miRNAs variants accuracy with sequencing platforms

Lorena Pantano, P.h. D

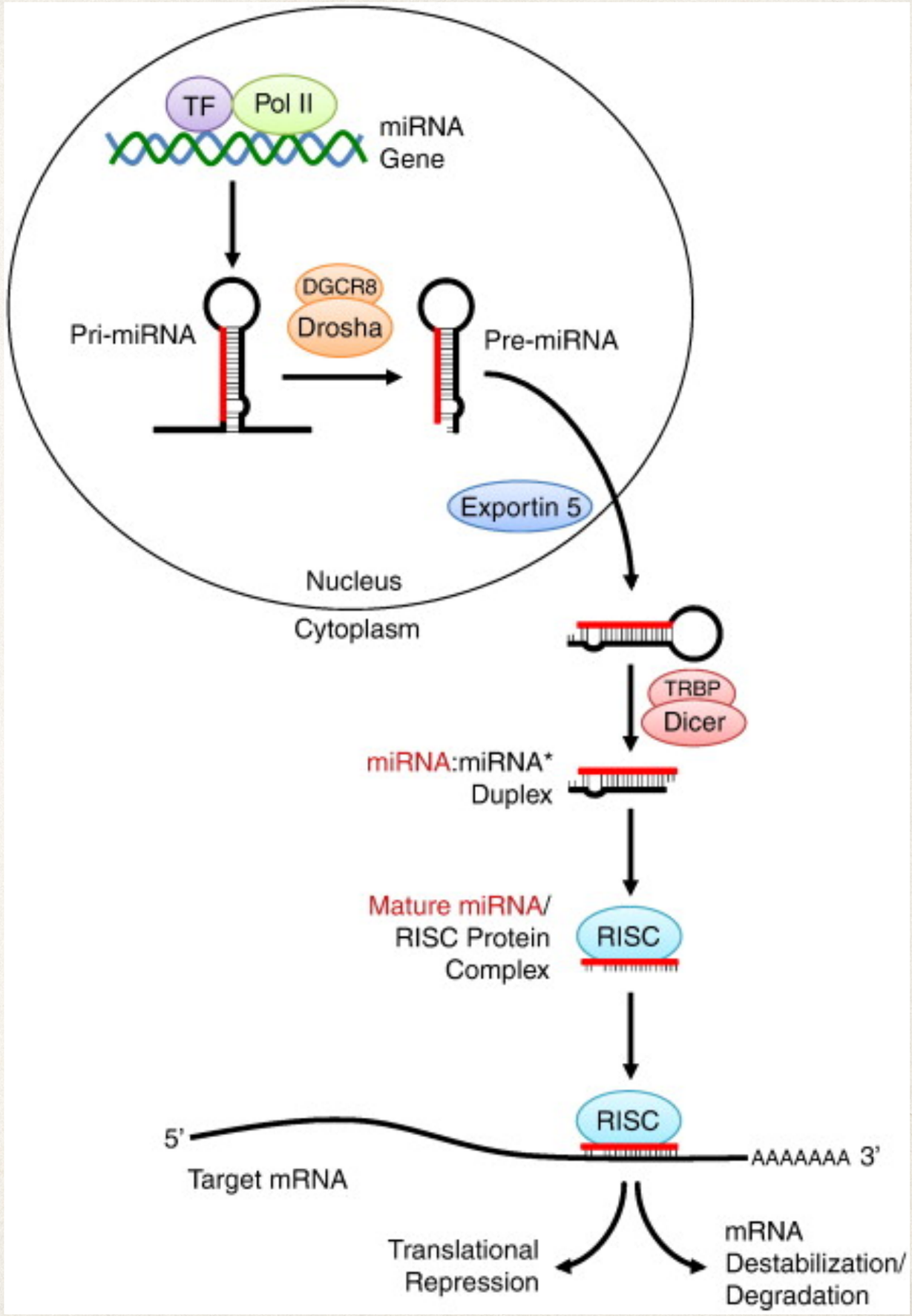
https://lpantano.github.io

@lopantano

03-18-2019

miRNAs

Gene regulation by imperfect complementary between seed region in miRNA and 3'UTR in the targeted RNA molecule.



https://www.sciencedirect.com/science/article/pii/S002228361300154X

isomiRs

GGGATGAGGTAGTAGGTTGTATAGTTTTAGG

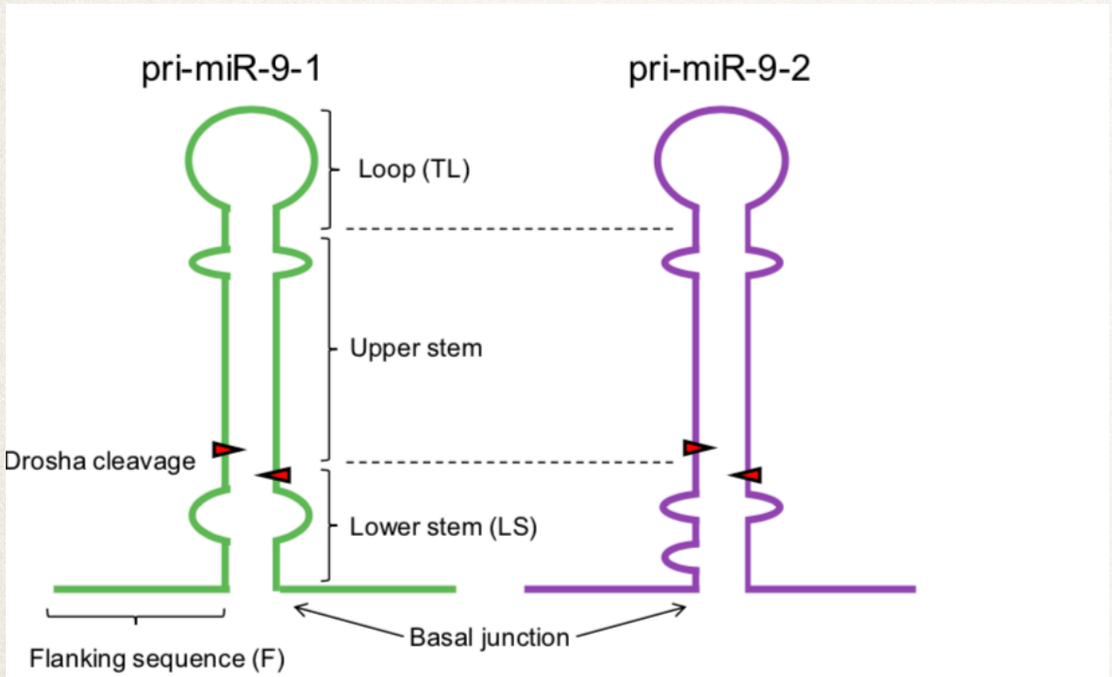
TGAGGTAGTAGGTTGTATAGTT

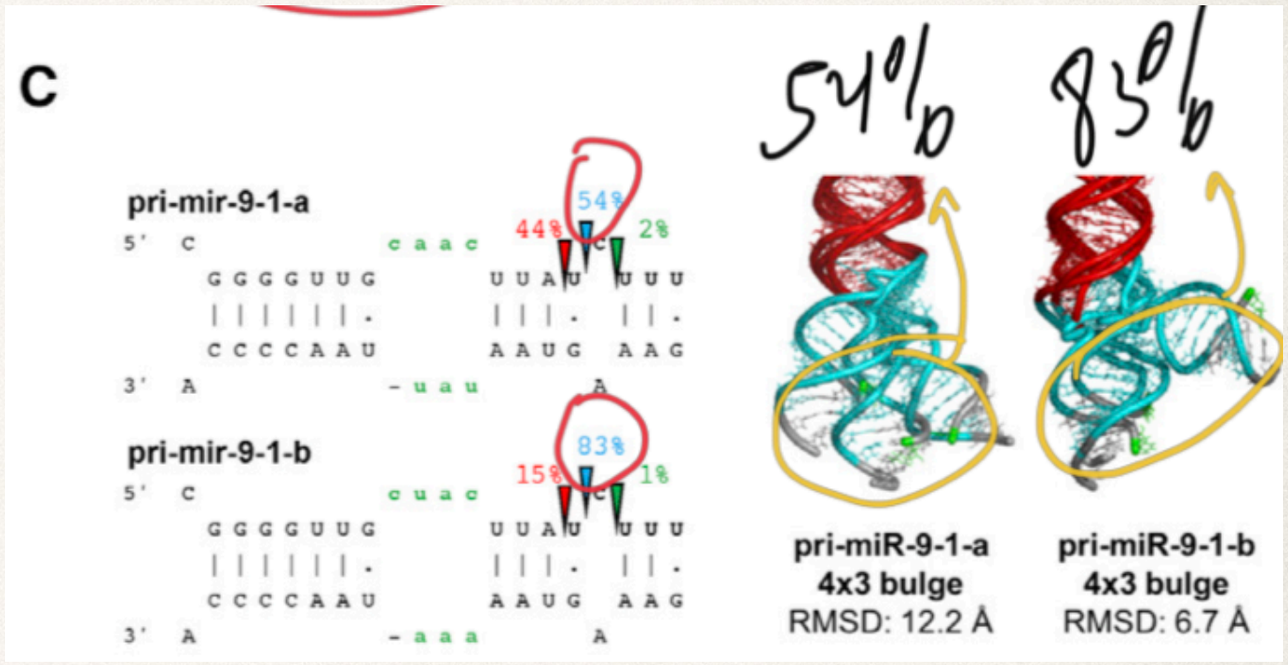
ATGAGGTAGTAGGTTGTATAGTTT

∓GAGGTAGTAGGTTGTATAGT∓

TGAGGTAGTAGGTTGTATAGTTAA

∓GAGGTAGTAGGTTGTATAGT∓AA





Bofill-De Ros et al., 2019, Cell Reports *26*, 447–459
January 8, 2019
https://doi.org/10.1016/j.celrep.2018.12.054

Biomarkers for the early-detection and monitoring of Huntington's Disease

David W. Salzman*, *Joli Bregu^, Nathan S. Ray*, and Richard H. Myers*^*
**sRNAlytics Inc. AstraZeneca BioHub Incubator, 35 Gatehouse Drive, Waltham MA, 02451*
^Boston University Medical School, Department of Neurology, 72 East Concord Street, Boston, MA, 02118

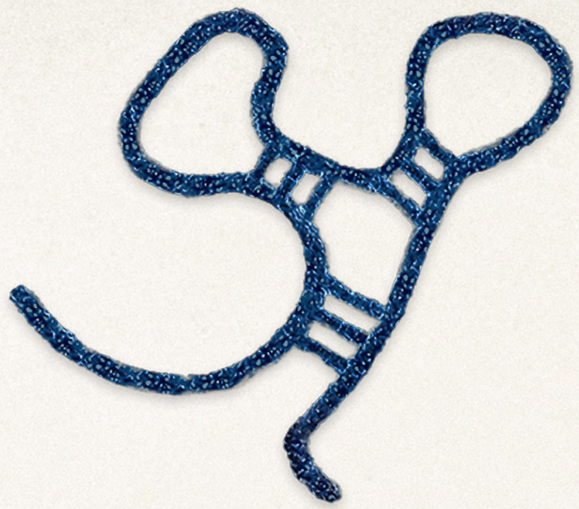
Huntington's Disease (HD) is a genetic disease caused by a CAG trinucleotide repeat in Exon1 of the huntingtin gene. Neurodegeneration results in the loss of cognitive and motor functions, and is caused by aggregation of mutant huntingtin protein in striatal neurons. Volumetric changes in the striatum can be detected decades before the manifestation of clinical phenotypes, indicating that therapeutic intervention would need to occur long before symptomatic presentation. In clinical practice and research settings, the Unified Huntington's Disease Rating Scale (UHDRS) is utilized to evaluate a patients overall physical and neurological health. UHDRS is also the most widely used outcome measure for establishing drug efficacy. However, ...

