Benchmarks of kallisto bus and bustools

Caltech Bioinformatics Symposium
February 14, 2019

kallisto single-cell family (sort by the last names)











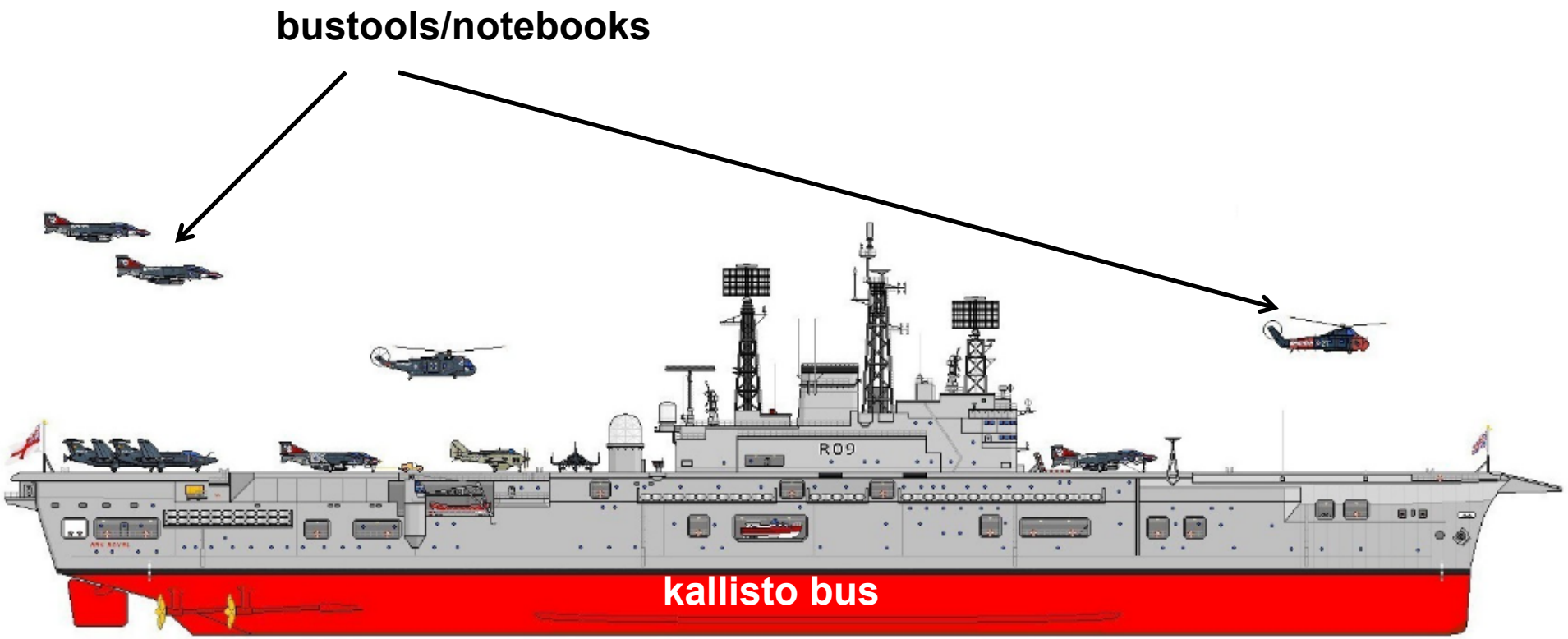

- Outline

- Benchmarks of kallisto bus

- Benchmarks of bustools/notebooks for single-cell analysis

- Detecting species-mixture using kallisto single-cell workflow

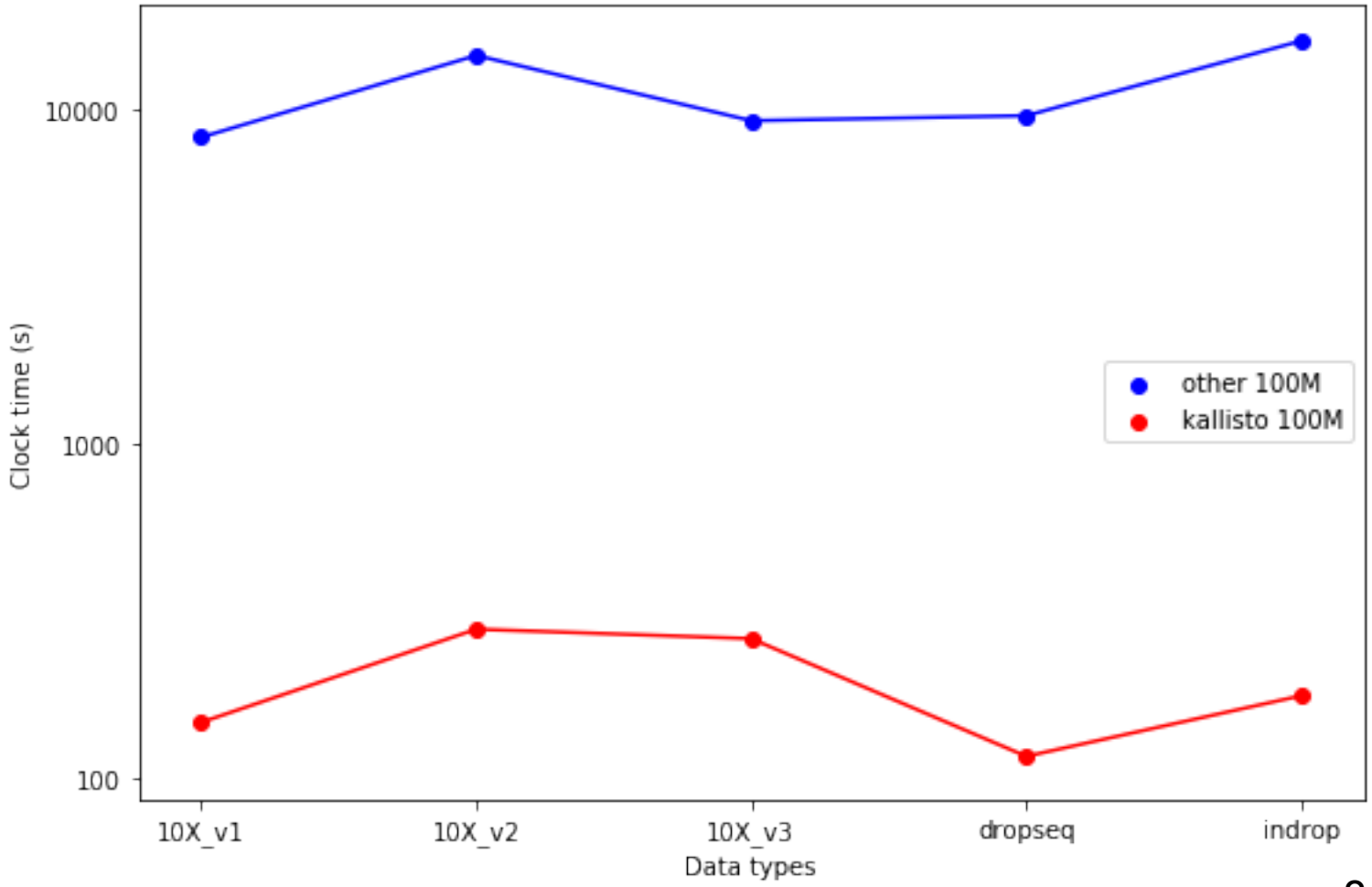
Pachter lab single-cell bioinformatics



Part I

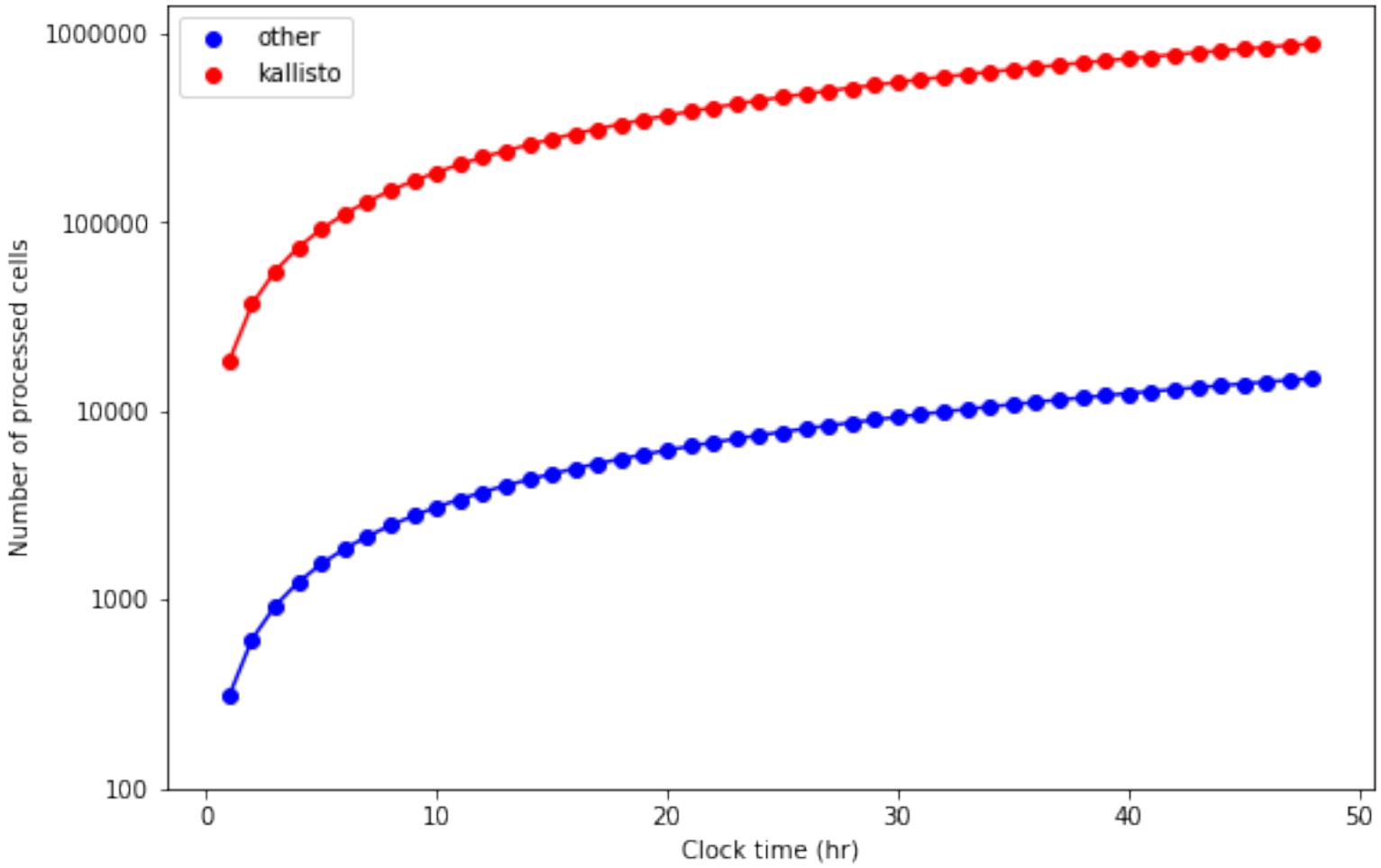
Benchmarks of kallisto bus

Benchmarks of clock time



8-thread

Number of single cells recovered (estimated)

