miRNA and isomiR annotation

Lorena Pantano
HSPH

Dec-15-2016

Agenda

- miRNA and isomiR definition
- small RNA-seq protocols comparison
- miRNA mapping comparison
- isomiR annotation from BAM files
- isomiR analysis in R

miRNA

RNA molecules of 18-36 nts long with regulation function

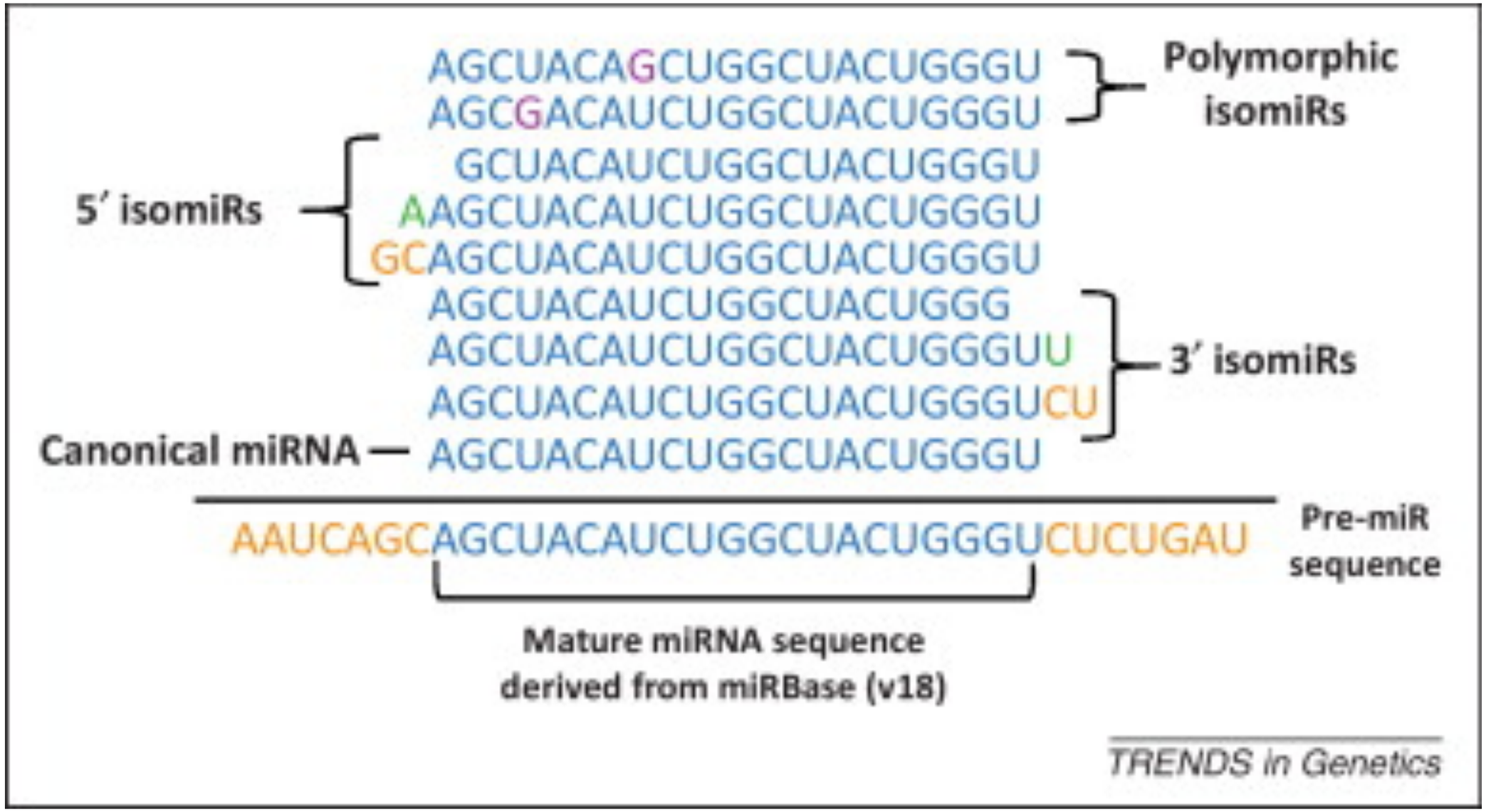
isomiRs

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

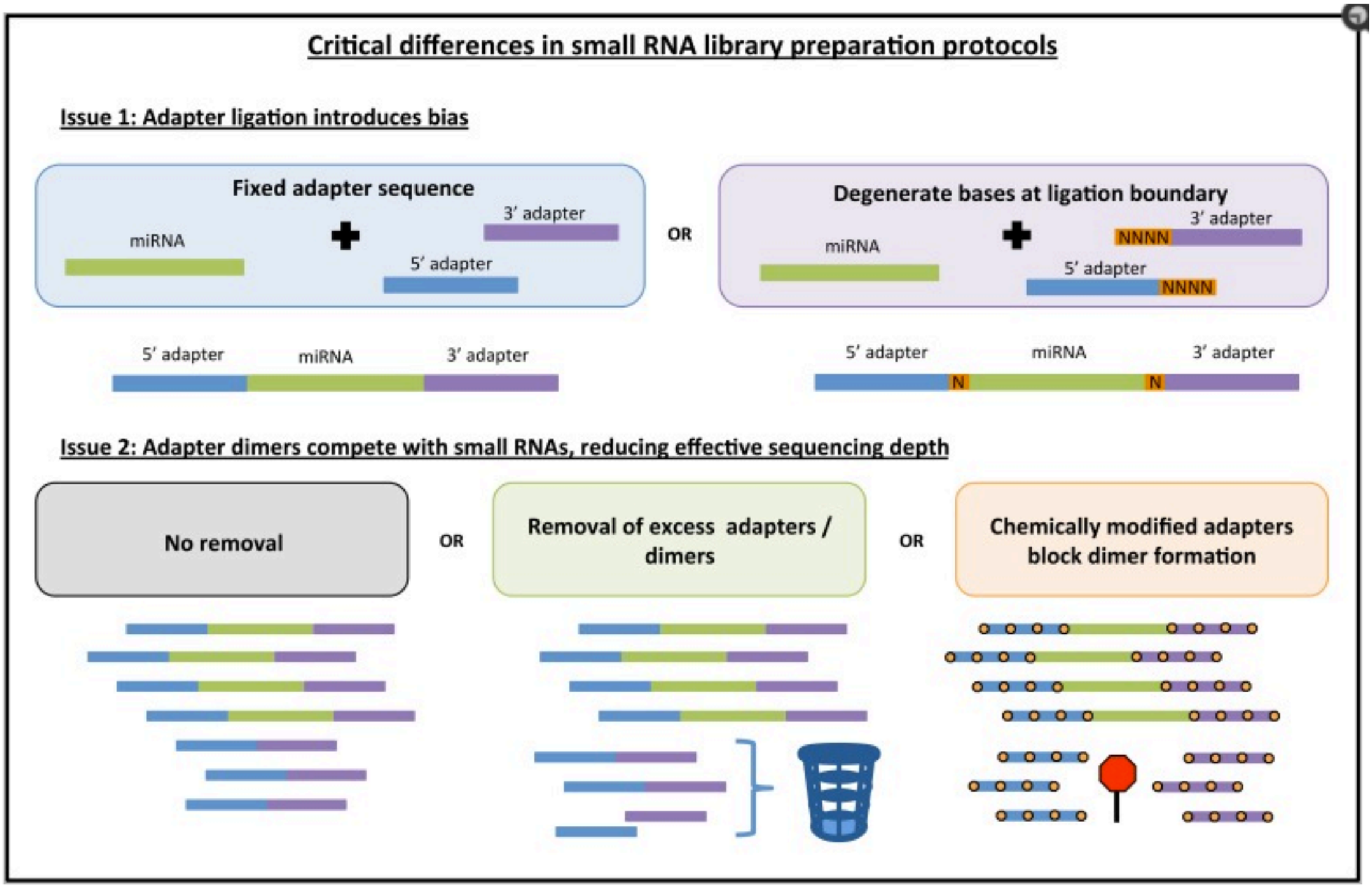
```
CUCCGGUGCCUACUGAGCUGAUAUCAGUUCUCAUUUUACACACUGGCUCAGUUCAGCAGGAACAGGAG
((((.((.(((.(((((((((..(((((.............)))))..))))))))).))).)).)))) (-26.32)
```

