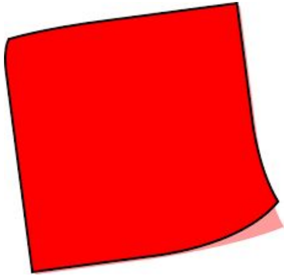


miRTOP: An open source community project for the development of a unified format file for miRNA data

miRTOP Group

https://lpantano.github.io

@lopantano

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

June 27th, 2018

GC BOSC Portland 2018

Team

Thomas Desvignes[2] , Karen Ellbeck[3], Ioannis S. Vlachos[4], Bastian Fromm[5], Marc K. Halushka[6], Michael Hackenberg[7], Gianvito Urgese[8], Elisa Ficarra[8], Shruthi Bandyadka[9], Jason Sydes[2], Peter Batzel[2], John H. Postlethwait[2], Phillipe Loher[10], Eric Londin[10], Aristeidis G. Telonis[10], Isidore Rigoutsos[10], Lorena Pantano[1]

[1] Harvard School TH Chan of Public Health, Boston, MA, USA. Email: lpantano@hsph.harvard.edu

[2] Institute of Neuroscience, University of Oregon, Eugene, OR, USA.

[3] University of Utah, Biomedical Informatics, UT, USA.

[4] Brigham & Women's Hospital, Broad Institute of MIT and Harvard, Harvard Medical School, Cambridge, MA, USA.

[5] Science for Life Laboratory, Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, S-10691 Stockholm, Sweden

[6] Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

[7] Bioinformatics Group. University of Granada, Spain.

[8] Politecnico di Torino, Italy.

[9] Partners Personalized Medicine, Cambridge, MA, USA.

[10] Computational Medicine Center, Thomas Jefferson University, Philadelphia, PA, USA

