# Gene regulatory network inference from single-cell data

**Supervisor:** Dr. Alireza Fotuhi Siahpirani

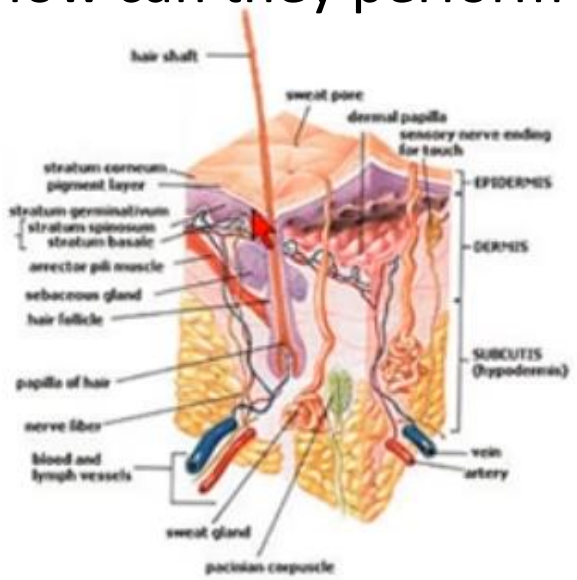
**Presenter:** Mahboobeh (Mariya) golchinpour leili
Bioinformatics Ph.D. student
Laboratory of Bioinformatics & Computational Genomics (LBCG)
Institute of Biochemistry & Biophysics (IBB)
University of Tehran, Iran

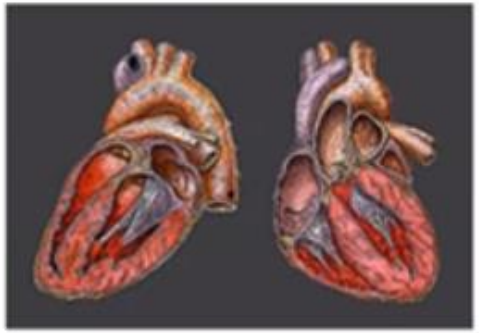
2024

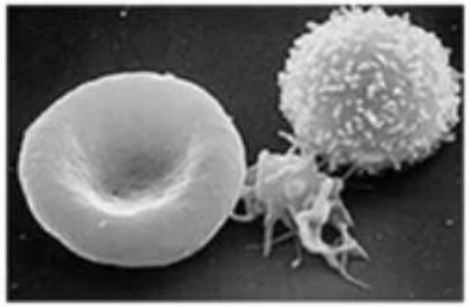
# One Genome –Many Cell Types

ACCAGTTACGACGGTCA
GGGTACTGATACCCCAA
ACCGTTGACCGCATTTA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTTAGCTACTGGTTTAG
CAATTTACCGTTACAAC
GTTTACAGGGTTACGGT
TGGGATTTGAAAAAAAG
TTTGAGTTGGTTTTTTC
ACGGTAGAACGTACCGT
TACCAGTA

How can they perform different functions?

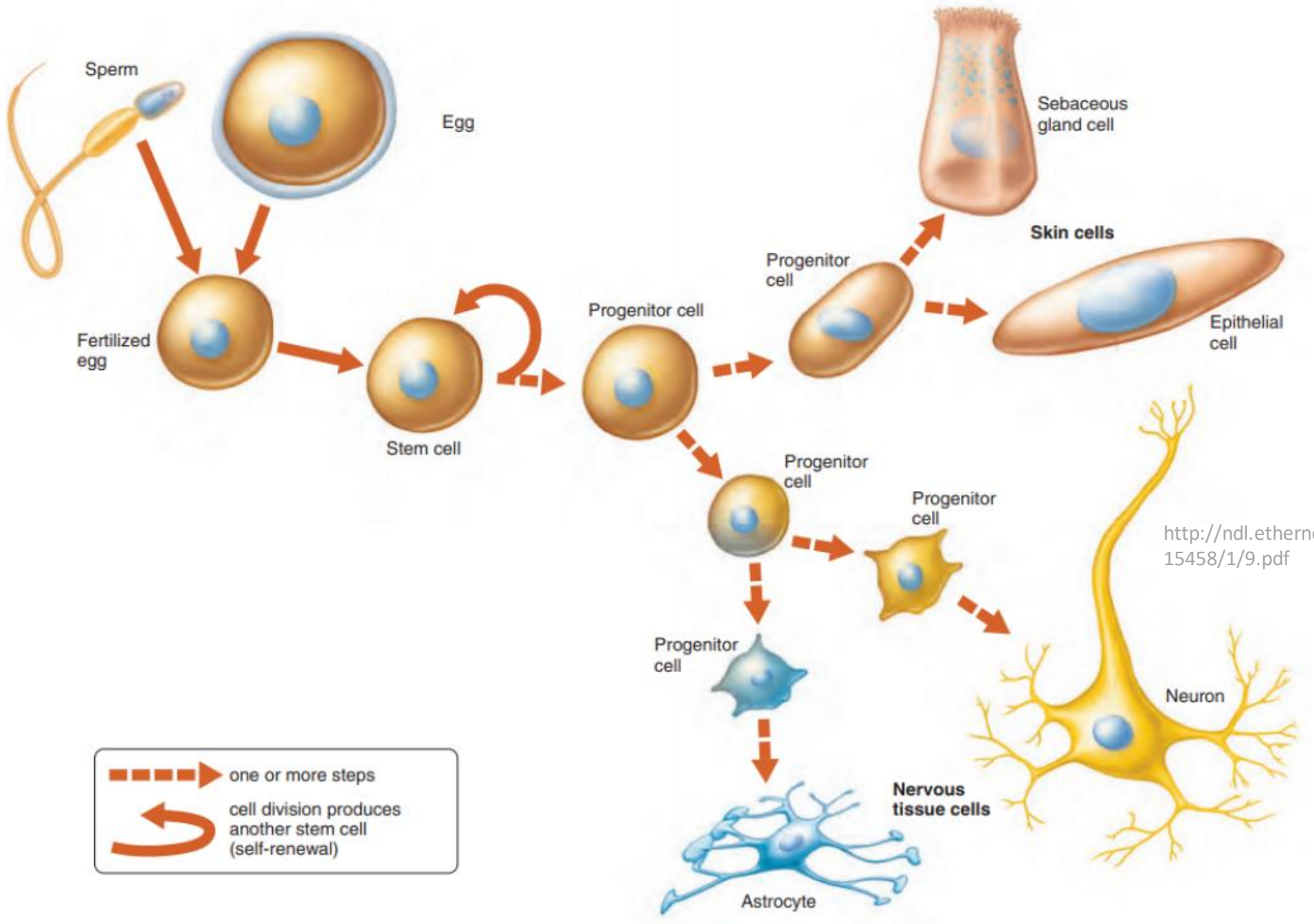




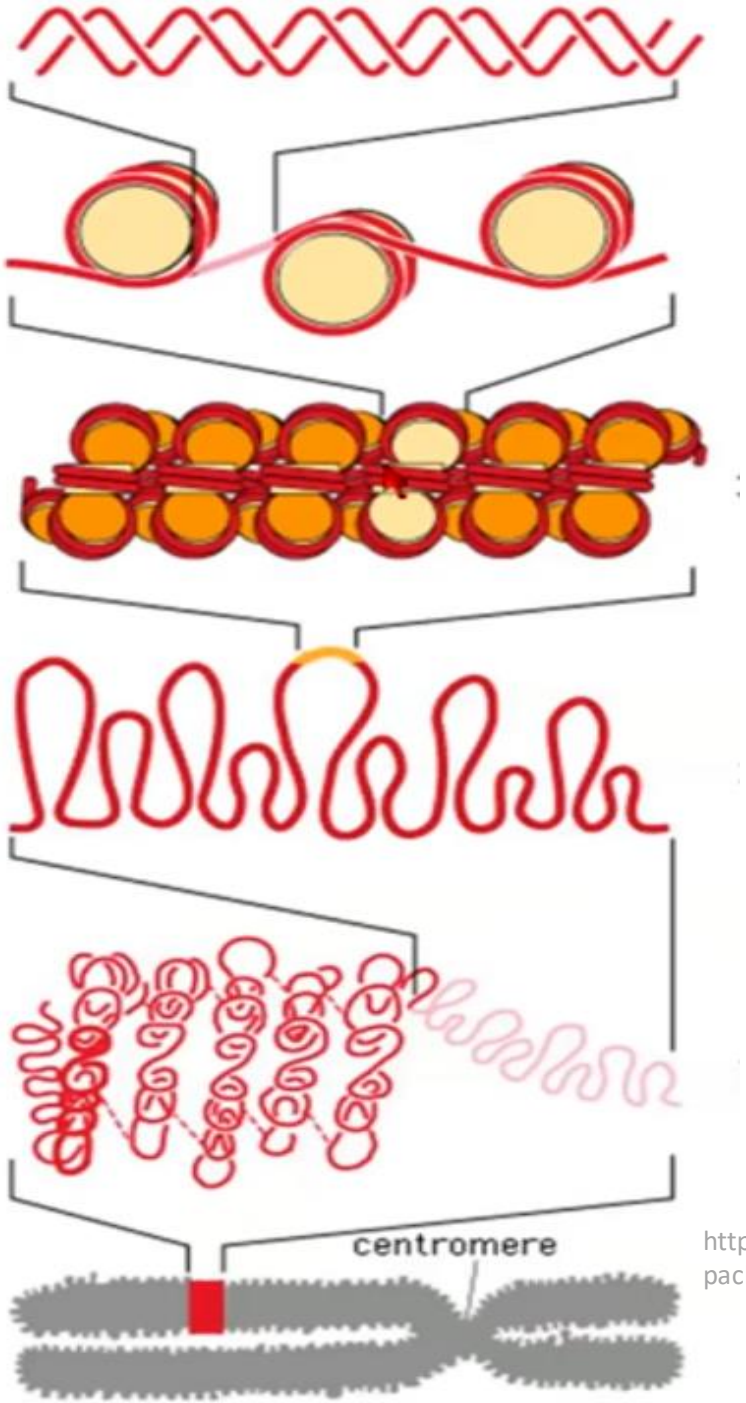

https://www.youtube.com/watch?v=6Ay5_NGCtPY

# Cell differentiates



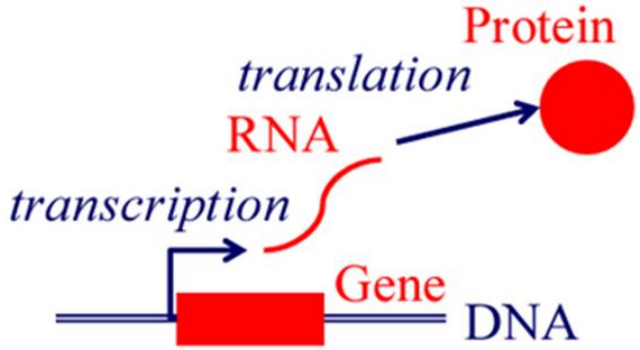

http://ndl.ethernet.edu.et/bitstream/123456789/15458/1/9.pdf

# DNA packaging And Gene regulation

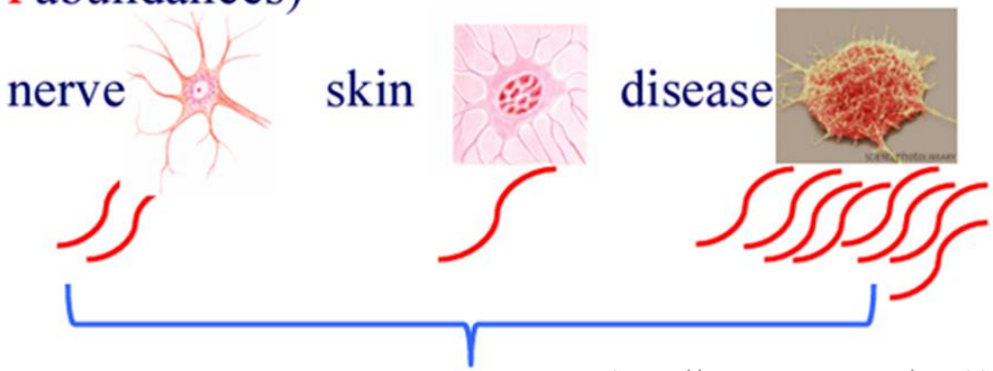



https://www.slideshare.net/slideshow/dna-packaging-137079008/137079008

Identical DNA but different gene expression



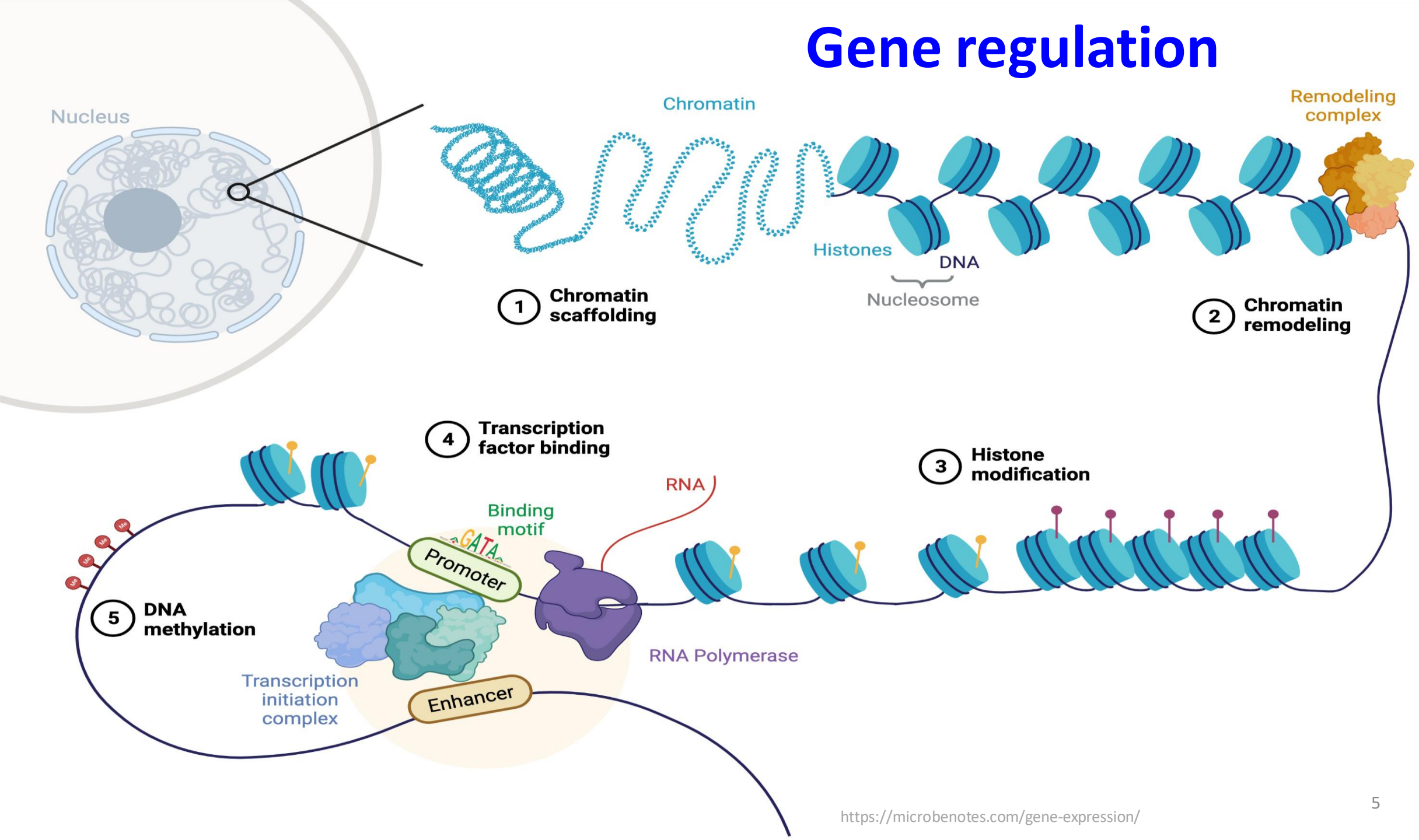

**Gene expression** levels (e.g., values to quantify RNA abundances)




https://www.nature.com/scitable/topicpage/gene-expression-14121669/

**Gene regulation**: which & how genes express?

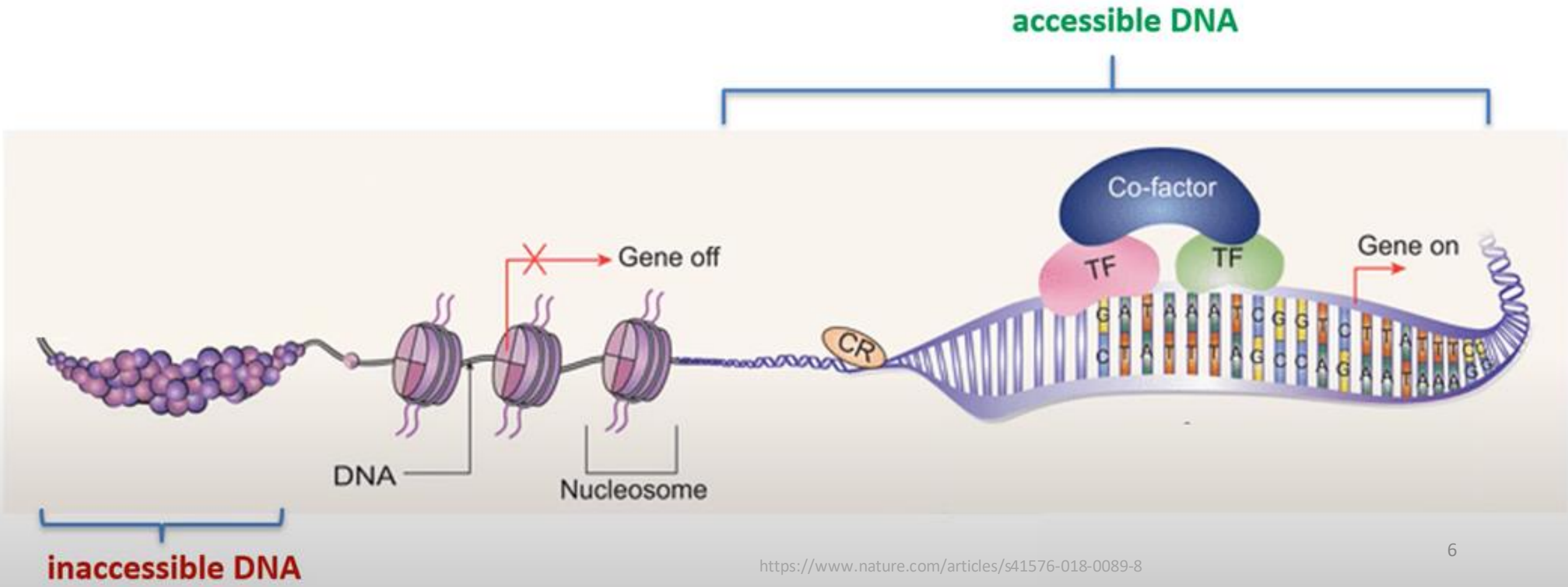
# Gene regulation




https://microbenotes.com/gene-expression/

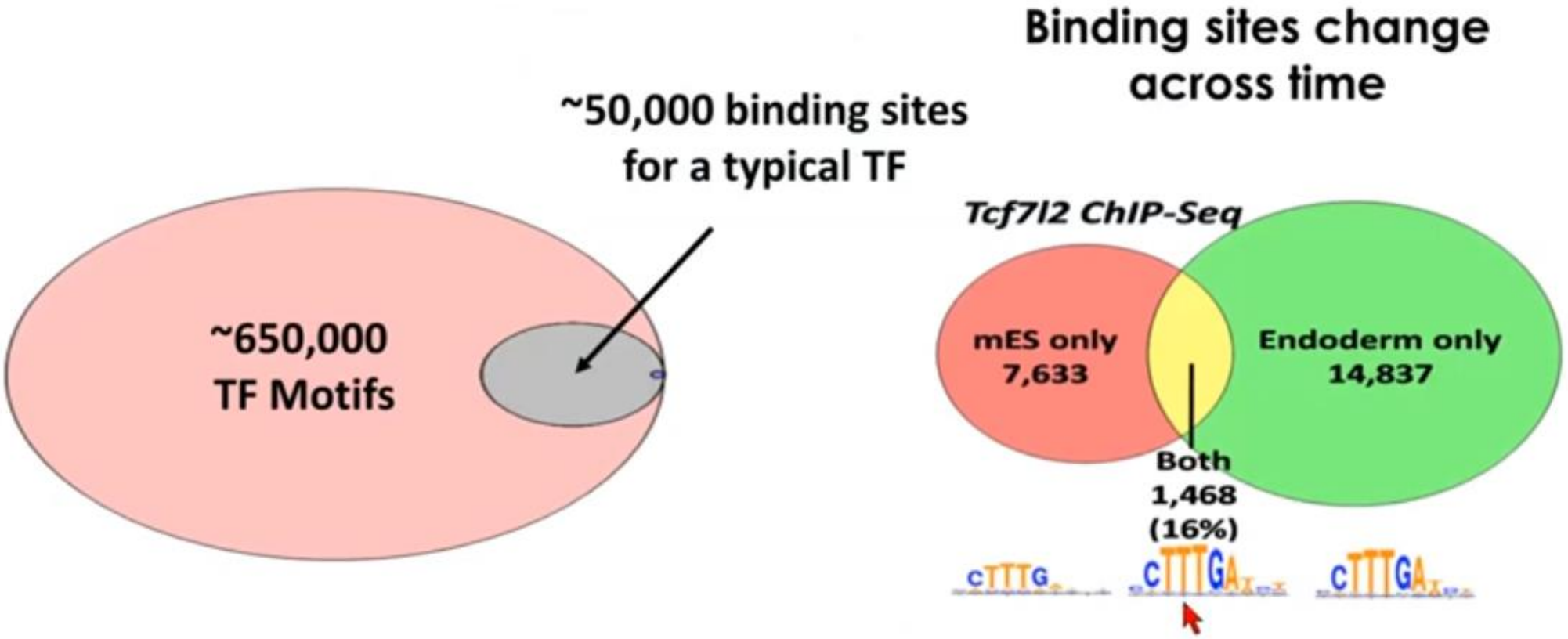
# Accessible chromatin and Transcription factor (TF) binding

- TFS binds to DNA at transcription factors binding sites (TFBSs)




https://www.nature.com/articles/s41576-018-0089-8

# Motifs can predict TF binding




https://www.youtube.com/watch?v=KCn_99lXCs8

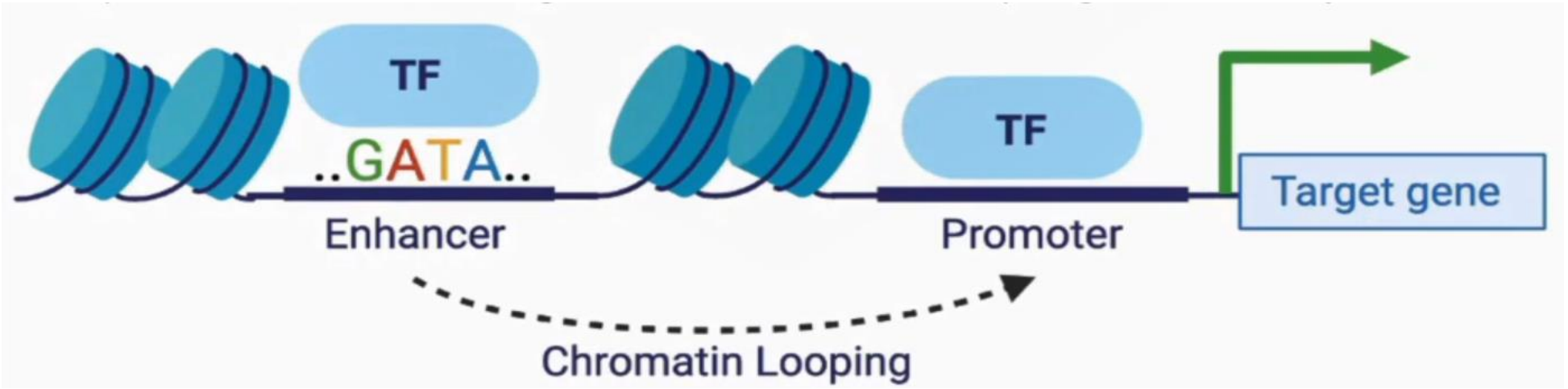
# Transcription Factors

"Transcription factors (TFs) directly interpret the genome" Lambert, Samuel A., et al (2018)

~1,600 TFs in human
100's expressed at any given time

1) TFs have a DNA-binding domain

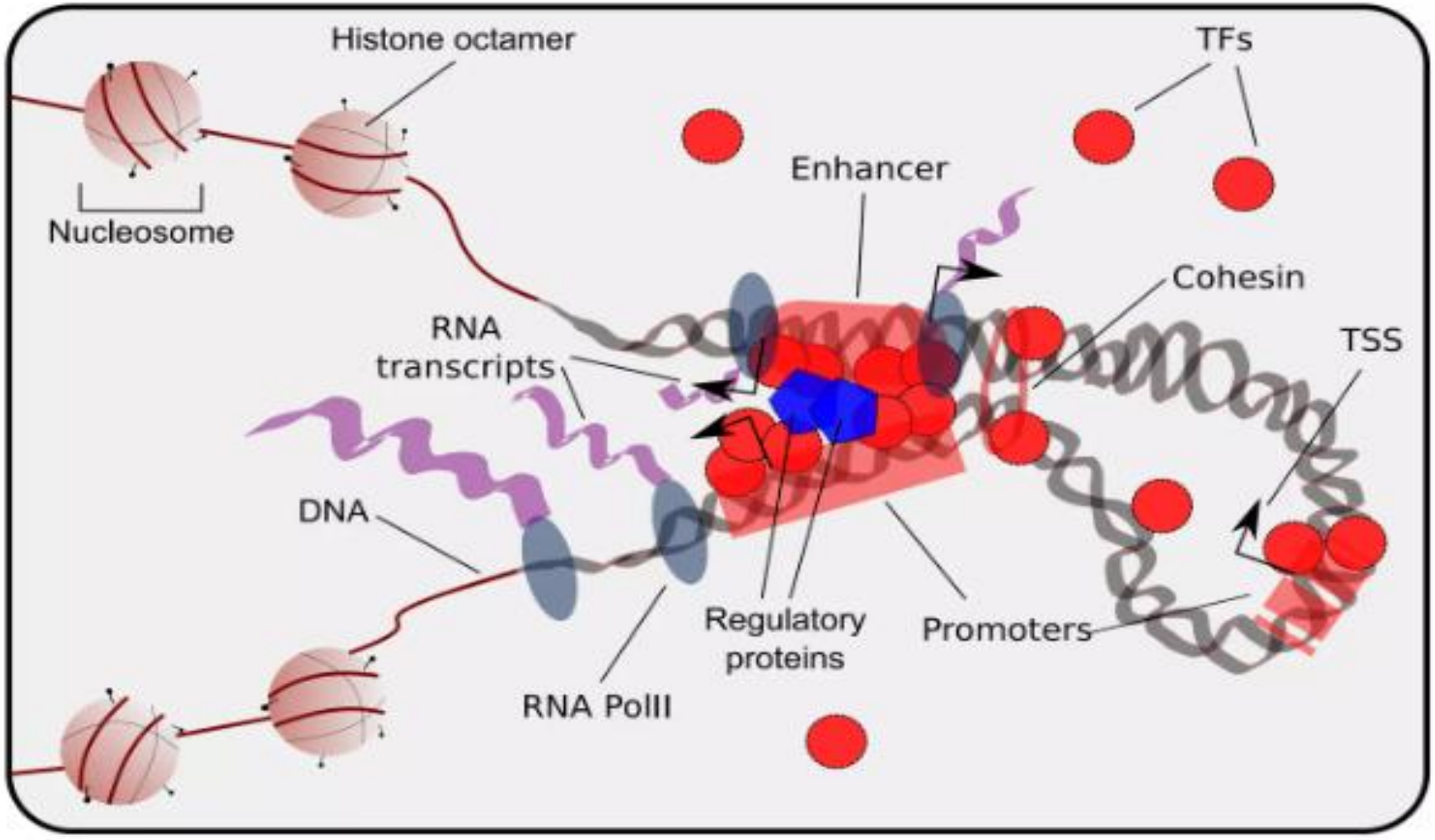
2) **Regulate transcription**




https://www.cincinnatichildrens.org/research/divisions/i/immunobiology/labs/miraldi/research