precursor

Types of isomiRs

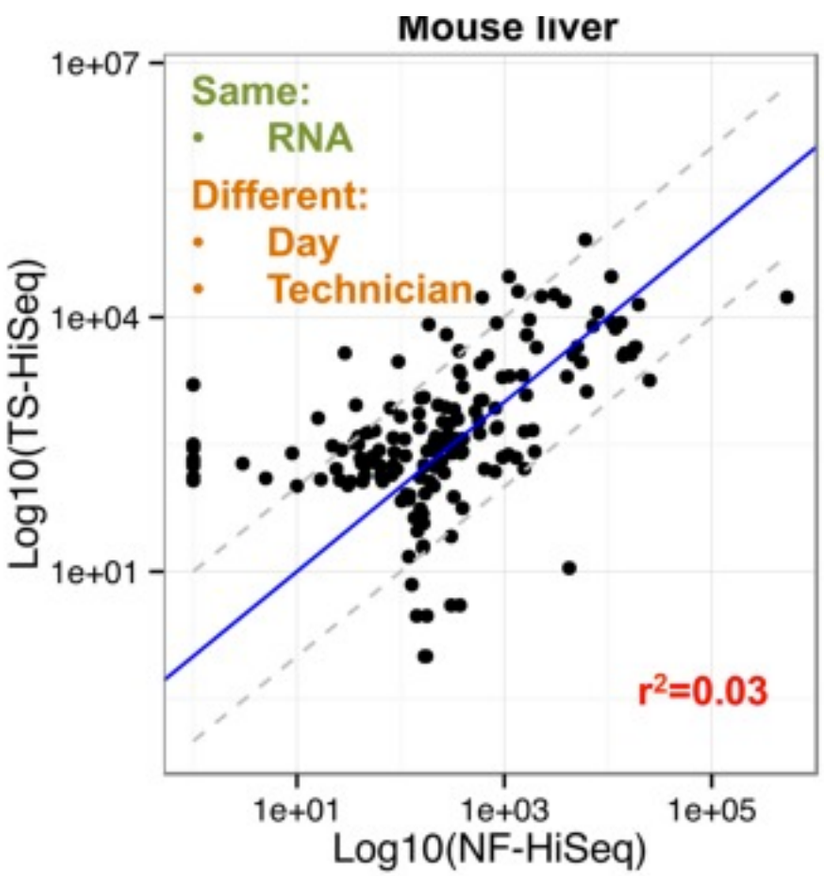


Protocols


Critical differences in small RNA library preparation protocols
Issue 1: Adapter ligation introduces bias
Fixed adapter sequence
miRNA
3' adapter
5' adapter
OR
Degenerate bases at ligation boundary
miRNA
3' adapter
NNNN
5' adapter
NNNN
5' adapter
miRNA
3' adapter
5' adapter
miRNA
3' adapter
N
N
Issue 2: Adapter dimers compete with small RNAs, reducing effective sequencing depth
No removal
OR
Removal of excess adapters / dimers
OR
Chemically modified adapters block dimer formation

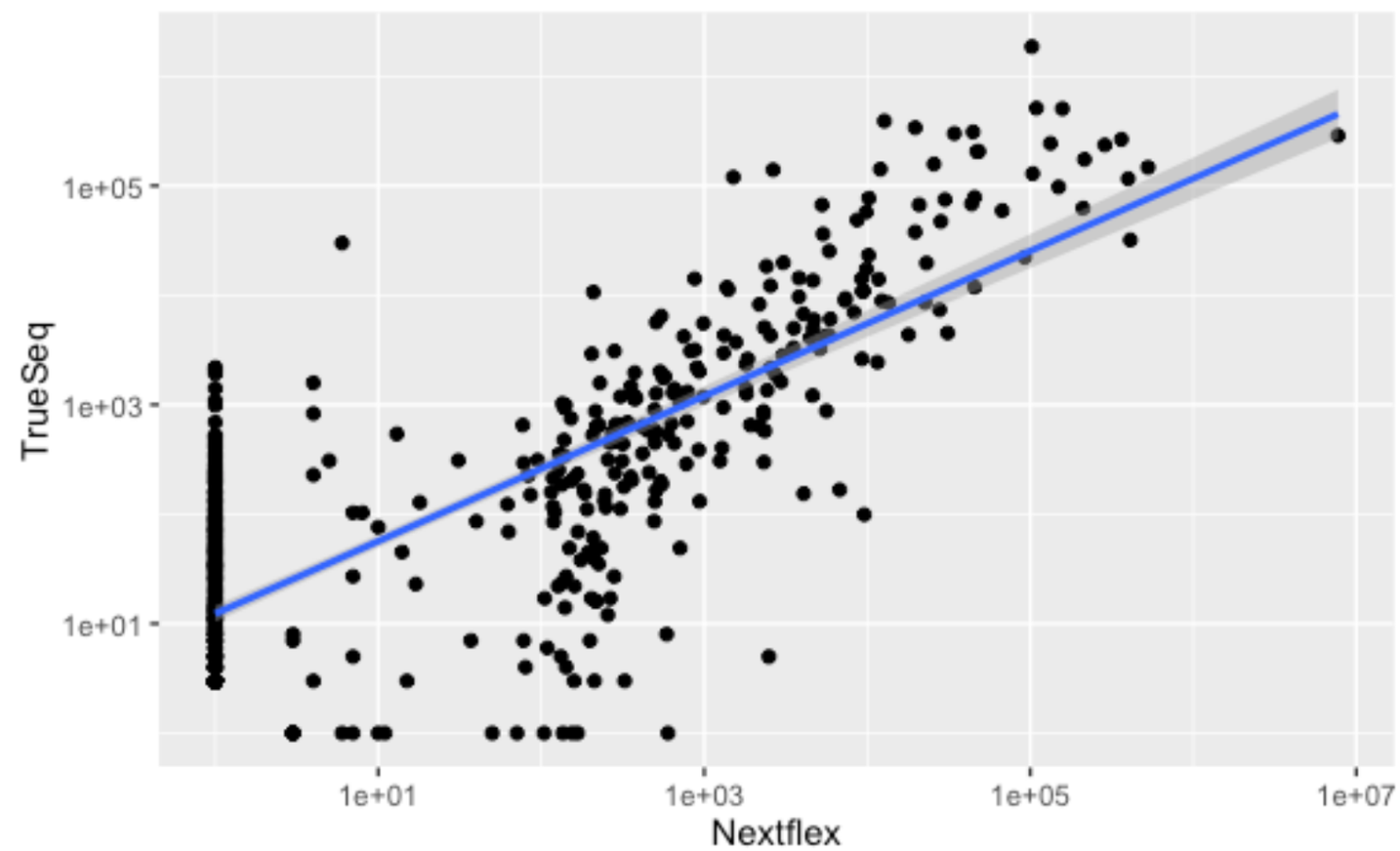
Protocol comparison

Paper figure

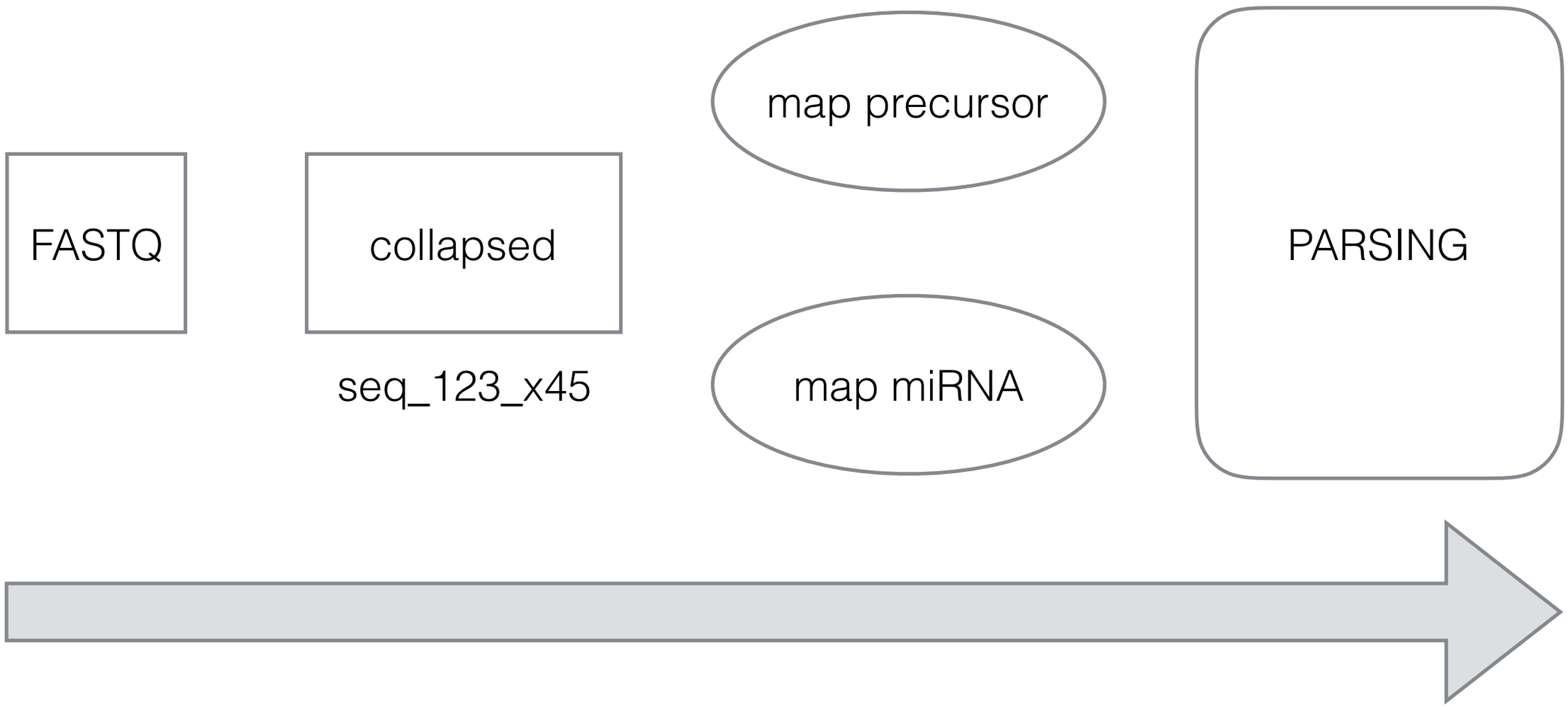


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4686641/

bcbio pipeline



miRNA mapping



Benchmark

- simulation of miRNAs/isomiRs (~ 16000)
- mapping with different tools
- compare miRNA detection and accuracy

