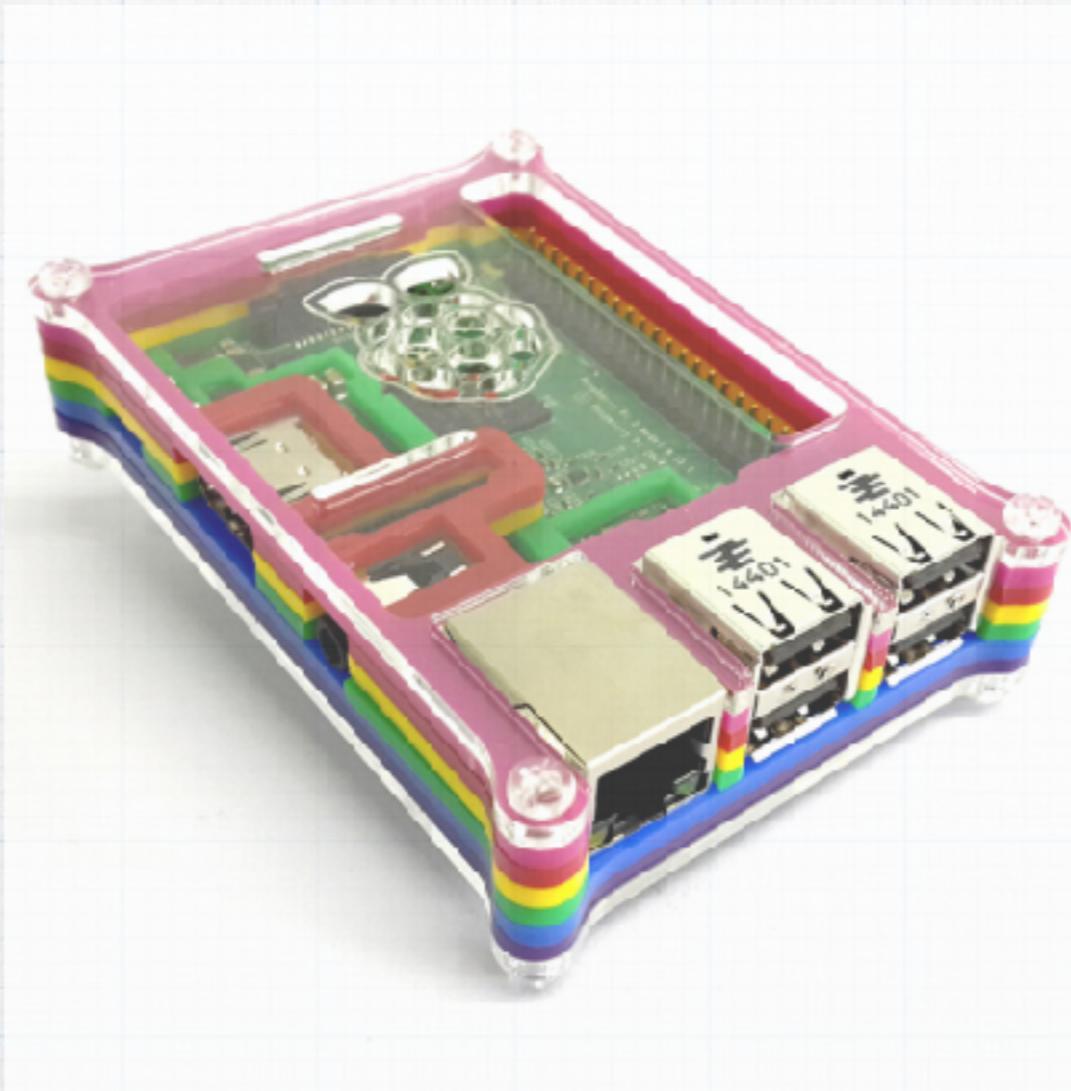
# Caveats

- \* TrueSeq Illumina: ligation bias
- \* NextFlex Bioo Scientific: generation of random sequences?. We lose the accuracy to detect isomiRs

# 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**<sup>®</sup>

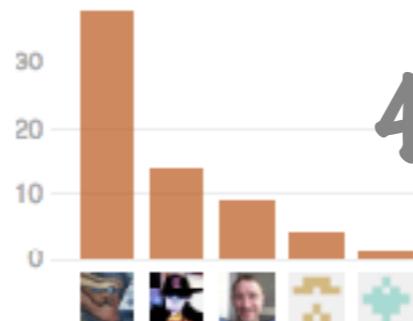
May 29, 2016 – June 29, 2016

Period: 1 month ▾

## Overview

		 	
2 Active Pull Requests		74 Active Issues	
 2 Merged Pull Requests	 0 Proposed Pull Requests	 62 Closed Issues	 12 New Issues

Excluding merges, **5 authors** have pushed **66 commits** to master and **66 commits** to all branches. On master, **68 files** have changed and there have been **1,085 additions** and **393 deletions**.



# 41 contributors

# small RNA-seq analysis

processing & QC

cutadapt  
fastqc  
qualimap  
multiqc

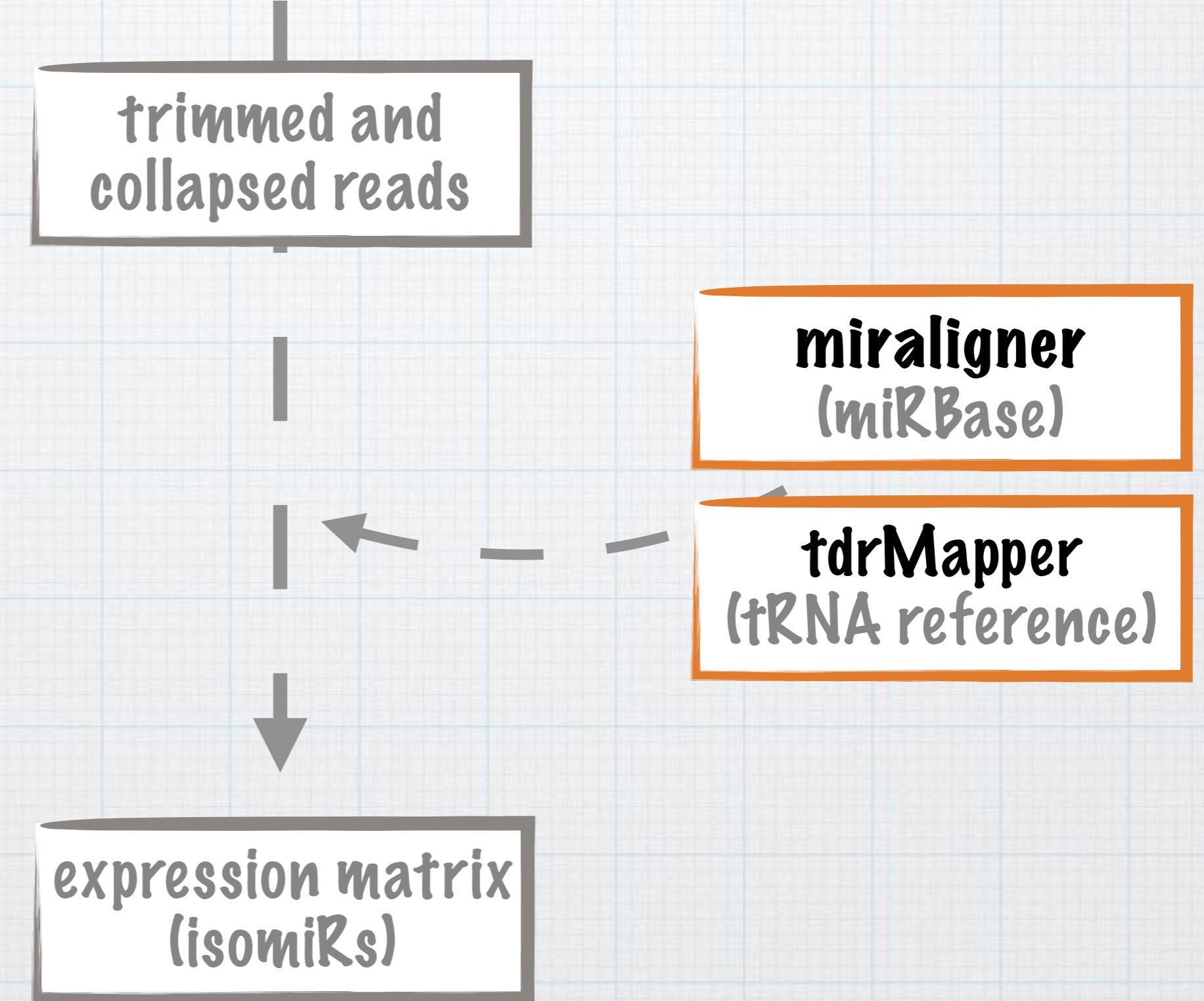
detection & annotation

miraligner  
tdrmapper

de-novo

seqcluster  
mirdeep2 for mirna  
protac for piRNA (next)

