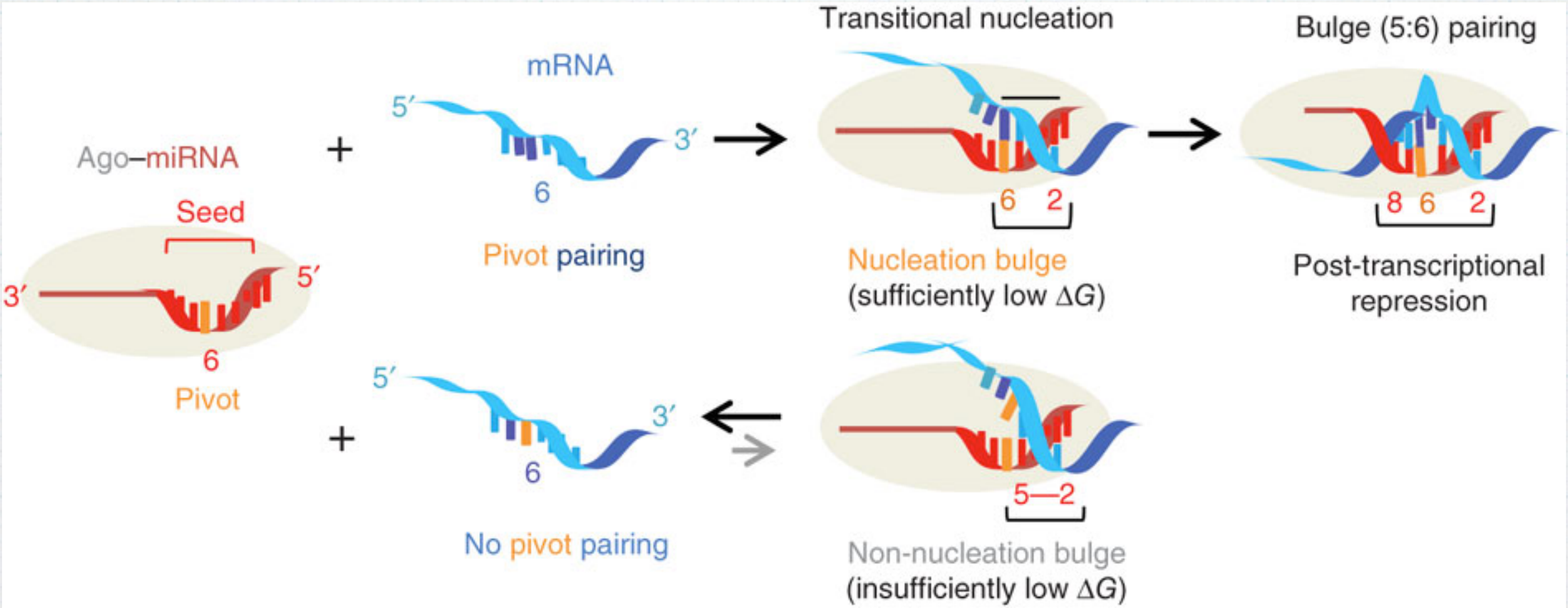
Characterization of the small RNA transcriptome using the bcbio-nextgen python framework

