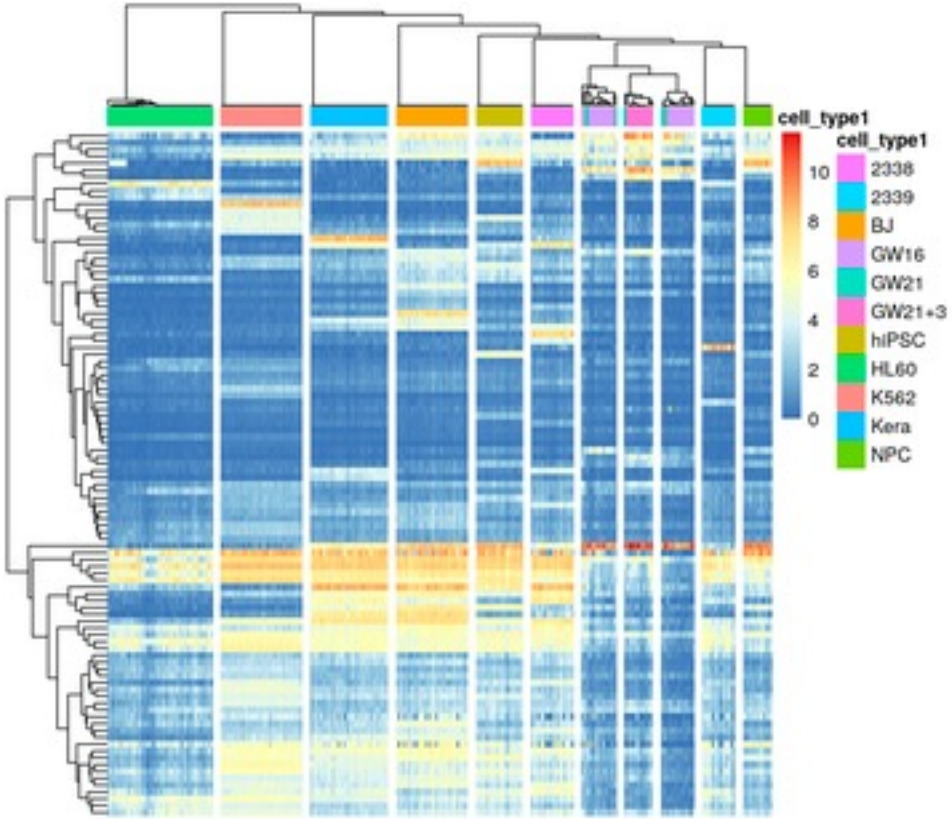
Analysis of single-cell ATAC-seq data

Orr Ashenberg, Jacques Serizay, Fabricio Almeida-Silva

November 7, 2024

Physalia
Courses

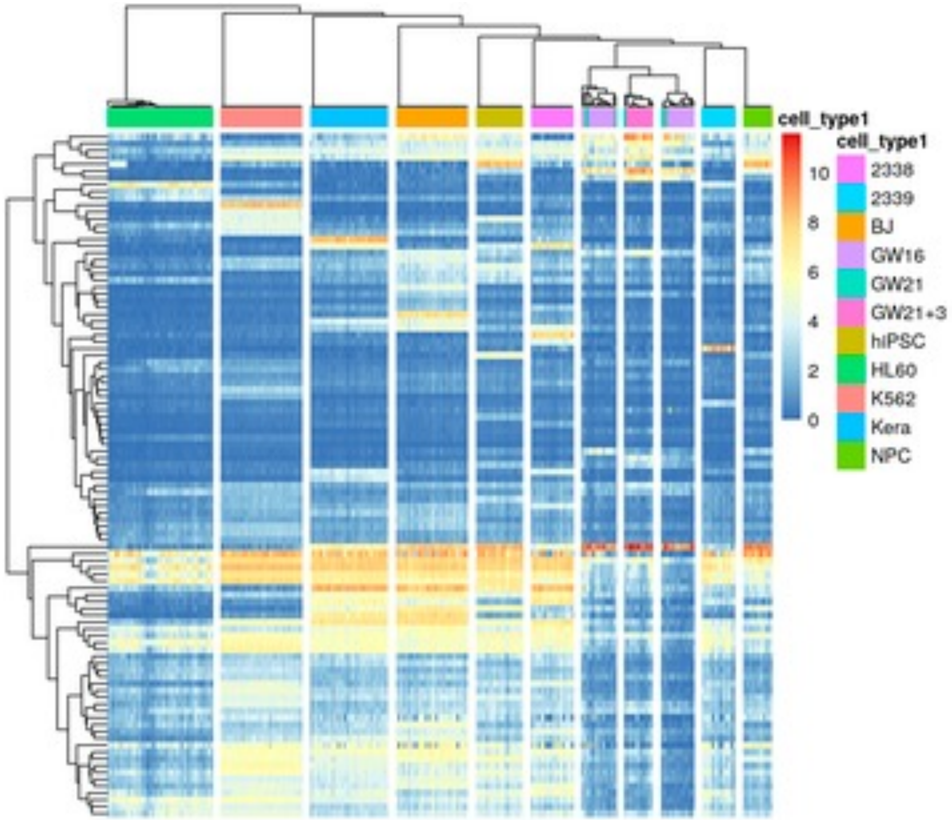


Analysis of single-cell ATAC-seq data

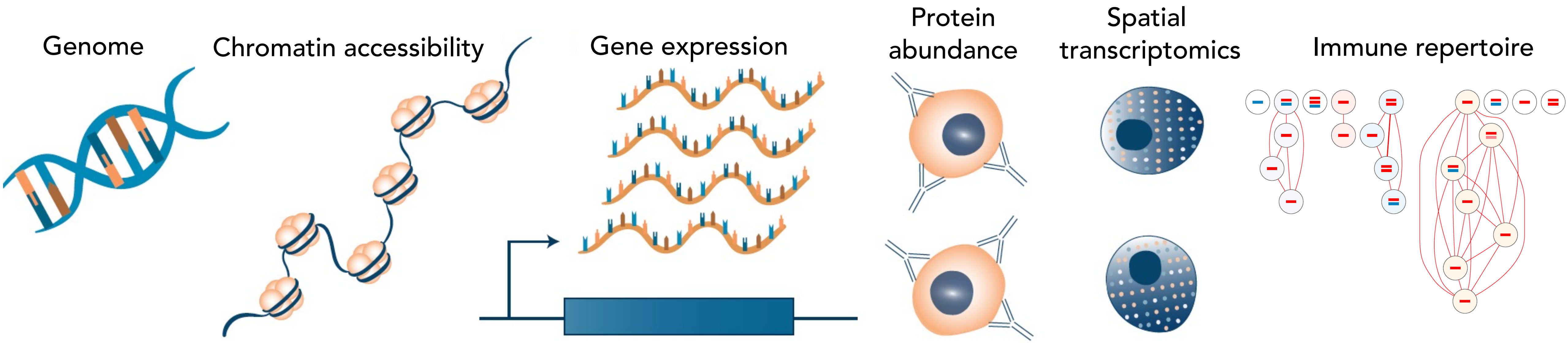
Orr Ashenberg, Jacques Serizay, Fabricio Almeida-Silva

November 7, 2024

Physalia
Courses



Multimodal measurements



Genome

Chromatin accessibility

Gene expression

Protein abundance

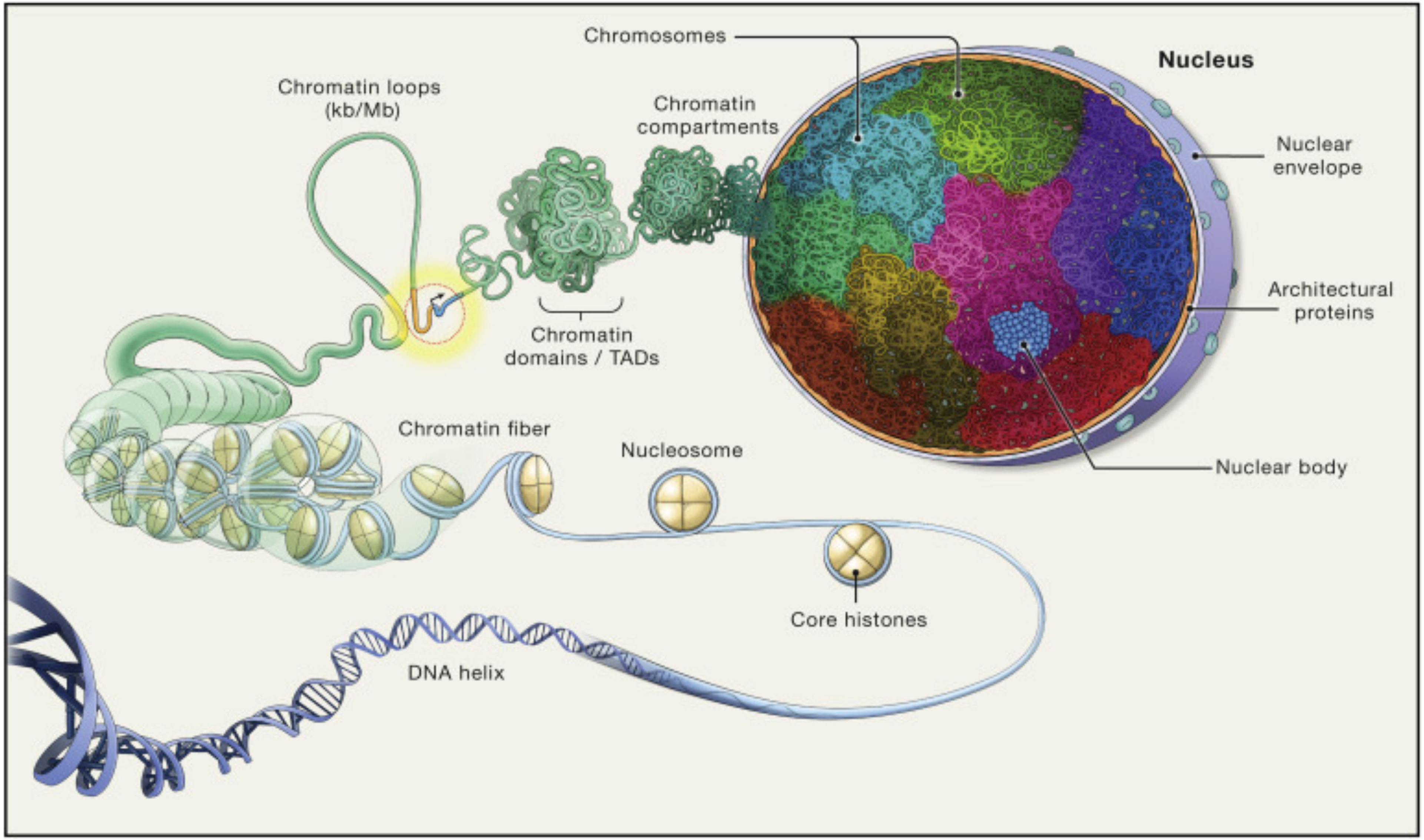
Spatial transcriptomics

Immune repertoire

Combining single-cell transcriptomic measurements with other data modalities can reveal gene function and gene regulation.