tools compared

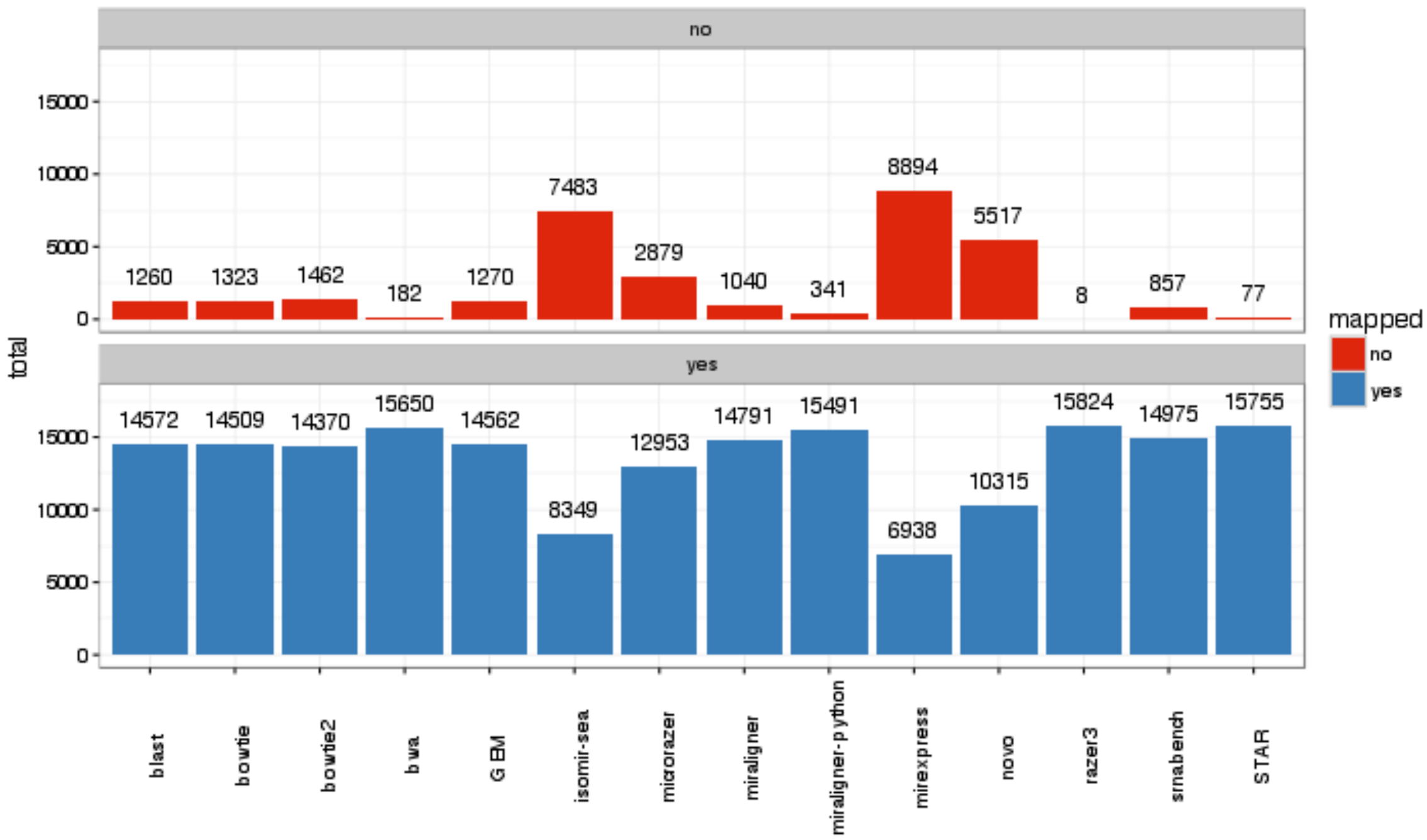
general mappers

bowtie, bowtie2, blast, GEM, microzer, novoaling, razer3, STAR, megablast,

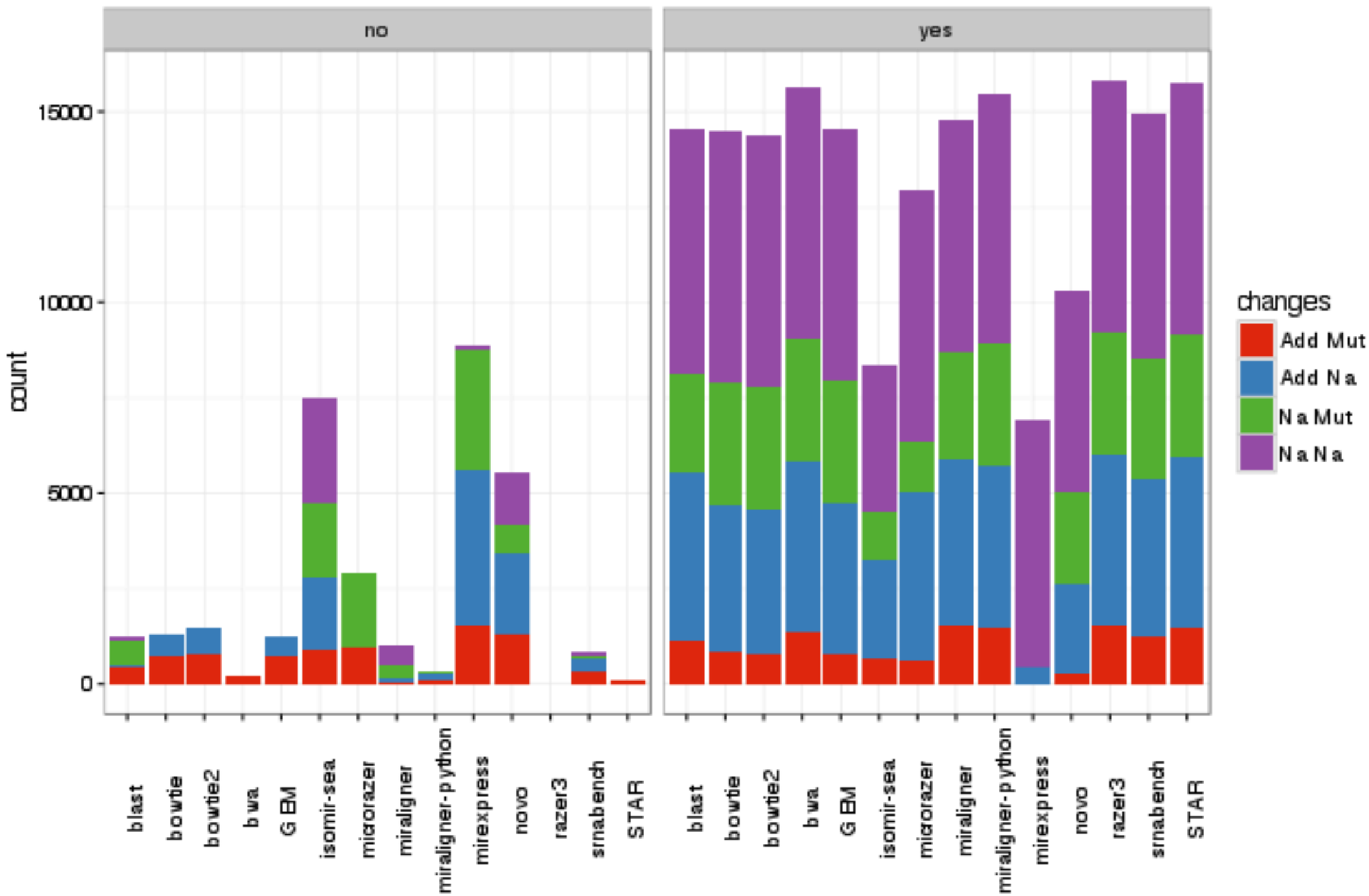
miRNA specialized mappers

miraligner, miraligner-python, srnabench, mirexpress

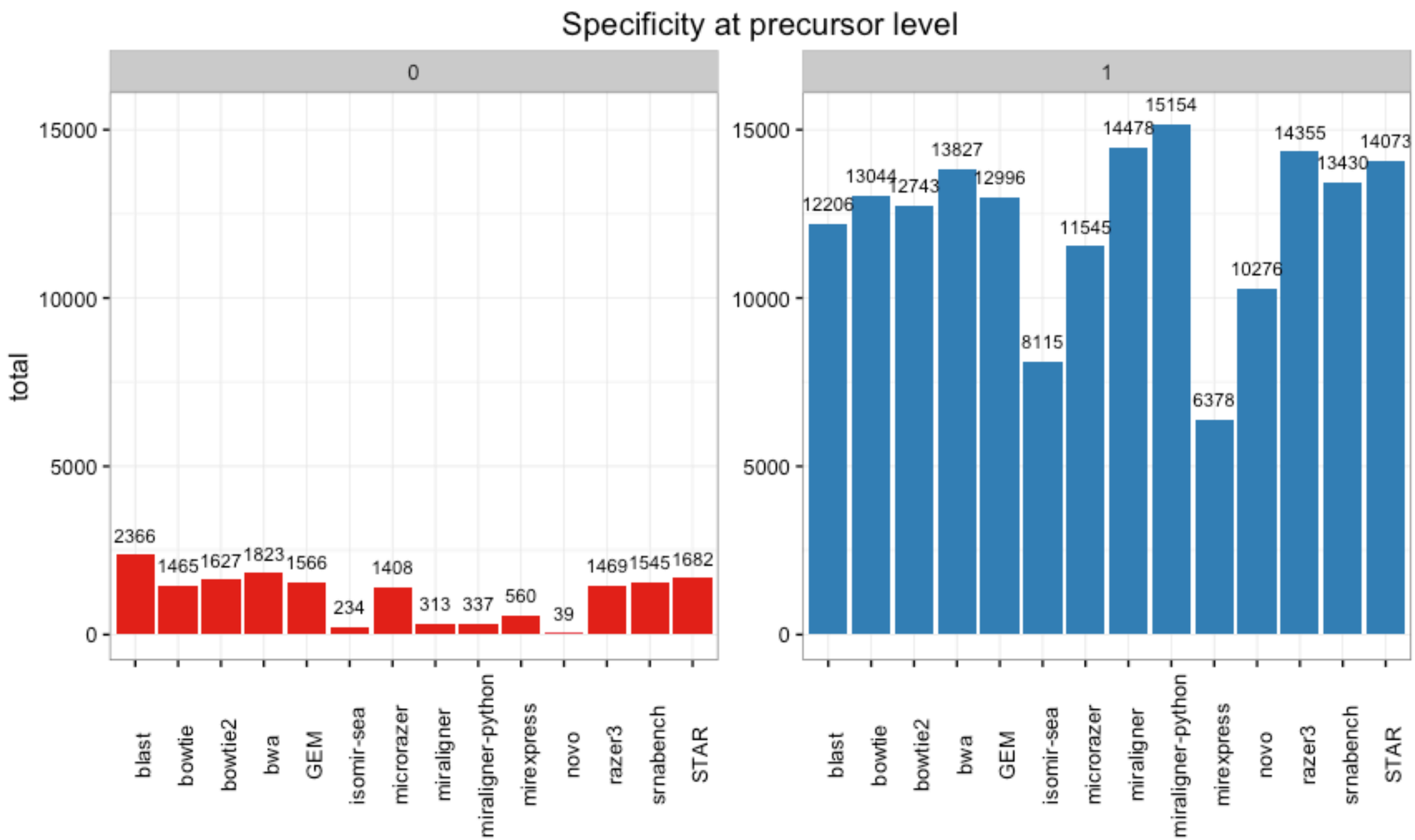
miRNA detection



Cause of missing



miRNA accuracy



isomiR annotation

miRNA in database

isomiR

UPPER CASE: addition

lower cases: deletion

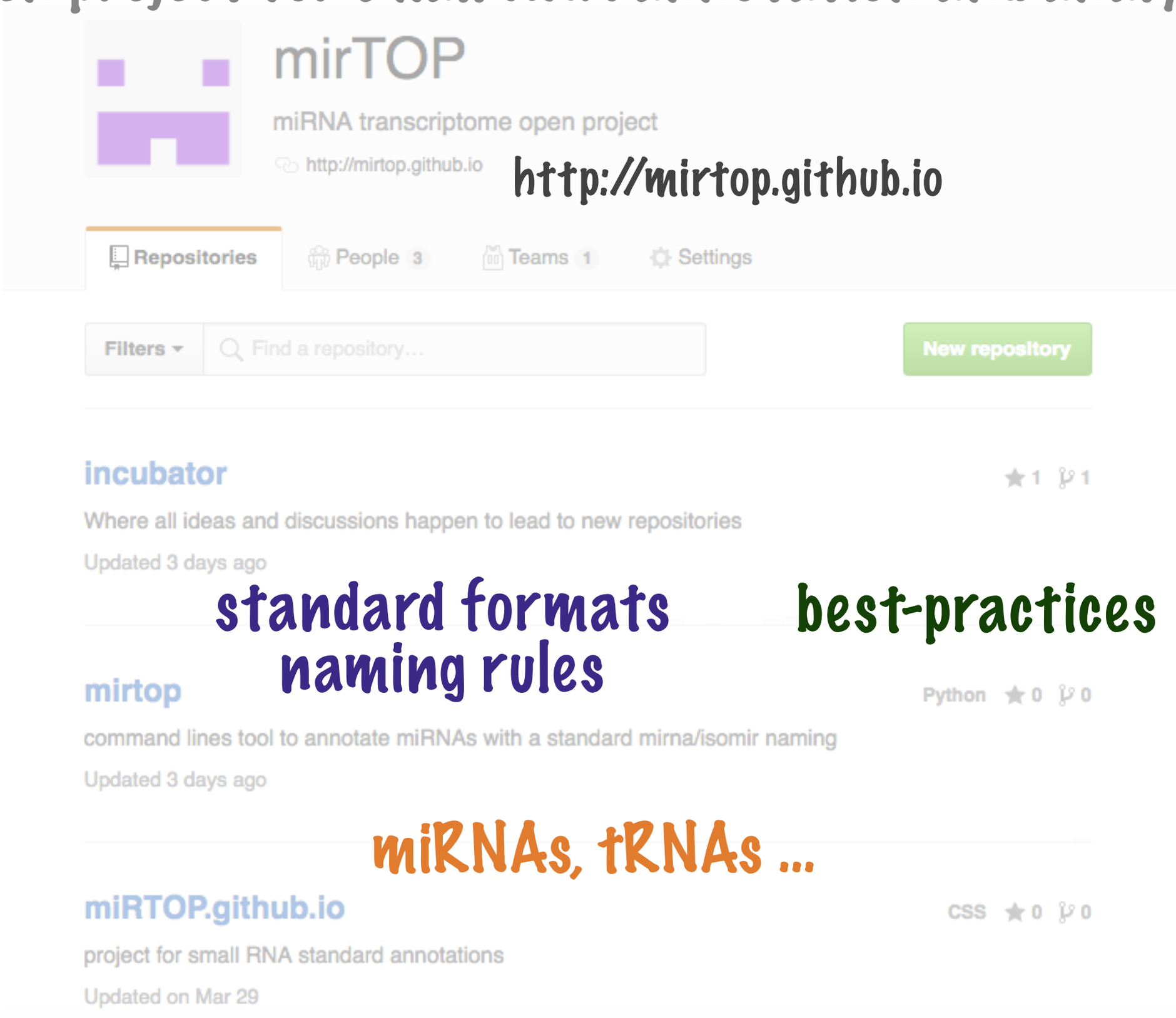
mismatch addition

trimming 5’ trimming 3’

miRNA_name:mismatch:addition:t5:t3

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

open project for small RNA annotation and analysis



http://mirtop.github.io

standard formats
naming rules

best-practices

miRNAs, tRNAs ...

