

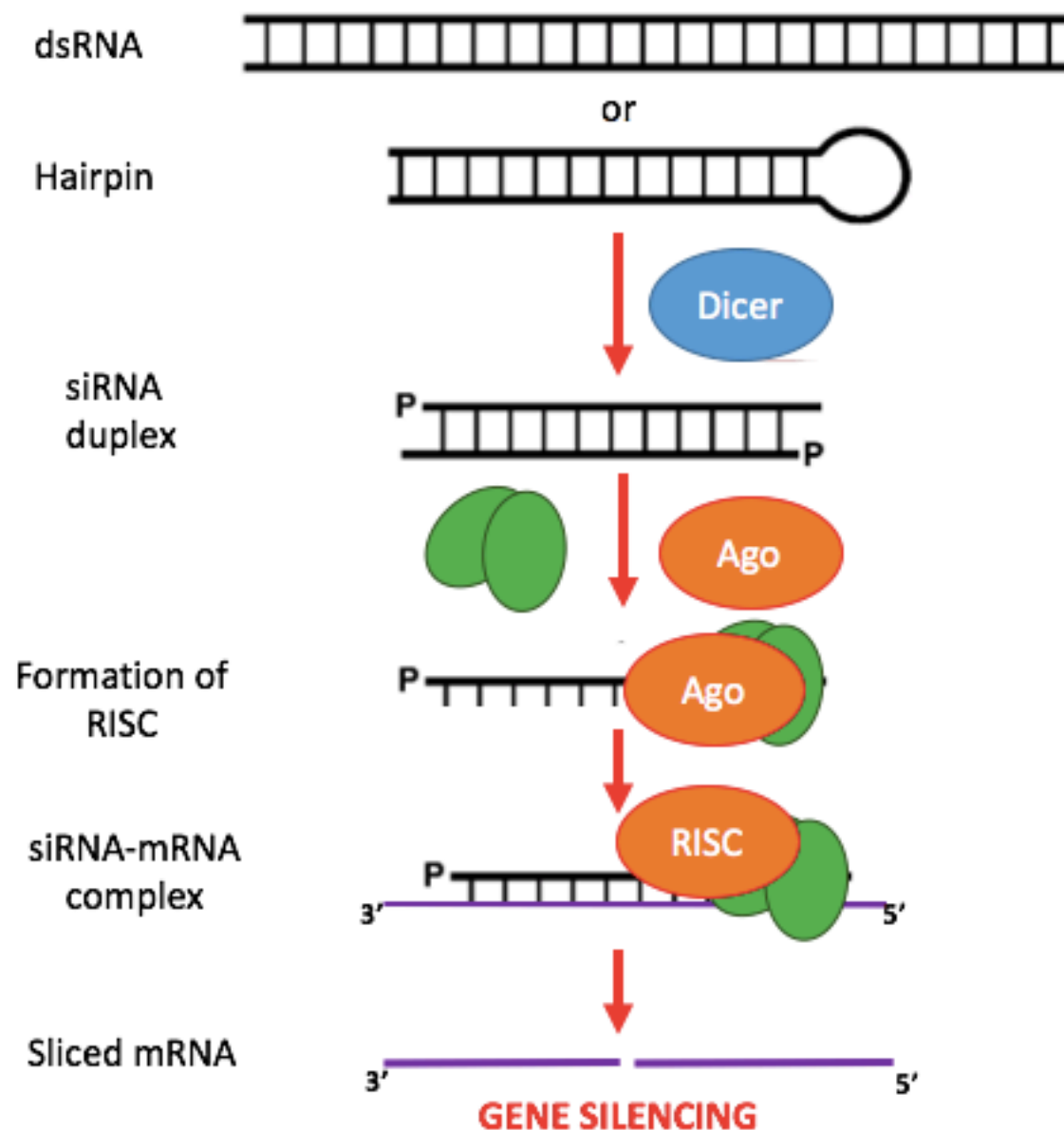
# Characterization of the small RNA transcriptome

Lorena Pantano  
@lopantano lpantano@hsph.harvard.edu  
Harvard TH Chan School of Public Health

<https://goo.gl/uZWG0E>

2017-06-22

# small interference RNA

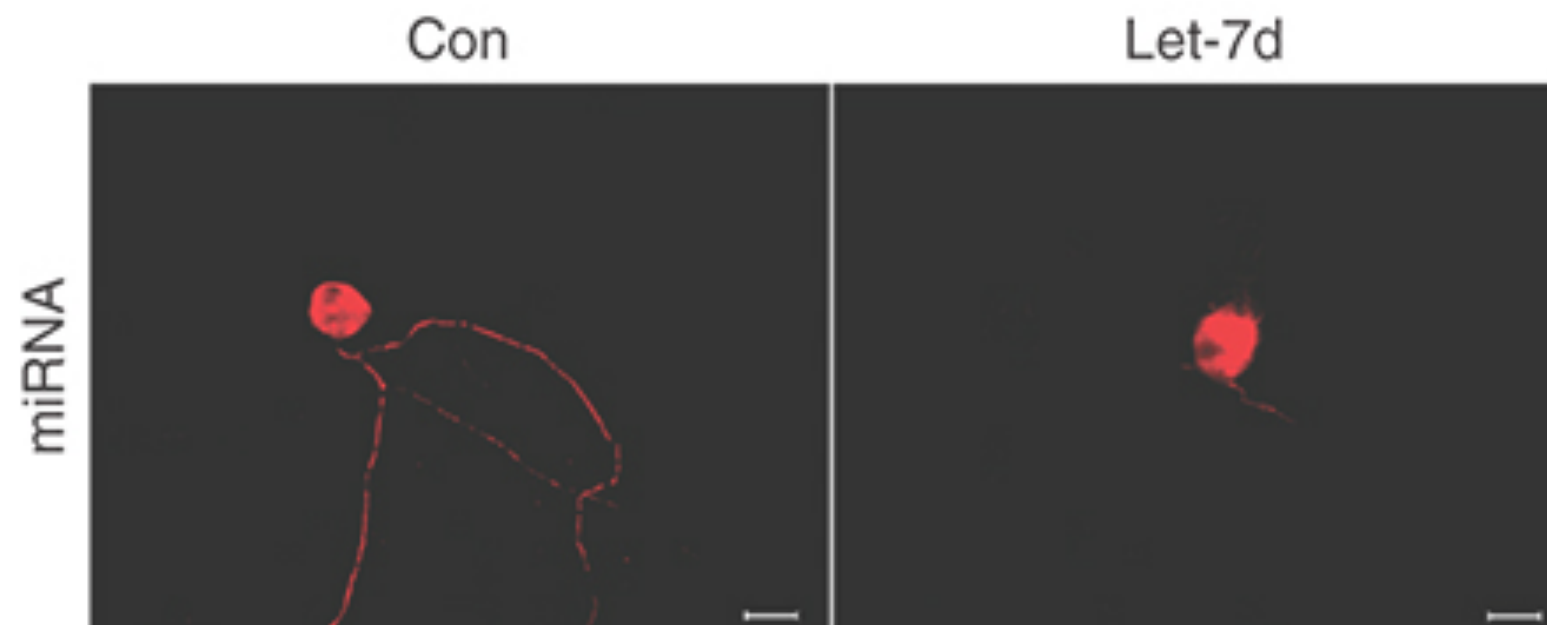


- miRNA (18-25nt)
- endo-siRNA (20-25nt)
- piRNA (25-33nt)

# 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

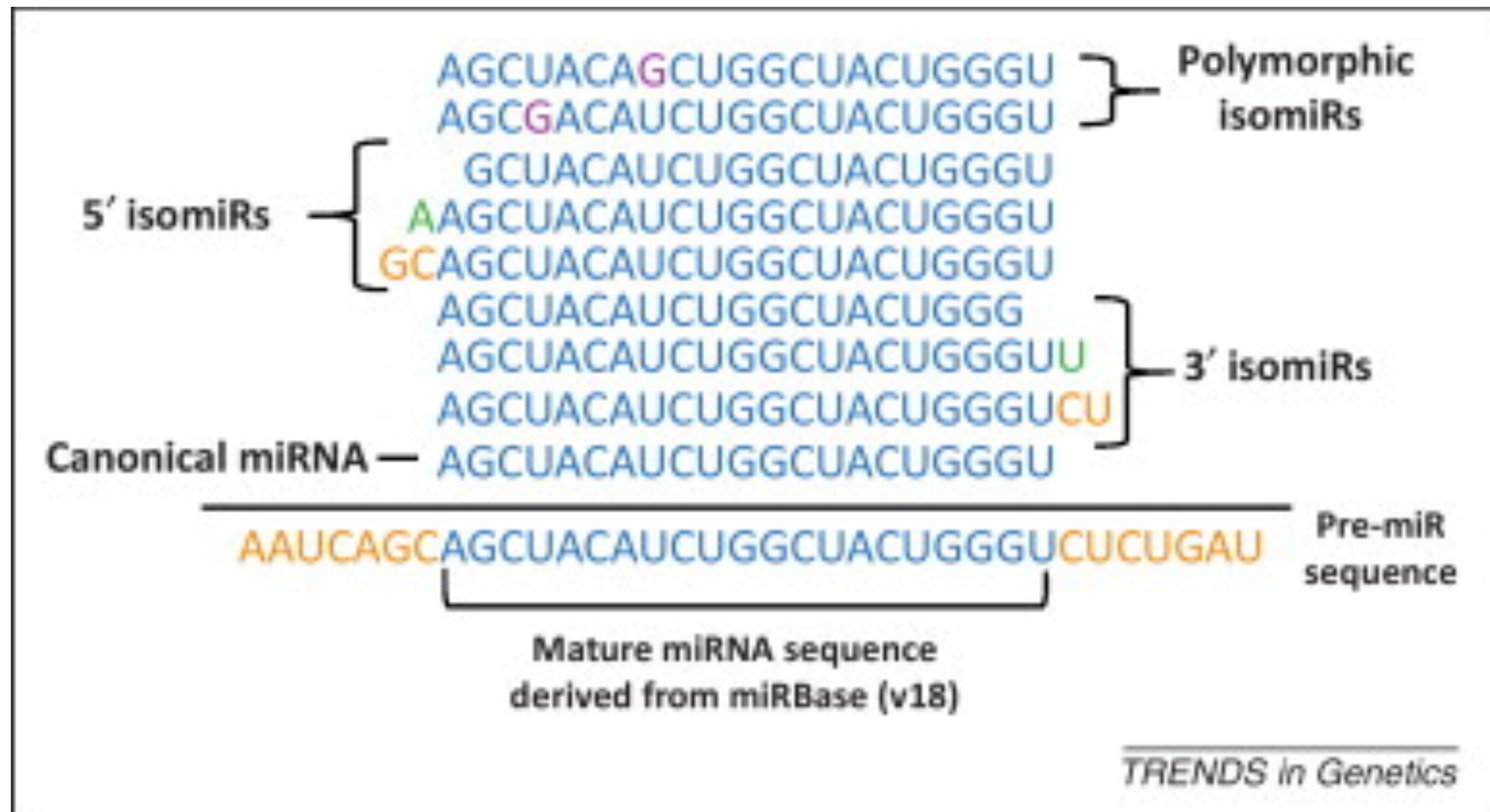
---

<u>hsa-miR-24-1-5p</u>	<u>hsa-miR-24-3p</u>
..... <u>GGUGCCUACUGAGCUGAUAUC</u> .....	
..... <u>GUGCCUACUGAGCUGAUAUCAGU</u> .....	
..... <u>GUGCCUACUGAGCUGAUAUCAG</u> .....	
..... <u>GUGCCUACUGAGCUGAUA</u> .....	
..... <u>UGCCUACUGAGCUGAUAUCA</u> .....	
..... <u>UGCCUACUGAGCUGAUAUCAGU</u> .....	
..... <u>UGCCUACUGAGCUGAUAUC</u> .....	
..... <u>UGCCUACUGAGCUGAUA</u> .....	
..... <u>CCUACUGAGCUGAUAUCA</u> .....	
..... <u>CCUACUGAGCUGAUAUCAGU</u> .....	
..... <u>CUACUGAGCUGAUAUCA</u> .....	
..... <u>CUACUGAGCUGAUAUC</u> .....	

CUCCGGUGCCUACUGAGCUGAUAUCAGUUCUCAUUUUACACACUGGCUCAGUUCAGCAGGAACAGGAG  
(((((((((.....)))))))).))))))((-26.32)

precursor

# types of isomiRs



# isomiRs

## Search results

Items: 1 to 20 of 146

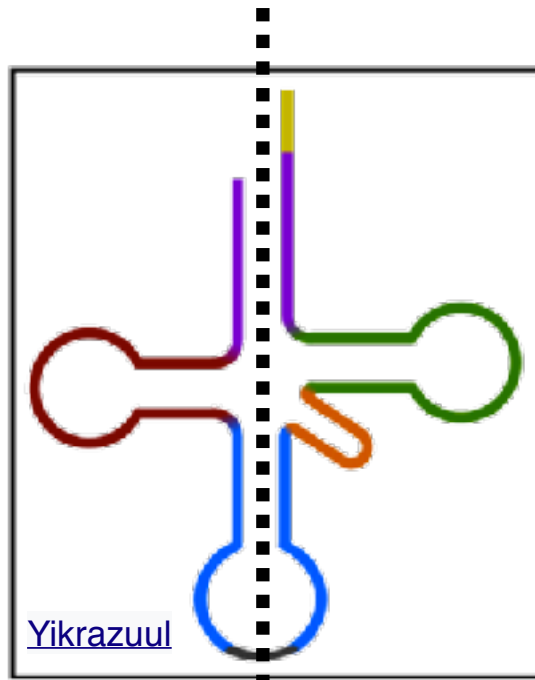
<< First < Prev Page 1 of 8 Next > Last >>

- ☐ [Chronic low-dose exposure to a mixture of environmental endocrine disruptors induces microRNAs/isomiRs deregulation in mouse concomitant with intratesticular estradiol reduction.](#)  
1. Buñay J, Larriba E, Moreno RD, Del Mazo J.  
Sci Rep. 2017 Jun 13;7(1):3373. doi: 10.1038/s41598-017-02752-7.  
PMID: 28611354  
[Similar articles](#)
  
- ☐ [Expression profile of Epstein-Barr virus and human adenovirus small RNAs in tonsillar B and T lymphocytes.](#)  
2. Assadian F, Kamel W, Laurell G, Svensson C, Punga T, Akusjärvi G.  
PLoS One. 2017 May 25;12(5):e0177275. doi: 10.1371/journal.pone.0177275. eCollection 2017.  
PMID: 28542273    **Free PMC Article**  
[Similar articles](#)
  
- ☐ [isomiR2Function: An Integrated Workflow for Identifying MicroRNA Variants in Plants.](#)  
3. Yang K, Sablok G, Qiao G, Nie Q, Wen X.  
Front Plant Sci. 2017 Mar 21;8:322. doi: 10.3389/fpls.2017.00322. eCollection 2017.  
PMID: 28377776    **Free PMC Article**  
[Similar articles](#)
  
- ☐ [3' Uridylation controls mature microRNA turnover during CD4 T-cell activation.](#)  
4. Gutiérrez-Vázquez C, Enright AJ, Rodríguez-Galán A, Pérez-García A, Collier P, Jones MR, Benes V, Mizgerd JP, Mittelbrunn M, Ramiro AR, Sánchez-Madrid F.  
RNA. 2017 Jun;23(6):882-891. doi: 10.1261/rna.060095.116. Epub 2017 Mar 28.  
PMID: 28351886  
[Similar articles](#)

