Single-Cell ATAC-seq

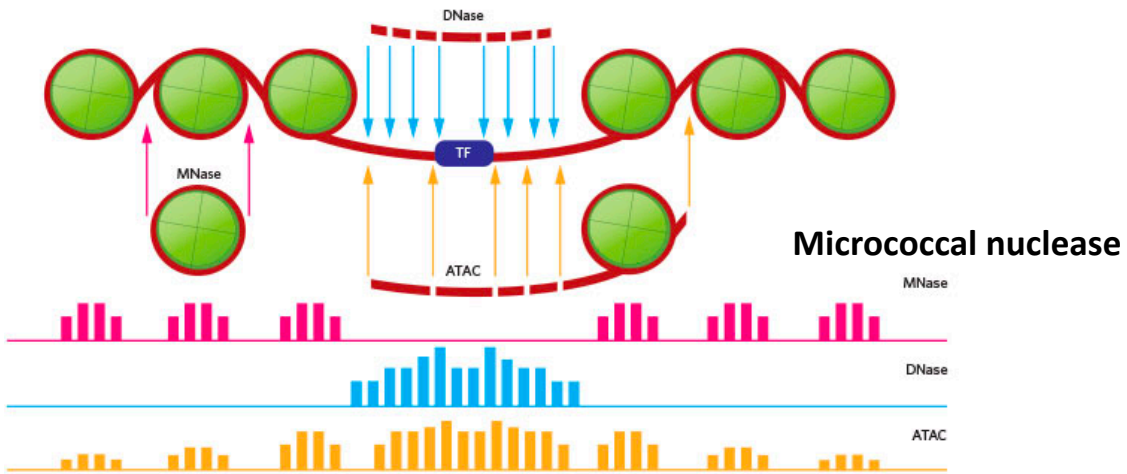
Ming (Tommy) Tang

Twitter: @tangming2005

https://divingintogeneticsandgenomics.rbind.io/

STAT115/215, BIO/BST282

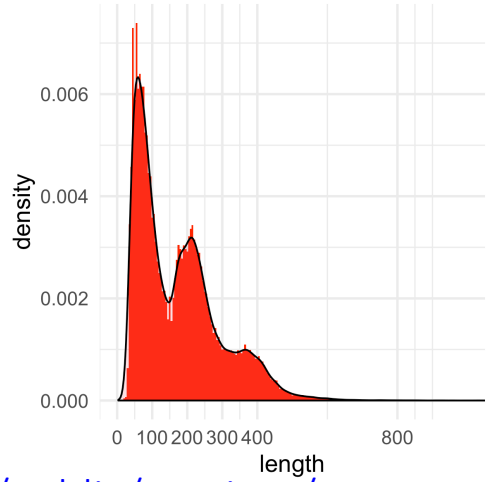
Why ATAC-seq for assay chromatin accessibility ?

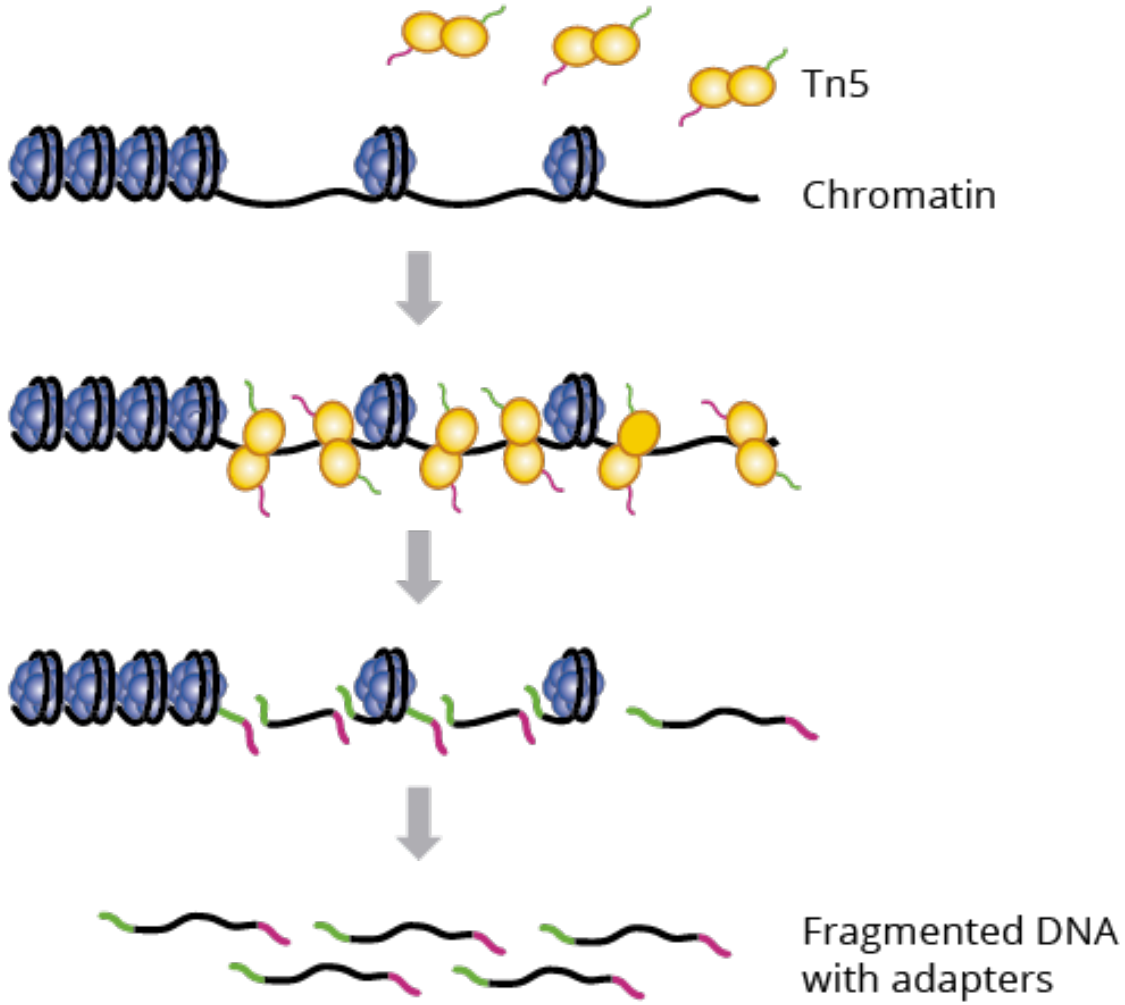


Micrococcal nuclease

Tn5 transposase

Fragment length Distribution

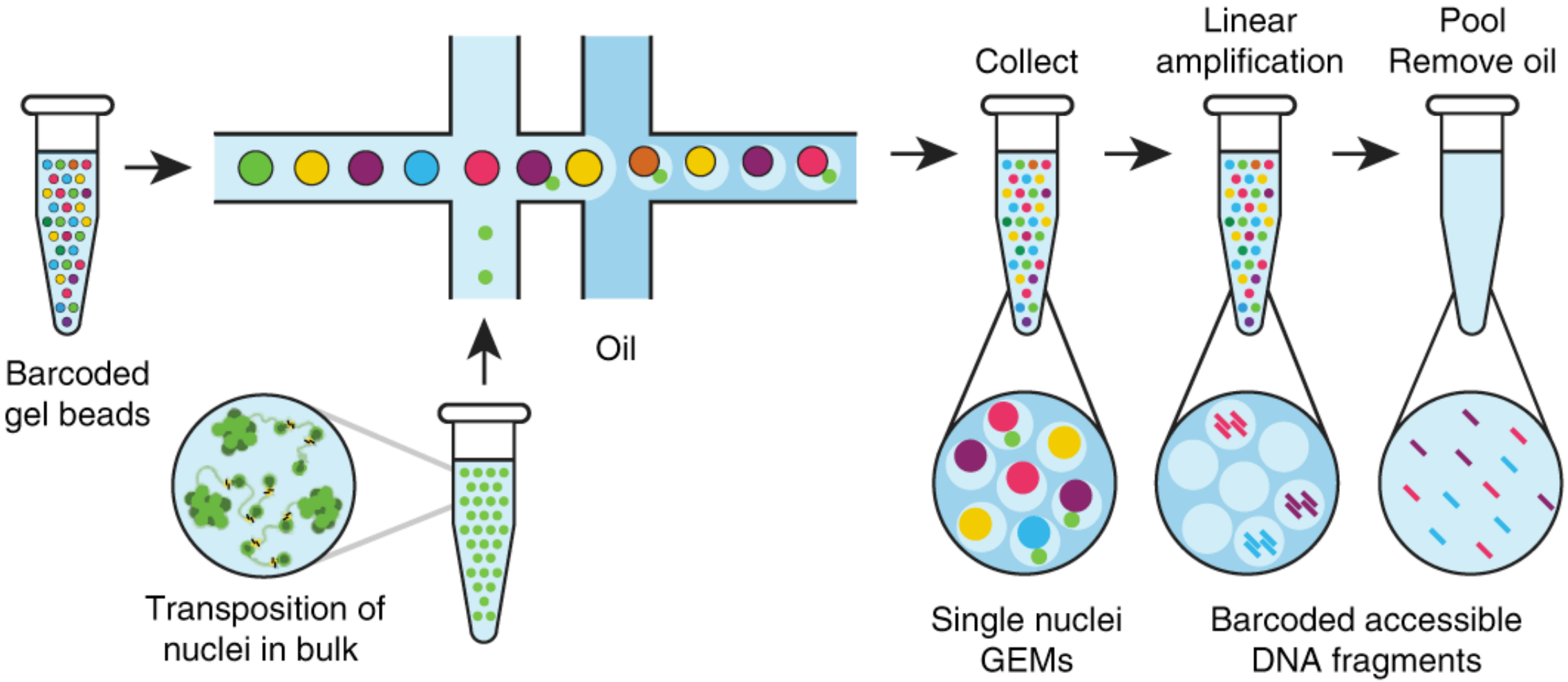




Buenrostro et al, Curr. Protoc. Mol. Biol. 2015

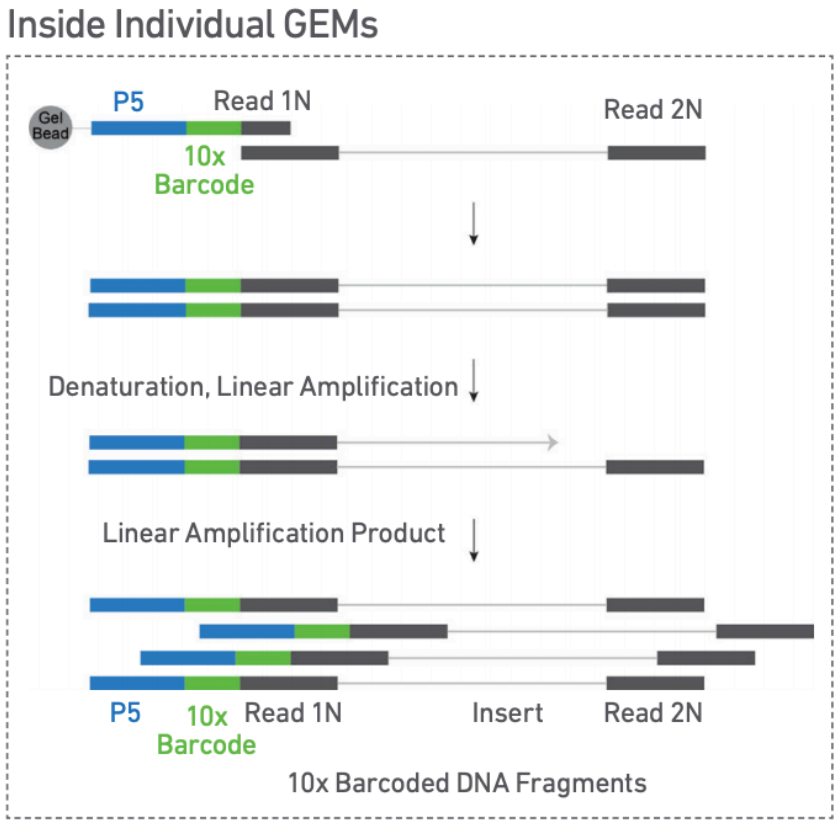
https://www.genewiz.com/Public/Services/Next-Generation-Sequencing/Epigenomics/ATAC-Seq/

scATAC-seq Experimental Procedure

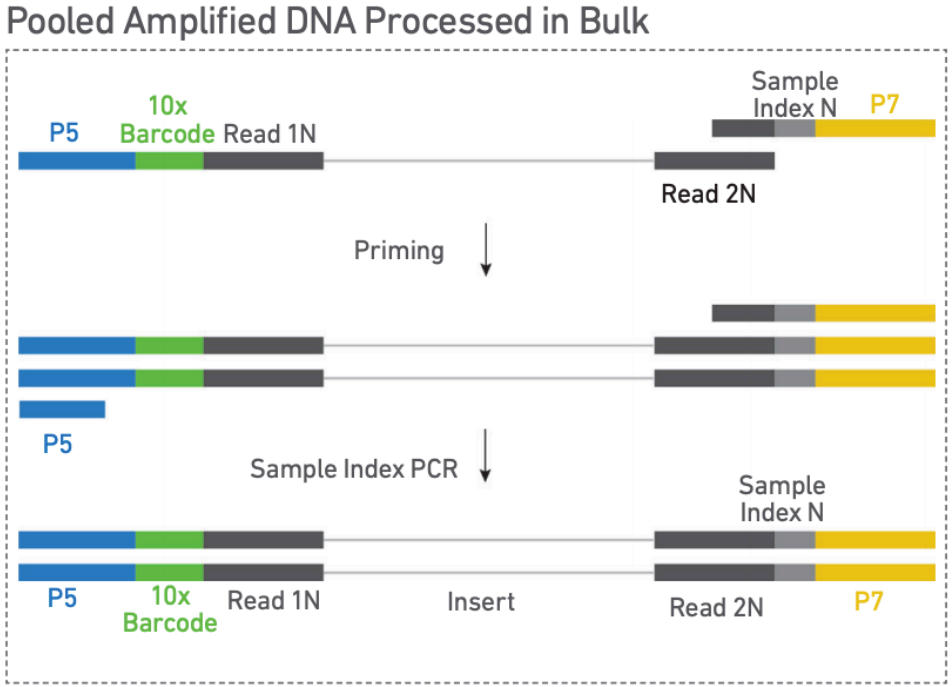


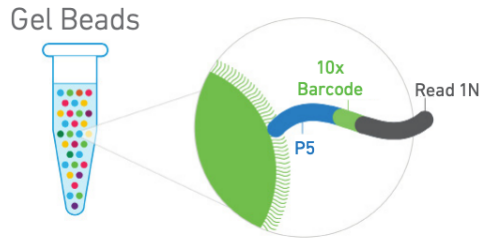
Satpathy et al, Nat Biotech 2019

scATACseq library construction



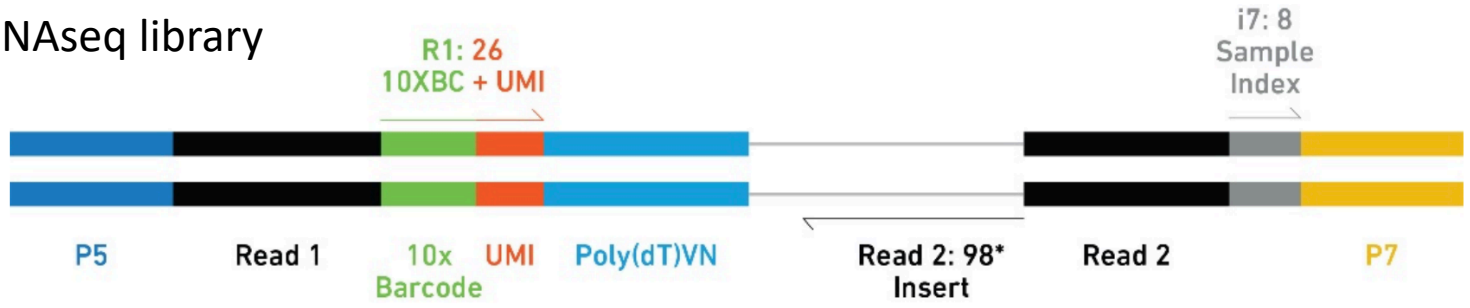
P7 and a sample index are added during library construction via PCR. The final libraries contain the P5 and P7 sequences used in Illumina® bridge amplification.