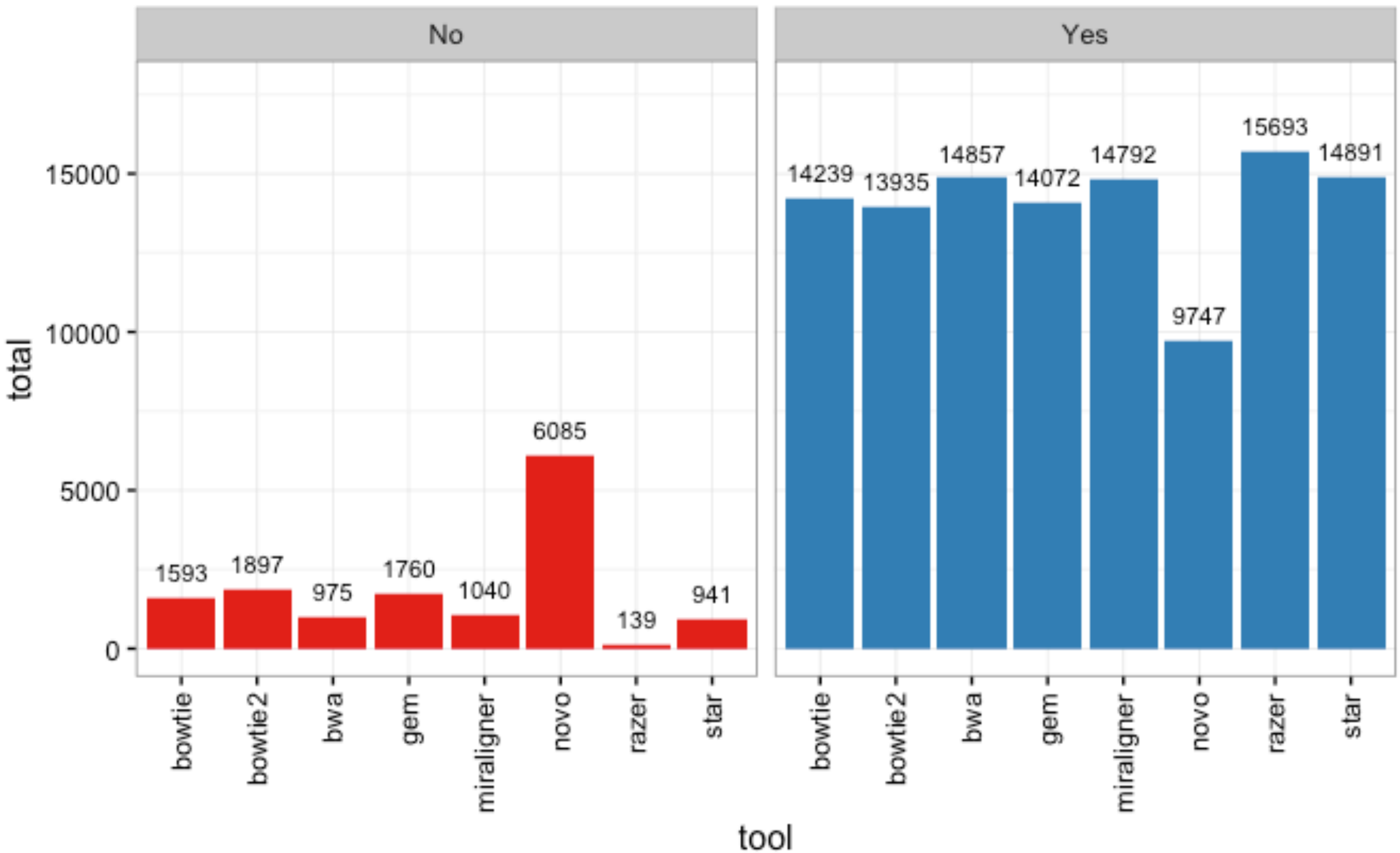
tools compared

bowtie, bowtie2, GEM, miraligner, novoaling, razer3, STAR

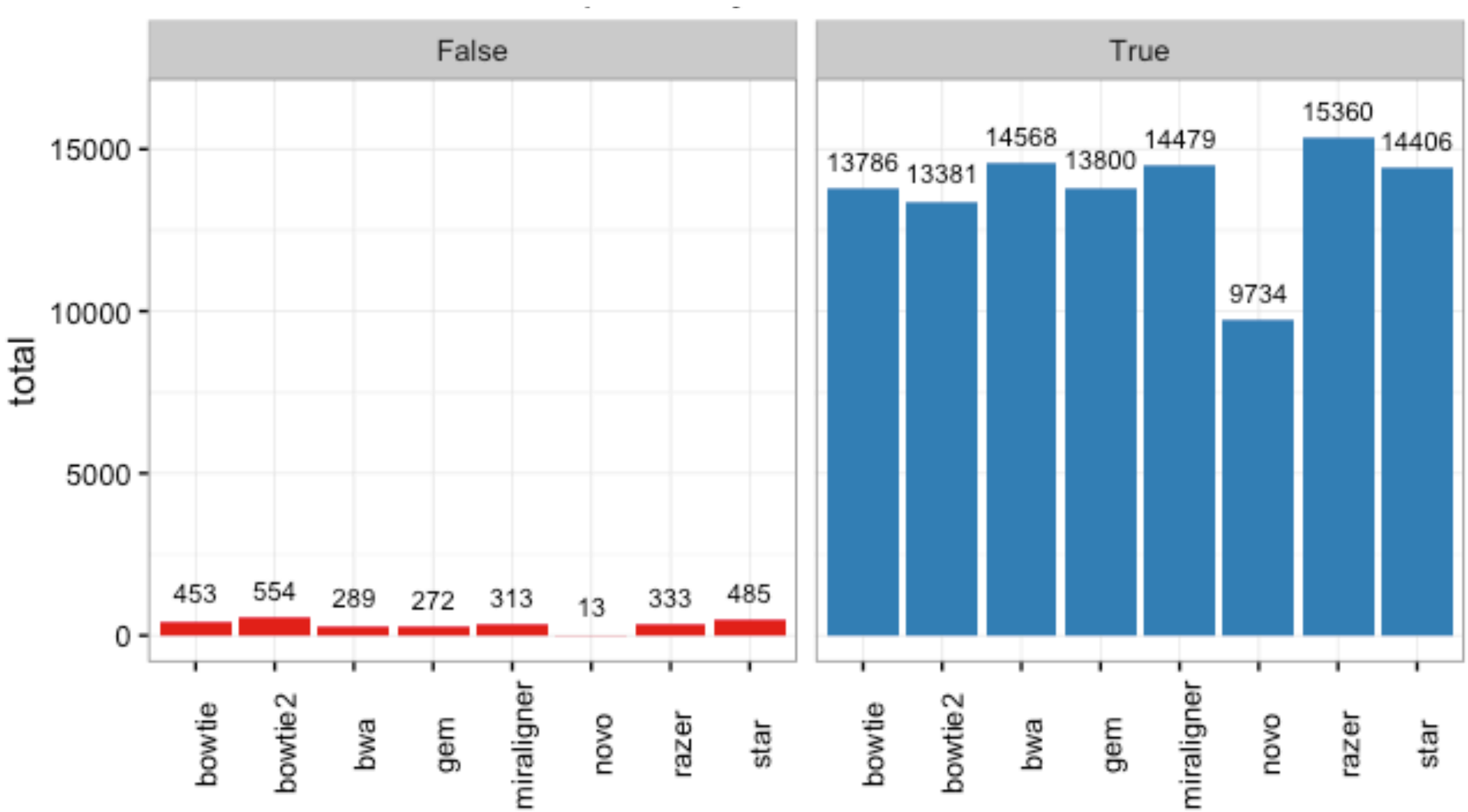
```
mirtop annotate --sps hsa
--hairpin ../hairpin.hsa.fa
--mirna ../miRNA.str
-o gem_out ../gem/sim.21.hsa.sam
```

