Modelling gene regulation via integrative analysis of single cell multi-omics data

07/03/2023
Single-Cell Plus – Data Science Challenges
in Single-Cell Research
Banff International Research Station

Zhana Duren
Assistant Professor
Center for Human Genetics
Department of Genetics and Biochemistry

Outline

1 Introduction

2 LINGER Model

3 Validation

4 Application

Outline

1 Introduction

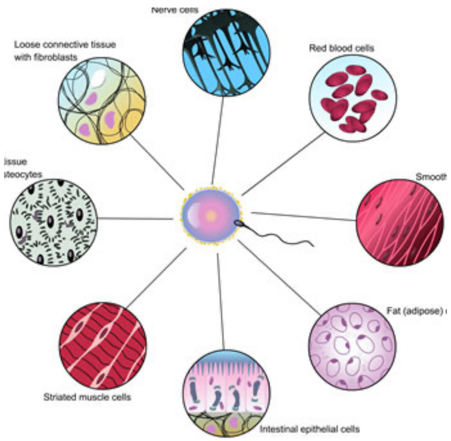
2 LINGER Model

3 Validation

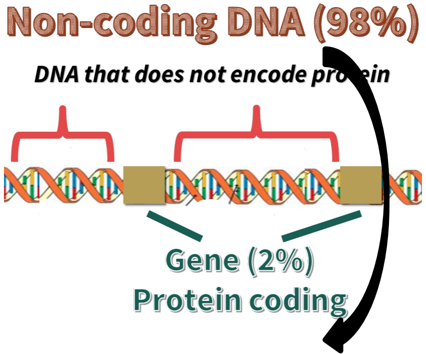
4 Application

What are biologists thinking?

Developmental biologists



Geneticists



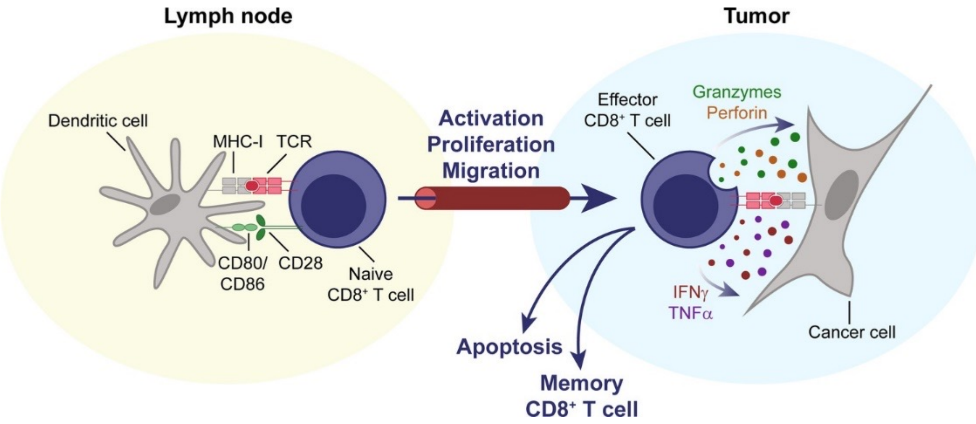
Non-coding DNA (98%)

DNA that does not encode protein

Gene (2%)
Protein coding

Diseases?

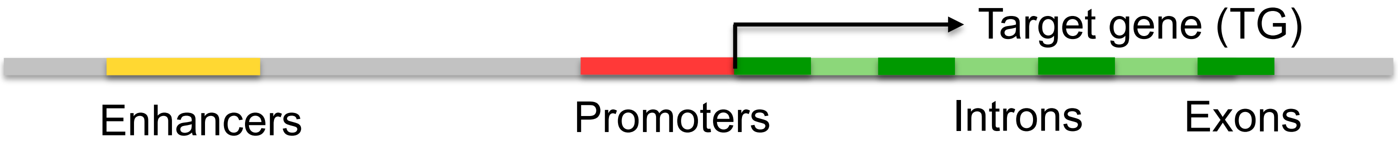
Medical health researchers



Turning on or off the expression of genes

Gene regulatory networks

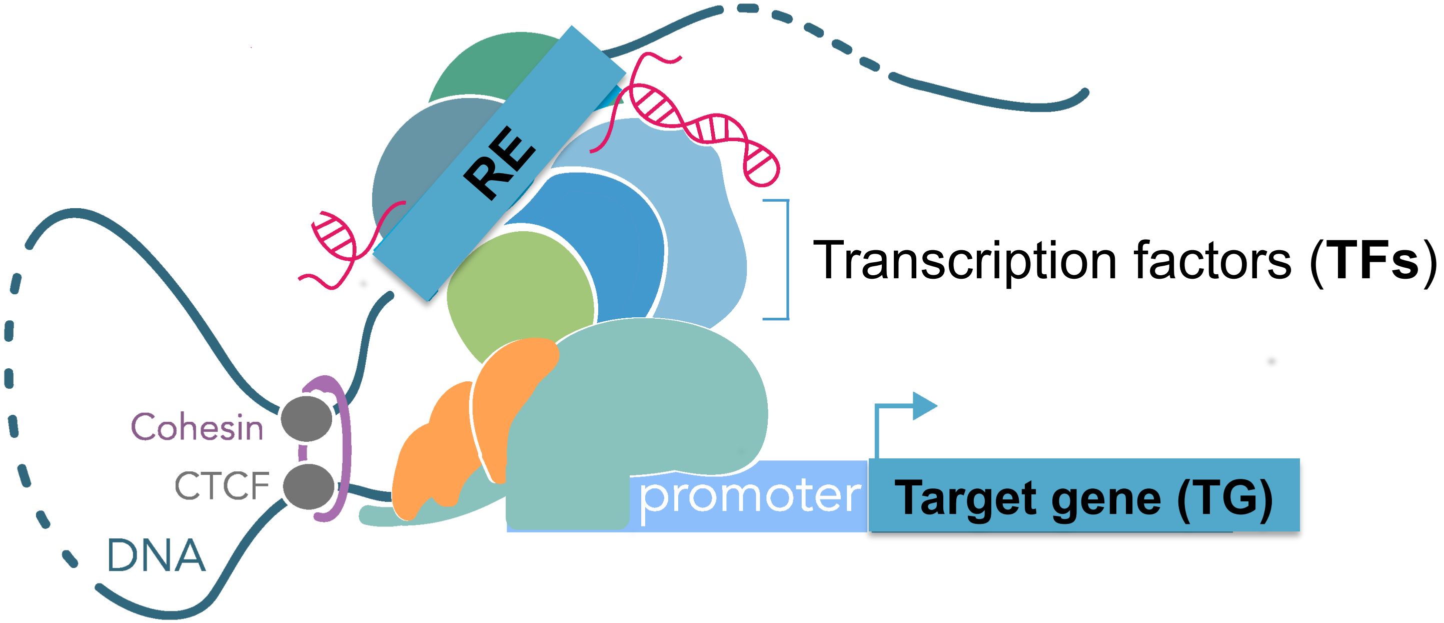
Gene regulation


Target gene (TG)

Enhancers Promoters Introns Exons

cis- regulatory elements **(REs)**
(Enhancers, Promoters)

Non-coding regions!


Transcription factors (**TFs**)

Challenges in GRN inference

- Non-coding region (98%) (ATAC-seq + RNA-seq)
- Mixture of cell types (single cell)
- Complex system, require large samples
- Incorporate knowledge (motif for protein-DNA binding)

