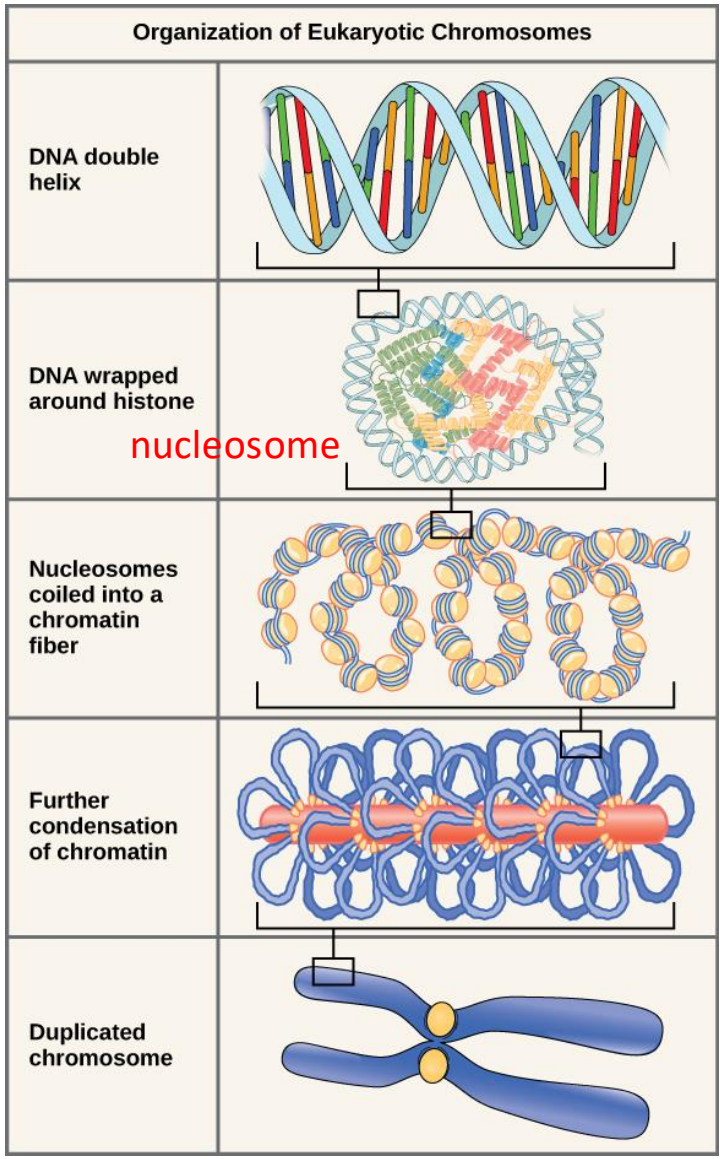
Lecture 11
scATAC-seq technology, preprocessing and dimension reduction

Outline

- scATAC-seq technology
- scATAC-seq preprocessing and quality control
  - Peak calling
  - Filtering low-quality cells
  - Doublet detection
  - Barcode multiplets
- Dimension reduction and feature transformations

Epigenomics and scATAC-seq



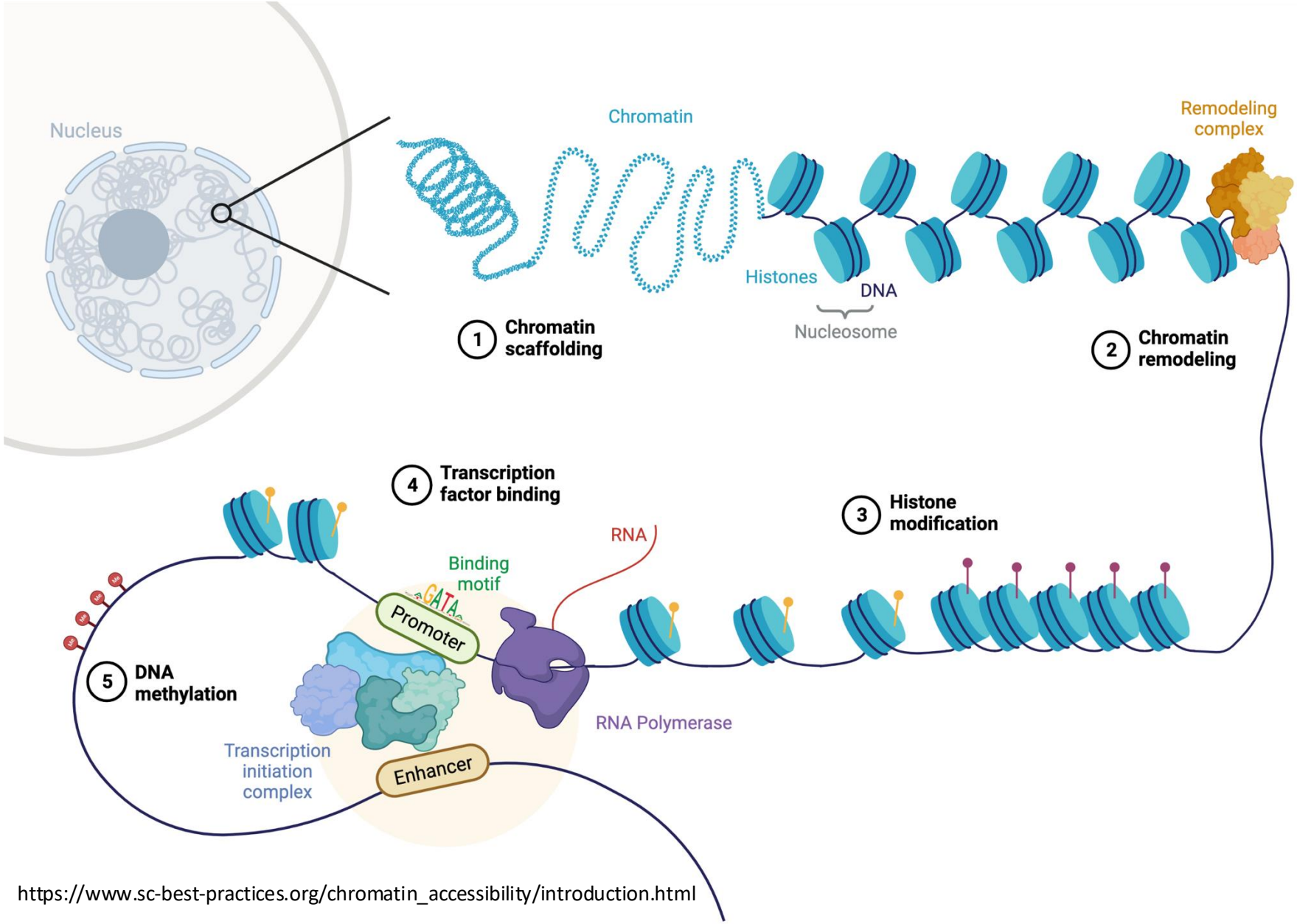
https://courses.lumenlearning.com/wm-biology1/chapter/reading-chromosome-structure/

- DNA is packaged inside the nucleus
  - Make DNA fit into the nucleus and stable
  - Controls the activity of DNA: inactive if tightly packed
- The basic unit is called nucleosome: DNA wrapped around 8 histone proteins

- Epigenomics:
  - Modification of DNA / histones that does not alternative the DNA sequence
  - Understand regulation of gene expression

- Single-cell ATAC-seq:
  - measure the open regions of DNA (chromatin accessibility)
  - Understand how nucleosome positioning regulates gene expression (transcriptional activation)

How is chromatin accessibility influenced?

