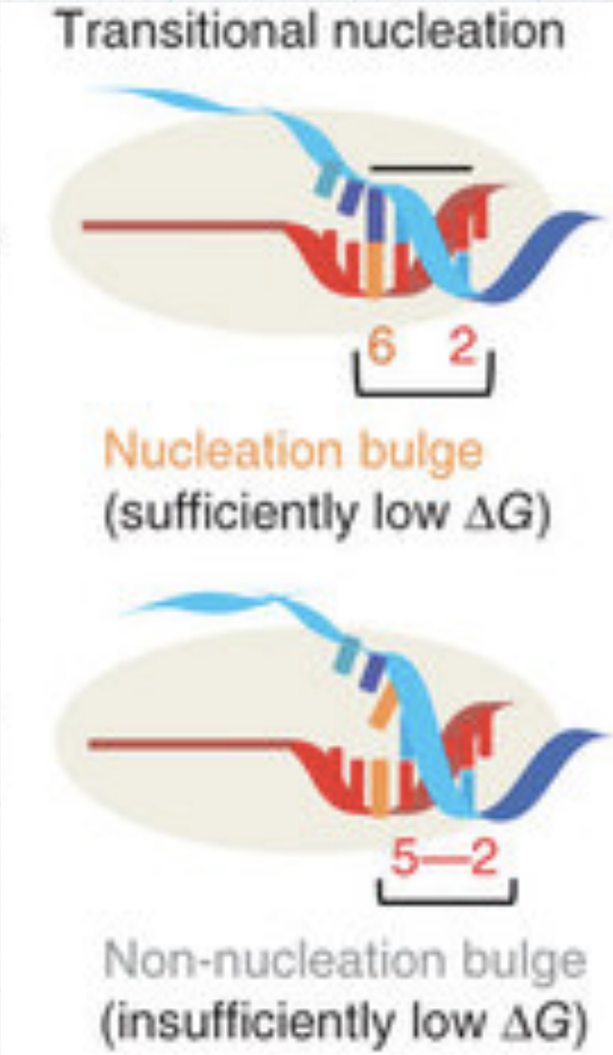
small RNAseq with bcbio

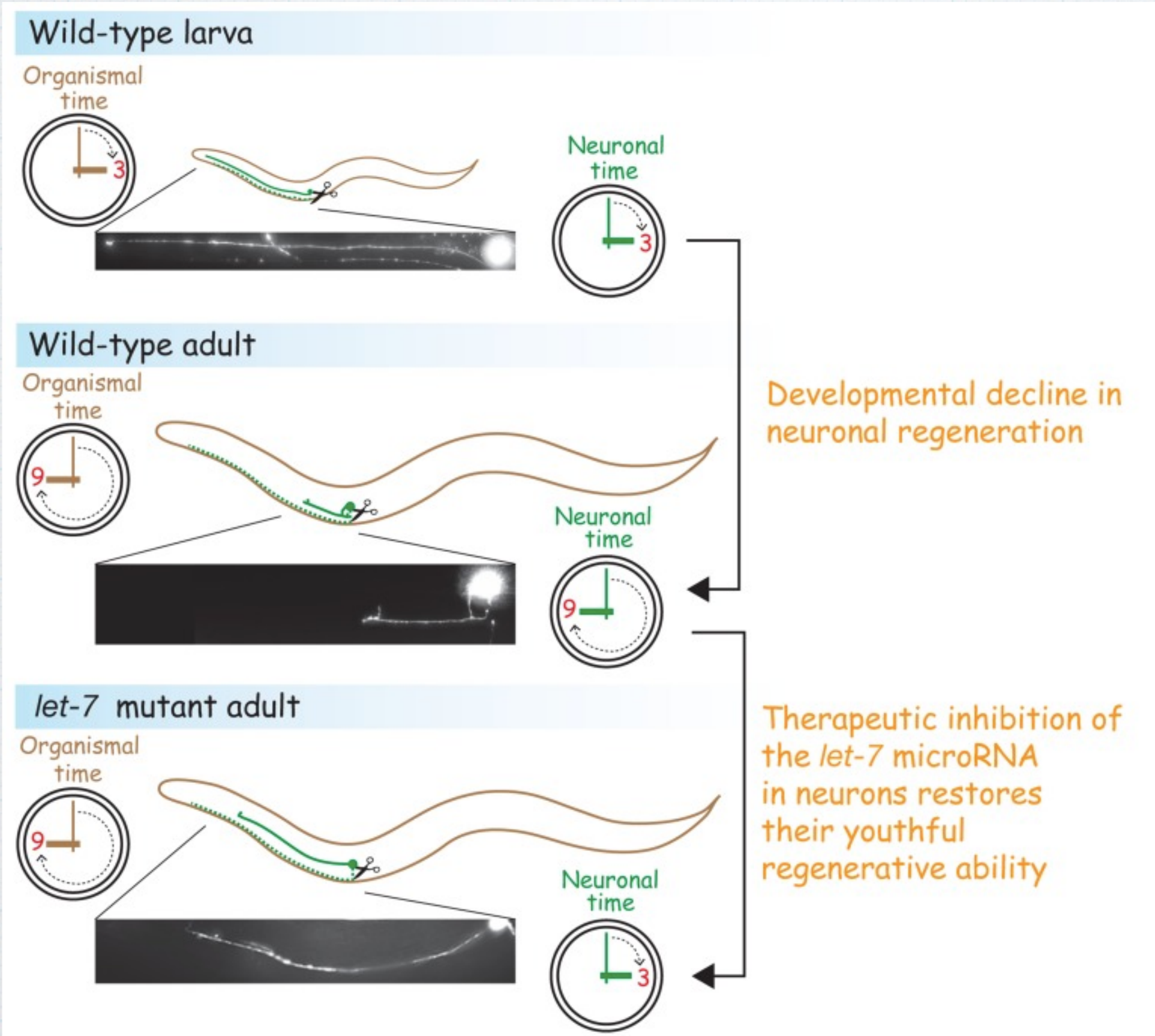
Lorena Pantano
Harvard TH Chan School of Public Health

2015-12-03

small RNA

RNA molecules of 18-36 its long with regulation function





Hui Chiu and Chieh Chang, Aging (Albany NY). 2013 Jul; 5(7): 485–486.

