Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

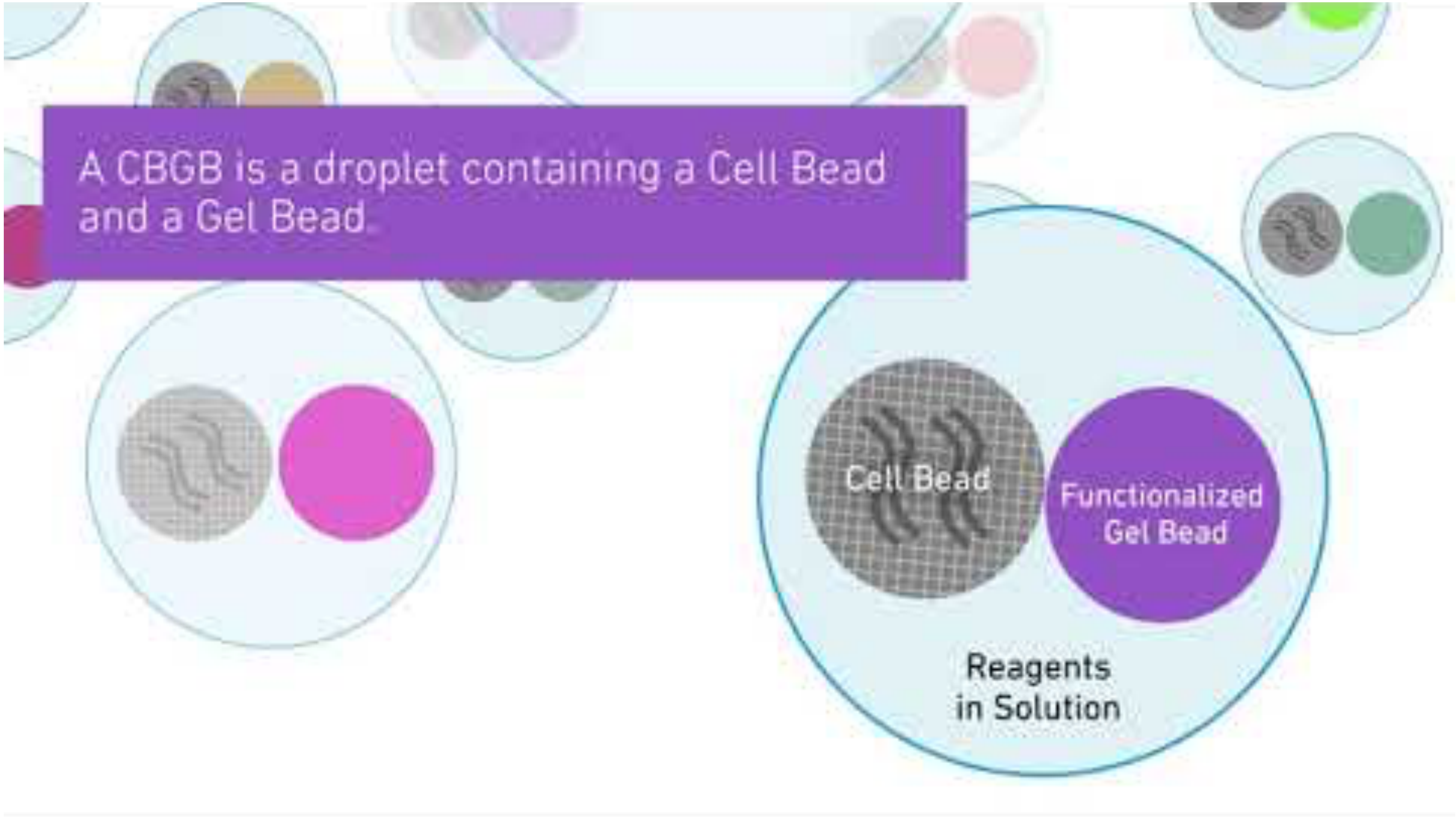
BTEP Presentation: scCNV + scATAC

Keyur Talsania
CCR-SF Bioinformatics Group
Advanced Biomedical and Computational Sciences
Biomedical Informatics and Data Science (BIDS) Directorate
Frederick National Laboratory for Cancer Research

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute
The Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

scCNV

10x Genomics scCNV



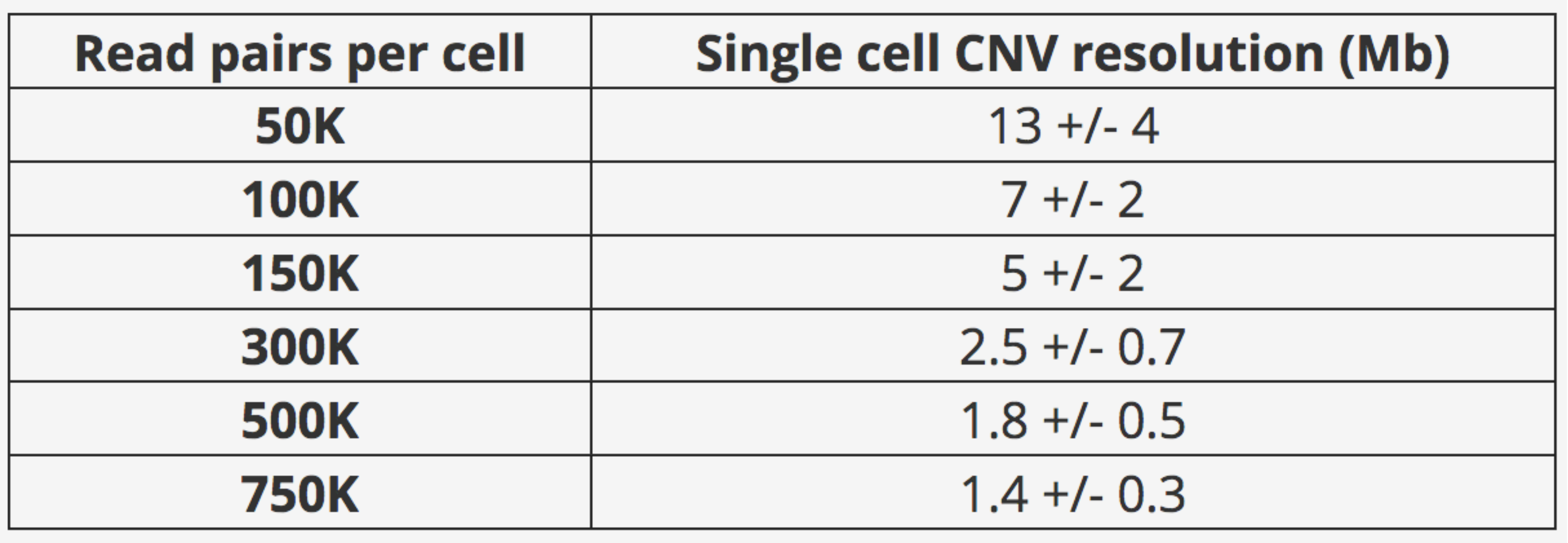
A CBGB is a droplet containing a Cell Bead and a Gel Bead.

scCNV – 10x Genomics

- **Sequencing Requirement:**

“ For an approximately diploid human sample we recommend a sequencing depth of 750,000 read-pairs per cell. At this depth, the metric median effective reads per 1Mbp is between 350-400, and we expect to be able to detect single cell copy number events in the size range 1-2 megabases (and upwards) with high sensitivity and positive predictive value. In groups of 10 or more cells we expect to be able to detect copy number events in the 100-200 kilobase (and upwards) with high sensitivity and positive predictive value.”

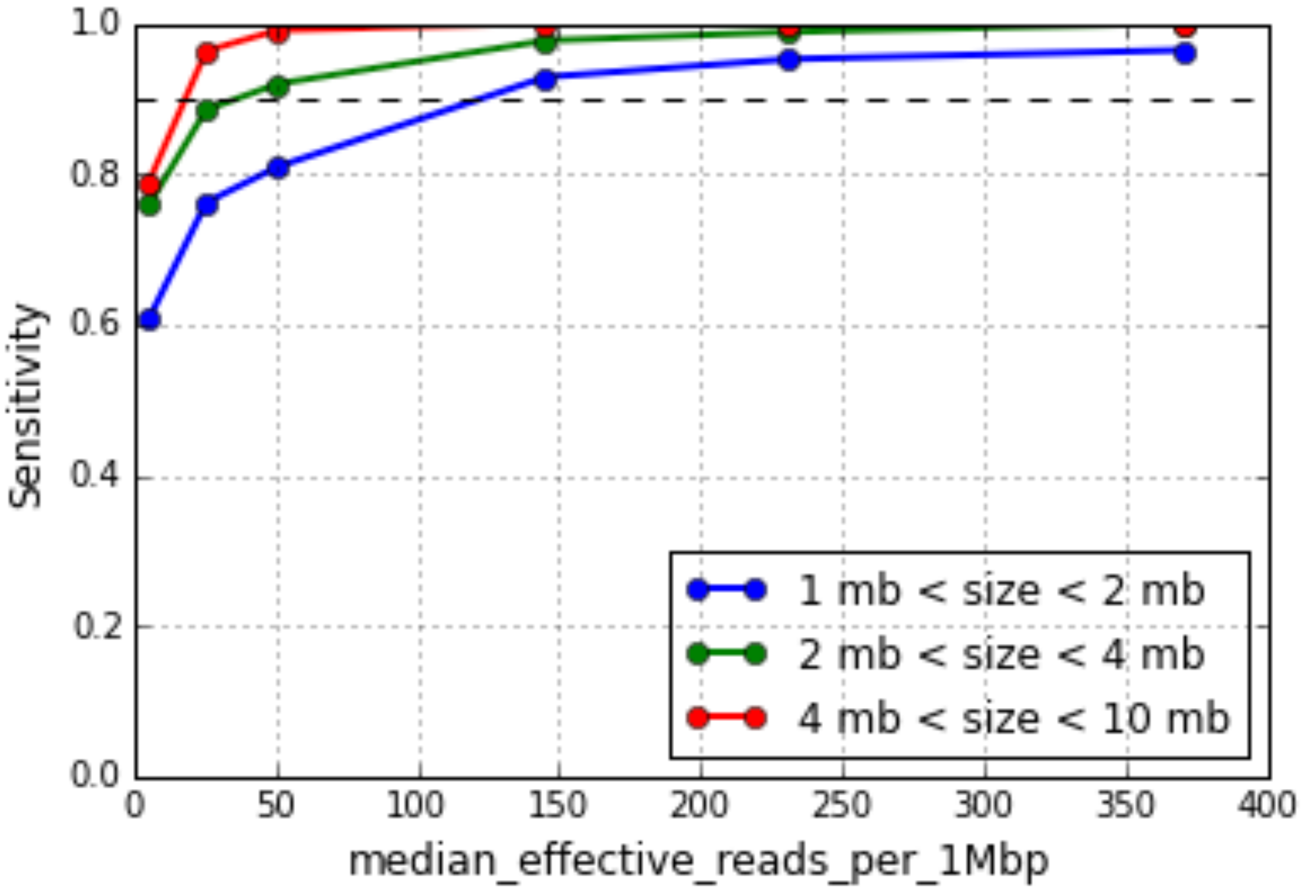
scCNV – 10x Genomics



Read pairs per cell	Single cell CNV resolution (Mb)
50K	13 +/- 4
100K	7 +/- 2
150K	5 +/- 2
300K	2.5 +/- 0.7
500K	1.8 +/- 0.5
750K	1.4 +/- 0.3

scCNV – 10x Genomics

Frederick National Laboratory for Cancer Research



scCNV – 10x Genomics

- Non-human organisms

We expect high quality CNV detection when median effective reads per 1Mbp is in the range 350-400, and the results in the graph above will likely translate across organisms. This level of coverage can be achieved by scaling the recommended coverage of 1.5-2.0 million reads per cell by the ratio of the organism genome size to the human genome size.

- Very high ploidy samples

For samples that contain cells with average ploidy significantly different from two, as is some times the case in cancer genomes, we recommend scaling the input coverage in proportion to the average ploidy / 2. In a tetraploid sample, for example, the extra coverage allows us to distinguish 4 -> 5 copy number changes and other n -> n+1 higher copy number transitions where the relative ploidy difference can be small.

scCNV – 10x Genomics

- Non-human organisms

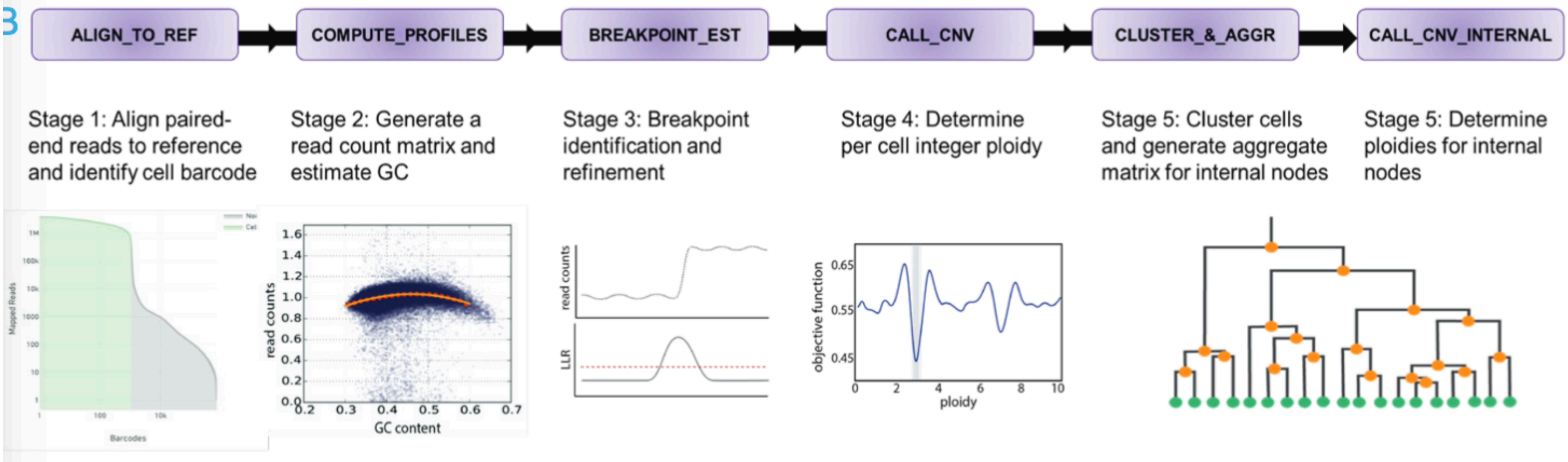
We expect high quality CNV detection when median effective reads per 1Mbp is in the range 350-400, and the results in the graph above will likely translate across organisms. This level of coverage can be achieved by scaling the recommended coverage of 1.5-2.0 million reads per cell by the ratio of the organism genome size to the human genome size.

- Very high ploidy samples

For samples that contain cells with average ploidy significantly different from two, as is some times the case in cancer genomes, we recommend scaling the input coverage in proportion to the average ploidy / 2. In a tetraploid sample, for example, the extra coverage allows us to distinguish 4 -> 5 copy number changes and other n -> n+1 higher copy number transitions where the relative ploidy difference can be small.

- Ploidy of 3 = 1.5m * (3/2) = 2.25m reads – 1.2m read pairs
- Ploidy of 3 = 1.5m * (4/2) = 3m reads – 1.5m read pairs

10x Genomis scCNV Pipeline



ALIGN_TO_REF → COMPUTE_PROFILES → BREAKPOINT_EST → CALL_CNV → CLUSTER_&_AGGR → CALL_CNV_INTERNAL

Stage 1: Align paired-end reads to reference and identify cell barcode

Stage 2: Generate a read count matrix and estimate GC

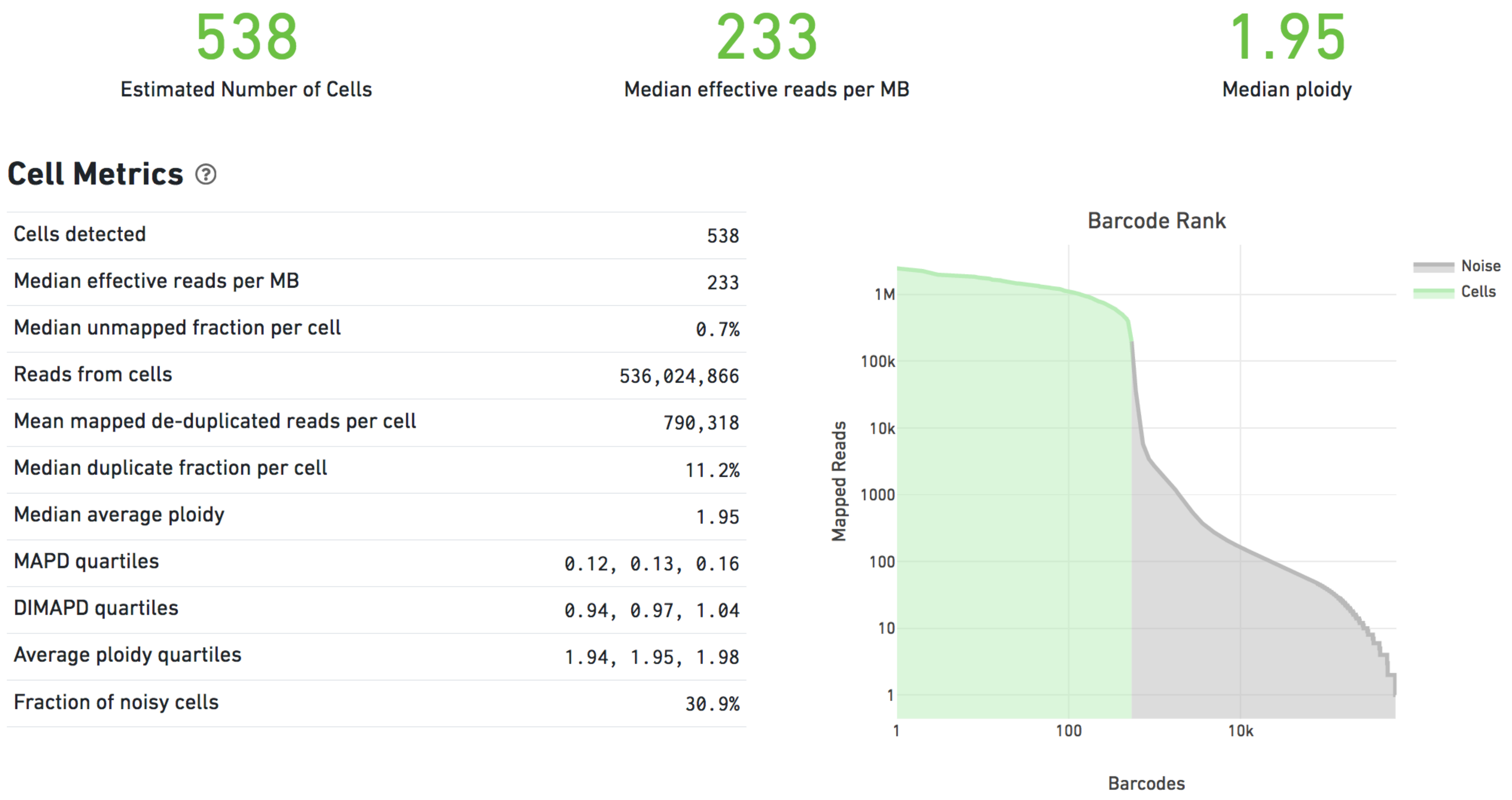
Stage 3: Breakpoint identification and refinement

Stage 4: Determine per cell integer ploidy

Stage 5: Cluster cells and generate aggregate matrix for internal nodes

Stage 5: Determine ploidies for internal nodes

ScCNV Result Summary



538

Estimated Number of Cells

233

Median effective reads per MB

1.95

Median ploidy

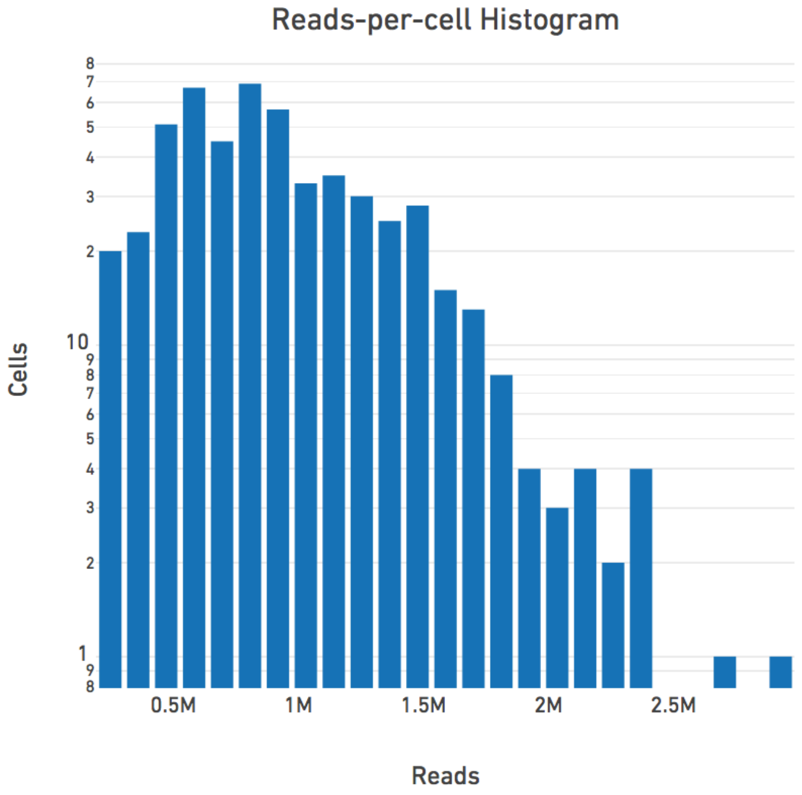
Cell Metrics ?