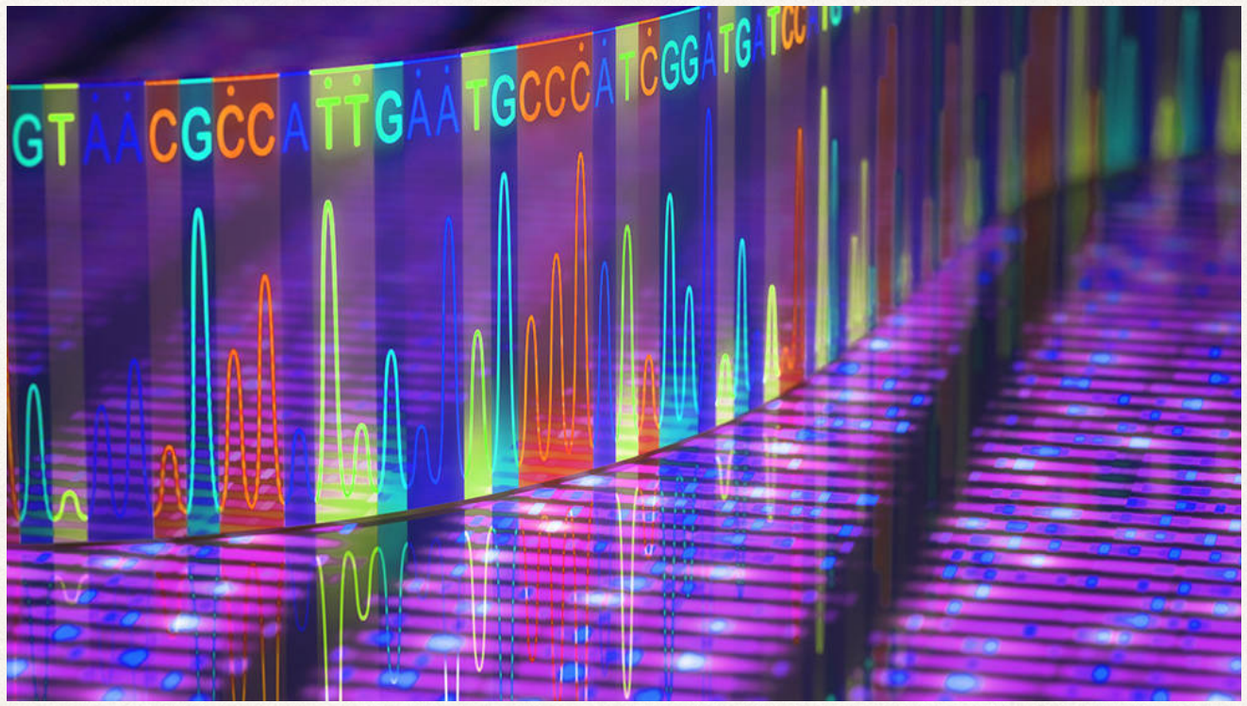
Nucleic Acids Res. 2017 Apr 7;45(6):2973-2985. doi: 10.1093/nar/gkx082.

Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types.

Telonis AG[1], Magee R[1], Loher P[1], Chervoneva I[2], Londin E[1], Rigoutsos I[1].

Author information



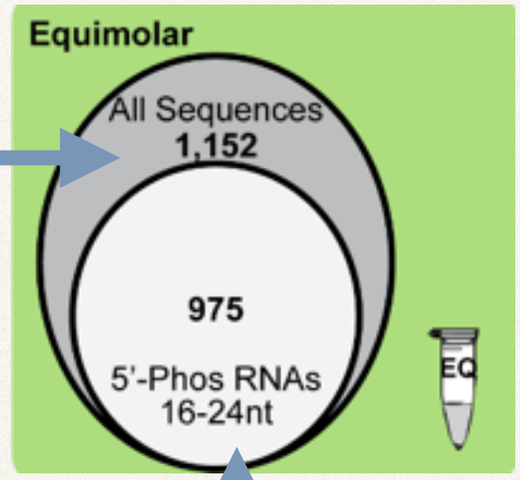


miRNAs

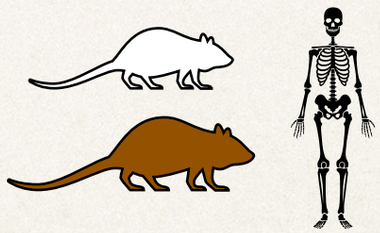
simplify idea of synthetic

Giraldez et al.

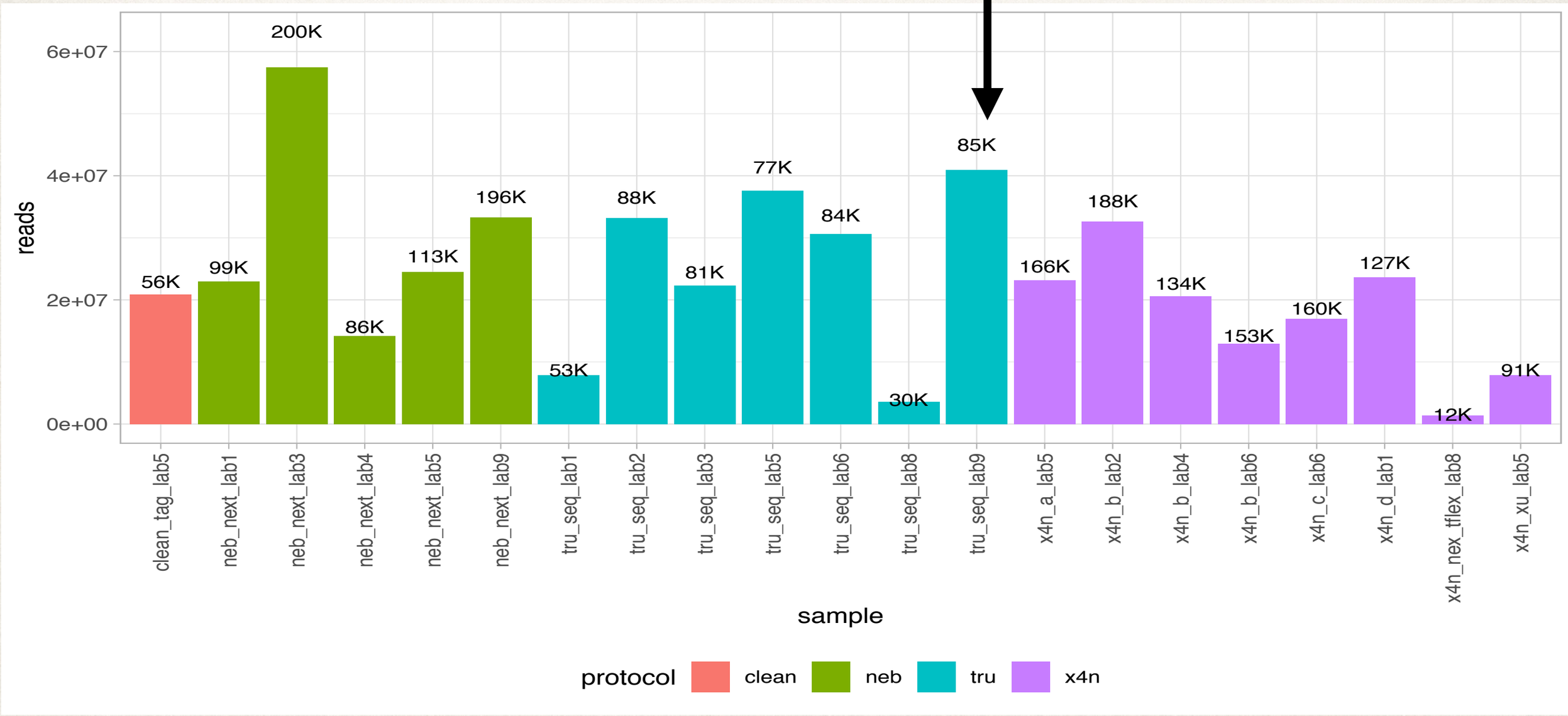
Tewari custom



Tewari synthetic



40 million reads and 85K different sequences.



mirTOP

miRNA transcriptome open project

http://mirtop.github.io

Repositories 5 | People 20 | Teams 2 | Projects 0 | Settings

Find a repository... | Type: All | Language: All | Customize pinned repositories | New

incubator

Where all ideas and discussions happen to lead to new repositories

R ★ 3 4 Updated a minute ago

miRTOP.github.io

project for small RNA standard annotations

HTML ★ 2 1 MIT 1 issue needs help Updated 4 days ago

mirtop

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

formatter mirna gff isomirs smallrna-seq

Python ★ 6 13 MIT 17 issues need help Updated 5 days ago

simulator

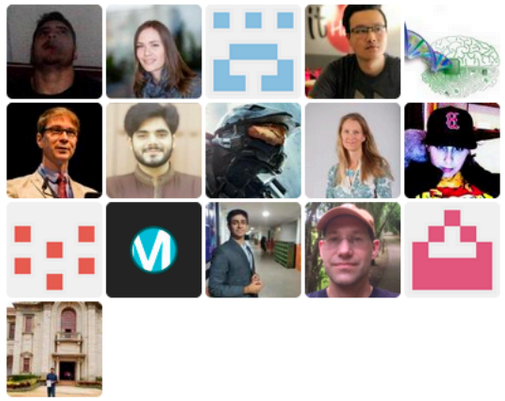
first ideas and brainstorming for small RNA simulator

C++ Updated 25 days ago

Top languages

C++ Python R HTML

People 20 >



Invite someone

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

UNIVERSITY OF
OREGON

CRG
Centre
for Genomic
Regulation

Harvard Medical School
Research Computing

UNIVERSIDAD
DE GRANADA

Jefferson
HOME OF SIDNEY KIMMEL MEDICAL COLLEGE

JOHNS HOPKINS
SCHOOL of MEDICINE

HEALTH
UNIVERSITY OF UTAH

BRIGHAM HEALTH
BWH BRIGHAM AND
WOMEN'S HOSPITAL

SciLifeLab

BROAD
INSTITUTE

NIH National Institutes of Health

NSF

POLITECNICO DI TORINO
1859 1906

BC
CAN
CER
Provincial Health Services Authority

Main projects:
- a format
- toolkit

By Kieran O'Neill

Format derived from GFF3: mirGFF3

a

```
## mirGFF3. VERSION 1.1
## source-ontology: miRBasev21 doi:10.25504/fairsharing.hmgte8
## COLDATA: sample1
```

b

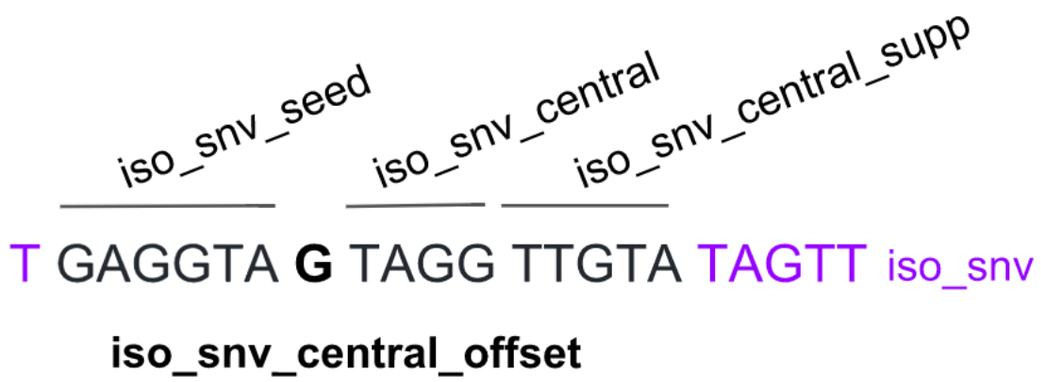
Read=GATGAGGTAGTAGGTTGTATAGTT -> UID=**iso-24-5URPV39QFE**

Read=ATGAGGTAGTAGGTTGTATAGTT -> UID=**iso-23-I0S31NSL0E**

c

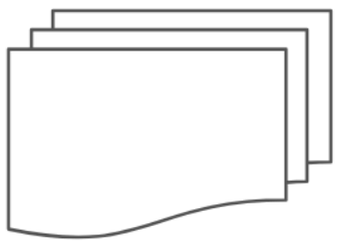
GGGATGAGGTAGTAGGTTGTATAGTTTTAGG	Precursor
TGAGGTAGTAGGTTGTATAGTT	User defined reference
ATGAGGTAGTAGGTTGTATAGTT**T**	iso_5p:-1, iso_3p:+1
~~T~~GAGGTAGTAGGTTGTATAGT~~T~~	iso_5p:+1, iso_3p:-1
TGAGGTAGTAGGTTGTATAGTT**AA**	iso_add:2
~~T~~GAGGTAGTAGGTTGTATAGT~~T~~**AA**	iso_5p:+1, iso_3p:-1, iso_add:2

d



e

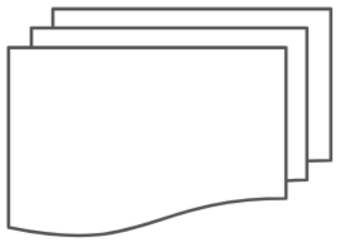
Python API: mirtop



Input:
seqbuster
miRge2.0
isomiR-SEA
sRNAbench
Prost!
mirGFF3
BAM

mirtop API

Importer
Converter
Manager
Compiler
Exporter



Output:
mirGFF3
isomiRs
VCF
Count matrix

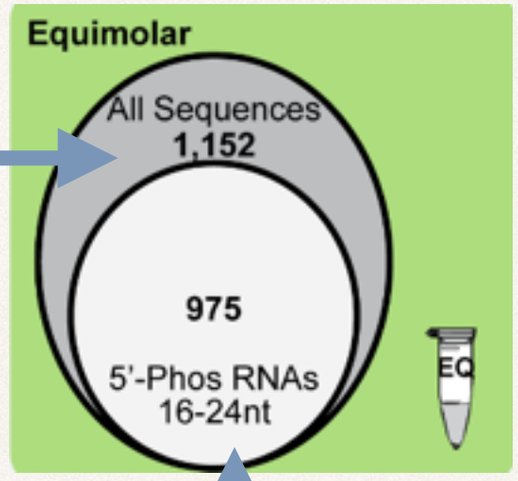
"How well NGS detect isomiRs?"

