

# Medical Genomics: Epigenomics & multi-omic integration

*N. Alcala*

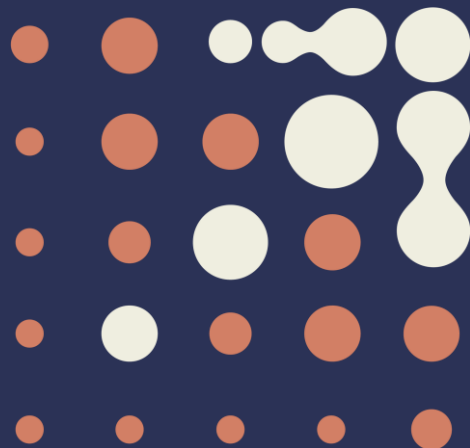
Computational Cancer Genomics Team

November 27th 2024

International Agency  
for Research on Cancer



RARE  
CANCERS  
GENOMICS



# Plan

## Part I. Epigenomics

- Concepts
- Techniques
- Processing
- Analysis

## Part II. Multi-omics

- **Concepts:** complementarity of 'omic layers
- **Analysis:** tools for integration

# Motivating example I

## What is the impact of COVID vaccination on patients immune systems?

- Longitudinal single-cell multiomics analysis

nature immunology



Article

<https://doi.org/10.1038/s41590-023-01808-9>

### Multimodal single-cell datasets characterize antigen-specific CD8<sup>+</sup> T cells across SARS-CoV-2 vaccination and infection

Received: 24 January 2023

Accepted: 31 July 2023

Published online: 21 September 2023

Check for updates

Bingjie Zhang<sup>1,2,3,4</sup>, Rabi Upadhyay<sup>1,4,5,6</sup>, Yuhao Hao<sup>1,2</sup>, Marie I. Samanovic<sup>5,6</sup>, Ramin S. Herati<sup>5,6</sup>, John D. Blair<sup>1,2</sup>, Jordan Axelrad<sup>5</sup>, Mark J. Mulligan<sup>5,6</sup>, Dan R. Littman<sup>3,4,7</sup> & Rahul Satija<sup>1,2</sup>✉

The immune response to SARS-CoV-2 antigen after infection or vaccination is defined by the durable production of antibodies and T cells. Population-based monitoring typically focuses on antibody titer, but there is a need for improved characterization and quantification of T cell responses. Here, we used multimodal sequencing technologies to perform a longitudinal analysis of circulating human leukocytes collected before and after immunization with the mRNA vaccine BNT162b2. Our data indicated distinct subpopulations of CD8<sup>+</sup> T cells, which reliably appeared 28 days after prime vaccination. Using a suite of cross-modality integration tools, we defined their transcriptome, accessible chromatin landscape and immunophenotype, and we identified unique biomarkers within each modality. We further showed that this vaccine-induced population was SARS-CoV-2 antigen-specific and capable of rapid clonal expansion. Moreover, we identified these CD8<sup>+</sup> T cell populations in scRNA-seq datasets from COVID-19 patients and found that their relative frequency and differentiation outcomes were predictive of subsequent clinical outcomes.

# Motivating example II

## Why Down syndrome predisposes to leukemia?

- single-cell multiomics of >1.1 million cells from 3 fetuses with disomy and 15 fetuses with trisomy

### Article

#### Single-cell multi-omics map of human fetal blood in Down syndrome


<https://doi.org/10.1038/s41586-024-07046-4>

Received: 24 February 2023

Accepted: 14 August 2024

Published online: 25 September 2024

Open access

 Check for updates

Andrew B. Marderstein<sup>1</sup>, Marco De Ziani<sup>2,3,4</sup>, Rebecca Moeller<sup>2</sup>, Jon Beznay<sup>5</sup>, Evlin M. Padhi<sup>6</sup>, Shuo Wang<sup>2,3,4</sup>, Yin H. H. Coorens<sup>7</sup>, Yilin Xie<sup>8</sup>, Haoliang Xue<sup>2,3,4</sup>, Stephen B. Montgomery<sup>2,3,4</sup> & Ana Cvejic<sup>2,3,4,9</sup>

Down syndrome predisposes individuals to haematological abnormalities, such as increased number of erythrocytes and leukaemia in a process that is initiated before birth and is not entirely understood<sup>1–3</sup>. Here, to understand dysregulated haematopoiesis in Down syndrome, we integrated single-cell transcriptomics of over 1.1 million cells with chromatin accessibility and spatial transcriptomics datasets using human fetal liver and bone marrow samples from 3 fetuses with disomy and 15 fetuses with trisomy. We found that differences in gene expression in Down syndrome were dependent on both cell type and environment. Furthermore, we found multiple lines of evidence that haematopoietic stem cells (HSCs) in Down syndrome are 'primed' to differentiate. We subsequently established a Down syndrome-specific map linking non-coding elements to genes in disomic and trisomic HSCs using 10X multiome data. By integrating this map with genetic variants associated with blood cell counts, we discovered that trisomy restructured regulatory interactions to dysregulate enhancer activity and gene expression critical to erythroid lineage differentiation. Furthermore, as mutations in Down syndrome display a signature of oxidative stress<sup>4,5</sup>, we validated both increased mitochondrial mass and oxidative stress in Down syndrome, and observed that these mutations preferentially fell into regulatory regions of expressed genes in HSCs. Together, our single-cell, multi-omic resource provides a high-resolution molecular map of fetal haematopoiesis in Down syndrome and indicates significant regulatory restructuring giving rise to co-occurring haematological conditions.

yahoo/actualités

SCIENCES AVENIR  
La Recherche

### Trisomie 21 et leucémie : on sait enfin pourquoi les deux sont liés

Sciences et Avenir

27 septembre 2024

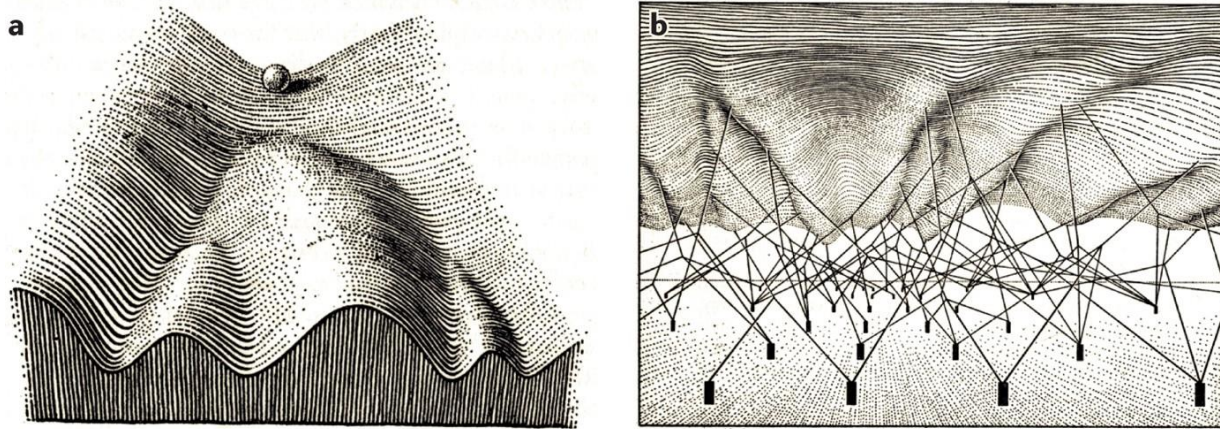


International Agency for Research on Cancer

# Part I. Epigenomics | *Concepts*

Although all cells in your body carry the same DNA sequence, huge diversity of cell types due to **epigenetic processes regulating gene expression**

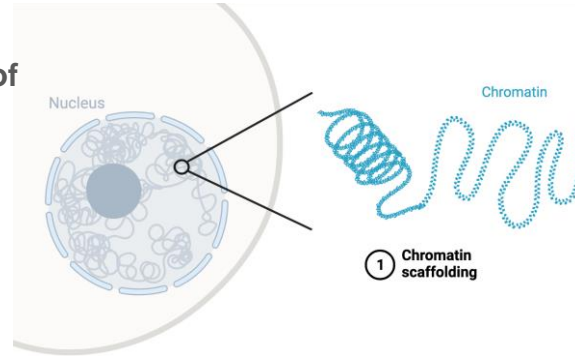
Thus, epigenetic processes play a key role in cell state determination, including diseased states



# Part I. Epigenomics | *Concepts*

Epigenetic processes

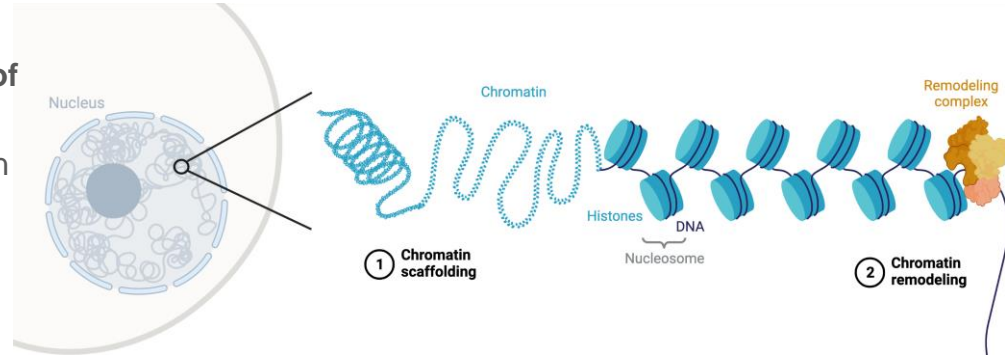
1. **Chromatin scaffolding:** formation of **backbone of chromosomes** from structural proteins



# Part I. Epigenomics | Concepts

Epigenetic processes

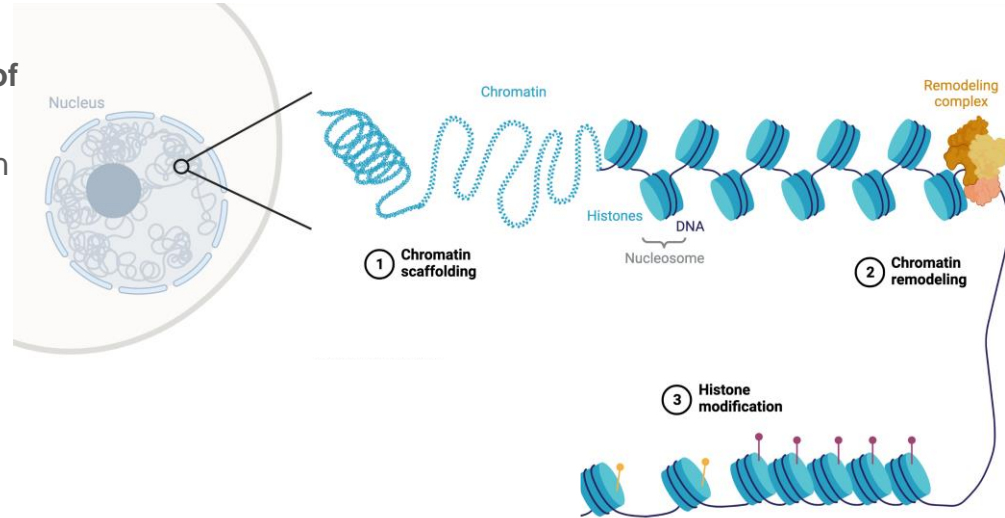
1. **Chromatin scaffolding:** formation of **backbone of chromosomes** from structural proteins
2. **Chromatin remodeling:** modification of chromatin architecture. heterochromatin (**compact**) regions are “**silent**” (no transcription, inactive genes), euchromatin regions are “**active**”