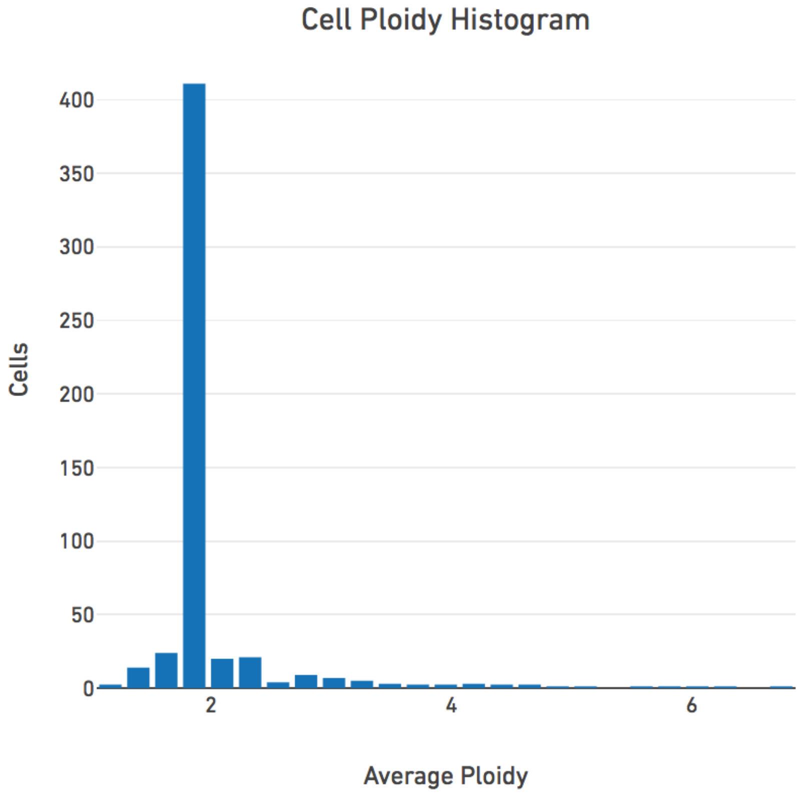
Cells detected	538
Median effective reads per MB	233
Median unmapped fraction per cell	0.7%
Reads from cells	536,024,866
Mean mapped de-duplicated reads per cell	790,318
Median duplicate fraction per cell	11.2%
Median average ploidy	1.95
MAPD quartiles	0.12, 0.13, 0.16
DIMAPD quartiles	0.94, 0.97, 1.04
Average ploidy quartiles	1.94, 1.95, 1.98
Fraction of noisy cells	30.9%

Barcode Rank

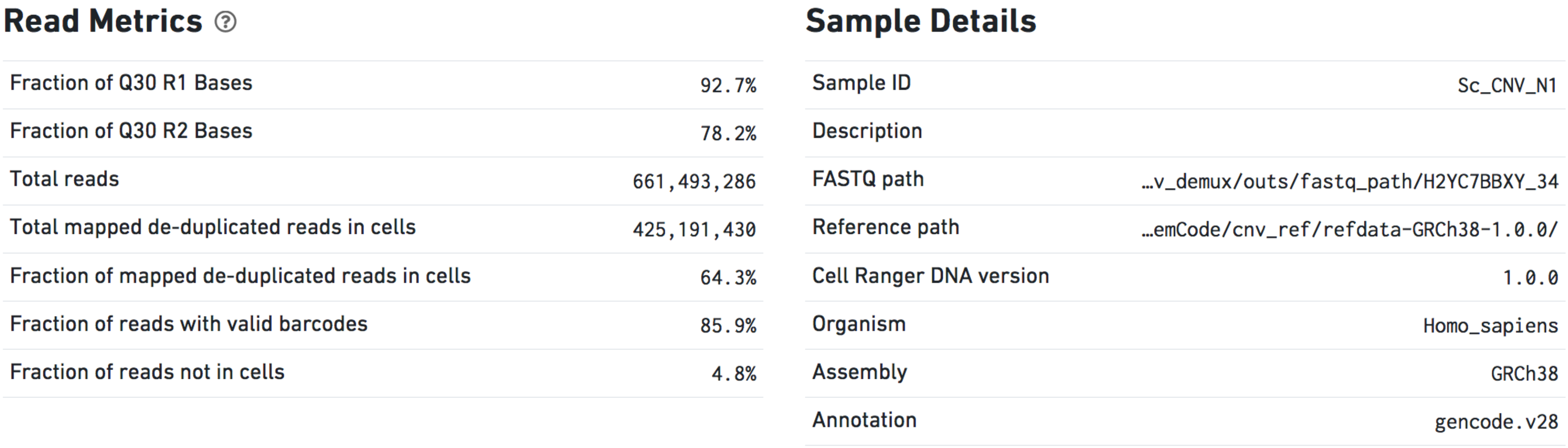
ScCNV Result Summary

Cell Plots





ScCNV Result Summary



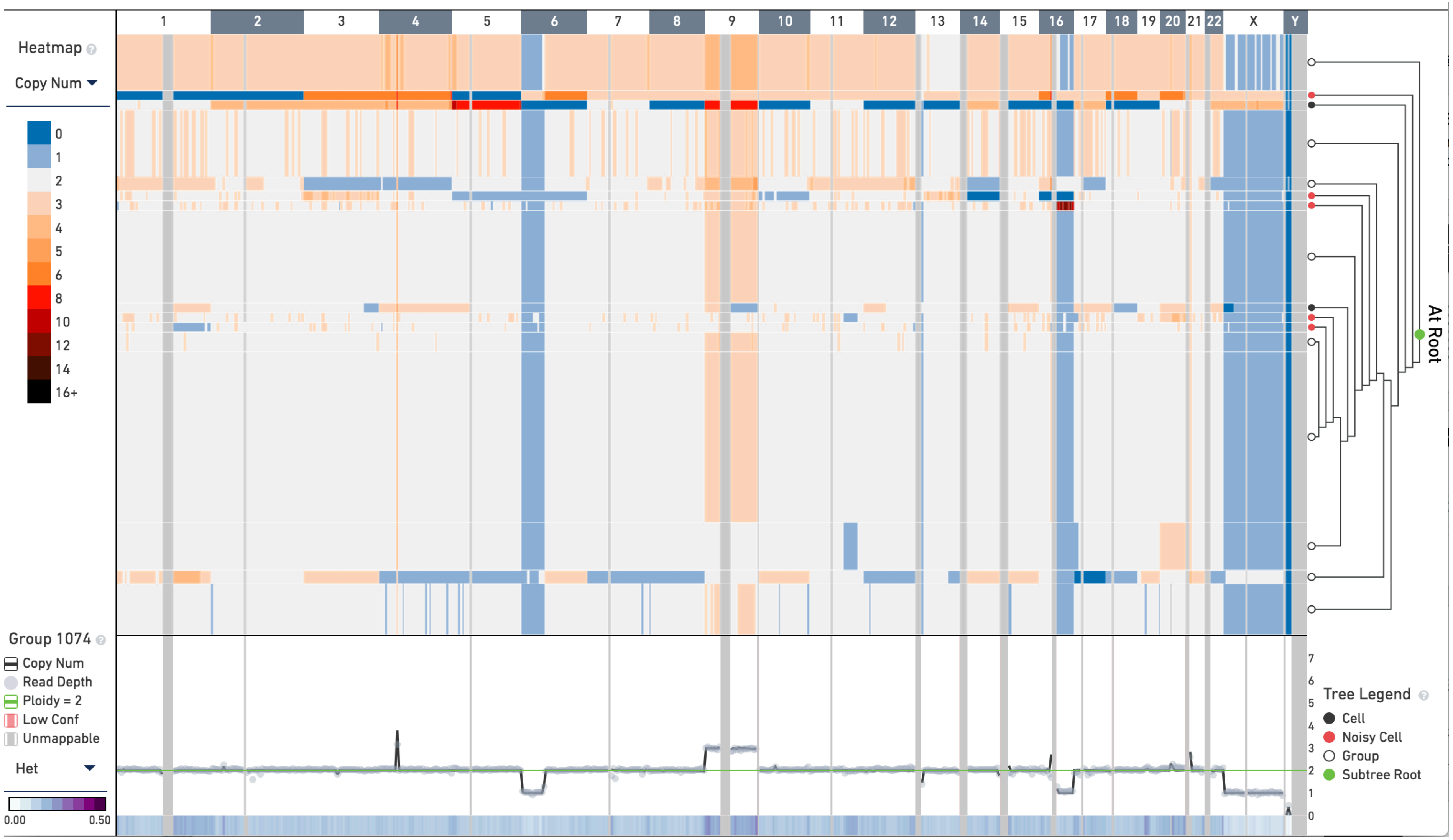
Read Metrics ?

Fraction of Q30 R1 Bases	92.7%
Fraction of Q30 R2 Bases	78.2%
Total reads	661,493,286
Total mapped de-duplicated reads in cells	425,191,430
Fraction of mapped de-duplicated reads in cells	64.3%
Fraction of reads with valid barcodes	85.9%
Fraction of reads not in cells	4.8%

Sample Details

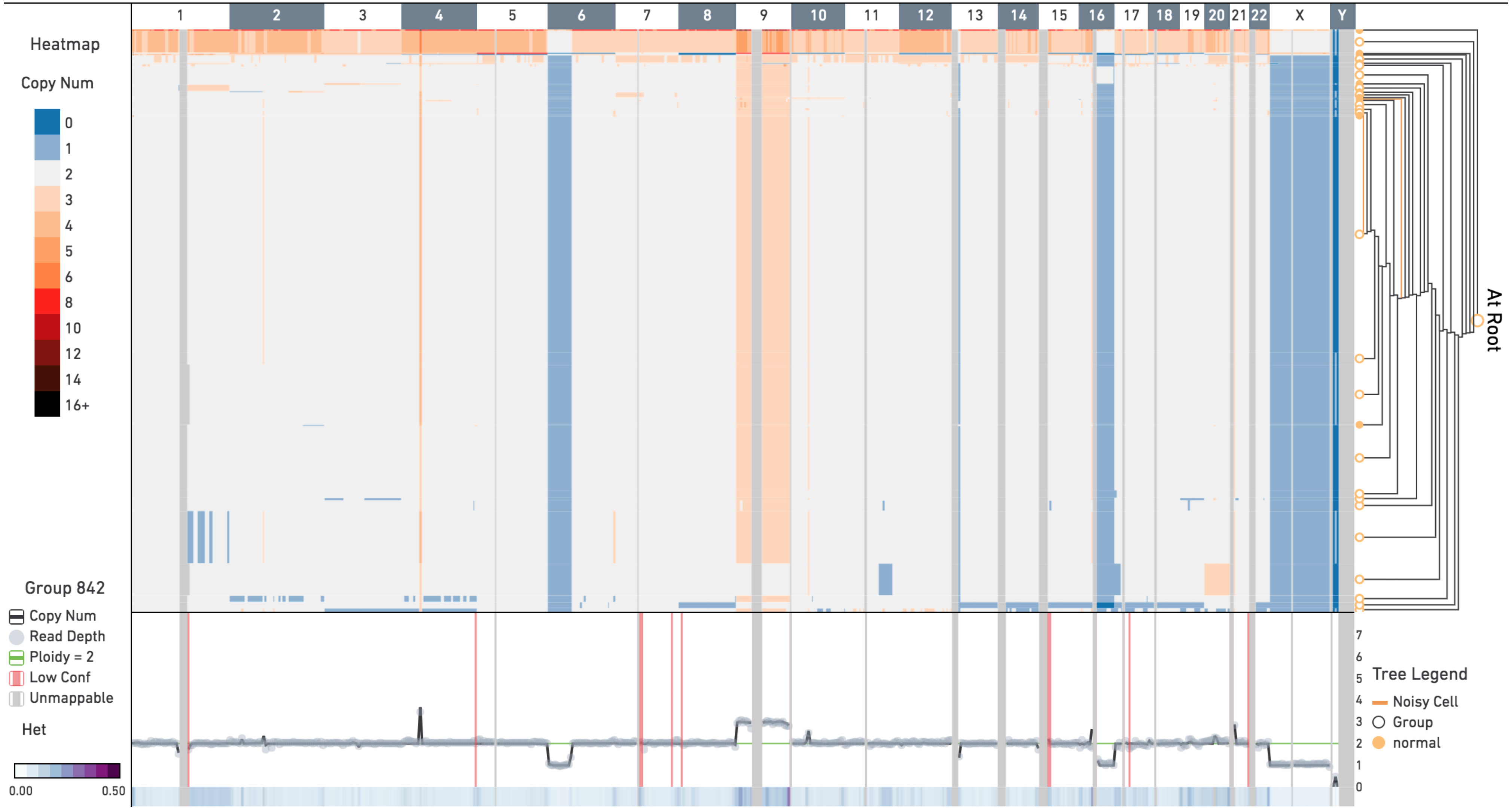
Sample ID	Sc_CNV_N1
Description	
FASTQ path	…v_demux/outs/fastq_path/H2YC7BBXY_34
Reference path	…emCode/cnv_ref/refdata-GRCh38-1.0.0/
Cell Ranger DNA version	1.0.0
Organism	Homo_sapiens
Assembly	GRCh38
Annotation	gencode.v28