https://mirtop.github.io

Outline

A. miRNA (data)

B. Current formats for the data

C. Motivation

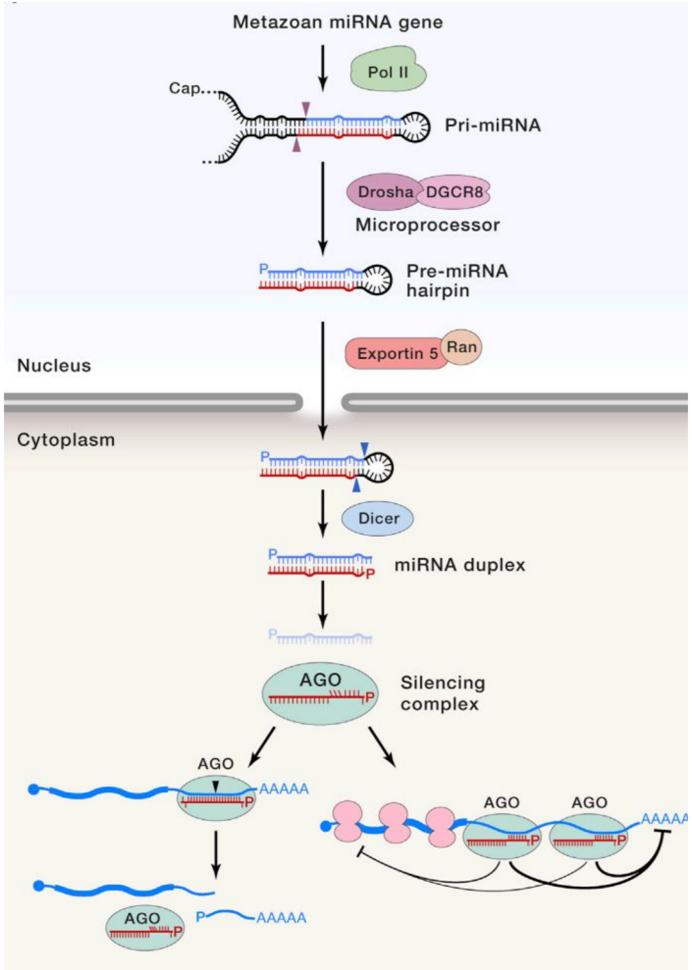
D. Format definition

E. Validation with public data

F. Conclusions

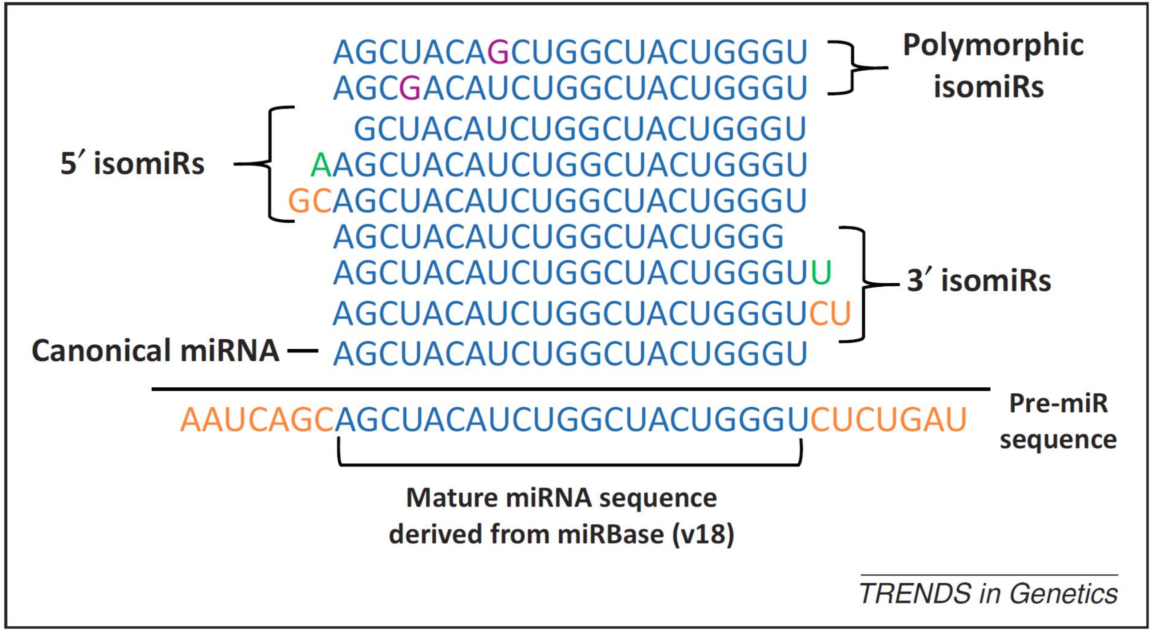
G. Future work

miRNAs



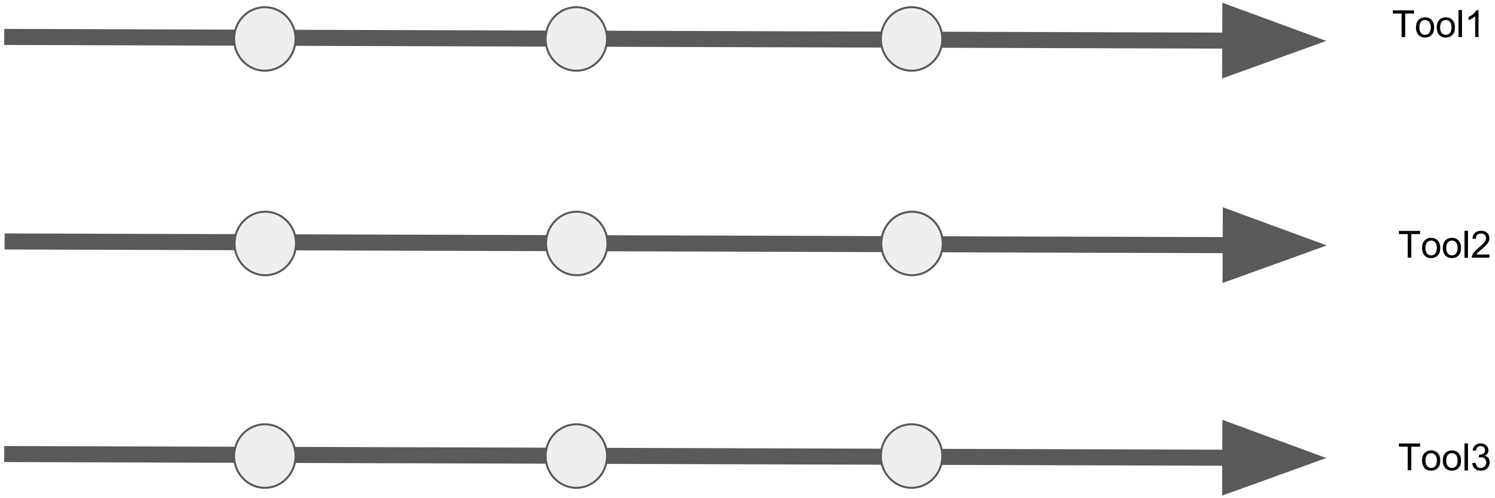
https://doi.org/10.1016/j.cell.2018.03.006