Lorena Pantano @lopantano
Harvard TH Chan School of Public Health

2016-07-09

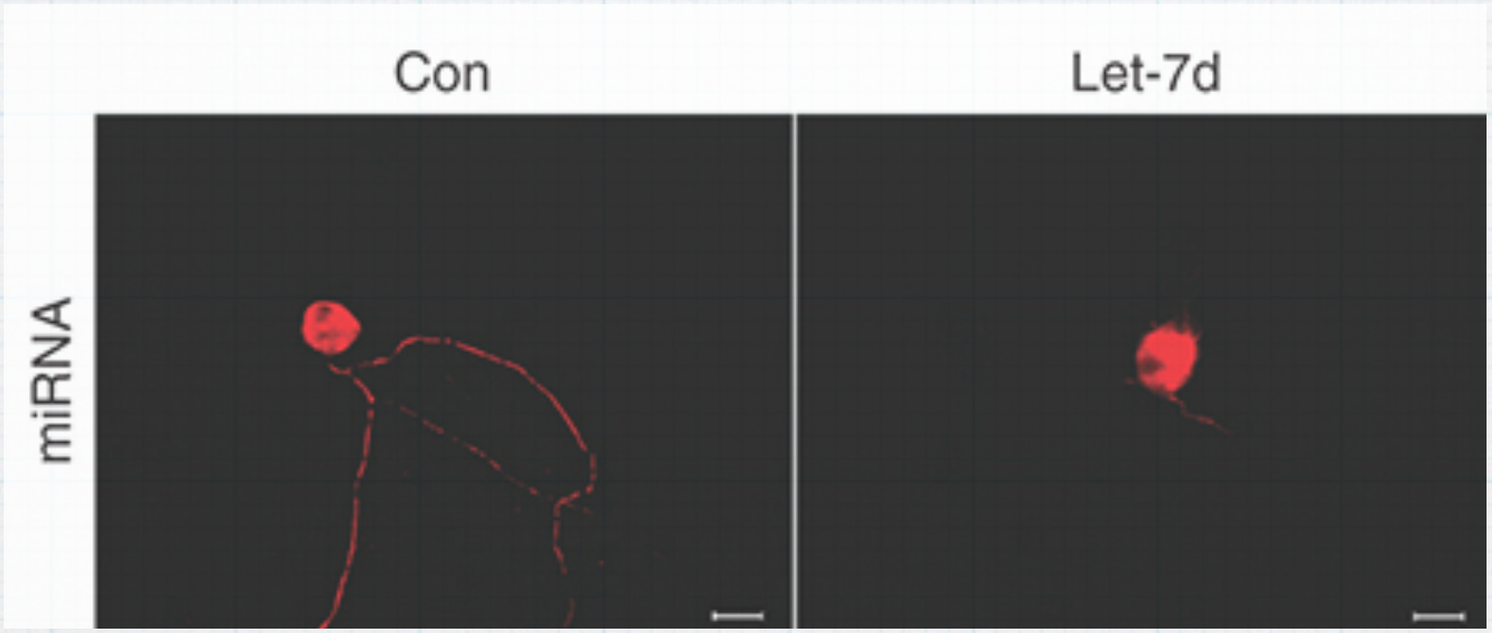
small RNA

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

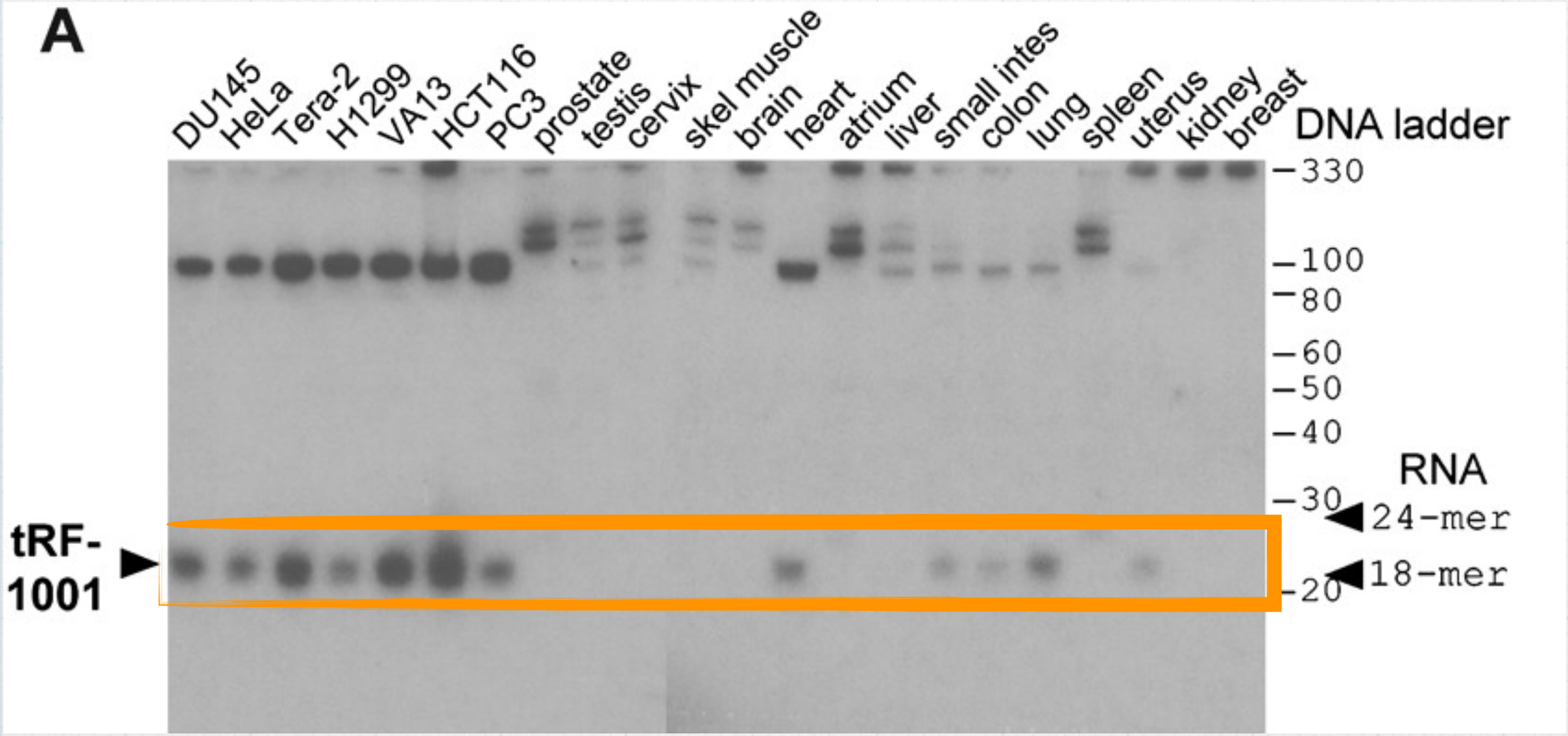
isomiRs

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