ScCNV: 10x scDNA Loupe



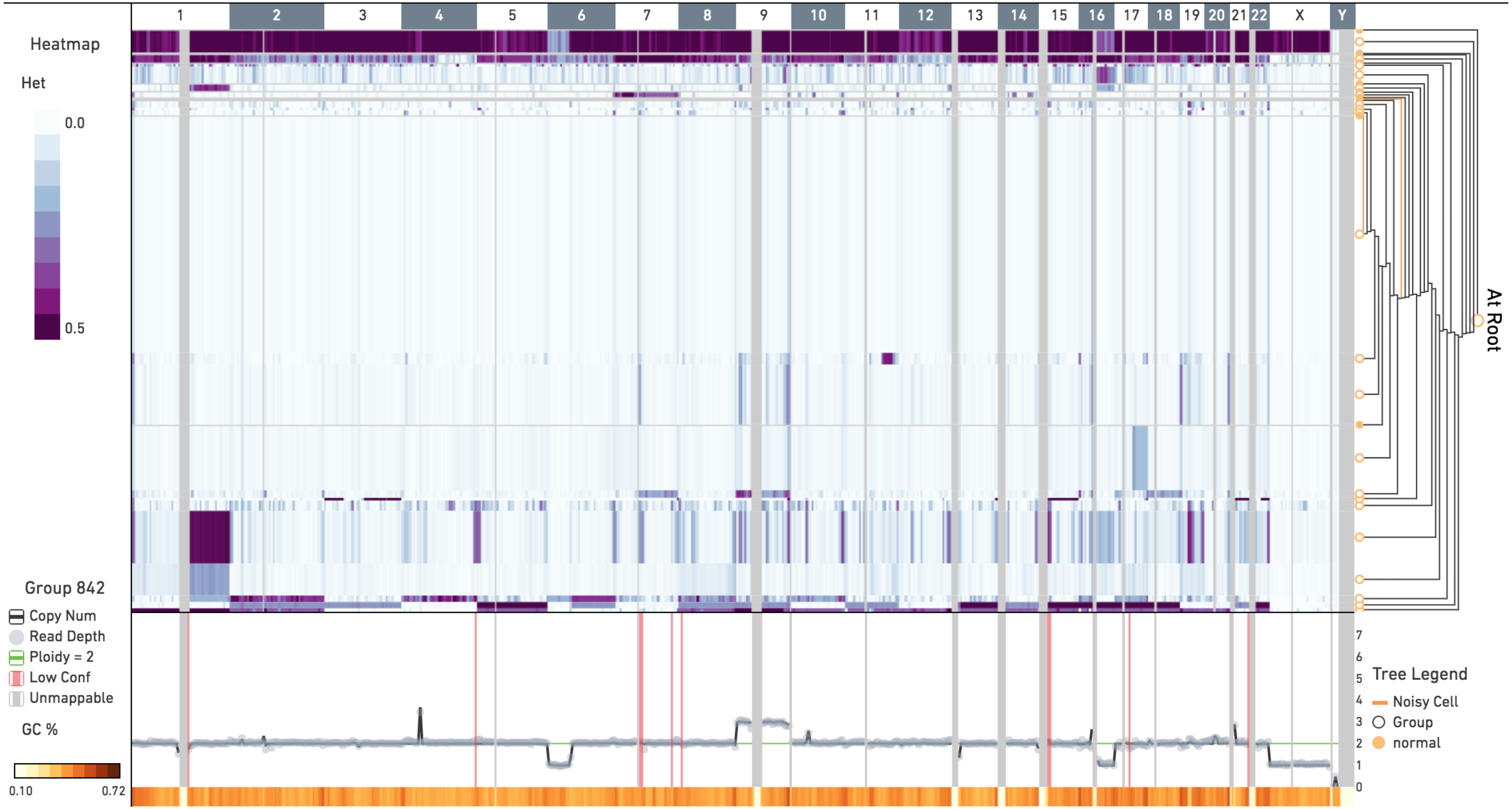
ScCNV: 10x scDNA Loupe

Frederick National Laboratory for Cancer Research



ScCNV: 10x scDNA Loupe

Frederick National Laboratory for Cancer Research



ScCNV: 10x scDNA Loupe Demo

Loupe scDNA Browser

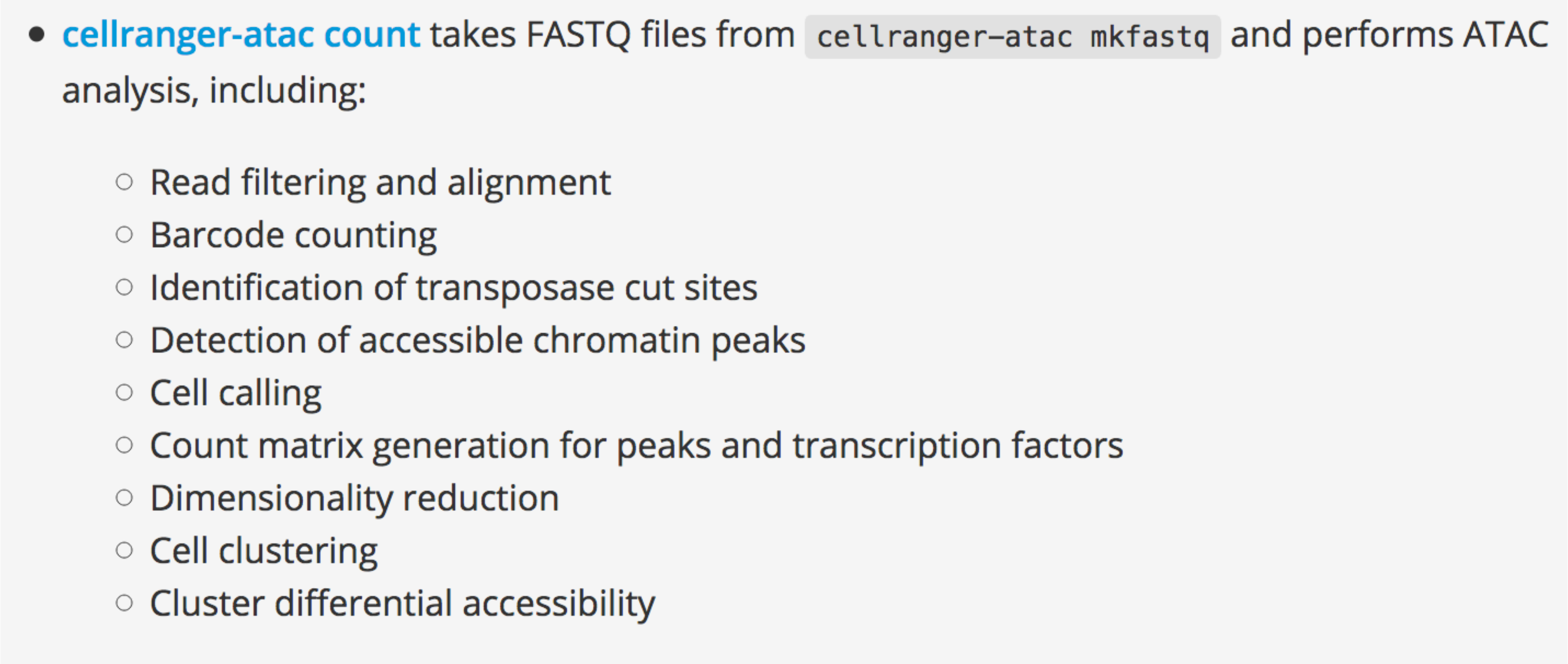
scATAC

scATAC – 10x Genomics

Frederick National Laboratory for Cancer Research

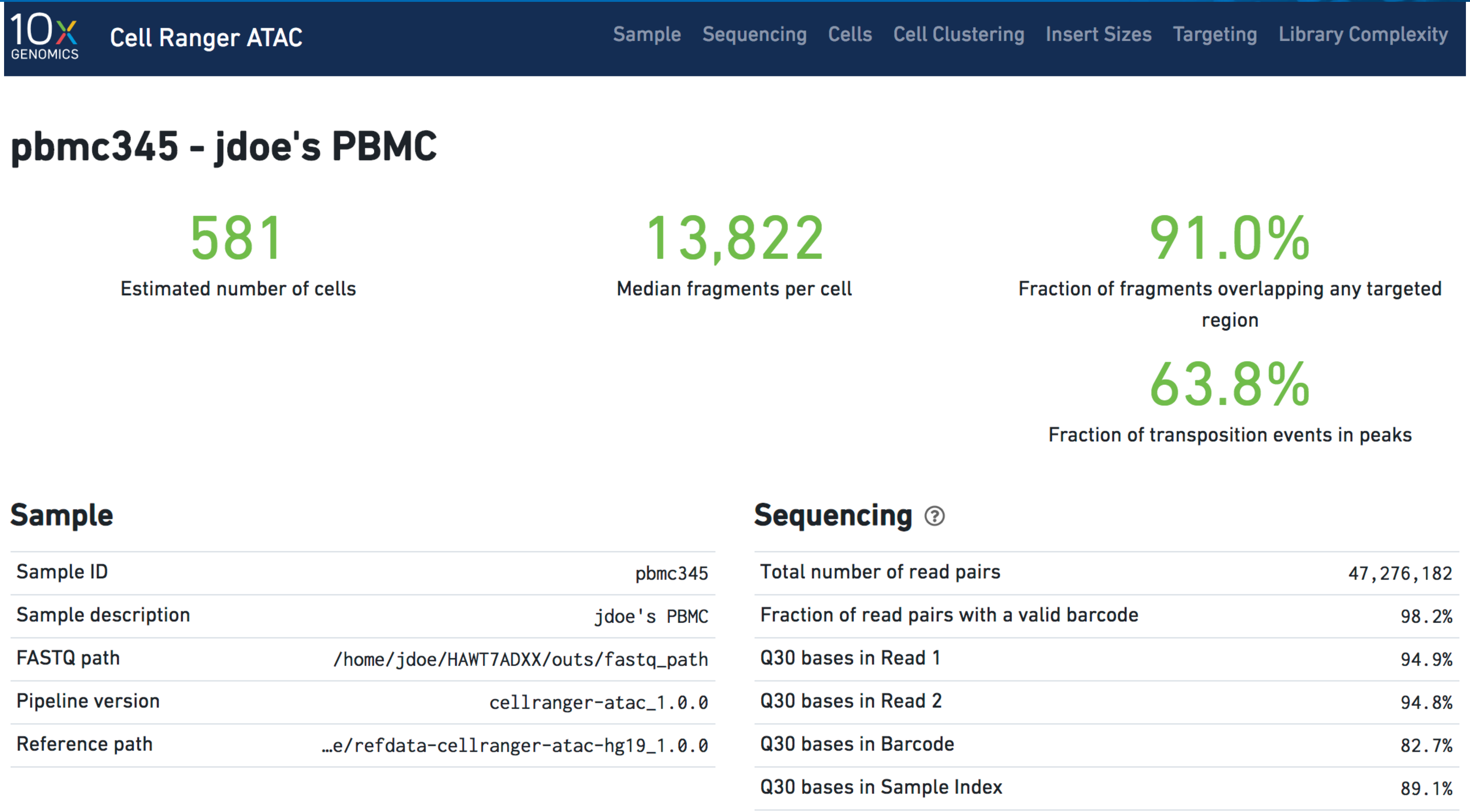
- How it works
- Results from 10x Genomics
- Other tools
- Data integration

Results from 10x Genomics



- cellranger-atac count takes FASTQ files from `cellranger-atac mkfastq` and performs ATAC analysis, including:
  - Read filtering and alignment
  - Barcode counting
  - Identification of transposase cut sites
  - Detection of accessible chromatin peaks
  - Cell calling
  - Count matrix generation for peaks and transcription factors
  - Dimensionality reduction
  - Cell clustering
  - Cluster differential accessibility

Summary of results



pbmc345 - jdoe's PBMC

581
Estimated number of cells

13,822
Median fragments per cell

91.0%
Fraction of fragments overlapping any targeted region

63.8%
Fraction of transposition events in peaks

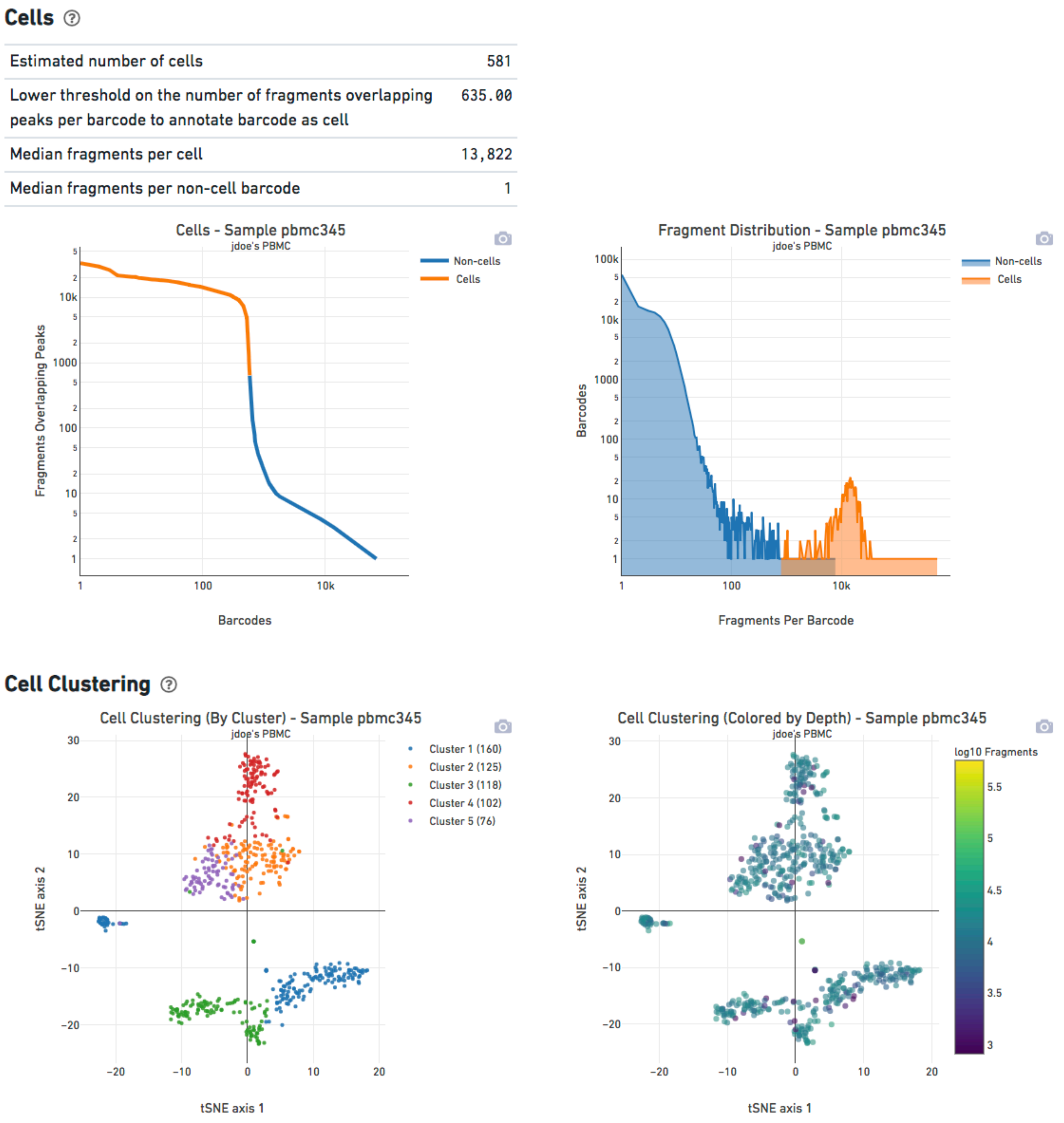
Sample

Sample ID	pbmc345
Sample description	jdoe's PBMC
FASTQ path	/home/jdoe/HAWT7ADXX/outs/fastq_path
Pipeline version	cellranger-atac_1.0.0
Reference path	…e/refdata-cellranger-atac-hg19_1.0.0

Sequencing ?

Total number of read pairs	47,276,182
Fraction of read pairs with a valid barcode	98.2%
Q30 bases in Read 1	94.9%
Q30 bases in Read 2	94.8%
Q30 bases in Barcode	82.7%
Q30 bases in Sample Index	89.1%

Summary of results

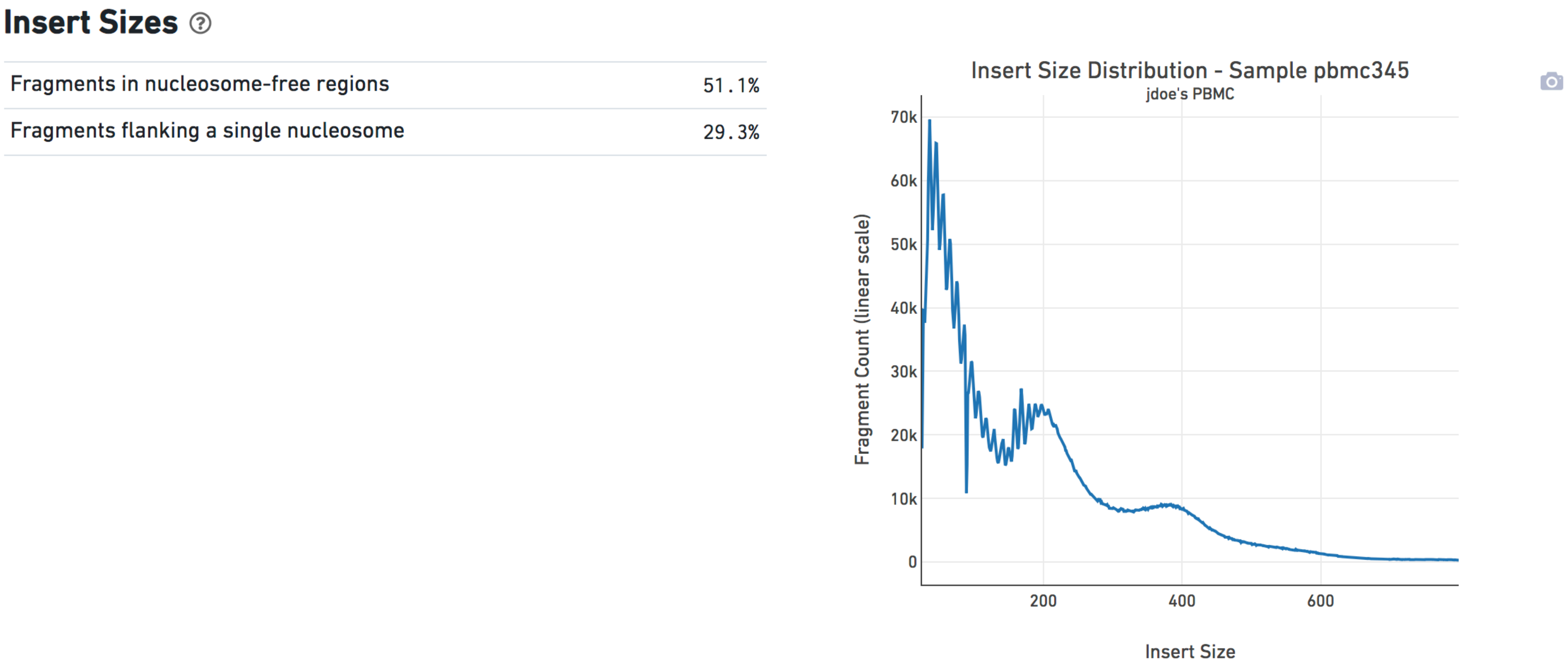


Cells ?

Estimated number of cells	581
Lower threshold on the number of fragments overlapping peaks per barcode to annotate barcode as cell	635.00
Median fragments per cell	13,822
Median fragments per non-cell barcode	1

Cell Clustering ?

Summary of results

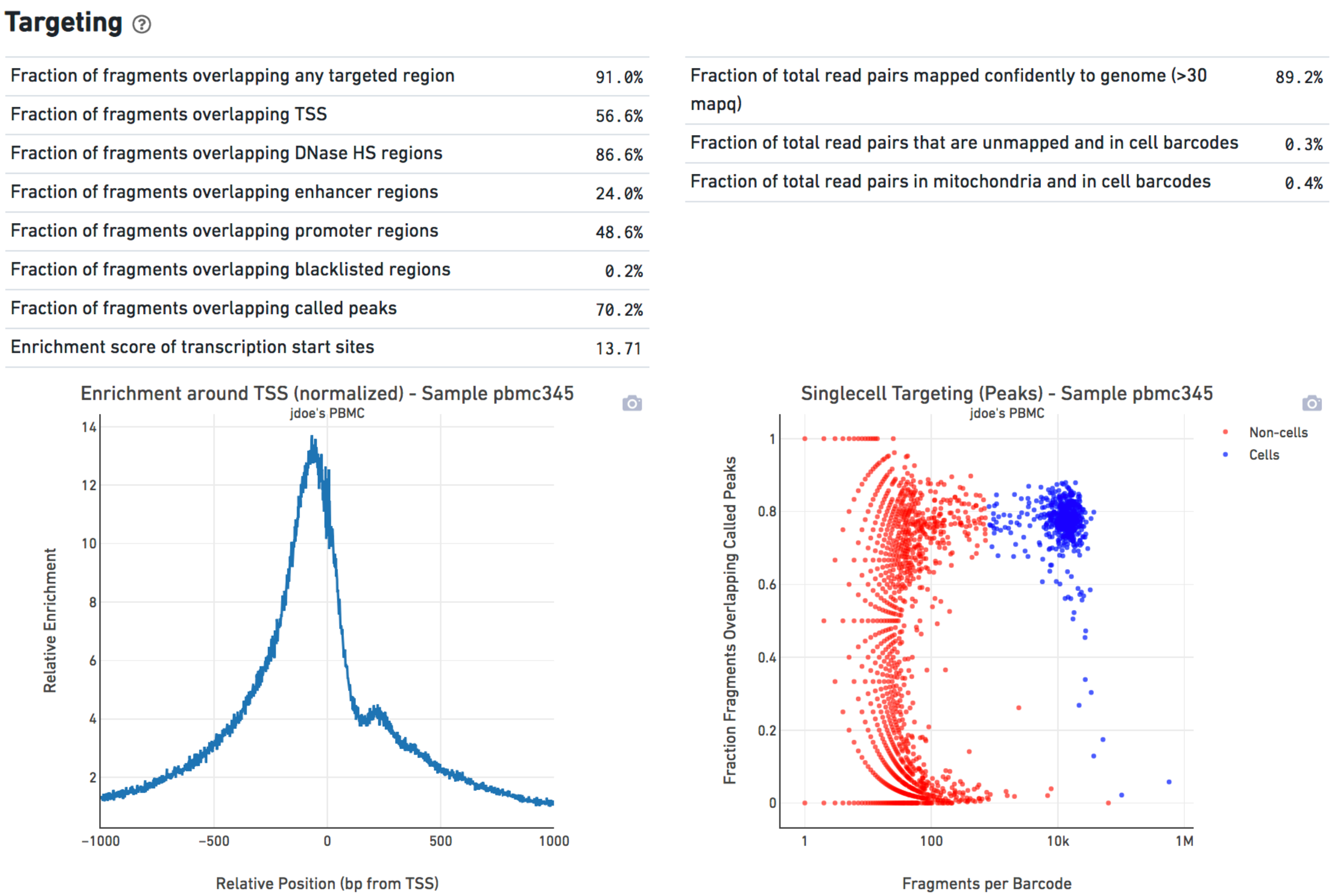


Insert Sizes ?

Fragments in nucleosome-free regions	51.1%
Fragments flanking a single nucleosome	29.3%

Insert Size Distribution - Sample pbmc345
jdoe's PBMC
Fragment Count (linear scale)
70k
60k
50k
40k
30k
20k
10k
0
200
400
600
Insert Size

Summary of results



Targeting ⓘ

Fraction of fragments overlapping any targeted region	91.0%
Fraction of fragments overlapping TSS	56.6%
Fraction of fragments overlapping DNase HS regions	86.6%
Fraction of fragments overlapping enhancer regions	24.0%
Fraction of fragments overlapping promoter regions	48.6%
Fraction of fragments overlapping blacklisted regions	0.2%
Fraction of fragments overlapping called peaks	70.2%
Enrichment score of transcription start sites	13.71

Fraction of total read pairs mapped confidently to genome (>30 mapq)	89.2%
Fraction of total read pairs that are unmapped and in cell barcodes	0.3%
Fraction of total read pairs in mitochondria and in cell barcodes	0.4%