sc-multiome: paired RNA-seq and ATAC-seq for the same cell

Outline

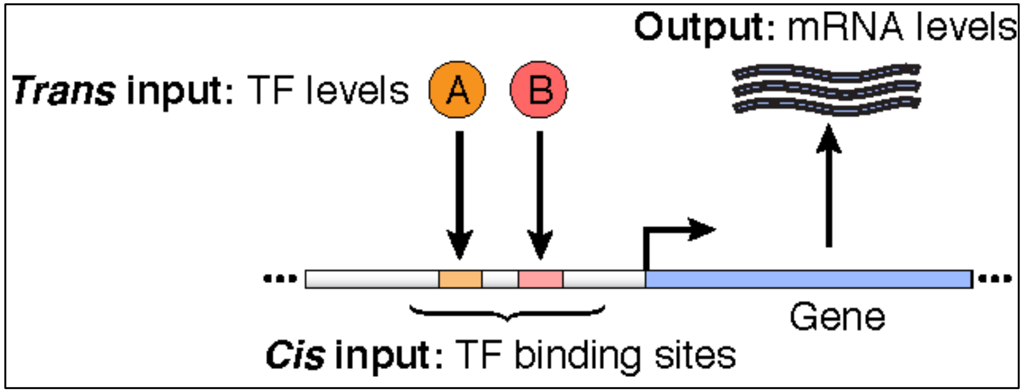
1 Introduction

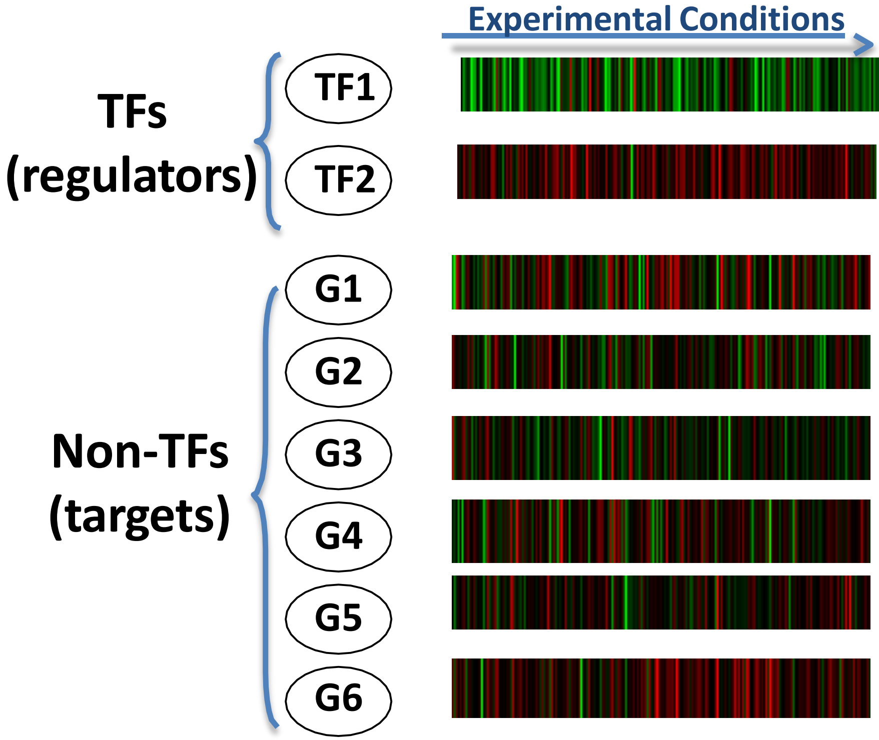
2 LINGER Model

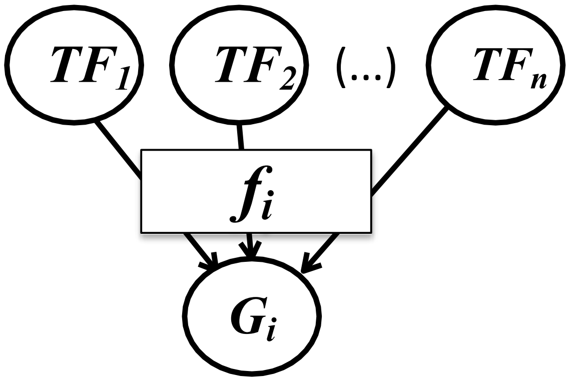
3 Validation

4 Application

GRN by linear regression model





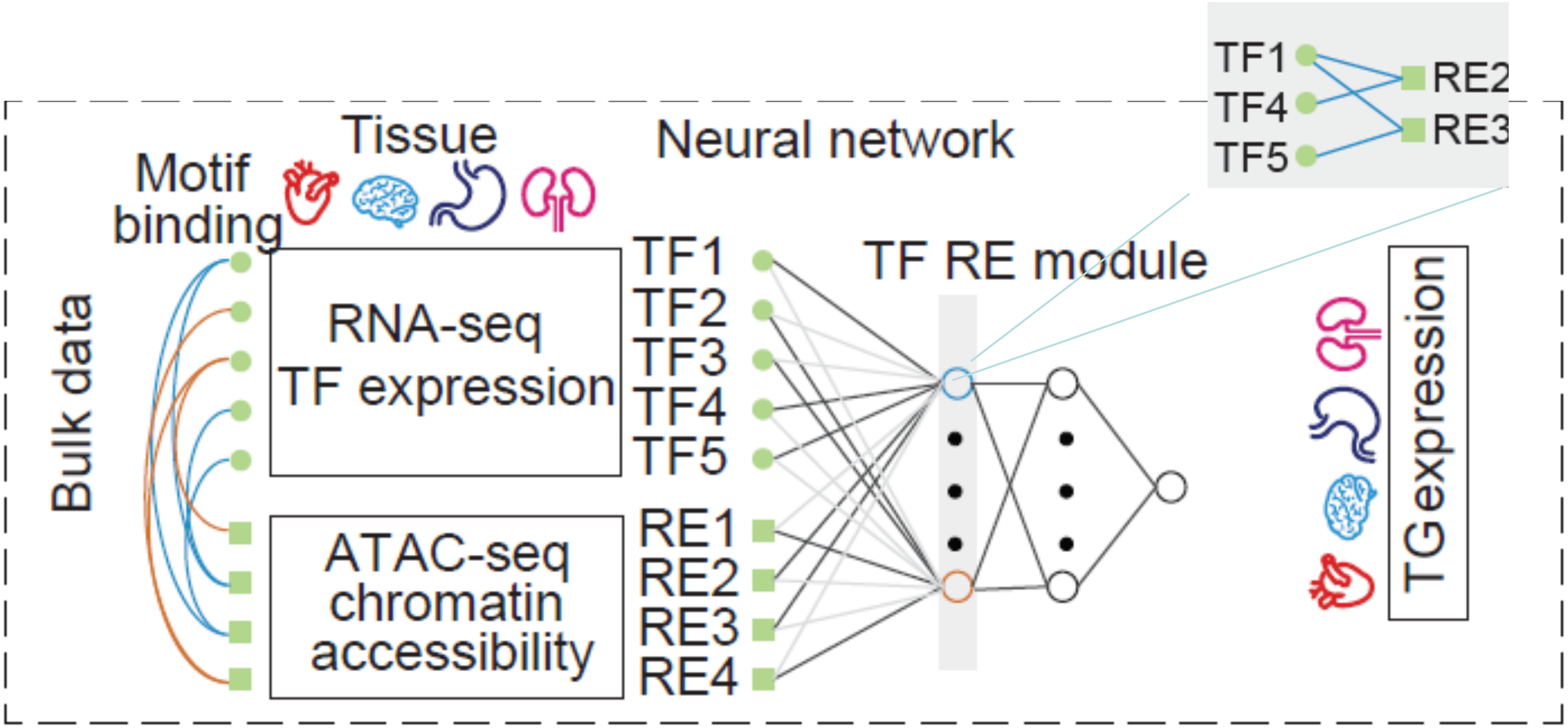


- Gene expression prediction:

G1=$\beta_{10} + \beta_{11}TF_1 + \beta_{12}TF_2 + \varepsilon_1$
G2=$\beta_{20} + \beta_{21}TF_1 + \beta_{22}TF_2 + \varepsilon_2$
G3=$\beta_{30} + \beta_{31}TF_1 + \beta_{32}TF_2 + \varepsilon_3$

- Non-linear
- Non-coding regions are not considered

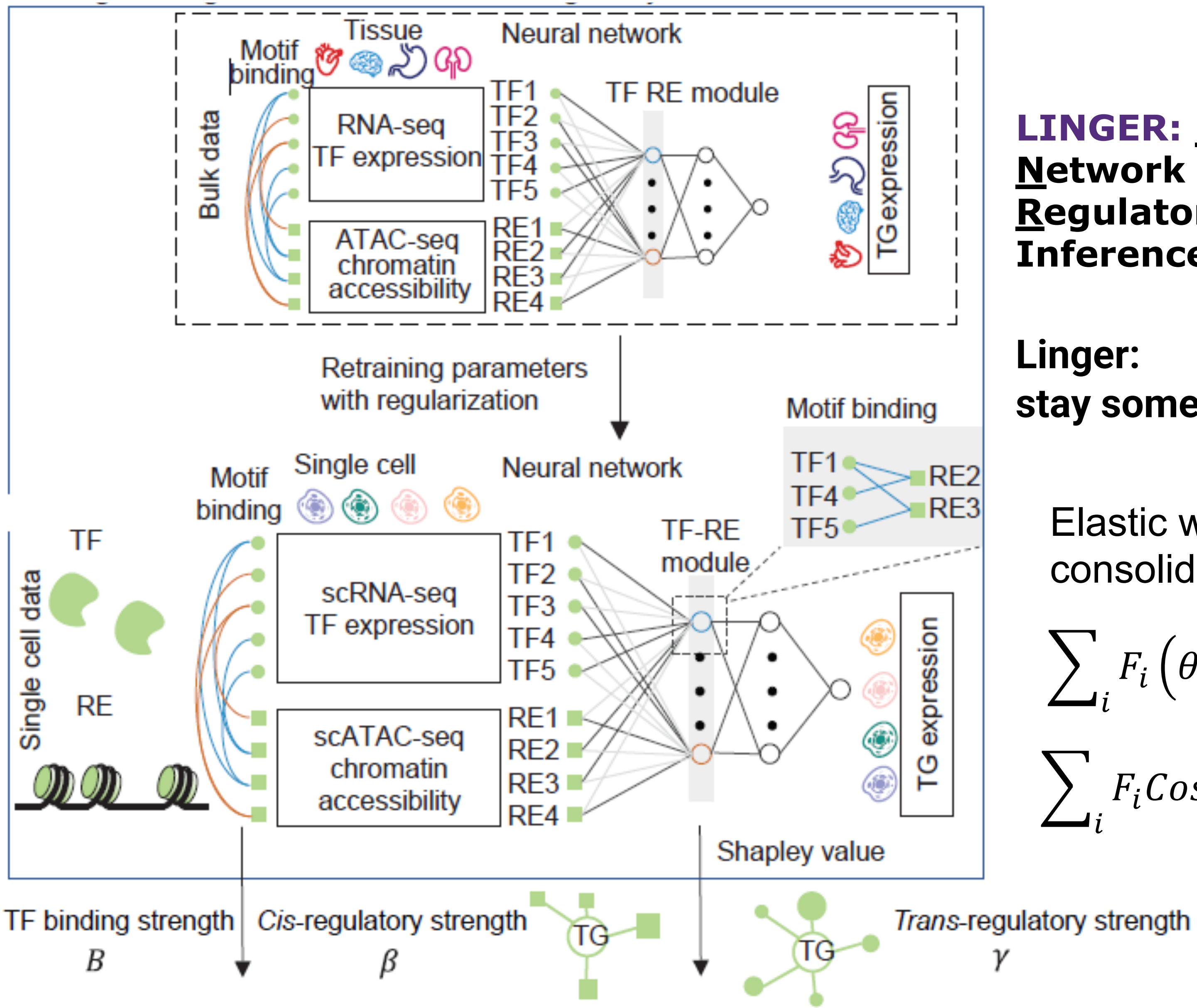
Fit TG by TF and RE on bulk data by neural network



Manifold regularization $tr(w^T L w)$

L: Laplacian matrix of TF-RE motif matching
w: weight matrix

lifelong learning on single cell data



LINGER: <u>Li</u>felong <u>N</u>eural <u>N</u>etwork for <u>Ge</u>ne <u>R</u>egulatory Network Inference

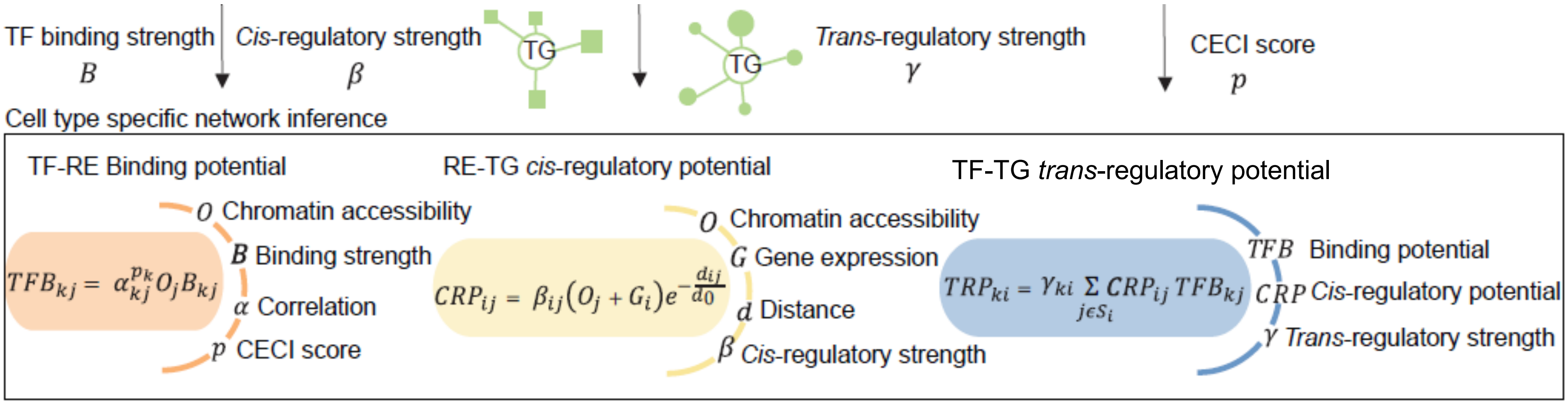
Linger:
stay somewhere longer

Elastic weight consolidation (EWC)

