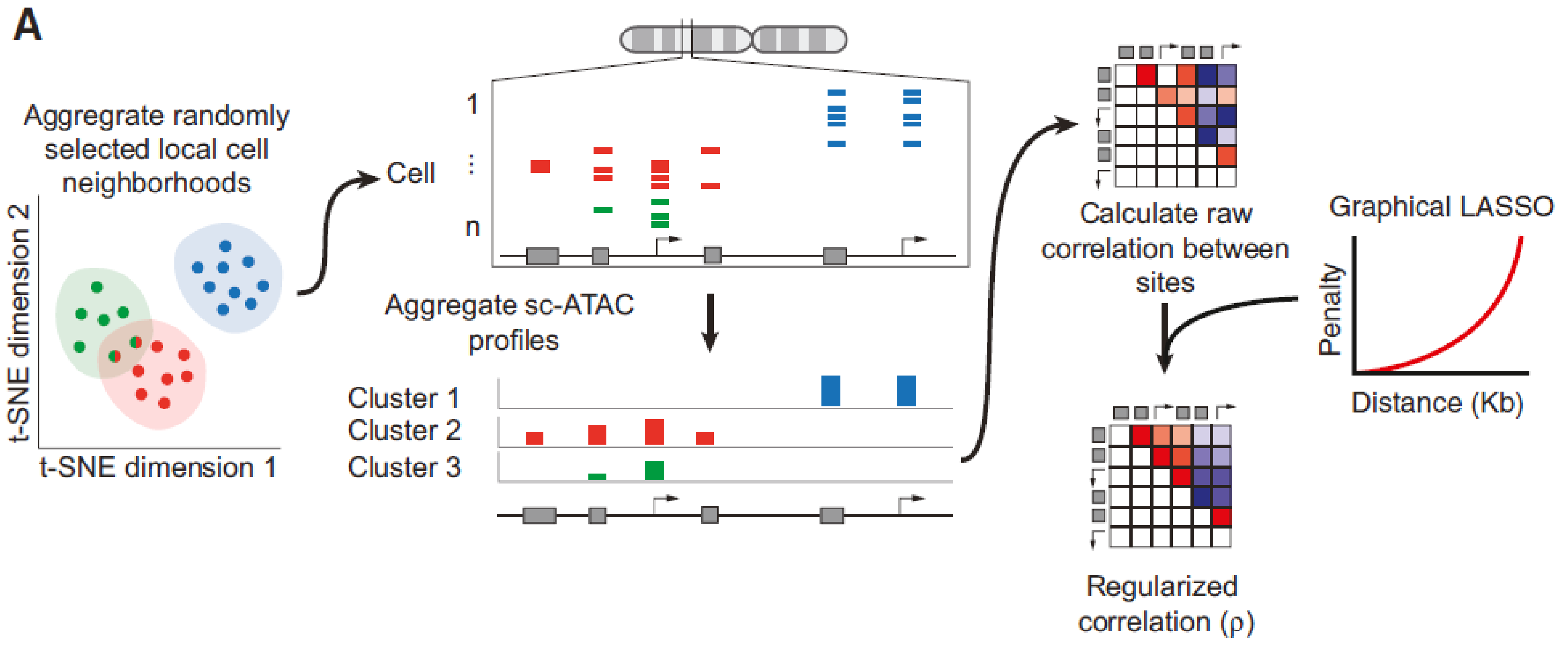
scATAC-seq

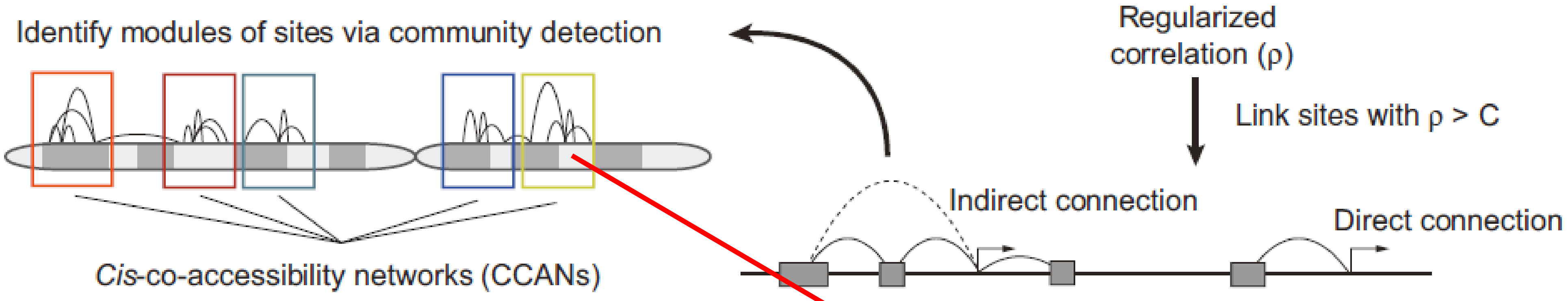
- Cicero (Co-accessible region analysis)



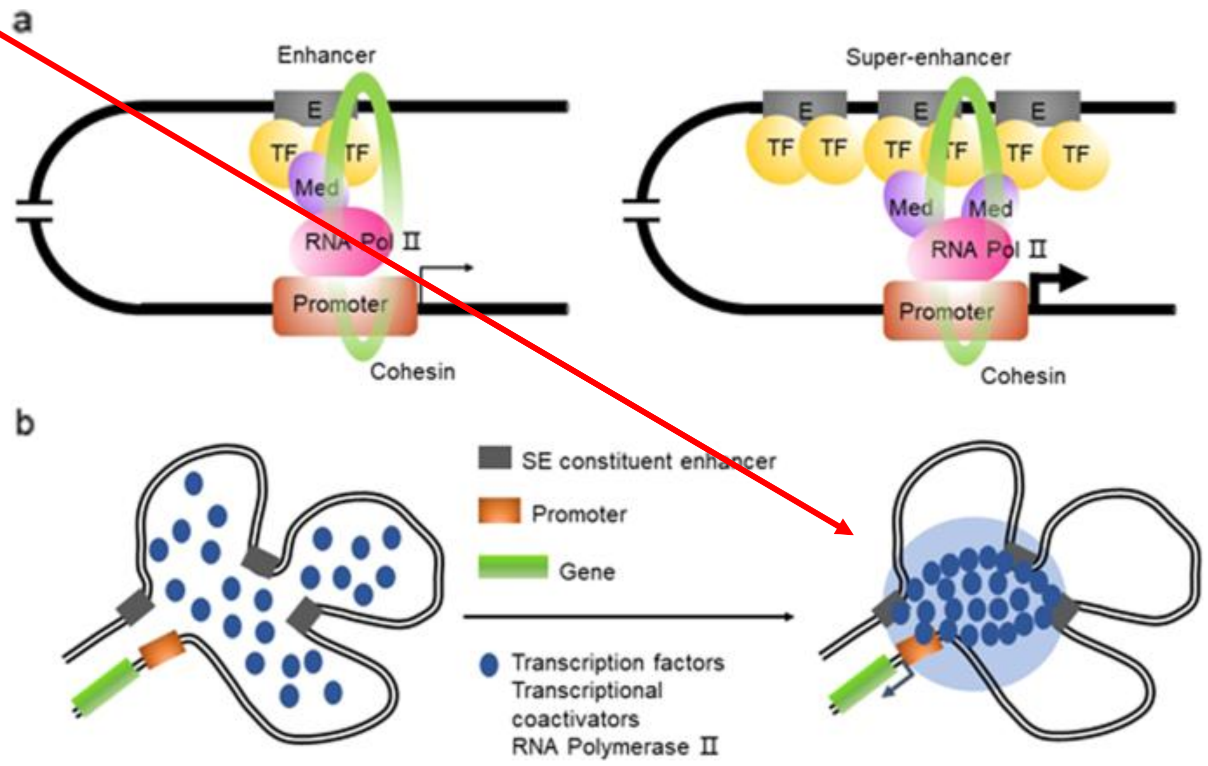
- cell → pseudotemoral ordering (merging 50 cells)
- Raw correlation (within 500kb): LASSO (distance penalty), Regularized correlation
-Sparse covariance matrix of each pair of peaks (within 500kb)
→ Correlation (if it agrees with a qualitative agreement)

Sanity check: distal site ~ gene activity or gene expression

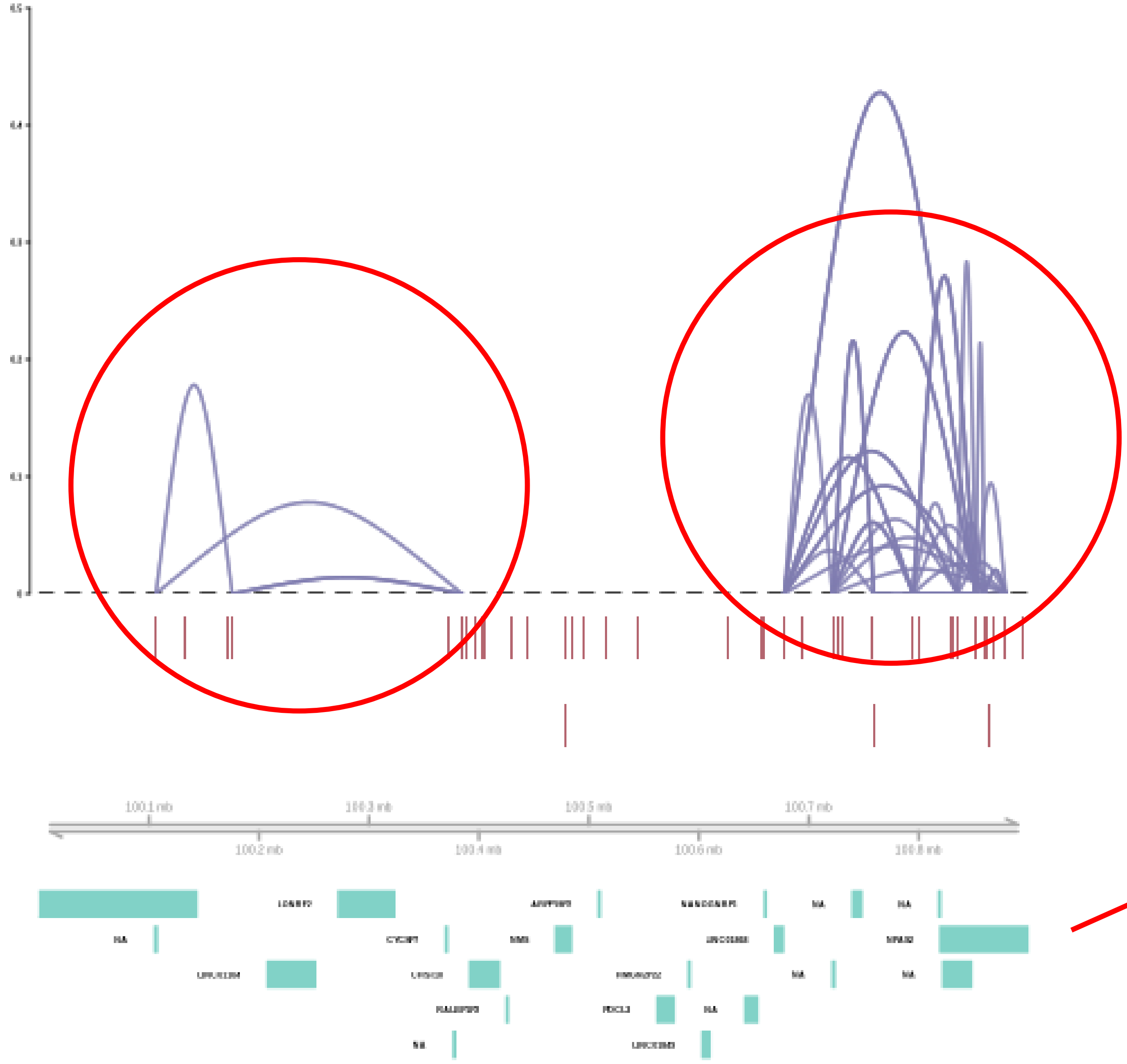
- # Cicero (Co-accessible region analysis)



-CCAN (cis-co-accessibility network)
→ Co-accessible score (edge)
→ Thresholding → graph → Louvain clustering
→ chromatin hub (looping interactions)
1: **common protein complex** (TFs)
2: same epigenetic modification
3: substantially regulates promoters
4: proximity

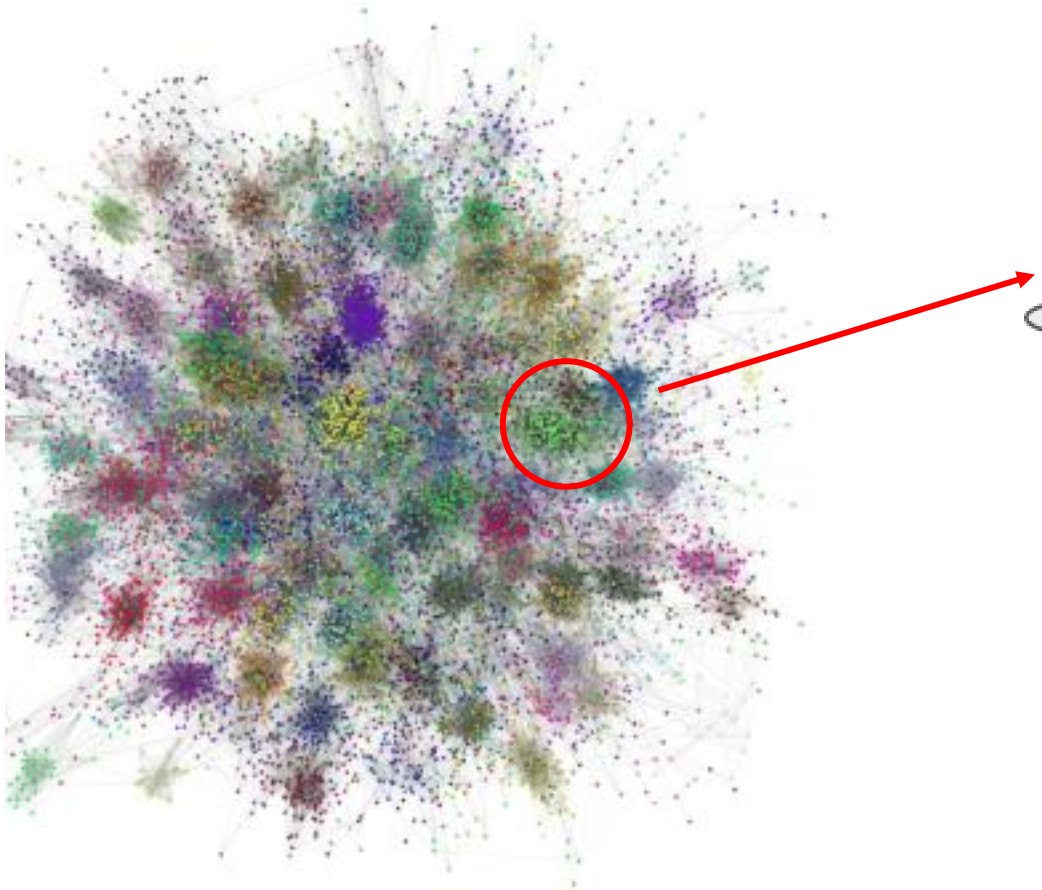


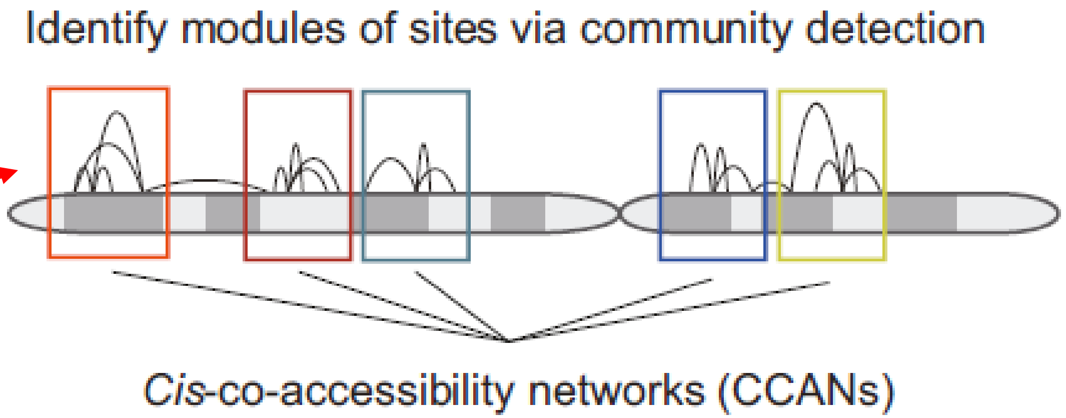
- Cicero (Co-accessible region analysis)