Enrichment around TSS (normalized) - Sample pbmc345

jdoe's PBMC

Singlecell Targeting (Peaks) - Sample pbmc345

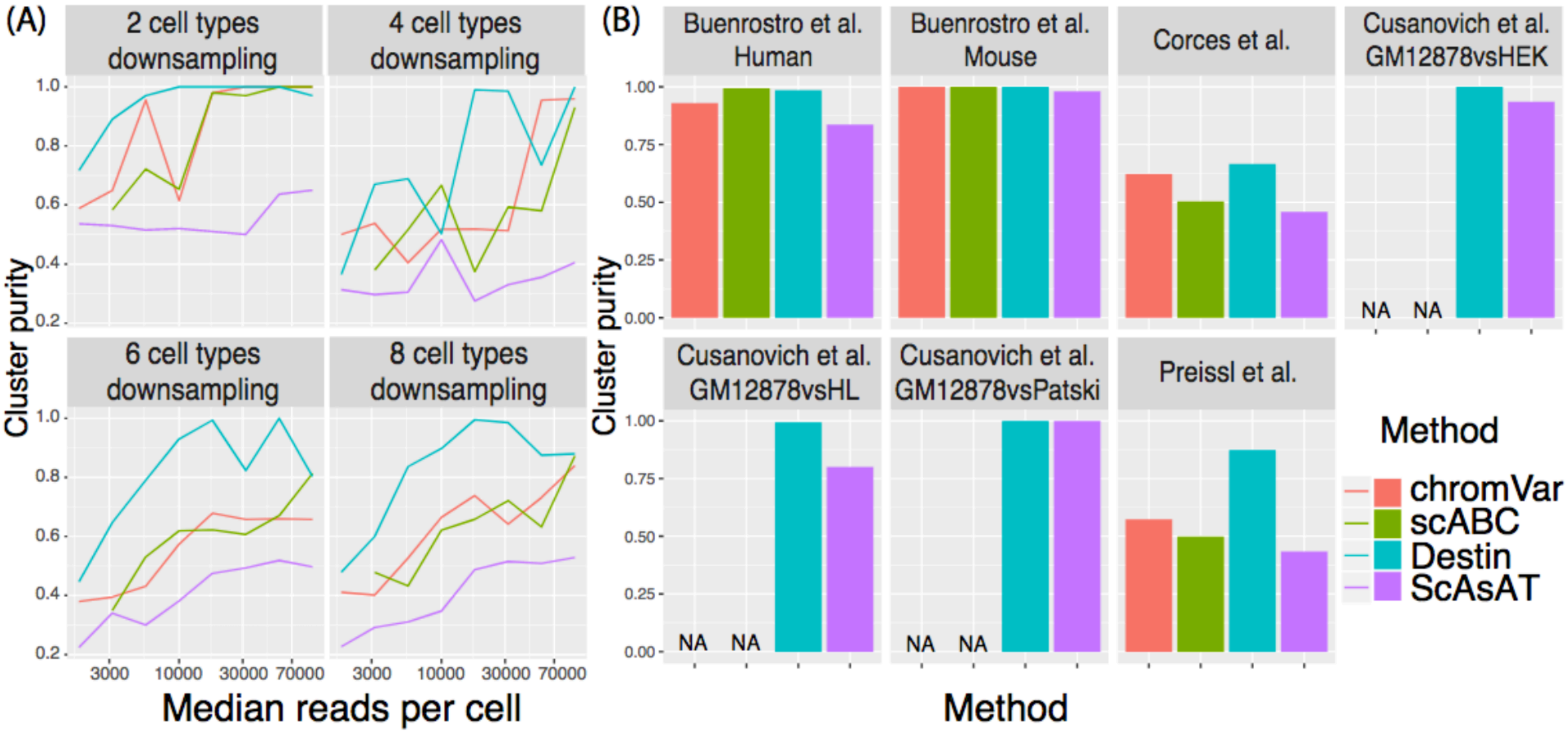
jdoe's PBMC

Other tools

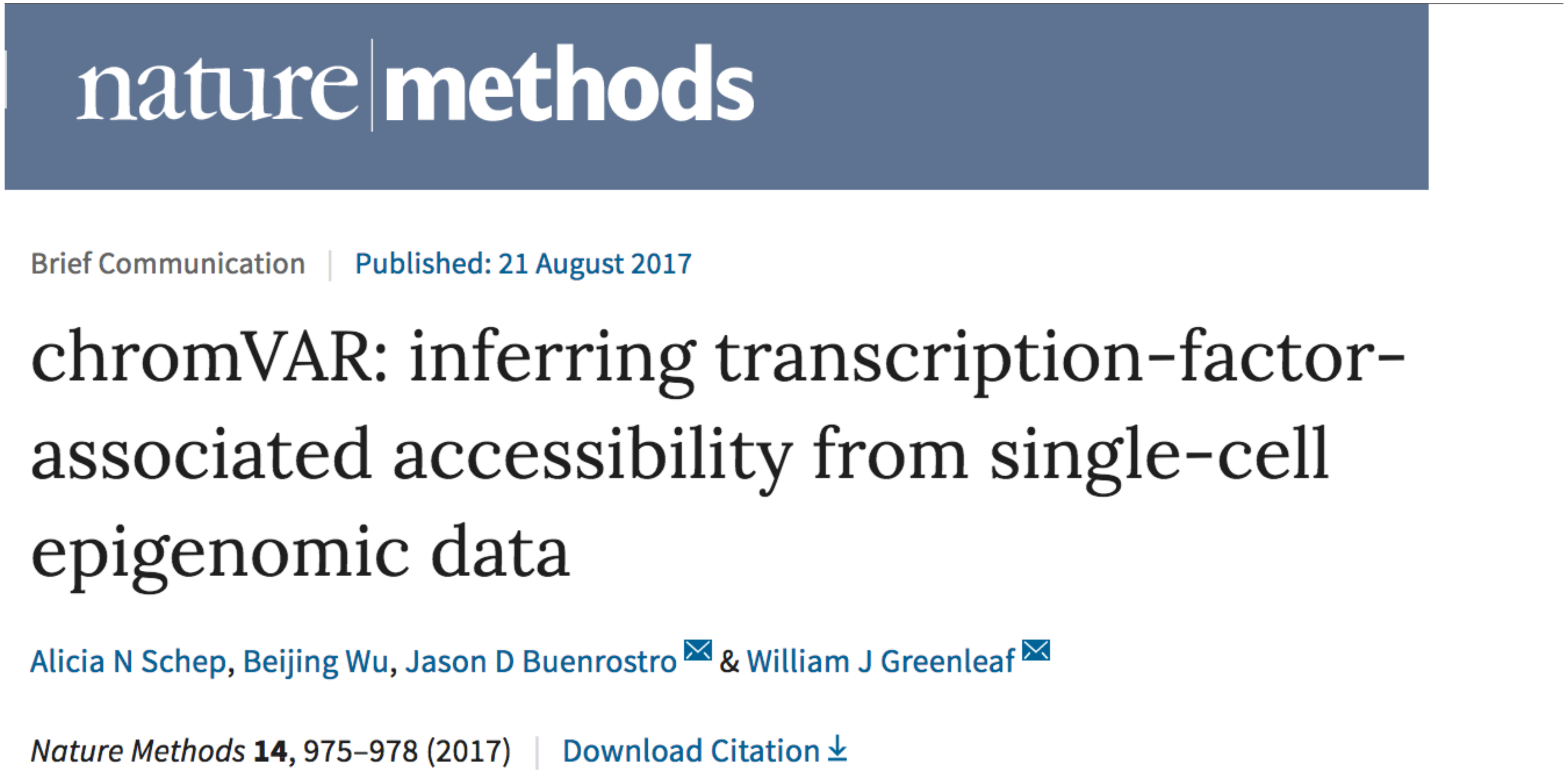
- scABC
- Destin
- ChromVAR
- Cicero
- CoupledNMF
- CisTopic
- Brockman
- SNAP-ATAC
- Signac

Other tools - Destin

Frederick National Laboratory for Cancer Research



ChromVAR



nature methods

Brief Communication | Published: 21 August 2017

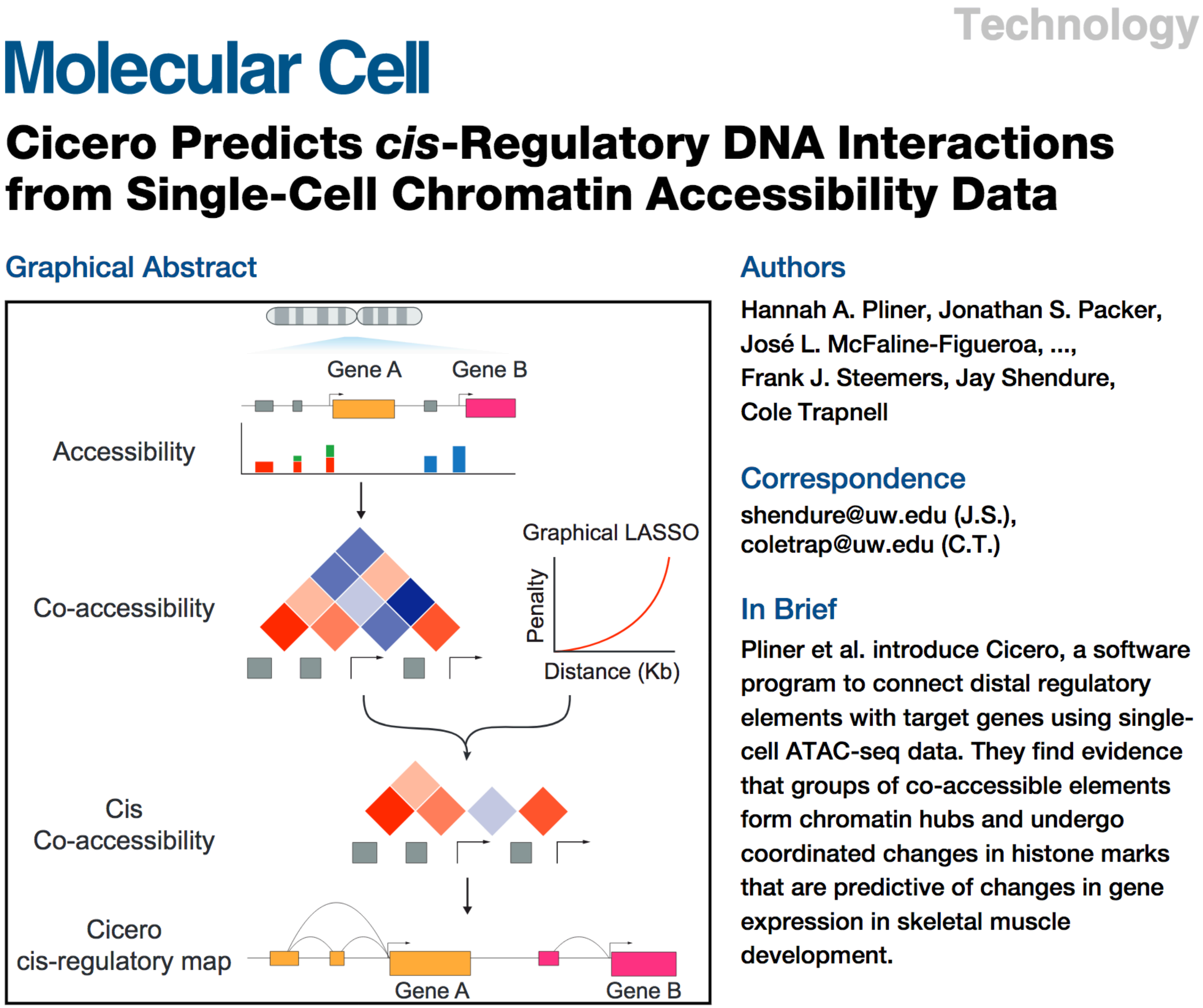
chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data

Alicia N Schep, Beijing Wu, Jason D Buenrostro ✉ & William J Greenleaf ✉

Nature Methods **14**, 975–978 (2017) | Download Citation ⤓

Cicero

Frederick National Laboratory for Cancer Research



Technology

Molecular Cell

Cicero Predicts *cis*-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data

Graphical Abstract

Authors

Hannah A. Pliner, Jonathan S. Packer, José L. McFaline-Figueroa, ..., Frank J. Steemers, Jay Shendure, Cole Trapnell

Correspondence

shendure@uw.edu (J.S.),
coletrap@uw.edu (C.T.)

In Brief

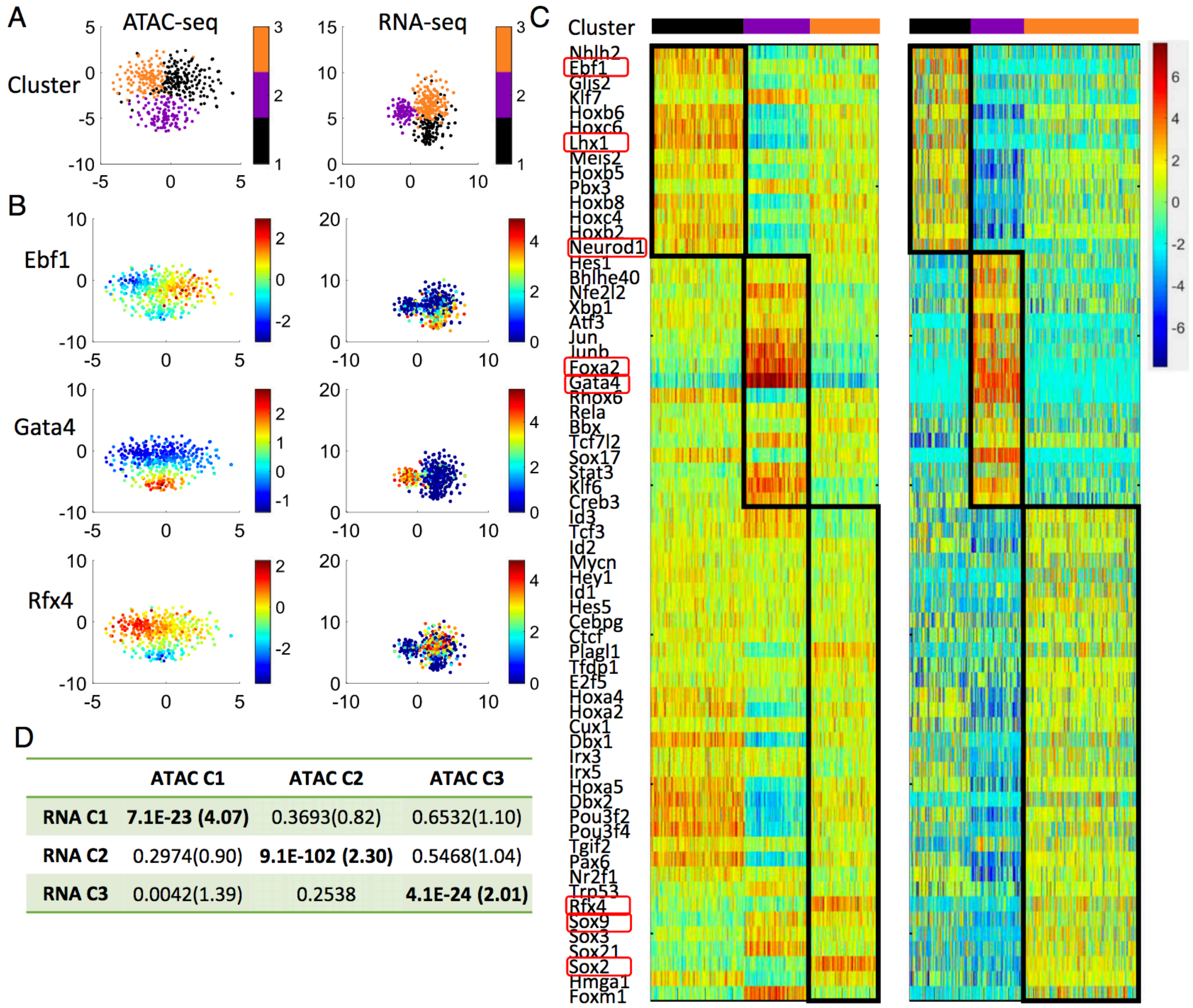
Pliner et al. introduce Cicero, a software program to connect distal regulatory elements with target genes using single-cell ATAC-seq data. They find evidence that groups of co-accessible elements form chromatin hubs and undergo coordinated changes in histone marks that are predictive of changes in gene expression in skeletal muscle development.

CoupledNMF

Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations

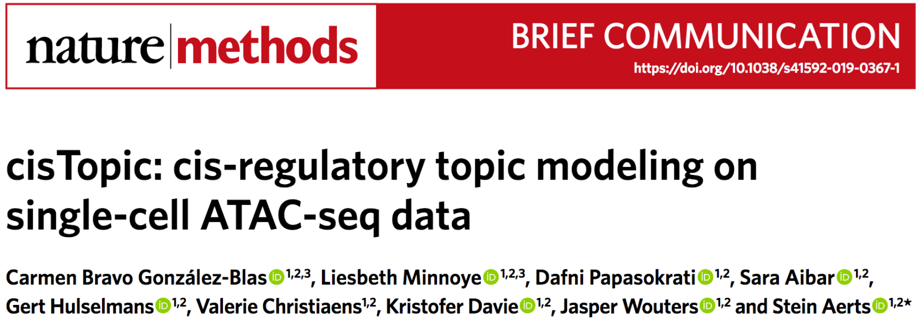
Zhana Duren[a,b,1], Xi Chen[a,b,1], Mahdi Zamanighomi[a,b,c,1], Wanwen Zeng[a,b,d], Ansuman T. Satpathy[c], Howard Y. Chang[c], Yong Wang[e,f], and Wing Hung Wong[a,b,c,2]

[a]Department of Statistics, Stanford University, Stanford, CA 94305; [b]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; [c]Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305; [d]Ministry of Education Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic & Systems Biology, Department of Automation, Tsinghua University, 100084 Beijing, China; [e]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 100080 Beijing, China; and [f]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, 650223 Kunming, China



	ATAC C1	ATAC C2	ATAC C3
RNA C1	**7.1E-23 (4.07)**	0.3693(0.82)	0.6532(1.10)
RNA C2	0.2974(0.90)	**9.1E-102 (2.30)**	0.5468(1.04)
RNA C3	0.0042(1.39)	0.2538	**4.1E-24 (2.01)**

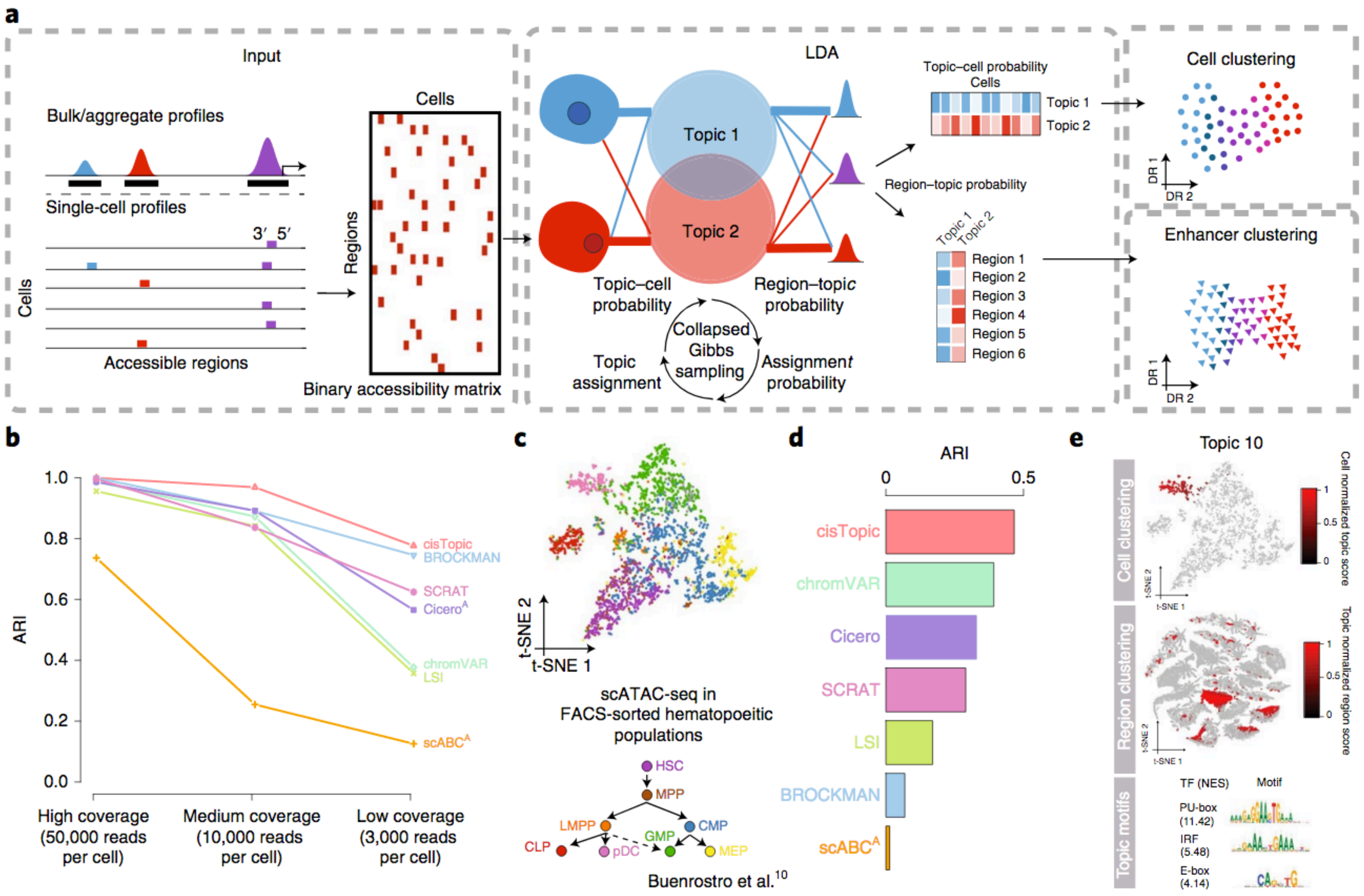
CisTopic



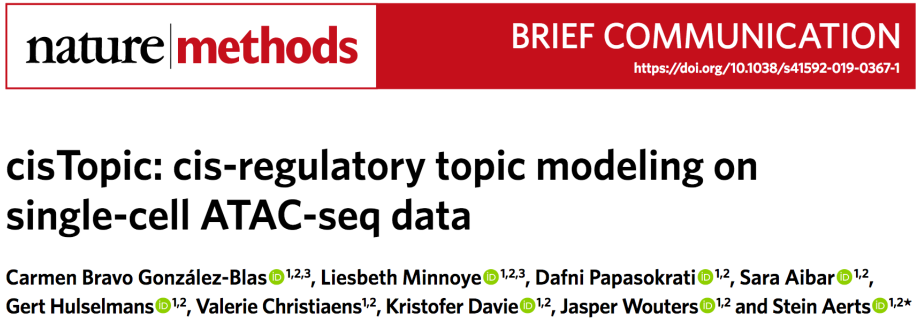
nature methods
BRIEF COMMUNICATION
https://doi.org/10.1038/s41592-019-0367-1

cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data

Carmen Bravo González-Blas[1,2,3], Liesbeth Minnoye[1,2,3], Dafni Papasokrati[1,2], Sara Aibar[1,2], Gert Hulselmans[1,2], Valerie Christiaens[1,2], Kristofer Davie[1,2], Jasper Wouters[1,2] and Stein Aerts[1,2*]