You can input multi-bam files at the same time

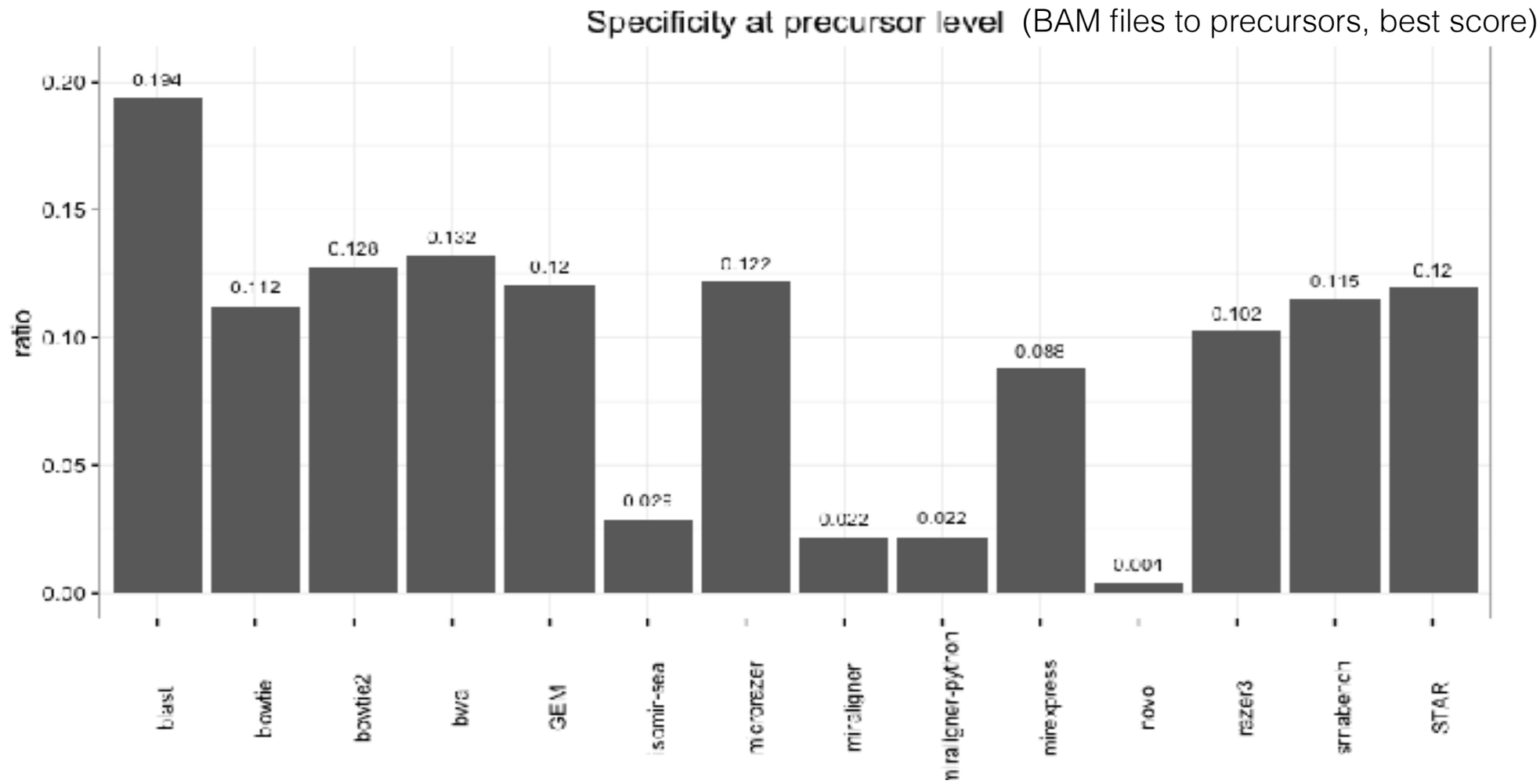
isomiR detection

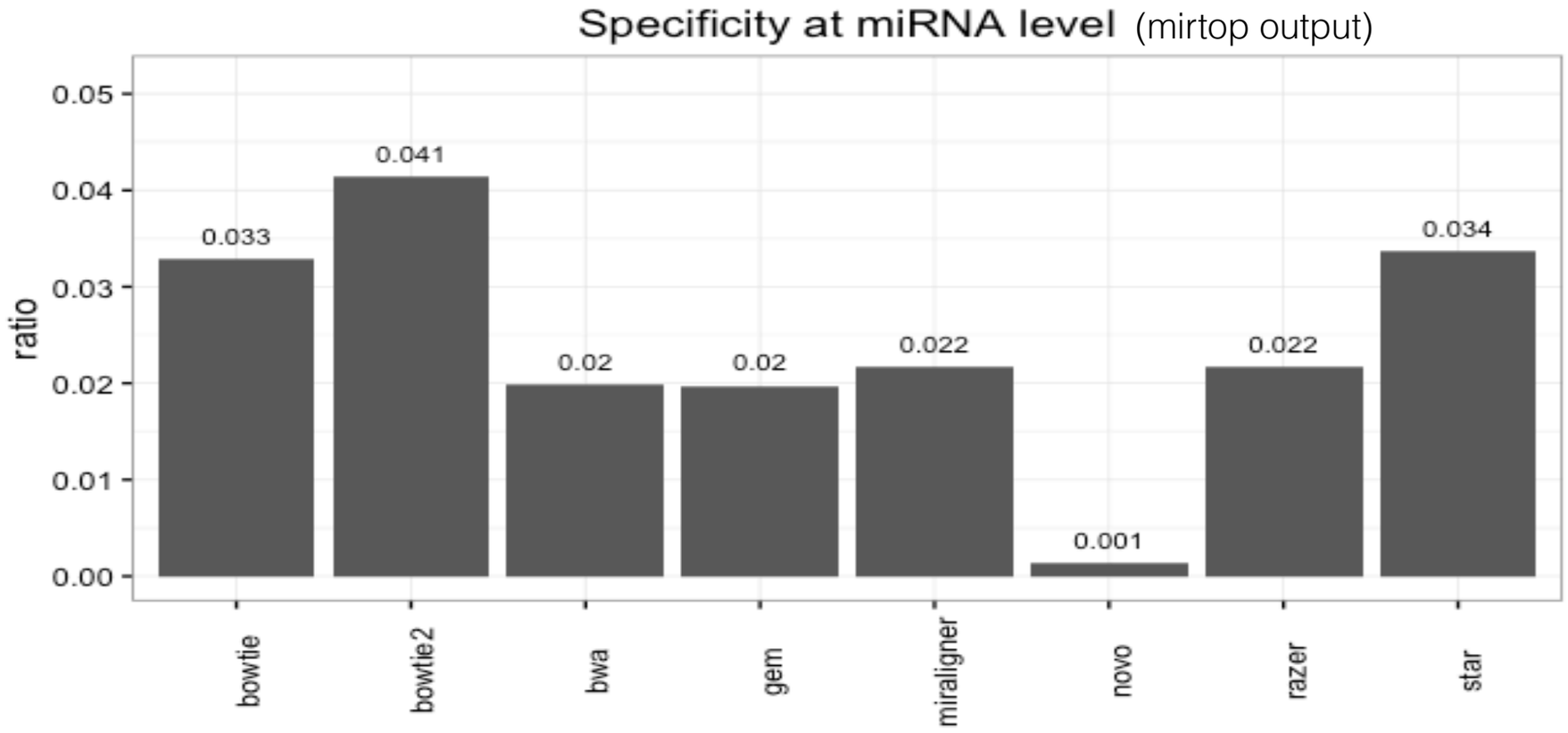


isomiR accuracy



FN/TP


Specificity at precursor level (BAM files to precursors, best score)
ratio
0.20
0.15
0.10
0.05
0.00
0.194
0.12
0.128
0.132
0.12
0.022
0.022
0.088
0.001
0.102
0.115
0.12
blast
bowtie
bowtie2
bwa
GEM
isomir-sea
microrezer
miraligner
miraligner-python
mirexpress
novo
razer3
STAR


Specificity at miRNA level (mirtop output)
ratio
0.05
0.04
0.03
0.02
0.01
0.00
0.033
0.041
0.02
0.02
0.022
0.001
0.022
0.034
bowtie
bowtie2
bwa
gem
miraligner
novo
razer
star

miRNA with R (isomiRs)

- what to consider as input for the DE tools
- isomiR characterization
- query the data
- supervised clustering with feature selection

Input

seq	name	freq	mir	start	end	mism	add	t5	t3	s5	s3	DB	precu
AGGTCGACCGTGTTATATTCG			seq_100056_x3		3	rno-miR-369-5p		14	34	3GA	0	0	c
TTGAAAGGCTGTTTCTTGGTT			seq_100058_x15		15	rno-miR-488-3p		49	68	0	T	0	c
TACTGCACTCGTCCCGGCCT			seq_100063_x3		3	rno-miR-92b-3p		52	71	3CT	0	0	cc
TTGAAAGGCTGTTTCTTGGTG			seq_100069_x33		33	rno-miR-488-3p		49	68	0	G	0	c
CTACTTCACAACACCAGGGTTA			seq_10011_x13		13	rno-miR-138-1-3p			64	83	0	TA	cgg
TGAGGTAGTAGTTTGTGCTGAT			seq_100122_x3		3	rno-let-7i-5p		6	25	0	AT	0	tt
TCTACAGTGCACGTGCCTCCA			seq_100131_x5		5	rno-miR-139-5p		7	27	16CT	0	0	g
ACGTCATCGTCGTCATCGTTA			seq_100132_x5		5	rno-miR-598-3p		49	69	0	0	t	0
TGTGACAGATTGATAACTGAAAG			seq_100147_x11		11	rno-miR-542-3p		49	71	0	0	0	G
CTGGCCCTCTCTGCCCTTCCGCAT			seq_100148_x9		9	rno-miR-328a-3p		48	68	0	CAT	0	