isomiRs

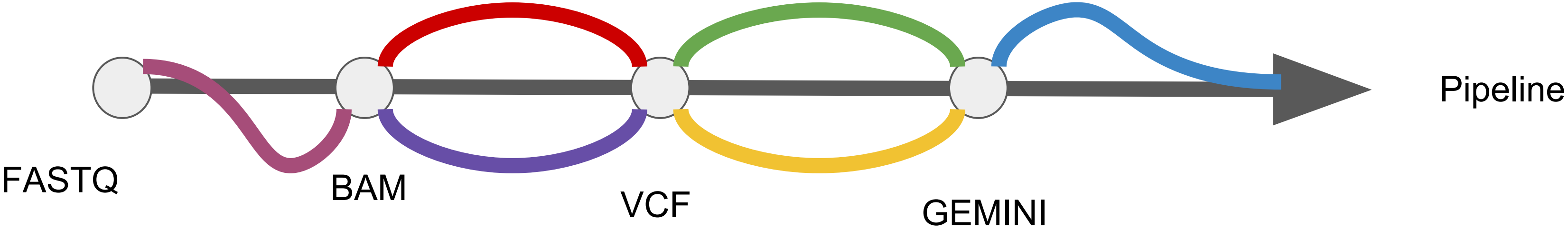


https://doi.org/10.1016/j.tig.2012.07.005

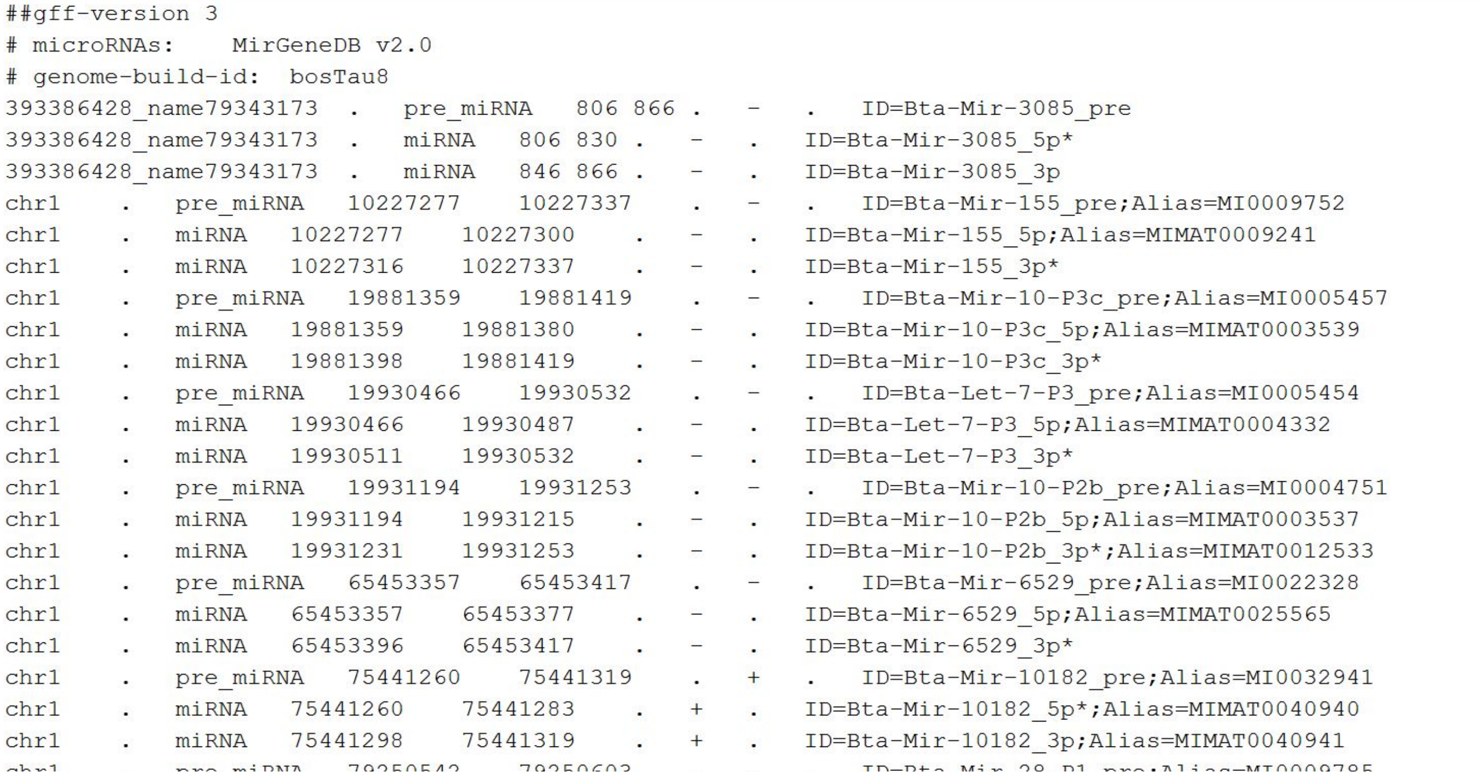
State of the art



Motivation



Features of the format (GFF3)



```
##gff-version 3
# microRNAs:     MirGeneDB v2.0
# genome-build-id:  bosTau8
393386428_name79343173  .   pre_miRNA   806 866 .   -   .   ID=Bta-Mir-3085_pre
393386428_name79343173  .   miRNA   806 830 .   -   .   ID=Bta-Mir-3085_5p*
393386428_name79343173  .   miRNA   846 866 .   -   .   ID=Bta-Mir-3085_3p
chr1    .   pre_miRNA   10227277    10227337    .   -   .   ID=Bta-Mir-155_pre;Alias=MI0009752
chr1    .   miRNA   10227277    10227300    .   -   .   ID=Bta-Mir-155_5p;Alias=MIMAT0009241
chr1    .   miRNA   10227316    10227337    .   -   .   ID=Bta-Mir-155_3p*
chr1    .   pre_miRNA   19881359    19881419    .   -   .   ID=Bta-Mir-10-P3c_pre;Alias=MI0005457
chr1    .   miRNA   19881359    19881380    .   -   .   ID=Bta-Mir-10-P3c_5p;Alias=MIMAT0003539
chr1    .   miRNA   19881398    19881419    .   -   .   ID=Bta-Mir-10-P3c_3p*
chr1    .   pre_miRNA   19930466    19930532    .   -   .   ID=Bta-Let-7-P3_pre;Alias=MI0005454
chr1    .   miRNA   19930466    19930487    .   -   .   ID=Bta-Let-7-P3_5p;Alias=MIMAT0004332
chr1    .   miRNA   19930511    19930532    .   -   .   ID=Bta-Let-7-P3_3p*
chr1    .   pre_miRNA   19931194    19931253    .   -   .   ID=Bta-Mir-10-P2b_pre;Alias=MI0004751
chr1    .   miRNA   19931194    19931215    .   -   .   ID=Bta-Mir-10-P2b_5p;Alias=MIMAT0003537
chr1    .   miRNA   19931231    19931253    .   -   .   ID=Bta-Mir-10-P2b_3p*;Alias=MIMAT0012533
chr1    .   pre_miRNA   65453357    65453417    .   -   .   ID=Bta-Mir-6529_pre;Alias=MI0022328
chr1    .   miRNA   65453357    65453377    .   -   .   ID=Bta-Mir-6529_5p;Alias=MIMAT0025565
chr1    .   miRNA   65453396    65453417    .   -   .   ID=Bta-Mir-6529_3p*
chr1    .   pre_miRNA   75441260    75441319    .   +   .   ID=Bta-Mir-10182_pre;Alias=MI0032941
chr1    .   miRNA   75441260    75441283    .   +   .   ID=Bta-Mir-10182_5p*;Alias=MIMAT0040940
chr1    .   miRNA   75441298    75441319    .   +   .   ID=Bta-Mir-10182_3p;Alias=MIMAT0040941
chr1        pre_miRNA   79250542    79250603                ID=Bta-Mir-28-P1_pre;Alias=MI0009785
```