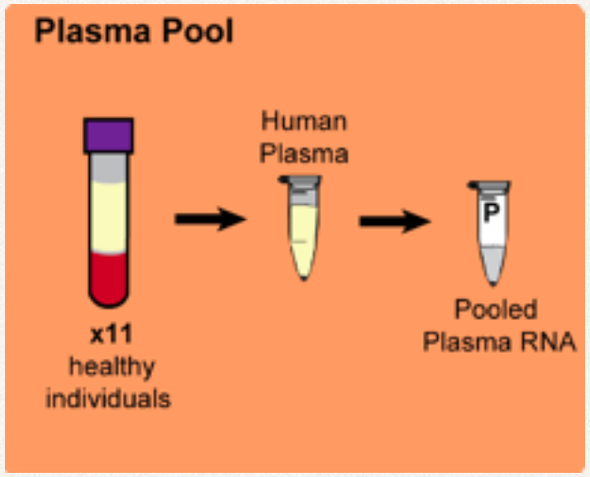
– Scientists

Samples suitable for benchmarking

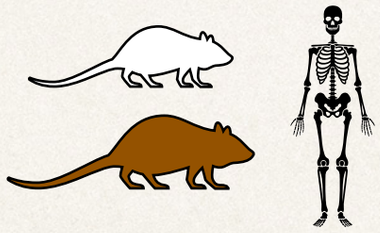
Giraldez et al.

Tewari custom

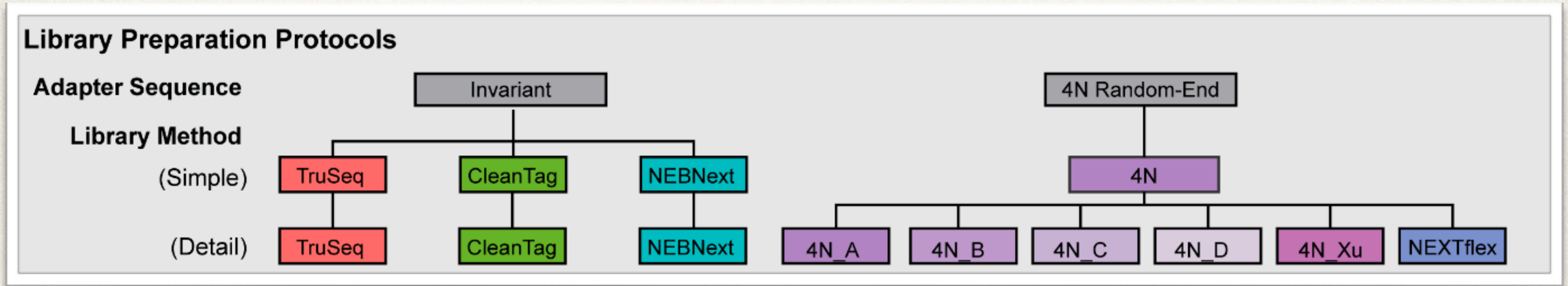




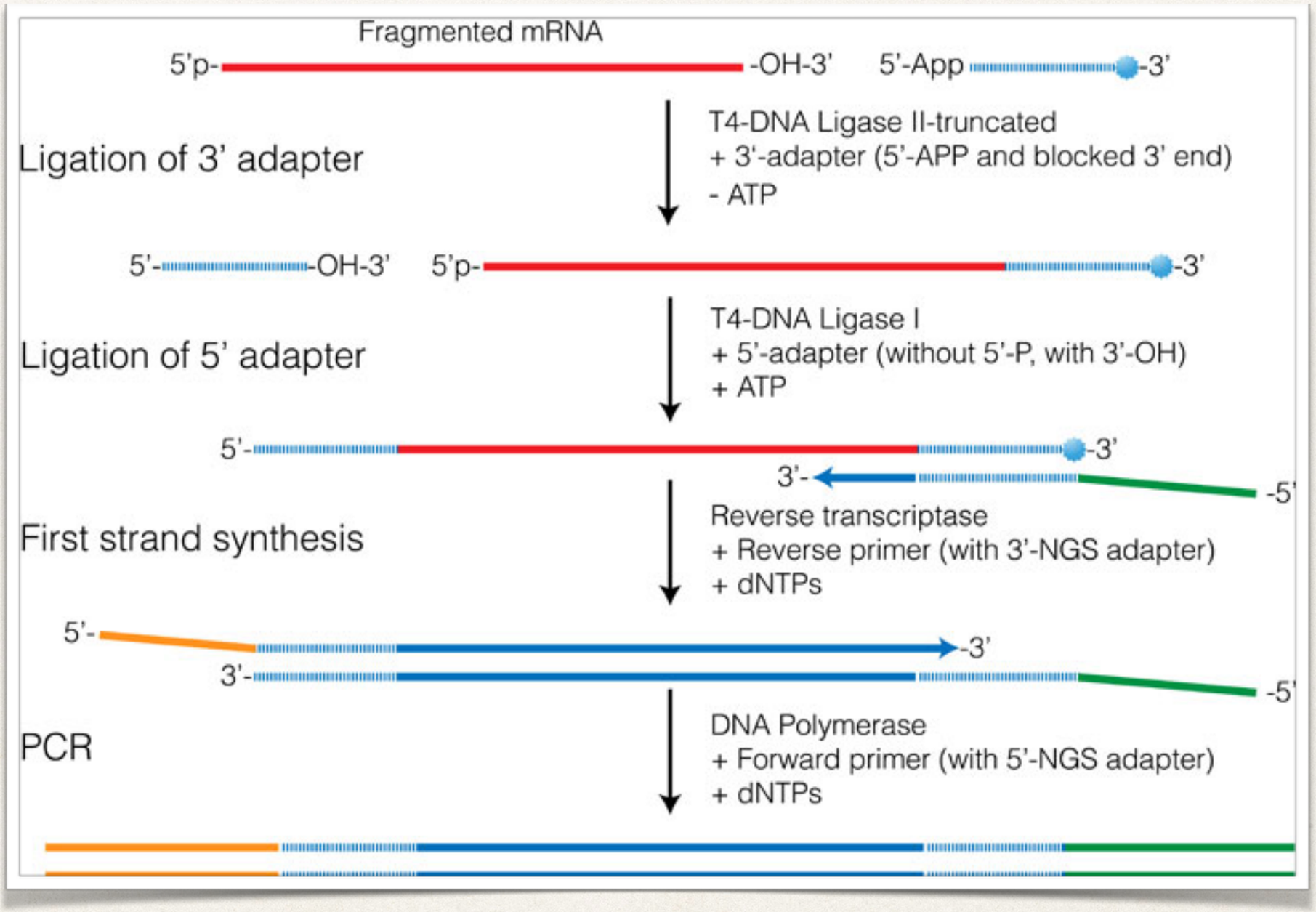
Tewari synthetic



Different protocols



Protocols



https://rnaseq.uoregon.edu/

Variations

	3 adapter	5 adapter	PEG
TrueSeq			
NEBNext			Yes
NextFlex	4N	4N	Yes
CleanTag	Methylphosphonate	2’0me	
CATS	poly-T	TSO with rX	
SMARTer	poly-T	TSO	