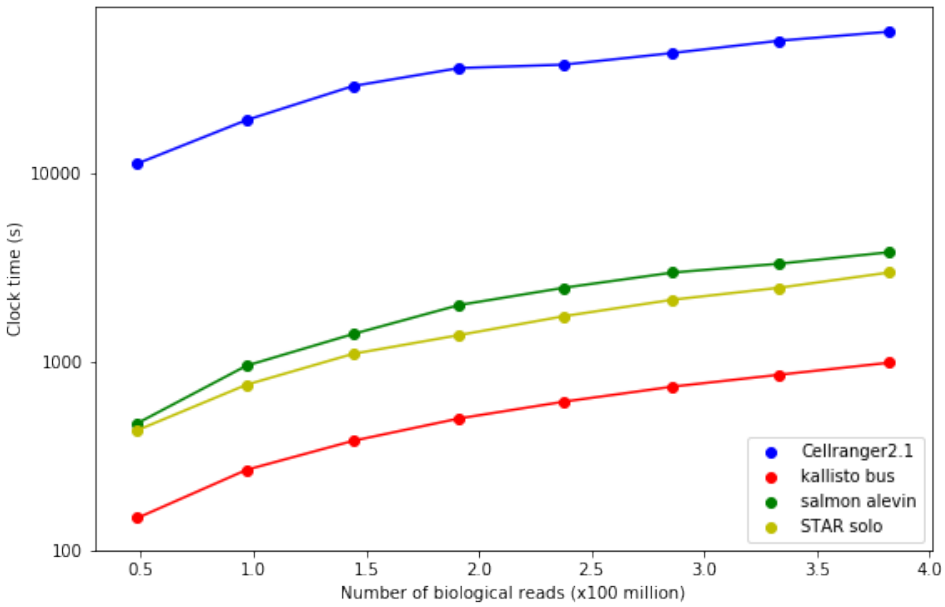


4,000 cells expected from 400M raw reads

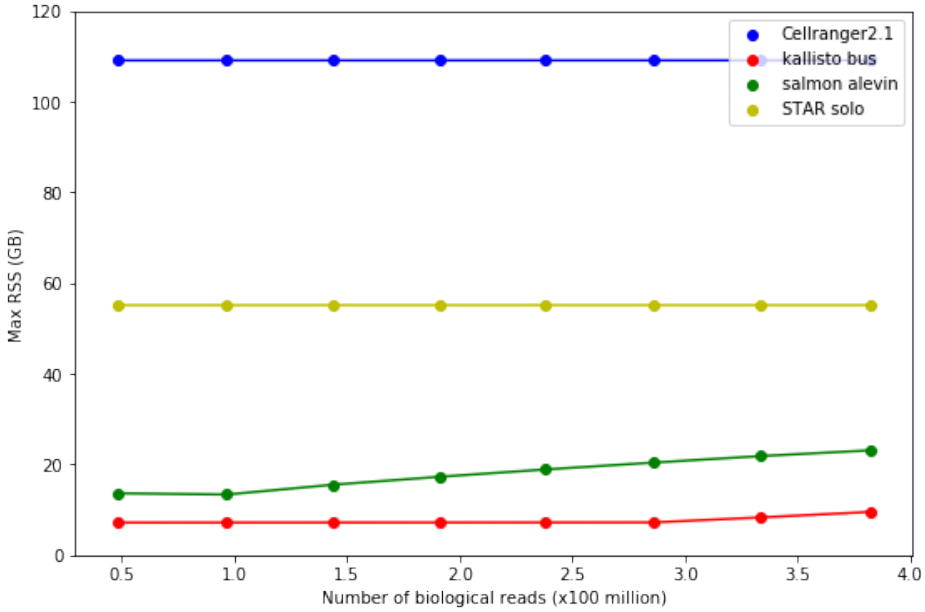
8-thread

Performance at different sequencing depth

clock time



memory usage



8-thread

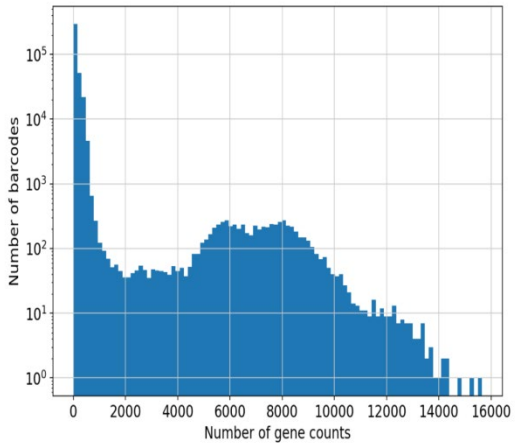
Part II

Benchmarks of bustools/ notebooks for single-cell analysis

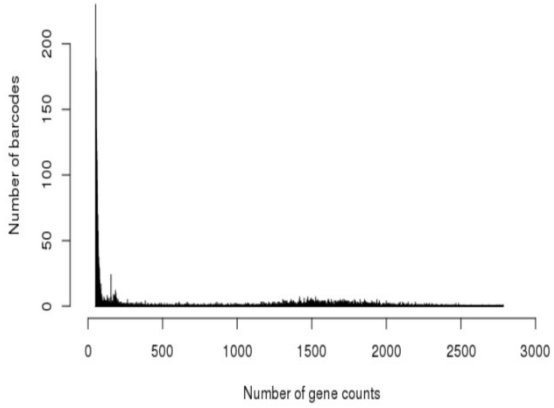
BUS notebooks
– Interactive computing protocols


```
In [27]: bcv = [x for b,x in barcode_hist.items() if x > 0]
         _ = plt.hist(bcv,bins=100, log=True)
         plt.rcParams["figure.figsize"] = [9,6]
         plt.xlabel("Number of gene counts")
         plt.ylabel("Number of barcodes")
         print(len(bcv))
```

```
381648
```

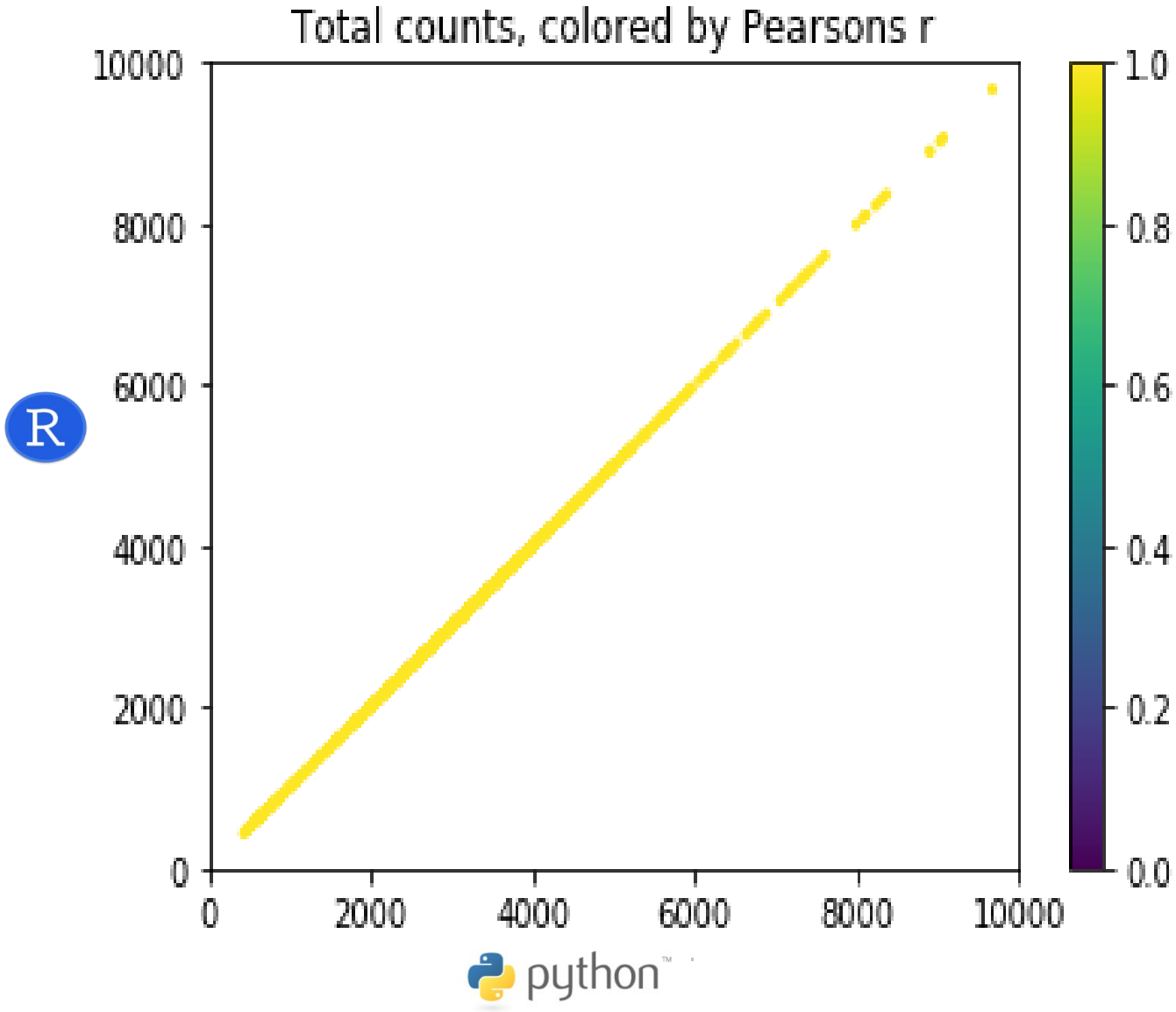



```
36 ```{r}
37 tot_umis <- Matrix::colSums(res_mat)
38 tot_genes <- Matrix::colSums(res_mat>0)
39 hist(tot_genes[tot_genes > 50], breaks = 100000, xlim=c(0,3000), main=NULL,
40      xlab="Number of gene counts", ylab="Number of barcodes")
41 ```
```



BUSparsers
– Batch processing

Consistency of gene quantification workflows



Barcode error correction

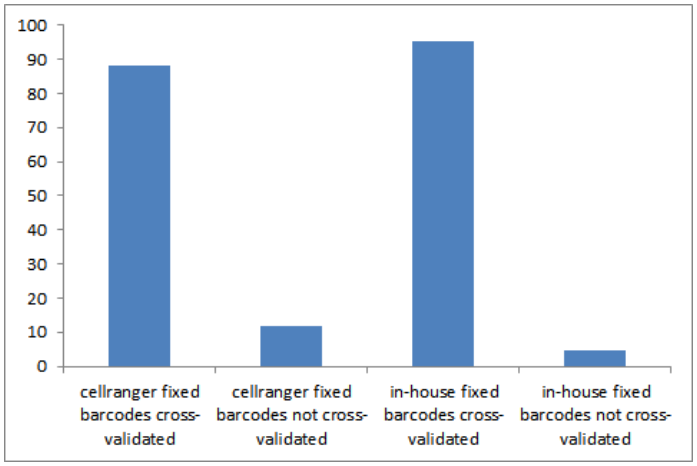
Input fastq:
@ST-K00126:308:HFLYFBBXX:1:1101:20496:1279 1:N:0:NGATGCAT
TGGCCAGCACGACACGTTGTCTTTTT
+
AAFFFJJFFJJF-AAJJAJJJFJJJJ

The in-house tool finds following potential references (hamming distance 1) from 10xV2 whitelist (737,280 barcodes):
TGGCCAGCAAGACACG (a)
TGGCCAGCACGAAACG (b)
TGGCCAGCACGACTCG (c)

Then the tool checks the base quality at positions 10 (a), 13 (b), 14 (c) and finds J, -, A; then position 13 of the raw reads is converted from C to A.

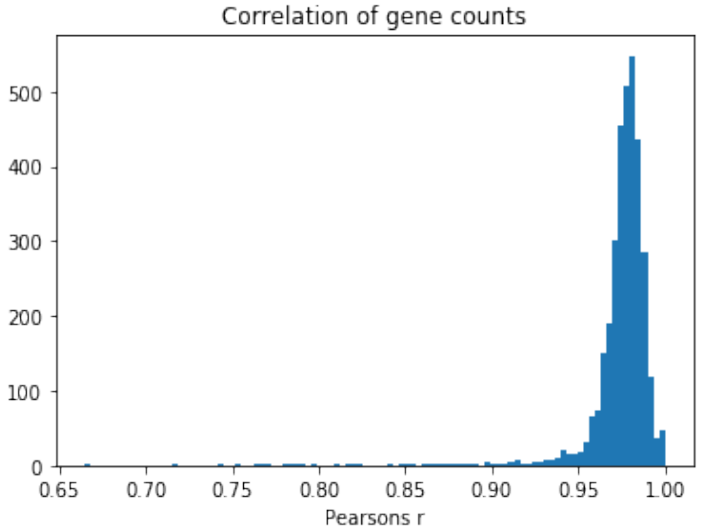
Below is the output fastq:
@ST-K00126:308:HFLYFBBXX:1:1101:20496:1279 1:N:0:NGATGCAT
TGGCCAGCACGAAACGTTGTCTTTTT
+
AAFFFJJFFJJF-AAJJAJJJFJJJJ

%



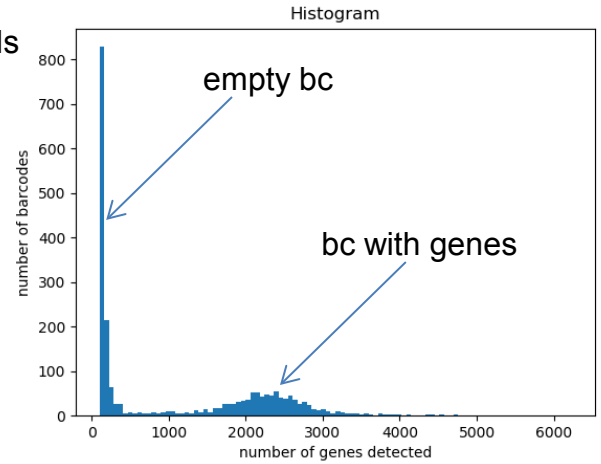
1.3% BC fixed

We are still optimizing this step!