$$\sum_i F_i \left(\theta_i - \theta_i^{(b)}\right)^2$$

$$\sum_i F_i CosineDis(\theta_i, \theta_i^{(bulk)})$$

Cell Type specific GRNs



TF binding strength B

Cis-regulatory strength β

Trans-regulatory strength γ

CECI score p

Cell type specific network inference

TF-RE Binding potential

$$TFB_{kj} = \alpha_{kj}^{p_k} O_j B_{kj}$$

O Chromatin accessibility

B Binding strength

α Correlation

p CECI score

RE-TG *cis*-regulatory potential

$$CRP_{ij} = \beta_{ij}(O_j + G_i)e^{-\frac{d_{ij}}{d_0}}$$

O Chromatin accessibility

G Gene expression

d Distance

β *Cis*-regulatory strength

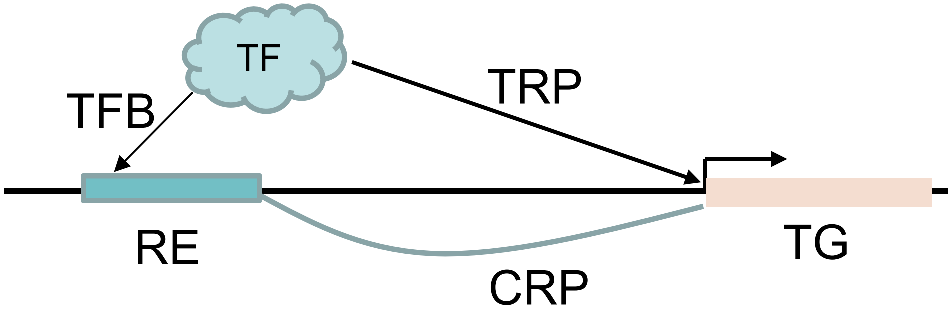
TF-TG *trans*-regulatory potential

$$TRP_{ki} = \gamma_{ki} \sum_{j \in S_i} CRP_{ij} TFB_{kj}$$

TFB Binding potential

CRP *Cis*-regulatory potential

γ *Trans*-regulatory strength



CECI assess the ability of TFs to regulate open chromatin.

Outline

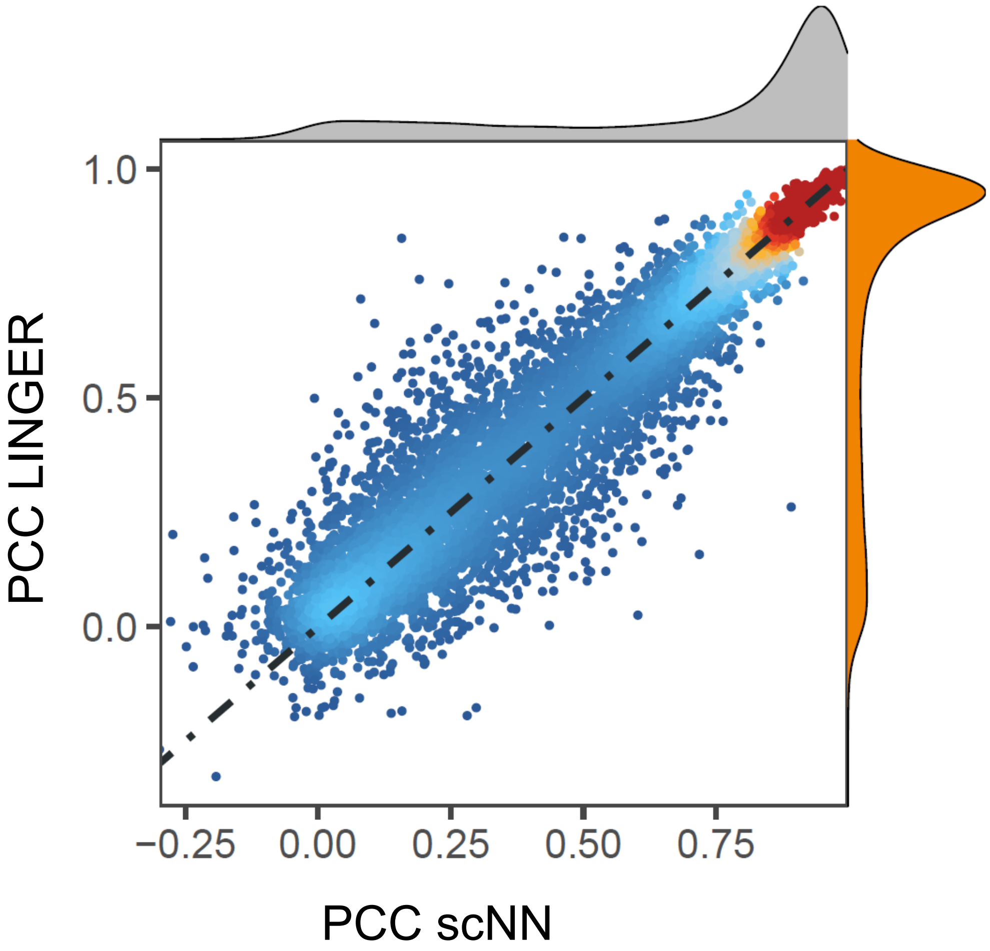
1 Introduction

2 LINGER Model

3 Validation

4 Application

Lifelong learning **cannot improve** TG expression predictions



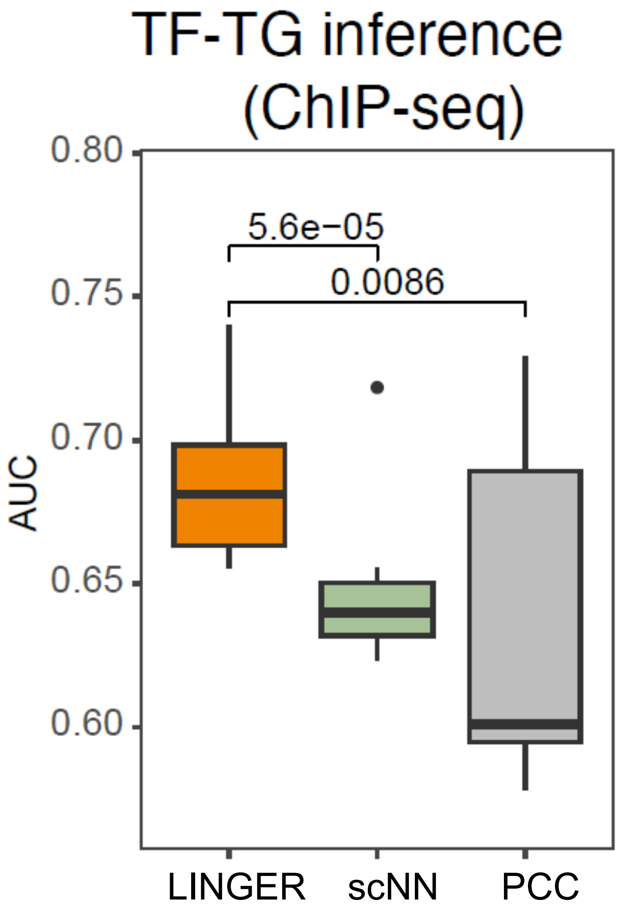
Data: PBMC sc-multiome data

Metric: Pearson Correlation of observed versus predicted gene expression

Dot: one gene

5-fold cross validation

Lifelong learning improves general GRN inference



Ground Truth:
ChIP-seq data of 10 TFs from PBMC
(different cell types)

Top 1,000 targets of each TF are considered as positive samples, and other are negative.

scNN: same neural network method without using lifelong learning.

PCC: Pearson correlation of TF-TG pairs on single cell data.

LINGER: Average of absolute value of the shapely value across cells.

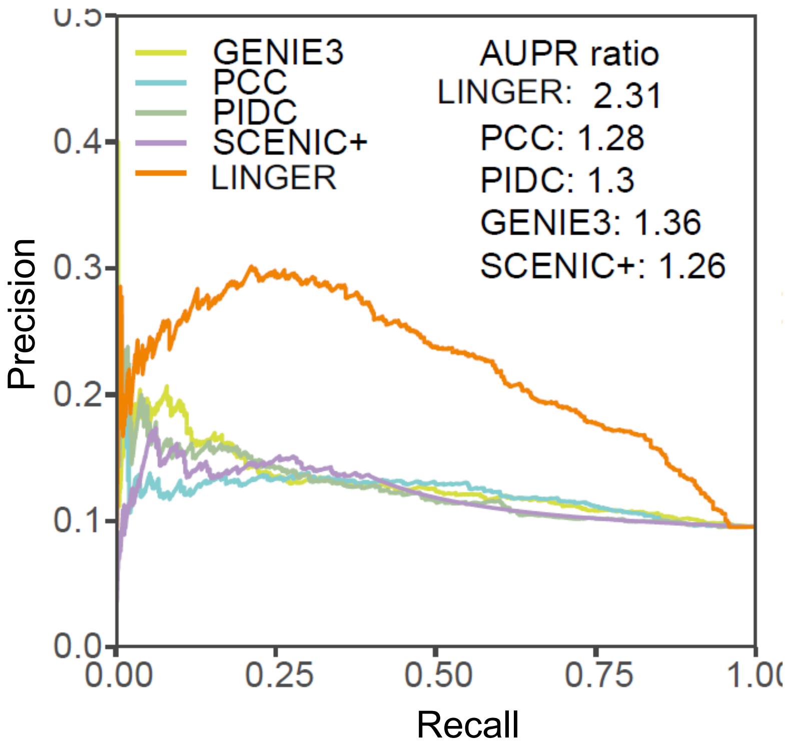
In terms of **TG expression prediction**, using external data information by lifelong learning **cannot improve the performance**.