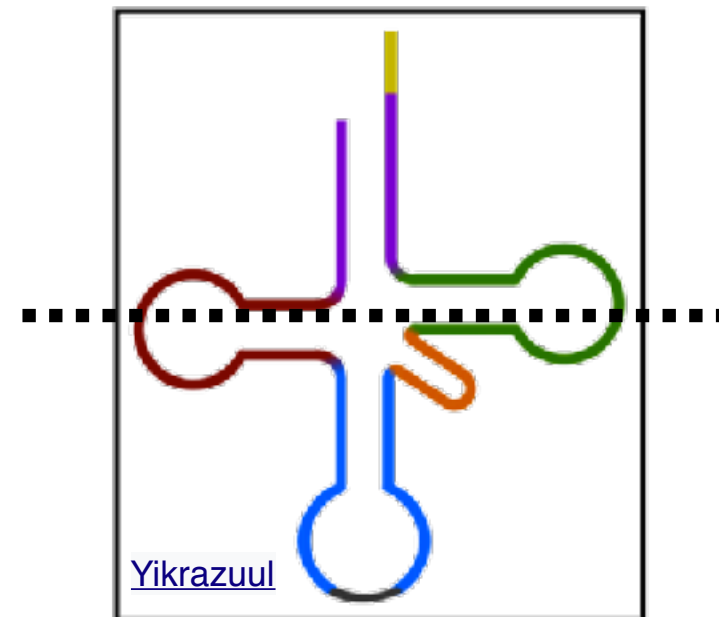


# tRNA derived fragments

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



tRNA-halves (30-33nt)



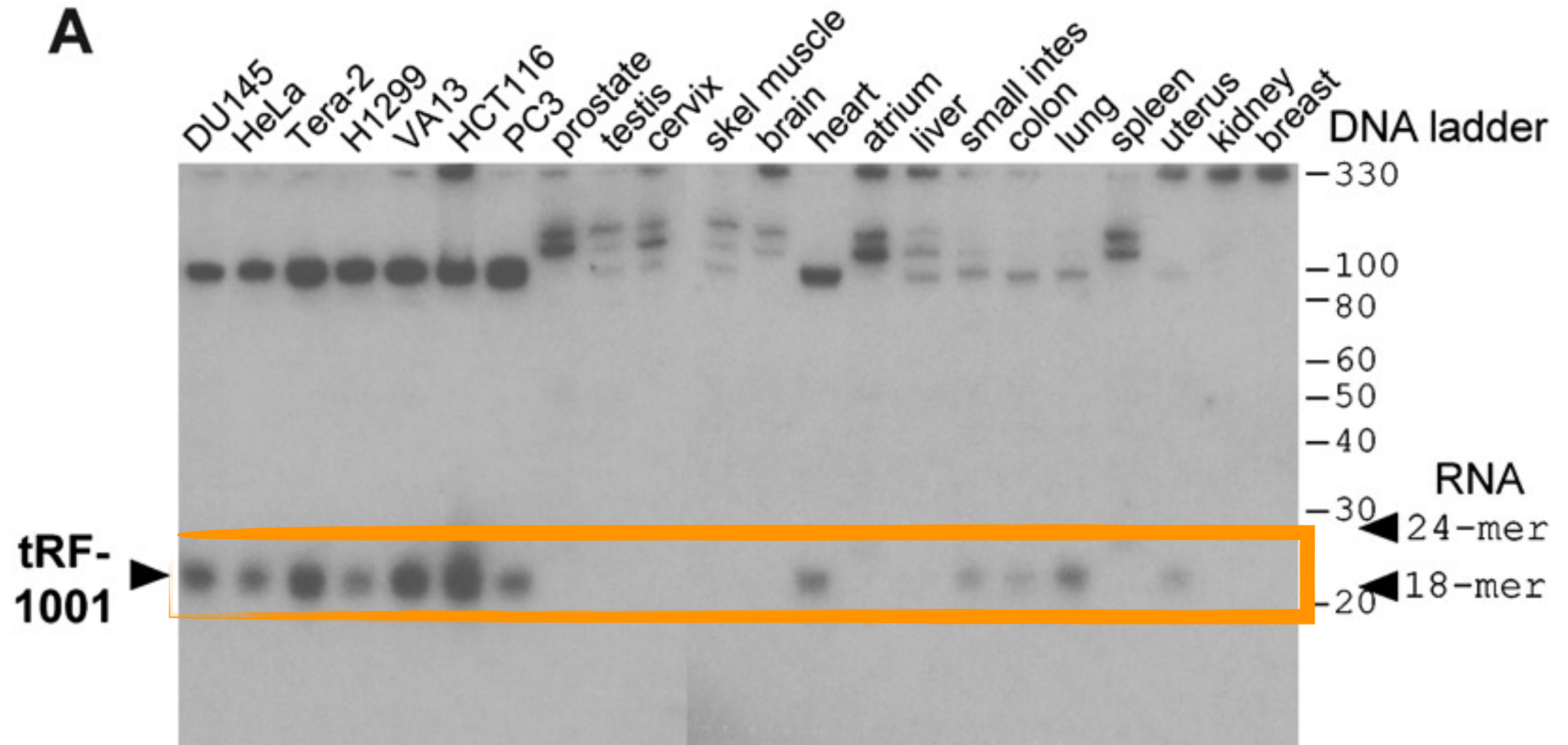
5'-tRNAs (18-22)    3'-tRNAs (18-22)

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



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



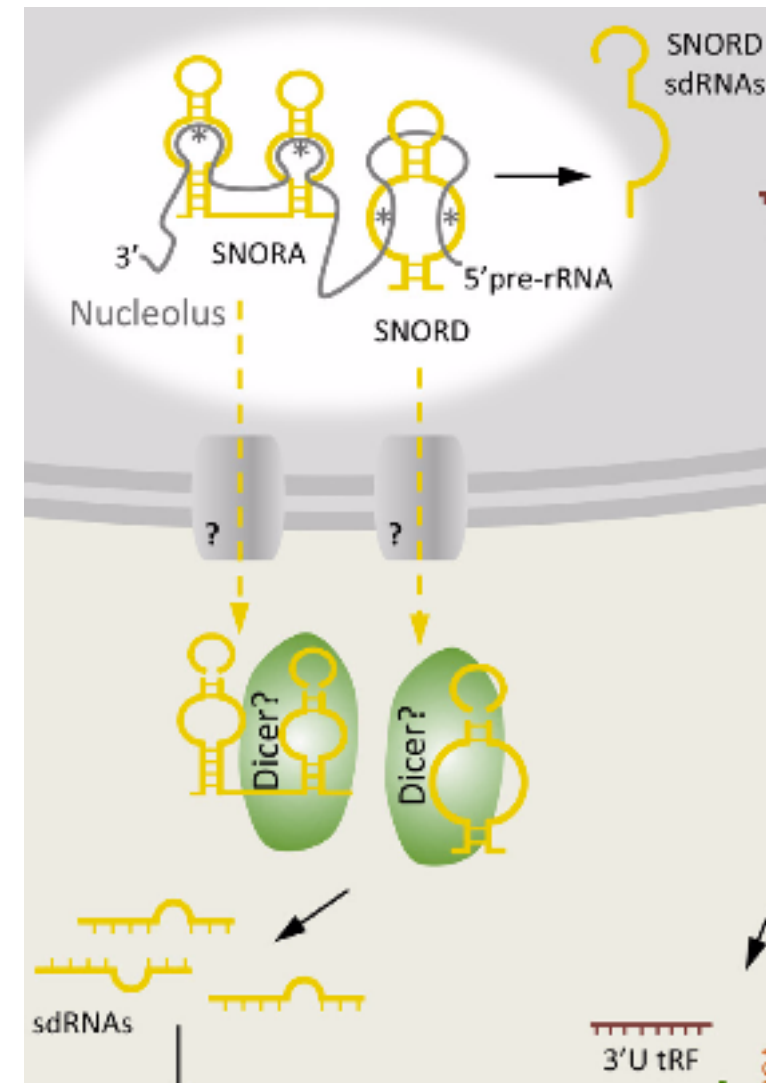
# sncRNA fragments

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)

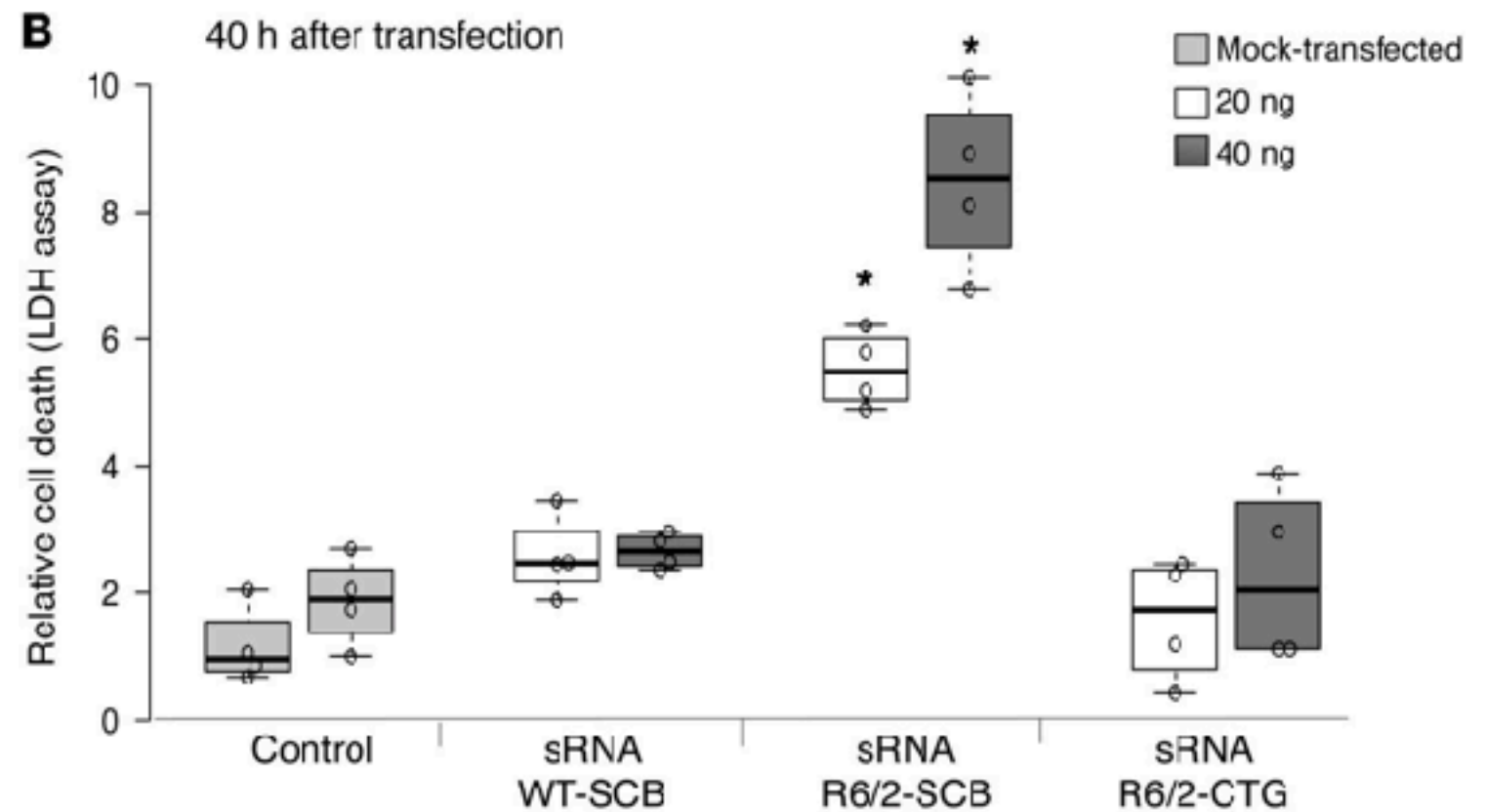
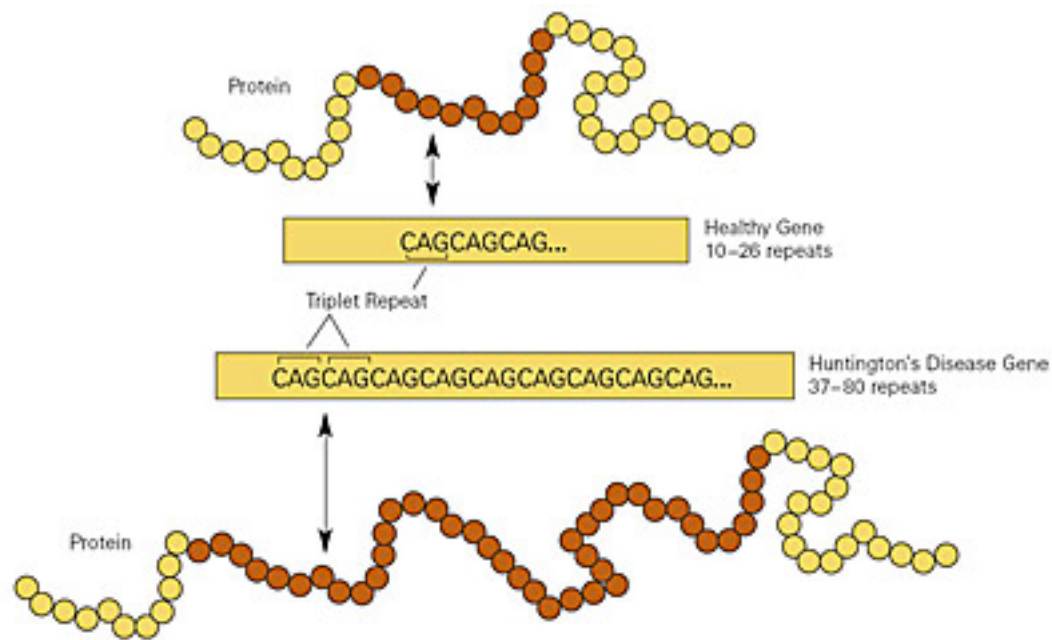
SNORD44/78 up-regulated in **prostate cancer**.



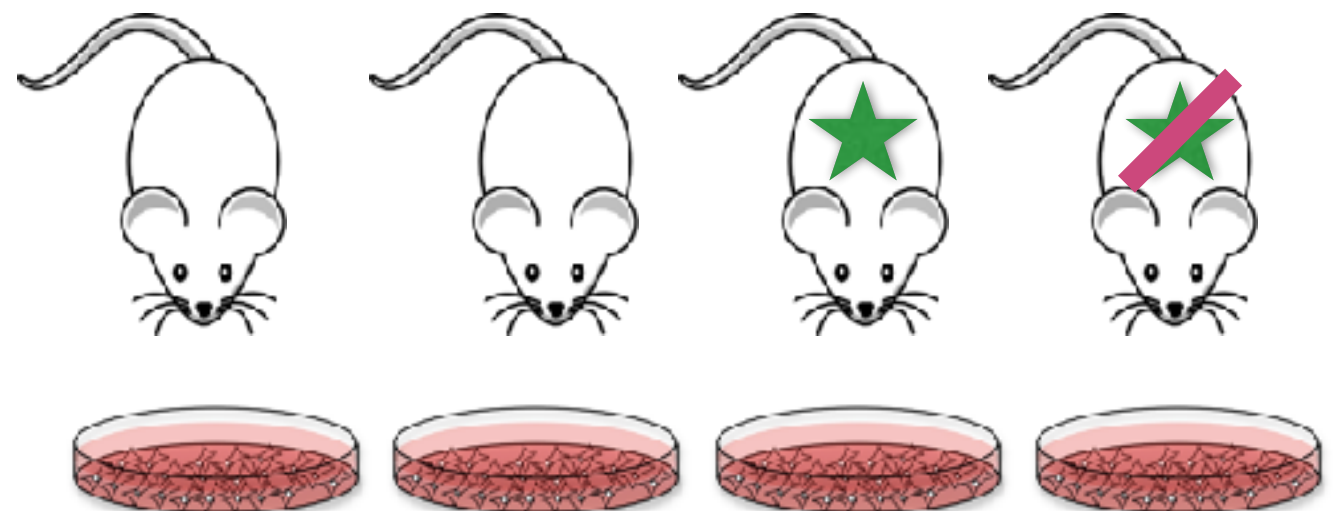
snoRNA-derived RNAs (sdRNAs)