Efremova M. et al. (2020) *Nature Methods.* 17:11-20.

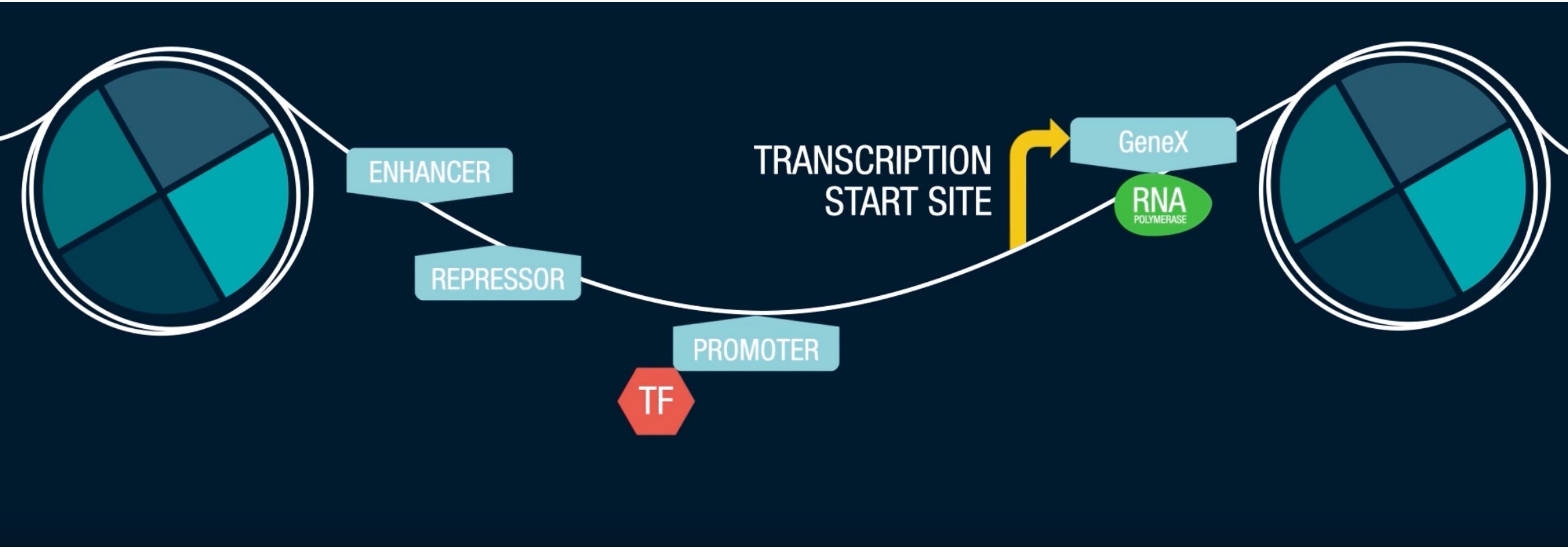
3D organization of DNA



Misteli T. et al. (2020) *Cell.* 183:28-45.

ATAC-Seq detects accessible chromatin regions

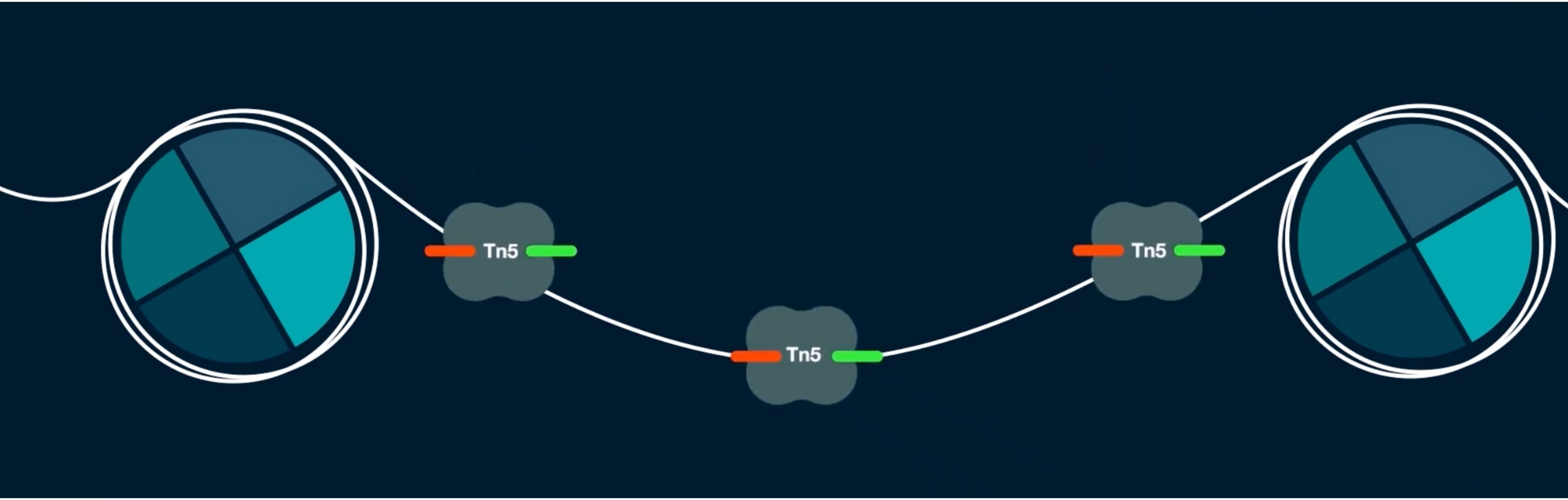
In the cell nucleus, the chromosomes contain tightly packed chromatin material. Part of the chromatin is open and accessible to many regulatory factors who control the expression and suppression of a variety of genes.



“How Single-Cell ATAC-Seq Works”, Bio-Rad Laboratories

ATAC-Seq detects accessible chromatin regions

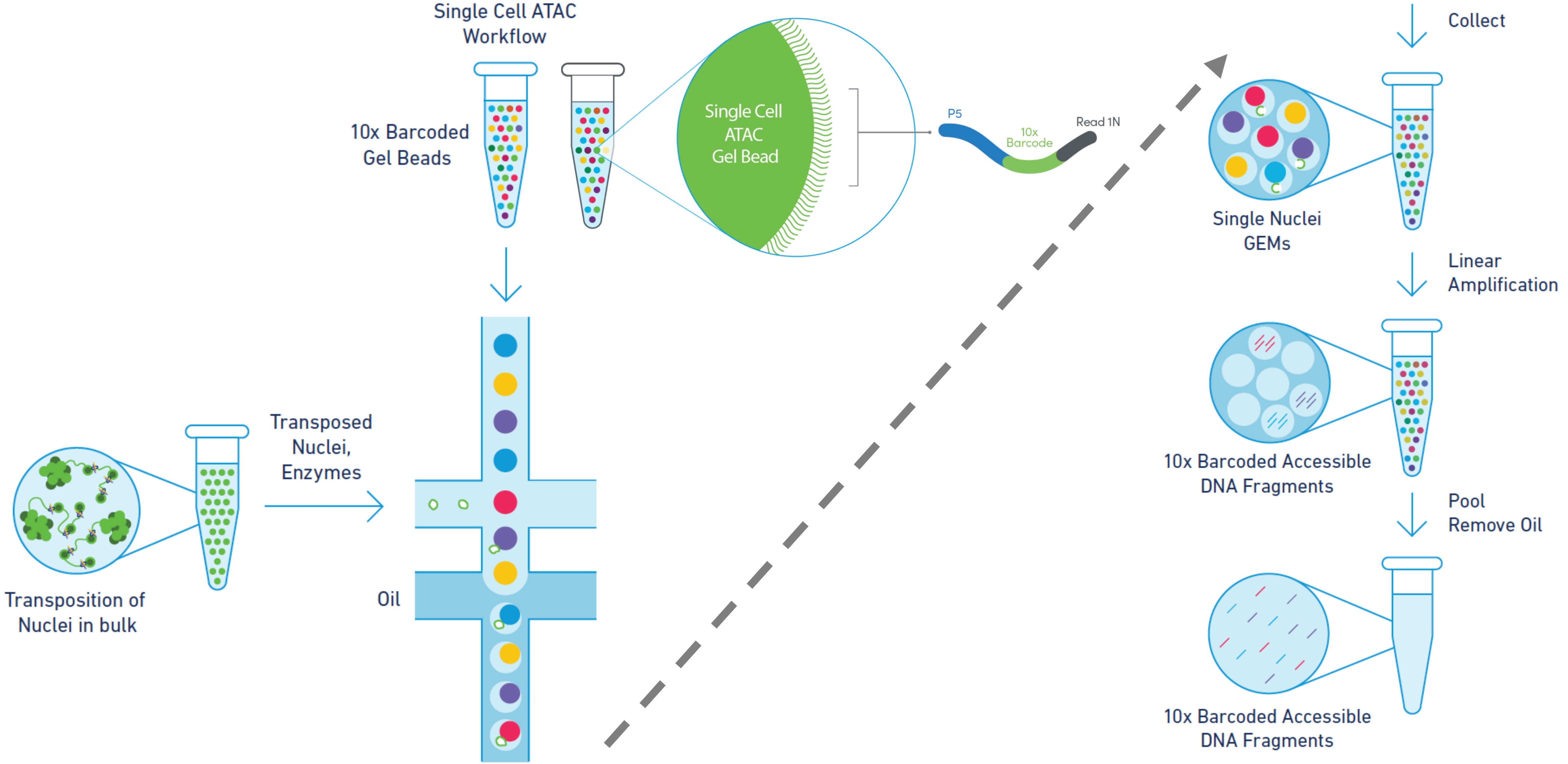
ATACseq (as well as scATACseq) measures how open this piece of DNA is. This openness is a proxy of how easily a transcription factor can bind these parts of the genome. ATACseq measures by using an enzyme called Tn5 transposase which binds open chromatin and inserts DNA sequencing adapters.