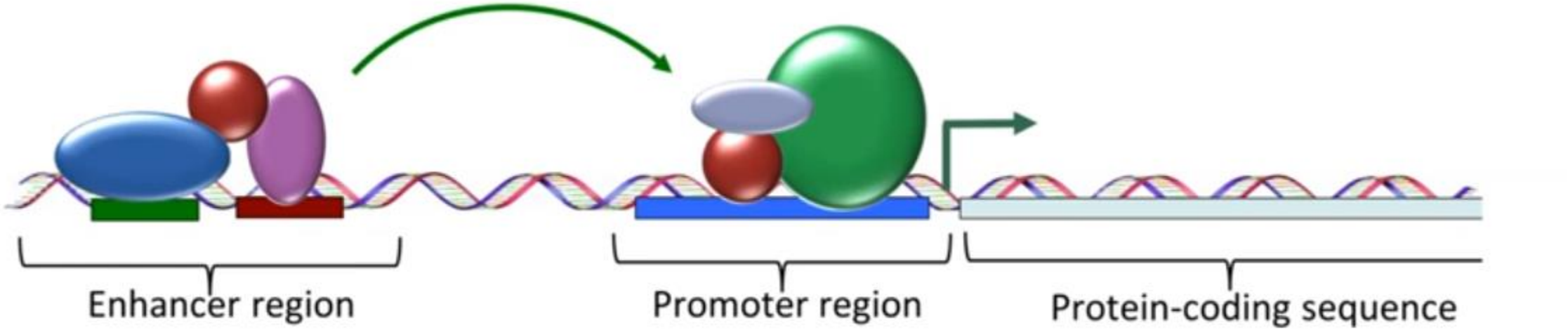
# Transcription factor Binding Site

- TFBS are often located in: Gene promoters ,Distal regulatory elements, such as: enhancers, silencers, insulators.




https://www.youtube.com/watch?v=aq8PAM5Sa0s

# Transcription factors control activation of cell-type –specific promoters and enhancers

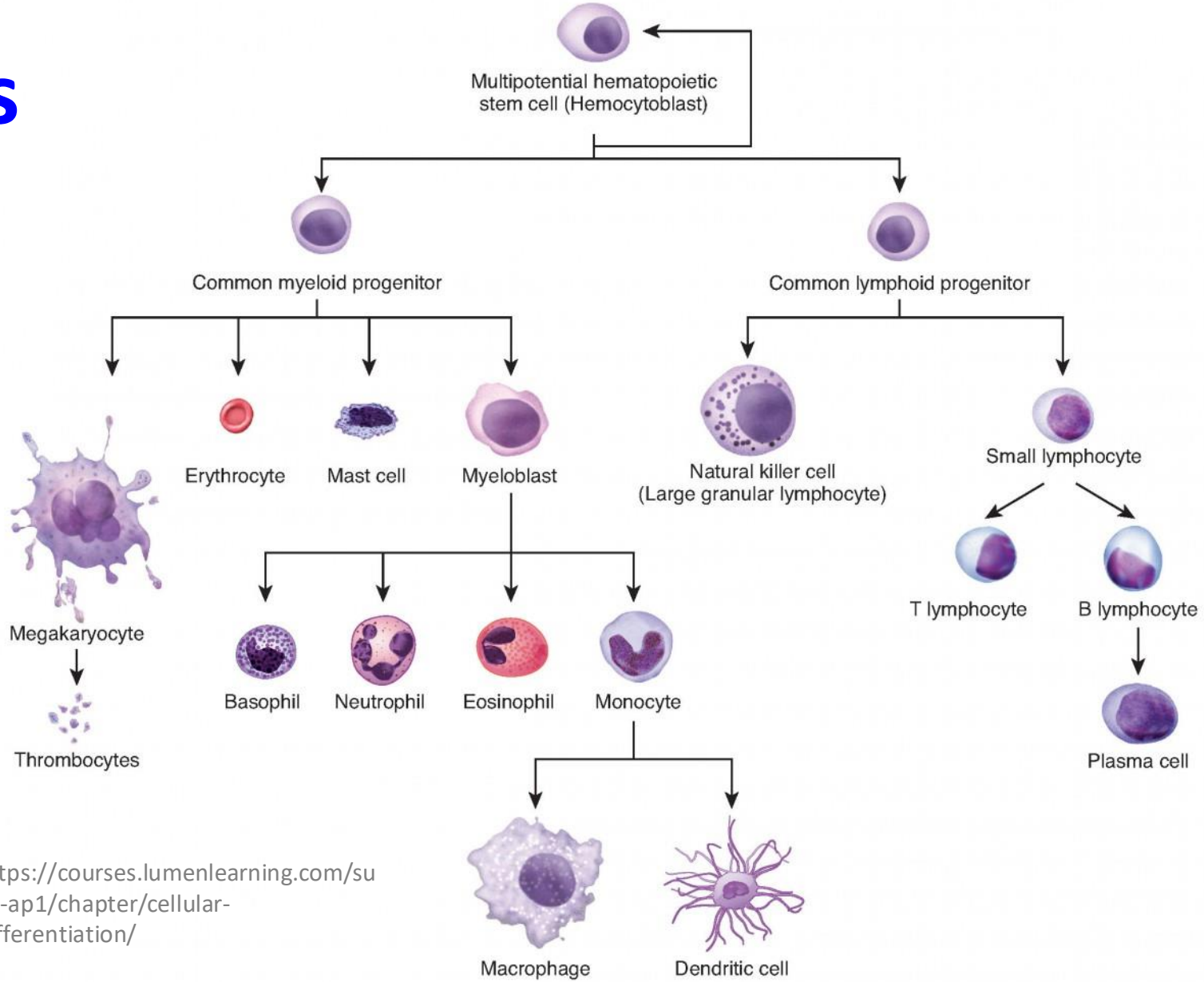



https://www.youtube.com/watch?v=KCn_99lXCs8

# Cell differentiates

**Hematopoiesis.** The process of hematopoiesis involves the differentiation of multipotent cells into **blood** and **immune** cells.

The multipotent hematopoietic stem cells give rise to many different cell types, including the cells of the immune system and red blood cells.

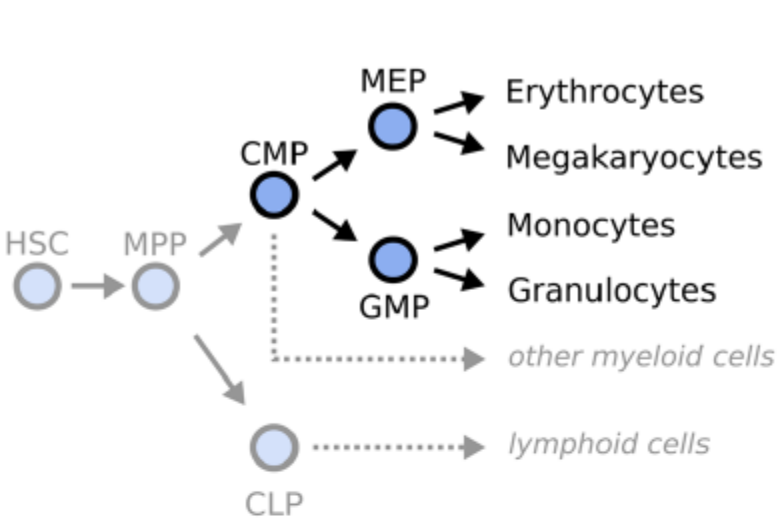


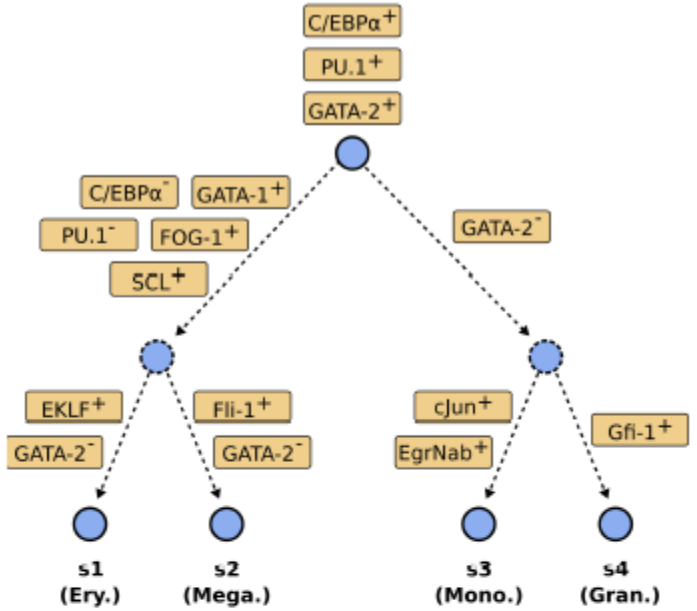

https://courses.lumenlearning.com/suny-ap1/chapter/cellular-differentiation/



https://courses.lumenlearning.com/suny-ap1/chapter/cellular-differentiation/

# Cell differentiates



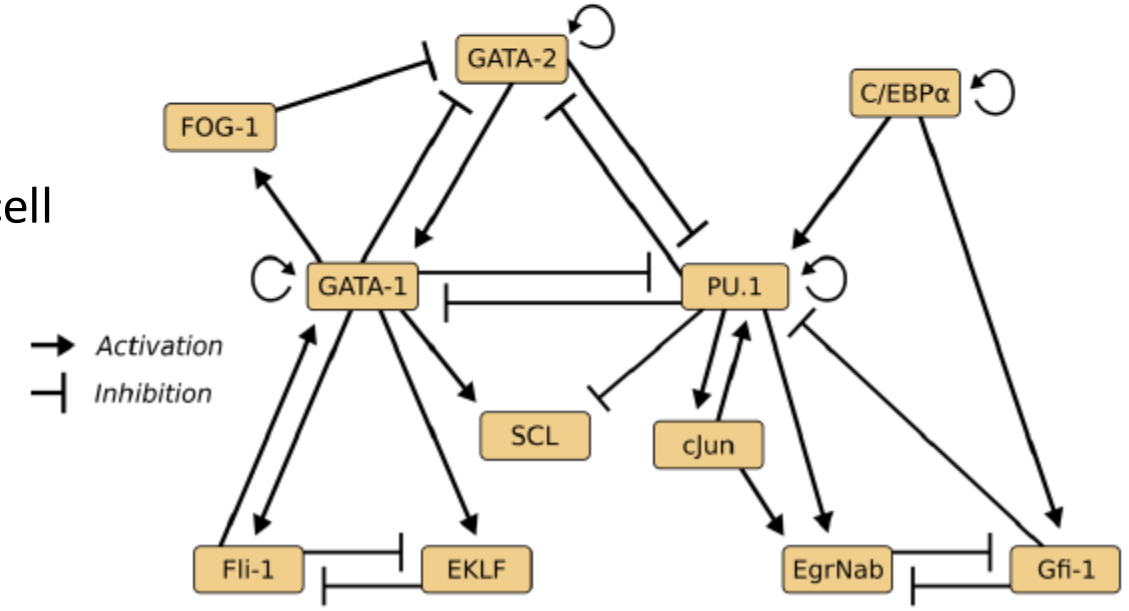




Cells in different states express different sets of genes

Cells move from one"state" to another.

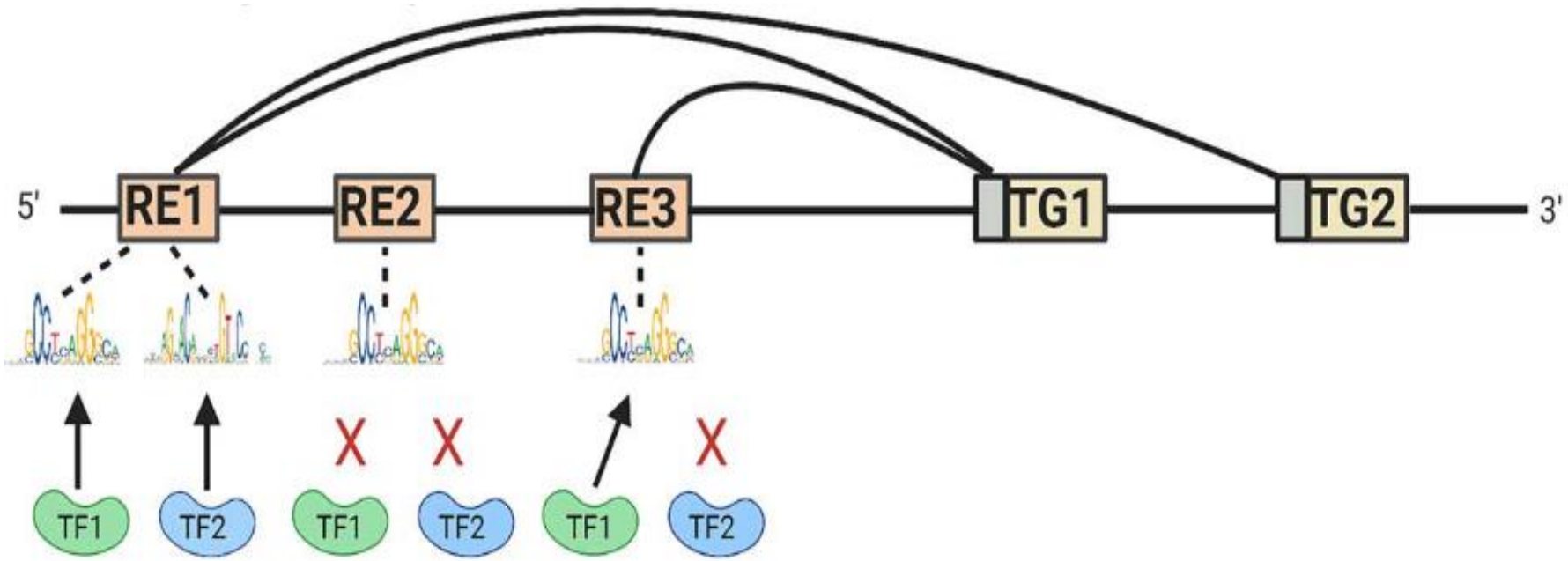
Transcription factors activate/inhibit genes to effect cell

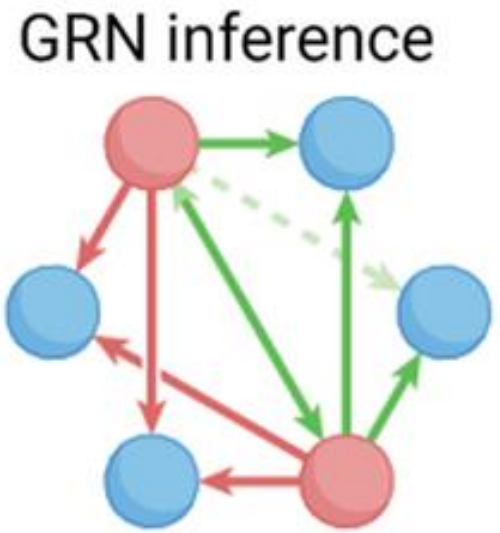
transition from one state to another.



# Gene regulatory networks (GRNs)

The interplay between transcription factors, chromatin and genes, generates complex regulatory circuits



# Subnetwork of gene regulatory network of NR3C1

Nodes: genes and TF

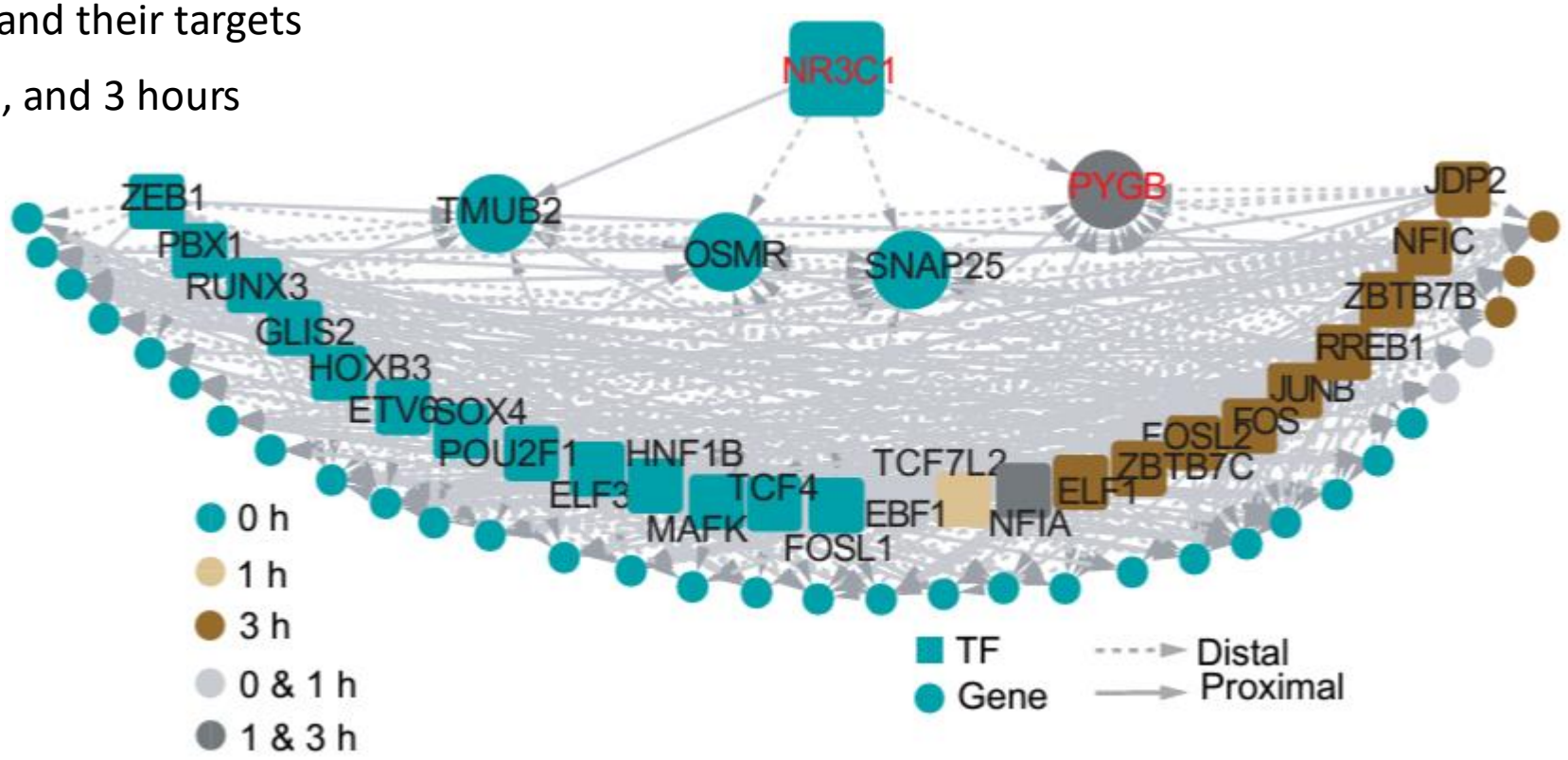
Edges: regulatory relations between TFs and their targets

Connections for NR3C1 signals across 0, 1, and 3 hours

Edges are directed or undirected (causality relationship between genes or lack of it)

Edges are signed (denoting the mode of regulation, positive or negative)

Edges are weighted or not (denoting the strength of the interaction)

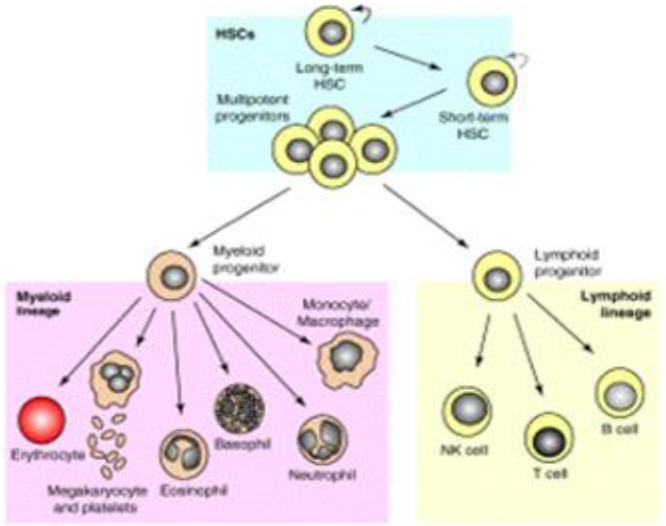



DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data

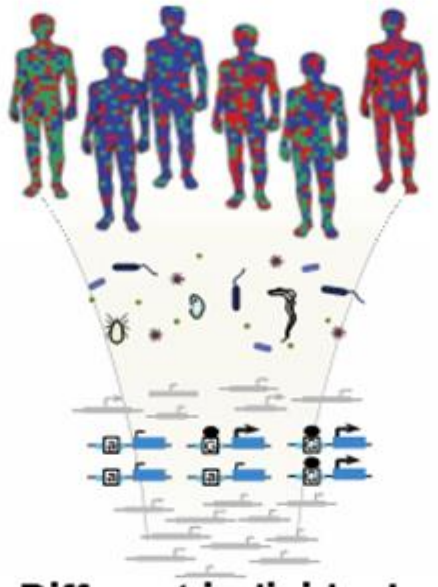
# Why GRNs is useful?

GRNs are context-specific

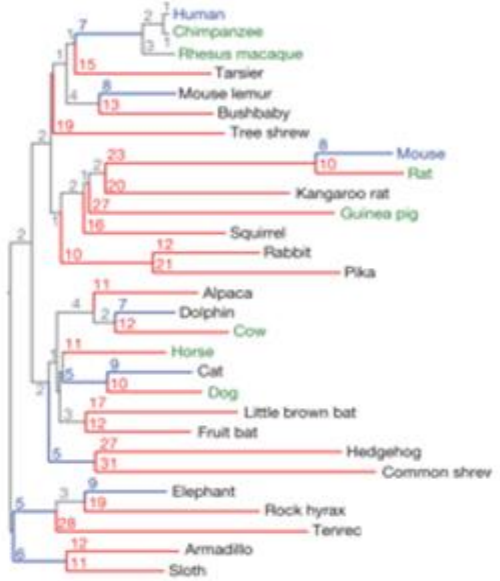
The study of GRNs is useful to understand how cellular identity is established, maintained and disrupted in different time, developmental stages, tissue, cell type, organ, disease and species.




Different cell types



Different individuals




Different species

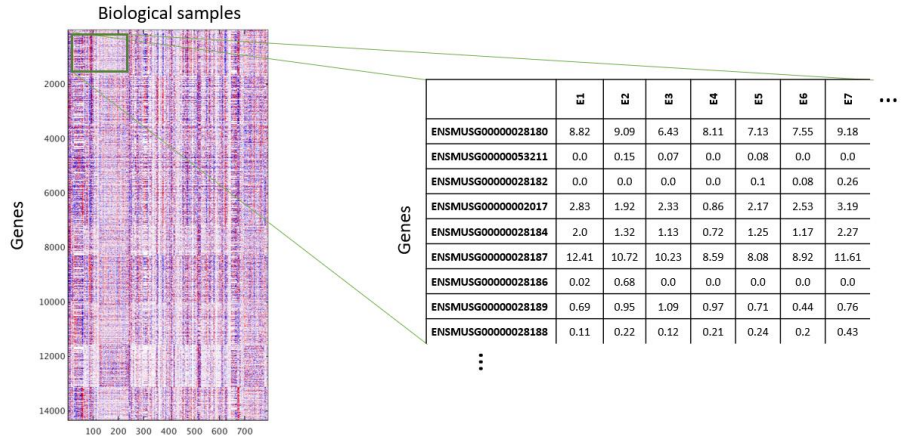

Lindblad-Toh et al. Nature, 2011.
Ye et al. Science, 2014.
Cabrita et al. Trends in Biotechnology 2003.

# Evolution of GRN inference and Sequencing Technologies

# Bulk RNA Sequencing

- Sequences RNA from a mixed population of cells as a single sample.
- Masks individual differences by averaging the expression across all cells.
- Fails to capture cellular heterogeneity and intermediate cell states.
- Difficult to experimentally purify or identify cells in transitional states.