# small RNA

## Huntington disease therapy



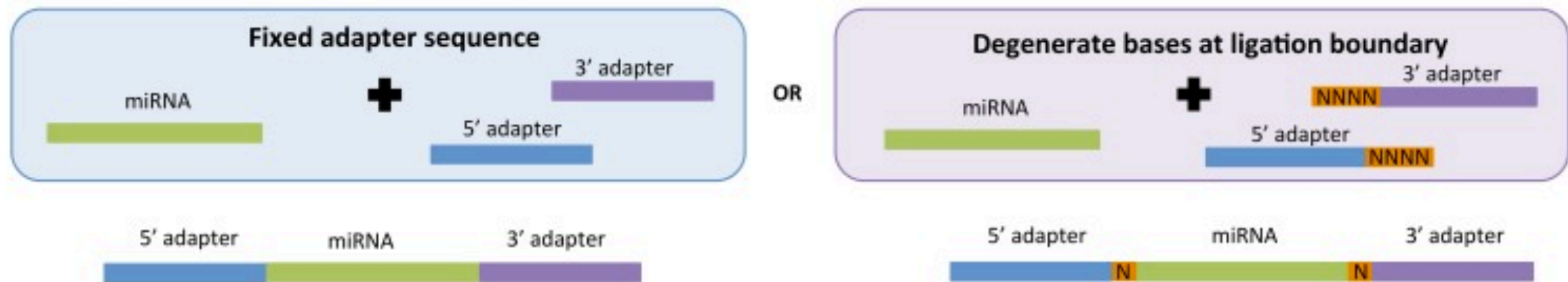
synthetic small RNA  
CTGCTGCTGCTGCTGCTGCT