isomiRs

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

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

isomiRs

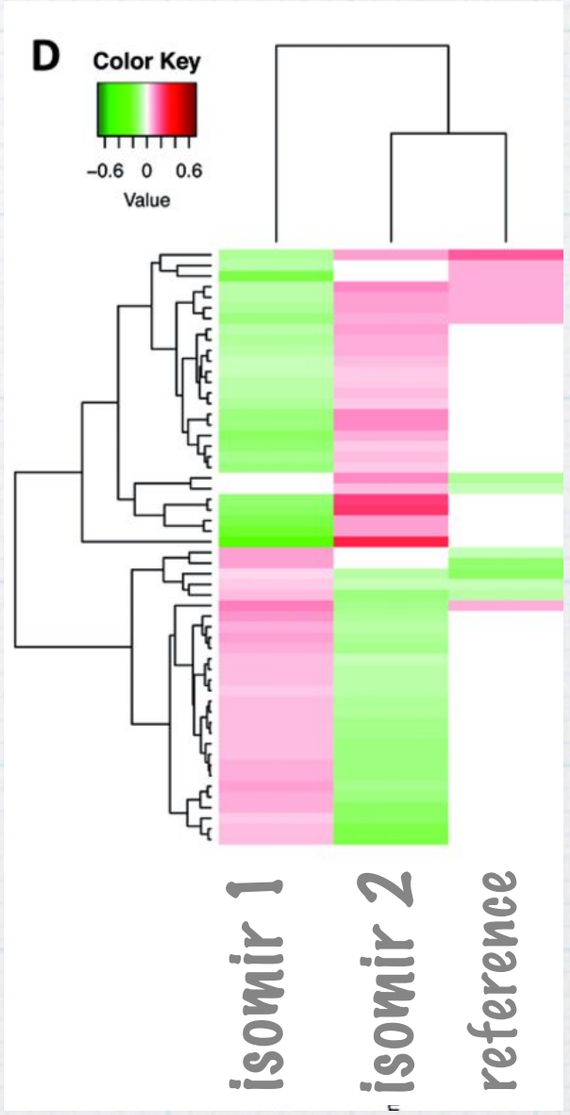
reference

isomir +2 nt

isomir -2 nt

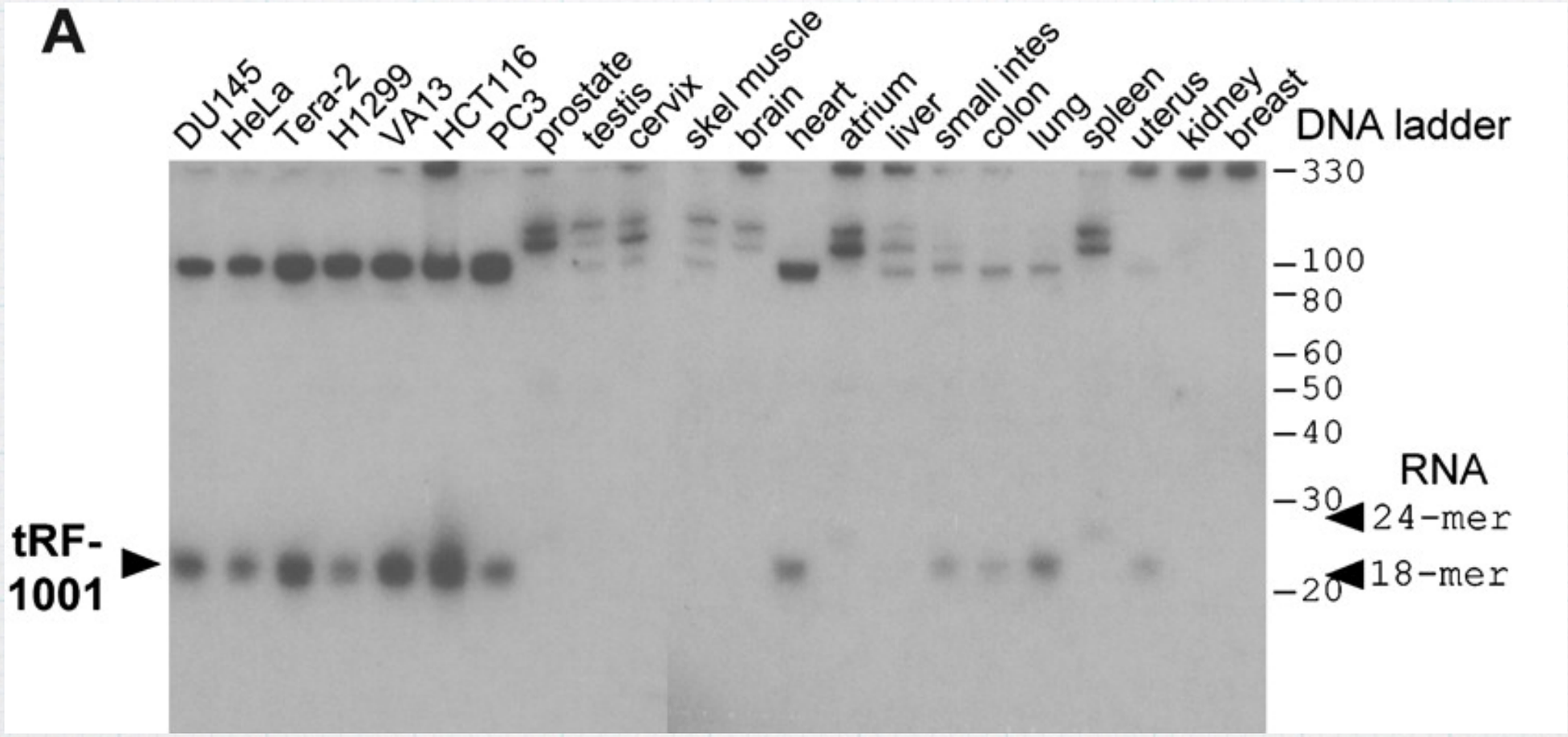
transfected mammary cells line
derived from metastatic site