But the **inferred GRN becomes more accurate!!!**

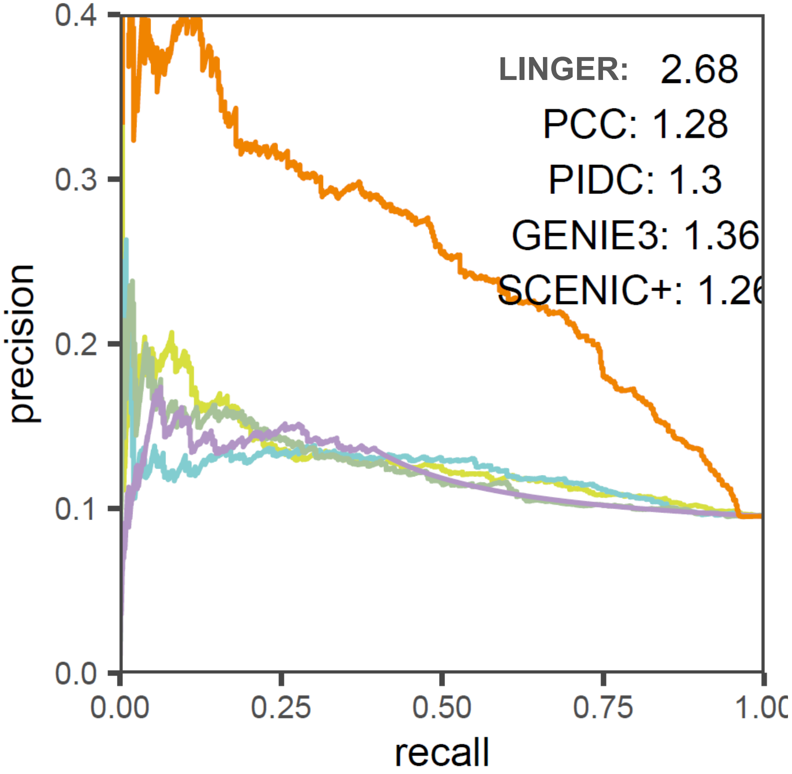
LINGER improve cell type specific GRN inference

Ground truth: STAT1 ChIP-seq on classic monocytes

TF-RE binding prediction



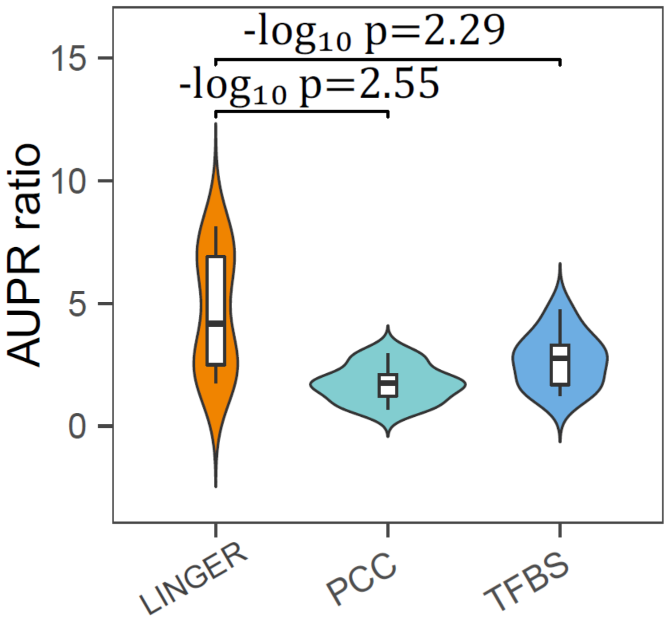
TF-TG prediction



LINGER improve cell type specific GRN inference

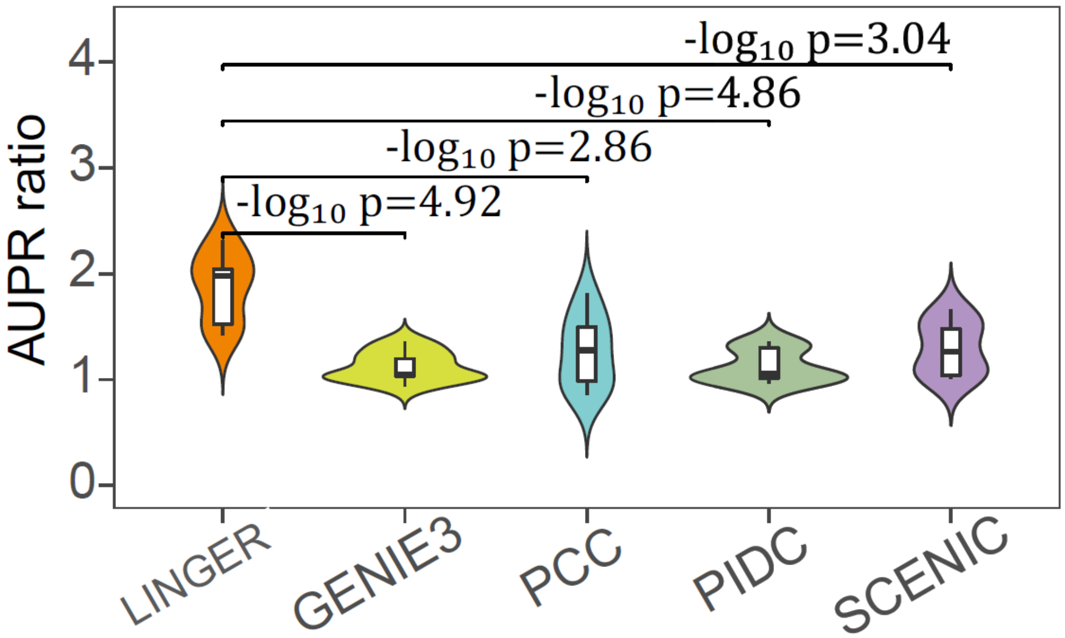
Ground truth: ChIP-seq of 10 TFs on PBMC

TF-RE binding prediction



3-fold

TF-TG prediction



2-fold

Outline

1 Introduction

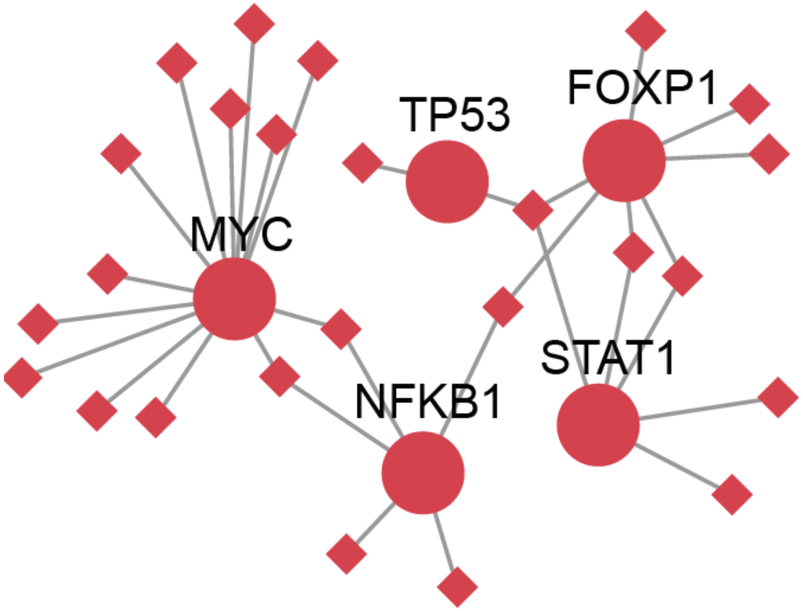
2 LINGER Model

3 Validation

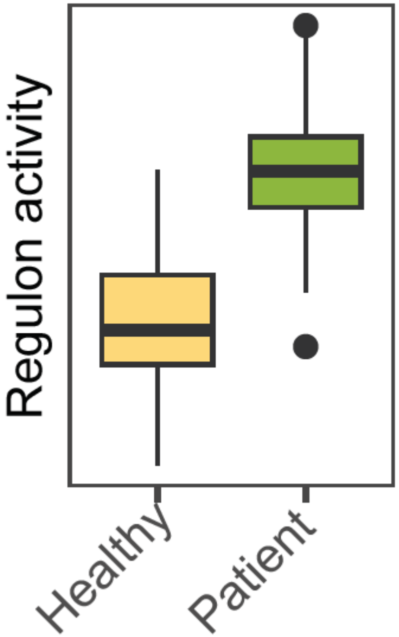
4 Application

Regulon activity

Regulon: Target genes of a given TF

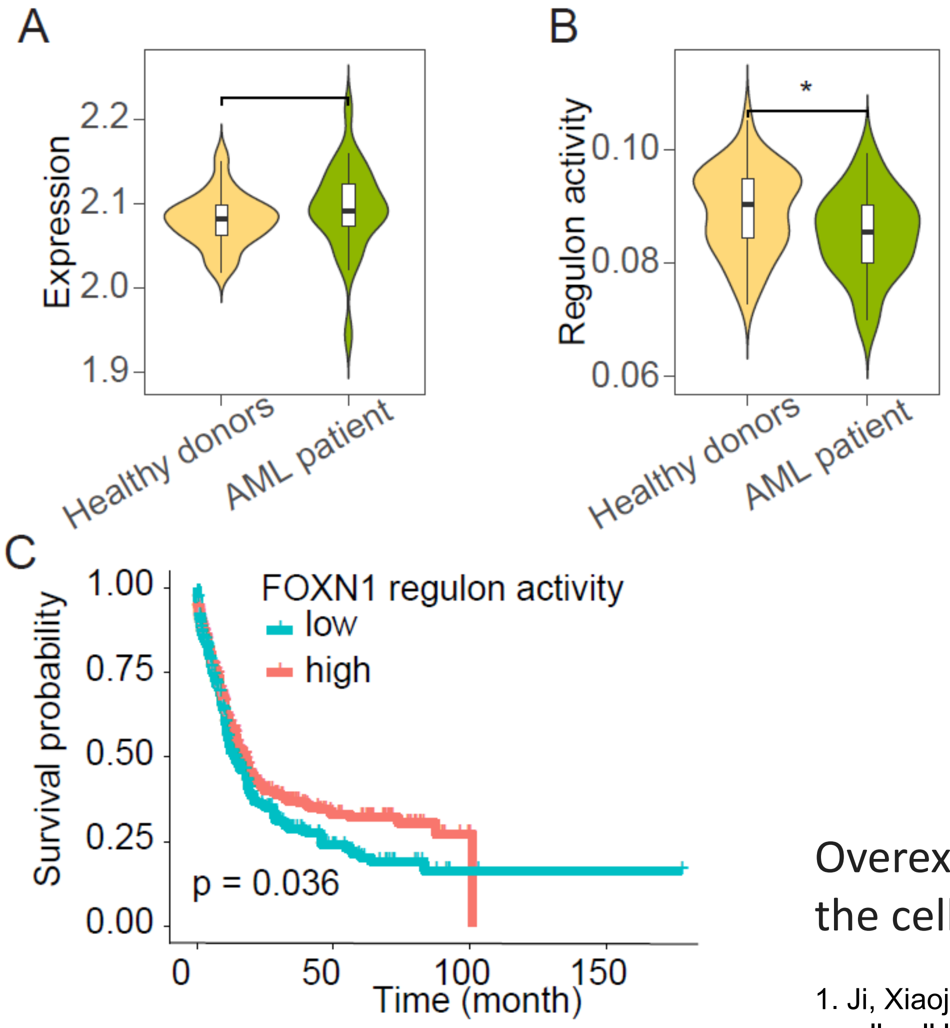


Regulon Activity for each sample: relative expression of target genes.



Input: GRN and gene expression data for each individual
Output: Regulon activity for each TF for each individual

Identify FOXN1 as a key regulator of Acute Myeloid Leukemia



Data: bulk microarray gene expression data
(38 healthy vs 26 AML)
Results: No expression change in FOXN1

But observe Regulon activity change.