Source of Variation

Synthesis

Shipment

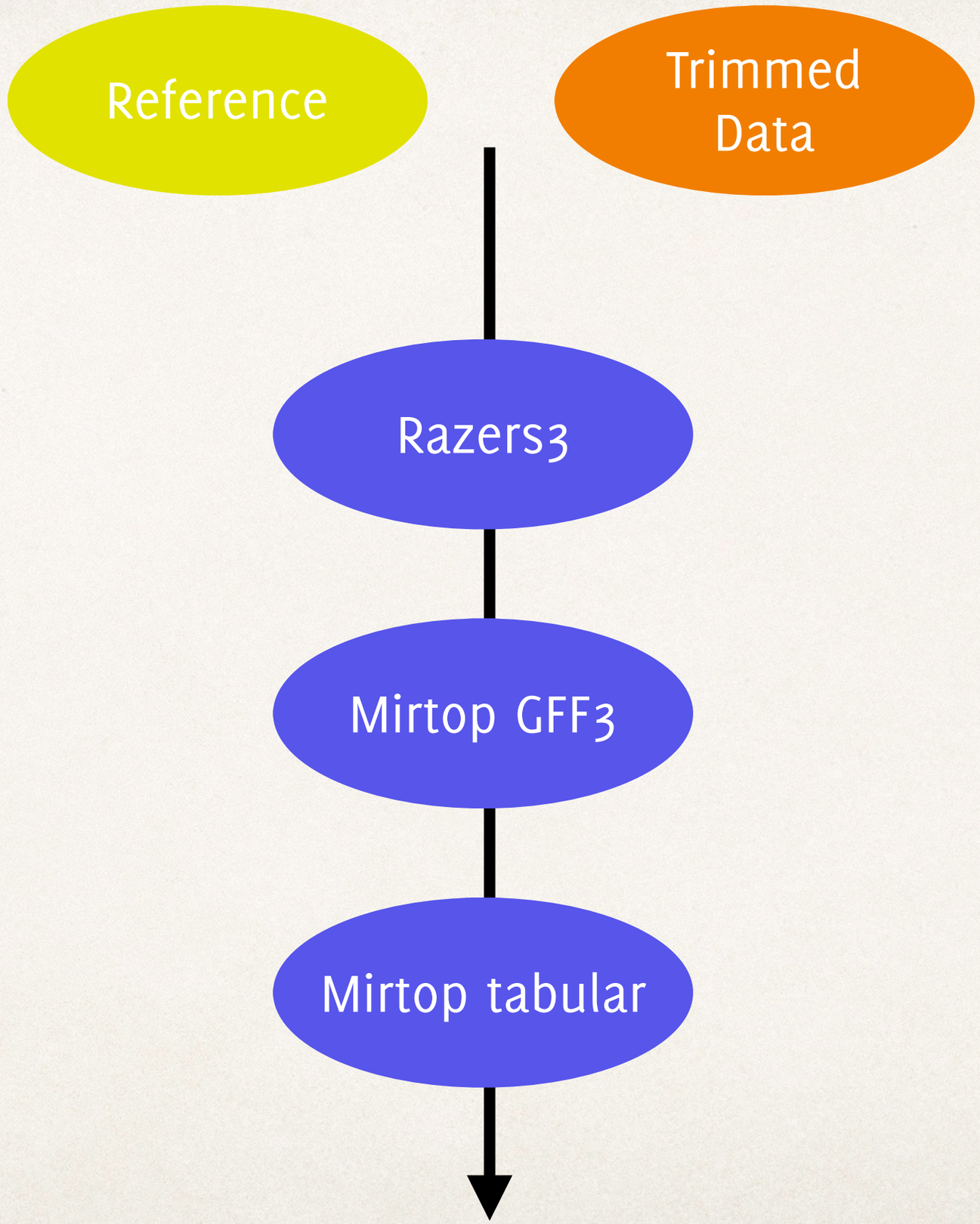
Library preparation

Sequencing

Analysis

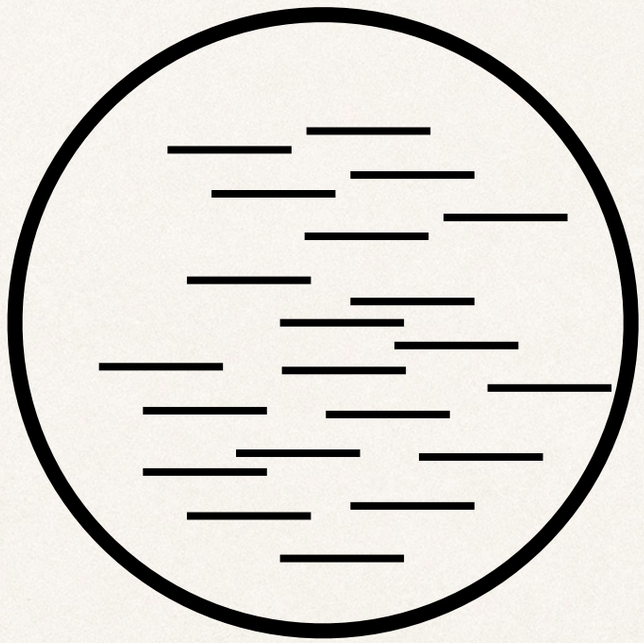
Pipeline razers3 + mirtop

snakemake


Reference
Trimmed Data
Razers3
Mirtop GFF3
Mirtop tabular

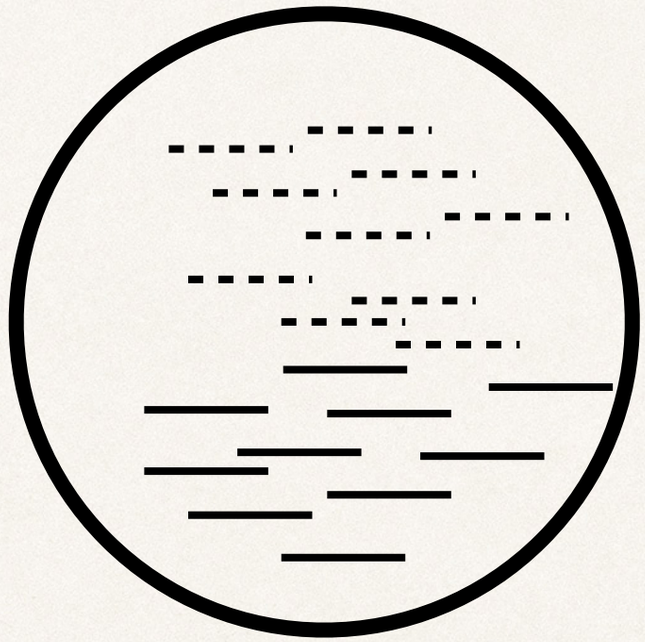
Methods and Metrics

Sample



Library size = 20

Sample

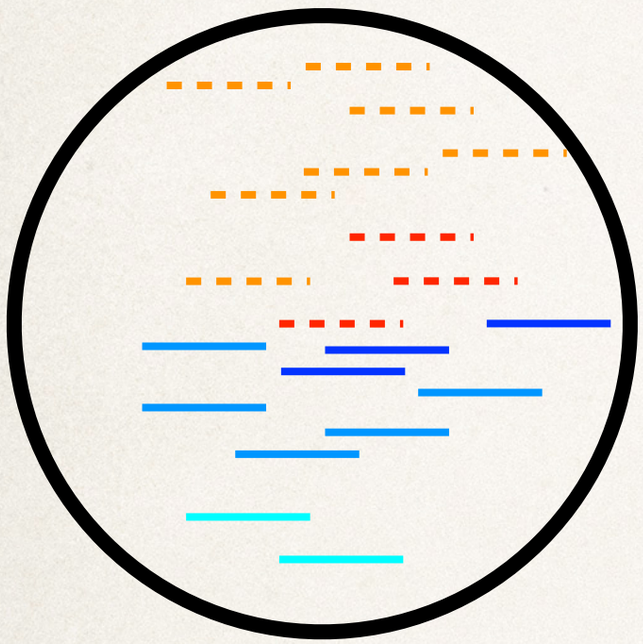


miRNA 1 = 10

miRNA 2 = 10

Library size = 20

Sample



Library size = 20

miRNA 1 = x10

isomiR 1.1 = x7

(match perfectly spike-in)

isomiR 1.2 = x3

(NOT match perfectly spike-in)

miRNA 2 = x10

isomiR 2.1 = x3

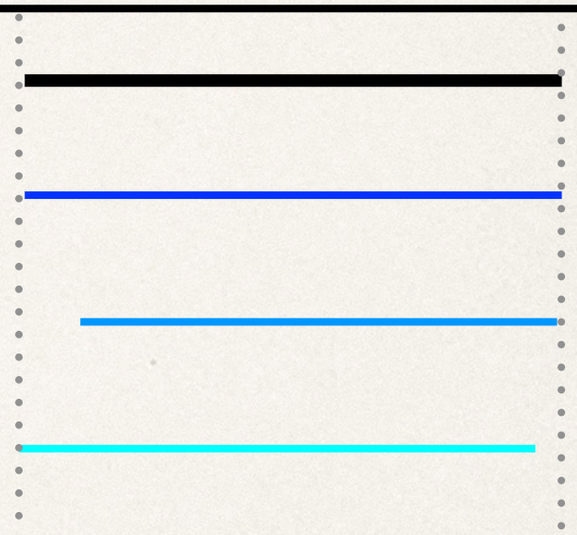
(match perfectly spike-in)

isomiR 2.2 = x5

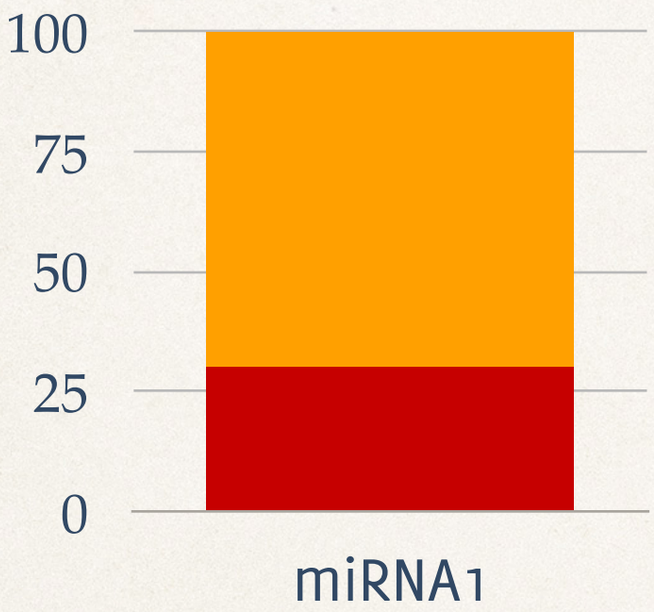
(NOT match perfectly spike-in)

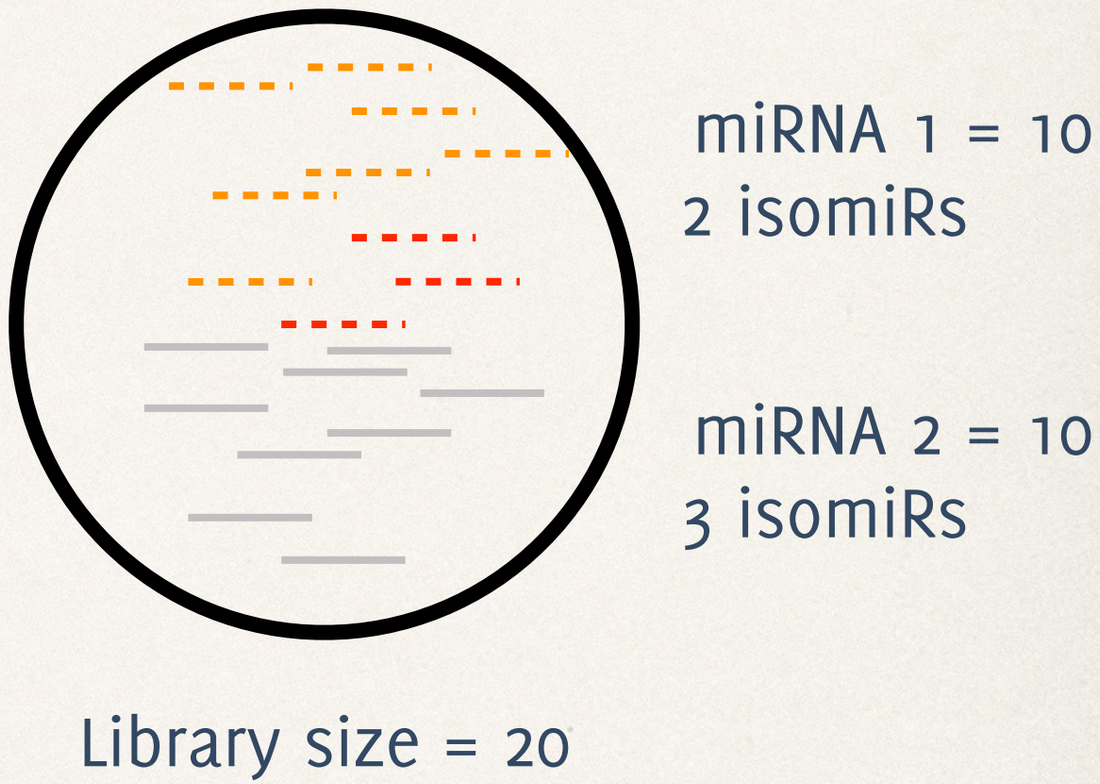
isomiR 2.2 = x2

(NOT match perfectly spike-in)



IMPORTANCE = isomiR 1.1 is 7/10=70%
of the miRNA1 reads





Data analysis - Filters

All sequences in a miRNA

Human?

NO

YES

Is the spike-in detected?

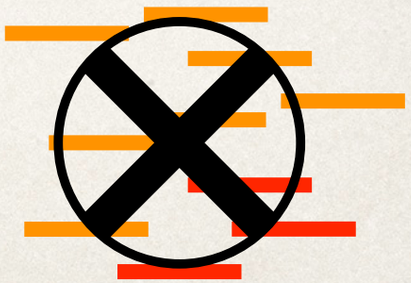
NO

YES

Any sequence in group mapped to other miRNA?

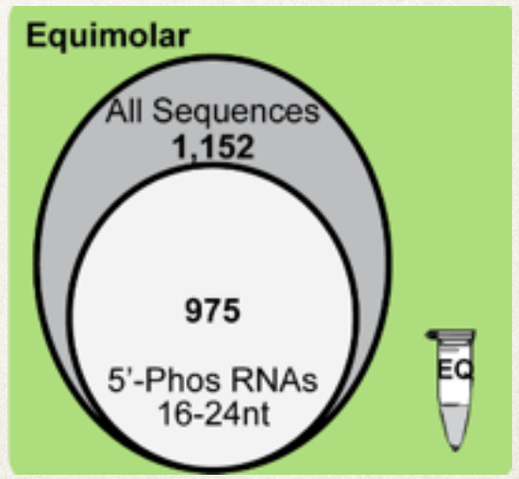
NO

YES

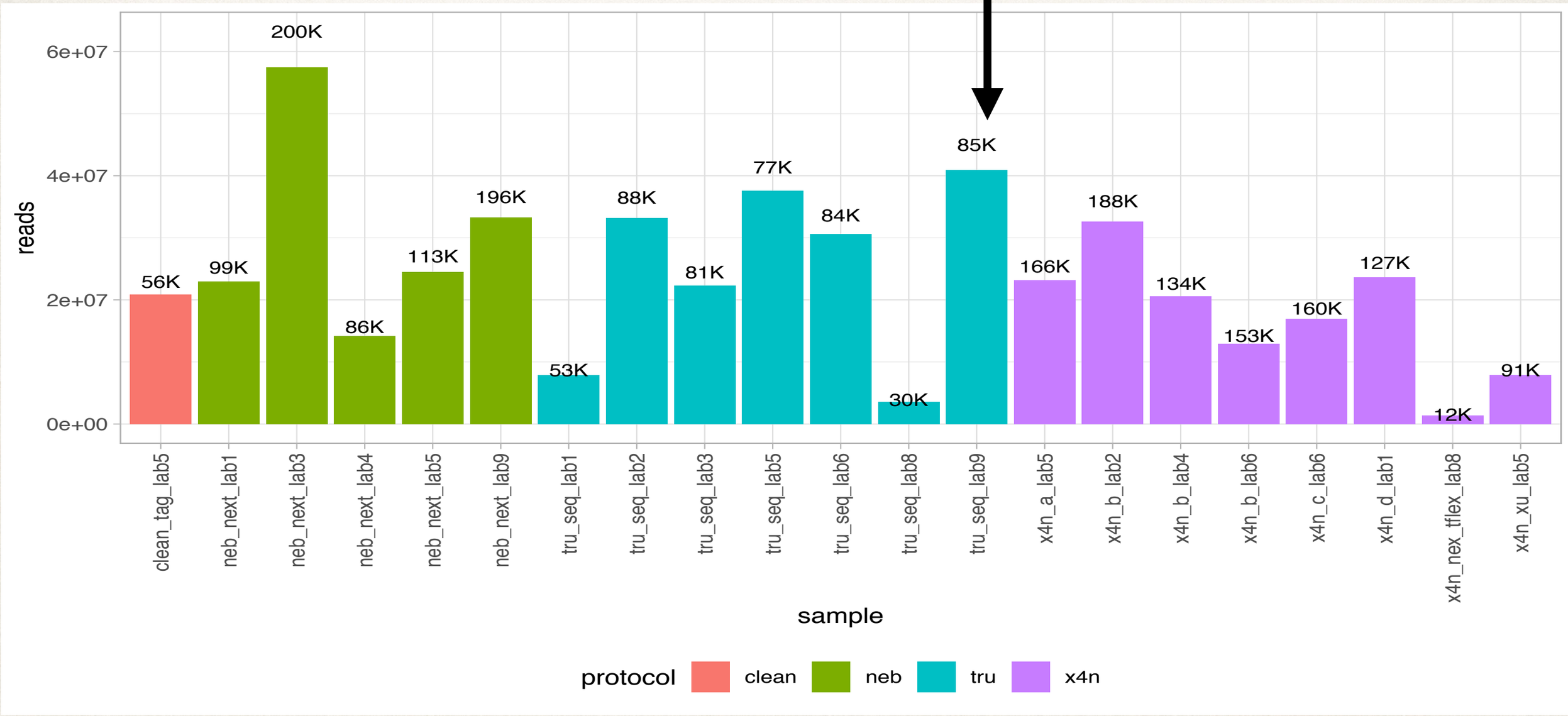


synthetic: library size

Giraldez et al.