Gene expression

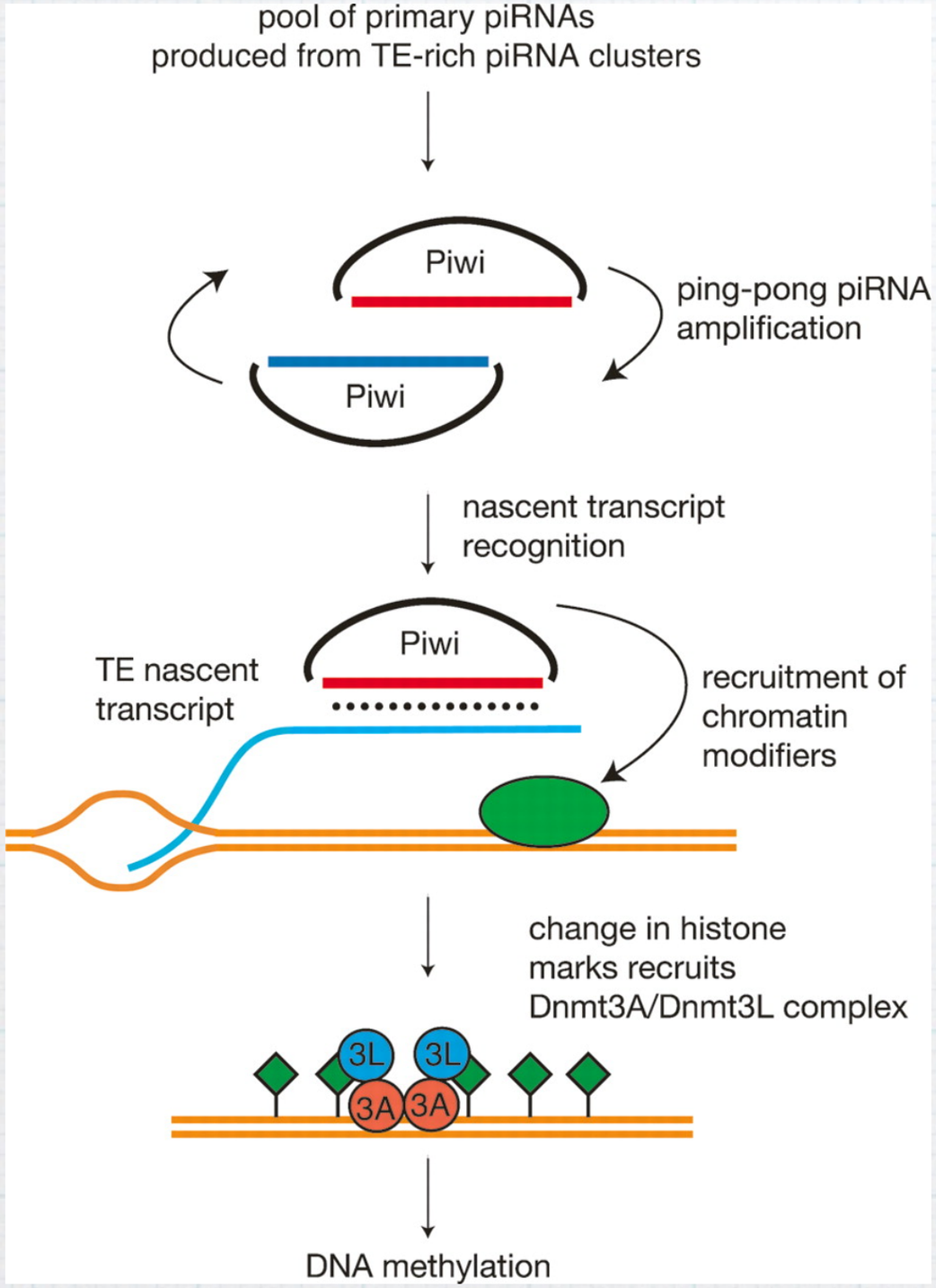


Aristeidis G. Telonis et al. Nucl. Acids Res. 2015;nar.gkv922

small tRNAs



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



Alexei A. Aravin, and Déborah Bourc'his Genes Dev. 2008;22:970-975

piRNAs

bcbio-nextgen

processing & QC

fastqc
qualimap

detection & annotation

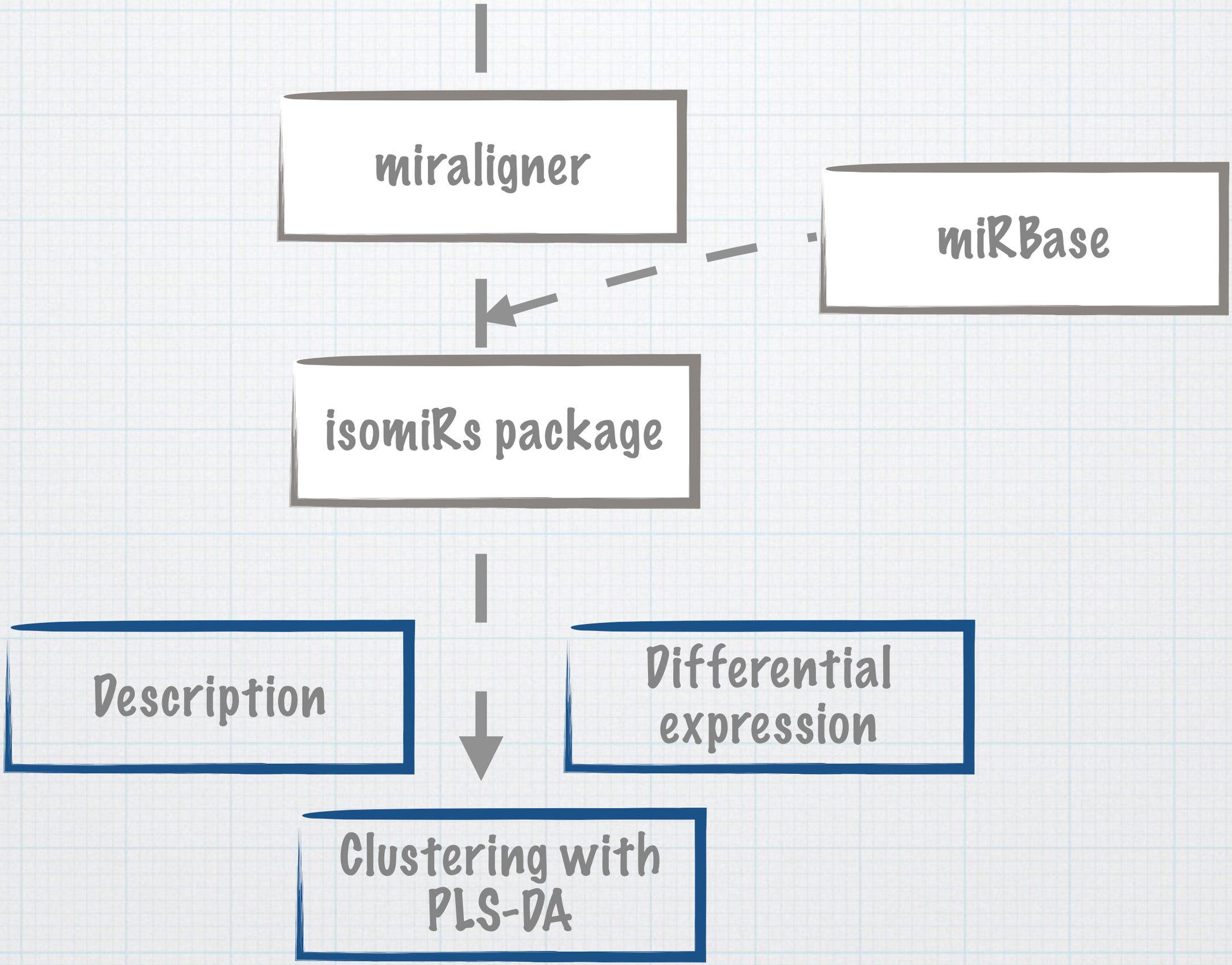
miraligner
seqcluster