**Gene expression profiling experiments produce expression matrices**







https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

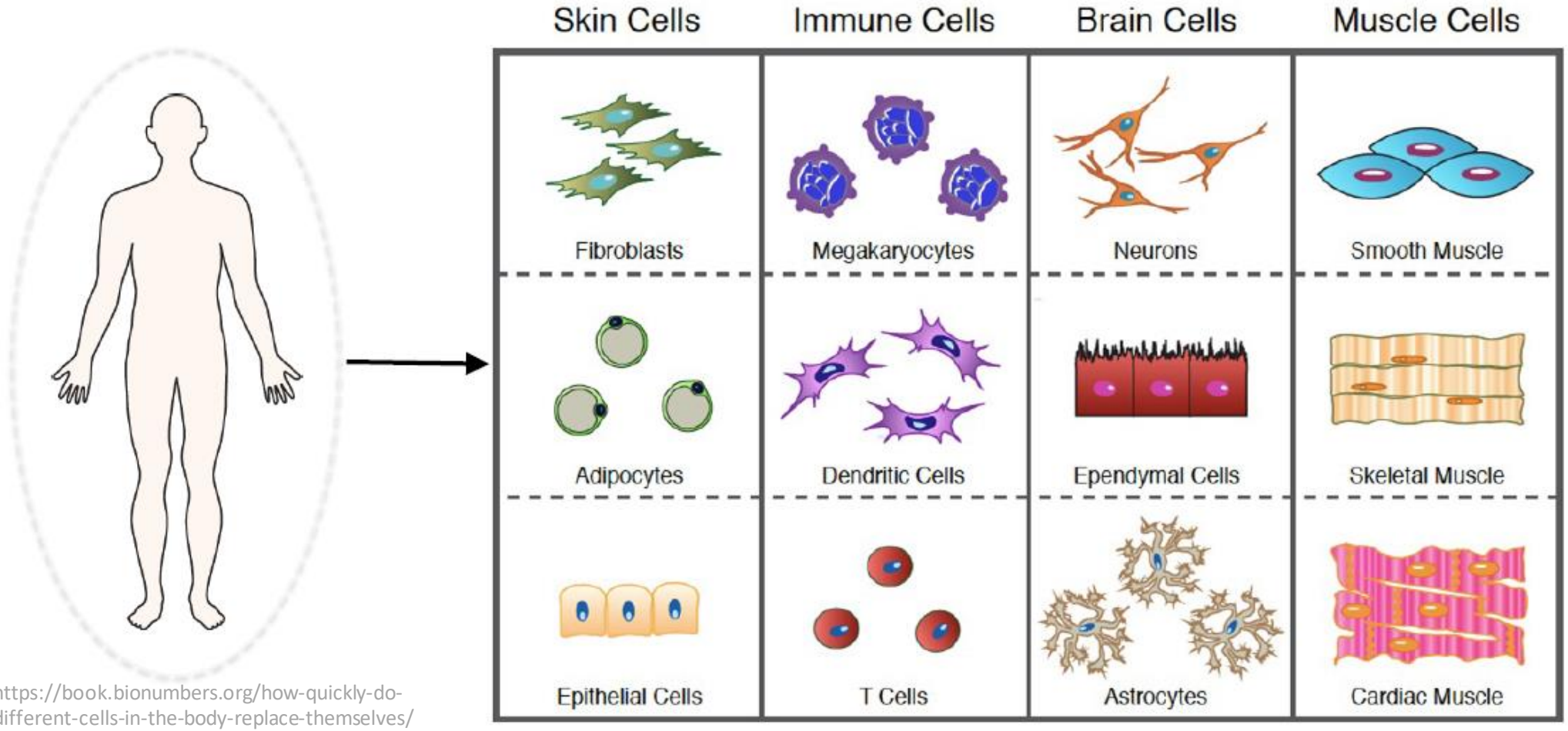
# Limitation of bulk omics data for GRN

GRN inference methods using bulk omics data have enabled the characterization of **genome-wide regulatory events** at the **tissue level**

But in the case of mixed samples such as tissues, **they cannot capture** the **cell type** or **state specificity** of GRNs **(limited in capturing cellular heterogeneity)**

# Why study single cells?

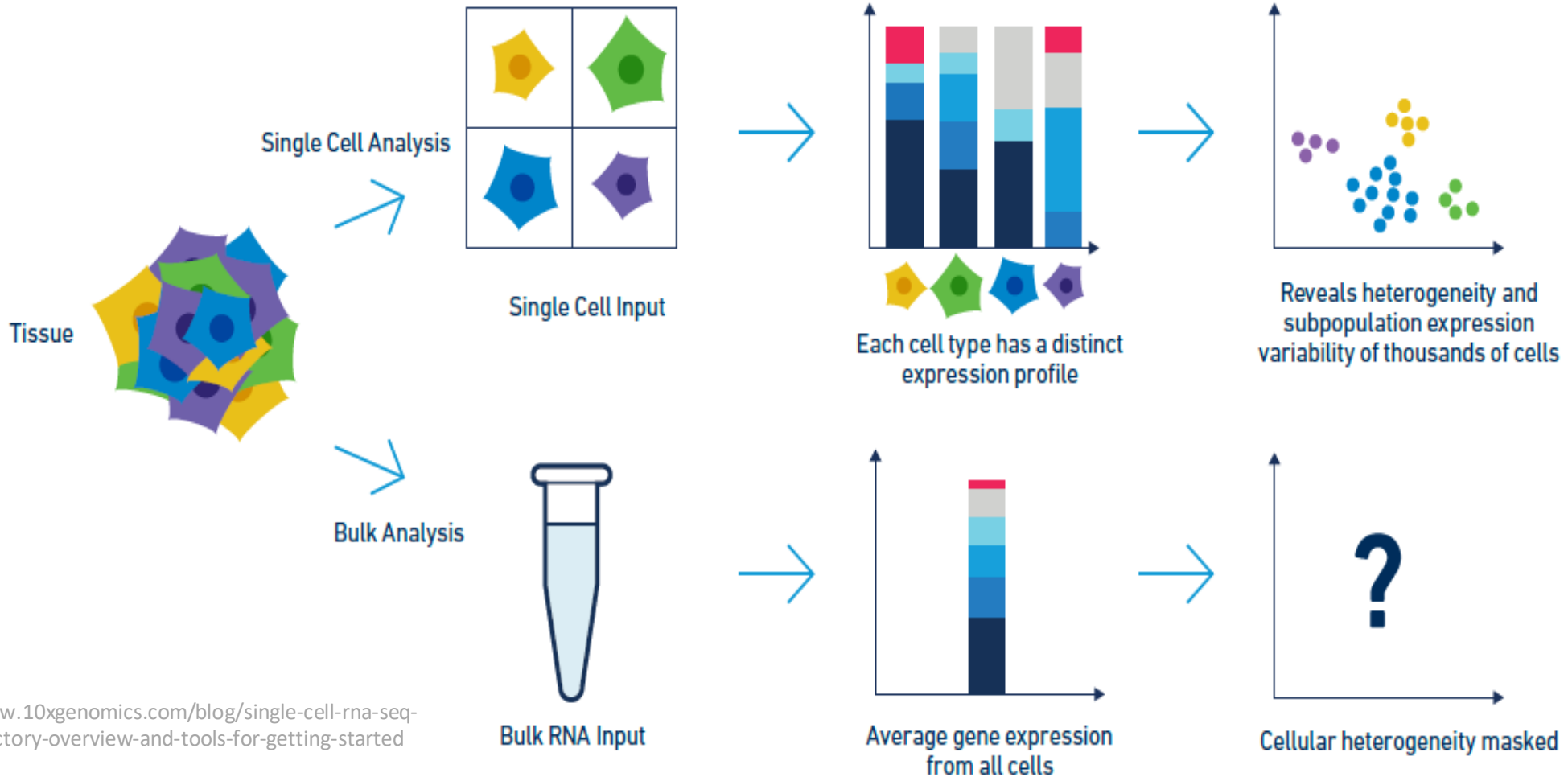
Cells are our constituents, are classified by characteristic molecules, structures, and functions




https://book.bionumbers.org/how-quickly-do-different-cells-in-the-body-replace-themselves/

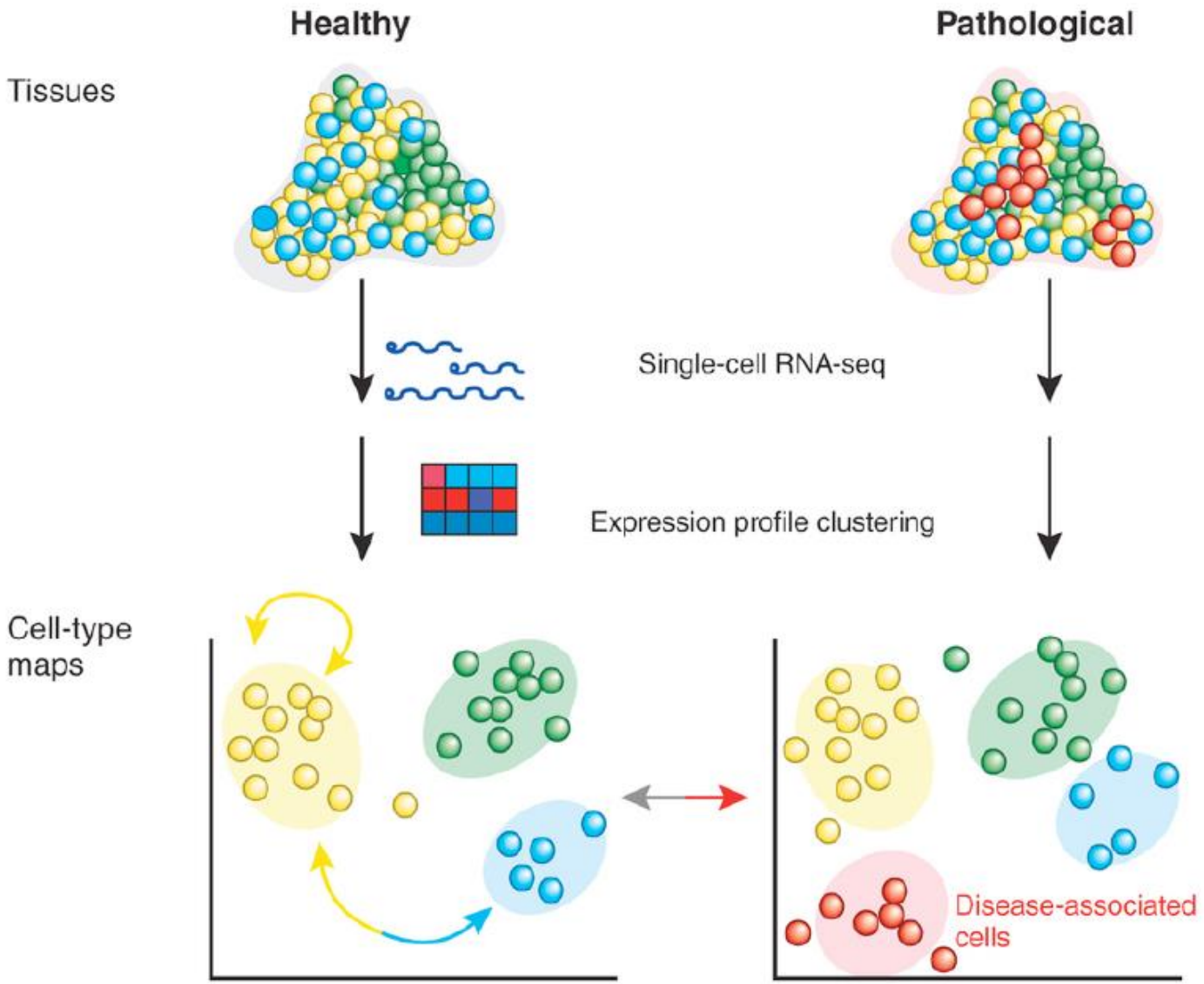
# Single-cell RNA Sequencing (scRNA-seq)

Single cell gene expression profiling reveals cellular heterogeneity that is masked by bulk RNA-seq methods.




https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

# Why study single cells?

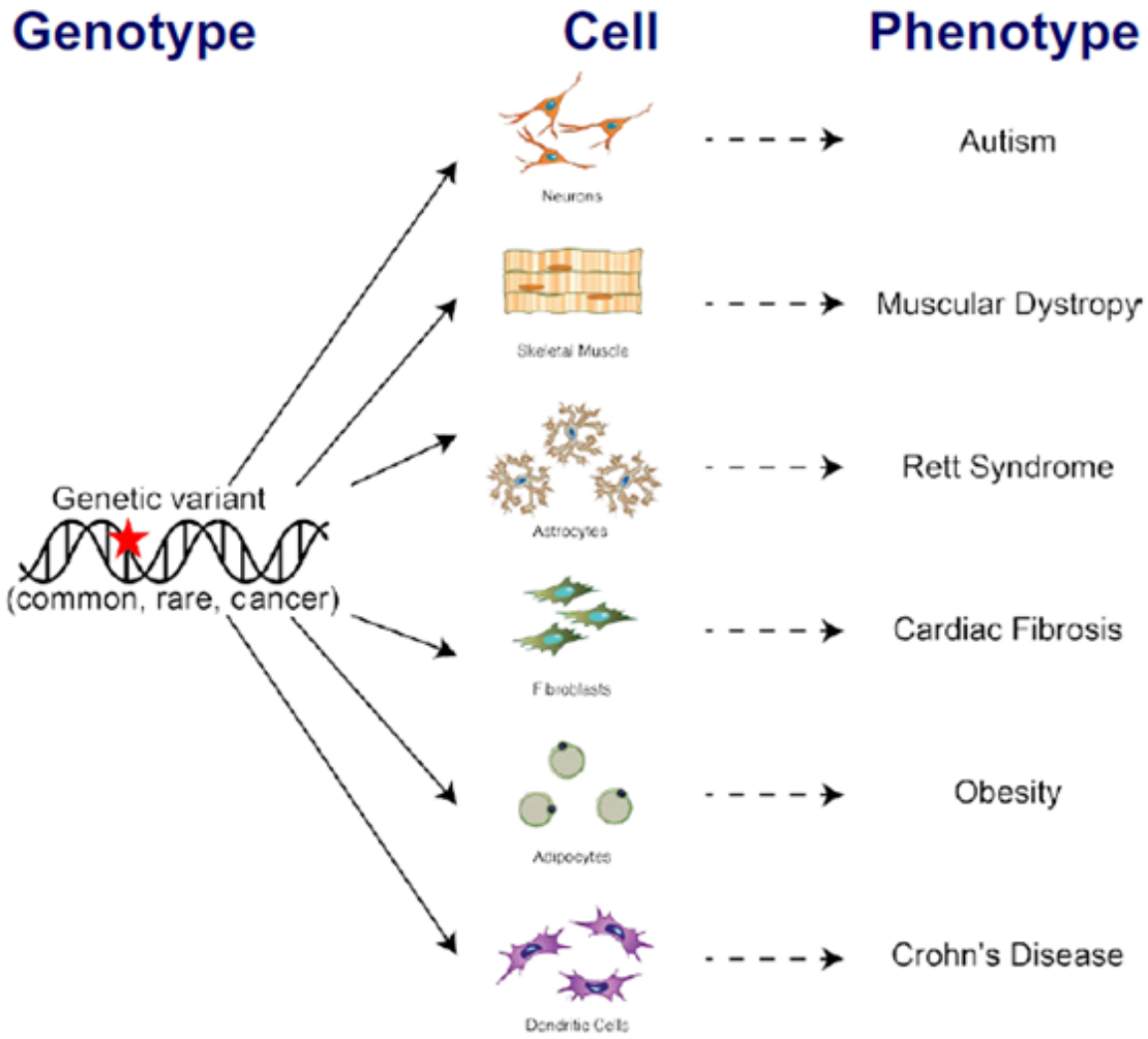


# Why study single cells?

How genetic variants, whether common, rare, or associated with cancer, affect different cell types?

Cells are key intermediate from genotype to phenotype.

Each cell type depicted can lead to a different phenotype or disease condition based on how these genetic variants manifest at the cellular level.




https://bioinformatics-core-shared-training.github.io/cruk-summer-school-2018/SingleCell/slides/2018-07-25_CRUK_CI_summer_school-scRNAseq.pdf

# Single cell limitation

- **High Noise Levels**: Single-cell data can exhibit a high level of noise due to biological variability

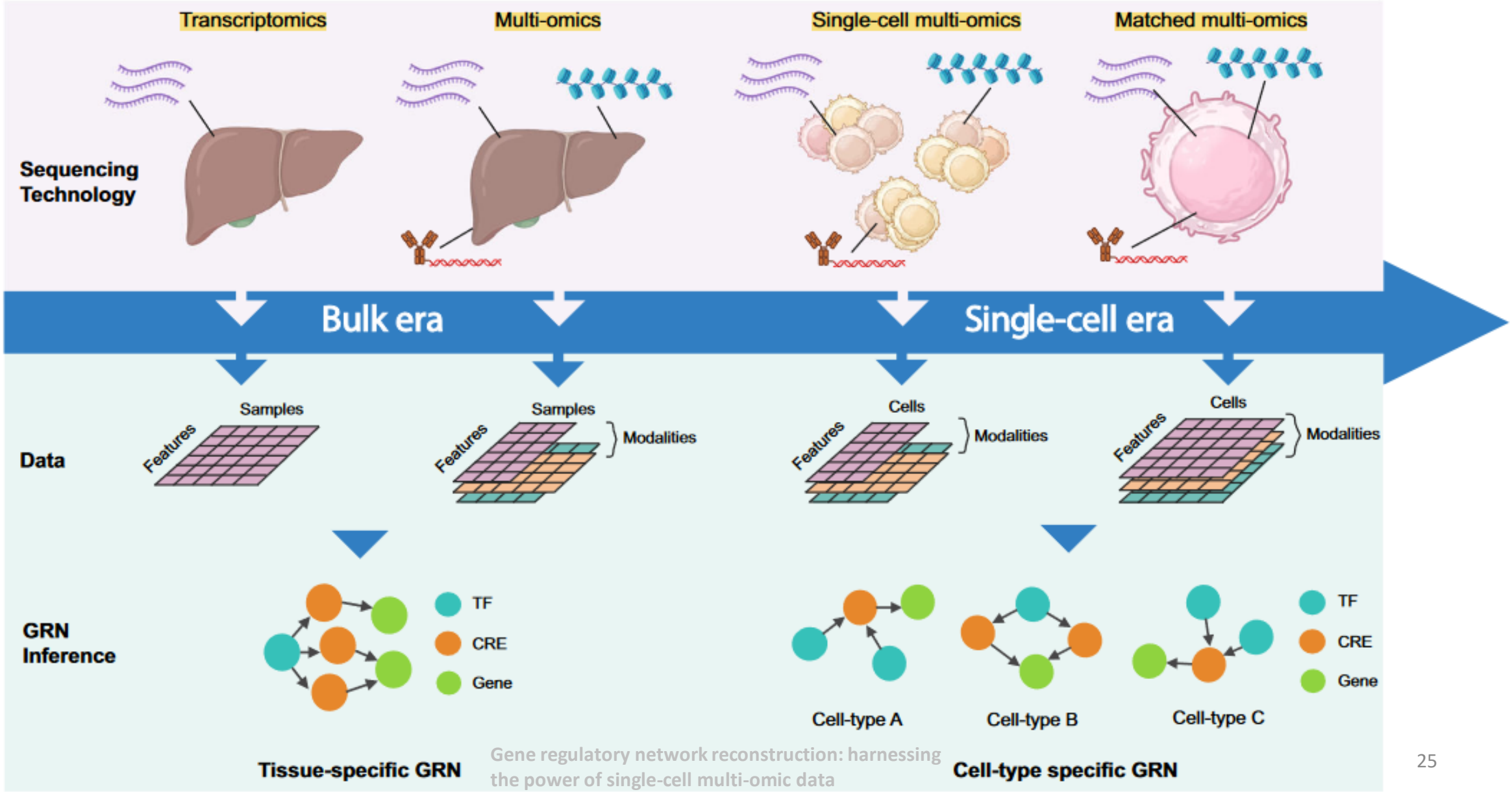
- **Dropout Events**: A common problem is dropout events, where certain genes or features are not detected in some cells due to technical limitations or low expression levels

# GRN inference and single-cell technologies

The emergence of single-cell technologies, **overcome** the limitation of bulk omics data, and has been revolutionized the field by:

- allowing the inference of GRNs across different cell types, differentiation trajectories and conditions.
- Enabling to uncover cellular heterogeneity at the single-cell resolution
- Enhances understanding of cellular dynamics and functions

# Evolution of GRN inference and sequencing technologies




Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data

# Single-cell technologies

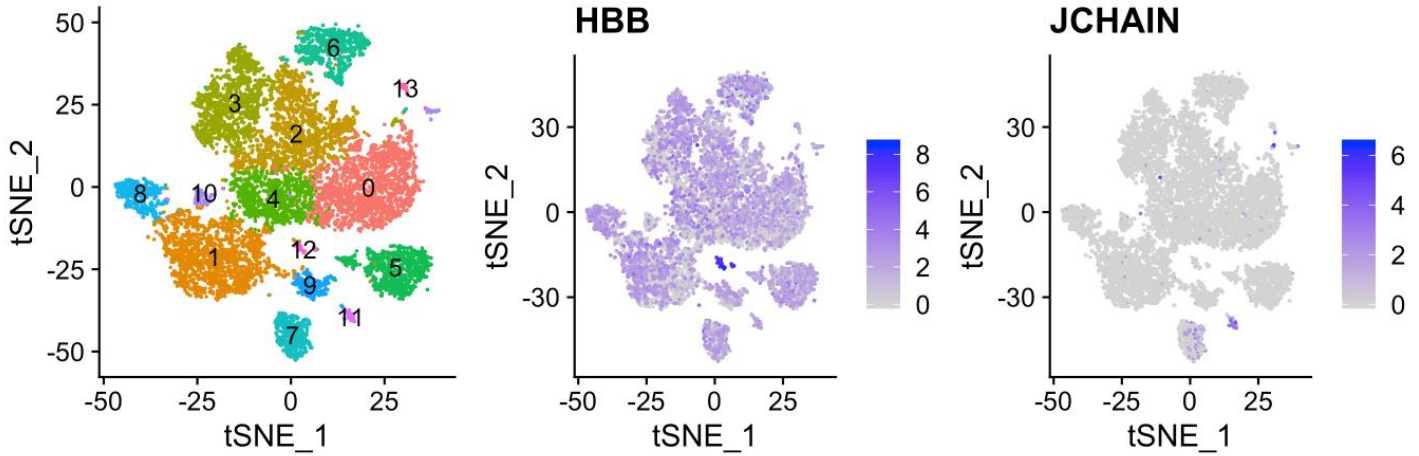
singlecell RNA-seq (scRNA-seq)

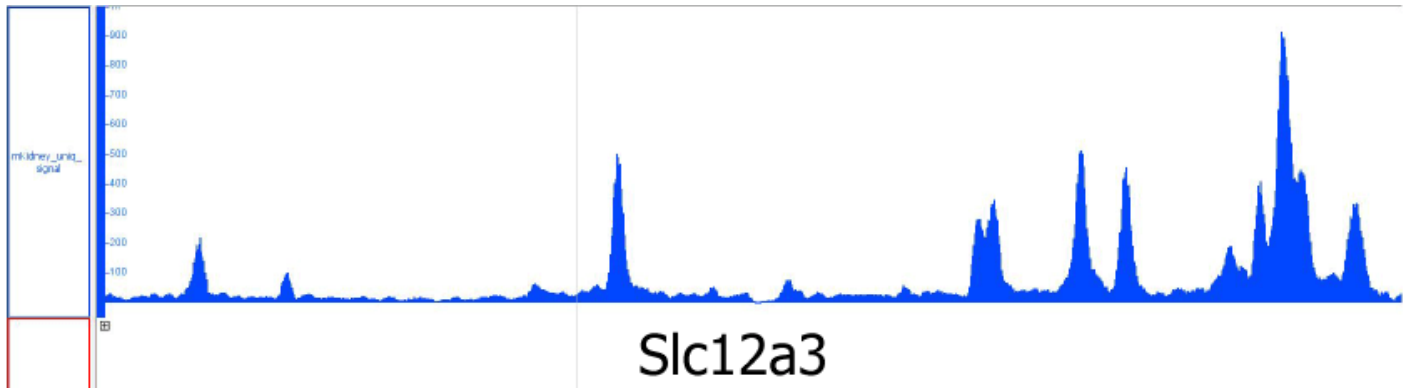
single-cell ATAC-seq (scATAC-seq)

single-cell Hi-C (scHi-C)

single-cell ChIP-seq (scChIP-seq)

single-cell DNA methylation




https://www.activemotif.com/catalog/1299/single-cell-atac-seq-services?__cf_chl_rt_tk=UwQgRFSi.vT.hMv6igwKmWWY3RVamH7minS12TLY06A-1734890453-1.0.1.1-PpD_PCtN2r9XqoCueZ24etaB6fv0zy8DFileELlMbt4

https://www.singlecellcourse.org/single-cell-rna-seq-analysis-using-seurat.html

# GRN inference methods

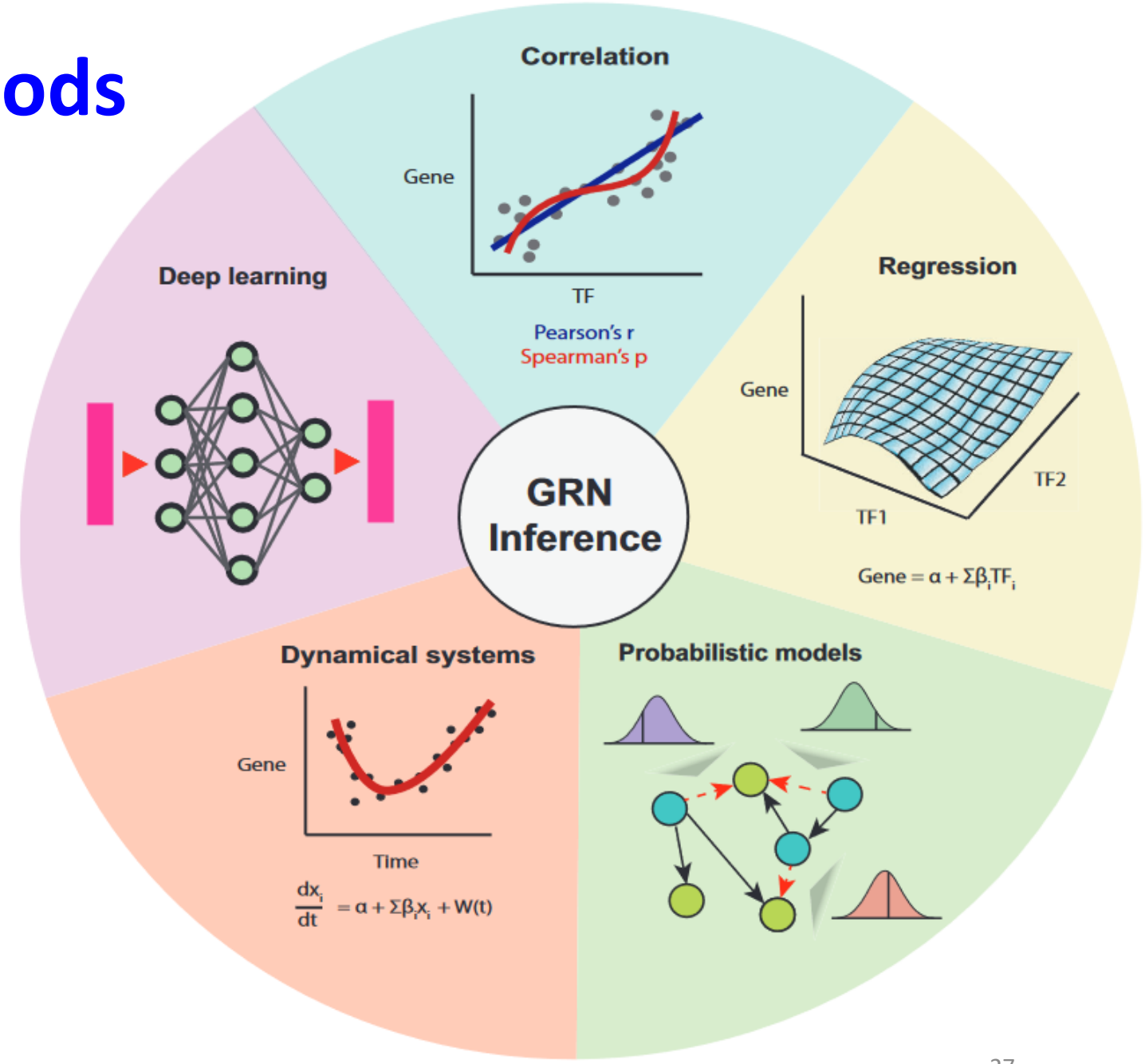
**Correlation-based methods** seek to identify pairs of variables (i.e., TF expression, gene expression or CRE accessibility) that vary similarly.

**Regression-based approaches** model the gene expression based on multiple predictor variables (i.e., TF expression and/or CRE accessibility).

