Single-cell ATAC-seq data analysis

BIML2020

2020/02/01

Insuk Lee, Yonsei University

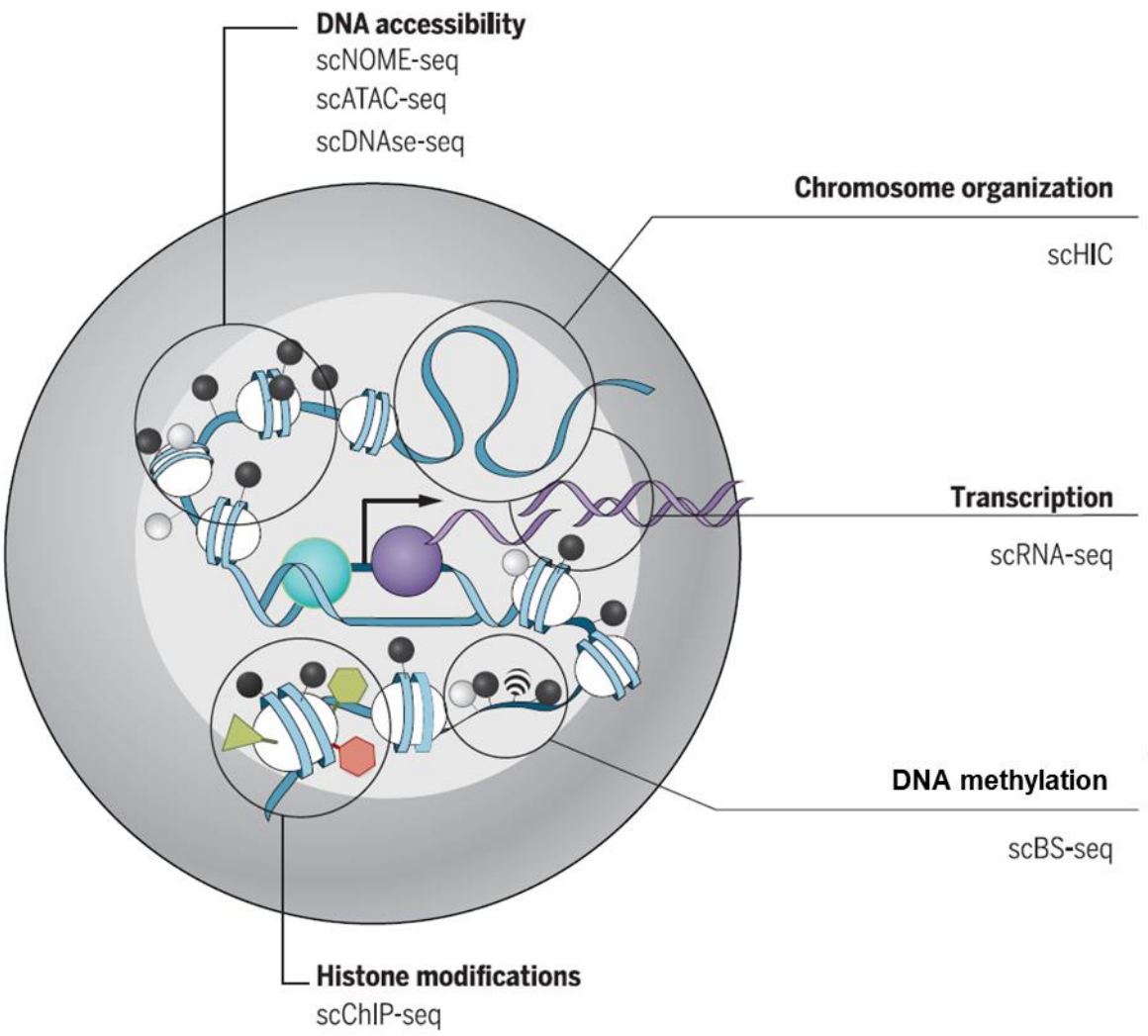
❖ Why epigenomics?

- Human has many different types of cells, specialized for different functions, and each of these cells carries essentially the same genome in its nucleus. The differences among cells are mainly determined by how and when different sets of genes are turned on or off.
- The transcriptional regulations are mediated by cis- and trans-elements. Epigenomic profiles reveal state of *cis*- and *trans*-regulatory elements for the cells.

❖ Five key epigenomic components (and profiling methods)

1. **DNA methylation (bisulfate sequencing; genotyping microarray):** Methylation of cytosines in CpG islands or other genomic DNA is associated with gene silencing.
2. **Chromatin accessibility (DNase-seq; ATAC-seq):** The genomic DNA region that is tightly bound by the nucleosomes is less accessible to transcription machinery and other regulatory proteins. Chromatin accessibility is regulated during cellular differentiation.
3. **Histone modification (ChIP-seq):** Chromatin is a dynamic structure that must respond to myriad stimuli to **regulate access to DNA**, and **chemical modification of histone** is a major means by which the cell modulates **nucleosome mobility and turnover**.
4. **Chromatin organization (Hi-C):** Spatial organization of chromatic regions is dynamic and may affect gene expression. **Spatial (3-D) proximity of chromosomal loci** can be experimentally determined.
5. **Non-coding RNA (RNA-seq):** Both short and long non-coding RNA species are involved in epigenome regulation by (i) **influencing expression and function of epigenetic regulators**, (ii) **recruiting epigenetic regulator** and (iii) **shaping 3D nuclear architecture.**

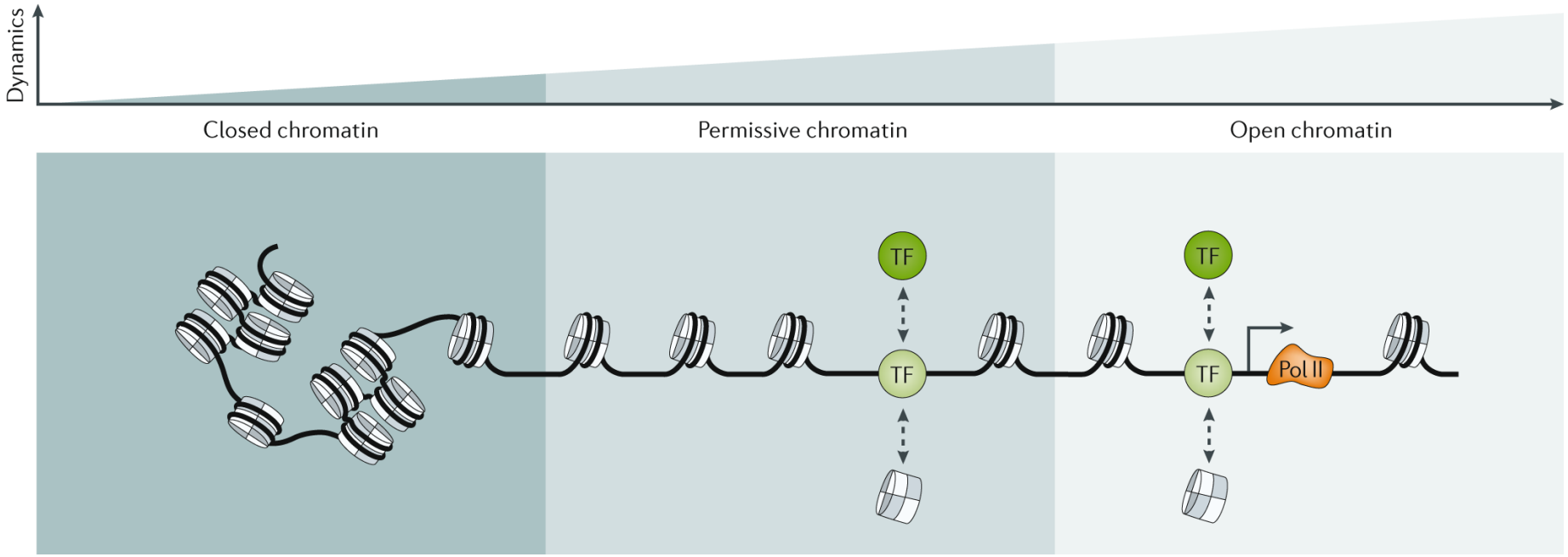
❖ **All epigenomic profiling methods are available at single-cell level**



Science 358: 69 (2017)

❖ Chromatin accessibility

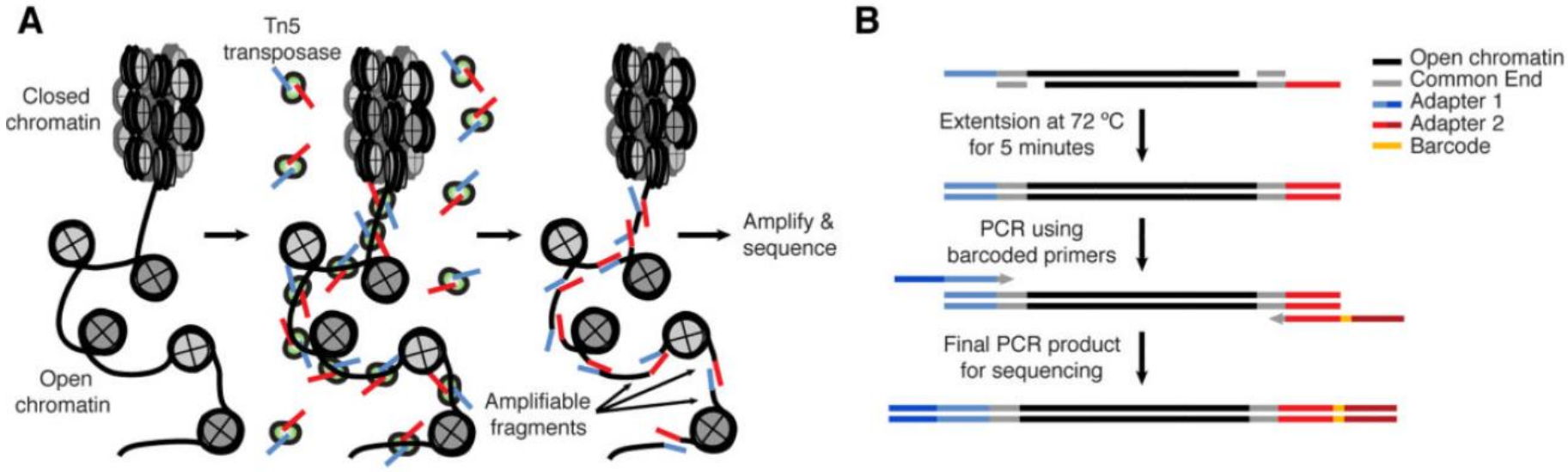
- Open chromatin = more transcription
- Chromatin accessibility varies for different cell types, states, and activity



S. Klemm, et al. Nature Reviews Genetics (2019)

❖ Methods for measuring Chromatin accessibility

- Dnase-seq (using Dnase I) and ATAC-seq (using mutated hyperactiveTn5 transposase)
- Others are MNase-seq (using Endo/exonuclease Mnase) and NOME-seq (using Methyltransferase).
- Currently, ATAC-seq is the method of choice, because it requires much less DNA sample and much shorter preparation time than other methods.
- In a process called "**tagmentation**" (tag + fragmentation), Tn5 transposase cleaves and tags dsDNA with sequencing adaptors.



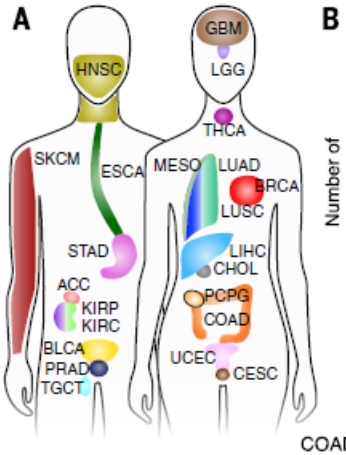
J. Buenrostro, et al. Curr Protoc Mol Biol(2015)

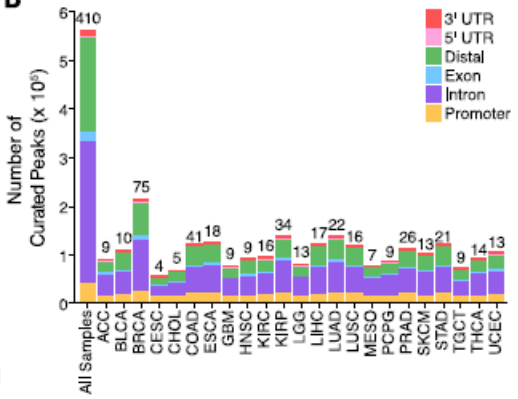
❖ Why (single-cell) chromatin accessibility?

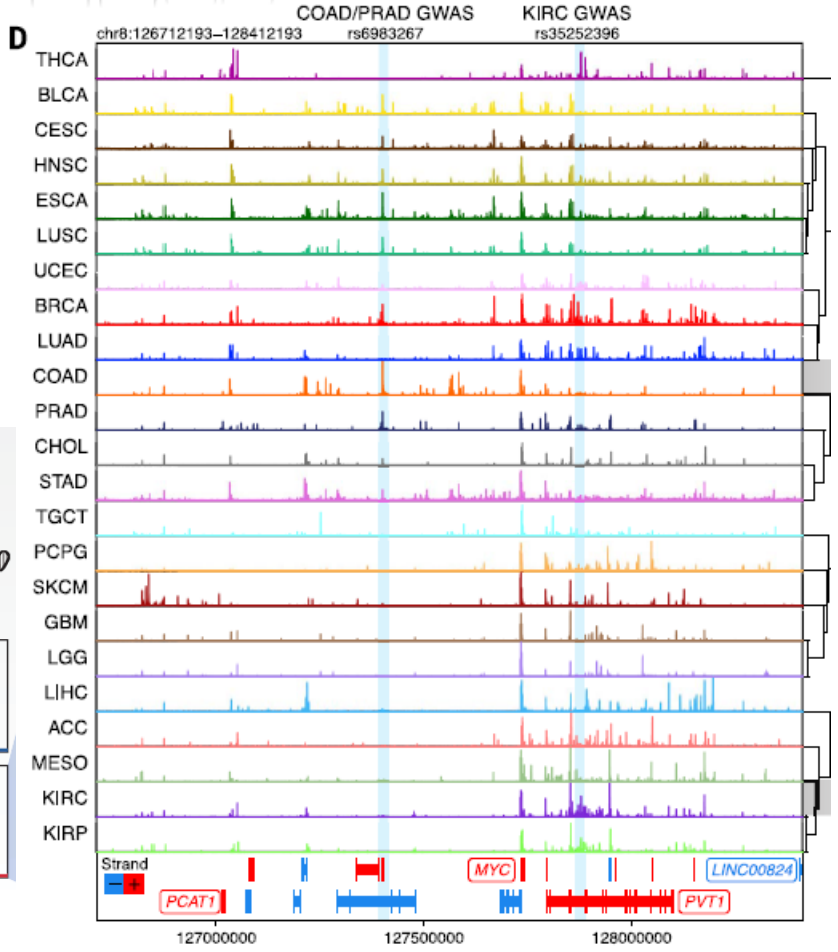
- Gene expression programs are tightly controlled by the concerted action of TFs, chromatin modifiers, chromatic accessibility and other regulatory factors. Genome-wide epigenomic analysis are therefore instrumental for **determining key regulators of gene expression** and **refining gene regulatory network (GRN) models**.
- **Unbiased chromatin accessibility profiles** allow to **discover novel *cis* regulators** (e.g., enhancers) and **identifying *master TFs*** with their activity profiles across cell states.
- The vast majority of disease-associated SNPs lie outside of coding regions. Chromatic accessibility information is essential for **relating noncoding genetic variation to regulatory mechanisms** underlying disease.
- With **pseudotemporal ordering based on single-cell chromatin accessibility profiles**, asynchronous cells can be ordered by their developmental progression to identify the **step-wise activation of key cis- and trans-effectors underlying cell differentiation and commitment**.
- Before single-cell technologies available, **ATAC-seq profiles for homogeneous cell population** (e.g., cancer cells for distinct type; FACS-sorted immune cells for each type) suggested the importance of **cell-type(state)-specific chromatin accessibility** information in understanding disease-associated regulatory mechanisms.