Adapted to miRNA data

`http://bit.ly/mirtop-gff3`

Column/Attribute	**Value**	**Example**
type	ref_miRNA | isomiR	isomiR
UID	Coded sequence	7II7B6
CIGAR	SAM CIGAR	20MT
Filter	PASS;user_defined	PASS,low-coverage,...
Variants	Variants code	**iso_add:+2,iso_5p:-2**
Changes	Variants code with NTs	iso_3p:tt,iso_add:GTC
Expression	Numeric vector	**3,0,4,0,0**

Mirtop - tool to work with the format

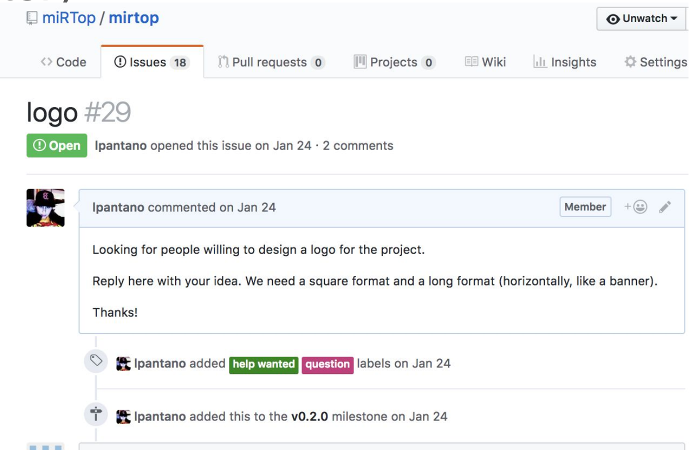
Python tool: (pypi and bioconda)

Importer: Helps to generate the GFF file

Exporter: Helps to generate inputs for downstream analysis

Helpers: Helps to work with the GFF file (stats, join, filter)

`https://gitter.im/mirtop/Lobby`



logo #29

Open lpantano opened this issue on Jan 24 · 2 comments

lpantano commented on Jan 24

Looking for people willing to design a logo for the project.

Reply here with your idea. We need a square format and a long format (horizontally, like a banner).

Thanks!