CisTopic



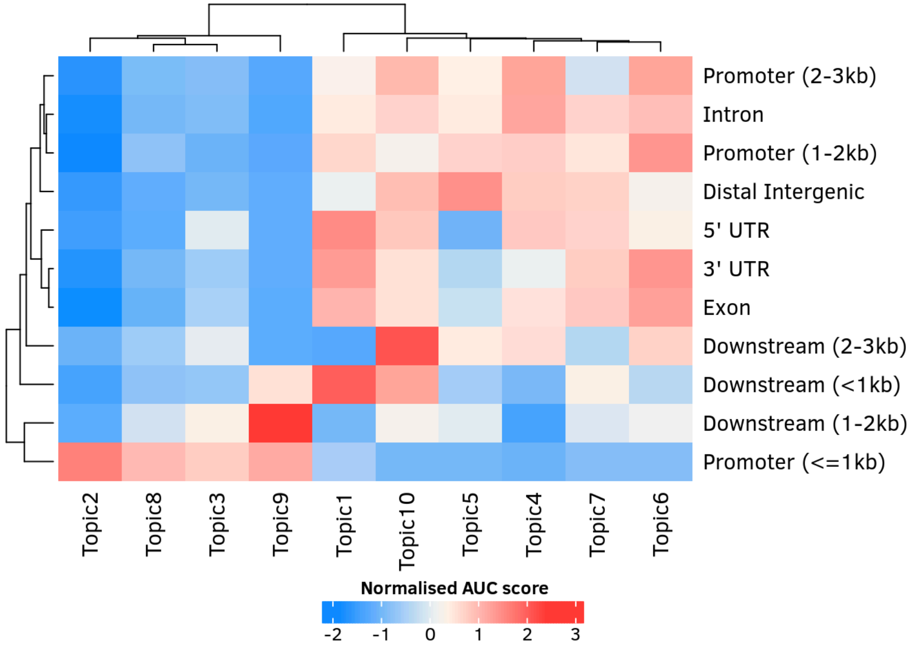
nature methods

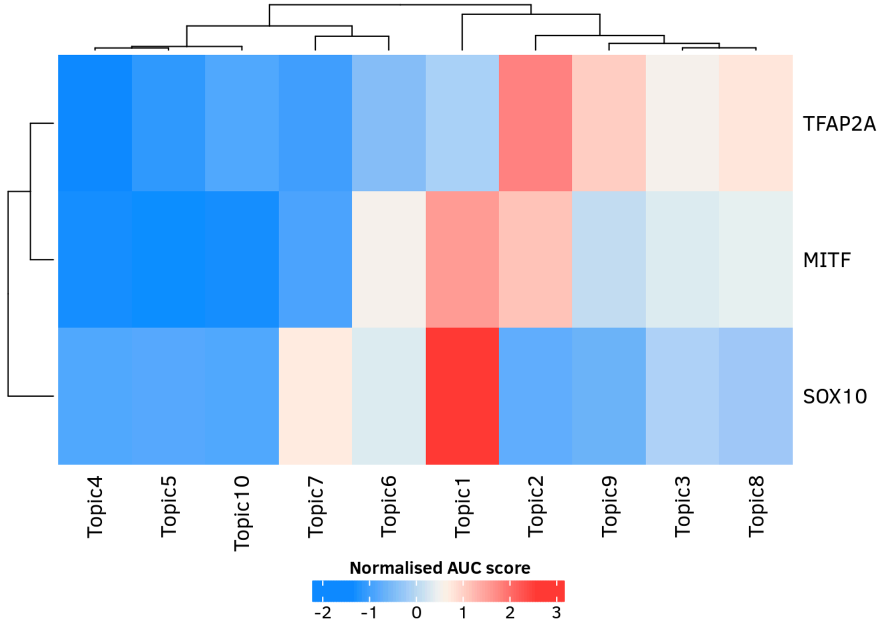
BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-019-0367-1

cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data

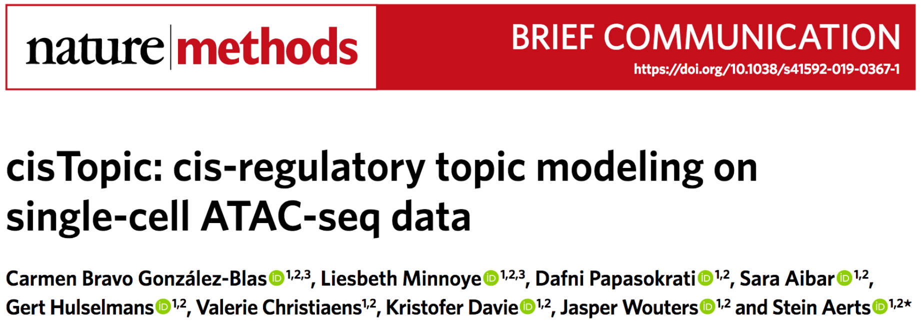
Carmen Bravo González-Blas[1,2,3], Liesbeth Minnoye[1,2,3], Dafni Papasokrati[1,2], Sara Aibar[1,2], Gert Hulselmans[1,2], Valerie Christiaens[1,2], Kristofer Davie[1,2], Jasper Wouters[1,2] and Stein Aerts[1,2*]





CisTopic

Frederick National Laboratory for Cancer Research



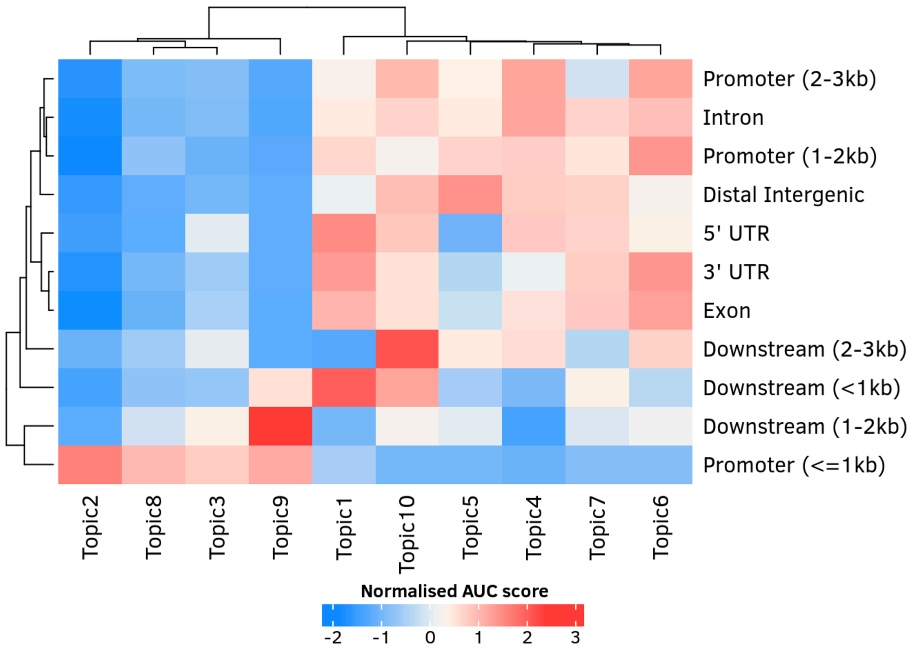
nature methods

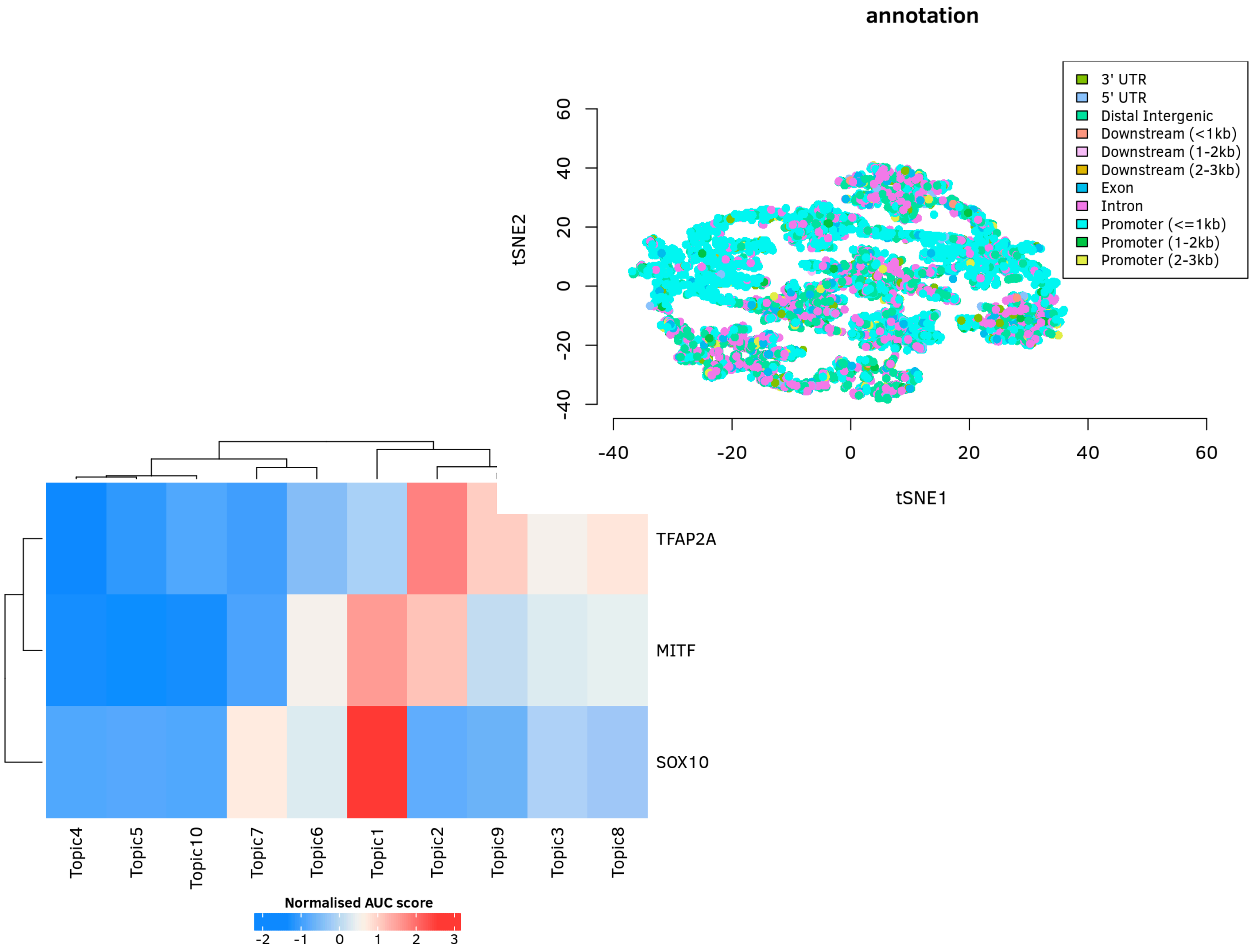
BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-019-0367-1

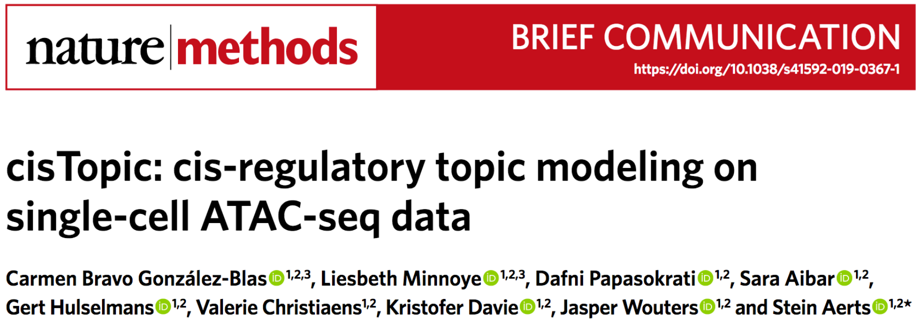
cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data

Carmen Bravo González-Blas[1,2,3], Liesbeth Minnoye[1,2,3], Dafni Papasokrati[1,2], Sara Aibar[1,2], Gert Hulselmans[1,2], Valerie Christiaens[1,2], Kristofer Davie[1,2], Jasper Wouters[1,2] and Stein Aerts[1,2*]





CisTopic



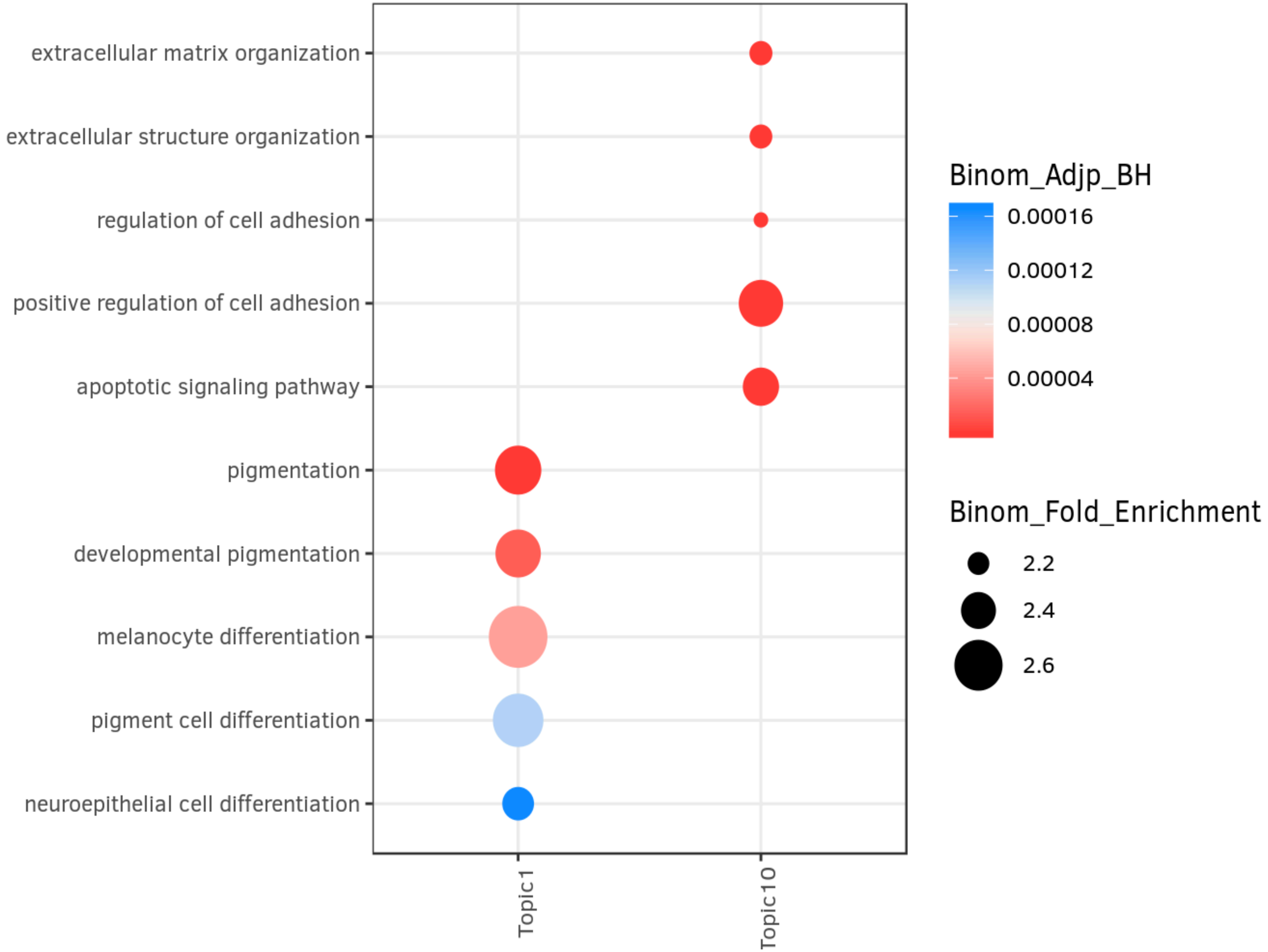
nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-019-0367-1

cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data

Carmen Bravo González-Blas[1,2,3], Liesbeth Minnoye[1,2,3], Dafni Papasokrati[1,2], Sara Aibar[1,2], Gert Hulselmans[1,2], Valerie Christiaens[1,2], Kristofer Davie[1,2], Jasper Wouters[1,2] and Stein Aerts[1,2*]

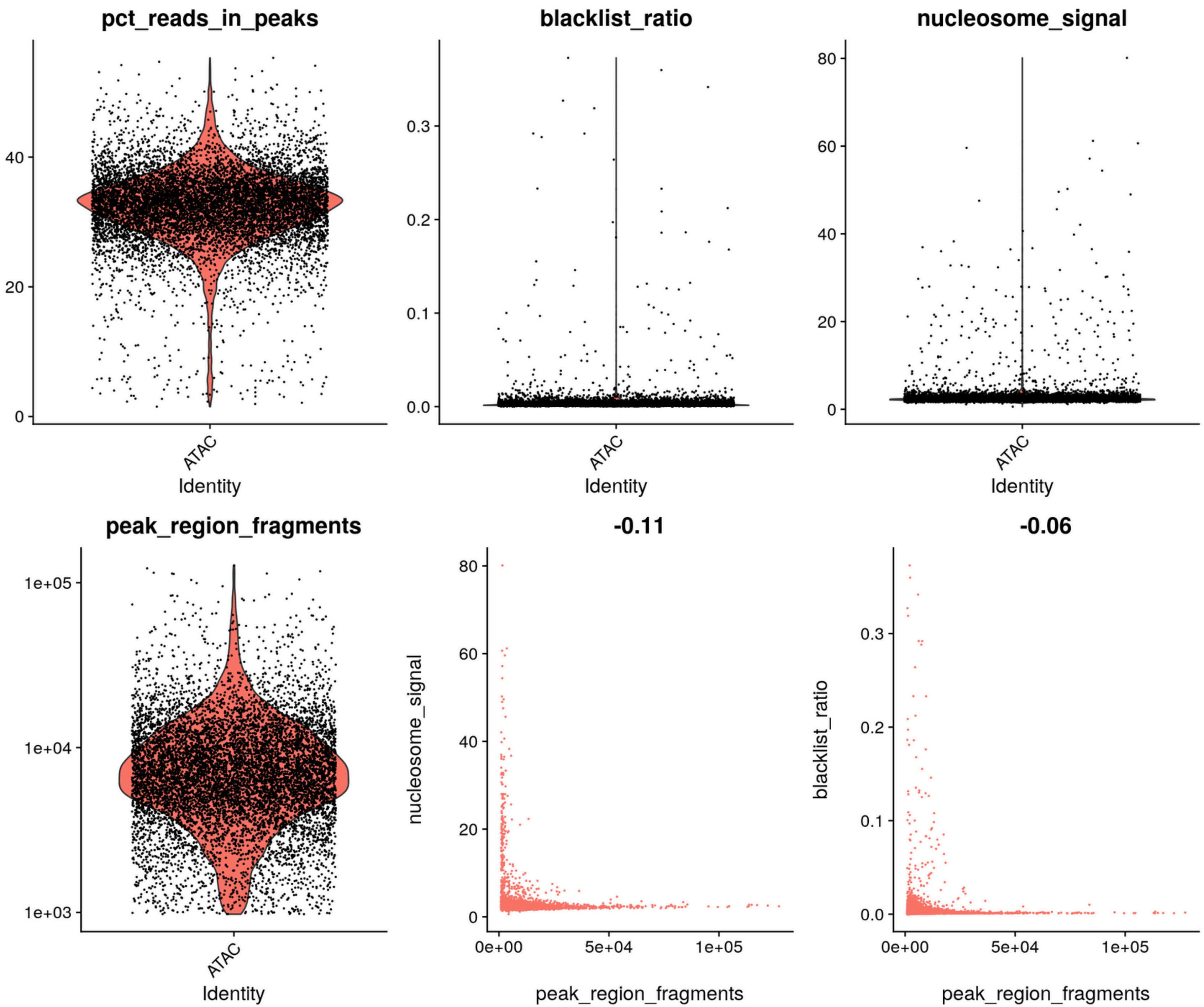


Signac

- Signac is an extension of Seurat for the analysis, interpretation, and exploration of single-cell chromatin datasets.
- Calculating single-cell QC metrics
- Dimensional reduction, visualization, and clustering
- Identifying cell type-specific peaks
- Visualizing ‘pseudo-bulk’ coverage tracks
- Integration of multiple single-cell ATAC-seq datasets
- Integration with single-cell RNA-seq datasets
- Motif enrichment analysis

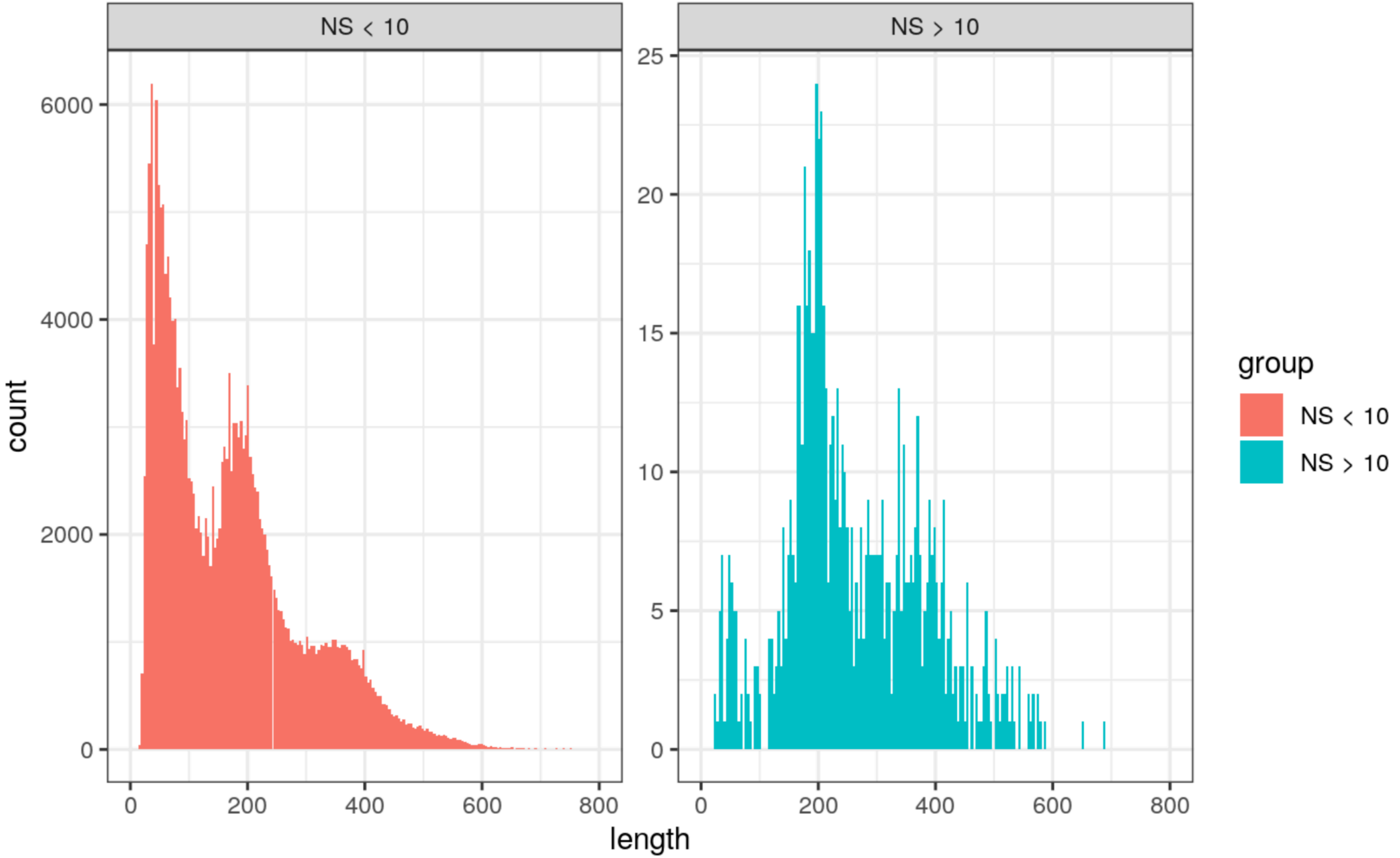
- Integration with other single cell tools (Harmony, Cicero, Chromvar)