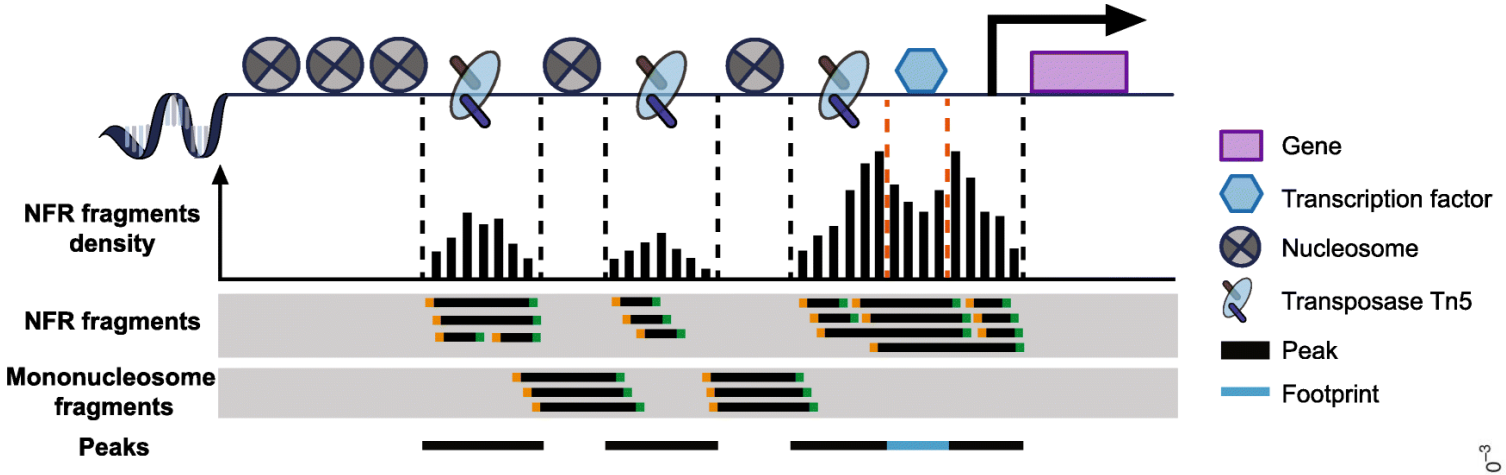


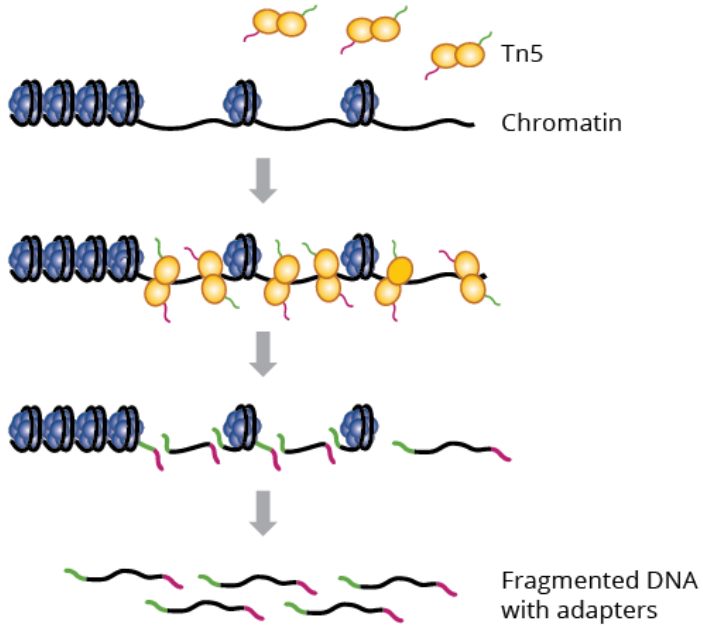
https://www.sc-best-practices.org/chromatin_accessibility/introduction.html

- Transcription factor (TF):
  - proteins that help turn specific genes "on" or "off" by binding to nearby DNA
- Promotor: a region of DNA upstream of a gene where relevant proteins such as TF bind to initiate transcription
- Enhancer: region of DNA that can be bound by proteins to increase likelihood of transcription

ATAC-seq for measuring chromatin accessiblity

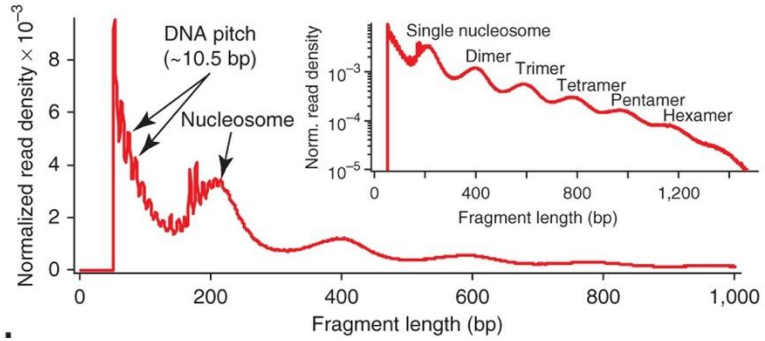
- ATAC-seq (**A**ssay for **T**ransposase-**A**ccessible **C**hromatin with high-throughput **seq**uencing) (Buenrostro et. al. Nature Methods 2015)
  - chromatin is fragmented and simultaneously tagmented with sequencing adapters using the Tn5 transposase
  - NFR fragments: represent the open chromatin nucleosome-bound fragments: reflect nucleosome position





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

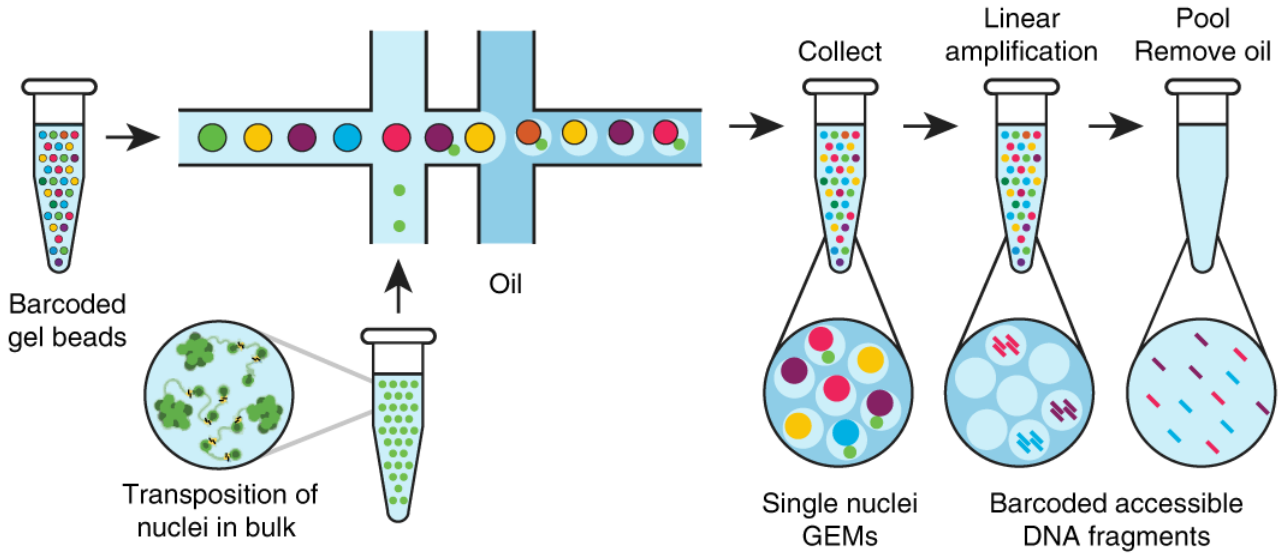
- Fragment length distribution

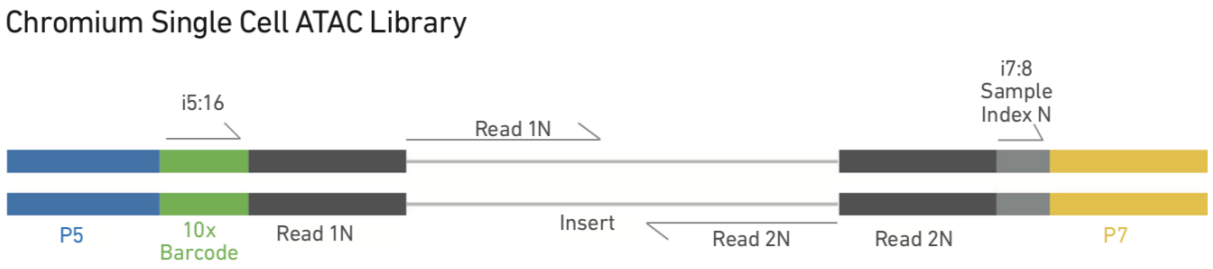


- Compared with other techniques, ATAC-seq requires few starting materials and less preparation time

scATAC-seq by 10x Genomics (Satpathy et. al., 2019)

- Nuclei are transposed (chromatin fragmented and simultaneously tagmented) in bulk before isolated in a suspension
- Transposed DNA are amplified inside each nuclei first before PCR amplification