Corresponding genes

- Cicero (Co-accessible region analysis)

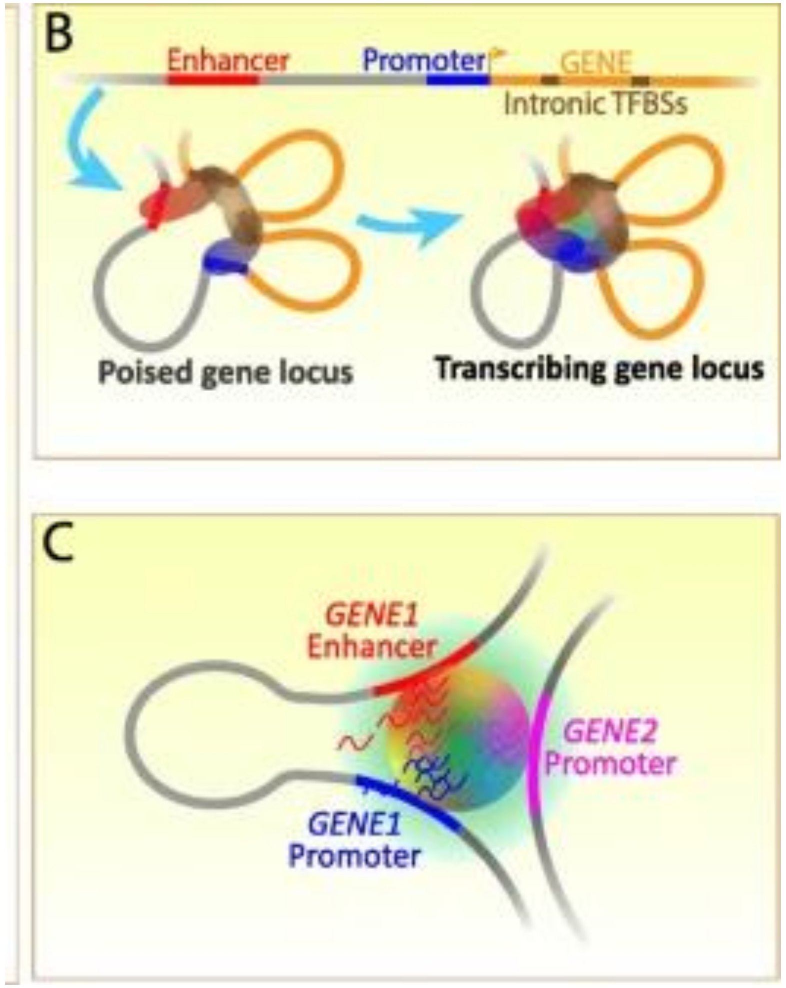


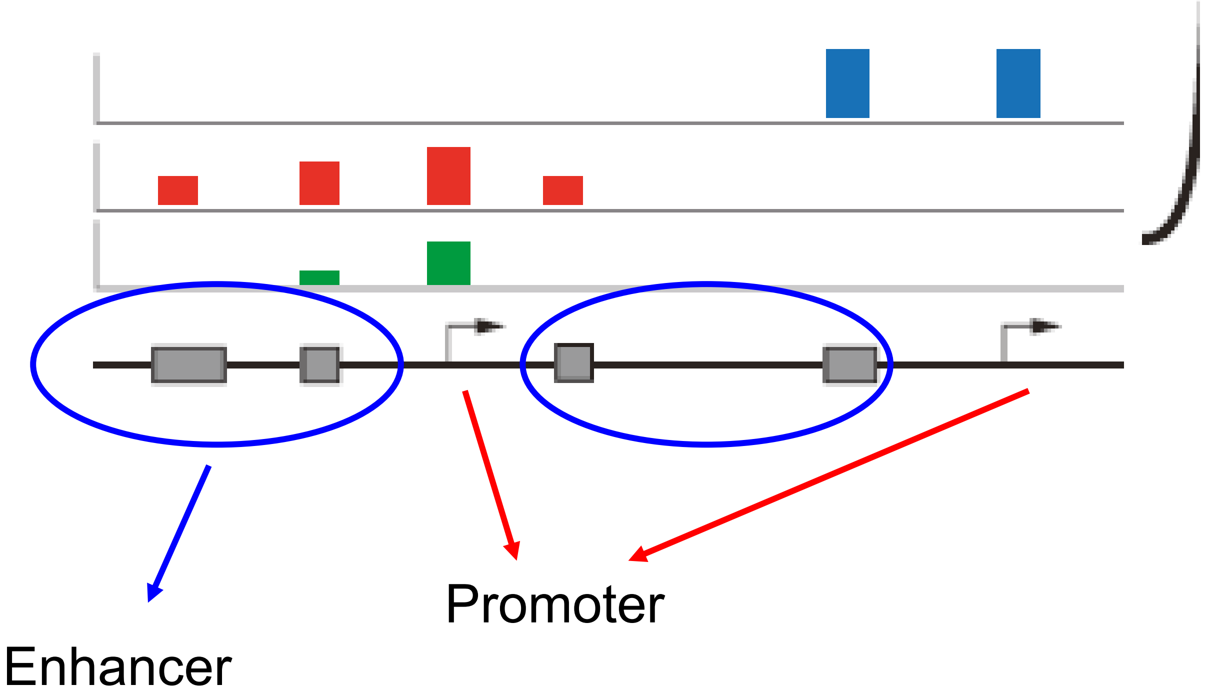


-CCAN → gene regulation module → gene expression network

Gene expression network

- Cicero (Co-accessible region analysis)





• Cicero (gene activity)

-find_overlapping_coordinates(fData(temp)$site_name, "chr1:3,000,200-3,090,000")

→ Coaccessible region → correlates with gene expression

→ Combine the scores of regional accessibility to a specific gene → infer gene activity

Cicero gene activity scores

We have found that often accessibility at promoters is a poor predictor of gene expression. However, using Cicero links, we are able to get a better sense of the overall accessibility of a promoter and it's associated distal sites. This combined score of regional accessibility has a better concordance with gene expression. We call this score the Cicero gene activity score, and it is calculated using two functions.

The initial function is called `build_gene_activity_matrix`. This function takes an input CDS and a Cicero connection list, and outputs an unnormalized table of gene activity scores. **IMPORTANT**: the input CDS must have a column in the fData table called "gene" which indicates the gene if that peak is a promoter, and `NA` if the peak is distal. One way to add this column is demonstrated below.

The output of `build_gene_activity_matrix` is **unnormalized.** It must be normalized using a second function called `normalize_gene_activities`. If you intend to compare gene activities across different datasets of subsets of data, then all gene activity subsets should be normalized together, by passing in a list of unnormalized matrices. If you only wish to normalized one matrix, simply pass it to the function on its own. `normalize_gene_activities` also requires a named vector of of total accessible sites per cell. This is easily found in the pData table of your CDS, called "num_genes_expressed". See below for an example. Normalized gene activity scores range from 0 to 1.

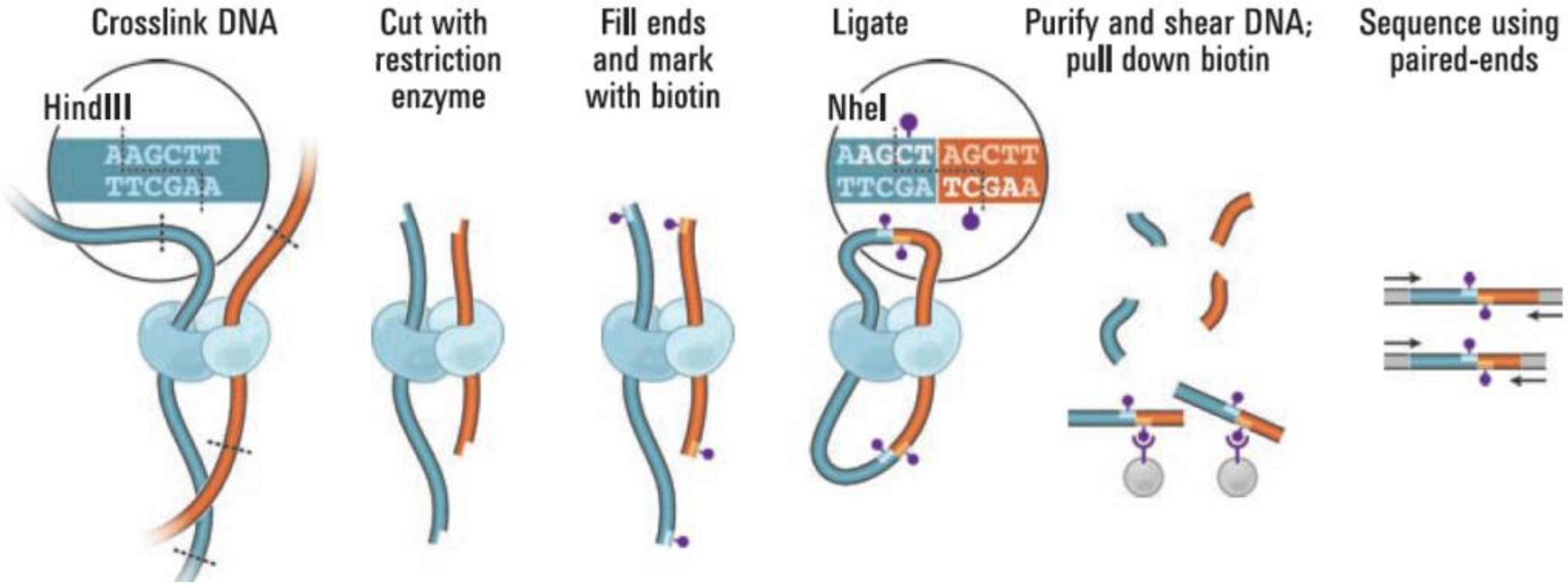
- No accessible region

→ Signac: based on the typical promoter (or gene body) region

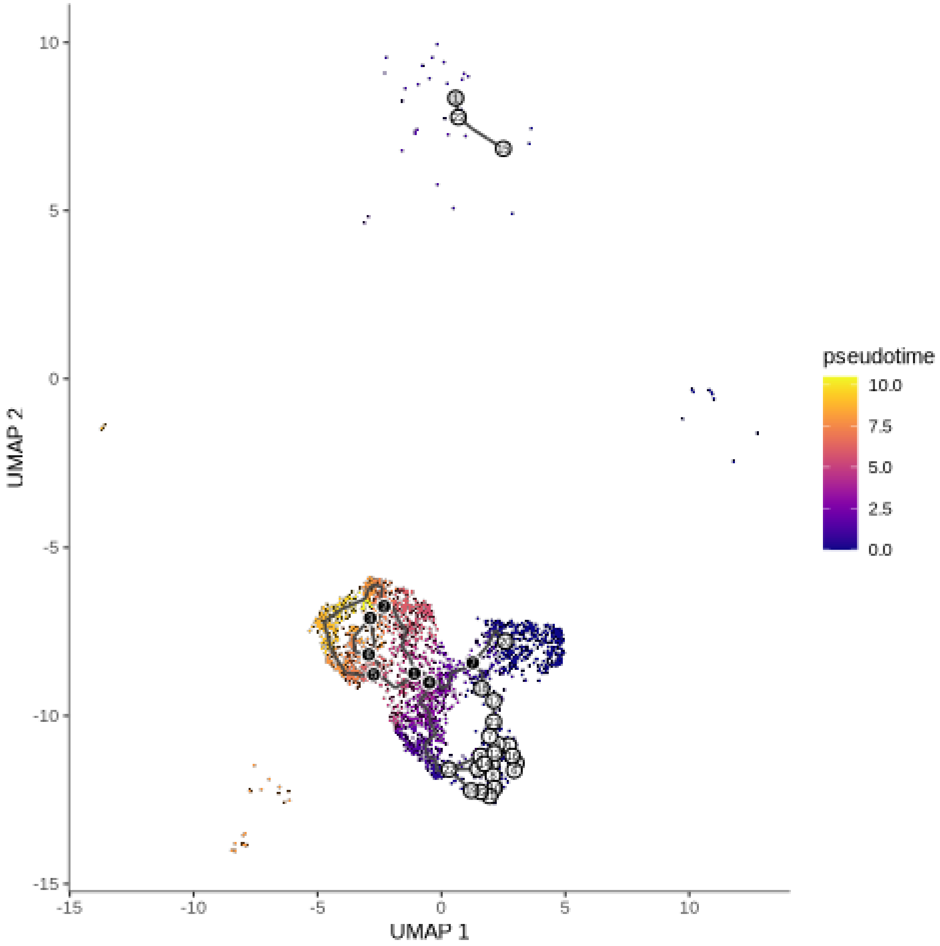
- Cicero (gene activity)

-However, coaccessible analysis is just inferrence
→ Prone to False positive
→ Arbitrary threshold, imperfect algorithm, poor data (sparsity)