Signac - QC



https://satijalab.org/signac/index.html

Signac - QC


NS < 10
NS > 10
count
6000
4000
2000
0
25
20
15
10
5
0
0
200
400
600
800
length
group
🟥 NS < 10
🟦 NS > 10

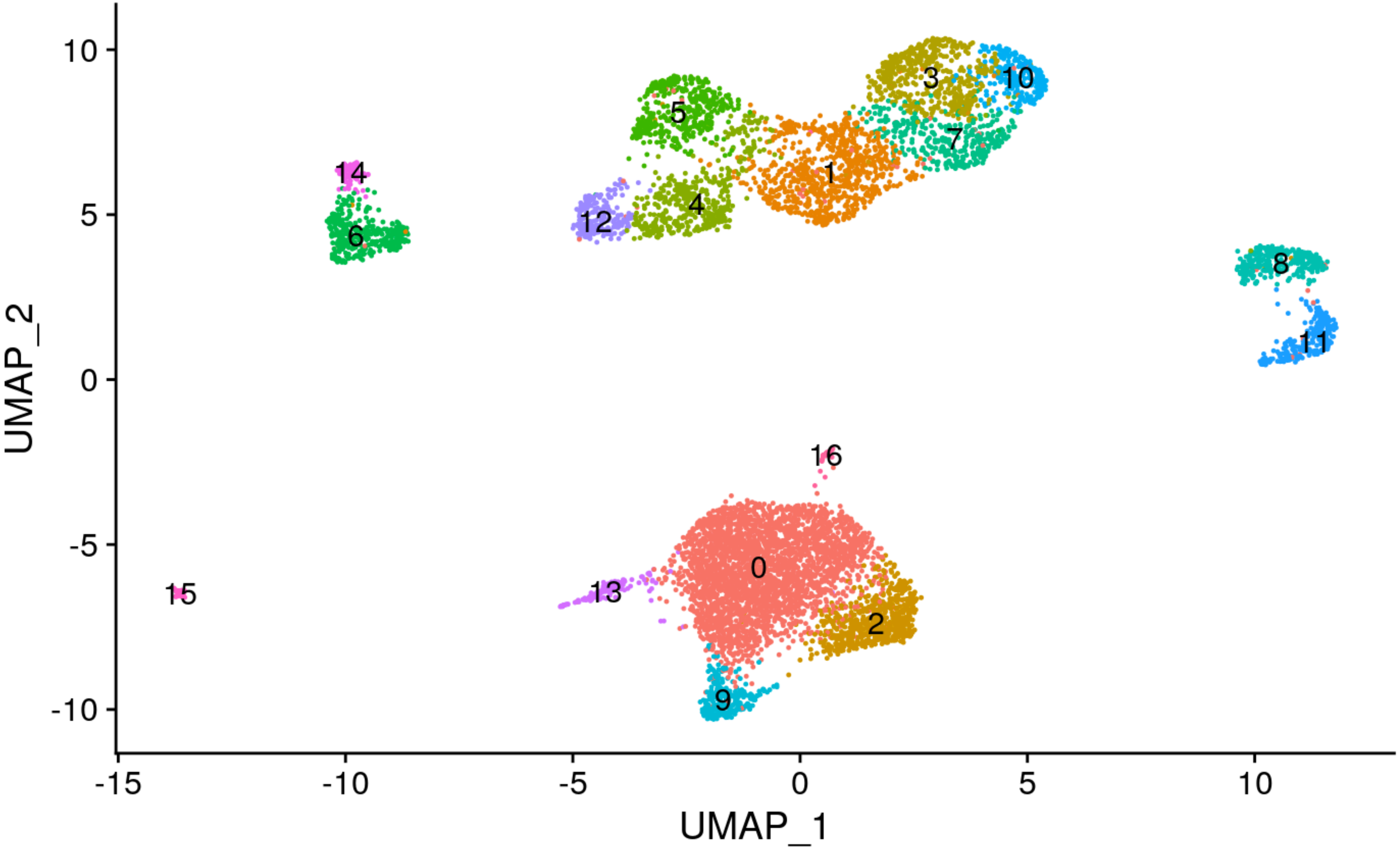
Signac- Normalization and linear dimensional reduction

- Normalization: Signac performs term frequency-inverse document frequency (TF-IDF) normalization.
  - two-step normalization procedure,
    - normalizes across cells to correct for differences in cellular sequencing depth & across peaks to give higher values to more rare peaks.

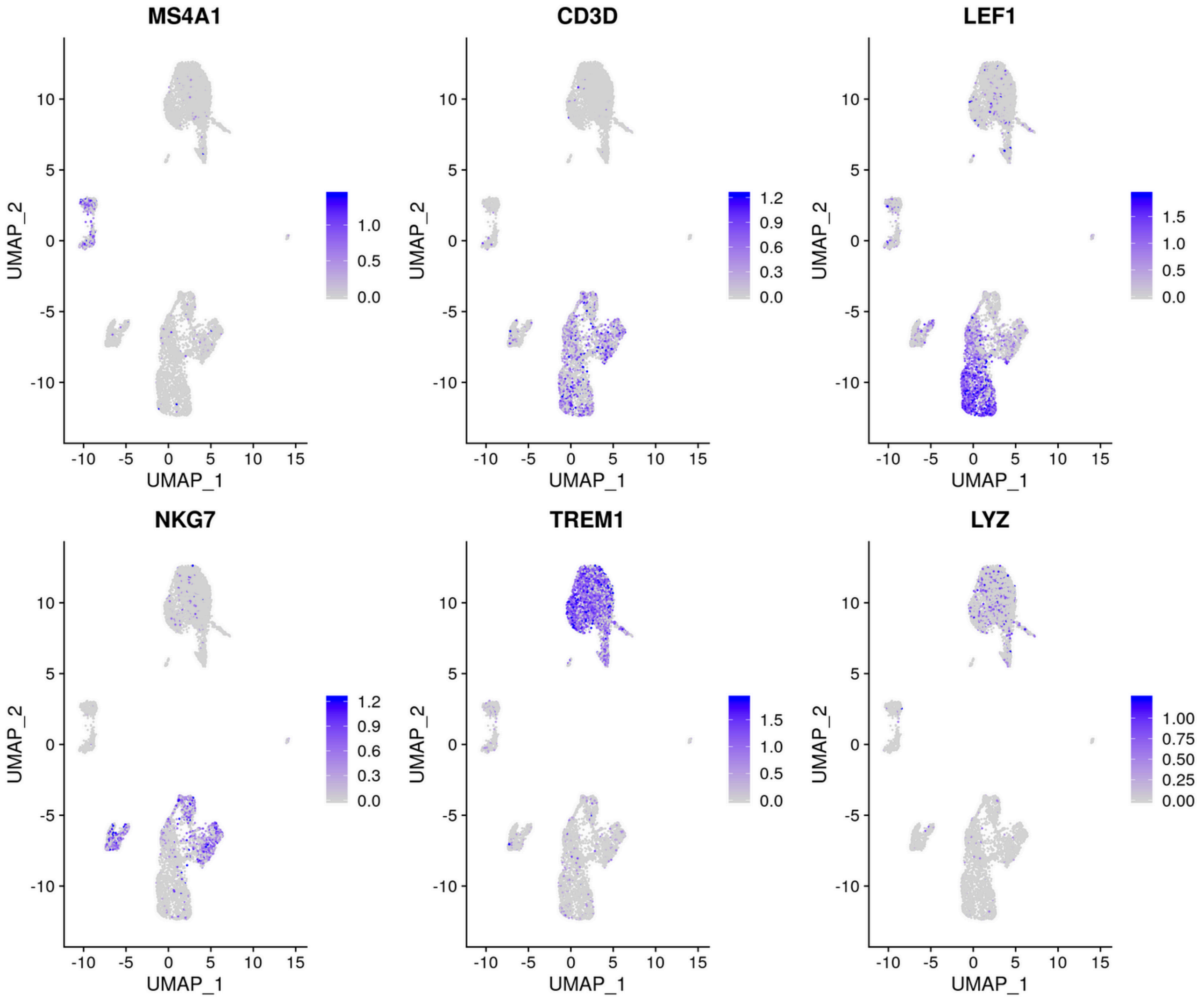
- Feature selection: Binary nature of scATAC-seq data makes it challenging to perform ‘variable’ feature selection, as we do for scRNA-seq.
  - Instead, use only the top *n*% of features (peaks) for dimensional reduction,
  - or remove features present in less that *n* cells with the FindTopFeatures function.

- Dimensional reduction: Singular value decomposition (SVD) on the TD-IDF normalized matrix, using the features (peaks) selected above. This returns a low-dimensional representation of the object (for users who are more familiar with scRNA-seq, you can think of this as analogous to the output of PCA)