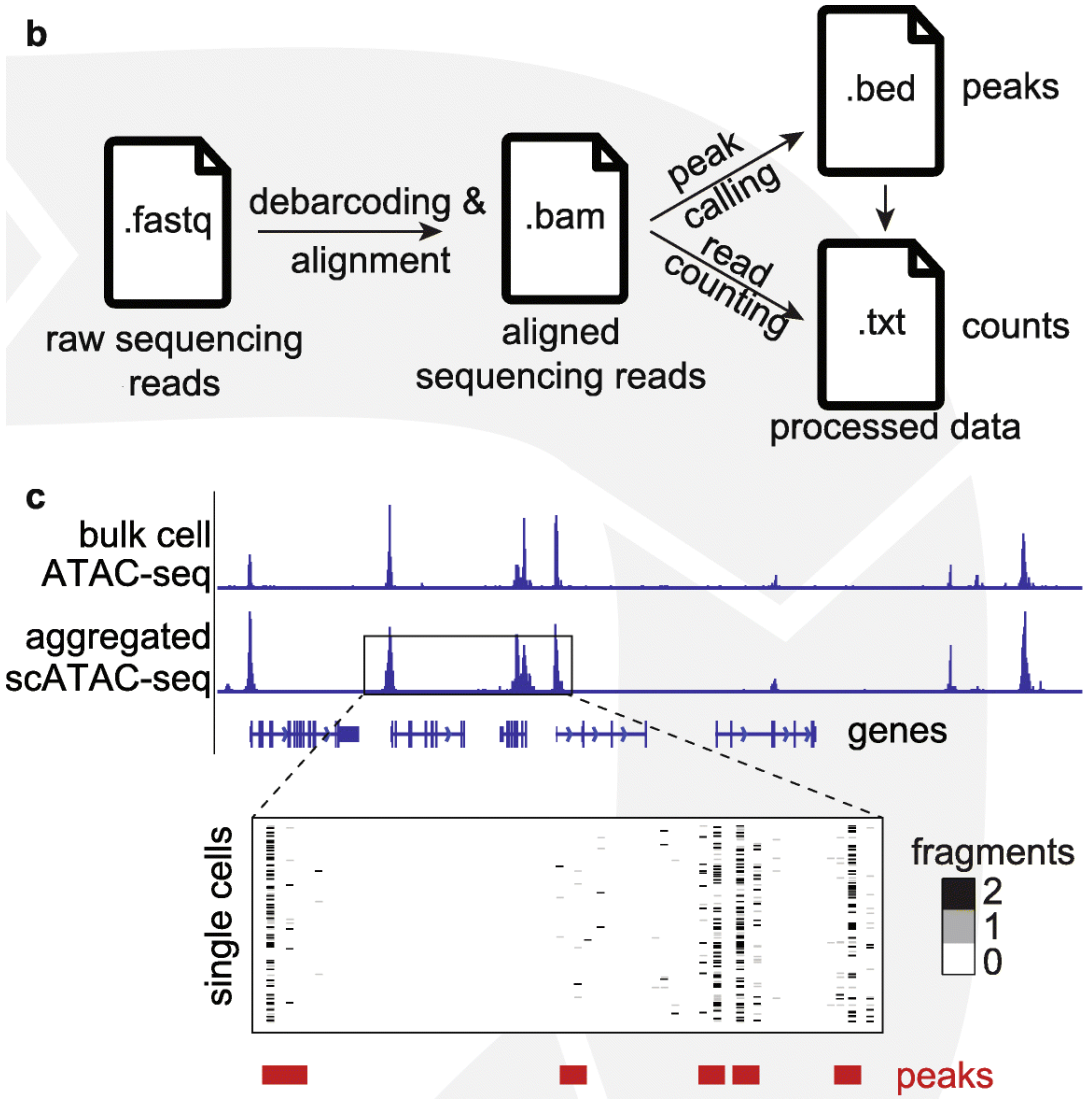




No UMI:

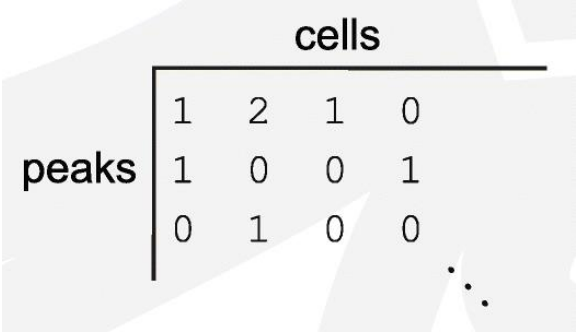
- One position on one chromatin can only be on/off
- One nuclei only have two copies of one chromosome
- Can easily remove duplicated fragments

How does scATAC-seq data look like?



- scATAC-seq preprocessing steps (Chen et. al. Genome Biology 2019)

- scATAC-seq peak by cell matrix is extremely sparse (much sparser than scRNA-seq)
  - DNA only have two copies per cell
  - 1-10% detection rate of accessible peaks
  - 10-20 times feature size than scRNA-seq
- Can have more than 2 fragments in a peak before amplification



Peak calling for scATAC-seq

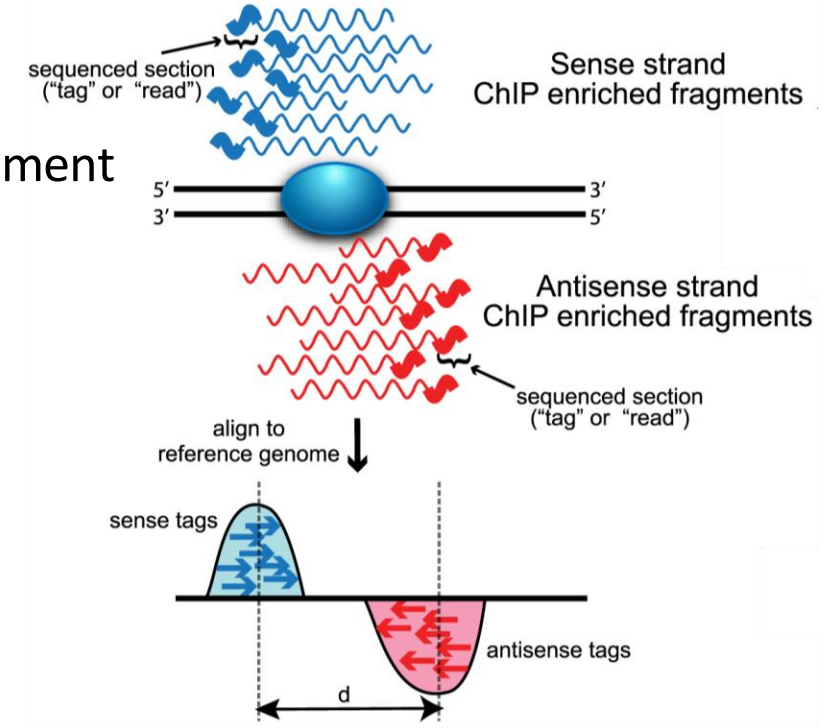
- Peak calling methods have been developed for a long time for other types of epigenetic data

- Various ways to detect peaks
  - Detect peaks based on a reference bulk ATAC-seq data
  - Detect peaks based on pseudo-bulk ATAC-seq data (ignore cell barcode to create a “bulk” dataset)
  - Perform clustering first and perform calling for each cluster of cells (SnapATAC, Fang et. al., Nature communications, 2021)
    - Aim to identify small peaks that only appear in small cell types
    - To perform clustering without peaks
      - Create cell-by-bin count matrix
        - Segment the genome into bins (5kb size by default)
        - Count the number of read in each bin and binarize the matrix

- One common method to detect peaks in MACS2

MACS2 (Zhang et. al. Genome Biology, 2008)

- Can work with both single-end reads and paired-end reads
  - scATAC-seq is paired-end, call peaks only use NFR fragments (fragment length less than 100bp)
  - Or use all reads and treat them as single-end
    - Need to recenter the reads



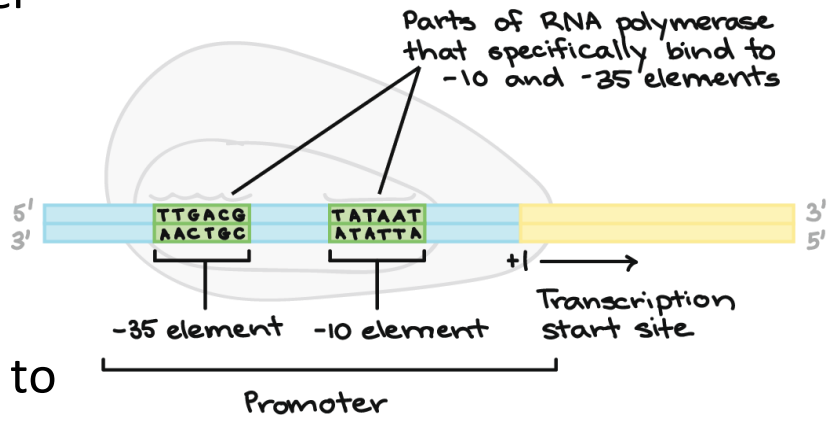
- Core steps when analyzing scATAC-seq
  - Remove duplicate reads: reads at the exact same location
  - Recenter the reads setting $d = 200$
  - Peak detection
    - Slide $2d$ window across the genome to find peaks
    - Given any window of the genome, assume number of reads follow a Poisson distribution with mean λ_{local} if there is no peak (null)

$$\lambda_{\text{local}} = \max(\lambda_{\text{BG}}, \lambda_{5\text{k}}, \lambda_{10\text{k}})$$