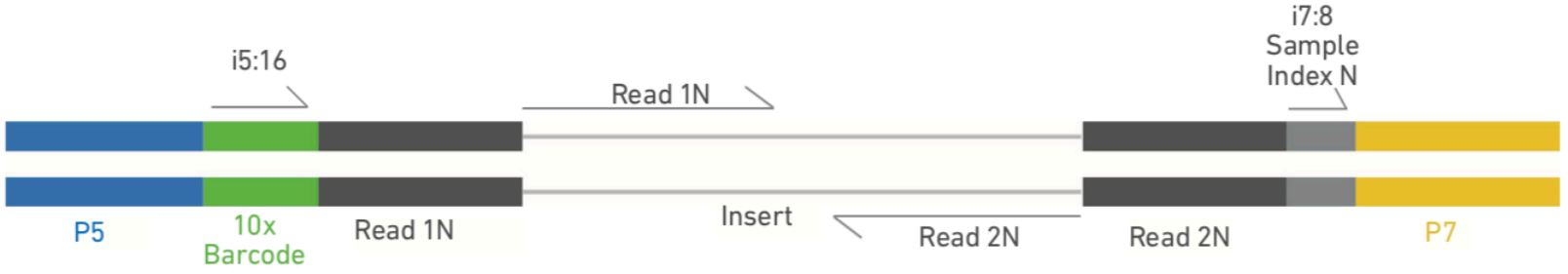
10x Single-cell ATAC-seq library

10x scRNAseq library



16 bp cell barcode
10bp (V2) vs 12bp (V3)
UMI

Chromium Single Cell ATAC Library



Question:
Why no UMI?

https://divingintogeneticsandgenomics.rbind.io/post/understand-10x-scrnaseq-and-scatac-fastqs/

Different scATAC-seq Techniques

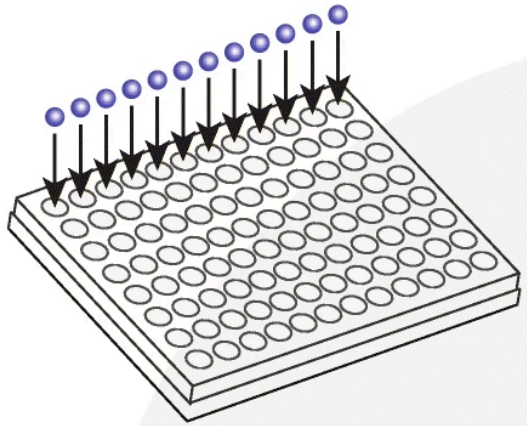
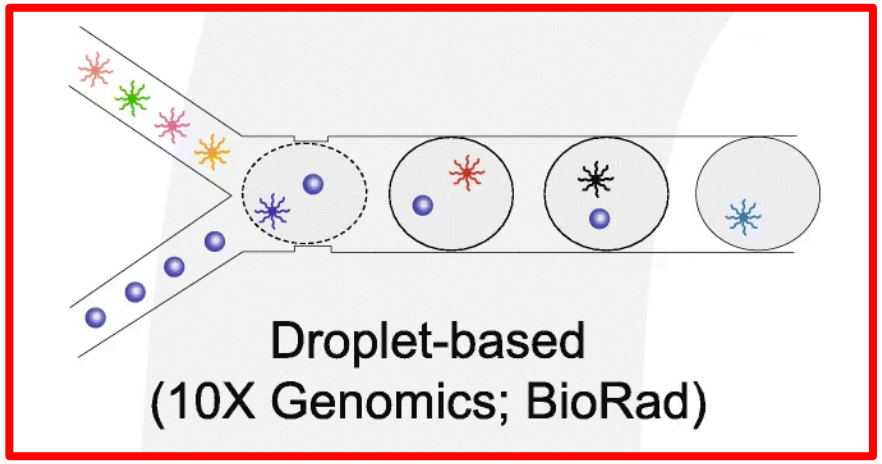
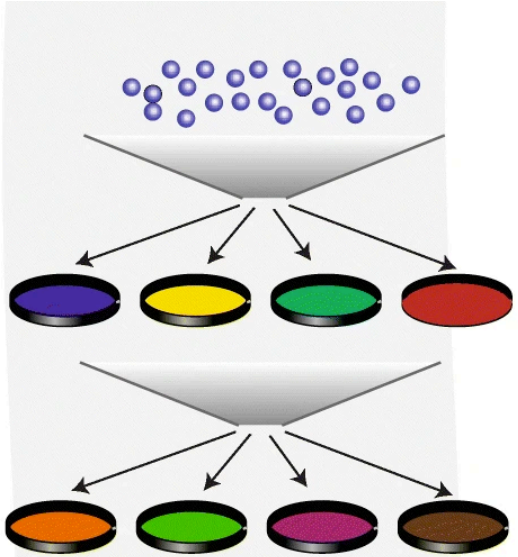


Plate or array
(ICELL8, Fluidigm C1)



Droplet-based
(10X Genomics; BioRad)

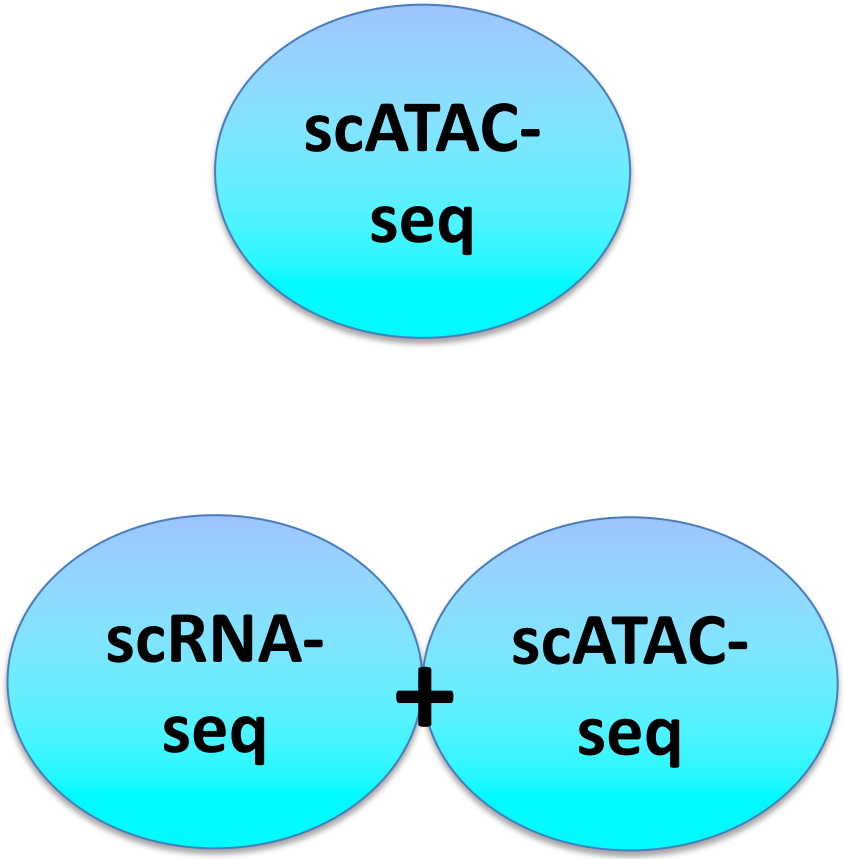


split-pool
(sciATAC-seq)

https://www.youtube.com/watch?v=WqaeZe7mKUc

Chen et al, Genome Biol 2019

Different scATAC-seq Applications



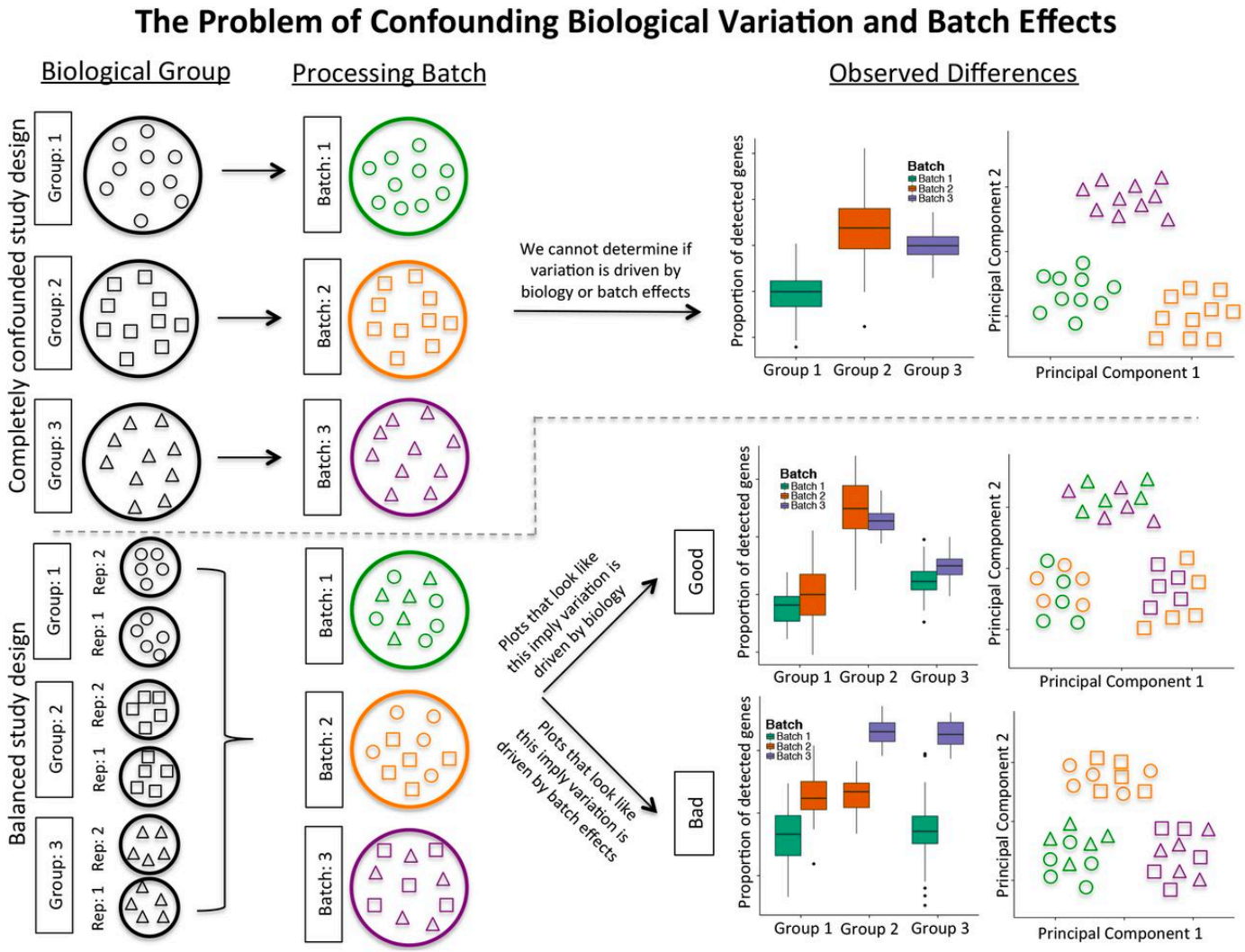
- Develop technology for single cell epigenetic profiling
- Study gene regulation at single cell resolution
  - Epigenetics in the context of gene expression
  - Same tissue but different cells
  - Same tissue and same cells

Zoom poll

What are the limitations for split-pool method?

