最適輸送理論の単一細胞データ解析への応用：細胞分化ダイナミクスの推定

谷地村 敏明 (Toshiaki Yachimura)
東北大学数理科学共創社会センター (MathCCS)

第19回学習物理領域セミナー＋第71回DLAP
2025 7/24

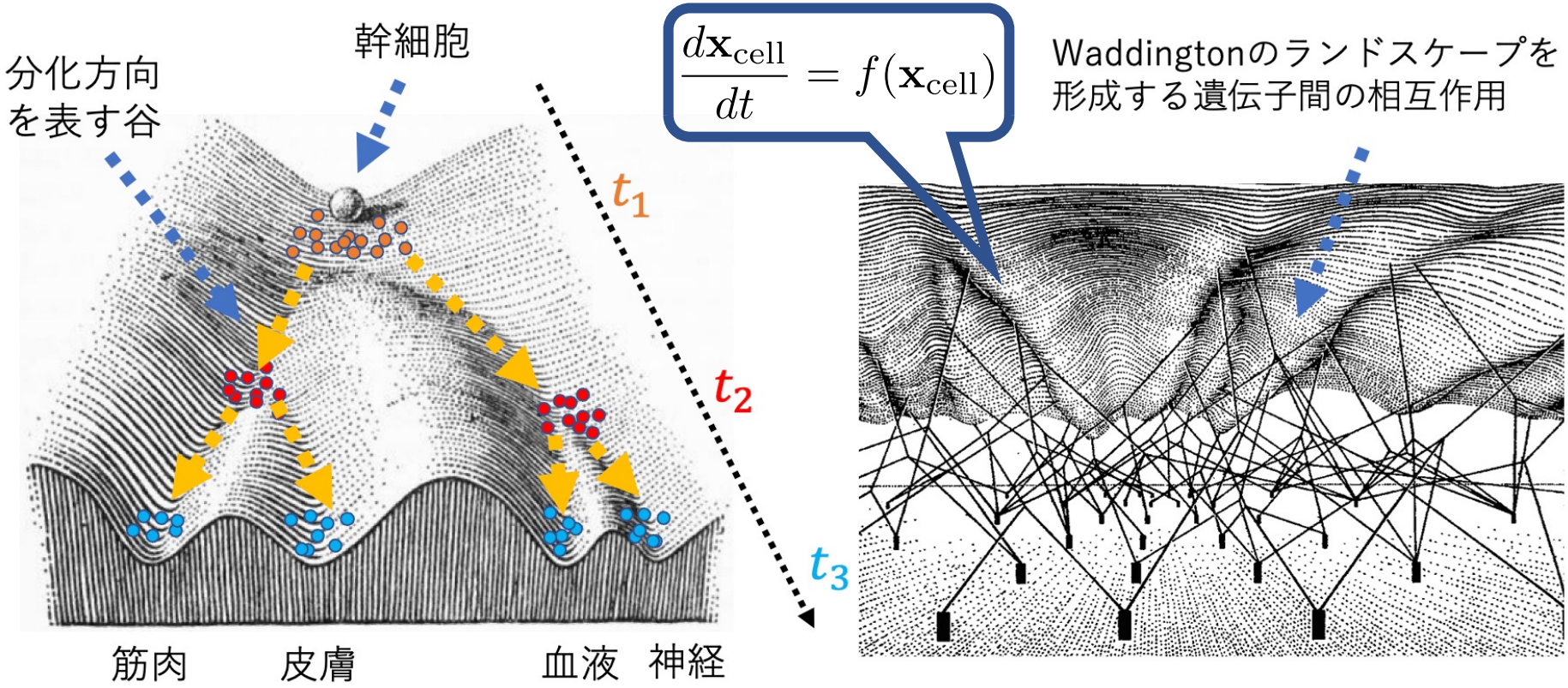
最適輸送理論×細胞分化ダイナミクス推定

T. Yachimura, et al., scEGOT: single-cell trajectory inference framework based on entropic Gaussian mixture optimal transport, BMC Bioinformatics, 2024.

github: https://github.com/yachimura-lab/scEGOT

細胞分化とWaddingtonのランドスケープ

Waddington’s landscape (1957)：細胞分化ダイナミクスの概念モデル



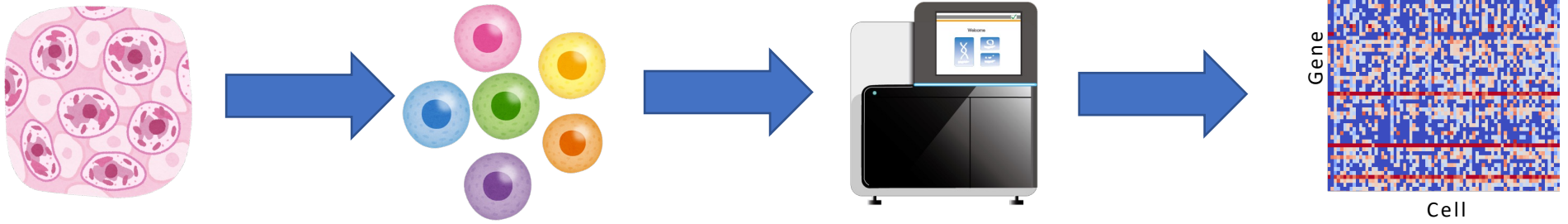
実際の細胞分化におけるWaddingtonのランドスケープを知ることができれば…

➔ 細胞分化を予測・制御することが可能

- 実際の細胞分化において，Waddingtonのランドスケープはどういった形状か?
- Waddingtonのランドスケープを形成するマーカー遺伝子と遺伝子制御ネットワークは何か？

scRNA-seqと細胞系譜推定問題

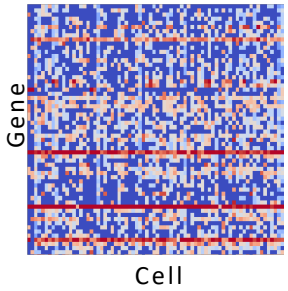
単一細胞RNAシーケンス技術 (scRNA-seq)

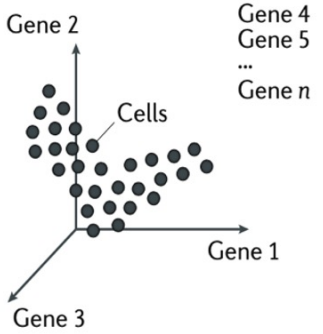


組織　　単一細胞に分離　　シーケンサ　　行列データ

(非負実数値)

Single-cell trajectory inference (細胞系譜(軌跡)推定問題)