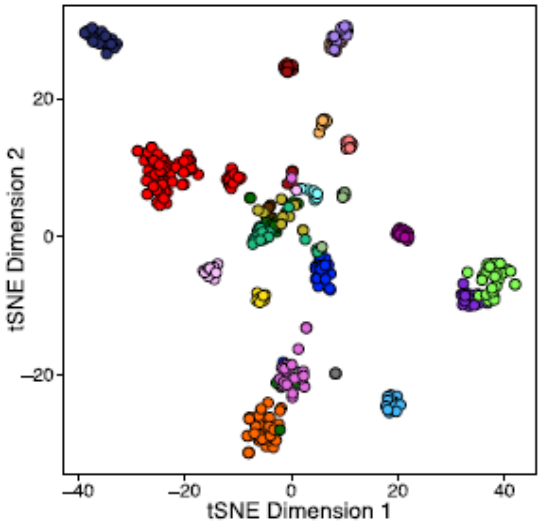
- **Chromatin accessibility landscape of primary human cancers** *Science* **362:**eaav1898, (2018)

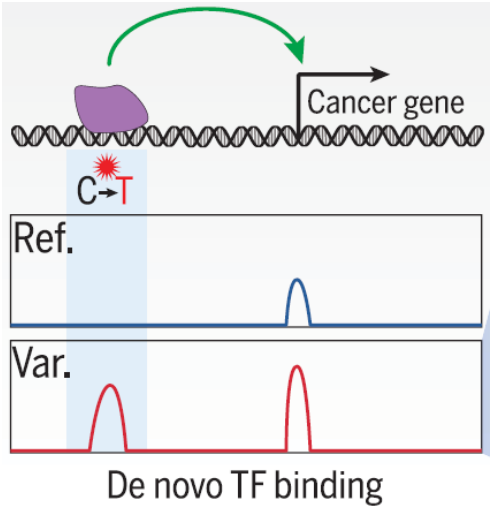
  - ATAC-seq for 23 cancer types, 410 tumor samples
  - Peaks for non-coding regions might contain active regulatory elements, suggesting <u>many novel enhancer elements</u>.
  - There are <u>many **cancer-type-specific peaks**</u>.
  - Dimension reduction (e.g., tSNE) of ATAC-seq data using most variable peaks confirms **cancer-type-specific chromatin accessibility** (e.g., active regulatory element).
  - Integration with mutation data **identify cancer-relevant noncoding mutations that are associated with altered gene expression.**



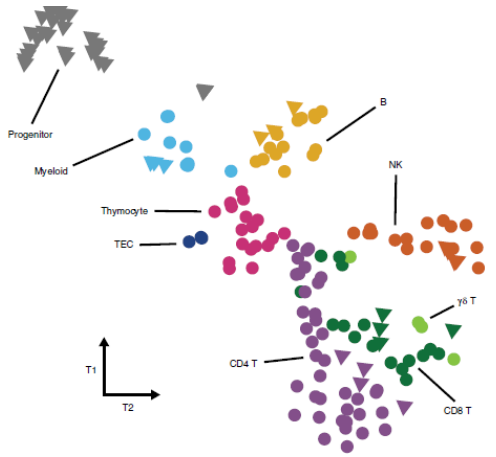


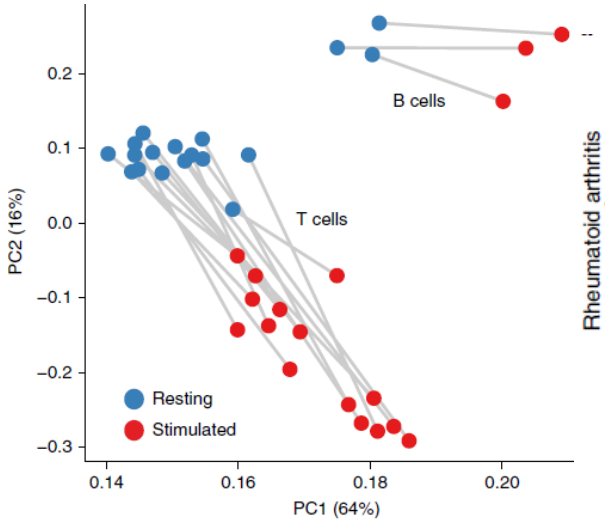


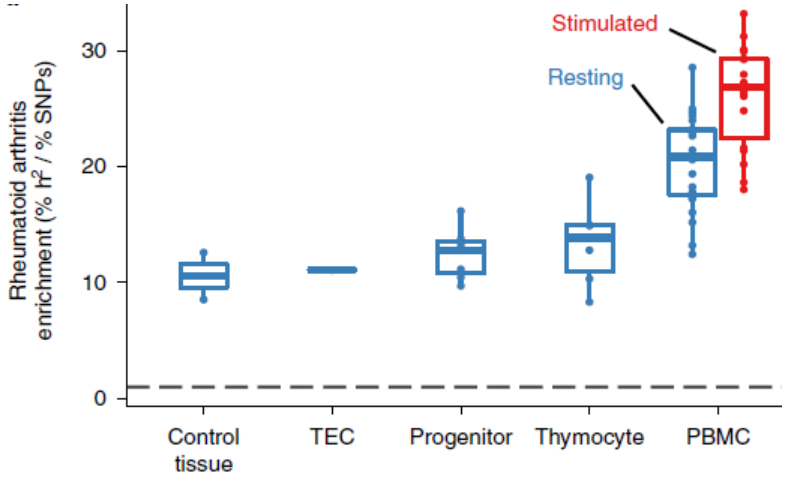




- **Chromatin accessibility landscape of resting and stimulated immune cells**
  Nature Genetics 51**:1494** (2019)
  - ATAC-seq for 25 types of immune cells at resting state and stimulated (activated) B, T, NK cells
  - GWAS have identified many SNPs that contribute to the risk of autoimmunity. ~90% of these signals lie in noncoding regions and thus presumably act by altering gene regulation; however, only ~25% of them could be explained through eQTL (SNP associated with expression variation of a certain gene).
  - It has been shown that some GWAS-eQTL overlap can remain hidden within stimulation-specific regulatory regions of immune cells, emphasizing the unique role of stimulation to autoimmunity.
  - Dimension reduction (e.g., tSNE) of ATAC-seq data using most variable peaks from cells in a resting state shows clusters of major immune cell types.
  - Overall, stimulation drives dramatic changes in the chromatin landscapes of B and T cells. **The chromatin differences due to stimulation were nearly as large as differences between cell lineages** (25% versus 32%, respectively).
  - For rheumatoid arthritis, **SNPs in accessible chromatin regions show much higher enrichment in stimulated cells compared with resting-state cells**. Thus, we now can identify eQTL SNPs that were net detectable from immune cells at resting state.

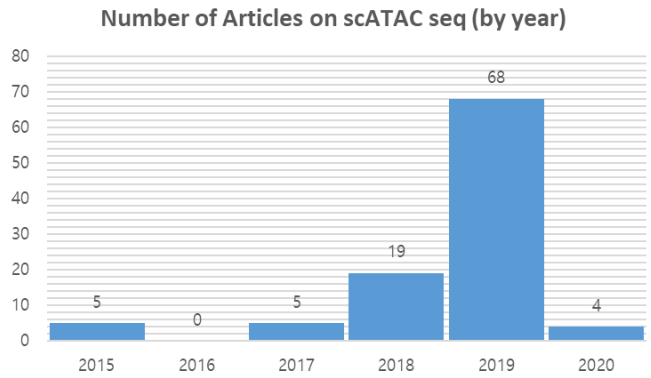


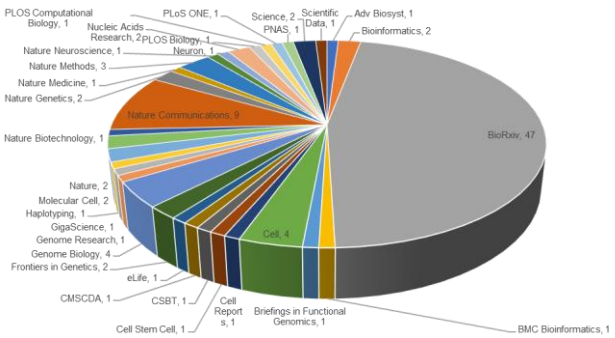


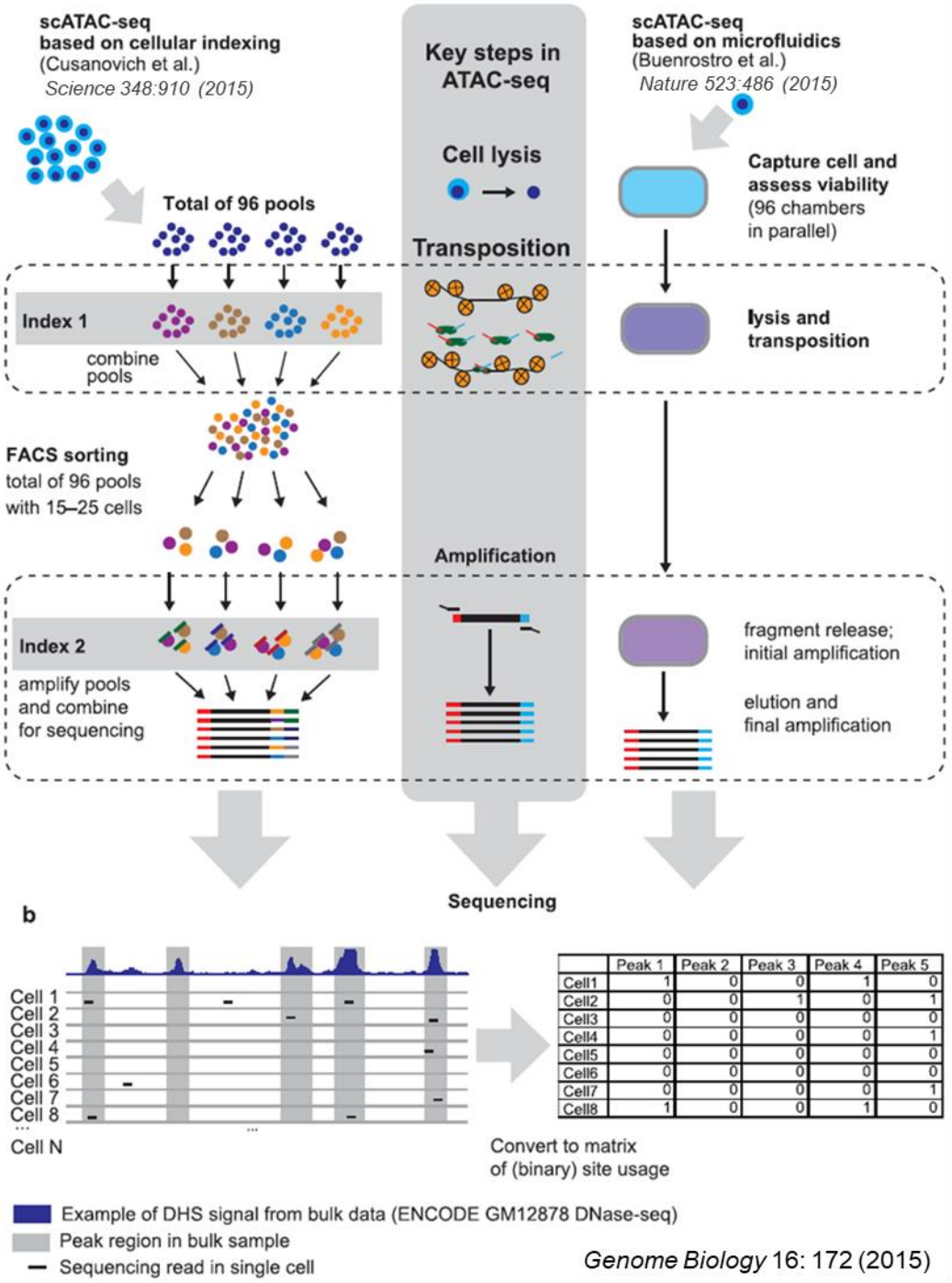


❖ Single-cell ATAC sequencing

- The first bulk ATAC-seq was published in 2013 (Nature Methods **10**:1213)
- The first scATAC-seq was published in 2015.
- Now >100 scATAC-seq papers







❖ Key articles for scATAC-seq research

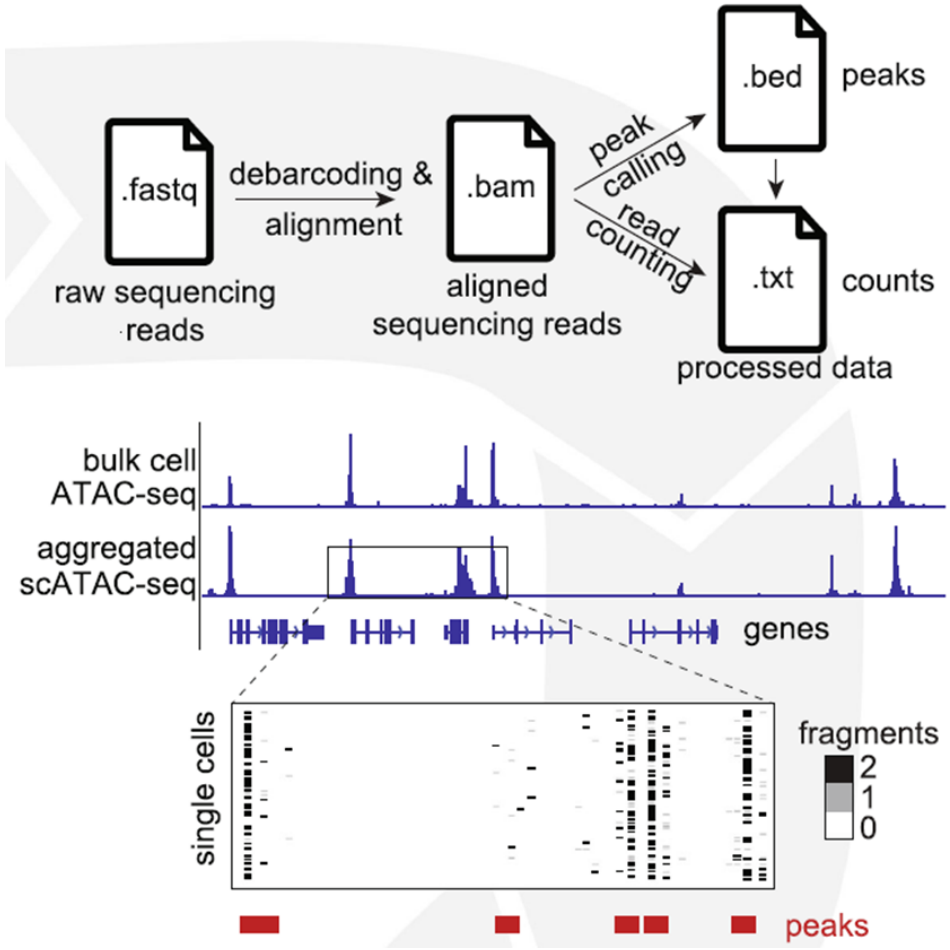
Development of bulk ATAC-seq	Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position Jason D. Buenrostro, et al. Nature Methods (2013)
Development of sciATAC seq (indexing)	Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing Darren A. Cusanovich, et al. Science (2015)
Development of scATAC-seq (droplet)	Single-cell chromatin accessibility reveals principles of regulatory variation Jason D. Buenrostro, et al. Nature (2015)
chromVAR (TF motif)	chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomics data Alicia N. Schep, et al. Nature Methods (2017)
sciATAC seq of 15,000+ mouse forebrain cells	Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation Sebastian Prelssl, et al. Nature Neuroscience (2018)
sciATAC seq of 20,000+ Drosophila embryos cells	The cis-regulatory dynamics of embryonic development at single-cell resolution Darren A. Cusanovich, et al. Nature (2018)
sciATAC seq of 100,000+ adult mouse cells	A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility Darren A. Cusanovich, et al. Cell (2018)
Development of Cicero (trajectory analysis)	Cicero Predicts cis-regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data Hannah A. Pliner, et al. Molecular Cell (2018)
scRNA + scATAC seq integration (Seurat)	Comprehensive Integration of Single-Cell Data Tim Stuart , et al. Cell (2019)
10x scATAC of 200,000+ PBMC & BCC cells	Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion Ansuman T. Satpathy, et al. Nature Biotechnology (2019)
Benchmark of scATAC-seq tools	Assessment of computational methods for the analysis of single-cell ATAC-seq data Huidong Chen, et al. Genome Biology (2019)

❖ Data analysis tools

	Tool name	Article name	Journal	Year
1	**SCRAT**	Single-cell regulome data analysis by SCRAT	Bioinformatics	May 2017
2	**Dr.seq2**	Dr.seq2: A quality control and analysis pipeline for parallel single cell transcriptome and epigenome data	PLoS ONE	July 2017
3	**chromVAR**	chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomics data	Nature Methods	October 2017
4	**scABC**	Unsupervised clustering and epigenetic classification of single cells (scABC)	Nature Communications	June 2018
5	**BROCKMAN**	BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization	BMC Bioinformatics	July 2018
6	**PRISM**	A Cosine Similarity-based Method to Infer Variability of Chromatin Accessibility at the Single-Cell Level (PRISM)	Frontiers in Genetics	August 2018
7	**Cicero**	Cicero Predicts cis-regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data	Molecular Cell	September 2018
8	**Scasat**	Classifying cells with Scasat, a single-cell ATAC-seq analysis tool	Nucleic Acids Research	October 2018
9	**SIP**	SIP: An Interchangeable Pipeline for scRNA-seq Data Processing	BioRxiv	October 2018
10	**fdapace**	Sparse functional data analysis accounts for missing information in single-cell epigenomics	BioRxiv	December 2018
11	**Destin**	Destin: toolkit for single-cell analysis of chromatin accessibility	Bioinformatics	March 2019
12	**ChromA**	Characterizing the epigenetic landscape of cellular populations from bulk and single-cell ATAC-seq information	BioRxiv	March 2019
13	**snapATAC**	Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types	BioRxiv	April 2019
14	**STREAM**	Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM	Nature Communications	April 2019
15	**cisTopic**	cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data	Nature Methods	May 2019
16	**EpiScanpy**	EpiScanpy: integrated single-cell epigenomic analysis	BioRxiv	May 2019
17	**APEC**	APEC: an accesson-based method for single-cell chromatin accessibility analysis	BioRxiv	May 2019
18	**genesorteR**	genesorteR: Feature Ranking in Clustered Single Cell Data	BioRxiv	June 2019
19	**XenoCell**	XenoCell: classification of cellular barcodes in single cell experiments from xenograft samples	BioRxiv	June 2019
20	**ChromSCape**	ChromSCape: a Shiny&R application for interactive analysis of single-cell chromatin profiles	BioRxiv	July 2019
21	**rCASC**	rCASC: reproducible classification analysis of single-cell sequencing data	GigaScience	August 2019
22	**scBFA**	scBFA: modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data	Genome Biology	August 2019
23	**scAEspy**	scAEspy: a unifying tool based on autoencoders for the analysis of single-cell RNA sequencing data	BioRxiv	August 2019
24	**SCALE**	SCALE method for single-cell ATAC-seq analysis via latent feature extraction	Nature Communications	October 2019
25	**DC3**	DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data	Nature Communications	October 2019
26	**scfind**	Fast searches of large collections of single cell data using scfind	BioRxiv	October 2019
27	**scATAC-pro**	scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data	BioRxiv	October 2019
28	**Garnett**	Supervised classification enables rapid annotation of cell atlases	Nature Methods	October 2019
29	**AtacWorks**	AtacWorks: A deep convolutional neural network toolkit for epigenomics	BioRxiv	November 2019
30	**dryclean**	Robust foreground detection in somatic copy number data	BioRxiv	November 2019
31	**scOpen**	scOpen: chromatin-accessibility estimation of single-cell ATAC data	BioRxiv	December 2019
32	**UniPath**	UniPath: A uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles	BioRxiv	December 2019
33	**Augur**	Cell type prioritization in single-cell data	BioRxiv	December 2019

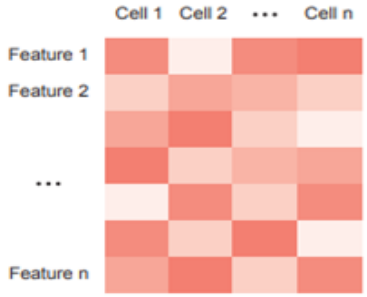
❖ scATAC-seq data are represent as various ‘cell-by-feature’ matrix

- After sequencing, the raw reads obtained in .fastq format for each single cell are mapped to a reference genome (by ***bowtie2***), producing aligned reads in .bam format.
- Finally, peak calling (by ***MACS2***) and read counting return the genomic position and the read count files in. bed and .txt format, respectively. Data in these file formats is then used for downstream analysis.