Signac- Clustering

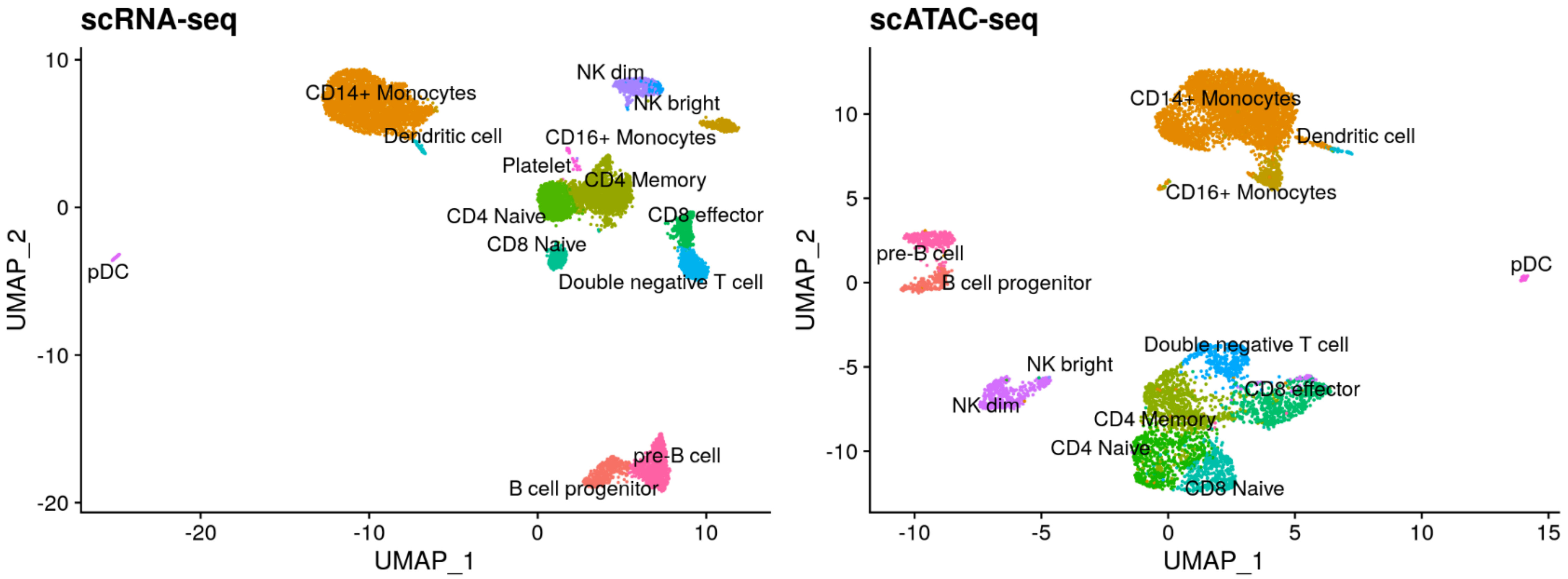


Signac- Cicero: *gene activity matrix*

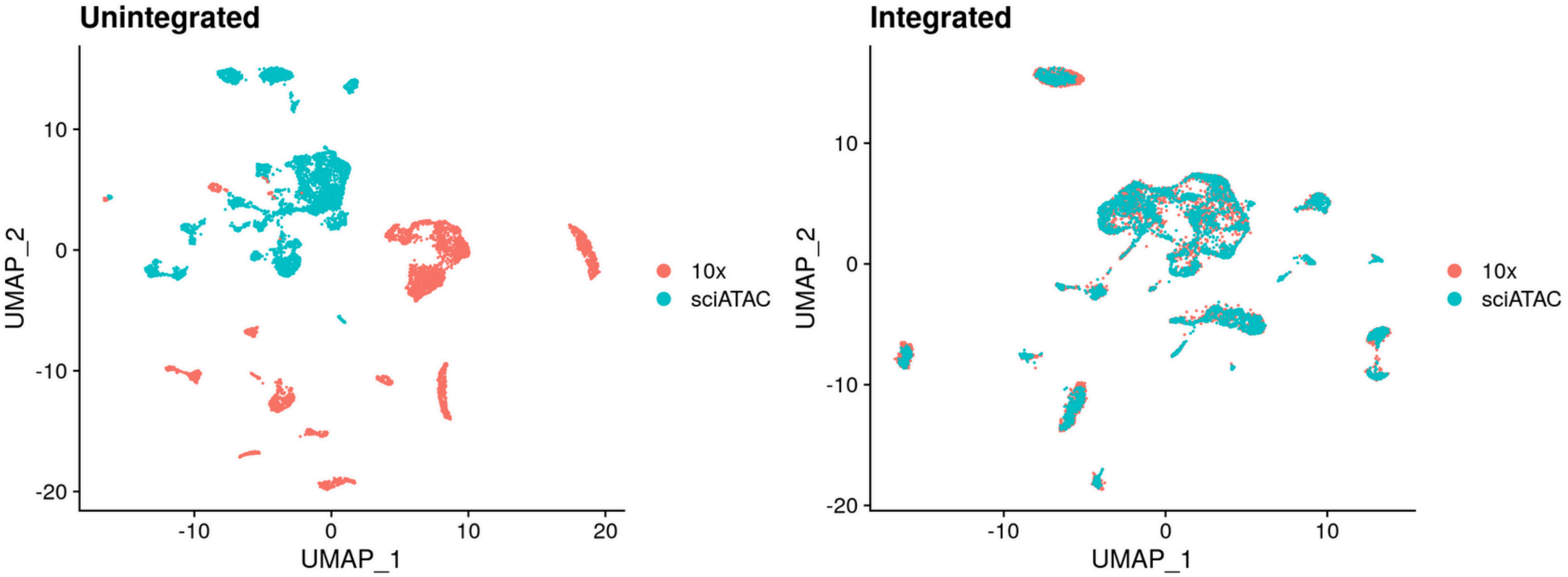


scATAC + scRNA integration

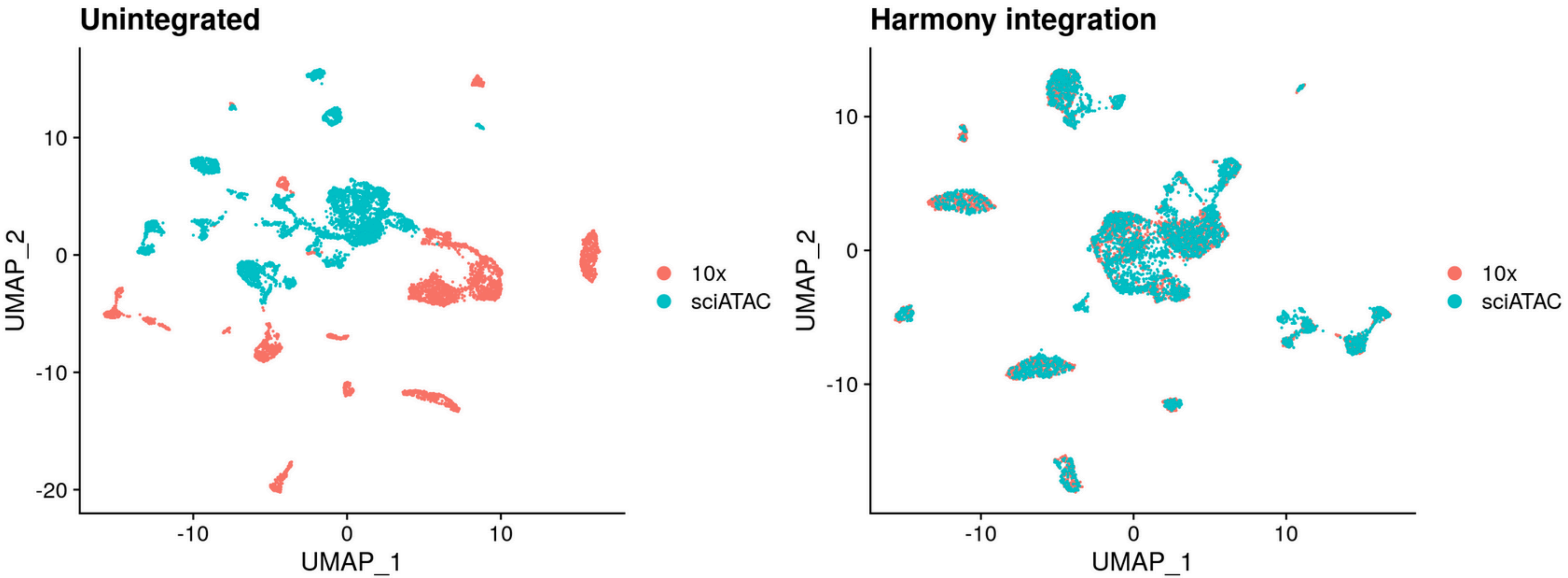
Signac- scATAC + scRNA using Seurat



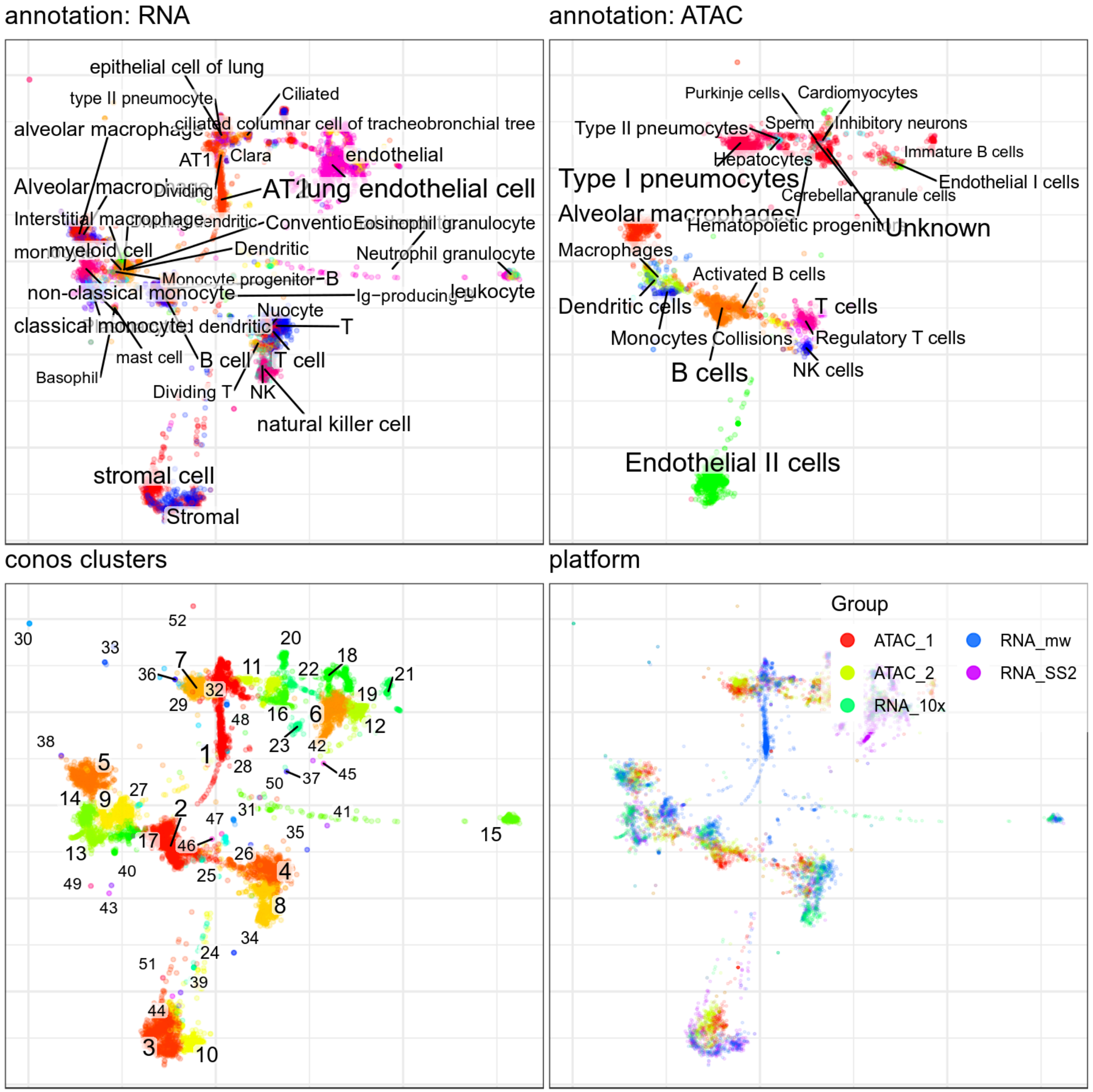
Signac- scATAC + scRNA using Seurat



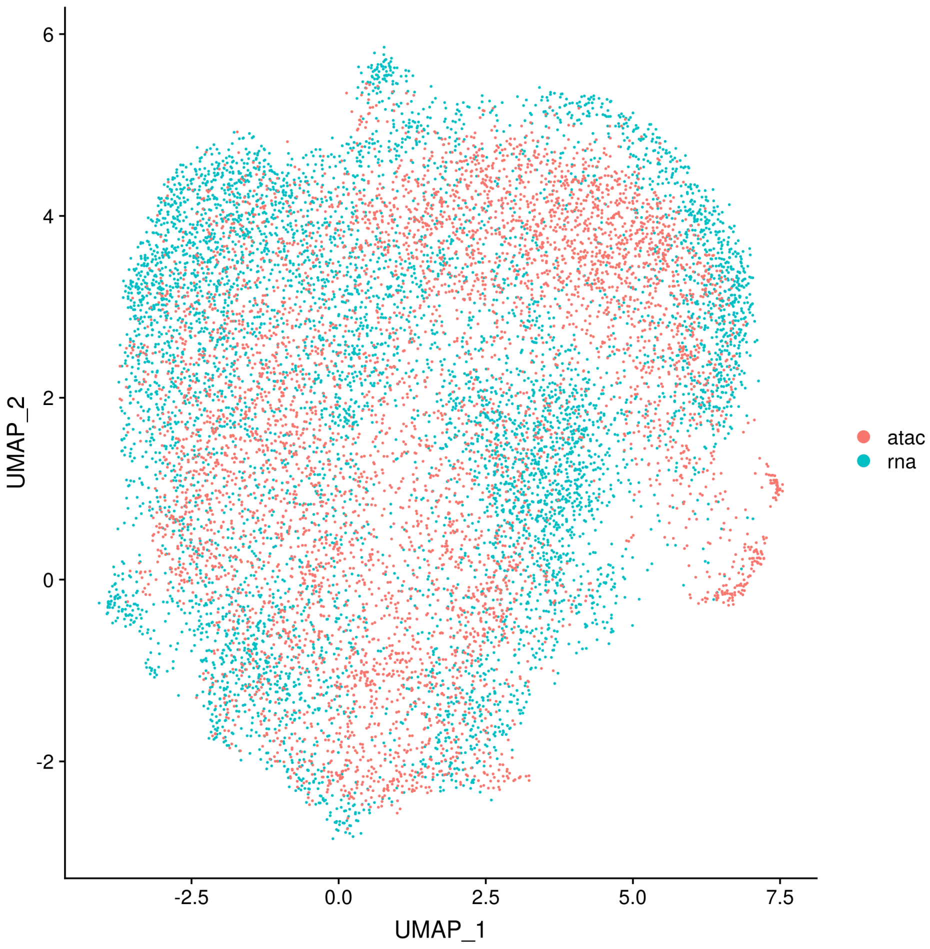
Signac- scATAC + scRNA using Harmony

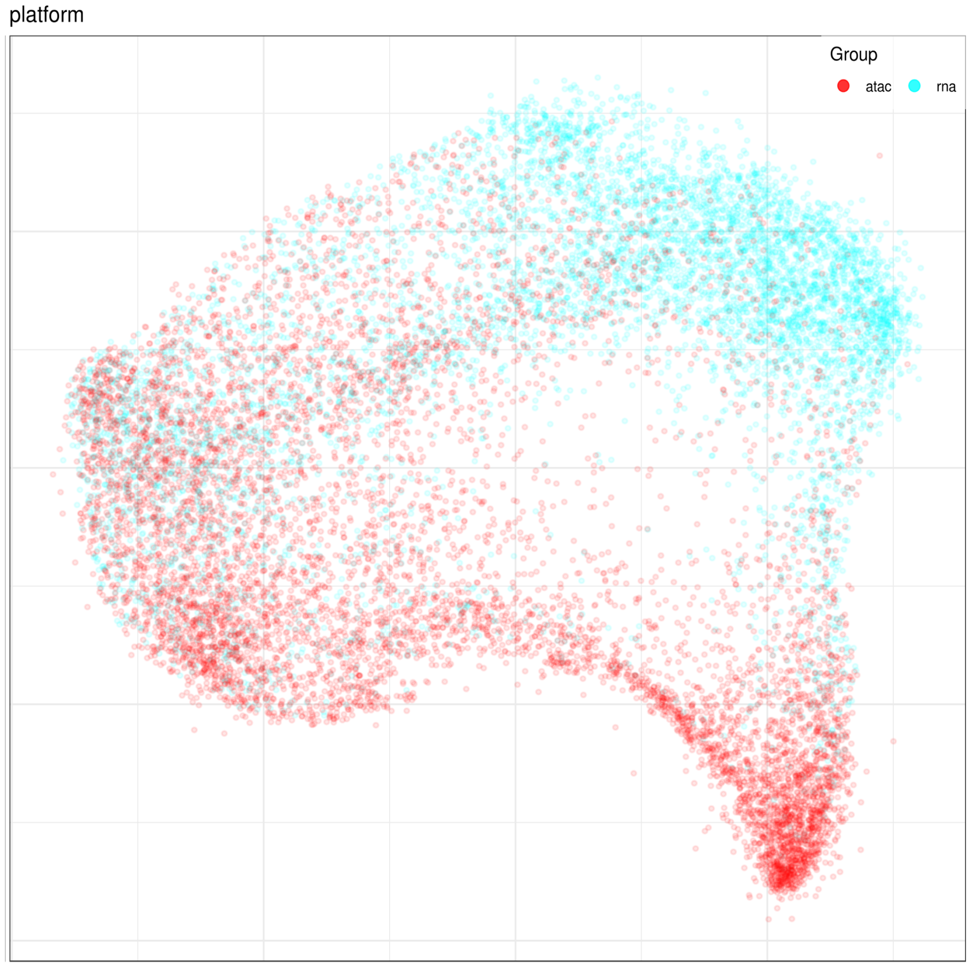


Signac- scATAC + scRNA using Conos



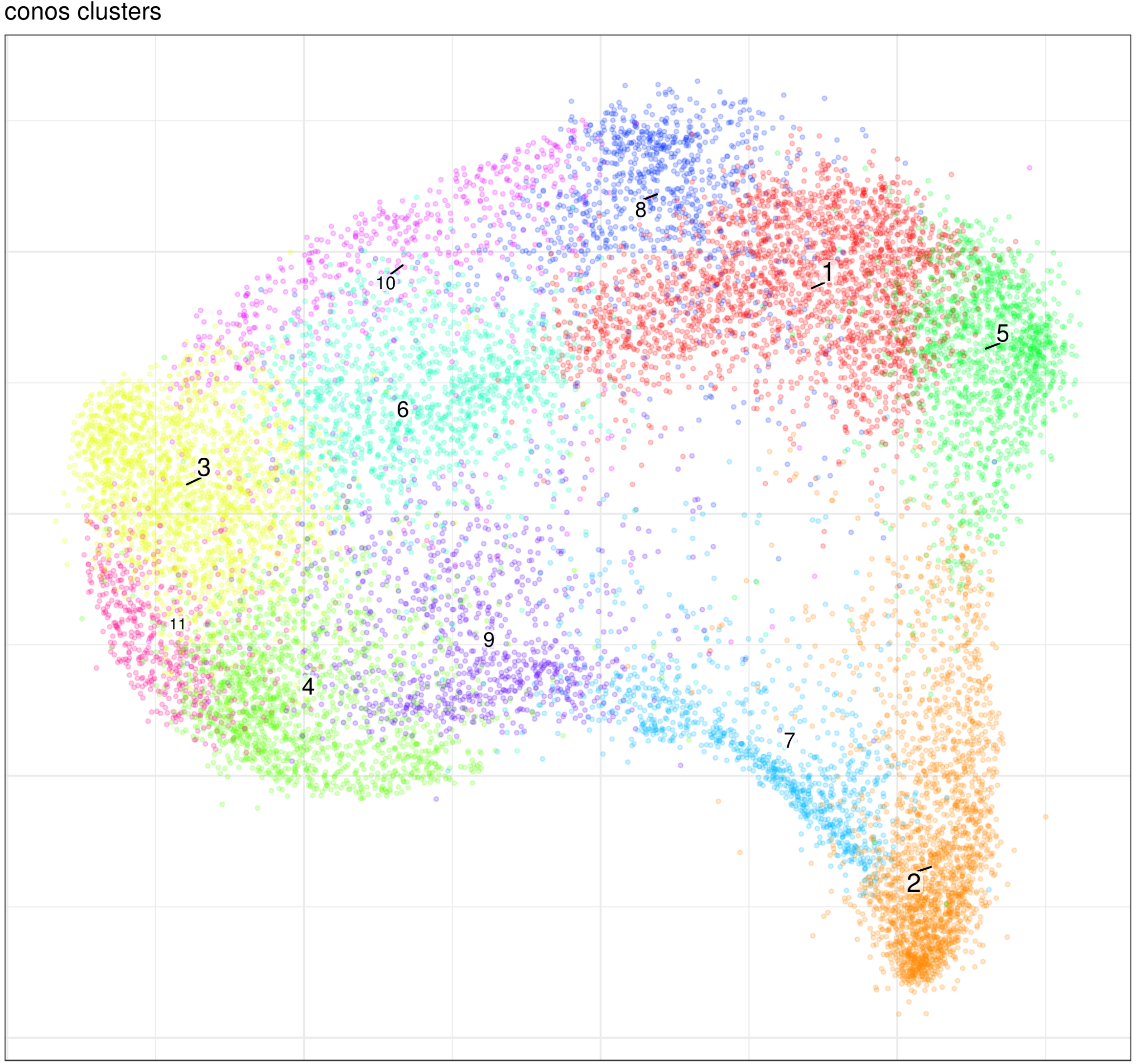
Only Seurat vs Seurat + Conos