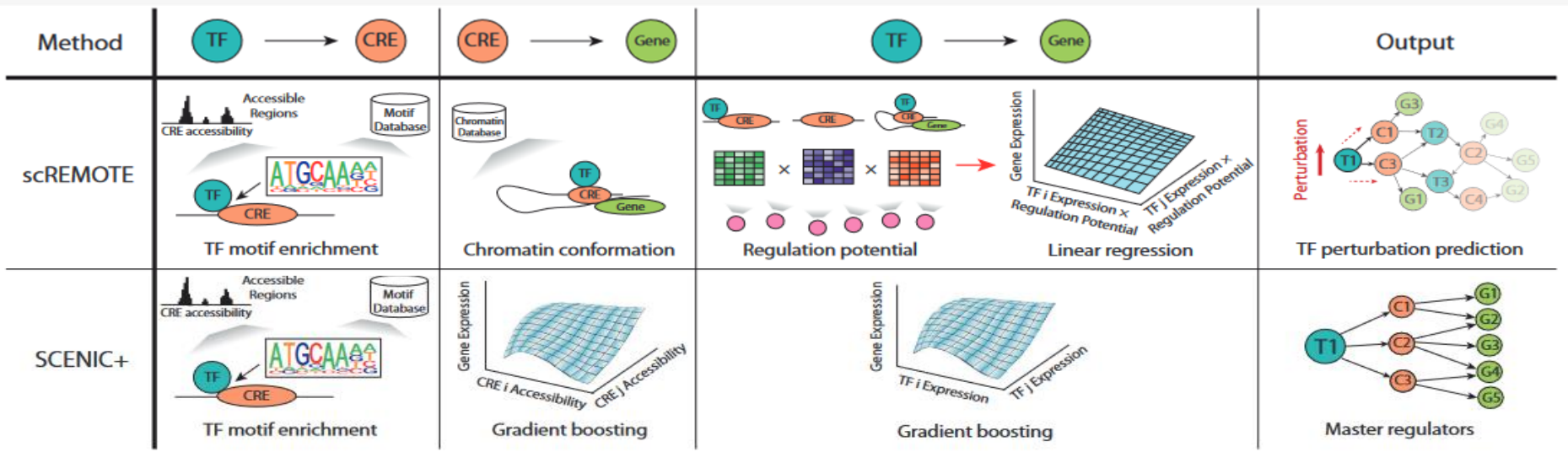
**Probabilistic models** aim to identify the most likely regulators for a gene.

**Dynamical systems-based approaches** model changes in gene expression based on biological factors (e.g., TF expression, cell cycle stage, general stochasticity).

**Deep learning-based approaches** use neural networks to infer complex relationships between TFs, CREs, genes and cells.



# Regression-based methods



Brief Communication | Published: 09 October 2017

# SCENIC: single-cell regulatory network inference and clustering

Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters & Stein Aerts

*Nature Methods* **14**, 1083–1086 (2017) | Cite this article

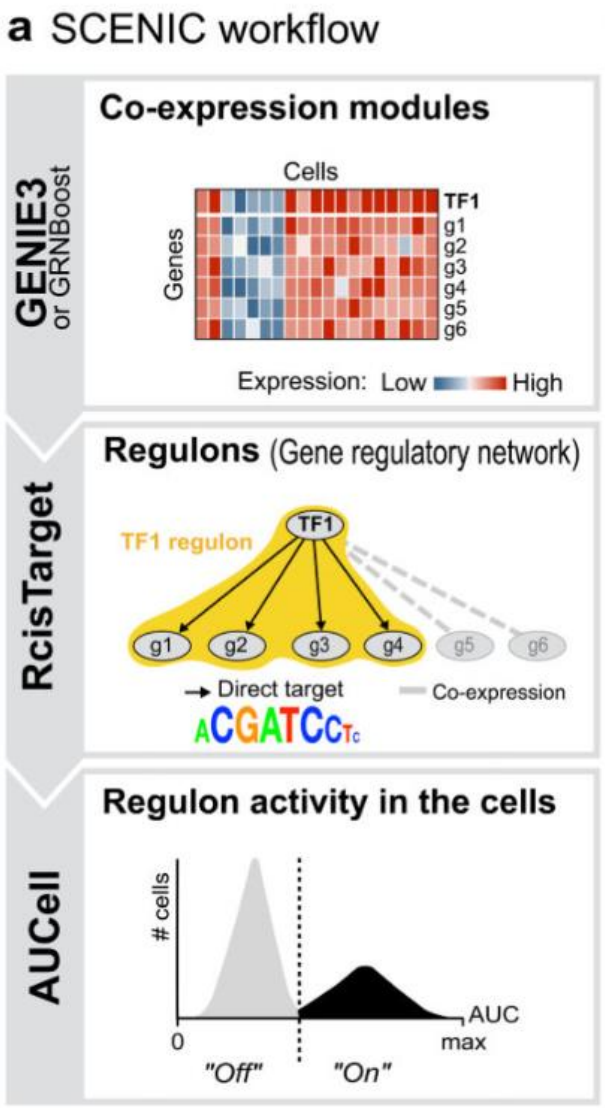
# SCENIC: Single-Cell rEgulatory Network Inference and Clustering

- Reconstruction of gene regulatory networks

And

- Identification of cell states.

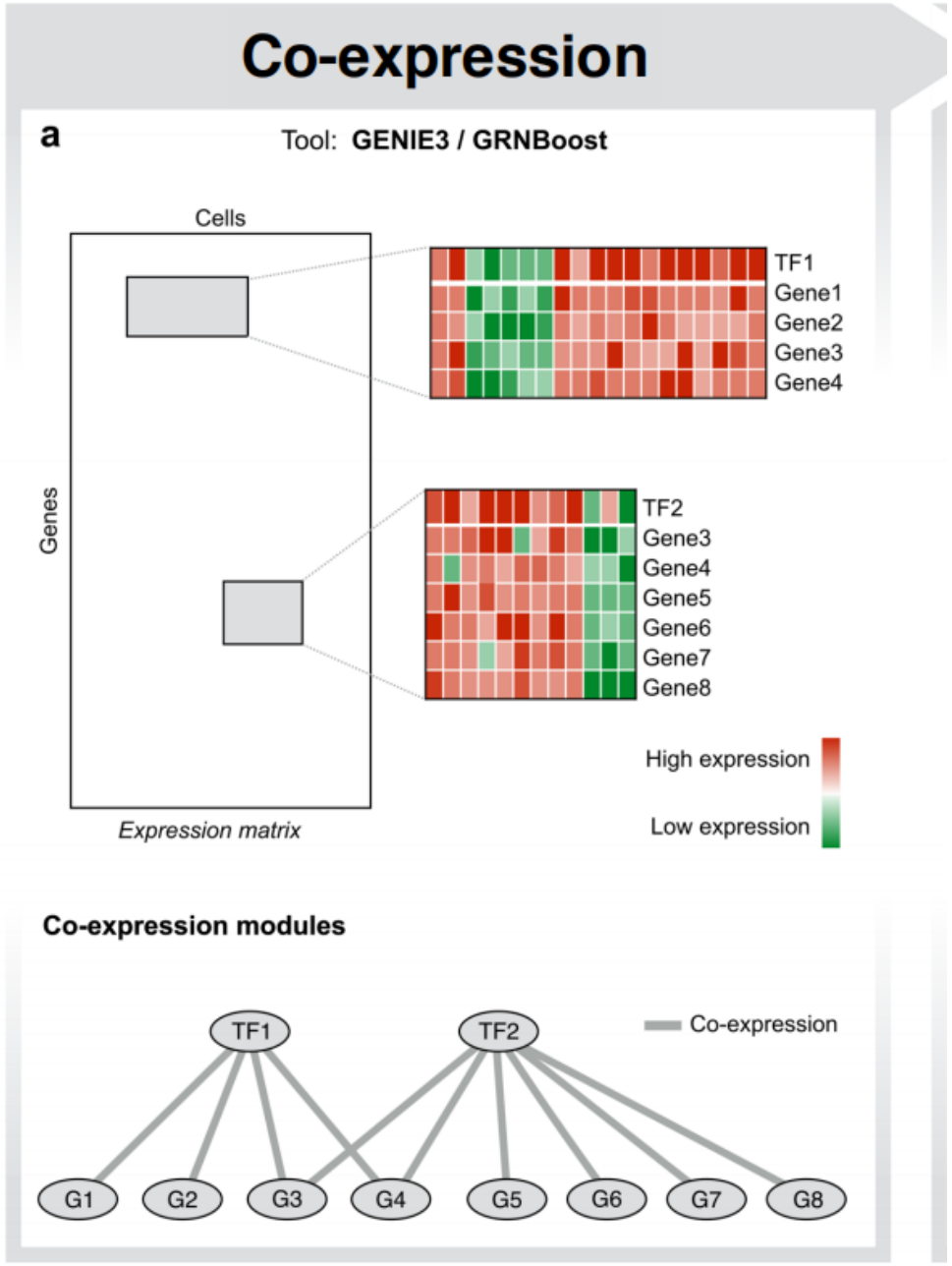
SCENIC is a three-step workflow based on three new R/bioconductor packages




SCENIC: single-cell regulatory network inference and clustering

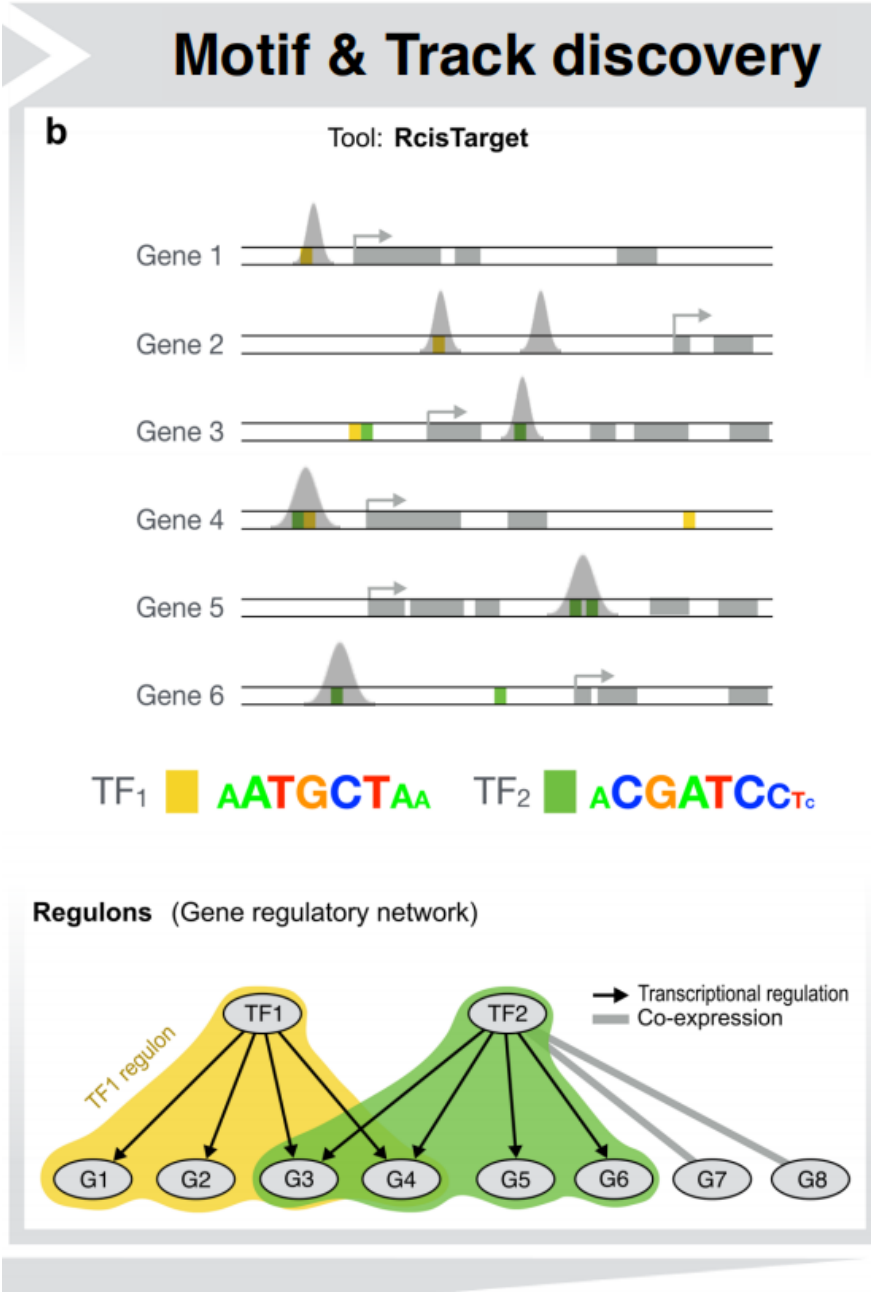
# SCENIC workflow

**(a) Inferring Co-expression modules** between transcription factors and candidate target genes using GENIE3 or GRNBoost



## b) identify putative direct-binding targets

- each co-expression module is analyzed for cis-regulatory motif analyses using RcisTarget
- RcisTarget identifies those modules only with significant motif enrichment of the correct upstream regulator are retained, and pruned to remove indirect target genes without motif support.
- RcisTarget creates regulons with only direct targets.

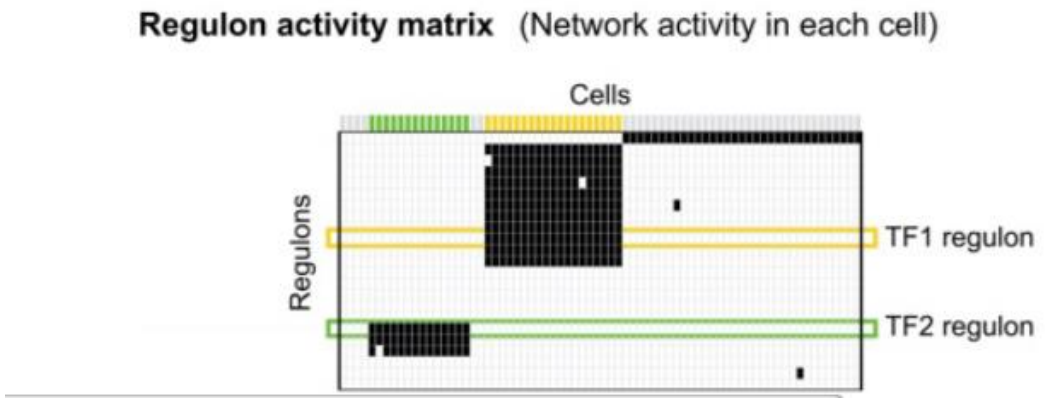


## c) scores Regulon activities in each cell

AUCell uses the “Area Under the Curve” (AUC) to calculate whether a critical subset of the input gene set is **enriched** within the expressed genes for each cell.

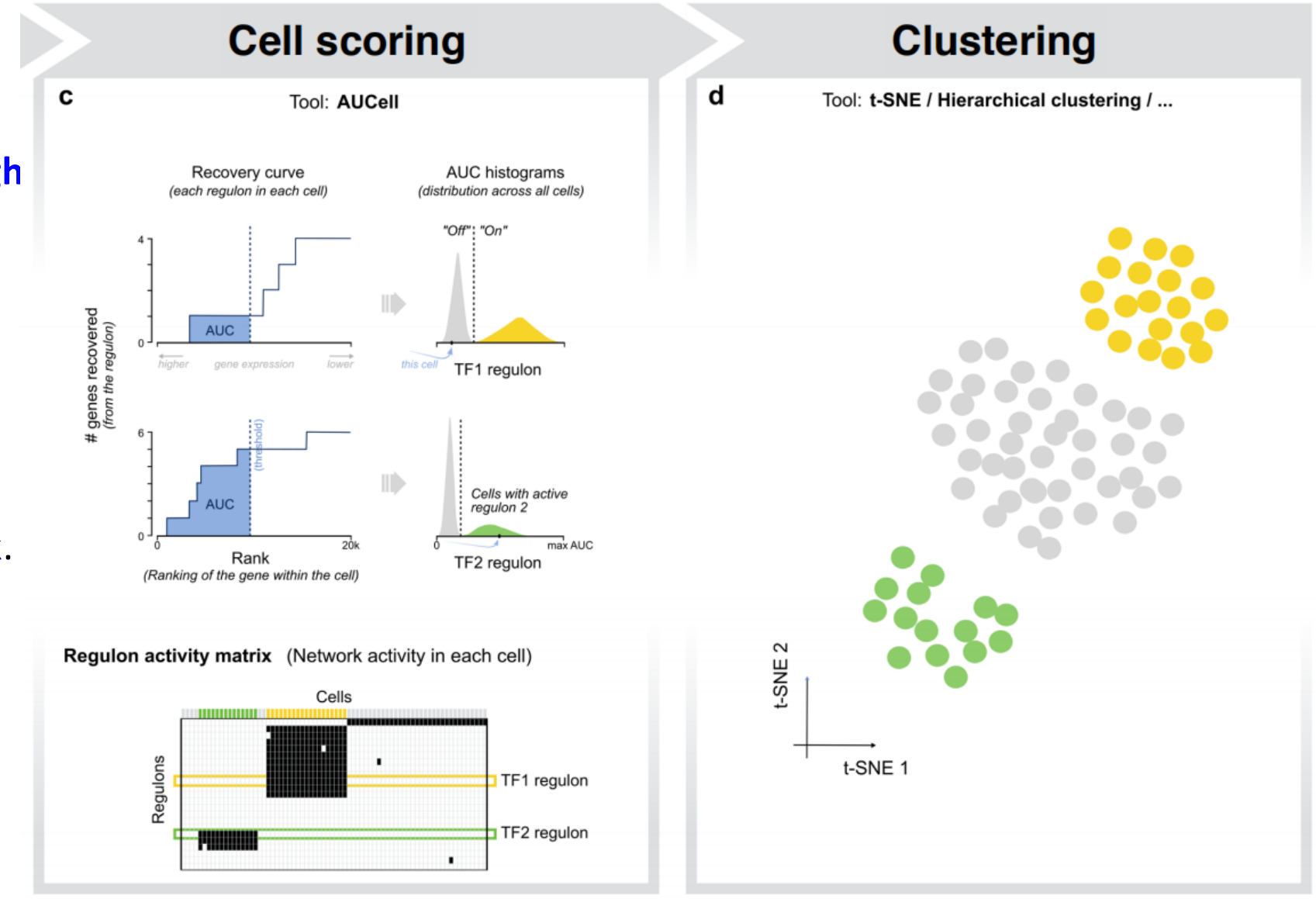
AUCell score: measure how active a regulon is in a cell

**AUCell Identifying cells with active gene-sets**



The relative scores of each regulon across the cells allow identifying **which cells have a significantly high sub-network activity.**

performing a clustering on this matrix allows identifying cell types and states based on the shared activity of a regulatory subnetwork.

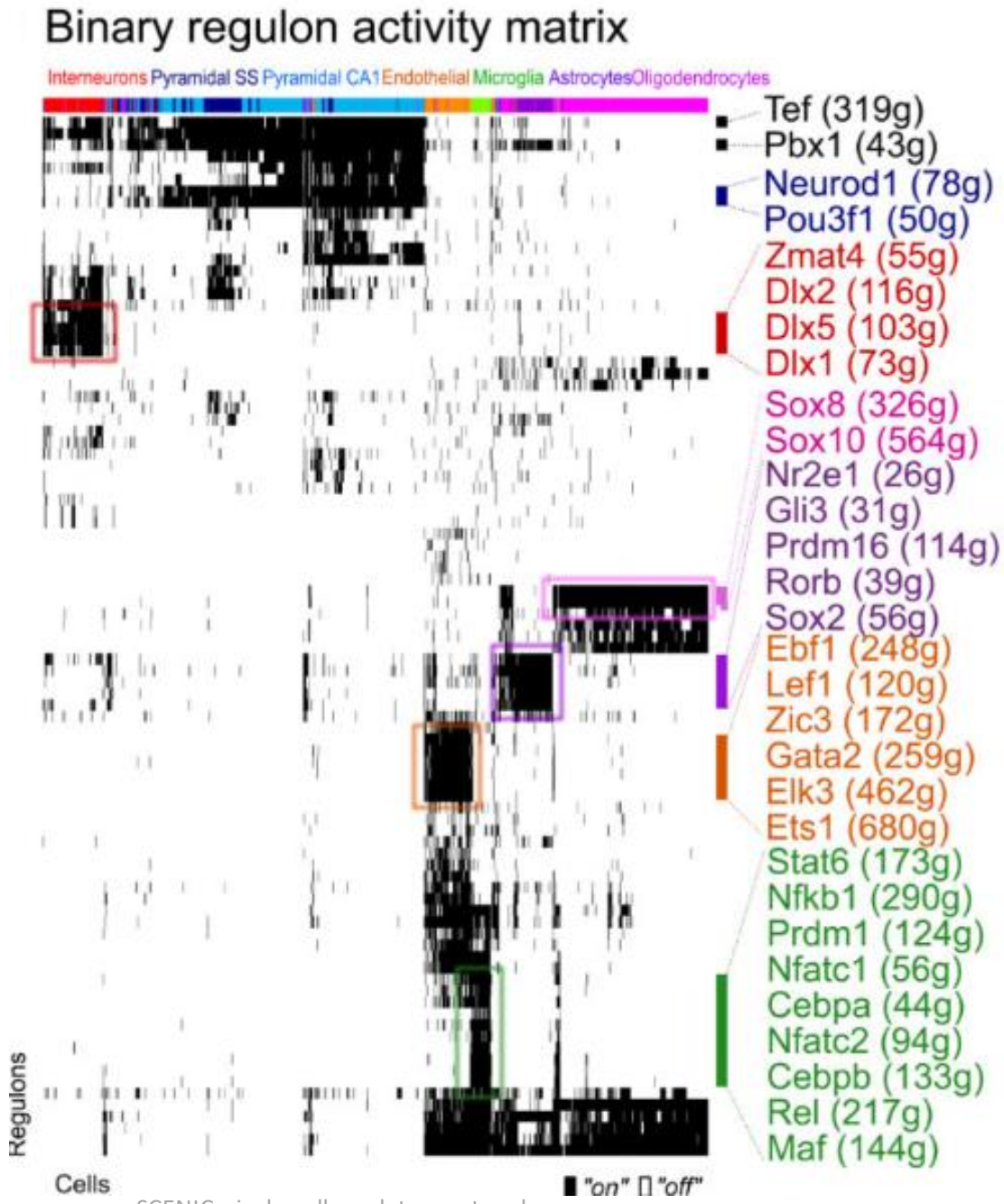



SCENIC: single-cell regulatory network inference and clustering

# Top regulons on the Mouse brain

Applied it to a scRNA-seq data set with well known cell types from the adult mouse brain

- Each row represents a different transcription factor (TF),
- Each column represents a single cell from the mouse brain data.
- Black Squares: Indicate that a particular regulon (set of genes regulated by a single TF) is active in that cell.
- White Squares: Indicate that the regulon is not active in that cell.
- This means the TF is not influencing gene activity in that particular cell.




SCENIC: single-cell regulatory network inference and clustering

Article | Open access | Published: 13 July 2023

# SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks

Carmen Bravo González-Blas, Seppe De Winter, Gert Hulselmans, Nikolai Hecker, Irina Matetovici, Valerie Christiaens, Suresh Poovathingal, Jasper Wouters, Sara Aibar & Stein Aerts

*Nature Methods* **20**, 1355–1367 (2023) | Cite this article

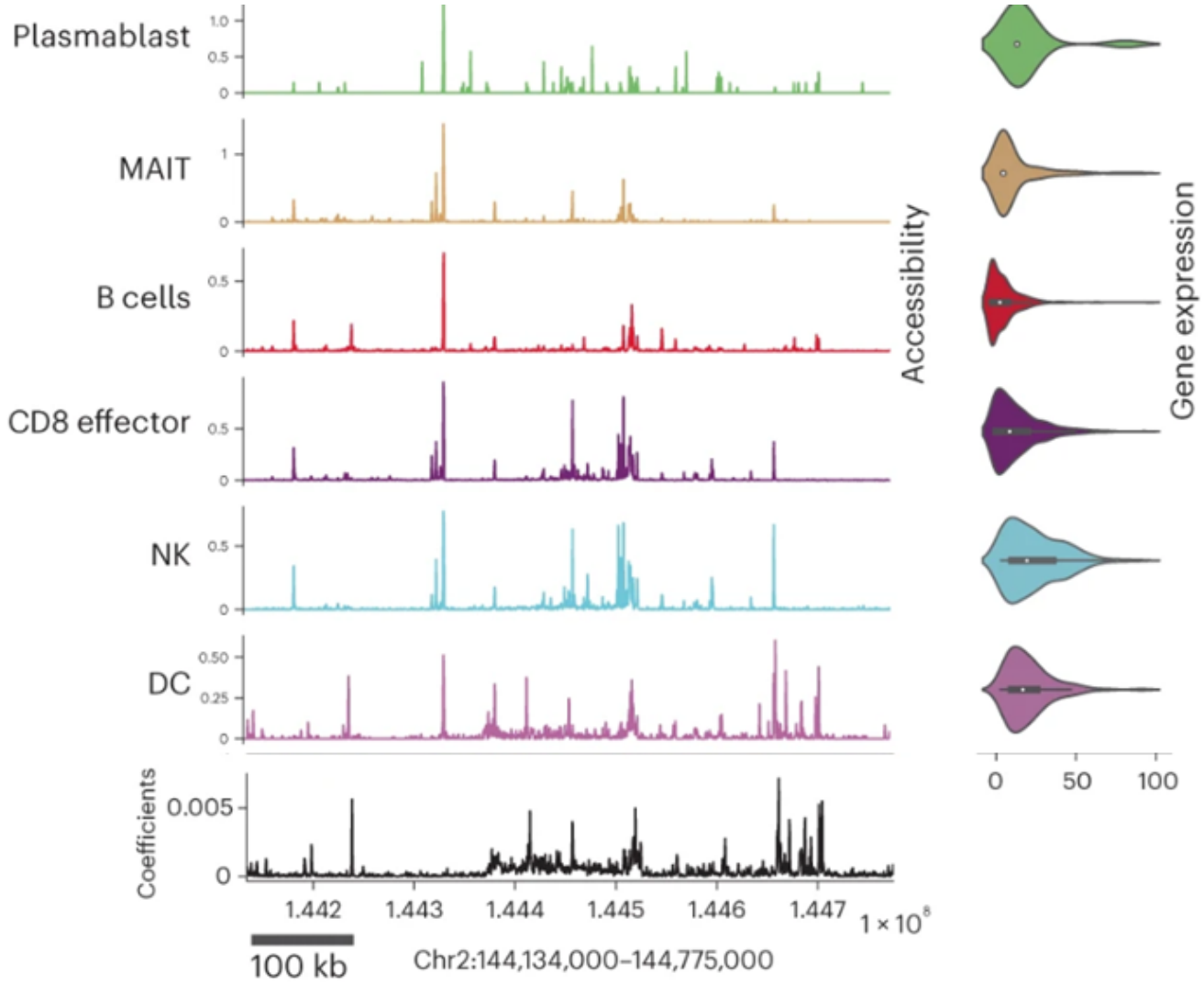
# SCENIC limitations

- SCENIC combines single-cell RNA-sequencing (scRNA-seq) coexpression networks with TF motif discovery
- it **cannot** identify the **exact CRE targeted by the TF**
- it only uses a small proportion of a gene's putative regulatory space

# SCENIC+

With single-cell chromatin-accessibility data, the accuracy of TFBS predictions can be improved substantially.