- **ATAC-seq peaks** in bulk samples can generally be recapitulated **in aggregated single-cell samples**, but not every single cell has a fragment at every peak.
- A **cell-by-peak matrix** can be constructed from single cells (e.g., by counting the number of reads at each peak for every cell).
- Other types of **cell-by-feature matrix** can be generated for downstream analysis: **cell-by-TF matrix, cell-by-gene matrix, cell-by-gene_activity matrix**

		cells	
	1	0	1
	1	1	2
peaks	1	2	1
	0	0	1
	1	1	0



❖ Challenges in scATAC-seq data analysis

Genome Biology **20:**241 (2019)

- Single-cell chromatin studies are **exceedingly sparse**, as normal diploid cells have two genomic copies and thus 0, 1 or 2 reads are observed per locus per cell. Due to low copy numbers (diploid in humans), lead to inherent data sparsity (**1–10% of peaks detected per cell**) compared to transcriptomic (scRNA-seq) data (10–45% of expressed genes detected per cell).
- While most **data analysis tools** share upstream pre-processing steps (i.e., alignment, peak calling, and counting), they **differ in obtaining a feature matrix for downstream analyses**.
- The potential feature set in scATAC-seq, which includes genome-wide regions of accessible chromatin, is **typically 10–20× the size of the feature set in scRNA-seq** experiments (which is defined and limited by the number of genes expressed). This **larger feature set** could be **valuable in distinguishing a wider variety of cell populations and inferring the dynamics** underlying cell organization into complex tissues.

❖ *Benchmarking scATAC-seq tools for clustering*

Genome Biology **20:**241 (2019)

- **SnapATAC**, ***Cusanovich2018***, and **cisTopic** outperform other methods in separating cell populations of different coverages and noise levels in both synthetic and real datasets.
- Notably, **SnapATAC** is the only method **able to analyze a large dataset** (> 80,000 cells).
- In addition, **SnapATAC** is easy to use and carry **many different functions**.
- Because of accuracy, scalability, and versatility, we recommend SnapATAC for pre-processing.

- The **feature matrix construction** can be roughly summarized into four different modules:

  (i) define regions, (ii) count features, (iii) transformation, (iv) dimensionality reduction

- ***Define regions***
  - An essential aspect of feature matrix construction is the selection of a set of regions to describe the data (e.g., putative regulatory elements such as peaks and promoters).
  - **Most methods** define regions based on **peak** calling from either a reference bulk ATAC-seq profile or an aggregated single-cell ATAC-seq profile.
  - ***Cusanovich2018* and SnapATAC** segment the genomes into fixed-size **bins** (windows) and count features within each bin.
  - *Cusanovich2018* first creates **pseudo-bulk** clades by performing hierarchical clustering on the transformed matrix using the top frequently accessible windows. Then, **peaks** are called **by aggregating cells within each pseudo-bulk clade**.

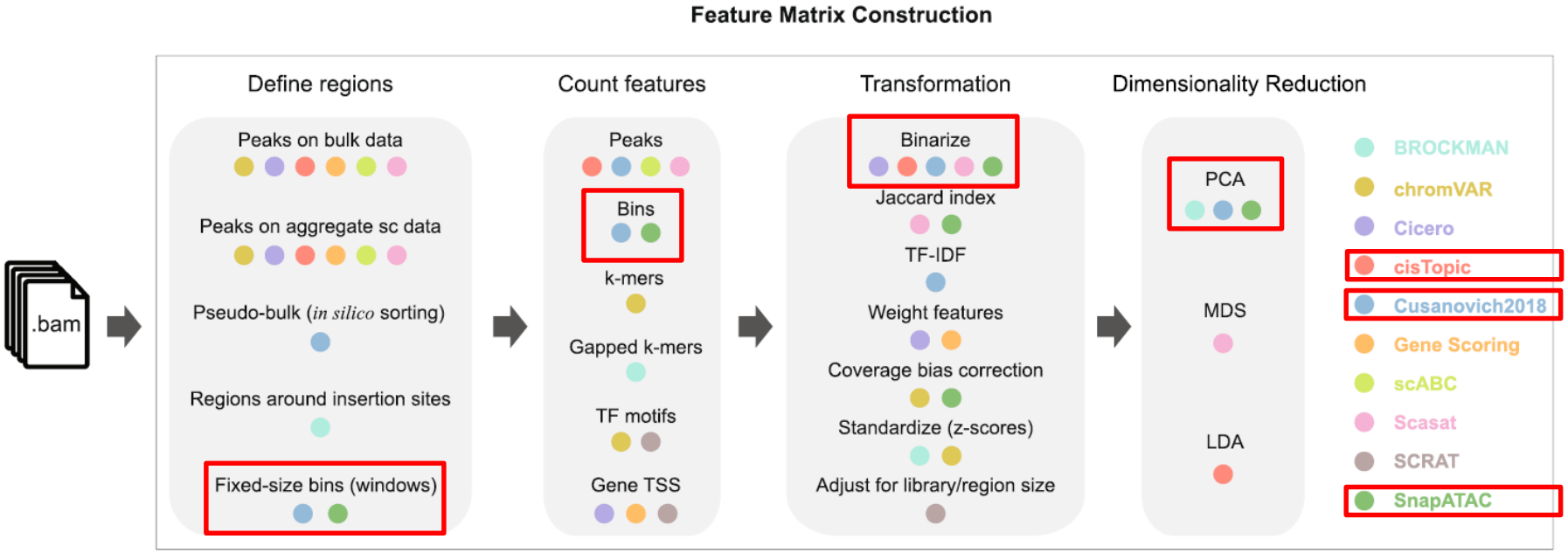
- ***Count features***
  - Raw features within the defined regions are counted: **Peaks, bins, k-mers, TF motifs, gene TSS**.
  - For cisTopic and *Cusanovich2018*, **reads overlapping peaks** are counted. For *Cusanovich2018* and SnapATAC, **reads overlapping bins** are counted. Similarly, for chromVAR, **reads overlapping TF motifs and k-mers** are counted.
  - If pre-defined genomic annotations such as **coding genes** are given, **Gene Scoring** and **Cicero** use gene TSSs as anchor points to calculate gene enrichment scores based on reads nearby or just within peaks nearby.

- ***Transformations***
  - To the initial raw feature matrix, different transformation methods can be performed.
  - **Binarization** of read counts is used by many tools including best performed ones. This step is based on the assumption that each site is present at most twice. Binarization is advantageous in alleviating technical noise arising from sequencing depth or PCR amplification artifacts.
  - SnapATAC convert the binary count matrix into a cell-pairwise **Jaccard index similarity** matrix.
  - *Cusanovich2018* normalizes the binary count matrix using the **TF-IDF** transformation.
  - Cicero **weights feature** sites by their co-accessibility, while Gene Scoring weights sites by a decaying function based on its distance to a gene TSS. → **Gene activity score**
  - Both chromVAR and SnapATAC perform a **read coverage bias correction** to account for the influence of sample depth. chromVAR creates “background” peaks consisting of an equal number of peaks matched for both average accessibility and GC content to calculate bias-corrected deviation while SnapATAC uses a regression-based method to mitigate the coverage difference between cells. chromVAR compute *z*-scores to measure the gain or loss of chromatin accessibility across cells.

- ***Dimensionality reduction***
  - In the final step before downstream analysis, several methods apply dimensionality reduction. This step can refine the feature space mitigating redundant features and potential artifacts and potentially reducing the computation time of downstream analysis.
  - **PCA** is the most commonly used method (used by SnapATAC, and *Cusanovich2018*). cisTopic uses **latent Dirichlet allocation (LDA)** to generate two distributions including topic-cell distribution and region-topic distribution. Choosing the top topics based on the topic-cell distribution reduces the dimensionality.

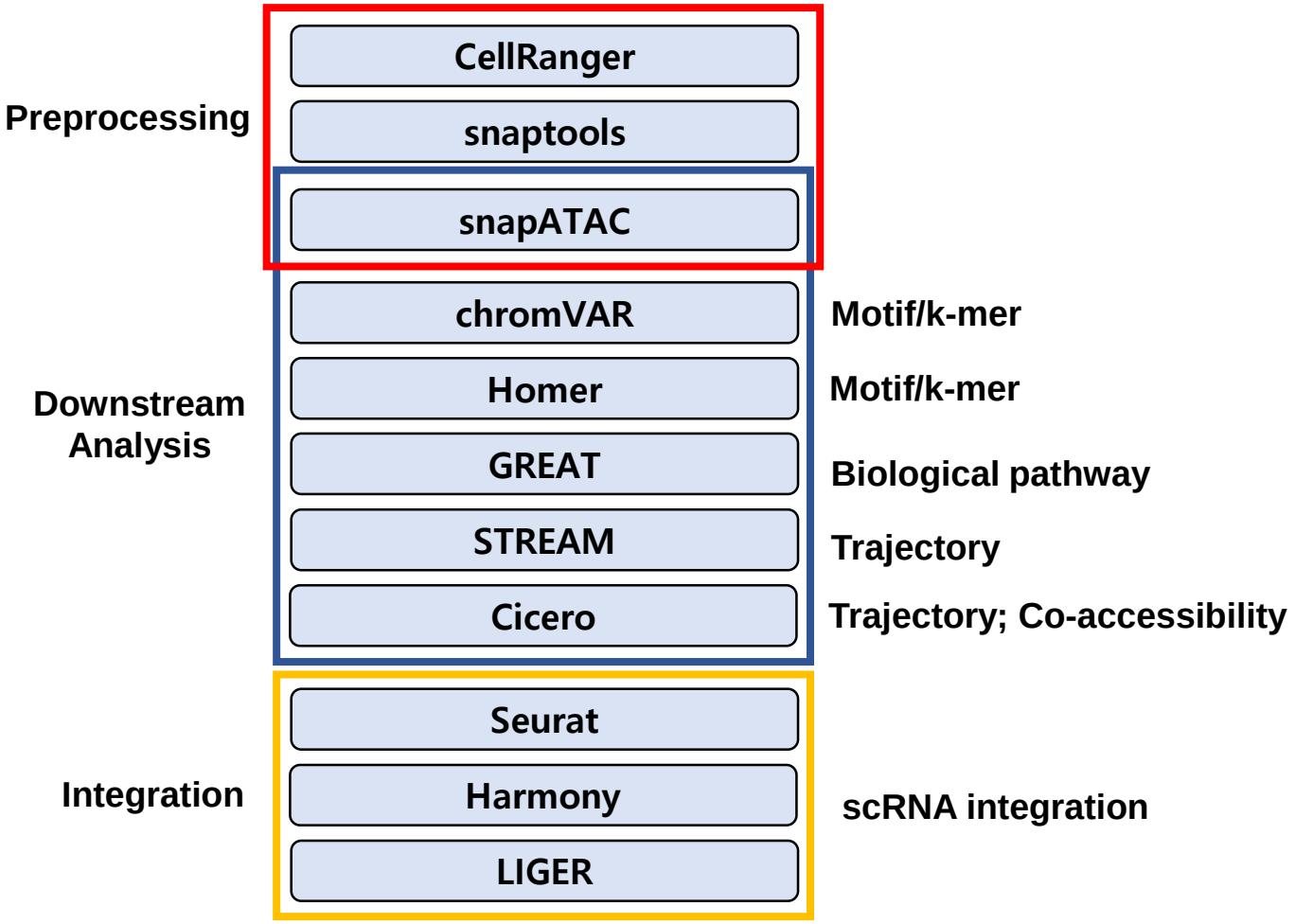
➢ *Conclusions from benchmarking study (for clustering efficiency)*

1. **Peak-level or bin-level feature** counting generally performs better. This may indicate that the complexity of gene regulatory circuits where precise enhancer elements may have distinct functions that cannot be sufficiently approximated by sequence context or proximity to gene bodies alone.
2. **Dimensionality reduction** step generally helps the separation of cell types, since this step may help to remove the redundancy between a large number of raw features and to mitigate the effect of noise.
3. The robustness of different methods to noise and coverage varies among different datasets. Among the top three methods, **cisTopic is the most penalized by low coverage**.
4. Inappropriate transformations, such as **log2 transformation** and **normalization based on region size** as implemented in SCRAT, may **impact negatively** clustering performance.
5. Louvain (**community detection-based clustering**) method overall performs more consistently and accurately than others for scATAC-seq data.



Genome Biology **20:**241 (2019)

❖ scATAC-seq data analysis software



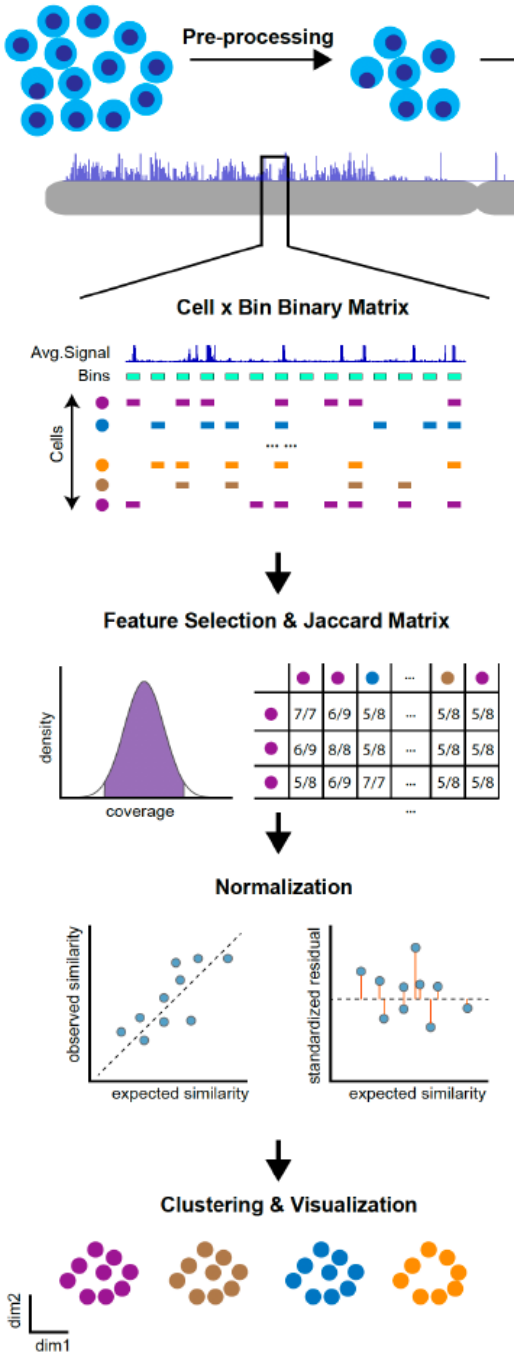
❖ snapATAC

Rongxin Fang, et al. *bioRxiv* (2019)

- The use of pre-defined accessibility peaks based on bulk data has at least three key limitations.

1. It requires sufficient number of single cell profiles to create robust aggregate signal for peak calling.
2. The cell type identification is biased toward the most abundant cell types in the tissues.
3. It lacks the ability to reveal regulatory elements in the rare cell populations which are underrepresented in the aggregate signal.

- SnapATAC does not require population-level peak annotation, and instead assembles chromatin landscapes by directly clustering cells based on the similarity of their genome-wide accessibility profile (profile based on uniform-sized bins that segmented the genome).
- **Binarization:** Each bin has value “1” if one or more reads fall within that bin. Otherwise, value “0” is assigned → **cell-by-bin binary matrix**
- **Feature selection**: remove invariant features (e.g., remove top 5% bins and bins with 0 reads)
- Binary vectors from all the cells is converted into a **Jaccard index matrix**. Because Jaccard Index can be influenced by sequencing depth, regression-based normalization procedure was applied to remove such confounding factor.
- The normalized matrix is subject to **Dimension reduction** and significant components are selected for clustering analysis.

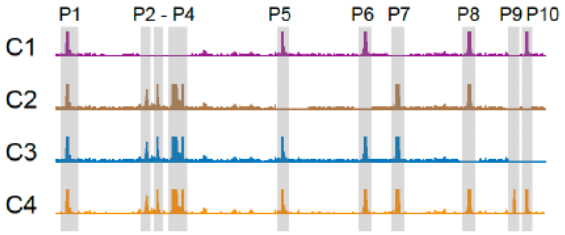


❖ Peak identification and downstream analysis

- **Louvain clustering** analysis groups cells with similar chromatin accessibility profiles.
- **Peak Calling** using MACS2: cells belonging to the same cluster are aggregated to create a representation of cell-type specific regulatory landscape for identification of candidate *cis*-regulatory elements *de novo*.
- **Peak Occurrence Frequency Matrix**: the frequency (number of cells out of the total) of a peak occurring in each cluster is calculated.
- **Cell-by-Peak matrix**: Merging peaks identified from each cluster, we create a reference list of regulatory elements. Using this reference map, we next create a cell-by-peak matrix.
- **Motif analysis**: identify overrepresented motif (or k-mer) within accessible regions for each cluster (cell type) using homer or chromVAR
- **Pathway analysis**: performed Genomic Region Enrichment Analysis (GREAT) to predict the function of each cluster.