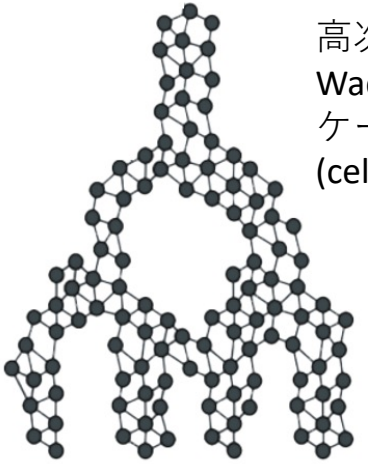




遺伝子発現空間 = $\mathbb{R}^{20000}$

再構成

グラフ理論/
次元削減手法



高次元遺伝子発現空間に
Waddingtonのランドスケープのような分化図
(cell state graph)を再構成

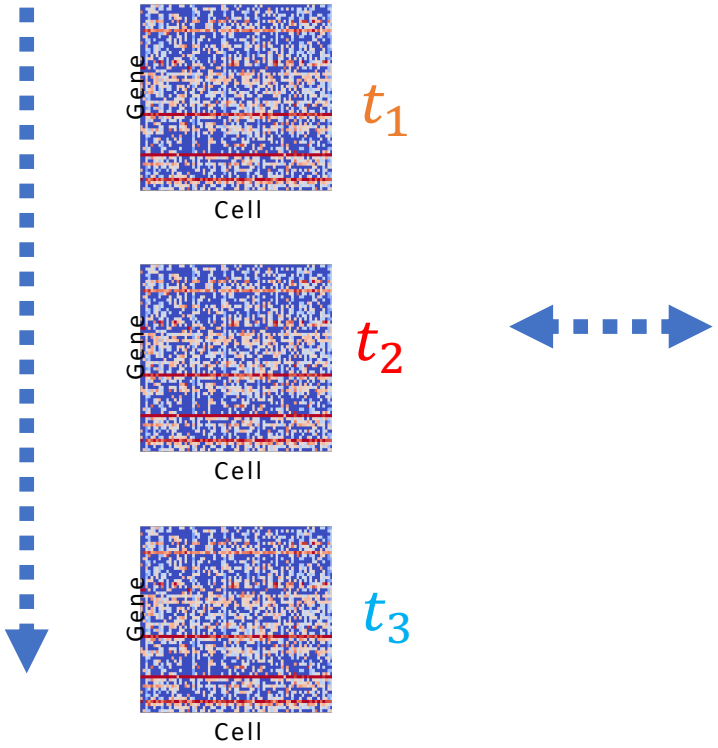
$\mathbb{R}^{20000}$

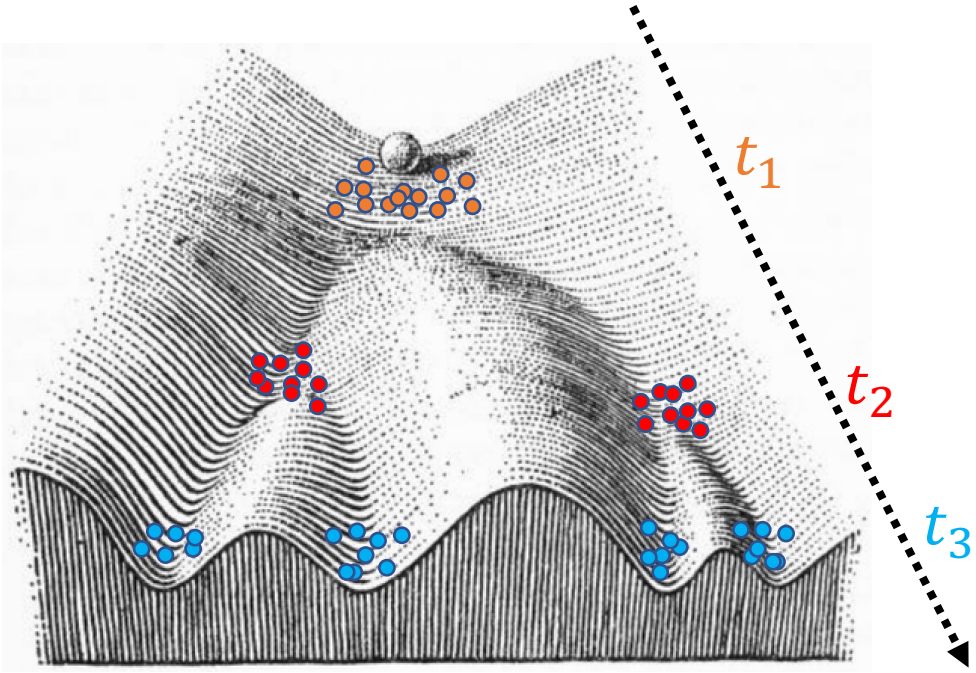
問題点

細胞分化過程全体におけるダイナミクス(分化の速度や中間状態など)を推定することは困難

最適輸送理論×単一細胞データ解析

時系列scRNA-seqはWaddingtonランドスケープを転がる細胞集団のスナップショット





数学的課題:

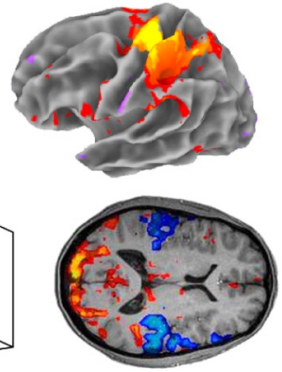
データから遺伝子発現のダイナミクス $\frac{d\mathbf{x}_{\mathrm{cell}}}{dt} = f(\mathbf{x}_{\mathrm{cell}})$ を推定すること

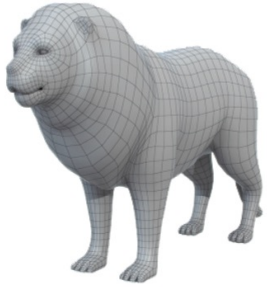
困難点: シーケンスにかける際に細胞を壊す必要がある(ラベルの欠如).

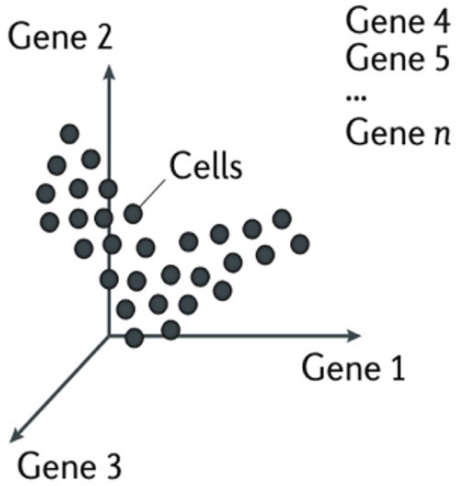
最適輸送理論を用いることでscRNA-seqデータ間をつなぎ，細胞分化ダイナミクスを推定する

最適輸送理論

画像, グラフィックス, 言語, scRNAseqデータ etc --> 多くのデータは高次元空間における確率分布(点群, ヒストグラム)と思える.

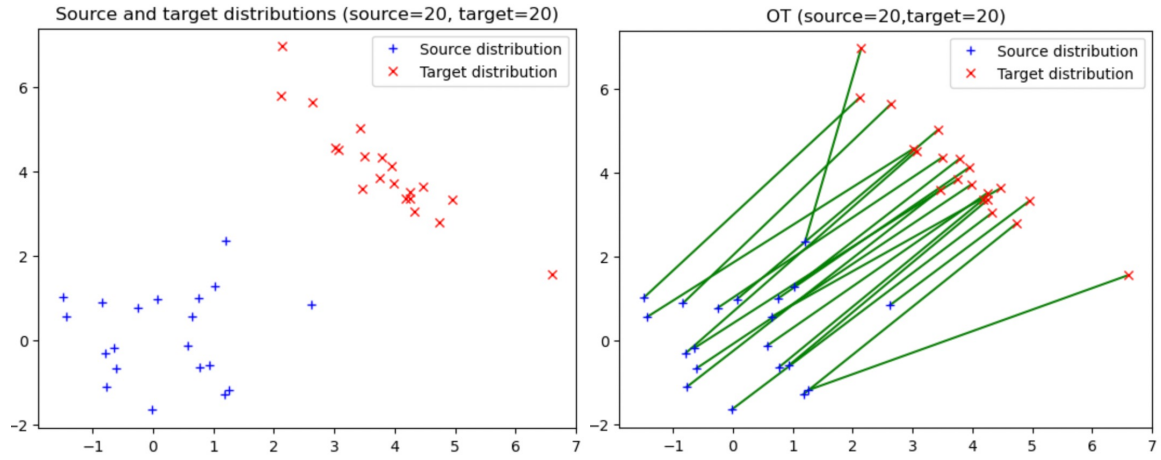


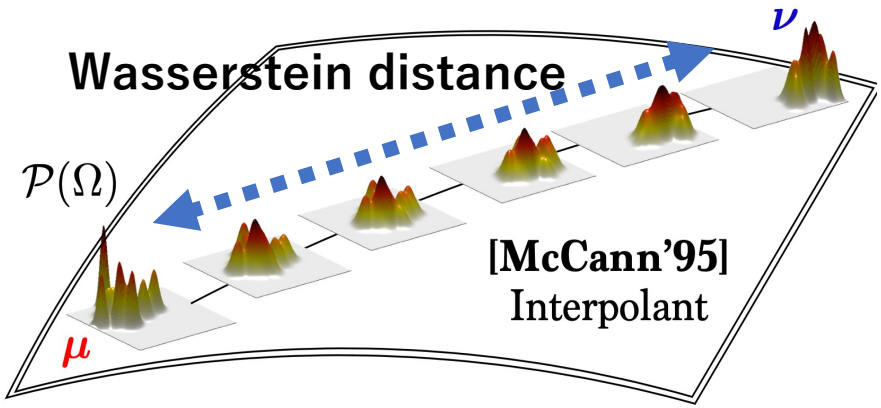




https://optimaltransport.github.io/

最適輸送理論: 確率測度に関する変分(最適化)問題で, 確率測度間における距離(**Wasserstein距離**)やその最適なマッチング (**最適輸送計画**)を与える数学理論.