## detection & annotation



# miRNA naming

miRNA in database

isomiR

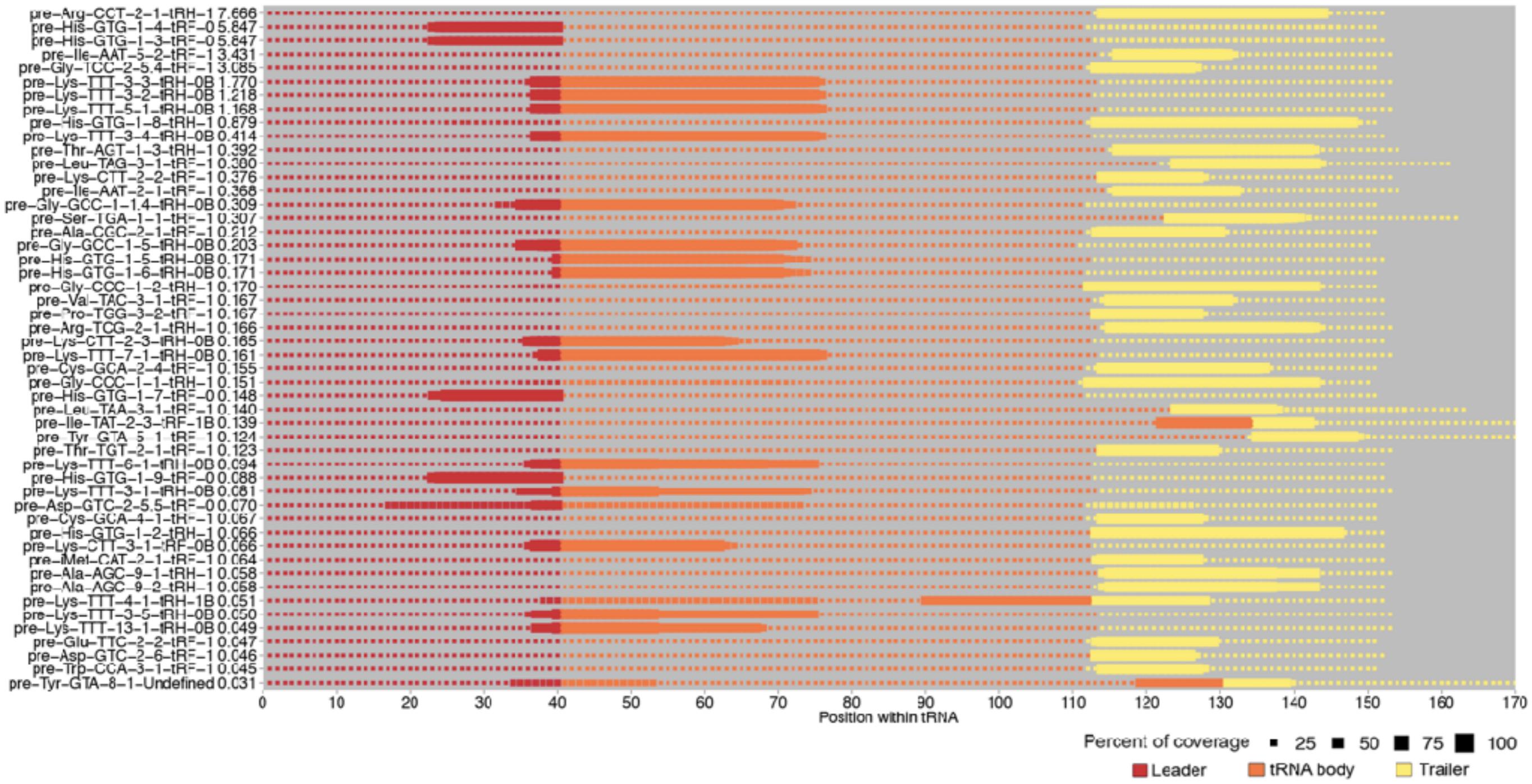
UPPER CASE: addition  
lower cases: deletion

mismatch      addition  
trimming 5'    trimming 3'

miRNA\_name:mismatch:addition:t5:t3

hsa-let-7a-5p:0:0:GT:t

# tRNA analysis



\*Pre-tRNA coverage map from NIH roadmap H1 derived mesendoderm cells, accession ID: GSM1296464

## de-novo detection

trimmed and collapsed reads

collapsing samples into one

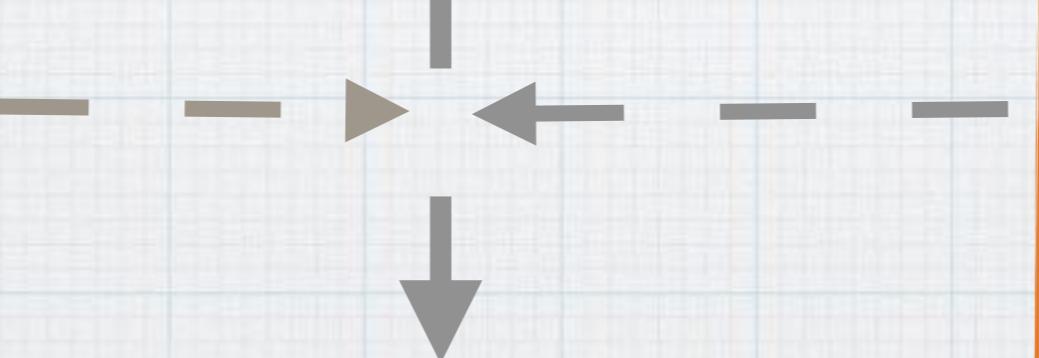
align to genome

**seqcluster**  
(genome and annotation)

**mirdeep2**  
(genome and annotation)

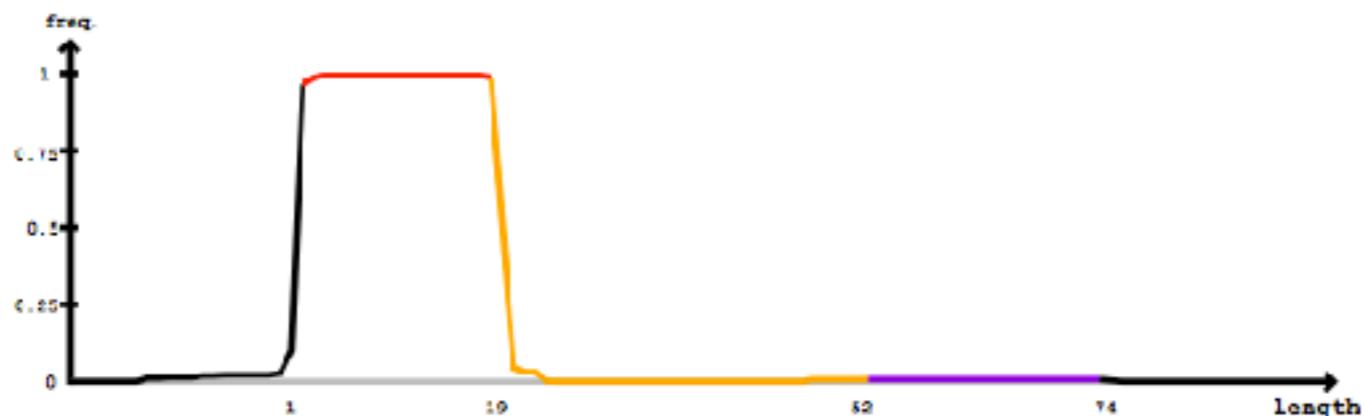
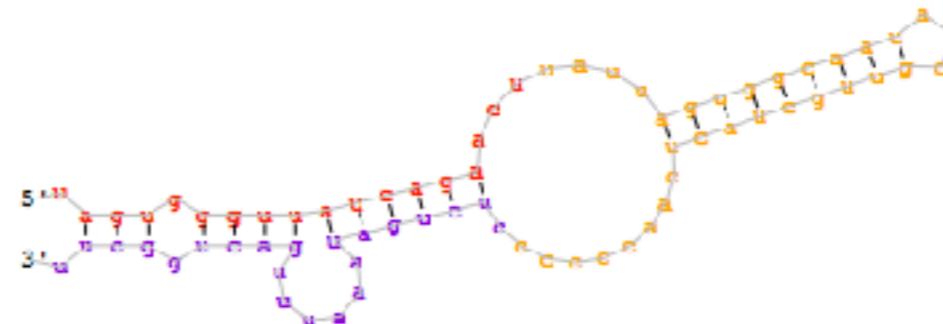
visualization & quantification

expression matrix  
(DESeq2)



