# Single-cell Developmental Trajectories Inference

Mengyu Li

Institute of Statistics and Big Data
Renmin University of China

April 24, 2022

# Outline

1. Introduction
2. Model
3. Results and Discussion

# Outline

1 Introduction

2 Model

3 Results and Discussion

## Brief Review

**Goal:** inferring cell developmental trajectories during reprogramming.

**Question:** given a cell at one time point, where will its descendants be at a later time point, and where are its ancestors at an earlier time point?

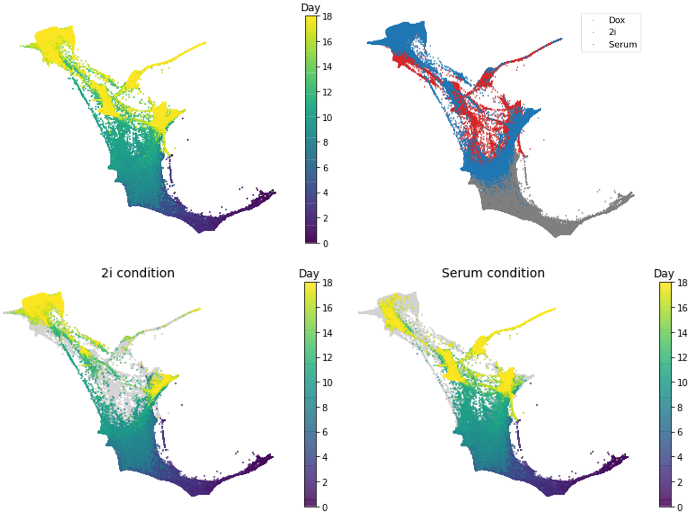
**Dataset:** scRNA-seq dataset collected across 18 days (39 time points) of reprogramming mouse embryonic fibroblasts (MEFs) into induced pluripotent stem (IPS) cells.

# Data Preparation

- Quality control analysis
  - cell-level filtering, gene-level filtering
- Exploratory data analysis
  - highly variable genes selection, dimensional reduction
- Data visualization
  - force-directed layout embedding

# Expression matrix Visualization

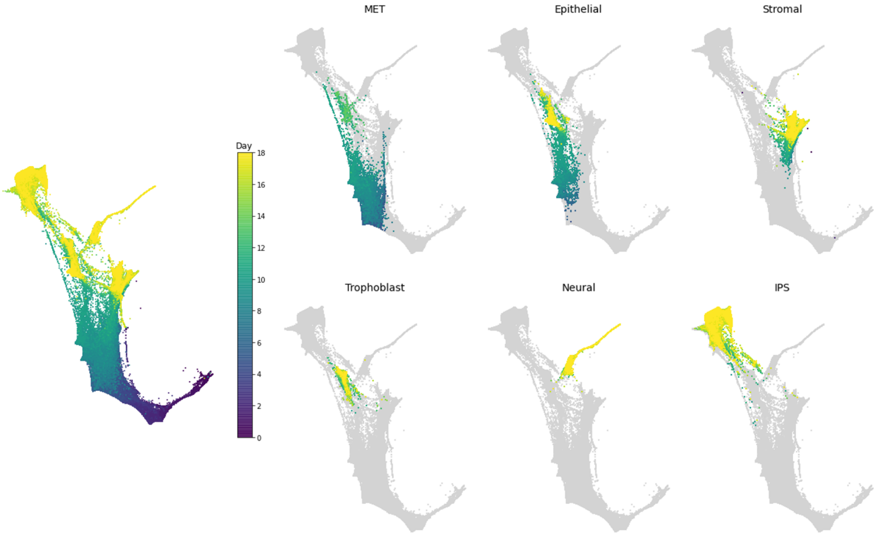
**Phase-1**: Dox; **Phase-2**: 2i and Serum.



2i condition

Serum condition

# Cell sets Visualization



# Outline

1 Introduction

2 Model

3 Results and Discussion

# Notation

- Developmental trajectory in gene expression space:

$$x : [0, T) \to \underbrace{\mathbb{R}^d \times \mathbb{R}^d \times \ldots \times \mathbb{R}^d}_{n(t) \text{ times}}.$$

- Expression profile of Cells at time $t$: $x(t) = \left(x_1(t), \ldots, x_{n(t)}(t)\right)$.
- Developmental process $\mathbb{P}_t$: a time-varying distribution (i.e. stochastic process) over trajectories.
- For example, the distribution of a set of cells $(x_1, \ldots, x_n)$ can be represented by

$$\mathbb{P} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}.$$

# Temporal coupling

A stochastic process is determined by its temporal dependence structure.

### Assumption

Cells don't change expression by large amounts over short time.
$\Longleftrightarrow$ Cells move short distances over short time periods.