# Part I. Epigenomics | Concepts

Epigenetic processes

1. **Chromatin scaffolding:** formation of **backbone of chromosomes** from structural proteins
2. **Chromatin remodeling:** modification of chromatin architecture. heterochromatin (**compact**) regions are “**silent**” (no transcription, inactive genes), **euchromatin regions are “active”**
3. Histone modification: molecular modifications affecting histone compaction (e.g., acetylation or phosphorylation)

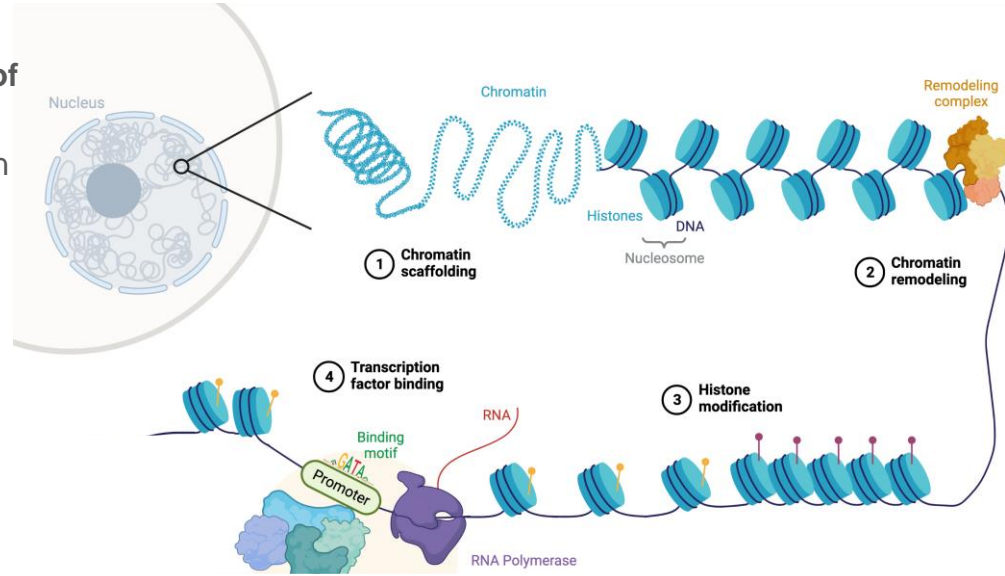




# Part I. Epigenomics | Concepts

Epigenetic processes

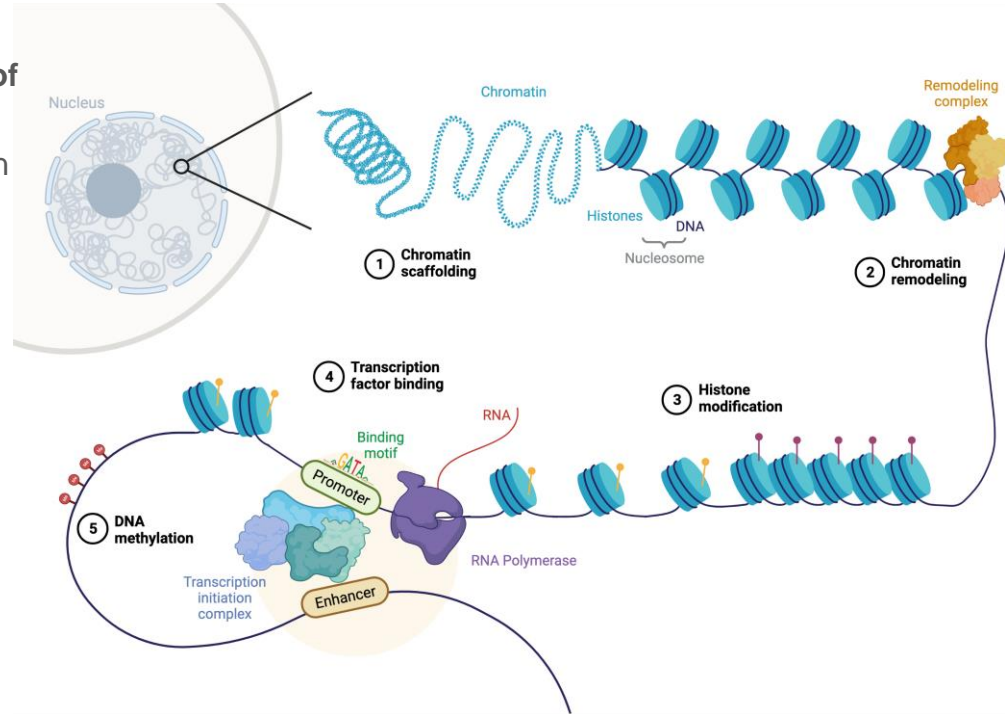
1. **Chromatin scaffolding**: formation of **backbone of chromosomes** from structural proteins
2. **Chromatin remodeling**: modification of chromatin architecture. heterochromatin (**compact**) regions are “**silent**” (no transcription, inactive genes), **euchromatin regions are “active”**
3. Histone modification: molecular modifications affecting histone compaction (e.g., acetylation or phosphorylation)
4. **Transcription factor (TF) binding** influence nucleosome position & activate gene expression



# Part I. Epigenomics | Concepts

Epigenetic processes

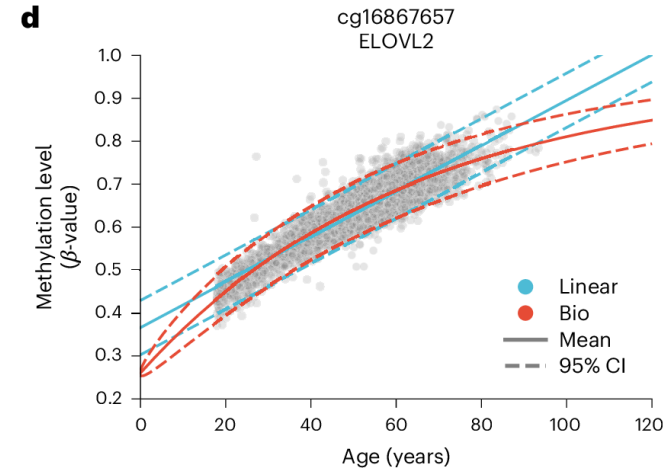
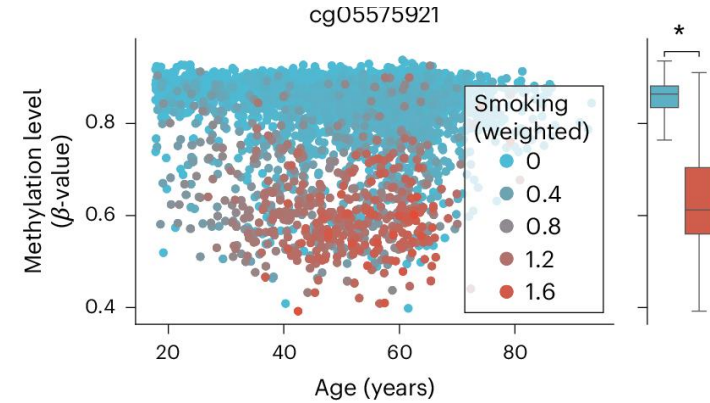
1. **Chromatin scaffolding:** formation of **backbone of chromosomes** from structural proteins
2. **Chromatin remodeling:** modification of chromatin architecture. heterochromatin (**compact**) regions are “**silent**” (no transcription, inactive genes), **euchromatin regions are “active”**
3. Histone modification: molecular modifications affecting histone compaction (e.g., acetylation or phosphorylation)
4. **Transcription factor (TF) binding** influence nucleosome position & activate gene expression
5. **Methylation:** a methyl group tags DNA at CpG sites, influencing binding affinity of TFs



# Part I. Epigenomics | Concepts

Epigenetic mechanisms are affected by

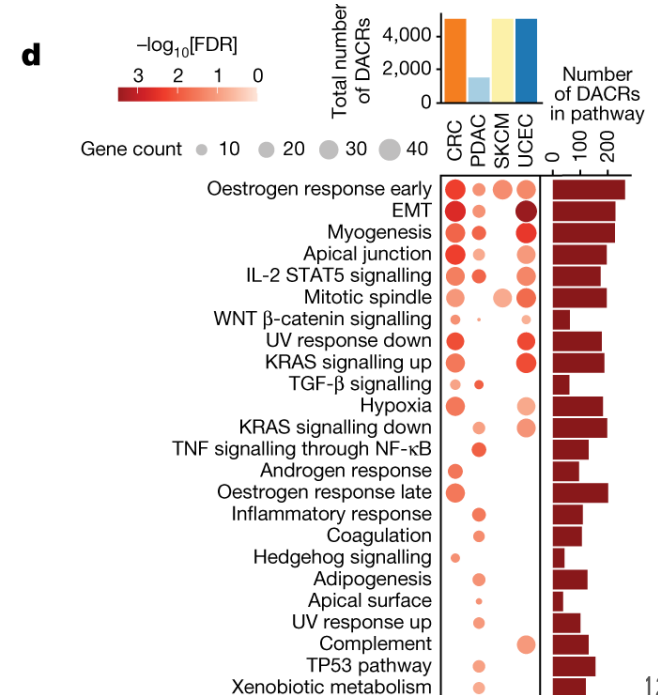
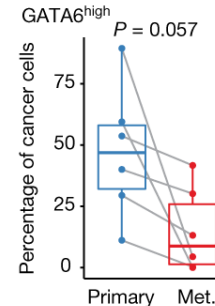
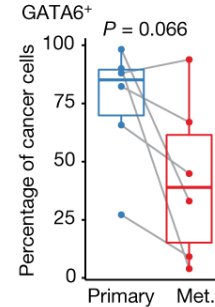
- Development / aging
- Lifestyle (diet)
- Drugs
- Environmental chemicals