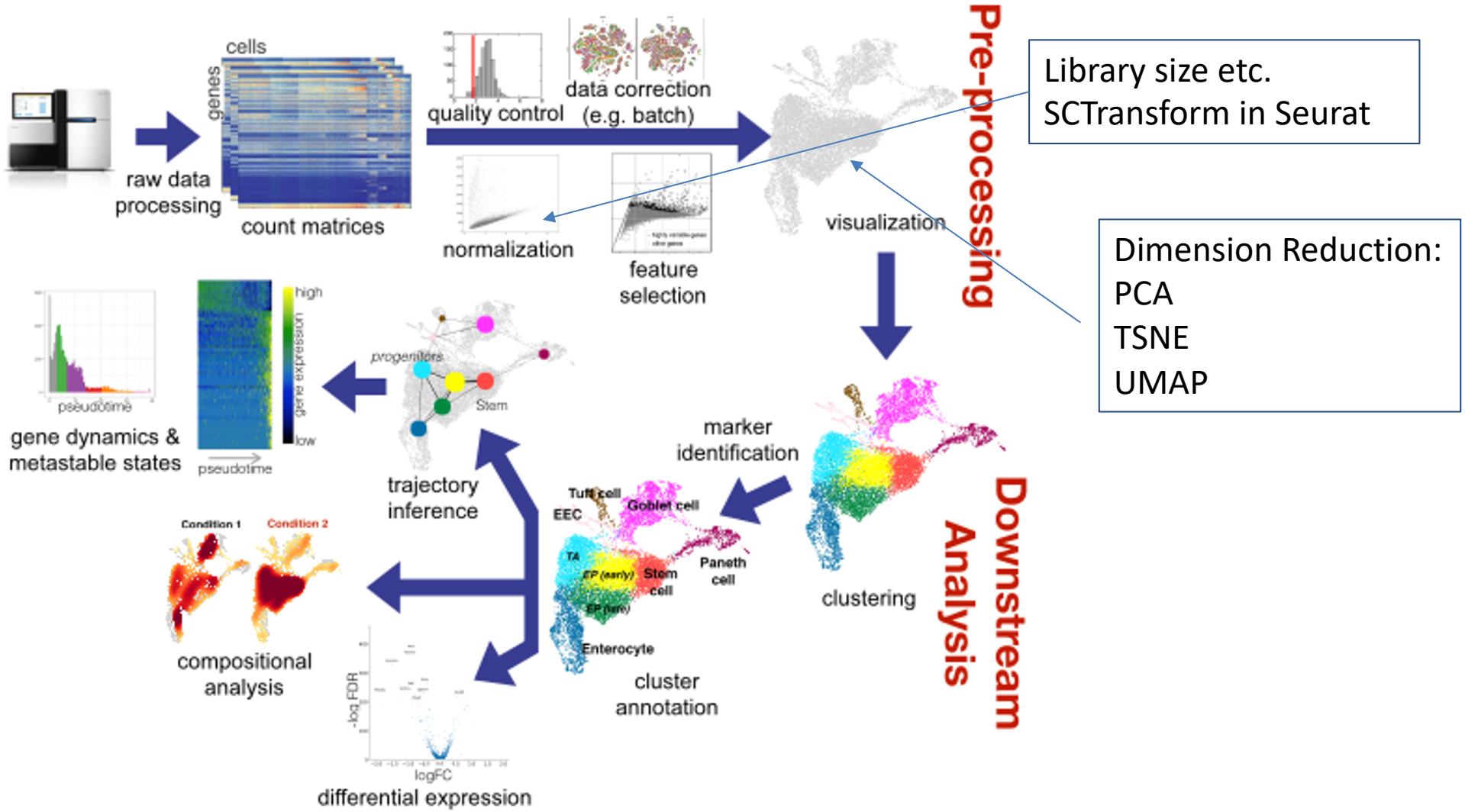
1. It will need a lot of cells to begin with
2. It is labor intensive
3. Both above

Avoid batch and confounding effects: experimental design



Hicks *et al.*, *Biostatistics*. 2018

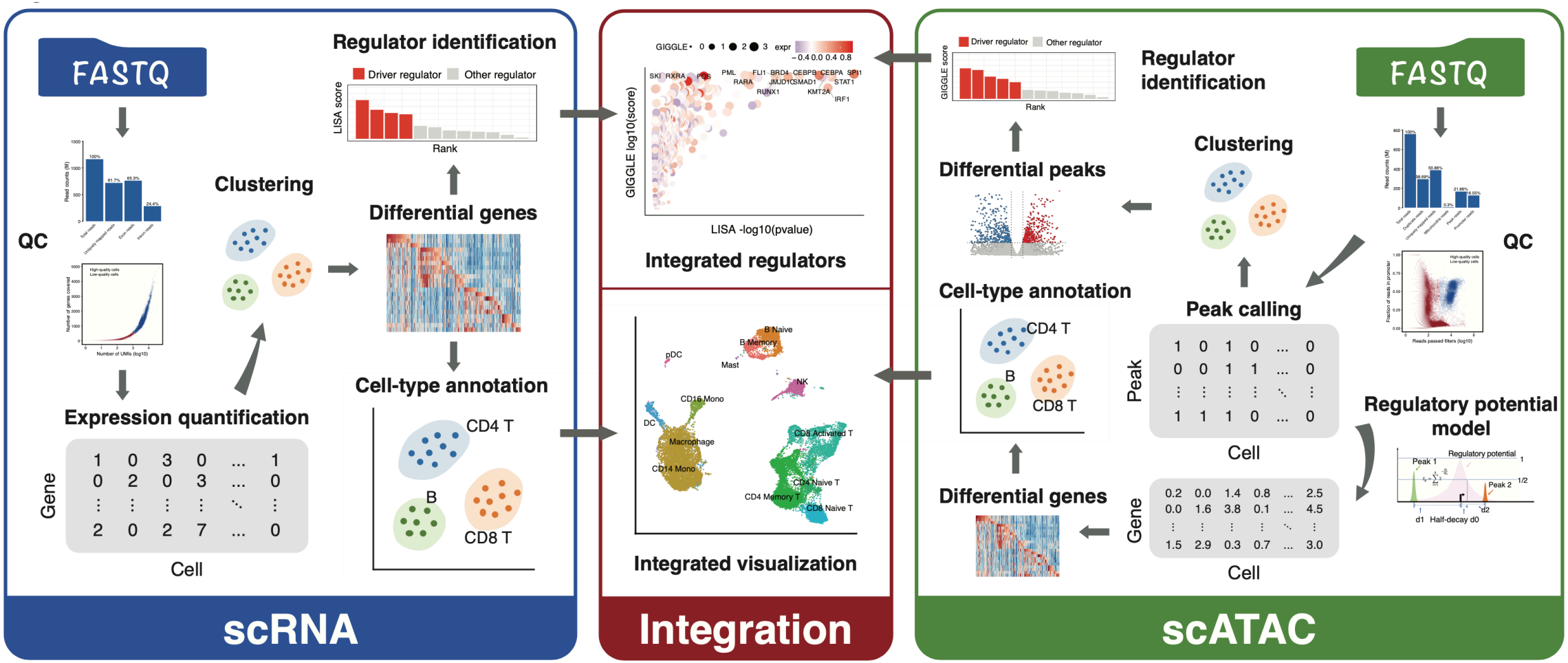
Workflow of a typical* scRNA-seq analysis



Credit to Peter Hickey

Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, (2019).

MAESTRO, an integrative analysis workflow based on Snakemake for scRNA-seq and scATAC-seq



https://github.com/liulab-dfci/MAESTRO

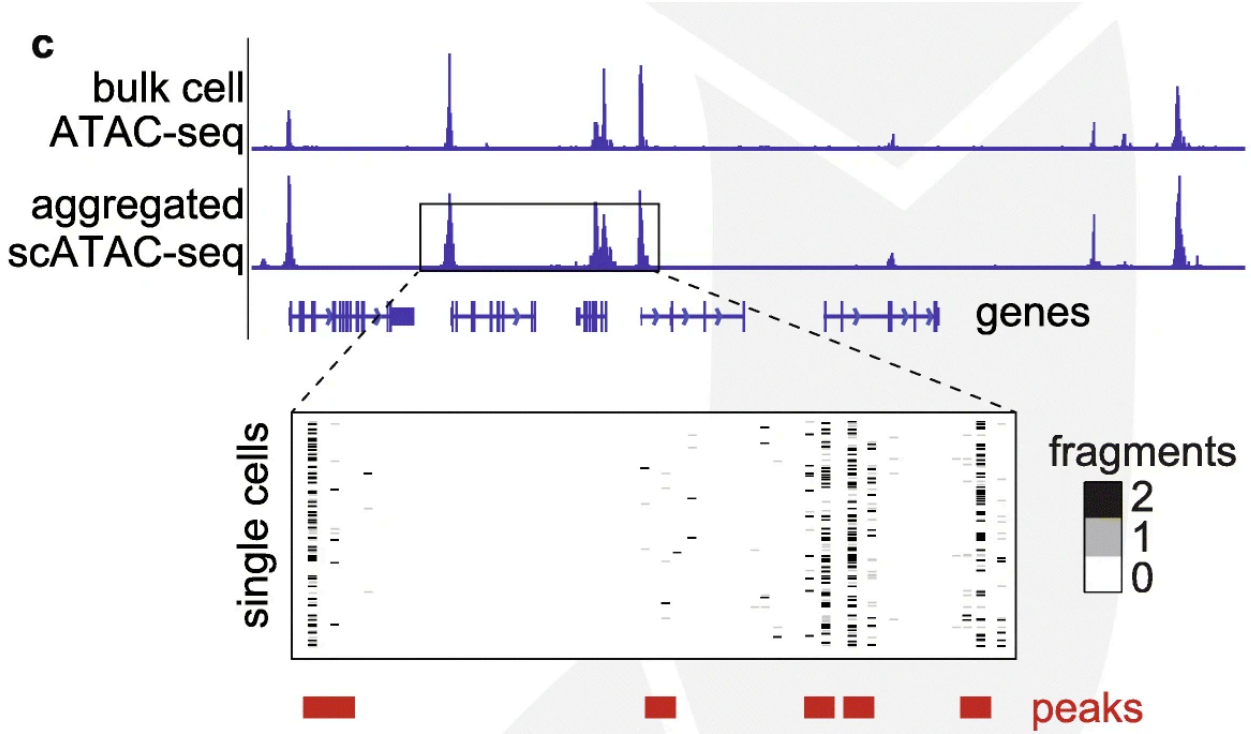
Wang et.al Genome Biol 2020

Read Alignment

- Cell Ranger (10X Genomics) solution
- RNA-seq: STAR
  - STARsolo (Blibaum et al, F1000 2019): 10X faster than CellRanger
- ATAC-seq: BWA
  - Minimap2 (Li, Bioinfo 2018): 15x faster than CellRanger
  - **Question**: can you use Salmon or Kallisto (pseudo-alignment tools) for scATACseq?
- Resolve cell barcode and correct barcode sequencing errors

Sample QC

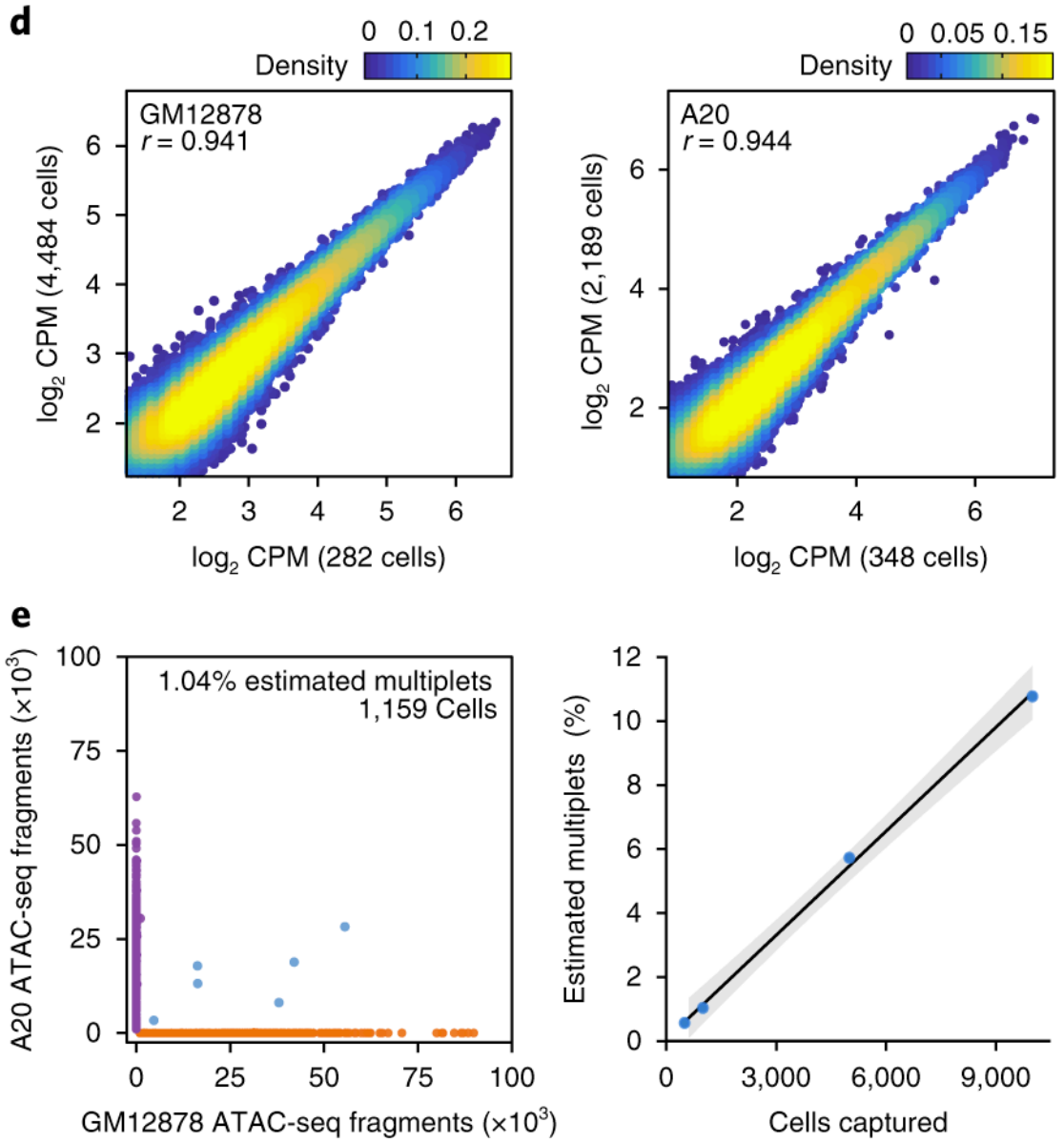
- Peak calling: MACS2
- RSeQC for RNA-seq
- ChIP-seq QC for ATAC-seq