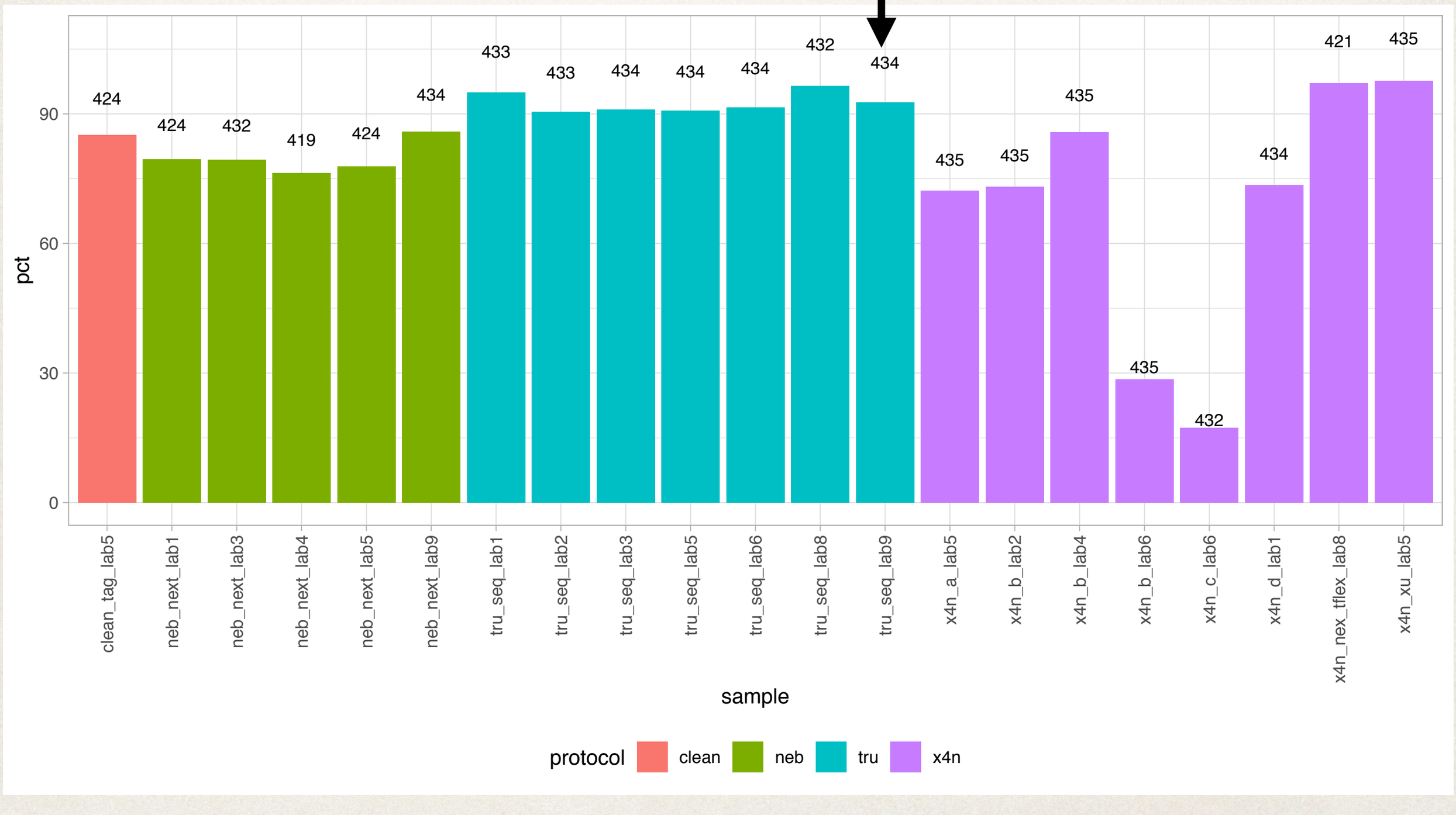


40 million reads and 85K different sequences.



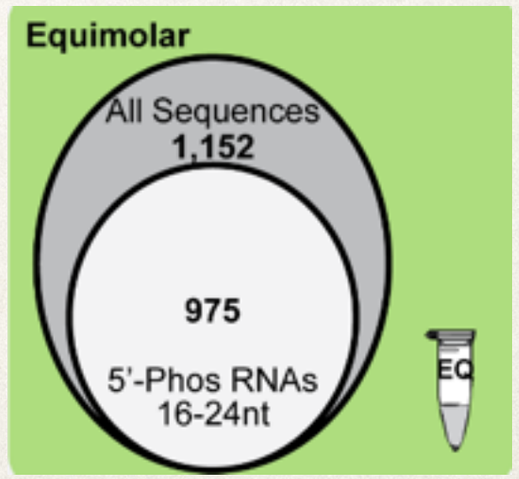
synthetic: spike-ins are the top sequence?

90% (434 being total detected miRNAs) of miRNAs with top sequence to be the expected one.



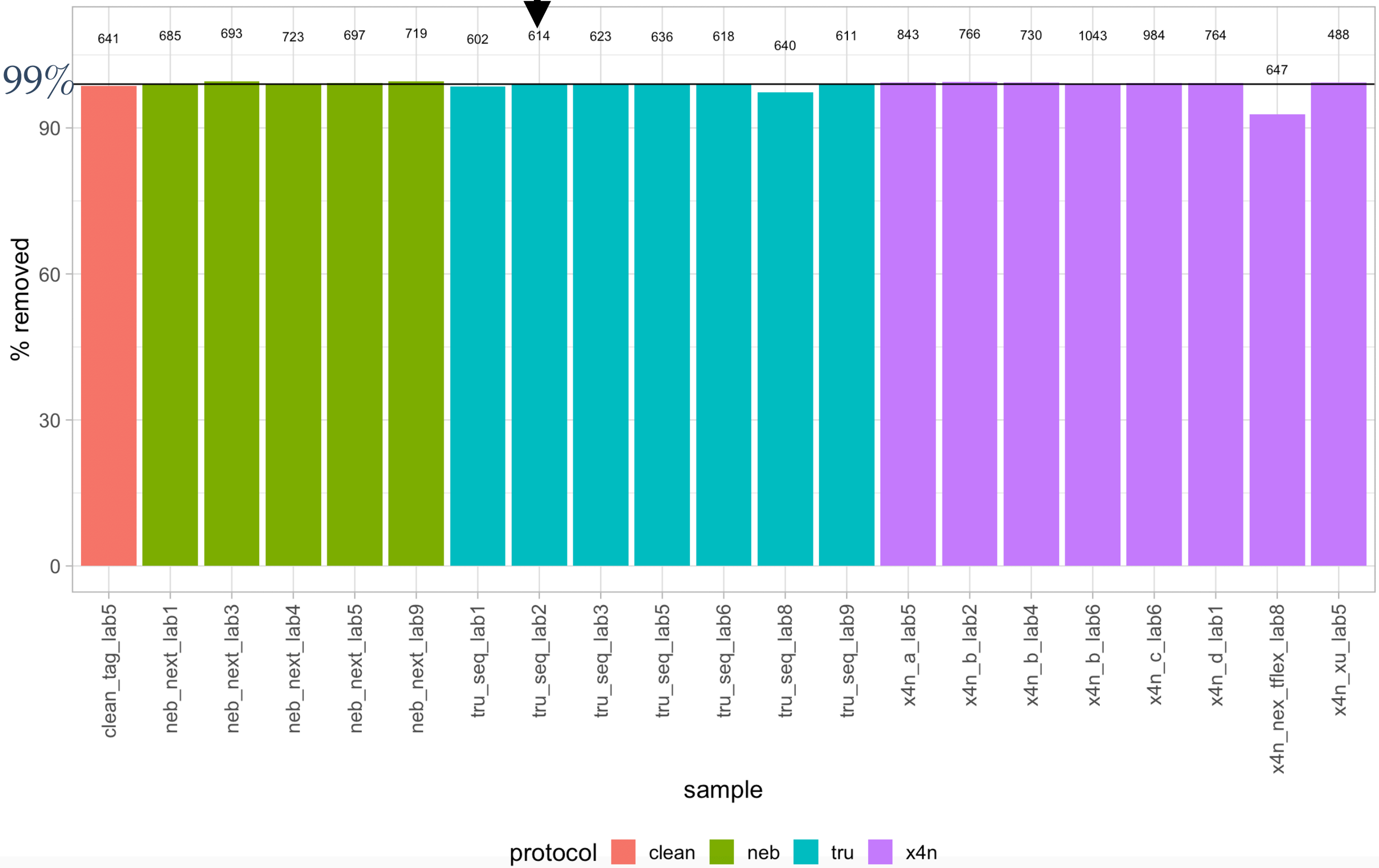
synthetic: Importance of the ‘isomiRs’ detected

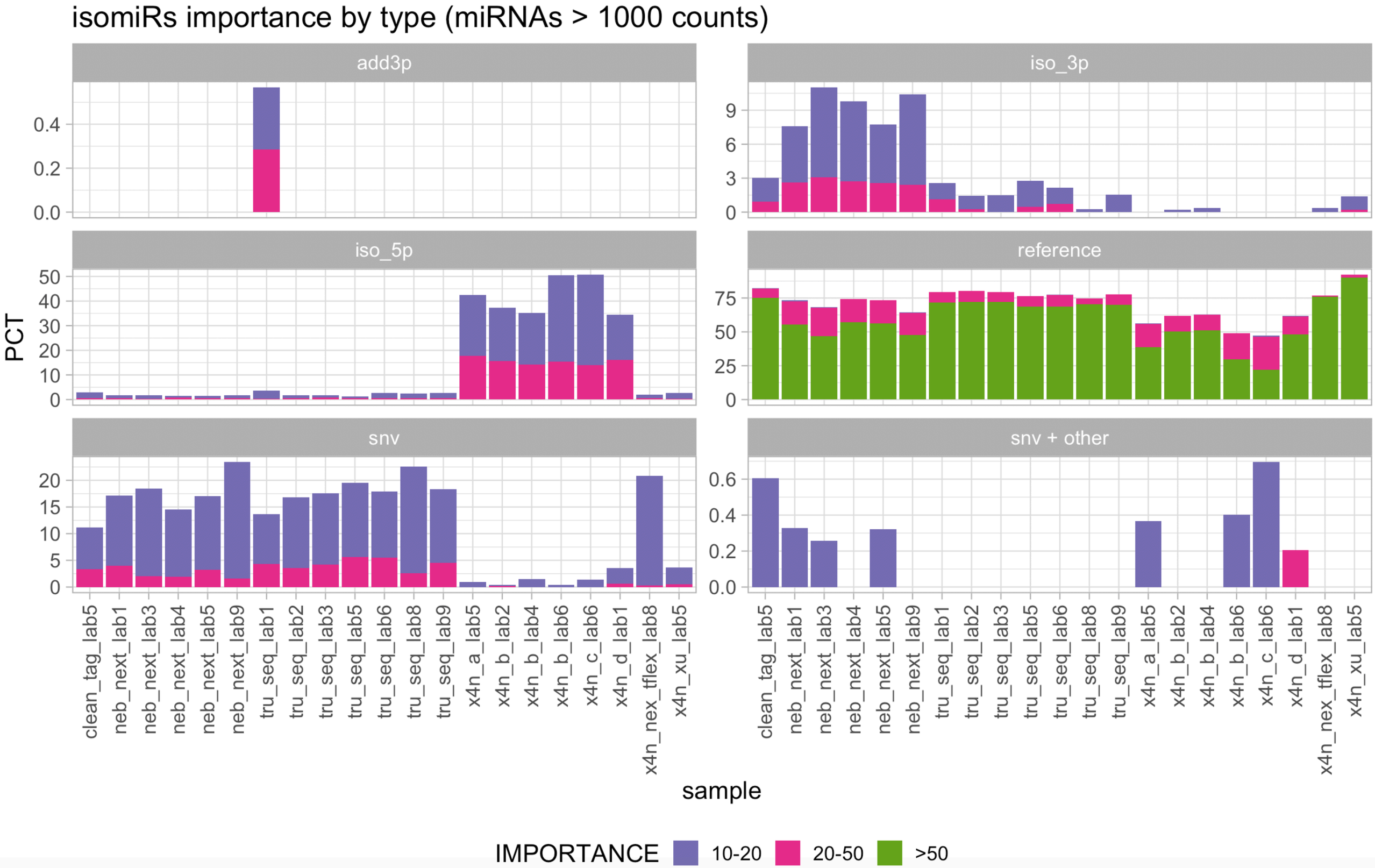
Giraldez et al.