# Motivating example III

## What are the epigenetic drivers of carcinogenesis?

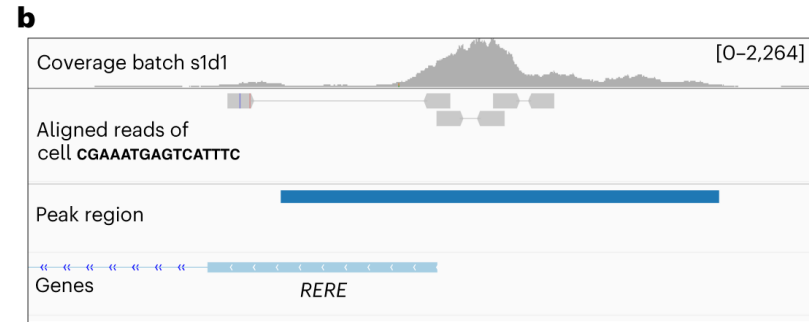
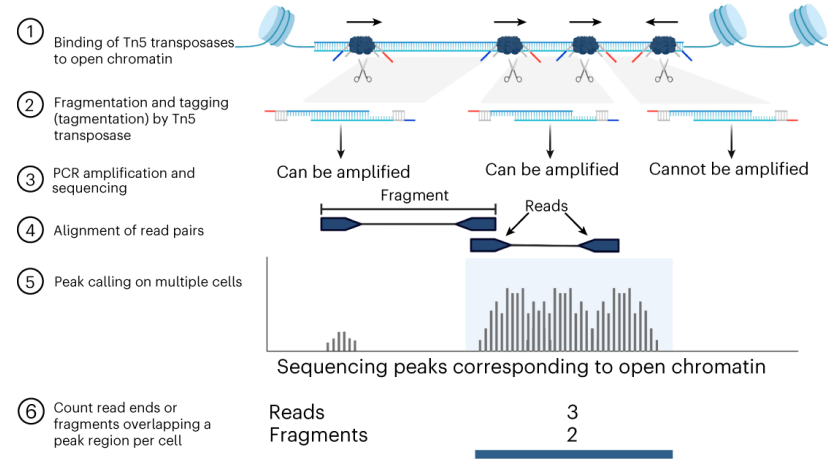
- Single-nucleus multiomics analysis of 225 tumors
- Identification of master regulators of cancer, e.g. loss of *GATA6* activity during pancreatic tumor metastasis
- snATAC trajectory analysis in colorectal and uterine cancers identified a linear path from normal cells to primary tumors to metastasis that correlated with epithelial-to-mesenchymal transition



# Part I. Epigenomics | *Techniques*

## Assay for transposase-accessible chromatin sequencing (ATAC-seq)

- Tn5 transposase cleaves and tags double-stranded DNA in open regions with sequencing adaptors
- Tagged DNA fragments are purified, PCR-amplified, and sequenced
- Number of reads => how open chromatin is
- **Informative about cell identity:** 25% of accessible chromatin regions are different between cells
- **Informative about regulatory mechanisms:** e.g., accessible TF binding sites



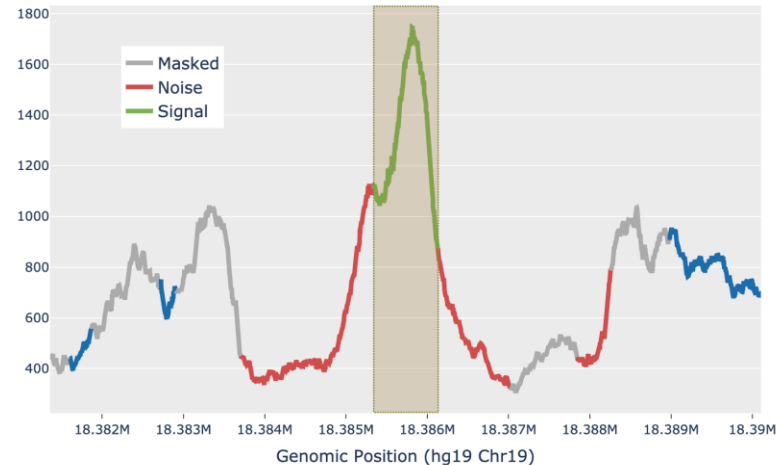
Source: Carter & Zhao, Nature Reviews Genetics 2021.

# Part I. Epigenomics | *Techniques*

## Single-cell sequencing: processing

Processing of fastq files starts similarly to scRNA-seq but has **extra steps to define features**

- barcode-aware alignment (e.g., CellRanger, STARsolo)
    - error-correction and demultiplexing of cell barcodes
    - standard mapping on reference genome
    - deduplication of UMIs
  - Quantification of UMIs
    - **Identification of transposase cut sites**
    - **Detection of accessible chromatin peaks**
    - **cell calling**
- => cell x peak matrix of read counts



Source: 10X genomics cellranger-atac

# Part I. Epigenomics | *Techniques*

## The peak matrix interpretation of features

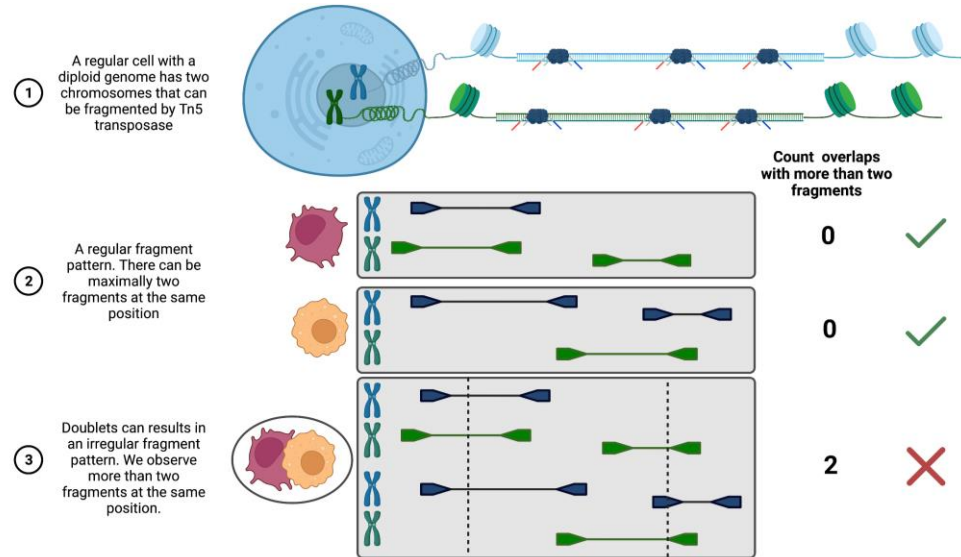
- Peaks in coding regions indicate that a gene might be transcribed
- Peaks in non-coding regions indicate that regulatory proteins (TFs) can bind (closest gene annotated by default by cellranger)
- Peaks are cell-type specific, thus requiring clustering of cells before calling

# Part I. Epigenomics | Processing

## Single-cell sequencing: QC

Some steps of quality control are similar to scRNA-seq , some are **specific to ATAC-seq**

- Doublet detection (simulation-based approach similar to scRNA-seq for heterotypic doublets, and ATAC-seq methods like AMULET for hetero and homotypic doublets)





# Part I. Epigenomics | *Processing*

## Single-cell sequencing: QC

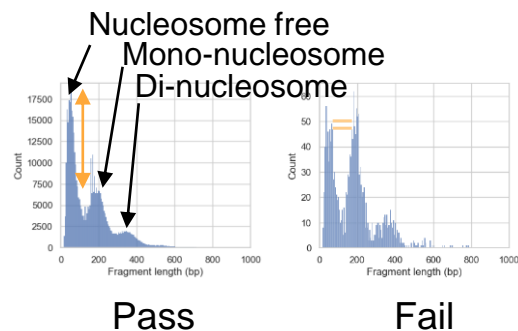
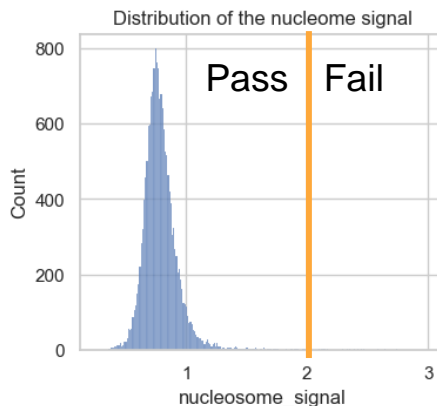
Some steps of quality control are similar to scRNA-seq , some are **specific to ATAC-seq**

- Doublet detection
- Detect low-quality cells using:

Total counts per cell : is there enough data for each cell?

Number of features per cell : is there enough data for each cell?

**Nucleosome signal** : ratio of mono-nucleosomal to nucleosome-free fragments, interpreted as signal-to-noise ratio



Source: single-cell best practices

# Part I. Epigenomics | Processing

## Single-cell sequencing: QC

Some steps of quality control are similar to scRNA-seq , some are **specific to ATAC-seq**

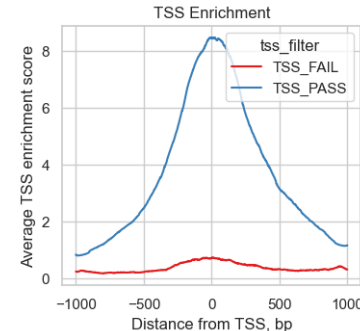
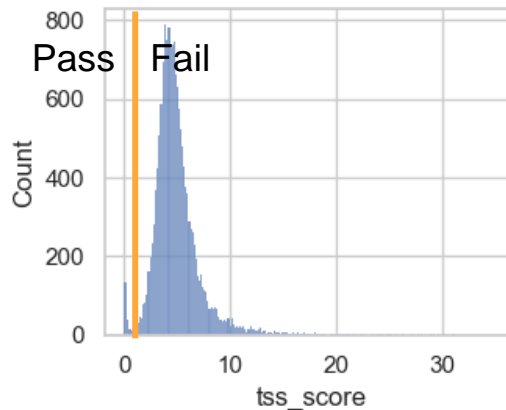
- Doublet detection
- Detect low-quality cells using:

Total counts per cell : is there enough data for each cell?

Number of features per cell : is there enough data for each cell?

**Nucleosome signal** : ratio of mono-nucleosomal to nucleosome-free fragments, interpreted as signal-to-noise ratio

**TSS enrichment** : ratio of fragments centered at transcription start site (TSS) to TSS-flanking regions, interpreted as signal-to-noise ratio



Source: single-cell best practices

# Part I. Epigenomics | *Processing*

## Single-cell sequencing: QC

Some steps of quality control are similar to scRNA-seq , some are **specific to ATAC-seq**

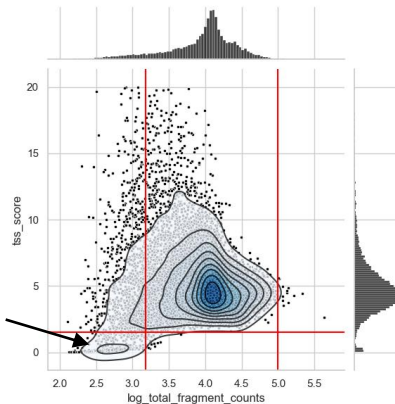
- Doublet detection
- Detect low-quality cells using:

Total counts per cell : is there enough data for each cell?

Number of features per cell : is there enough data for each cell?