lpantano added help wanted question labels on Jan 24

lpantano added this to the v0.2.0 milestone on Jan 24

Mirtop - Current supported tools

`http://bit.ly/mirtop-dev`

A. BAM
B. Bcbio (seqbuster)
C. sRNAbench
D. isomiR-SEA
E. miRge2.0
F. PROST!

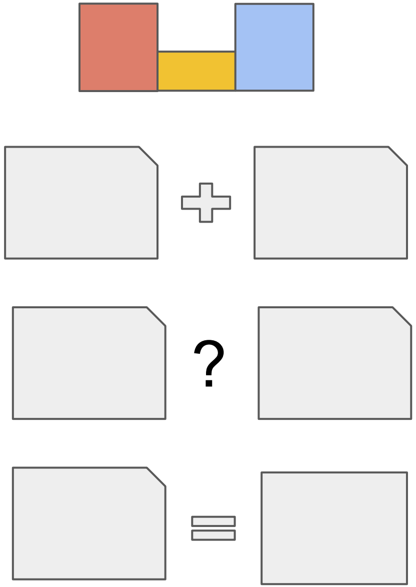
Mirtop - current supported actions

Stats

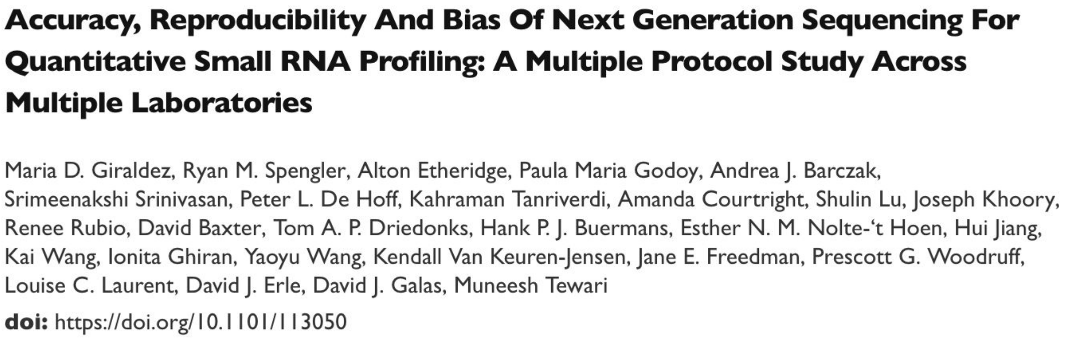
Join

Compare

Counts



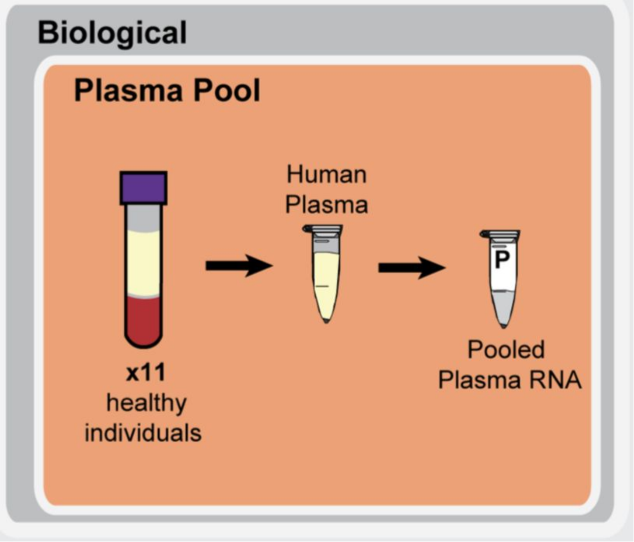
Application



Accuracy, Reproducibility And Bias Of Next Generation Sequencing For Quantitative Small RNA Profiling: A Multiple Protocol Study Across Multiple Laboratories

Maria D. Giraldez, Ryan M. Spengler, Alton Etheridge, Paula Maria Godoy, Andrea J. Barczak, Srimeenakshi Srinivasan, Peter L. De Hoff, Kahraman Tanriverdi, Amanda Courtright, Shulin Lu, Joseph Khoory, Renee Rubio, David Baxter, Tom A. P. Driedonks, Hank P. J. Buermans, Esther N. M. Nolte-‘t Hoen, Hui Jiang, Kai Wang, Ionita Ghiran, Yaoyu Wang, Kendall Van Keuren-Jensen, Jane E. Freedman, Prescott G. Woodruff, Louise C. Laurent, David J. Erle, David J. Galas, Muneesh Tewari

doi: https://doi.org/10.1101/113050



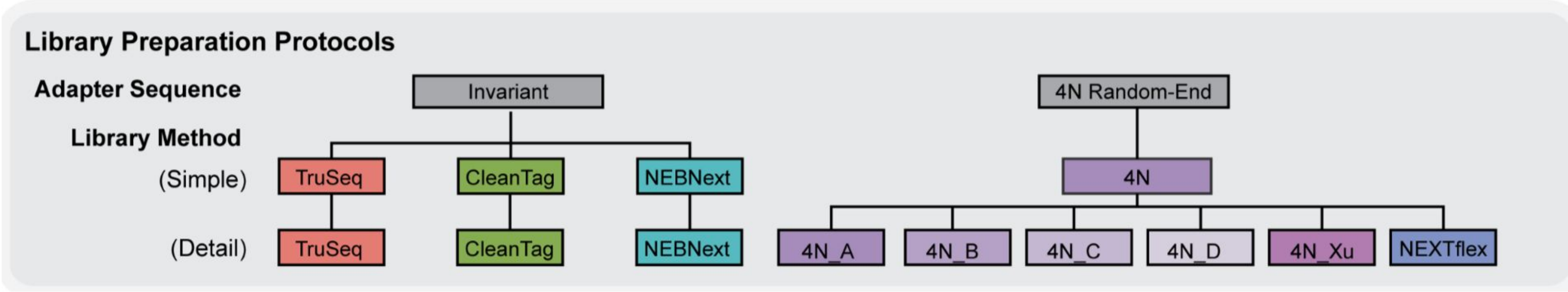
Biological

Plasma Pool

x11 healthy individuals

Human Plasma

Pooled Plasma RNA



Library Preparation Protocols

Adapter Sequence

Library Method

(Simple)

(Detail)

