Dissecting the Gene Regulatory Grammar at a Single-cell Resolution

Jing Zhang, Ph.D.
Assistant Professor
Department of Computer Science
University of California, Irvine

Email: zhang.jing@uci.edu
Website: https://www.ics.uci.edu/~jingz31/
**Research Background**: Understand Gene Regulation and Link Dysregulations to Disease

**Goal**: link **variants** to **disease**

- **Key questions**
  - Which **cell types** contribute most?
  - How to identify key dysregulation at:
    - **variant** level
    - **regulatory element** level
    - **network** level
    - **micro-environment** level

**Approach**: Computational modeling to link **dysregulation** to **disease**

- **scATAC-seq**
- **scRNA-seq**
- **scHi-C**

Within a cell

Within a microenvironment

**Efficient Clustering**

- **ForestFire** *(Nat. Comm. 22’)*
- **SAILER** *(ISMB 21’)*
- **Translator** *(JCB 21’)*
- **DeepVelo** *(Sci. Adv. 23’)*
- **Venus** *(Plos CB. 22’)*
- **SCAN-ATAC-sim** *(bioinfo. 20’)*

**Cell Embedding**

- **Cell fate decision**
- **Viral detection**

**Cross-modality generation Perturbation**

**GRN inference Network Comparison**

**Imputation Segmentation**

**SAILERX** *(NAR 22’)*

**DirectNet** *(Sci. Adv. 22’)*

**SCAN-IT** *(BMVC. 21’)*

**Impeller** *(under-review’)*

**EX-ADGNN** *(APSIPA, 23’)*

**GRN inference Network Comparison**

**Image adapted from https://doi.org/10.1002/glia.24343**
scATAC-seq

Bins: regions in the genome (m>100K)

- **Sparse:** ~ 90% to 99% of data are 0s
- **Noisy:**
  - 0s include true 0s and missing data (majority)
  - 1s mean open or technical noise
- **Nonlinear:** lots of long-distance nonlinear interactions across the genome

Goal: learn robust & biologically meaningful cell representations

**Confounding factor 1: batch effect**

**Confounding factor 2: gender, age, etc.**
Variational Autoencoders (VAE)
- Desired $z$: cell state/cell type
- Real $z$: cell state/cell type + confounding factors

Invariant representation learning
- To learn a representation $z$ independent of confounding factor $c$

**Objective**
$$L_{VAE} + \lambda I(z, c)$$

$I(z, c)$ is **mutual information** penalty between latent representation and confounding factors

Cao et. al, ISMB, 2021
The Zhang Lab

Table 2: Evaluation results on the mouse atlas dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>ARI</th>
<th>NMI</th>
<th>$I(z,c)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAILER</td>
<td>0.575</td>
<td>0.799</td>
<td>0.040</td>
</tr>
<tr>
<td>SnapATAC</td>
<td>0.538</td>
<td>0.748</td>
<td>0.127</td>
</tr>
<tr>
<td>SCALE</td>
<td>0.315</td>
<td>0.557</td>
<td>0.279</td>
</tr>
</tbody>
</table>

Mouse Atlas Data (81,173 adult mouse cells)

Merging two mouse brain datasets (batch effect)

Cao et. al, ISMB, 2021
SAILERX: noise-aware single cell multiome analysis

Modeling Intuition:

• Using scRNA-seq as the **baseline** due to its “higher” sensitivity

• Encourage similar **cell-to-cell similarities** from both modalities

• Perform **batch-wise** C2C alignment for large-scale analysis

• **Weight** each cell by $\delta$ for the ATAC-seq modality

• **Invariant** representation learning for batch effect correction

Cao et. al, NAR, 2022
SAILERX: noise-aware single cell multiome analysis

Cao et. al, NAR, 2022
**Question:** Can we generate realistic genomic profiles using known modality?