    - Compute a p-value for each window, selection all windows with small p-values (10^{-5})
    - Merge nearby peak regions and identify peak center as the "summit" – extend each read from its center by d so that reads can pile up

Quality control for scATAC-seq

- Detect low-quality cells
  - General metrics: total fragment counts, number of features per cell
  - Transcription starting site (TSS) enrichment
    - scATAC-seq fragments should be enriched near the TSS
      - Select a random subset of TSS
      - For each TSS, compare number of overlapping fragments (+- 2000bp window) with nearby windows to calculate an enrichment score



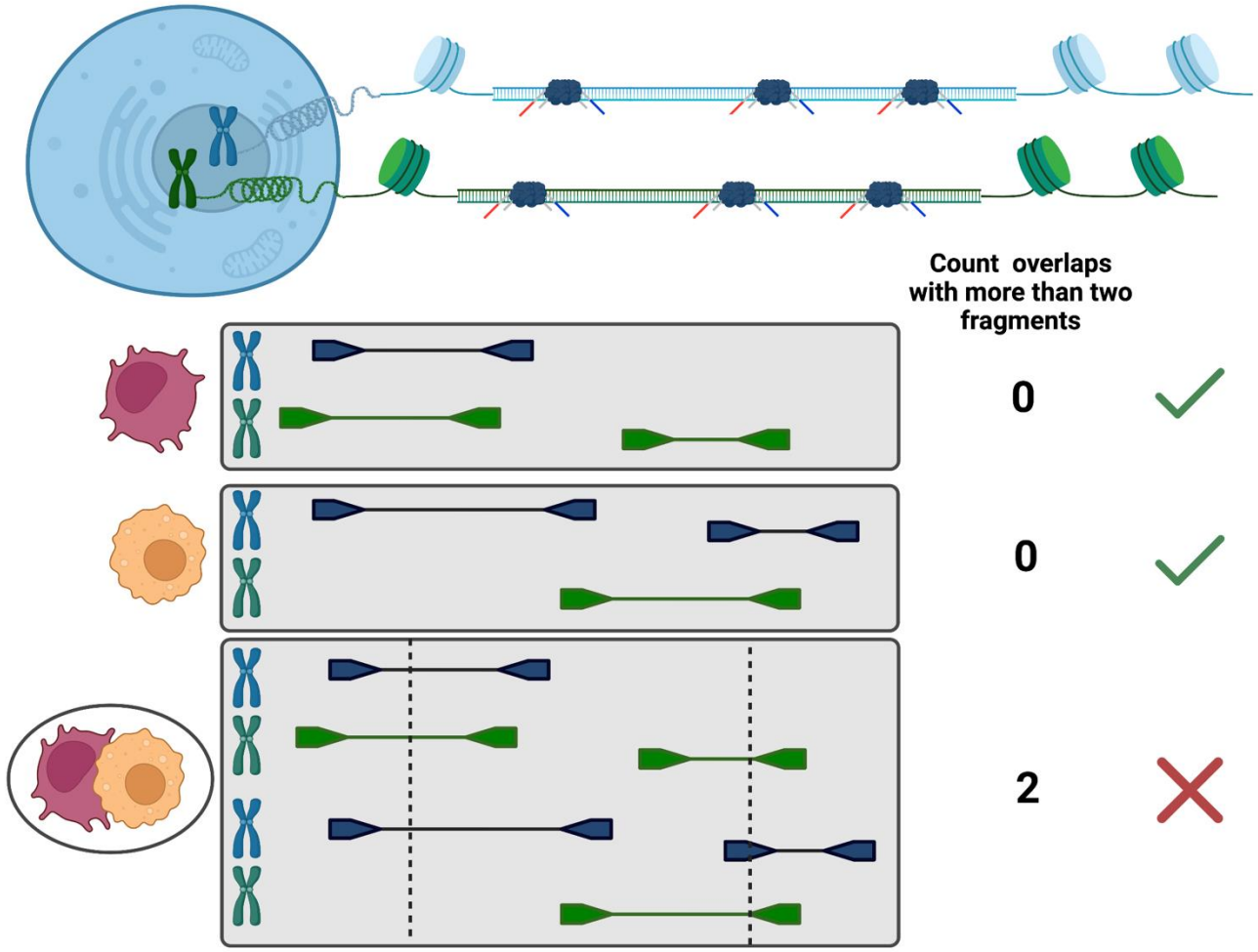
- Doublet detection
  - Much more challenging as the scATAC-seq has much higher sparsity
  - If we use similar idea as in scRNA-seq, need to aggregate correlated features (https://www.sc-best-practices.org/chromatin_accessibility/quality_control.html)
  - A different idea: At most two fragments detected per location in a single cell

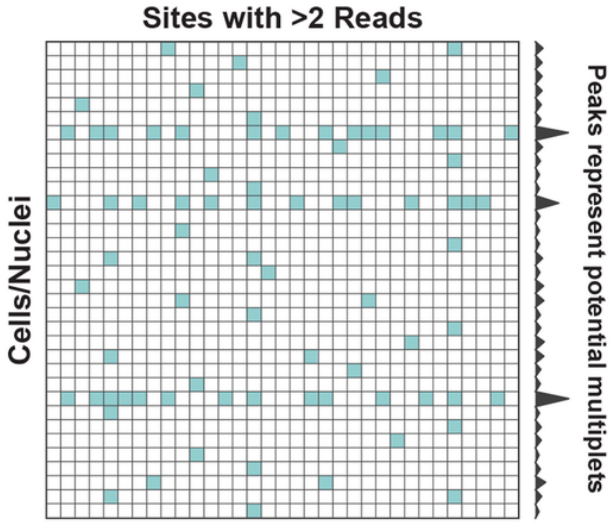
AMULET (Thibodeau et. al., Genome Biology 2021)