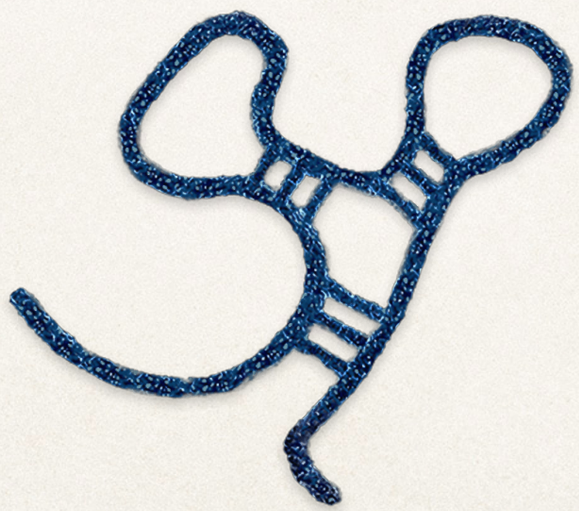
isomiRs importance by type (miRNAs > 1000 counts)

From 85K to 614 unique sequences.
99% of sequences are removed.




isomiRs importance by type (miRNAs > 1000 counts)
add3p
iso_3p
iso_5p
reference
snv
snv + other
PCT
sample
IMPORTANCE
10-20
20-50
>50
clean_tag_lab5
neb_next_lab1
neb_next_lab3
neb_next_lab4
neb_next_lab5
neb_next_lab9
tru_seq_lab1
tru_seq_lab2
tru_seq_lab3
tru_seq_lab5
tru_seq_lab6
tru_seq_lab8
tru_seq_lab9
x4n_a_lab5
x4n_b_lab2
x4n_b_lab4
x4n_b_lab6
x4n_c_lab6
x4n_d_lab1
x4n_nex_tflex_lab8
x4n_xu_lab5

We assume real molecules will be detected across replicates



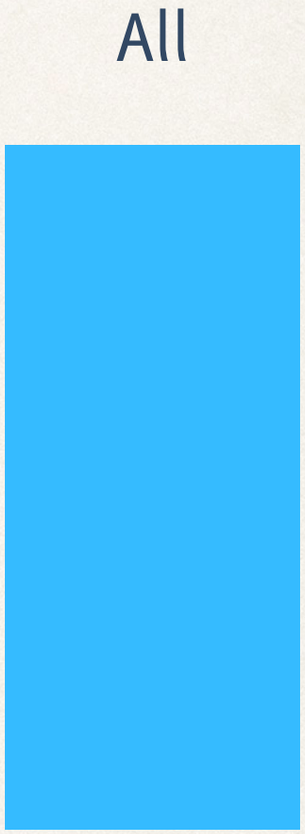
Sample 1

Sample 2

Sample 3

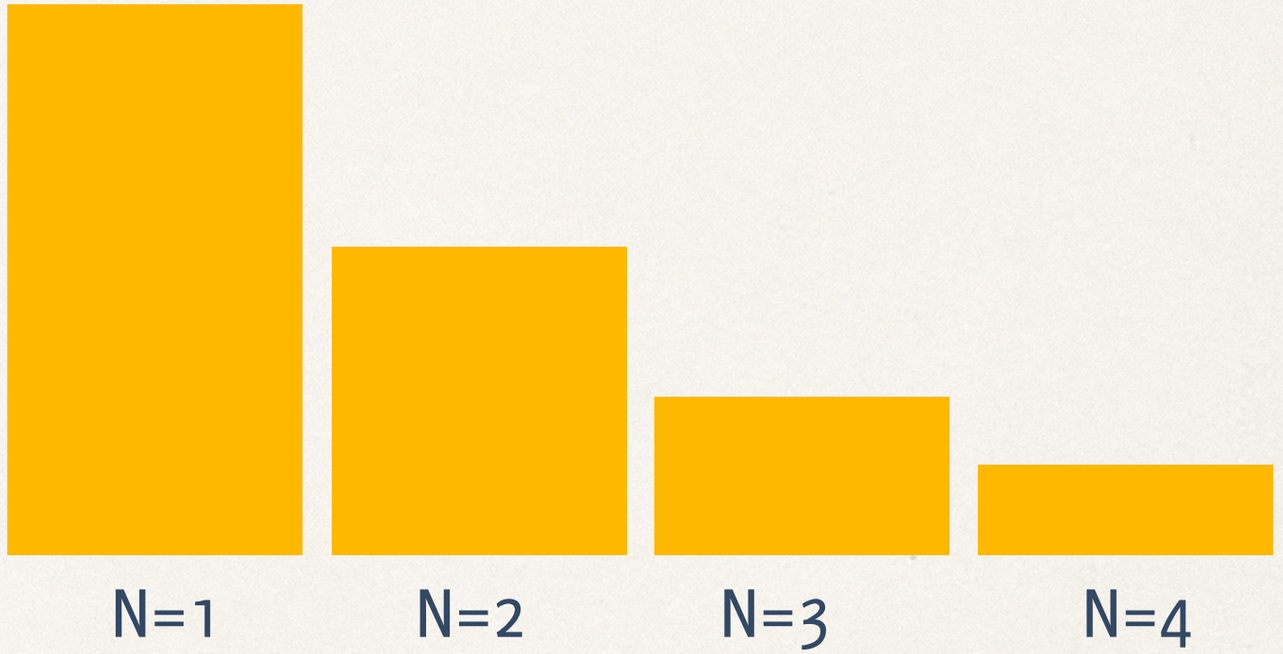
Sample 4

Synthetic miRNAs



Input Noise

Downstream Noise



Narry Kim Data

Protocol 1

30 spikeins x 3 replicates

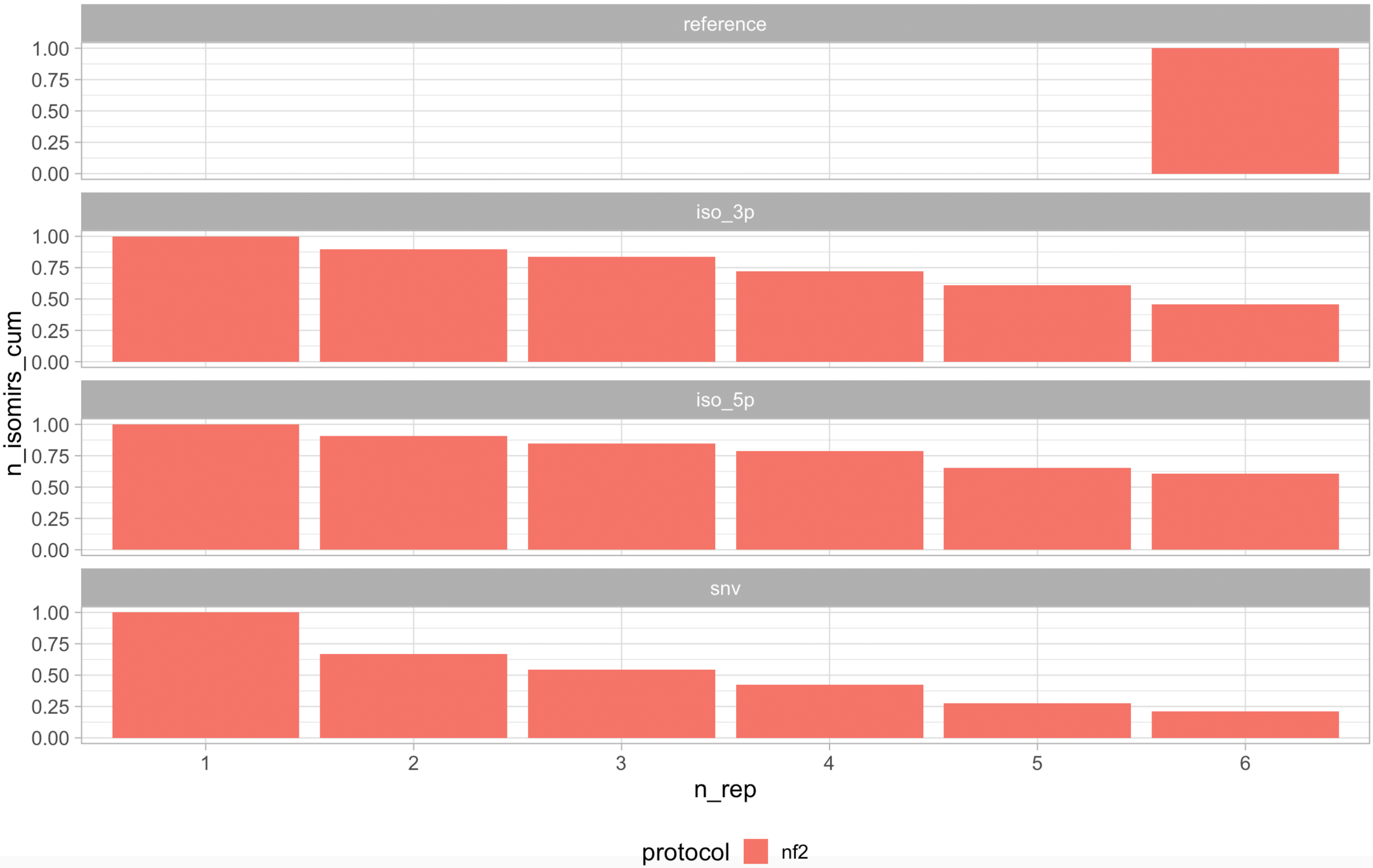
Protocol 2

30 spikeins x 4 replicates


reference
iso_3p
iso_5p
snv
1.00
0.75
0.50
0.25
0.00
n_isomirs_cum
1
2
3
4
n_rep
protocol
_aq_
_ts_

Van Dijk Data

12 spikeins x 6 replicates


reference
iso_3p
iso_5p
snv
1.00
0.75
0.50
0.25
0.00
n_isomirs_cum
1
2
3
4
5
6
n_rep
protocol
nf2