The Tn5 transposase ideally cuts DNA just once between the neighboring nucleosomes.

“How Single-Cell ATAC-Seq Works”, Bio-Rad Laboratories

Chromium Single Cell ATAC-Seq (10x)


Single Cell ATAC Workflow
10x Barcoded Gel Beads
Single Cell ATAC Gel Bead
P5
10x Barcode
Read 1N
Collect
Single Nuclei GEMs
Linear Amplification
10x Barcoded Accessible DNA Fragments
Pool Remove Oil
10x Barcoded Accessible DNA Fragments
Transposed Nuclei, Enzymes
Oil
Transposition of Nuclei in bulk

Single cell resolution reveals cell-type specific regulatory elements

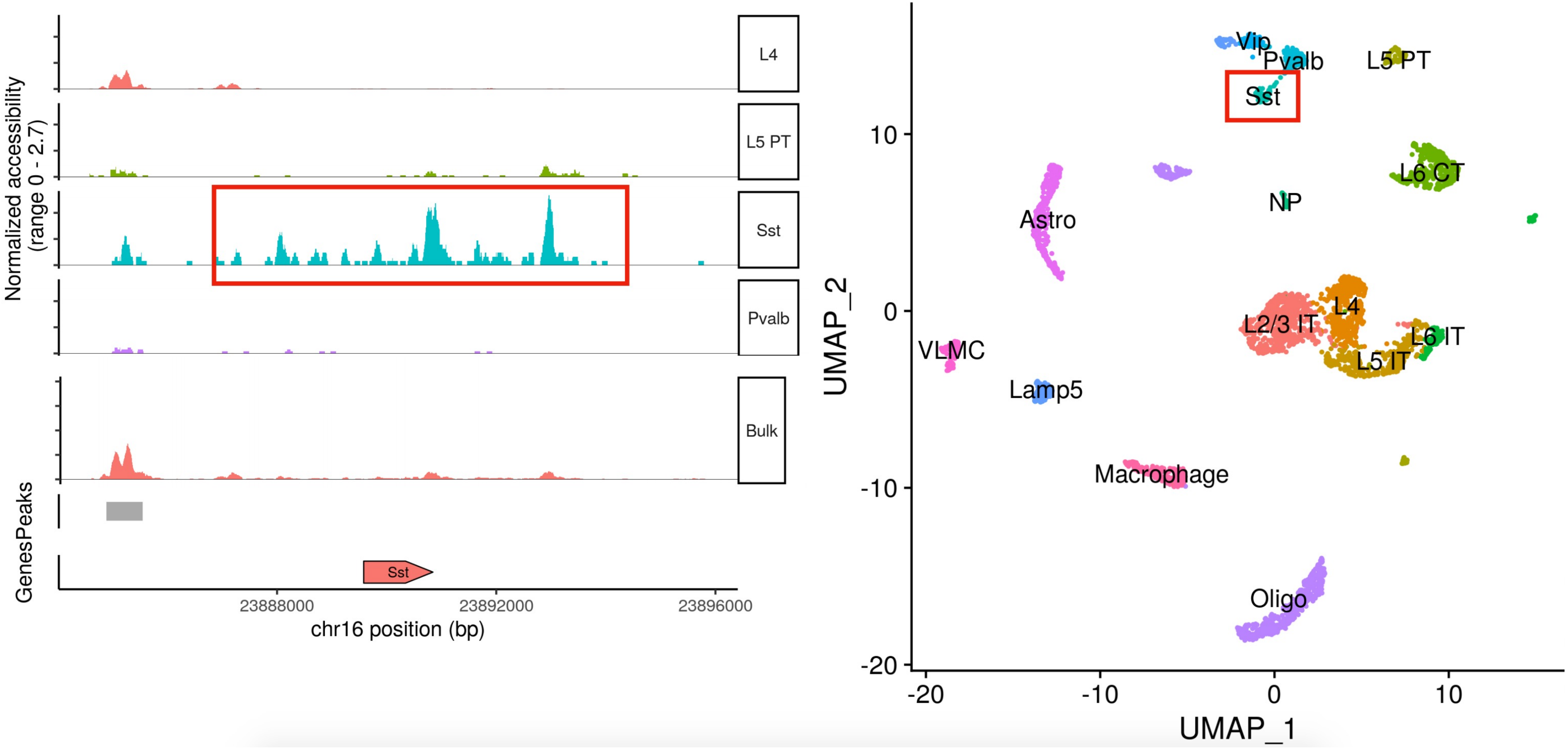
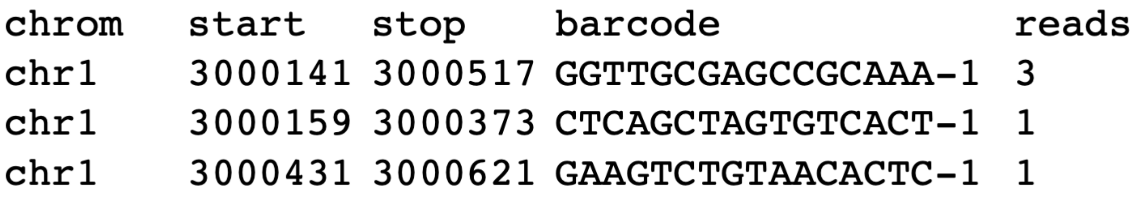


Figure adopted from “Analyzing single-cell ATAC-seq datasets” lecture by Tim Stuart

Pre-processing generates a fragment file and a peak/cell matrix

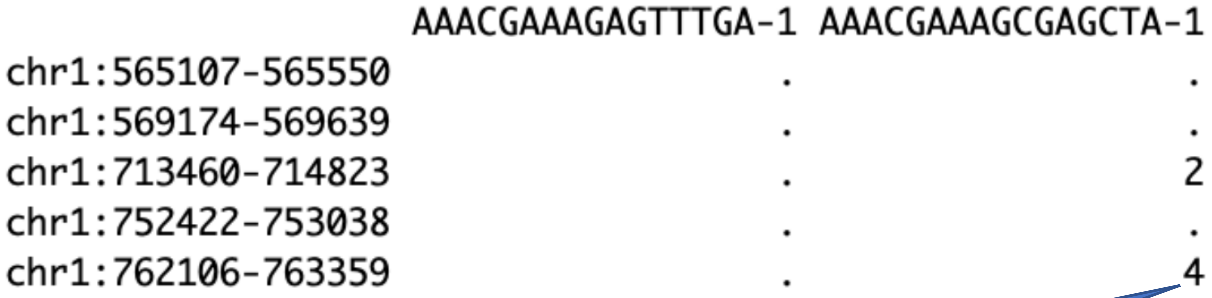
A full list of **all** unique fragments across all single cells, as opposed to only reads that map to peaks.

1. Indexed fragment file



```
chrom    start   stop    barcode              reads
chr1     3000141 3000517 GGTTGCGAGCCGCAAA-1   3
chr1     3000159 3000373 CTCAGCTAGTGTCACT-1   1
chr1     3000431 3000621 GAAGTCTGTAACACTC-1   1
```

2. Large sparse matrix



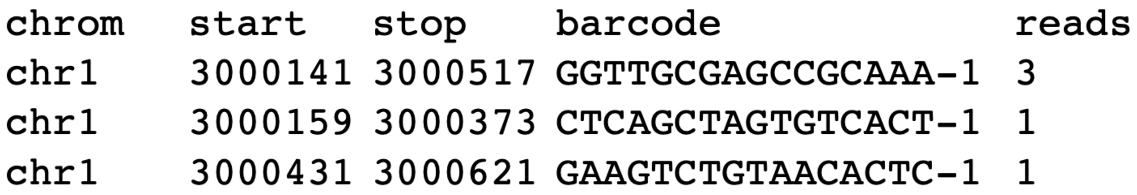
```
                    AAACGAAAGAGTTTGA-1 AAACGAAAGCGAGCTA-1
chr1:565107-565550                   .                  .
chr1:569174-569639                   .                  .
chr1:713460-714823                   .                  2
chr1:752422-753038                   .                  .
chr1:762106-763359                   .                  4
```

Each value in the matrix represents the number of Tn5 cut sites for each single barcode (i.e. cell) that map within each peak

Figure adopted from "Analyzing single-cell ATAC-seq datasets" lecture by Tim Stuart

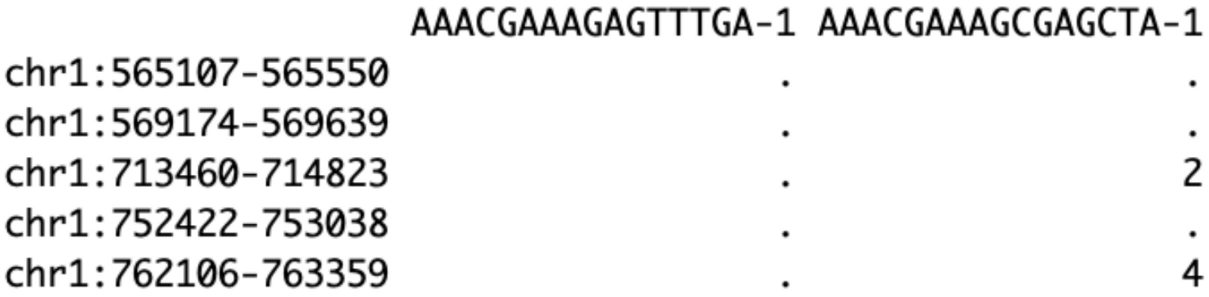
scATAC-Seq data is highly sparse

1. Indexed fragment file



```
chrom   start   stop    barcode             reads
chr1    3000141 3000517 GGTTGCGAGCCGCAAA-1  3
chr1    3000159 3000373 CTCAGCTAGTGTCACT-1  1
chr1    3000431 3000621 GAAGTCTGTAACACTC-1  1
```