1. A regular cell with a diploid genome has two chromosomes that can be fragmented by Tn5 transposase

2. A regular fragment pattern. There can be maximally two fragments at the same position

3. Doublets can results in an irregular fragment pattern. We observe more than two fragments at the same position.

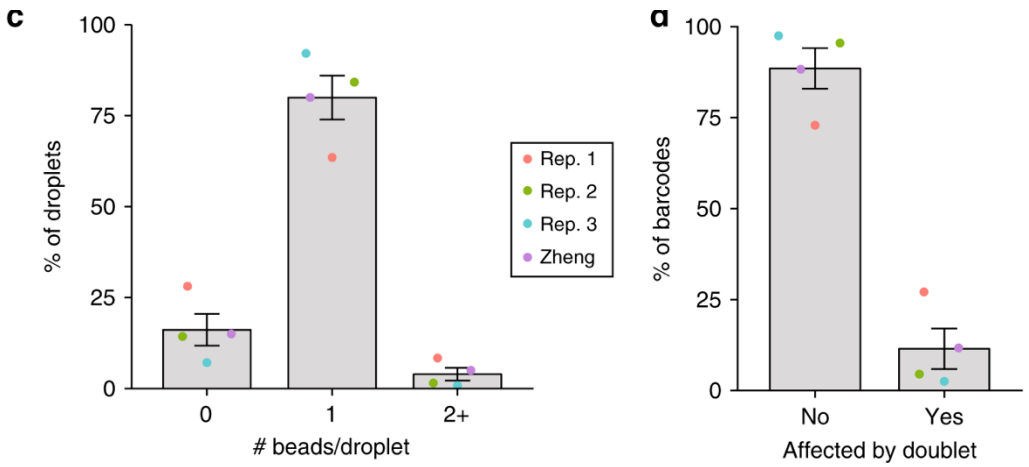


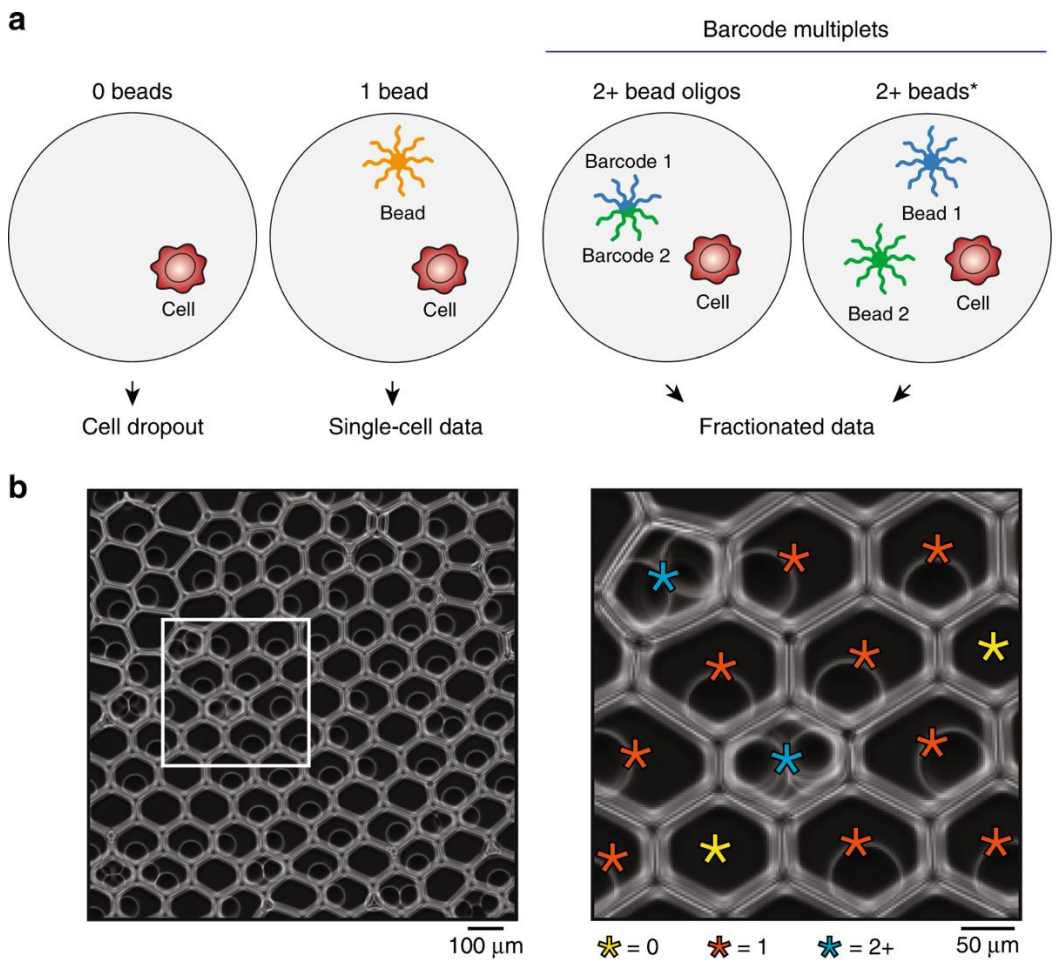


- Requires a relatively large library size
- Can identify homotypic doublets

Barcode multiplets (Lareau et. al., Nature Communications 2019)

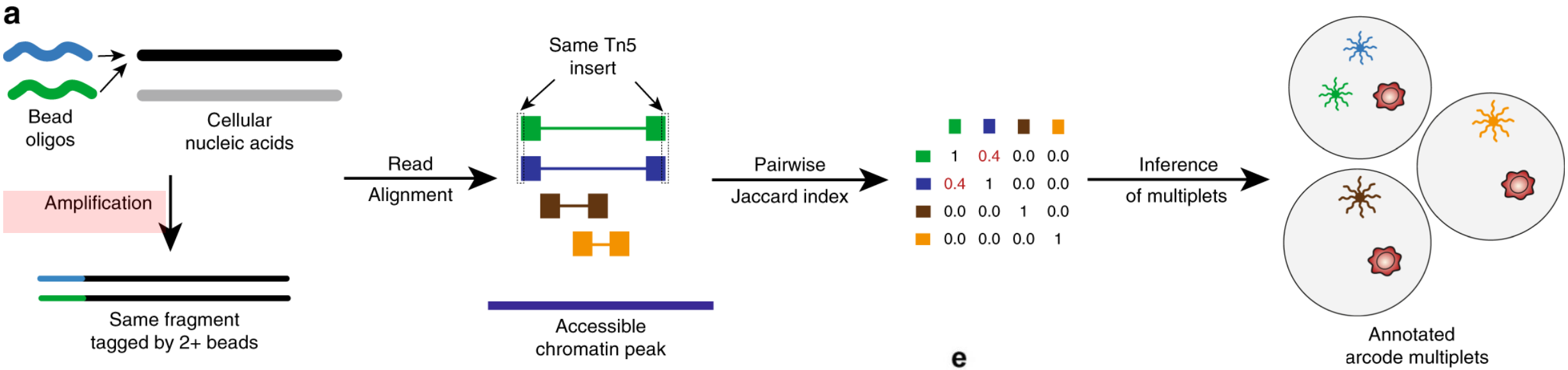
- Though the bead softgel follows super Poisson distribution in 10X genomics, there can still be barcode multiplets
  - About 5% barcode multiplets in 10X scATAC-seq (80% single-bead droplets)





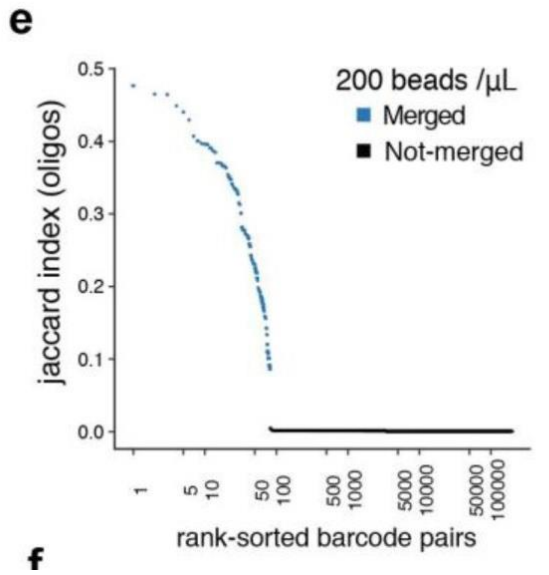
- Barcode multiplets in scRNA-seq can be challenging to detect

Barcode multiplets (Lareau et. al., Nature Communications 2019)



- Compute Jaccard index over the insertion positions of reads, providing a measure of how similar the Tn5 insertions were for any pair of bead barcodes

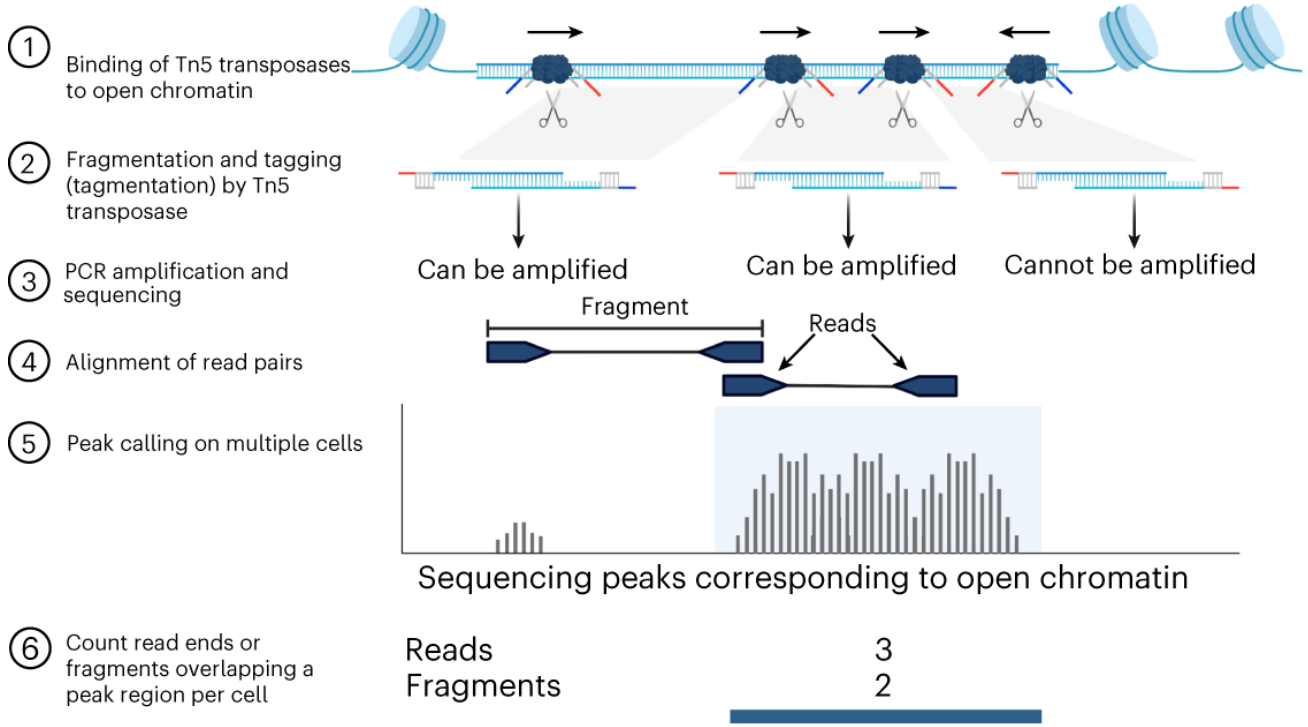
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