- Cross-modality generation using **cyclic training**
- Using **neural networks** to map cell embedding in the latent space
- Use **adversarial training** via weakly linked features

### Diagram:

**Cells from scRNA-seq**

**Generated scRNA-seq from scATAC-seq**

---

**scCarpo: Cross Modality Generation via Cyclic Training**

Liu et. al., Under-review
Cell Type Prioritization for Different Diseases using Single-cell Epigenome Data

105 samples, 935,371 cells
38 CON, 32 MDD, 35 PTSD
**Research Background: Understand Gene Regulation and Link Dysregulations to Disease**

**Goal:** link **variants** to disease

- Within a cell
- Within a microenvironment

**Key questions**

- Which **cell types** contributes most?
- How to identify key dysregulation at
  - **variant** level
  - regulatory **element** level
  - **network** level
  - micro-**environment** level

**Approach:** Computational modeling to link **dysregulation** to disease

- scATAC-seq
- scRNA-seq
- scHi-C
- Chromatin Accessibility
- 3D structure
- Transcription
- TF binding
- Methylation

**Approach Diagram**

- **ForestFire** (Nat. Comm. 22’)
- **SAILER** (ISMB 21’)
- **Translator** (JCB 21’)
- **Venus** (Plos CB. 22’)
- **DeepVelo** (Sci. Adv. 23’)
- **SCAN-IT** (BMVC. 21’)
- **Venus** (Plos CB, 23’)
- **iHerd** (Plos CB, 23’)
- **EX-ADGNN** (APSIPA, 23’)
- **SAILERX** (NAR 22’)
- **DirectNet** (Sci. Adv. 22’)
- **SCANT-ATAC-sim** (bioinfo. 20’)
- **Impeller** (under review’)
- **Cell Embedding**
- **Cell fate decision**
- **Cross-modality generation**
- **GRN inference**
- **Gene Expression**
- **Data alignment**
- **Viral detection**
- **Network Comparison**
- **Imputation Segmentation**

**Image adapted from [https://doi.org/10.1002/glia.24343](https://doi.org/10.1002/glia.24343)**
**Key Question:** Cis-regulatory element identification

Identify noncoding regions (scATAC-seq) that regulate mRNA expression changes (scRNA-seq)

### Computational Challenges

- No additional functional genomics data in most cases
- Both $X$ and $Y$ are very sparse
- Multiple $x$ regulate $y$
- Regulation relationship is usually nonlinear (e.g., gene regulation redundancies)

<table>
<thead>
<tr>
<th>Genomic Regions (&gt;200k)</th>
<th>Cells</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1 0 0 0 1 1 0 0 0 ...</td>
</tr>
<tr>
<td></td>
<td>1 1 1 0 0 1 1 0 0 0 ...</td>
</tr>
<tr>
<td></td>
<td>0 0 0 0 0 0 0 0 1 1 ...</td>
</tr>
<tr>
<td></td>
<td>1 0 0 0 0 0 0 0 0 0 ...</td>
</tr>
<tr>
<td></td>
<td>0 0 0 1 0 0 0 1 0 0 ...</td>
</tr>
<tr>
<td></td>
<td>0 0 0 0 1 0 1 0 0 1 ...</td>
</tr>
<tr>
<td></td>
<td>0 1 0 0 0 1 1 0 0 0 ...</td>
</tr>
<tr>
<td></td>
<td>0 0 0 0 0 0 0 0 0 0 ...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genes (&gt;20k)</th>
<th>Cells</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 0 1</td>
</tr>
<tr>
<td></td>
<td>0 8 0 0 ...</td>
</tr>
<tr>
<td></td>
<td>0 0 0 1 ...</td>
</tr>
<tr>
<td></td>
<td>10 0 1 5 ...</td>
</tr>
<tr>
<td></td>
<td>5 0 2 3 ...</td>
</tr>
<tr>
<td></td>
<td>0 3 0 0 ...</td>
</tr>
<tr>
<td></td>
<td>5 0 0 0 ...</td>
</tr>
<tr>
<td></td>
<td>1 0 0 10 ...</td>
</tr>
</tbody>
</table>

$X$: scATAC-seq, chromatin accessibility

$Y$: scRNA-seq, mRNA amount

**DIRECT-NET:** Dissect Cis-regulatory Grammar Using Single-cell Multiomic Data

Zhang et. al., Science Advances, 2022
DIRECT-NET: Dissect Cis-regulatory Grammar Using Single-cell Multiomic Data

DIRECT-NET: an efficient machine learning method to discover cis-regulatory elements from single cell multi-omics data