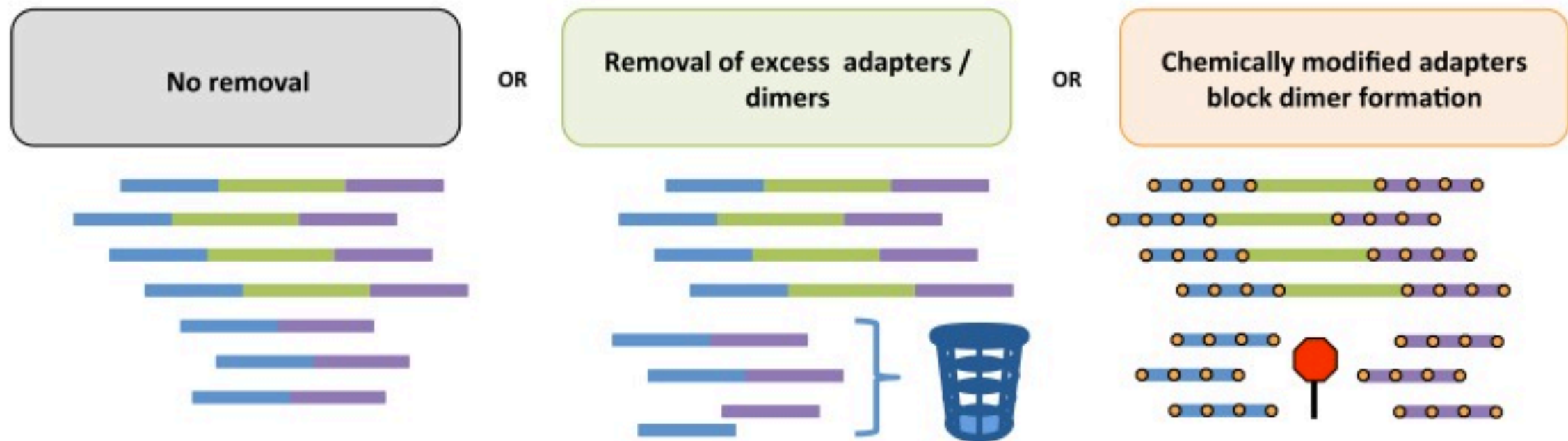
# Protocols

## Critical differences in small RNA library preparation protocols

### Issue 1: Adapter ligation introduces bias

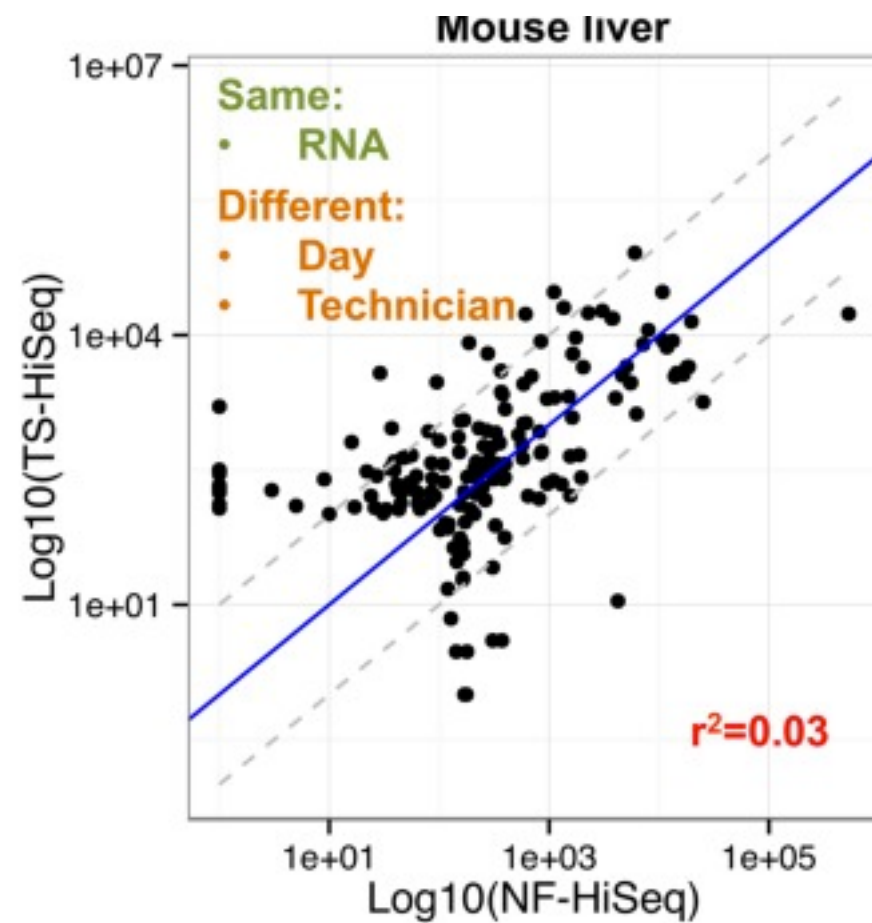


### Issue 2: Adapter dimers compete with small RNAs, reducing effective sequencing depth

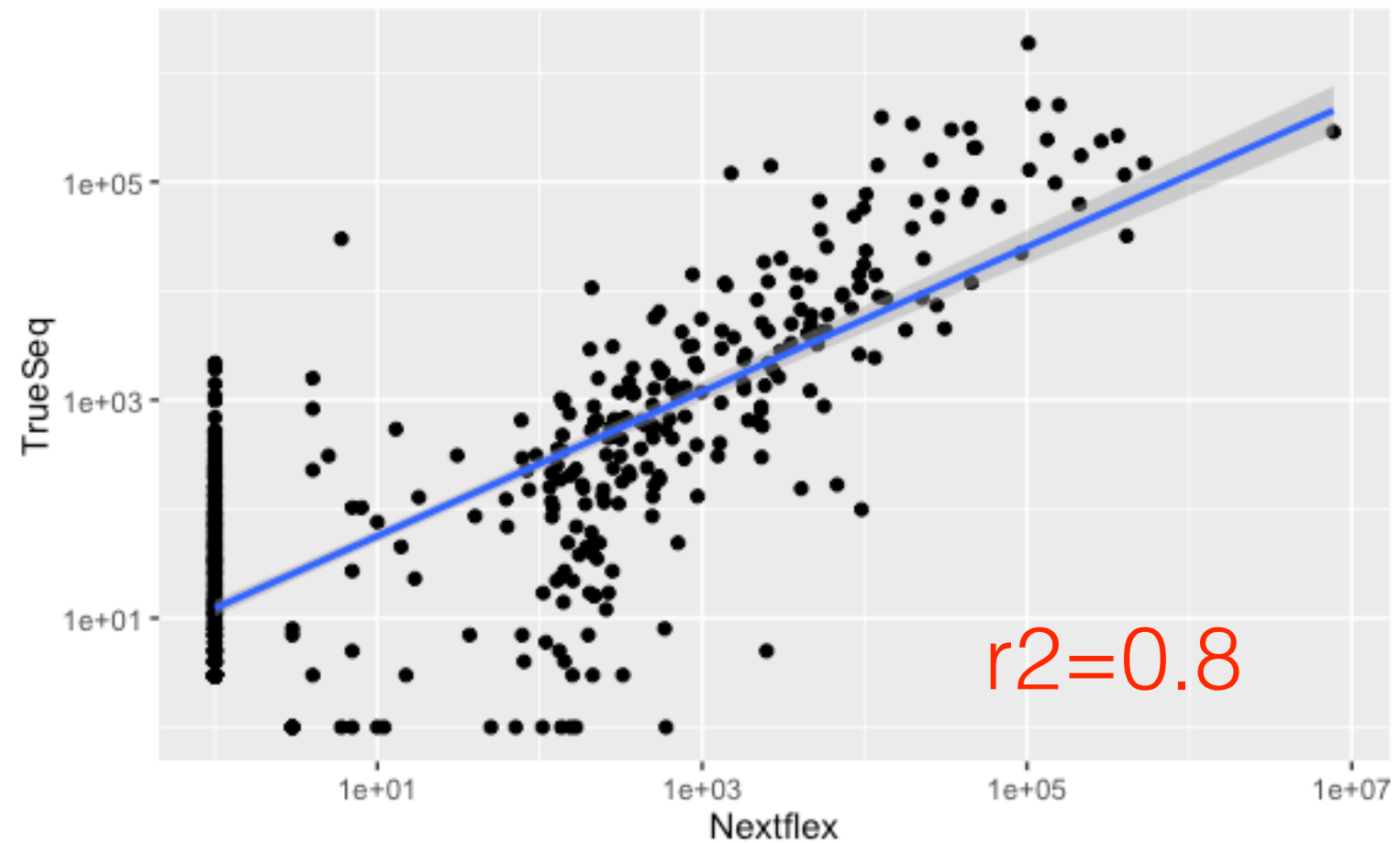


# Protocol correlation

Paper figure

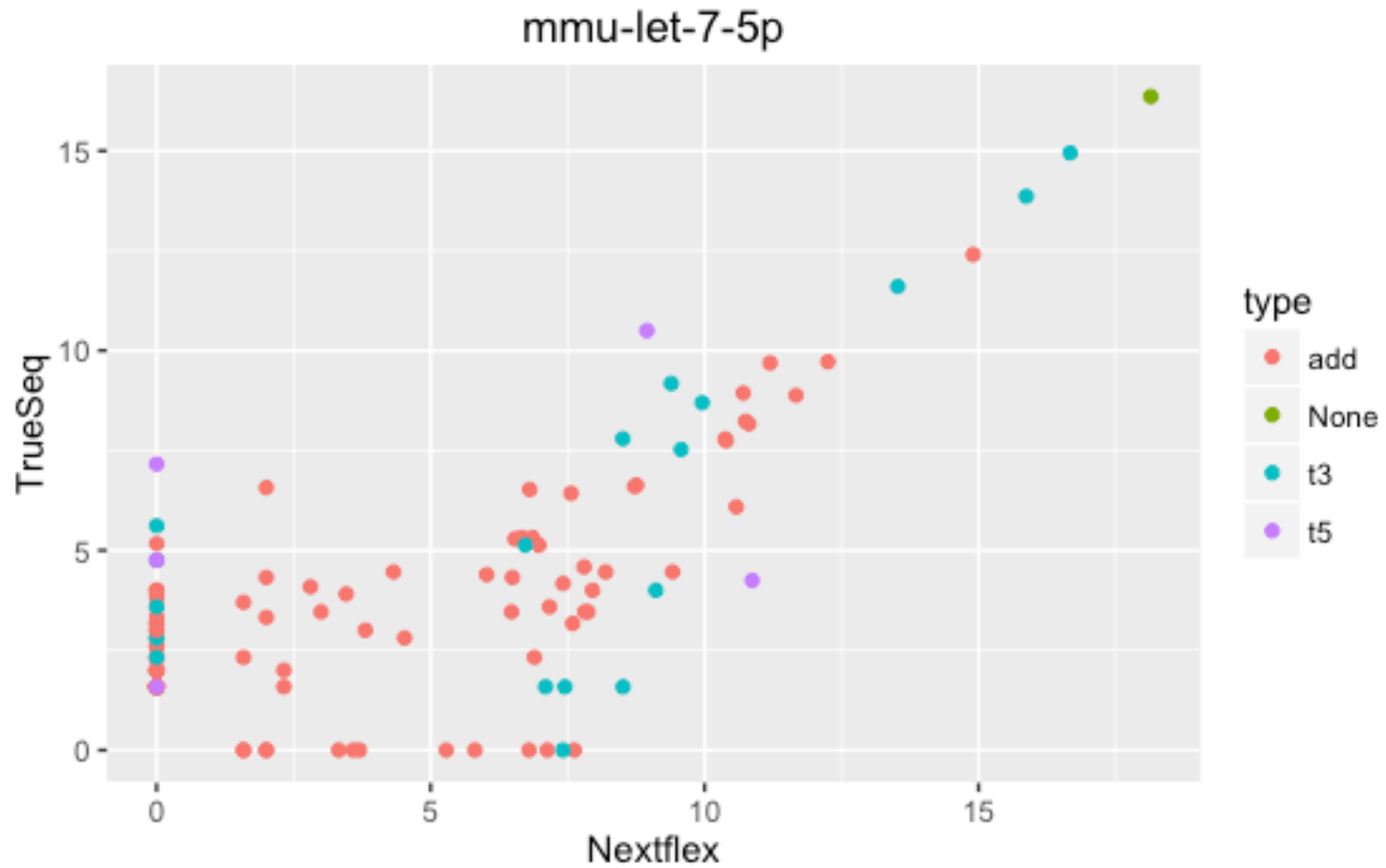


bcbio pipeline



# let7-a-5p miRNA

---



# Caveats

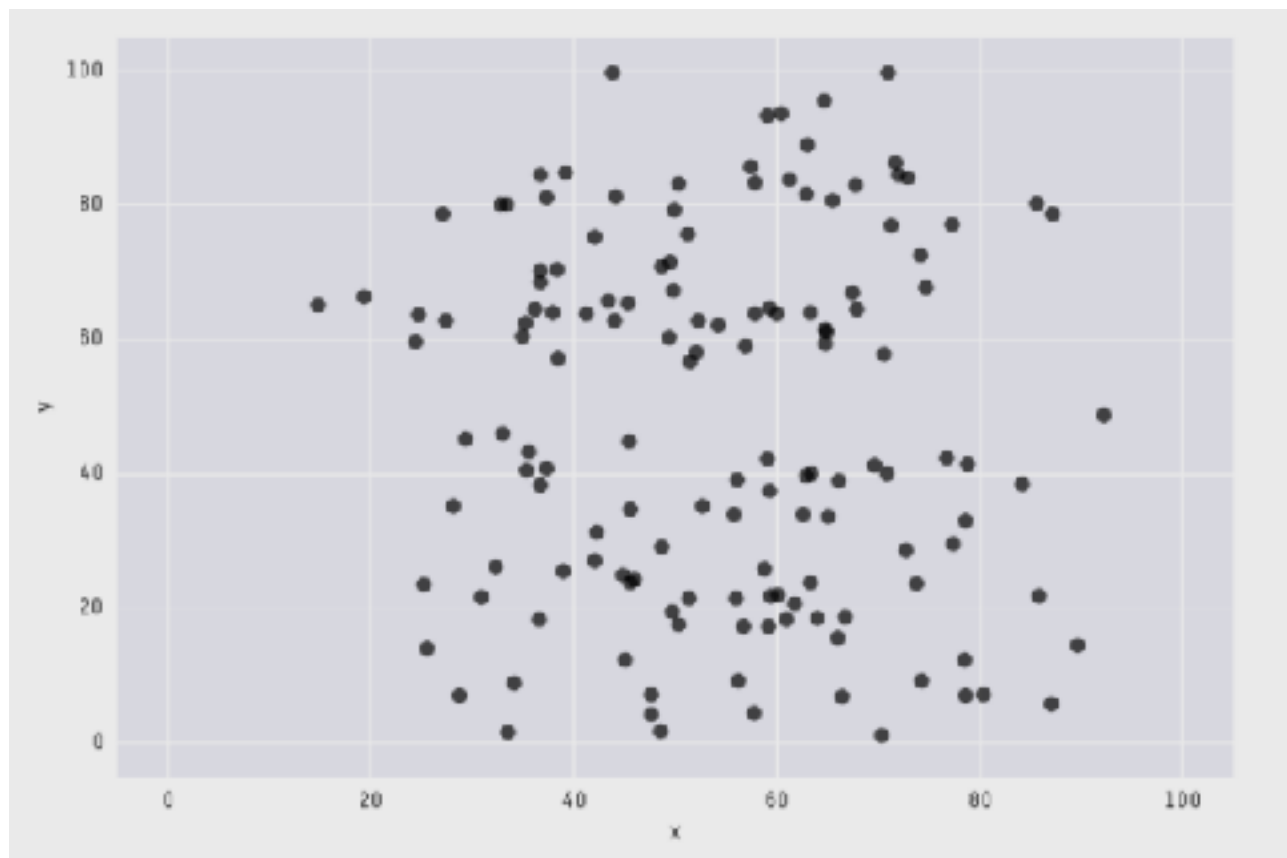
---

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

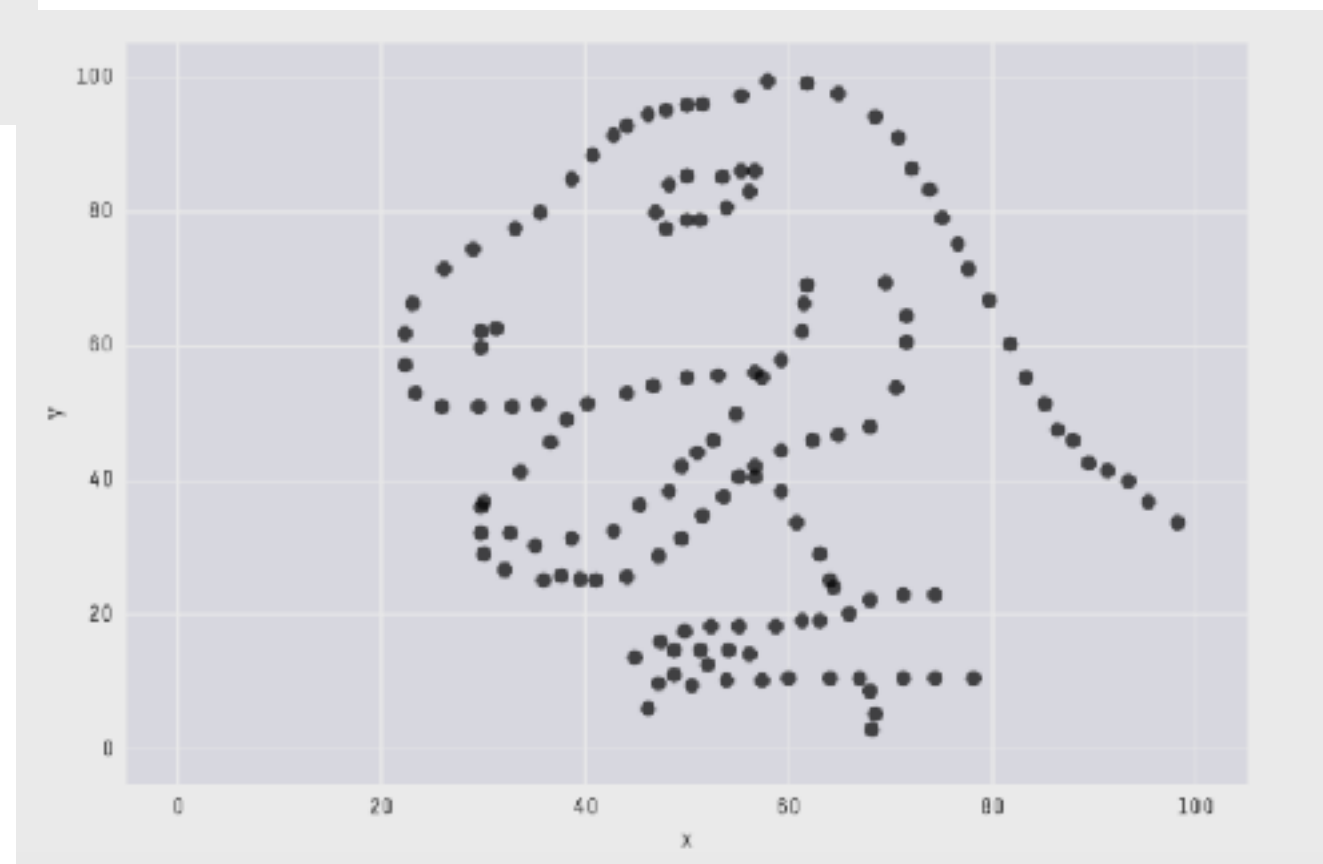
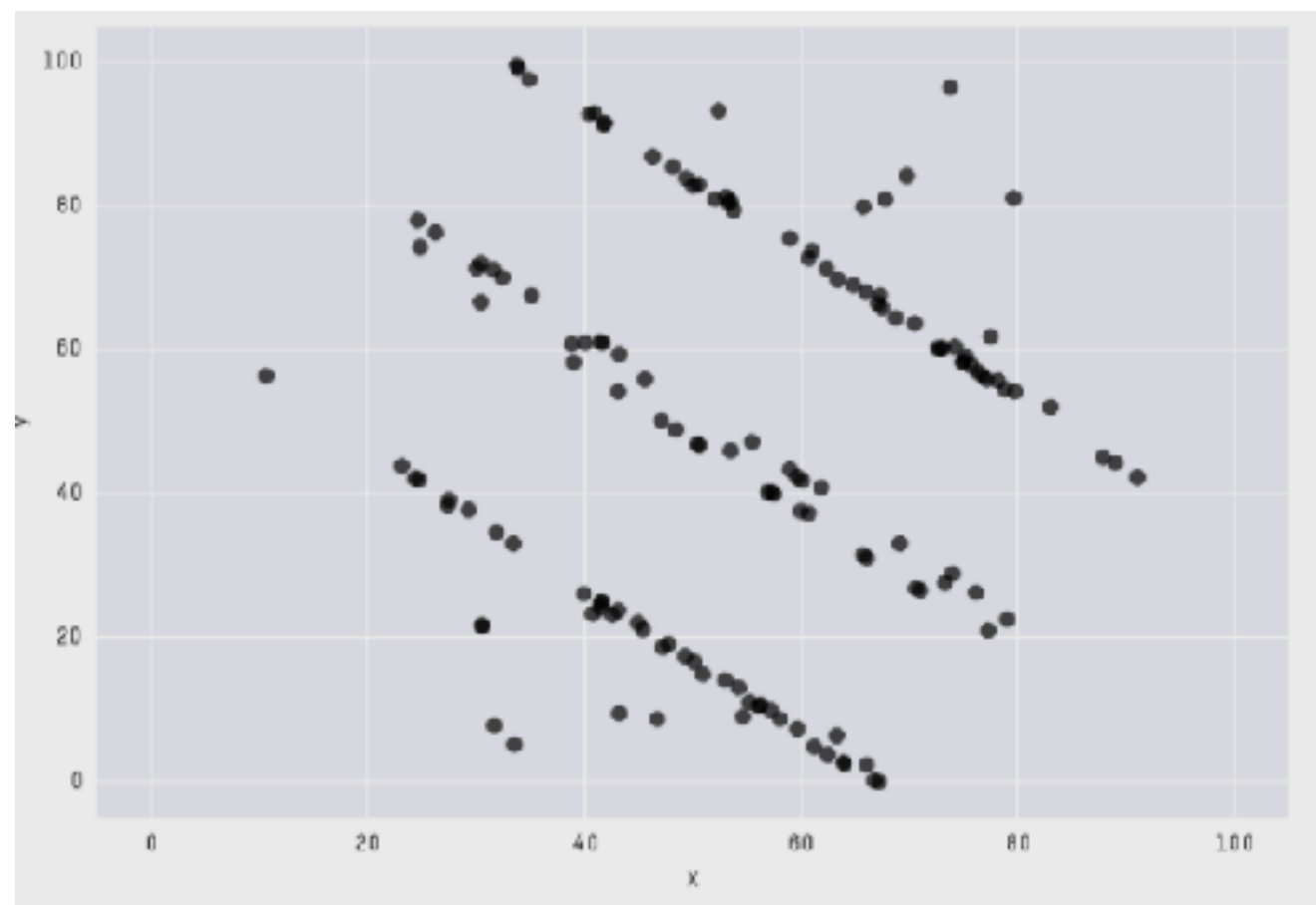


# Same Stats, Different graphs

---



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



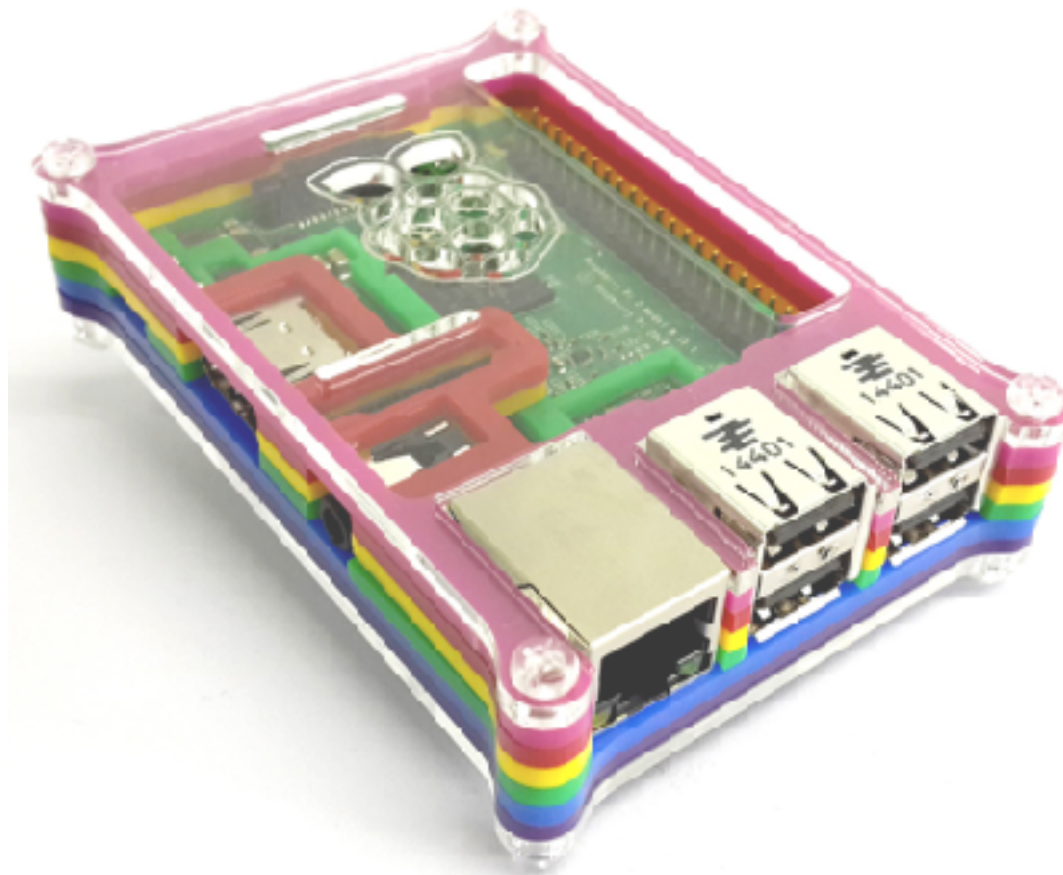
# challenges

---

- 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

---



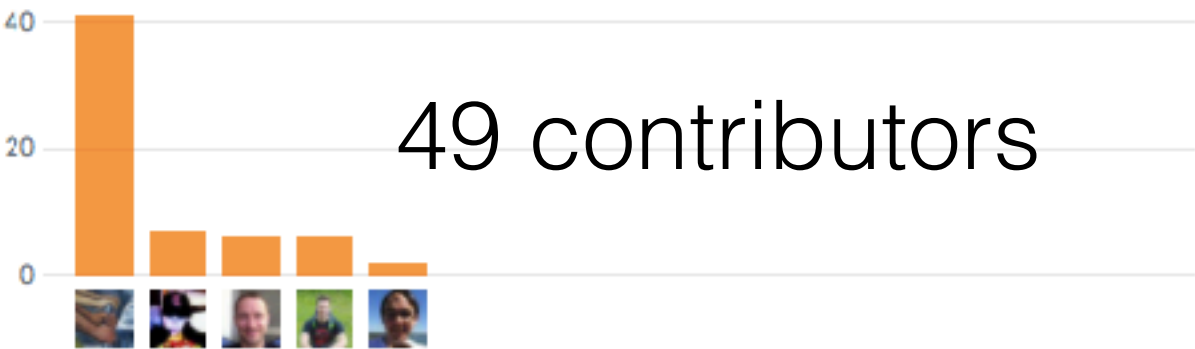
Variant calling, RNA-seq, small RNA-seq  
over 200 peer reviewed tools **BIOCONDA**<sup>®</sup>

May 20, 2017 – June 20, 2017

Period: 1 month ▾

Overview			
<div><div></div></div> <div>6 Active Pull Requests</div>		<div><div></div></div> <div>39 Active Issues</div>	
<div><div></div>6</div> <div>Merged Pull Requests</div>	<div><div></div>0</div> <div>Proposed Pull Requests</div>	<div><div></div>29</div> <div>Closed Issues</div>	<div><div></div>10</div> <div>New Issues</div>

Excluding merges, **5 authors** have pushed **62 commits** to master and **62 commits** to all branches. On master, **91 files** have changed and there have been **3,836 additions** and **536 deletions**.