2. Large sparse matrix



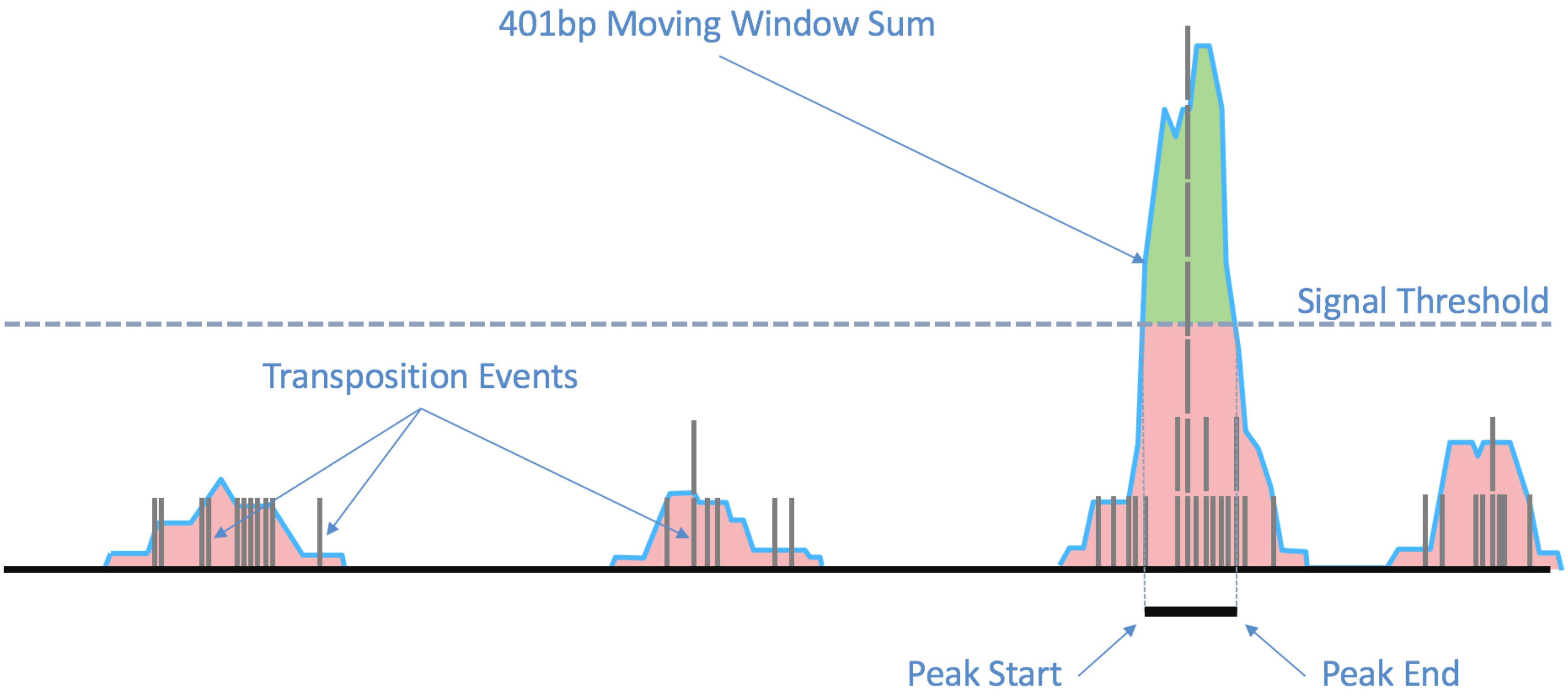
```
                     AAACGAAAGAGTTTGA-1 AAACGAAAGCGAGCTA-1
chr1:565107-565550                    .                  .
chr1:569174-569639                    .                  .
chr1:713460-714823                    .                  2
chr1:752422-753038                    .                  .
chr1:762106-763359                    .                  4
```

Challenges in comparison to scRNA:

1. More sparse
2. Near-binary data
3. Non-fixed feature set
4. Order of magnitude more features

Figure adopted from "Analyzing single-cell ATAC-seq datasets" lecture by Tim Stuart

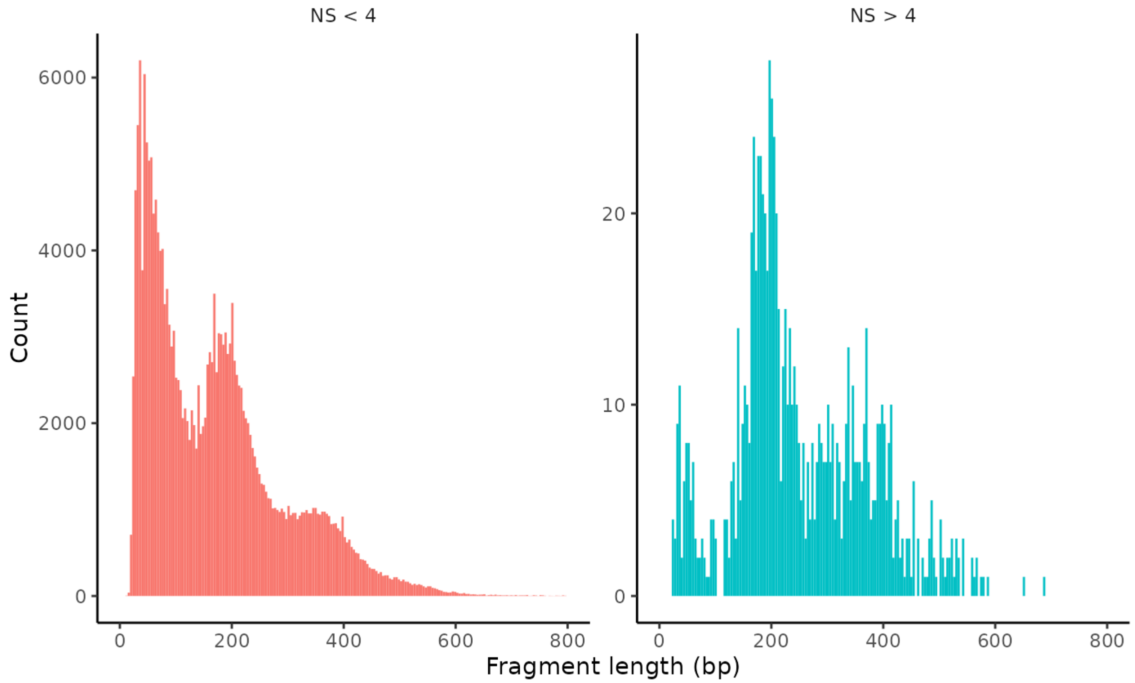
Peak calling: from chromatin accessible fragments to peaks

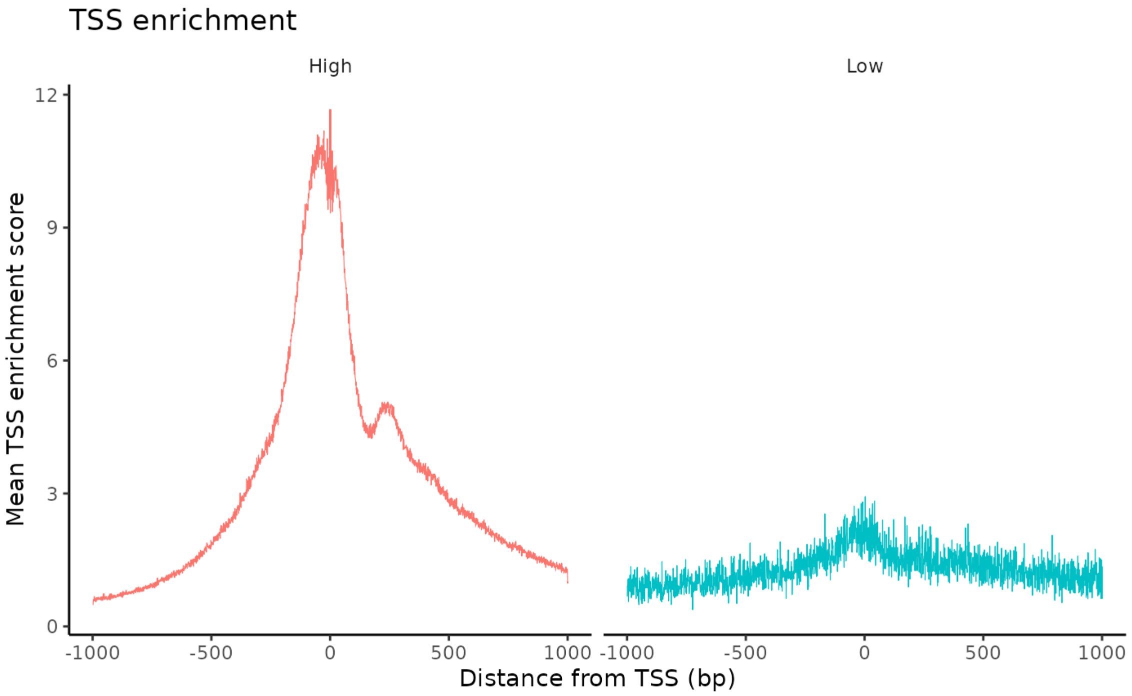


https://support.10xgenomics.com/single-cell-atac/software/pipelines/latest/algorithms/overview#peaks

Quality control metrics for scATAC-seq data

1. Nucleosome banding pattern
2. Transcriptional start site (TSS) enrichment
3. Total number of fragments in peaks
4. Fraction of fragments in peaks