Optimal transport can be used to find the coupling $\pi_{t_1,t_2}$ between $\mathbb{P}_{t_1}$ and $\mathbb{P}_{t_2}$ $(t_1 < t_2)$, i.e.,

$$\begin{aligned} \pi_{t_1,t_2} = \operatorname*{argmin}_{\pi} \iint \|x-y\|^2 \pi(x,y)dxdy \\ \text{s.t.} \int \pi(\cdot,y)dy = \mathbb{P}_{t_1} \\ \int \pi(x,\cdot)dx = \mathbb{P}_{t_2} \end{aligned}$$

# Temporal coupling (cont.)

**Modification 1: Account for growth.** Rescale the source distribution $\mathbb{P}_{t_1}$ using the relative growth rate $g(x)$:

$$\mathbb{Q}_{t_1}(x) = \mathbb{P}_{t_1}(x) \frac{g^{t_2 - t_1}(x)}{\int g^{t_2 - t_1}(z) d\mathbb{P}_{t_1}(z)}$$

**Modification 2: Relax the marginal constraints.**

$$\begin{aligned} \pi_{t_1,t_2} = \underset{\pi}{\operatorname{argmin}} & \iint \|x-y\|^2 \pi(x,y) dxdy \\ & + \lambda_1 \mathrm{KL}\left( \int \pi(\cdot, y) dy \| \mathbb{Q}_{t_1}(x) \right) + \lambda_2 \mathrm{KL}\left( \int \pi(x, \cdot) dx \| \mathbb{P}_{t_2}(y) \right) \end{aligned}$$

## Remark

- $\hat{g}(x)$ can be estimated by the output row-sums of $\hat{\pi}_{t_1,t_2}$.
- Take $\lambda_2 \gg \lambda_1$.

# Interpretation

Consider a set of cells $C \subset \mathbb{R}^d$ with $\mathbb{P}_{t_j}(x) = \begin{cases} \frac{1}{|C|} & x \in C, \\ 0 & \text{otherwise.} \end{cases}$

### Descendants

The descendants of $C$ at time $t_{j+1}$ are obtained by pulling $C$ through $\pi_{t_j,t_{j+1}}$, i.e., $\mathbb{P}_{t_{j+1}}^\top = \mathbb{P}_{t_j}^\top \pi_{t_j,t_{j+1}}$.

### Ancestors

The ancestors of $C$ at time $t_{j-1}$ are obtained by pulling $C$ back through $\pi_{t_{j-1},t_j}$, i.e., $\mathbb{P}_{t_{j-1}} = \pi_{t_{j-1},t_j}\mathbb{P}_{t_j}$.

### Trajectory

The trajectory of a cell set $C$ is the sequence of ancestor distributions at earlier time points and descendant distributions at later time points.

# Outline

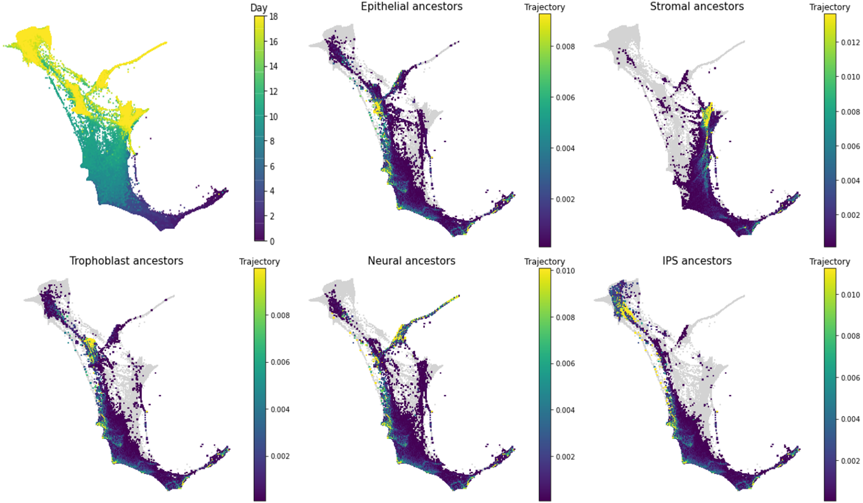
1 Introduction

2 Model

3 Results and Discussion

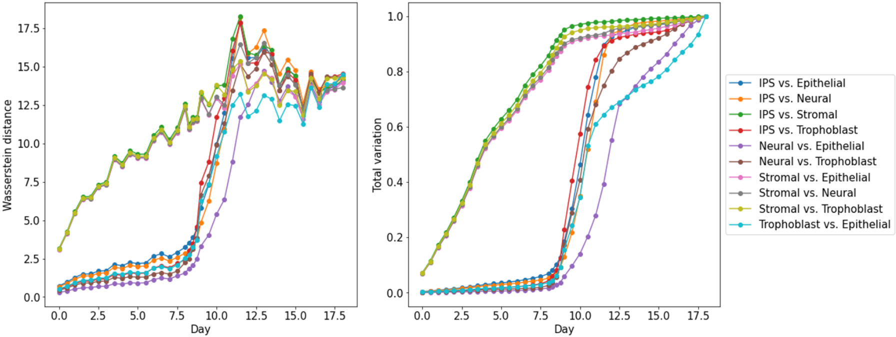
# Developmental Trajectory

Major cell sets at day 18 and their ancestors:



# Shared ancestors

For a pair of cell sets, whether they shared the same ancestry and when they diverged from a common set of ancestors?