In fact, genomic regions that are specifically accessible in a cell type often represent enhancers and are enriched for TFBS combinations

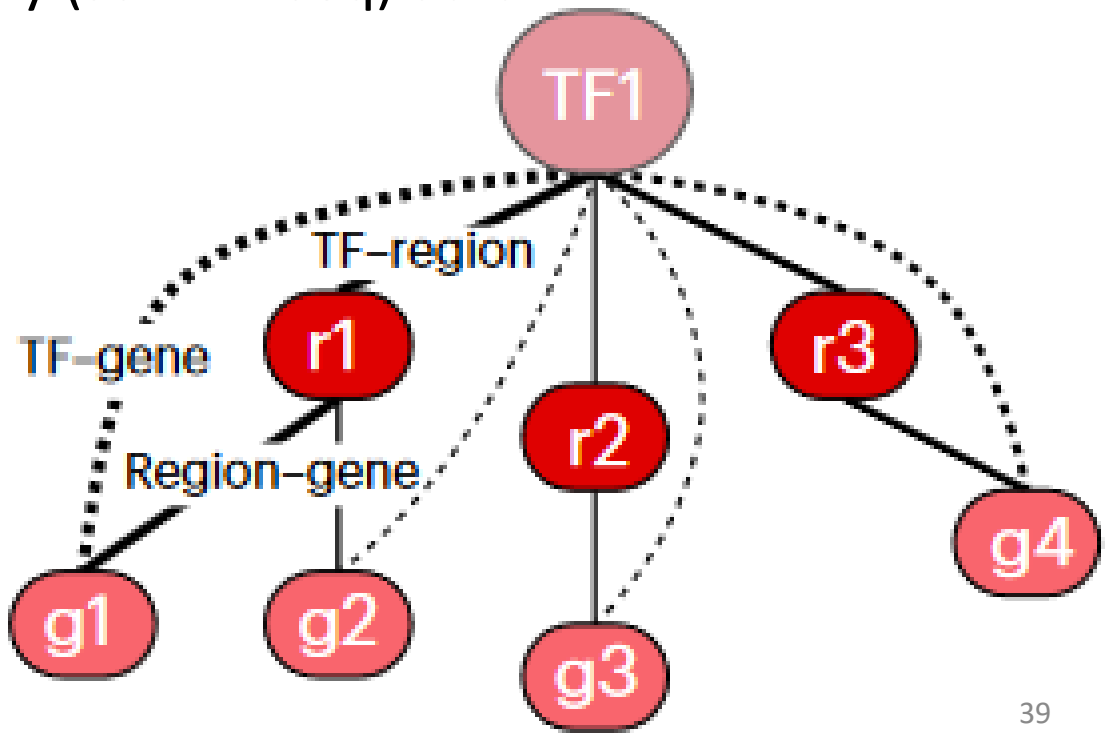



https://www.nature.com/articles/s41588-024-01689-8/figures/1

# What is SCENIC+?

SCENIC+ is a sophisticated method for inferring **enhancer-driven gene regulatory networks (eGRNs)** from

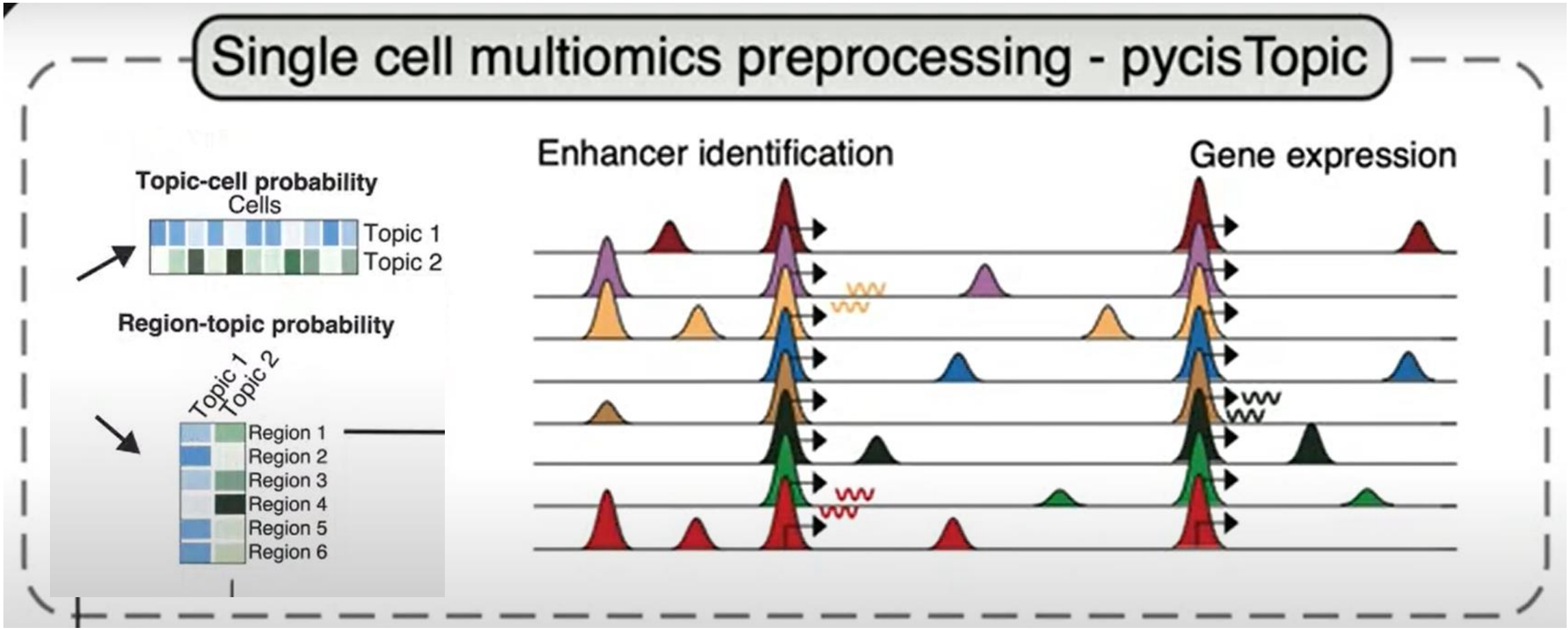
- single-cell gene expression (scRNA-seq)
- and single-cell chromatin accessibility (scATAC-seq) data
- with motif discovery



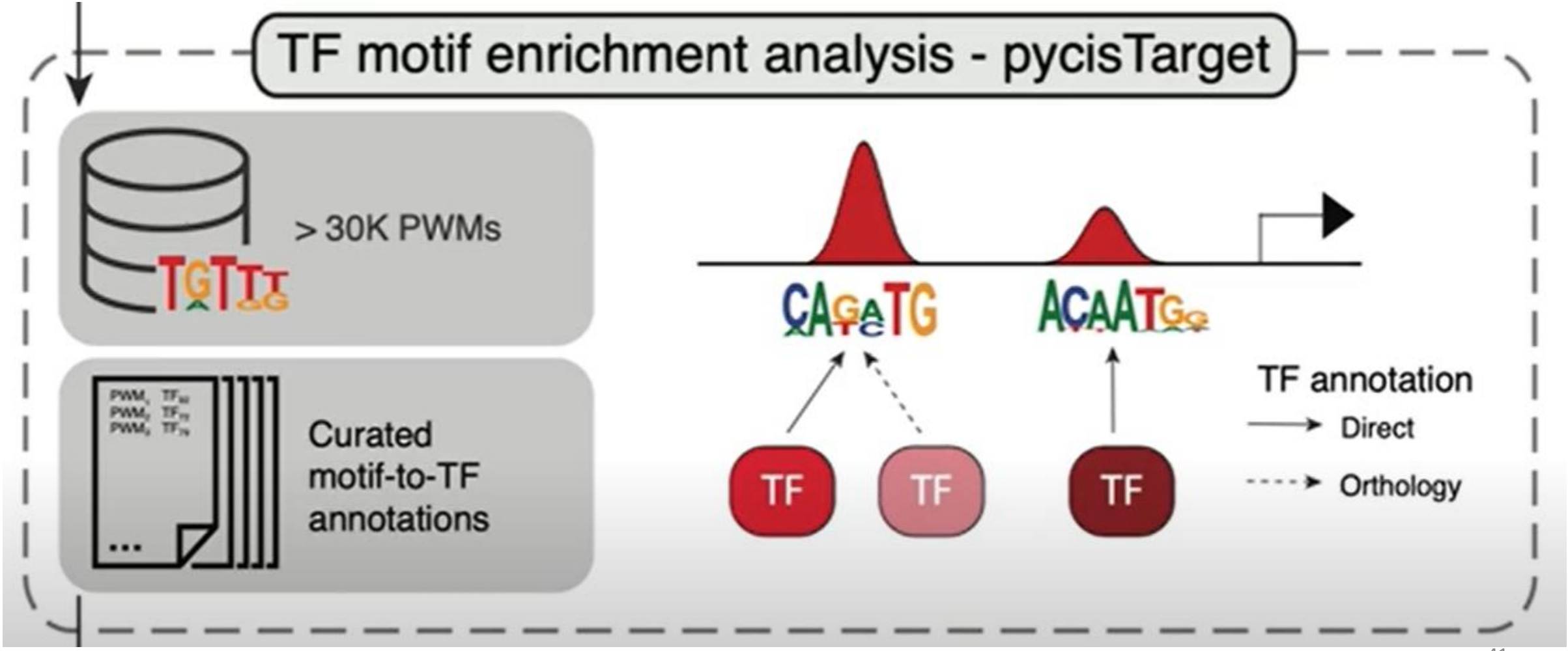



SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks

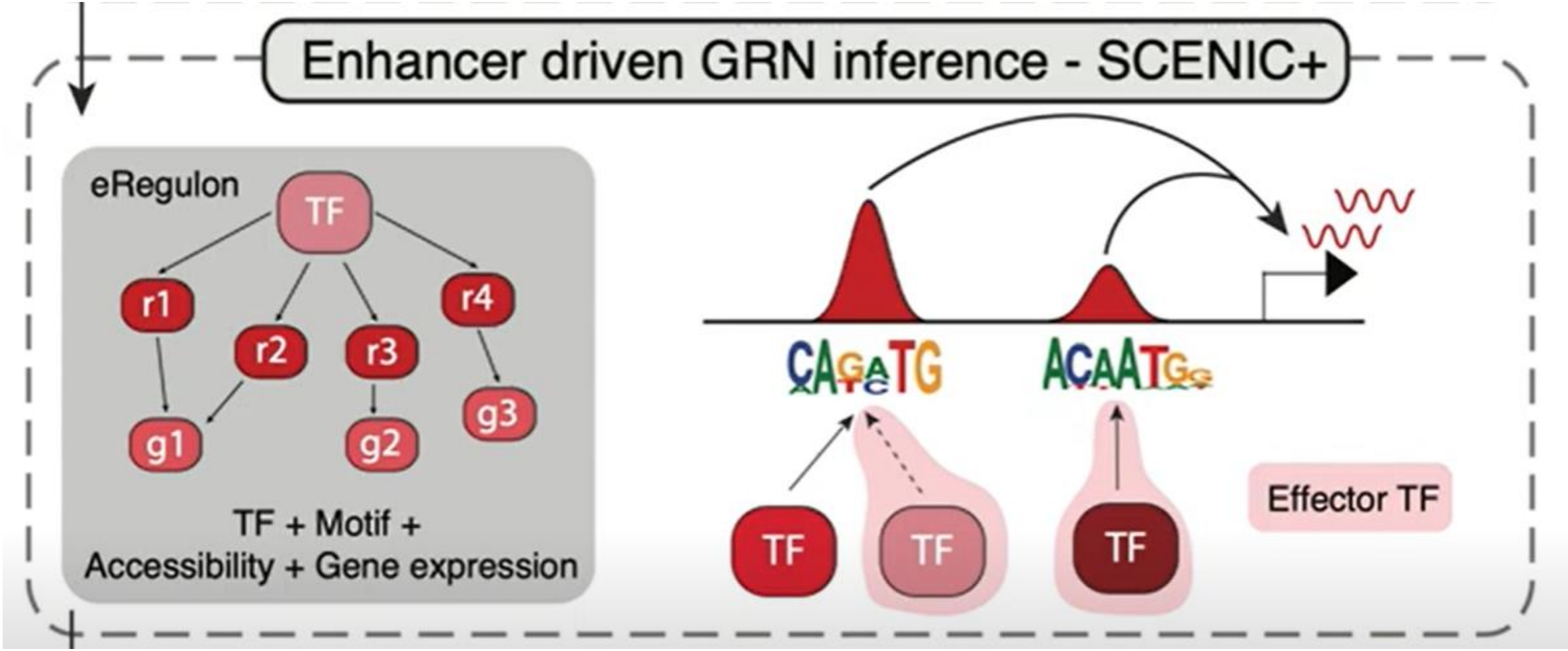
# 1. Identify Enhancer Candidate by analysisng sc-ATAC-seq data using pycisTopic



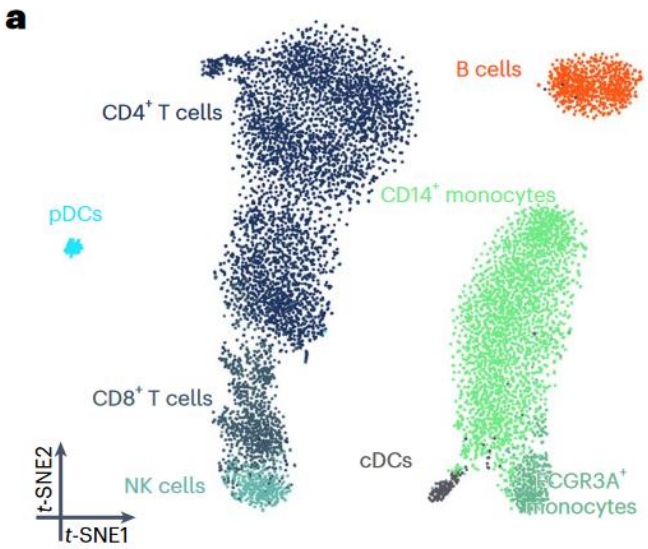
# 2. Enhancer sequence is scan to the presents of candidate motif

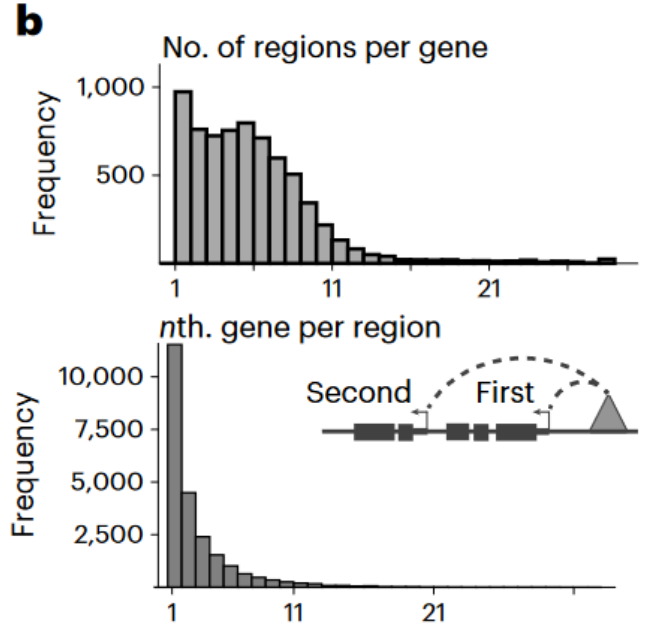


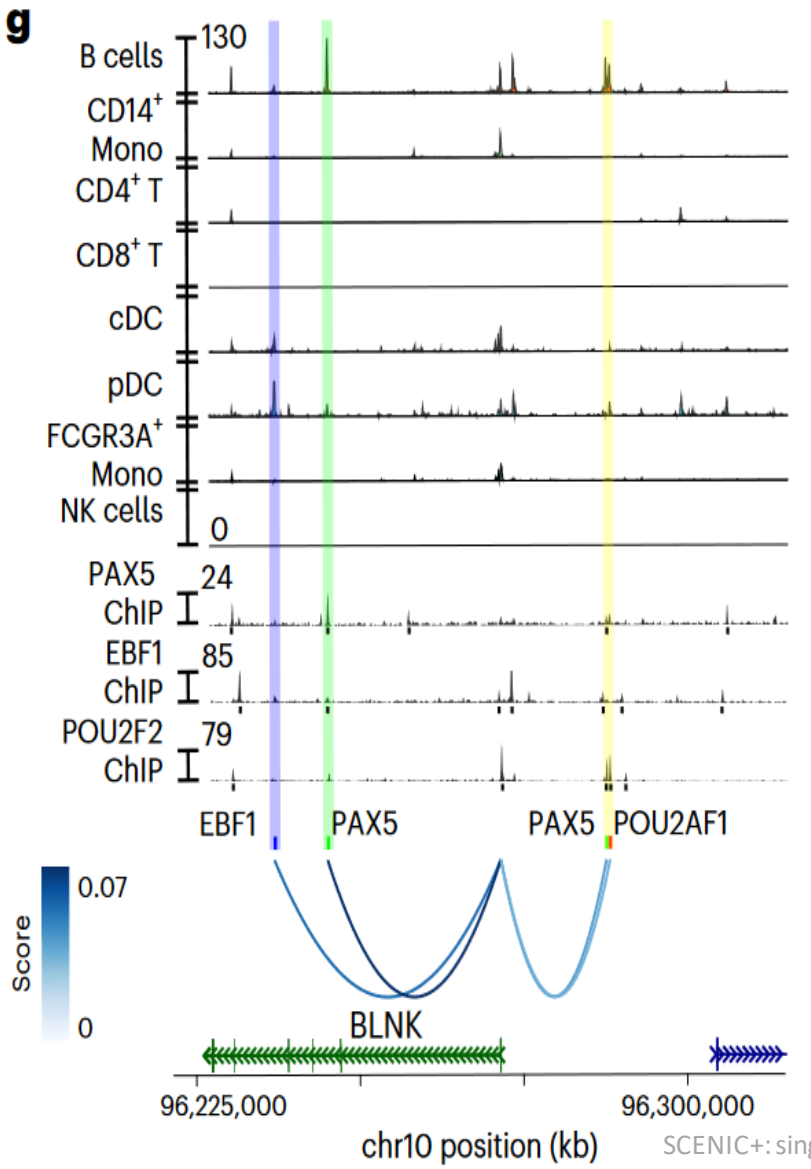
# 3. Combine data from step1,2 to gene expression data to link TF to Regions to Target Genes (eRegulons)

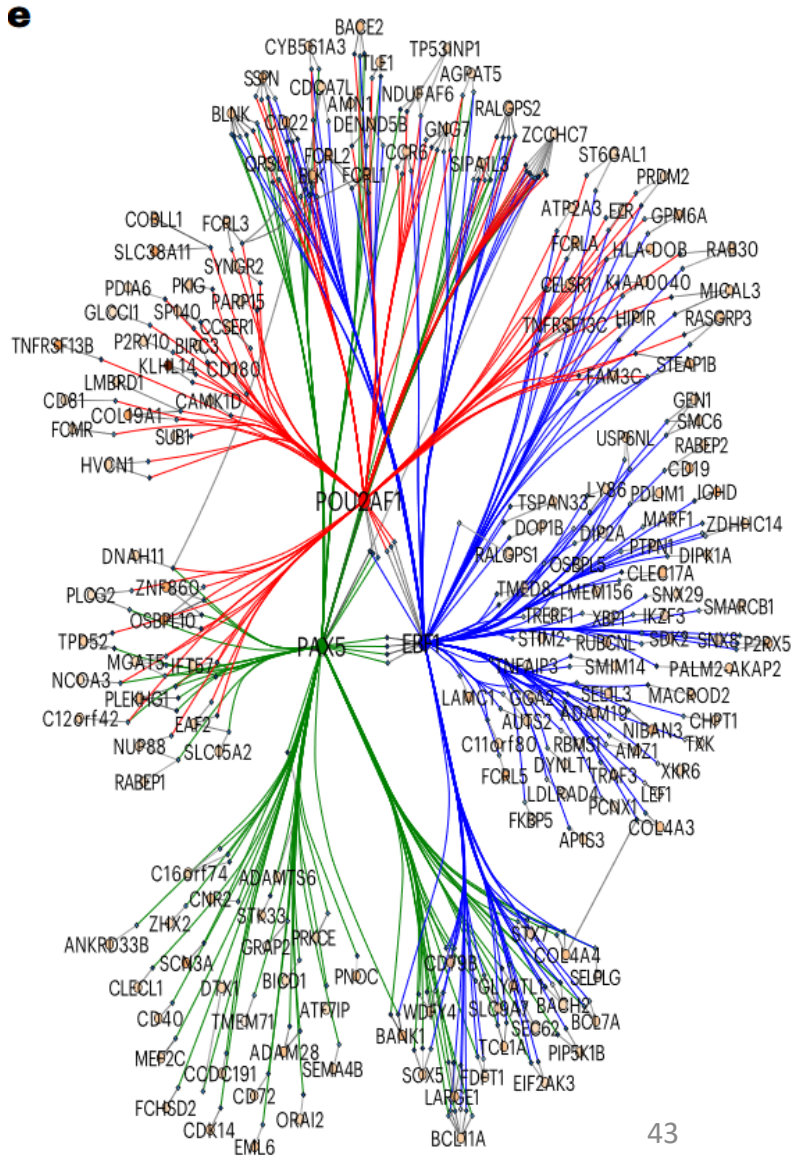


# SCENIC+ analysis on peripheral blood mononuclear cells (PBMC)






SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks

*Published online 15 March 2022*

*NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 1*
*https://doi.org/10.1093/nargab/lqac02*

# scREMOTE: Using multimodal single cell data to predict regulatory gene relationships and to build a computational cell reprogramming model

**Andy Tran [1,2], Pengyi Yang [1,2,3], Jean Y.H. Yang [1,2] and John T. Ormerod[1,*]**

[1]School of Mathematics and Statistics, The University of Sydney, Camperdown NSW 2006, Australia, [2]Charles Perkins Centre, The University of Sydney, Camperdown NSW 2006, Australia and [3]Children's Medical Research Institute, Westmead NSW 2145, Australia

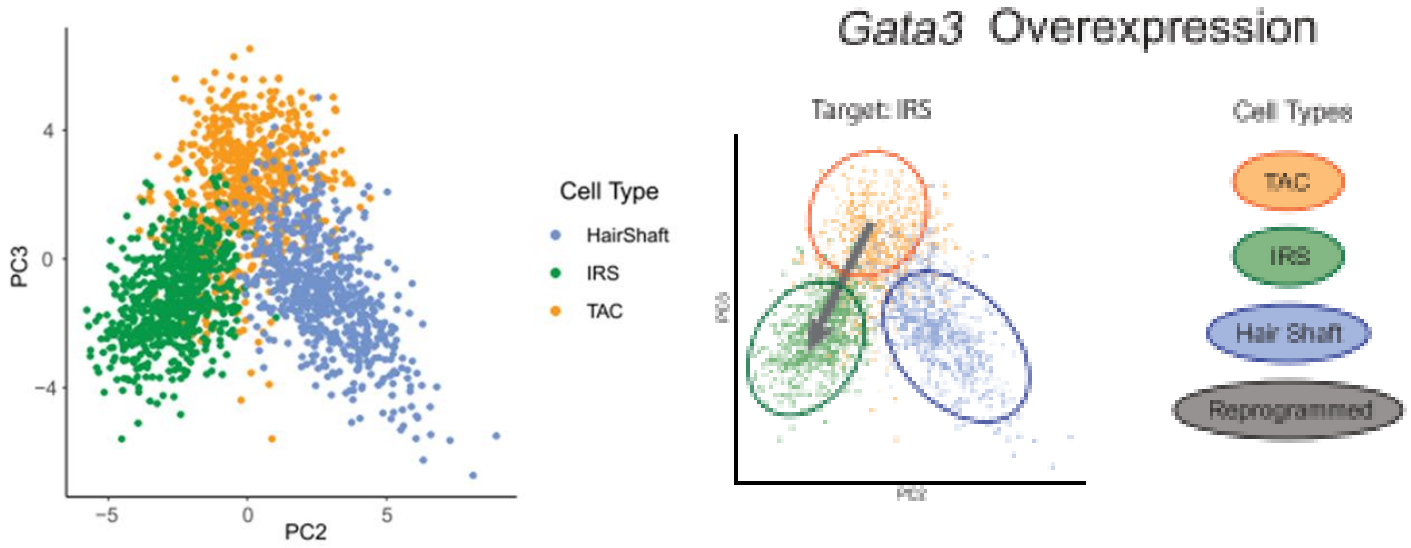
Received June 25, 2021; Revised February 22, 2022; Editorial Decision February 25, 2022; Accepted March 10, 2022

# scREMOTE (single-cell REprogramming MOdel Through cis-regulatory Elements)

scREMOTE: a computational model to infer gene regulation and cell reprogramming

scREMOTE is designed to predict how transcription factor (TF) perturbations can affect gene expression at the single-cell level.

It leverages multimodal single-cell data (scRNA-seq and scATAC-seq data) to predict long-term gene expression changes and model cell reprogramming (hair follicle development )



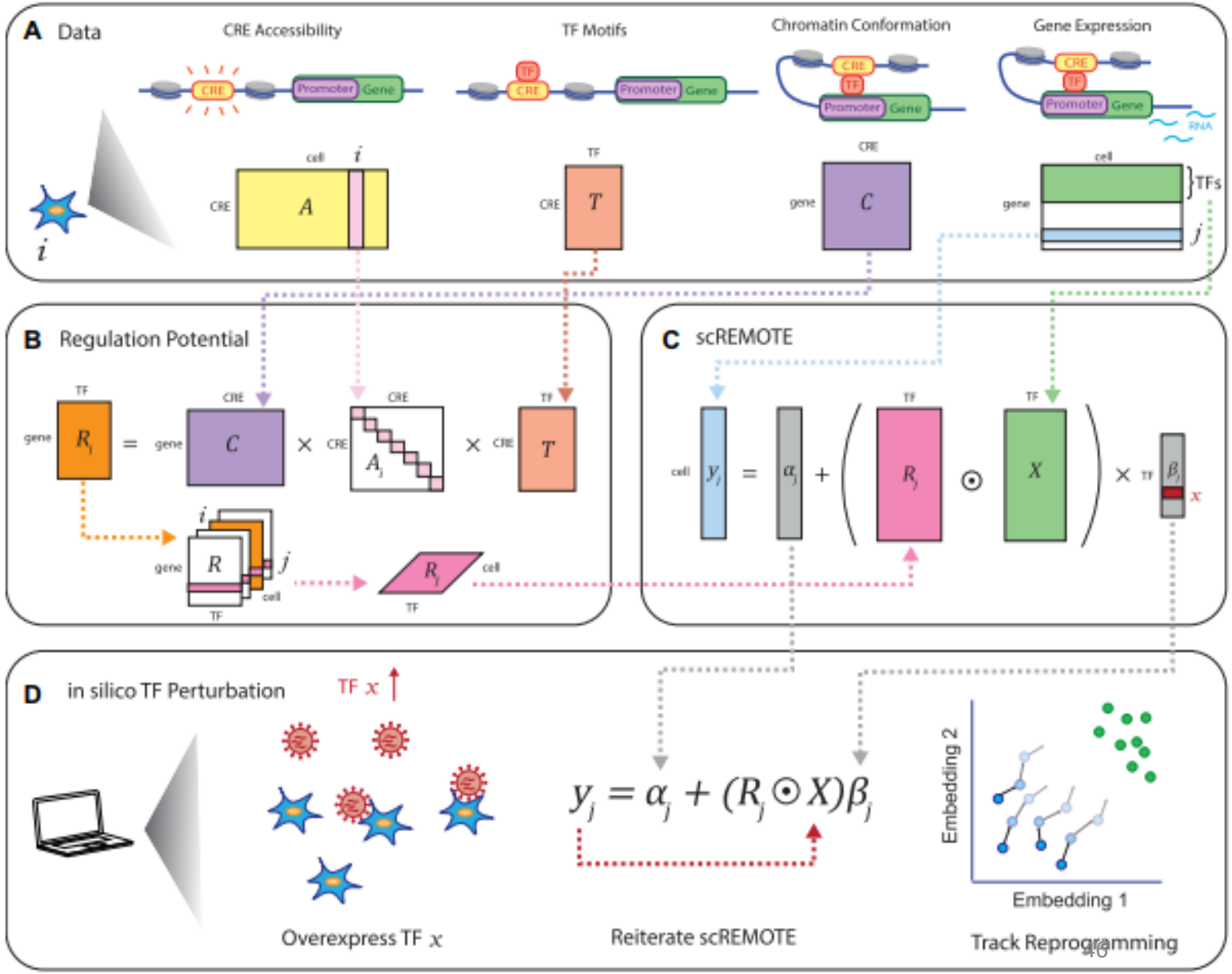
# Schematic of scREMOTE

**(A)** The data inputs to scREMOTE.

**(B)** Calculation of binding potential.

**(C)** Calculation of fitted coefficients.

**(D)** In silico overexpression of TF x



# (A)The data inputs to scREMOTE.

**scREMOTE models four key components of gene regulation as:**

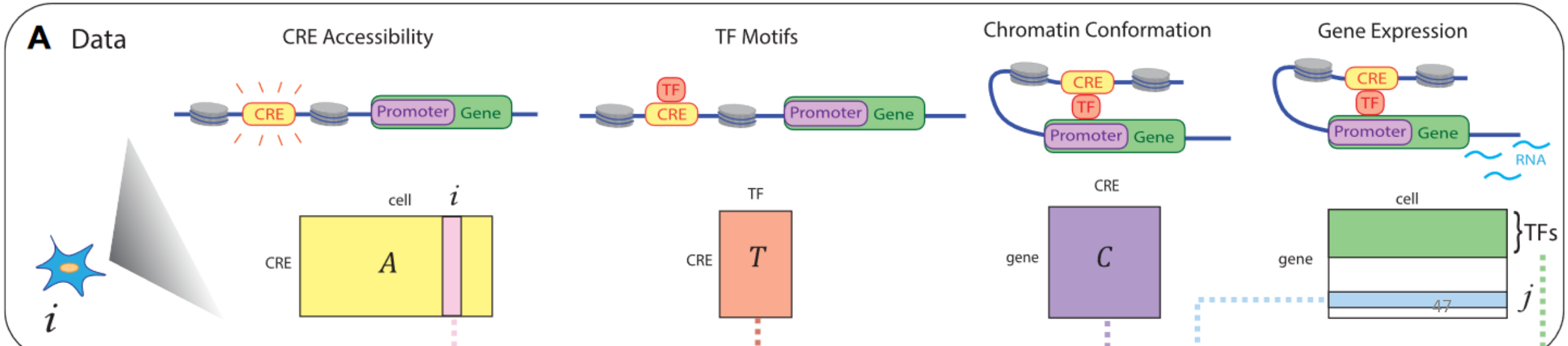
(1) CRE accessibility (a CRE × cell matrix A), where TFs can only bind to regions of the genome that are accessible;

(2) TF motifs (a CRE × TF matrix T), where TFs need a matching motif in order to bind to a CRE;

(3) Chromatin conformation(a gene × CRE matrix C), where CREs need to be able to form a DNA loop with the promoter of the target gene;

(4) Gene expression (a gene × cell matrix E, where the TFs are a subset of the genes), which we expect to vary based on the previous three factors.

we can reasonably assume that the motifs a TF recognises remain the same between cells.