Hi-C sequencing



- Monocle3: trajectory analysis



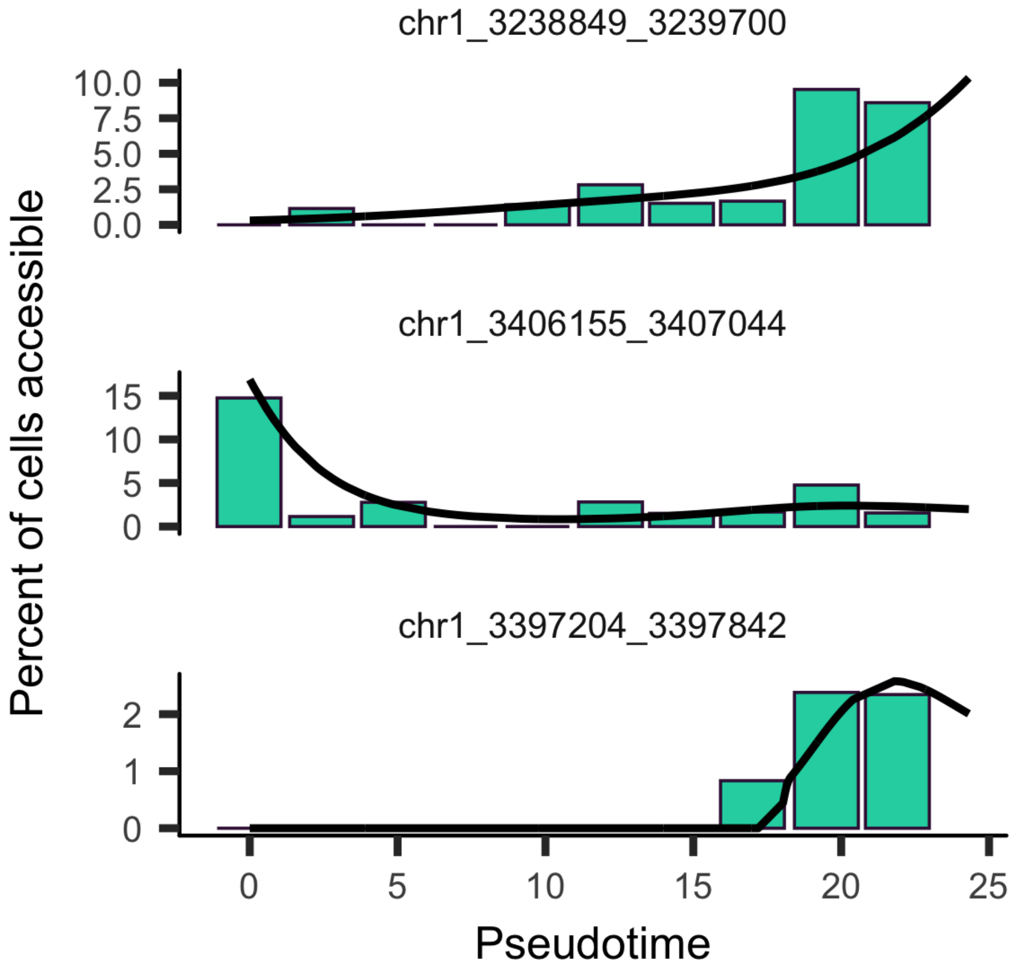
-Feature: it does not have to be gene activity
→ mathematically, it does not matter

- Monocle3 & Cicero

object <- monocle (aligned alongside pseudotime)

→ Differential path associated with epigenome

→ Alteration of accessible region across pseudotime



- Monocle3 & Cicero

- fit_models: regression analysis btw pseudotime

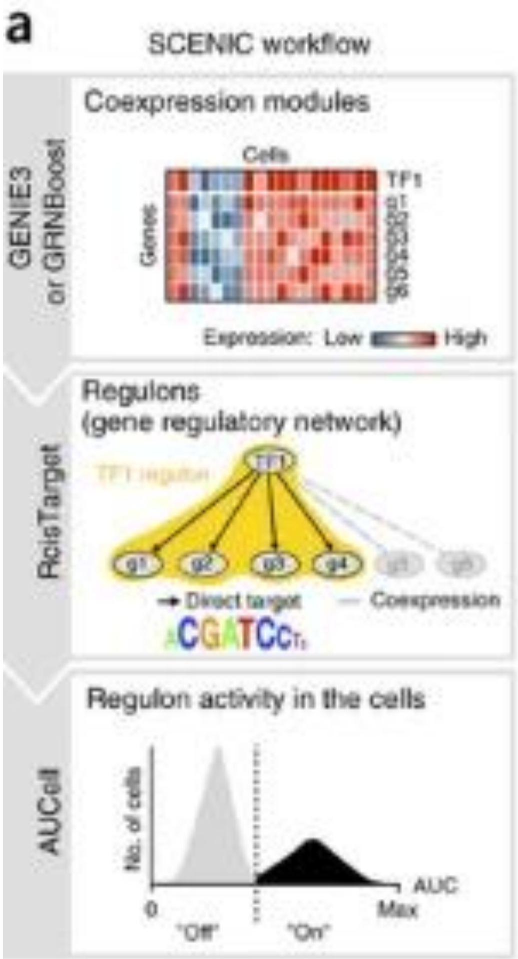
→ Generalized linear model (you can put your own formula)

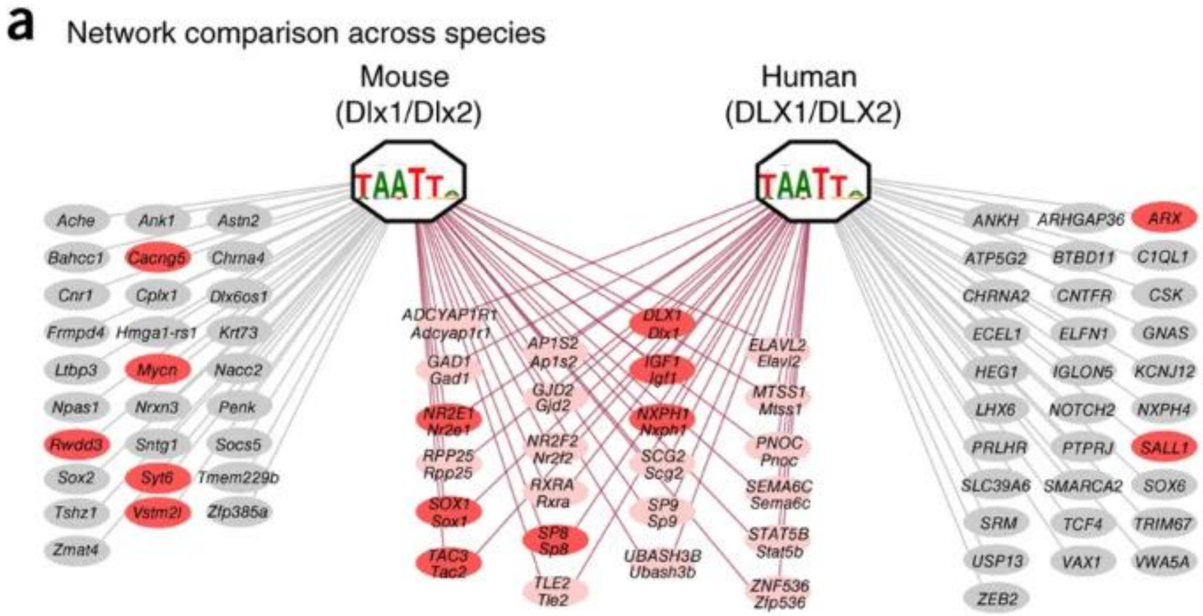
$$log(y_i) = \beta_0 + \beta_t x_t$$

- aggregate_by_cell_bin: reduce computational cost + overcome higher sparsity (binning; default:10)

```
> head(b)
                site_name num_cells_expressed use_for_ordering
1 chr2-200603380-200604940                   6            FALSE
2 chr2-200603380-200604940                   6            FALSE
3 chr2-200603380-200604940                   6            FALSE
4 chr1-221356415-221356913                   3            FALSE
5 chr1-221356415-221356913                   3            FALSE
6 chr1-221356415-221356913                   3            FALSE
                  gene_id        model model_summary status
1 chr2-200603380-200604940 c(`(Inte....  speedglm....     OK
2 chr2-200603380-200604940 c(`(Inte....  speedglm....     OK
3 chr2-200603380-200604940 c(`(Inte....  speedglm....     OK
4 chr1-221356415-221356913 c(`(Inte....  speedglm....     OK
5 chr1-221356415-221356913 c(`(Inte....  speedglm....     OK
6 chr1-221356415-221356913 c(`(Inte....  speedglm....     OK
                term      estimate      std_err   test_val   p_value
1        (Intercept)  3.765692e+00 2.901629e+00  1.2977853 0.2354868
2         Pseudotime -2.391297e-01 1.611598e-01 -1.4838053 0.1814313
3 num_genes_expressed -2.524447e-05 3.567382e-05 -0.7076468 0.5020386
4        (Intercept) -2.561857e+00 6.263759e+00 -0.4089966 0.6947660
5         Pseudotime  2.659335e-02 2.604892e-01  0.1020900 0.9215481
6 num_genes_expressed  2.840407e-05 7.688879e-05  0.3694175 0.7227384
  normalized_effect model_component q_value
1      0.000000e+00           count       1
2     -3.449011e-01           count       1
3     -3.641164e-05           count       1
4      0.000000e+00           count       1
5      3.401582e-02           count       1
6      3.627702e-05           count       1
```

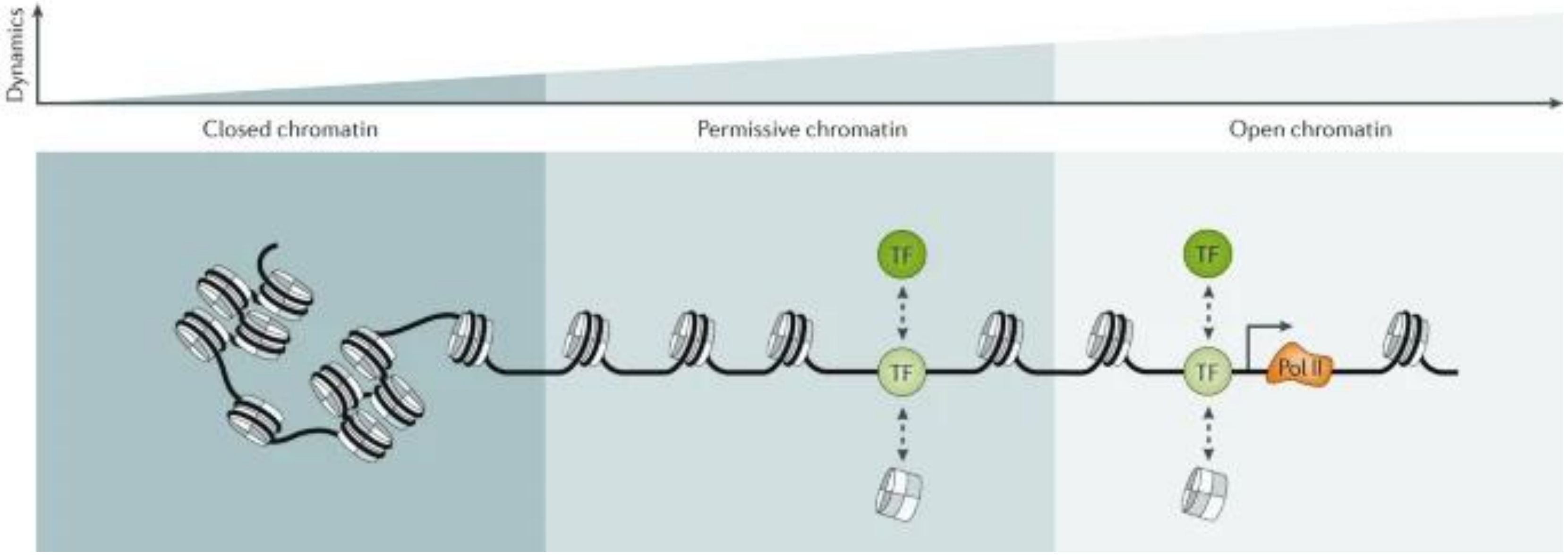
- SCENIC+ (gene regulatory network with open chromatin region)

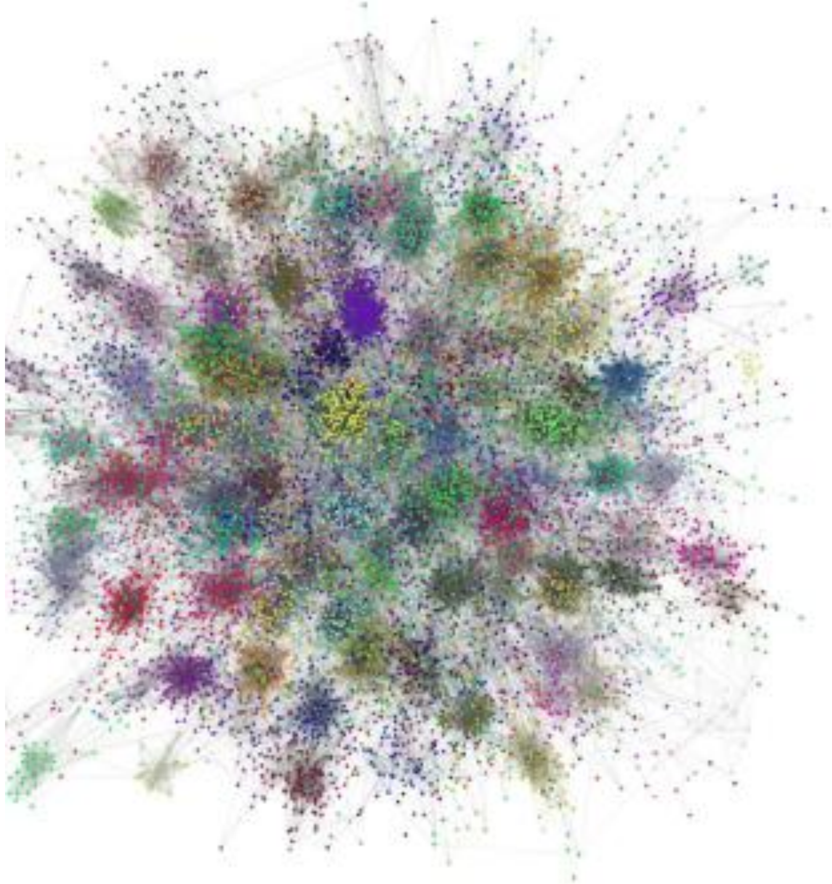




-SCENIC from scRNA-seq
All the TF that express can be a candidate

- SCENIC+ (gene regulatory network with open chromatin region)