最適輸送理論の応用例 (グラフィックスの補間)

- 2つ(以上)の確率測度の自然な補間が作れる (McCannの変位補間，Wasserstein重心)

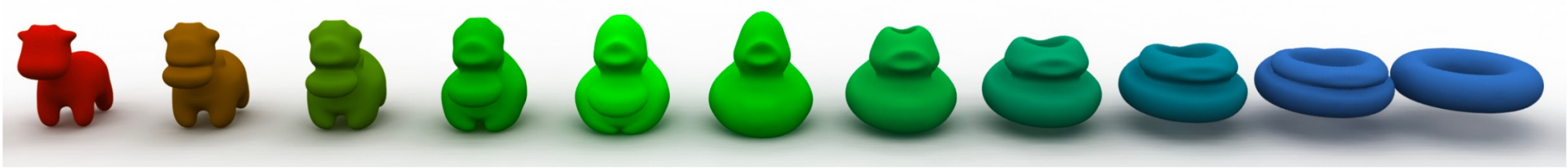
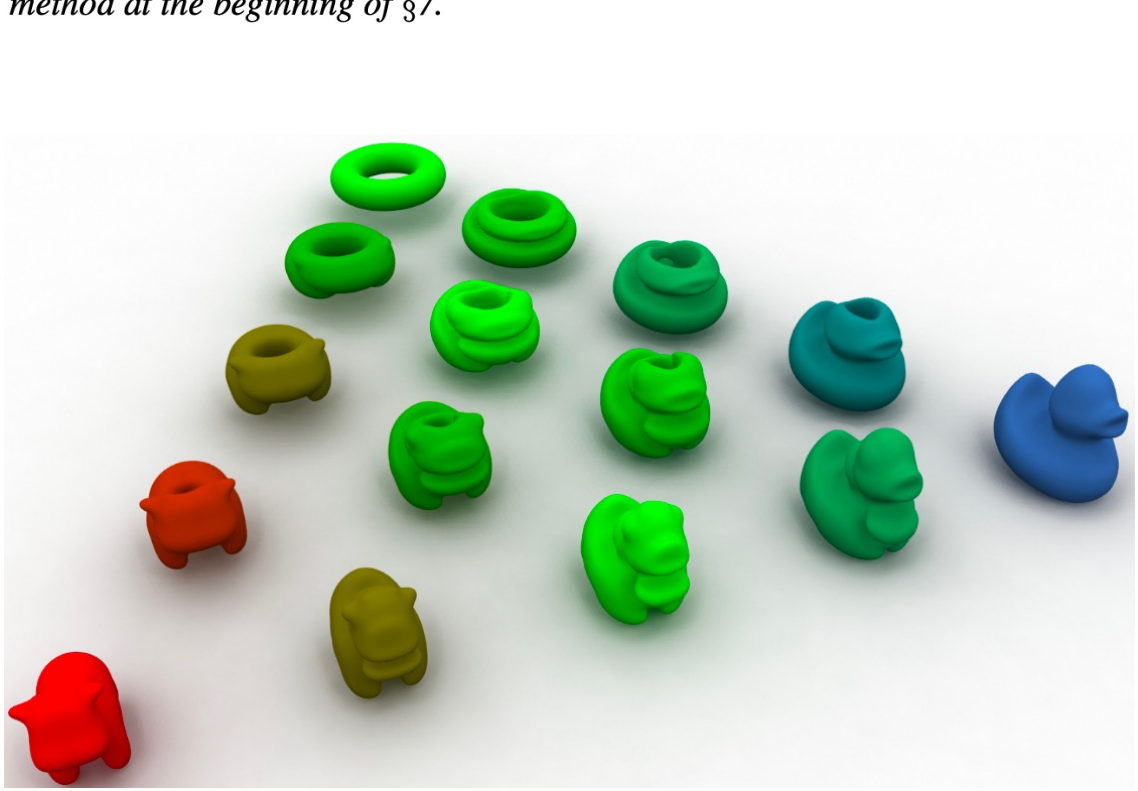
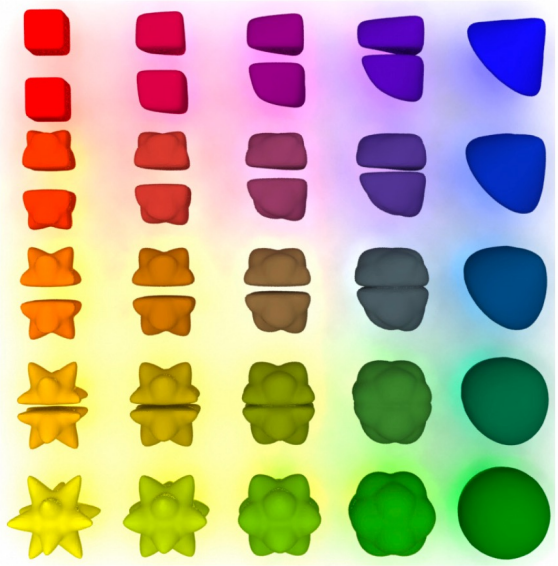
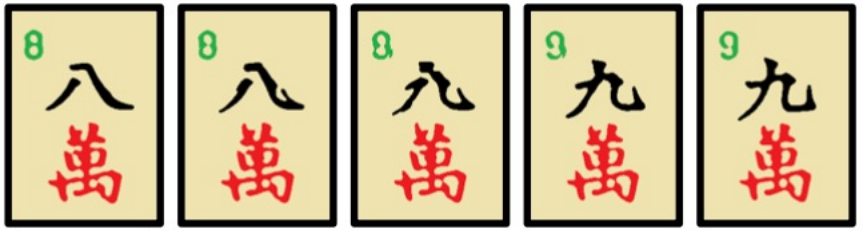


Figure 1: *Shape interpolation from a cow to a duck to a torus via convolutional Wasserstein barycenters on a* $100 \times 100 \times 100$ *grid, using the method at the beginning of* §7.