Clustering & Visualization

Peak Calling



Peak Occurrence Frequency

peaks →
clusters ↓

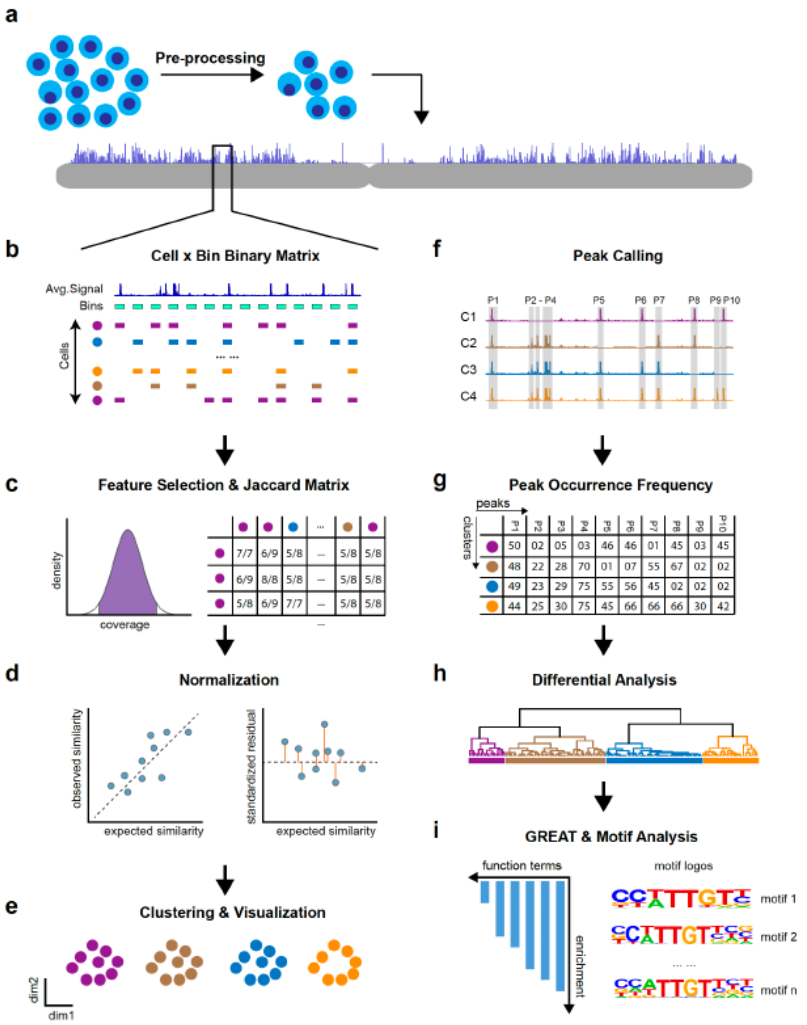
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
●	50	02	05	03	46	46	01	45	03	45
●	48	22	28	70	01	07	55	67	02	02
●	49	23	29	75	55	56	45	02	02	02
●	44	25	30	75	45	66	66	66	30	42

Cell-by-peak matrix

		cells	
	1	0	1
	1	1	2
peaks	1	2	1
	0	0	1
	1	1	0

❖ snapATAC overview

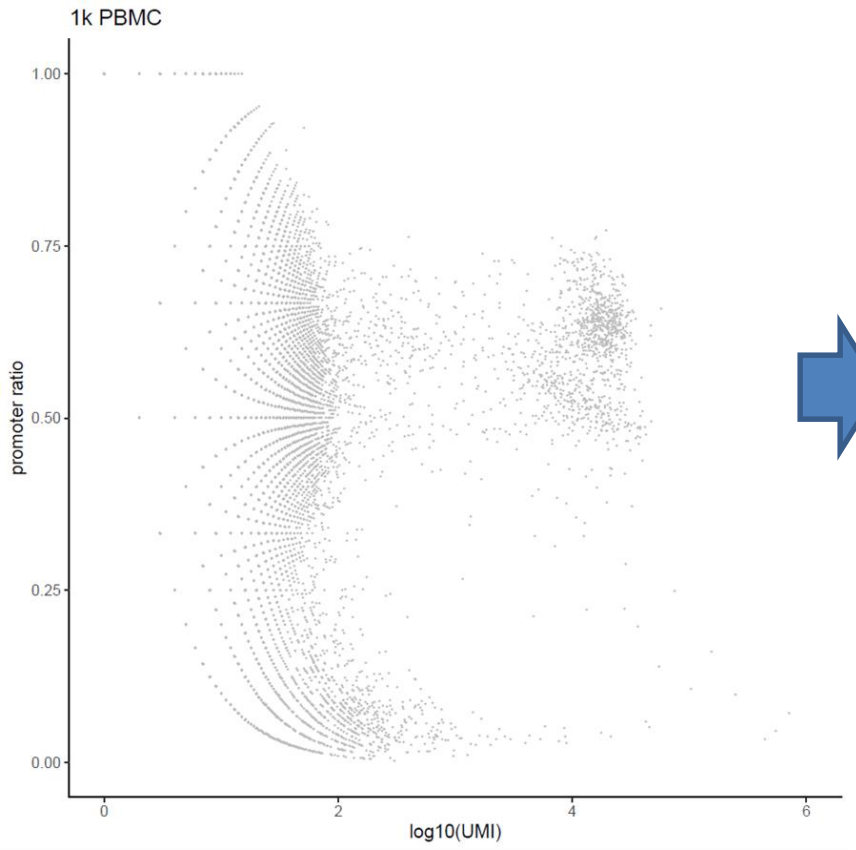
- **SnapTools (python)**
  - Index Reference Genome
  - Fastq Alignment
  - Create snap file from bam file
  - Cell by Bin matrix
- **SnapATAC (R)**
  - Barcode selection
  - Cell-by-bin matrix: **bmat**
  - Bin filtering
  - Dimensionality reduction
  - Clustering
  - Visualization
  - Gene-based annotation / Seurat variable Cell-by-gene matrix: **gmat**
- **SnapTools + MACS (python)**
  - Peak calling of each cluster
  - Create combined peaks (R)
  - Cell-by-peak matrix: **pmat**
- **SnapATAC (R)**
  - Identify differentially accessible regions(DAR)
  - Motif analysis – master regulators using Homer & chromVAR motif variability analysis
  - GREAT analysis for identifying biological pathways

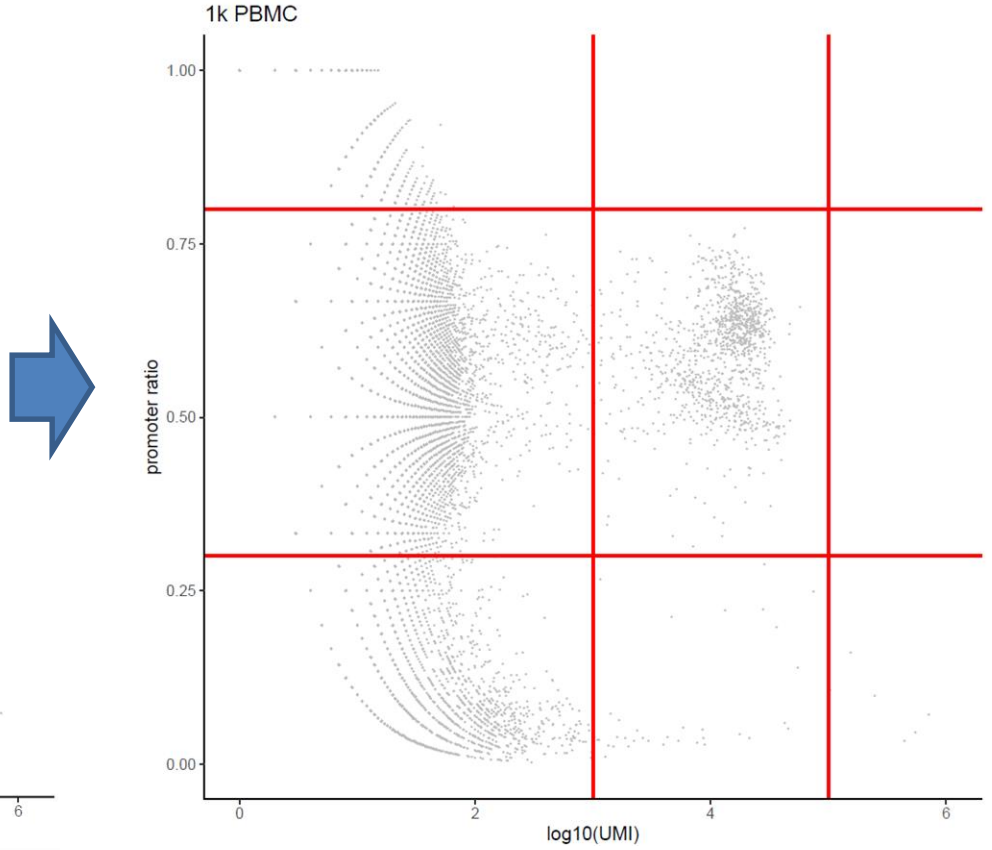


❖ Preprocessing

▪ Barcode selection (cell selection)

- Calculate promoter ratio and log(UMI)
- Plot cells with promoter ratio and log(UMI)
- Filter cells with read count 1k ~ 100k & promoter read ratio: 30% ~ 80% (adjusted by the plot)





- **Loading Cell-by-Bin matrix, and Binarization**

- **Quality Control: Bin filtering**
  - Blacklist genes (using Encode hg19 consensus signal artifact regions)
  - Unwanted chromosome (mitochondrial chromosome)
  - Top 5% bins that overlap with invariant features (house-keeping gene promoter)
  - Low bin coverage: Cells with bin coverage less than 1000

- **Dimension Reduction with DiffusionMaps**
  - Determine significant components (how many dimensions to include)

- **Clustering cells based on significant components**
  - Louvain clustering (graph-based clustering)

❖ Cell identity annotation with scATAC-seq data

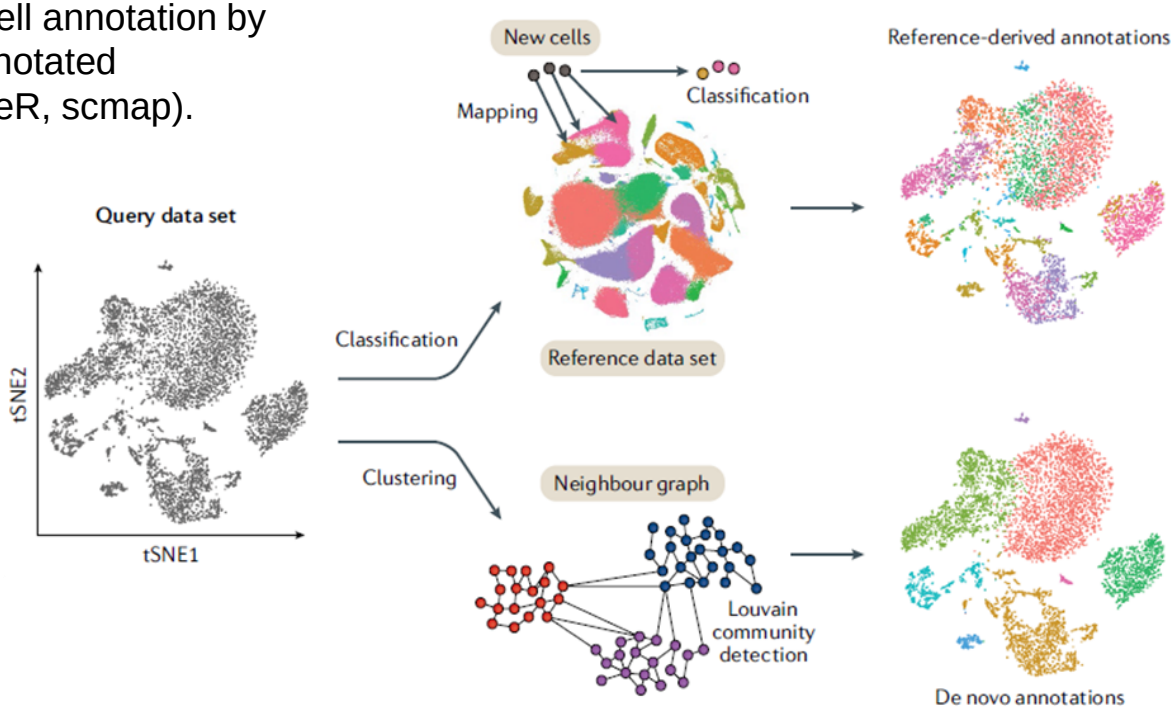
- Methods for scRNA-seq are similarly useful for annotating scATAC-seq datasets based on scATAC-seq-derived "**gene activity scores**" (using Cicero).

▪ *De novo* annotation: clustering → label with marker genes

- Each cluster is annotated by meaningful biological label such as **marker genes** for cell types.
- However, definition of cell type is not clear. Furthermore, cells of the same cell type in different states may be detected in separate clusters. For these reasons, it is best to use the term "**cell identities**" rather than "cell types".
- External sources of information of marker genes (e.g., databases and literature) can be used to annotate clusters.

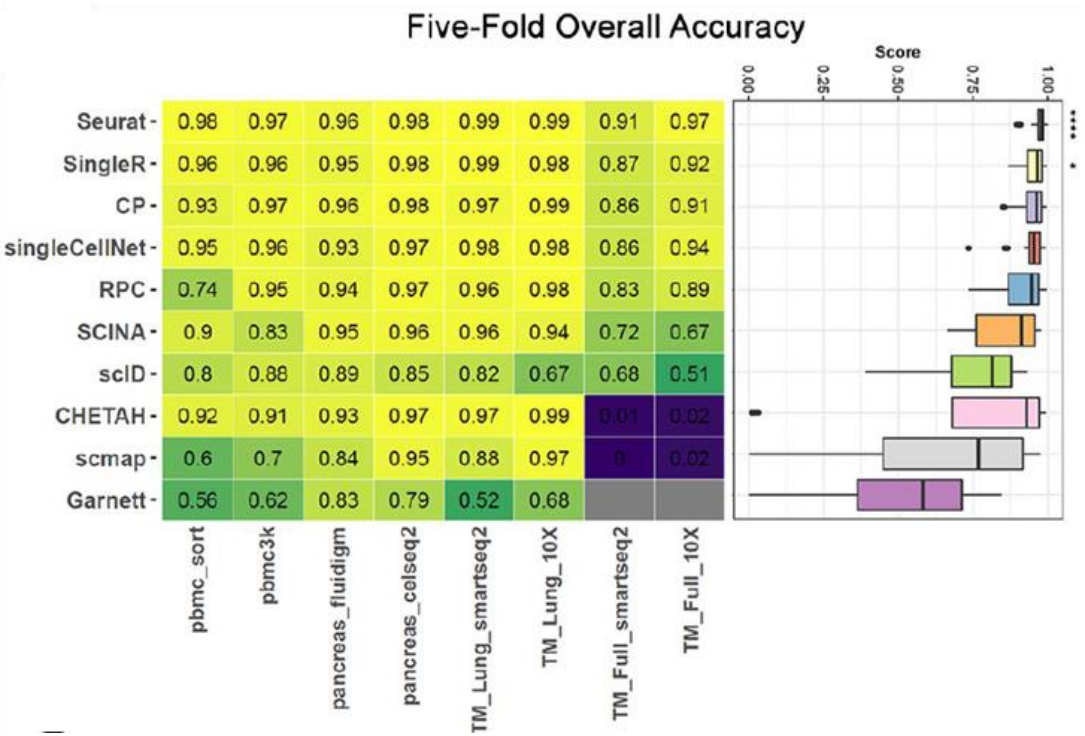
▪ Reference-derived annotation: classification based on cell atlases

- There are tools for automated cell annotation by classifying individual cells to annotated **reference cell atlas** (e.g., singleR, scmap).
- With well-established cell atlases (e.g. HCA), this approach could be more powerful.



Nature Reviews Genetics **20:** 257 (2019)

- **Evaluation of Cell Type Annotation software on scRNA-seq Data**
  - Overall, **Seurat**, **SingleR**, CP, and SingleCellNet performed well.
  - **Seurat** works the **best at annotating major cell types**.
  - However, **Seurat** does have a major **drawback at predicting rare cell populations**, and it is suboptimal **at differentiating similar cell types**, while **SingleR** and CP are much **better in these aspects**.
  - **Seurat**, **SingleR** and CP are more robust against down-sampling.



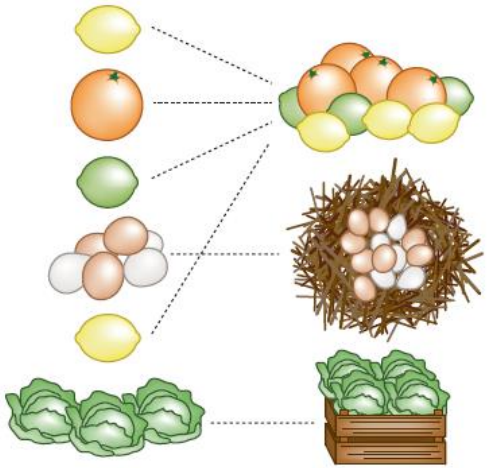
Software	Method/Algorithm	Bulk/Single Reference
SingleR	Correlation-based with Iterative Tuning	Bulk
CP	Reference-based method using Constrained Projection	Bulk
RPC	Reference-based Robust Partial Correlations	Bulk
Garnett	Elastic net Multinomial Regression	Single
SCINA	Bimodal Distribution assumption for marker genes	Single
Seurat	Define anchor with CCA, L2-norm and MNN	Single
singleCellNet	Multi-Class Random Forest	Single
CHETAH	Correlation-based with Hierarchical Classification	Single
scmap	K-nearest-neighbor classification with cosine similarity	Single
scID[a]	Fisher's Linear Discriminant Analysis-like Framework	Single

https://www.biorxiv.org/content/10.1101/827139v1

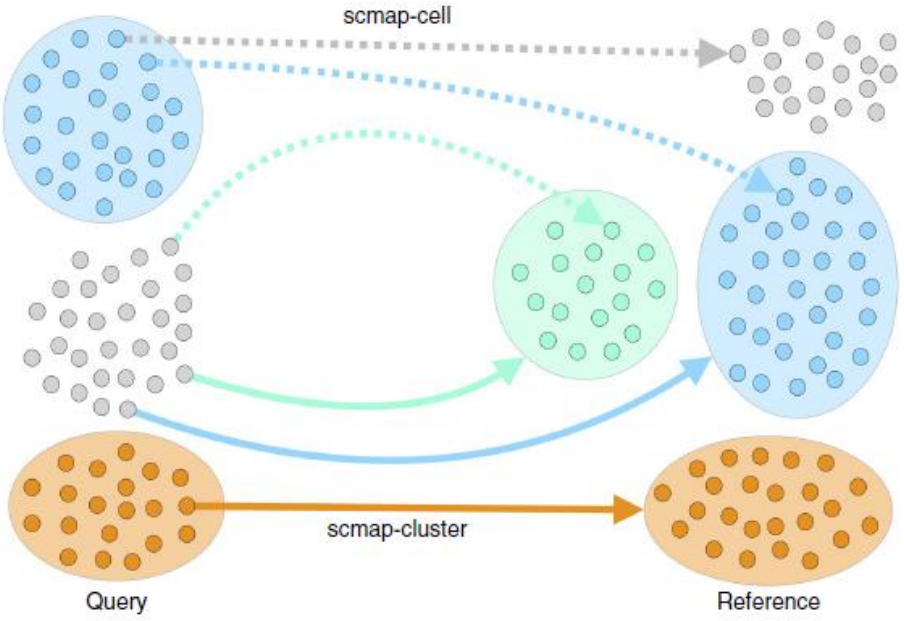
- ***scmap*** *Nature Methods* **15**: 359 (2018) – **first-in-class software**

  - As large references such as the Human Cell Atlas (HCA) become available, it will be important to project cells from a new sample (e.g., disease tissues) onto the references to characterize differences in composition or detect new cell types. Conceptually, such projections are **similar to the BLAST search**, which quickly finds the closest match in a database of nucleotide or amino acid sequences.
  - scmap can **map individual cells from a query sample to cell types in the reference** (scmap-cluster) **or to individual cells in a reference** (scmap-cell). Scmap-cluster is more robust and faster than scmap-cell.
  - In scmap-cluster, **each cluster is represented by its centroid** (a vector of the median value of the expression of each gene) and measure the similarity between a new cell, *c,* and each cluster centroid or cell. The nearest cluster can be searched for exhaustively because the number of clusters is typically much smaller than the number of cells in the reference.