tdrmapper

de-novo

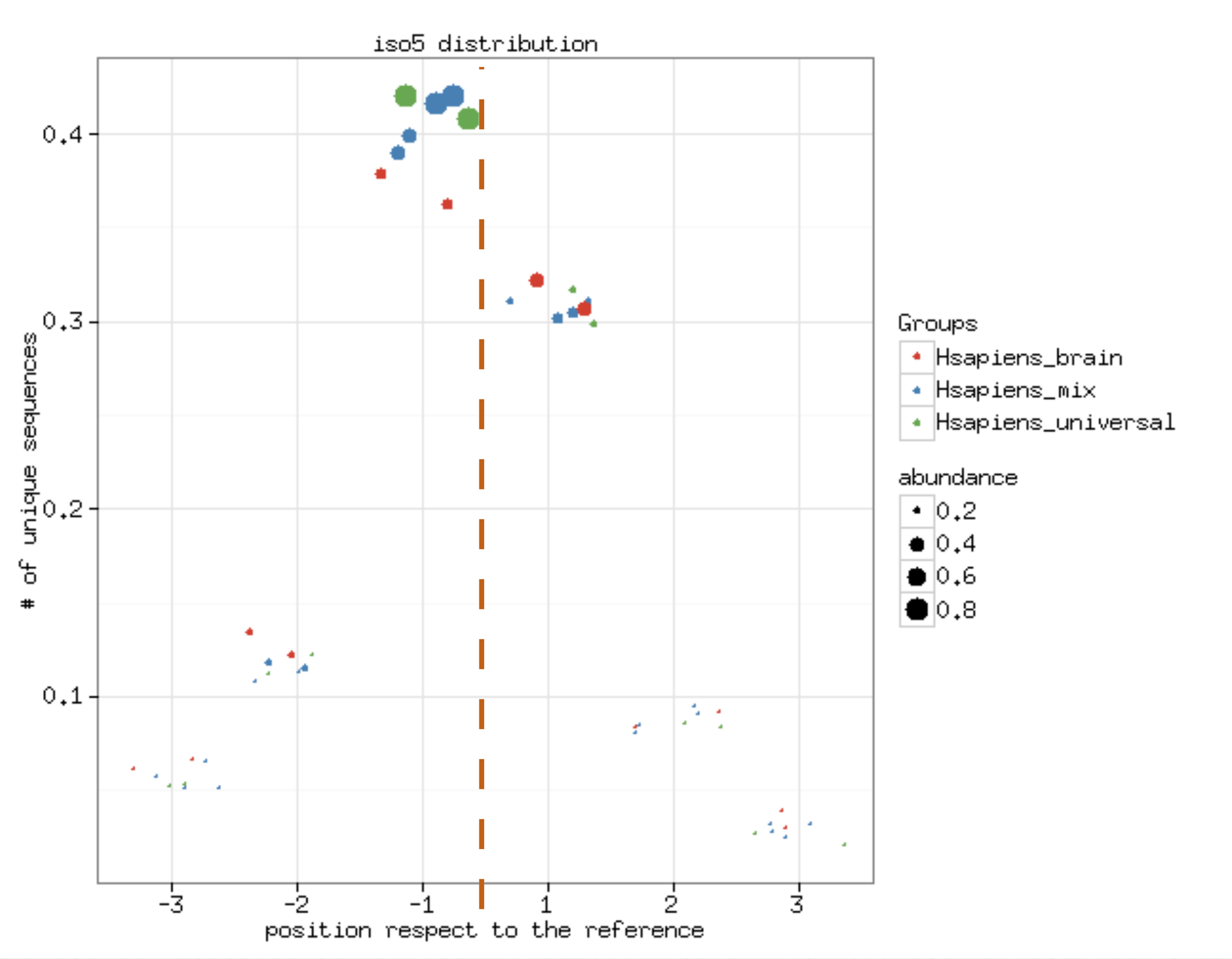
mirdeep2 for mirna (current)
protac for pirna (next)

challenges

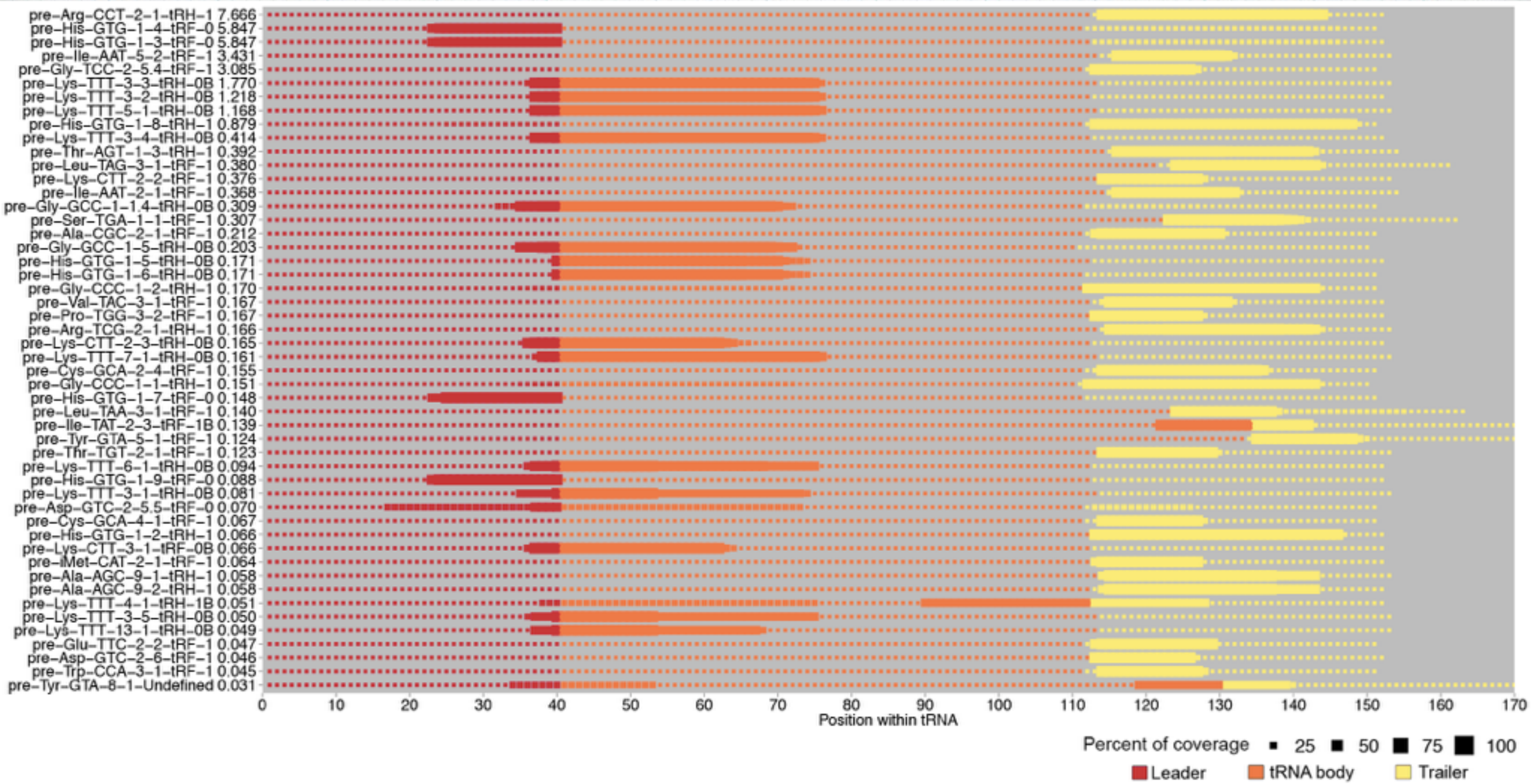
* isomiRs
* small RNAs coming from multiple precursors over the genome (multi-mapped reads can be **40%** of the data.)
* differentiate degradation and functional molecules
* non-model organism
* high variability among cell types/individuals