**Nucleosome signal** : ratio of mono-nucleosomal to nucleosome-free fragments, interpreted as signal-to-noise ratio

**TSS enrichment** : ratio of fragments centered at transcription start site (TSS) to TSS-flanking regions, interpreted as signal-to-noise ratio



Low-quality mode

Source: single-cell best practices

# Part I. Epigenomics | *Processing*

## Single-cell sequencing: QC

Some steps of quality control are similar to scRNA-seq , some are **specific to ATAC-seq**

- Doublet detection
- Detect low-quality features:

Cells per feature : is there enough cells with this peak to perform analyses? ~15 cells

# Part I. Epigenomics | Analysis

		Python			R			
		Muon	snappyATAC 2.0	pyCisTopic	Signac	ArchR	PeakVI	PoissonVAE
Stage	Dimensionality reduction	Method	Latent Semantic Indexing (LSI)	Spectral embedding of Jaccard similarity	Latent Dirichlet Allocation (LDA)	LSI	Iterative LSI	
	Annotation	Visualization	UMAP/ TSNE	UMAP/ TSNE	UMAP/ TSNE	UMAP/ TSNE		
	Visualization	Clustering	Leiden	Leiden	Leiden	Louvain	Louvain	
Step	Feature for gene activity computation	Gene body and upstream of TSS (2000 bp)	Gene body	Gene body and up-/ downstream of TSS (exponentially decaying and avoiding gene boundaries)	Gene body and upstream of TSS (2000 bp)	Gene body and upstream of TSS (exponentially decaying and avoiding gene boundaries)		
	Differentially accessible regions	T test (possibility for Logistic regression or Wilcoxon test)	Logistic regression	Wilcoxon test	Logistic regression	Wilcoxon test		
	Gene activity imputation	✗	✓ (Using MAGIC)	✓ (Using topics)	✗	✓ (Using MAGIC)		
	Track plotting	✗	✗	✗	✓	✓		
	Interactive genome browser	✗	✗	✗	✓	✓		
	Interpretation	Motif enrichment	✗	✓ (Using pycisTarget)	✓	✓		
	chromVAR motif deviations	✗	✗	✗	✓	✓		
	Footprinting	✗	✗	✗	✓	✓		
	Co-accessibility	✗	✗	✗	✓ (Using Cicero)	✓		
	Trajectory inference	✗	✗	✗	✓ (Using Monocle 3)	✓		

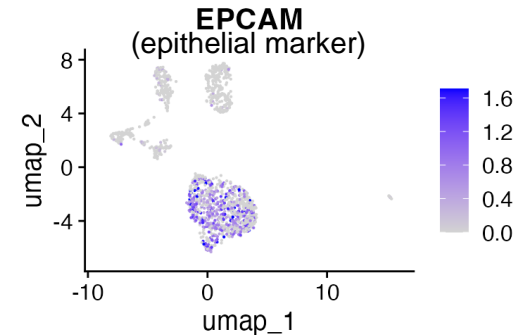
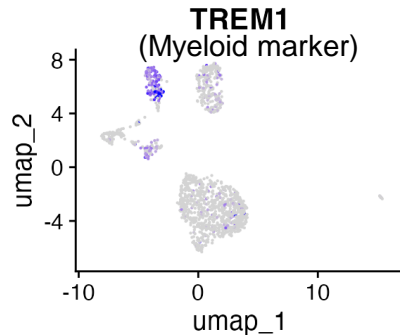
Source: single-cell best practices

# Part I. Epigenomics | *Analysis*

## Annotation

Objective: identify cells types

- Clustering of the cell x peak matrix (Louvain)
- Assess gene activity (e.g. impute gene expression) by summing accessibility in gene body and promoter region
- Use cell type markers or databases to assign labels to clusters



Source: single-cell best practices

# Part I. Epigenomics | *Analysis*

## Differentially accessible regions

Objective: identify regions responsible for regulation of cells state

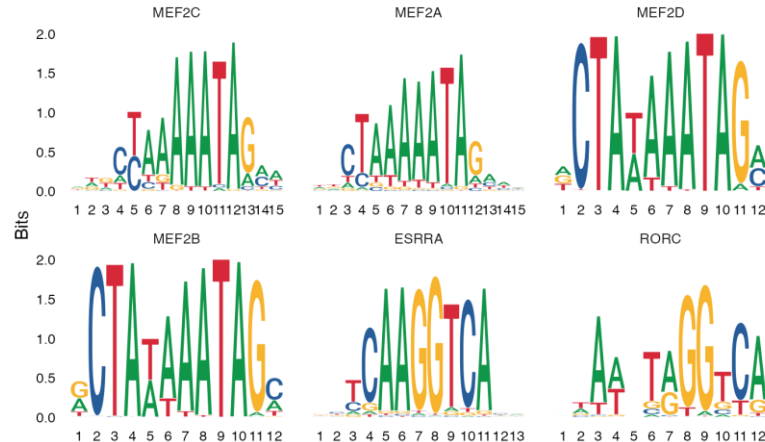
- Statistical test (Wilcoxon rank sum test) of accessibility as a function of cell state / cluster
- Computes log fold change of features between conditions (similar to scRNA-seq)

# Part I. Epigenomics | *Analysis*

## Motif analysis

Objective: identify regulatory DNA sequence motifs (short patterns recruiting TFs or microRNAs)

- Hypergeometric test of the probability of observing a motif at observed frequency compared to background peaks with matching GC content



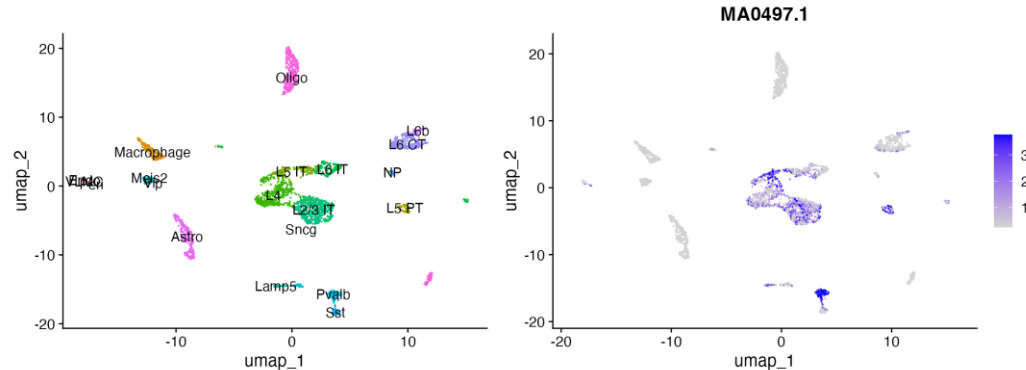


# Part I. Epigenomics | *Analysis*

## Motif analysis

Objective: identify regulatory DNA sequence motifs (short patterns recruiting TFs or microRNAs)

- chromVAR: motifs associated with variability in chromatin accessibility between cells => identify motifs and regulatory mechanisms responsible for cell state

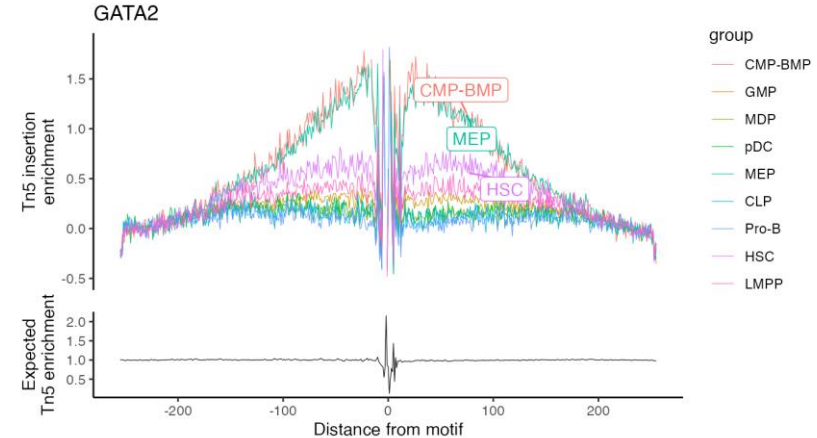
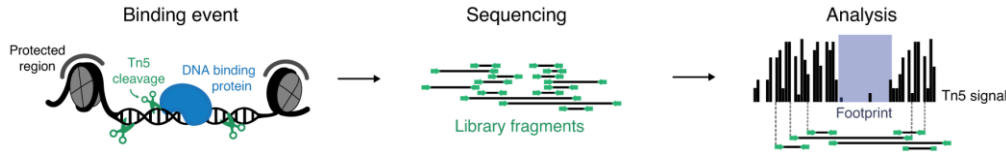


# Part I. Epigenomics | Analysis

## TF footprinting

Objective: identify TF binding at cis-regulatory elements based on DNA motifs

- Compare observed and expected Tn5 insertion frequency around a motif => dip indicates TF binding

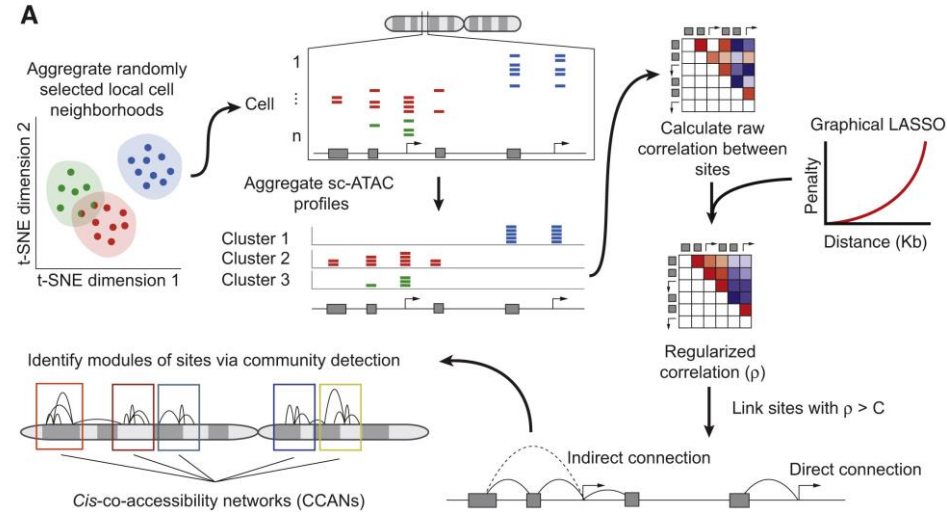


# Part I. Epigenomics | Analysis

## Co-accessibility

Objective: Link regulatory DNA elements to their target genes, which may be located >100kb away

- Cicero: identify cis-regulatory elements through co-accessibility networks
- Networks are “chromatin hubs”, physically close, interacting with common set of TFs, and undergoing coordinated changes in histone marks affecting expression

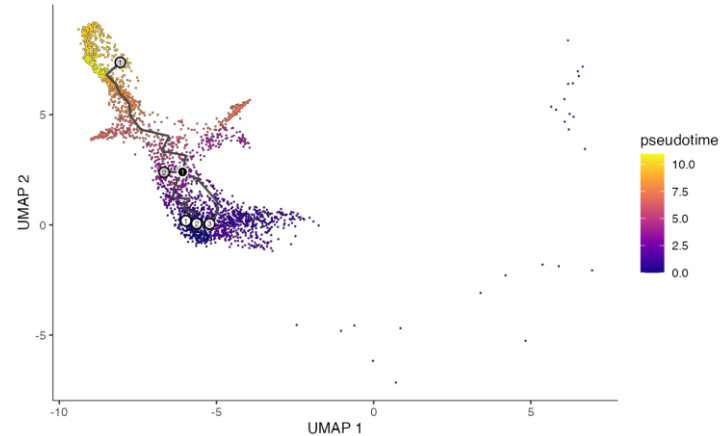
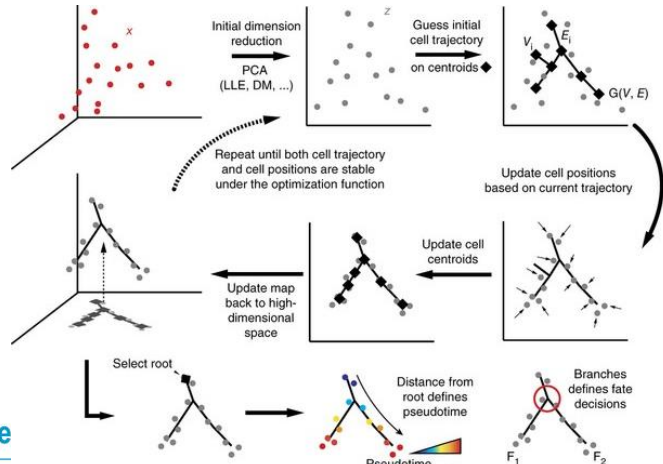