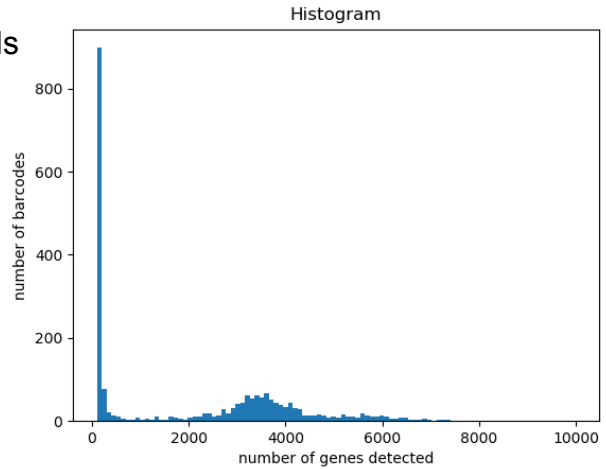
kallisto single-cell pipeline works with both 10X v2 & v3 data

~39M reads



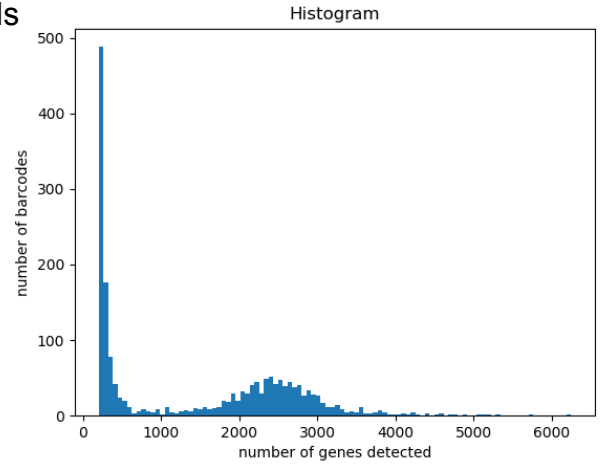
10X PBMC(1k_v2) data

~33M reads



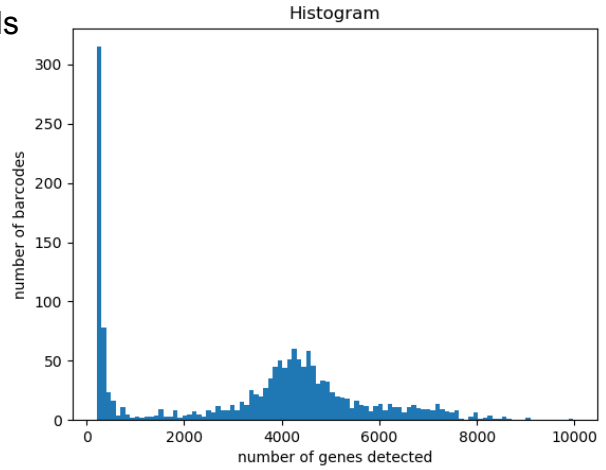
10X PBMC(1k_v3) data

~77M reads



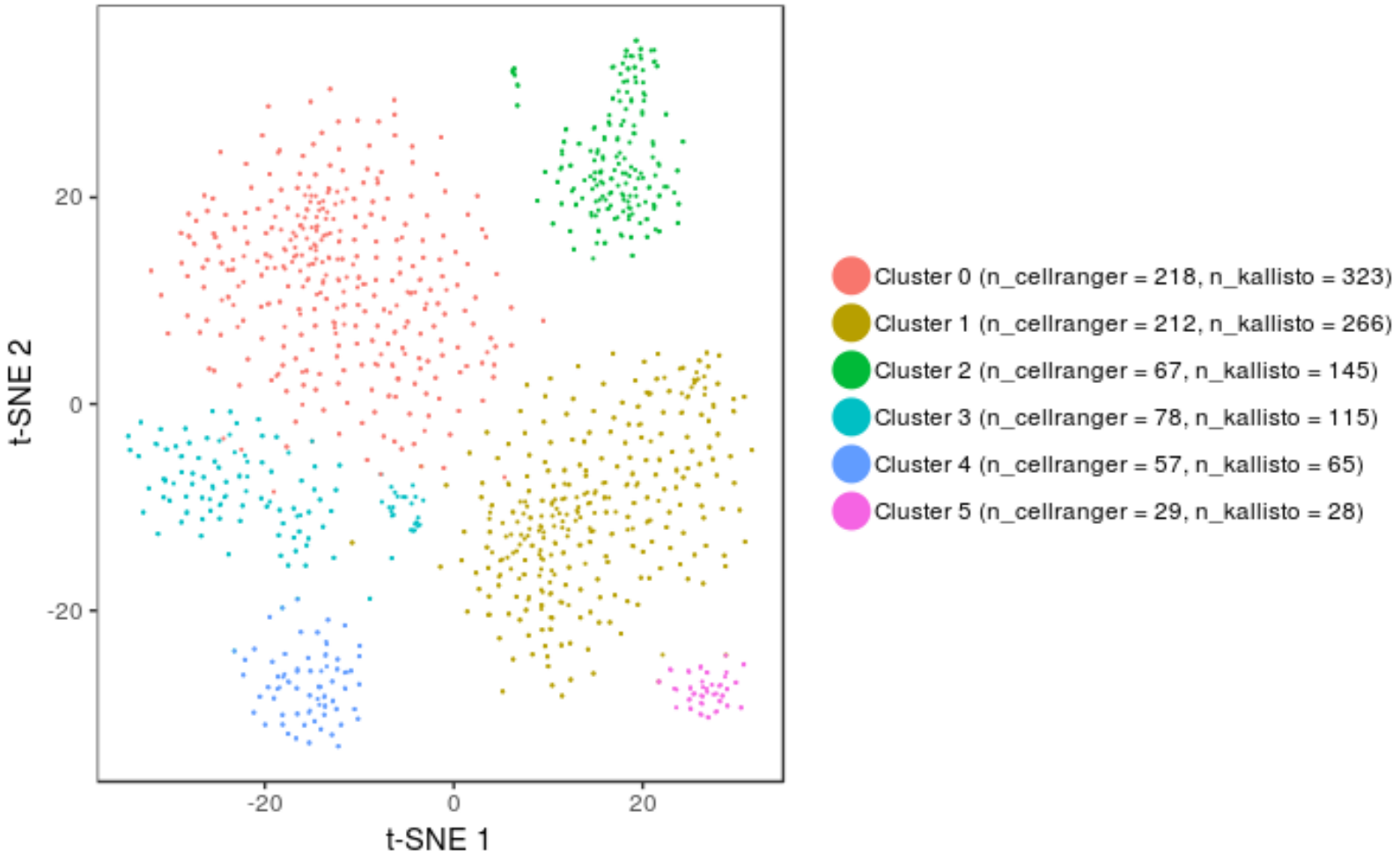
10X PBMC(1k_v2) data

~67M reads



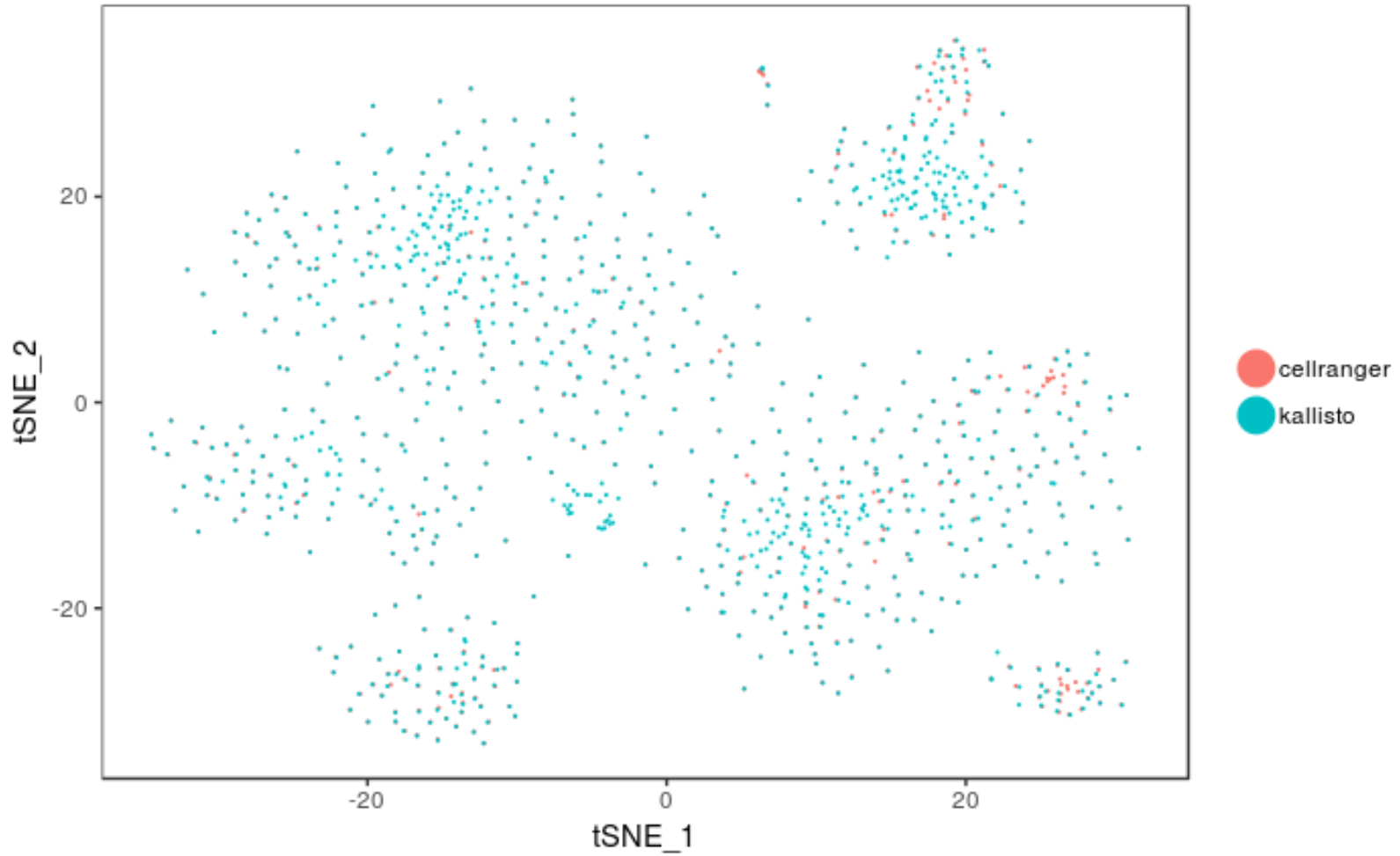
10X PBMC(1k_v3) data

Merge of cellranger & kallisto count matrices for clustering



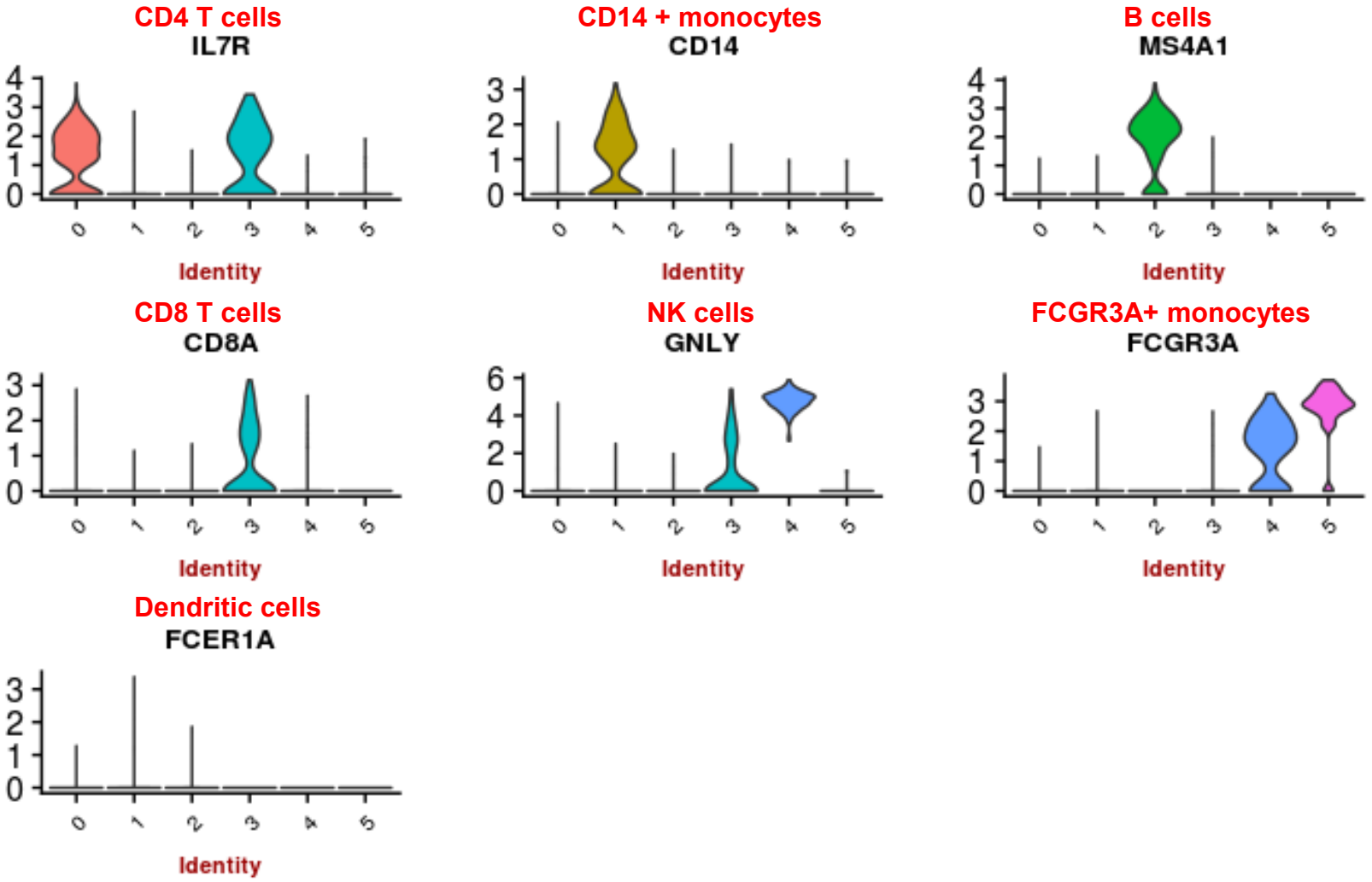
10X PBMC(1k_v2) data
CCA subspace alignment