Add distance constraint
Aggregation to overcome data sparsity
Nonlinear predictive model
Regulatory Networks

\[ y \sim f(x_2, x_3, x_4) \]

Gradient Boosting Machine

Zhang et. al., Science Advances, 2022
DIRECT-NET: Dissect Cis-regulatory Grammar Using Single-cell Multiomic Data

High-conf CREs: HF strong impacts to any gene’s expression
Middle-conf. CREs: limited impacts to any gene’s expression
Non-CREs: no impacts to any gene’s expression

High confidence CREs are highly cell-type-specific

Zhang et. al., Science Advances, 2022
DIRECT-NET: Dissect Cis-regulatory Grammar Using Single-cell Multiomic Data

Independent Validation 1:
- H3k27ac ChIP-seq signals indicate functional elements in the human genome
- True functional elements should overlap with cell-type-specific ChIP-seq signals

High confidence CREs overlap best with cell-type-matched ChIP-seq data
DIRECT-NET: Dissect Cis-regulatory Grammar Using Single-cell Multiomic Data

**Independent Validation 2:**

- CRE($X$)-gene($Y$) linkages using cell-type-specific Hi-C
- **Fraction** of predicted linkage by validated Hi-C
- strictly controlled by CRE($X$)-gene($Y$) distances

Zhang et. al., *Science Advances*, 2022
DIRECT-NET: Dissect Cis-regulatory Grammar Using Single-cell Multiomic Data

- snATAC-seq data of 70631 individual nuclei
- 7 brain regions of adult brain
- 18 distinct clusters

- FEV is required for both development and function of serotonergic neurons
- FEV is a TF associated with Autism
- FEV is highly expressed in striatal inhibitory 2 and regulated key markers genes
- Autism SNPs are located near one HF CRE in Striatal inhibitory 2

Autism Associated Variants

- Enrichment of SNPs in Striatal inhibitory 1 and Striatal inhibitory 2
- Autistic SNPs are located near one HF CRE in Striatal inhibitory 2

Zhang et al., Science Advances, 2022
**Goal:** link **variants** to **disease**

**Key questions**

Which **cell types** contributes most? How to identify key dysregulation at

- **variant** level
- regulatory **element** level
- **network** level
- **micro-environment** level

**Approach:** Computational modeling to link **dysregulation** to **disease**

**Within a cell**

- scATAC-seq
- scRNA-seq
- scHi-C
- methylations
- chromatin
- transcription
- TF binding
- 3D structure

**Within a microenvironment**

- spatial-omics
- network

---

**Research Background:** Understand Gene Regulation and Link Dysregulations to Disease

---

**ForestFire** (Nat. Comm. 22’)

**SAILER** (ISMB 21’)
Translating
JCB 21’

**Translator**

**SCAN-ATAC-sim**

**Bioinfo. 20’**

**Efficient Clustering**

**Cell Embedding**

**Data alignment**

**Transfer learning**

**Cell fate decision**

**Viral detection**

---

**DeepVelo** (Sci. Adv. 23’)

**Venus** (Plos CB. 22’)

**SAILERX** (NAR 22’)

**scCarpo** (under review)

**scGust** (under review)

**DirectNet** (Sci. Adv. 22’)

**iHerd** (Plos CB, 23’)

**EX-ADGNN** (APSIPA, 23’)

**SCAN-IT** (BMVC. 21’)

**Impeller** (under-review’)

---

**GRN inference**

**Network Comparison**

---

**Imputation**

**Segmentation**

---

**Goal:** link **variants** to **disease**

Which **cell types** contributes most? How to identify key dysregulation at

- **variant** level
- regulatory **element** level
- **network** level
- **micro-environment** level

---

**Approach:** Computational modeling to link **dysregulation** to **disease**

**Within a cell**

- scATAC-seq
- scRNA-seq
- scHi-C
- methylations
- chromatin
- transcription
- TF binding
- 3D structure