# small RNA-seq analysis

---

## processing & QC

cutadapt  
fastqc  
qualimap  
multiqc

## de-novo

seqcluster  
mirdeep2 for miRNA  
protac for piRNA

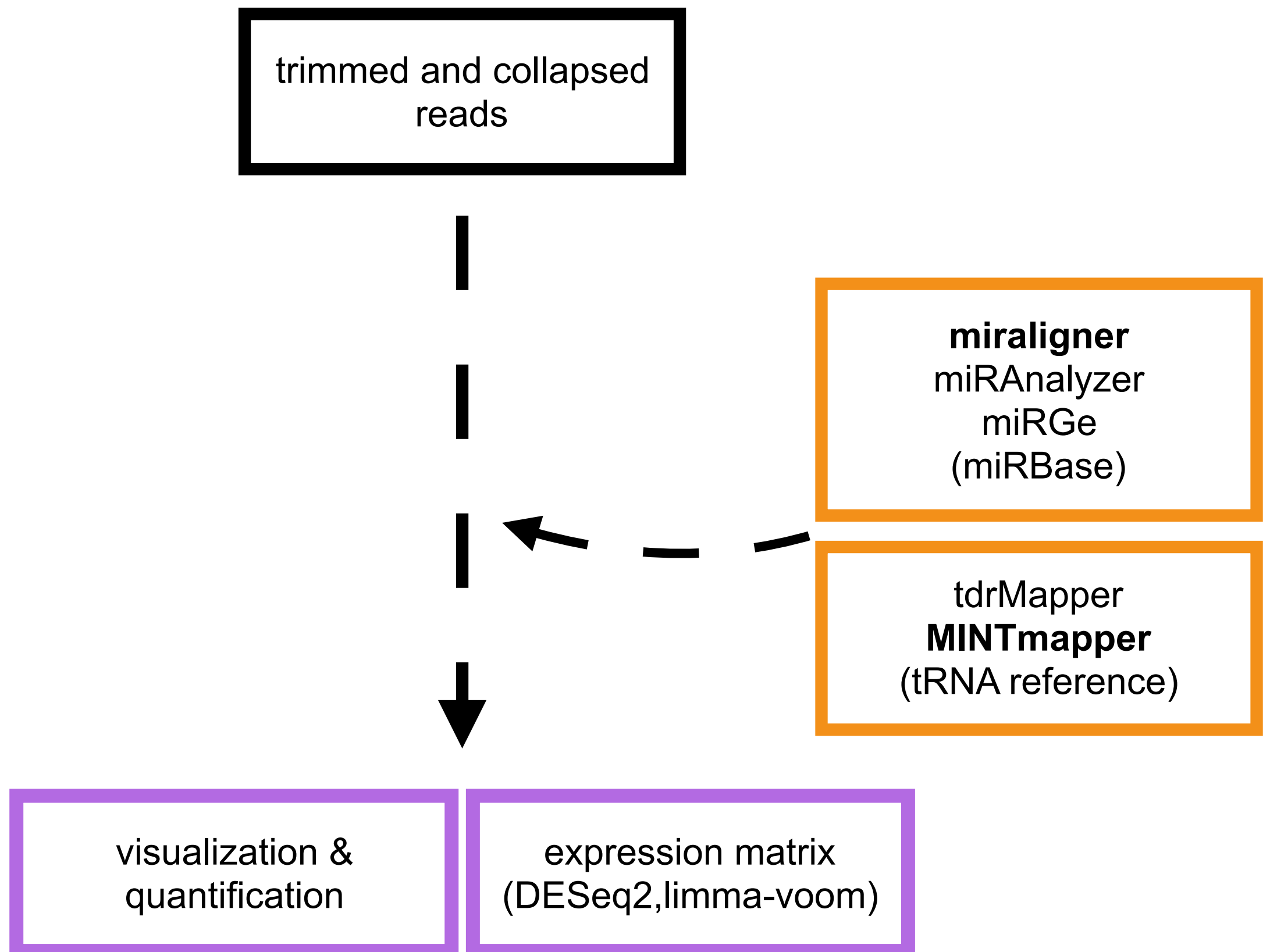
## detection & annotation

miraligner  
miRAnalyzer  
miRGe  
tdrmapper  
MINTmapper



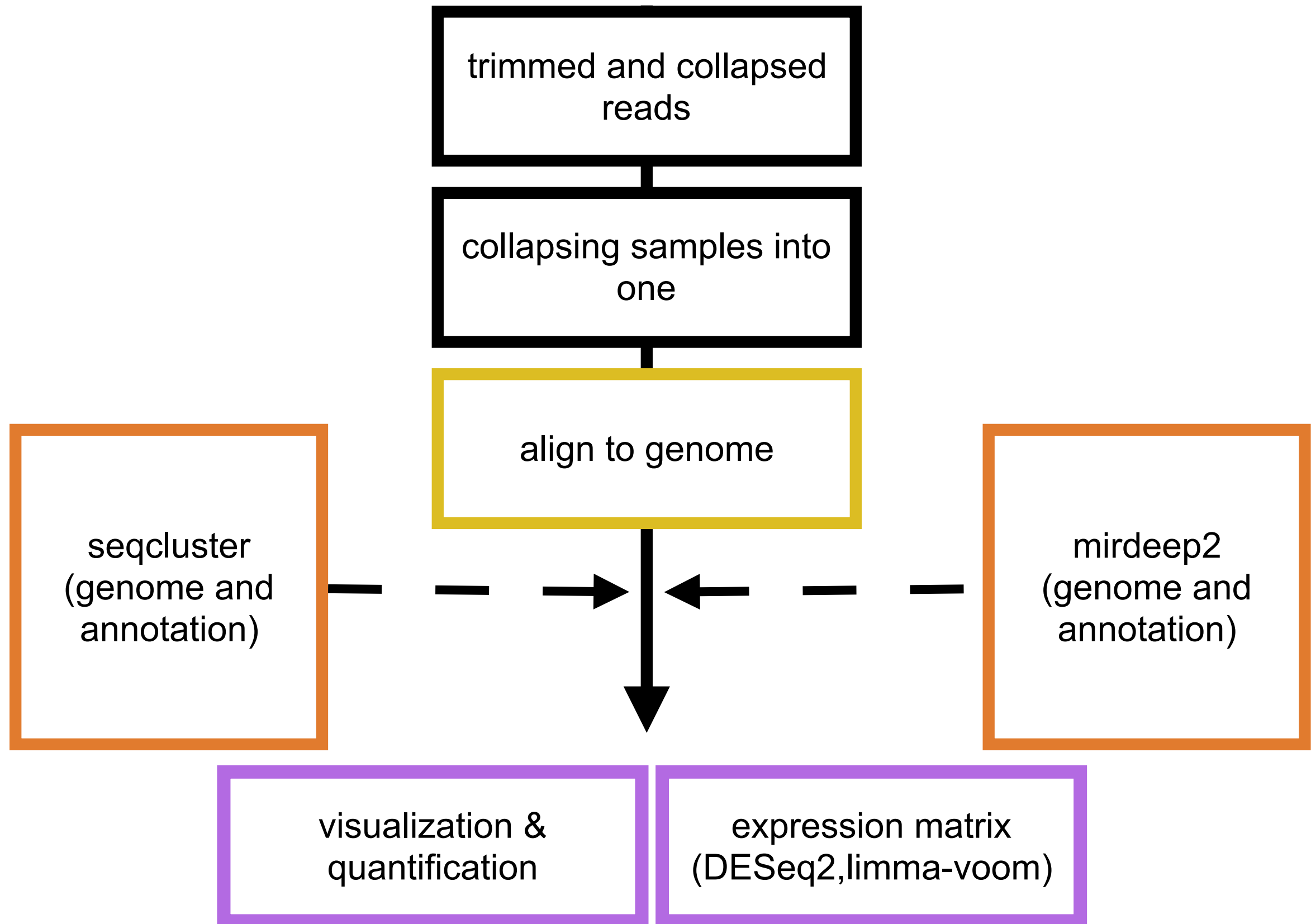
# Detection & Annotation

---



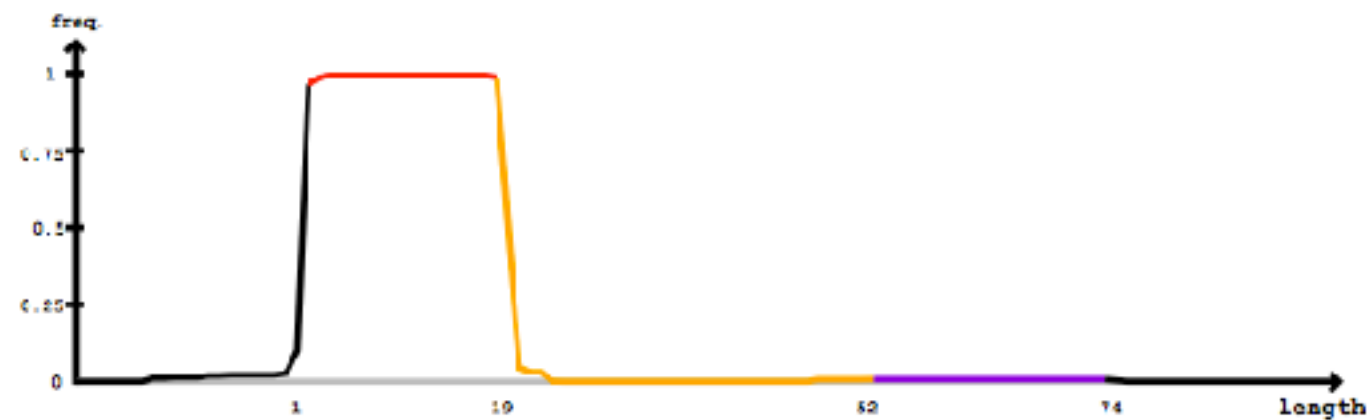
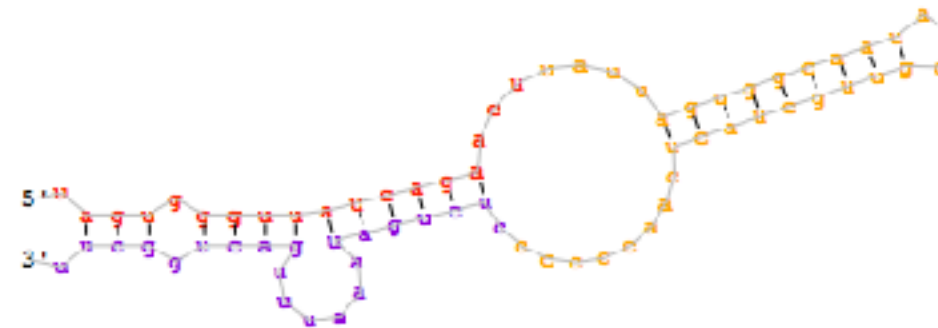
# De-novo detection

---



# 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

[illegible]