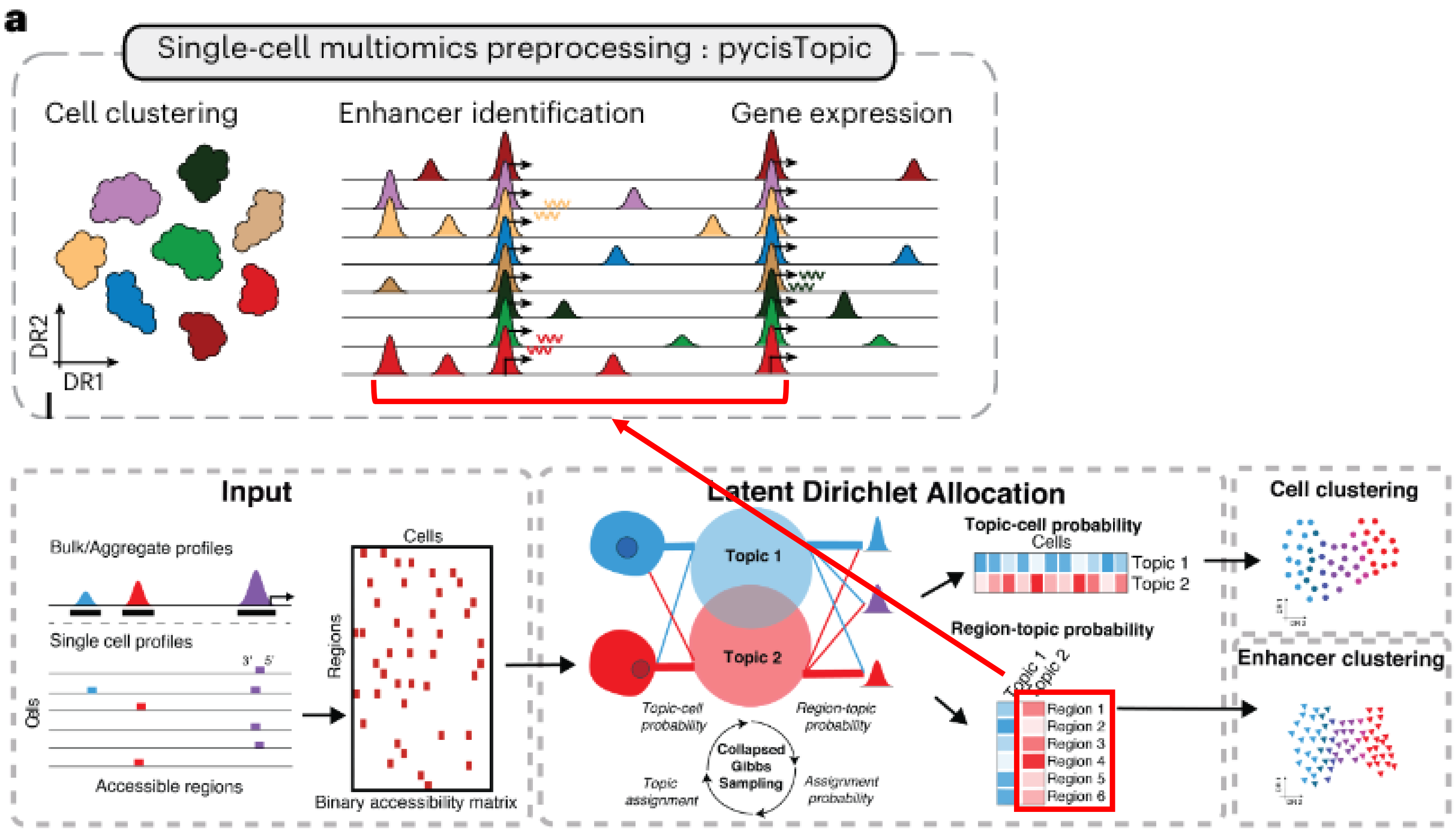




Gene regulatory network

-TF should bind to the promoter to regulate the gene expression
→ The promoter region must be opened
→ ATAC: assess the open region whether TF can really bind
→ Search for the TF motif in open regions of the promoter
+ Enhancer regions (coaccessibility + motif)

• SCENIC+ (gene regulatory network with open chromatin region)



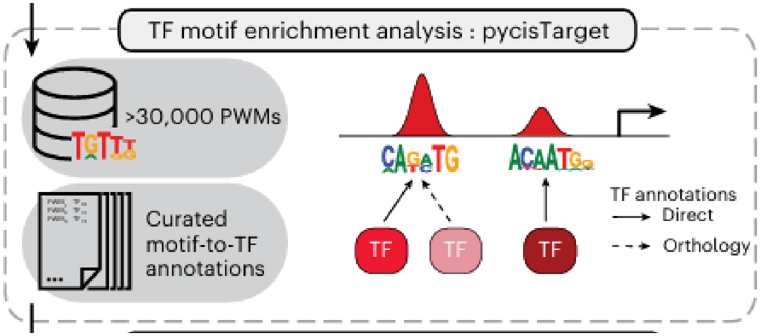
-Sparse peak matrix → text-mining algorithm (like TF-IDF)
-Cell~region (peak) matrix → LDA → merge regions into "Topic" (similar to NMF)
-Cell cluster (by topic) → which cluster has which topic → which topic has which enhancer
→ cell: which enhancer (enhancer identification)

• SCENIC+ (gene regulatory network with open chromatin region)

-pycisTarget: largest motif DB → motif enrichment analysis
→ Motif to TF

From the obtained enhancer list
→ Motif enrichment
→ Which TF is binding to a given enhancer

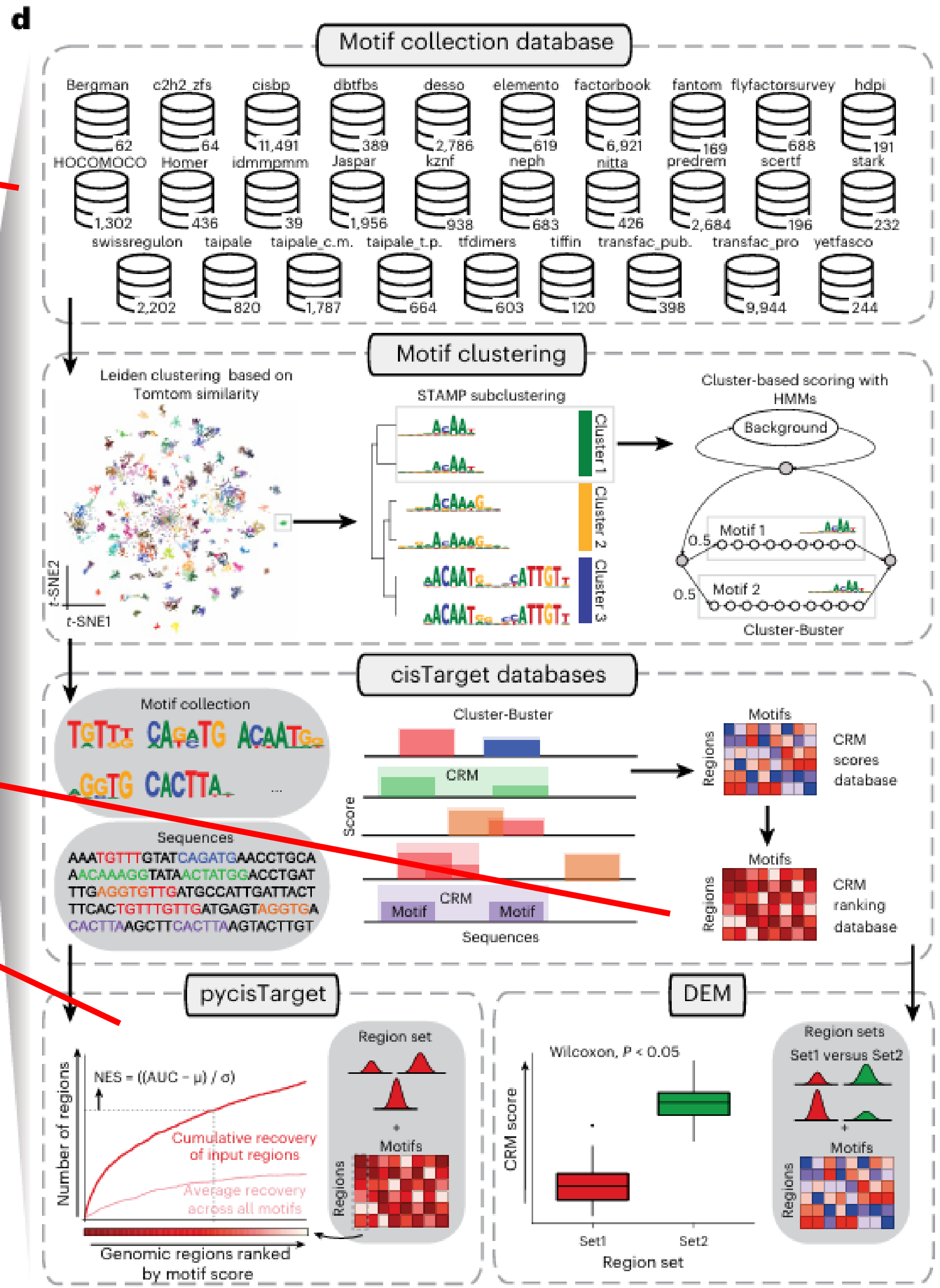
*procedure
-Motif clustering (HMM)



Each topic: each peak → ranked by which motif cluster (TF)
pycisTarget: AUC method → which motif cluster (TF) ranks the highest among the set of peaks in a given "topic"

Peaks from each topic → which "motif cluster" matches the best → which TF regulates the "topic"

DEM: regions in the topic vs other regions→
Wilcoxon by motif cluster (TF) score



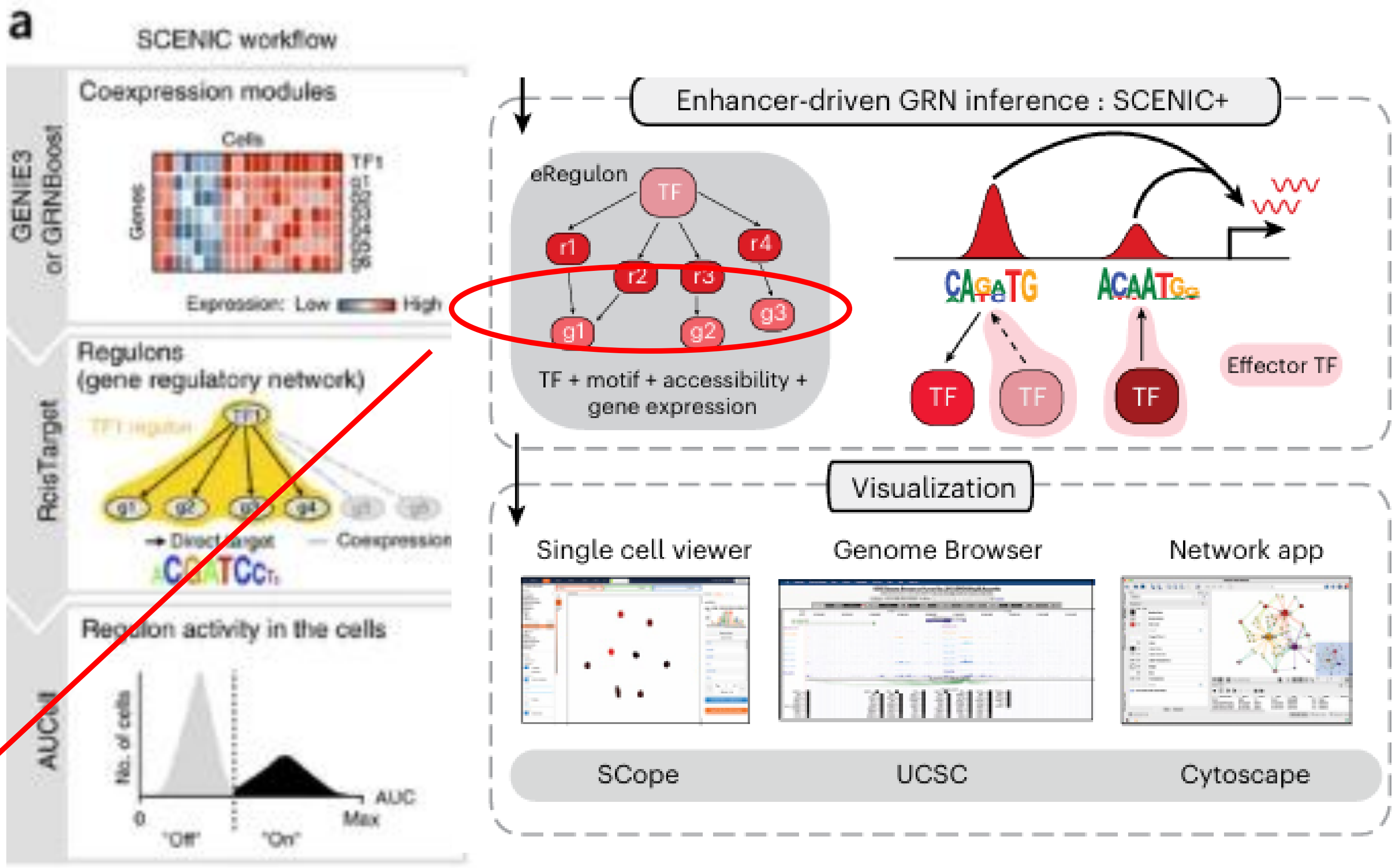
• SCENIC+ (gene regulatory network with open chromatin region)

-GRNboost2
Based on GENIE3 (random forest)
Output gene ~ other genes
→ TF → gene (Same)
Or region (enhancer; <150kbp
for a given gene) → gene

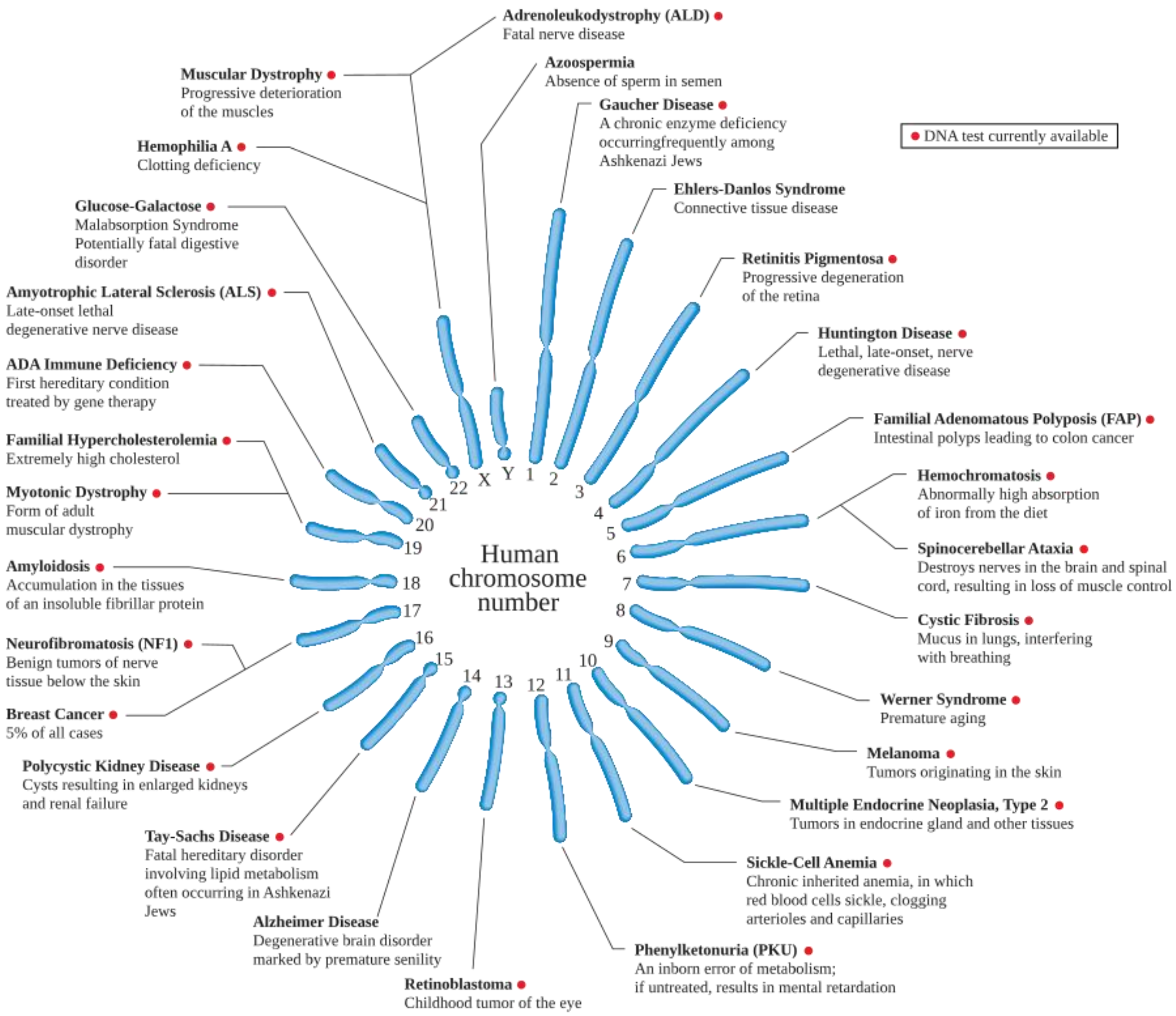
-TF~region~gene network
→ TF ~ region by motif enrichment
→ Just collecting all the edges
(after thresholding)
Direction: based on the correlation