# 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 mfc	:	-1
Score for rand/fold	:	
Score for cons. seed	:	
Total read count	:	3673
Mature read count	:	3670
Loop read count	:	0
Star read count	:	3



# seqcluster



cluster at position 1

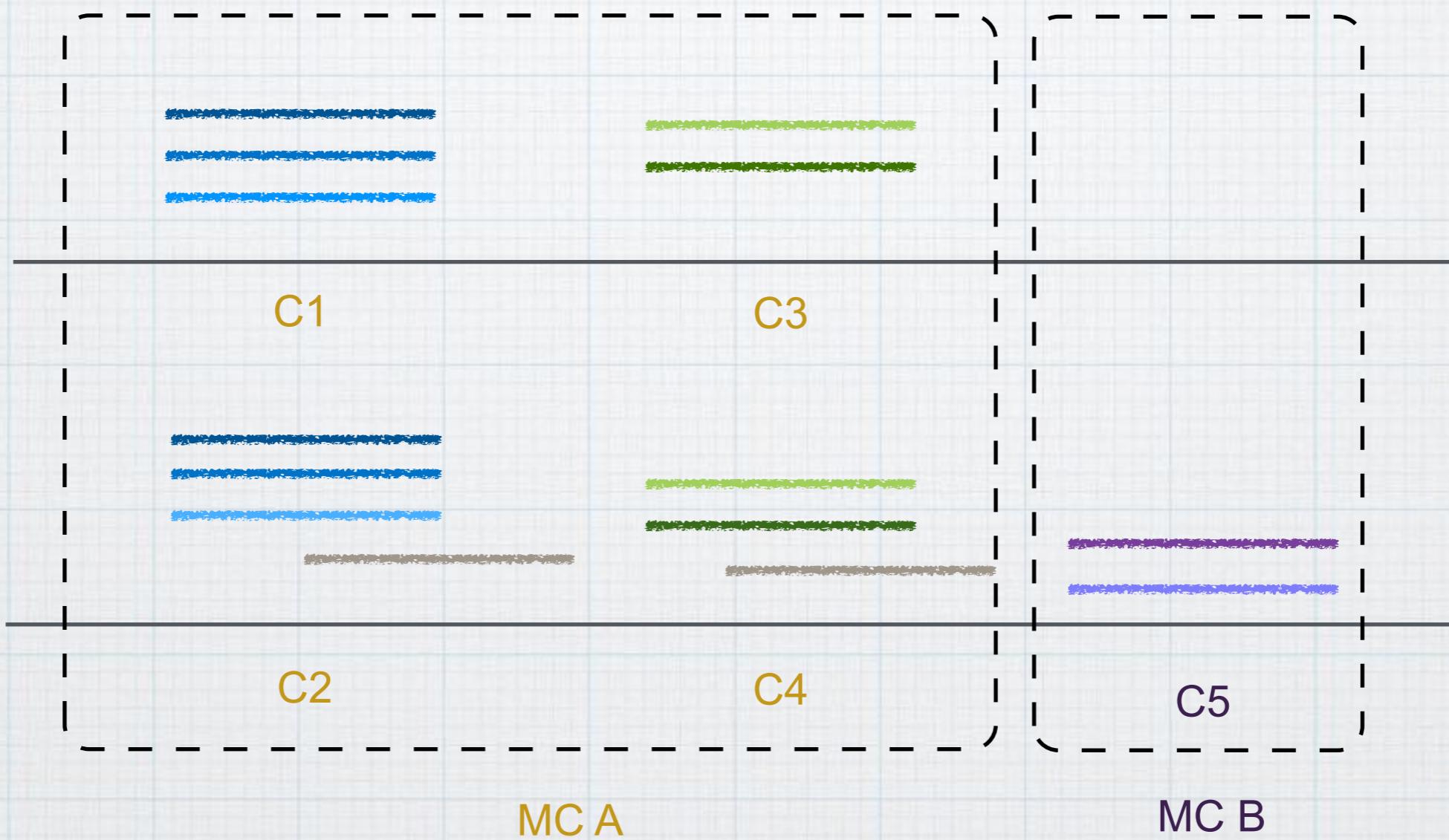


cluster at position 2

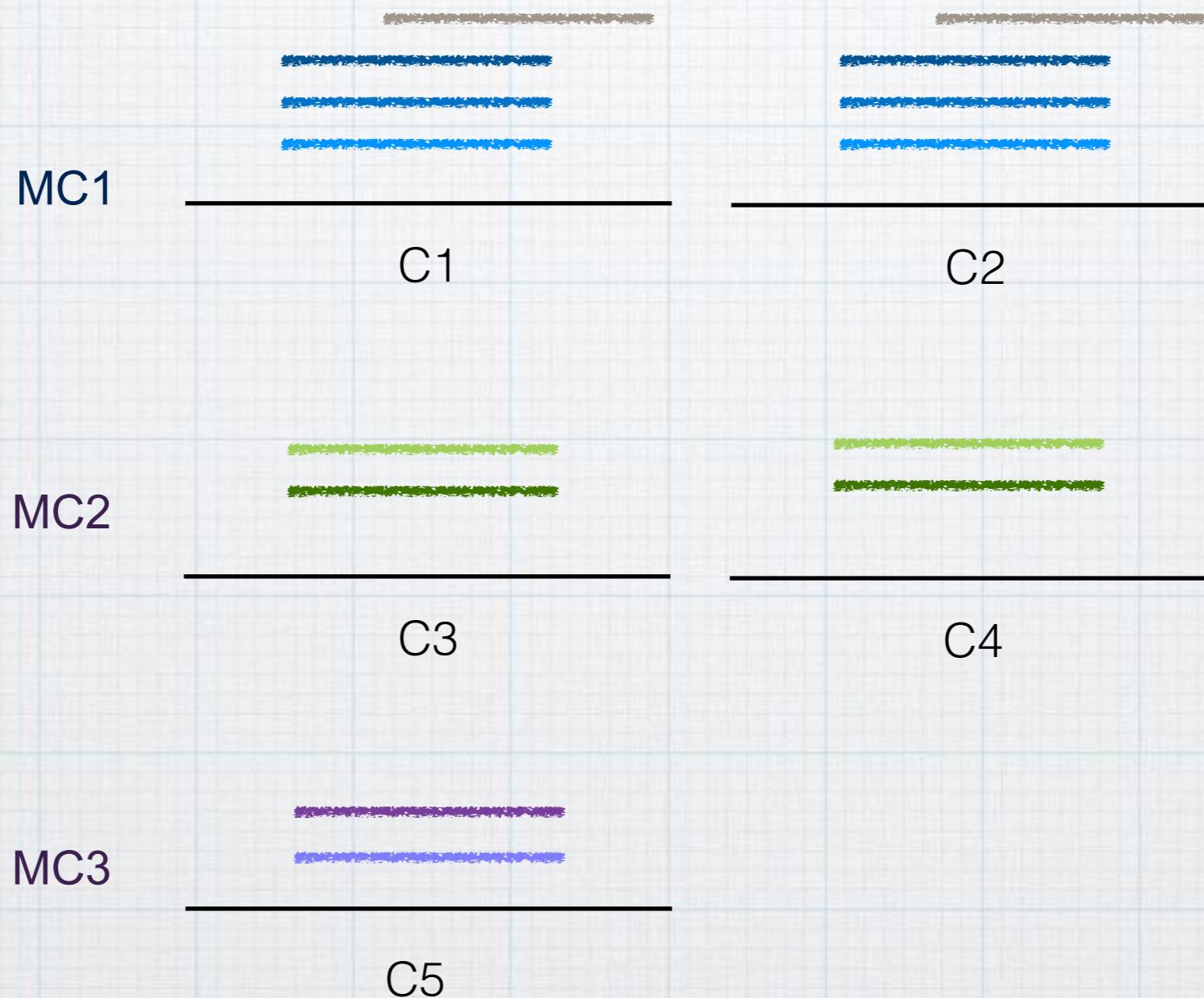
meta-cluster

seqcluster deals with multi-mapped reads

# Step 1: clustering



# Step 2: cleaning



# seqcluster visualization

The screenshot shows a web browser window with the title "clusters information". The address bar displays the URL "file:///Users/centano/repos/seqclusterViz/reader.html". The browser's toolbar includes standard icons for back, forward, search, and file operations. Below the toolbar, the OS X menu bar shows "FileC 3.3: BUILD/Check", "Files - OneDrive", "Home - Dropbox", "Timesheet - HBC - Ho", "Minor Categories - Mo", "Build your own Images", "Home - PubMed - NC", and "Deep e Scholar".