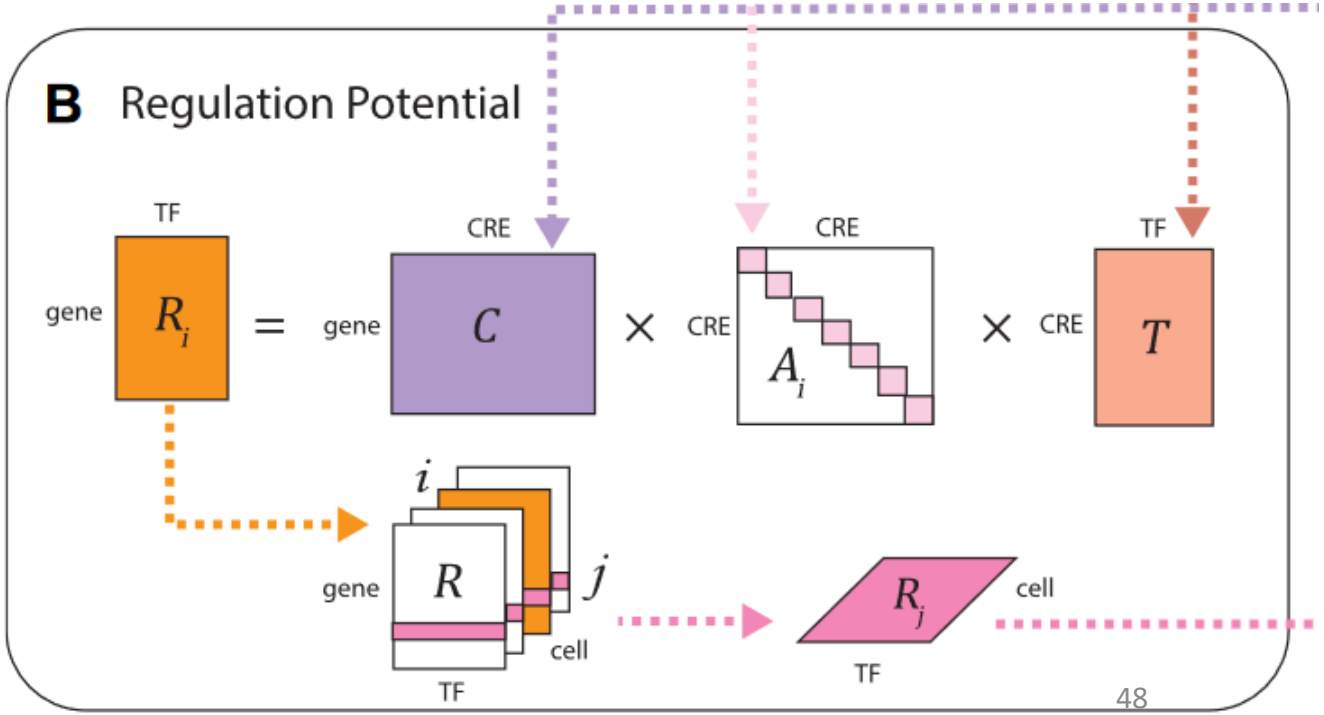


# (B) Calculation of Regulation Potential

The regulatory effect of TFs on genes is calculated by

- C is a matrix representing chromatin interactions between CREs and promoters.
- Ai is a matrix containing chromatin accessibility scores for cell I
- T is a matrix indicating TF motif enrichment in CREs.

That is for a regulatory potential to be positive, the CRE must be
(1) accessible, and
(2) enriched of the TF's motif,
(3) able to form a DNA loop with the gene's promoter.



# (C) Calculation of fitted coefficients.

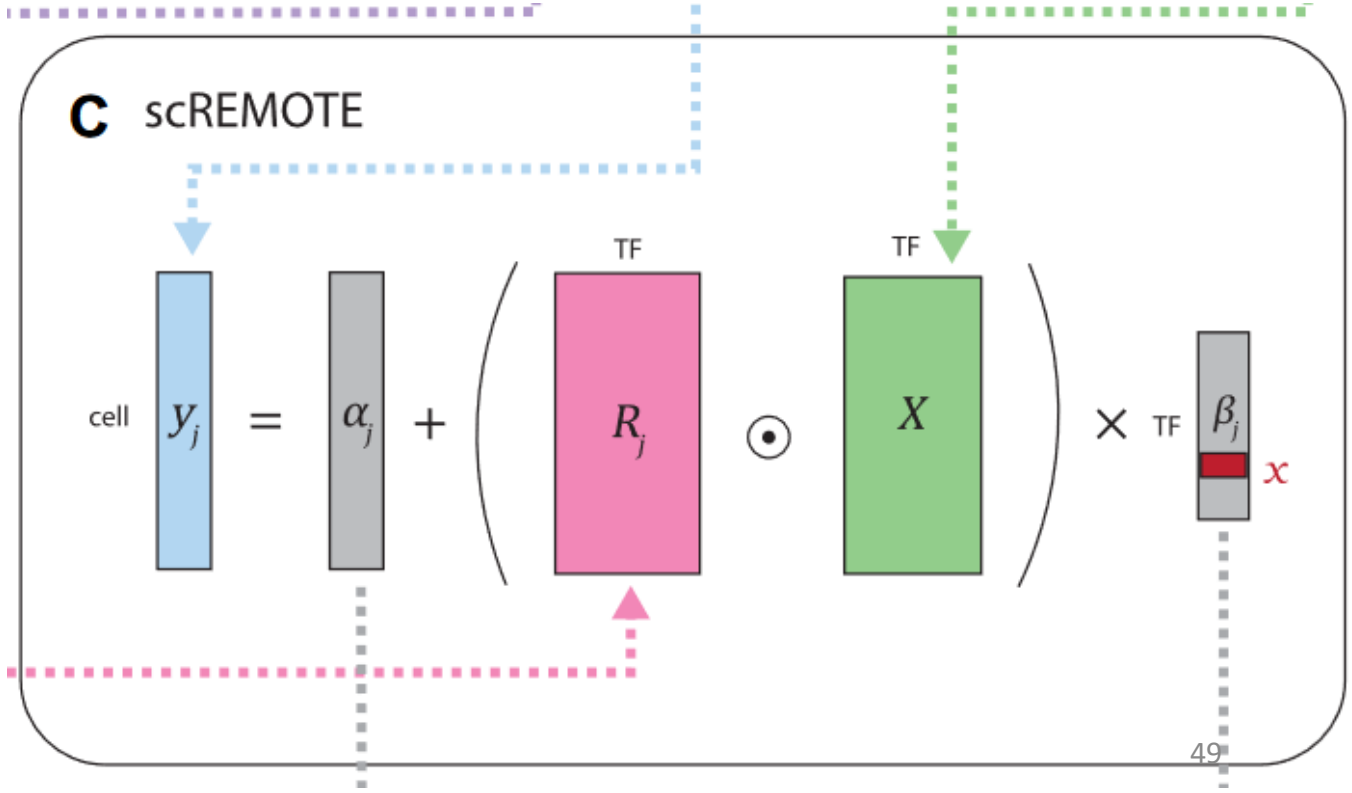
scREMOTE estimates how a cell will respond to a perturbation in TF expression.

fitting a linear regression model with the cell's state, represented by its gene expression, as the response. We incorporate both the gene expression data and regulation potential into the predictor of our model

$$y_j = \alpha_j 1 + (R_j \odot X)\beta_j + \epsilon_j$$

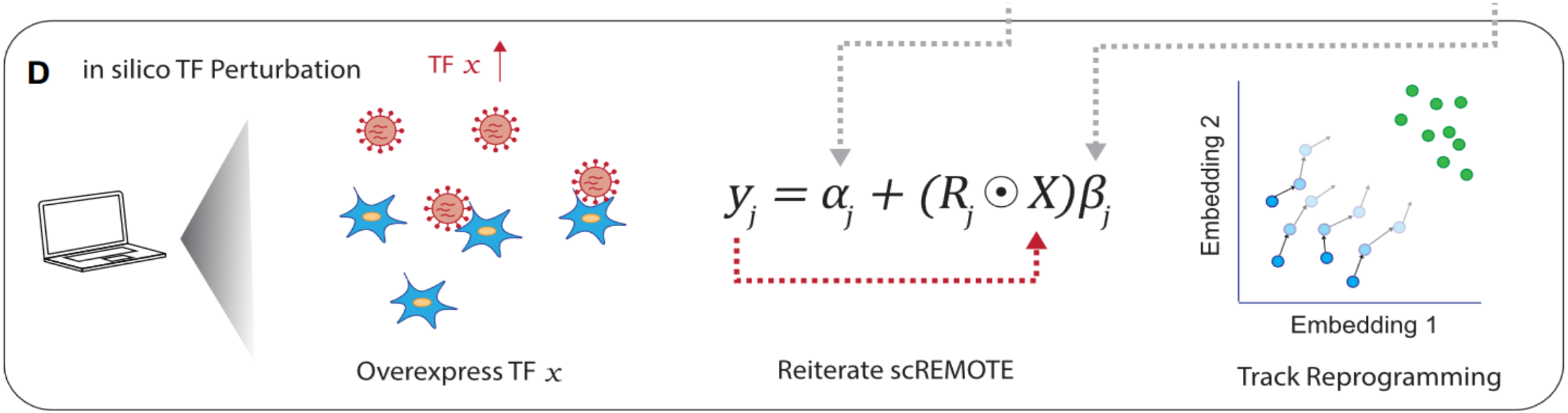
where:

- $y_j$ is the expression of gene $j$,
- $R_j$ is the regulatory potential for gene $j$,
- $X$ is the expression of TFs,
- $\alpha_j$ and $\beta_j$ are regression coefficients,
- $\epsilon_j$ is the residual error,
- $\odot$ represents the element-wise (Hadamard) product.



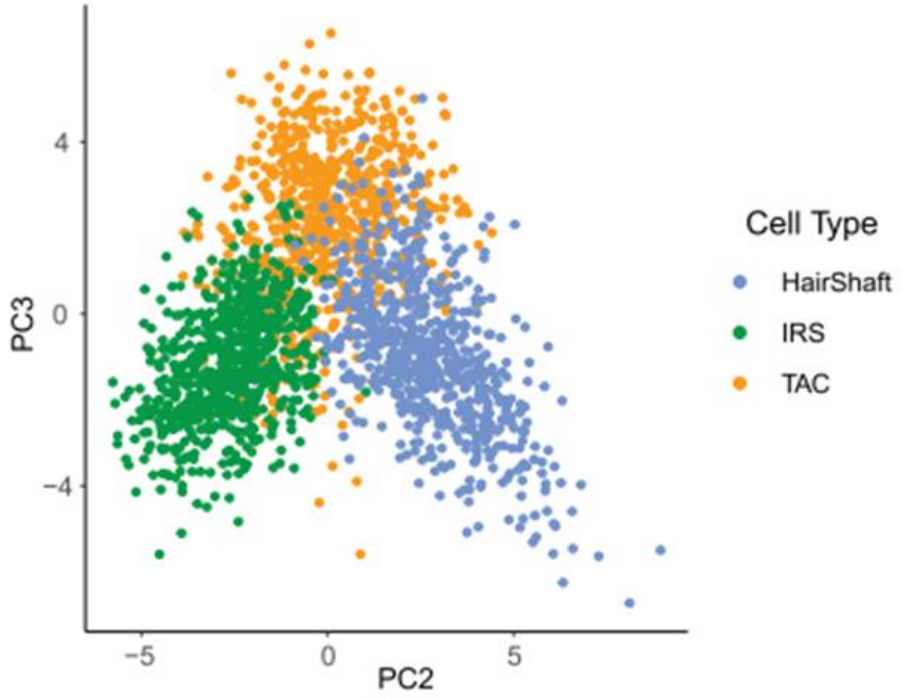
# (D) In silico overexpression of TF x.

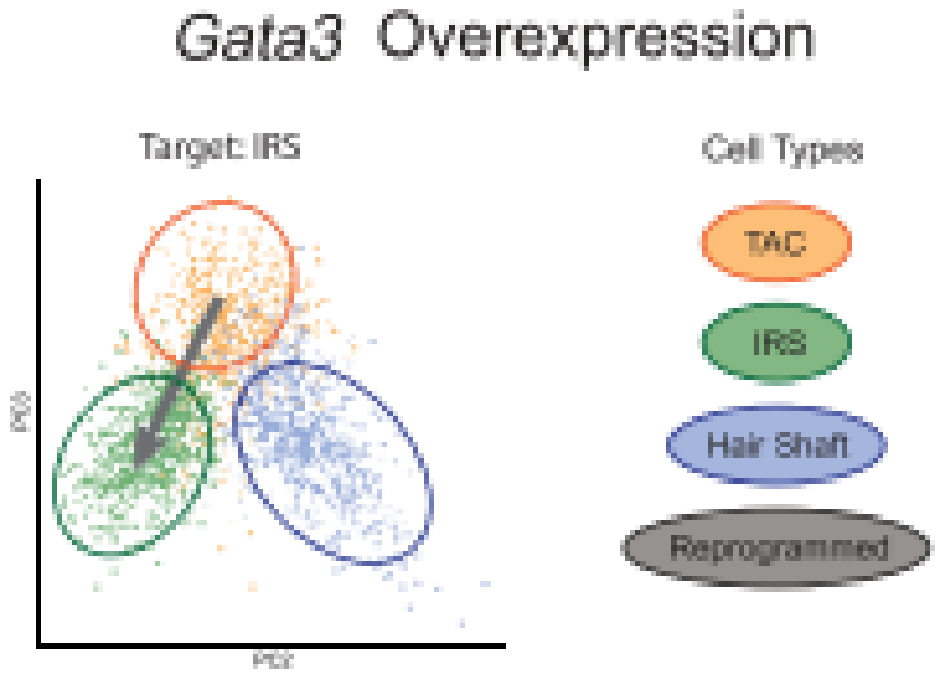
- The final step in scREMOTE is to **perturb a TF's expression** (or a combination of TF expressions
- TF perturbations can be of different types, including: Overexpression, Knockdown or Knockout, Mutations



# Model Validation

- The model was tested using experimental data from mouse **hair follicle development**
- a system where specific TFs like **Gata3** and **Runx1** drive the differentiation of **Transit-Amplifying Cells (TACs)** into either the **Inner Root Sheath (IRS)** or **Hair Shaft** lineages.



# Perturbation of TFs in hair follicle development

**Gata3 overexpression** : scREMOTE and the Coexpression Model predict similar short-term outcomes (early and middle time points), where TACs are pushed towards the IRS fate (green cluster).

**Runx1 overexpression:** both models predict short-term shifts from TACs toward the Hair Shaft fate (blue cluster).

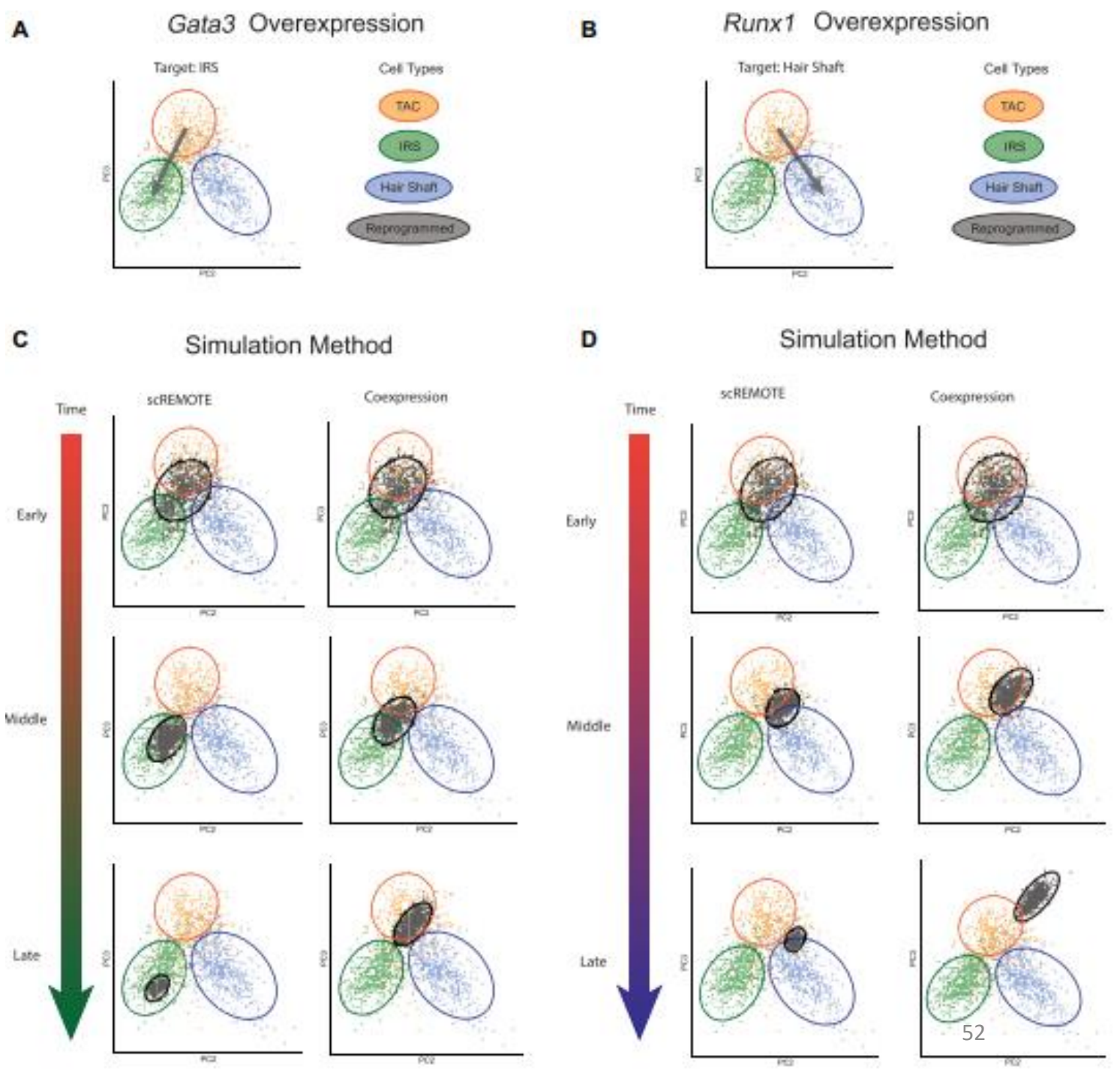
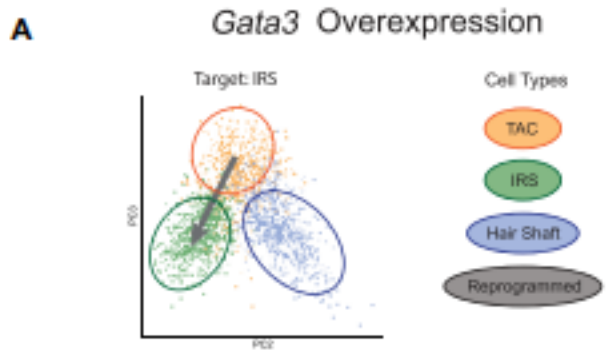
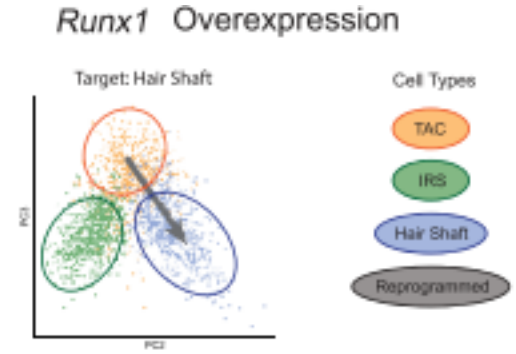
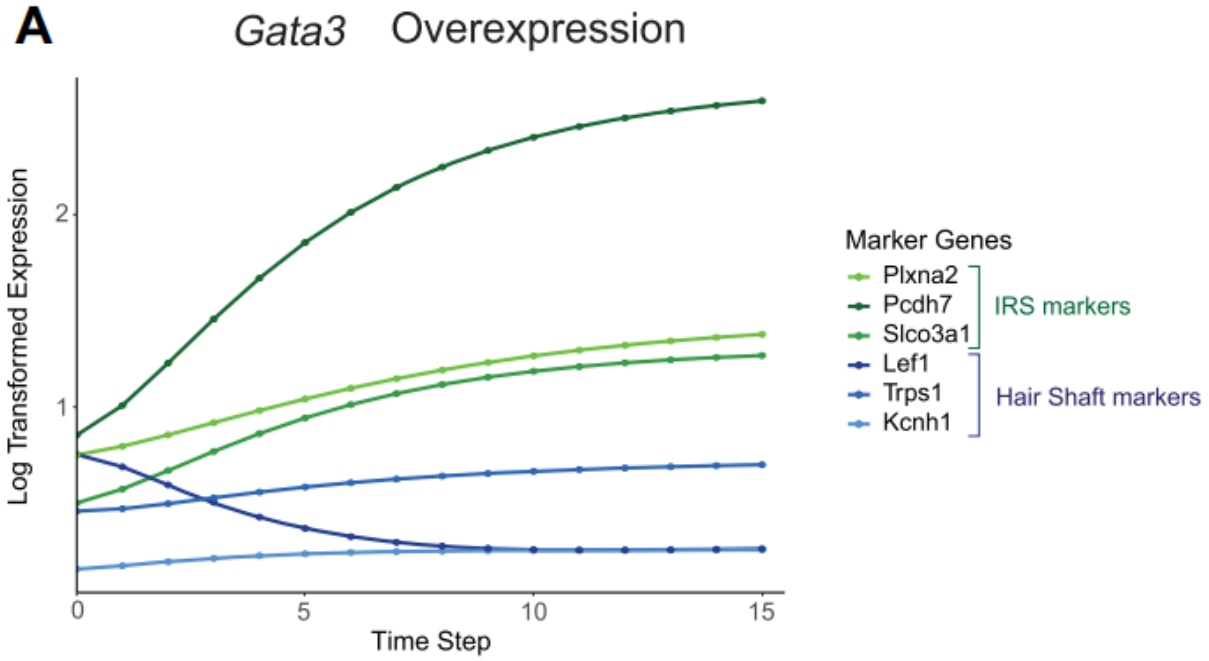
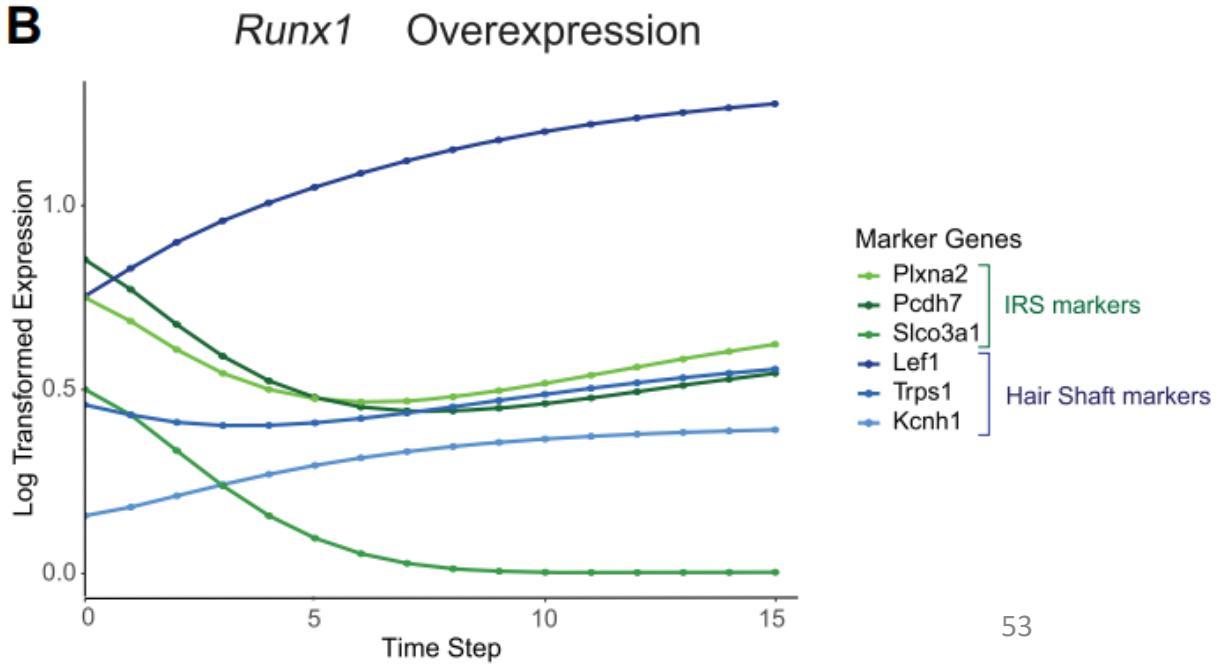


**Figure 4.** Tracking marker genes. (**A)** Marker genes during *Gata3* overexpression. Genes colored in green represent IRS markers and genes colored in blue represent Hair Shaft markers. **(B)** Marker genes during *Runx1* overexpression