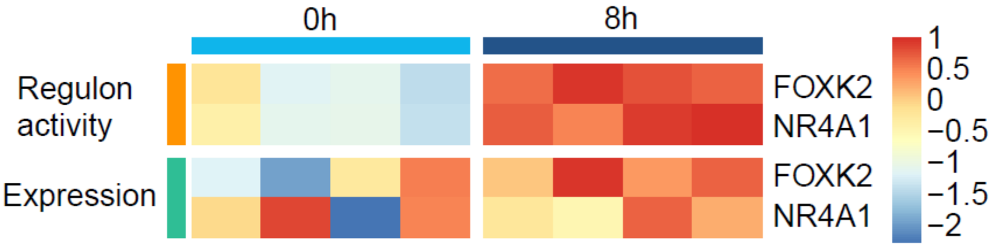
Protect from AML.

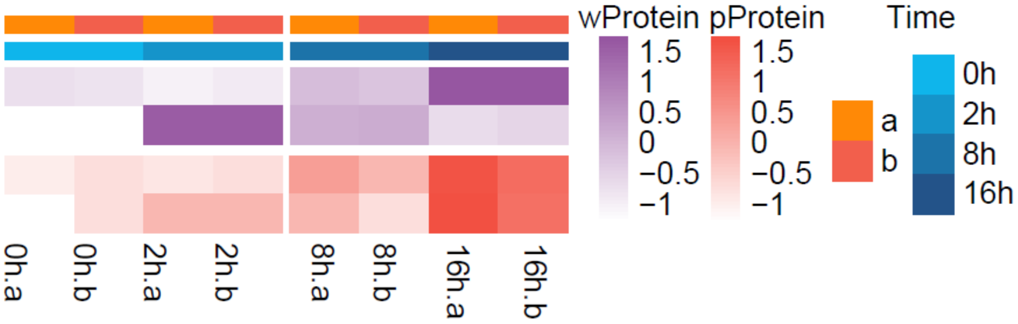
Validated by Survival analysis

Overexpression of FOXN1 evidently suppressed the cell growth (1).

1. Ji, Xiaojian, et al. "Forkhead box N1 inhibits the progression of non-small cell lung cancer and serves as a tumor suppressor." Oncology letters 15.5 (2018): 7221-7230.

FOXK2/NR4A1 respond to TCR stimulation in CD4T cells





Data: bulk RNA-seq data TCR stimulation at 8 h vs. 0 h.

Results: No significant expression change in FOXK2/NR4A1

But observe Regulon activity change.

Validated by whole protein and Phosphoproteome data.

Evidence of NR4A1 (1): transgenic mice expressing GFP, GFP was up-regulated after TCR stimulation.

1. Moran, Amy E., et al. "T cell receptor signal strength in Treg and iNKT cell development demonstrated by a novel fluorescent reporter mouse." *Journal of Experimental Medicine* 208.6 (2011): 1279-1289.

Summary

Modelling of GRN by Lifelong learning

Inconsistency between gene expression fitness and GRN accuracy.

Validation of GRNs

Regulon activity comparison reveal key regulators in disease/healthy and stimulation studies.

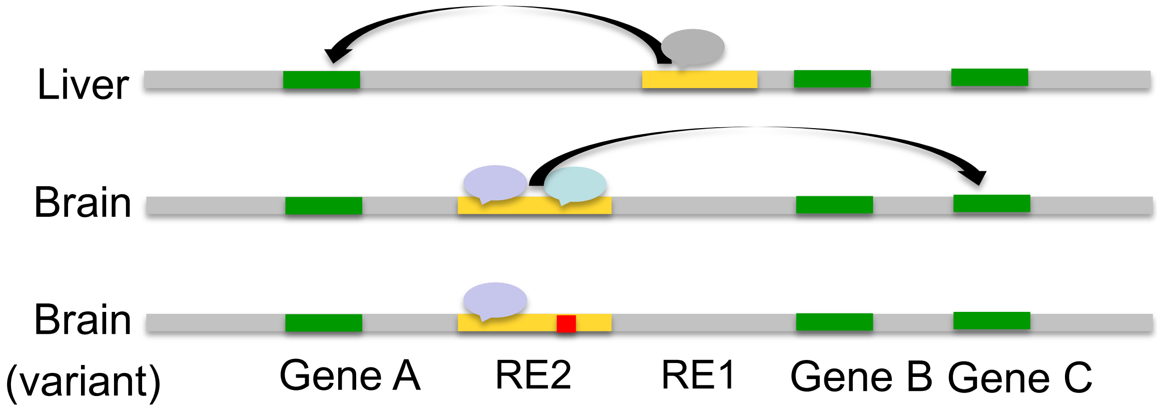
Acknowledgement

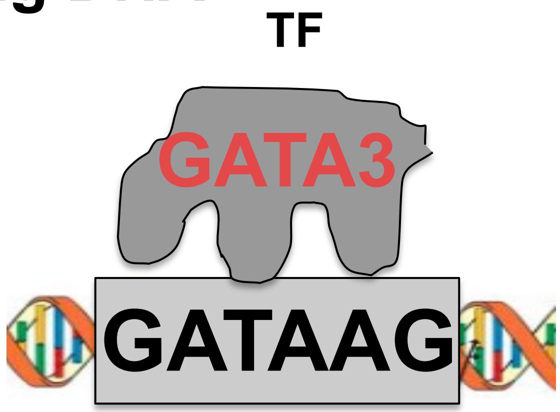
Credit:
Significant contribution from Dr. Qiuyue Yuan (postdoc in Duren lab)

Funding:
NIH COBRE grant: P20 GM139769

CLEMSON® UNIVERSITY
CENTER FOR
HUMAN GENETICS

Regulatory mechanism of non-coding DNA





1, Which gene?

2, In which cellular contexts?

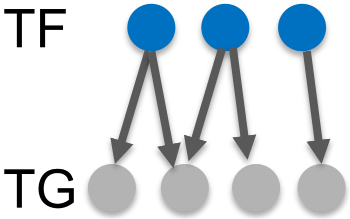
3, Which TF ?

4, How does a mutation affect?

Gene regulatory network inference from genomics data

★ Gene expression

1995 microarray
2008 RNA-seq
2009 scRNA-seq

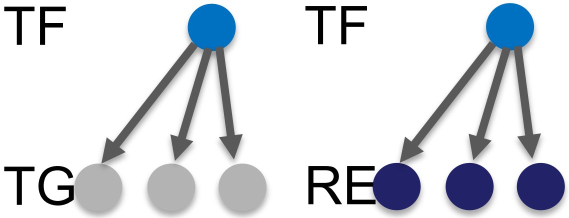


Not include RE

Context non-specific

Protein-DNA

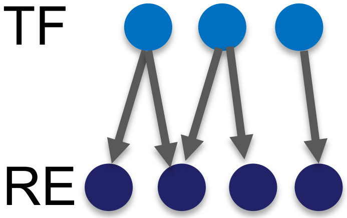
2000 ChIP-chip
2008 ChIP-seq



Single TF

★ Chromatin accessibility

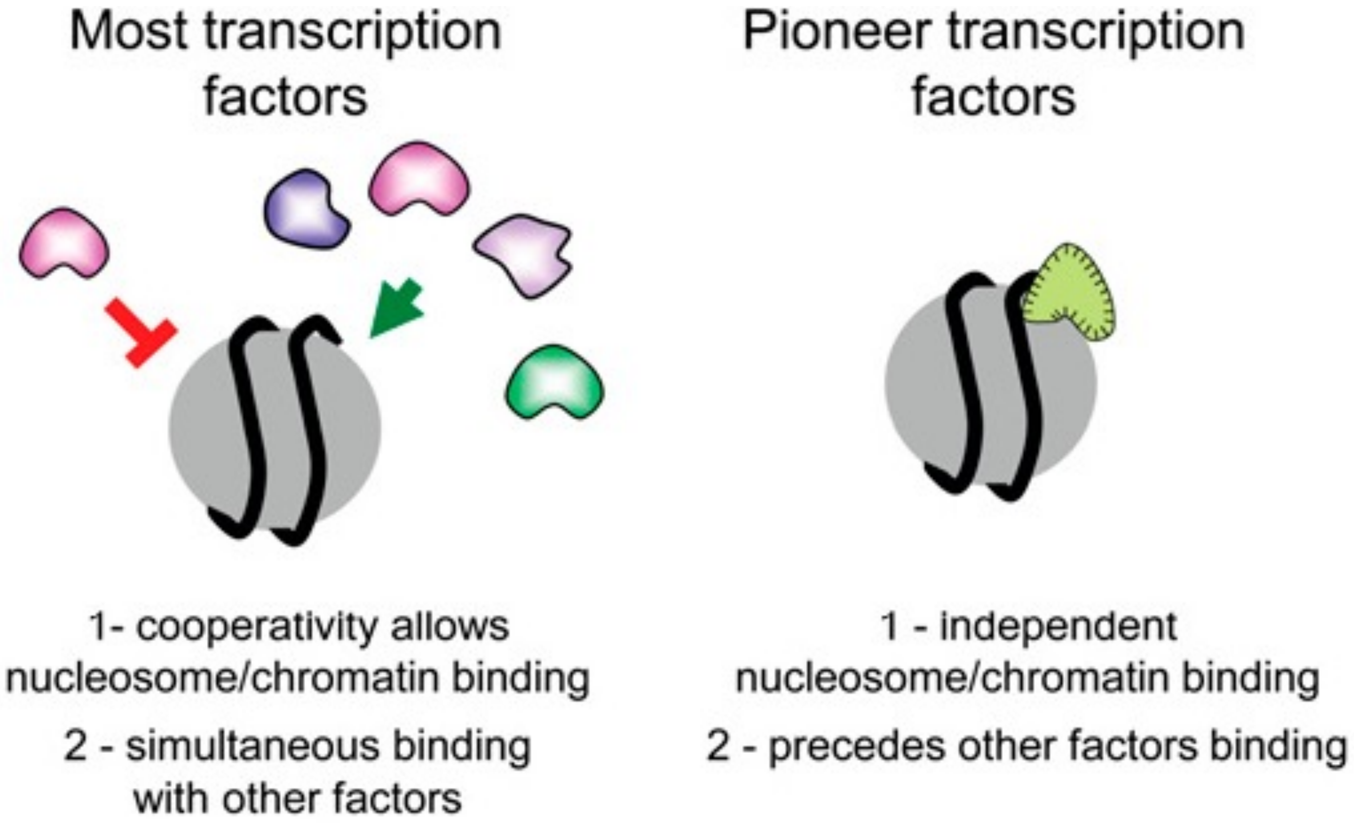
2008 DNase-seq
2013 ATAC-seq
2015 scATAC-seq



Not include TG

★ **Data we use**

How TFs bind to closed chromatin?