NGAATTGTGGCTGGACATCTGT			seq_100185_x4		4	rno-miR-219a-2-3p			62	83	1NA	0	0
GGAAGACTAGTGATTTTATTGT			seq_100227_x5		5	rno-miR-7a-5p		20	41	18AG	0	t	0
AACATTTATTGCTGTCGGTGGGT			seq_100277_x8		8	rno-miR-181b-5p		15	37	7TC	0	0	0

Processing annotation

```
<<package-plot-iso,message=FALSE,eval=FALSE>>=
ids <- IsomirDataSeqFromFiles(fn_list, design=de)
@
```

Order in fn_list should be the same than in the design data.frame

```
> fn_list
[1] "/Library/Frameworks/R.framework/Versions/3.3/Resources/library/isomiRs/extra/sample1.m
irna"
[2] "/Library/Frameworks/R.framework/Versions/3.3/Resources/library/isomiRs/extra/sample2.m
irna"
```

```
> de
   condition
f1   newborn
f2   newborn
```

count matrix

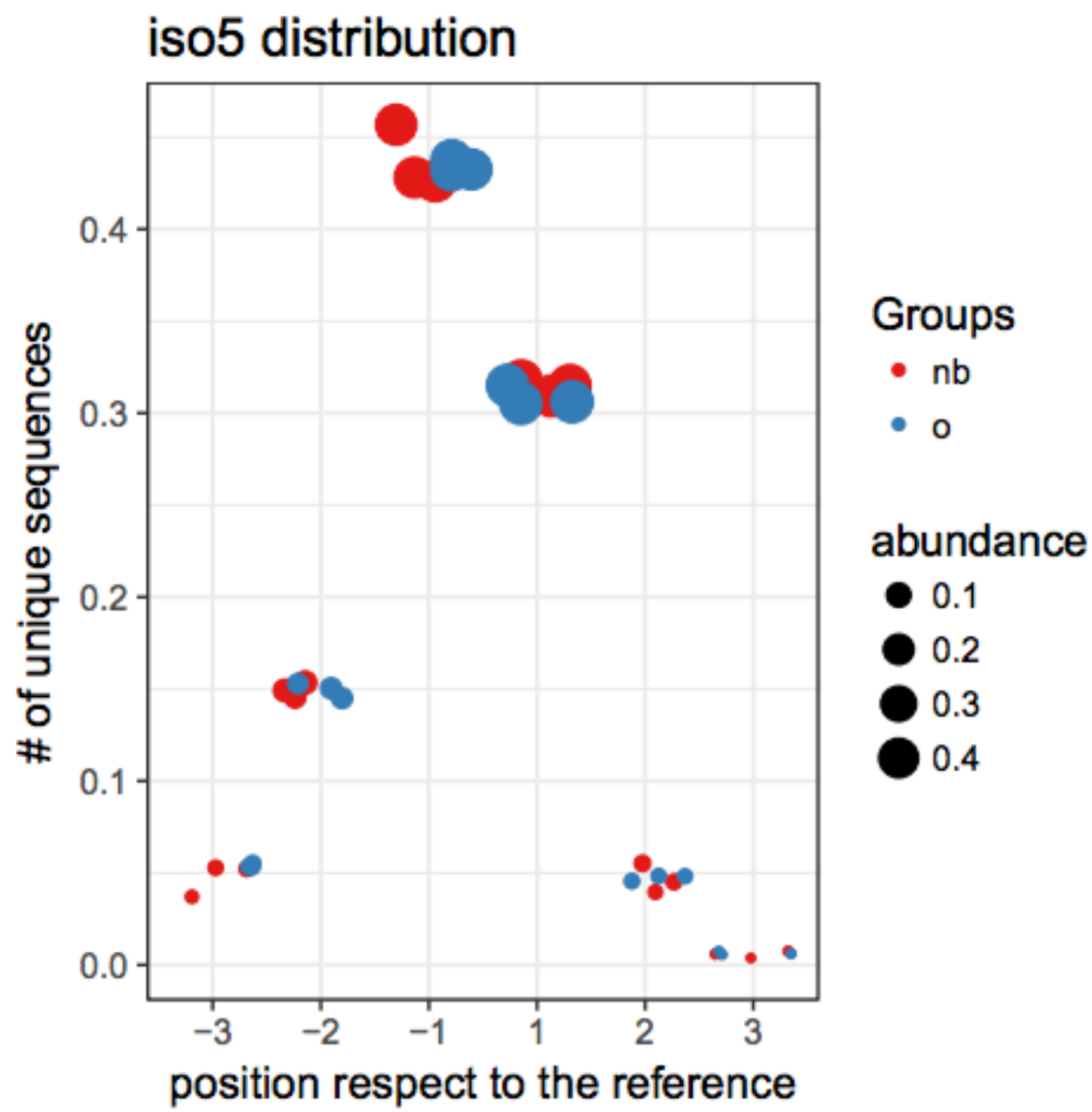
```
> head(counts(ids))
                 nb1    nb2    nb3     o1     o2     o3
hsa-let-7a-3p     24     70     23     47     26     65
hsa-let-7a-5p 427615 544663 427219 556660 325845 625602
hsa-let-7b-3p     12     38     17     24     27     33
hsa-let-7b-5p 109767 188394 125986 150227 104593 160253
hsa-let-7c-3p      0      1      2      0      0      3
hsa-let-7c-5p 481931 462630 363116 425470 272375 434007
```

```
> head(counts(isoCounts(ids, iso5 = T, add = T)))
                              nb1 nb2 nb3 o1 o2 o3
hsa-let-7a-3p.t5:0.ad:0     6  17   5 16  3 13
hsa-let-7a-3p.t5:0.ad:A     0   2   1  1  0  3
hsa-let-7a-3p.t5:0.ad:T    11  24   9 13  9 21
hsa-let-7a-3p.t5:0.ad:TT    0   2   0  0  0  1
hsa-let-7a-3p.t5:c.ad:0     1   8   4  7  4  5
hsa-let-7a-3p.t5:c.ad:A     0   4   1  3  0  3
```