miraligner
miRBase
isomiRs package
Description
Differential expression
Clustering with PLS-DA

isomiRs at 5' end of the miRNAs



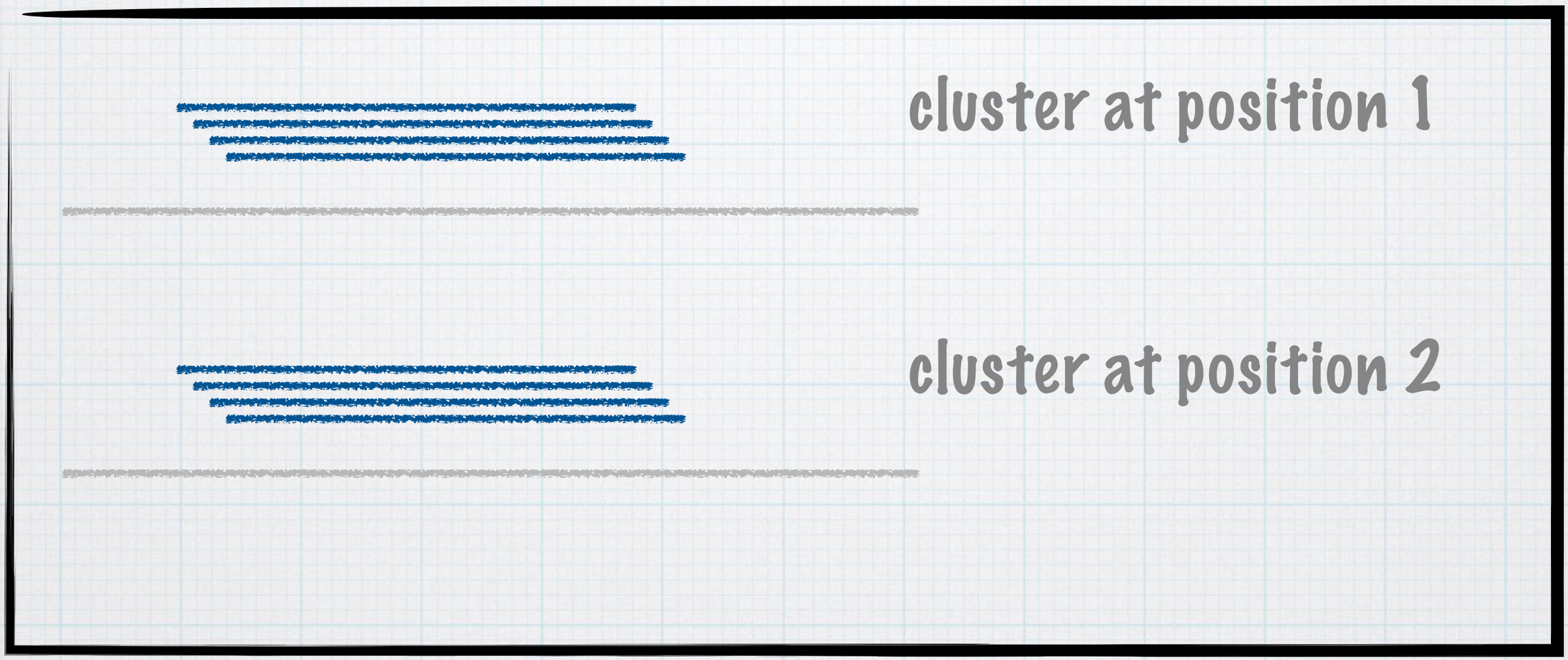
tRNA analysis



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

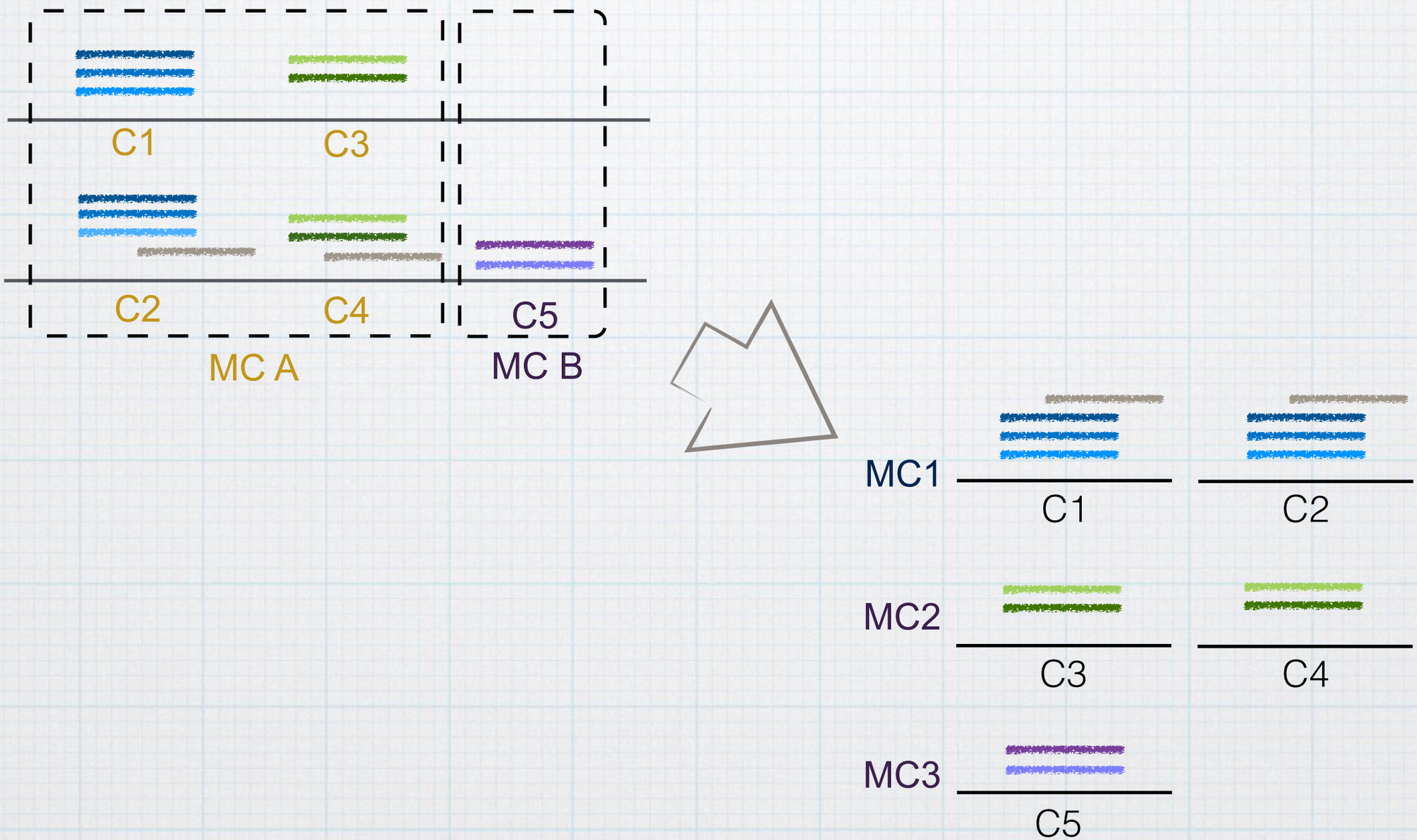
seqcluster deals with multi-mapped reads

seqcluster naming



meta-cluster

multi-mapped reads



annotation

Only for well annotated genomes

meta-cluster

	C1	C2	C3	C4
tRNA	YES	YES	YES	NO