Chen et al, Genome Biol 2019

Technology Development QC

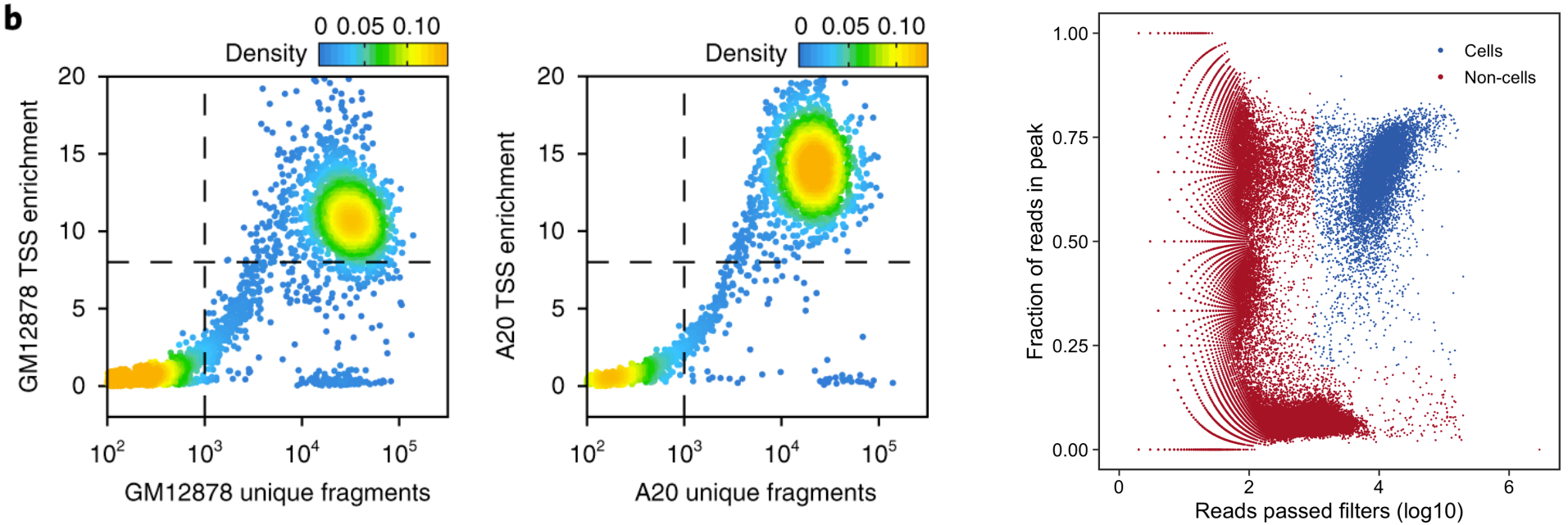
- Agreement between replicates of different cells
- Low multiplets mixing of human and mouse cells
- Not necessarily used in common experiments



Satpathy et al, Nat Biotech 2019

scATAC-seq Cell QC

- % reads in promoters / peaks (good) or mitochondria (bad), often set empirically depending on experiments / platforms



- scATAC-seq pseudo-bulk often has better FRiP after cell QC

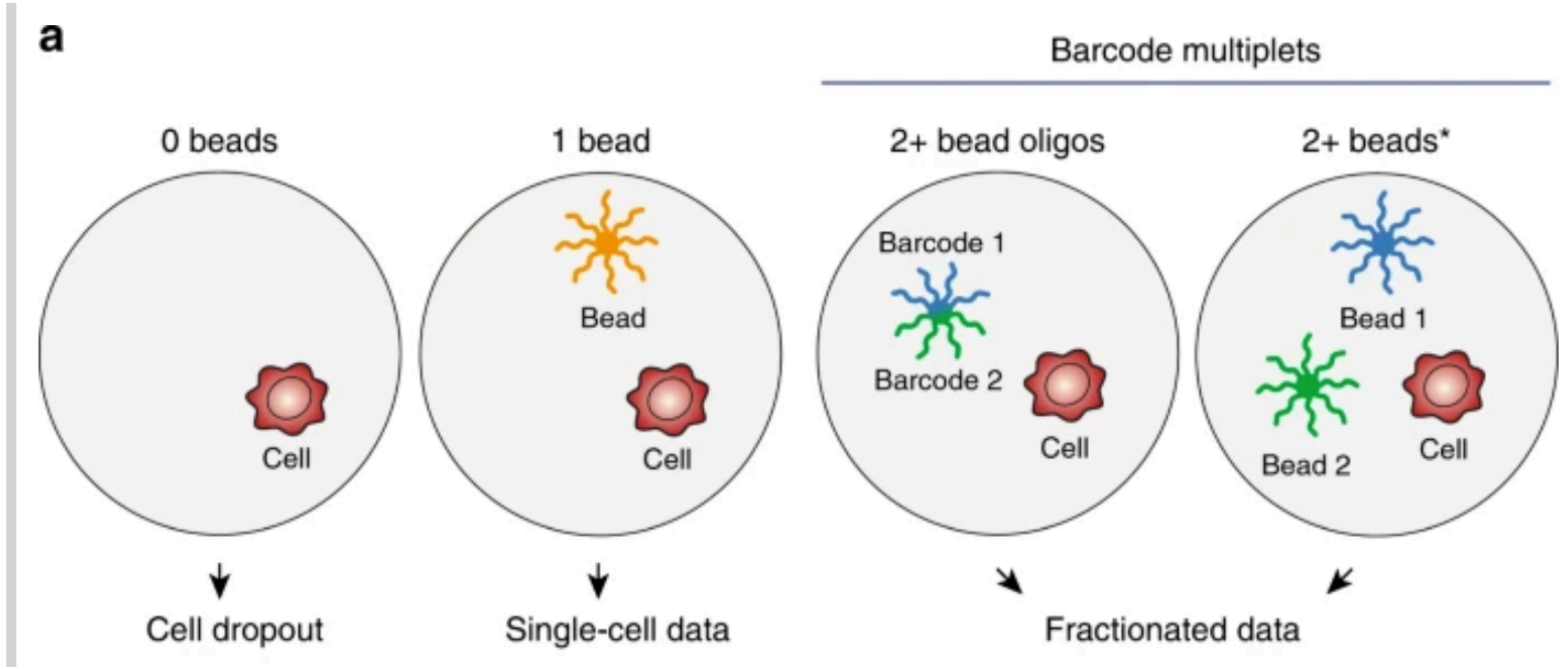
https://divingintogeneticsandgenomics.rbind.io/post/calculate-scatacseq-tss-enrichment-score/

Satpathy et al, Nat Biotech 2019

https://github.com/liulab-dfci/MAESTRO

What else could go wrong?

- Doublet (more than two cells in the same droplet)
- Different barcodes can be from a single droplet

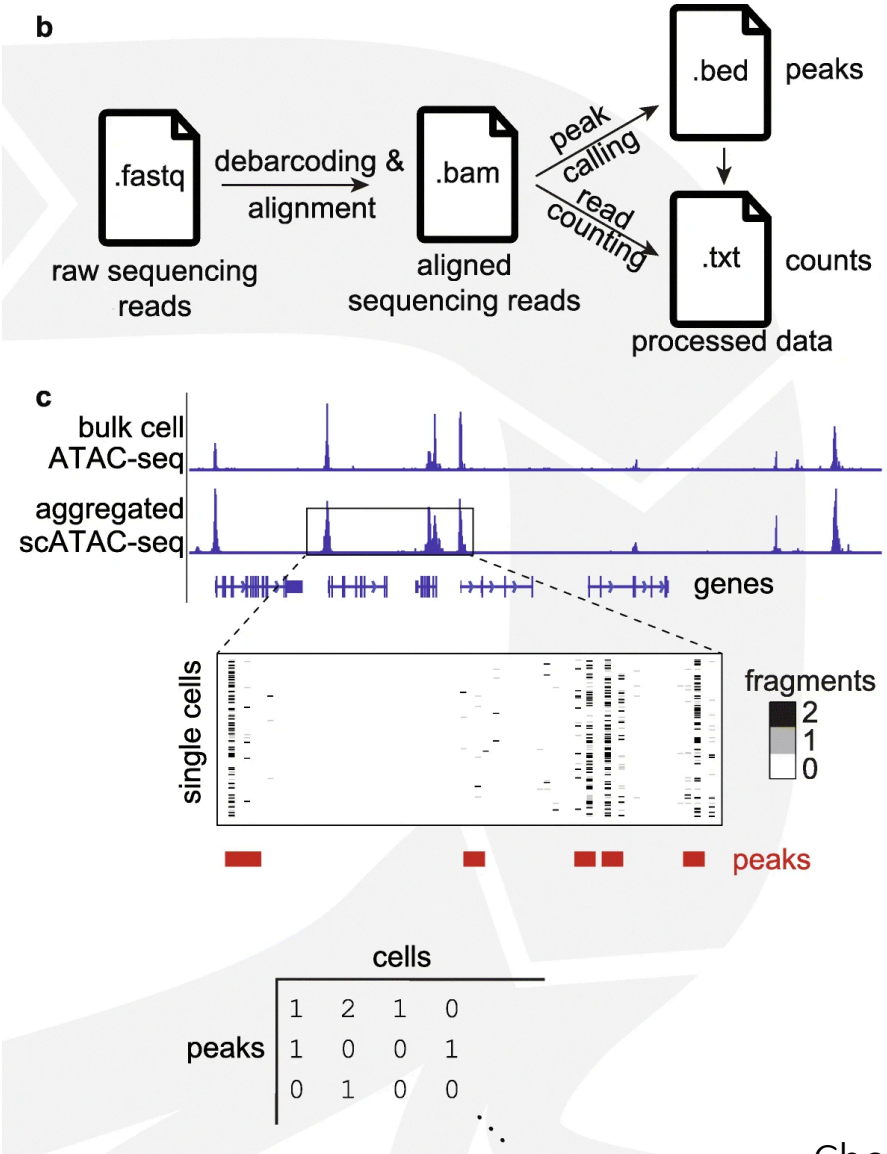


Lareau et.al 2019 Nature Communications

Zoom poll

- Which one is better for QC?
- 1. Fraction of reads in peaks
- 2. Fraction of reads in promoters

Generate Count Matrix



Chen et al, Genome Biol 2019

Technical considerations

- Only look at reads on peaks in each cell. (peak based)
- SnapATACseq uses bin-based method. (count reads in 5kb tiling bin across the genome)
- Generate peak by cell count matrix or bin by cell count matrix
- Very sparse, no UMI
- Could also be binary matrix
- TF-IDF transformation.
- Downstream analysis starts here

TF-IDF transformation

matrices using a term **frequency-inverse document frequency ("TF-IDF") transformation.** To do this, we first weighted all the sites for individual cells by the total number of sites accessible in that cell ("term frequency"). We then multiplied these weighted values by log(1 + the inverse frequency of each site across all cells), the "inverse document frequency." We then used singular value decomposition on the TF-IDF matrix to generate a lower dimensional representation of the data by only retaining the 2nd through 10th dimensions (because the first dimension tends to be highly correlated with read depth).

Darren et.al Cell 2018

```
TF.IDF.custom <- function(data, verbose = TRUE) {
  if (class(x = data) == "data.frame") {
    data <- as.matrix(x = data)
  }
  if (class(x = data) != "dgCMatrix") {
    data <- as(object = data, Class = "dgCMatrix")
  }
  if (verbose) {
    message("Performing TF-IDF normalization")
  }
  npeaks <- Matrix::colSums(x = data)
  tf <- t(x = t(x = data) / npeaks)
  # log transformation
  idf <- log(1+ ncol(x = data) / Matrix::rowSums(x = data))
  norm.data <- Diagonal(n = length(x = idf), x = idf) %*% tf
  norm.data[which(x = is.na(x = norm.data))] <- 0
  return(norm.data)
}


mat<- TF.IDF.custom(mat)
```

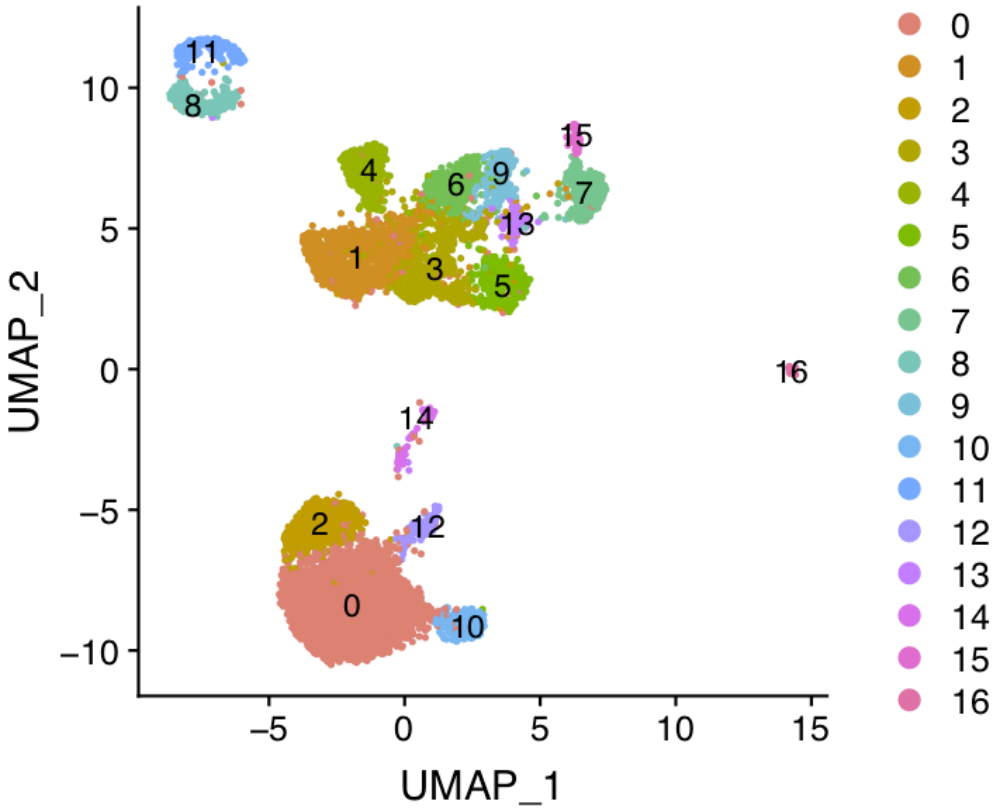
https://divingintogeneticsandgenomics.rbind.io/post/clustering-scatacseq-data-the-tf-idf-way/