The main content area contains several sections:

- Browse...**: A button for navigating files.
- Clusters Filter:** An input field for filtering cluster results.
- Clusters Id:** An input field for specifying a specific cluster ID.
- Table with clusters**: A table with columns: Sel., I.D., and Description. No data is present.
- Table with Locus**: A table with columns: I.D., Index, and Locus. No data is present.
- Abundance profile along precursor**: A placeholder text label.
- Secondary structure**: A placeholder text label.

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

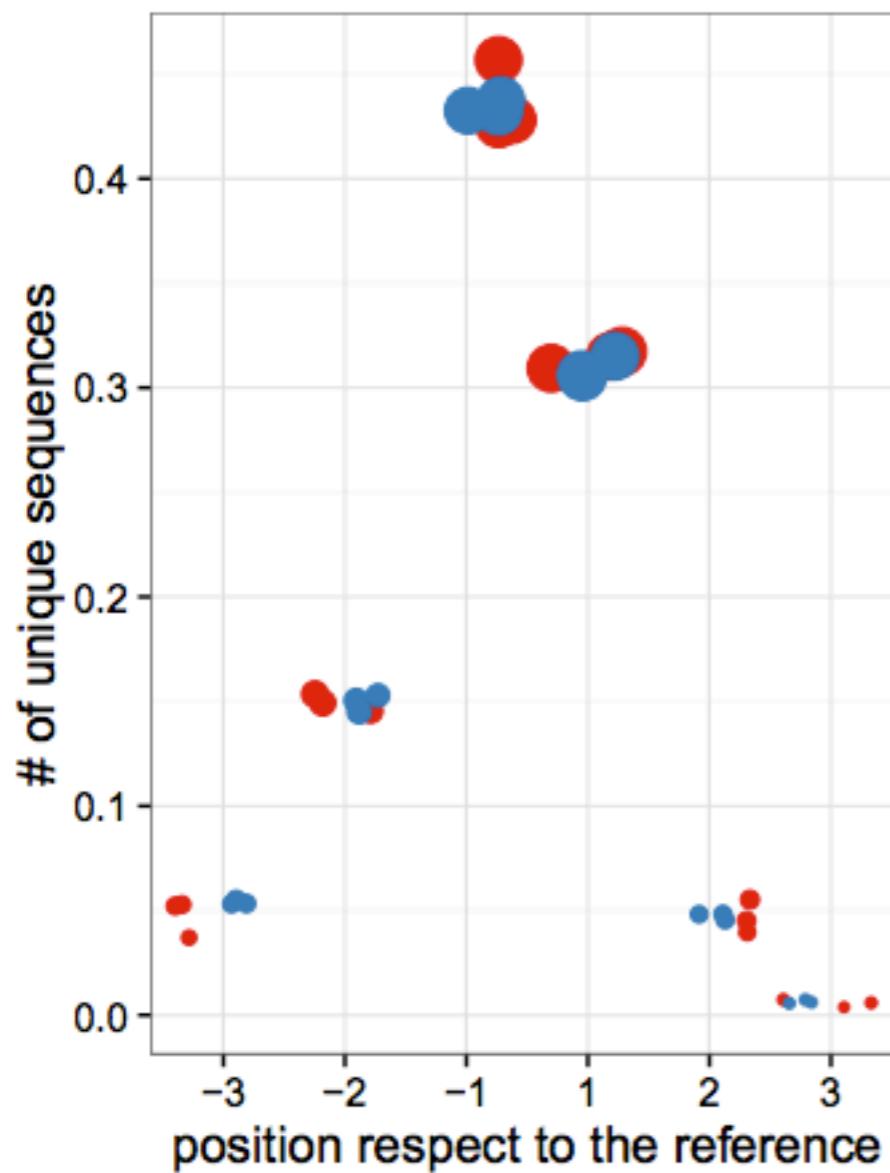
# isomiRs: R package

- \* General Characterization of isomiRs
- \* Collapsing isomiRs in different ways
- \* Supervised clustering analysis to detect important miRNAs (PLS-DA)
- \* RNAseq and miRNA time serie data

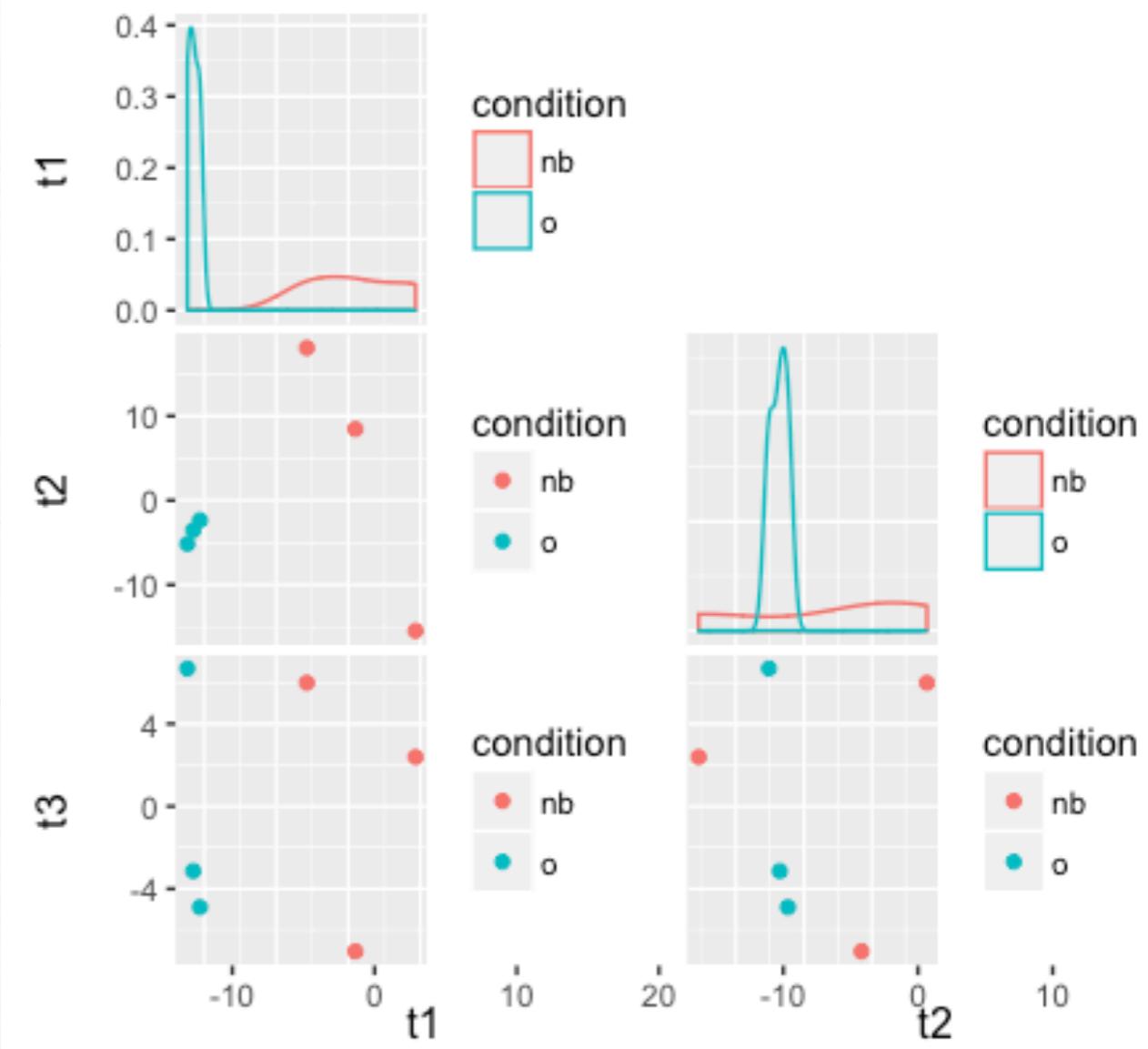
[http://bioconductor.org/packages/release/bioc/html/  
isomiRs.html](http://bioconductor.org/packages/release/bioc/html/isomiRs.html)