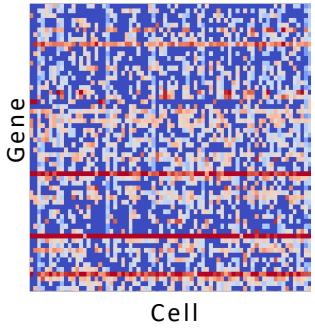


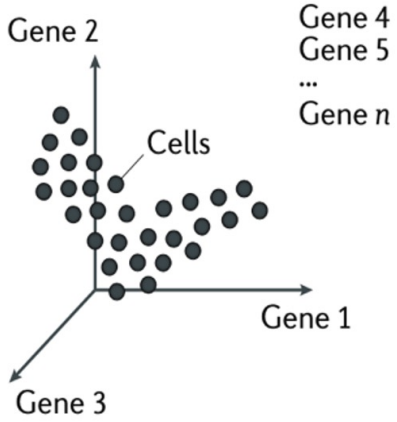




J. Solomon et al., Convolutional wasserstein distances, 2015.

最適輸送理論の単一細胞データ解析への応用例

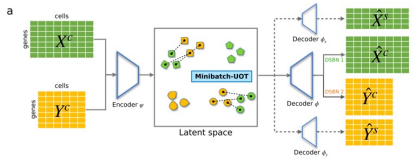


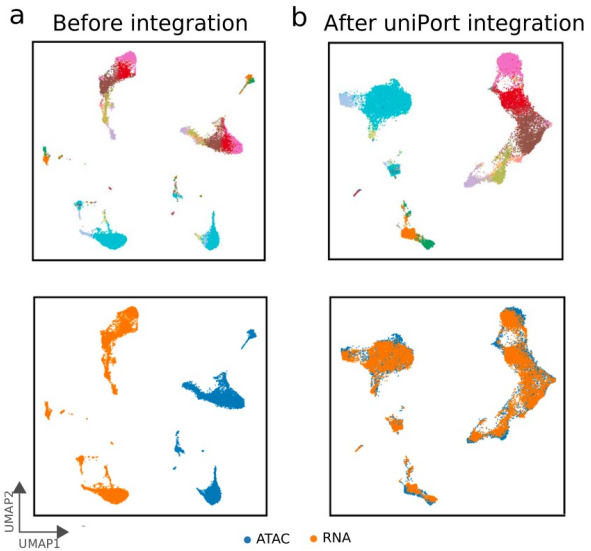


scRNAseqデータ

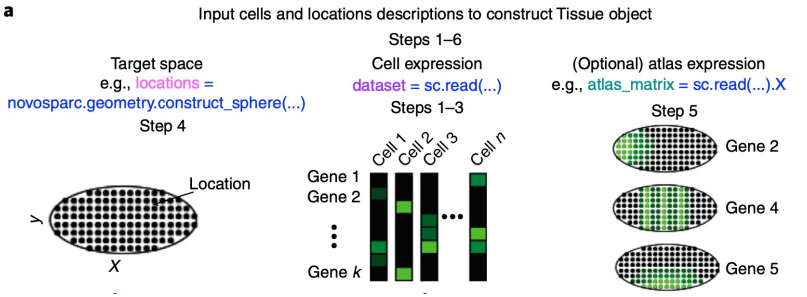
高次元空間における点群(離散確率測度)

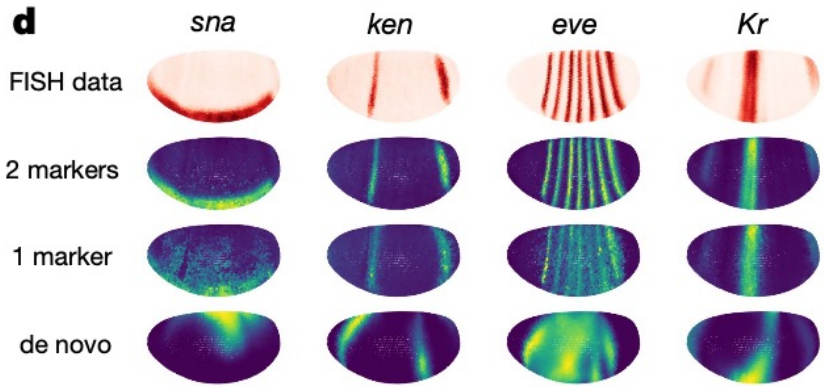
VAEと最適輸送を用いたdata integration(scRNAとscATAC), Cao et.al., Nat. Commun., 2022.



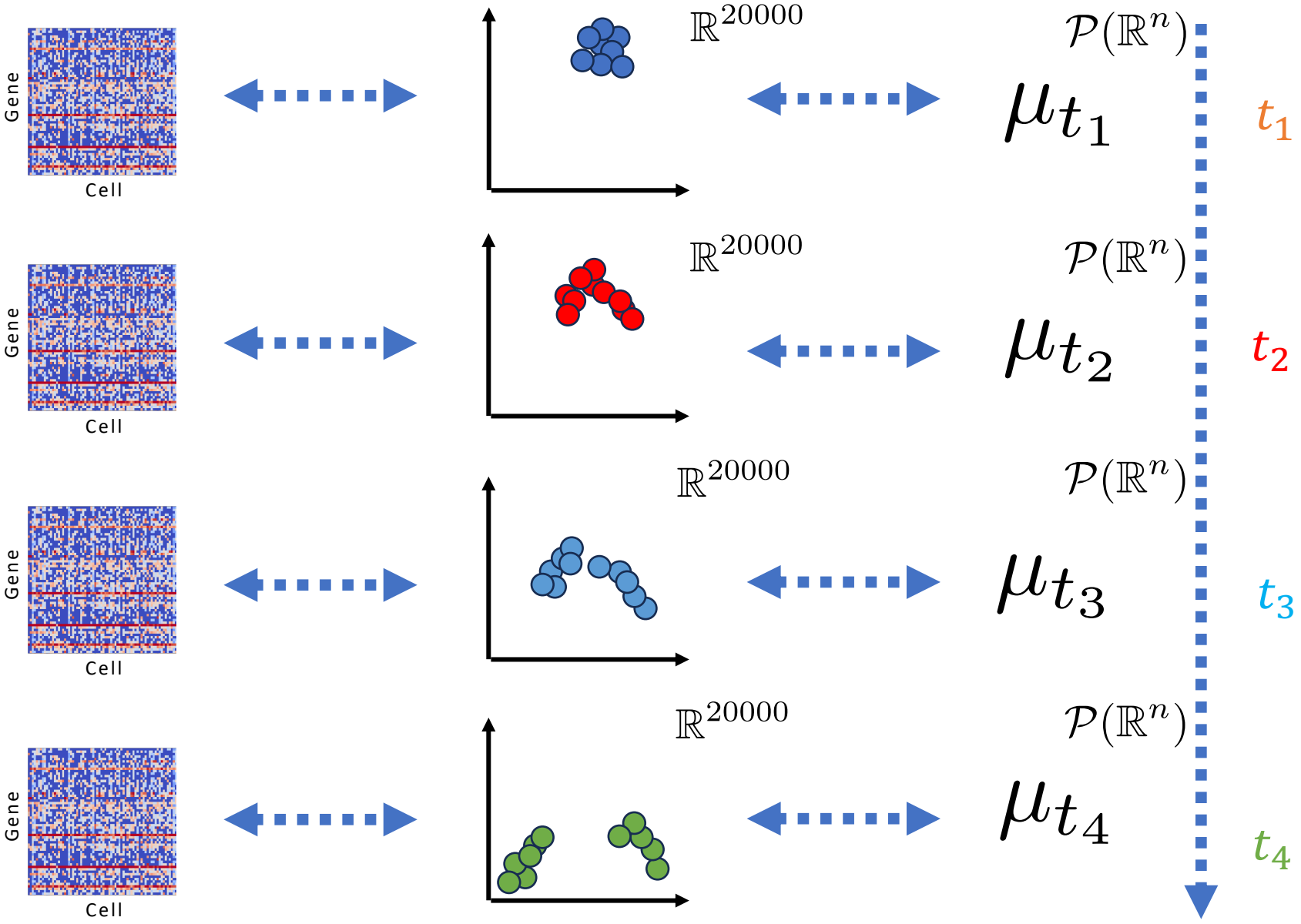


最適輸送による単一細胞遺伝子発現の空間再構成, Nitzan et. al., Nature, 2019.