Overview of scATAC-Seq analysis

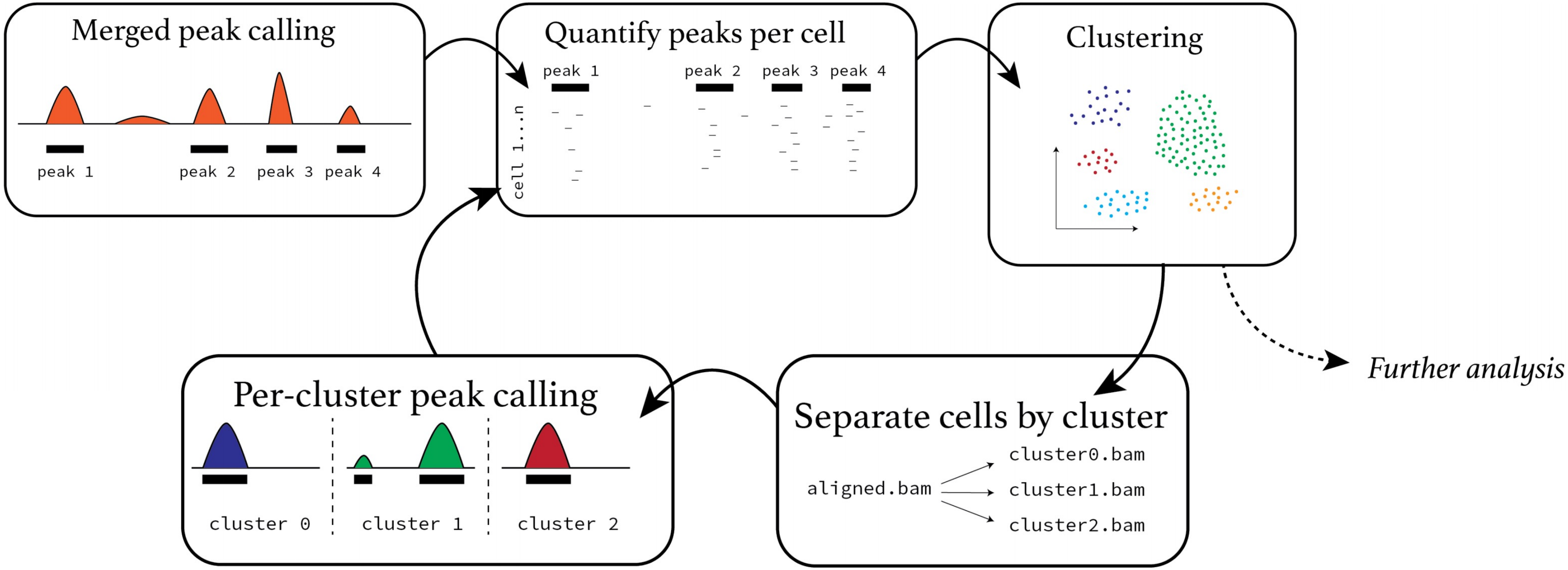


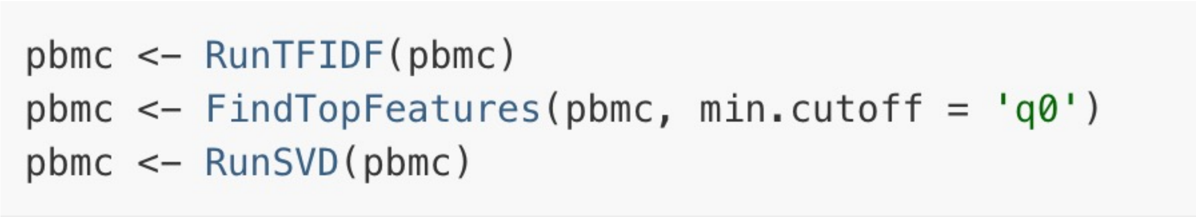
Figure adopted from "Analyzing single-cell ATAC-seq datasets" lecture by Tim Stuart

scATAC-Seq often uses latent semantic indexing (LSI) for dimensionality reduction

- Originally developed for topic modeling / natural language processing (Deerwester et al. 1990) and first applied to scATAC-Seq in 2015 (Cusanovich et al. *Science*)

- term frequency-inverse document frequency (TF-IDF) normalization followed by singular value decomposition (SVD)

- Cell = document and peak = term



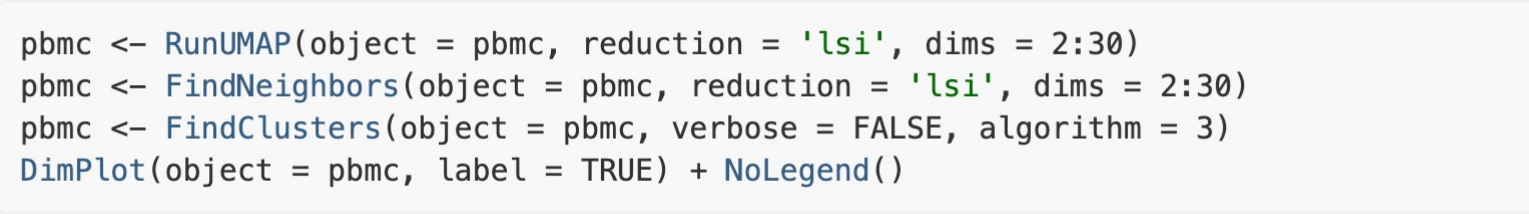
```
pbmc <- RunTFIDF(pbmc)
pbmc <- FindTopFeatures(pbmc, min.cutoff = 'q0')
pbmc <- RunSVD(pbmc)
```

- Term frequency: normalize across cells to correct for differences in sequencing depth

- Inverse document frequency: give higher values to more rare peaks

Figure adopted from "Analyzing single-cell ATAC-seq datasets" lecture by Tim Stuart

Visualize clusters in scATAC-Seq



```
pbmc <- RunUMAP(object = pbmc, reduction = 'lsi', dims = 2:30)
pbmc <- FindNeighbors(object = pbmc, reduction = 'lsi', dims = 2:30)
pbmc <- FindClusters(object = pbmc, verbose = FALSE, algorithm = 3)
DimPlot(object = pbmc, label = TRUE) + NoLegend()
```

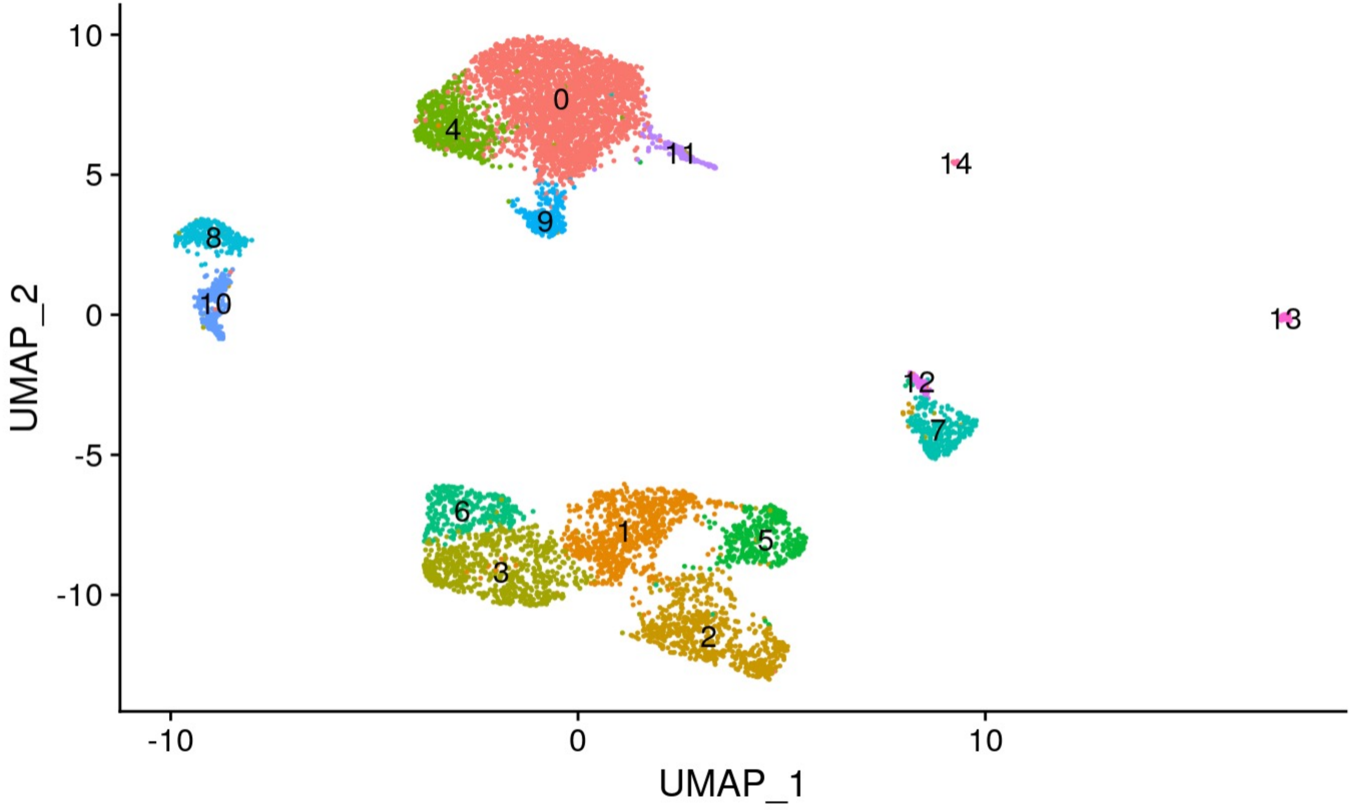


Figure adopted from signac tutorial, https://satijalab.org/signac/articles/pbmc_vignette.html

Fragment file helps infer gene "activity" and annotate clusters

Quantify the activity of each gene in the genome by assessing the chromatin accessibility associated with each gene: count the number of fragments for each cell that map to the promoter + gene body