# Part I. Epigenomics | Analysis

## Trajectory inference

Objective: identify cell differentiation trajectories

- Monocle: learn sequence of gene expression changes each cell goes through to differentiate
- Learn the trajectory graph (parsimonious tree), select root and compute pseudotime as geodesic dist



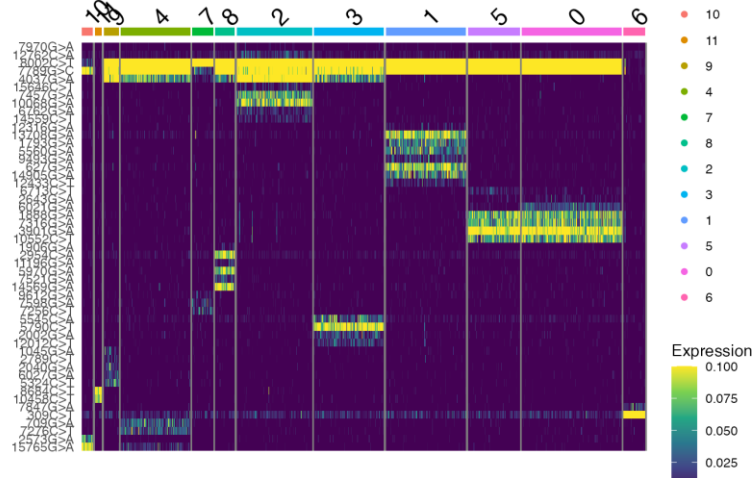
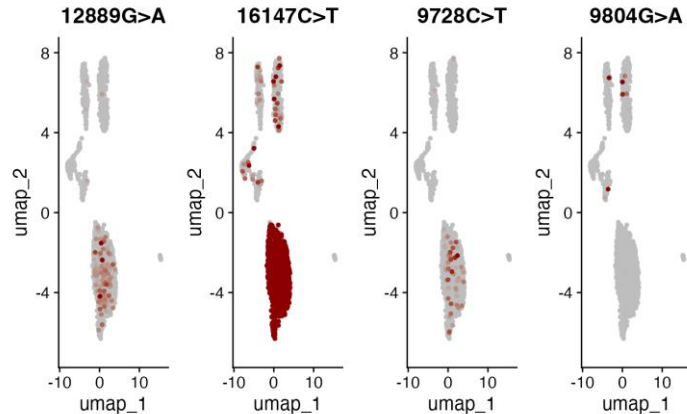
Source: Qiu et al. Nat Methods 2017 28

# Part I. Epigenomics | Analysis

## Phylogenetic inference

Objective: reconstruct cell genealogies using (mitochondrial) somatic alterations

- Somatic alterations in cell subsets are natural barcodes to infer cell differentiation trajectories
- Mitochondrial DNA is often used because variable and good coverage



# Conclusions

- ATAC-seq data has specificities (difficult feature definition, high sparsity, noise) that require specific processing methods and QCs
- Single-cell epigenetics allows to perform similar analyses than RNA-seq (cell annotation, differentiation trajectory inference)
- Allows to suggest gene regulatory mechanisms for each cell state

## Part II. Multi-omics | Concepts

FOCUS | EDITORIAL

### Simultaneous measure of multiple modalities

**Goal:** Quantify the level of expression of genes and transcripts of each individual cell of a tissue

- Track cell differentiation
- Quantify tissue heterogeneity
- Quantify diversity of microbiome

Different methods

- **Droplet based (10X genomics multiome)**
- Plate-based with unique molecular identifiers (UMIs): CEL-seq, MARS-seq
- Plate-based with reads: Smart-seq2

### Method of the Year 2019: Single-cell multimodal omics

Multimodal omics measurement offers opportunities for gaining holistic views of cells one by one.



## Part II. Multi-omics | Concepts

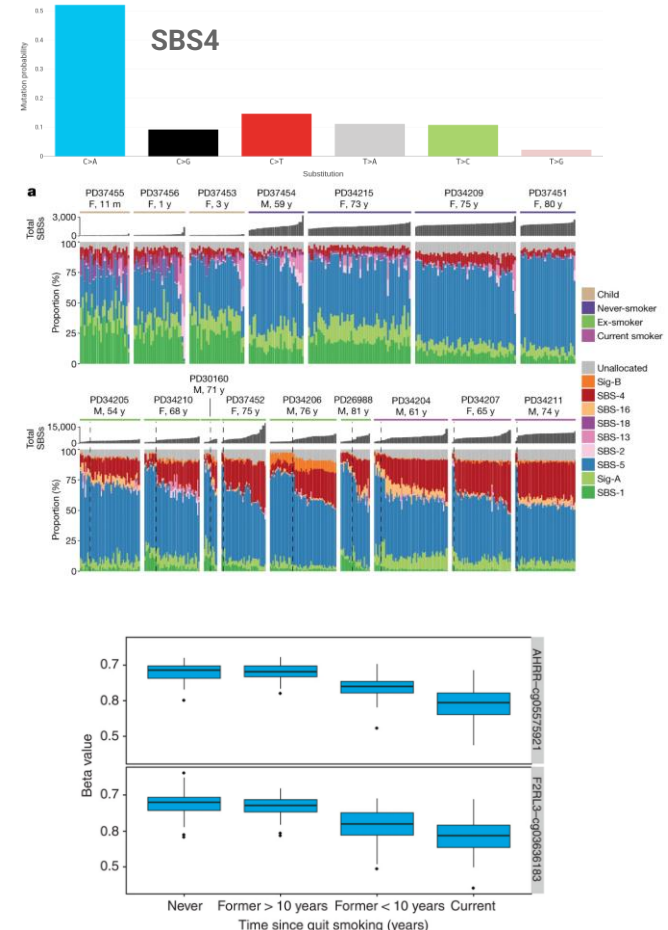
### Interactions between 'omic layers

Alterations in one 'omic layer impact other layers, for instance

- eQTLs: genomic variants -> transcriptome & proteome
- Epigenome -> transcriptome & proteome

Processes of interest impact multiple layers

- **Environmental exposures** can leave mutational signature (genome) and leave epigenetic marks that impact gene regulation (transcriptome/proteome)

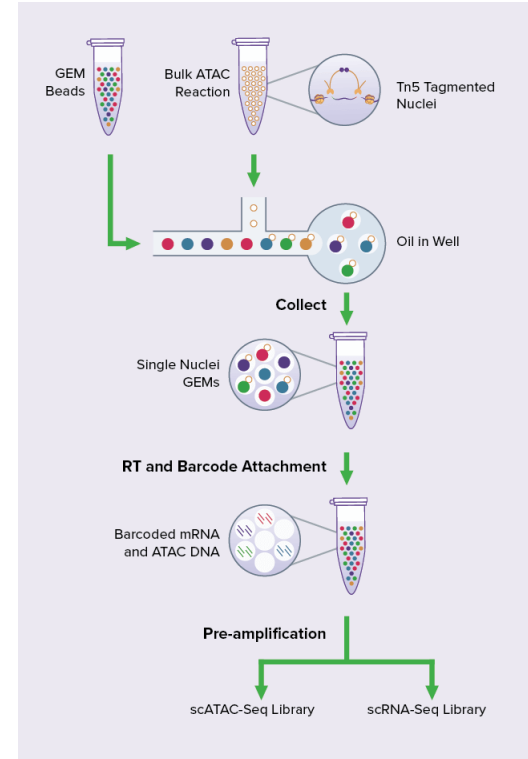
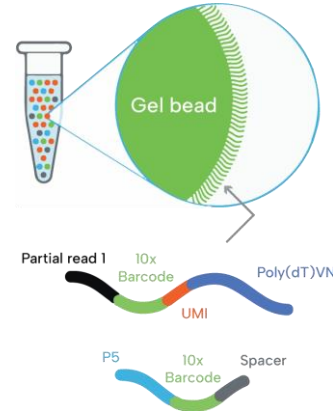




## Part II. Multi-omics | *Techniques*

### 10X genomics multiome (single nuclei RNA+ATAC-seq)

- Preparation of nuclei
- Tn5 tagmentation of open chromatin regions
- Nuclei loaded onto Chromium Controller
- Barcoding of mRNA and ATAC DNA in the same nucleus
- Library generation
- Sequencing



## Part II. Multi-omics | Resources

## Databases of multiome

- Processed data (read counts) open-access

## SCA HUMAN ATLAS

## RNA-Sequencing

### Adult Body - Lung

## Omics Types

scRNA-Seq

scATAC-Seq

## scImmune Profiling

## Spatial

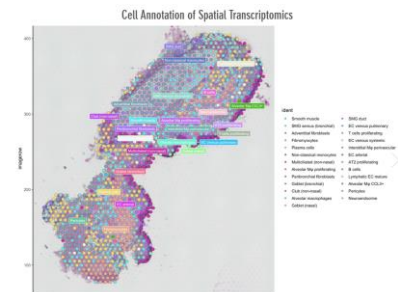
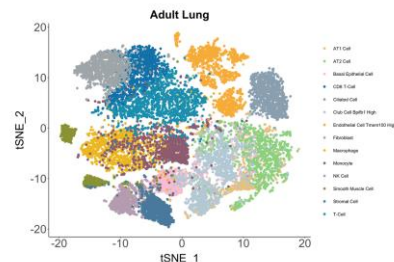
## RNA-Seq

CyTOF(Not Avail)

Flow

The following contains tissue-specific phenotypes for lung in adult body. Refer to the left panel to navigate to other omics within the same tissue type.