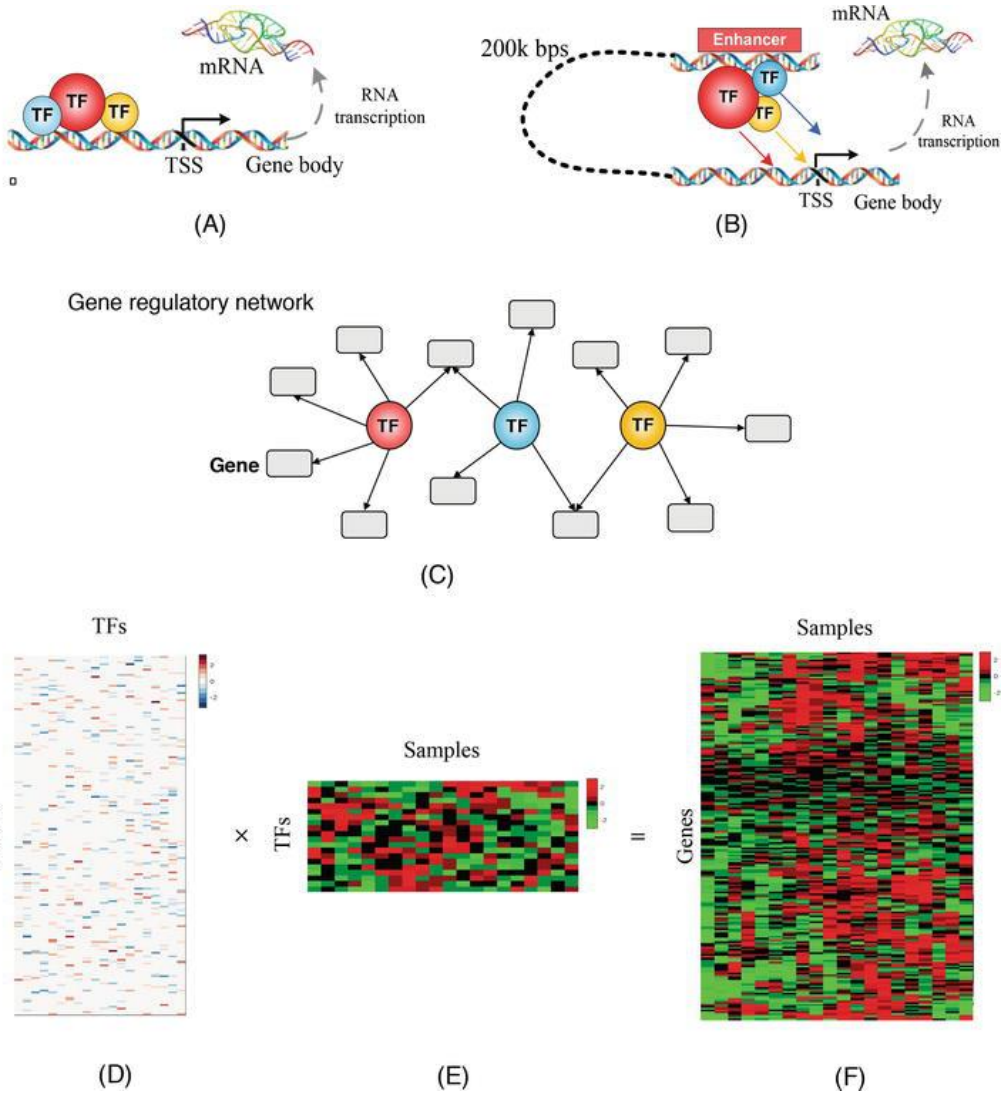



# Conclusion

- GRN inference
- Bulk and single cell
- Methods used for GRN inference
- papers




https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.intechopen.com%2Fchapters%2F68821&psig=AOvVaw1LMfcZMQaU5u0rKii91LEi&ust=1734979727339000&source=images&cd=vfe&opi=89978449&ved=0CBcQjhxqFwoTCOCTkK6FvIoDFQAAAAAdAAAAABAE

# References


1. Aibar, S.; González-Blas, C.B.; Moerman, T.; Huynh-Thu, V.A.; Imrichova, H.; Hulselmans, G.; Rambow, F.; Marine, J.-C.; Geurts, P.; Aerts, J.; et al. SCENIC: Single-Cell Regulatory Network Inference and Clustering. *Nat Methods* 2017, *14*, 1083–1086, doi:10.1038/nmeth.4463.

2. Bravo González-Blas, C.; De Winter, S.; Hulselmans, G.; Hecker, N.; Matetovici, I.; Christiaens, V.; Poovathingal, S.; Wouters, J.; Aibar, S.; Aerts, S. SCENIC+: Single-Cell Multiomic Inference of Enhancers and Gene Regulatory Networks. *Nat Methods* 2023, *20*, 1355–1367, doi:10.1038/s41592-023-01938-4.

3. Tran, A.; Yang, P.; Yang, J.Y.H.; Ormerod, J.T. scREMOTE: Using Multimodal Single Cell Data to Predict Regulatory Gene Relationships and to Build a Computational Cell Reprogramming Model. *NAR Genomics and Bioinformatics* 2022, *4*, lqac023, doi:10.1093/nargab/lqac023.

# Thanks

# Dataset- Validation

## scRNA-seq Data

- mouse brain
- Cancer cell
  - oligodendroglioma (4043 cells from six tumors)
  - melanoma 13 (1252 cells from fourteen lesions)

*Bioinformatics Advances*, 2023, vbad003
https://doi.org/10.1093/bioadv/vbad003
Advance Access Publication Date: 12 January 2023
Application Note

Gene expression

# scMEGA: single-cell multi-omic enhancer-based gene regulatory network inference

**Zhijian Li[1], James S. Nagai [1], Christoph Kuppe [2,3], Rafael Kramann [2,3,4] and Ivan G. Costa [1,*]**

# Methods

There are three major steps in scMEGA namely

1. multimodal data integration
2. identification and filtering of candidate TFs and genes
3. GRN assembly and analysis.

## Tools

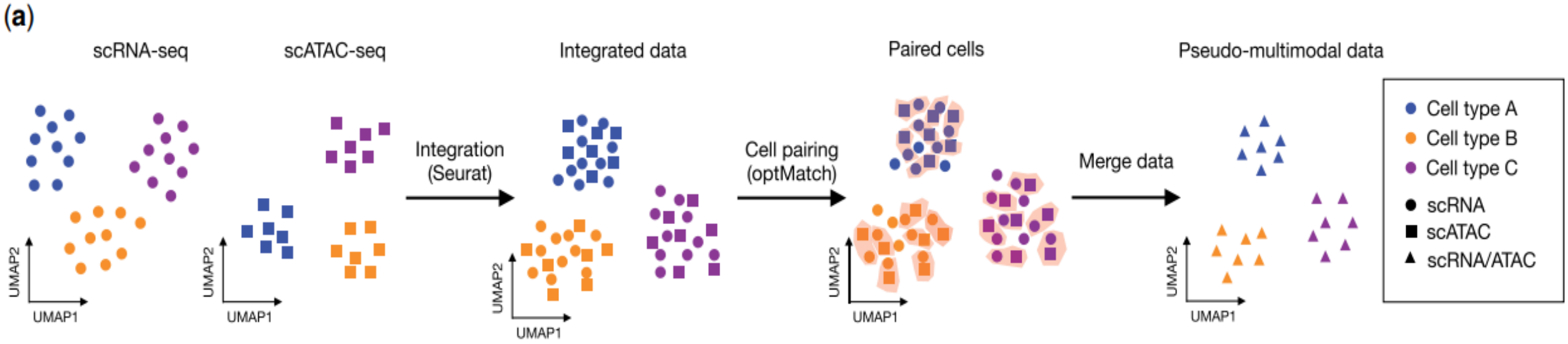
- **Seurat** for scRNA-seq analysis and data integration (Stuart *et al.*, 2019),
- **OptMatch** for Cell Pairing
- **ArchR** for scATAC-seq analysis and trajectory inference (Granja *et al.*, 2021),
- **chromVAR** for TF activity estimation (Schep *et al.*, 2017)
- **igraph** (Csardi and Nepusz, 2006) for network analysis

# 1. Single-cell multimodal data integration

**Integrate** single-cell RNA sequencing (scRNA-seq) and scATAC-seq data.

scMEGA uses canonical correlation analysis (CCA) from Seurat to project cells into a shared embedding space, combining gene expression from scRNA-seq and gene activity scores from scATAC-seq.

**Cell Pairing**: One-to-one matching between scRNA-seq and scATAC-seq cells is performed using the OptMatch algorithm.

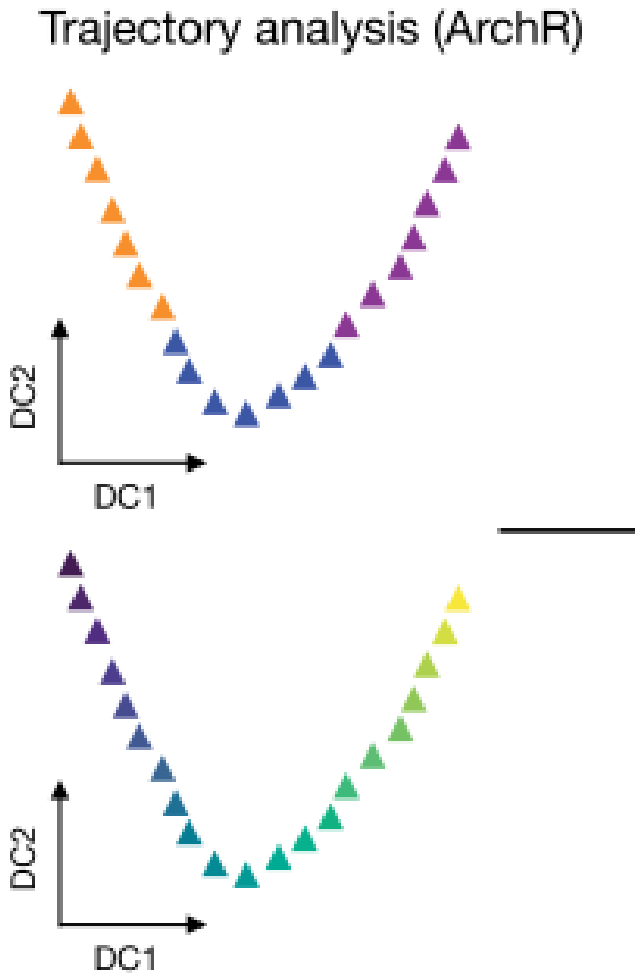


# scMEGA infers a pseudotime trajectory to characterize the underlying dynamic process of a given cell type using ArchR

Pseudotime refers to an inferred temporal ordering of cells along a biological process, like differentiation, without actual time point measurements.

It allows researchers to understand how gene expression or chromatin accessibility changes as cells progress through a dynamic process.

This analysis helps researchers' study how regulatory mechanisms (like transcription factors binding to enhancers) change as cells transition between different states.

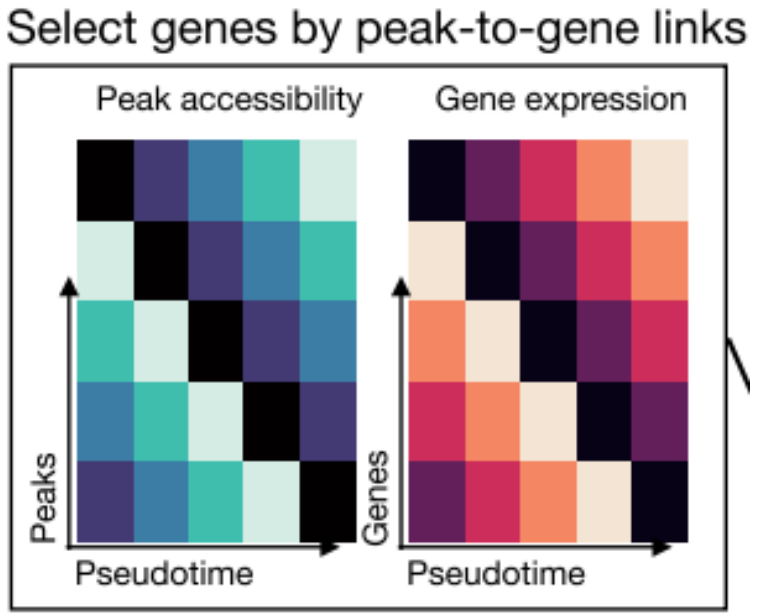


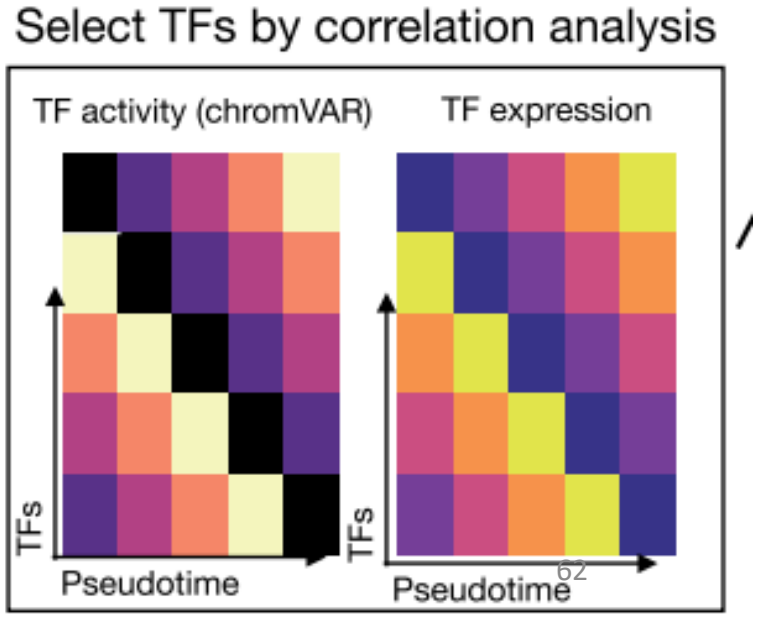
# scMEGA is filtering both genes and transcription factors (TFs)

scMEGA is filtering both genes and transcription factors (TFs) based on two correlation analyses:

- one between chromatin accessibility (scATAC-seq peaks) and gene expression
- and the other between TF binding activity (chromVAR- scATAC-seq ) and TF expression.

This step is crucial for selecting only the relevant genes and TFs to construct a meaningful gene regulatory network (GRN).

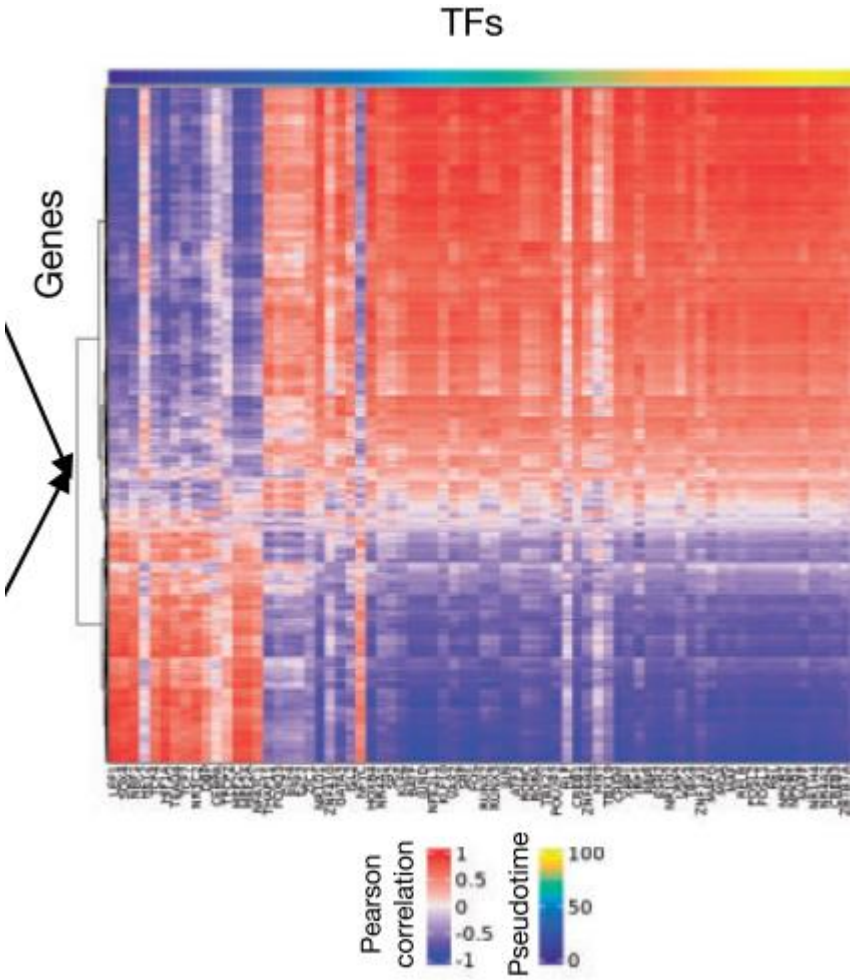


# Calculating the correlation between selected transcription factors (TFs) and genes.

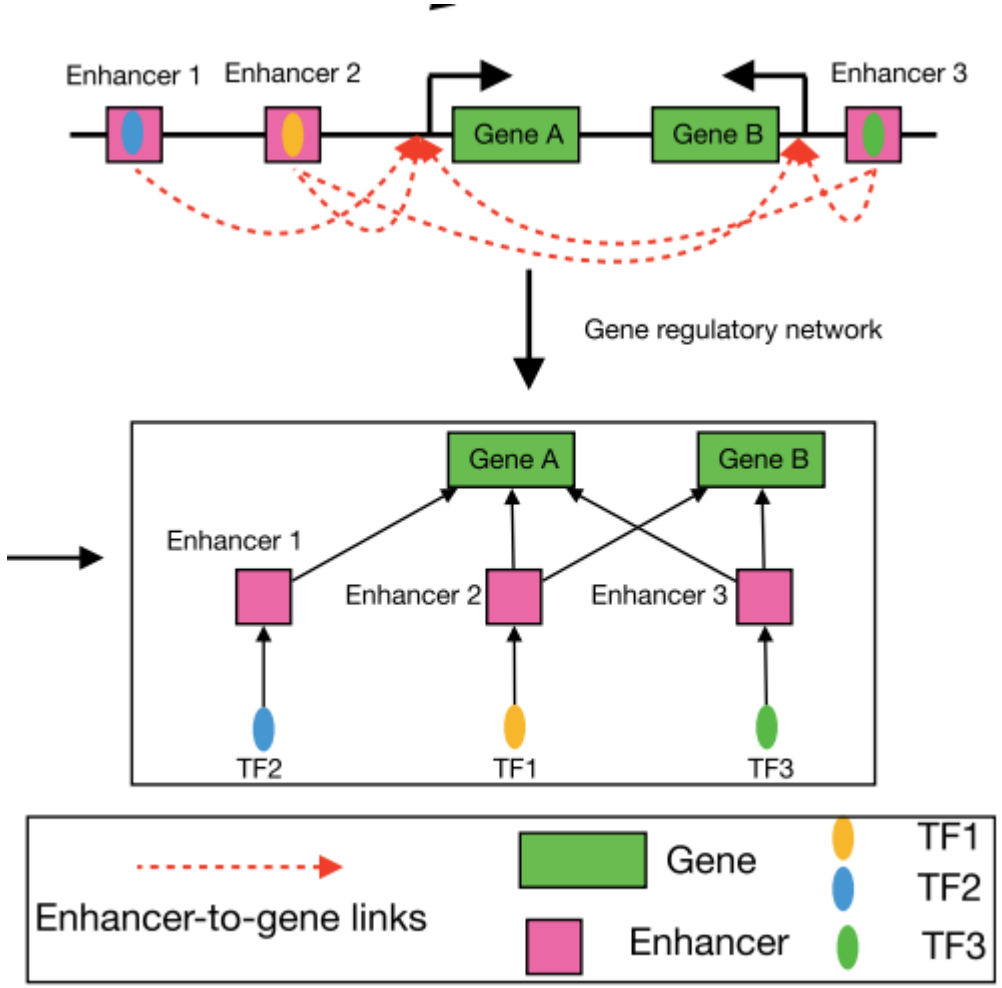
**Heatmap Explanation:**

- The **x-axis** represents TFs, and the **y-axis** represents genes.
- Each cell shows the **Pearson correlation** between a TF and a gene, with **red** indicating a strong positive correlation (TF likely activates the gene) and **blue** showing a negative correlation (TF likely represses the gene).
- Pseudotime (depicted by the color bar on the top) is another dimension in this heatmap, which might be used to align TFs and genes based on their behavior over time. The colors (blue to yellow) represent pseudotime progression, with blue indicating earlier time points and yellow indicating later time points.
- This step estimates **how strongly TFs regulate genes based on expression correlations**. The results form a **GRN**, highlighting key regulatory relationships over time.



scMEGA uses enhancer-to-gene links and motif matching to find enhancer-based TF-to-gene interactions. These are used to filter the previously defined quantitative GRN as shown in (d)

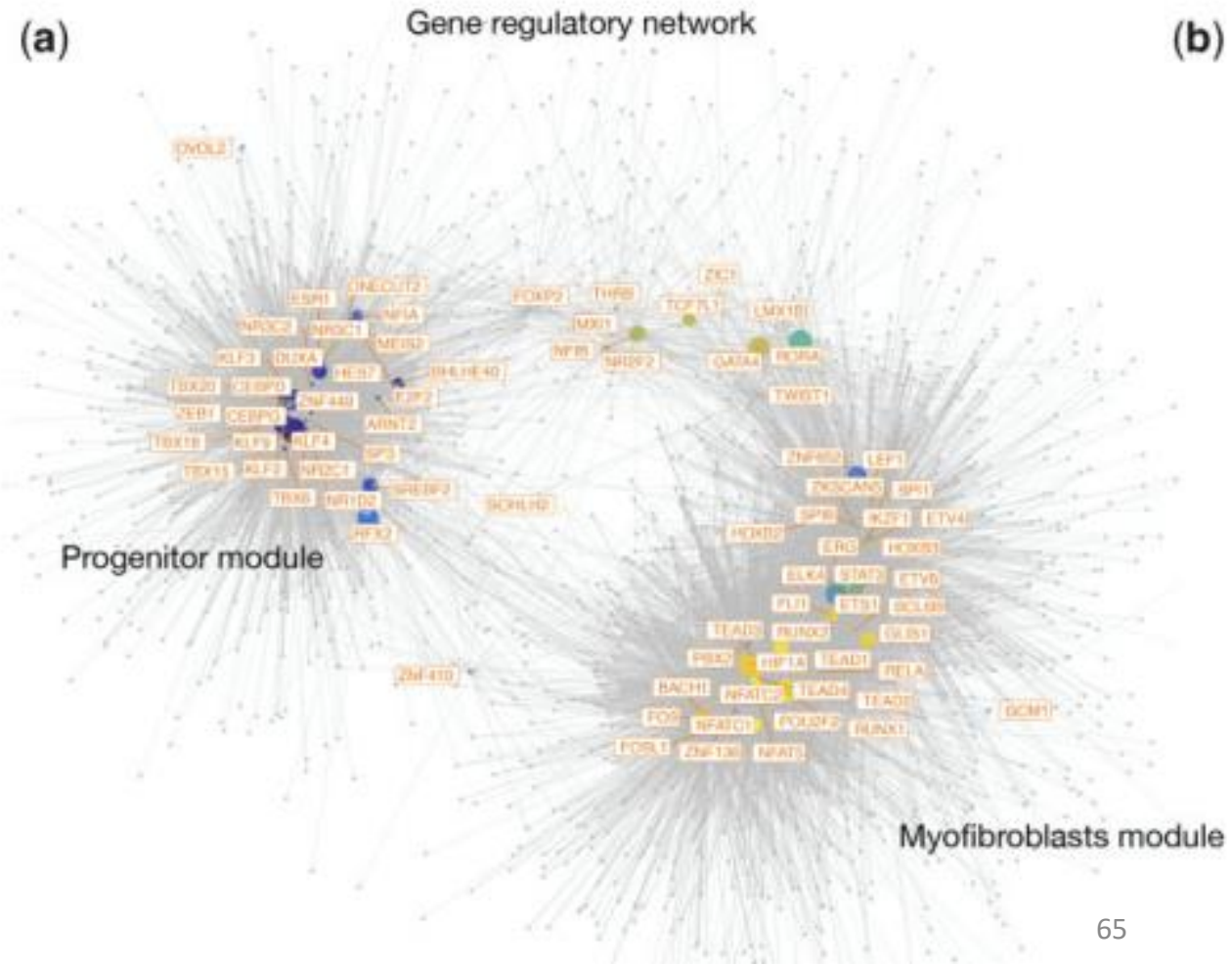
1. Enhancers are defined as peaks that are at least 2k base pairs (bp) from the transcription start site of a gene.
2. Link enhancers to genes by correlating enhancer accessibility with gene expression.
3. Link TFs to enhancers through motif matching (TF binding motifs in enhancers).
    - chromVAR is used to estimate TF binding activity by identifying the presence of TF motifs in the accessible chromatin regions (enhancers).
4. Create TF-to-gene links by combining enhancer-to-gene and TF-to-enhancer interactions.



# Case study on myocardial infarction known as "heart attack"

## (a) Gene Regulatory Network (GRN) Visualization:

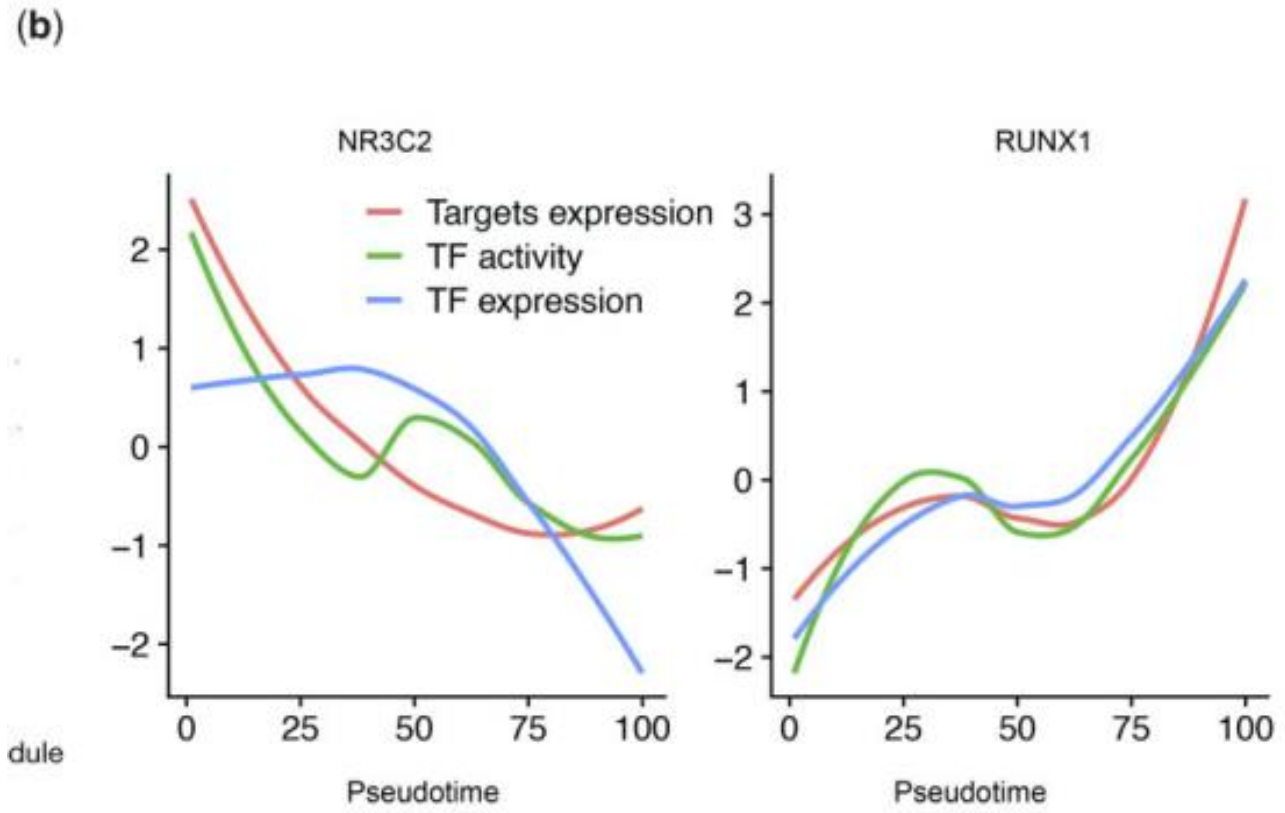
- This network represents the transcription factors (TFs) and genes involved in myofibroblast differentiation.
- Each **node** represents either a **TF (regulator)** or a **gene (target)**. The connections between nodes represent regulatory interactions.
- The nodes are divided into two major modules:
  - **Progenitor module**: Likely containing TFs regulating early-stage progenitor cells.
  - **Myofibroblast module**: Contains TFs and genes involved in later stages of fibroblast differentiation into myofibroblasts.
- **Coloring**: The TFs are colored based on the point in pseudotime at which they have the highest activity (estimated using chromVAR).