-eRegulon: target genes by enhancers
GSEA of the “importance score” for each
Enhancer-gene

(SCENIC: AUCell)



- # Genetic association



-Genetic diseases

→ Genetic mutations can affect disease

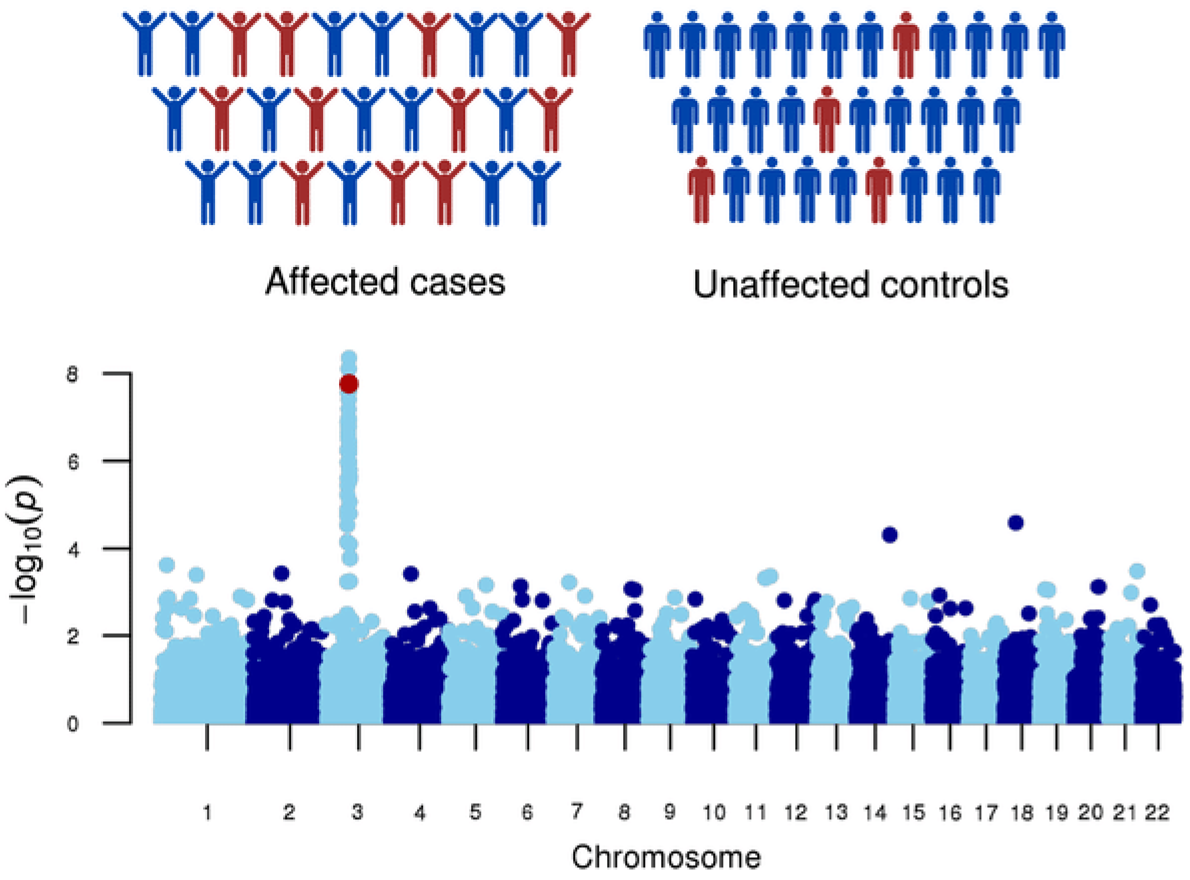
- Genetic association

Sequence Variation

Disease

GWAS

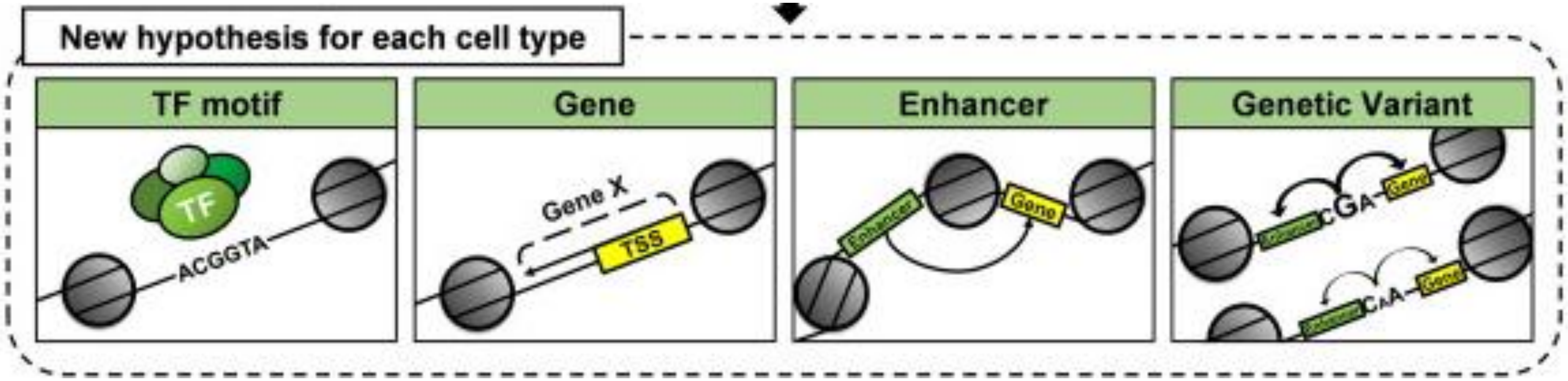
ATGCCAGTGTTTCAAGATGCTTGGCCAGCTGGACGAGGGCGATGAC
ATGCCAGTGTTTCAAGATGTTTGGCCAGCTGGACGAGGGCGATGAC



-GWAS
Genome-wide association study

→ All the genomic region
→ Associated with disease

- Genetic association

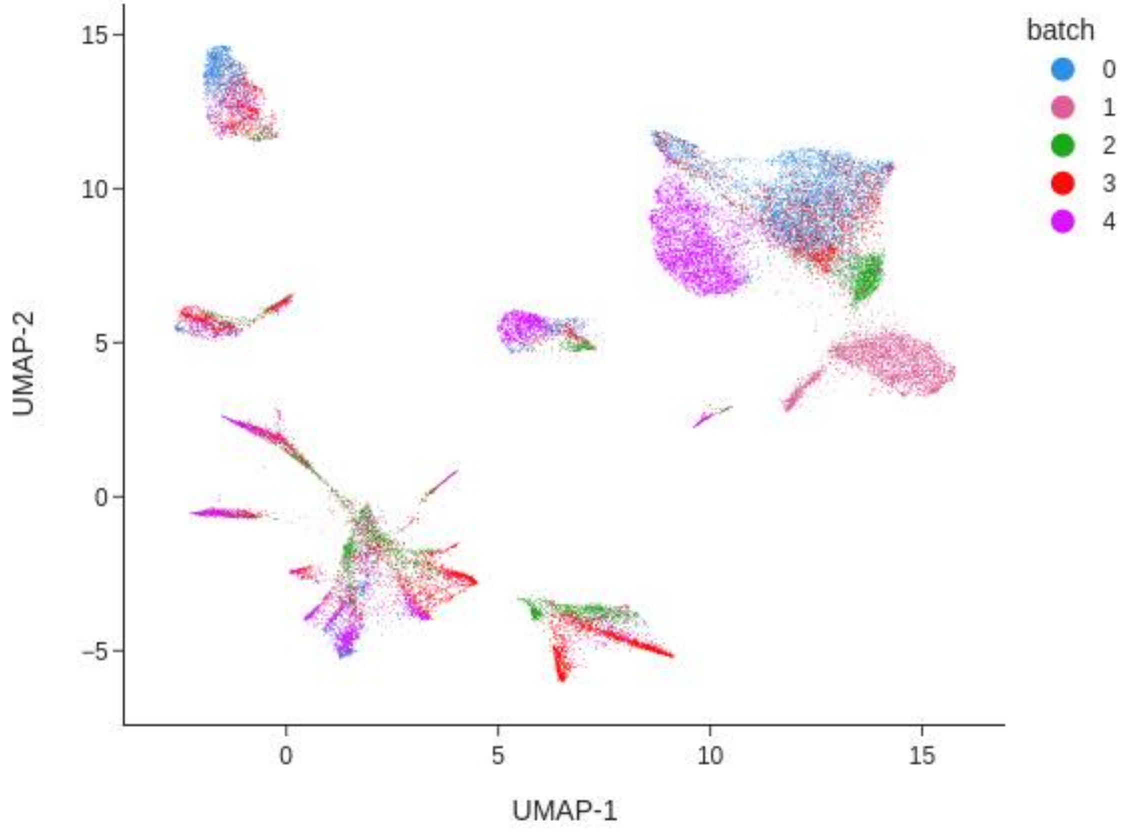


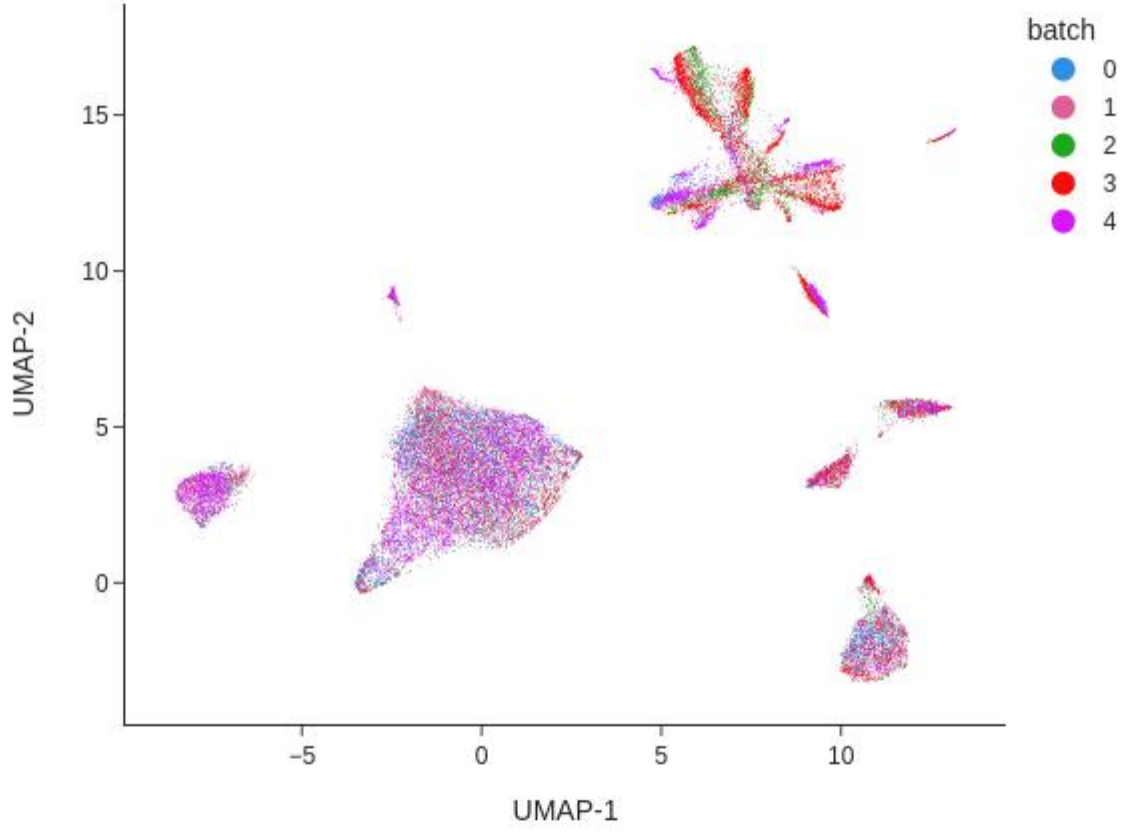
- Disease-associated open chromatin region (especially enhancer: it can regulate the gene expression)
- Even though a certain enhancer is opened from both disease and healthy group, if it encompasses genetic variants → Regulation can be altered (ex: TF binding by motif alteration)
- Known GWAS (SNP) overlaps with a given open chromatin region