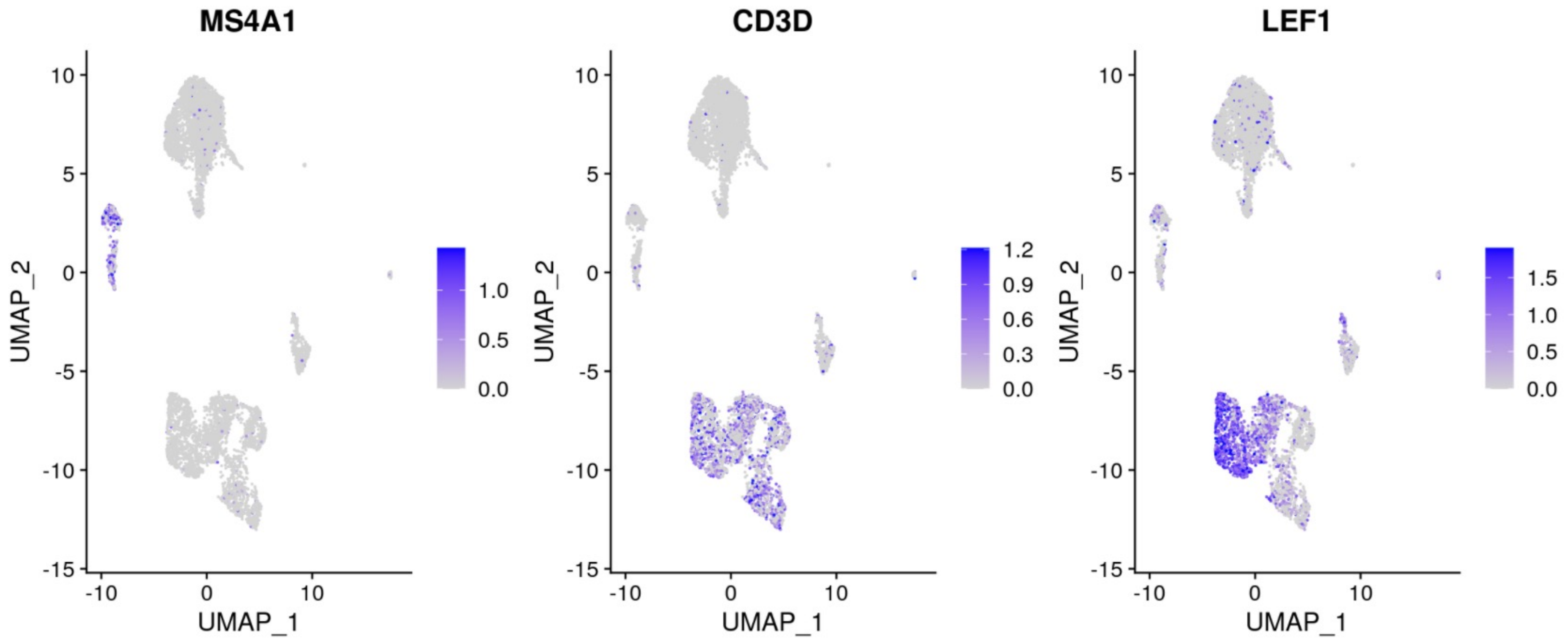
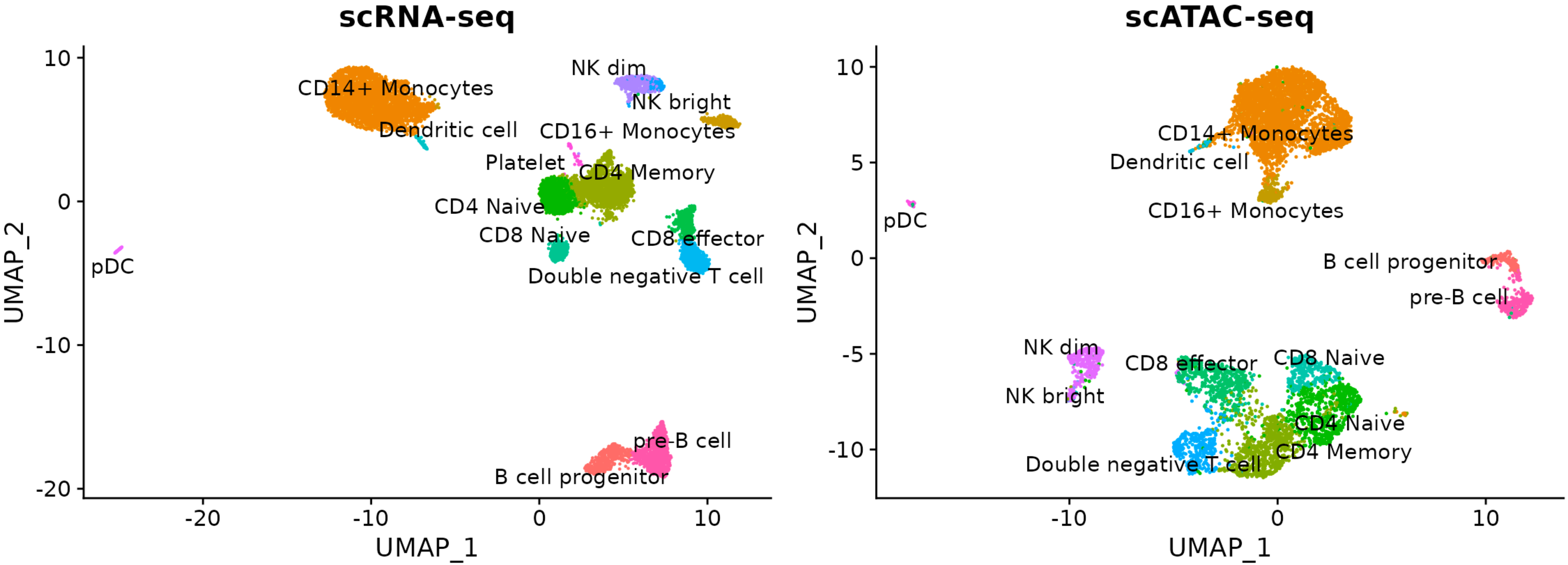


Figure adopted from signac tutorial, https://satijalab.org/signac/articles/pbmc_vignette.html

Integrating with scRNA-seq data using CCA + MNN (Seurat v4)



Critical assumption: there is generally a positive correlation between chromatin accessibility and gene expression!!!!

Figure adopted from signac tutorial, https://satijalab.org/signac/articles/pbmc_vignette.html

Finding overrepresented motifs

To identify potentially important cell-type-specific regulatory sequences, signac searches for DNA motifs that are overrepresented in a set of peaks that are differentially accessible between cell types.

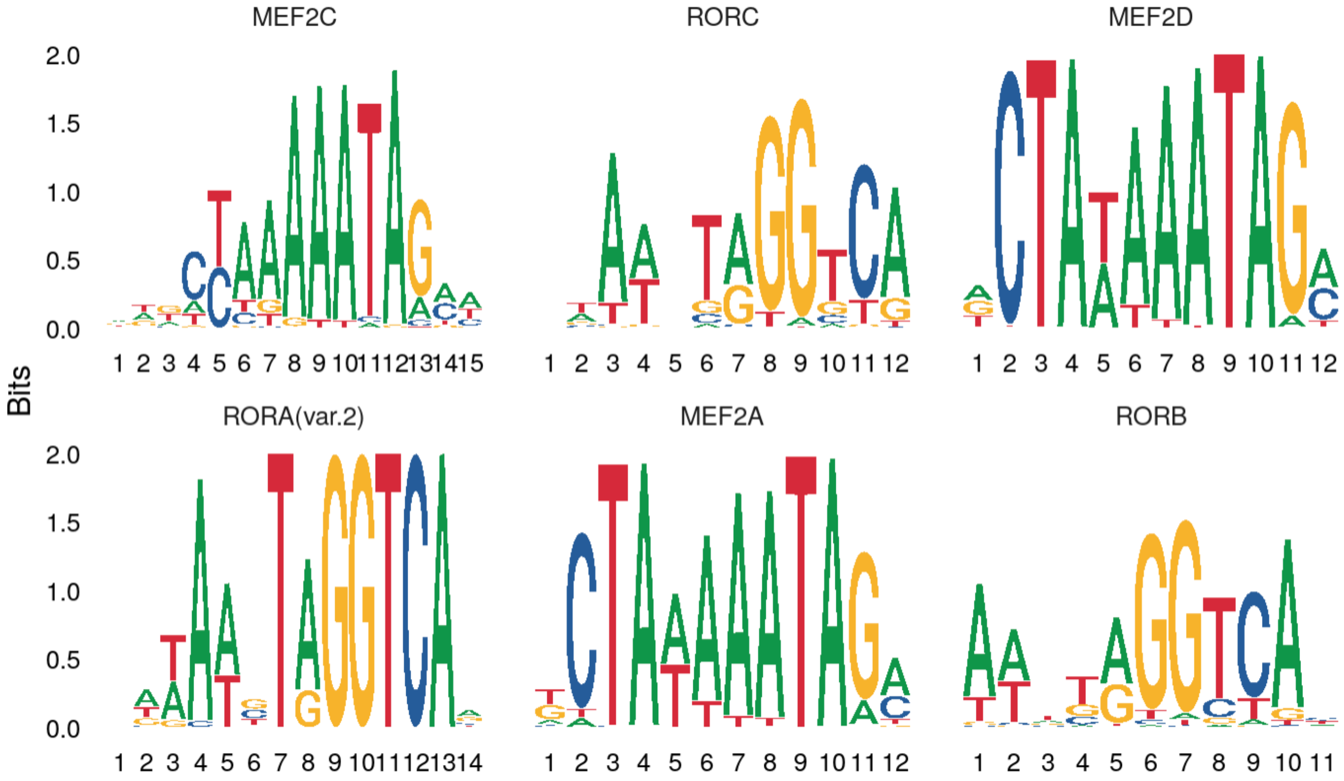


Figure adopted from signac tutorial, https://satijalab.org/signac/articles/pbmc_vignette.html

Computing motif activities

ChromVAR identifies motifs associated with variability in chromatin accessibility between cells.