Overlay of cellranger & kallisto cells



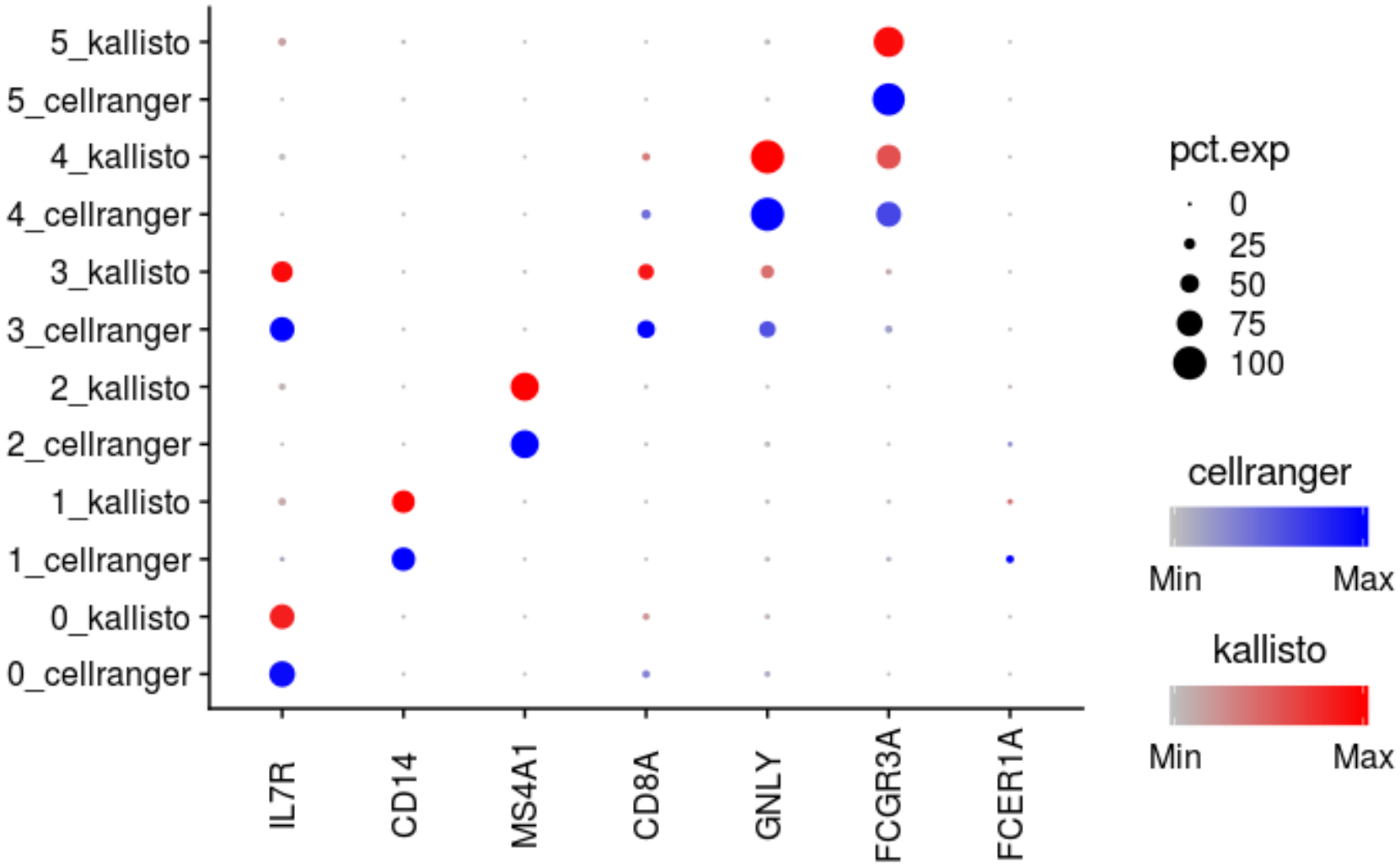
10X PBMC(1k_v2) data

Cell identity revealed by signature genes



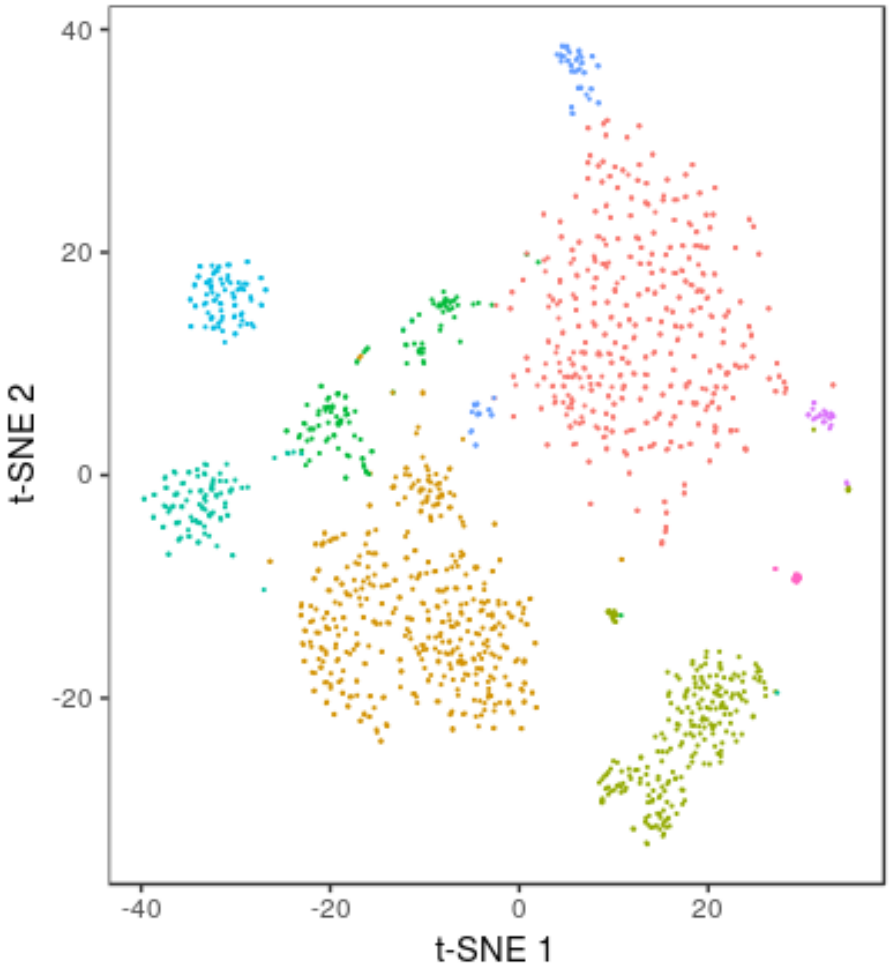
10X PBMC(1k_v2) data

Percentage of cellranger & kallisto cells expressing a marker gene



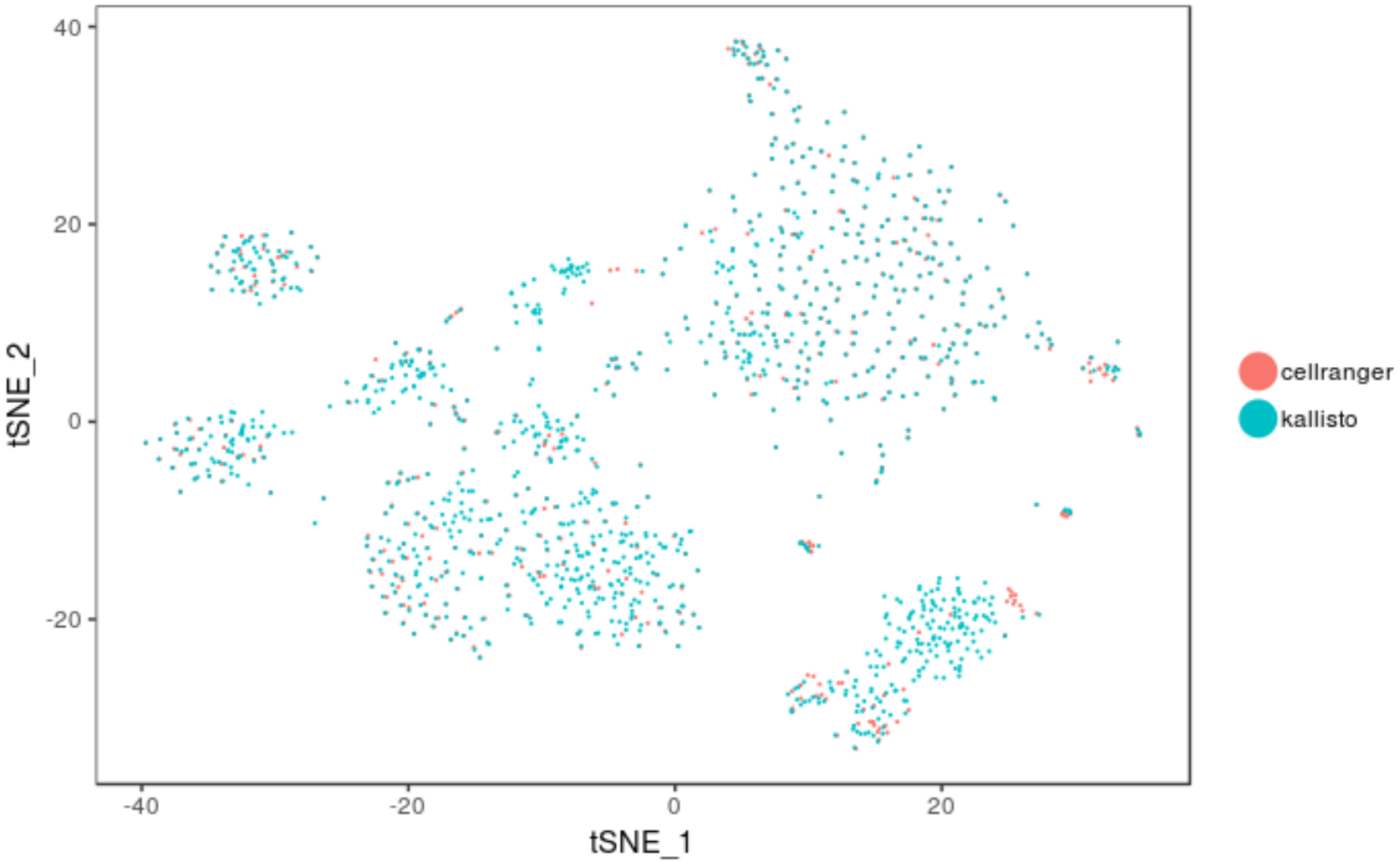
10X PBMC(1k_v2) data

Merge of cellranger & kallisto count matrices for clustering



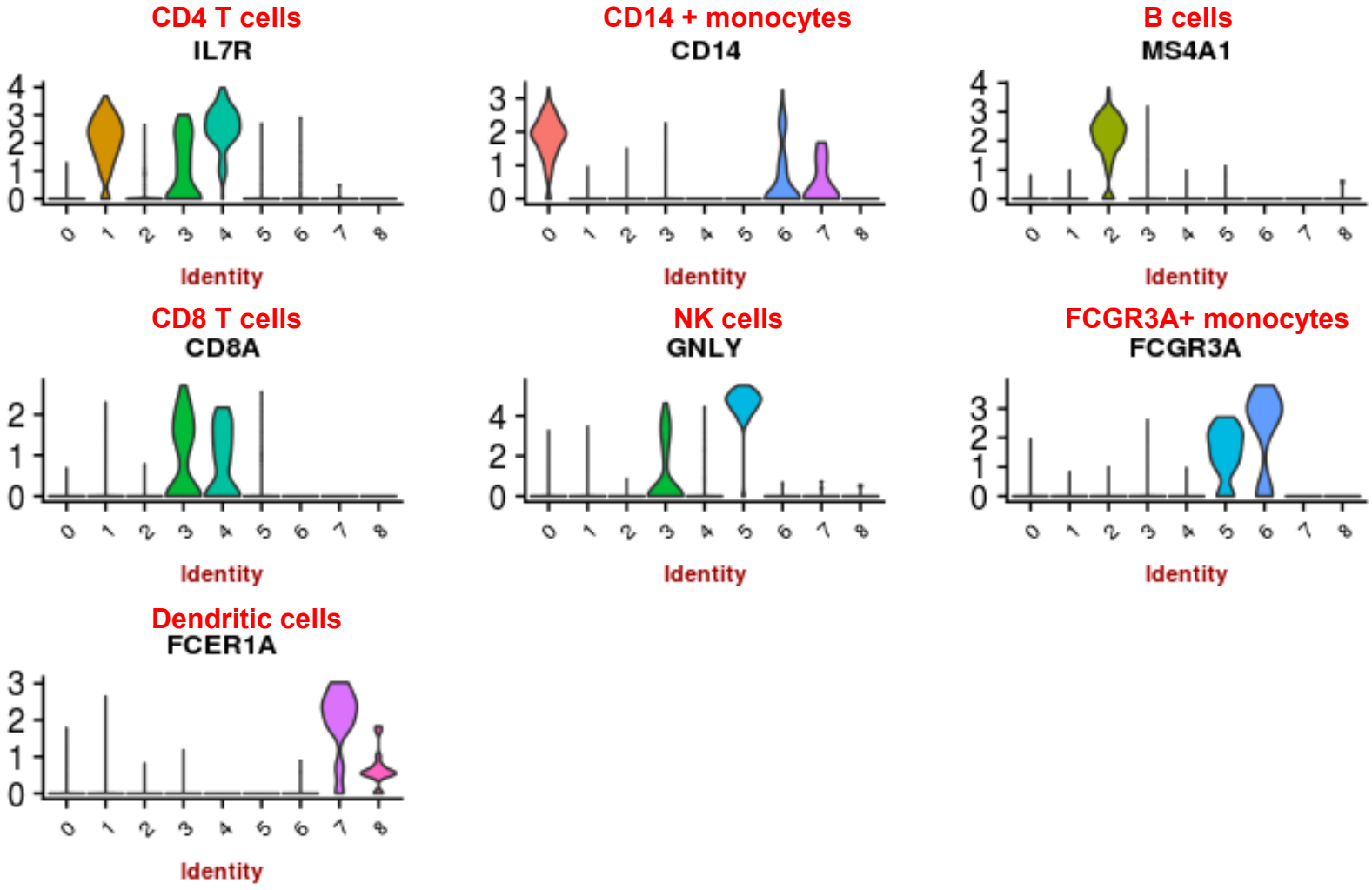
10X PBMC(1k_v3) data
CCA subspace alignment