Clustering

- Dimension reduction

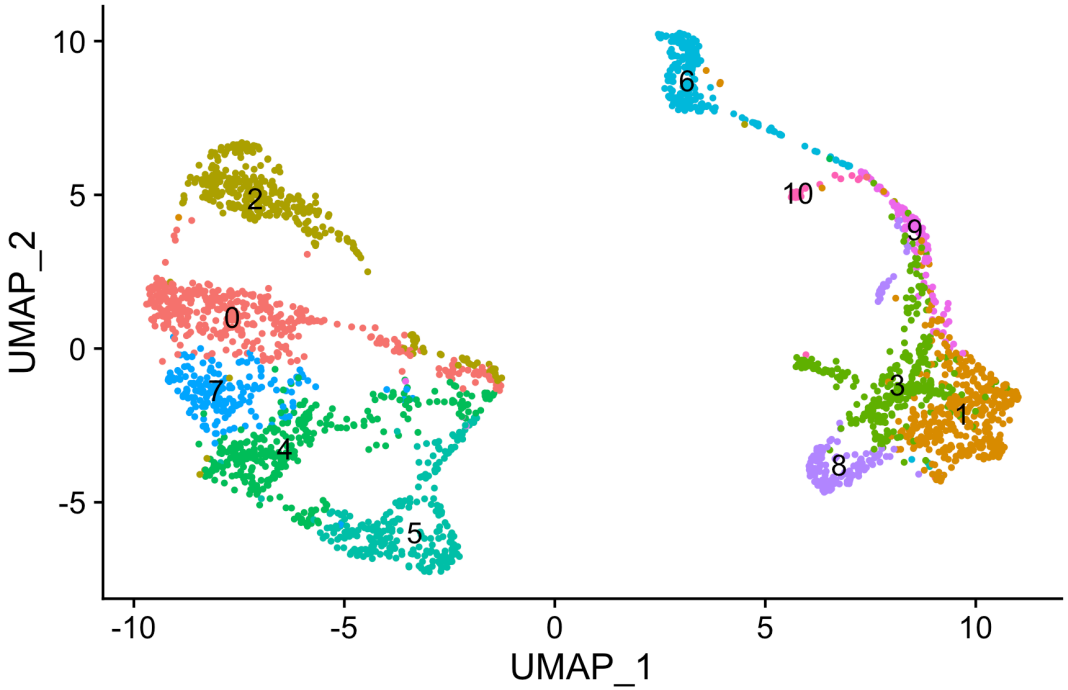
  - Latent semantic indexing (LSI): SVD applied to term-document matrix, i.e., peak-cell count matrix after TF-IDF transformation.

- Cluster cells (reduced dimension) using graph-based method in Seurat v3 (Stuart et al, Cell 2019). KNN graph + community detection algorithm
- Can visualize using t-SNE / UMAP

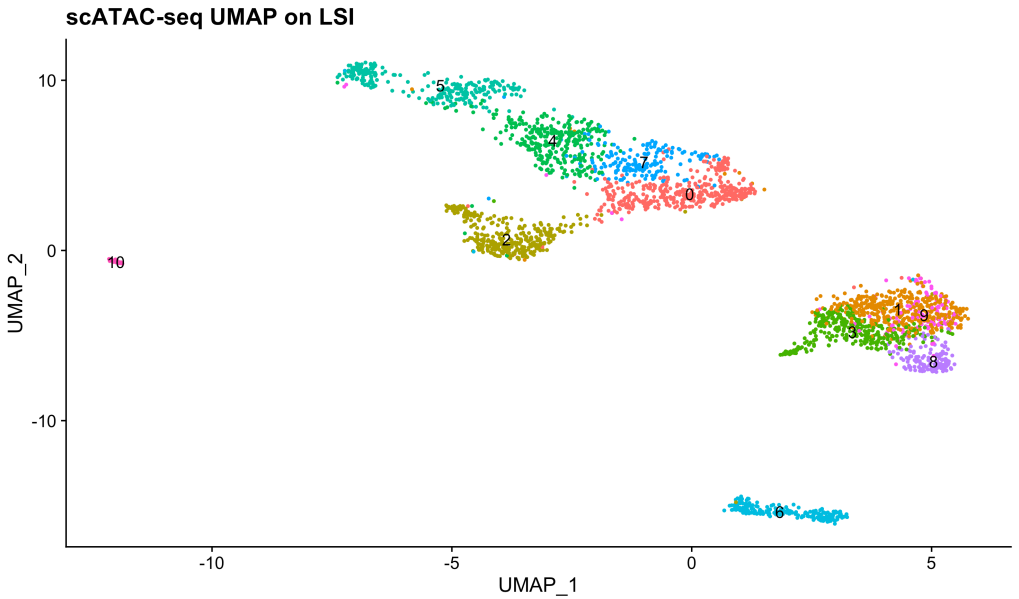


scATAC umap with and without TF-IDF

No TF-IDF



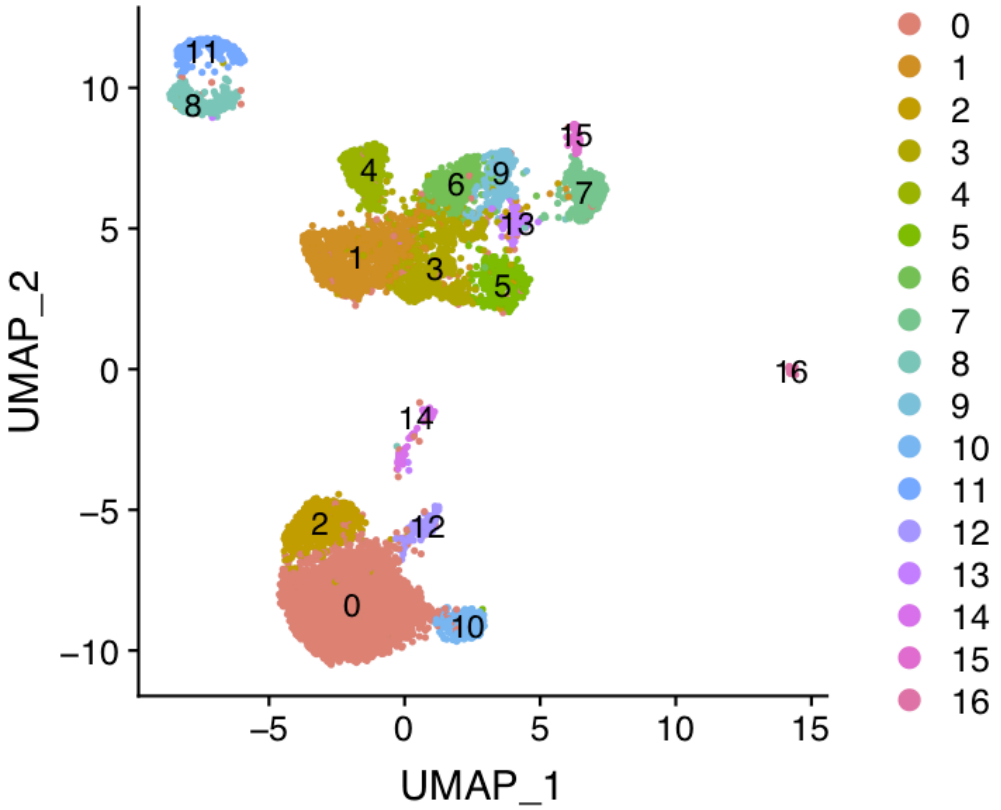
With TF-IDF



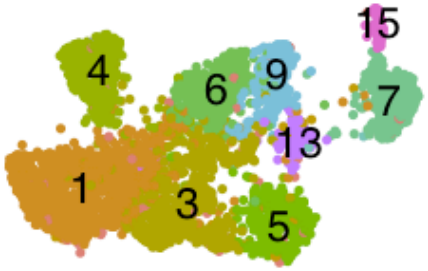
Latent semantic indexing (LSI) = TF-IDF + SVD

Peak Calling #2

- Optional step
- Assign cells to clusters
- Call peaks on original reads from cells in each cluster
- Sometimes can call some (small percentage) new peaks in minor clusters
- Merge peaks from clusters and regenerate count matrix



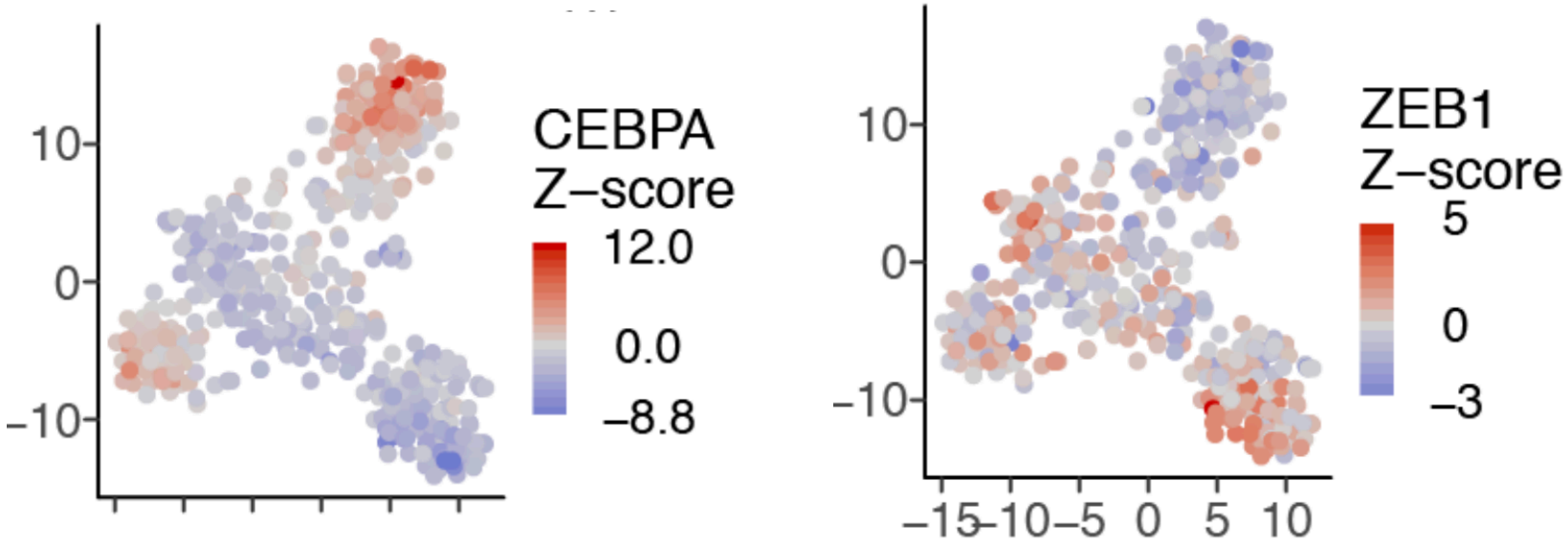
Differential Peak Calling



- At each peak (row), calculate differential enrichment between cluster X and NOT (cluster X)
- Mann–Whitney U test / Wilcoxon rank-sum test on peak-cell count matrix
- To overcome ties (0, 1) in peak-cell count matrix, normalize data in each cell (col), scale to 10K (reads / cell)
- Presto (Korsunsky et al, https://www.biorxiv.org/content/10.1101/653253v1): implementation of Wilcoxon test 1000 times faster than in Seurat.

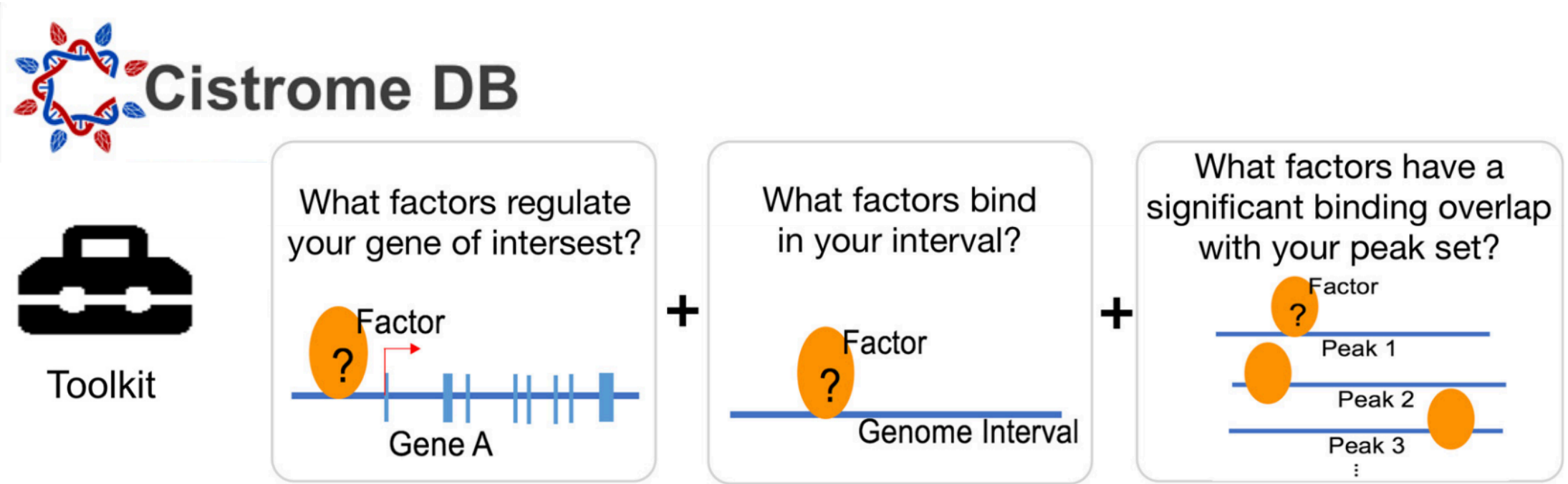
Annotate Relevant TFs with Motifs

- ChromVar (Schep et al, Nat Meth 2017) to find (a few hundred) TF motifs enriched in the peak count vector of each cell and visualize the results on tSNE / UMAP