→ Mechanism for how GWAS loci result in disease

- Batch correction

-Seurat or Harmony, etc

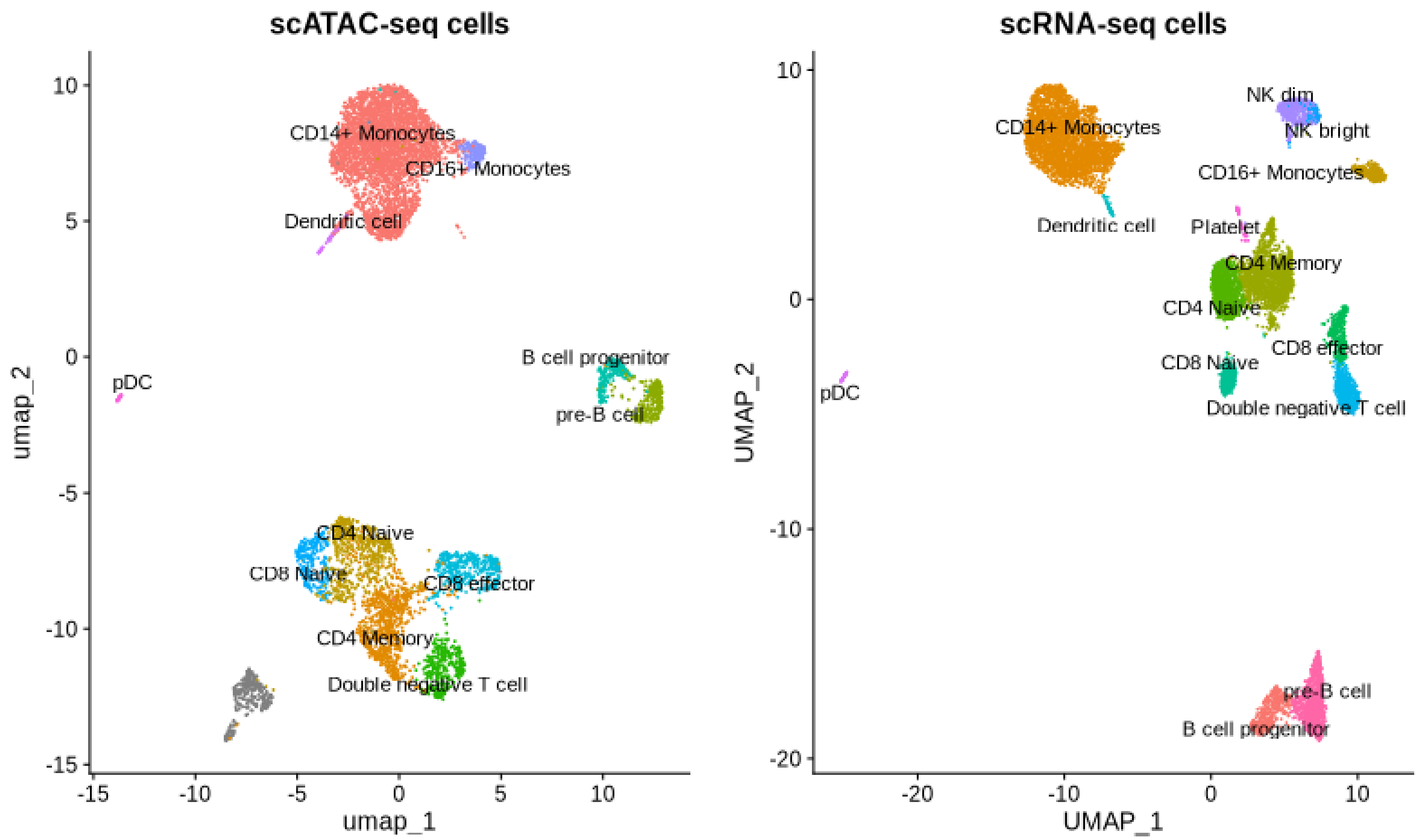




• Multi-modal analysis

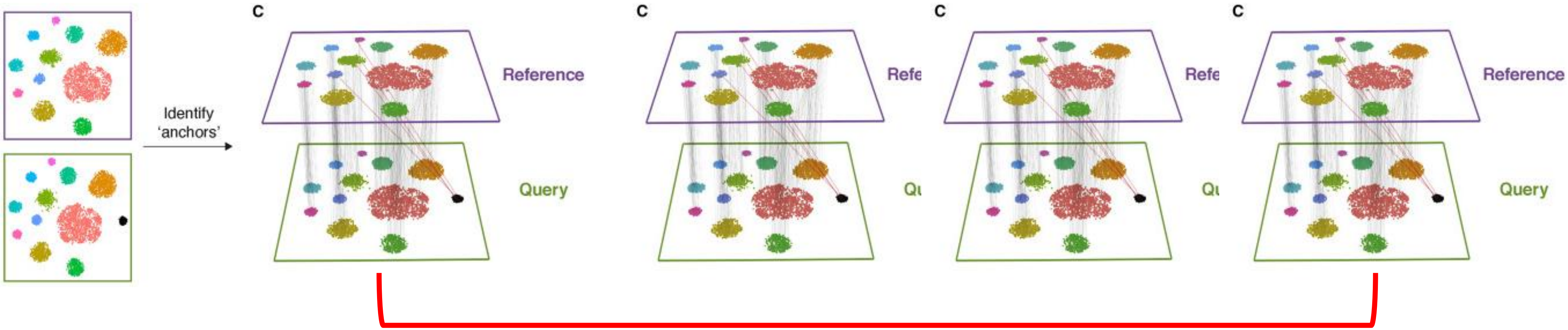
-Usually, scRNA-seq and scATAC-seq are conducted in a different sample

→ We need to merge both information together



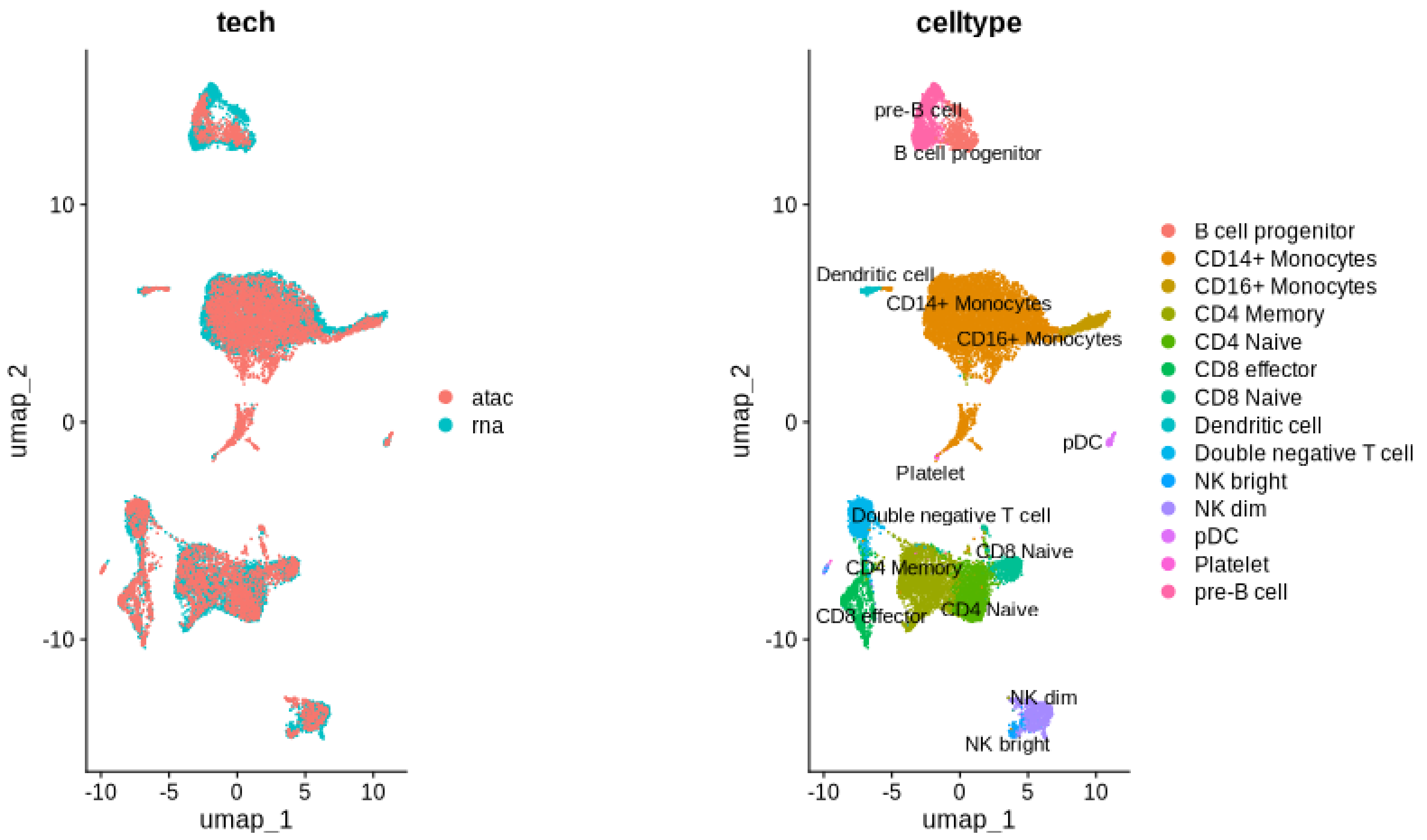
• Multi-modal analysis

-Label transfer method
Although ATAC has more features, high drop-out, worse for cell type annotation
(since gene activity is not an gene expression)
-Should use gene activity for label transfer (to match between feature name)

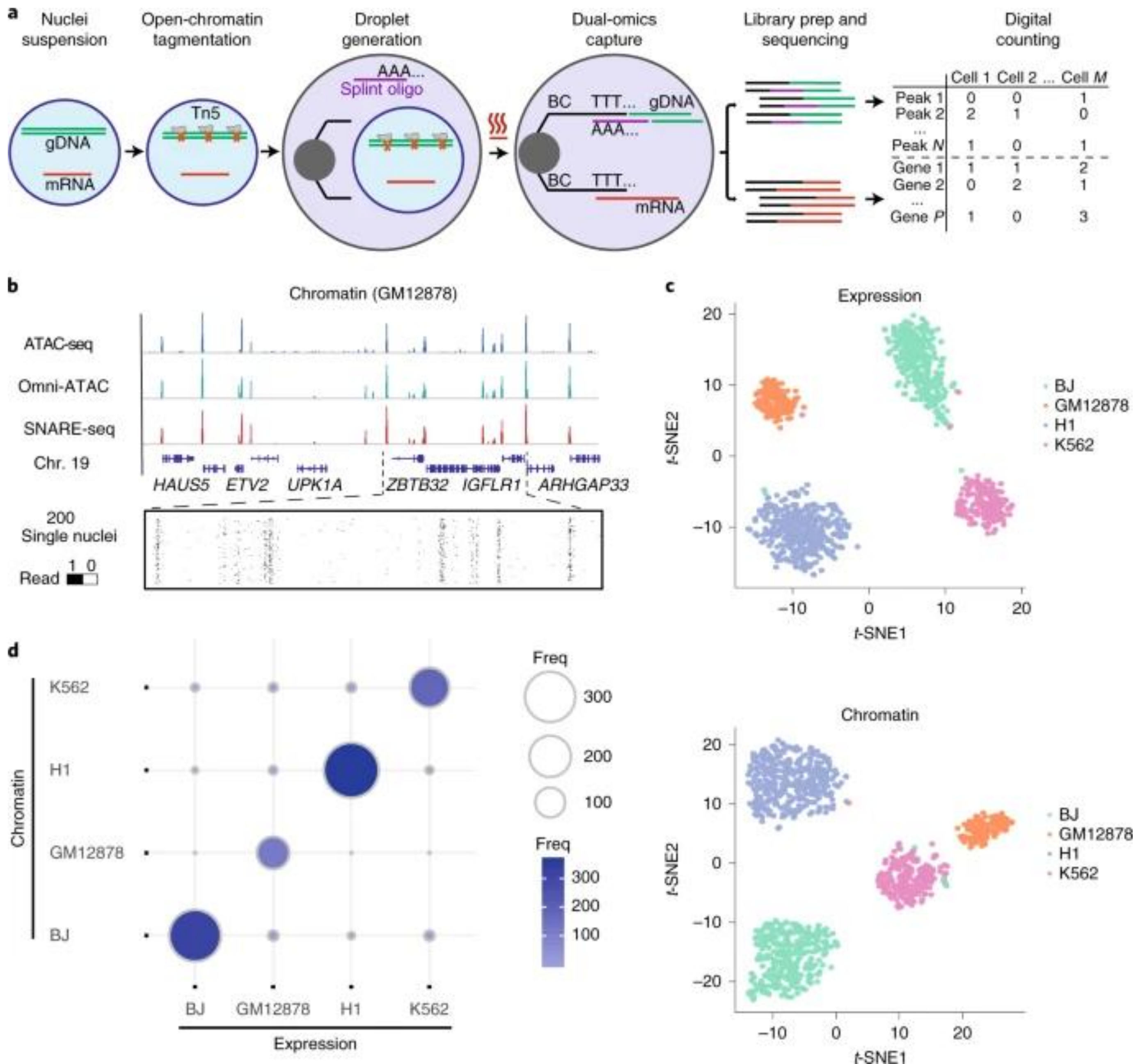


-Reference → query1, query2, query3 … (independently)
Reference: scRNA-seq
Query: scATAC-seq sample1,2,3 …

- Multi-modal analysis



• SNARE-seq



-scRNA + scATAC-seq from the same cell!