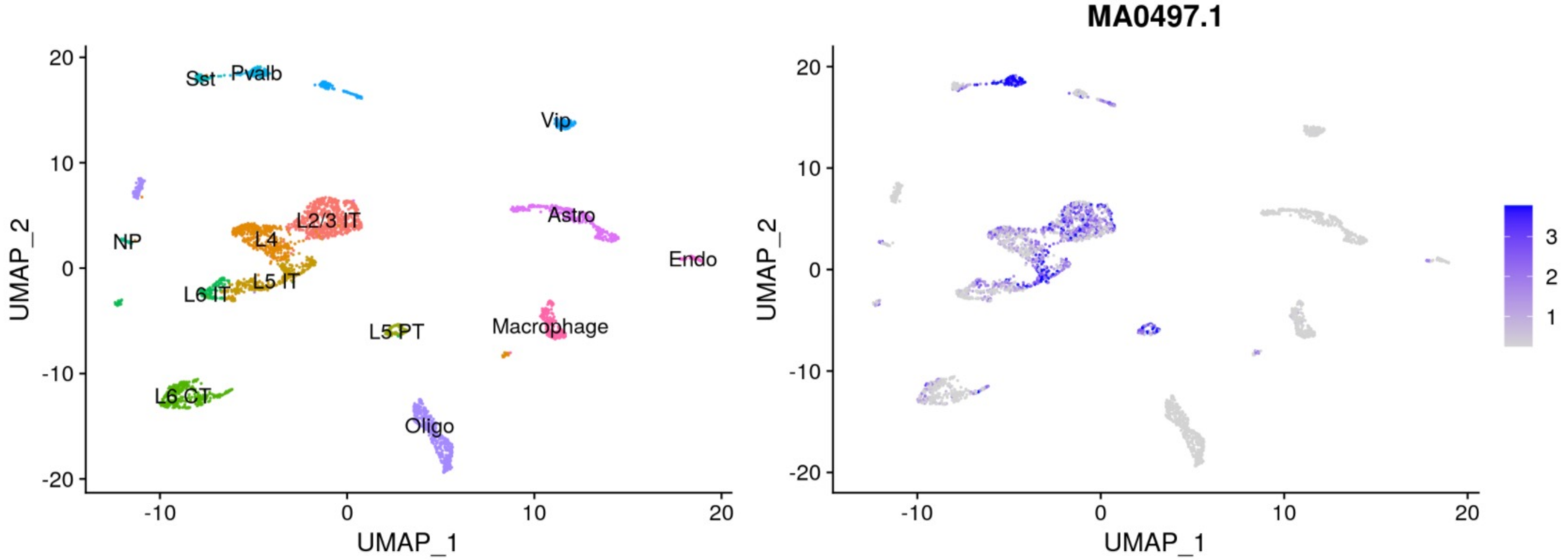
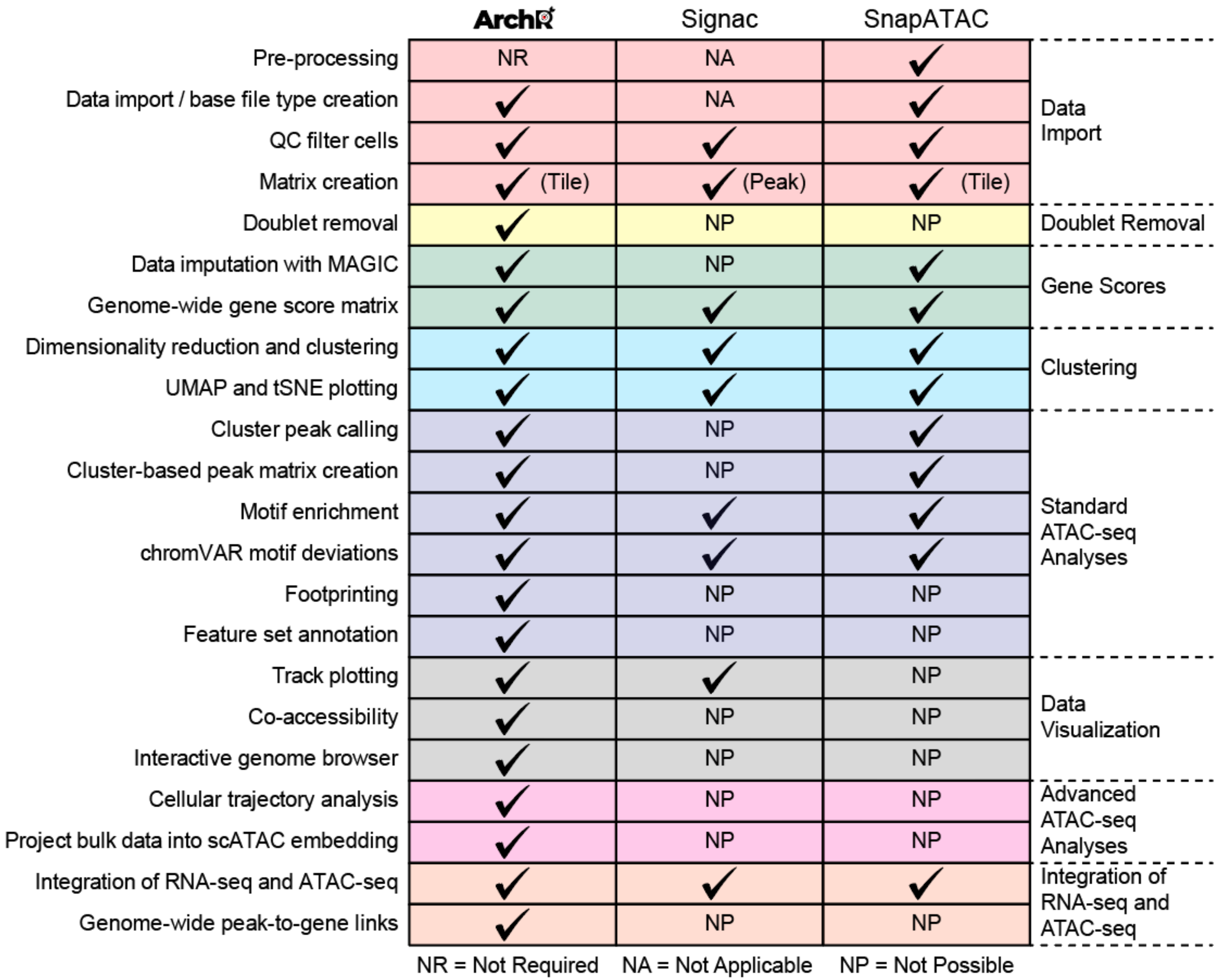


Figure adopted from signac tutorial, https://satijalab.org/signac/articles/pbmc_vignette.html

Software for scATAC-Seq analysis

Signac	*https://satijalab.org/signac/index.html*
SnapATAC	*https://github.com/r3fang/SnapATAC*
ArchR	*https://www.archrproject.com/*
cisTopic	*https://github.com/aertslab/cisTopic*
chromVAR	*https://github.com/GreenleafLab/chromVAR*
CICERO	*https://cole-trapnell-lab.github.io/cicero-release/*
episcanpy	*https://episcanpy.readthedocs.io/en/latest/*

Software for scATAC-Seq analysis



	ArchR	Signac	SnapATAC	
Pre-processing	NR	NA	✓	Data Import
Data import / base file type creation	✓	NA	✓	
QC filter cells	✓	✓	✓	
Matrix creation	✓ (Tile)	✓ (Peak)	✓ (Tile)	
Doublet removal	✓	NP	NP	Doublet Removal
Data imputation with MAGIC	✓	NP	✓	Gene Scores
Genome-wide gene score matrix	✓	✓	✓	
Dimensionality reduction and clustering	✓	✓	✓	Clustering
UMAP and tSNE plotting	✓	✓	✓	
Cluster peak calling	✓	NP	✓	Standard ATAC-seq Analyses
Cluster-based peak matrix creation	✓	NP	✓	
Motif enrichment	✓	✓	✓	
chromVAR motif deviations	✓	✓	✓	
Footprinting	✓	NP	NP	
Feature set annotation	✓	NP	NP	
Track plotting	✓	✓	NP	Data Visualization
Co-accessibility	✓	NP	NP	
Interactive genome browser	✓	NP	NP	
Cellular trajectory analysis	✓	NP	NP	Advanced ATAC-seq Analyses
Project bulk data into scATAC embedding	✓	NP	NP	
Integration of RNA-seq and ATAC-seq	✓	✓	✓	Integration of RNA-seq and ATAC-seq
Genome-wide peak-to-gene links	✓	NP	NP	

NR = Not Required NA = Not Applicable NP = Not Possible

Some approaches to multiome data (scRNA-seq and scATAC-seq)

ArchR: https://greenleaflab.github.io/ArchR_2020/Ex-Analyze-Multiome.html

Signac (same people who built Seurat): https://satijalab.org/signac/articles/pbmc_multiomic.html

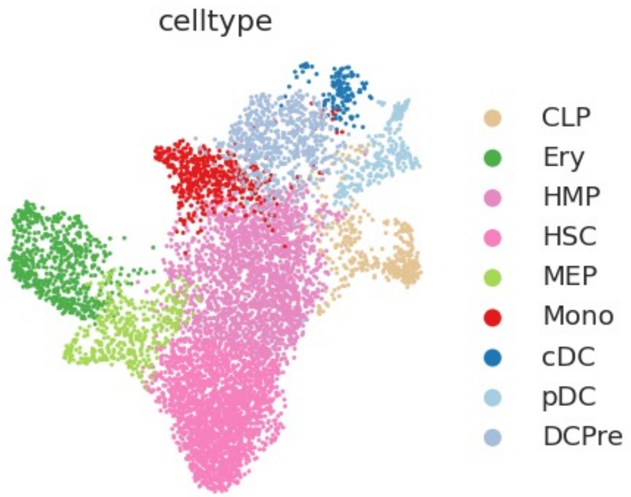
FigR: https://buenrostrolab.github.io/FigR

Methods to overcome sparsity in ATAC-seq data

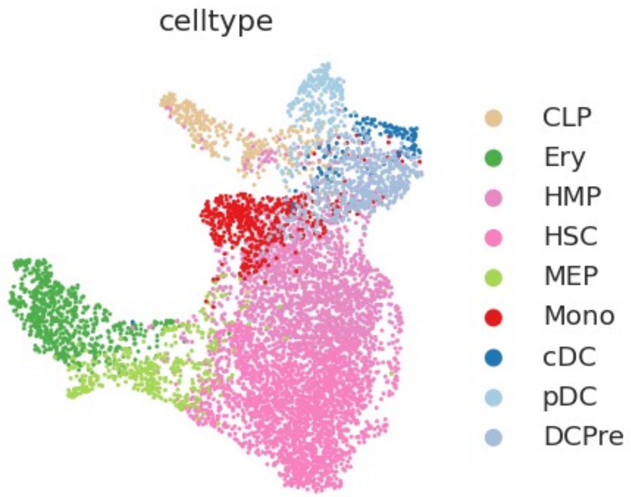
Computing meta-cells is on methodology used to overcome sparsity in scATAC-seq data

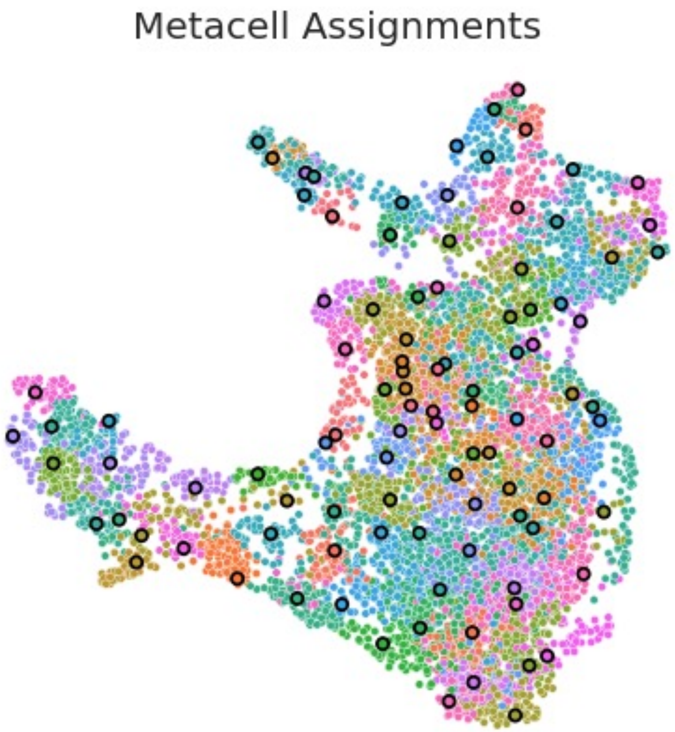
Computing meta-cells (e.g. SEACells algorithm) can improve computation of peak-gene associations

scRNA-seq



scATAC-seq





Persad S., et al. Nature Biotechnology (2023).