Overlay of cellranger & kallisto cells



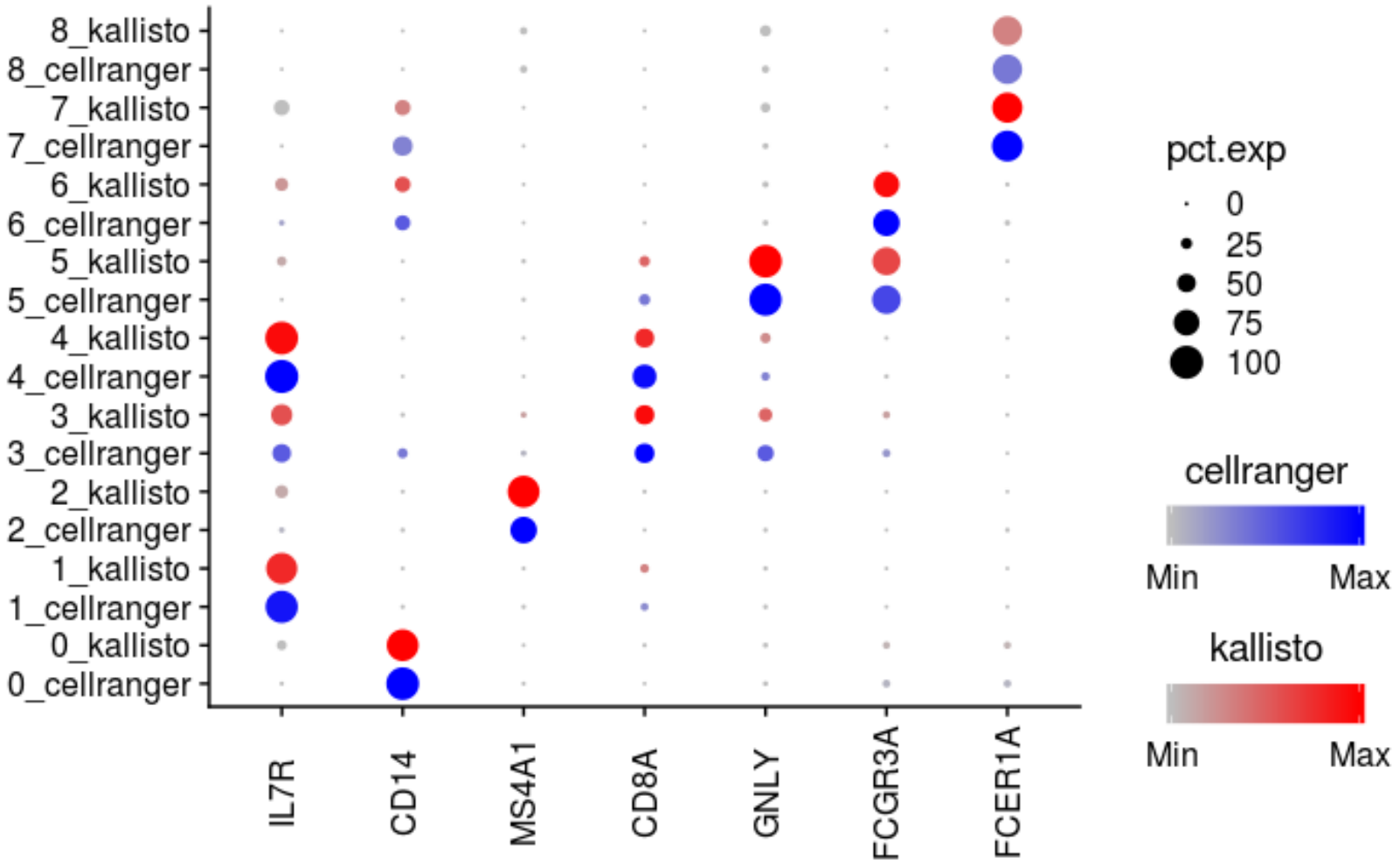
10X PBMC(1k_v3) data

Cell identity revealed by marker genes



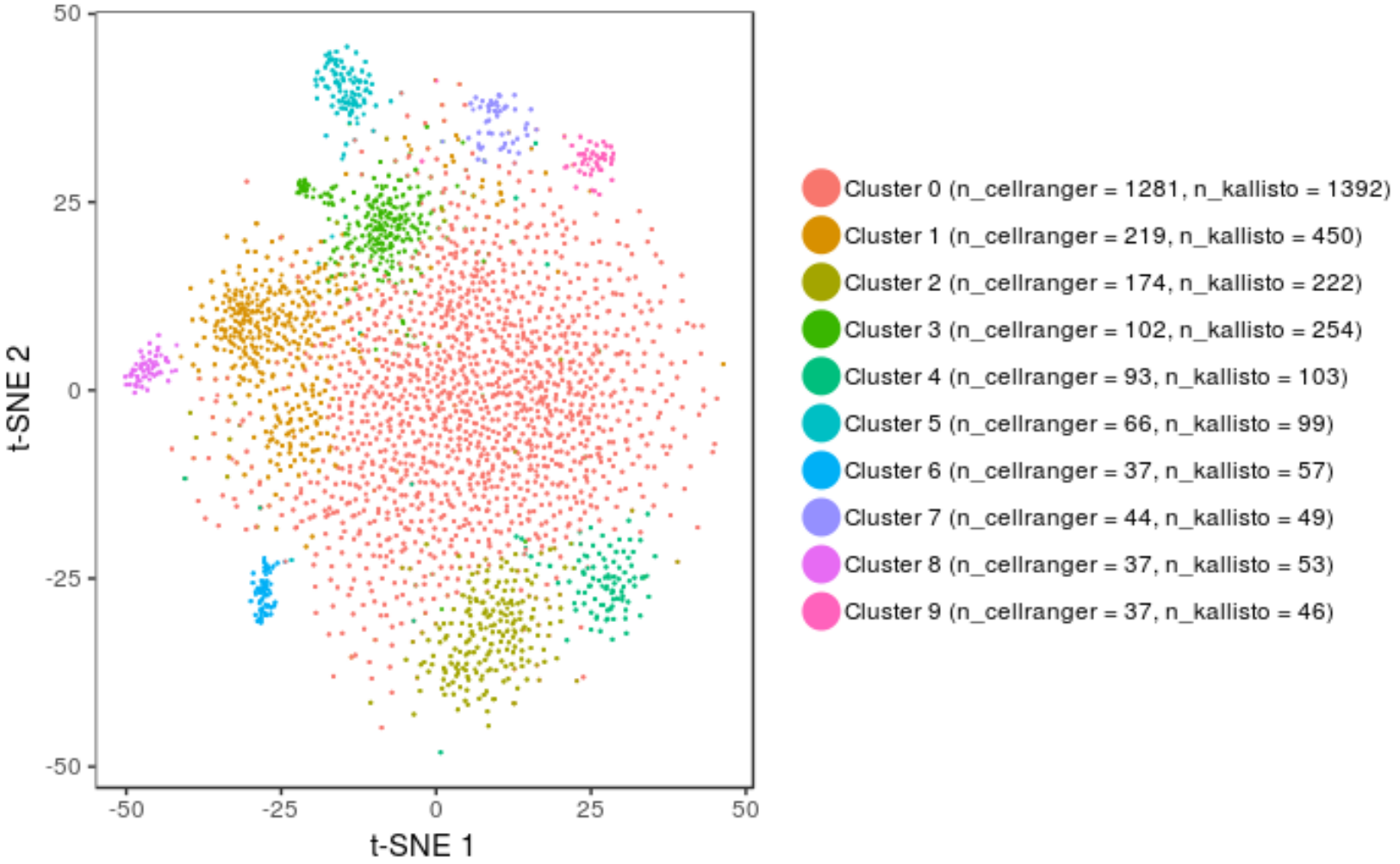
10X PBMC(1k_v3) data

Percentage of cellranger & kallisto cells expressing a marker gene



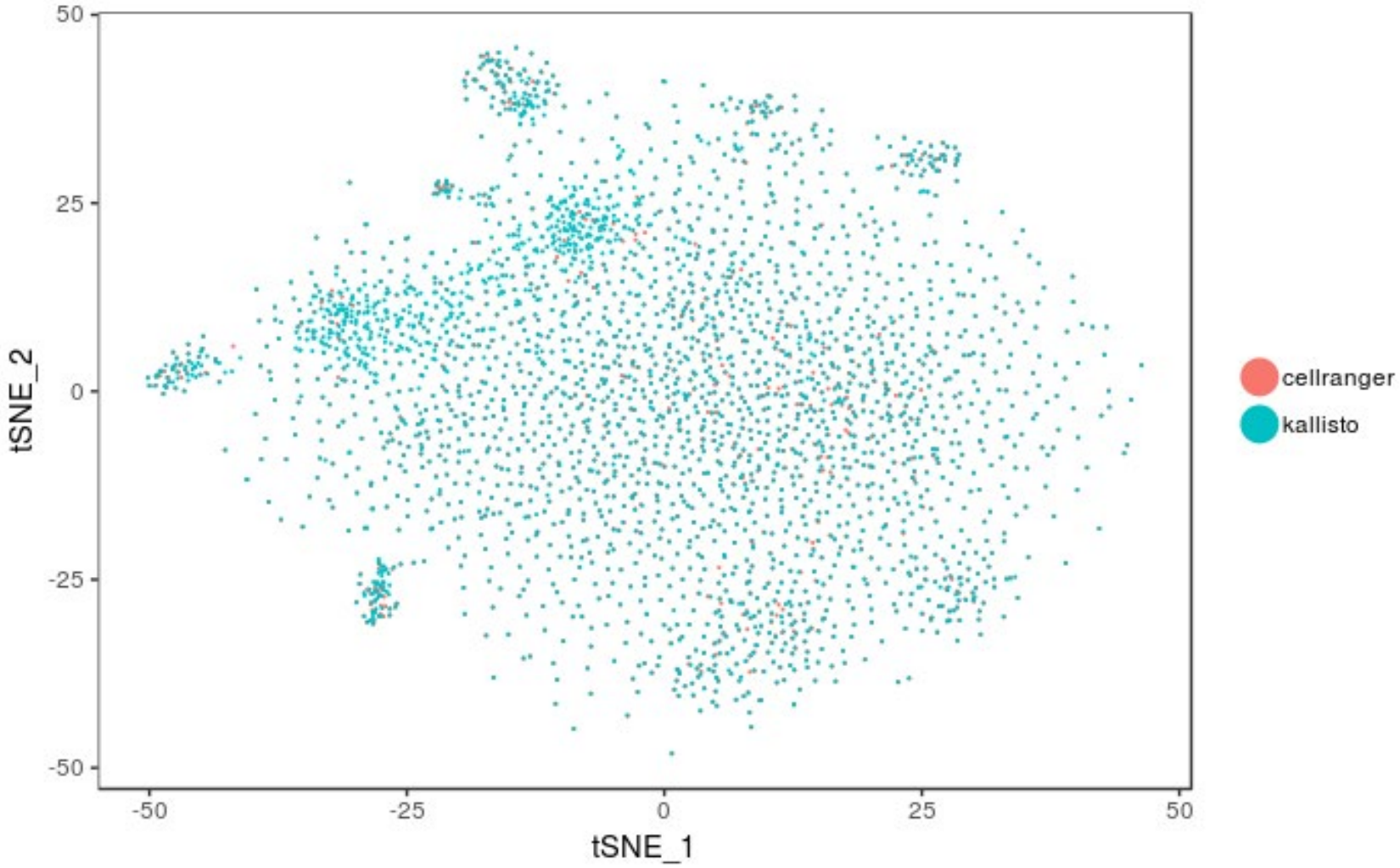
10X PBMC(1k_v3) data

Merge of cellranger & kallisto count matrices for clustering



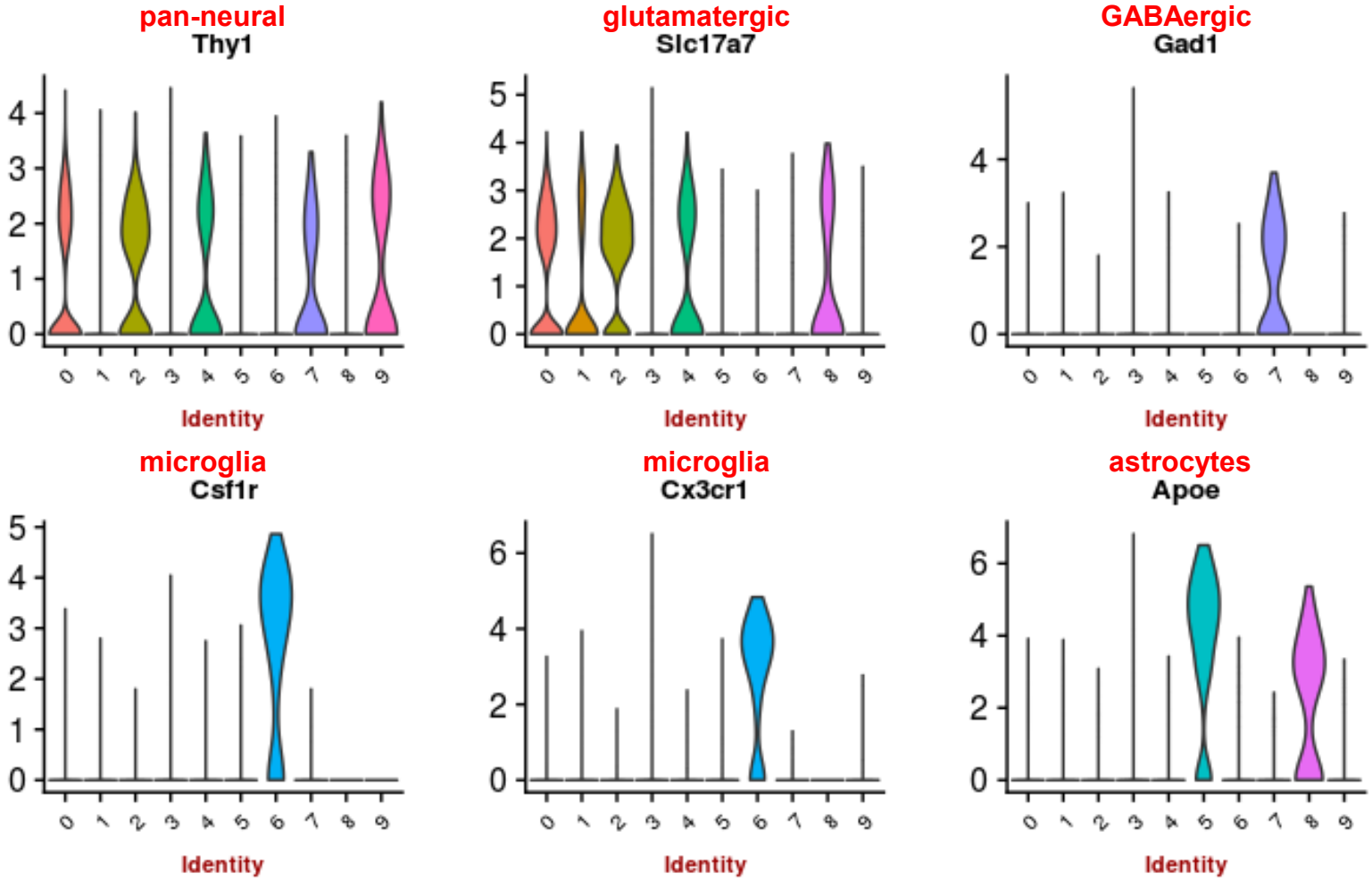
10X brain(2k_v2) data
CCA subspace alignment