miRNA	NO	NO	NO	NO
repeat	YES	NO	NO	NO

...

Most-voted strategy

miRQC project

Quantitative PCR (PCR)

EX	miRCury (Exiqon)
OA	OpenArray (Life Technologies)
TM	TaqMan Cards (Life Technologies) *
TMp	TaqMan Cards preAmp (Life Technologies)
QI	miScript (Qiagen)
QU	qScript (Quanta BioSciences)
WA	SmartChip (WaferGen)

Hybridization (HYB)

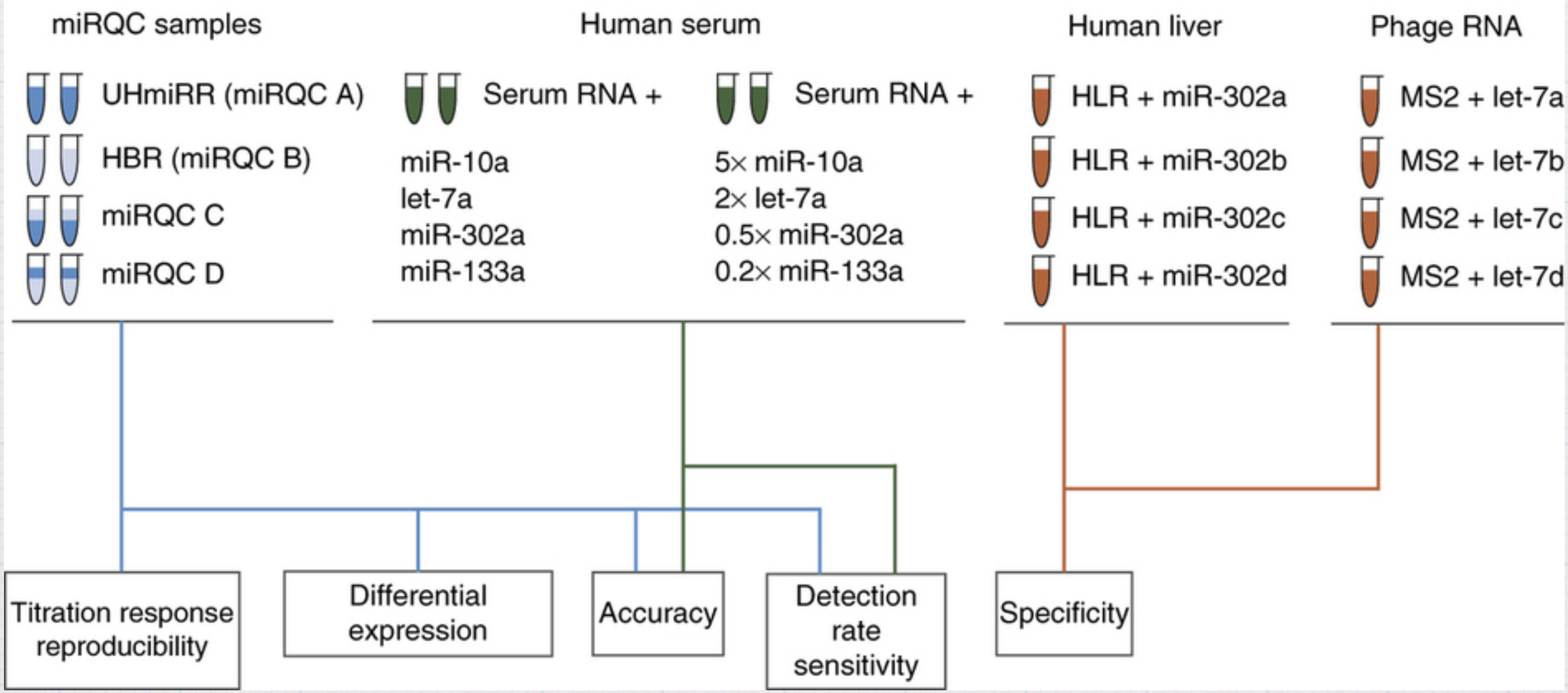
AF	microarray (Affymetrix) *
AG	microarray (Agilent)
NS	nCounter (Nanostring) *