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

precursor

small tRNAs

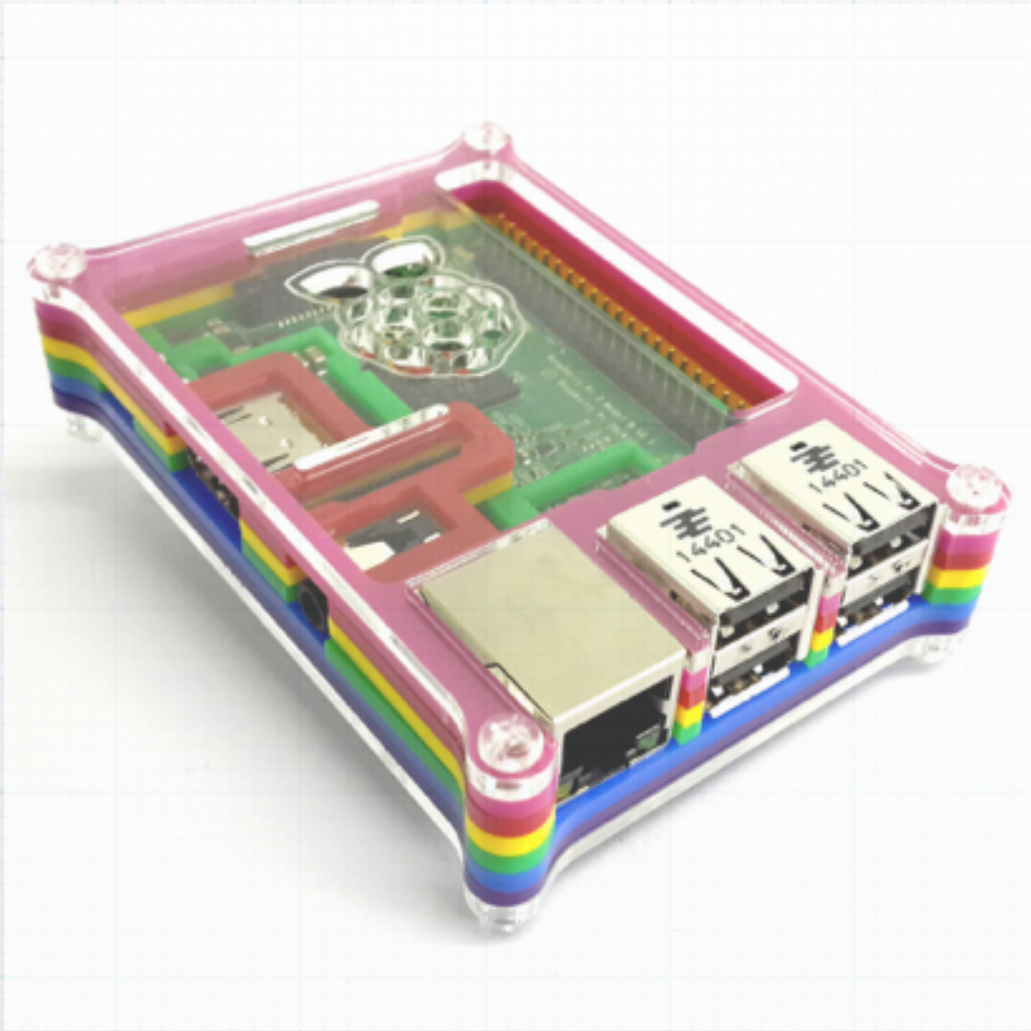


Yong Sun Lee et al. Genes Dev. 2009;23:2639-2649

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

detection & annotation

miraligner
tdrmapper

de-novo

seqcluster
mirdeep2 for mirna
protac for pirna (next)