Invariant

TruSeq

CleanTag

NEBNext

TruSeq

CleanTag

NEBNext

4N Random-End

4N

4N_A

4N_B

4N_C

4N_D

4N_Xu

NEXTflex

Measure isomiR consistency

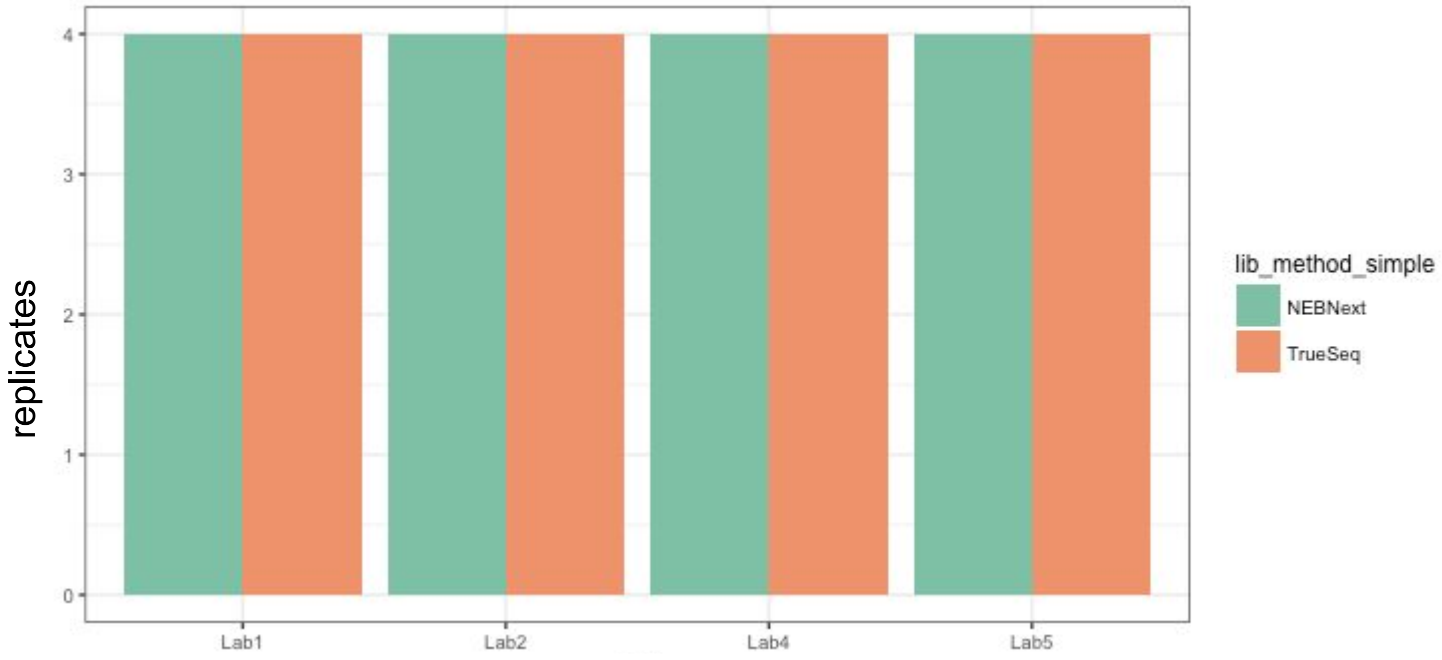
- How reproducible are the sequences coming from same sample,
  - same lab and same protocol
  - same lab but different protocol
  - different lab but same protocol
- Filtering conditions that increase reproducibility
- Benchmarking of tools