Overlay of cellranger & kallisto cells



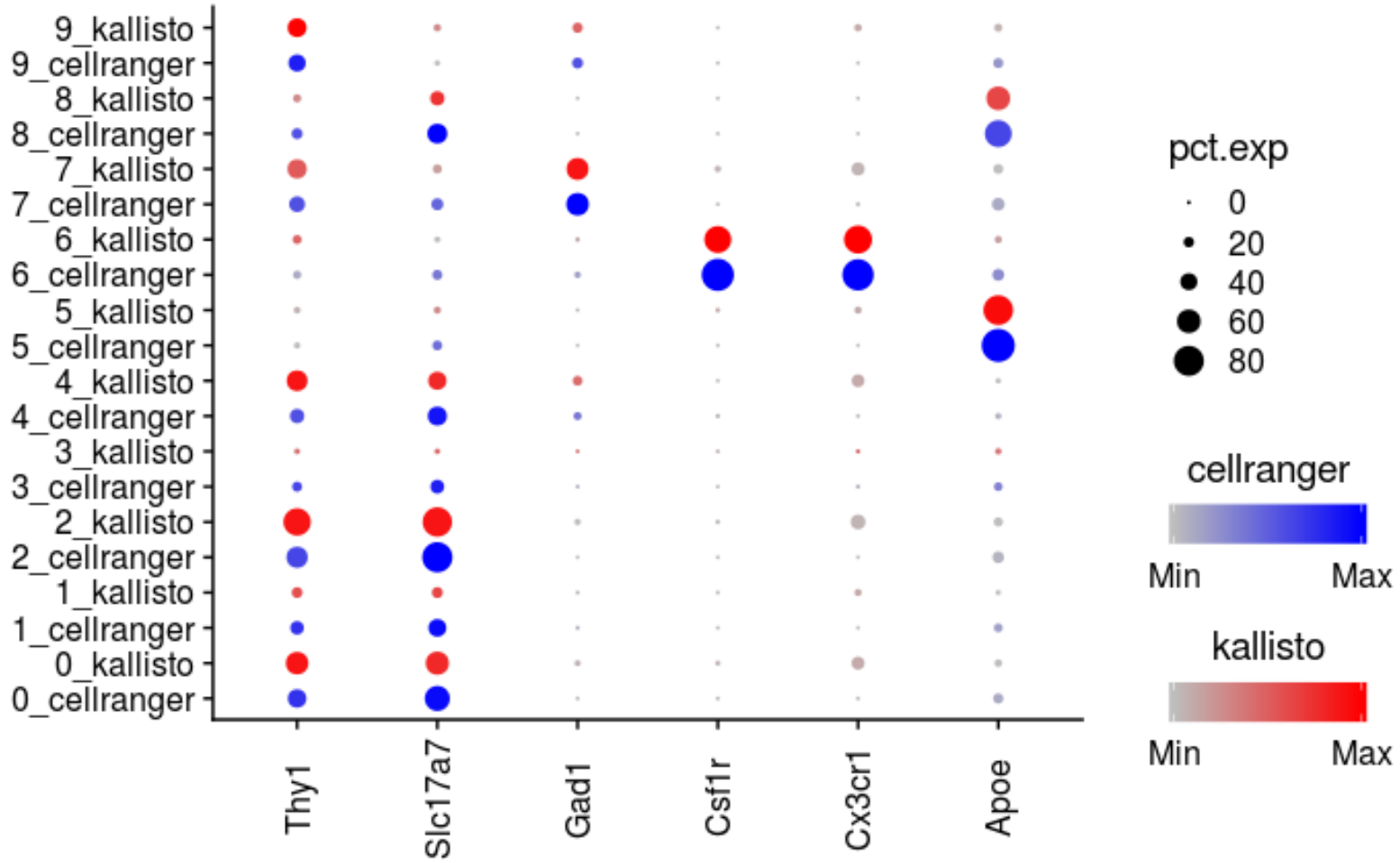
10X brain(2k_v2) data

Cell identity revealed by marker genes



10X brain(2k_v2) data

Percentage of cellranger & kallisto cells expressing a marker gene

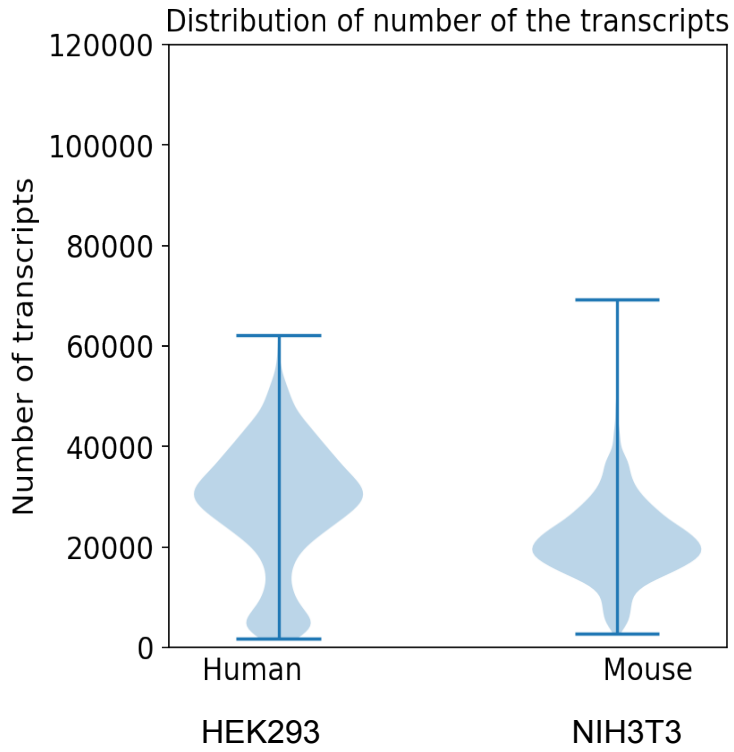


10X brain(2k_v2) data

Part III

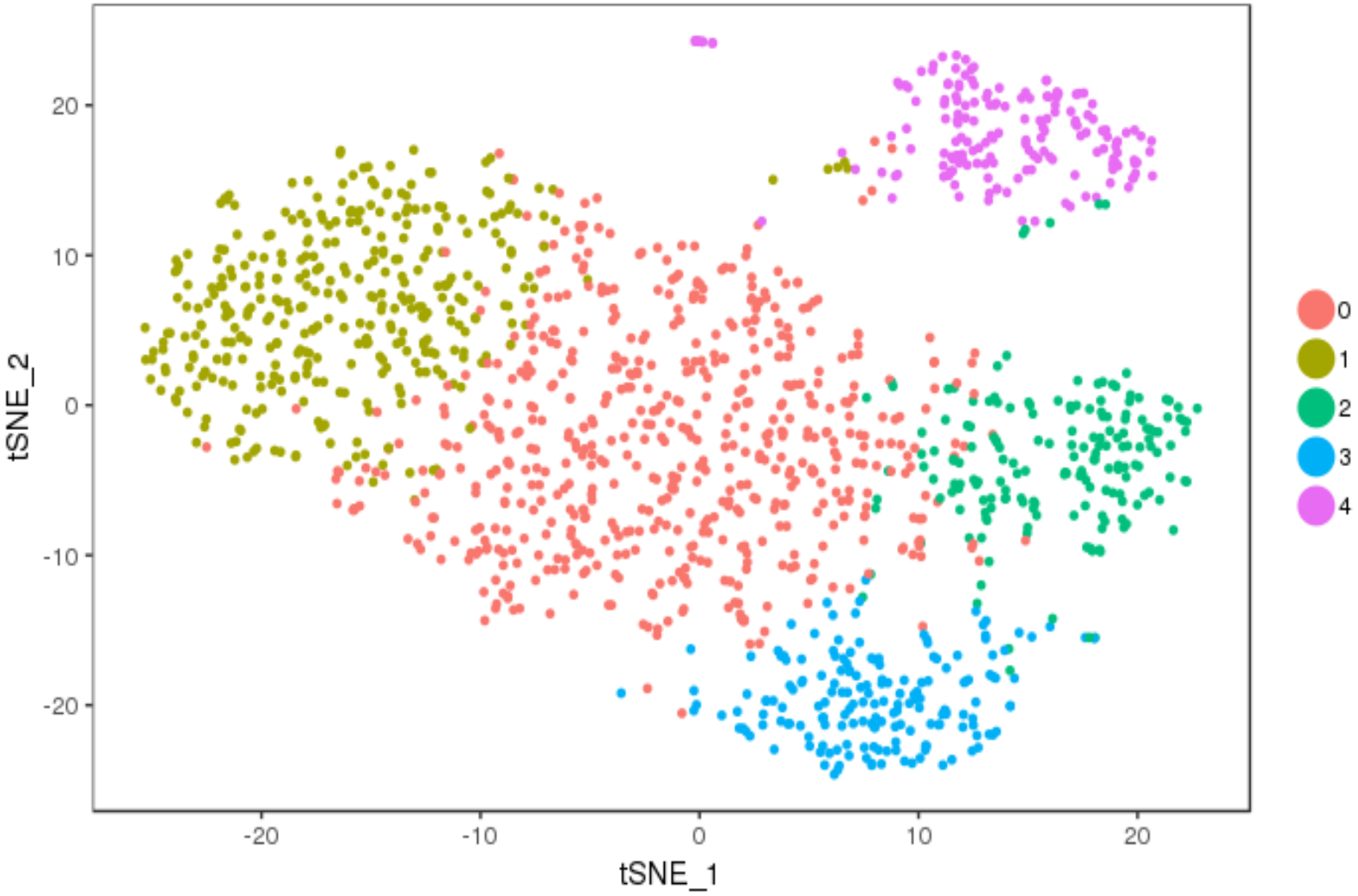
Detecting mixed-species