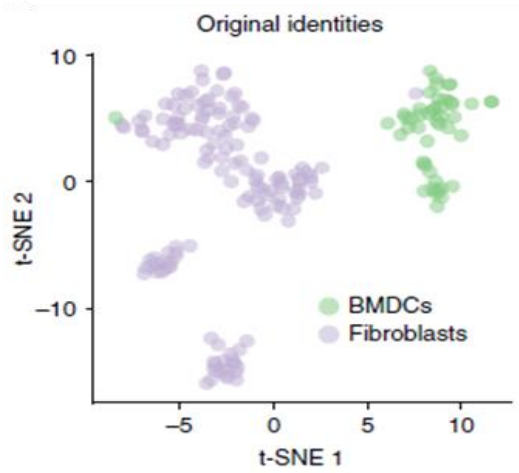
Concept behind scmap

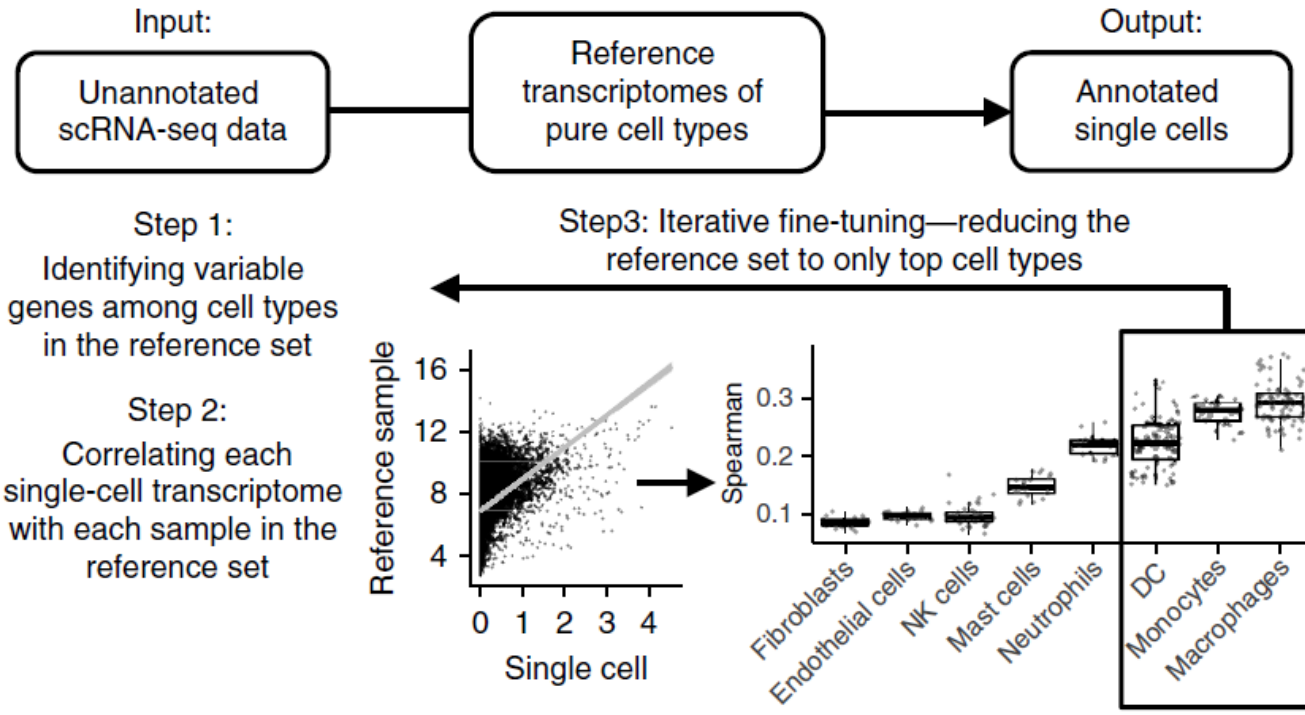


Nature Methods **15**: 321 (2018)

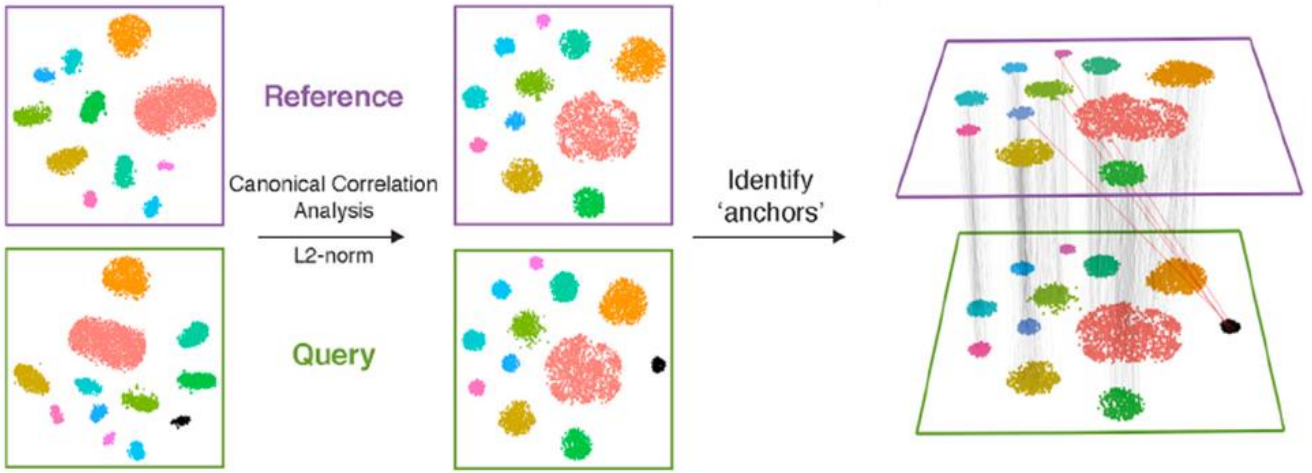


- ***SingleR*** *Nature Immunology* **20**: 163 (2019)
  - De novo annotation of clusters of cells using known marker genes is performed manually. This strategy suffers from subjectivity and limits adequate differentiation of closely related cell subsets. We need unbiased cell type recognition of scRNA-seq.
  - SingleR leverages **reference transcriptomic datasets of pure cell types** (human and mouse bulk RNA-seq samples) to infer the cell of origin of <u>each of the single cells independently</u>.





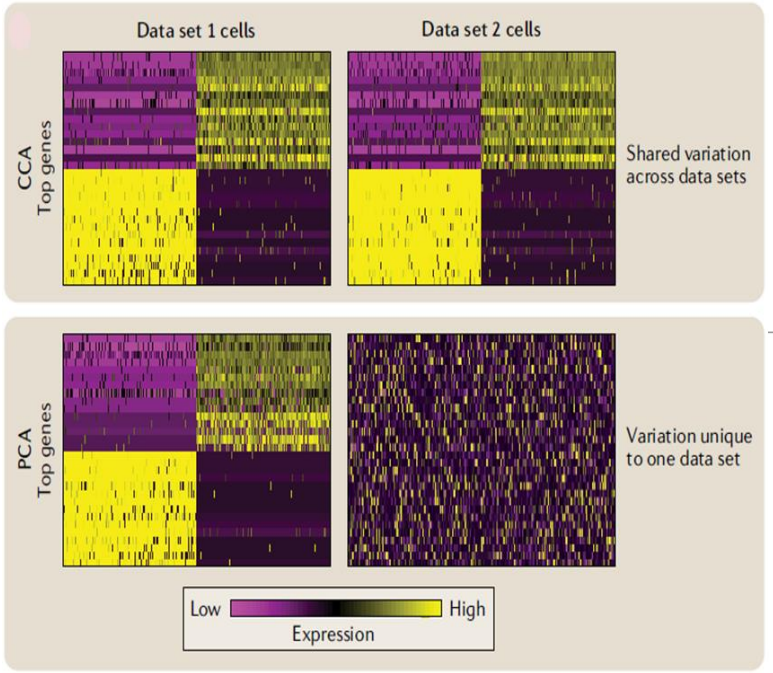
- **Seurat v3:** *Canonical Correlation Analysis* (CCA) + *Mutual Nearest Neighbors* (MNN)
  - *CCA → L2 normalization* of canonical correlation vectors → *Project the datasets into* a subspace defined by *shared correlation structure across datasets*.
  - In the shared space, **identify pairs of MNNs across reference and query cells**. These should represent cells in a shared biological state across datasets (gray lines) and serve as ***anchors*** to guide dataset integration.
  - While MNNs have previously been identified using L2-normalized gene expression, significant differences across batches can obscure the accurate identification of MNNs, particularly when the batch effect is on a similar scale to the biological differences between cell states. To overcome this, we first jointly reduce the dimensionality of both datasets using diagonalized CCA, then apply L2-normalization to the canonical correlation vectors.
  - We next search for MNNs in this shared low-dimensional representation. We refer to the resulting cell pairs as anchors, as they encode the cellular relationships across datasets that will form the basis for all subsequent integration analyses.
  - Anchors can successfully recover matching cell states even in the presence of significant dataset differences, as CCA can effectively identify shared biological markers and conserved gene correlation patterns. However, cells in non-overlapping populations should not participate in anchors.

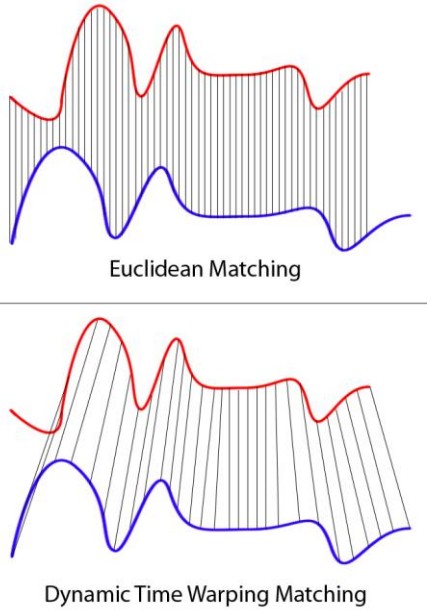


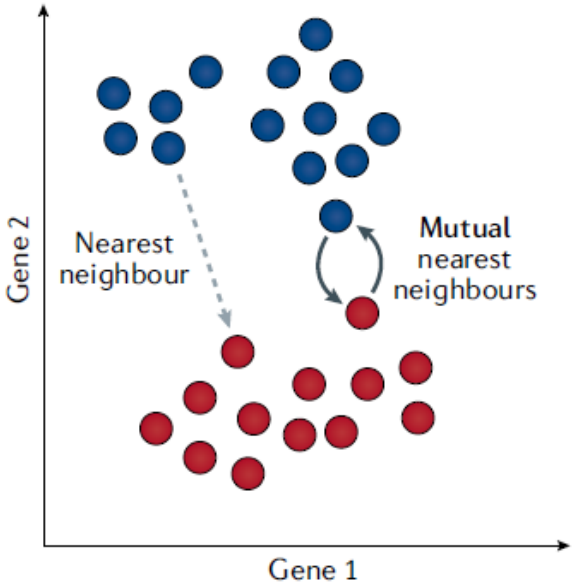
Cell **177:** 1888 (2019)

- **CCA aims to identify a set of variables that are maximally correlated between two data sets**. By contrast, methods such as principal component analysis (PCA) aim to find orthogonal variables that maximize the variance explained in a single dataset.
- Next, these canonical correlation vectors are aligned across data sets using ***dynamic time warping*** (a method for locally stretching or compressing two 1D vectors to correct for lag in one vector relative to another), a nonlinear transformation that corrects for differences in cell population density.

- **MNN** aims to **identify cells that are mutually nearest to one another in a space**, defined by the gene expression profiles of the cells, allows the identification of biologically equivalent cells.
- Once equivalent cells have been identified across data sets, this information can be used to compute a transformation of the original expression data that would remove data-set-specific expression patterns.

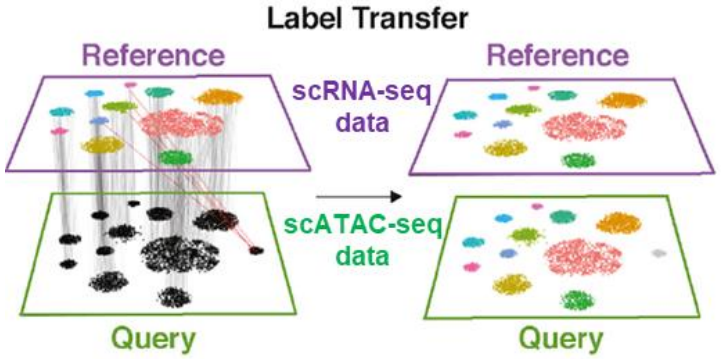


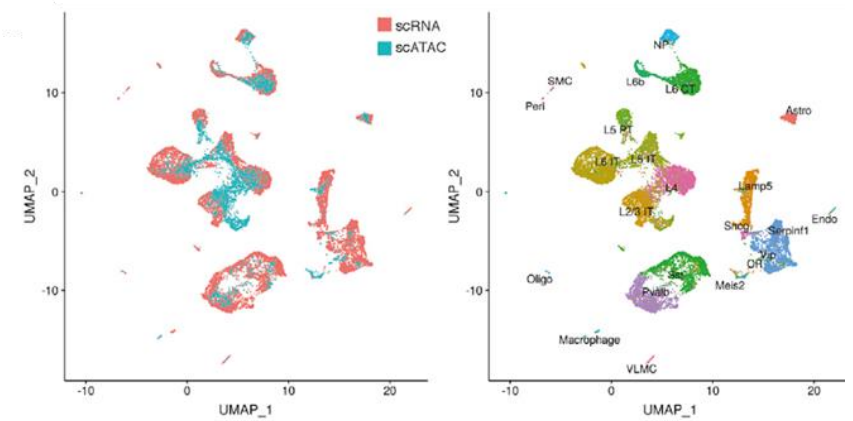


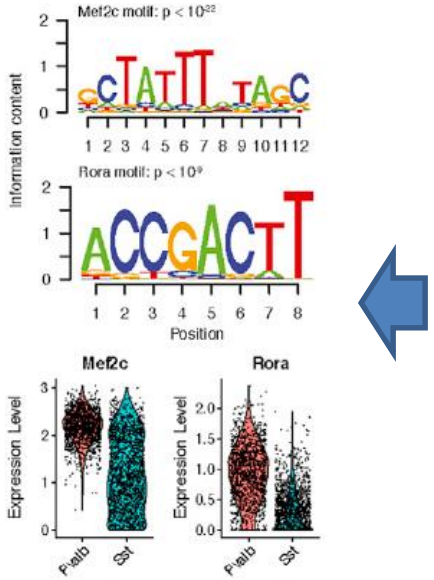


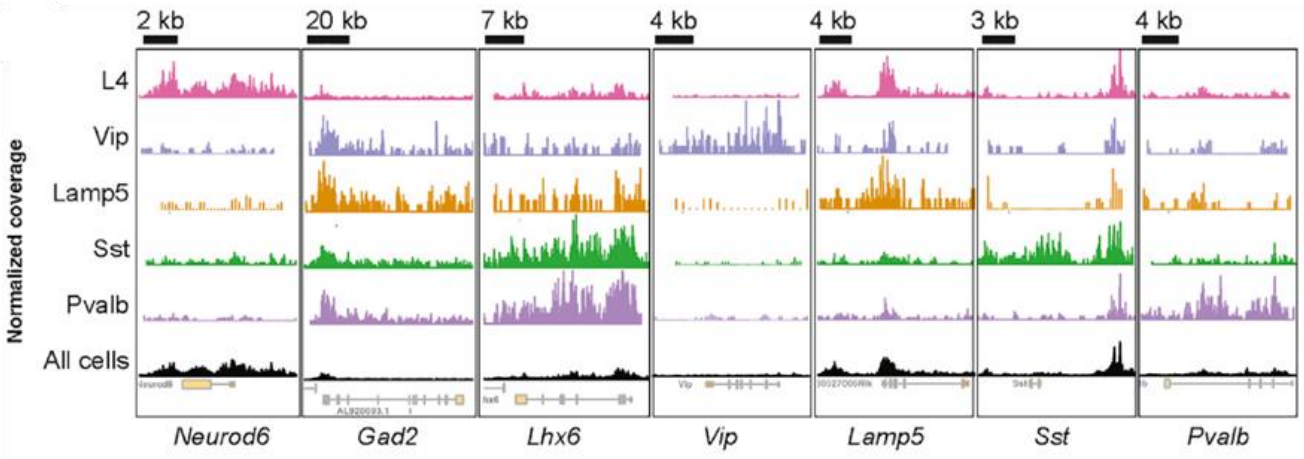
Nature Reviews Genetics **20:** 257 (2019)

- **Cell type annotation using scRNA-seq reference and Seurat v3**
  - Seurat v3 can transfer cell label from scRNA-seq-based reference cells to scATAC-seq cells via data integration. The integration with reference improves cell type annotation by scATAC-seq alone.
  - We create pseudo-bulk ATAC-seq profiles by pooling together cells within each cell type. Each cell type showed enriched accessibility near canonical marker genes.
  - Then, we identify overrepresented DNA motifs in cell-type-specific accessibility peaks (e.g., Mef2c motif overrepresented in PV-specific accessibility peaks and Mef2c expression is upregulated).