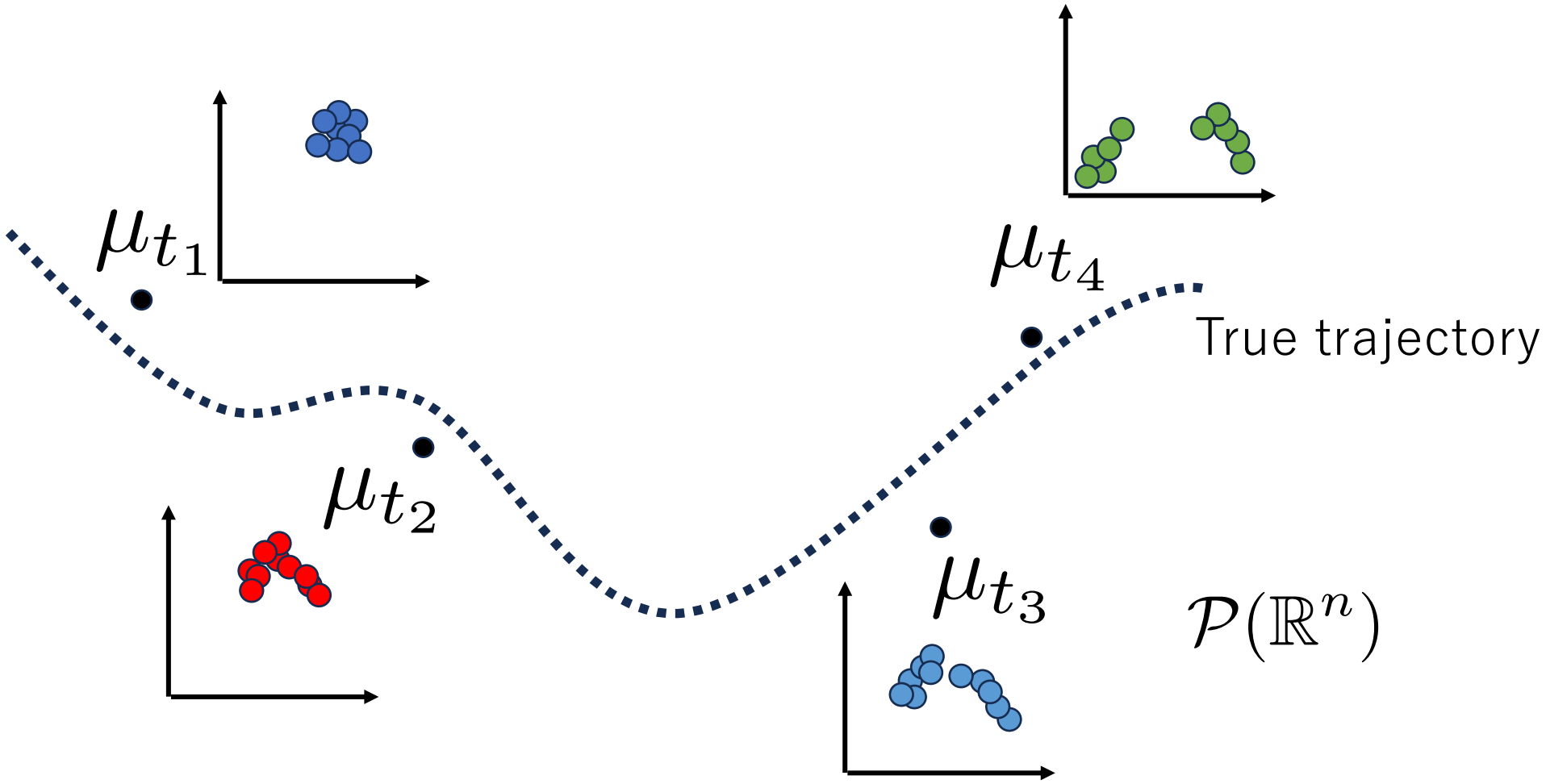




最適輸送理論からみた細胞系譜推定問題



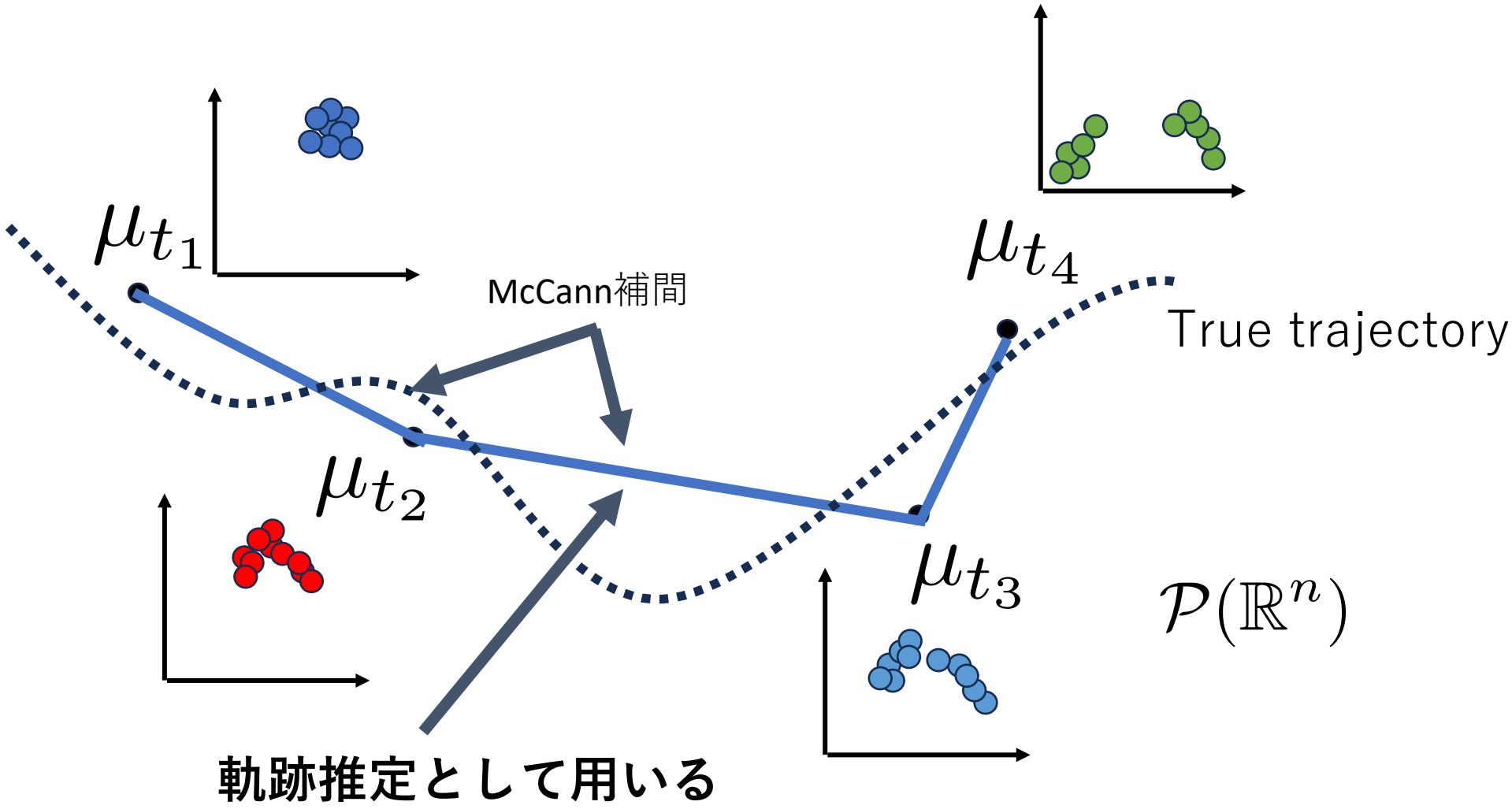
最適輸送理論からみた細胞系譜推定問題



最適輸送理論からみた細胞系譜推定問題:

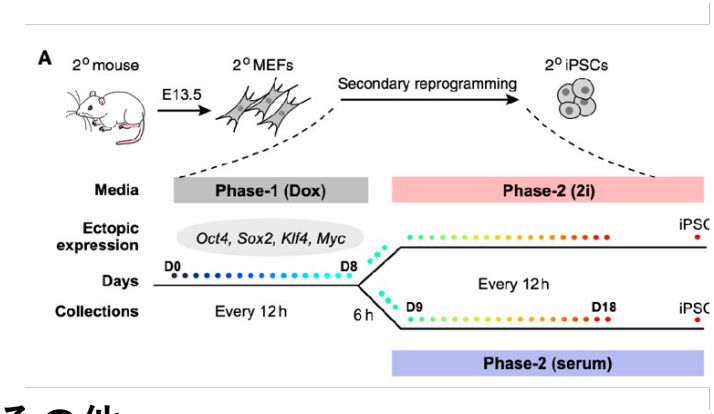
与えられたデータ μ_{t_i} を通る(または近い)Wasserstein曲線 μ_t を推定する