Detecting mixed-species - human & mouse single cells



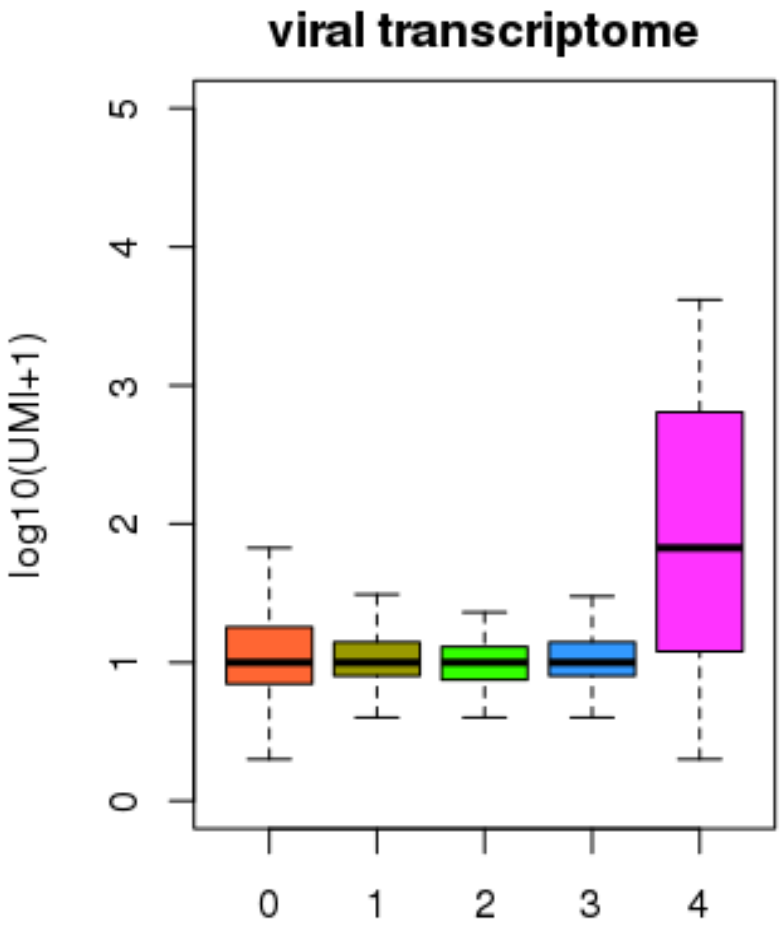
10X hgmm_6k v2 data

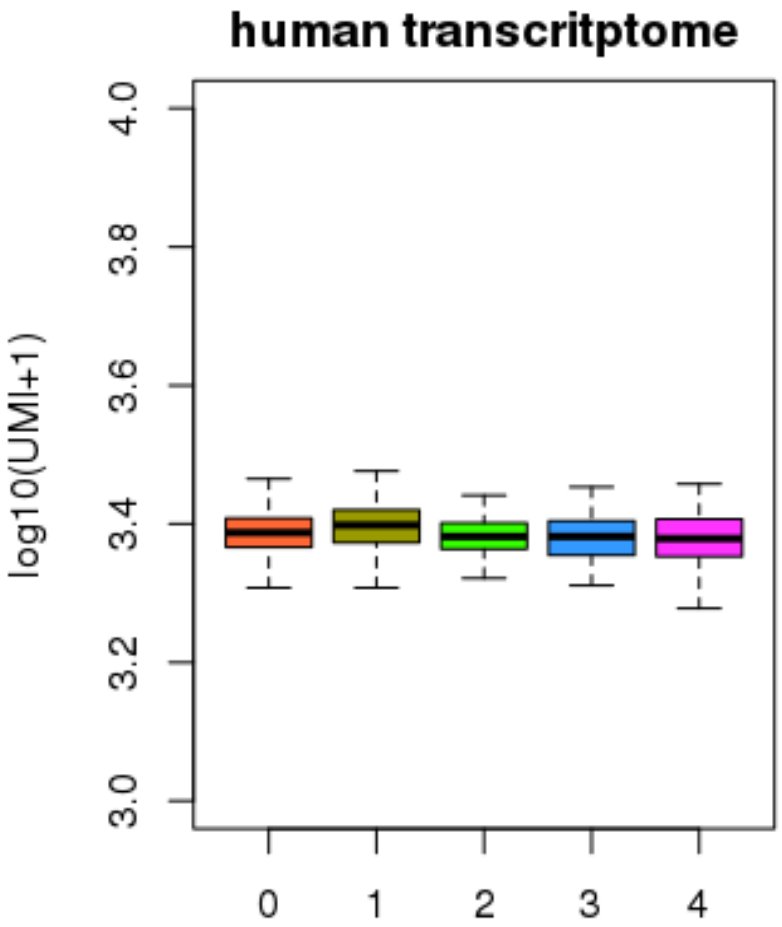
Detecting mixed-species - host pathogen interaction



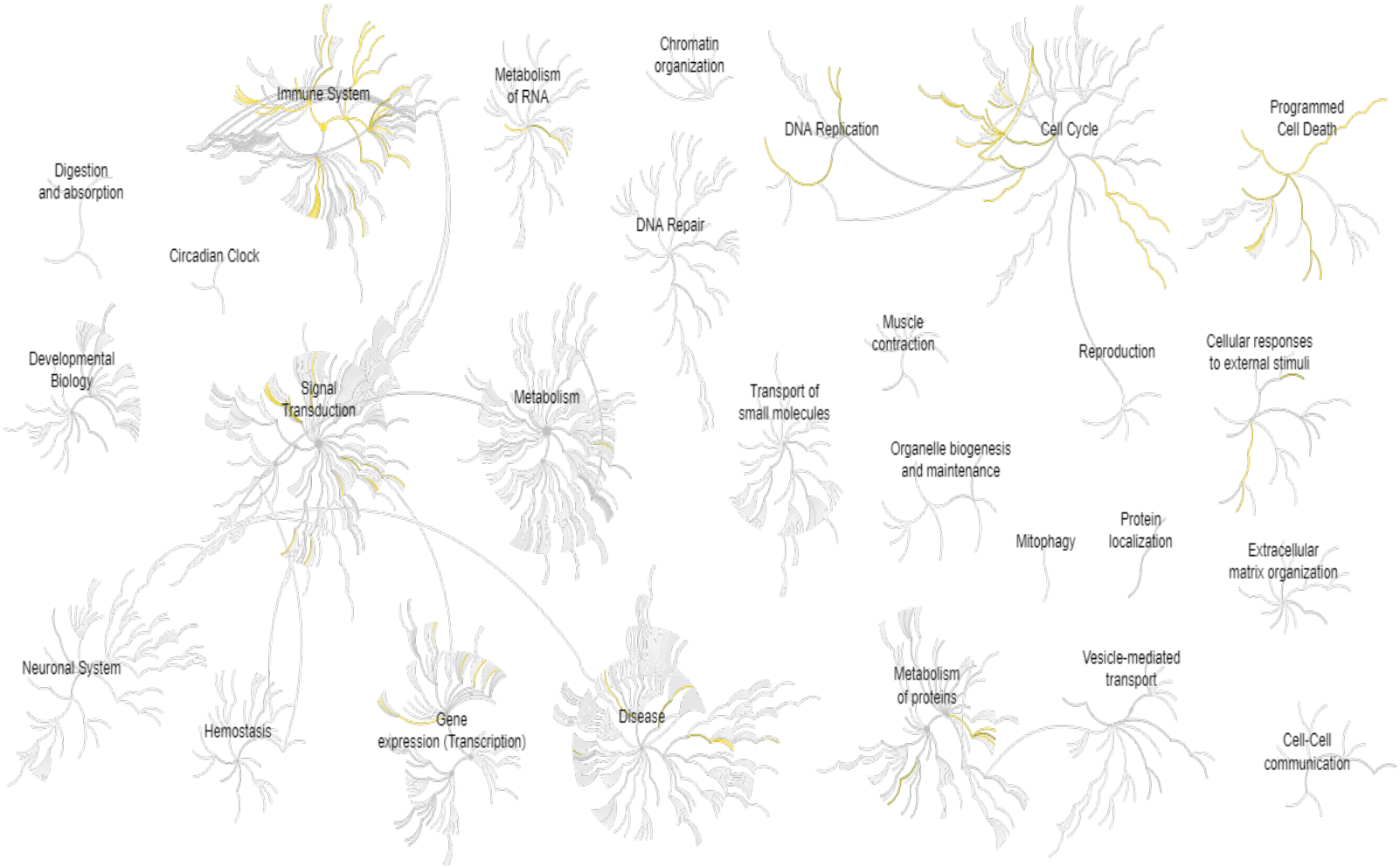
A549 (human) cells infected with A/WSN/1933(H1N1) influenza virus for 13hr (SRR8186084 10X_v2 data).