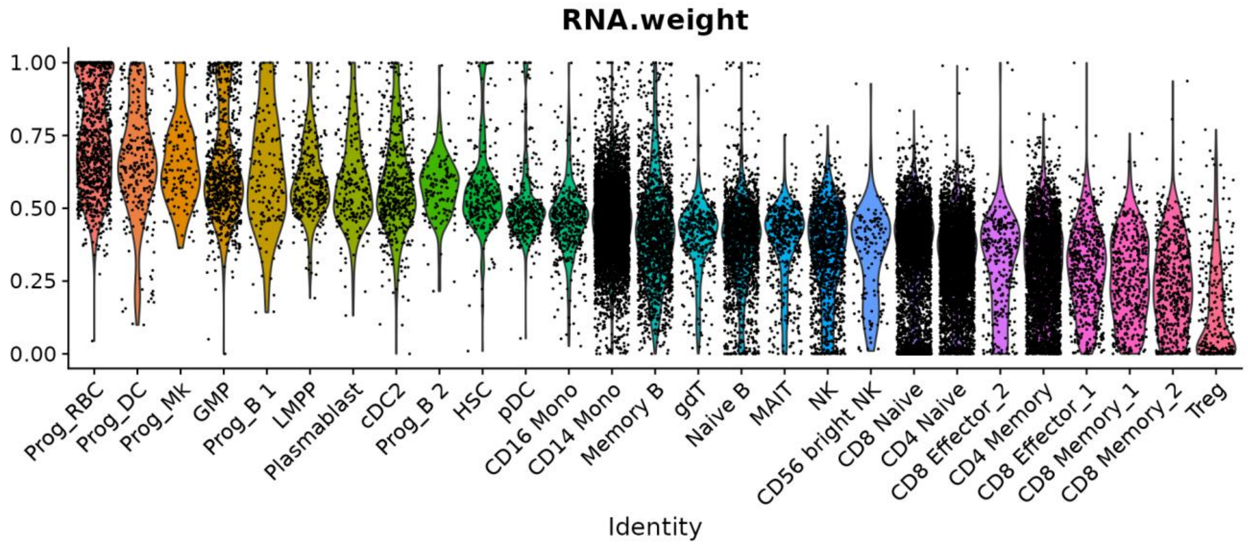
-Multi-modal analysis

-No need to multi-modal integration

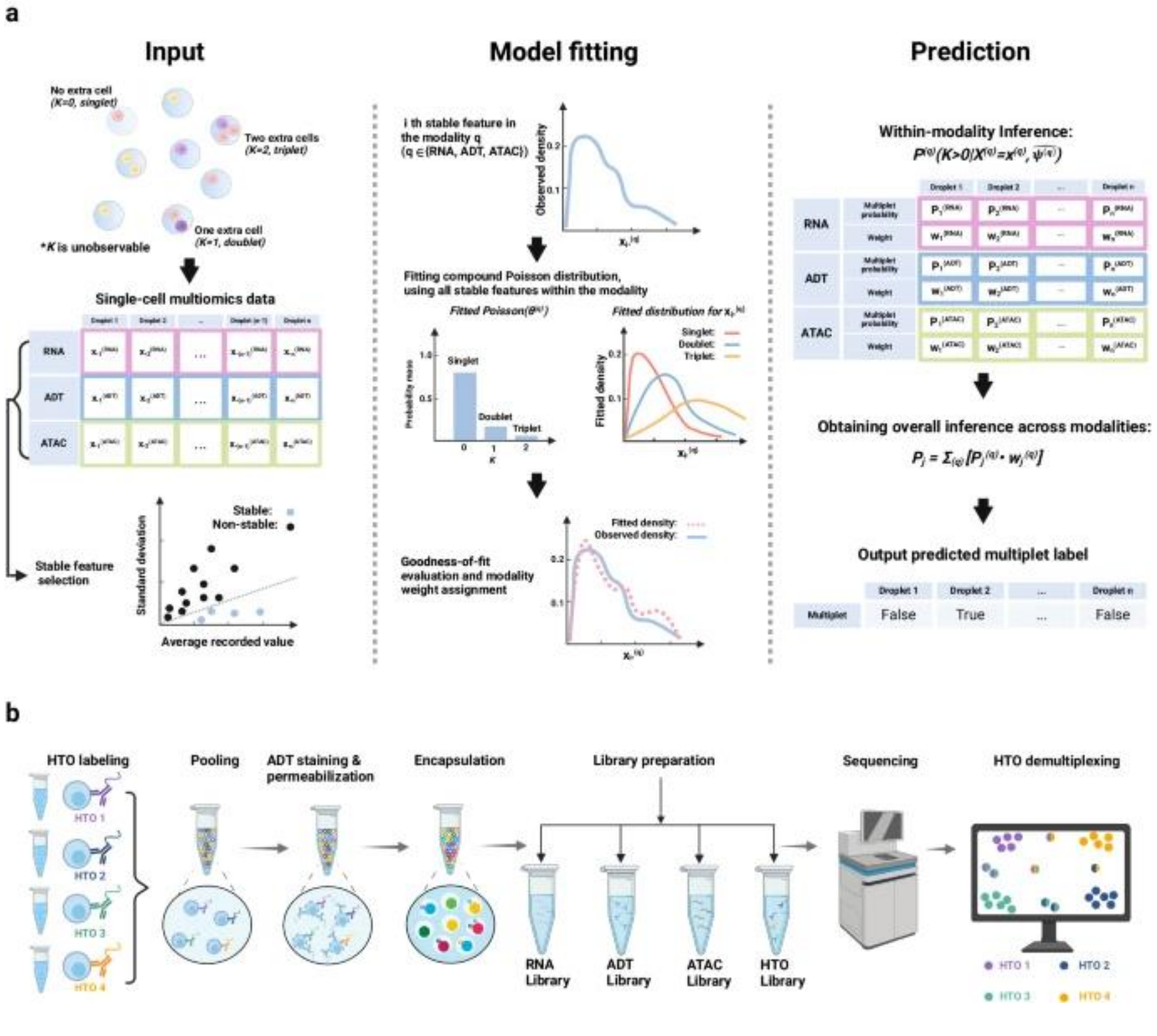
- SNARE-seq

*Weighted nearest neighbor analysis
-Mulitmodal integration analysis
-Incorporate (two) modalities (cell-specific weight)

*CITE-seq (protein) → scATAC-seq (open region)
→ Merging two modalities together



• Multi-modal analysis

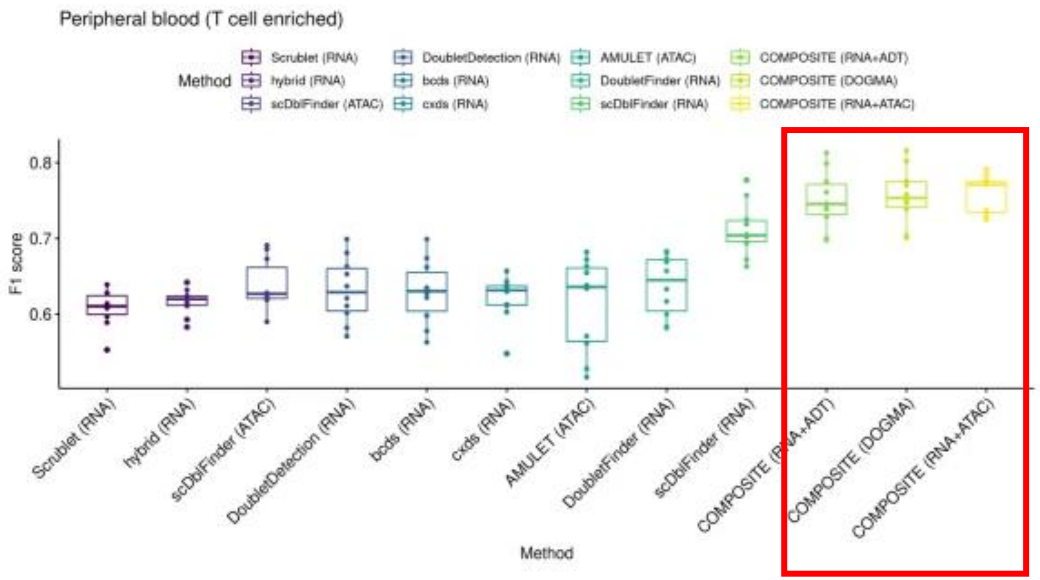


A unified model-based framework for doublet or multiplet detection in single-cell multiomics data

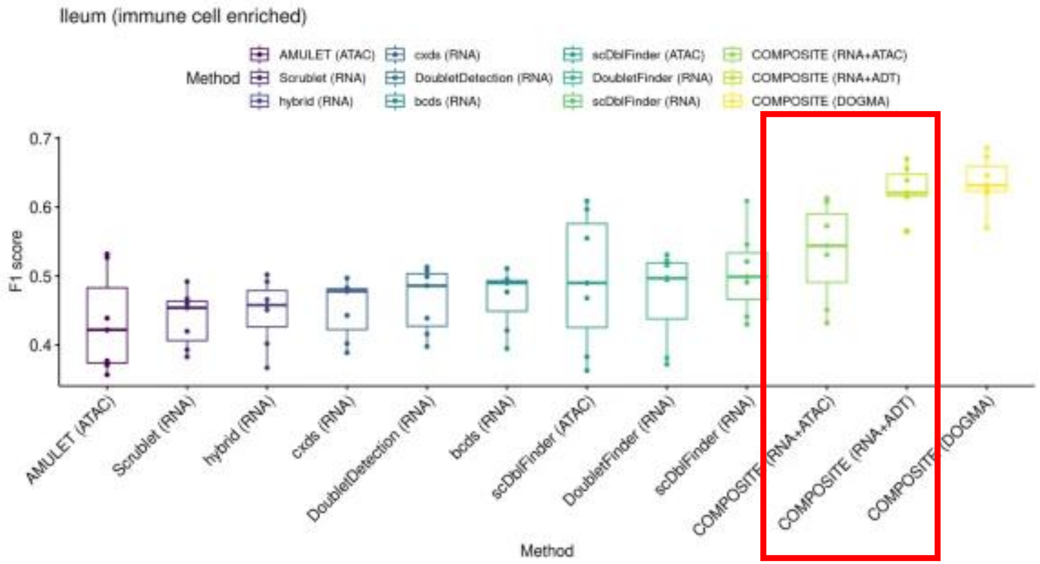
Doublet detection
COMpound POiSson multIplet deTEction
(COMPOSITE)

Stable gene selection
(less variable gene across cells)
→ Poisson distribution based model