# seqcluster deals with multi-mapped reads

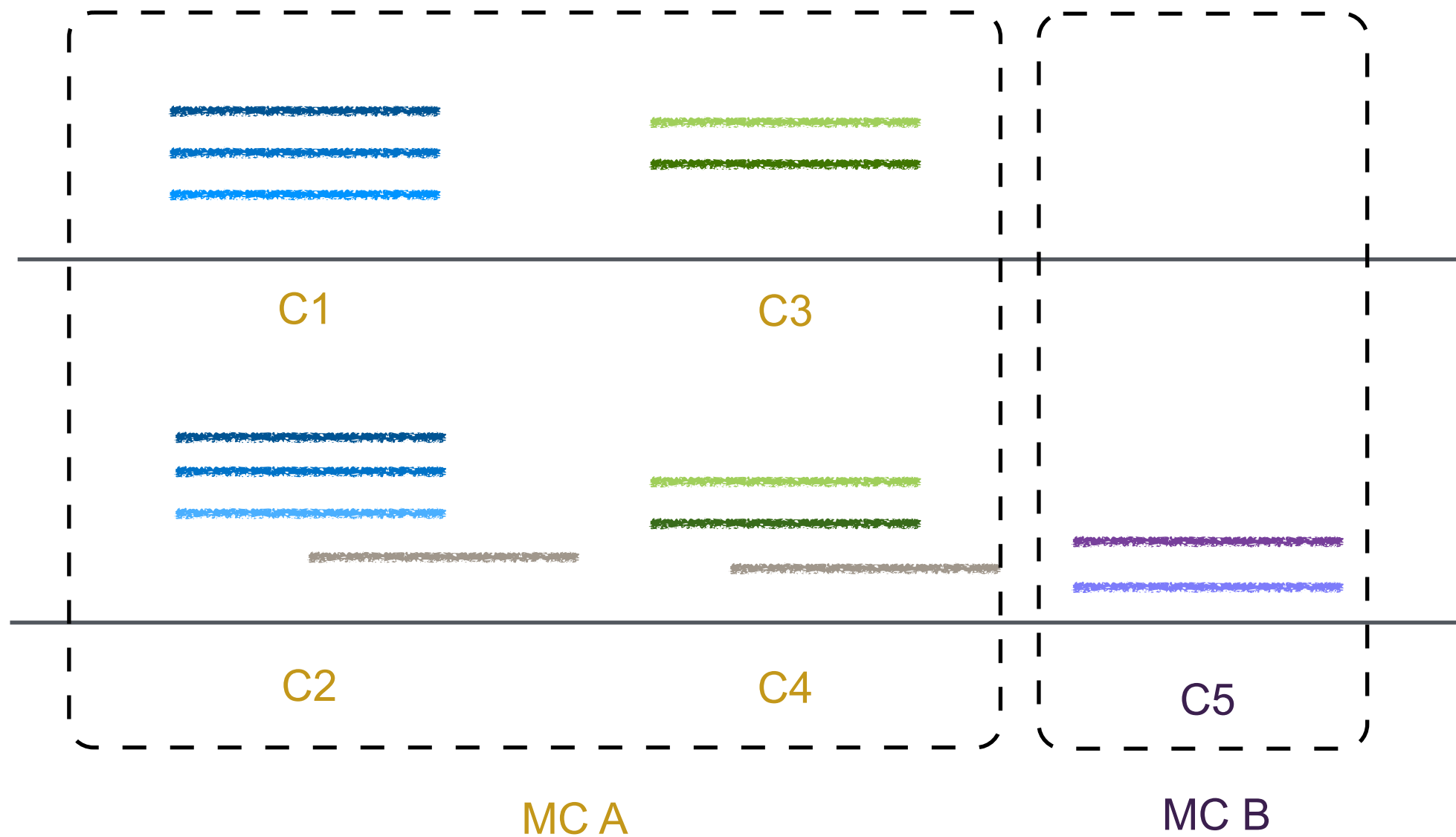
---



**meta-cluster**

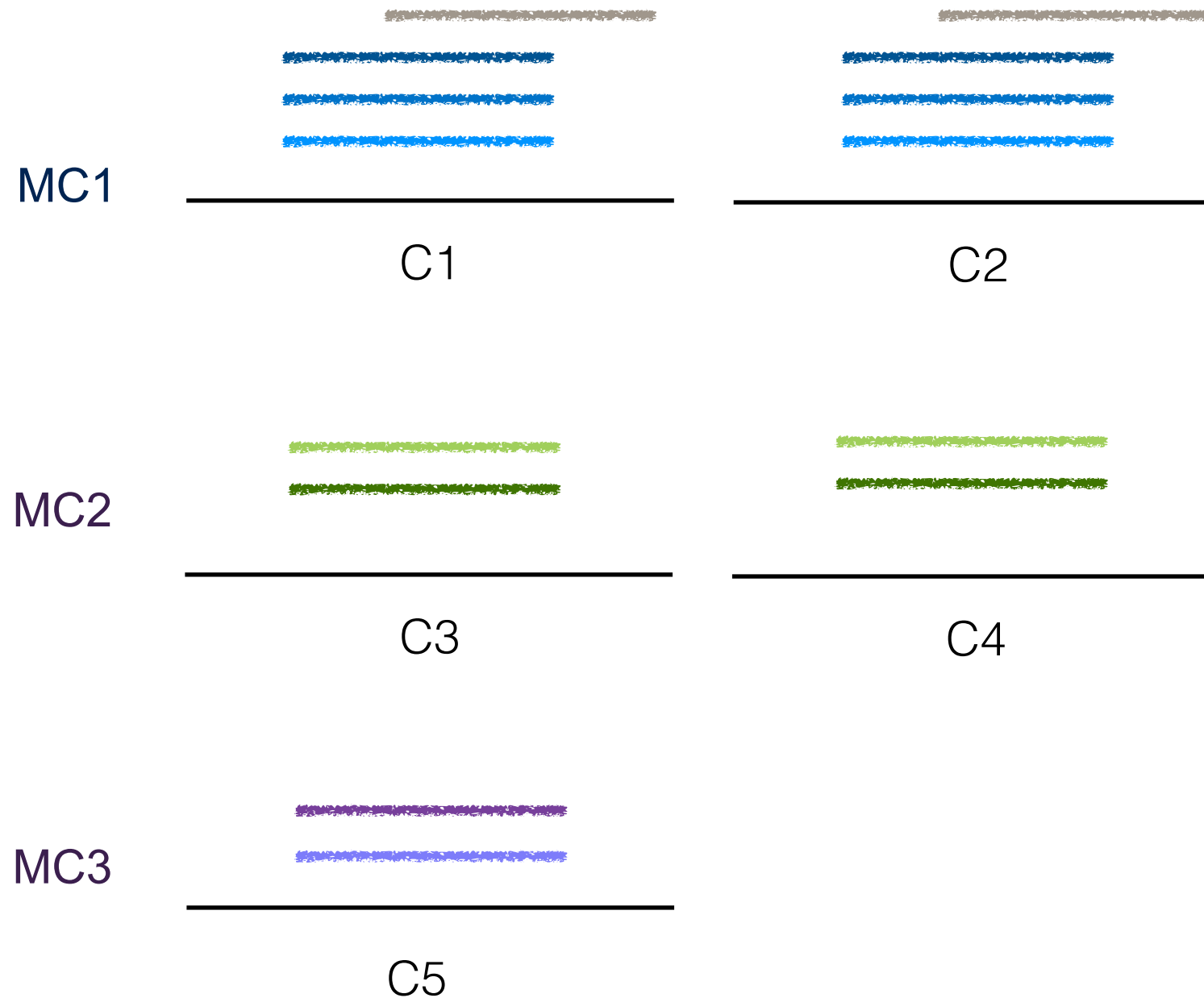
# Step 1: clustering

---



# Step 2: cleaning

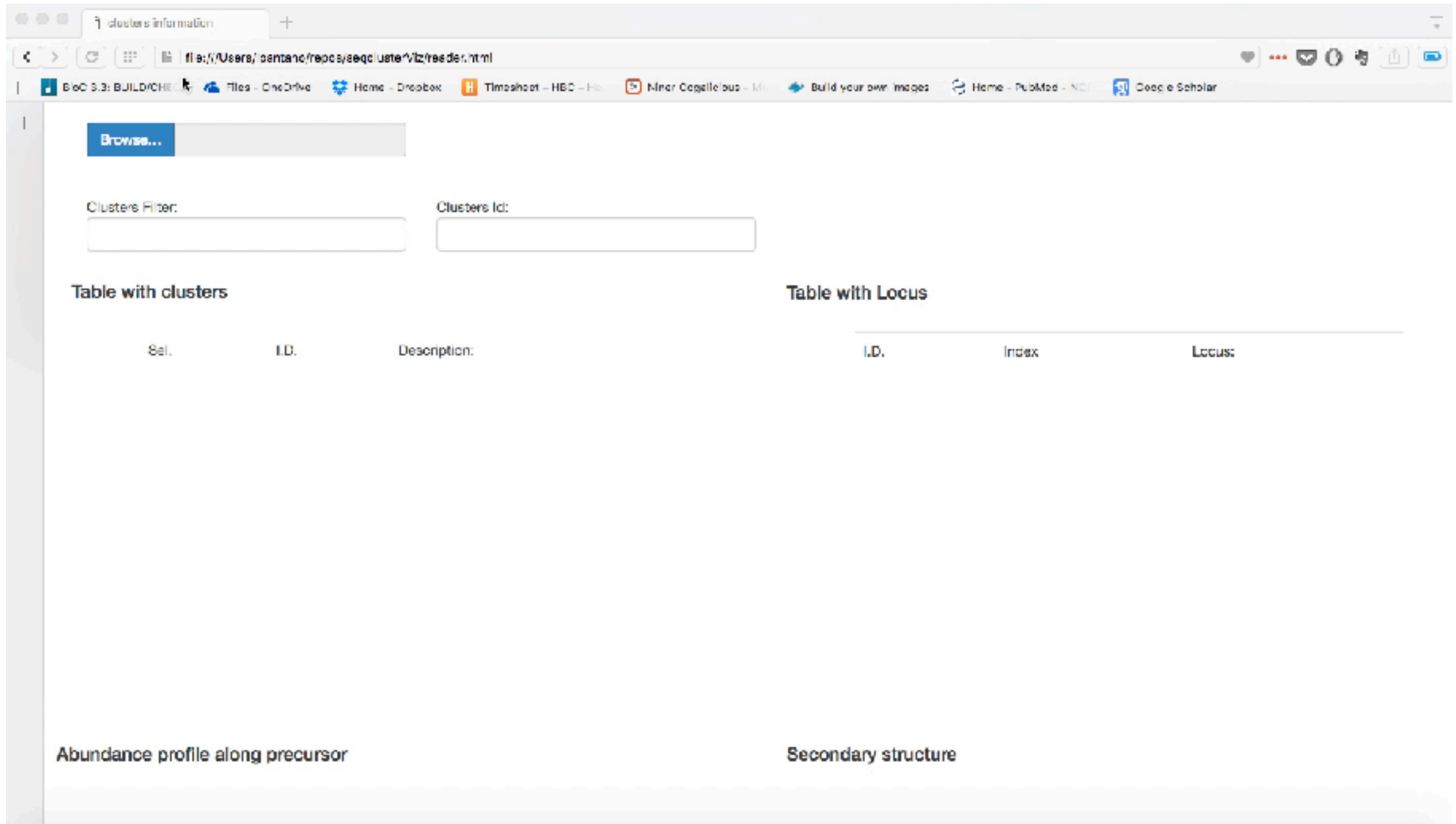
---



[seqcluster.readthedocs.io](http://seqcluster.readthedocs.io)



# seqcluster visualization



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

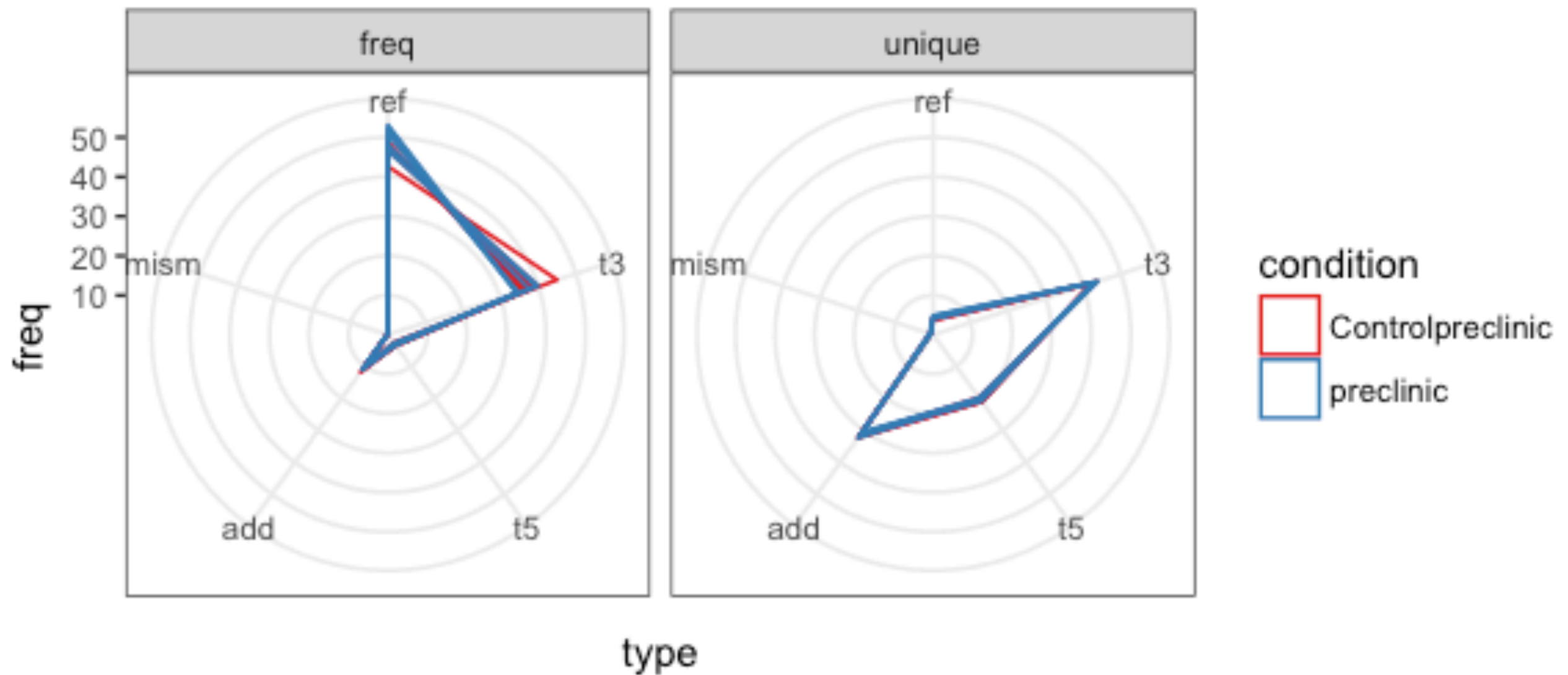
# isomiRs: R package

---

- 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



# miRNA naming

---

————— miRNA in database

.....

isomiR

**UPPER CASE:** addition

**lower cases:** deletion

mismatch

addition

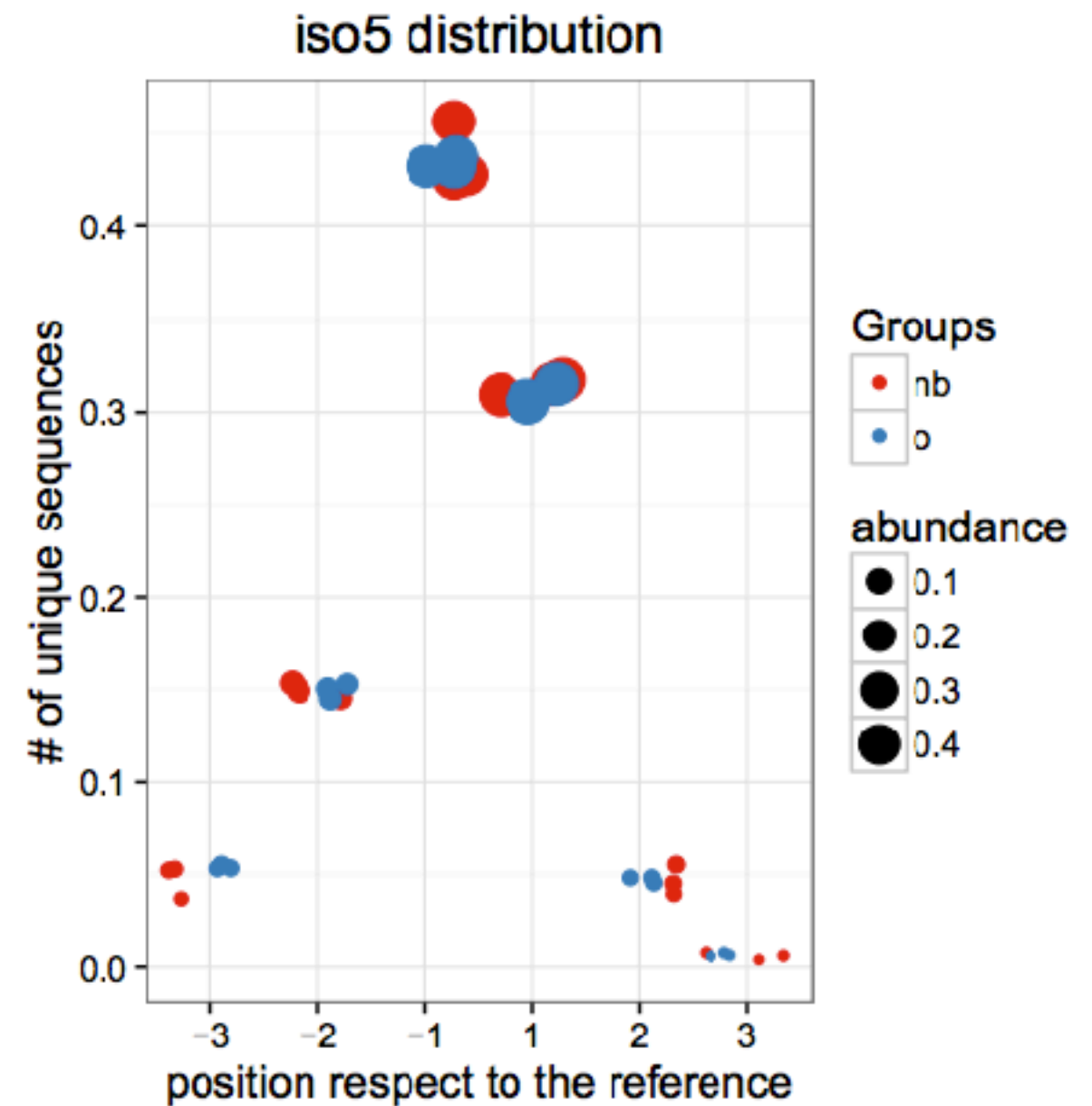
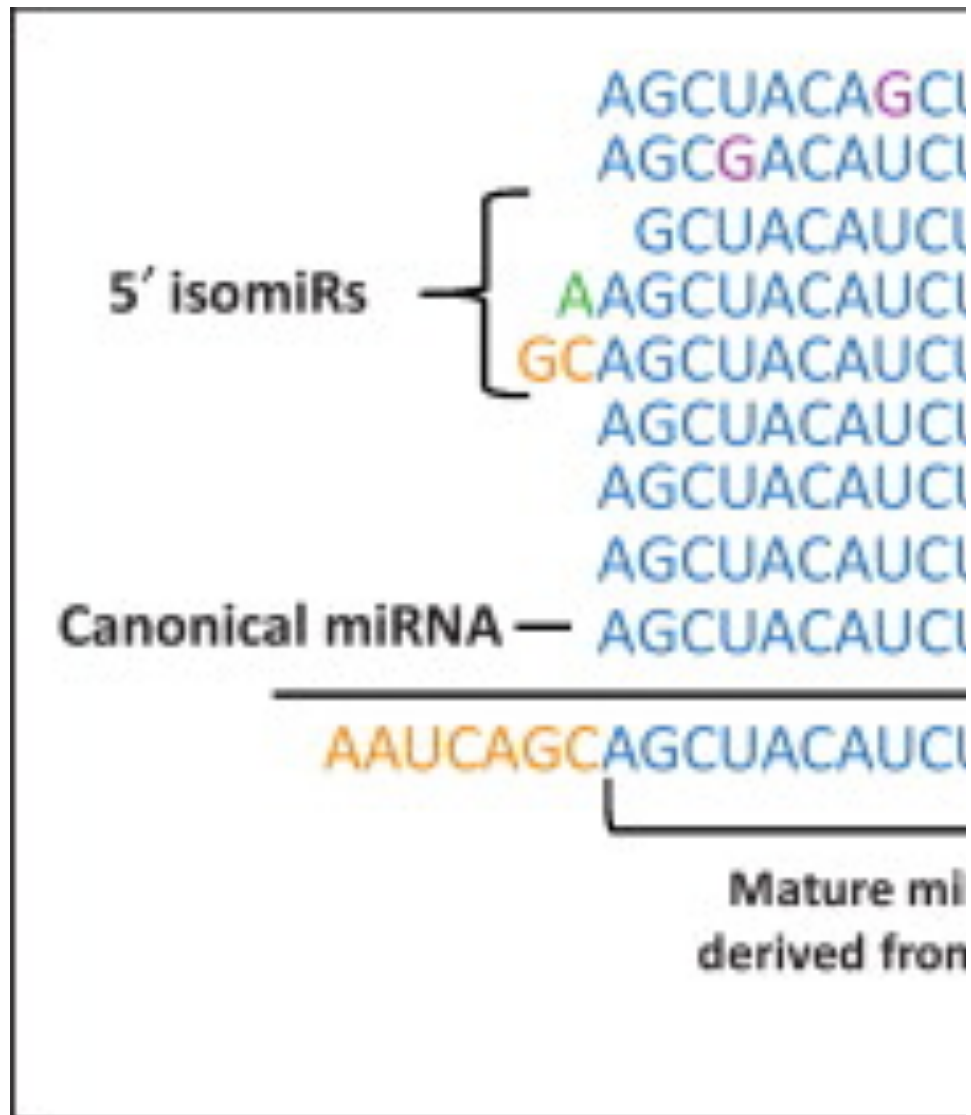
trimming 5'

trimming 3'