detection & annotation

trimmed and collapsed reads

miraligner (miRBase)

tdrMapper (tRNA reference)

isomiRs package

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

seqcluster

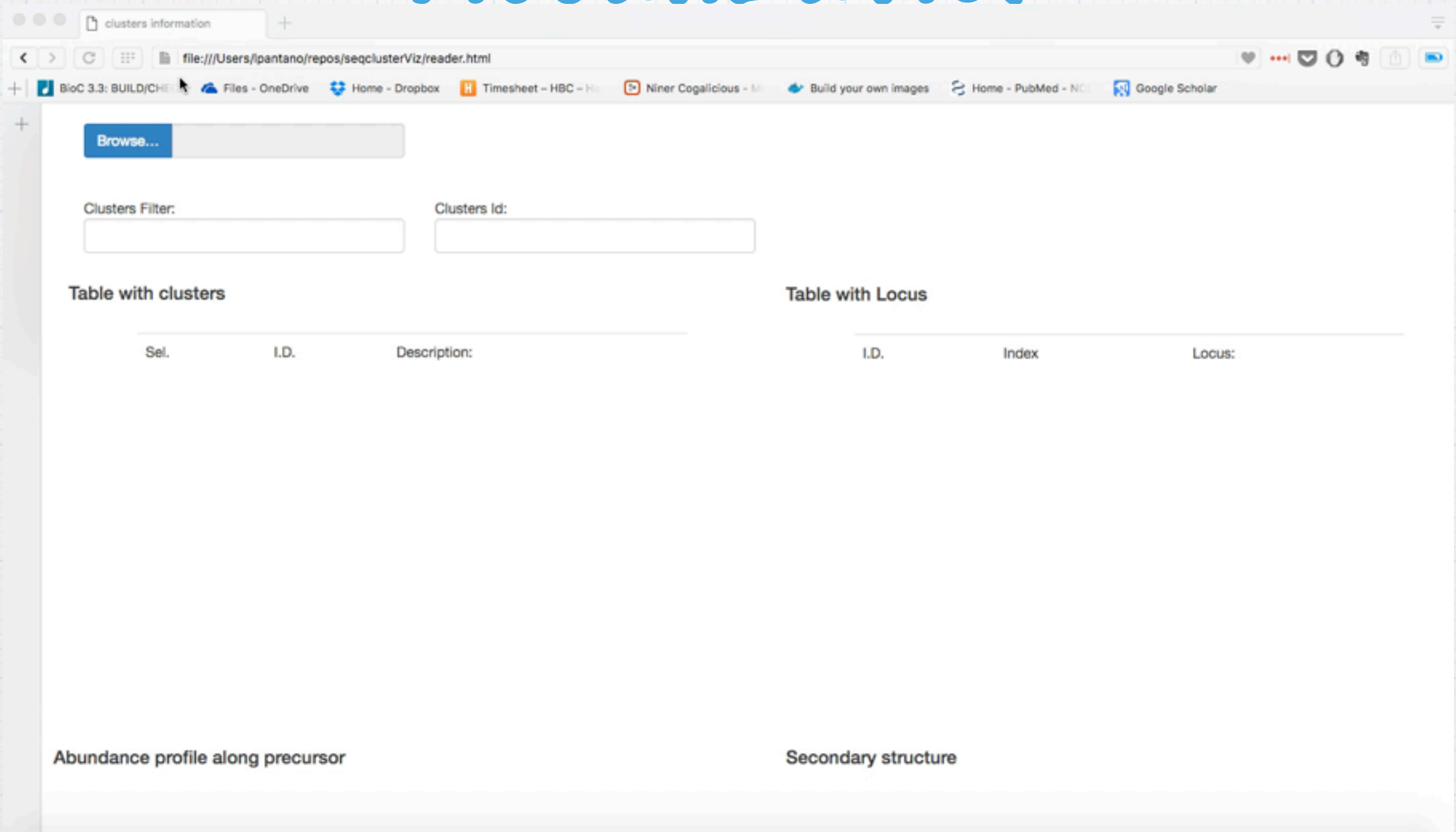
cluster at position 1

cluster at position 2

meta-cluster

seqcluster deals with multi-mapped reads