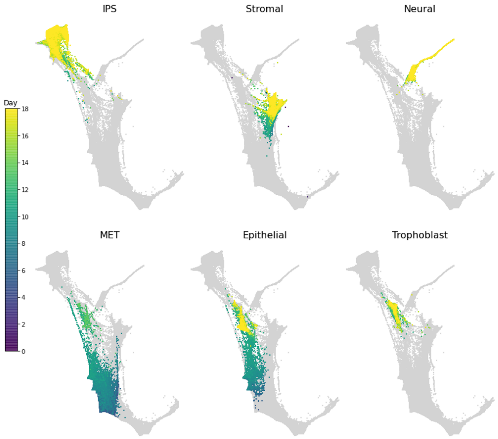


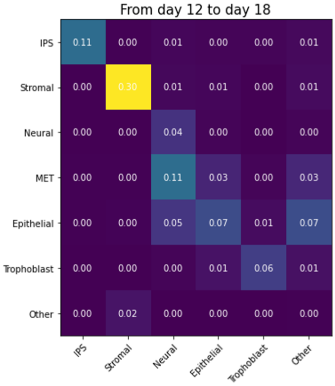
IPS, Epithelial, Neural and Trophoblast had the same ancestry, and they diverged from the start of Phase 2.

# Descendants: transition table

Consider cell sets $C_1, \ldots, C_m$ at time $t_k$ and cell sets $D_1, \ldots, D_n$ at time $t_{k+\Delta}$, then

$$\text{mass transported from } C_i \text{ to } D_j = \sum_{x \in C_i} \sum_{y \in D_j} \pi_{t_k, t_{k+\Delta}}(x, y).$$



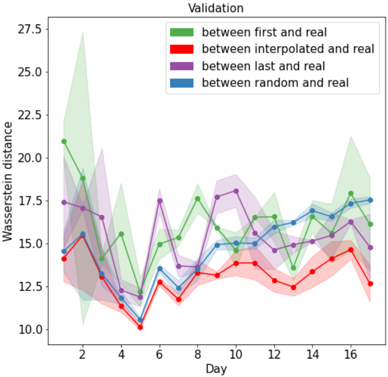

From day 12 to day 18

| | IPS | Stromal | Neural | Epithelial | Trophoblast | Other |
|---|---|---|---|---|---|---|
| IPS | 0.11 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 |
| Stromal | 0.00 | 0.30 | 0.01 | 0.01 | 0.00 | 0.01 |
| Neural | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 |
| MET | 0.00 | 0.00 | 0.11 | 0.03 | 0.00 | 0.03 |
| Epithelial | 0.00 | 0.00 | 0.05 | 0.07 | 0.01 | 0.07 |
| Trophoblast | 0.00 | 0.00 | 0.00 | 0.01 | 0.06 | 0.01 |
| Other | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 |

# Validation via Interpolation

Considering consecutive time points $(t_i, t_{i+1}, t_{i+2})$ with distributions $(\mathbb{P}_{t_i}, \mathbb{P}_{t_{i+1}}, \mathbb{P}_{t_{i+2}})$,

1. estimate the coupling $\pi_{t_i,t_{i+2}}$ between $t_i$ and $t_{i+2}$;
2. compute an interpolating distribution at time $t_{i+1}$, i.e., $\hat{\mathbb{P}}_{t_{i+1}}$;
3. compare $\hat{\mathbb{P}}_{t_{i+1}}$ to $\mathbb{P}_{t_{i+1}}$ by the Wasserstein distance.



# Discussion

### Summary

- Only a small subset of cells during reprogramming became IPS cells; while others mainly went to Epithelial, Neural, Trophoblast and Stromal cells.
- IPS, Epithelial, Neural and Trophoblast cells had the same ancestry.

From "linear paths" to "non-linear paths" ?

- Existing method: continuous normalizing flows + dynamic optimal transport.
- Nonparametric method, e.g. smoothing splines.

# References

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., ... & Lander, E. S. (2019)
Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming
*Cell*, 176(4), 928-943.

Tong, A., Huang, J., Wolf, G., Van Dijk, D., & Krishnaswamy, S. (2020)
Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics
In *International Conference on Machine Learning* (pp. 9526-9536). PMLR.

*Thanks!*