isomiR figures

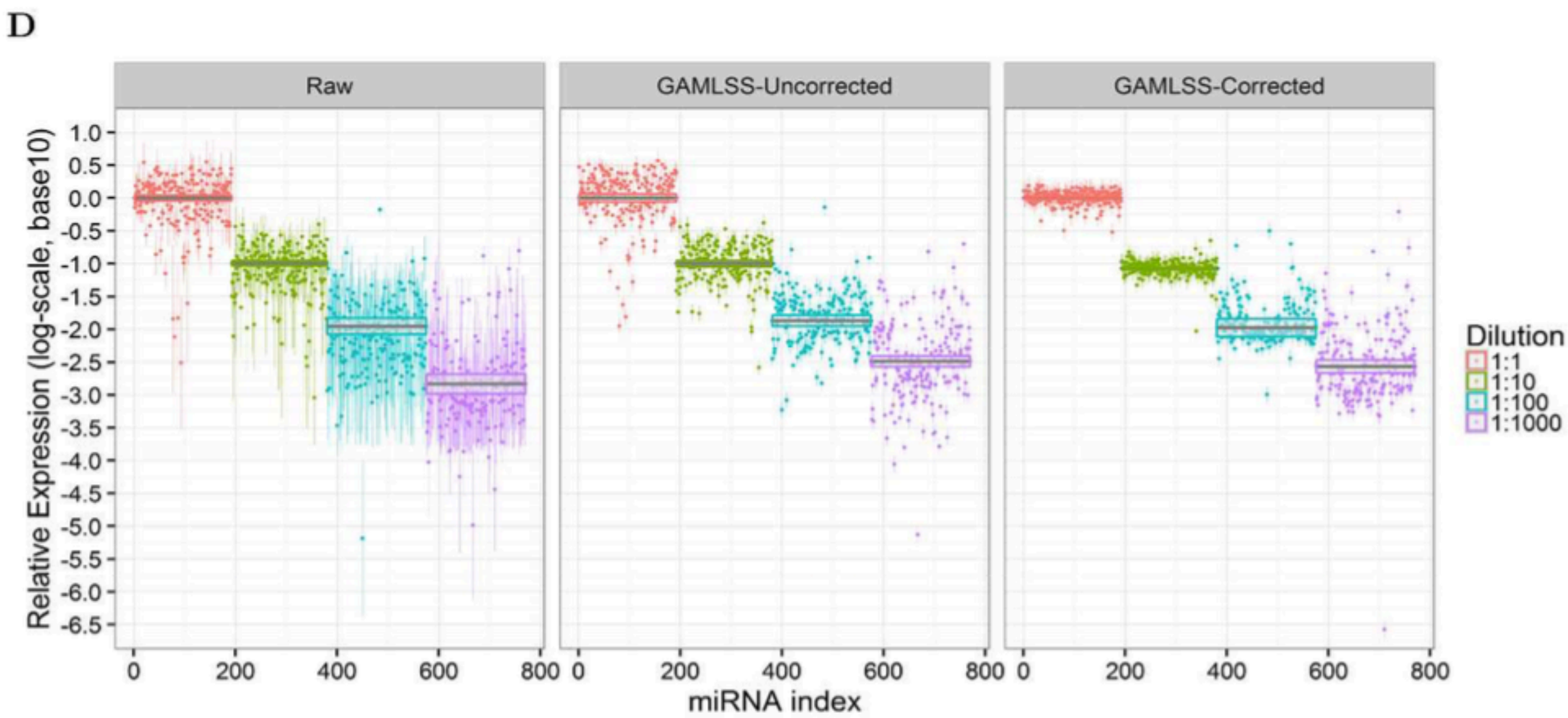
Higher in figure means different sequences with that isomiR type

Bigger the size of the dot means expression of that isomiR type is higher



Correcting quantification

PCR amplification and ligase bias correction factors



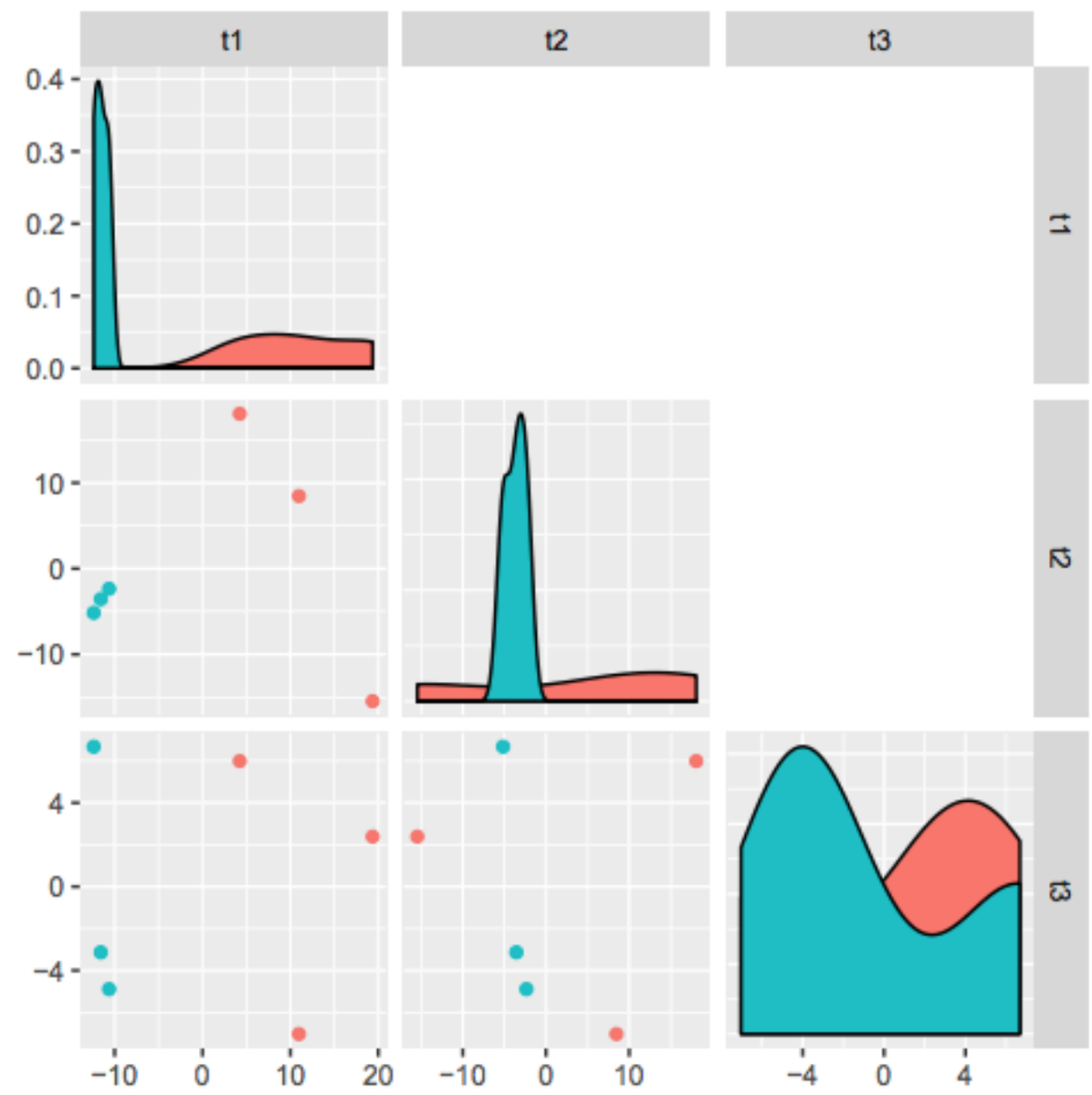
DE analysis

- DESeq2 as in RNAseq
- Sometimes filtering miRNA by group can help to increase power.(`keep miRNAs with counts in 80% of samples in any group`)
- limma-voom strategy should work equally