Human plasma

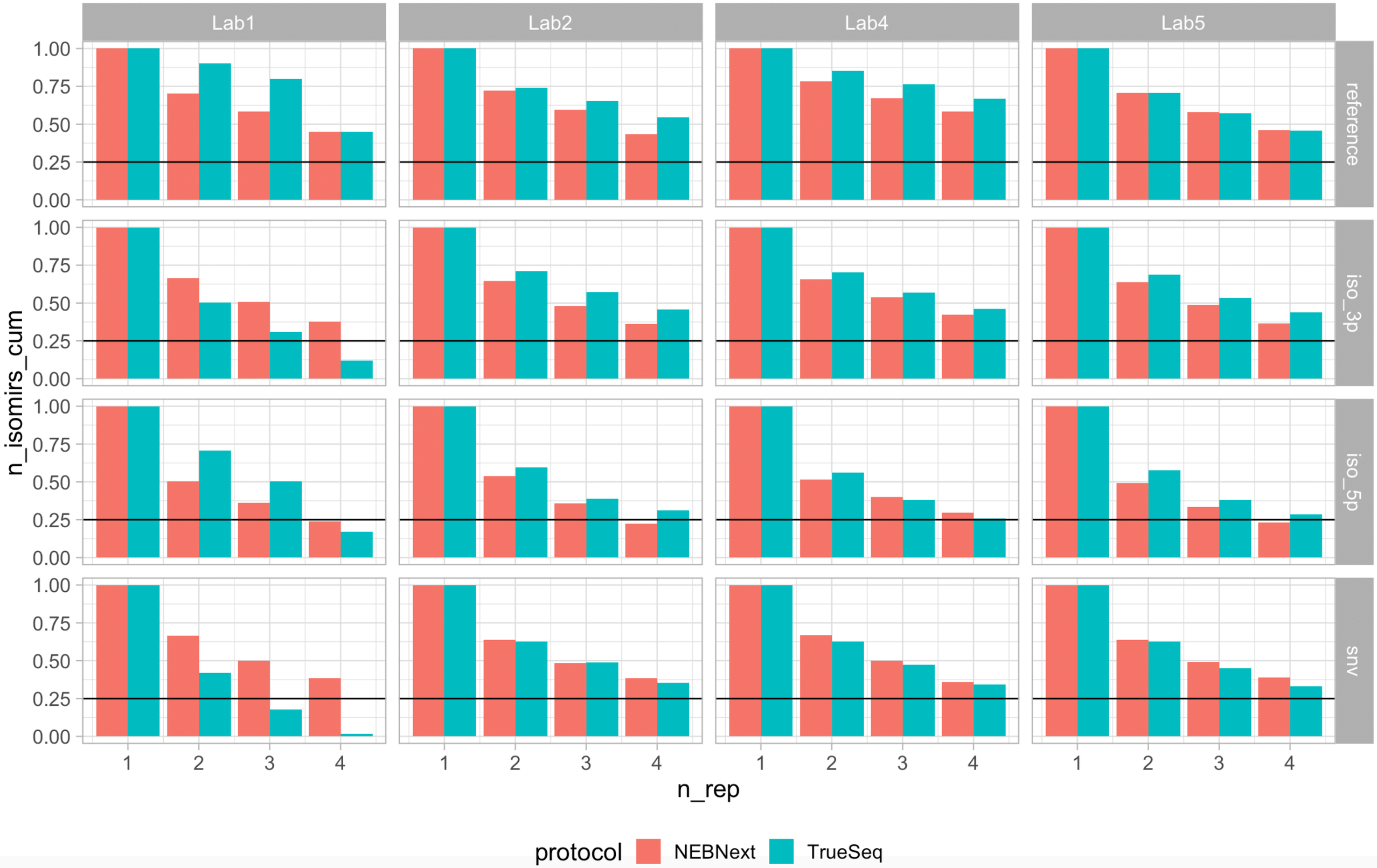
4 laboratories

Protocol 1

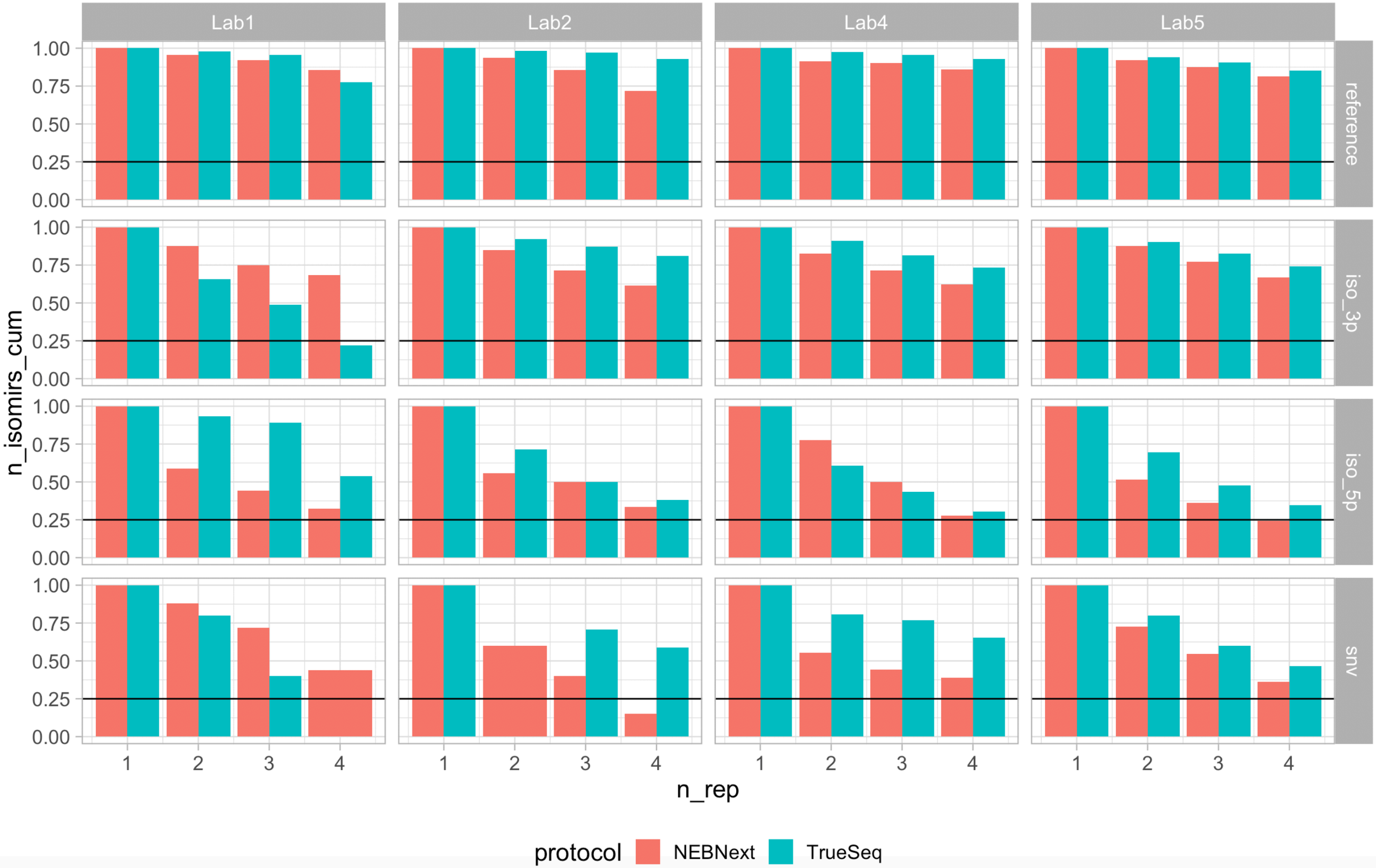
x4 replicates

Protocol 2

x4 replicates


Lab1
Lab2
Lab4
Lab5
reference
iso_3p
iso_5p
snv
irs_cum
n_isom
1.00
0.75
0.50
0.25
0.00
1
2
3
4
n_rep
protocol
NEBNext
TrueSeq

Remove isomiRs > 10% & total > 100


Lab1
Lab2
Lab4
Lab5
reference
iso_3p
iso_5p
snv
n_isom
irs_cum
1.00
0.75
0.50
0.25
0.00
1
2
3
4
n_rep
protocol
NEBNext
TrueSeq

Summary

- ✓ Each miRNA generates a diversity of isomiRs:
  - ✓ 99% contribute to < 10% of the miRNA abundance
    - ✓ Independent of pipelines and data sets
- ✓ >90% of miRNAs affected
- ✓ Custom 4N protocols perform worst, BUT NEXTFLEX shows good performance

Guidelines

- ✓ NEXFlex/TrueSeq shows less unexpected sequences
- ✓ Removing isomiRs <10% of importance may clean your data
- ✓ Increase your replicates to 6 if possible
- ✓ Don’t filter by sample individually

https://github.com/miRTop/incubator/tree/master/projects/tewari_equimolar

https://github.com/miRTop/mirGFF3

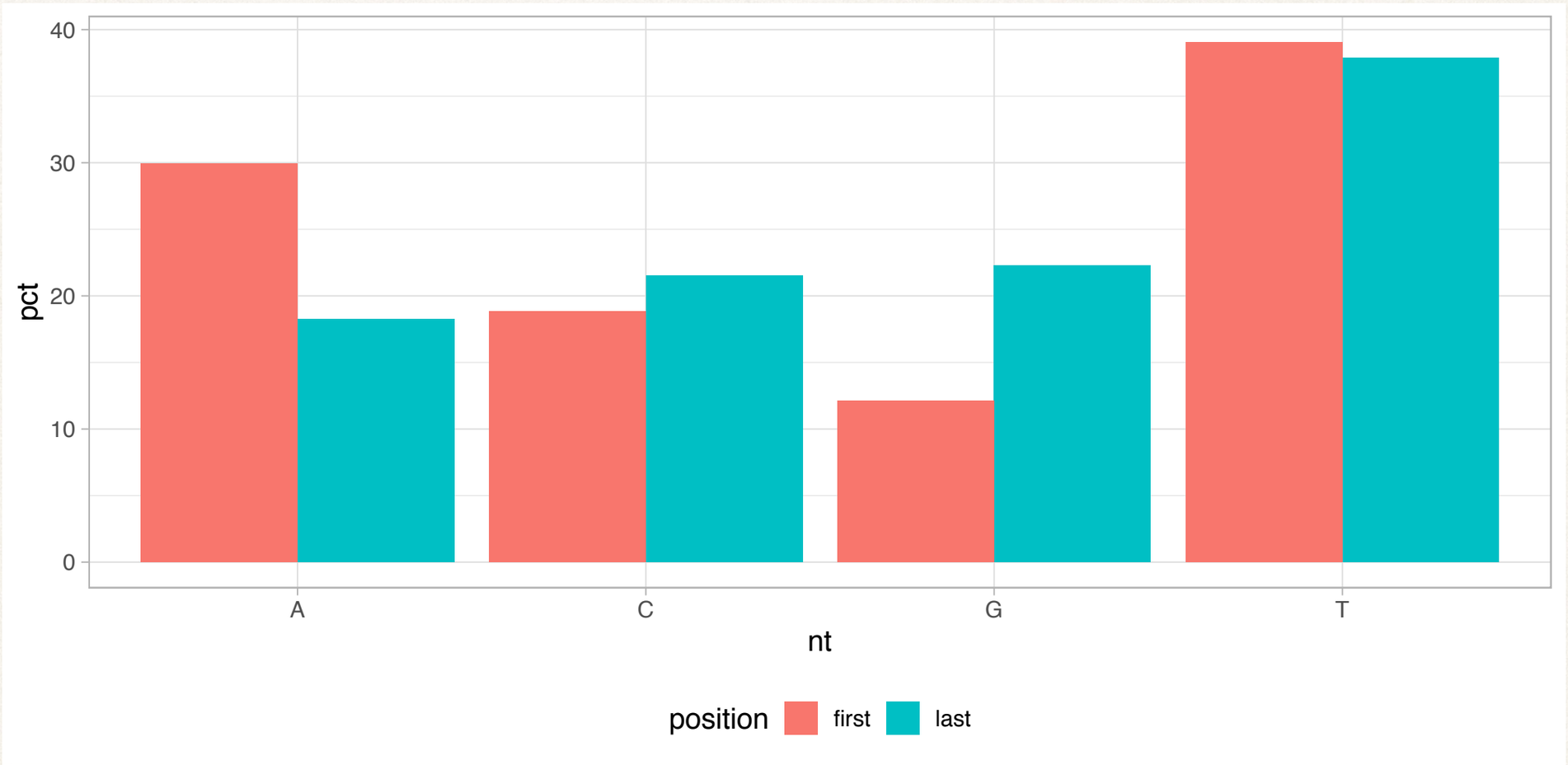
https://github.com/miRTop/mirtop

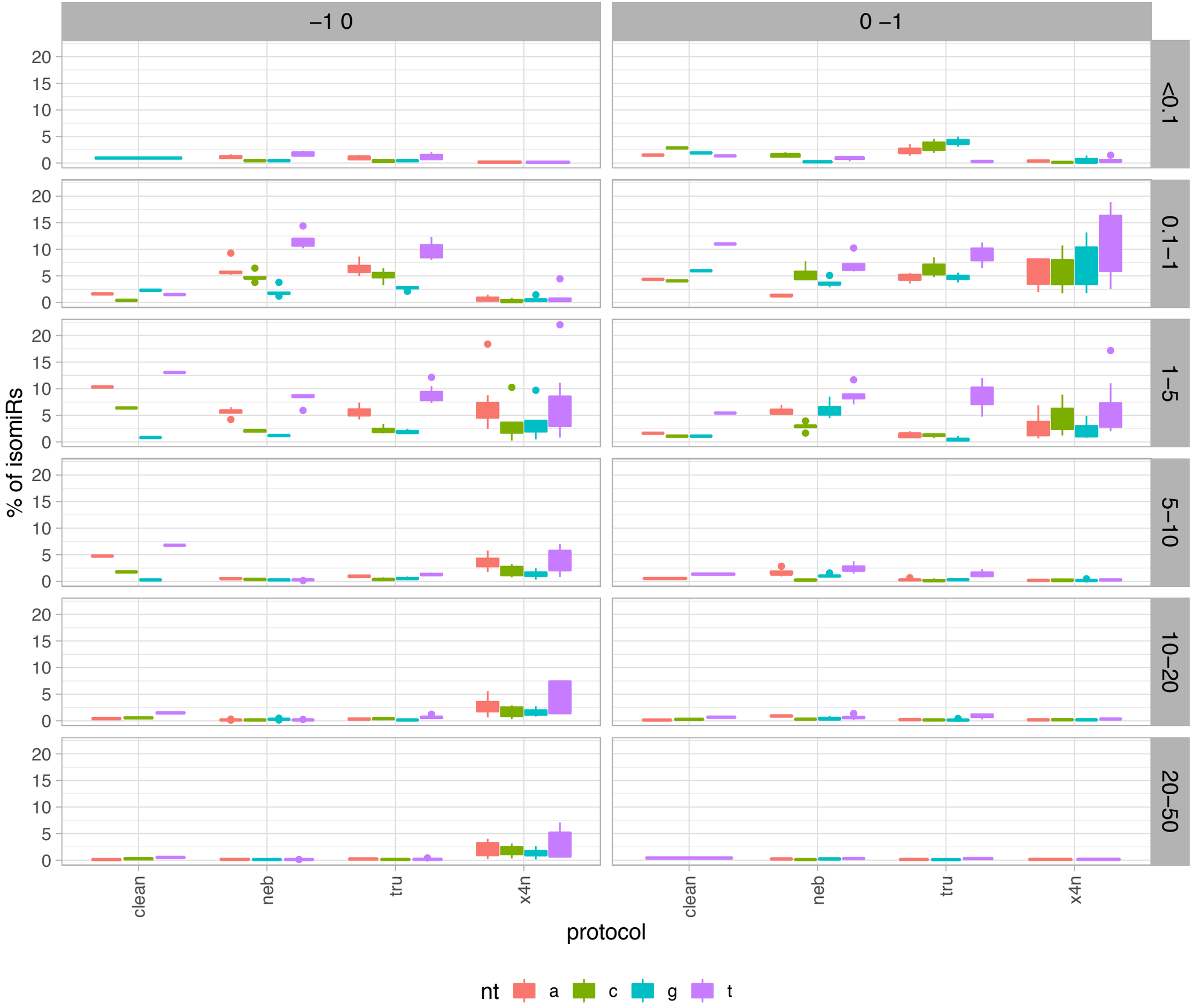
Thomas Desvignes
Phillipe Loher
Karen EIlbeck
Bastian Fromm
Gianvito Urgese
Isidore Rigoutsos
Michael Hackenberg
Ioannis S. Vlachos
Marc K. Halushka