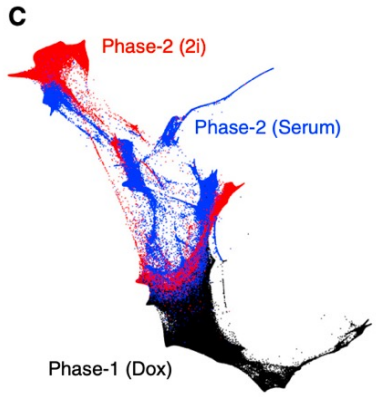
最適輸送理論からみた細胞系譜推定問題

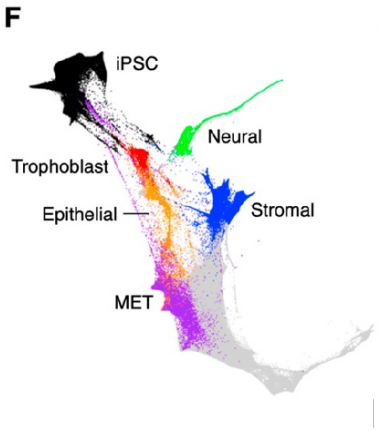


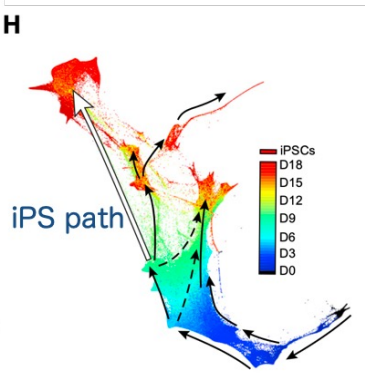
先行研究: Waddington-OT(Schiebinger, et al., Cell, 2019)

方法: 離散型最適輸送(**Entropic unbalanced OT with growth rate of cell**)
データ: MEFsからiPSCsへのリプログラミングに関する時系列scRNA-seqデータ (A,C,F)
結果: 1. 分化経路の発見(**iPS path**) (H)
2. リプログラミングに関連するマーカー遺伝子の発見(ex. *Obox6*)









その他:

- **TrajectoryNet (A. Tong, et al., ICML, 2020)** : Benamou--Brenier's dynamic formulation and normalizing flow
- **JKONET (C. Bunne, et al., ICML, 2021)** : Jordan--Kinderlehrer--Otto scheme and Neural Net
- **PRESCIENT (GHT Yeo, et al., Nat. communications, 2021)** : stochastic dynamics and Neural Net

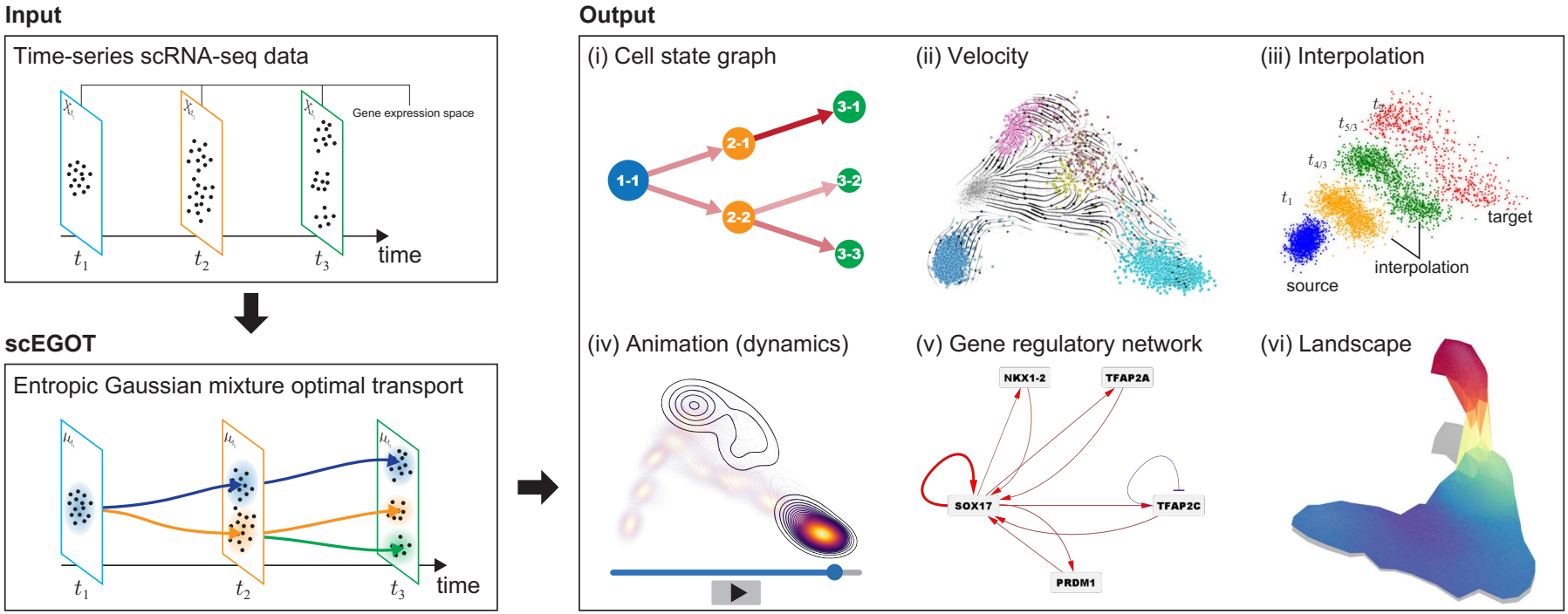
ニューラルネットを用いた細胞分化ダイナミクスの推定や中間状態の生成が可能

問題点:

- 計算コストの高さ
- ニューラルネットワークの技術を用いることによるブラックボックス化と生物学的な解釈の不透明性

新規ソフトウェア

scEGOT: single cell trajectory inference framework by Entropic Gaussian mixture Optimal Transport の開発 (github: https://github.com/yachimura-lab/scEGOT**)**



- 細胞集団はsub population (cluster) を形成しつつ分化していくはずである.

-> クラスタ間における最適輸送

- 細胞集団に関するデータが**混合ガウス分布**だと仮定して，**混合ガウス分布間のOT**を考える.

利点:

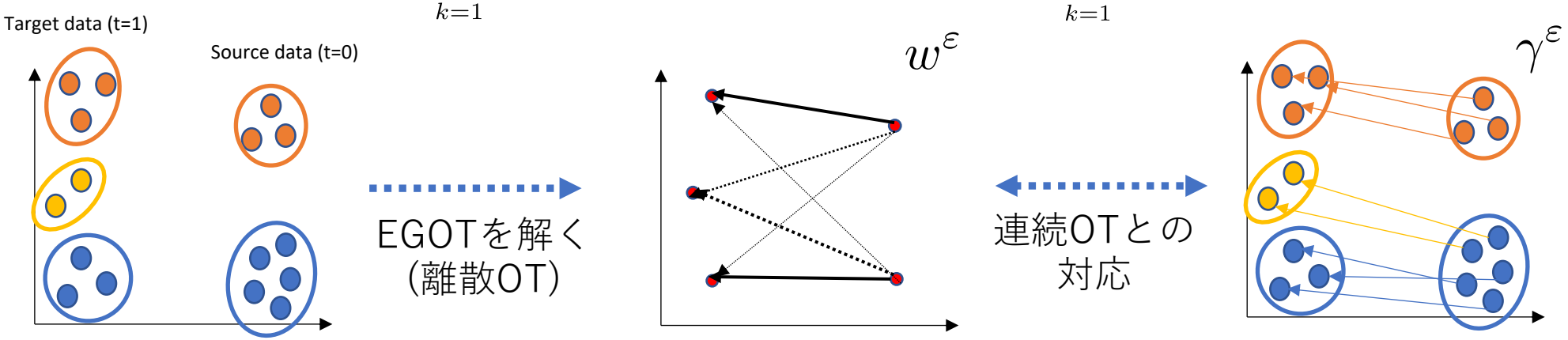
- クラスタ間における最適輸送であるため計算コストが著しく低い.
- 連続OT(生成モデル)との明確な対応.
- 生物学的な高い解釈性(各ガウス分布=各細胞種)

Entropic Gaussian mixture OT(EGOT)