Transcriptome size for different cell clusters



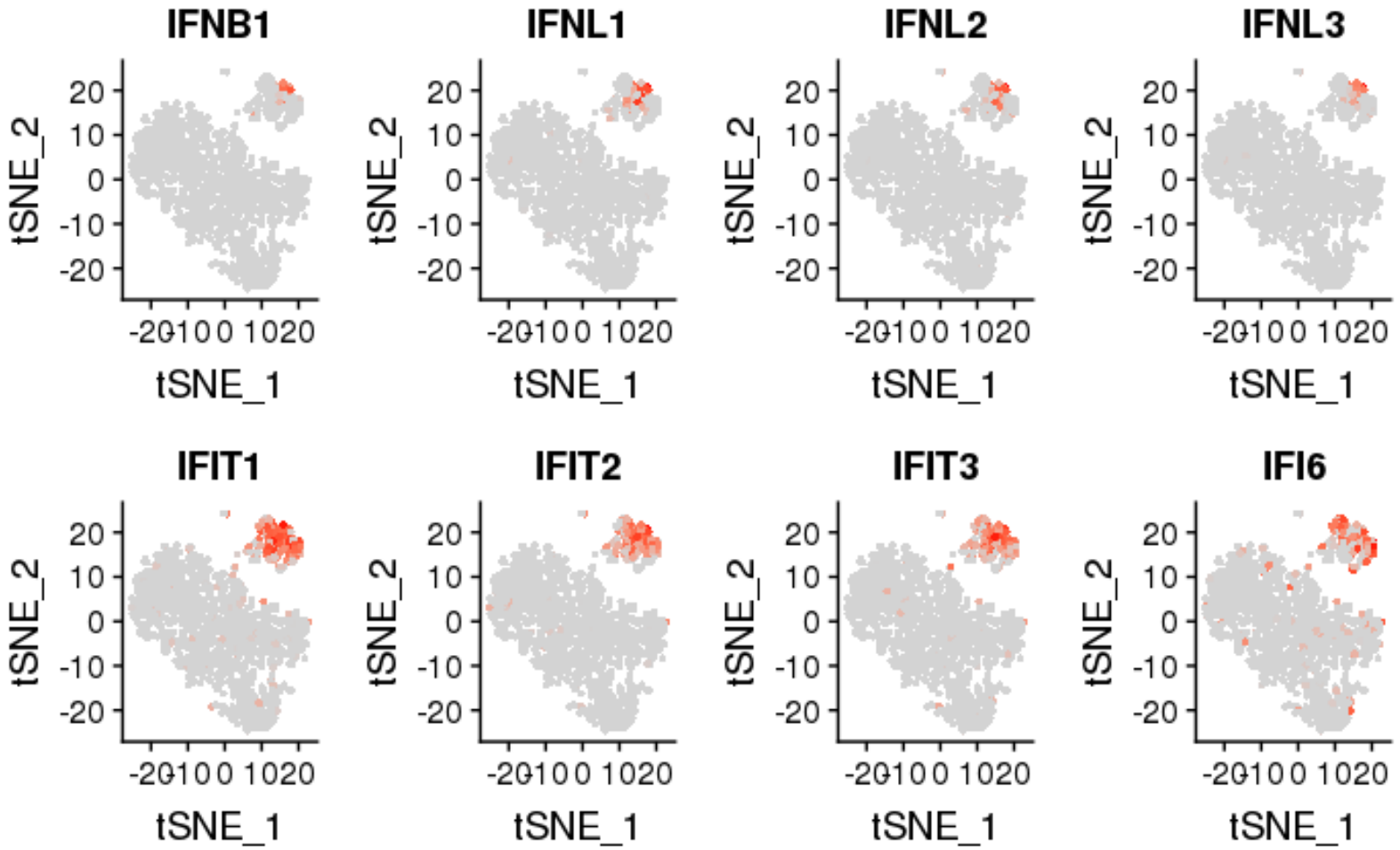


Biological pathways activated in response to viral infection

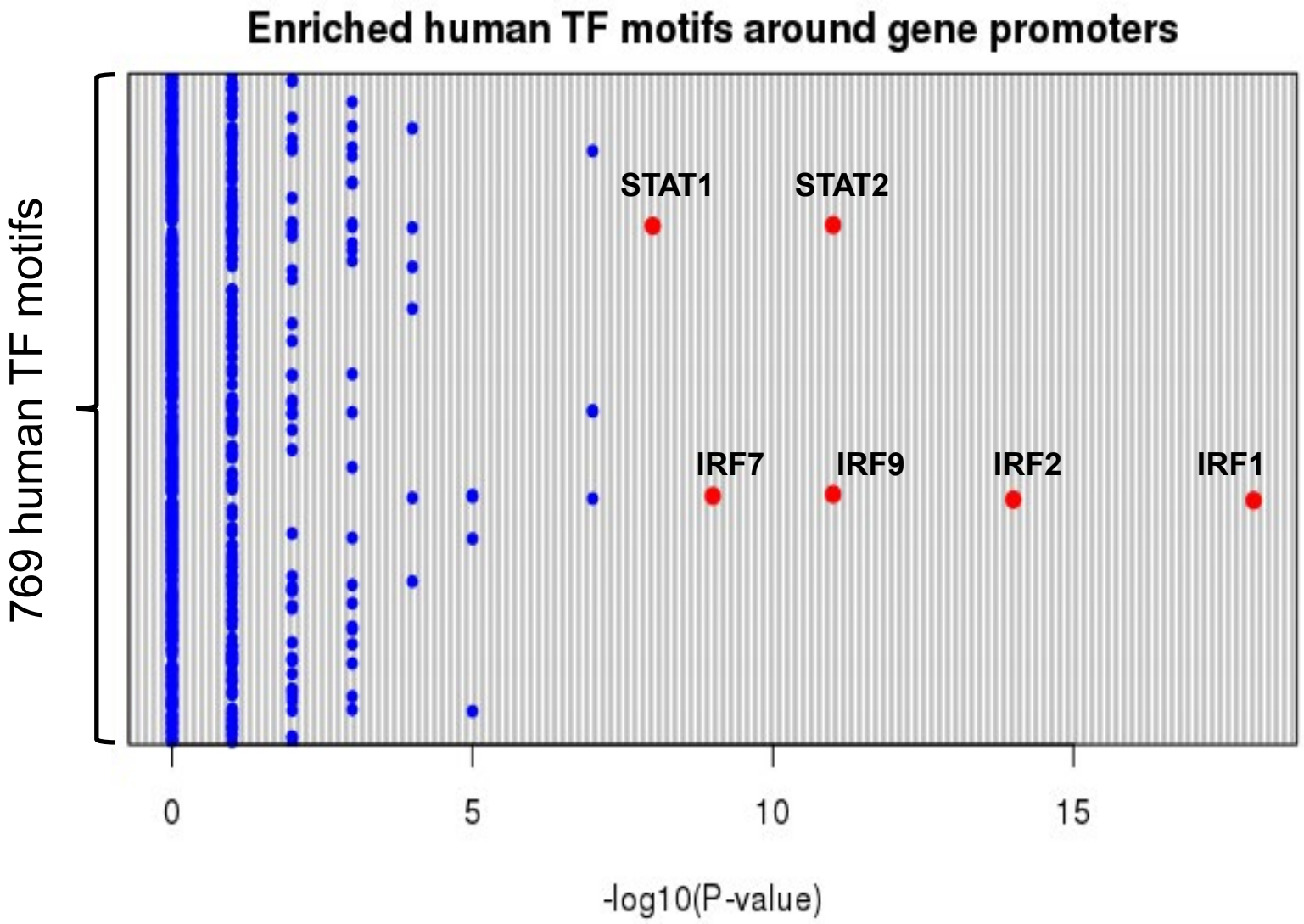


Reactome analysis of genes upregulated in cluster4 cells

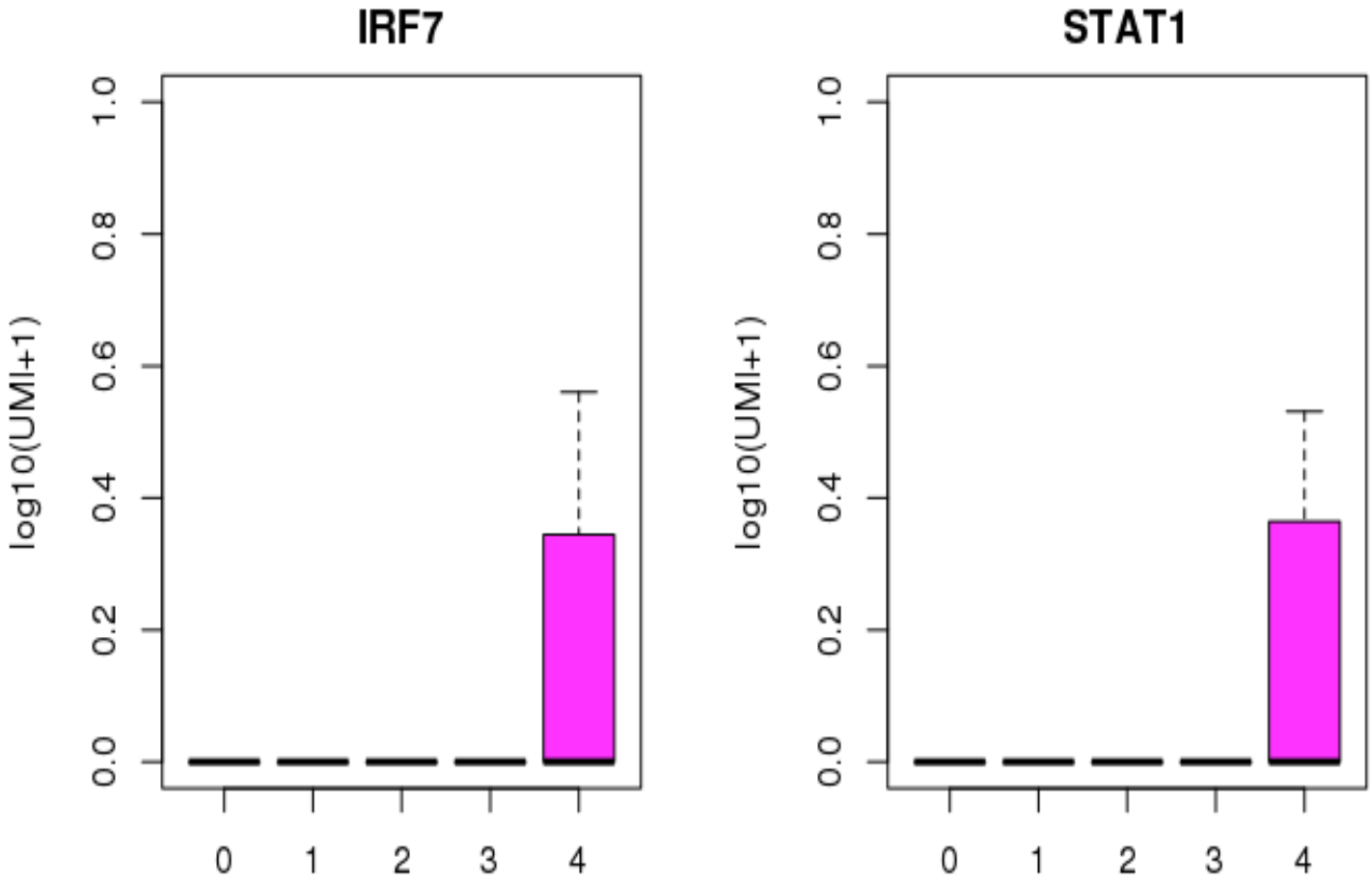
Viral infection induces interferon responses



Prediction of transcription regulators



IRF7 & STAT1 are candidate master regulators

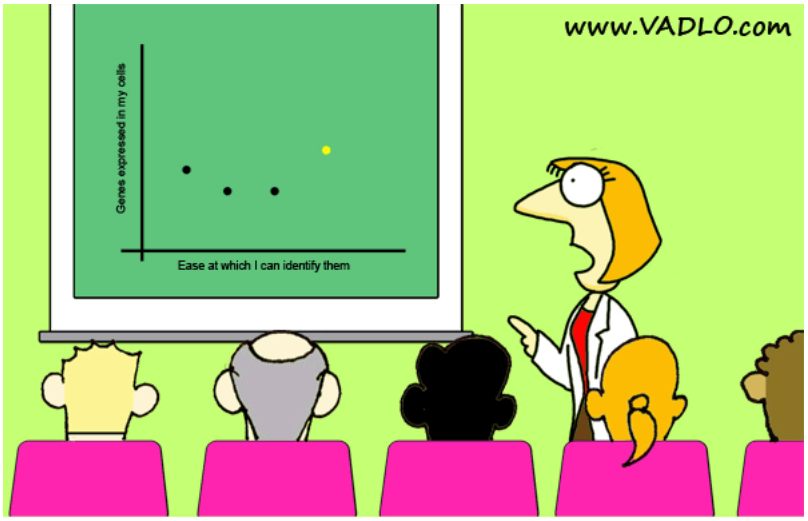


Summary

- Pachter lab has developed a cutting-edge single-cell RNA-Seq workflow.
- CBRC has benchmarked the workflow.
- Together with Pachter lab, CBRC will provide hands-on workshops to our students and postdocs.

Training materials https://caltech-bioinformatics.github.io/

CBRC (http://bioinformatics.caltech.edu)



"Same graph as last year,
but now I have an additional dot."

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☐ A description of all covariates tested
☐	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	*Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.*
Data analysis	*Provide a description of all commercial, open source and custom code used to analyse the data in this study, specifying the version used OR state that no software was used.*

Happy Valentine’s Day