- # Multi-modal analysis



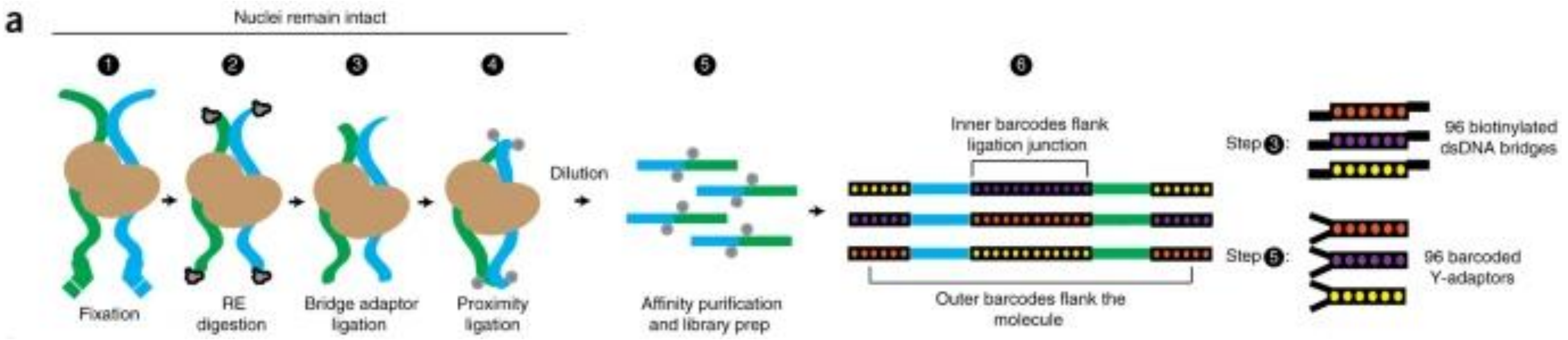
d

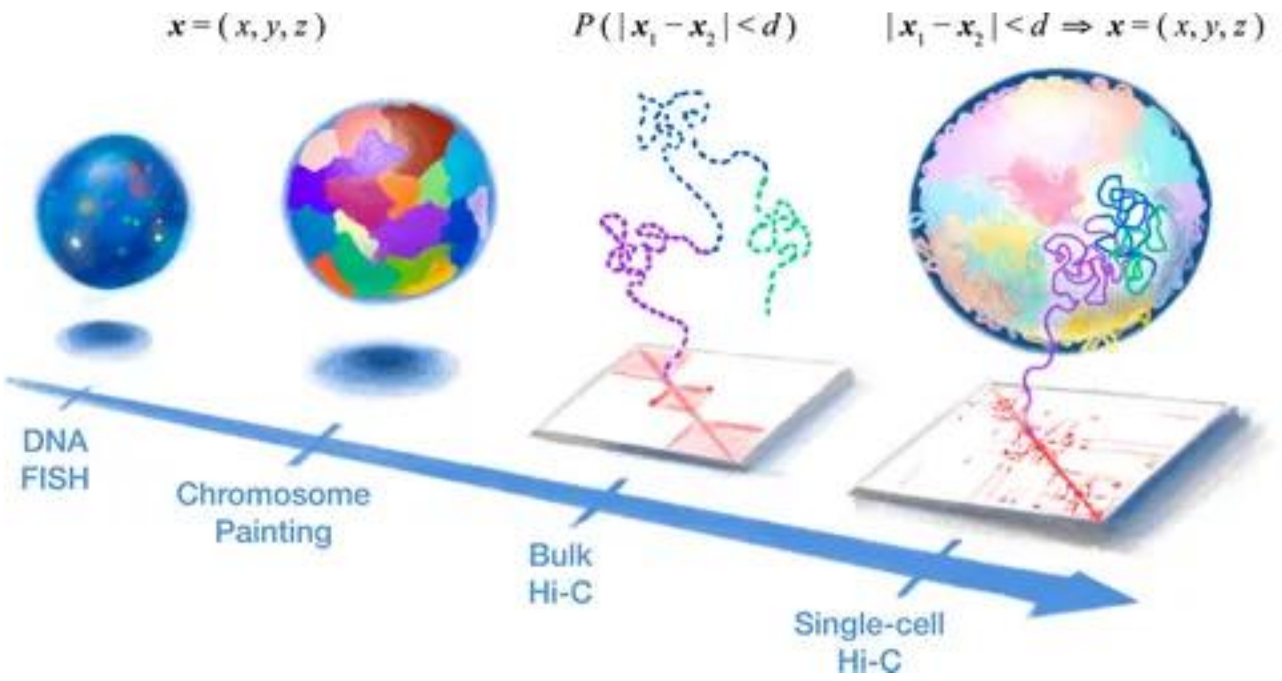


Multimodal approach
→ Better doublet detection

Other modalities

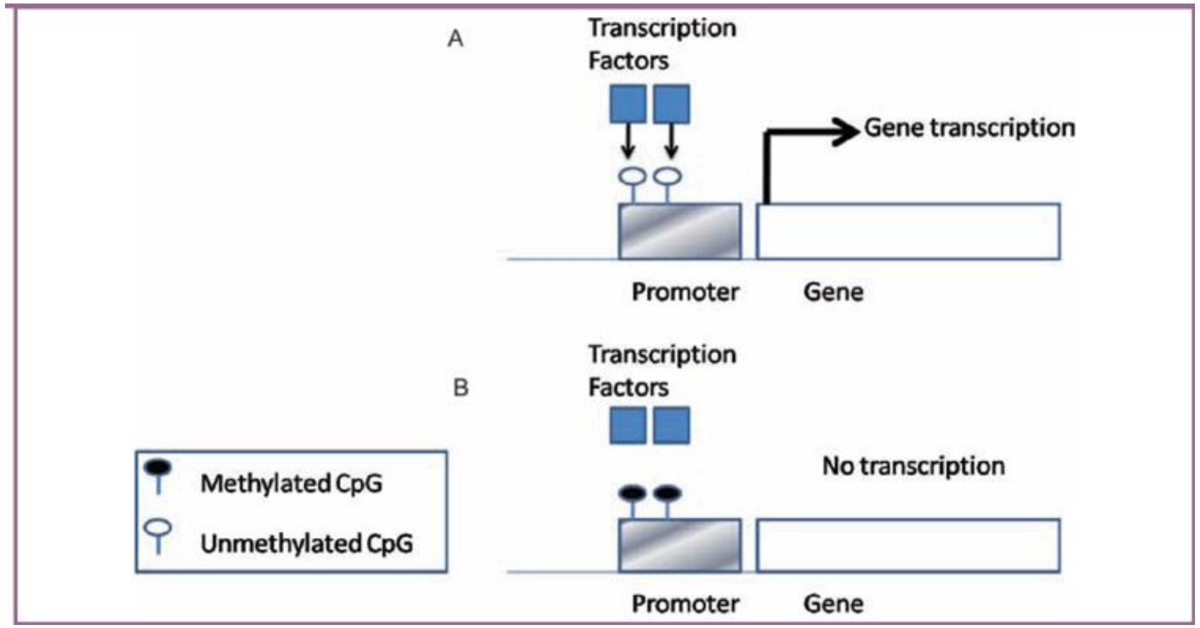
- scHi-C

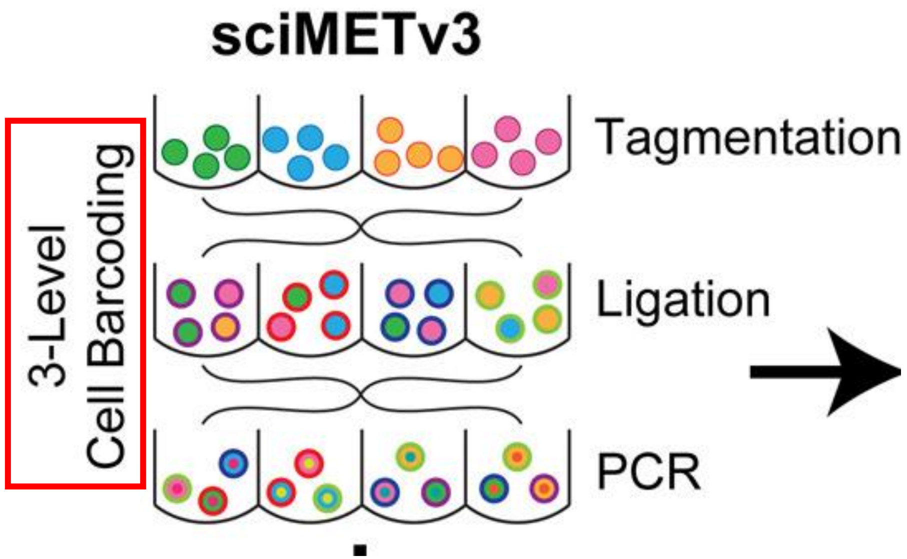


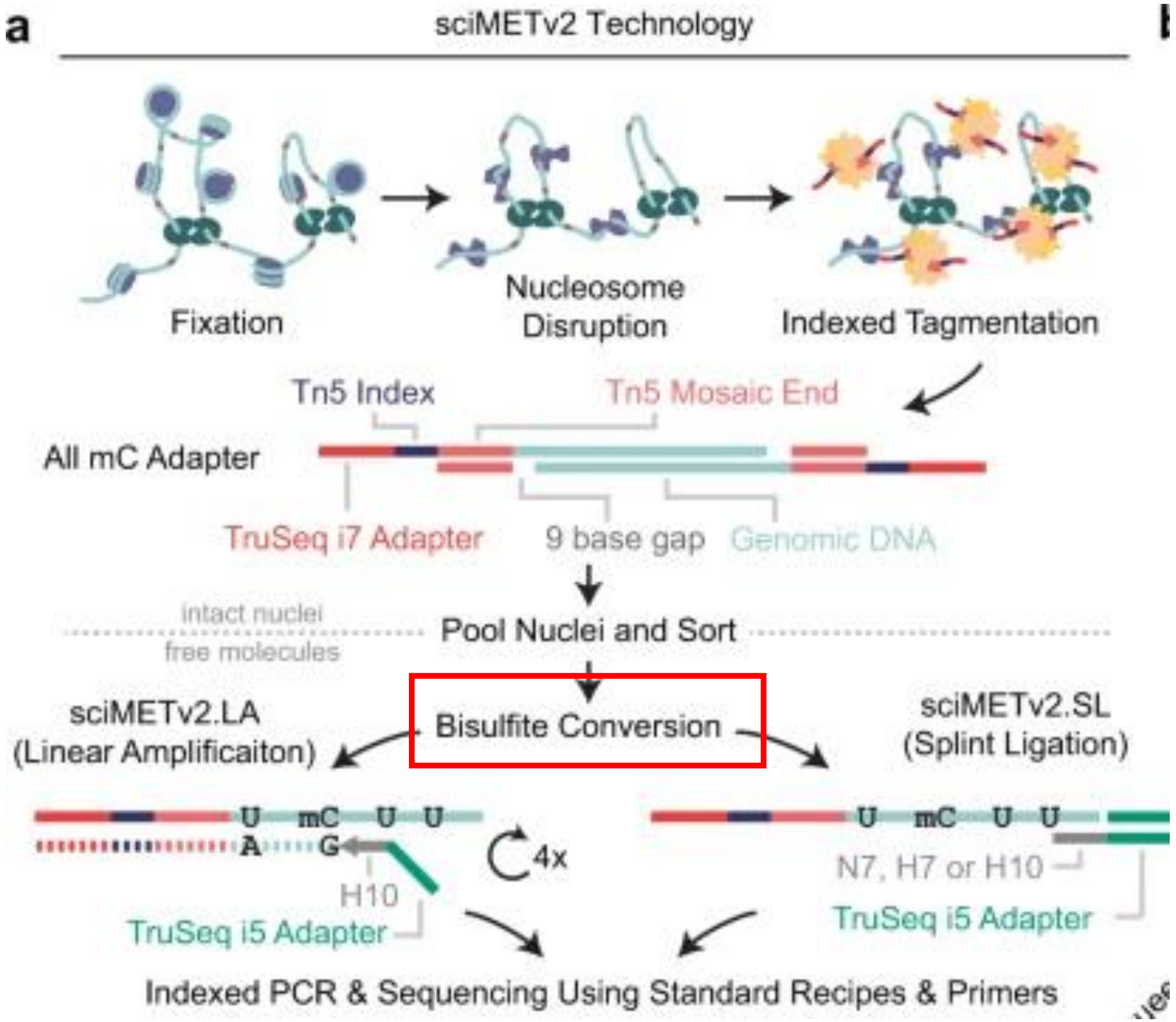


Barcoding during ligation
→ Multiplexed profiling
→ Highthroughput Single-cell performance

- scMethylation

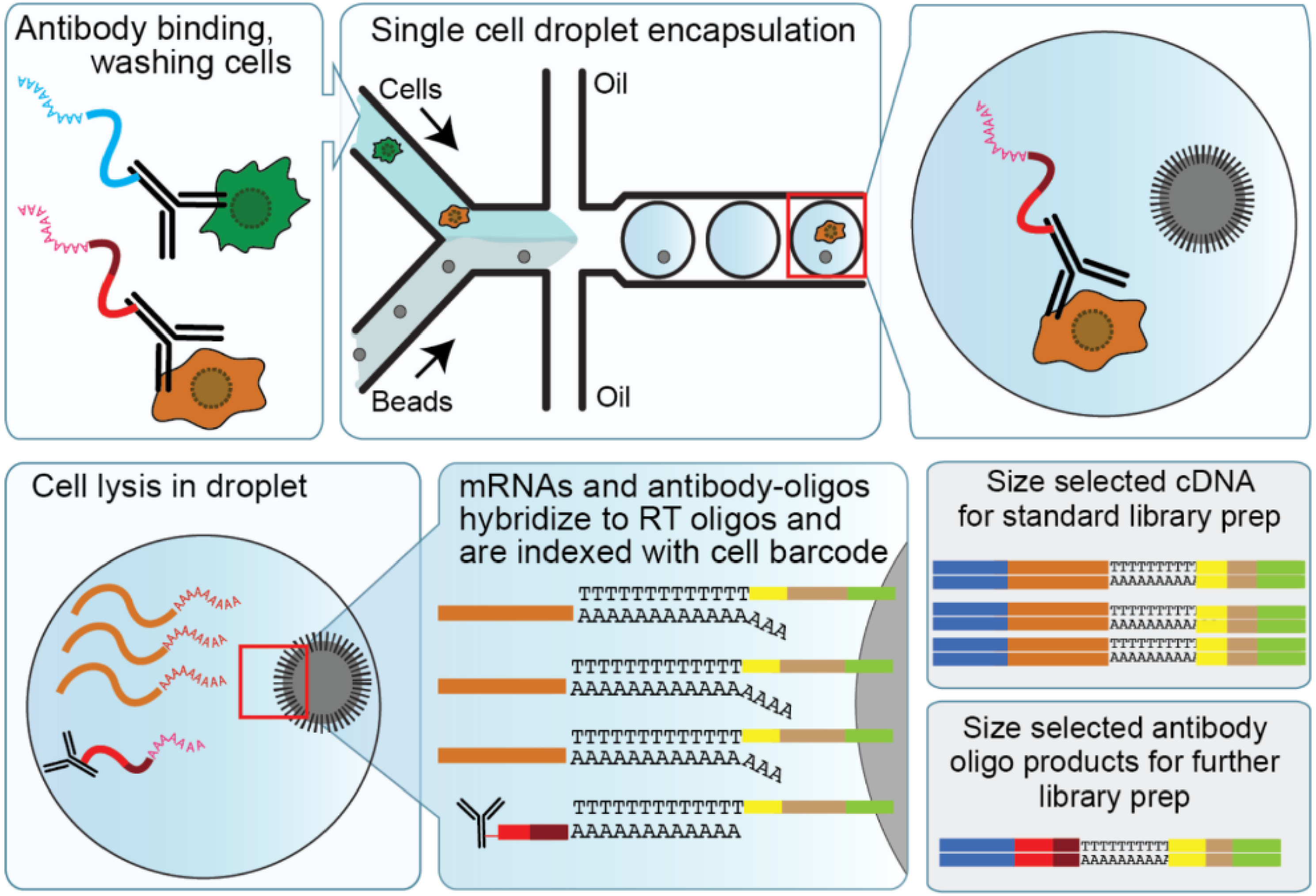






Bisulfite conversion →
distinguish methylation at the CpG island

- scProteomics



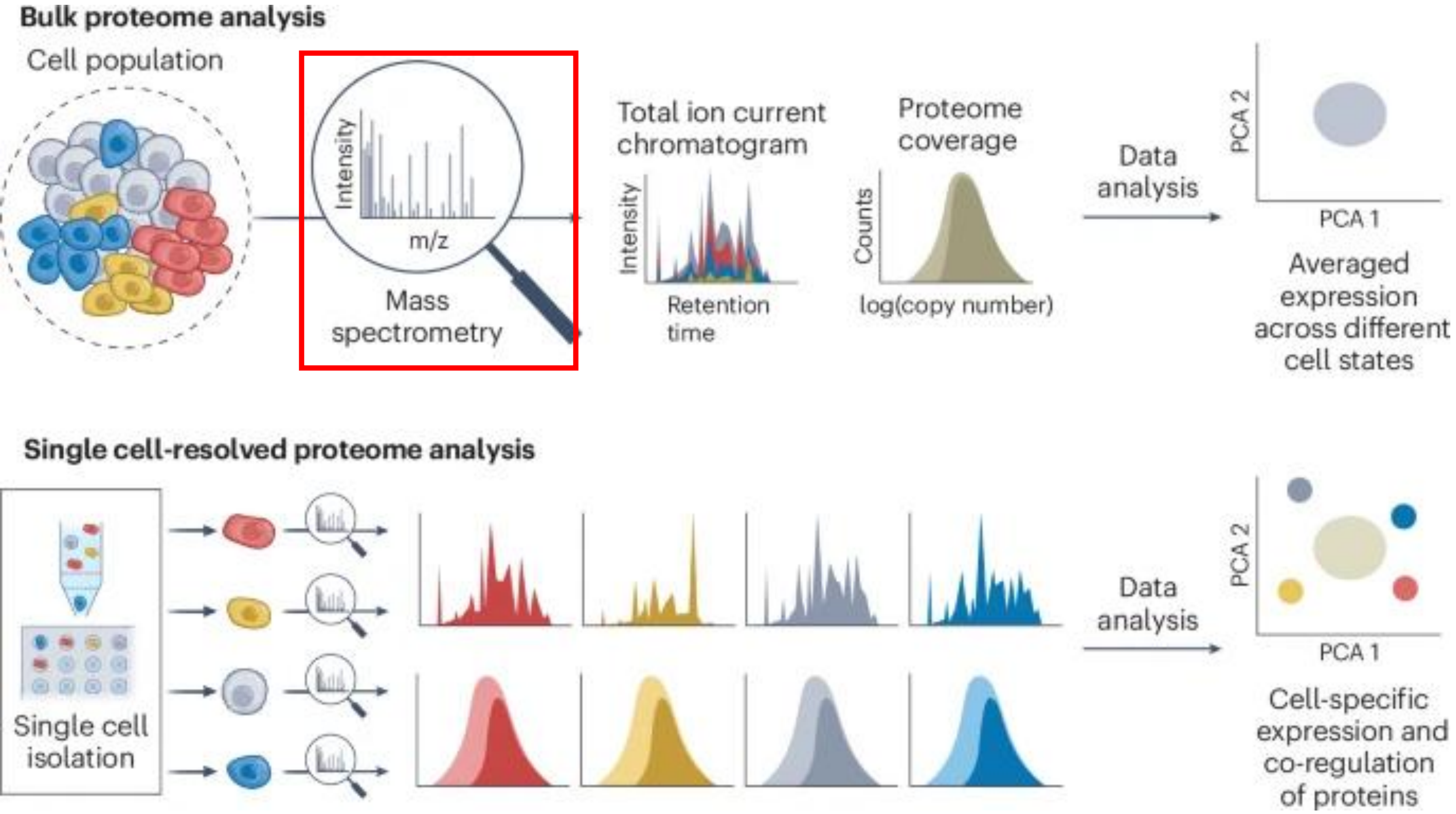
CITE-seq

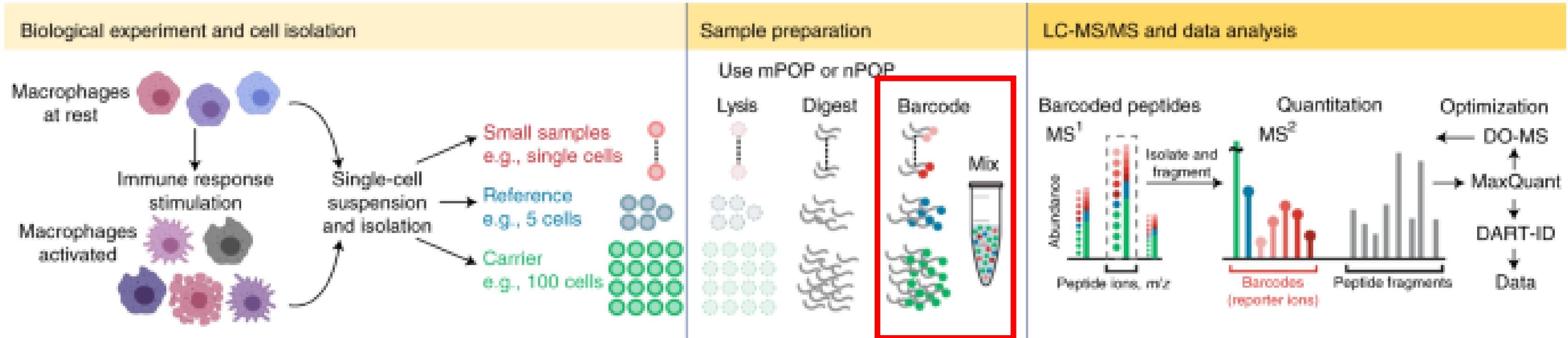
→ Can only probe ~10^2

→ Protein > Transcriptome

→ Quite limited ...

- scProteomics by Mass spectrometry





Time-of-Flight (TOF)

~ m/z

→ Z: is to define (integer)

→ M: distinguish peptides

→ Highthroughtput proteomics

+ single-cell barcoding