visualization



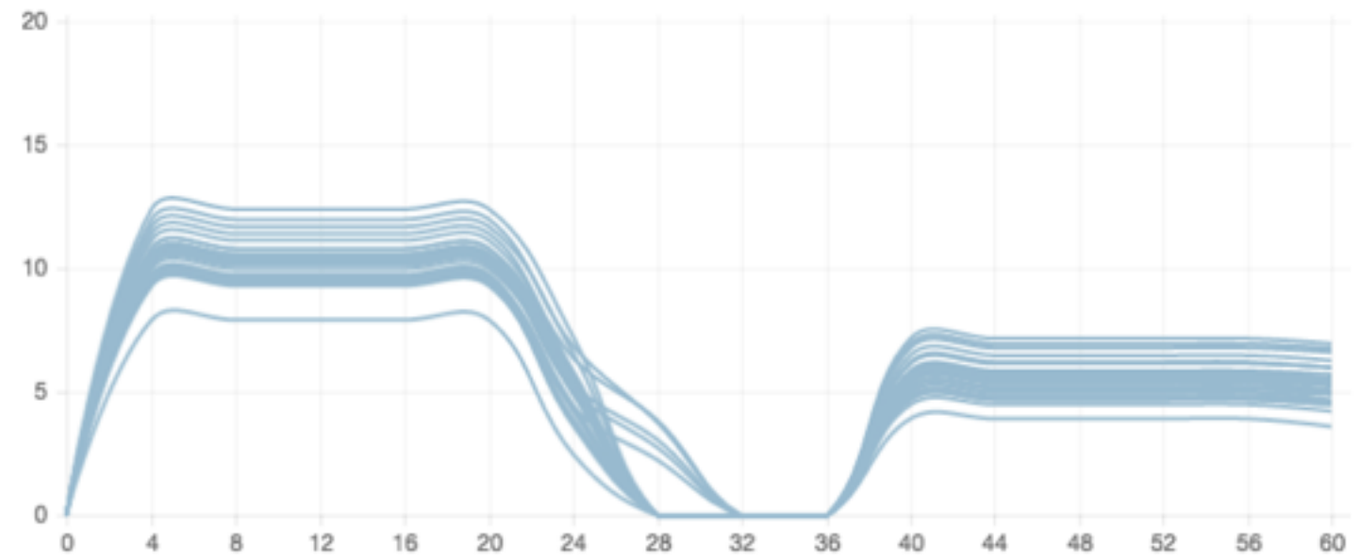
https://github.com/lpantano/seqclusterViz

visualization

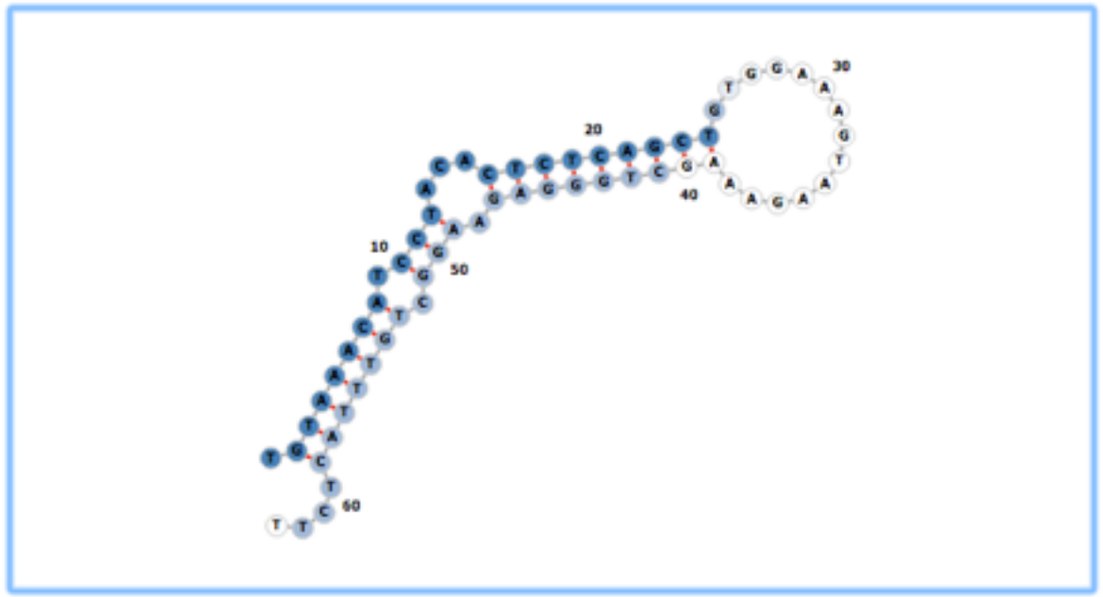
https://github.com/lpantano/seqclusterViz

visualization

Abundance profile along precursor



Secondary structure



Add filter

Select samples to change color.