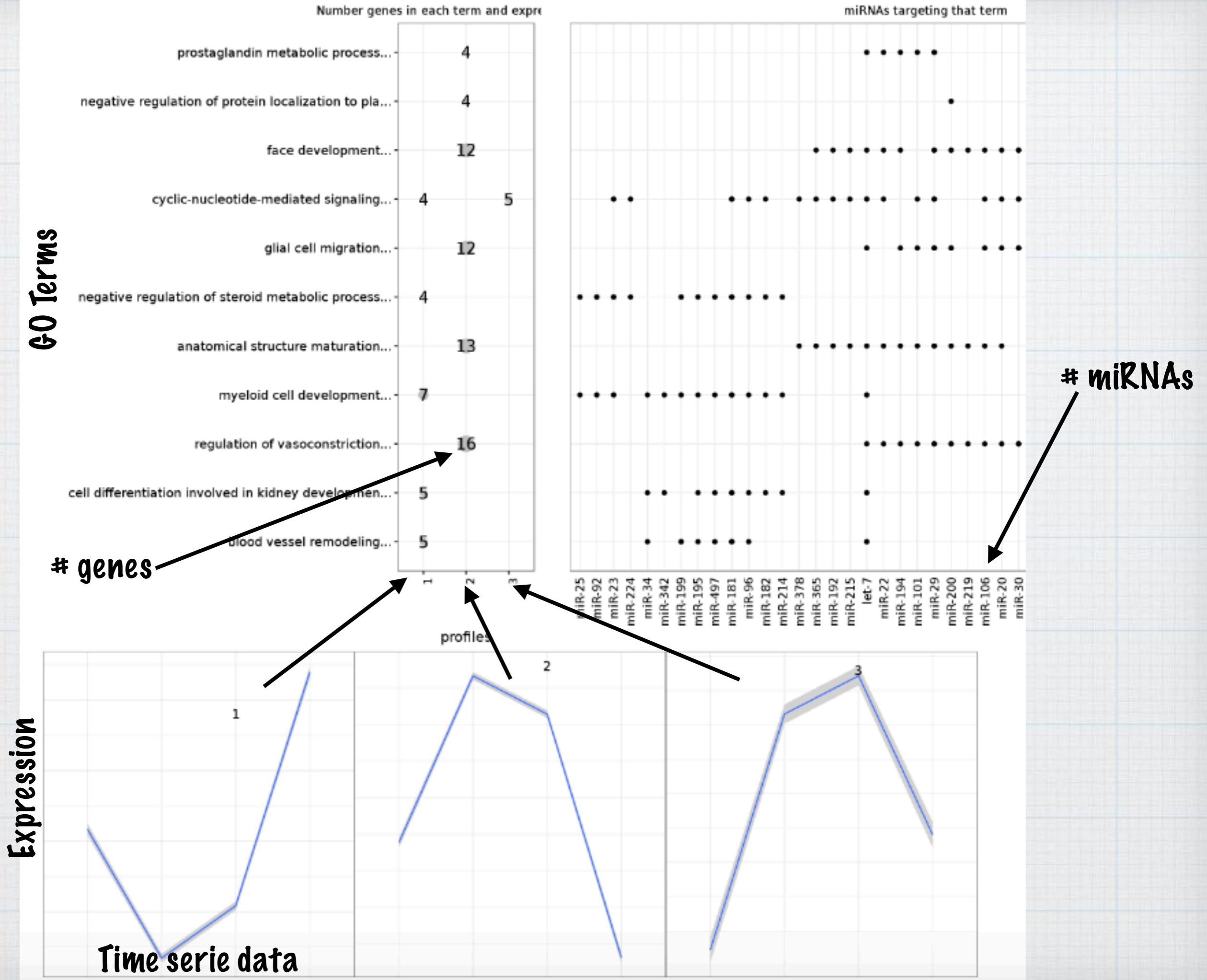
# isomiRs: R package

iso5 distribution



PLS-DA





# MultiQC



**Phil Ewels**

ewels

Bioinformatician working with next generation sequencing data.

Science for Life Laboratory

Stockholm, Sweden

phil.ewels@scilife.se

<http://phil.ewels.co.uk>

Joined on Nov 3, 2010

**48**

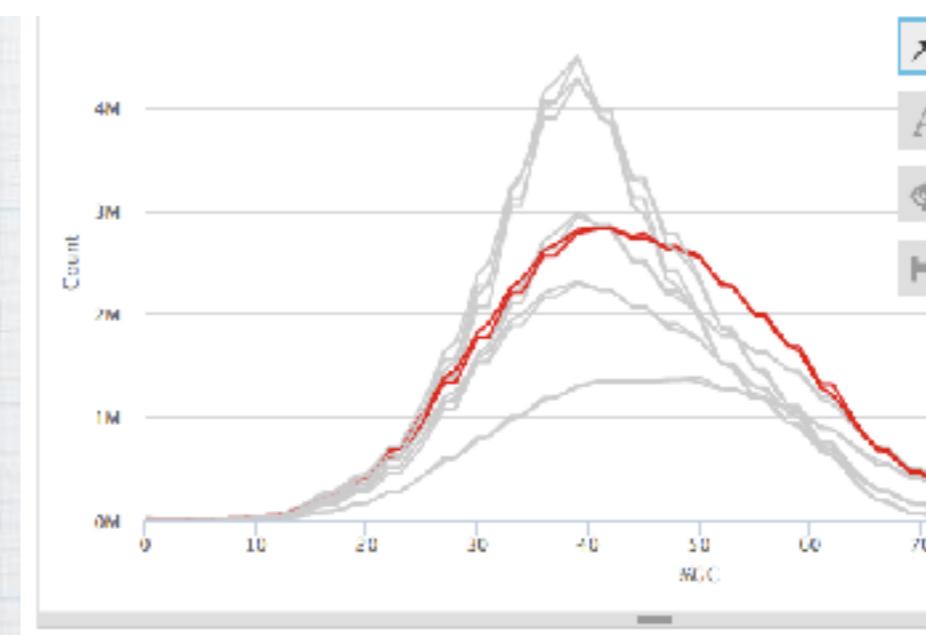
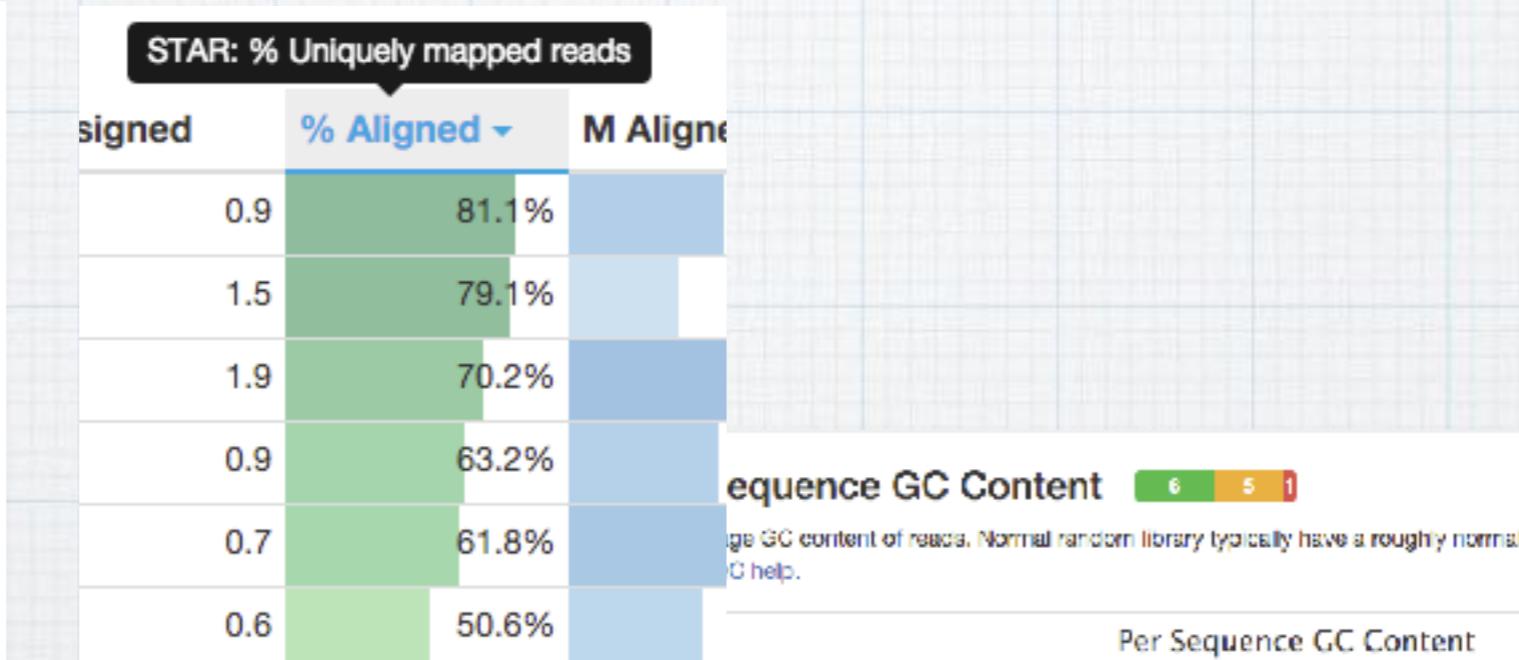
**21**