Most transcription factors

1- cooperativity allows nucleosome/chromatin binding

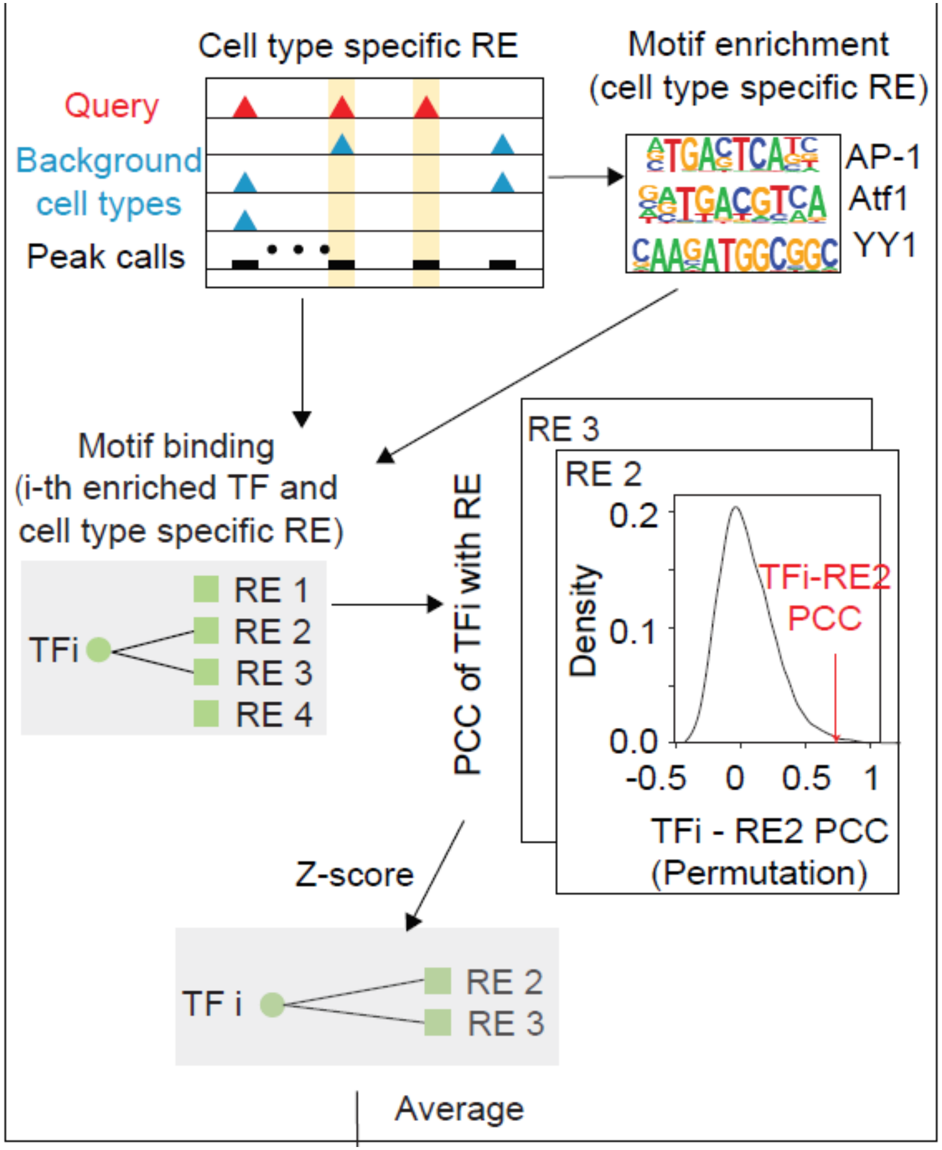
2 - simultaneous binding with other factors

Pioneer transcription factors

1 - independent nucleosome/chromatin binding

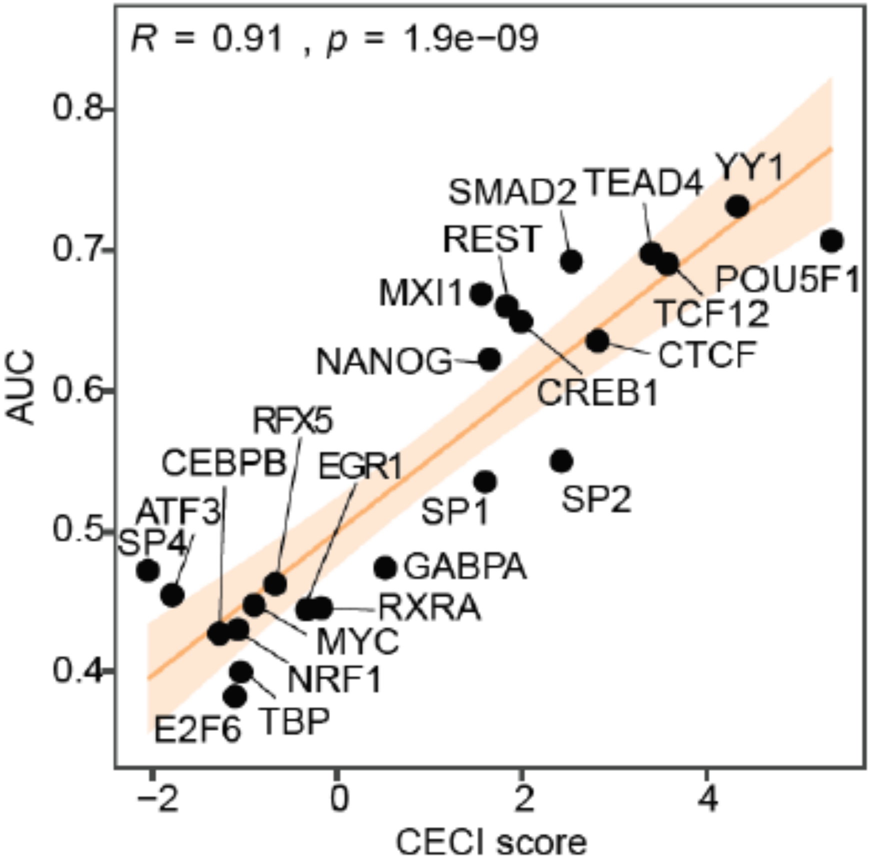
2 - precedes other factors binding

TFs Controlling Epigenetic Cell Identity (CECI)



Pioneer factors:
TF expression are correlated to the binding REs.

The CECI score captures pioneer TFs

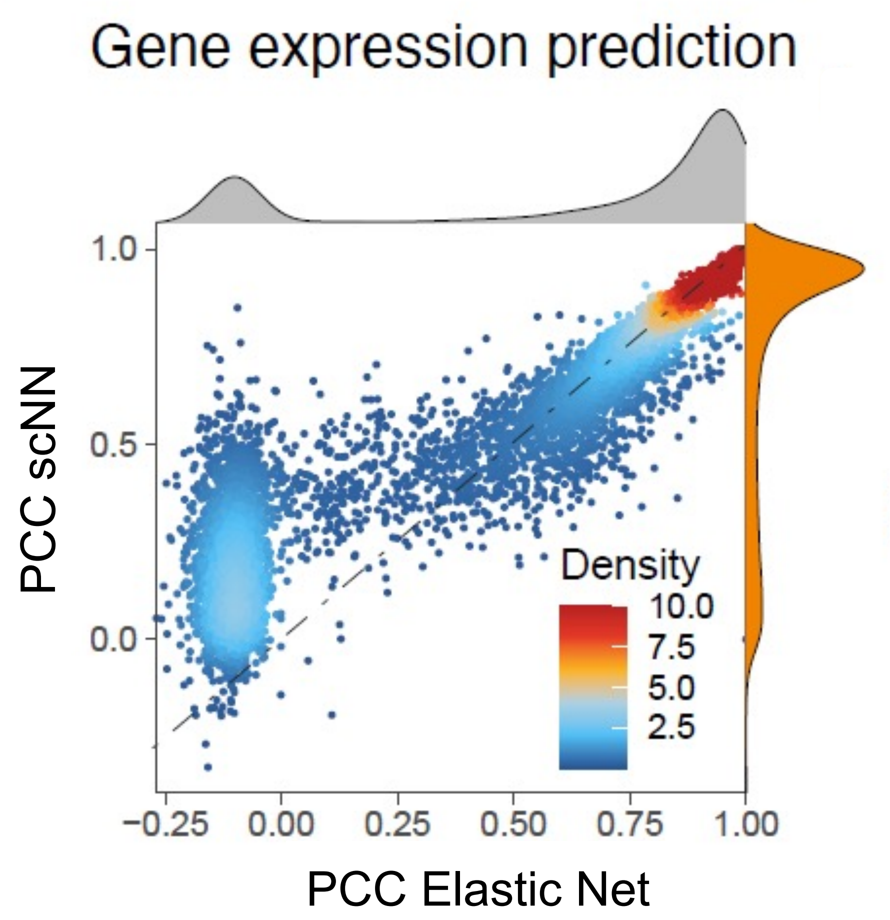


AUC: AUC of TF-RE binding prediction from TF-RE correlation taking the ChIP-seq data as ground truth.

High AUC means pioneer TFs

CECI: based on TF-RE correlation and motif binding information.

Neural network improve TG expression predictions



Data: PBMC sc-multiome data

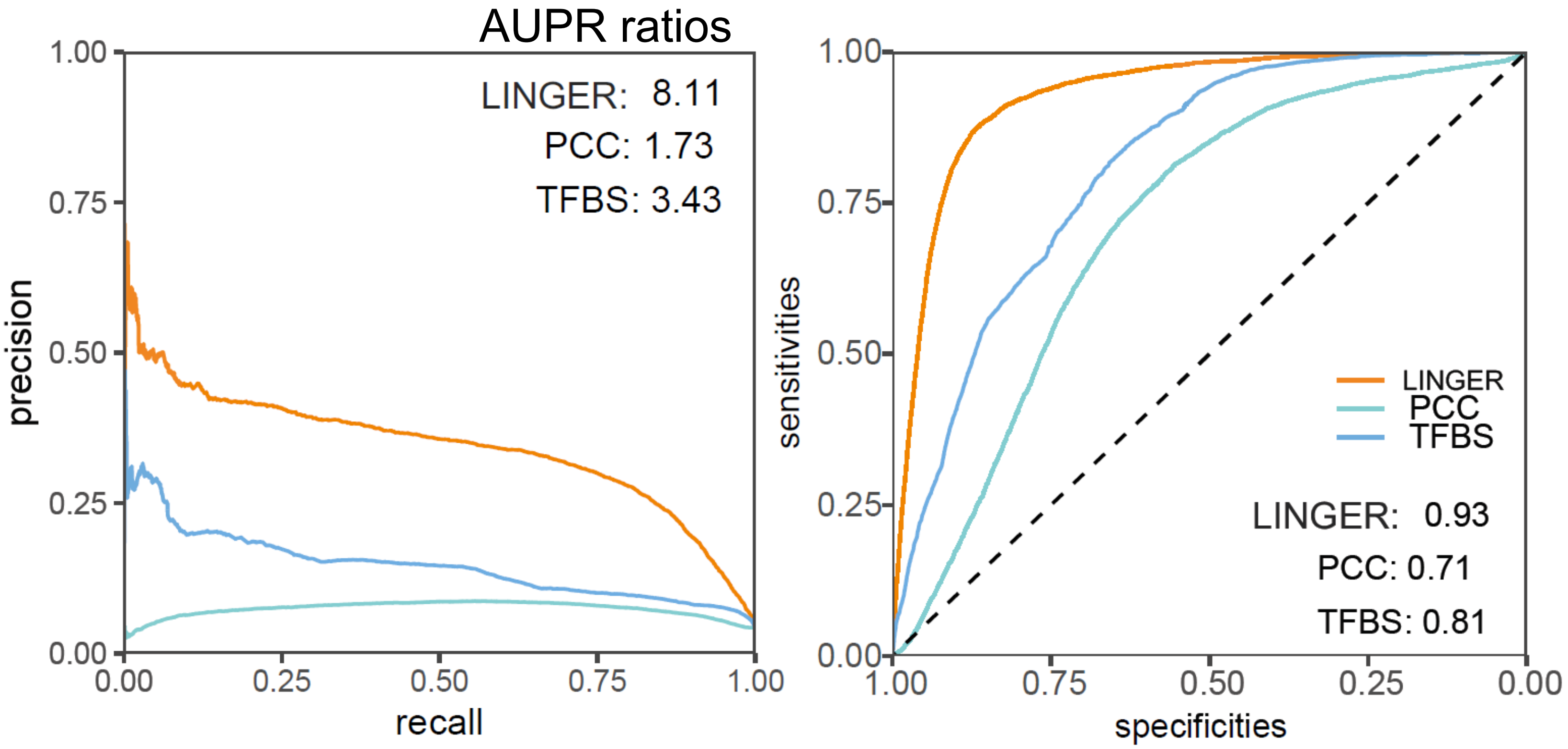
Metric: Pearson Correlation of observed versus predicted gene expression