miRNA\_name:mismatch:addition:t5:t3

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

# isomiRs: R package

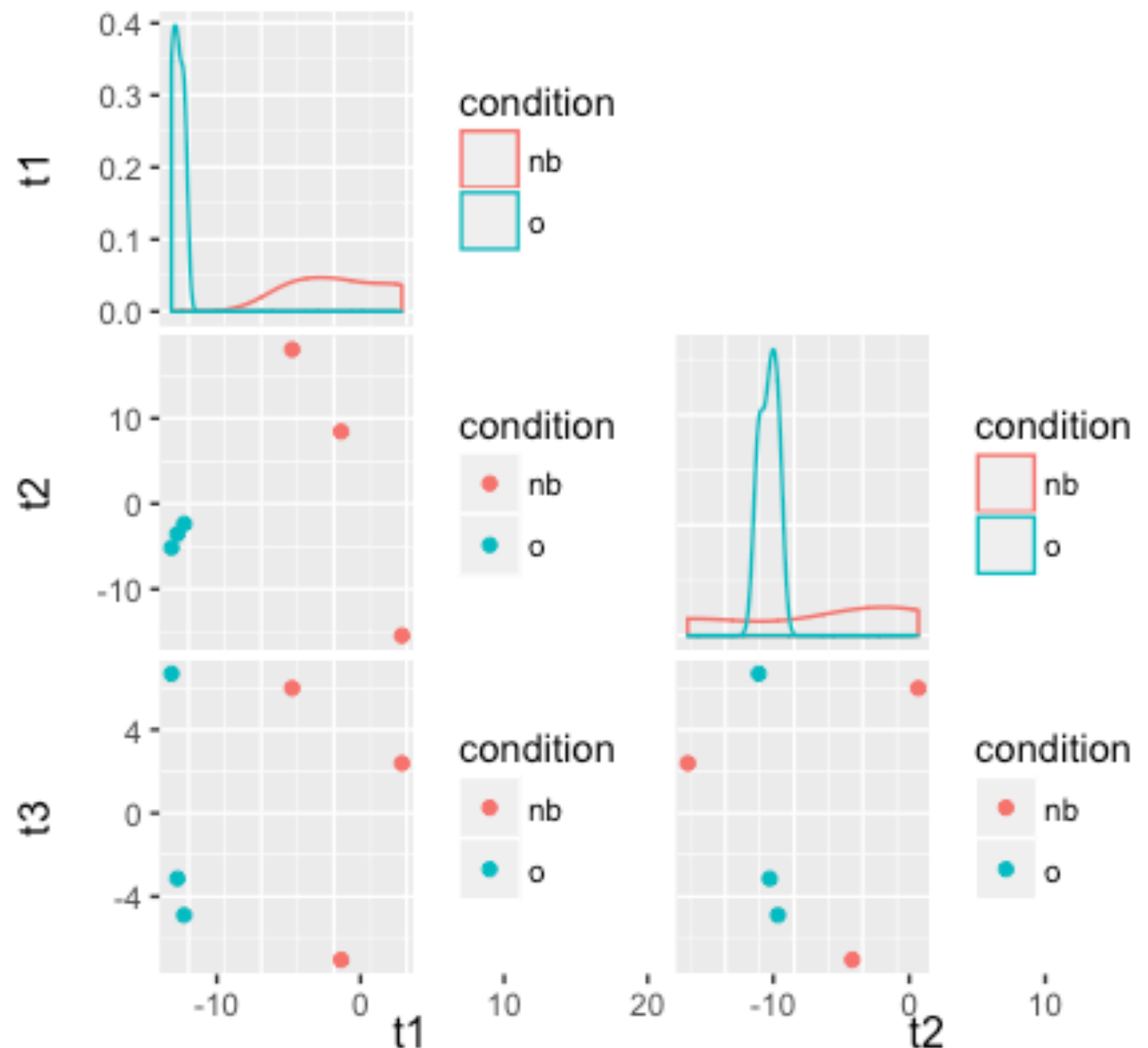


# PLS-DA

Similar to PCA

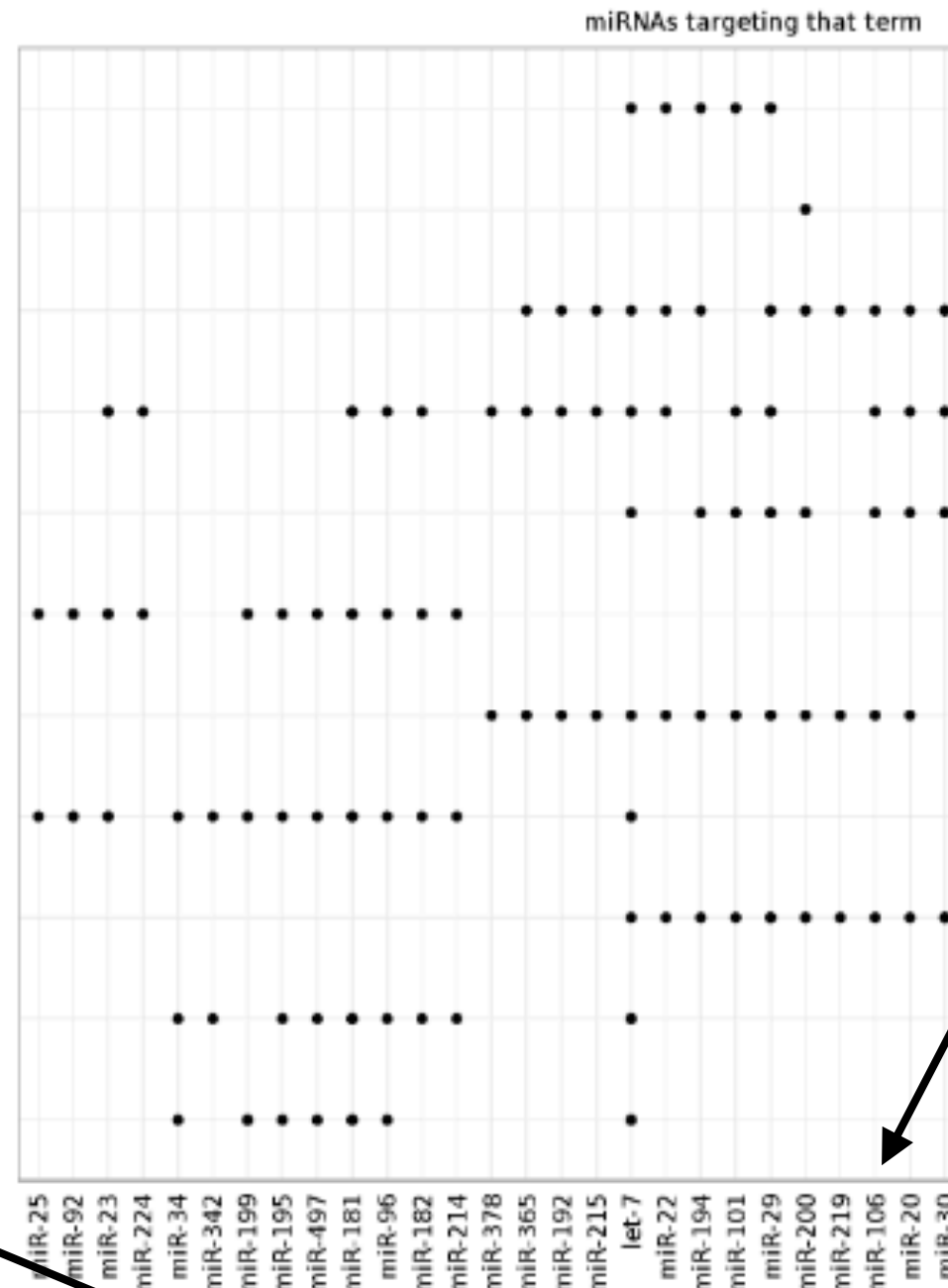
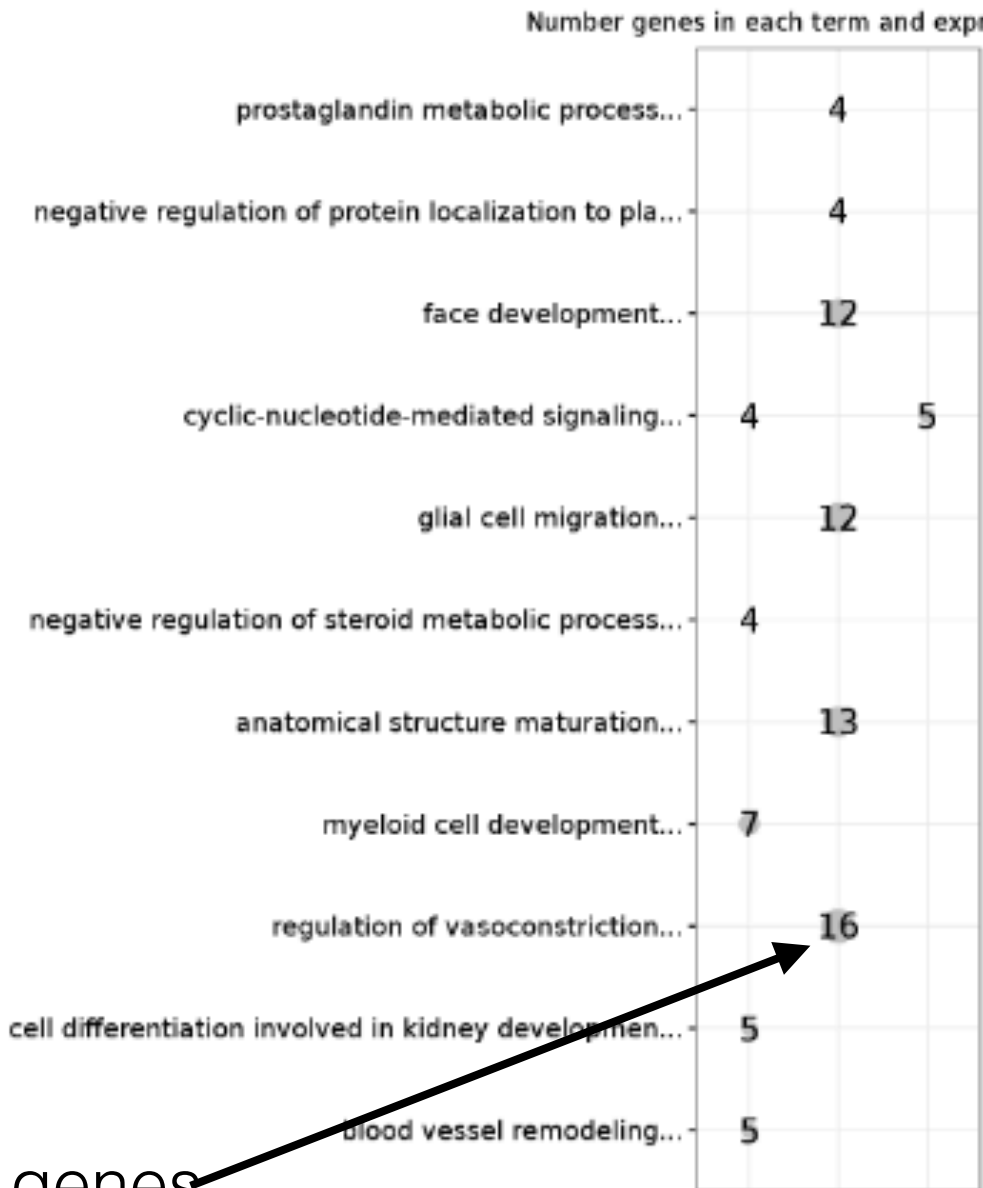
It outputs the most important miRNAs

It is compatible with correlation of miRNAs





GO Terms



# miRNAs

# genes

profiles

Expression



# MultiQC



**Phil Ewels**  
ewels

Bioinformatician working with next generation sequencing data.

 Science for Life Laboratory  
 Stockholm, Sweden  
 [phil.ewels@scilife-ab.se](mailto:phil.ewels@scilife-ab.se)  
 <http://phil.ewels.co.uk>  
 Joined on Nov 3, 2010

48

Followers

21

Starred

23

Following

STAR: % Uniquely mapped reads

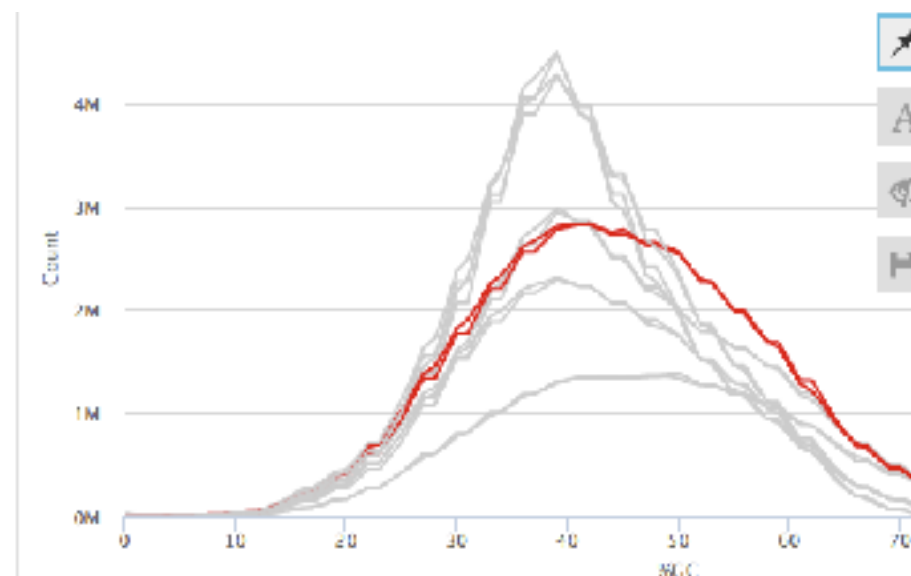
signed	% Aligned	M Aligned
0.9	81.1%	
1.5	79.1%	
1.9	70.2%	
0.9	63.2%	
0.7	61.8%	
0.6	50.6%	

## Sequence GC Content

6 5 1

Sequence GC content of reads. Normal random library typically have a roughly normal distribution of GC help.

## Per Sequence GC Content



## MultiQC Toolbox

### Highlight Samples

6:0

Regex mode ☐

11 1:14

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

Pieter Mestdagh, Nicole Hartmann, Lukas Baeriswyl, Ditte Andreassen, 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)

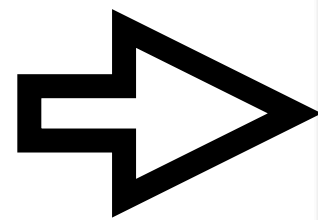
[Less...](#)

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 ...
```

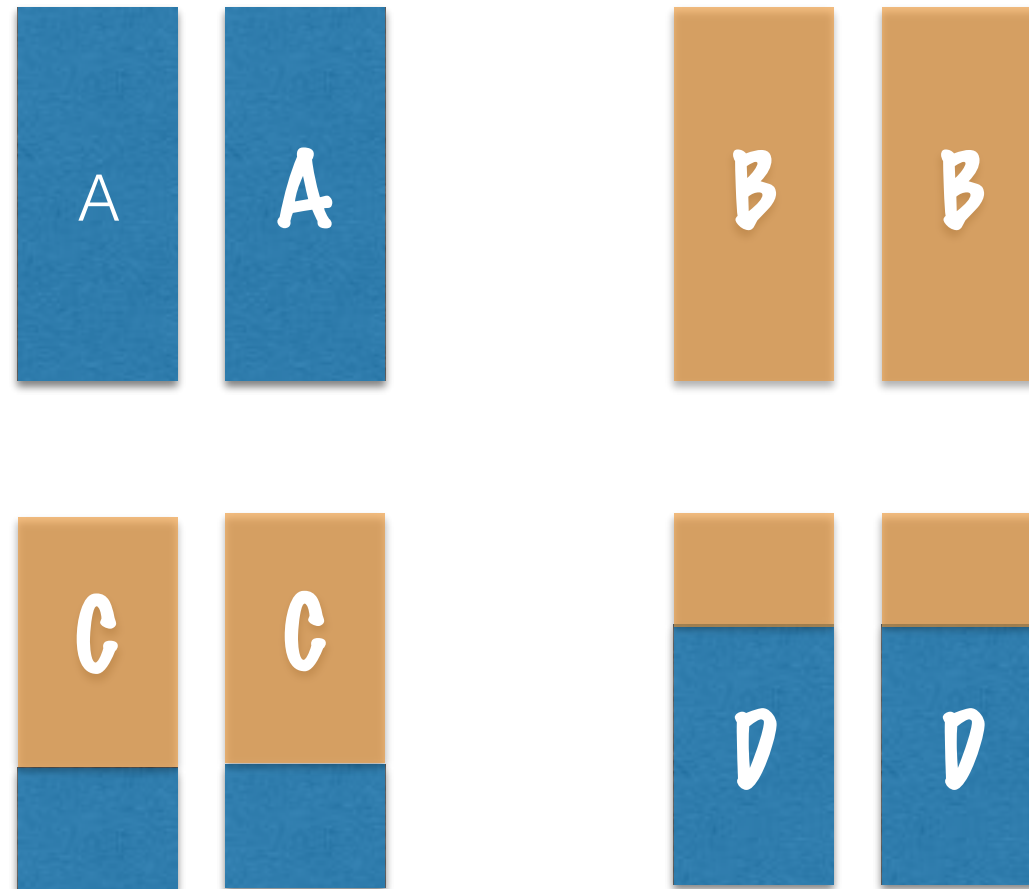
# Support of remote files

---