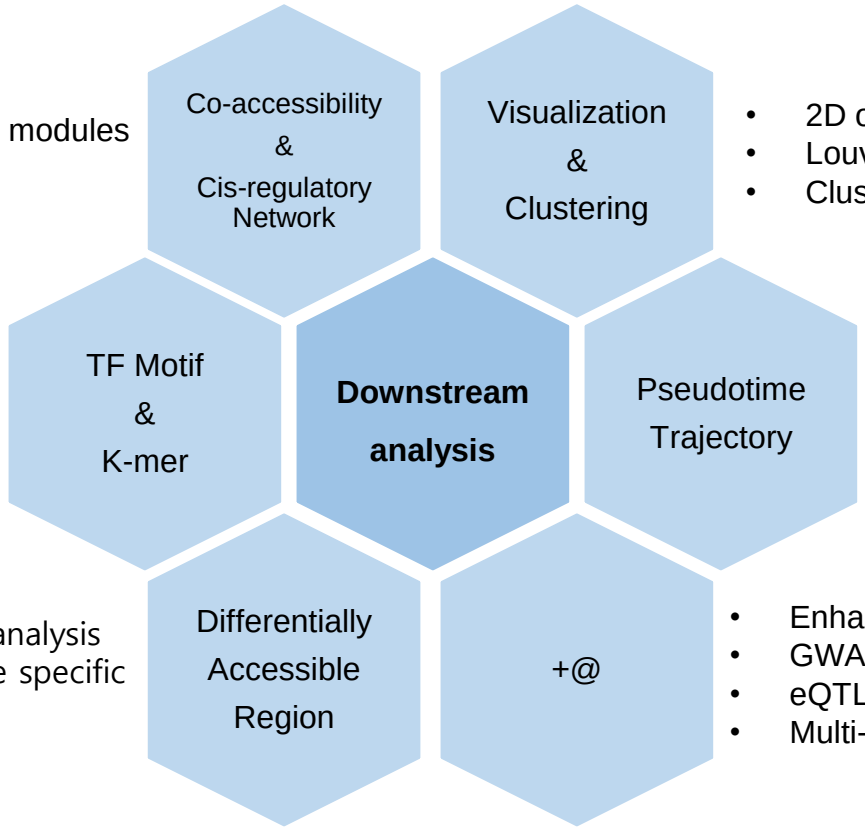








❖ Downstream analysis



- Highly co-accessible modules
- Gene-activity score
- Cicero

- 2D or 3D plots
- Louvain algorithm
- Cluster annotation (cell types)

- Enriched TF motifs & novel motifs with k-mer within accessible regions
- chromVAR, Homer

- STREAM
- Cicero

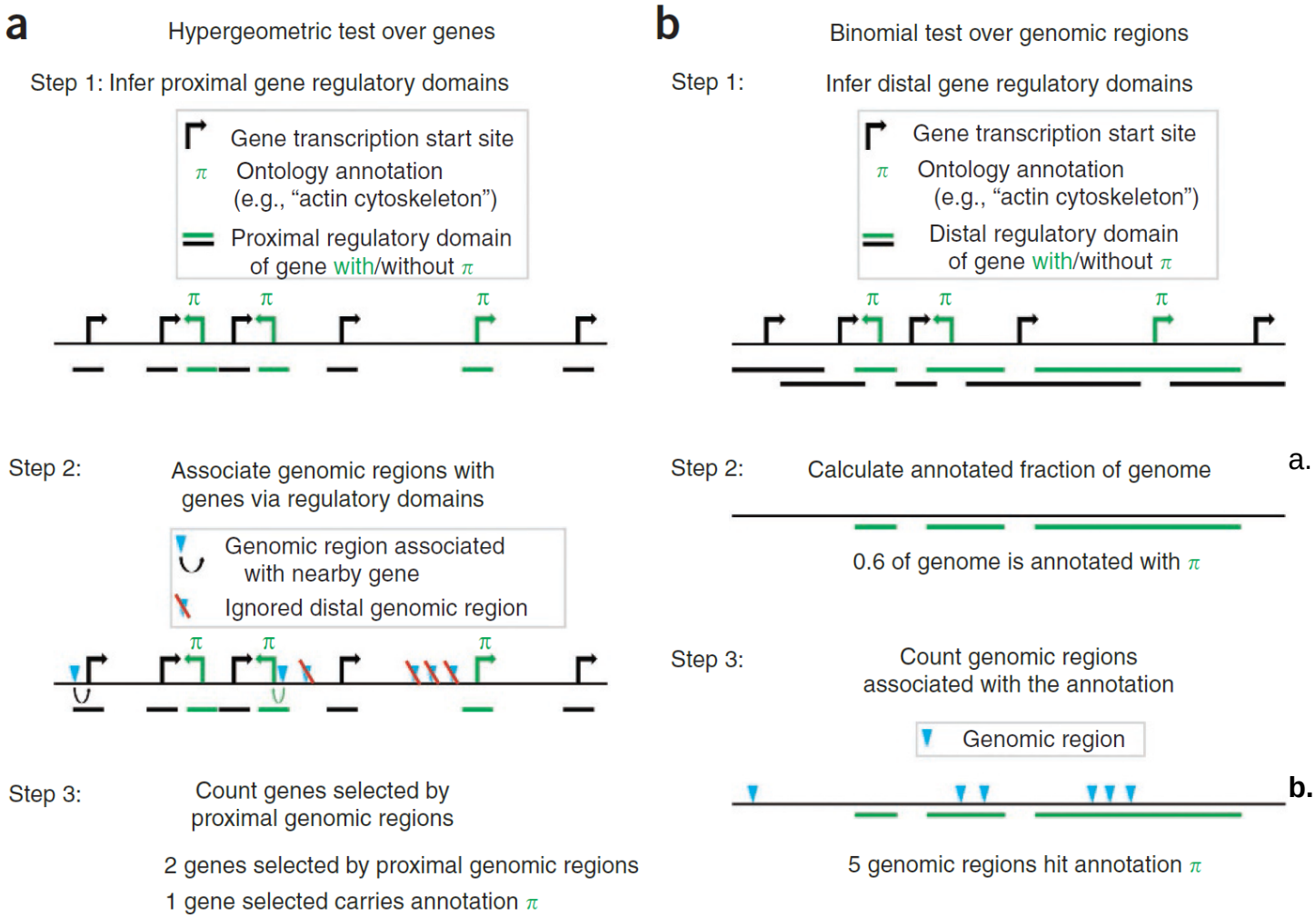
- Similar to DEG analysis
- Cell type / stage specific

- Enhancer
- GWAS (SNP)
- eQTL
- Multi-omics Integration

❖ GREAT (Genomic Regions Enrichment of Annotations Tool)

Nat. Biotech **28**:495 (2010)

- Previously, functional significance analysis of cis-regulatory regions identified from ATAC (or ChIP, Hi-C) analysis across an entire genome used only proximal to genes and performs a gene-based hypergeometric test.
- However, associating only proximal binding events (for example, under 2–5 kb from the transcription start site) typically discards over half of the observed chromosomal regions.
- The standard approach to capturing distal events—associating each binding site with the one or two nearest genes—introduces a strong bias toward genes that are flanked by large intergenic regions (e.g., GO term 'multicellular organismal development' is associated with 14% of human genes, but the 'nearest genes' approach associates >33% of the genome with these genes).
- GREAT associates genomic regions with genes by defining a '**regulatory domain**' for each gene in the genome.
- GREAT assigns each gene a regulatory domain **consisting of a basal domain** that extends 5 kb upstream and 1 kb downstream from its transcription start site (denoted below as 5+1 kb), **and an extension** up to the basal regulatory domain of the nearest upstream and downstream genes within 1 Mb.
- Given a set of input genomic regions and an ontology of gene annotations, GREAT computes ontology term enrichments using a **binomial test** that explicitly accounts for variability in gene regulatory domain size by measuring the total fraction of the genome annotated for any given ontology term and counting how many input genomic regions fall into those areas.
- Therefore, the longer the regulatory domain of any gene—and, by extension, of any ontology term—the greater the expected number of regions associated with this term by chance. The binomial statistic markedly **reduces the number of false positive enriched terms** even when very large regulatory domains are used.



a. <u>Previous methods</u> associates only proximal binding events with genes and performs a gene-list test of functional enrichments using **hypergenometric test**.

b. **GREAT's binomial approach** over genomic regions uses the total fraction of the genome associated with a given ontology term (green bar) as the expected fraction of input regions associated with the term by chance.

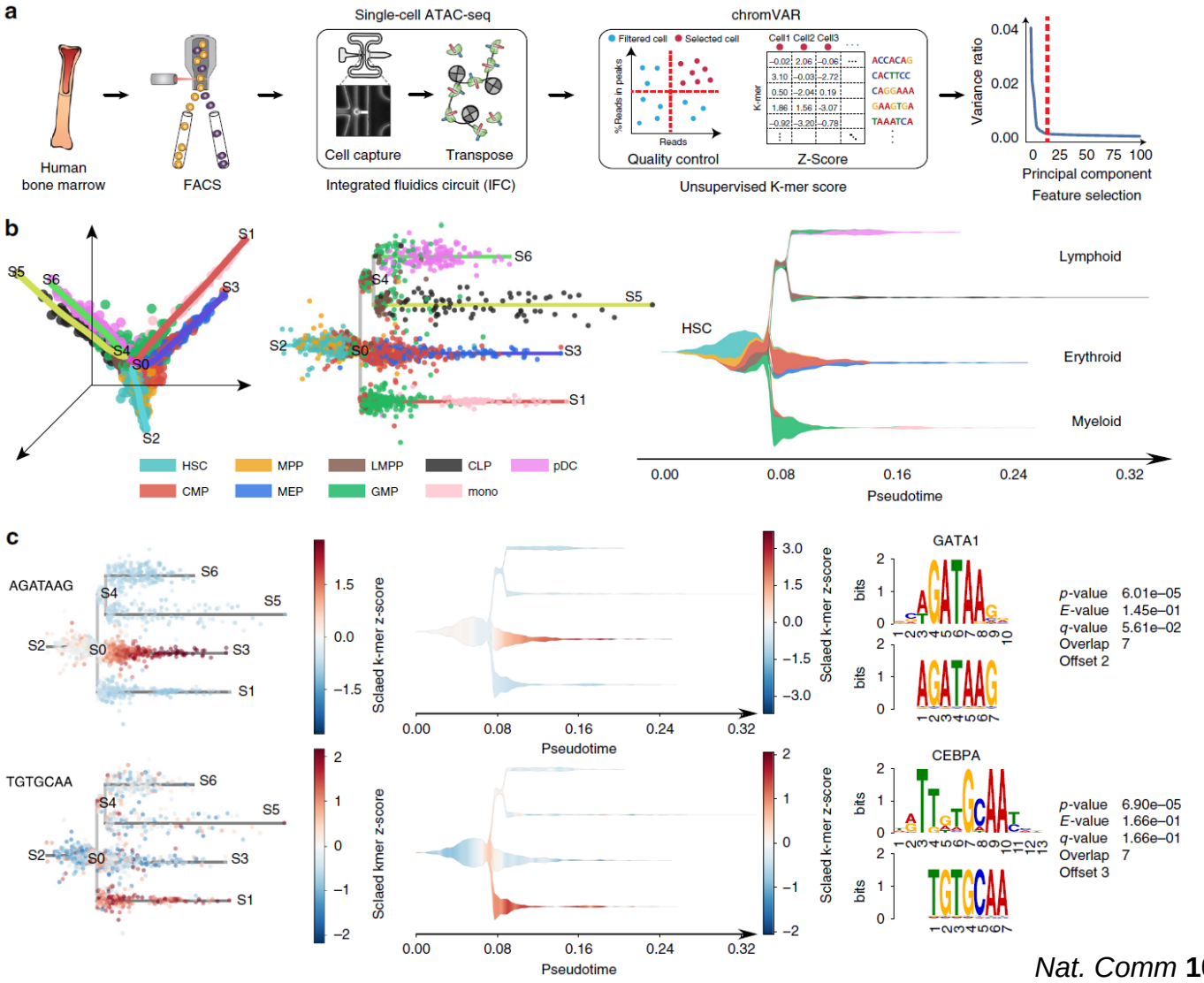
Nat. Biotech **28**:495 (2010)

❖ STREAM (single-cell trajectories reconstruction, exploration and mapping)

a. It uses **cell-by-k-mer matrix** with z-scores calculated from chromVAR. **Use top PCs** for the trajectory analysis

b. Subway map plot of mouse hematopoietic cells.

c. It **identifies significant 7-mers for each branch**. These 7-mers are **then mapped to known TF motifs**.

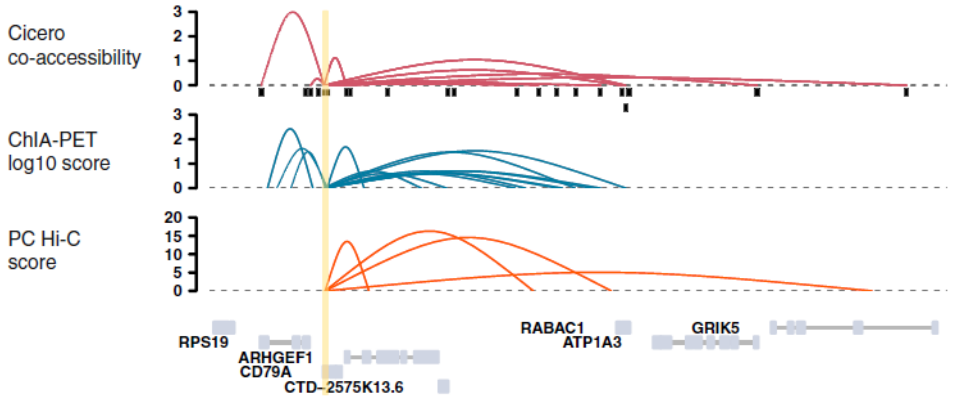


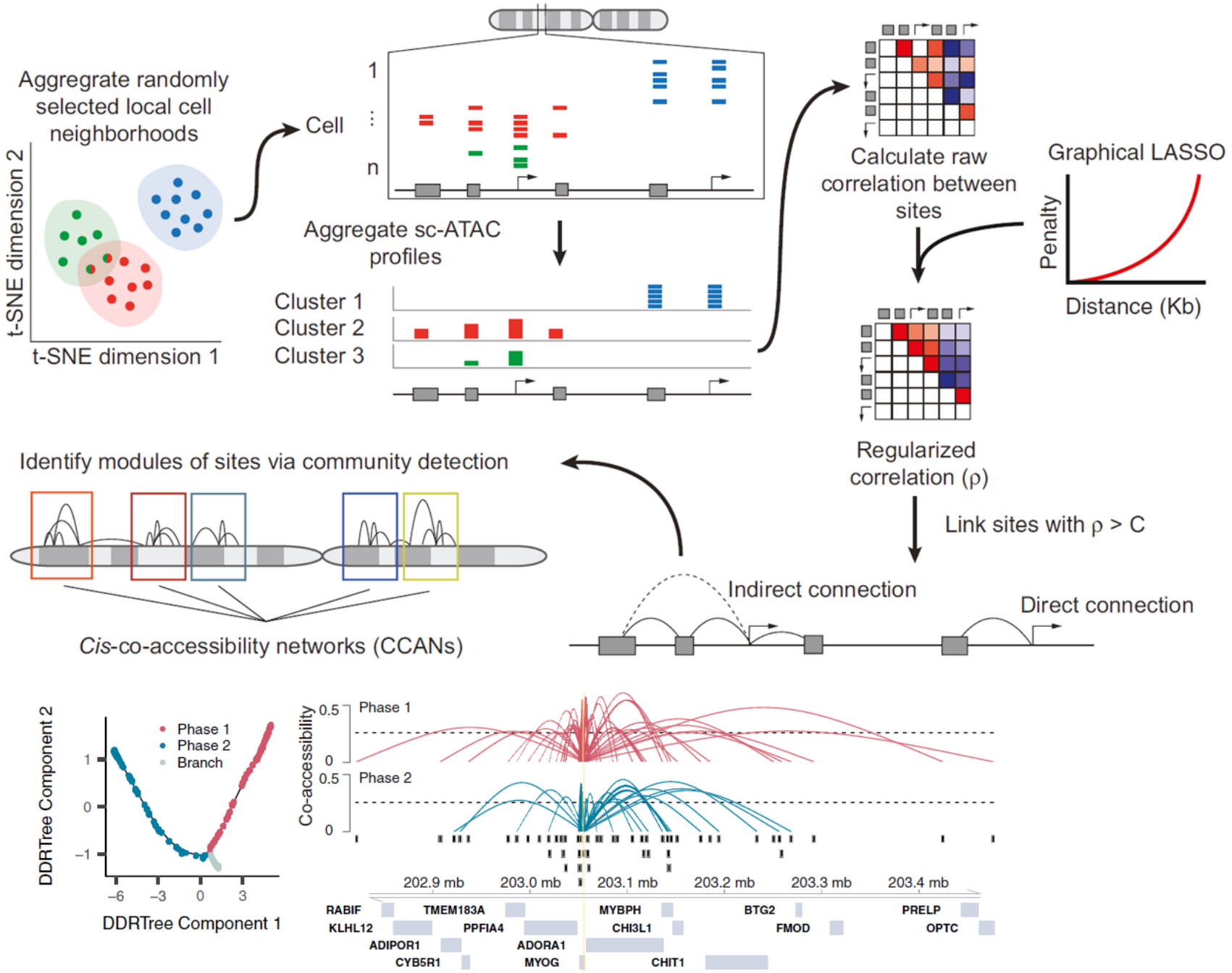
Nat. Comm **10**:1903 (2019)

❖ Cicero (co-accessibility analysis)

Molecular Cell **71**:858 (2018)

- Cicero aims to identify all pairs of **co-accessible sites**.
- Because of the sparsity of single-cell data, cells must be aggregated by similarity to allow robust correction for various technical factors in the data. Cicero does this using a k-NN approach which creates overlapping sets of cells. Cicero constructs these sets based on a reduced dimension coordinate map of cell similarity, for example, from a tSNE or DDRTree map.
- First, **groups of highly similar cells are sampled** using the clustering or pseudotemporal ordering, and their binary profiles are aggregated into integer counts.
- Cicero computes the raw covariances between each pair of sites within overlapping windows of the genome. Within each window, Cicero estimates a **regularized correlation matrix** using the graphical LASSO, **penalizing pairs of distant sites more than proximal sites**.
- These overlapping covariance matrices are ‘‘reconciled’’ to produce a single estimate of the co-accessibility across groups of cells.
- Co-accessibility network can be visualized.
- Often, it is useful to compare Cicero connections to other datasets with similar kinds of links. For example, you might want to compare the output of Cicero to ChIA-PET ligations. To do this, Cicero includes a function called **compare_connections**.
- User can extract **modules of co-accessibility networks** by first specifying a minimum co-accessibility score and then using the Louvain community detection algorithm on the subgraph induced by excluding edges below this score.