Sequencing (SEQ)

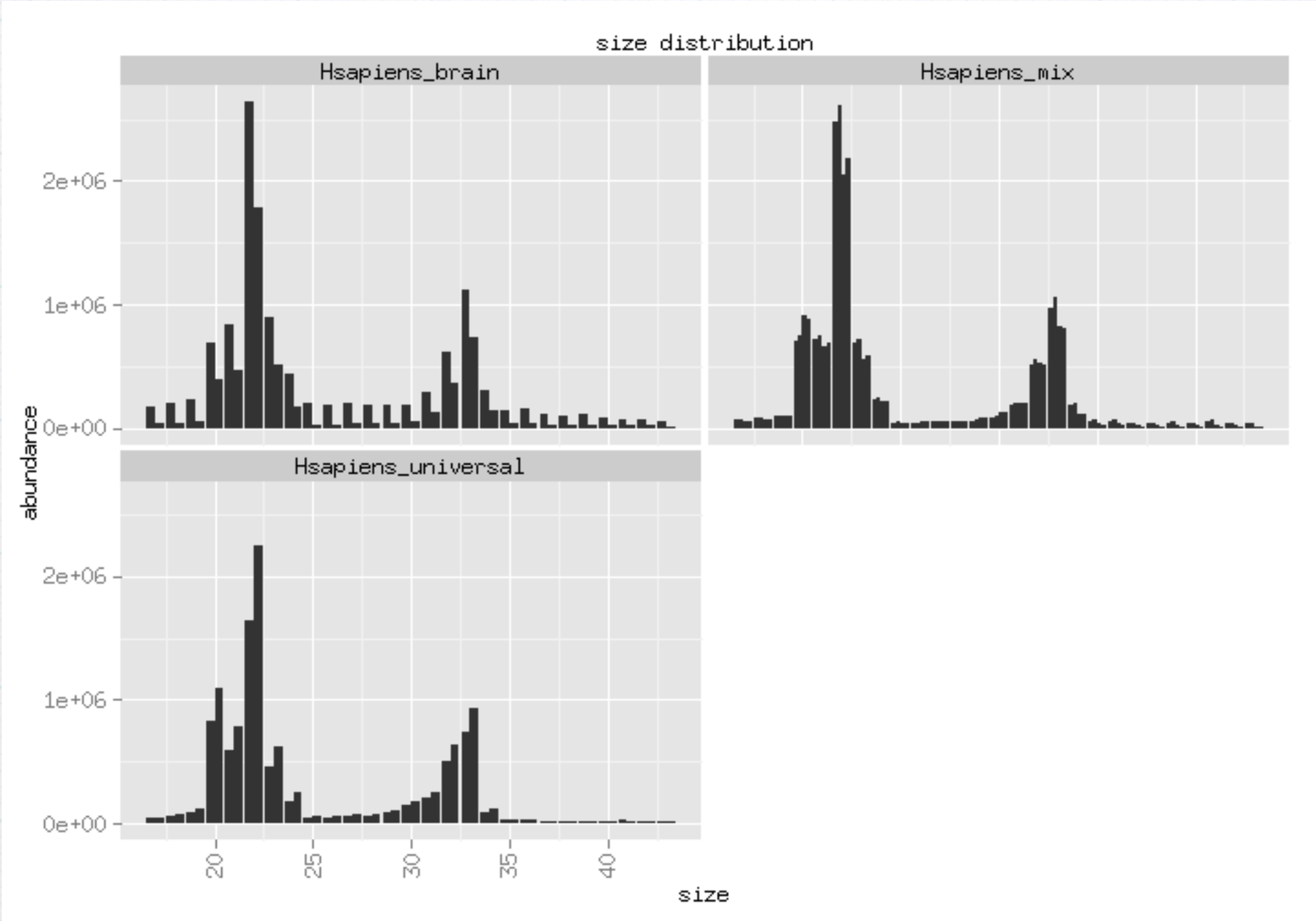
IL	TruSeq (Illumina)
IT	Ion Torrent (Life Technologies)

RNA input (ng)

EX	40	AF	400
OA	100	AG	100
TM	350	NS	100
TMp	50		
QI	500	IL	1,000
QU	800	IT	1,000
WA	1,000		



Good samples



Quantification

* A: universal human RNA sample
* B: human brain sample
* C: 25% of A + 75% of B
* D: 25% of B + 75% of A