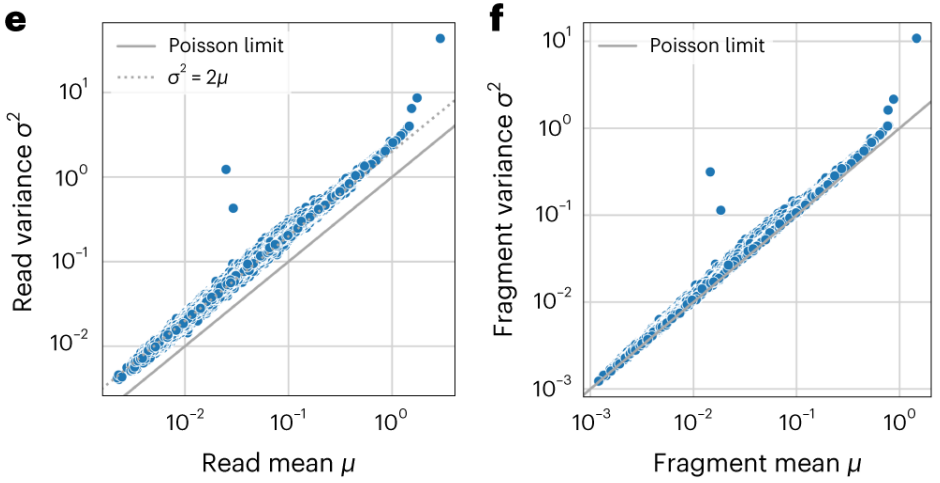


scATAC-seq count matrix (Martens et. al. Nature Methods 2024)

- Peak * cell count matrix
  - Count number of reads or number of fragments that overlap with peak region
  - Should count number of fragments to avoid bias
  - For a specific region, number of fragments follow Poisson distribution across cells

- Binarize the peak * cell matrix:
  - Entry = 1 if there are any fragments detected overlap with the peak region
  - Binarization is shown to hide quantitative information and not helpful



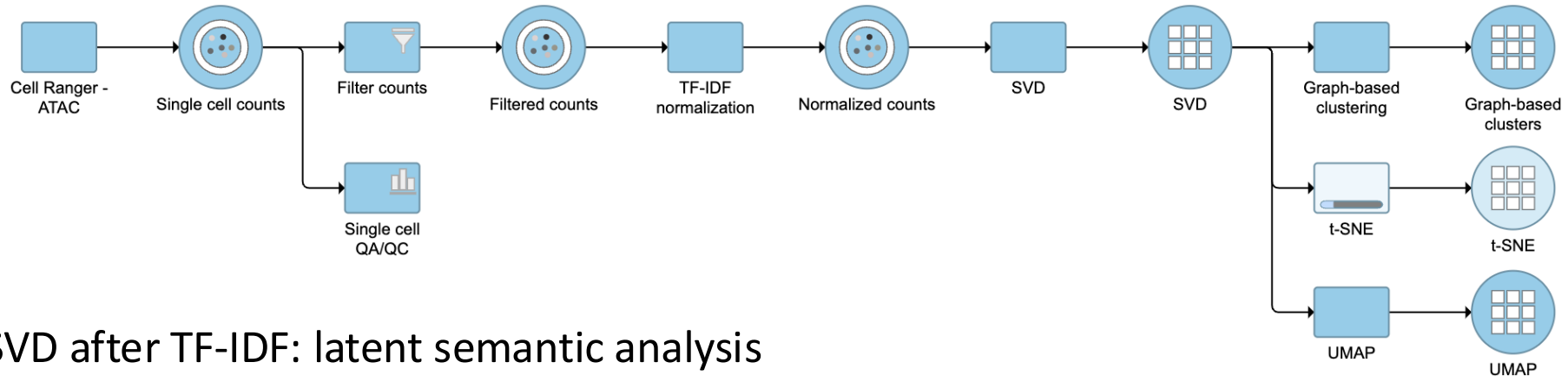


scATAC-seq normalization

- Normalize by total number of fragments per cell

- TF-IDF matrix transformation (Cusanovich et. al., Cell, 2018)
  - Normalize by gene and cell at the same time
  - Term Frequency (TF) - Inverse Document Frequency (IDF)
  - TF: total fragment normalization per cell $\mathbf{TF} = C_{ij}/F_j$
    - F_j: total number of fragments in cell j
  - IDF: $\log(1 + N/N_i)$ N_i total counts per peak across all cells
    - Can also directly use N/N_i as IDF (Stuart et. al., Nature Methods 2021)
  - TF-IDF
    - Can take log transformation if needed

$$TF\text{-}IDF = TF * IDF$$



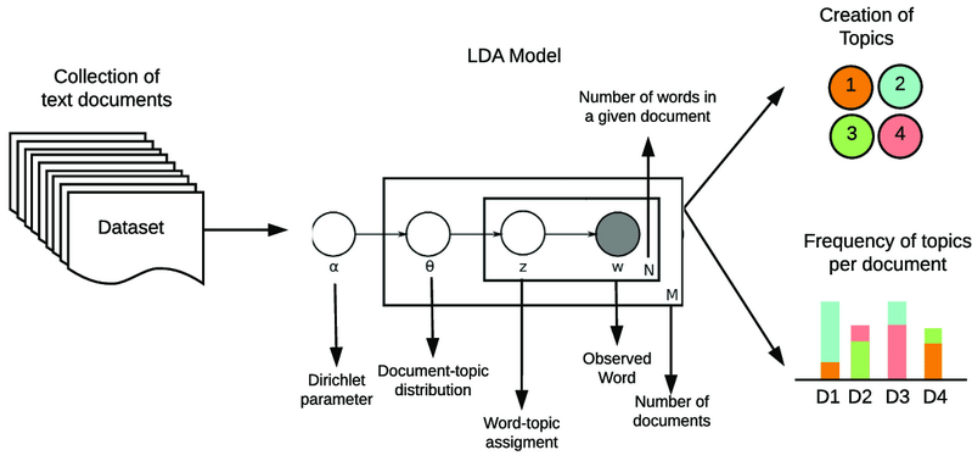
  - SVD after TF-IDF: latent semantic analysis

cisTopic (Gonzalez-Blas et. al., Nature Methods, 2019)

- Core steps:
  - Binarize the count matrix
  - Topic modeling using Latent Dirichlet Allocation (LDA, Blei et. al., JMLR 2003)
    - Generative process

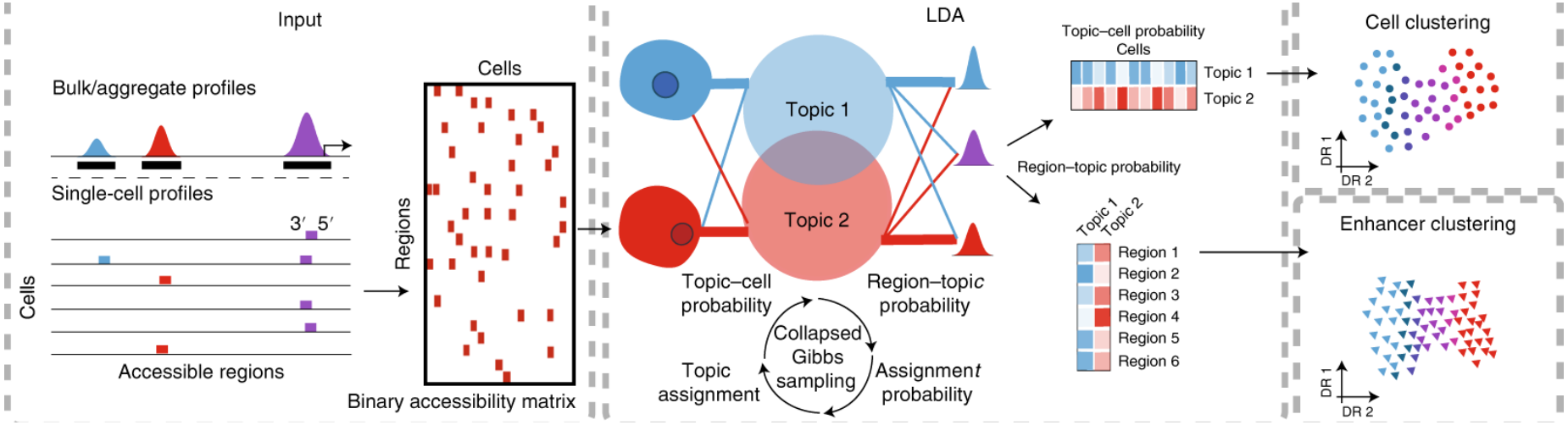
1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