Table with Sequences

Sequence filter:

Click on columns to sort sequences.

I.D	Index	Seqs.	cc1	cc2	cc3	cc4	cc5	cc6	cc7	ct1	ct2	ct3	ct4	ct5	ct6	ct7	pc1	pc2	pc3	pc4
1	0	TGTAAACATCCTACACTCTTAGCT	61	0	0	0	0	0	0	0	7	22	0	20	8	0	81	0	20	36
1	1	TGTAAACATCCTACACTCTCAGT	272	62	65	226	118	54	114	0	25	74	60	77	46	35	343	61	63	272

https://github.com/lpantano/seqclusterViz

MultiQC

Phil Ewels

ewels

Bioinformatician working with next generation sequencing data.

- Science for Life Laboratory
- Stockholm, Sweden
- phil.ewels@scilifelab.se
- http://phil.ewels.co.uk
- Joined on Nov 3, 2010

48	21	23
Followers	Starred	Following

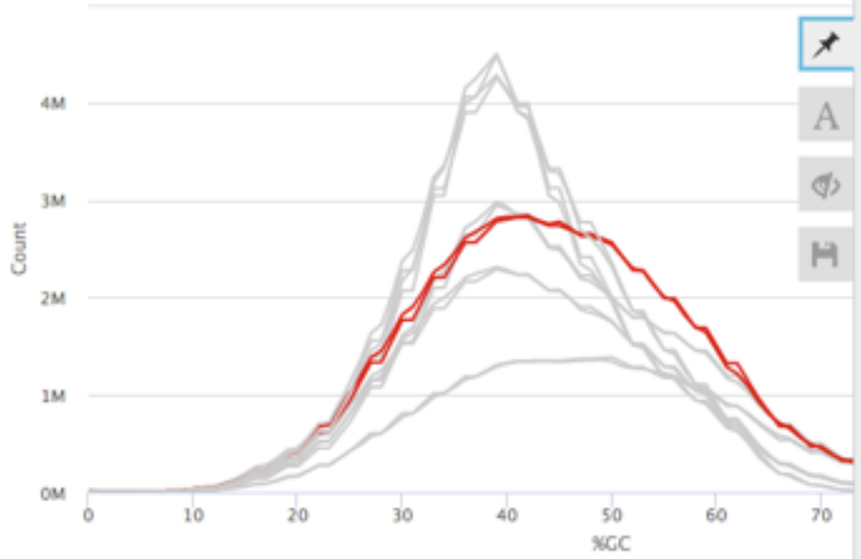
STAR: % Uniquely mapped reads

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

equence GC Content

ge GC content of reads. Normal random library typically have a roughly normal di
C help.

Per Sequence GC Content



MultiQC Toolbox

Highlight Samples

510

Regex mode off

514

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.*

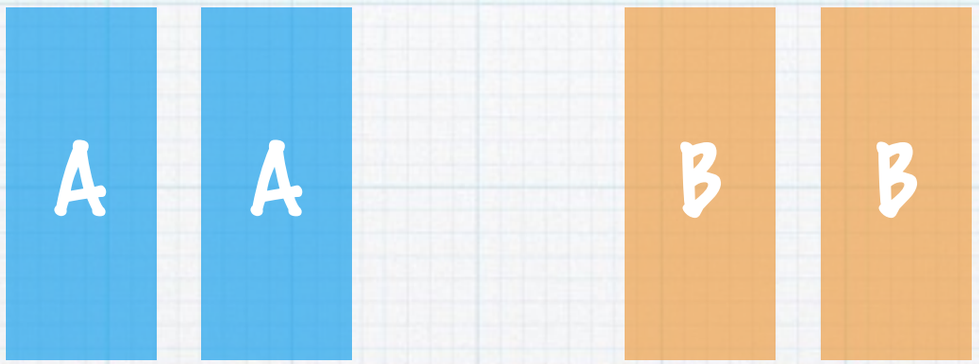
Affiliations | **Contributions** | **Corresponding author**