For each miRNA:
* If A > B then A > D > C > B
* If B > A then A < D < C < B

miRNA quantification

miRNAs > 5 counts in average

upper quantile normalization

miRNAs which A > B are **111**, and **all** of them follows A > D > C

miRNAs which B > A are **181** and **174** follows B > C > D

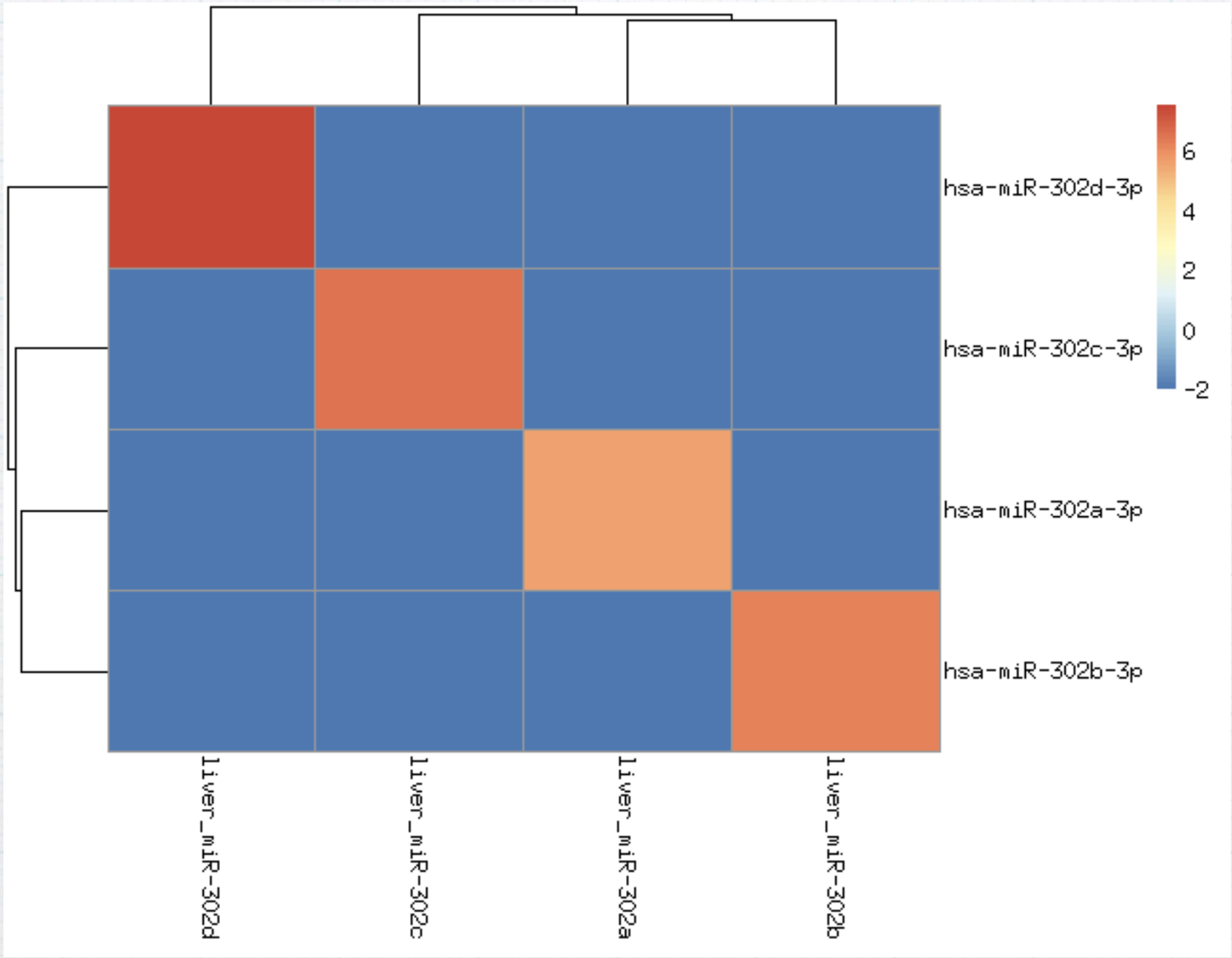
others' quantification

expression > 5 counts in average

upper quantile normalization

* clusters which A > B are **147**, where **139** (75 are known miRNAs) follow D > C
* clusters which B > A are **230**, where **222** (129 are known miRNAs) follow D > C

Specificity

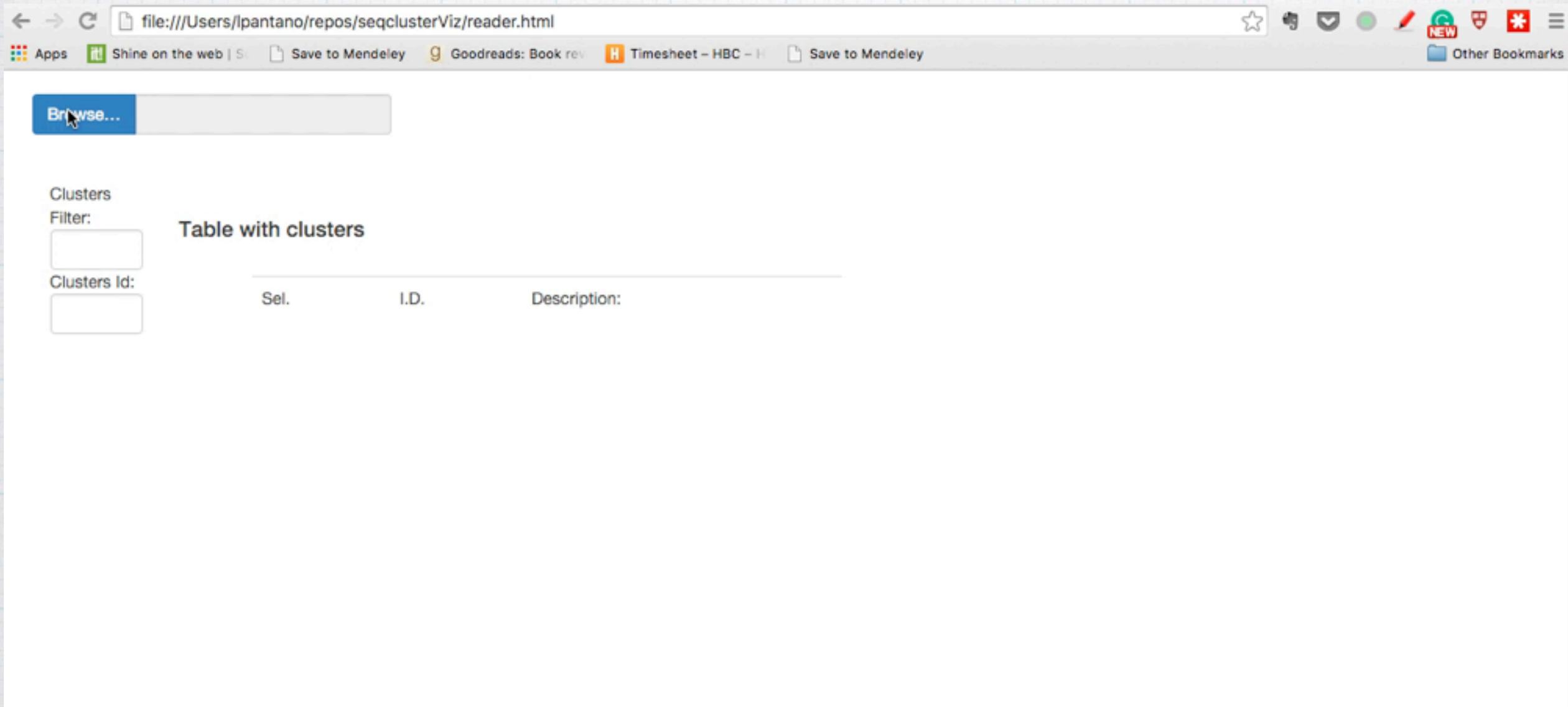


Resources

Total	3:19	total cores	total memory GB
organize samples	0	1	1
trimming & miRNA	0:21	8	20
prepare	0:01	1	8
alignment	0:07	6	42.1
cluster	2:49	1	8
quality control	0:01	8	20
report	0	1	1

The time for 8 samples with 6 millions reads each was 3 hours and 19 minutes.

visualization



https://raw.githubusercontent.com/lpantano/seqcluster/master/doc/slides/seqclusterViz.gif

open project for small RNA annotation and analysis

standard formats

naming rules

best-practices

miRNAs, tRNAs ...

thanks

* **Harvard T.H. Chan School of Public Health** for supporting the integration of small RNAseq pipeline in bcbio. Special thanks to **@roryk** and **@chapmanb**.

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

* Special thanks to the author of that papers to make data available. I encourage to use this data for any tool that analyzes small RNA data.