      - Each position in a document independently choose a topic
      - Each topic has a topic-specific Multinomial distribution of words
    - Solve the model:
      Gibbs sampling / variational Bayes / Expectation-propagation

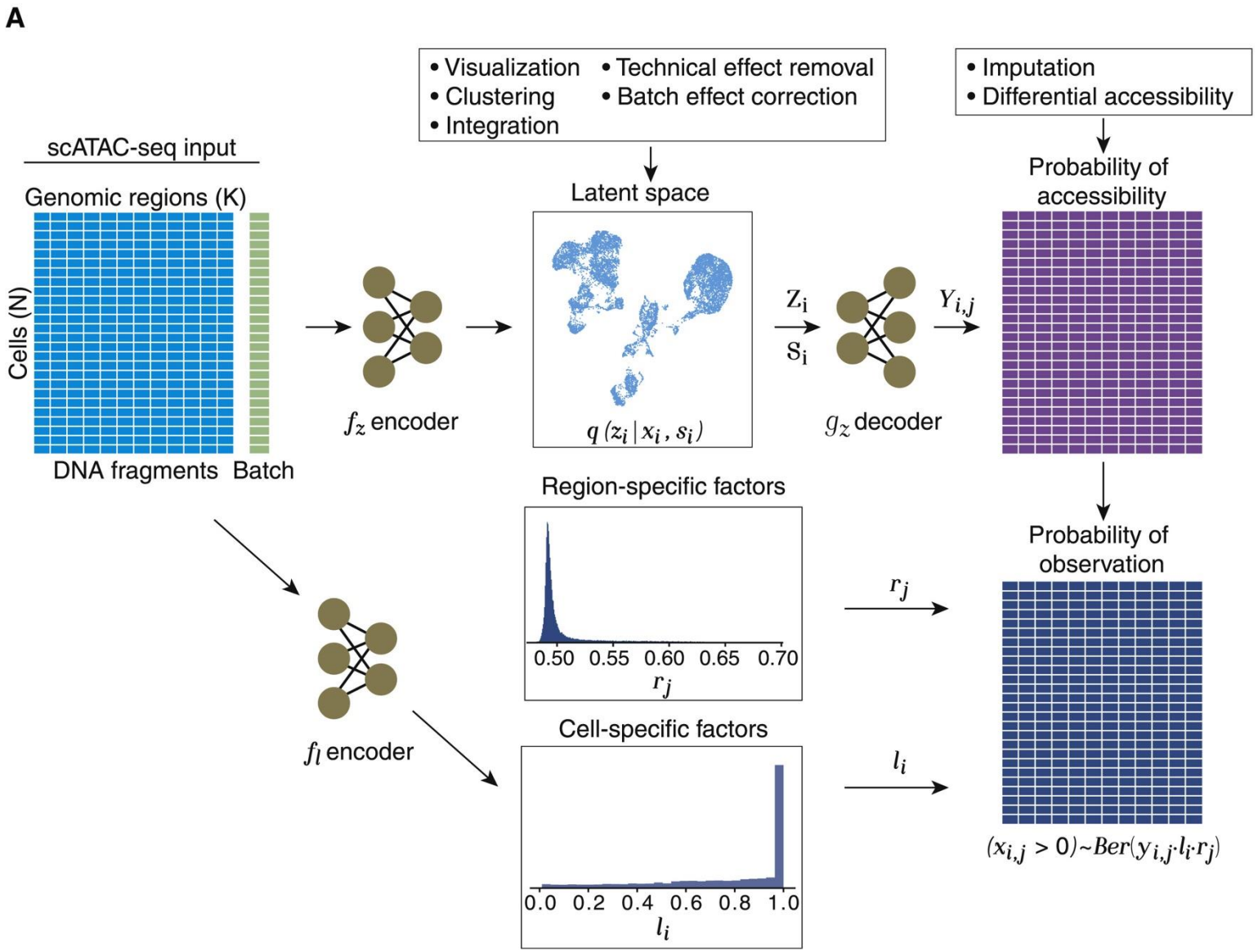


cisTopic (Gonzalez-Blas et. al., Nature Methods, 2019)

- Core steps:
  - Binarize the count matrix
  - Topic modeling using Latent Dirichlet Allocation (LDA, Blei et. al., JMLR 2003)
    - Treat each cell as a document and each region (peak) as a word
    - Use Gibbs sampler to iteratively optimize two probability distributions:
      - Region-topic distribution: the probability of a region belonging to a topic
      - Topic-cell distribution: the contribution of a topic within a cell
    - Determine the hyperparameters
      - Number of topics K: fit a model with different K, find the smallest K that stabilize the log-likelihood
      - Dirichlet distribution hyperparameters



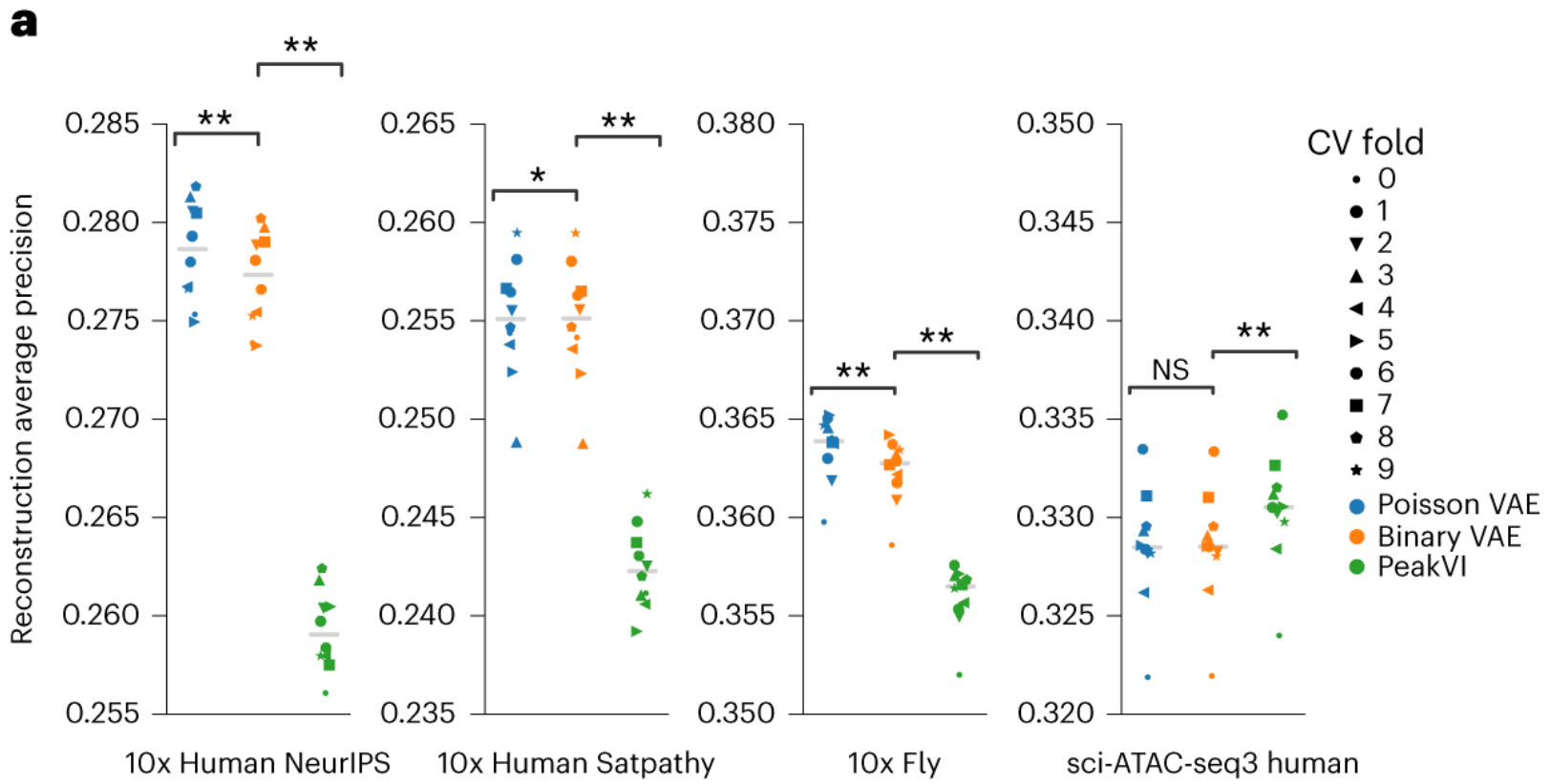
peakVI (Gonzalez-Blas et. al., Nature Methods, 2019)