Aggregate randomly selected local cell neighborhoods
t-SNE dimension 2
t-SNE dimension 1
Cell
1
n
Aggregate sc-ATAC profiles
Cluster 1
Cluster 2
Cluster 3
Calculate raw correlation between sites
Graphical LASSO
Penalty
Distance (Kb)
Regularized correlation (ρ)
Link sites with ρ > C
Indirect connection
Direct connection
Identify modules of sites via community detection
Cis-co-accessibility networks (CCANs)
DDRTree Component 2
DDRTree Component 1
Phase 1
Phase 2
Branch
Co-accessibility
0.5
0
Phase 1
Phase 2
202.9 mb
203.0 mb
203.1 mb
203.2 mb
203.3 mb
203.4 mb
RABIF
TMEM183A
MYBPH
BTG2
PRELP
KLHL12
PPFIA4
CHI3L1
FMOD
OPTC
ADIPOR1
ADORA1
CYB5R1
MYOG
CHIT1

❖ Cicero (single-cell accessibility trajectories)

Molecular Cell **71**:858 (2018)

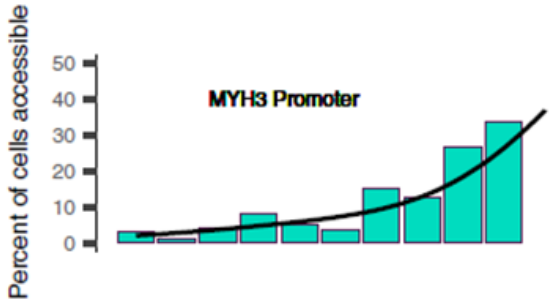
- The second major function of the Cicero is to **extend Monocle** for use with scATAC-seq data.
- **To overcome sparsity** of single-cell chromatin accessibility data is the sparsity, **nearby peaks are aggregated**. The function **aggregate_nearby_peaks** finds sites within a certain distance of each other and aggregates them together by summing their counts. Distance parameter can be adjusted by the density of the data (e.g., 1kb-10kb).
- To order the cells by progress through differentiation, we determined which aggregated peaks were relevant to the time course by fitting the following model:

$$\ln(M_i) = \beta_0 + \beta_T + \beta_S S$$

Where *Mi* is the mean of a negative binomially distributed random variable for the number of reads overlapping the aggregate region *i*, *T* encodes the times and *S* is the total number of accessible sites in each cell. We compared this full model to the following reduced model by likelihood ratio test.

$$\ln(M_i) = \beta_0 + \beta_S S$$

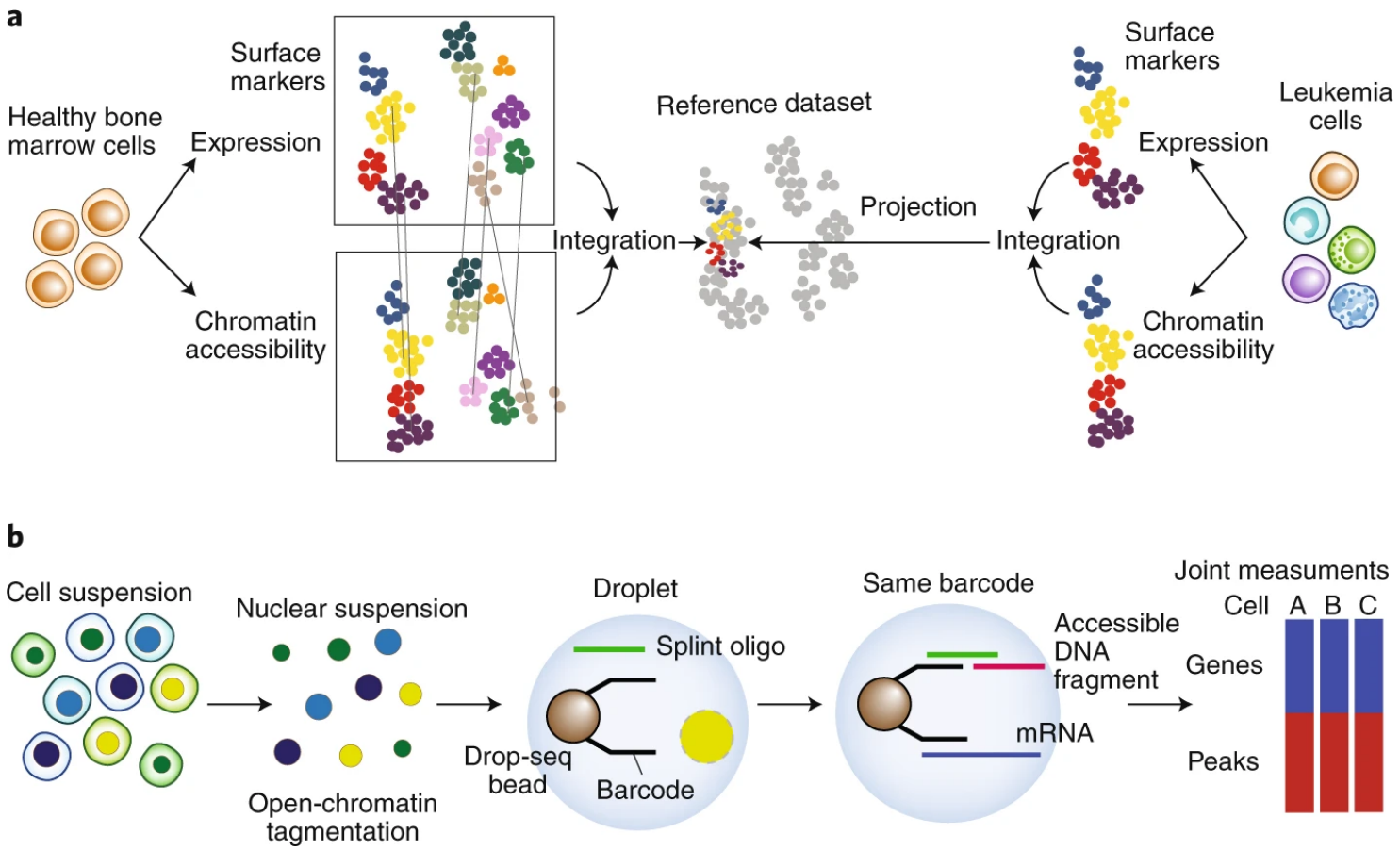
- **Sites determined by this method to be time dependent** and which were accessible in less than 10% of cells were then **used to reconstruct the pseudotime trajectory** using Monocle.
- To visualize accessibility across pseudotime, cicero first **grouped cells at similar positions in pseudotime** by *k*-means clustering along the pseudotime axis (*k* = 10). These clusters were further subdivided such into groups containing at least 50 and no more than 100 cells.
- Next, the binary accessibility profiles of the cells in each group are aggregated into a matrix *A*, so that A_{ij} contains the **number of cells in group *j* for which DNA element *i* is accessible**. The average pseudotime and average overall cell-wise accessibility for cells in each group *j* are preserved for use during differential analysis.



❖ Two approaches to connect scRNA-seq and scATAC-seq

a. **Integration-based multi-omics approach**. scRNA-seq and scATAC-seq **data from same tissue are integrated** (e.g., A dataset on healthy cells is used as a reference to decipher cancer-specific mechanisms in the leukemia dataset)

b. **Multimodal single-cell omics profiling approach**. Both scRNA-seq and scATAC-seq **data are generated from the same cell**s (nuclei). Lower throughput and no commercial platforms available.



Nature Biotechnology 37:1421 (2019)

❖ Motivation for integration of scRNA-seq and scATAC-seq data

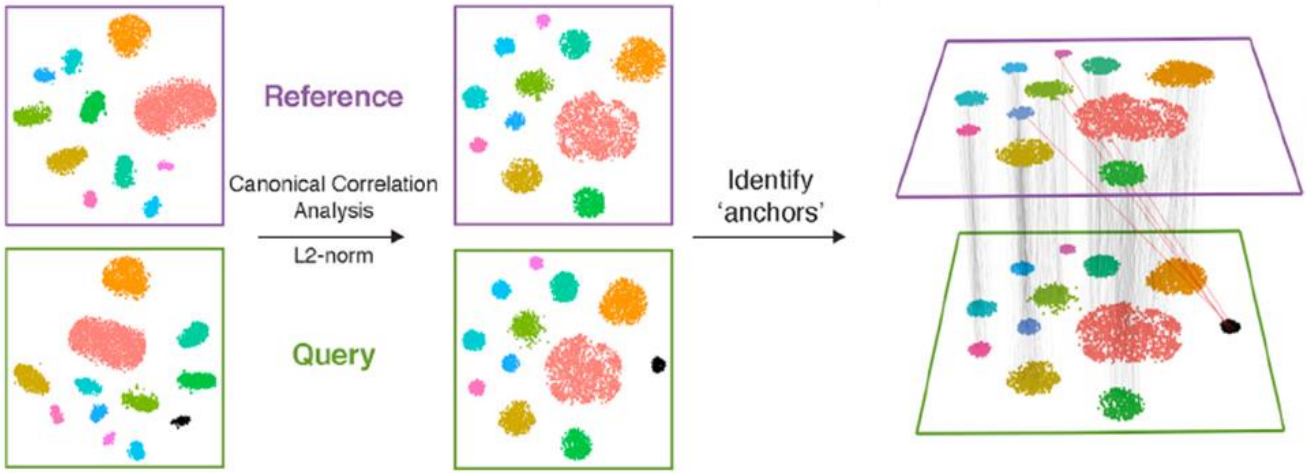
- **Cell identity annotation**: Integrated gene expression and chromatin accessibility data may improve cell identity annotation.
- **Reconstruction of regulatory networks**: Joint analysis of transcriptomics and chromatin accessibility using the integration methods can reveal the existence of novel cell states and enable to connect TF activity and enhancer elements that underlie those states.
- **Dynamic analysis of regulatory network**: Manifold alignment for inferring a shared pseudotime latent variable can reveal connections among transcriptomics and epigenetic changes and the underlying regulatory mechanisms driving dynamic processes such as differentiation, development and tumorigenesis.

Nature Biotechnology 36:411 (2018)

❖ Data integration

- The differing environments experienced by the cells can have an effect on the measurement of the transcriptome. The resulting effects exist on multiple levels: **between groups of cells** in an experiment, **between experiments** performed in the same laboratory or **between datasets** from different laboratories. **Correction for effects between experiments or between datasets** is considered as ***data integration***.
- Data integration can be applied for **multimodal** single-cell data analysis (e.g., scRNA-seq data and scATAC-seq data from the same sample). Diverse single-cell technologies each measure distinct elements of cellular identity and are characterized by unique sources of bias, sensitivity, and accuracy. As a result, measurements across datasets may not be directly comparable.
- Non-linear approaches for data integration methods such as ***Canonical Correlation Analysis*** (***CCA***; Butler *et al*, 2018), ***Mutual Nearest Neighbors*** (***MNN***; Haghverdi *et al*, 2018), and ***Harmony*** (Korsunsky *et al*, 2019) have been developed to overcome this issue.

- **Seurat v3:** *Canonical Correlation Analysis* (CCA) + *Mutual Nearest Neighbors* (MNN)
  - *CCA → L2 normalization* of canonical correlation vectors → *Project the datasets into* a subspace defined by *shared correlation structure across datasets*.
  - In the shared space, **identify pairs of MNNs across reference and query cells**. These should represent cells in a shared biological state across datasets (gray lines) and serve as ***anchors*** to guide dataset integration.
  - While MNNs have previously been identified using L2-normalized gene expression, significant differences across batches can obscure the accurate identification of MNNs, particularly when the batch effect is on a similar scale to the biological differences between cell states. To overcome this, we first jointly reduce the dimensionality of both datasets using diagonalized CCA, then apply L2-normalization to the canonical correlation vectors.
  - We next search for MNNs in this shared low-dimensional representation. We refer to the resulting cell pairs as anchors, as they encode the cellular relationships across datasets that will form the basis for all subsequent integration analyses.
  - Anchors can successfully recover matching cell states even in the presence of significant dataset differences, as CCA can effectively identify shared biological markers and conserved gene correlation patterns. However, cells in non-overlapping populations should not participate in anchors.



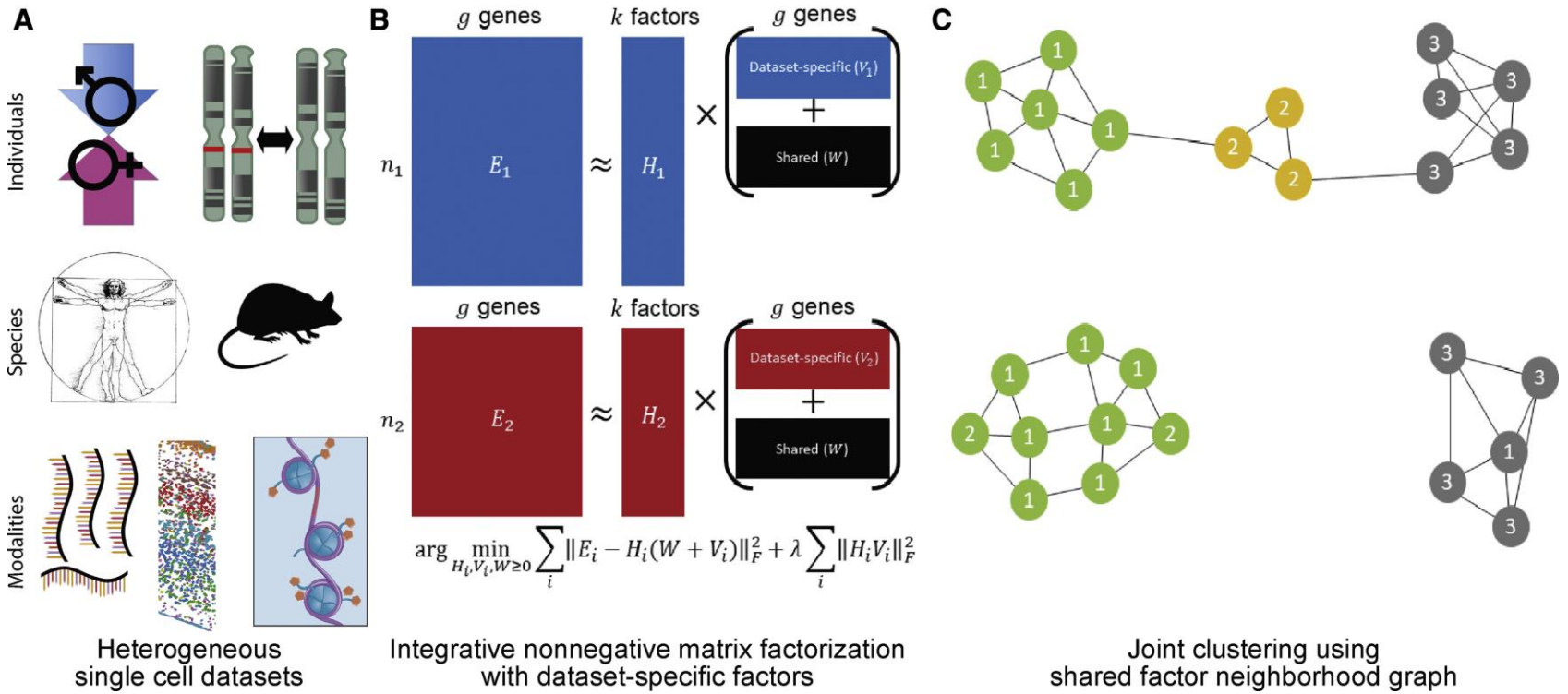
Cell **177:** 1888 (2019)

- **LIGER (**linked inference of genomic experimental relationships) *Cell* **177:** 1873 (2019)

A. LIGER identifies shared cell types across individuals, species, and multiple modalities as well as dataset-specific features, offering a unified analysis of heterogeneous single-cell datasets.