http://bit.ly/mirtop-tewari-data

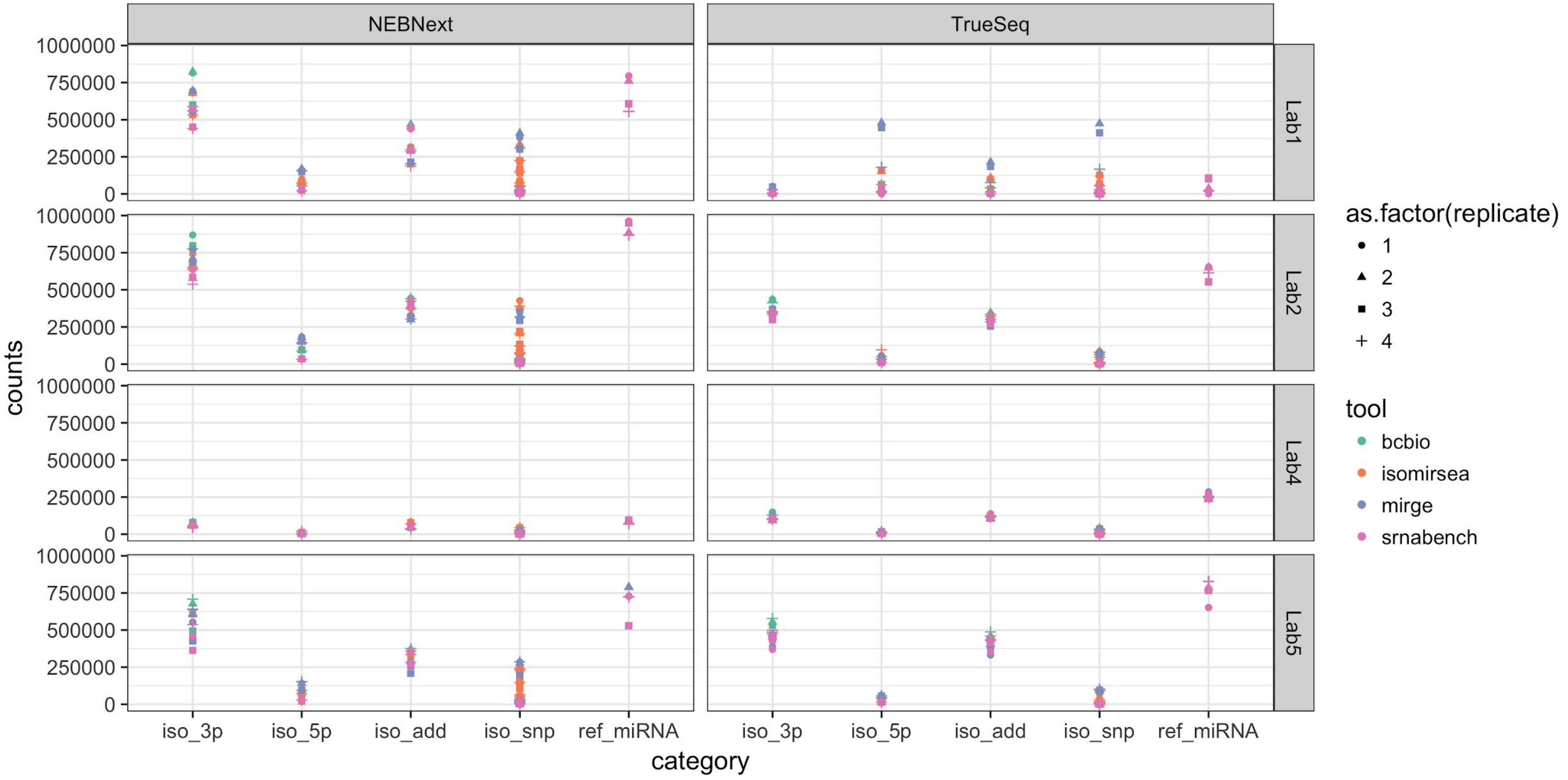
Limitations

1. miRge2.0 did the trimming by itself
2. bcbio did the trimming by itself and shared the trimmed files to be used by isomiR-SEA and sRNAbench
3. bcbio has a internal cutoff of a minimum of 2 counts to be annotated
4. isomiR-SEA considers only iso_5p:+/-1
5. sRNAbench labels some sequences as **mv** for isomiRs, these are lost for now in the conversion to GFF3 format.

Results - Pilot selection



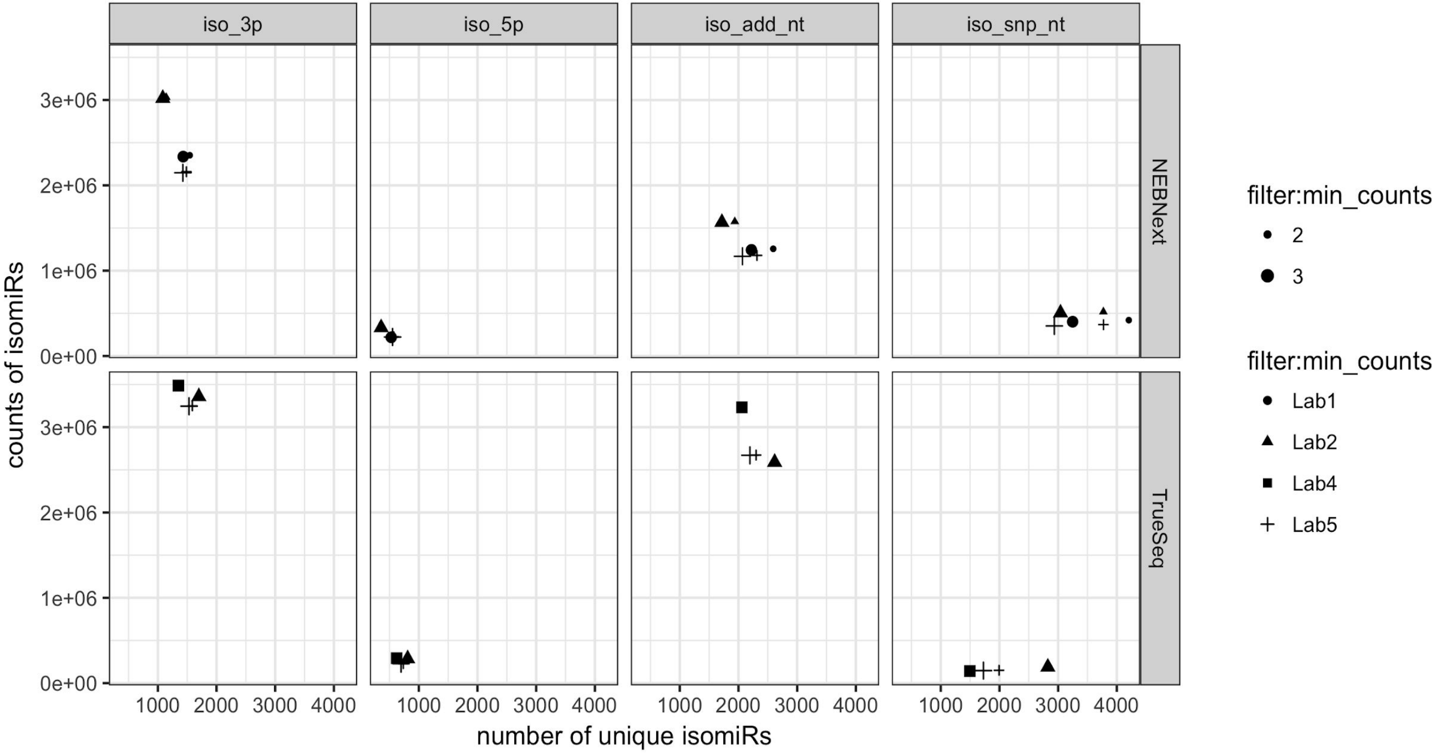
Results - Total number of miRNA reads



Results - Lab reproducibility (one tool)