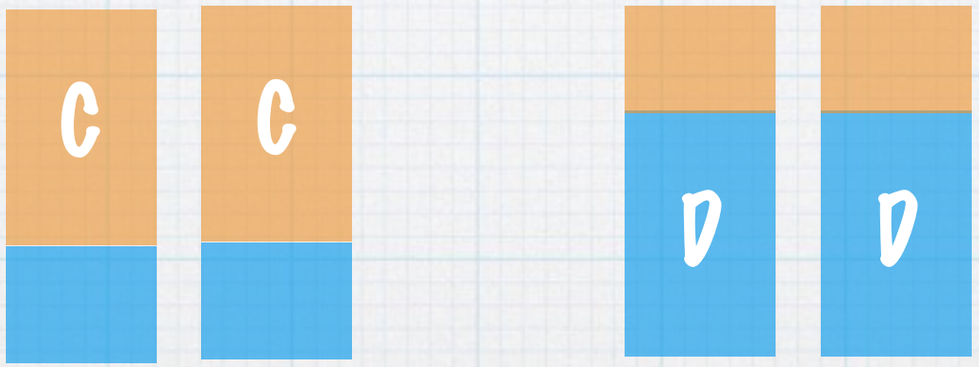
Nature Methods **11**, 809–815 (2014) | doi:10.1038/nmeth.3014

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

| Corrected online **30 July 2014**

Quality Control samples





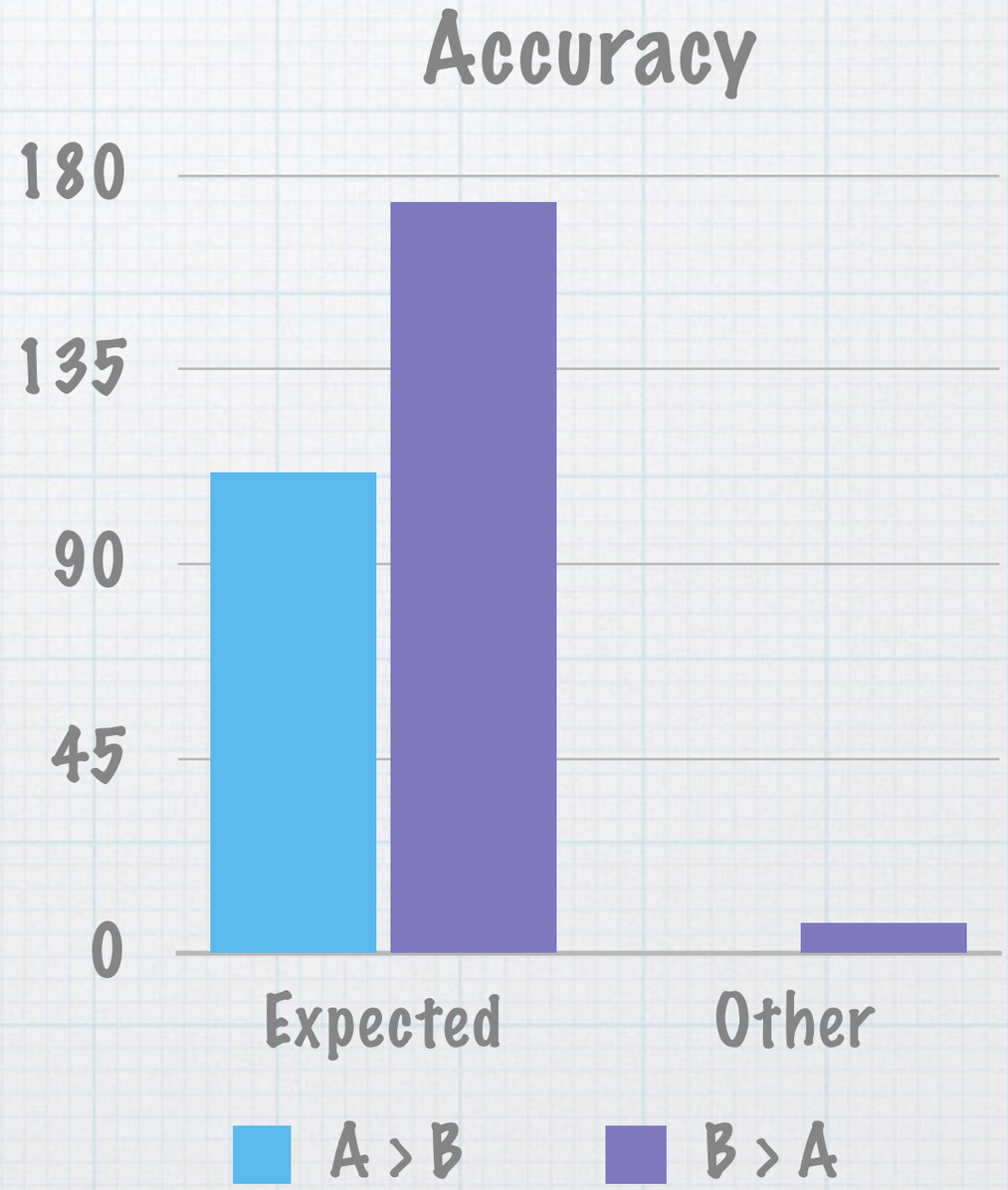
For each molecule:

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

miRNA quantification

miRNAs > 5 counts in average

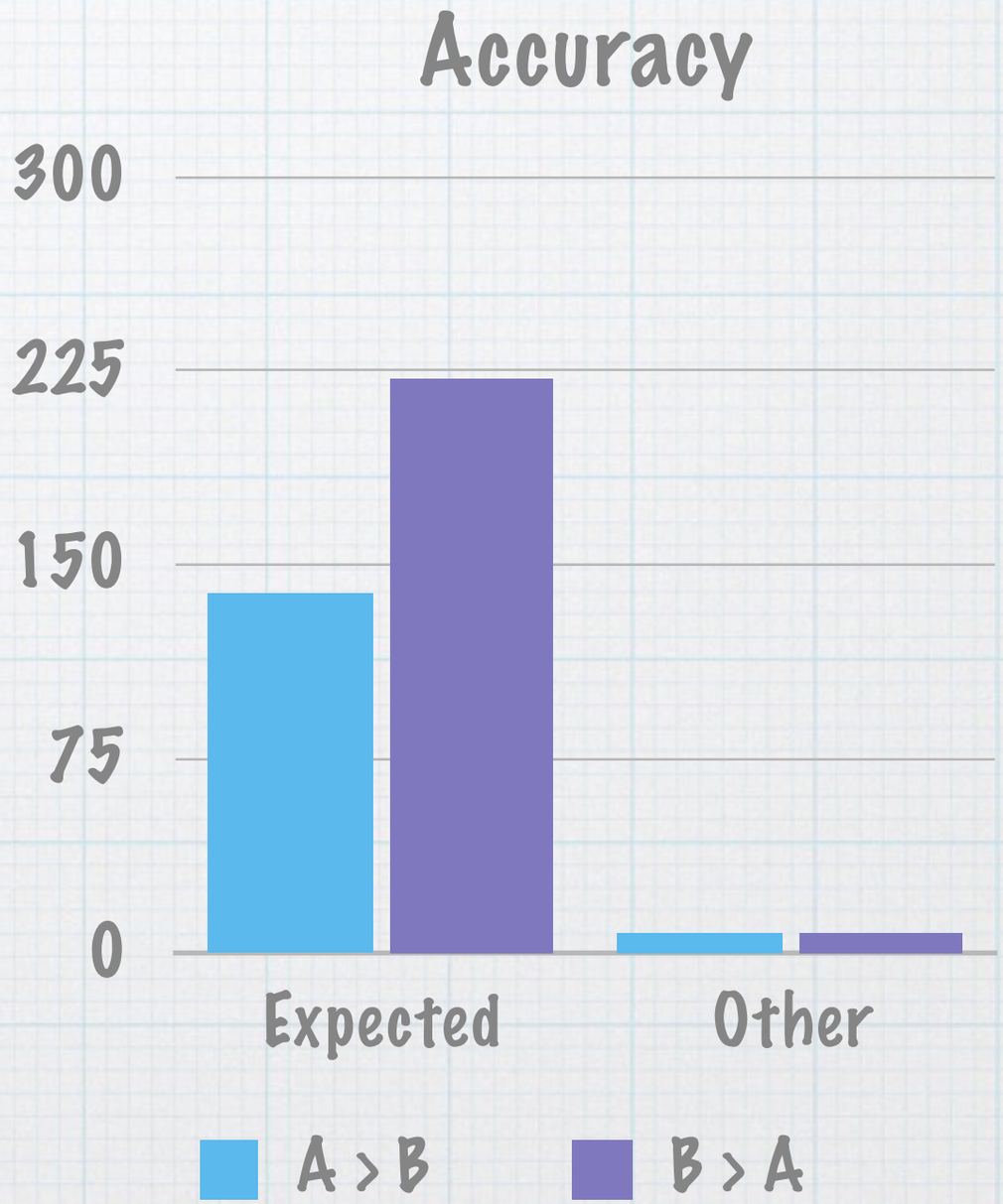
upper quantile normalization



clusters quantification

expression > 5 counts in average

upper quantile normalization

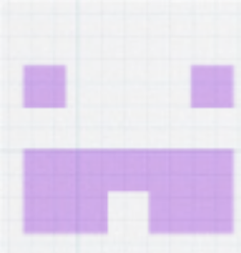


Resources

	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

standard formats
naming rules

best-practices

mirtop

Python ★ 0 | 0

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

Updated 3 days ago

miRNAs, tRNAs ...

miRTOP.github.io

CSS ★ 0 | 0

project for small RNA standard annotations

Updated on Mar 29

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.