- 全ての細胞集団 ＝混合ガウス分布
- 細胞種 ＝各ガウス分布

Source : $\mu_{t_i} = \sum_{k=1}^{K_i} \pi_i^k \mu_i^k$

Target : $\mu_{t_{i+1}} = \sum_{k=1}^{K_{i+1}} \pi_{i+1}^k \mu_{i+1}^k$



Entropic Gaussian mixture OT (EGOT)

$$d_G^\varepsilon(\mu_{t_i}, \mu_{t_{i+1}}) = \min_{w \in \Pi(\pi_i, \pi_{i+1})} \sum_{k,l} w_{k,l} W_2^2(\mu_i^k, \mu_{i+1}^l) - \varepsilon H(w),$$

where

$$\Pi(\pi_i, \pi_{i+1}) = \{w \in \mathcal{P}_{K_i \times K_{i+1}} : w \mathbf{1}_{K_{i+1}} = \pi_i, w^T \mathbf{1}_{K_i} = \pi_{i+1}\}$$

ガウス分布同士の
Wasserstein距離

Gaussianを一つの点とみなす
(Gaussian測度全体の空間における離散型正則化最適輸送を考える)

エントロピー
(KL情報量)

Entropic transport plan

$$\gamma^\varepsilon(x, y) = \sum_{k,l} w_{k,l}^\varepsilon g_{m_i^k, \Sigma_i^k}(x)\, \delta_{y = T_{k,l}(x)},$$

where $g_{m_i^k, \Sigma_i^k}$ is the multivariate Gaussian distribution

$$g_{m_i^k, \Sigma_i^k}(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_i^k|}} e^{-\frac{1}{2}(x - m_i^k)^T (\Sigma_i^k)^{-1} (x - m_i^k)}$$

Mixture Wasserstein distance (Delon–Desolneux (2020))

$$MW_2^2(\mu_i, \mu_{i+1}) = \inf_{\gamma \in \hat{\Pi}(\mu_i, \mu_{i+1}) \cap GMM_{2n}(\infty)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|_2^2 \, d\gamma(x, y)$$

Entropic Gaussian mixture OT(EGOT)

正則化輸送計画を用いることで，**entropic barycenter(分布の中間状態)**を作ることができる.

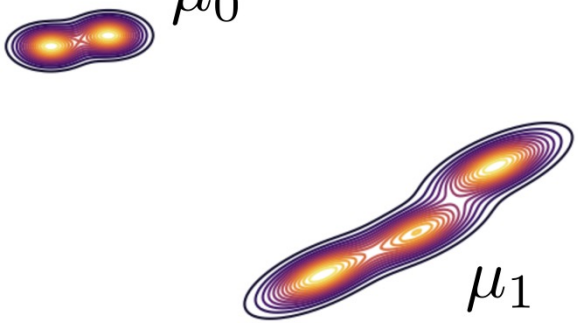
entropic barycenter

$$\mu^{\varepsilon}_{t_{i+s}} = \sum_{k,l} w^{\varepsilon}_{k,l} \mu^{k,l}_{t_{i+s}}$$

Entropic barycenter (2D, 40000分割)

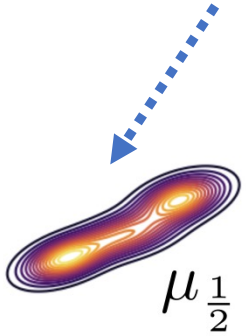
クラスタ性の保存

A



t = 0: クラスタ数2
t = 1: クラスタ数3

B



t=0.5, EGOT
Comp Time = 41ms

C

t=0.5, usual OT
Comp Time = 1h 44min 49s

Entropic Gaussian mixture OT(EGOT)

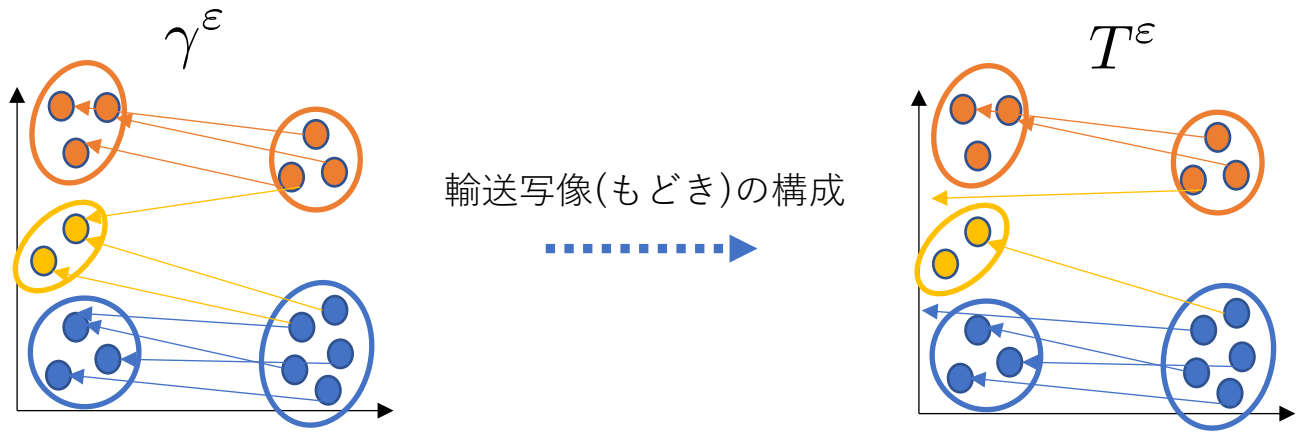
正則化輸送計画(点xから点yへどの程度の量運ぶか)を用いることで，**entropic projection map(点xをどの位置へ運べばよいか)**を作ることができる．

Entropic projection map T^{ε}

$$T^{\varepsilon}(x) = \int_{\mathbb{R}^n} y \, d\gamma_x^{\varepsilon}(y)$$

正則化輸送計画から，entropic projection mapを具体的に計算することができる．

$$T^{\varepsilon}(x) = \sum_{k,l} P_{k,l}^{\varepsilon}(x) T_{k,l}(x) \qquad P_{k,l}^{\varepsilon}(x) = \frac{w_{k,l}^{\varepsilon} g_{m_i^k, \Sigma_i^k}(x)}{\sum_j \pi_i^j g_{m_i^j, \Sigma_i^j}(x)}$$



scEGOT: Single cell trajectory inference framework based on Entropic Gaussian mixture Optimal Transport

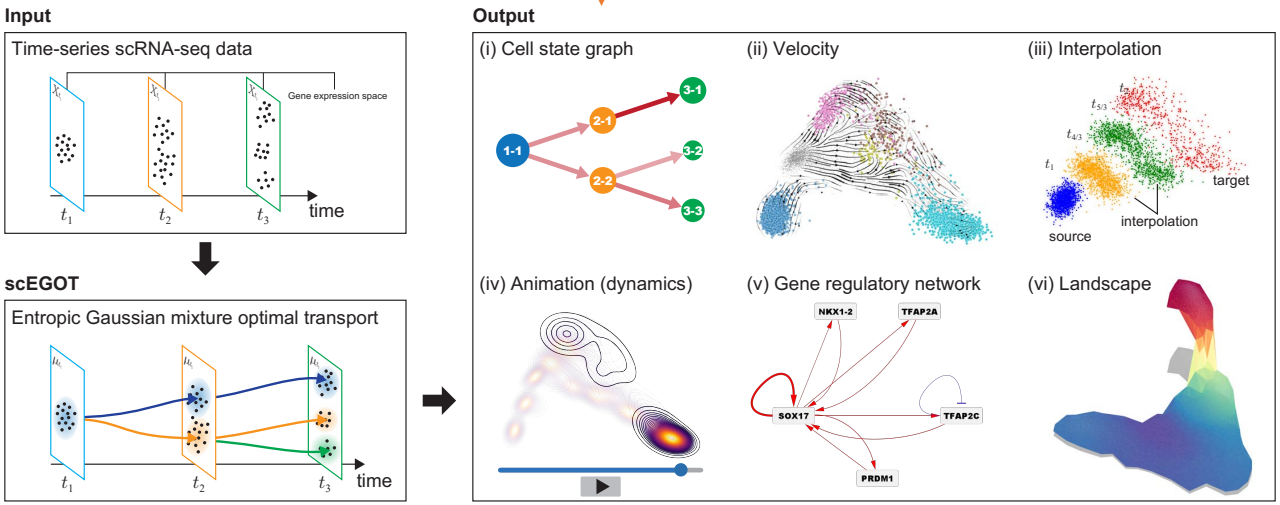
Math → **Biology**

- EGOTの解 (混合ガウスの重みの輸送) : w^{ε} → 細胞分化グラフ
- 正則化最適輸送計画 γ^{ε} とその重心: → 遺伝子発現のダイナミクス (アニメーション)
- Entropic barycentric projection map T^{ε} : → 遺伝子発現の速度 (cell velocity)