BACK →



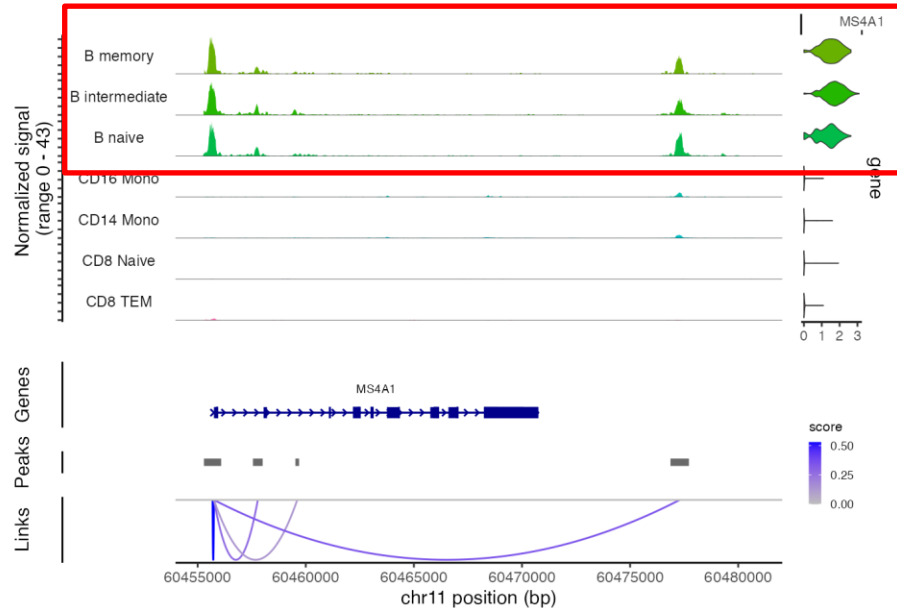
**Web interface of the single-cell atlas.** scRNA-seq and spatial transcriptomics of adult lung tissue (t-SNE). Source: <https://www.singlecellatlas.org/>

## Part II. Multi-omics | Analysis

### Link peaks to genes

Goal: identify genes and regulatory regions explaining cell states

- Correlation between expression and peaks near TSS, taking into account GC content



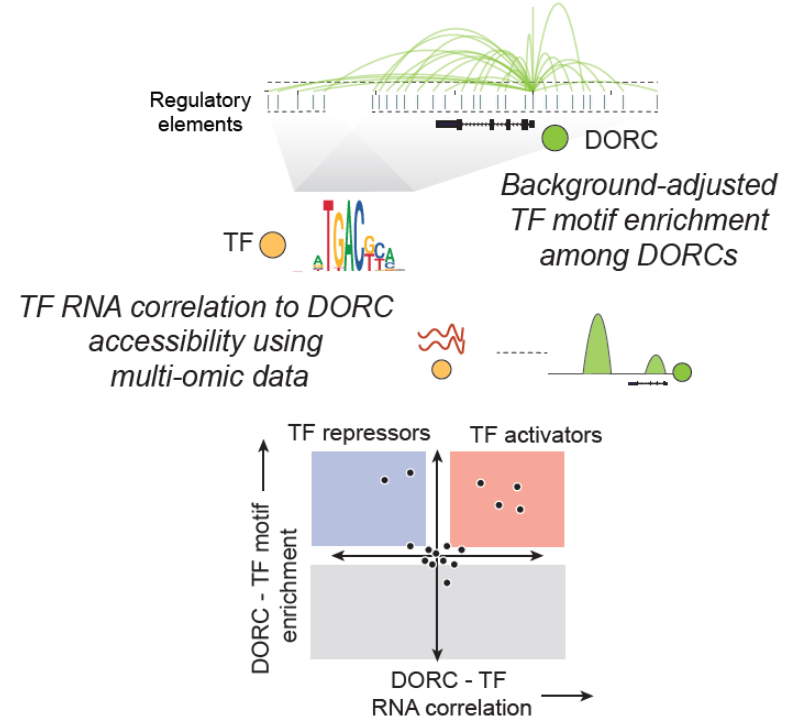
Cell types with gene expressed have peaks  
⇒ Chromatin compaction responsible for expression

## Part II. Multi-omics | Analysis

### Gene regulatory network inference

Goal: identify TFs regulating many genes (master regulators)

- Regroup all peaks associated with a given gene → domains of regulatory chromatin (DORCs)
- Find TF binding motifs associated with each DORC
- Test correlation with TF expression

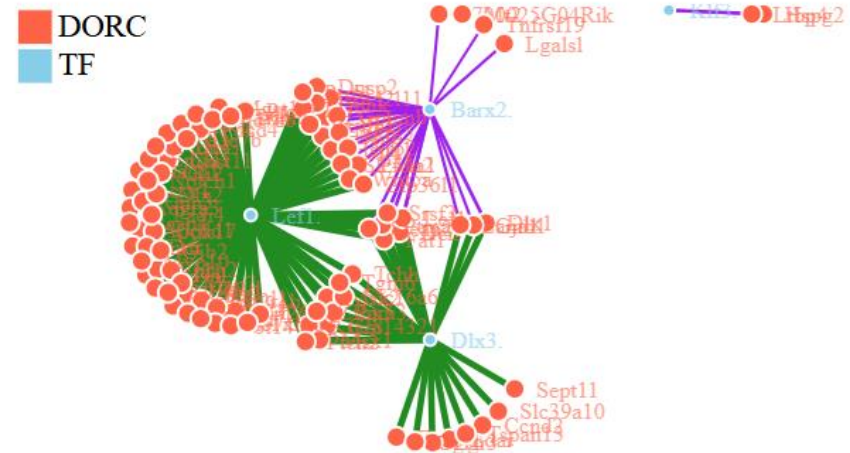


## Part II. Multi-omics | *Analysis*

### Gene regulatory network inference

Goal: identify TFs regulating many genes (master regulators)

- Regroup all peaks associated with a given gene → domains of regulatory chromatin (DORCs)
- Find TF binding motifs associated with each DORC
- Test correlation with TF expression
- Plot TF—DORC networks



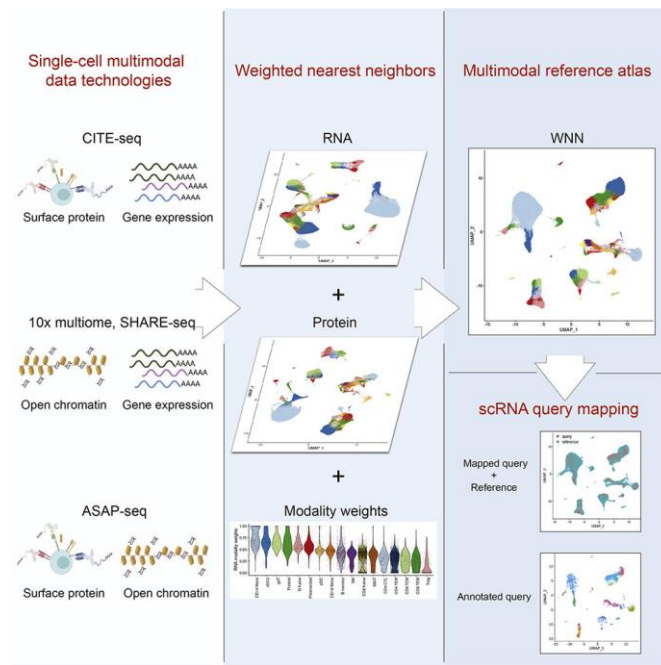
## Part II. Multi-omics | Analysis

### Coembeddings

Goal: produce a low-dimensional representation of both modalities at the same time

- Weighted nearest neighbor method (combines neighbor graphs from each modality)
- UMAP (also constructs graphs for each modality and can combine them)
- MultiVI (deep learning probabilistic model)
- MOFA+ (joint matrix factorization of the 2 matrices)

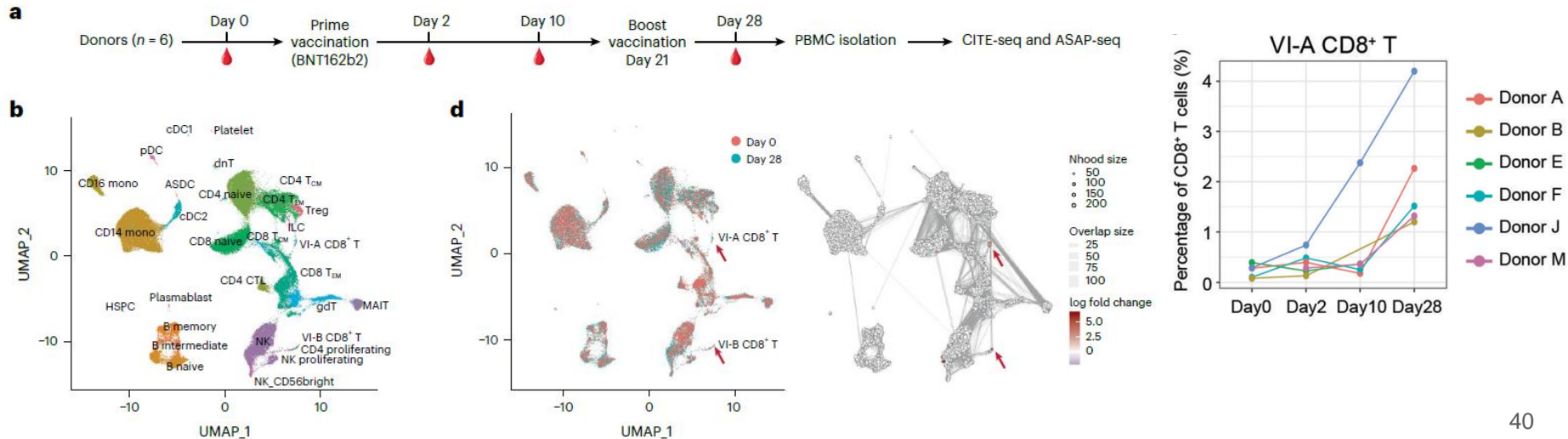
Can also allow to impute missing modalities for cells / samples



# Motivating example I

## What is the impact of COVID vaccination on patients immune systems?

- Longitudinal single-cell multiomics analysis
- Identify 2 vaccine-specific cell states

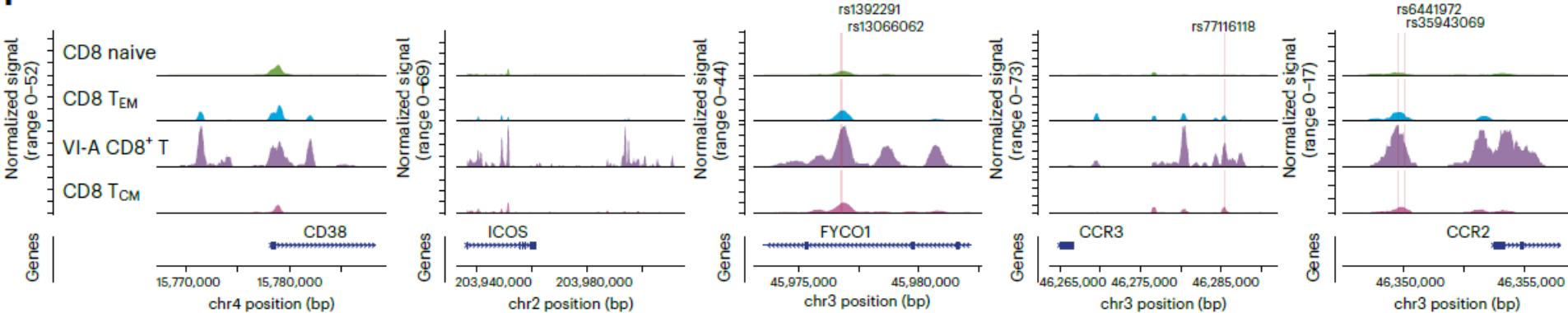


# Motivating example I

## What is the impact of COVID vaccination on patients immune systems?

- Longitudinal single-cell multiomics analysis
- Identify 2 vaccine-specific cell states and regulation mechanisms using chromatin states