**23**

Followers

Starred

Following



## MultiQC Toolbox

### Highlight Samples

6:0

Regex mode off

II 6:0

X

# miRQC project

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

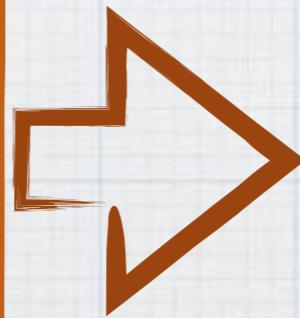
Samples (20)  
≡ Less...

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  
fastq/\*fastq.gz



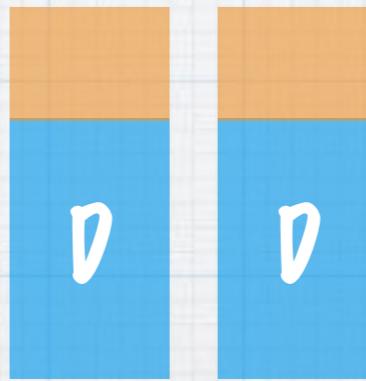
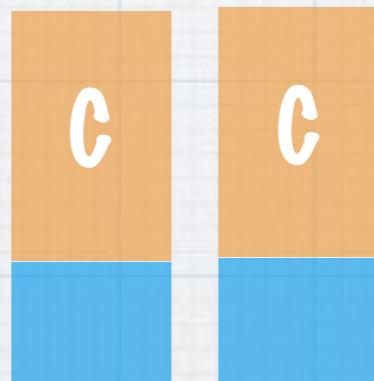
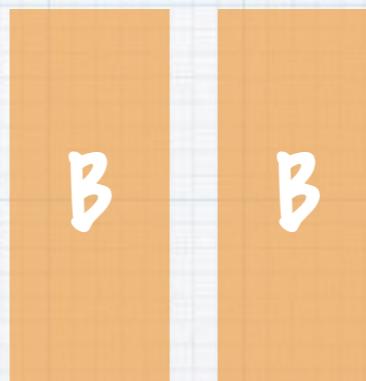
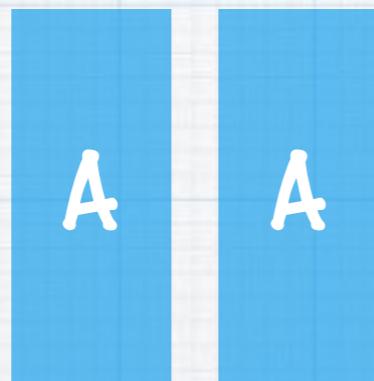
```
bcbio_nextgen.py -w template ...  
bcbio_nextgen.py config.yaml ...
```

# bcbio\_setup\_genome.py

## Current genomes

```
index mm10      /groups/bcbio/bcbio/genomes/Mmusculus/mm10/seq/mm10.fa
index Zv9       /groups/bcbio/bcbio/genomes/Drerio/Zv9/seq/Zv9.fa
index GRCh37    /groups/bcbio/bcbio/genomes/Hsapiens/GRCh37/seq/GRCh37.fa
index WBcel235   /groups/bcbio/bcbio/genomes/Celegans/WBcel235/seq/WBcel235.fa
index hg19      /groups/bcbio/bcbio/genomes/Hsapiens/hg19/seq/hg19.fa
index hg19-mt   /groups/bcbio/bcbio/genomes/Hsapiens/hg19-mt/seq/hg19-mt.fa
index mm9       /groups/bcbio/bcbio/genomes/Mmusculus/mm9/seq/mm9.fa
index greenberg-mm9 /groups/bcbio/bcbio/genomes/Mmusculus/greenberg-mm9/seq/greenberg-mm9.fa
index MB2409    /groups/bcbio/bcbio/genomes/Ecoli/MB2409/seq/MB2409.fa
index MG1655    /groups/bcbio/bcbio/genomes/Ecoli/MG1655/seq/MG1655.fa
index MB0009    /groups/bcbio/bcbio/genomes/Ecoli/MB0009/seq/MB0009.fa
index NC_912.3   /groups/bcbio/bcbio/genomes/Ecoli/NC_000913.3/seq/NC_000913.3.fa
index MG1655_v2  /groups/bcbio/bcbio/genomes/Ecoli/MG1655_v2/seq/MG1655_v2.fa
index MG1655_virus /groups/bcbio/bcbio/genomes/Ecoli/MG1655_virus/seq/MG1655_virus.fa
index MB2455    /groups/bcbio/bcbio/genomes/Ecoli/MB2455/seq/MB2455.fa
index k12       /groups/bcbio/bcbio/genomes/Ecoli/k12/seq/k12.fa
index UMD3.1    /groups/bcbio/bcbio/genomes/Btaurus/UMD3.1/seq/UMD3.1.fa
index ASM294v2   /groups/bcbio/bcbio/genomes/spombe/ASM294v2/seq/ASM294v2.fa
index ASM284v2.25 /groups/bcbio/bcbio/genomes/Spombe/ASM284v2.25/seq/ASM284v2.25.fa
index DQ900900.1  /groups/bcbio/bcbio/genomes/haD37/DQ900900.1/seq/DQ900900.1.fa
index hg19-ercc   /groups/bcbio/bcbio/genomes/Hsapiens/hg19-ercc/seq/hg19-ercc.fa
index FGSC_A4    /groups/bcbio/bcbio/genomes/Anidulans/FGSC_A4/seq/FGSC_A4.fa
index loxAfr3    /groups/bcbio/bcbio/genomes/Lafricana/loxAfr3/seq/loxAfr3.fa
index BDGP6     /groups/bcbio/bcbio/genomes/Dmelanogaster/BDGP6/seq/BDGP6.fa
index rn6       /groups/bcbio/bcbio/genomes/Rnorvegicus/rn6/seq/rn6.fa
```