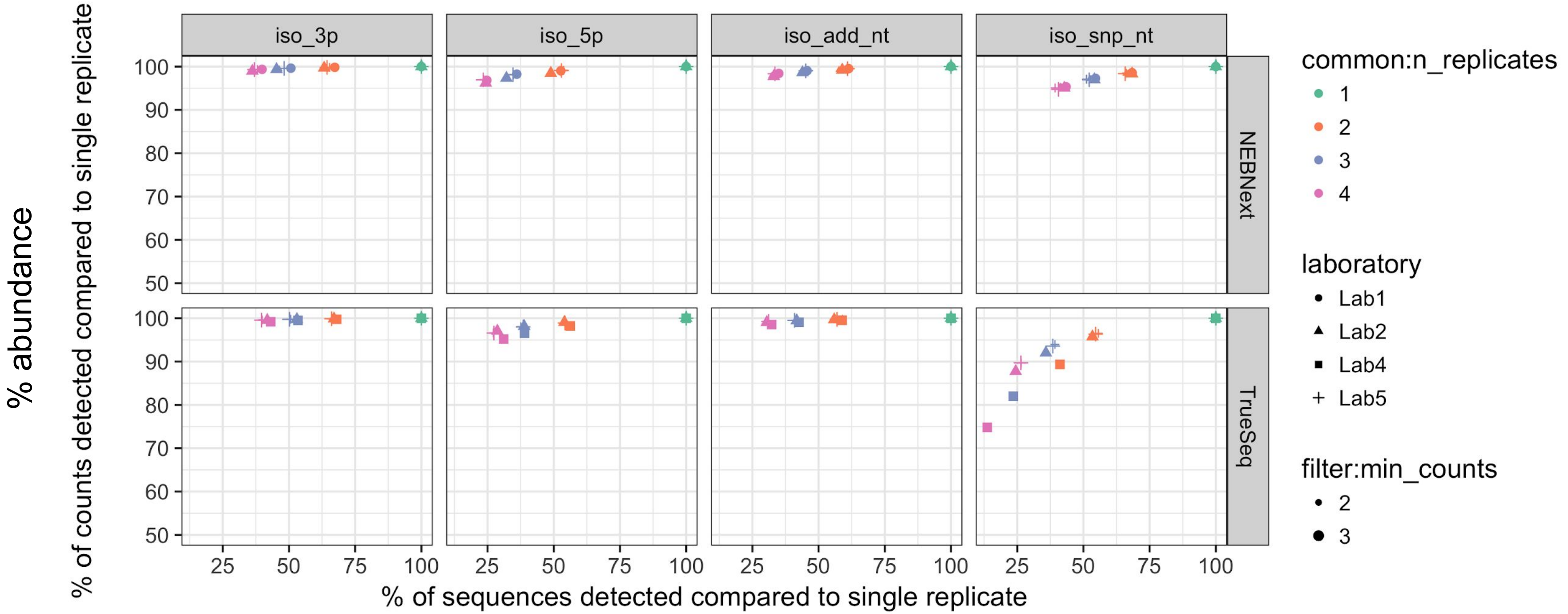
complexity

abundance



Results - Lab reproducibility (one tool)

% complexity



Conclusion up to now

- Diversity on coverage among lab and protocol
- Iso_add and Iso_3p are the most abundant
- 25% of unique sequences are detected by all 4 replicates for each lab/protocol
- 90% of the reads are detected by all 4 replicates for each lab/protocol
- Iso_snp are the most variable, probably due to error in sequencing
- TrueSeq is less reproducible than NEBNext for iso_snp.
- Tools are consistently showing similar general results

Future direction

- Definition of reproducible isomiRs using replicates
- Comparison among labs
- Comparison among protocols
- Comparison among tools
- Definition of a filter to remove non reliable sequences

Everybody is welcome: `http://bit.ly/mirtop-tewari-data`

- Standardized isomiR reporting across all alignment tools
- The effect of incorrect annotations in public repositories

Everybody is welcome: `http://bit.ly/mirtop-incubator`

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

UNIVERSITY OF
OREGON

GCC BOSC
Portland 2018

VE RI TAS
Harvard Medical School
Research Computing

UNIVERSITAS · GRANATENSIS · 1531
PLUS ULTRA
UNIVERSIDAD
DE GRANADA

Jefferson
HOME OF SIDNEY KIMMEL MEDICAL COLLEGE
TM

JOHNS HOPKINS
SCHOOL *of* MEDICINE

U HEALTH
UNIVERSITY OF UTAH

BRIGHAM HEALTH
BWH
BRIGHAM AND
WOMEN'S HOSPITAL

SciLifeLab

BROAD
INSTITUTE

POLITECNICO DI TORINO
1859 1906

NIH National Institutes of Health

NSF