f

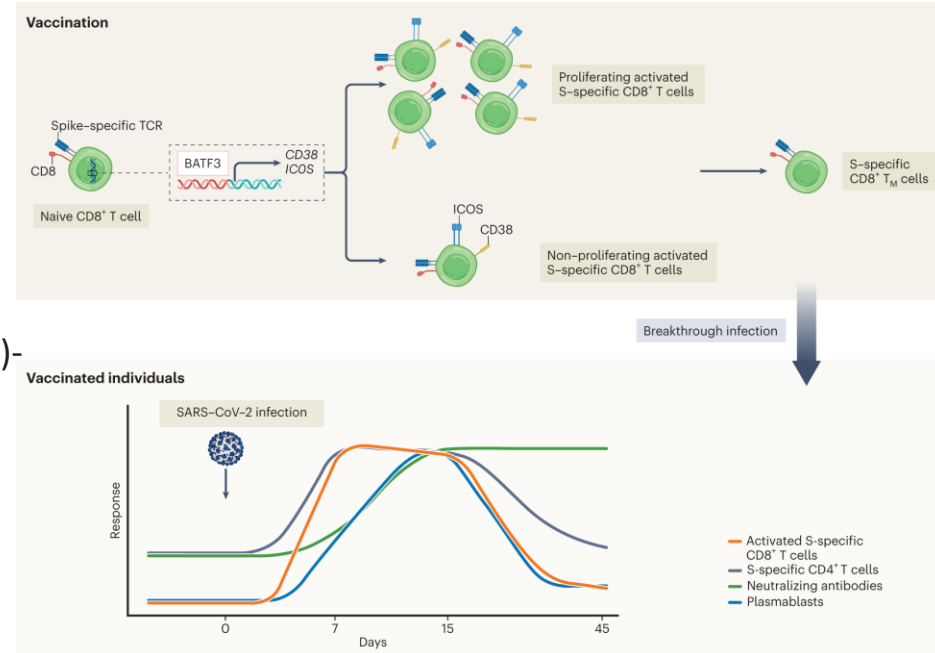




# Motivating example I

## What is the impact of COVID vaccination on patients immune systems?

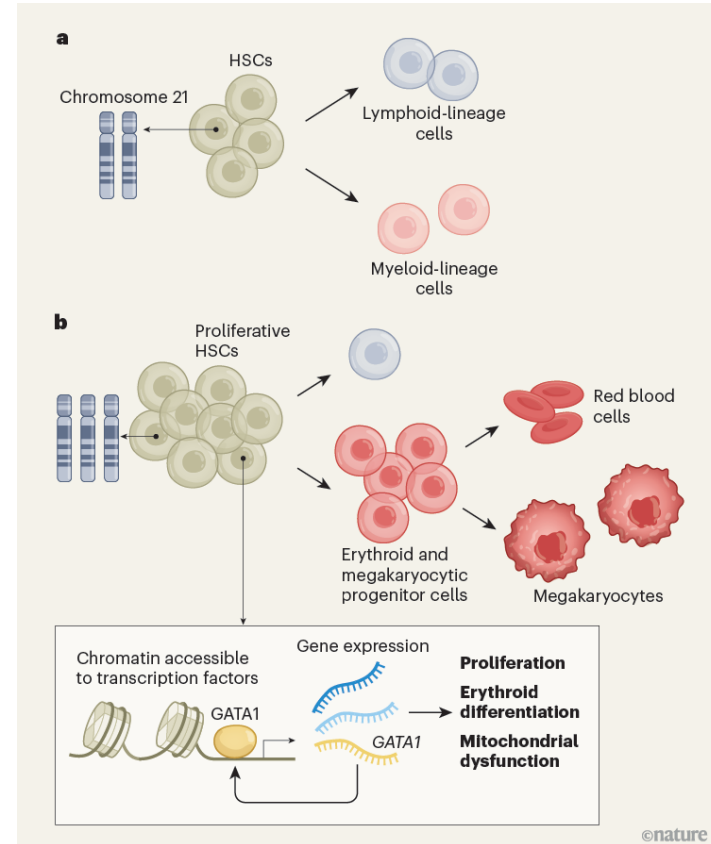
- Longitudinal single-cell multiomics analysis
- Vaccination activates naive CD8<sup>+</sup> T cells
- They differentiate into 2 vaccine-specific subsets: proliferating CD8<sup>+</sup>CD38<sup>+</sup>Ki67<sup>+</sup>KLRG1<sup>-</sup> and non-proliferating CD8<sup>+</sup>CD38<sup>+</sup>Ki67<sup>-</sup>KLRG1<sup>-</sup> cells
- These vaccine-induced T cell subsets generate spike (S)-specific memory CD8<sup>+</sup> T cells (T<sub>M</sub> cells)



# Motivating example II

## Why Down syndrome predisposes to leukemia?

- single-cell multiomics of >1.1 million cells from 3 fetuses with disomy and 15 fetuses with trisomy
- In people with 2 copies of chr21, haematopoietic stem cells (HSCs) produce lymphoid-lineage cells and myeloid-lineage cells
- Trisomy 21 makes chromatin accessible to TFs (e.g. GATA1)
- This makes HSCs more proliferative and biased towards erythroid and megakaryocytic lineage, that could drive the initiation of myeloid leukaemia



# Conclusions

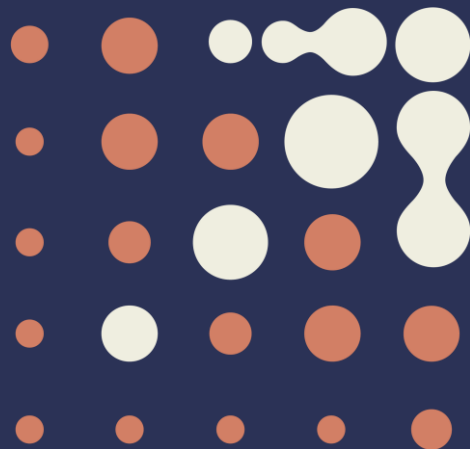
- Single-cell multiomics allows to better resolve cell states
- Combining epigenetics (methylation, chromatin compaction) + transcriptomics provides information about gene regulation and master regulators responsible for cell fate and disease

# Appendices

International Agency  
for Research on Cancer



World Health  
Organization



# Part I. Transcriptomics | Concepts

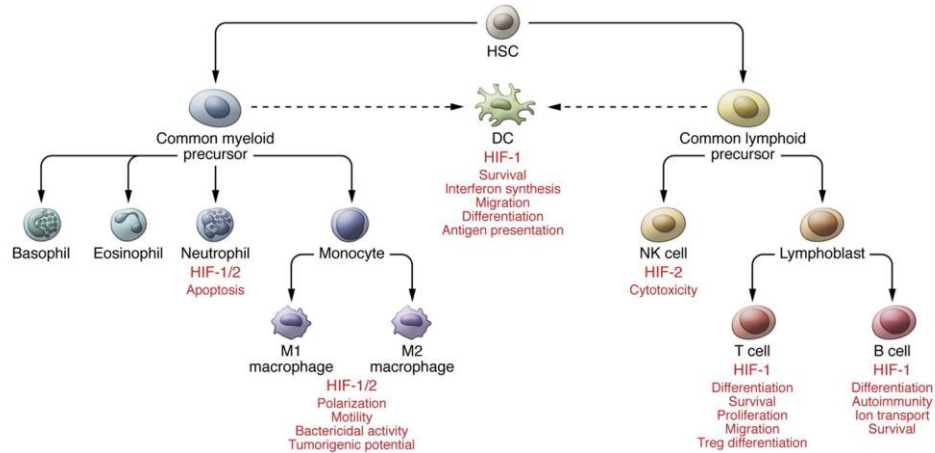
## Tissue heterogeneity: Stroma and Microenvironment

### Stromal cells (connective tissue cells)

- **Fibroblasts:** synthesize the extracellular matrix and collagen, initiate inflammation and immune response

### Immune cells

- **Dendritic cells:** present antigens
- **Macrophages:** perform phagocytosis
- **T cells:** cytotoxic (CD8+), helper (CD4+)
- **Neutrophils:** promote inflammation, phagocytosis



Immune cell differentiation. Source: Taylor et al. *J. Clin Invest* 2016.

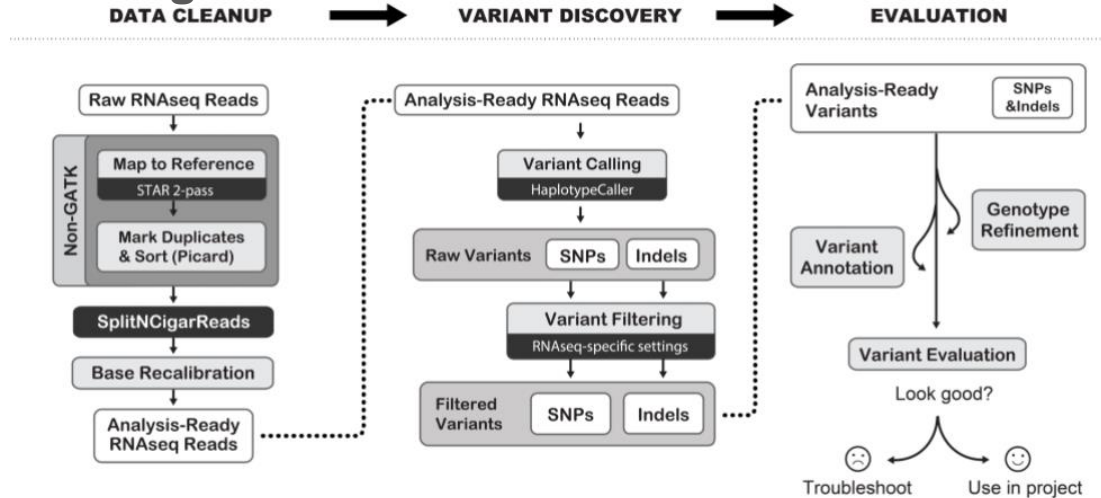
# Part I. Transcriptomics | Analysis

## Variant discovery: small variant calling

**Goal:** discover (or validate) small somatic variants (single nucleotide polymorphism or indels)

**Medical relevance:** many diseases are driven by small variants

**Methods:** Mapping to reference, and heavy filtering using estimated sequencing error rates and databases of known germline variants



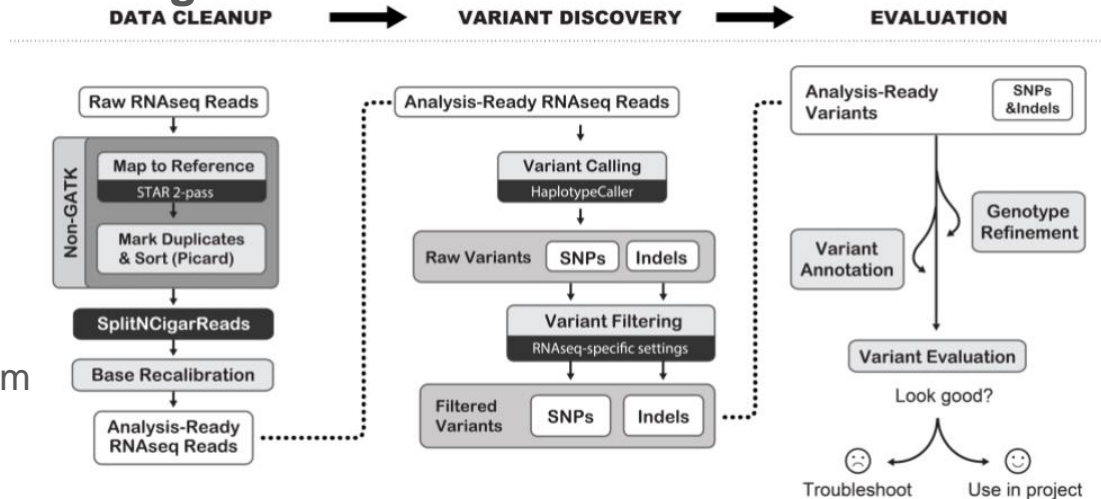
Schematic of the Genome Analysis ToolKit (GATK) best practices for small variant discovery from RNA-seq; Source: <https://gatk.broadinstitute.org>