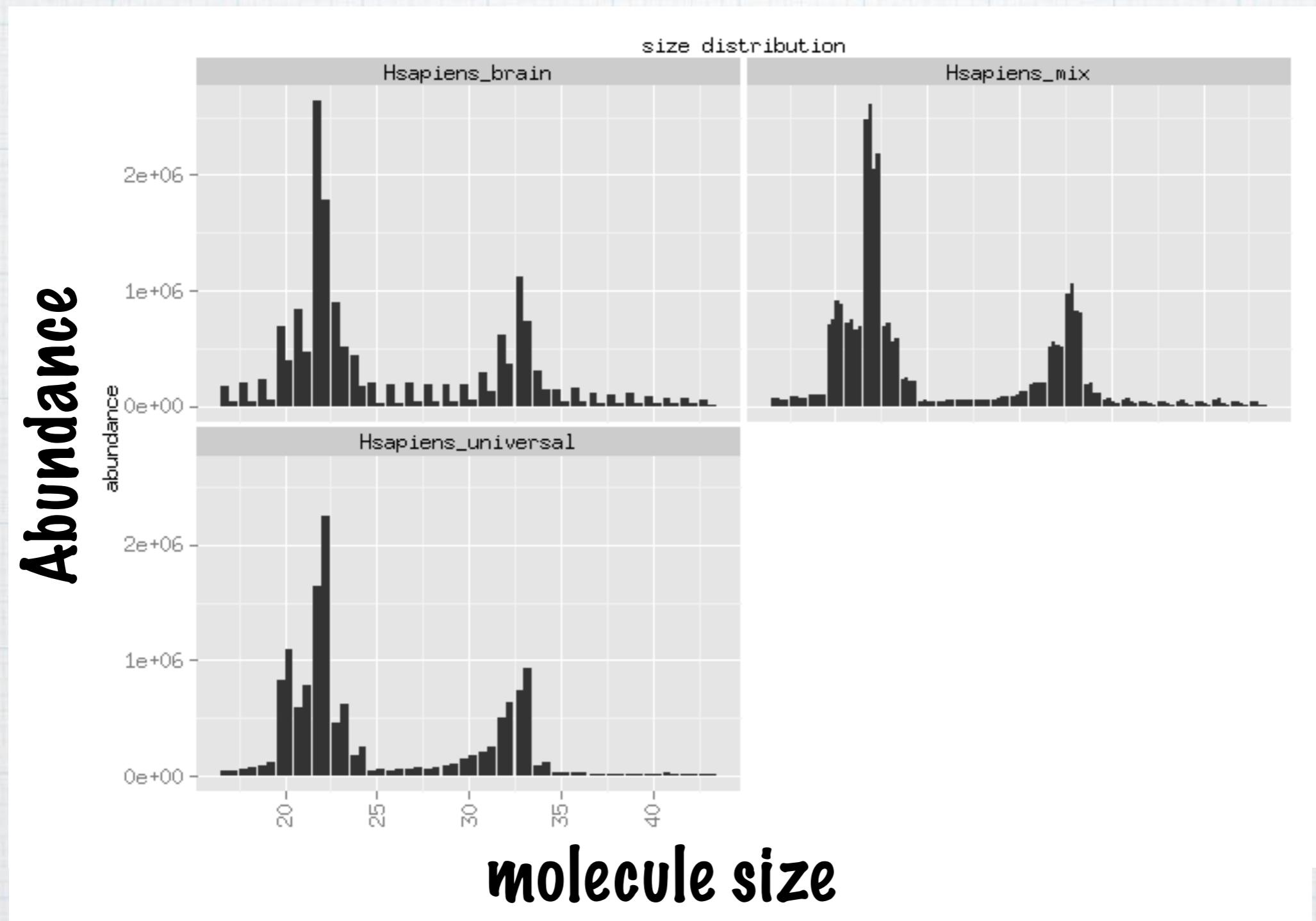
# Quality Control samples



For each molecule:

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

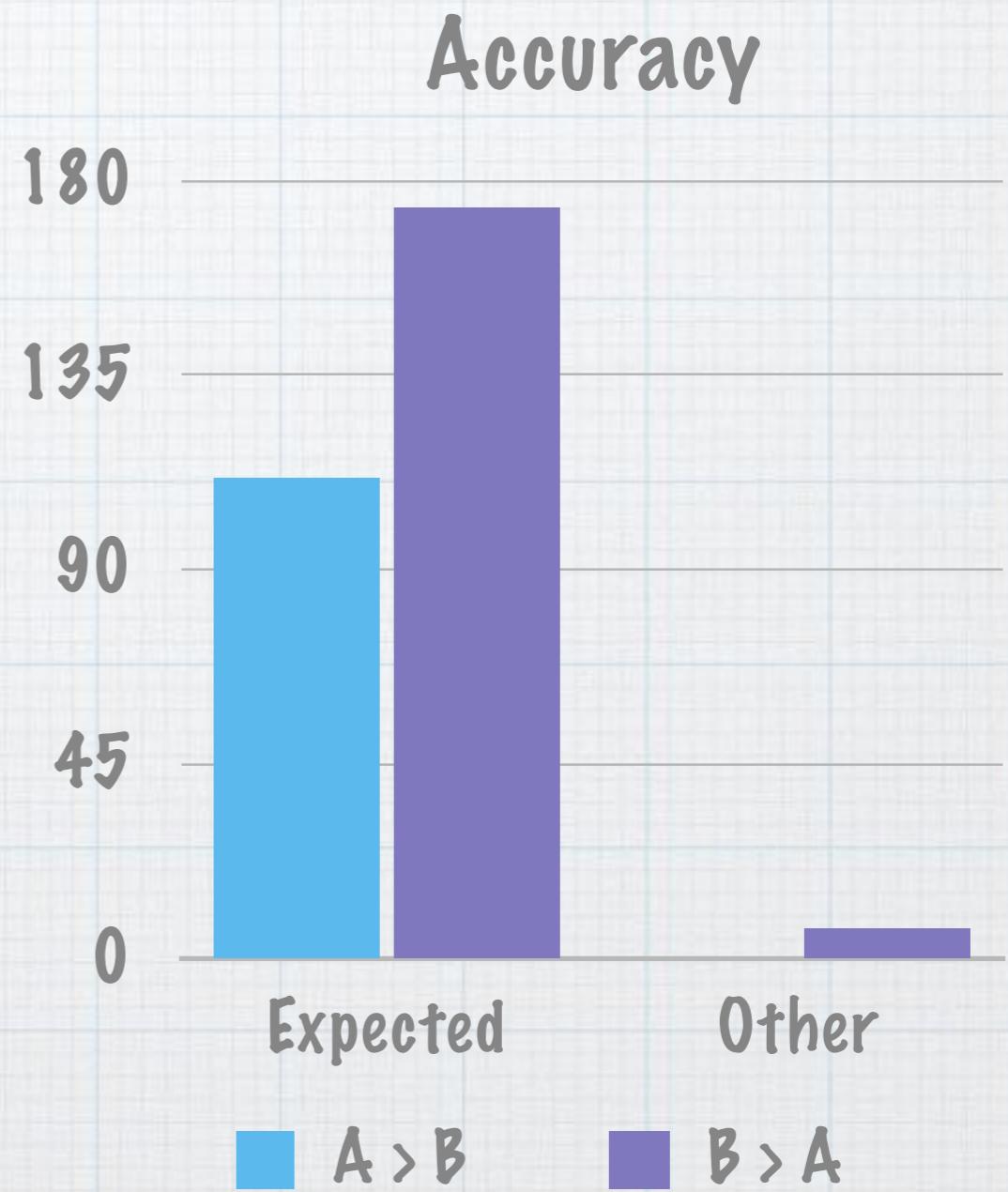
# Good samples



# miRNA quantification

miRNAs > 5 counts in average

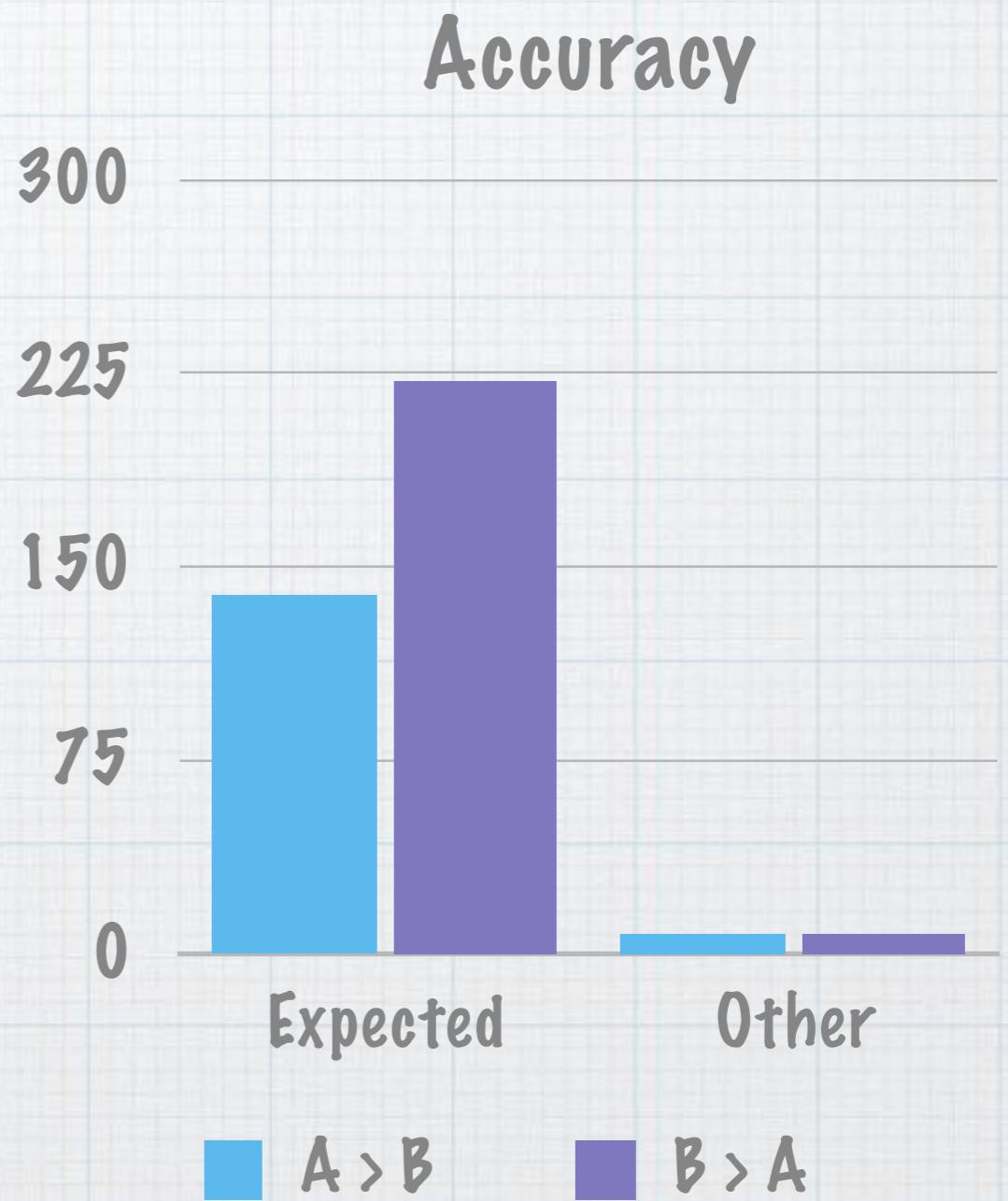
upper quantile normalization



# clusters quantification

expression > 5 counts in average

upper quantile normalization



# Positive controls

A

A

A

A

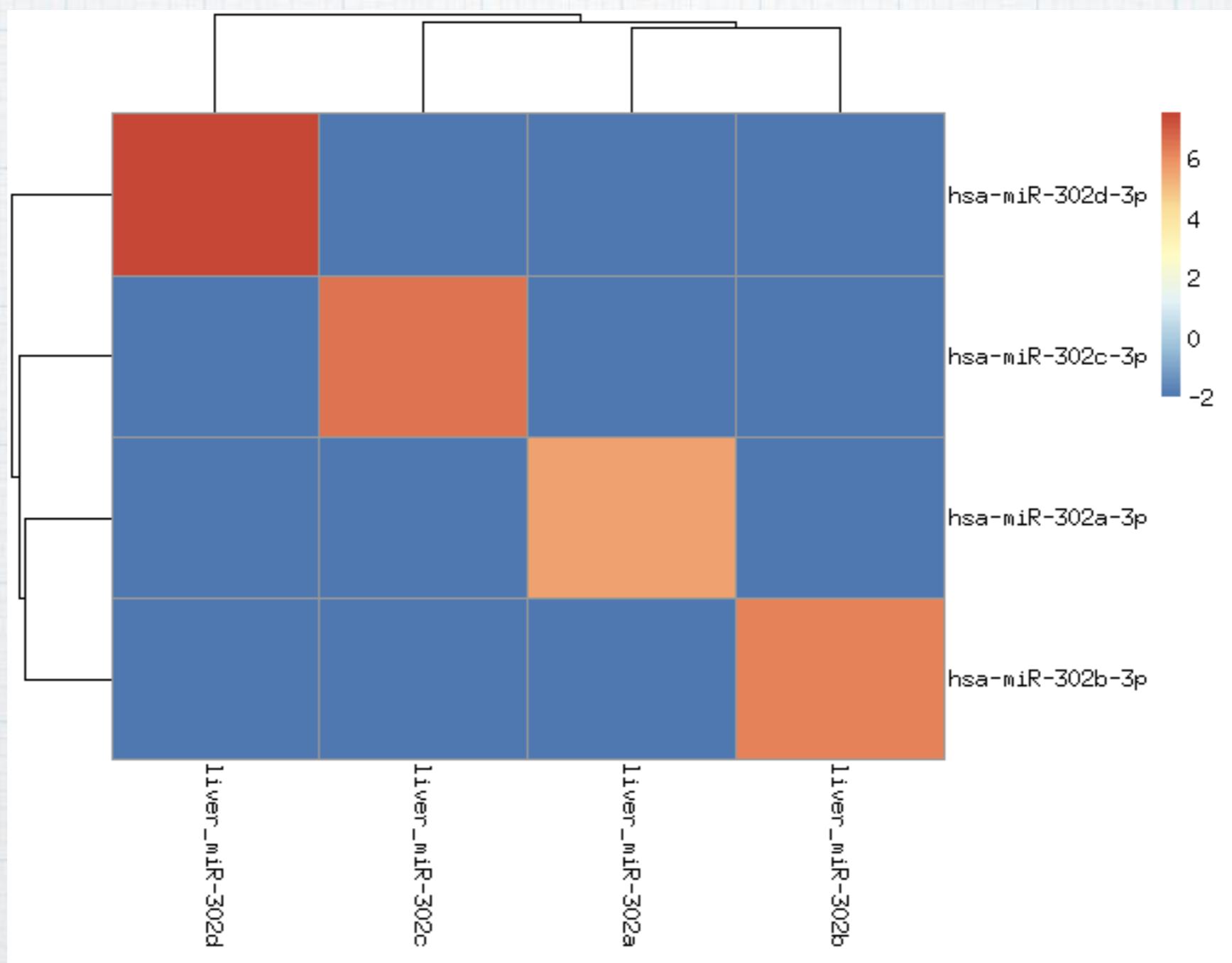
*miR-302a*

*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.

# open project for small RNA annotation and analysis

## mirTOP



miRNA transcriptome open project

http://mirtop.github.io

<http://mirtop.github.io>

Repositories

People 3

Teams 1

Settings

Filters ▾

Find a repository...

New repository

### incubator

★ 1 ⚡ 1

Where all ideas and discussions happen to lead to new repositories

Updated 3 days ago

mirtop

## standard formats naming rules

## best-practices

Python ★ 0 ⚡ 0

command lines tool to annotate miRNAs with a standard mirna/isomir naming

Updated 3 days ago

[mirtop.github.io](http://mirtop.github.io)

miRNAs, tRNAs ...

CSS ★ 0 ⚡ 0

project for small RNA standard annotations

Updated on Mar 29

# Cambridge Women BioInformatics Meetup

[Home](#)[Members](#)[Sponsors](#)[Photos](#)[Pages](#)[Discussions](#)[More](#)[Groups](#)

Cambridge, MA

Founded Mar 27, 2015

[About us...](#)[+ Invite friends](#)

minians 286

Group reviews 4

Upcoming Meetups 1

Past Meetups 15

Our calendar

Help support your Meetup

[Chip in](#)

## Welcome!

+ Schedule a new Meetup

[Upcoming \(1\)](#) Past Draft (1) Calendar

### Rshiny app to browse RNAseq data

Harvard University: Countway Library

10 Shattuck St, Boston, Ma ([map](#))



Hi, Join us in the last meeting of the year to create an easy app to browse RNAseq data. The goal is to have a small working code to visualize the expression of selected...

[Learn more](#)

Hosted by: [Lorena Pantano](#) (Organizer)

Tue Dec 13

5:45 PM

[I'm going](#)

16 going

4 spots left

0 comments

# 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.
- \* Special thanks to the authors of those papers to make data available.