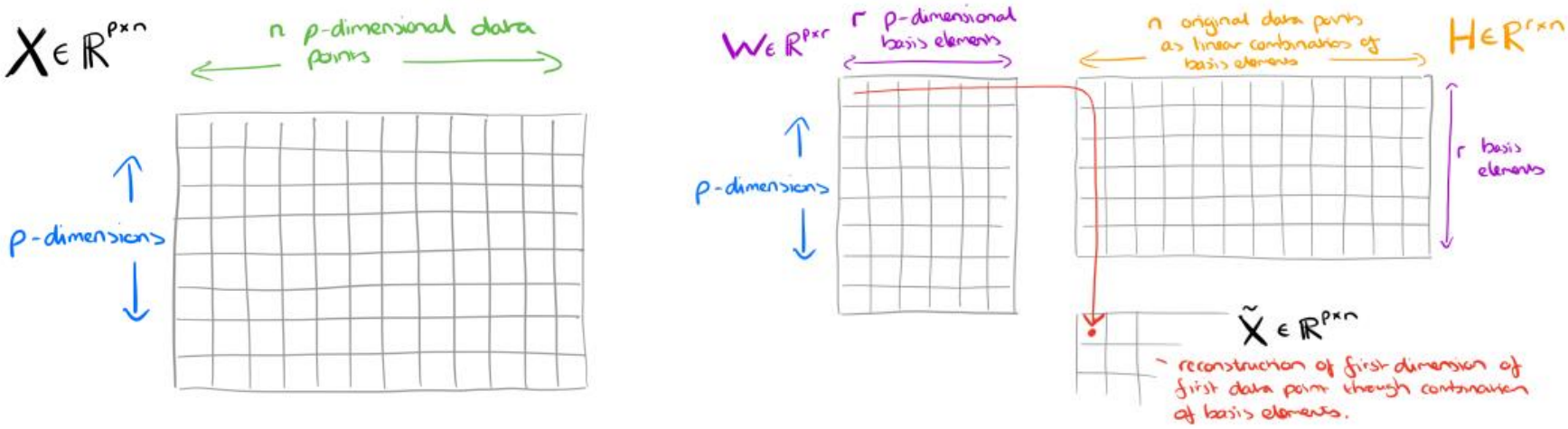
B. LIGER employs **iNMF** to learn a low-dimensional space in which each cell is defined by one set of **dataset-specific factors**, or **metagenes**. **Each factor** often **corresponds to a biologically interpretable signal**—like the genes that define a particular cell type. A tuning parameter λ allows adjusting the size of dataset-specific effects to reflect the divergence of the datasets being analyzed.

C. After performing iNMF, we **assign each cell a label based on the maximum factor loading** and then build a shared factor neighborhood graph, in which we connect cells that have similar factor loading patterns, to prevent the spurious integration of divergent cell types across datasets (such as the yellow cells shown).



- **NMF (nonnegative matrix factorization)**

- NMF approximates a matrix *X* with a low-rank matrix approximation such that ***X*** ≈ ***WH***.
- The **columns of *W*** are the **basis vectors**, '**building blocks**' from which we can reconstruct approximations to all of the original data points.
- The **rows of *H*** are **coefficient vectors**, which describe how strongly each 'building block' is present in the data.
- **Each element in *W* and *H* must be ≥ 0**. Thus, a key feature of NMF is the ability to identify nonsubtractive patterns that together explain the data as a linear combination of its basis vectors.
- NMF can automatically extract sparse and easily interpretable factors.
- In other words, NMF decomposed **original data (with original large dimension)** into **weighted sum of building blocks (with reduced dimension)**.



https://blog.acolyer.org/2019/02/18/the-why-and-how-of-nonnegative-matrix-factorization/

- **NMF example for text mining**
  - In text mining, the bag-of-words matrix: row corresponds to a word, and column to a document. The **columns of *W*** can be interpreted as **basis documents** (**bags of words**), which represent *topics*! Sets of words found simultaneously in different documents. *H* tells us how to sum contributions from different topics to reconstruct the word mix of a given original document.
  - Therefore, given a set of documents, NMF identifies topics and simultaneously classifies the documents among these different topics (i.e., decompose each document into a weighted sum of topics). We cannot interpret what it means to have a "negative" weight of the food topic.

$$\underbrace{X(:,j)}_{j\text{th document}} \approx \sum_{k=1}^{r} \underbrace{W(:,k)}_{k\text{th topic}} \underbrace{H(k,j)}_{\substack{\text{importance of } k\text{th topic} \\ \text{in } j\text{th document}}}, \qquad \text{with } W \geq 0 \text{ and } H \geq 0.$$

- **NMF example for image processing**
  - For image of a face containing *p* pixels, and squash the data into a single vector such that the *ith* entry represents the value of the *ith* pixel. The columns of *W* can be interpreted as images (the **basis images**), and *H* tells us how to sum up the basis images in order to reconstruct an approximation to a given face.
  - **In the case of facial images**, the basis images are features such as **eyes, noses, moustaches, and lips**, while the columns of *H* indicate which feature is present in which image.
  - It's difficult to interpret what it means for a face to have a "negative" component.

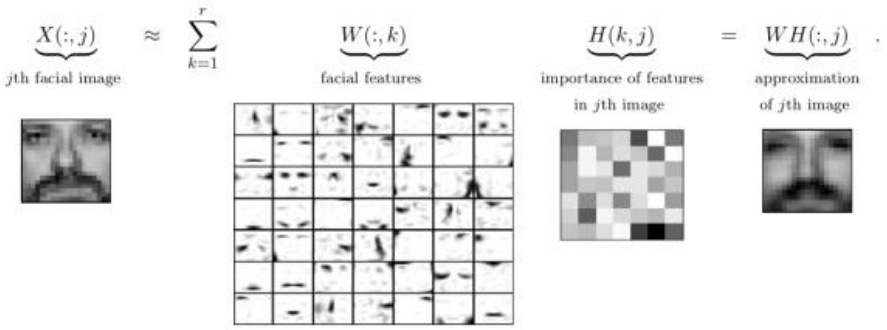
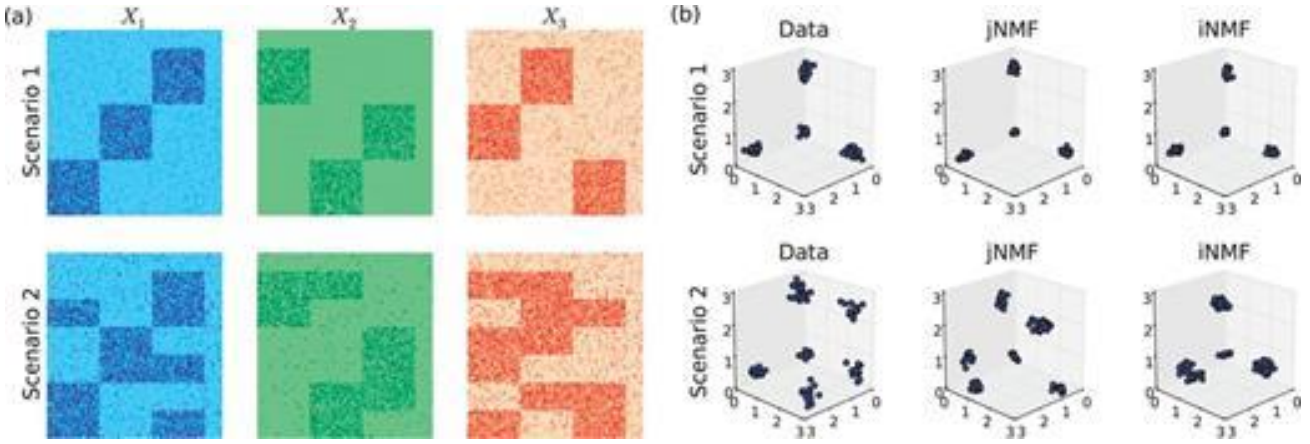


Figure 1: Decomposition of the CBCL face database, MIT Center For Biological and Computation Learning (2429 gray-level 19-by-19 pixels images) using $r = 49$ as in [79].

jNMF: *Nucleic Acids Research 40:9379 (2012)*
iNMF: *Bioinformatics 32: 1 (2016)*

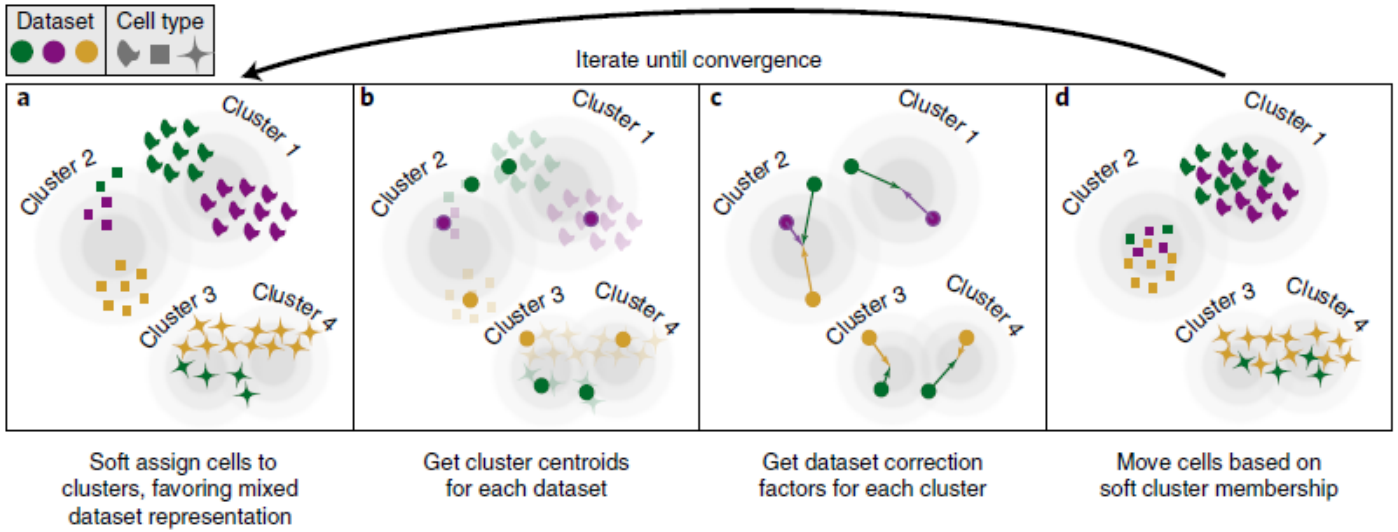
Joint NMF and integrative NMF

- The jNMF can be described as **multiple NMFs subject to a shared factor matrix**. It **can detect coordinated activity across multiple genomic variables** in the form of multi-dimensional modules. This serves as a useful preliminary step to reduce the dimensionality of the problem.
- In multiple datasets, the signal of interest is typically common among all sources (homogeneous), while extraneous effects tend to differ across sources (heterogeneous).
- While **jNMF** is very effective for detecting homogeneous effects, its factorization structure leaves no room for heterogeneous approximations. As a result, jNMF is sensitive to random noise and confounding effects, because they typically differ in structure across sources.
- While jNMF considers homogeneous effects, **iNMF** additionally considers heterogeneous effects.
- The iNMF separate the homogeneous and heterogeneous effects among the sources to extract the coordinated signal from extraneous noise via a partitioned factorization structure that captures homogeneous and heterogeneous effects.



(**a**) Multi-dimensional modules across three different data sources. Scenario 2 contains the same data with added random noise and confounding effects. (**b**) The modules are clearly detected by both methods in Scenario 1 but only by iNMF in Scenario 2.

- **Harmony is a fast and memory efficient tool.**



30-200 times faster,
30-50 times less memory than CCA/MNN

- PCA embeds cells into a space with reduced dimensionality. Harmony accepts the cell coordinates in this reduced space and runs an iterative algorithm to adjust for dataset specific effects.
- **a**, Harmony uses soft clustering to assign cells to potentially multiple clusters, to account for smooth transitions between cell states. Clusters serve as surrogate variables, rather than discrete cell types.
- **b**, Harmony calculates a global centroid and dataset-specific centroids for each cluster.
- **c**, Within each cluster, Harmony calculates a correction factor for each dataset based on the centroids. **Cluster-specific correction factors** correspond to individual cell-type and cell-state specific correction factors. In this way, Harmony learns a simple **linear adjustment function**.
- **d**, Finally, Harmony corrects each cell with a **cell-specific factor**: a linear combination of dataset correction factors weighted by the cell's soft cluster assignments made in step **a**. Since each cell may be in multiple clusters, each cell has a potentially unique correction factor.
- Harmony repeats steps **a** to **d** until convergence. The dependence between cluster assignment and dataset diminishes with each round.

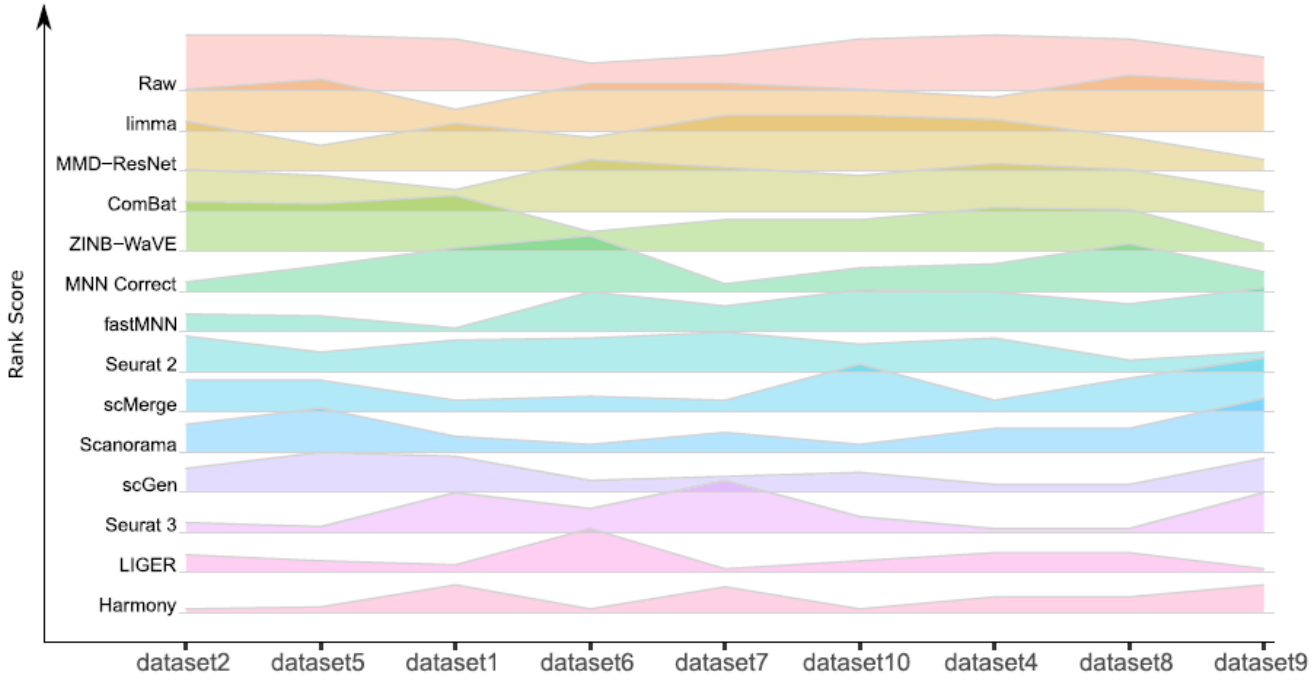
Nature Methods **16:** 1289 (2019)

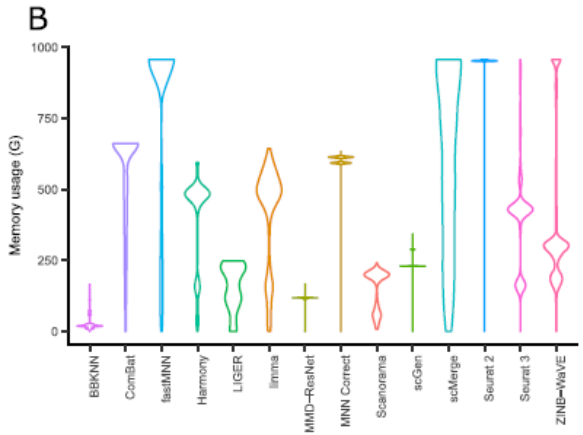
- **Comparison of 14 batch-effect correction methods for scRNA-seq data**

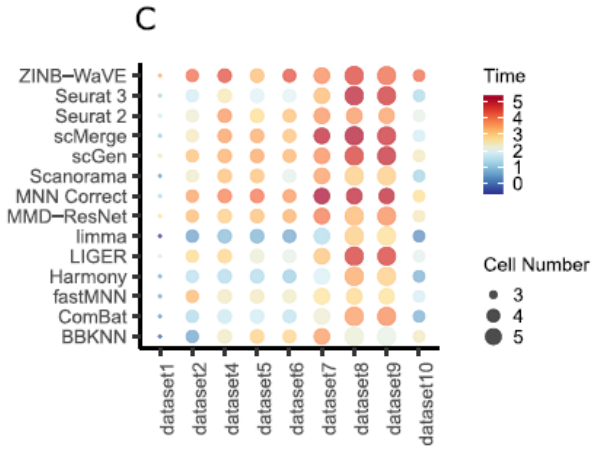
Tools	Programming language	Batch-effect-corrected output	Methods
Seurat 2 (CCA, MultiCCA)	R	Normalized canonical components (CCs)	Canonical correlation analysis and dynamic time warping
Seurat 3 (Integration)	R	Normalized gene expression matrix	Canonical correlation analysis and mutual nearest neighbors-anchors
Harmony	R	Normalized feature reduction vectors (Harmony)	Iterative clustering in dimensionally reduced space
MNN Correct	R	Normalized gene expression matrix	Mutual nearest neighbor in gene expression space
fastMNN	R	Normalized principal components	Mutual nearest neighbor in dimensionally reduced space
ComBat	R	Normalized gene expression matrix	Adjusts for known batches using an empirical Bayesian framework
limma	R	Normalized gene expression matrix	Linear model/empirical Bayes model
scGen	Python	Normalized gene expression matrix	Variational auto-encoders neural network model and latent space
Scanorama	Python/R	Normalized gene expression matrix	Mutual nearest neighbor and panoramic stitching
MND-ResNet	Python	Normalized principal components	Residual neural network for calibration
ZINB-WaVE	R	Normalized feature reduction vectors (ZINB-WaVE)/normalized gene expression matrix	Zero-inflated negative binomial model, extension of RUV model
scMerge	R	Normalized gene expression matrix	Stably expressed genes (scSEGs) and RUVIII model
LIGER	R	Normalized feature reduction vectors (LIGER)	Integrative non-negative matrix factorization (iNMF) and joint clustering + quantile alignment
BBKNN	Python/R	Connectivity graph and normalized dimension reduction vectors (UMAP)	Batch balanced *k*-nearest neighbors

Genome Biology 21**:** 12 (2020)

- **Benchmark of batch-effect correction methods for scRNA-seq data**
- Benchmarking on 9 scRNA-seq datasets with different sources and technologies.
- Based on rank score, memory usage, and runtime, **Harmony, LIGER, Seurat 3** were recommended.







Genome Biology 21**:** 12 (2020)