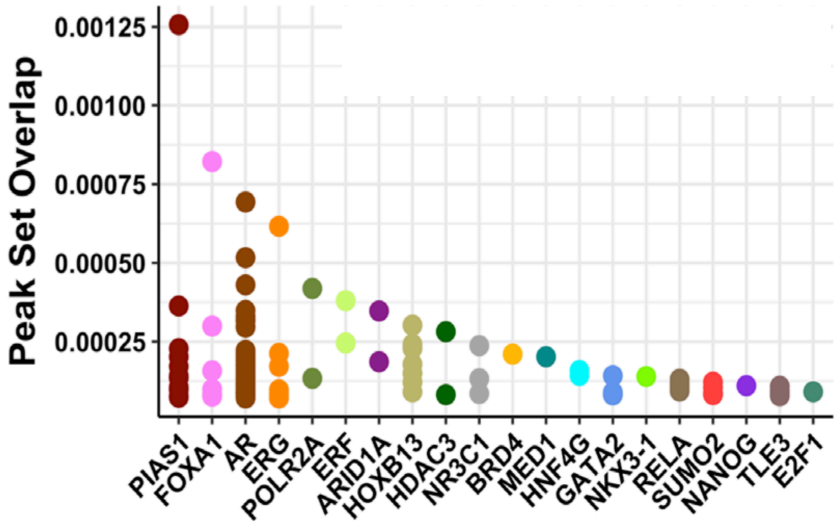
Annotate Relevant TFs with ChIP-seq

- CistromeDB Toolkit function (Zheng el al, NAR 2019)
- Uses Giggle (Layer et al, Nat Meth 2018) to find significant overlap with all public ChIP-seq data



Annotate Relevant TFs with ChIP-seq

- CistromeDB Toolkit function (Zheng el al, NAR 2019)
- Input (differential) ATAC-seq peaks (in bed file) in a cluster
- Output public TF ChIP-seq data with best overlap



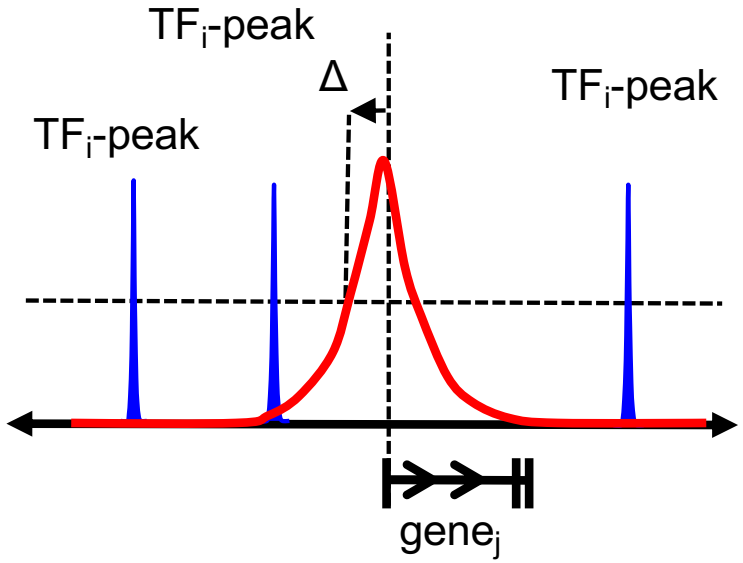
Zoom Poll

- What are the challenges with the scATACseq count matrix?
- 1. very sparse
- 2. no UMI

- What could be the issues using the peaks called from all cells?
- 1. peaks in a rare cell type can be missed
- 2. the clustering will miss the rare population

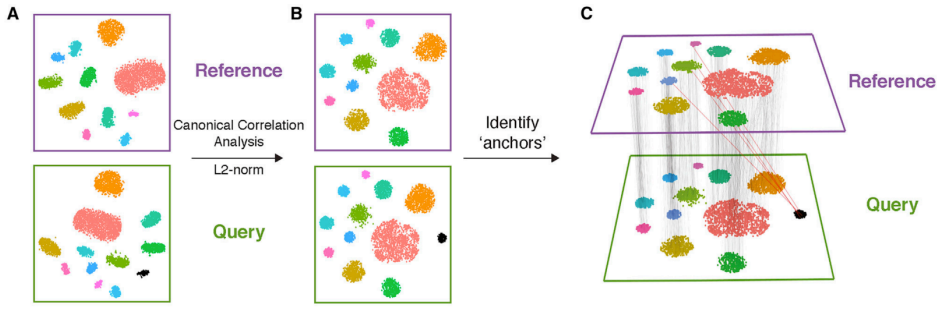
Annotate Cell Types

- Convert peak-by-cell count matrix (scATAC-seq) to gene-by-cell expression matrix (similar to scRNA-seq)
  - Promoter openness not a good proxy for gene expression
  - Gene body read coverage, could smooth data with K-nearest neighbor cells
  - Regulatory potential (MAESTRO, Wang et al https://github.com/liulab-dfci/MAESTRO)

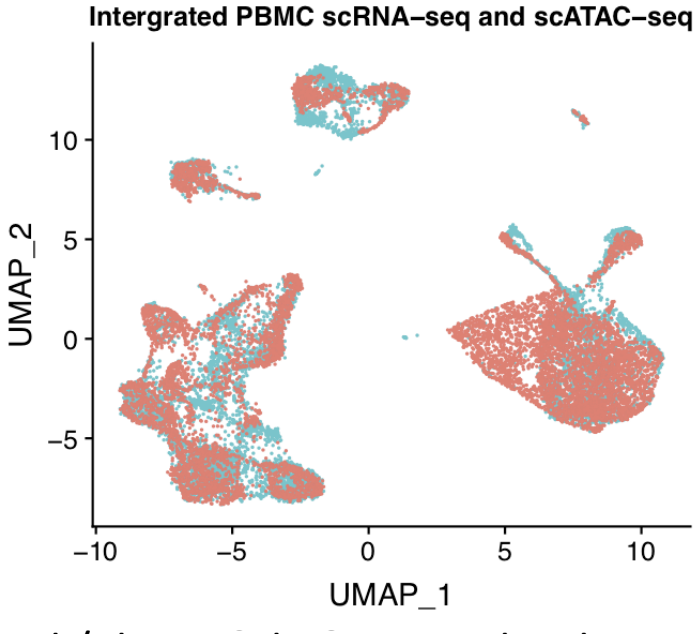


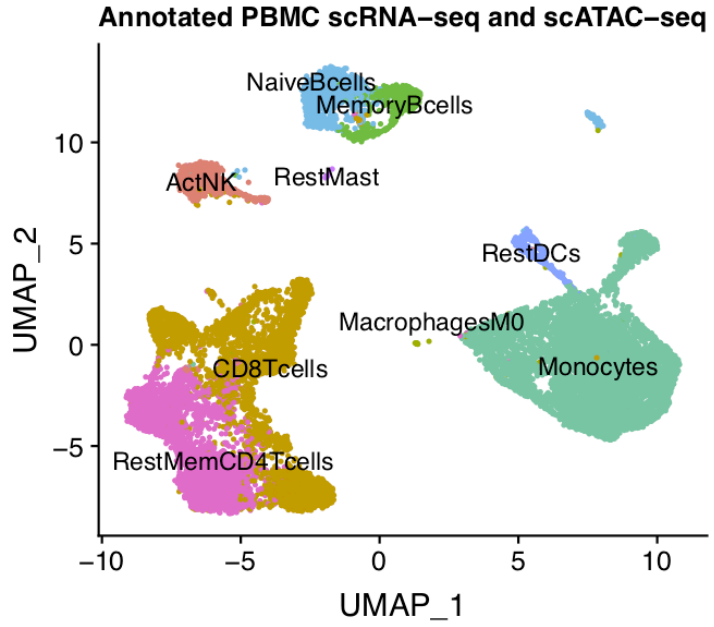
Integrate scATAC-seq with scRNA-seq

- Convert peak-by-cell count matrix (scATAC-seq) to gene-by-cell expression matrix (similar to scRNA-seq) using regulatory potential
- Use CCA to combine gene by cell matrices from scATAC-seq and scRNA-seq, treating these two matrices like two batches
- Transfer cell annotation from scRNA-seq cluster to scATAC-seq



Stuart et.al. 2019 Cell

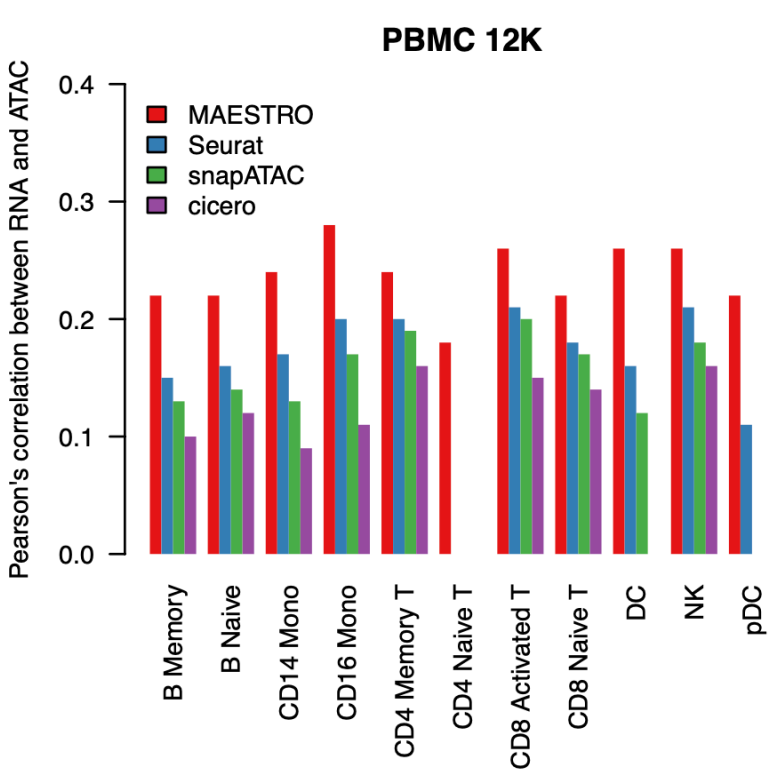


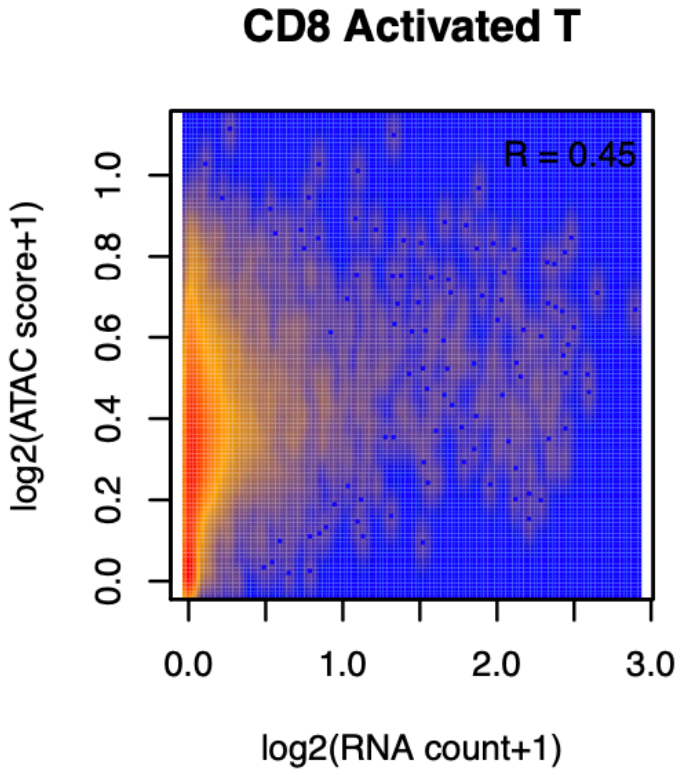


CCA: http://web.stanford.edu/class/bios221/book/Chap-MultivaHetero.html

Annotate Cell Types

- Convert peak-by-cell count matrix (scATAC-seq) to gene-by-cell expression matrix (similar to scRNA-seq) using regulatory potential
- Annotate cell type and find marker genes similarly to scRNA-seq