```
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

---

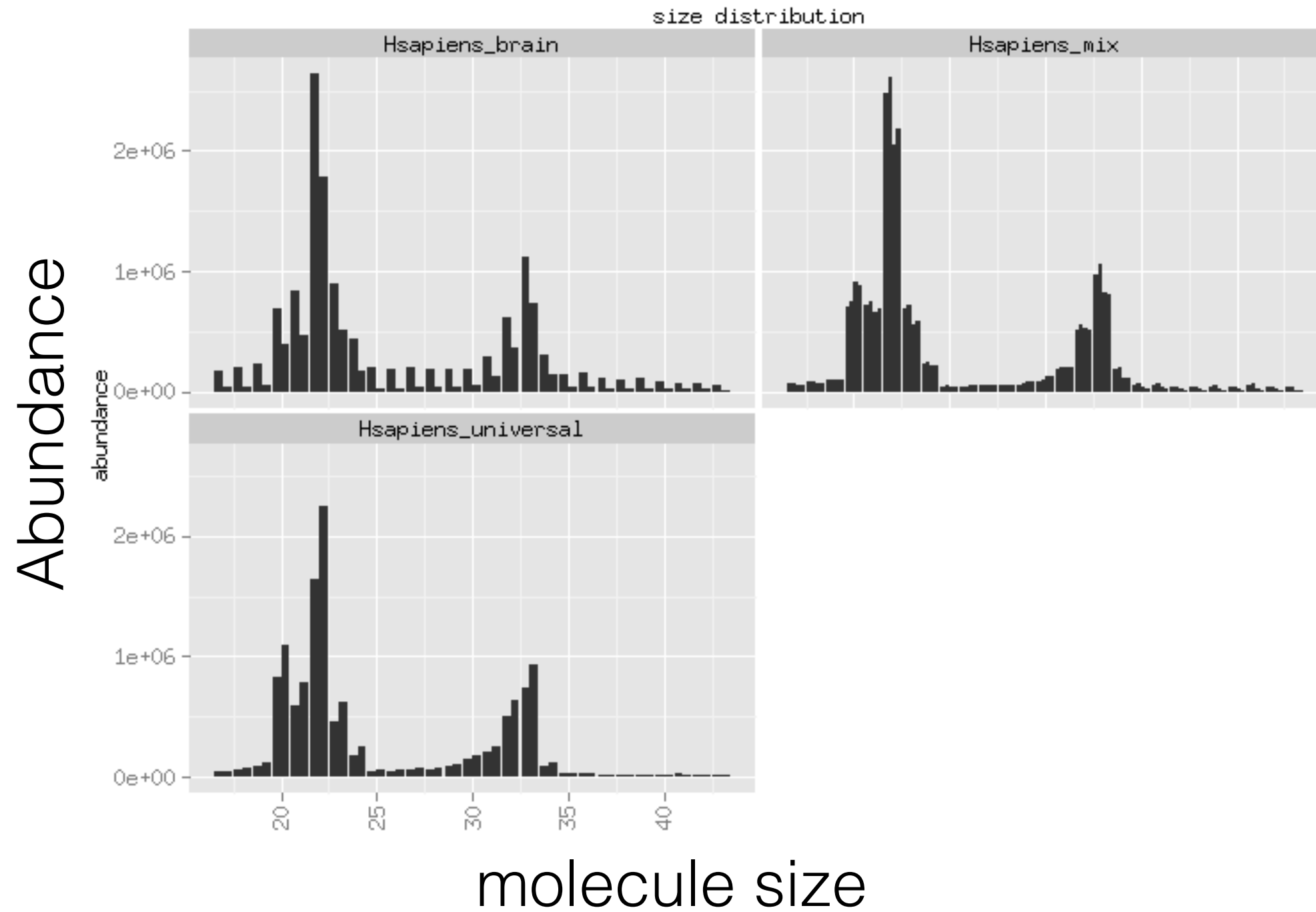


For each molecule:

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

# Good samples

---

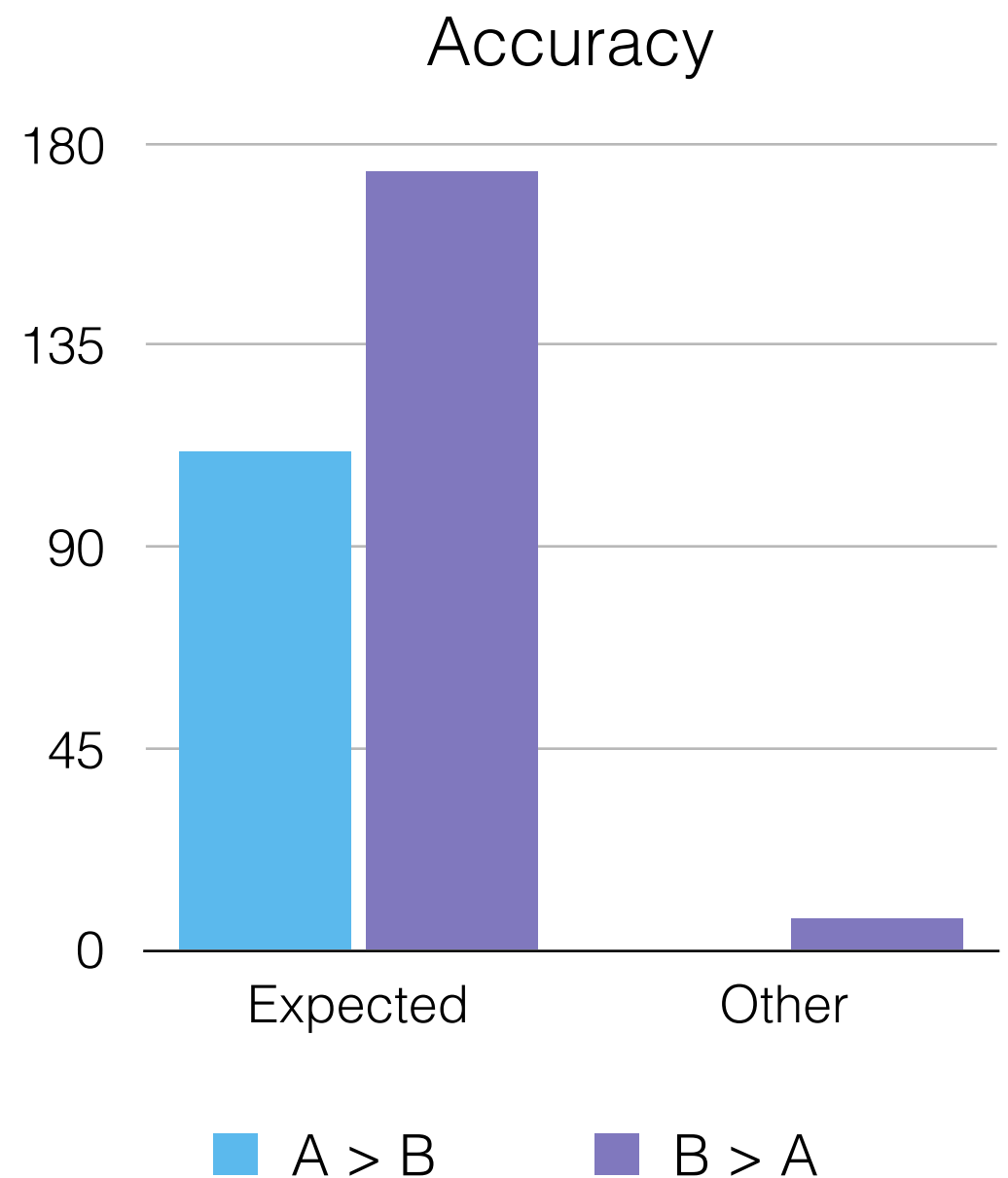




# miRNA quantification

---

miRNAs  $> 5$  counts in average  
upper quantile normalization

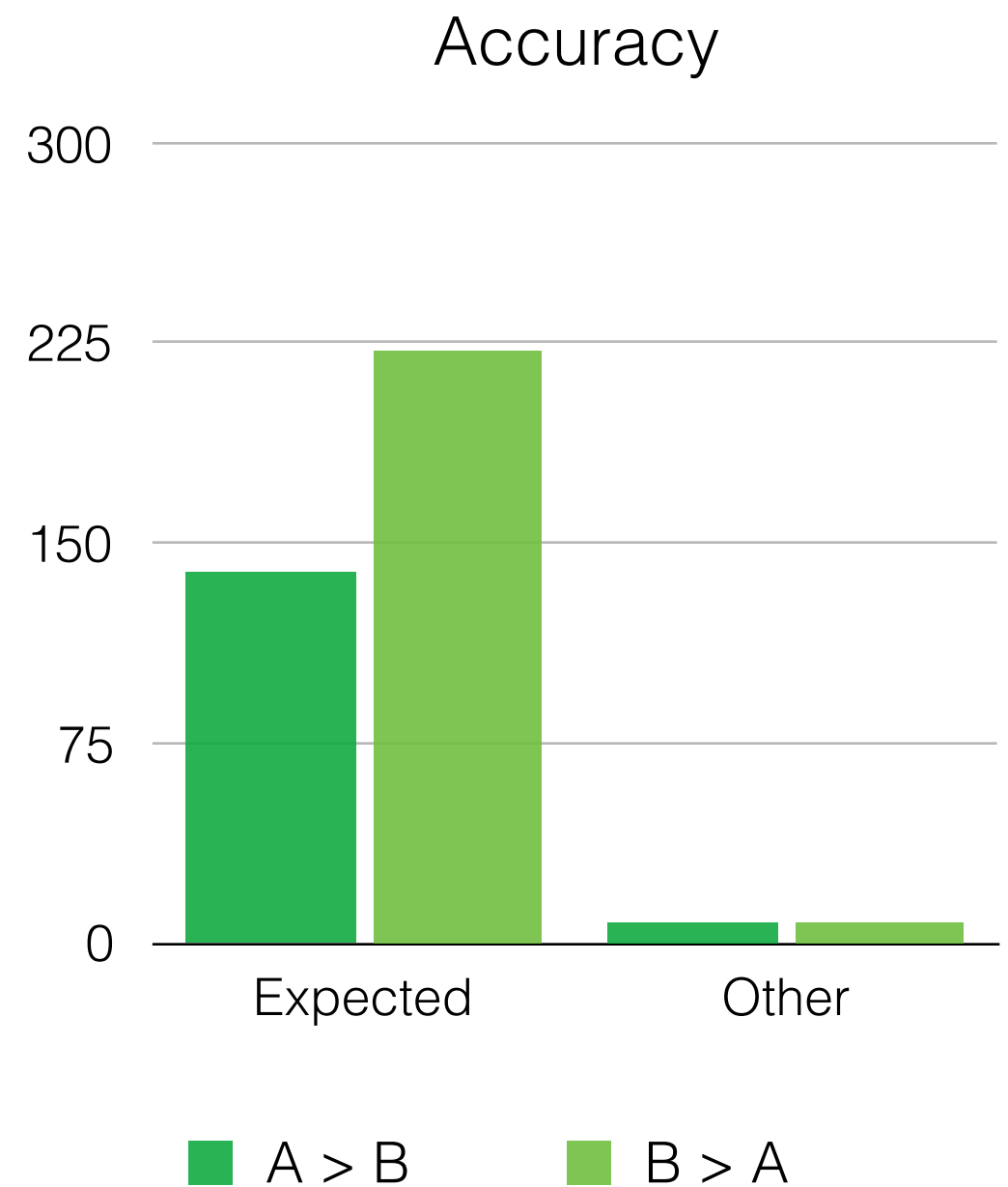




# clusters quantification

---

expression > 5 counts in average  
upper quantile normalization



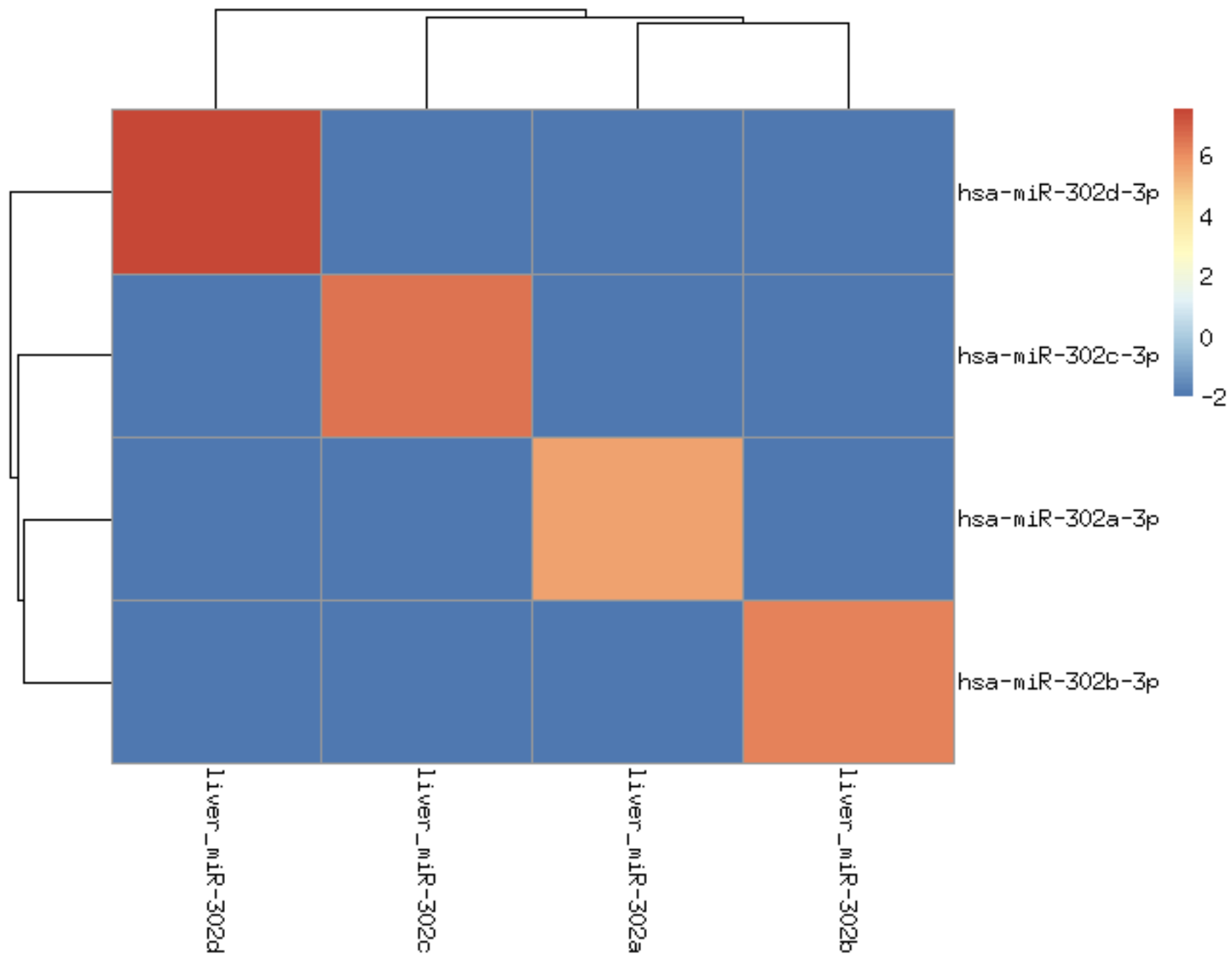
# Positive controls

---



# 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

---

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

The screenshot shows the GitHub repository page for mirTOP. The repository name is "mirTOP" with the description "miRNA transcriptome open project" and the URL "http://mirtop.github.io". The page includes navigation tabs for "Repositories", "People 3", "Teams 1", and "Settings". Below these is a search bar "Find a repository..." and a "New repository" button. The repository list shows three items:

- incubator**: "Where all ideas and discussions happen to lead to new repositories", updated 3 days ago, 1 star, 1 fork.
- mirtop**: "command lines tool to annotate miRNAs with a standard mirna/isomir naming", updated 3 days ago, Python, 0 stars, 0 forks.
- miRTOP.github.io**: "project for small RNA standard annotations", updated on Mar 29, CSS, 0 stars, 0 forks.

Annotations are overlaid on the image:

- standard formats** and **naming rules** (in dark blue) are positioned over the "mirtop" repository description.
- best-practices** (in dark green) is positioned to the right of the "mirtop" repository.
- miRNAs, tRNAs ...** (in orange) is positioned over the "miRTOP.github.io" repository description.

# 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 1

Meetups

Past Meetups 15

Our calendar 

Help support your Meetup

[Chip in](#)

## Welcome!

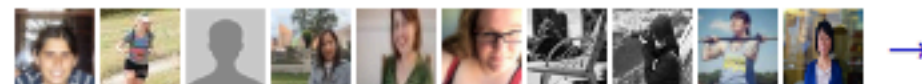
[+ Schedule a new Meetup](#)

[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.