# (b) Line Plots of TF Activity, Expression, and Target Expression:

These plots show the changes in TF activity, TF expression, and target gene expression along the pseudotime trajectory for two important TFs: NR3C2 and RUNX1.

- X-axis (pseudotime): Represents the progression of differentiation from progenitor cells to myofibroblasts.
- Y-axis (z-score): Shows standardized values for TF activity, TF expression, and target gene expression.
- The green, blue, and red lines correspond to TF activity, TF expression, and target gene expression, respectively.
- NR3C2: Its expression and target activity decrease over pseudotime.
- RUNX1: Its expression and target activity increase over pseudotime, suggesting that it plays a more active role in the later stages of differentiation.



# Cell oracle

nature > articles > article

Article | Open access | Published: 08 February 2023

## Dissecting cell identity via network inference and in silico gene perturbation

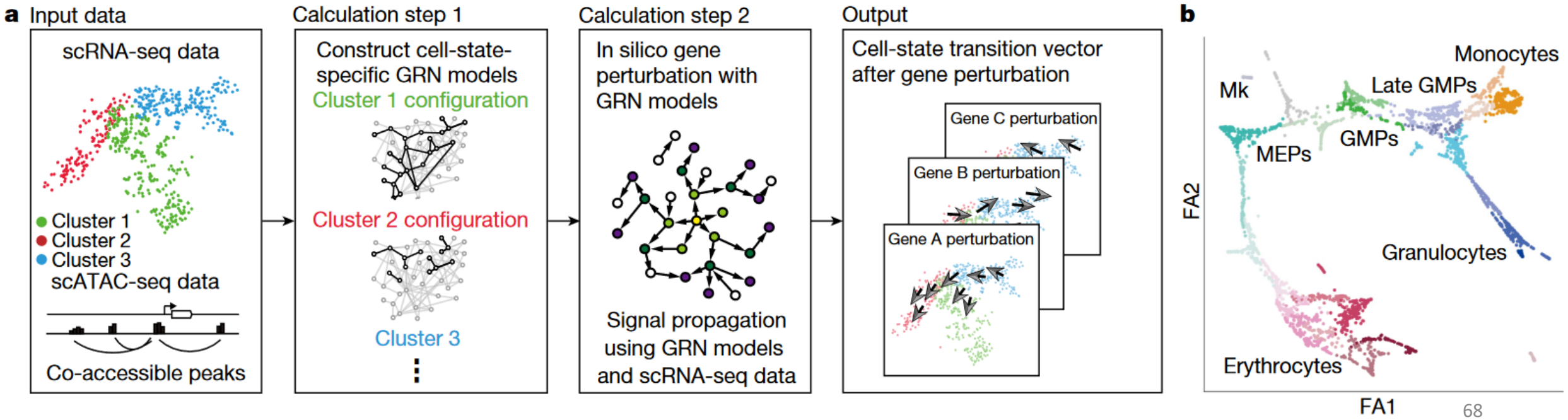
Kenji Kamimoto, Blerta Stringa, Christy M. Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel & Samantha A. Morris

*Nature* **614**, 742–751 (2023) | Cite this article

# Cell oracle

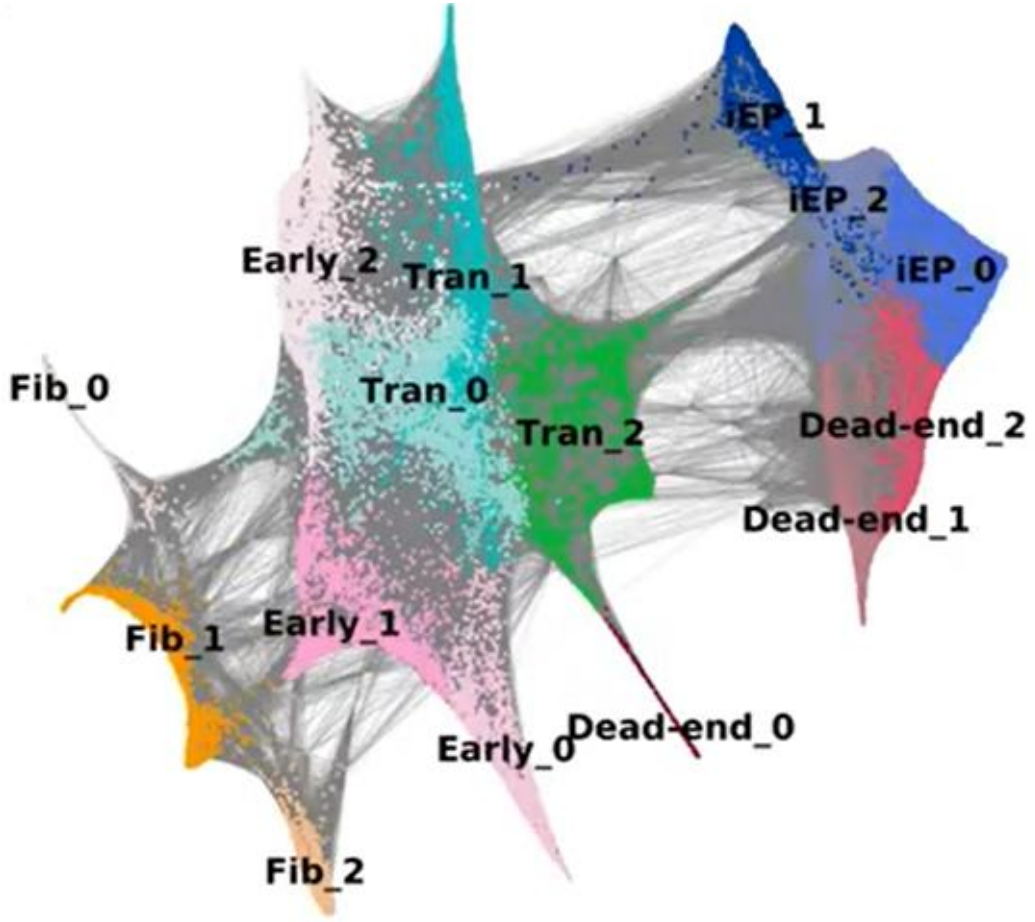
- CellOracle, a computational method to infer GRNs from single-cell transcriptome and epigenome data.

- Using inferred GRNs, CellOracle simulates gene expression changes in response to TF perturbation



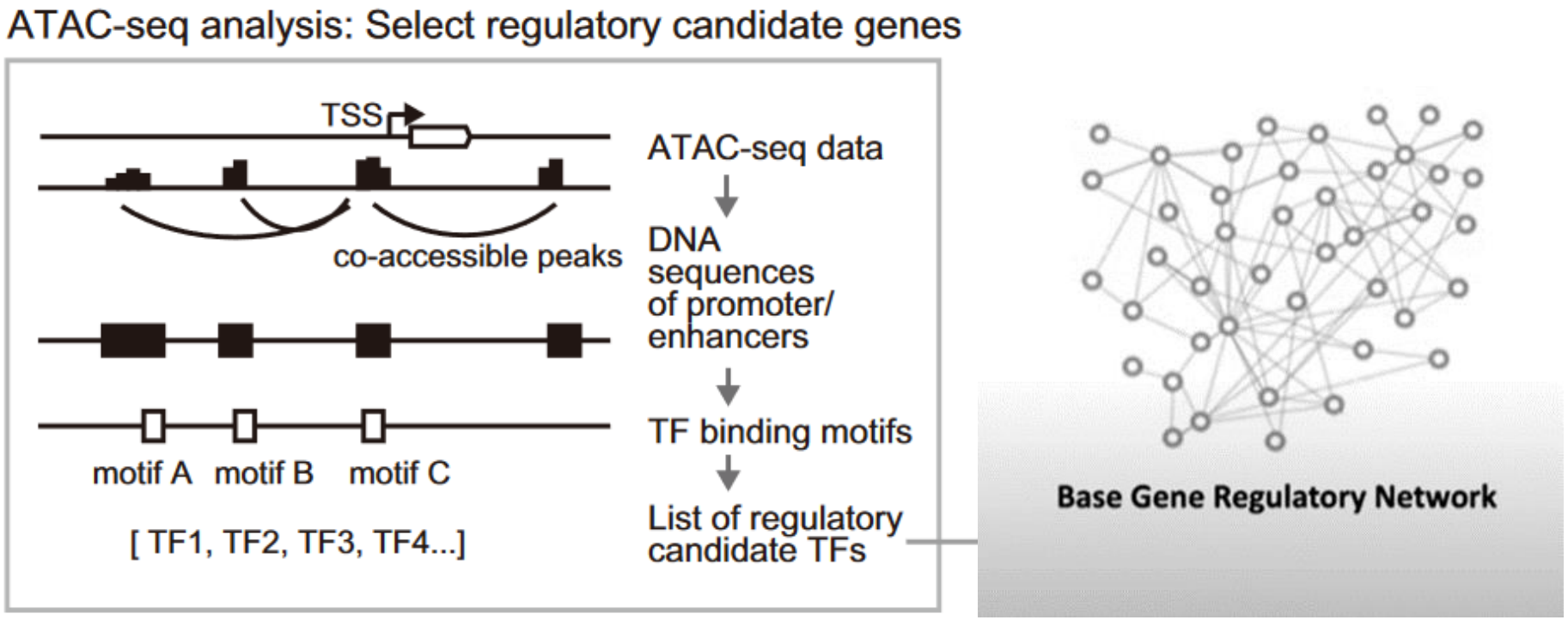
# CellOracle: Dissecting cell identity via network inference

- Using scRNA-seq and scATAC-seq data to infer gene regulatory network(GRN) **for each cluster**



# step 1: building a base GRN

- Using scATAC-seq data
- Identify accessible promoter /enhancer **(Cicero)**
- Scan these regulatory elements for TF binding motifs
- List of all potential regulatory connections between a TF and its target genes



# Co-accessible peak find with cicero

Cicero, an algorithm that identifies co-accessible pairs of DNA elements using single-cell chromatin accessibility data

and **connects regulatory elements to their putative target genes**.

## 6. Run Cicero

In [10]:

```
# Run the main function
conns <- run_cicero(cicero_cds, chromosome_length) # Takes a few minutes to run

# Save results (Optional)
#saveRDS(conns, paste0(output_folder, "/cicero_connections.Rds"))

# Check results
head(conns)
```

A data.frame: 6 × 3

| | Peak1 <chr> | Peak2 <fct> | coaccess <dbl> |
|---|---|---|---|
| 1 | chr10_100006139_100006389 | chr10_99774288_99774570 | -0.003546179 |
| 2 | chr10_100006139_100006389 | chr10_99825945_99826237 | -0.027536333 |
| 3 | chr10_100006139_100006389 | chr10_99830012_99830311 | 0.009588013 |
| 4 | chr10_100006139_100006389 | chr10_99833211_99833540 | -0.008067111 |
| 5 | chr10_100006139_100006389 | chr10_99941805_99941955 | 0.000000000 |

https://github.com/morris-lab/CellOracle/blob/master/docs/notebooks/01_ATAC-seq_data_processing/option1_scATAC-seq_data_analysis_with_cicero/01_atacdata_analysis_with_cicero_and_monocle3.ipynb

## Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data

Hannah A Pliner [1], Jonathan S Packer [1], José L McFaline-Figueroa [1], Darren A Cusanovich [1], Riza M Daza [1], Delasa Aghamirzaie [1], Sanjay Srivatsan [1], Xiaojie Qiu [2], Dana Jackson [1], Anna Minkina [1], Andrew C Adey [3], Frank J Steemers [4], Jay Shendure [5], Cole Trapnell [6]

Affiliations + expand

PMID: 30078726 PMCID: PMC6582963 DOI: 10.1016/j.molcel.2018.06.044

https://pubmed.ncbi.nlm.nih.gov/30078726/

# step 2: identify active connections from scRNA-seq

- Using scRNA-seq
- Build a regularized machine learning model that predicts TF-target gene relationships
- ML model fitting results –> certainty of connections as a distribution
- Removal of weak/inactive connections from the base GRN

scRNA-seq data analysis: GRN model fitting

Feature selection with regularized regression model
(Bayesian ridge regression or Bagging ridge)

ML prediction model with scRNA-seq data

$$y \sim \mathcal{N}(\mu = \alpha + X\beta, \epsilon)$$

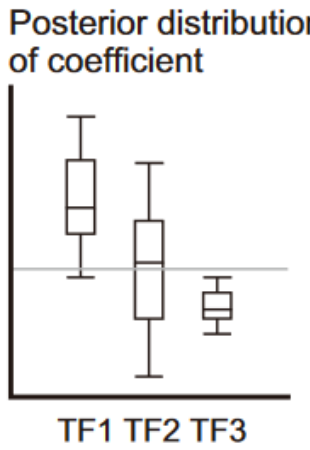
y: Target gene  X: Candidate regulatory TFs

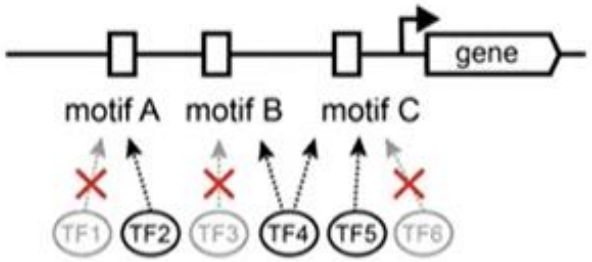
where

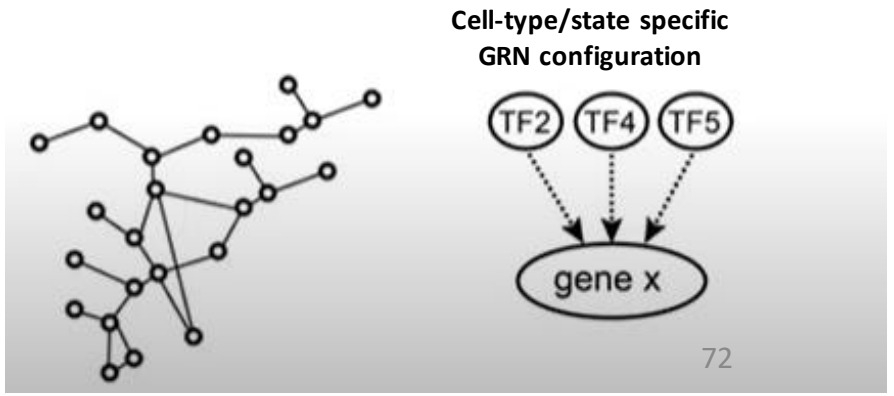
$$X\beta = \sum_{i=1}^{n} \beta_i x_i = \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_m x_m$$

$$\beta \sim \mathcal{N}(0, \sigma_\beta)$$

Regularizing prior distribution for coefficient β



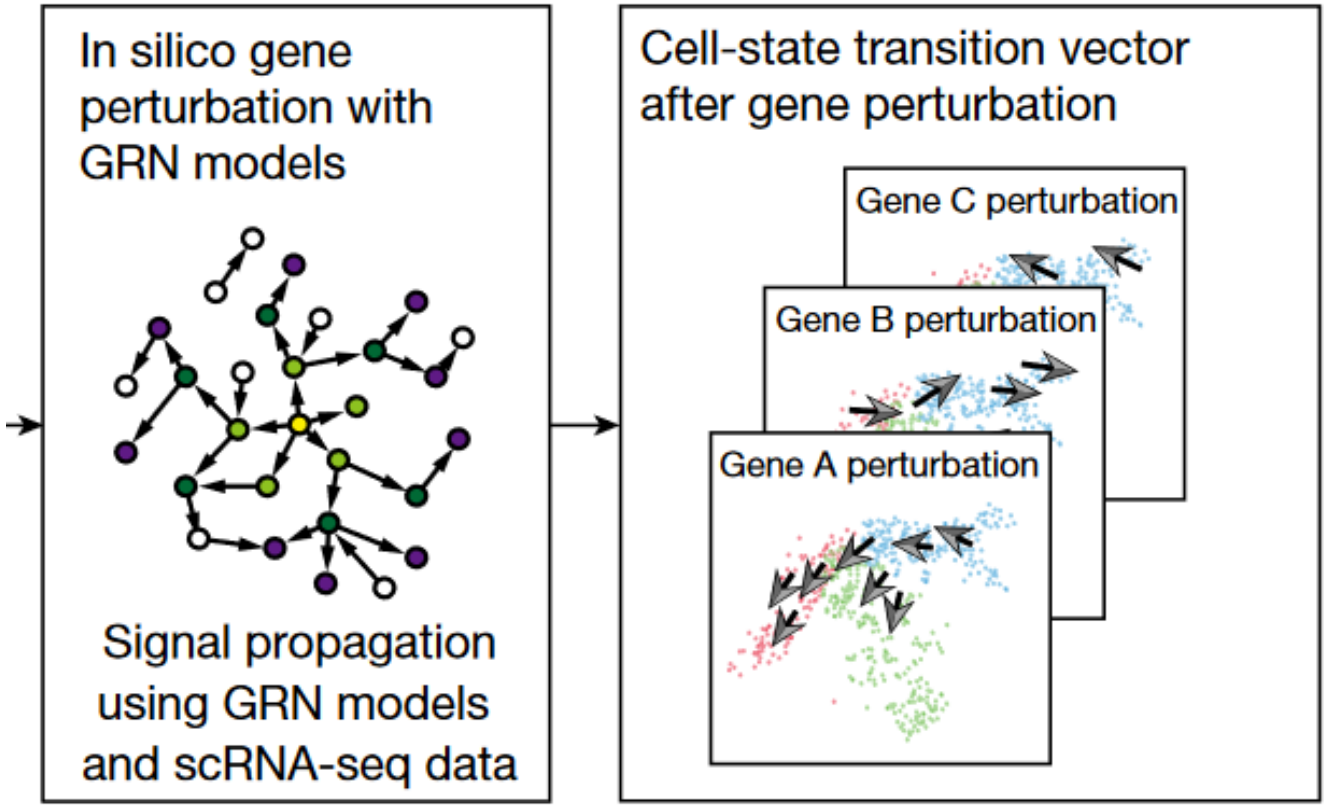



# Step3: simulation of transcription factor perturbation

(simulate global downstream shifts in gene expression following knockout (KO) or overexpression of TFs.)

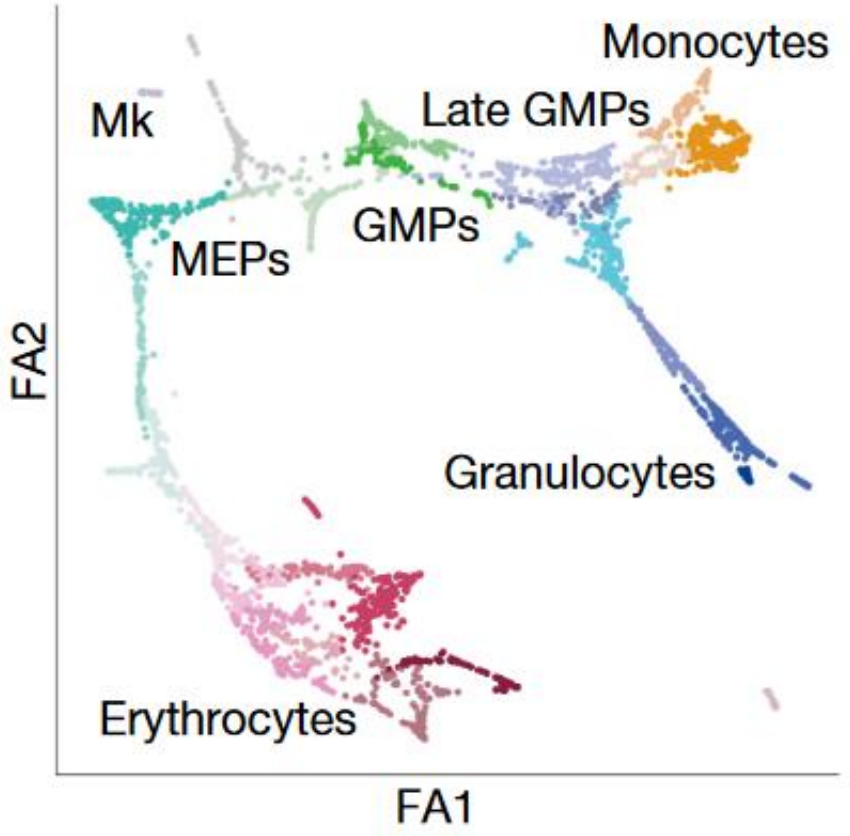
- CellOracle simulates the change in cell state in response to a TF perturbation
- projecting the results onto the cell trajectory map (right).

- CellOracle output: intuitive network analysis using graph theory –> initial prioritization of interesting TFs
- Use Ml model that predicts TF-target gene relationships to simulate TF perturbation
- **Propagate** this initial perturbation effect within the GRN to simulate indirect, global transcriptional effects

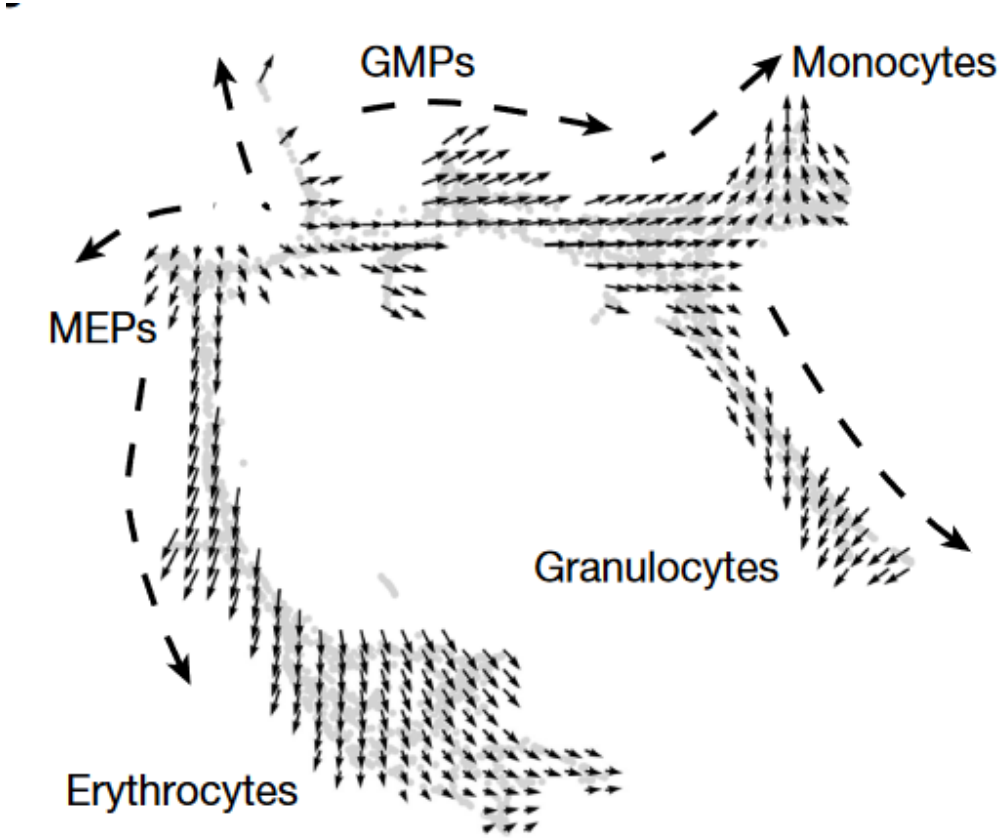


- Force-directed graph of 2,730 myeloid progenitor cells from Paul et al.16.
- Twenty-four cell clusters (Louvain clustering) were organized into **six main cell types**. Mk, megakaryocytes.

precursors of red blood cells

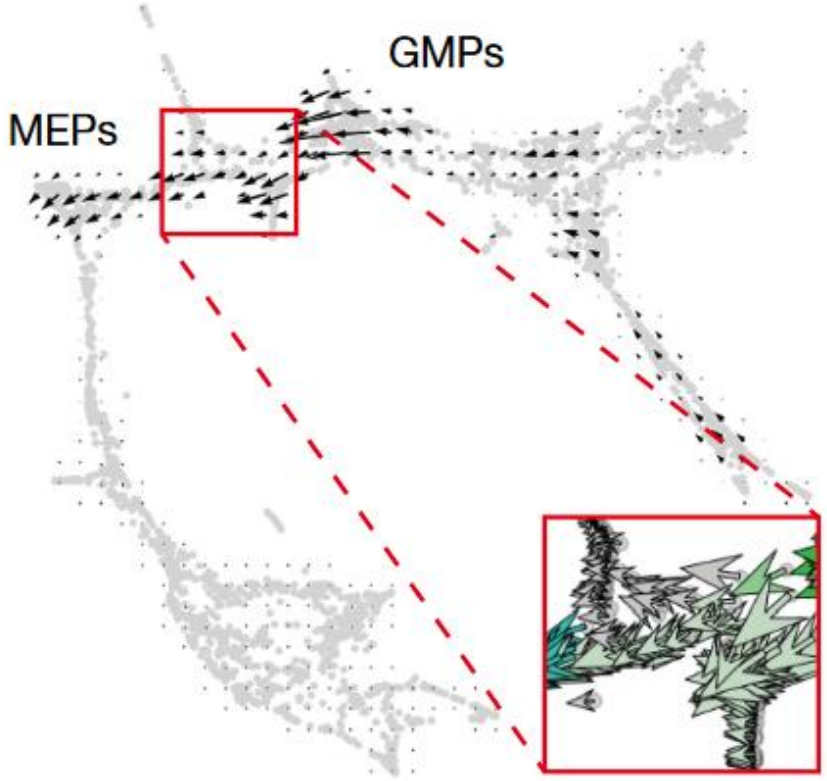



- Differentiation vectors for each cell projected onto the force-directed graph



- CellOracle simulation of cell-state transition in Spi1 KO simulation.

- Spi1 (SPI1) Spi1 promotes GM lineage differentiation.




- Spi1 KO simulation vector field with perturbation scores (PSs).

- The inhibition of Spi1 shifts cell identity from the GM to the ME lineage