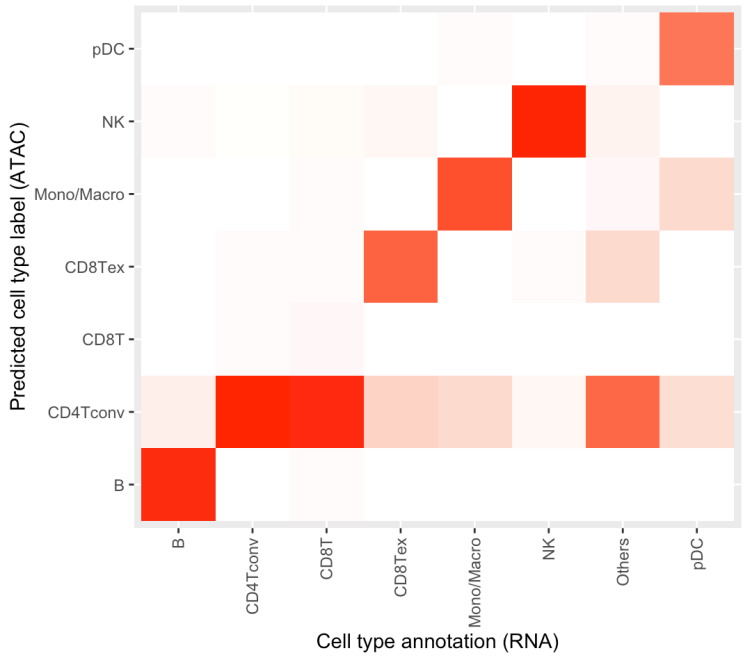


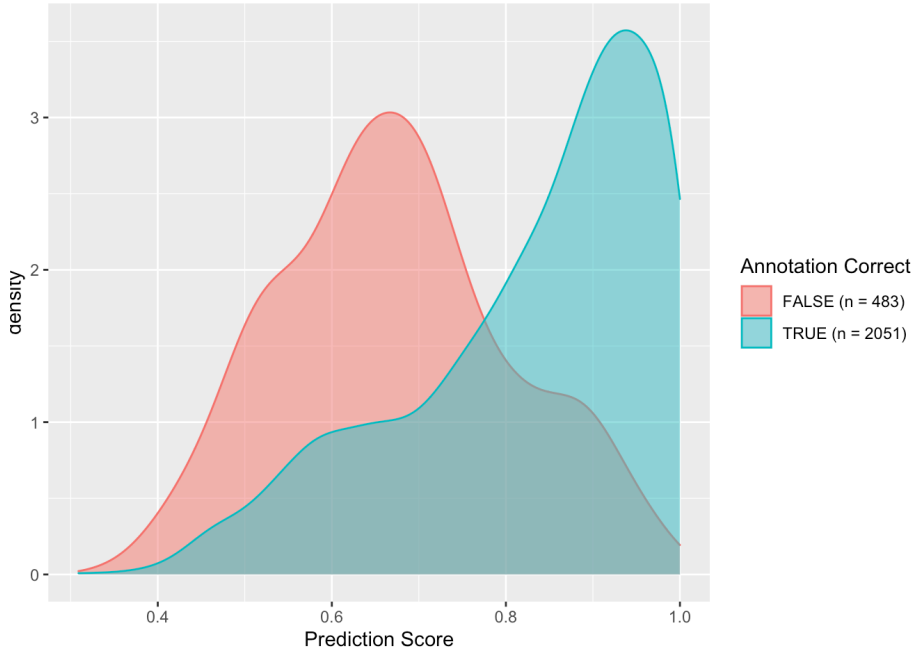
PCC = 0.26
SCC = 0.45

Using multiome data to evaluate label transfer accuracy

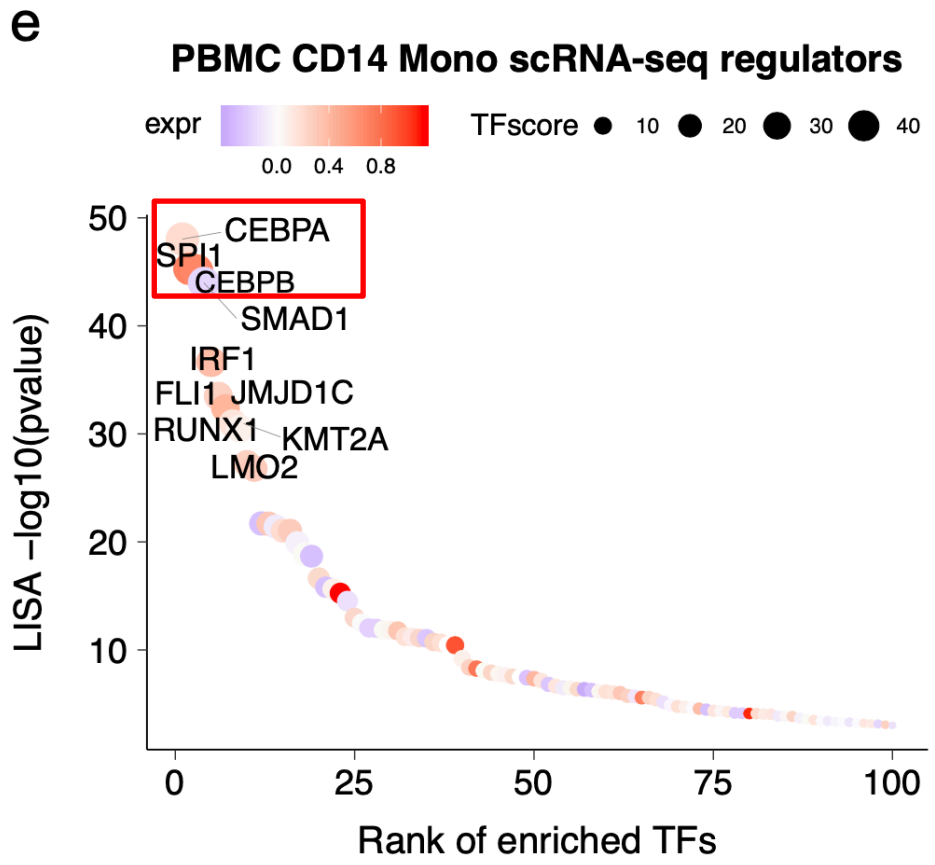
cells	scRNA	scATAC
cell1	CD4	CD4
cell2	CD8	CD8
cell3	CD4	CD8
cell4	NK	NK
...		

Label transfer

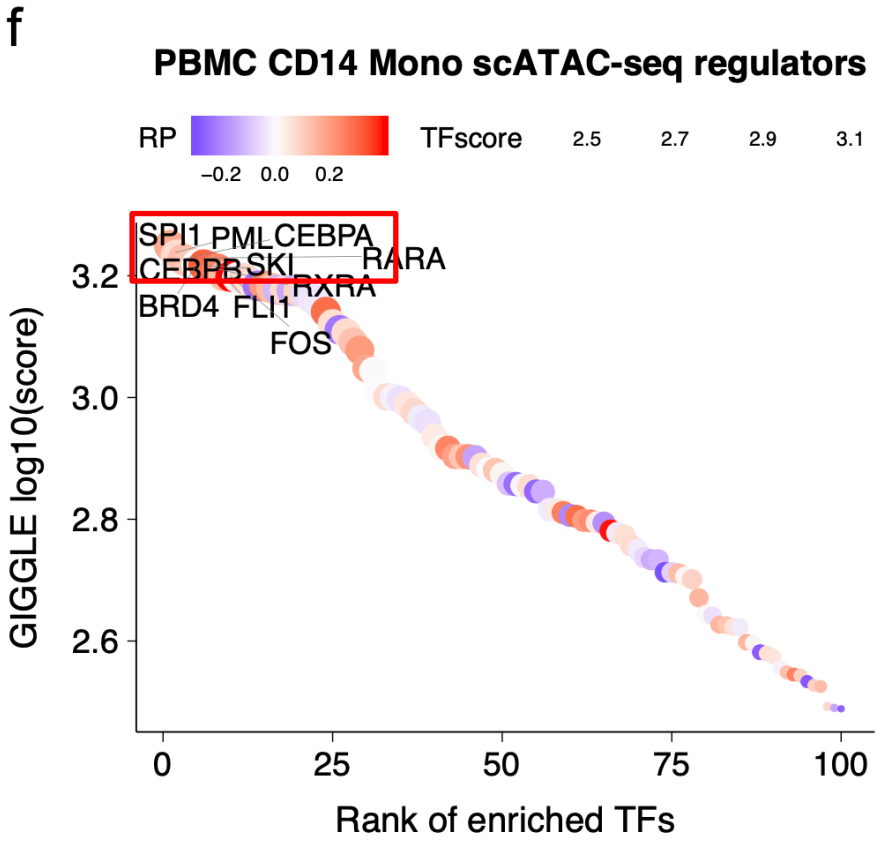




Visualize driver TFs for scRNA/ATAC-seq



Based on up-regulated genes in each cluster



Based on positive peaks in each cluster

LISA@ http://lisa.cistrome.org/

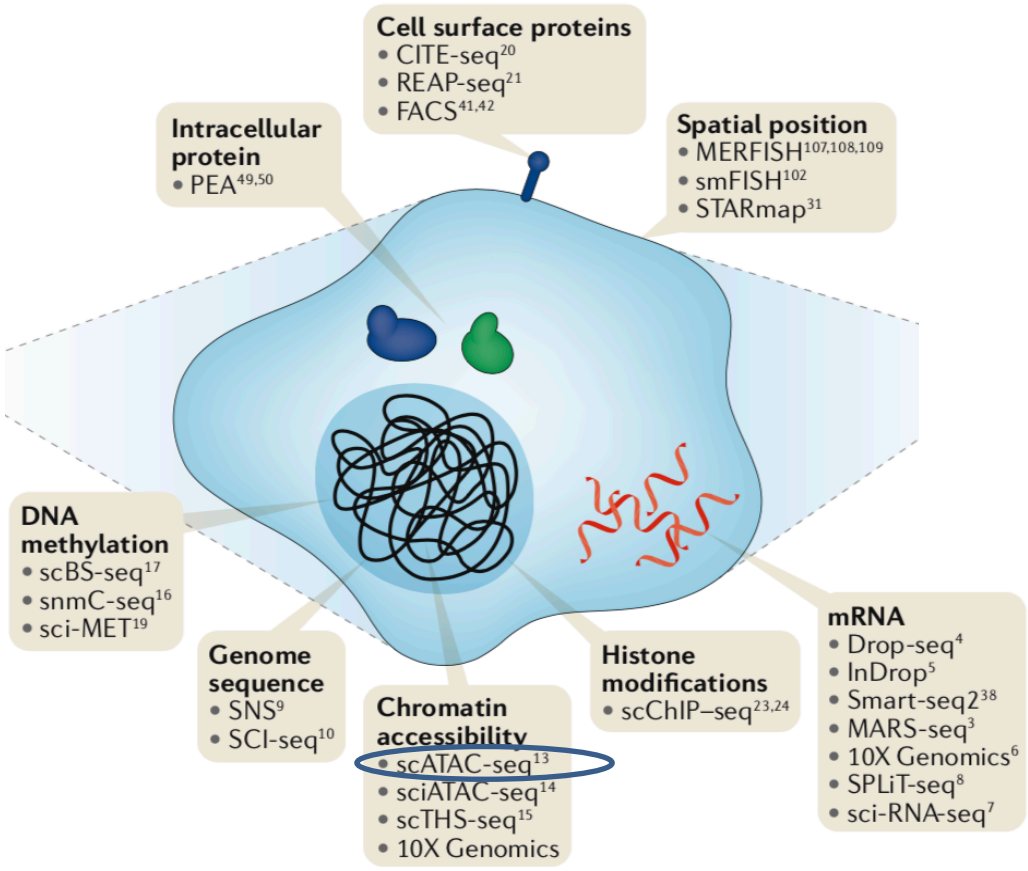
Cistrome Data Browser

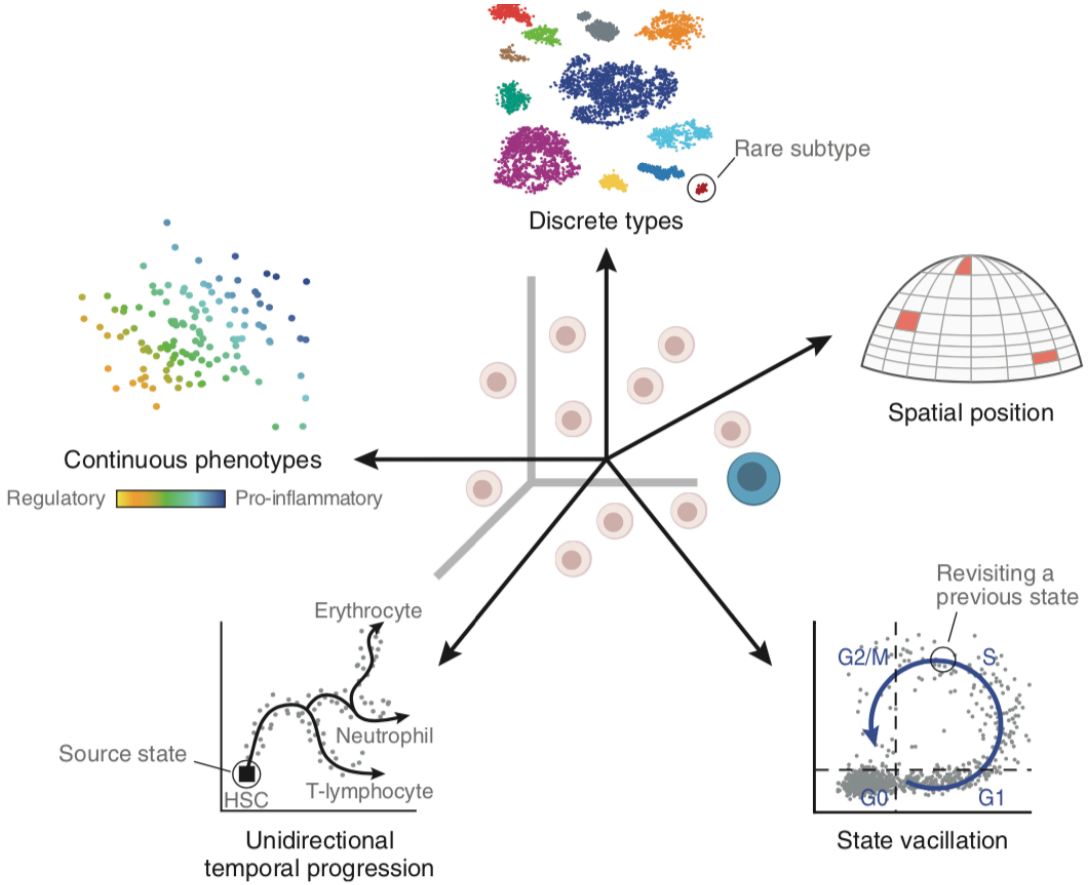
http://cistrome.org/db/#/
http://dbtoolkit.cistrome.org/

Summary

- Read mapping, peak calling
- Sample and cell QC, peak – cell matrix
- Clustering and visualization
- Differential peaks and TF annotation
- Use ATAC-seq regulatory potential as proxy for gene expression
- Integrate scATAC-seq with scRNA-seq and annotate cells
- Identify driver TFs
- Both technologies and computational methods still fast evolving