**Within a microenvironment**

- spatial-omics
- network

---

**Research Background:** Understand Gene Regulation and Link Dysregulations to Disease

---

**ForestFire** (Nat. Comm. 22’)

**SAILER** (ISMB 21’)
Translating
JCB 21’

**Translator**

**SCAN-ATAC-sim**

**Bioinfo. 20’**

**Efficient Clustering**

**Cell Embedding**

**Data alignment**

**Transfer learning**

**Cell fate decision**

**Viral detection**

---

**DeepVelo** (Sci. Adv. 23’)

**Venus** (Plos CB. 22’)

**SAILERX** (NAR 22’)

**scCarpo** (under review)

**scGust** (under review)

**DirectNet** (Sci. Adv. 22’)

**iHerd** (Plos CB, 23’)

**EX-ADGNN** (APSIPA, 23’)

**SCAN-IT** (BMVC. 21’)

**Impeller** (under-review’)

---

**GRN inference**

**Network Comparison**

---

**Imputation**

**Segmentation**

---
DeepVelo: Neural ODE to model cell transitions using single-cell RNA-seq data

**Observed:** snapshot of expression profiles

scRNA-seq, mRNA amount

<table>
<thead>
<tr>
<th>Genes (&gt;20k)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cells</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

Computational Goal:

- How to model gene regulatory relationships?
- How to model cell state transition dynamics?
DeepVelo: Neural ODE to model cell transitions using single-cell RNA-seq data

**Observed:** snapshot of expression profiles

<table>
<thead>
<tr>
<th>Cells</th>
<th>Genes (&gt;20k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 2 ...</td>
<td></td>
</tr>
<tr>
<td>0 8 0 0 ...</td>
<td></td>
</tr>
<tr>
<td>0 0 0 1 ...</td>
<td></td>
</tr>
<tr>
<td>10 0 1 5 ...</td>
<td></td>
</tr>
<tr>
<td>5 0 2 3 ...</td>
<td></td>
</tr>
<tr>
<td>0 3 0 0 ...</td>
<td></td>
</tr>
<tr>
<td>5 0 0 0 ...</td>
<td></td>
</tr>
<tr>
<td>1 0 0 10 ...</td>
<td></td>
</tr>
</tbody>
</table>

**scRNA-seq, mRNA amount**

**Computational Goal:**
- How to model gene regulatory relationships?
- How to model cell state transition dynamics?

**Existing solution:**
linear ordinary differential equations (ODE)

\[
\frac{dx}{dt} = Ax
\]

**Problems:**
- Nonlinear regulatory relationship?
- Out of sample predictions?
DeepVelo: Neural ODE to model cell transitions using single-cell RNA-seq data

- Using **VAE** to model **nonlinear** regulatory relationships
- Using **generative model** for out-of-sample predictions
DeepVelo: Neural ODE to model cell transitions using single-cell RNA-seq data

Modeling cell state transitions

\[ \vec{x}_{t+1} = \vec{x}_t + f_A(\vec{x}_t) \]

A. Developing mouse neocortex

- Observed RNA velocity

B. Latent layer

- Predicted cell state transitions

Chen et. al., Sci. Adv. 22'
Defining **cell criticality index** (CCI) to describe the instability of single-cell states

\[
\vec{x}_{t+1} = \vec{x}_t + f_A(\vec{x}_t) \quad \Rightarrow \quad CCI(\vec{x}) = \sum_{t=0}^{T} KL(\vec{x}_{i+1}^x \| \vec{x}_t^x)
\]
DeepVelo: Neural ODE to model cell transitions using single-cell RNA-seq data

DeepVelo can distinguish **driver** vs **passenger** genes for cell fate decisions

---

**Baseline control**

**Candidate genes**

**Perturbation Results**

- 7:3 ratio
- 6:4 ratio

---

Chen et al., Sci. Adv. 22'
Acknowledgement

Matthew Girgenti  Xiaohui Xie  Qing Nie

We are hiring postdocs and PhDs!

Siwei Xu  Ahyeon Hwang  Flynn Chen  Ziheng Duan  Cheyu Lee  Yingxin Cao  Laiyi Fu  Lihua Zhang

the Zhang Lab Members

NIH National Institute of Mental Health  NIH National Institute on Drug Abuse

NIH National Human Genome Research Institute  NIH National Institute of Neurological Disorders and Stroke

The Zhang Lab https://www.ics.uci.edu/~jingz31/