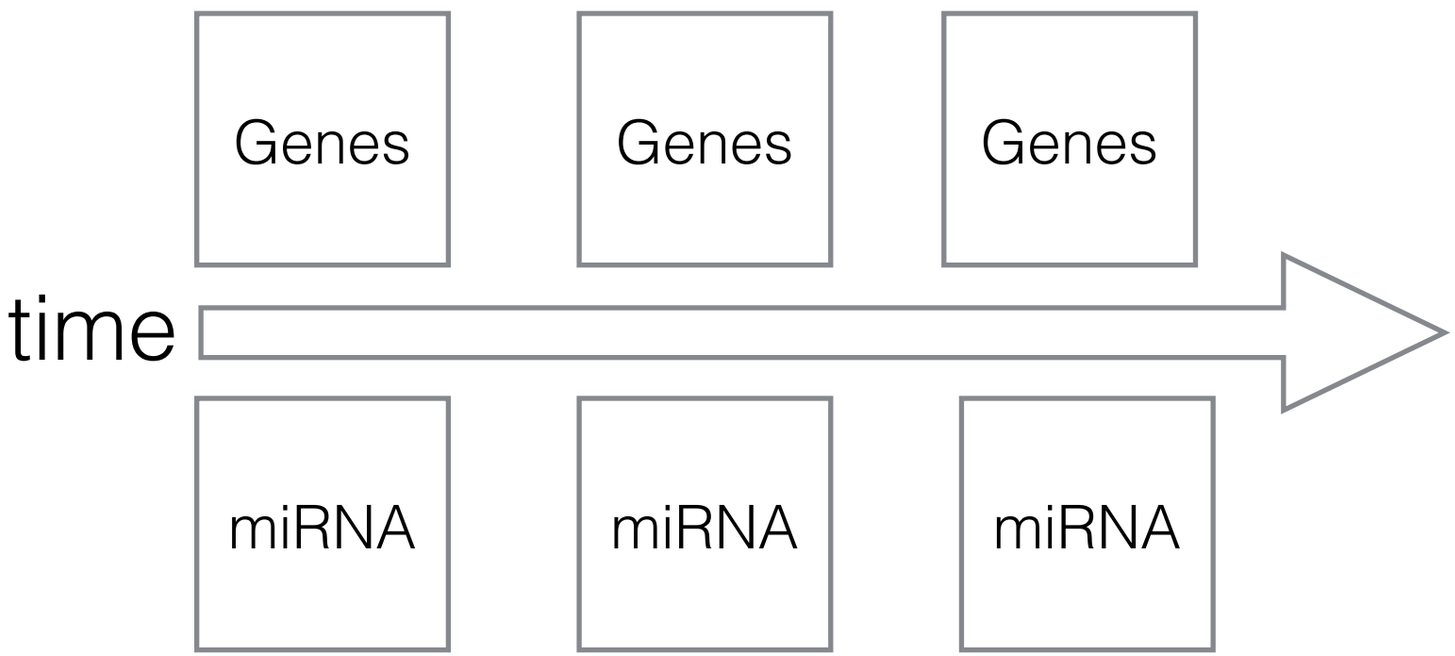
Supervised Clustering

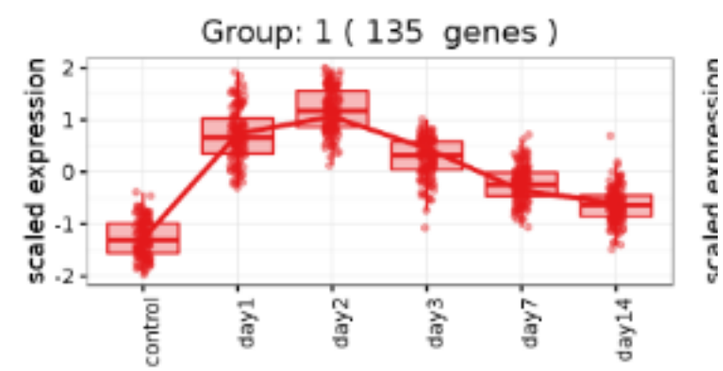
```
ids = isoCounts(ids, iso5=TRUE, minc=10, mins=6)
ids = isoNorm(ids)
pls.ids = isoPLSDA(ids, "condition", nperm = 2)
df = isoPLSDAplot(pls.ids)
```

```
> head(pls.ids$vip)
                                variable      VIP
hsa-let-7c-5p.t5:GT hsa-let-7c-5p.t5:GT 1.518223
hsa-let-7d-5p.t5:0   hsa-let-7d-5p.t5:0 1.533554
hsa-let-7f-5p.t5:tg hsa-let-7f-5p.t5:tg 1.421619
hsa-let-7i-5p.t5:0   hsa-let-7i-5p.t5:0 1.356090
hsa-let-7i-5p.t5:t   hsa-let-7i-5p.t5:t 1.525162
hsa-miR-1.t5:0           hsa-miR-1.t5:0 1.383350
```



mRNA-miRNA interaction





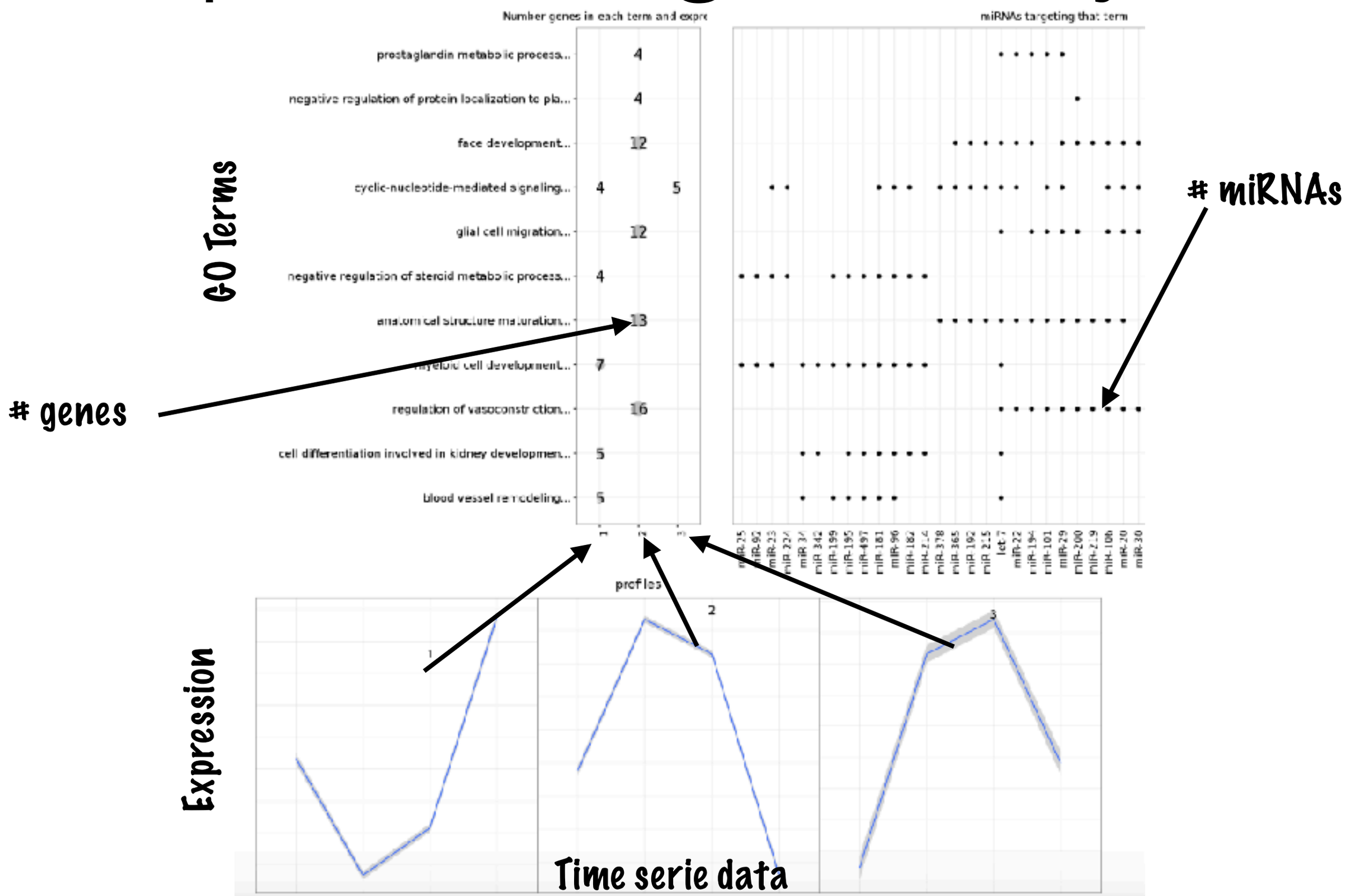
+

GO enrichment

targetscan targets

negative regulation

Output of target analysis



Conclusion

- decide wisely about the protocol to use
- mapping to precursor and parsing with mirtop
- participate in the open project for miRNA annotation
- analyze isomiRs as well (isomiRs)
- DESeq2 for differential expression (my experience)
- mRNA-miRNA paired data helps incredible for downstream analysis