- Adaptation of scVI to correct batch effects, denoising, and perform dimension reduction
- Main change: distributional assumption on count data

peakVI (Gonzalez-Blas et. al., Nature Methods, 2019)

- (Martens et. al. Nature Methods 2024) finds that using the Poisson model with counts without binarization and use observed region-specific and cell-specific factors instead of estimated factors can improve performance of peakVI



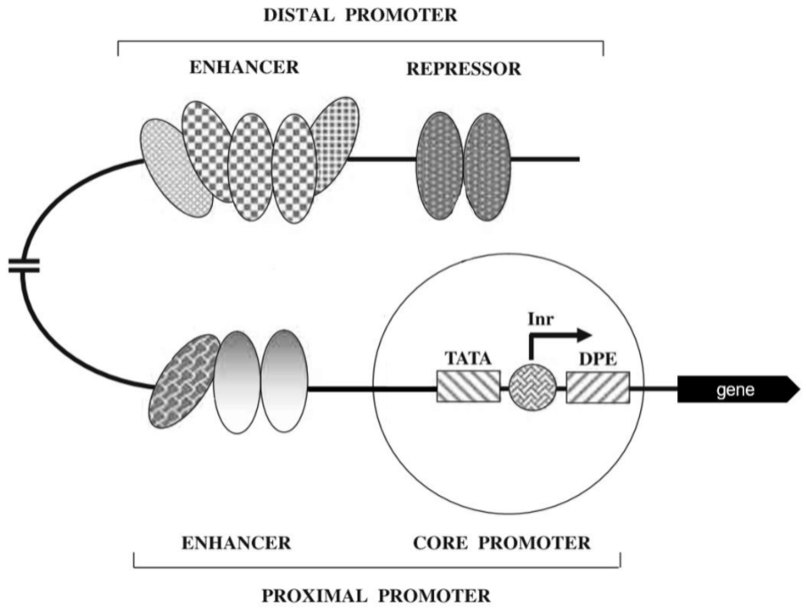
Gene activity score

- Transfer the peak * cell matrix to gene * cell matrix
  - Aggregate peaks around promoter region of a gene

- Cicero (Pliner et. al. Molecular Cell, 2018)
  - Overall measure linked to each gene k using peaks that belong to proximal or distal sites of gene TSS

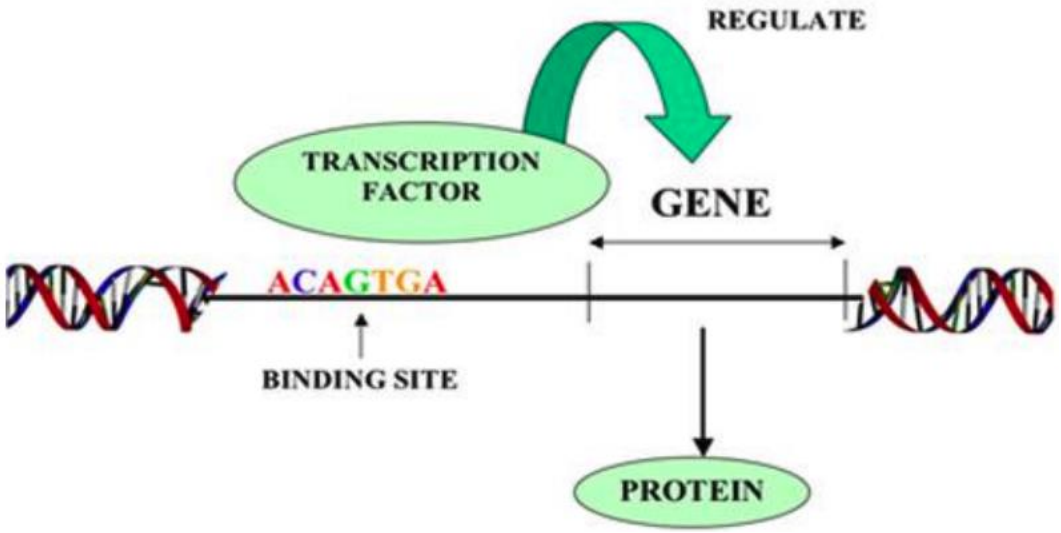
$$R_{ki} = \sum_{p \in P} \sum_{j \in D_p} A_{ji} \frac{u_{pj}}{\sum_{k \in D_p} u_{pk}} + A_{pi}$$

Where P indexes the promoter proximal sites of k, D_p indexes distal sites linked to proximal site p, and u is the Cicero co-accessibility score linking distal site j to proximal site p, and A is the binary score for accessibility at site j or p in cell i. In principle, D_p could include all distal sites linked to p, but here we restrict the set to distal sites that are differentially accessible (FDR< 1%) across pseudotime.



- Signac (Stuart et. al. Nature Methods, 2018)
  - Count number of fragments overlapping the gene body and a 2-kb upstream region for each gene in each cell
- Apply scRNA-seq methods on gene activity score matrix

Transcriptional factor activity matrix



- Motifs: DNA biding sites (has a specific structure)

- ChromVAR (Schep et. al., Nature Methods 2017)
- Motif enrichment of each cell
  - Motif matching matrix W: motif by peak matrix
  - for a list of motifs, calculate the frequency of each motif within any peak regions

$$Y = \frac{M \times X^T - M \times E^T}{M \times E^T} \qquad E = \frac{\Sigma_{i=1} x_{i,j}}{\Sigma_{j=1}\Sigma_{i=1} x_{i,j}} \times \sum_{j=1} x_{i,j}$$

    - Can adjust for other peaks that contain similar motifs (background peaks) to adjust for local bias
- Transcriptional factor activity of each cell
  - For each TF, select a representative subset of motifs

Related papers

- Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., McDermott, G. P., ... & Chang, H. Y. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nature biotechnology, 37(8), 925-936.
- Chen, H., Lareau, C., Andreani, T., Vinyard, M. E., Garcia, S. P., Clement, K., ... & Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome biology, 20, 1-25.
- Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., ... & Ren, B. (2021). Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nature communications, 12(1), 1337.
- Feng, J., Liu, T., Qin, B., Zhang, Y., & Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. Nature protocols, 7(9), 1728-1740.
- Thibodeau, A., Eroglu, A., McGinnis, C. S., Lawlor, N., Nehar-Belaid, D., Kursawe, R., ... & Ucar, D. (2021). AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. Genome biology, 22, 1-19.
- Lareau, C. A., Ma, S., Duarte, F. M., & Buenrostro, J. D. (2020). Inference and effects of barcode multiplets in droplet-based single-cell assays. Nature Communications, 11(1), 866.
- Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., ... & Shendure, J. (2018). A single-cell atlas of in vivo mammalian chromatin accessibility. Cell, 174(5), 1309-1324.
- Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., ... & Aerts, S. (2019). cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nature methods, 16(5), 397-400.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
- Ashuach, T., Reidenbach, D. A., Gayoso, A., & Yosef, N. (2022). PeakVI: A deep generative model for single-cell chromatin accessibility analysis. Cell reports methods, 2(3).
- Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., ... & Trapnell, C. (2018). Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Molecular cell, 71(5), 858-871.
- Stuart, T., Srivastava, A., Madad, S., Lareau, C. A., & Satija, R. (2021). Single-cell chromatin state analysis with Signac. Nature methods, 18(11), 1333-1341.
- Schep, A. N., Wu, B., Buenrostro, J. D., & Greenleaf, W. J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nature methods, 14(10), 975-978.