Dot: one gene

5-fold cross validation

LINGER improve cell type specific TF binding

AUPR and AUROC: MYC ChIP-seq on naïve B cells

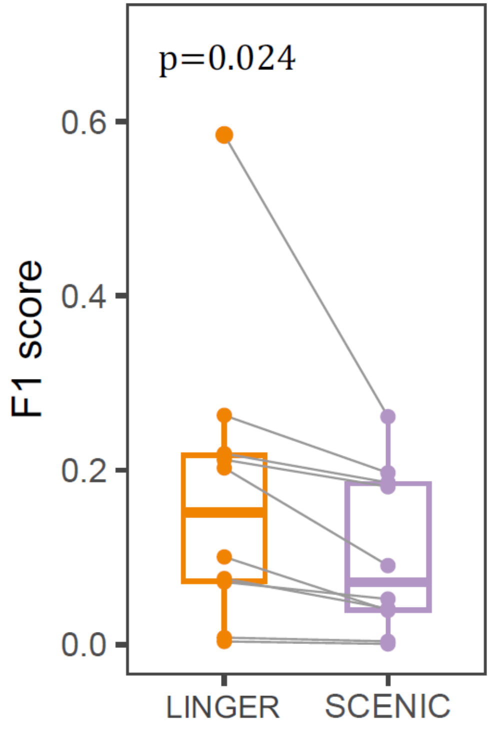


REs overlapped with ChIP-seq peaks are considered as positive, and others are negative.

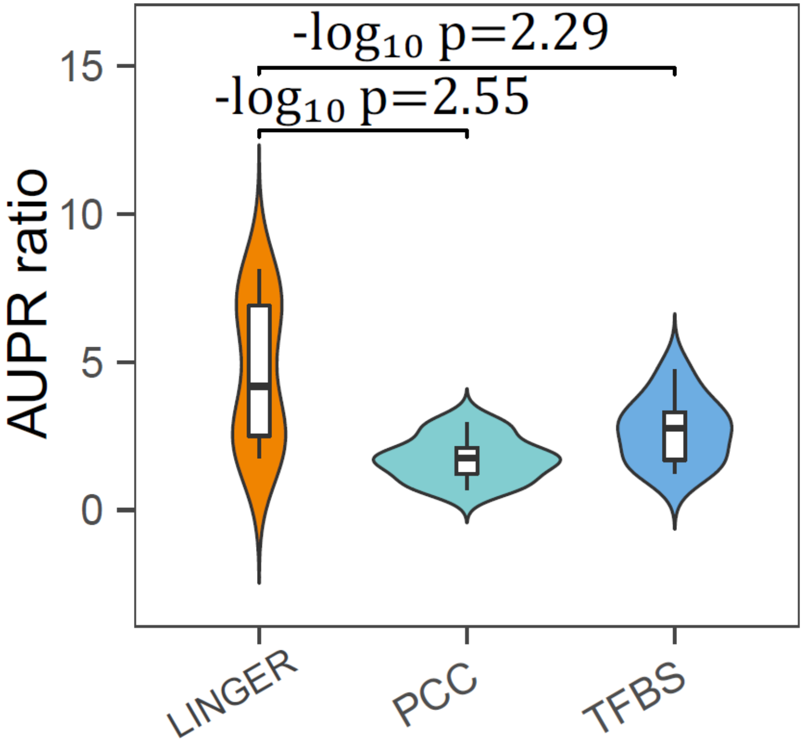
TFBS: motif scanning in open region.
PCC: TF-RE correlation.

LINGER improve cell type specific TF binding

LINGER vs SCENIC (ChIP-seq of 10 TFs)



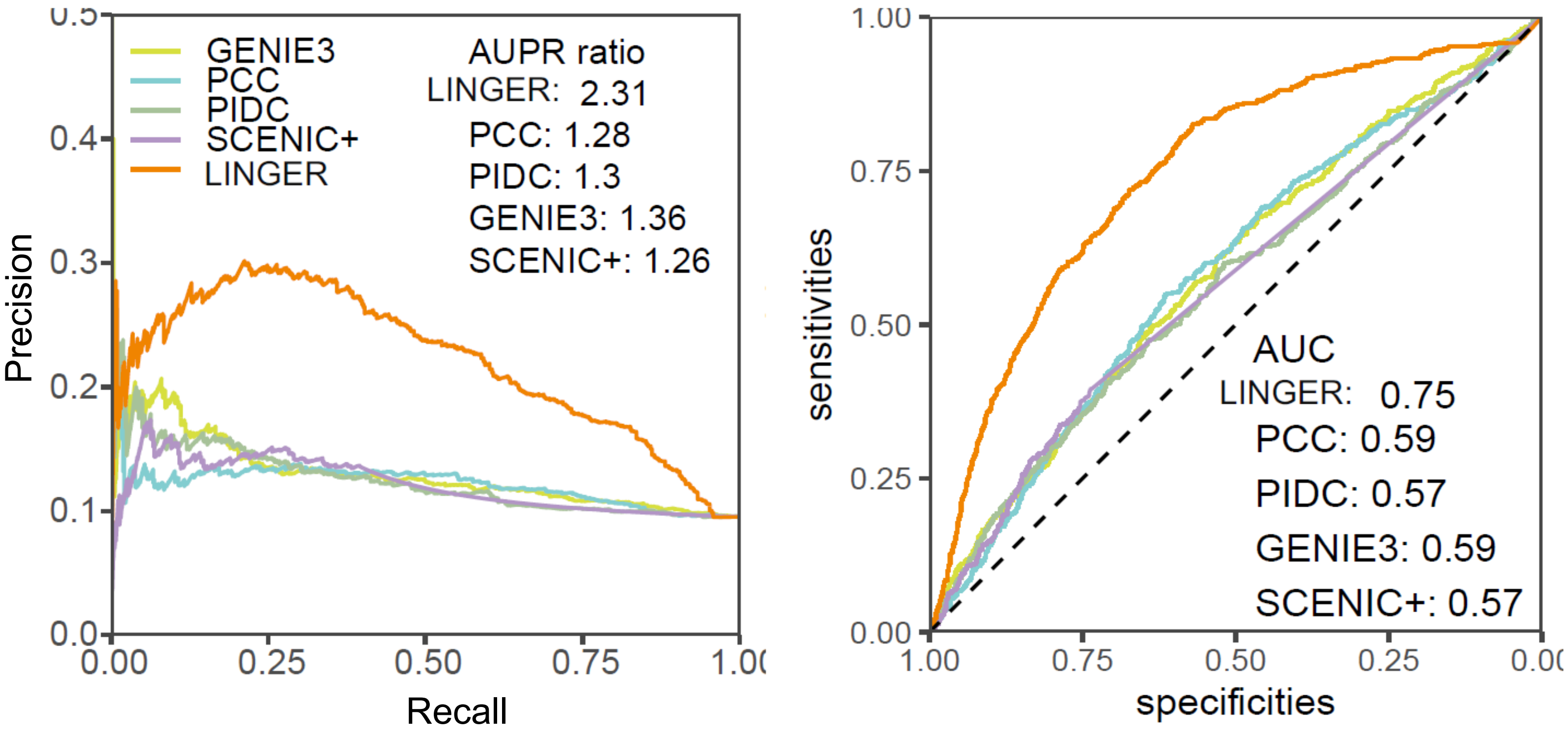
LINGER vs others (ChIP-seq of 10 TFs)



SCENIC. Sara Aibar,…..Stein Aerts. Nature method 2017.
SCENIC+ BioRxiv 2022.

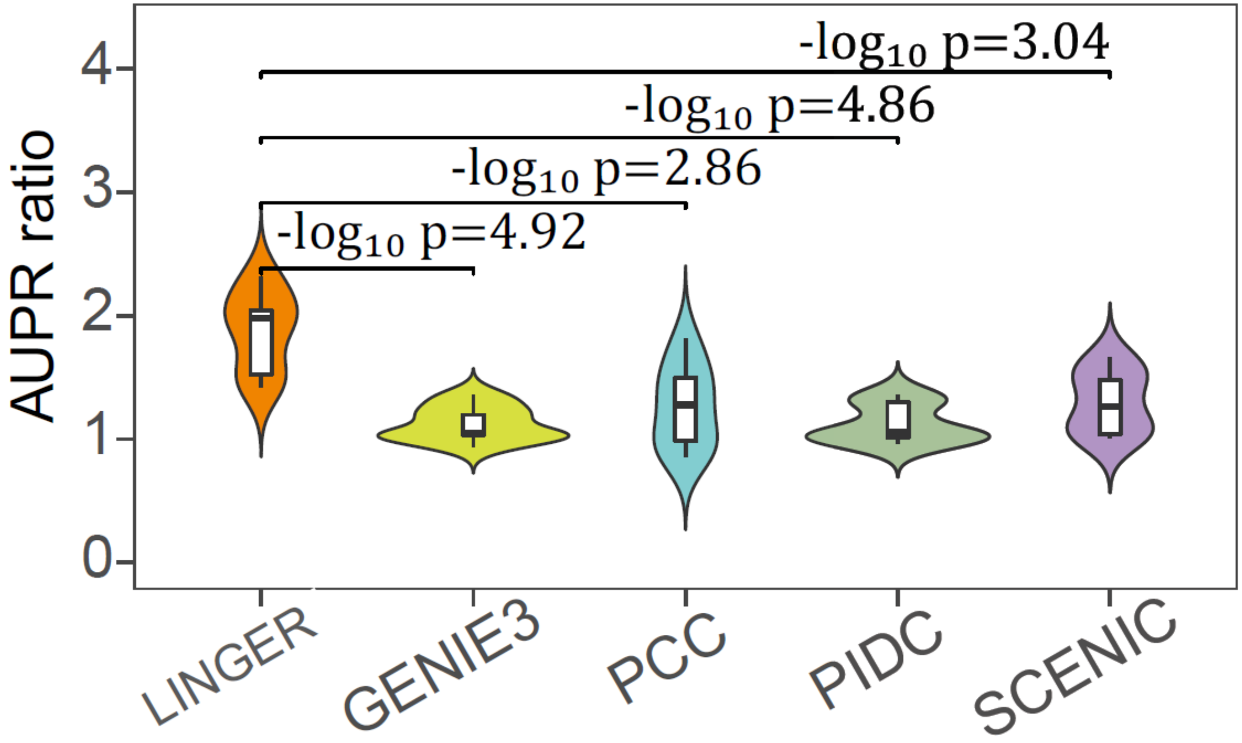
LINGER improve cell type specific TF-TG inference