Methods to overcome sparsity in ATAC-seq data

Generate summarized ATAC and RNA metacells

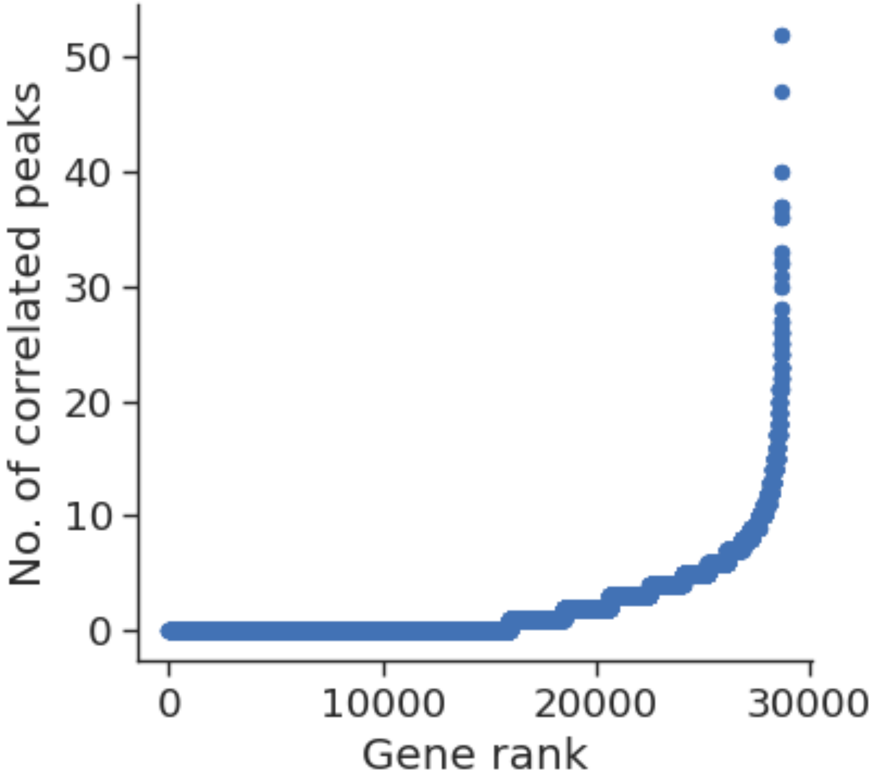
Used the paired multiome metacells to compute the correlation of gene expression and accessibility of peaks within the vicinity of the gene

Pick out highly regulated genes that are correlated with multiple peaks

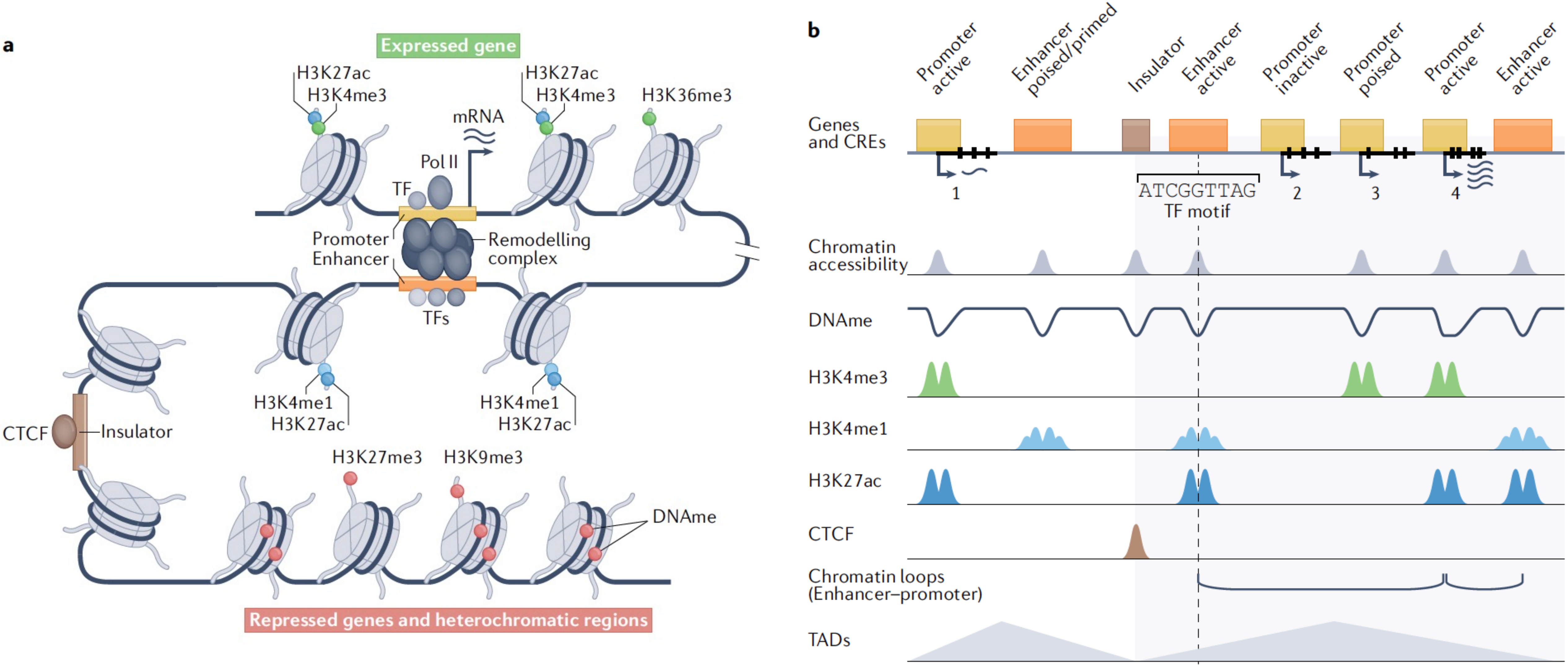
Calculate gene scores as the weighted sum of the accessibility of correlated peaks

To calculate gene accessibility metrics identify the subset of peaks that are open in each metacell

Open peaks are used to compute a gene accessibility metric which represents the fraction of correlated open peaks

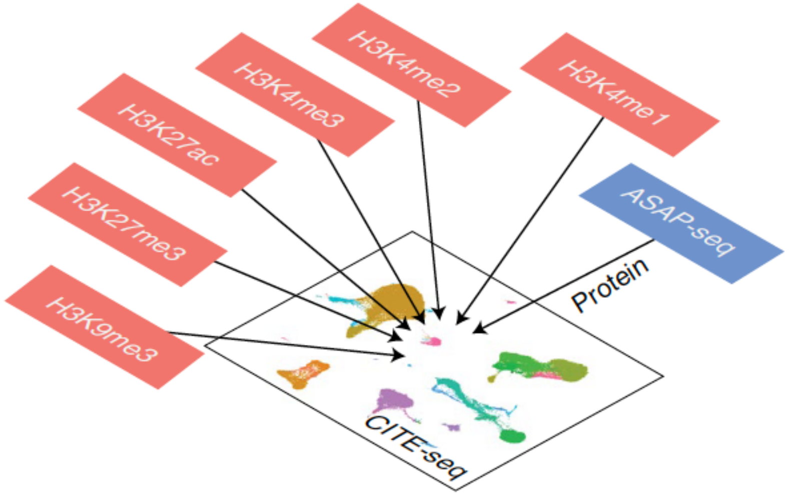


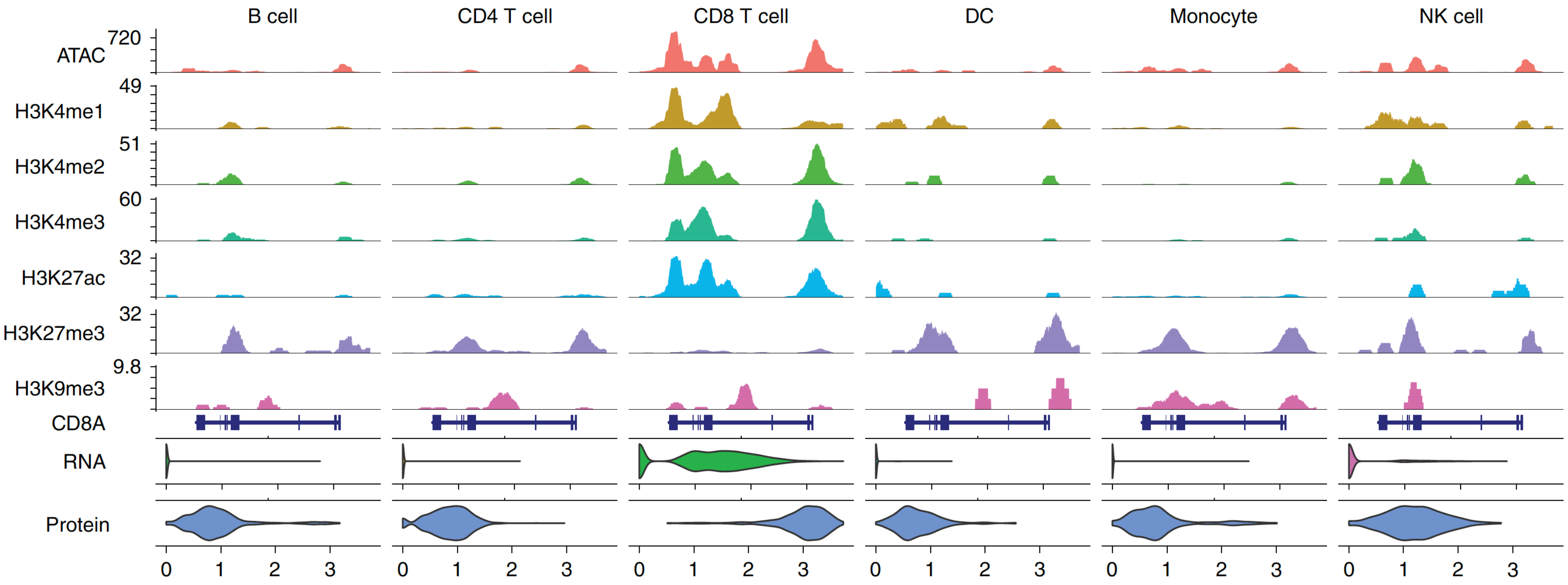
Studying gene regulation using single-cell epigenomics



Preissl S. et al. (2022) *Nature Reviews Genetics.*

Studying gene regulation using single-cell epigenomics





Zhang B. et al. (2022) *Nature Biotechnology.* 40:1220-1230.