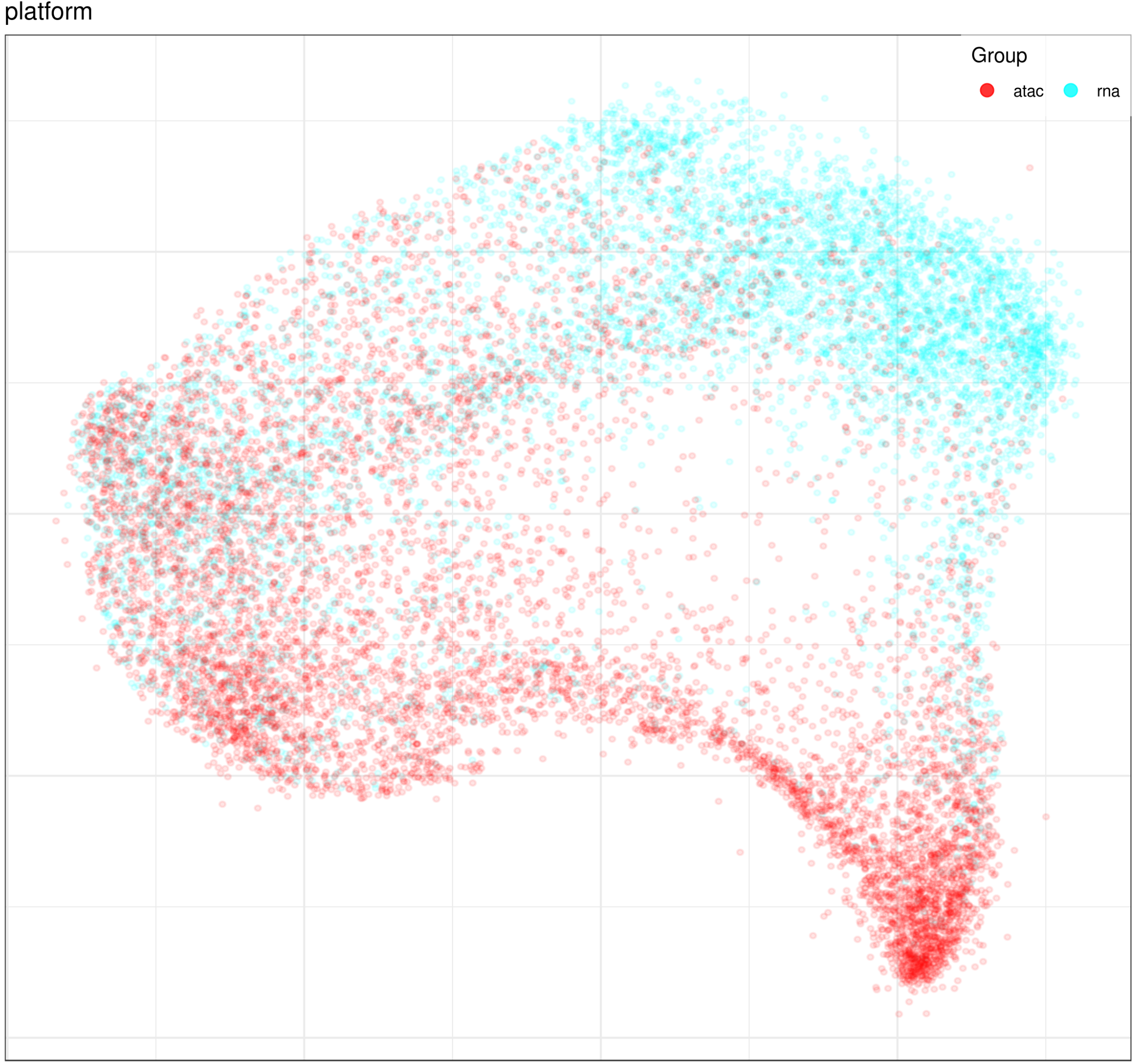




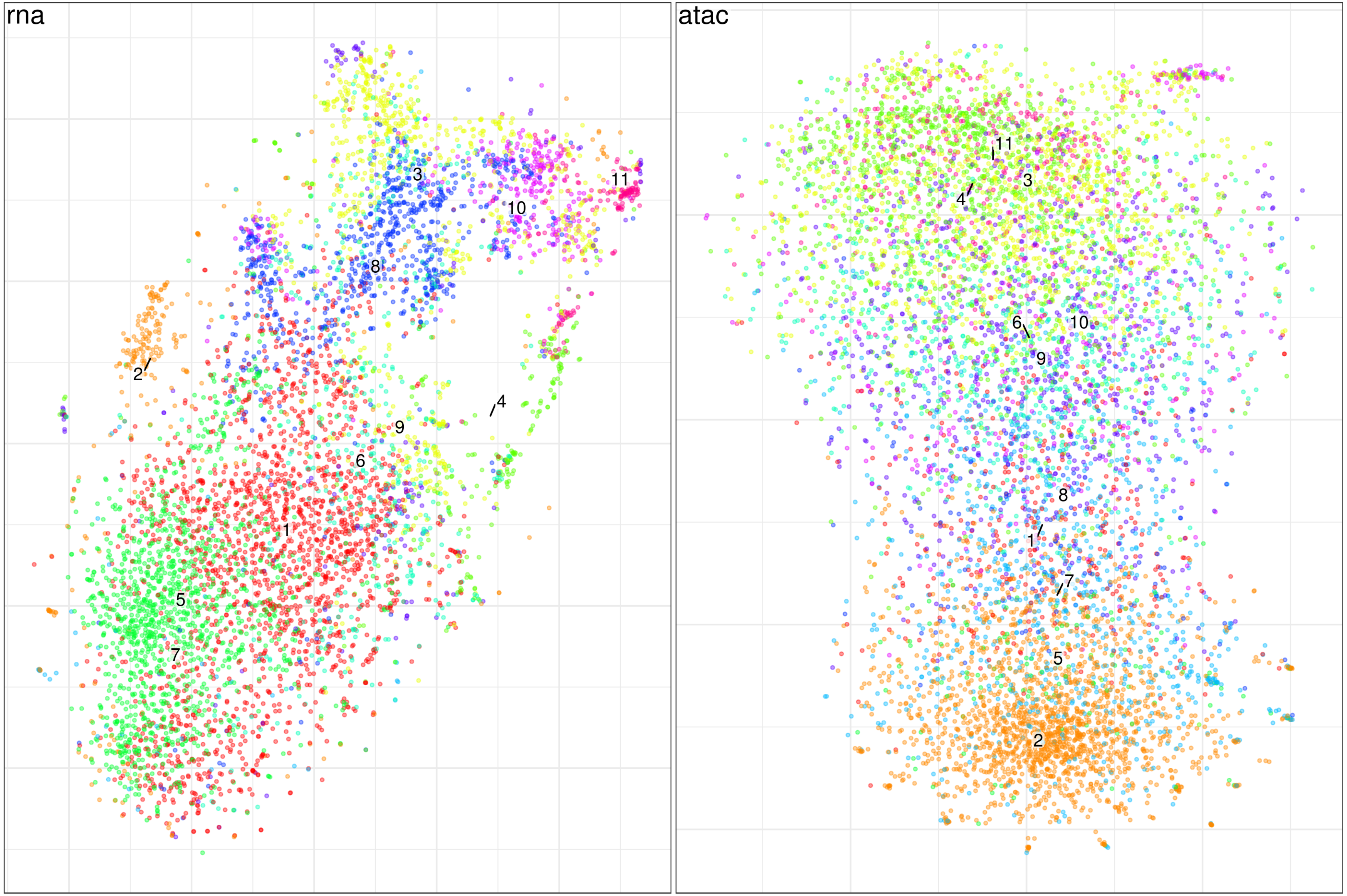
Seurat + Conos

Frederick National Laboratory for Cancer Research

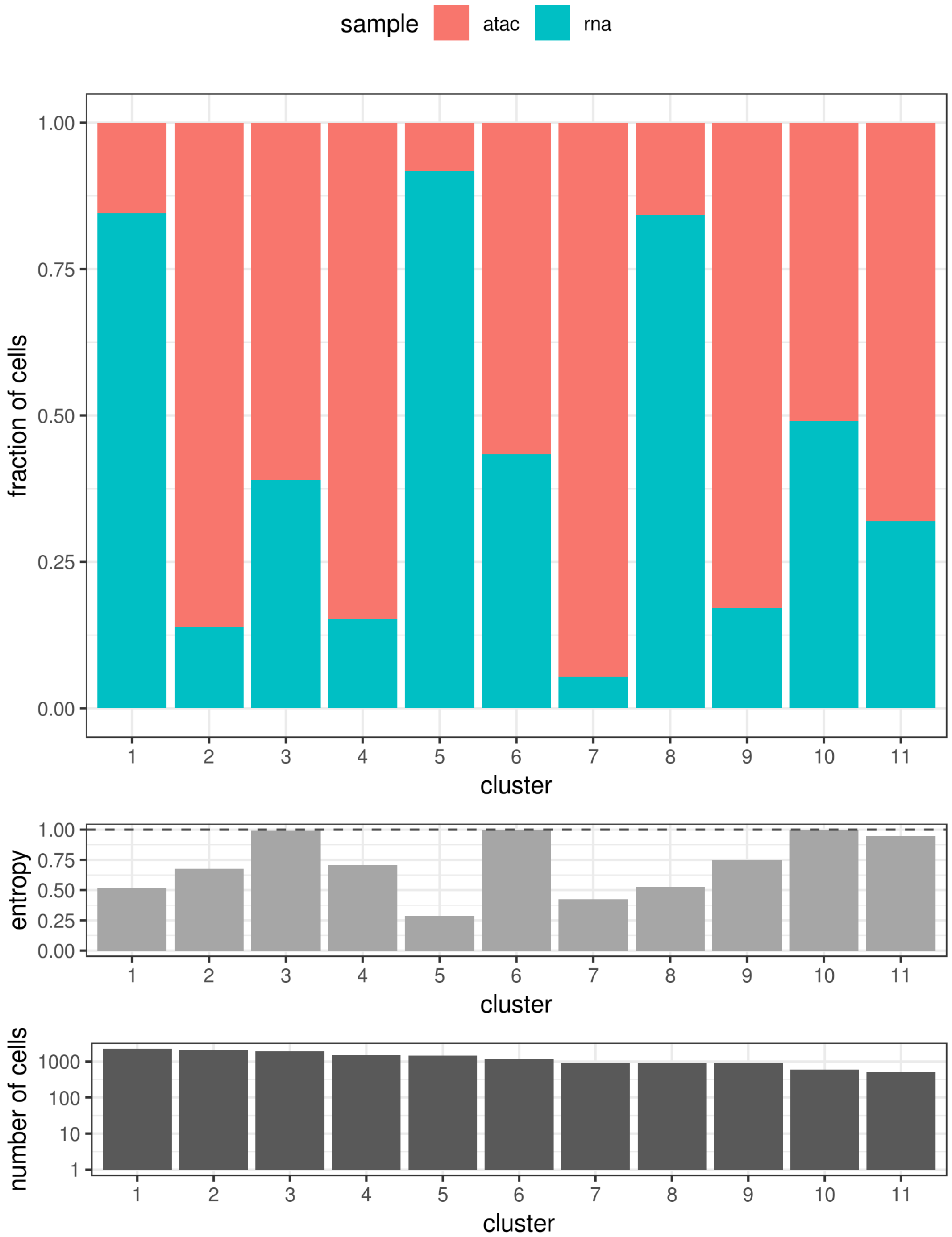




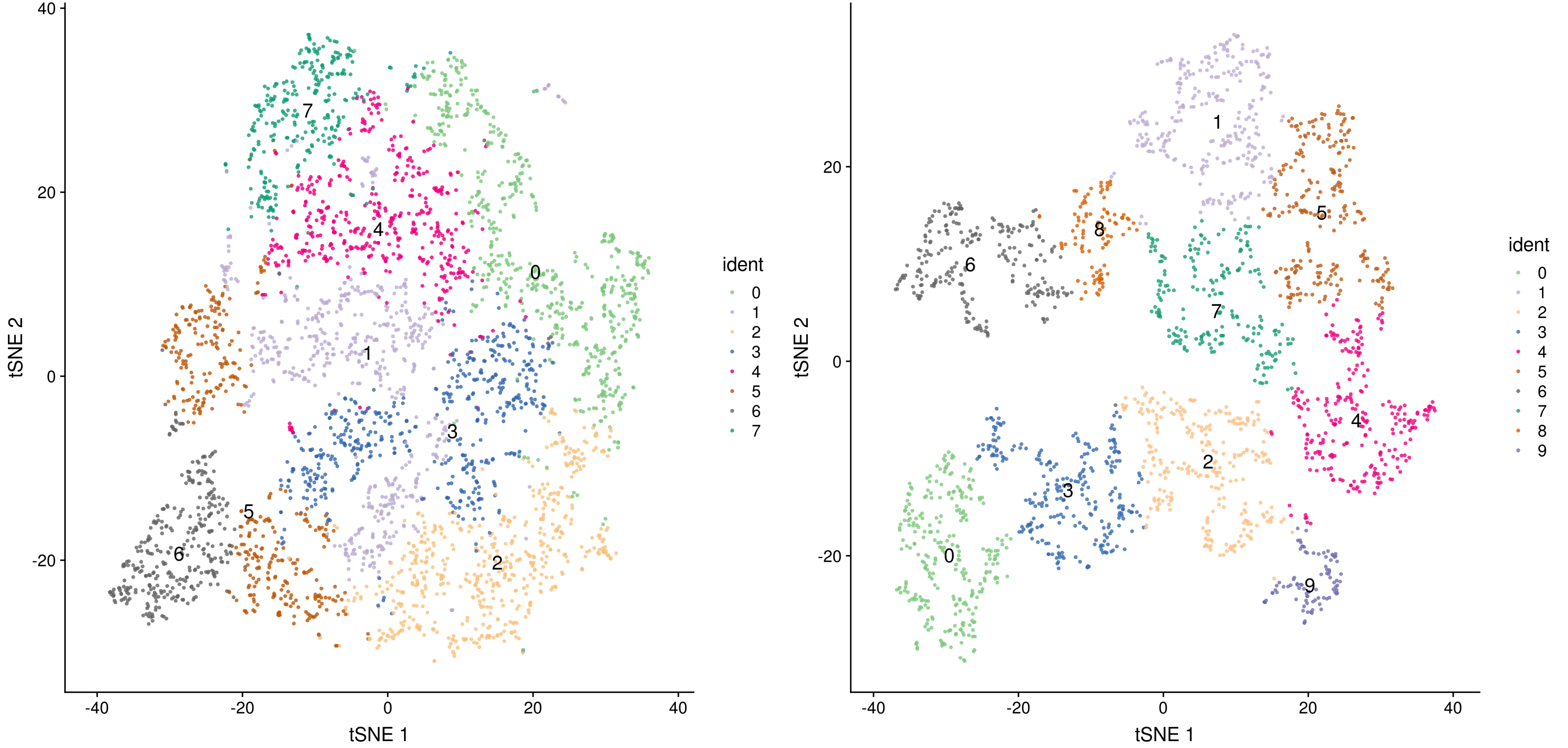
Seurat + Conos



Seurat + Conos

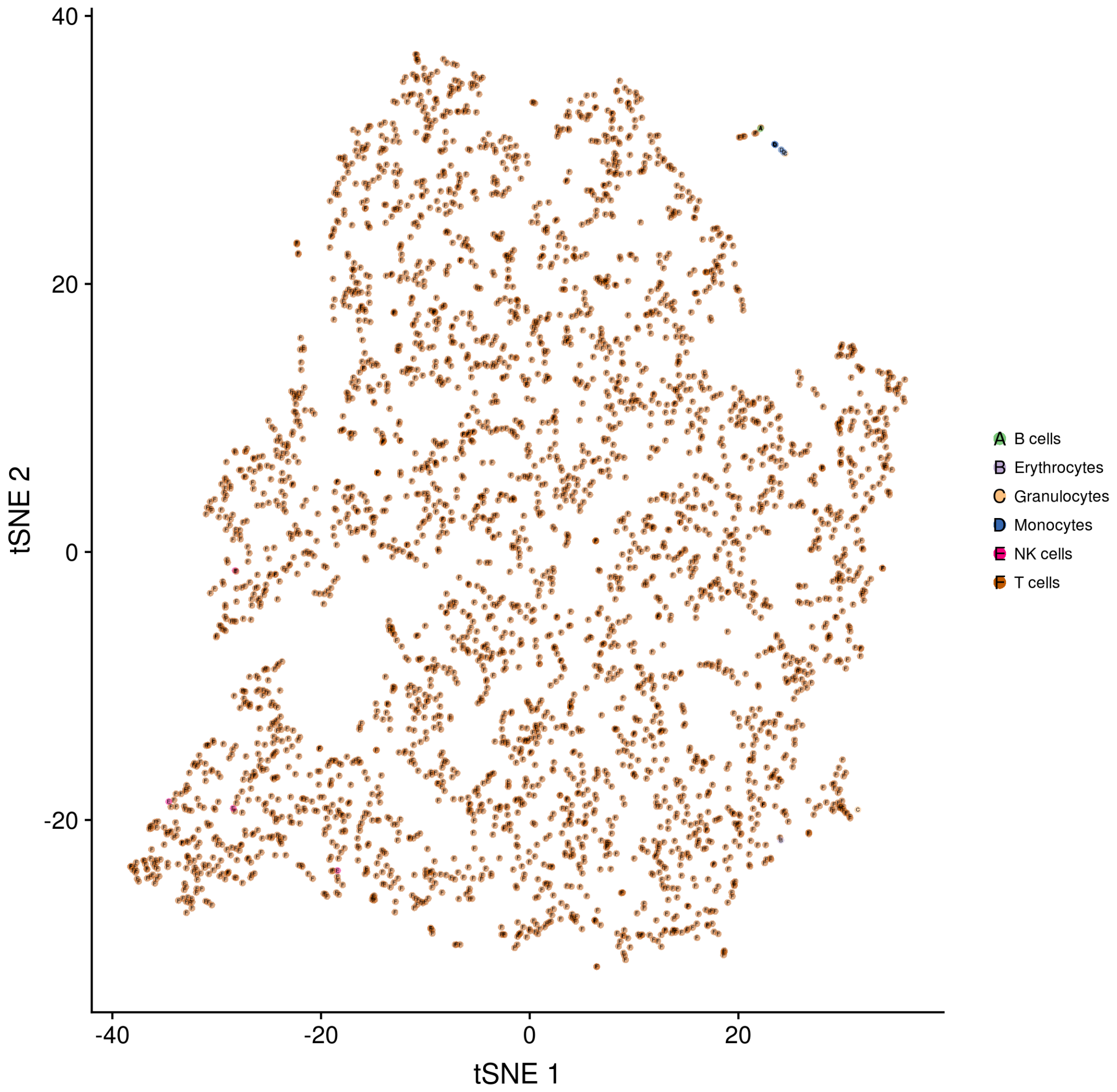


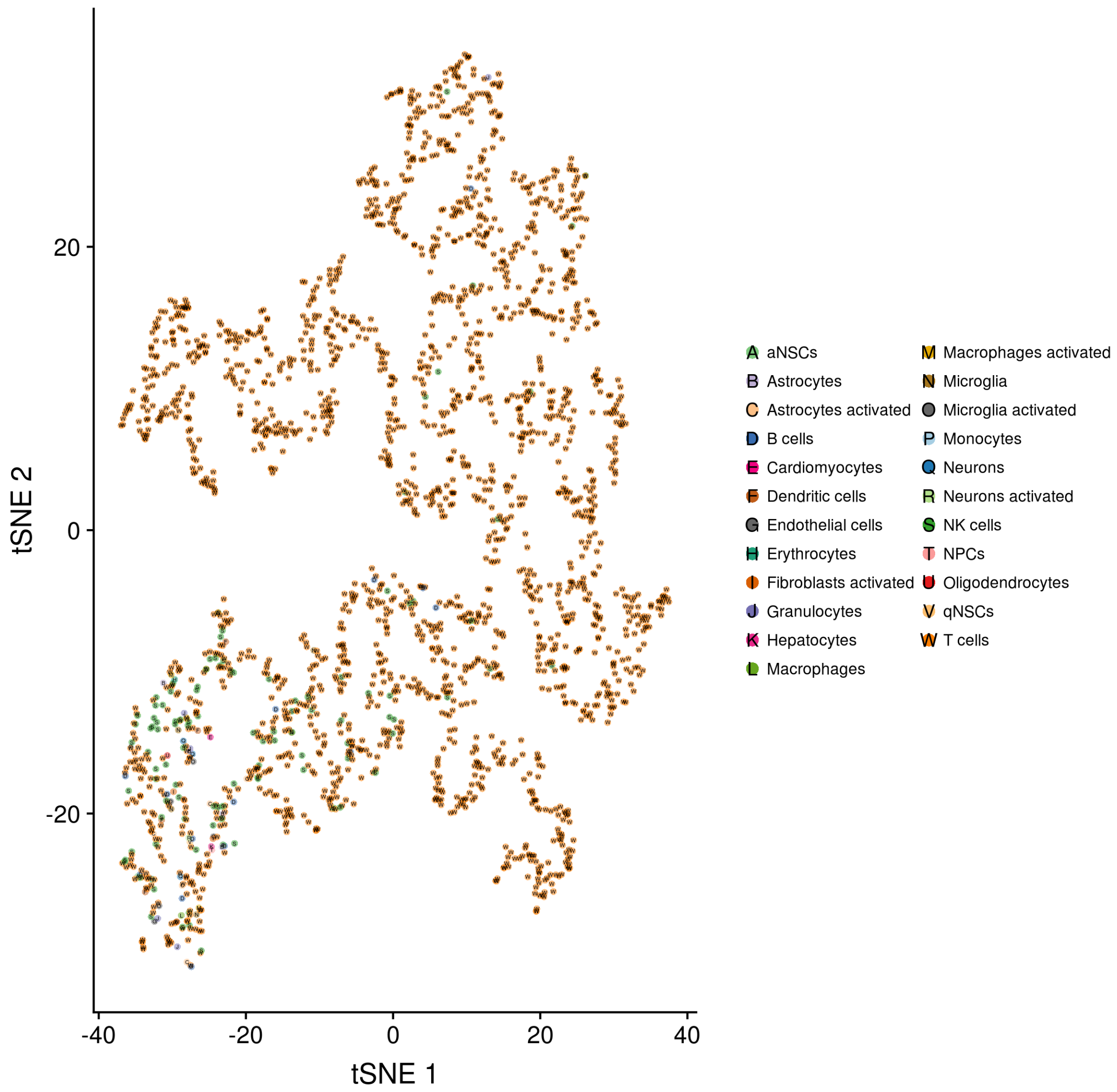
scRNA + scATAC - Compare



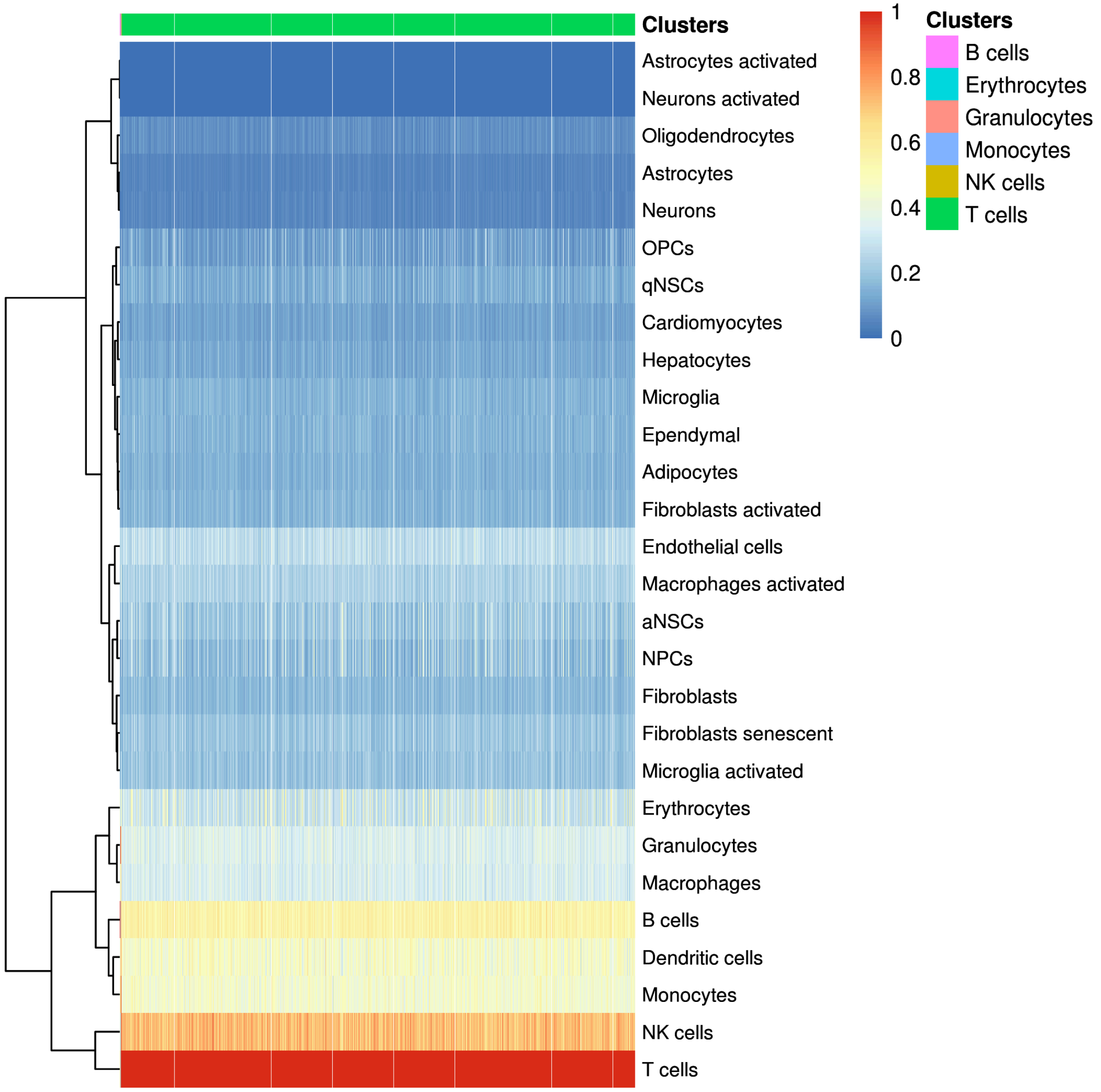
scRNA + scATAC – Compare - SingleR

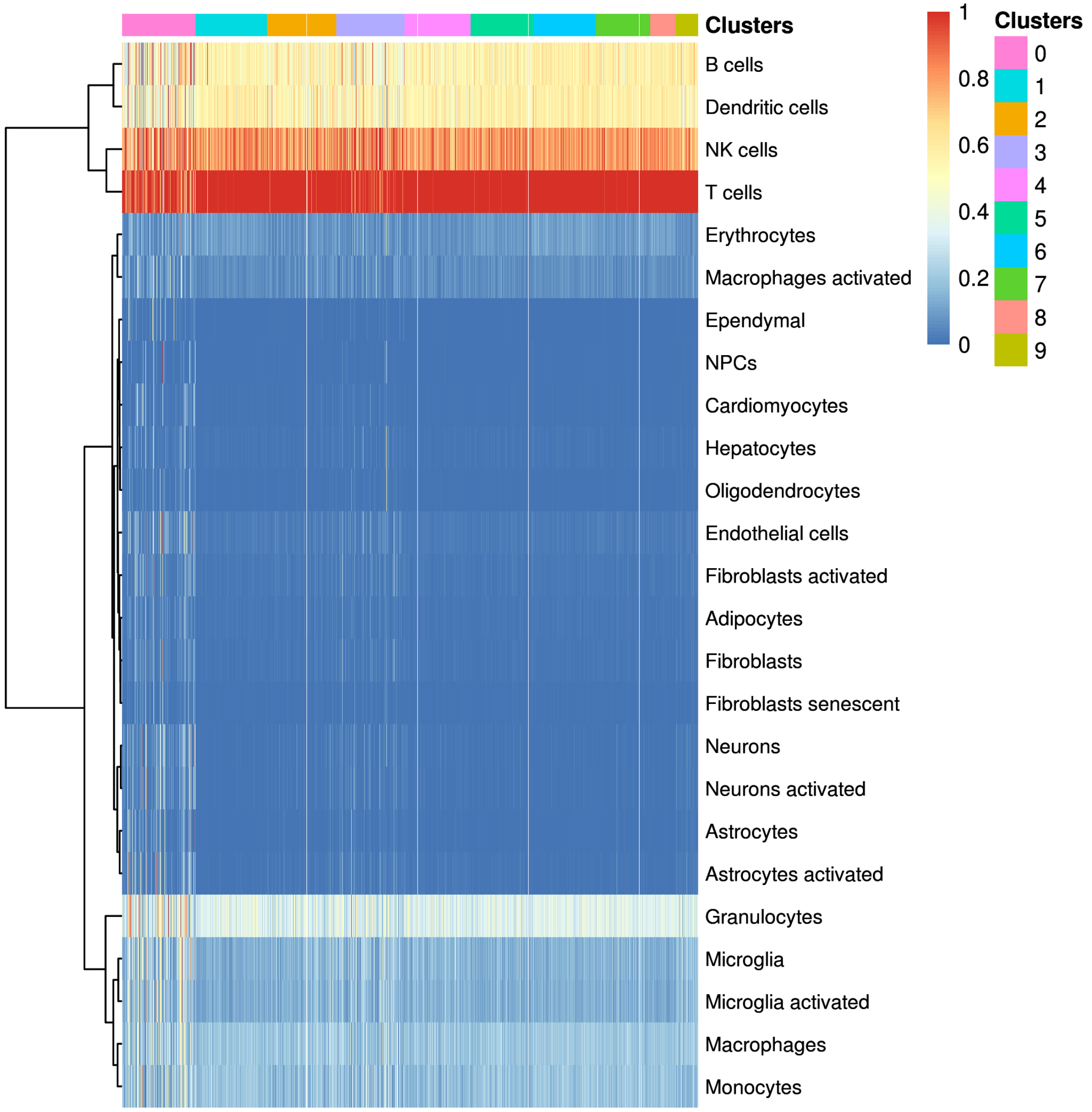
Frederick National Laboratory for Cancer Research





scRNA + scATAC – Compare - SingleR





Question?