# Part I. Transcriptomics | Analysis

## Variant discovery: small variant calling

### Caveats:

- High false positive rate (due to sequencing error and high depth at some locations)
- High false positive rate (due to variants in low-expression genes)
- Useful for validation of mutations from WGS/WES
- Useful for allele specific expression quantification



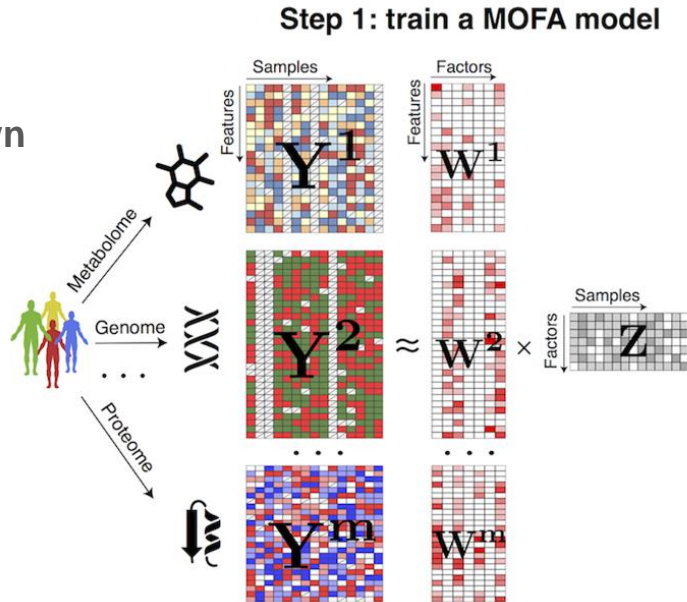
Schematic of the Genome Analysis ToolKit (GATK) best practices for small variant discovery from RNA-seq; Source: <https://gatk.broadinstitute.org>

## Part II. Multi-omics | *Analysis*

### Tools for integration: unsupervised analyzes

#### Multi-Omics Factor Analysis (MOFA)

- Identify **latent factors (unknown continuous variables)** representing biological variation **shared between modalities (e.g., genome, transcriptome)**



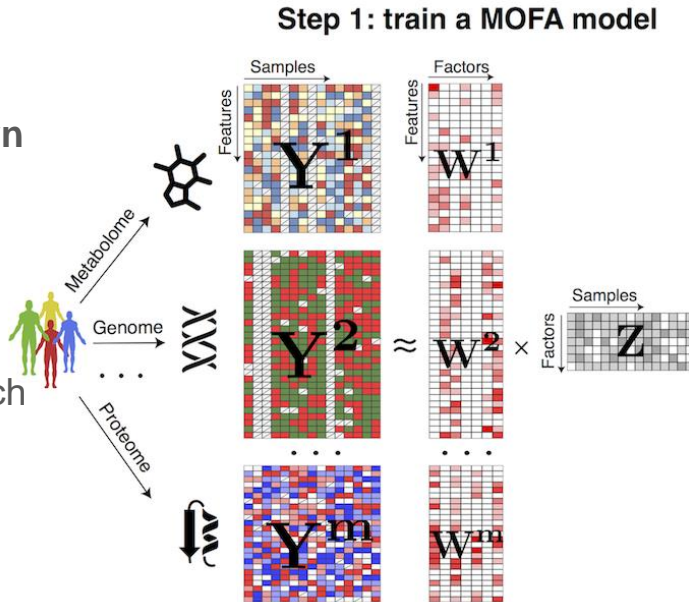


## Part II. Multi-omics | Analysis

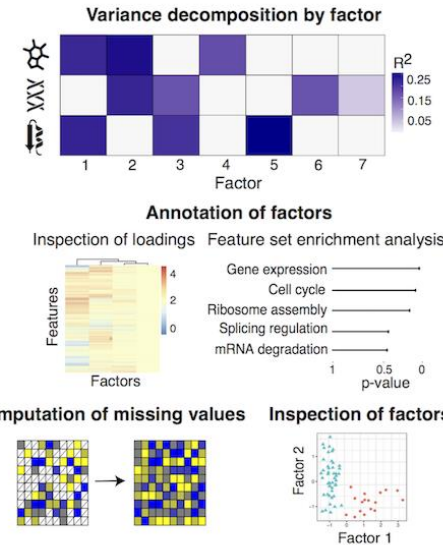
### Tools for integration: unsupervised analyzes

#### Multi-Omics Factor Analysis (MOFA)

- Identify **latent factors (unknown continuous variables)** representing biological variation **shared between modalities (e.g., genome, transcriptome)**
- Identify in which 'omic' layer each factor is active
- Downstream analysis to **understand what each factor represents**



#### Step 2: downstream analysis

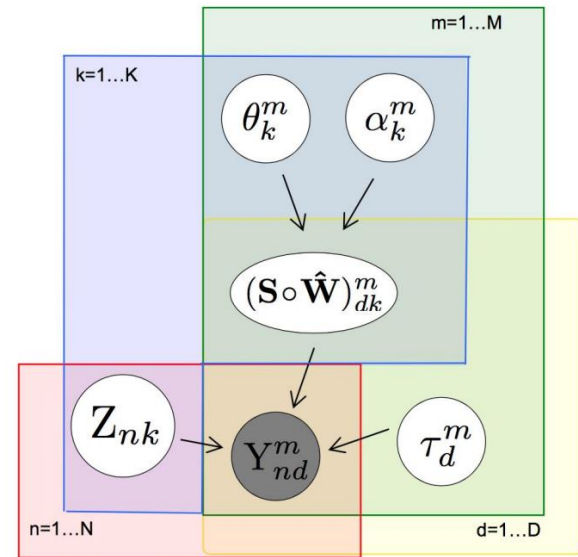


## Part II. Multi-omics | Analysis

### Tools for integration: unsupervised analyzes

#### Multi-Omics Factor Analysis (MOFA)

- **Generalization of Principal Component Analysis to multiple modalities  $M$**
- model  $\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^m + \boldsymbol{\varepsilon}^m$ ,
- where  $\mathbf{Y}^m$  is the matrix of observations for each sample  $n$  (rows) and each feature  $d$  (columns) for modality  $m$  (e.g., genomic alterations, expression)
- $\mathbf{Z}$  is the latent factors matrix ( $N$  by  $K$ ) shared by all modalities  $m$
- $\mathbf{W}^m$  is the weights (loadings) matrix ( $K$  by  $M$ ) of  $m$
- $\boldsymbol{\varepsilon}^m$  is the residual noise (column vector of size  $N$ )



MOFA directed acyclic graph. Source: Argelaguet et al. Mol Syst Biol 2018.

## Part II. Multi-omics | Analysis

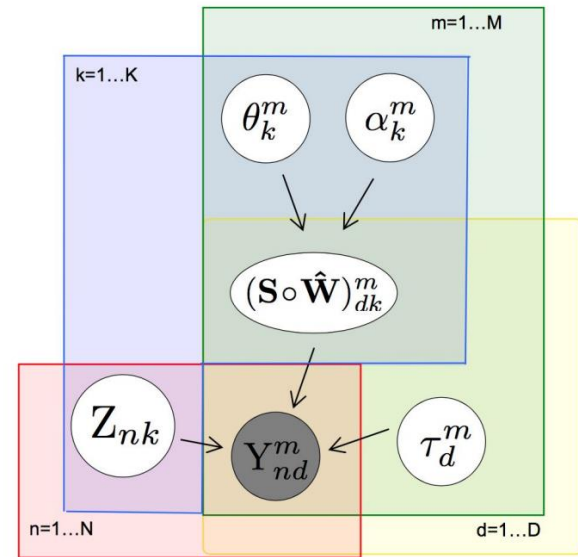
### Tools for integration: unsupervised analyzes

#### Multi-Omics Factor Analysis (MOFA)

- Model  $\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^m + \boldsymbol{\varepsilon}^m$

Important points:

- Because  $\mathbf{Z}$  is estimated from all 'omic' layers  $m$  and features  $d$ , the **model handles missing data naturally**
- The sparsity assumptions perform **automatic feature and factor selection**
- Technical artifacts**, usually restricted to a single modality  $k$ , are separated from variation with **evidence from multiple modalities**
- Correlations between modalities** are found (e.g.,



MOFA directed acyclic graph. Source: Argelaguet et al. Mol Syst Biol 2018.

## Part II. Multi-omics | Analysis

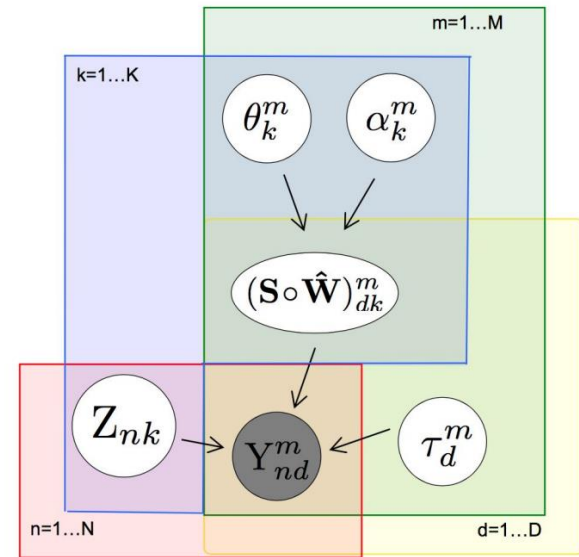
### Tools for integration: unsupervised analyzes

#### Multi-Omics Factor Analysis (MOFA)

- Model  $\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^m + \boldsymbol{\varepsilon}^m$

Important points:

- the likelihood formulation implicitly gives more weight to modalities with many features, so **beware of imbalance between input data matrices** (e.g., a mutation matrix of 20 features will not influence much  $\mathbf{Z}$  if an expression matrix with 10,000 features is also provided)



MOFA directed acyclic graph. Source: Argelaguet et al. Mol Syst Biol 2018.

## Part II. Multi-omics | Analysis

### Tools for integration: unsupervised analyzes

#### Other tools

- **Integrative clustering** (iCluster+; Mo *et al.* PNAS 2013):
  - integrative latent factors identification (similar to MOFA) for dimensionality reduction
  - then clustering in reduced space ( $K$ -means)
  - *Specificities*: latent factors are not directly interpreted; **emphasis on clustering rather than continuous analyses**