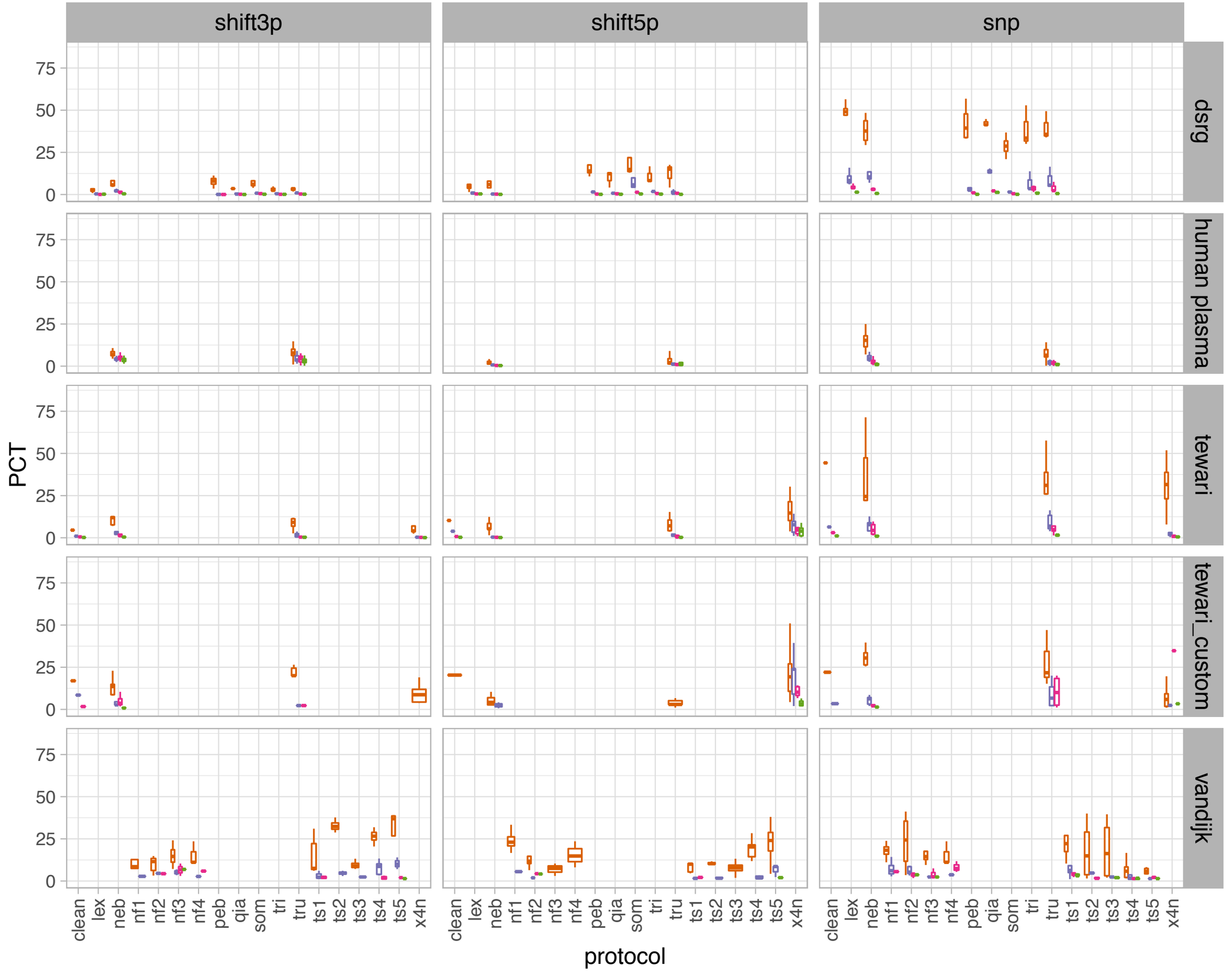
Jeffery Ma, Jason Sydes, Yin Lu, Ernesto Aparicio-Puerta,
Shruthi Bandyadka, Victor Barrera, Peter Batzel,
Rafa Allis, Roderic Espin

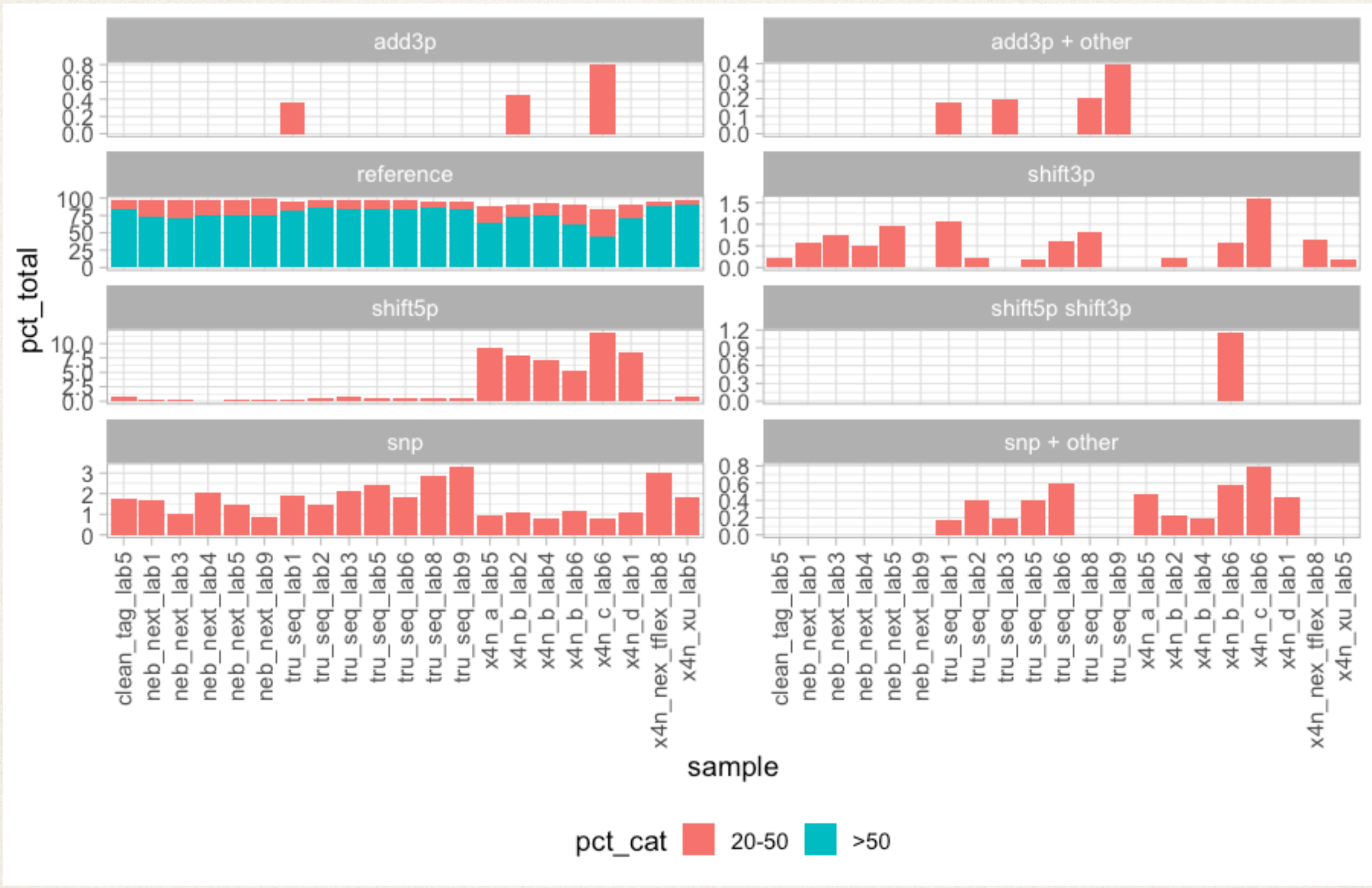
Kieran O'Neill, Eric Londin, Aristeidis G. Telonis,
Elisa Ficarra, John H. Postlethwait,

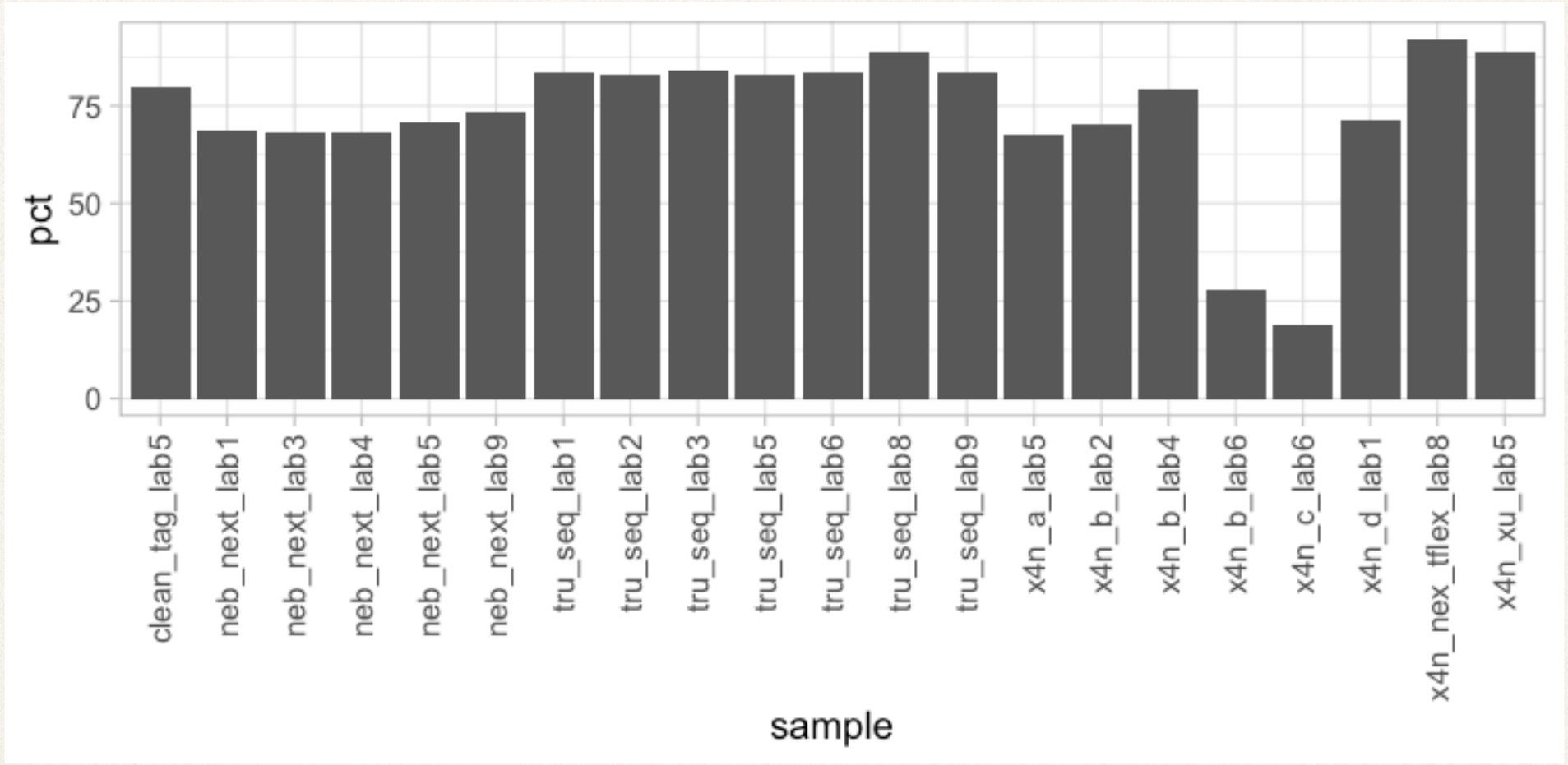
thank you!


40
30
20
10
0
pct
A
C
G
T
nt
position
first
last


−1 0
0 −1
<0.1
0.1–1
1–5
5–10
10–20
20–50
% of isomiRs
0
5
10
15
20
clean
neb
tru
x4n
protocol
nt
a
c
g
t


shift3p
shift5p
snp
dsrg
human plasma
tewari
tewari_custom
vandijk
PCT
0
25
50
75
clean
lex
neb
nf1
nf2
nf3
nf4
peb
qia
som
tri
tru
ts1
ts2
ts3
ts4
ts5
x4n
protocol
IMPORTANCE
1–5
5–10
10–20
20–50


add3p
add3p + other
reference
shift3p
shift5p
shift5p shift3p
snp
snp + other
pct_total
sample
pct_cat
20-50
>50
clean_tag_lab5
neb_next_lab1
neb_next_lab3
neb_next_lab4
neb_next_lab5
neb_next_lab9
tru_seq_lab1
tru_seq_lab2
tru_seq_lab3
tru_seq_lab5
tru_seq_lab6
tru_seq_lab8
tru_seq_lab9
x4n_a_lab5
x4n_b_lab2
x4n_b_lab4
x4n_b_lab6
x4n_c_lab6
x4n_d_lab1
x4n_nex_tflex_lab8
x4n_xu_lab5


pct
75
50
25
0
clean_tag_lab5
neb_next_lab1
neb_next_lab3
neb_next_lab4
neb_next_lab5
neb_next_lab9
tru_seq_lab1
tru_seq_lab2
tru_seq_lab3
tru_seq_lab5
tru_seq_lab6
tru_seq_lab8
tru_seq_lab9
x4n_a_lab5
x4n_b_lab2
x4n_b_lab4
x4n_b_lab6
x4n_c_lab6
x4n_d_lab1
x4n_nex_tflex_lab8
x4n_xu_lab5
sample