AUPR and AUROC: STAT1 ChIP-seq on classic monocytes



LINGER improve cell type specific TF-TG inference

ChIP-seq data of 10 TFs from PBMC



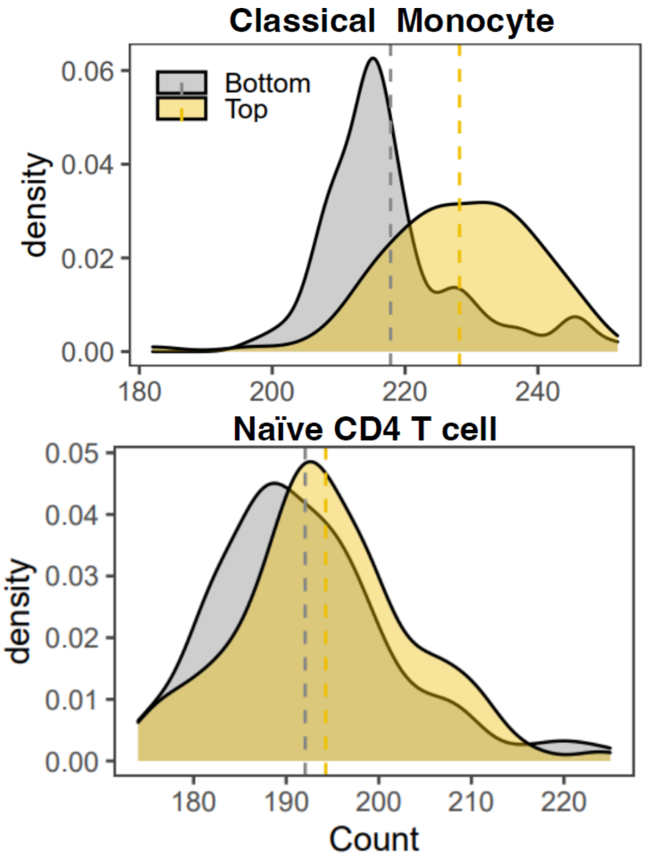
Application of inflammatory Bowel Disease (IBD)

Data used:

- List of Disease Associated Genes (DAG) from GWAS summary statistics
- PBMC sc-multiome data from healthy donor
- List of differential expressed genes between diseased and healthy in PBMC bulk data.

Which cell types are relevant to IBD?

Hypothesis: DAGs should be enriched in TGs of top CECI TFs in disease relevant cell types.



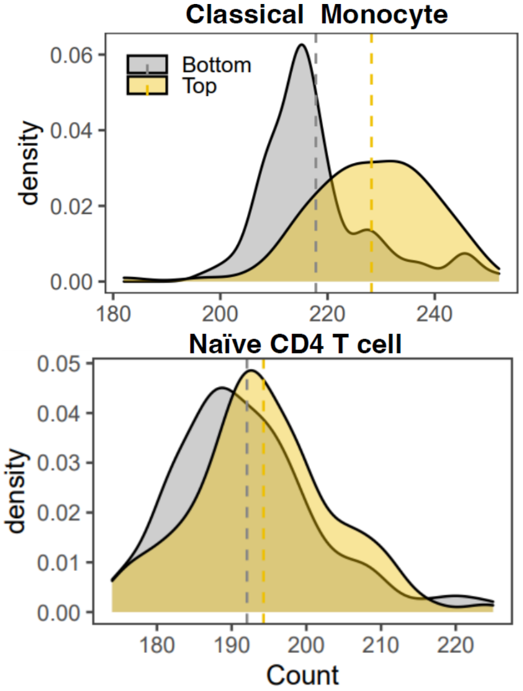
Top/Bottom: The highest/lowest 100 TFs based on the CECI score.

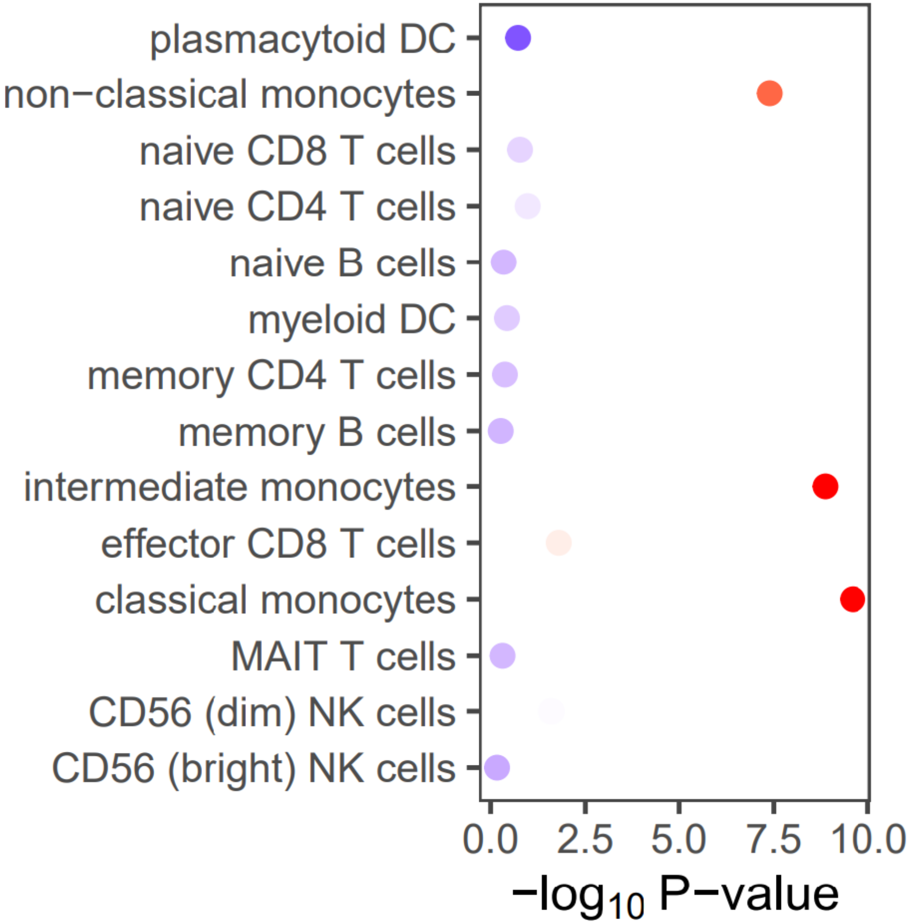
Count: # of DAGs in top 1000 TGs.

DAGs are enriched in TGs of top CECI TFs in disease relevant cell types

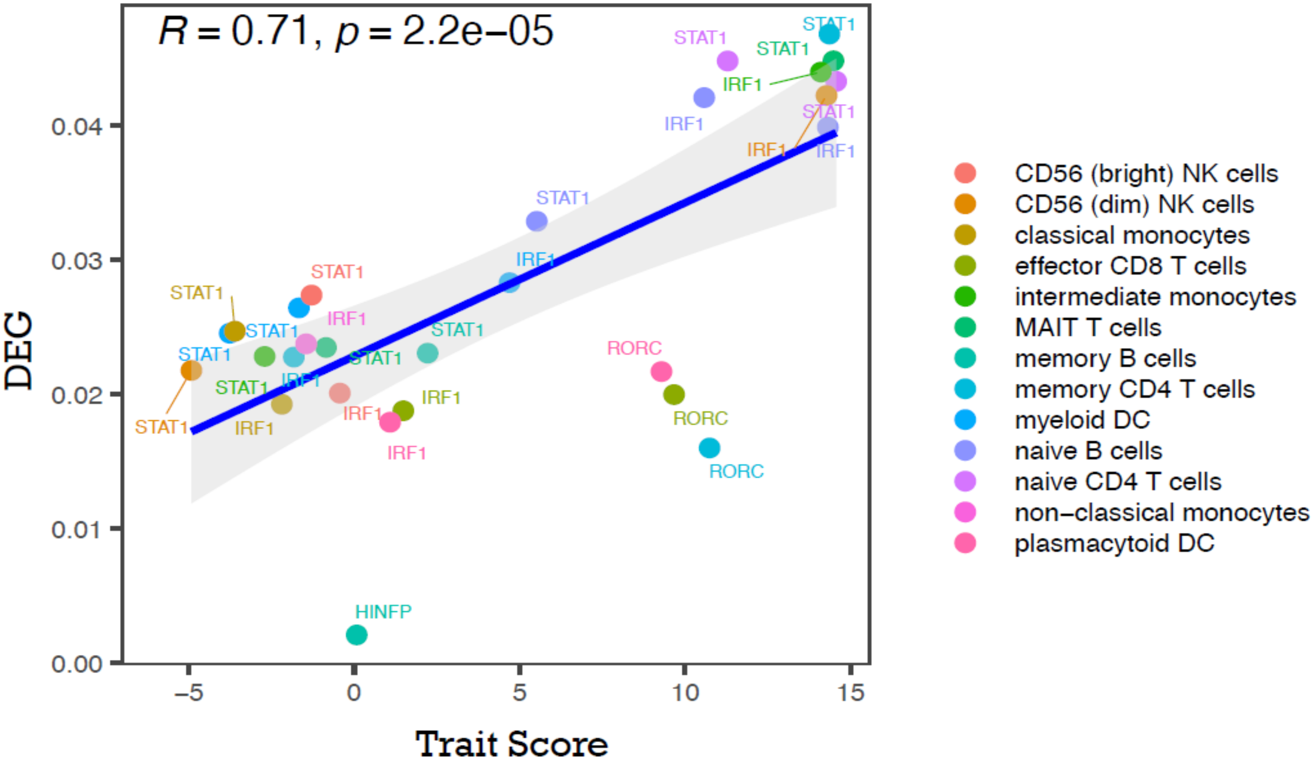
Top/Bottom: The highest/lowest 100 TFs based on the CECI score.

Count: # of DAGs in top 1000 TGs.





TGs of pioneer TFs in right cell types are differentially expressed



Trait score: CECI score + Z-score(Count)
Only show TFs with genome wide significant SNPs