Other single-cell related areas under fast development





- Multi-omics
- Spatial transcriptome/epigenome
- Single-cell DNA-seq

Tim Stuart & Rahul Satjia, *Nat Rev Genet*, 2019

Wager et al, *Nat Biotech*, 2016

Acknowledgements

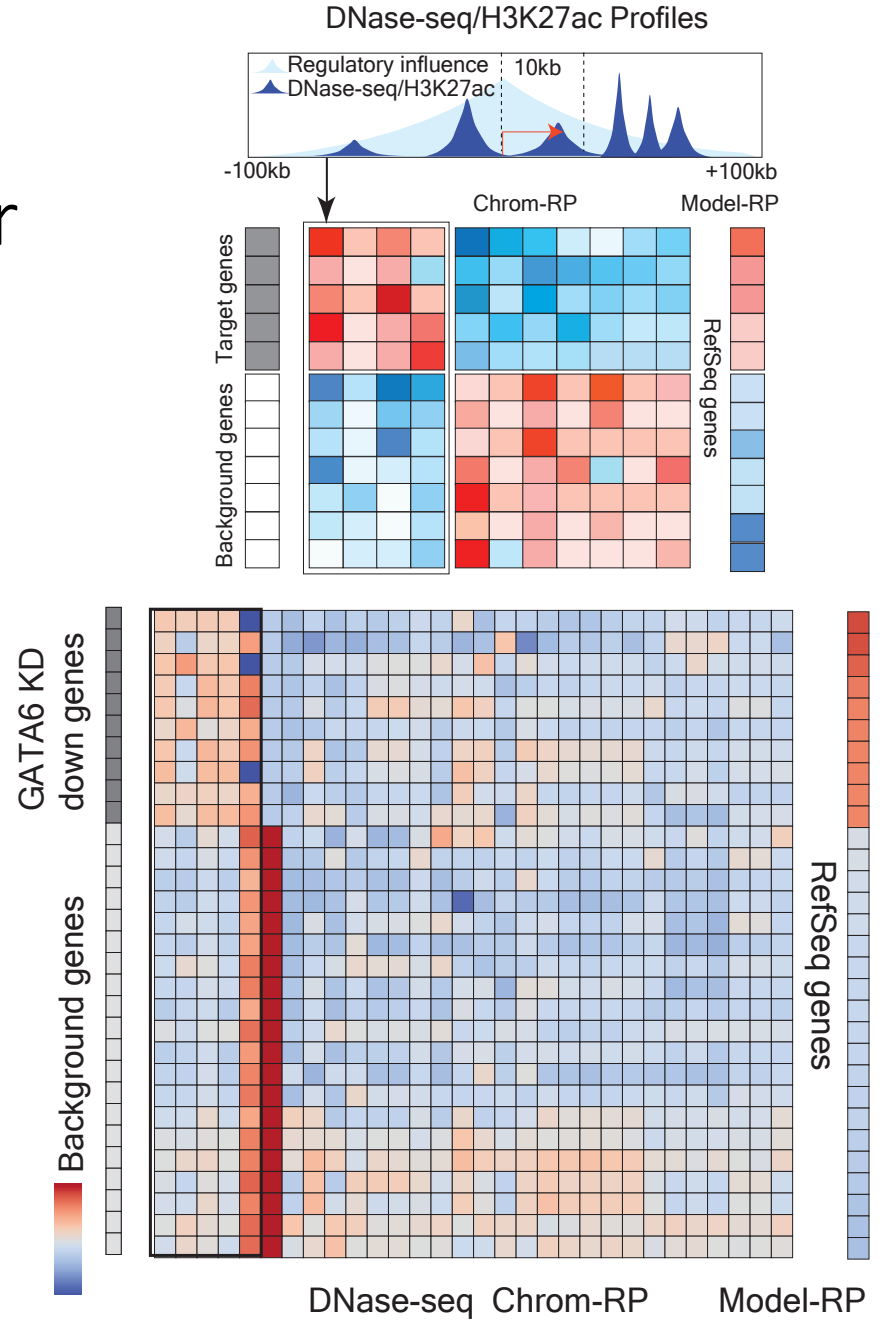
- Shirley X Liu
- Gali Bai
- Luciano Martelotto
- Chenfei Wang
- Dongqing Sun
- Luca Pinello
- Jingyu Fan
- Qian Qin
- Cliff Meyer

references

- https://www.sciencedirect.com/science/article/pii/S2001037020303019 Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation
- https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1854-5 Assessment of computational methods for the analysis of single-cell ATAC-seq data
- https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1929-3 From reads to insight: a hitchhiker's guide to ATAC-seq data analysis
- https://www.nature.com/articles/s43586-020-00008-9 Chromatin accessibility profiling methods

LISA Chromatin Model

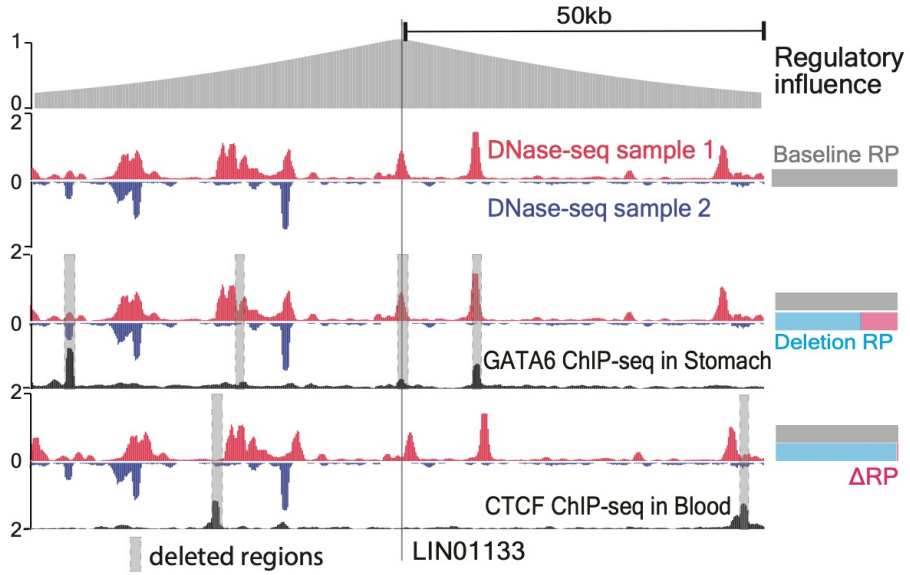
- Calculate regulatory potential (RP) for all public H3K27ac or DNase-seq data
- Given a list (few hundred) of up or down regulated genes (from scRNA-seq cluster, or any differential expression data)
- Use LASSO regression to select and weigh relevant publicly available H3K27ac ChIP-seq or DNase-seq profiles based on RP

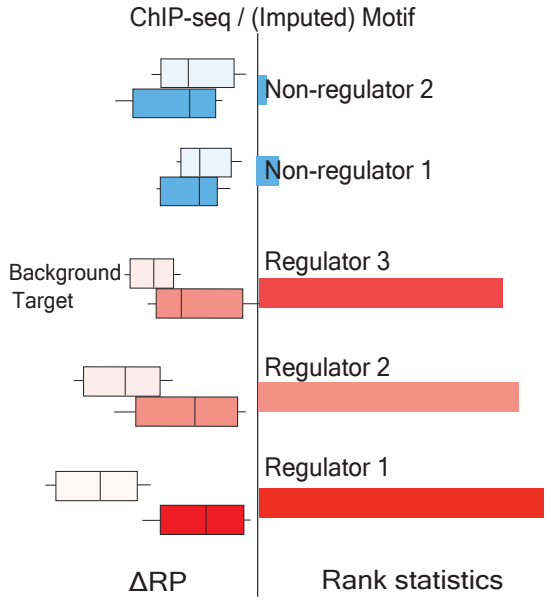


http://lisa.cistrome.org/ (Qin et al, Genome Biol 2020)

LISA TF Driver Inference

- Examine each TF ChIP-seq and motif (if without ChIP-seq)
- In silico delete binding from selected H3K27ac / DNase profiles
- Look at whether changes on regulatory potential are enriched on the user input differential genes

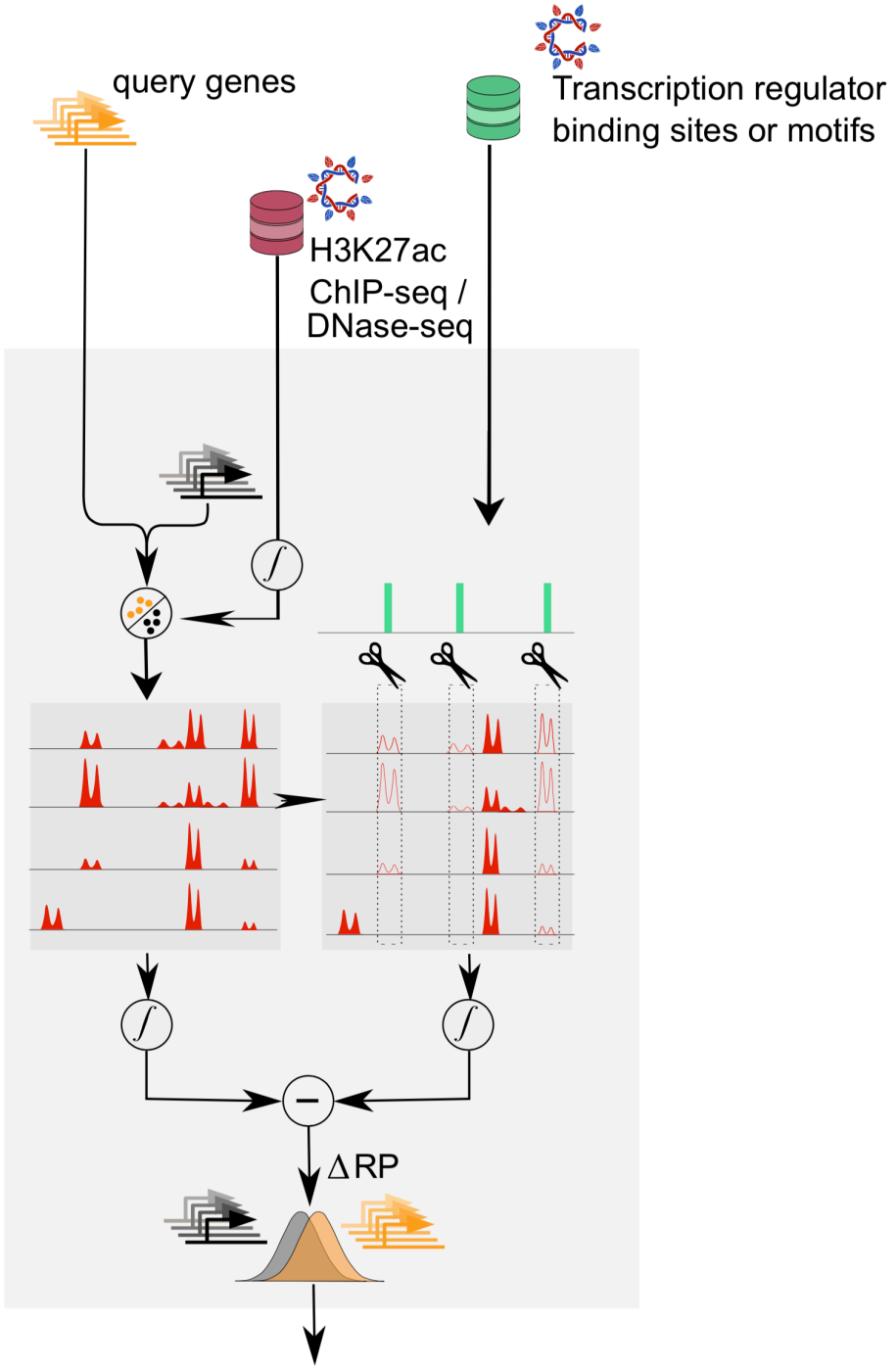




http://lisa.cistrome.org/ (Qin et al, Genome Biol 2020)

Predict Driver TFs from Gene Set

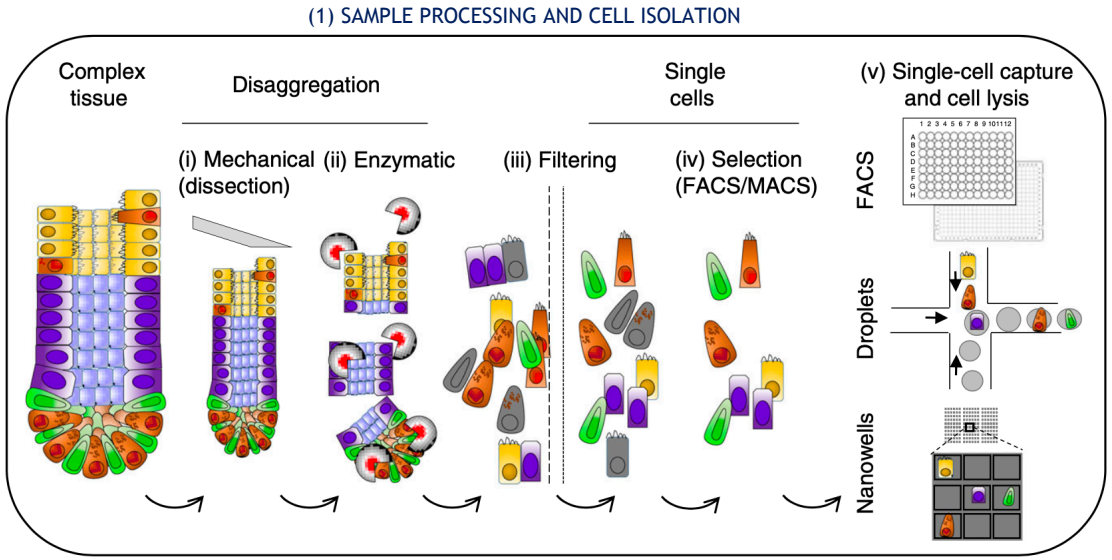
- Based on (e.g. differential expression) gene set, infer an epigenetic profile by selecting and weighing relevant publicly available H3K27ac ChIP-seq, DNase/ATAC-seq profiles
- Identify TF ChIP-seq or enriched TF motifs in the epigenetic profiles most relevant to the gene set



http://lisa.cistrome.org/ (Qin et al, Genome Biol 2020)

Single cell RNA-seq experiments: sample processing

Goal → live single cell suspensions



High level considerations

- ✓ Species
- ✓ Solid tissue (what organ?)
- ✓ Liquid biopsy (e.g. blood)
- ✓ Biopsy size/volume (e.g. cellularity)
- ✓ Cell composition
- ✓ Cellular heterogeneity
- ✓ ECM composition
- ✓ Post-mortem
- ✓ Naïve or treated (e.g. pre/post chemo)
- ✓ Viability
- ✓ Preservation and storage
- ✓ Rare cell population?
- ✓ All cells or a fraction of them?
- ✓ Total cells vs. aimed cells
- ✓ Known protocols for tissue dissociation?
- ✓ Methodology/Technology (i.e. droplet vs. plate)
- ✓ Joint profiling? What other application?

Different barcodes can be from a single droplet

Article | Open Access | Published: 13 February 2020

Inference and effects of barcode multiplets in droplet-based single-cell assays

Caleb A. Lareau, Sai Ma, Fabiana M. Duarte & Jason D. Buenrostro

Nature Communications **11**, Article number: 866 (2020) | Cite this article

Abstract

A widespread assumption for single-cell analyses specifies that one cell's nucleic acids are predominantly captured by one oligonucleotide barcode. Here, we show that ~13–21% of cell barcodes from the 10x Chromium scATAC-seq assay may have been derived from a droplet with more than one oligonucleotide sequence, which we call "barcode multiplets". We demonstrate that barcode multiplets can be derived from at least two different sources. First, we confirm that approximately 4% of droplets from the 10x platform may contain multiple beads. Additionally, we find that approximately 5% of beads may contain detectable levels of multiple oligonucleotide barcodes. We show that this artifact can confound single-cell analyses, including the interpretation of clonal diversity and proliferation of intra-tumor lymphocytes. Overall, our work provides a conceptual and computational framework to identify and assess the impacts of barcode multiplets in single-cell data.

Zoom poll

- Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling
- https://www.nature.com/articles/s41587-020-0645-6

- 1. There is only one copy of mitochondrial DNA
- 2. There is no normal control