$$v_{\mathrm{cell}}(x) = T^{\varepsilon}(x) - x$$

etc…

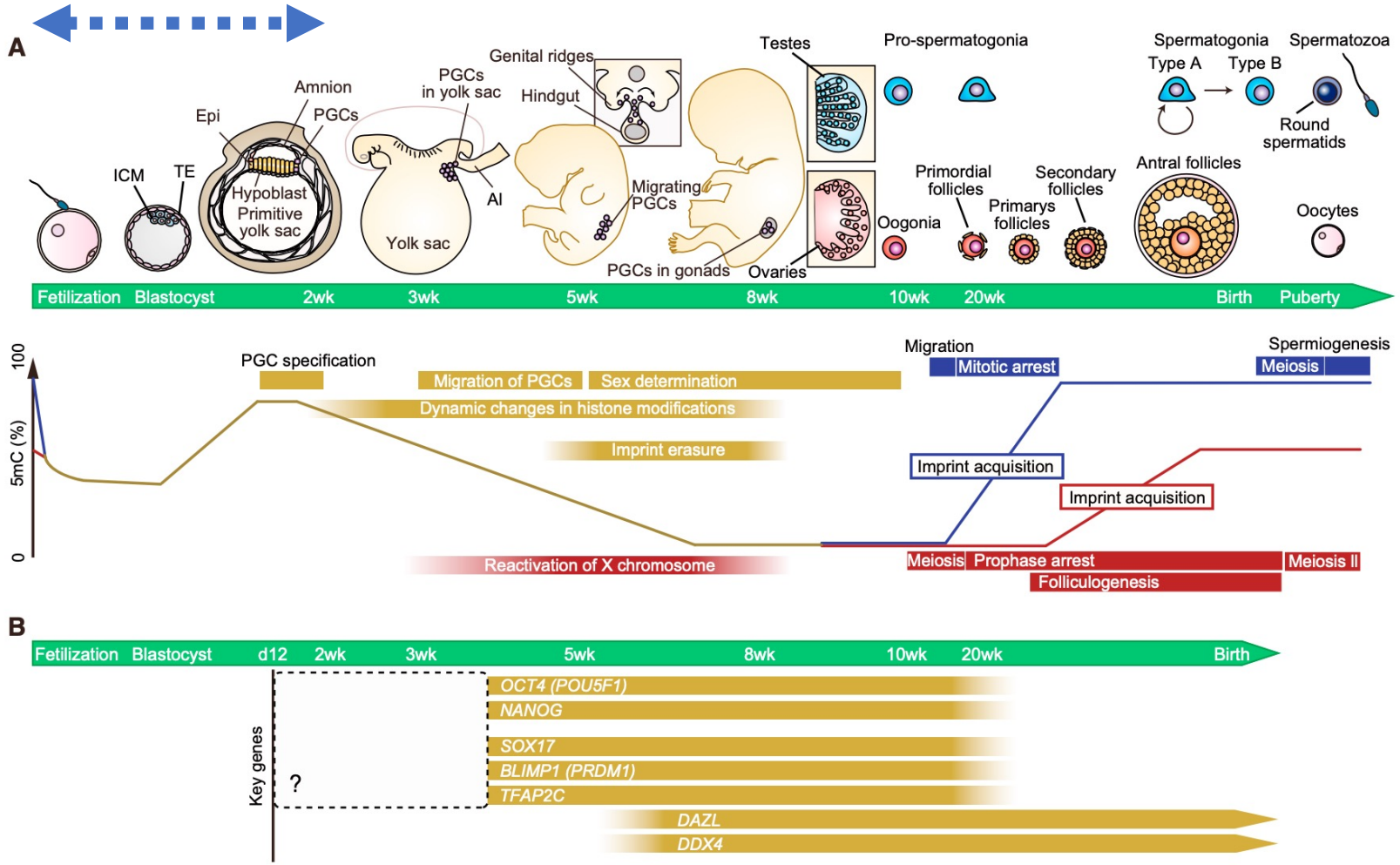
scEGOT: single cell trajectory inference framework by Entropic Gaussian mixture Optimal Transport



scEGOT×単一細胞データ解析（始原生殖細胞データ）

データ: iPS細胞からヒト始原生殖細胞(hPGCLC)への誘導系に関する時系列scRNAseqデータ(ASHBi斎藤グループから提供)

Day	0	0.5	1	1.5	2
Cells	4943	1981	551	2562	1753

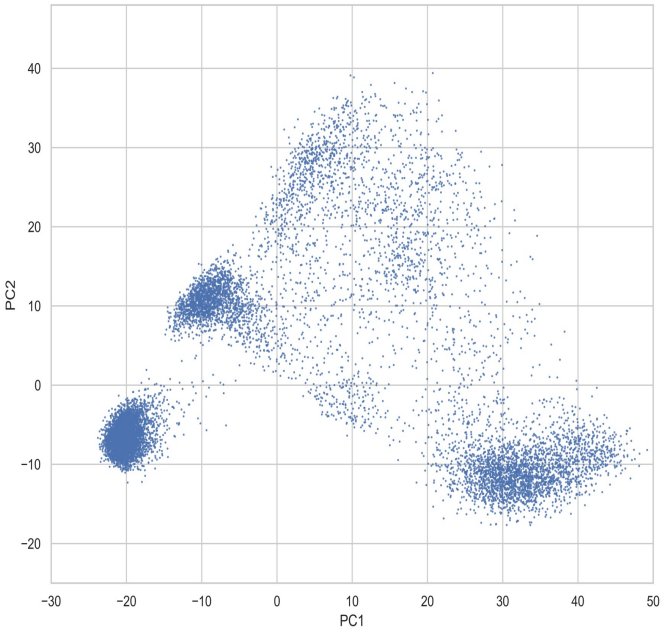


Saitou—Miyauchi, Cell Stem Cell, 2016

scEGOT×単一細胞データ解析（始原生殖細胞データ）

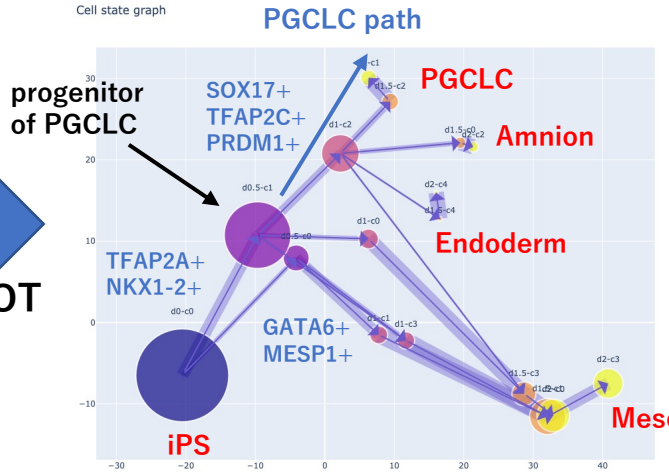
データ: iPS細胞からヒト始原生殖細胞(hPGCLC)への誘導系に関する時系列scRNA-seqデータ(ASHBi斎藤グループから提供)

Day	0	0.5	1	1.5	2
Cells	4943	1981	551	2562	1753

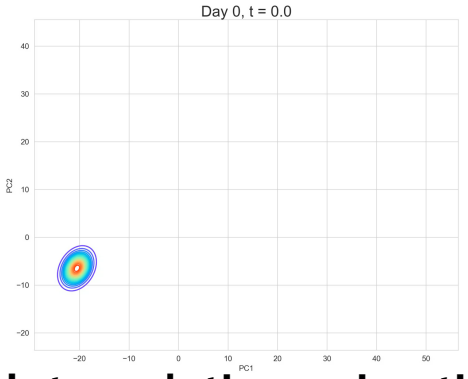


PC1/PC2 plot (11,790 cells)

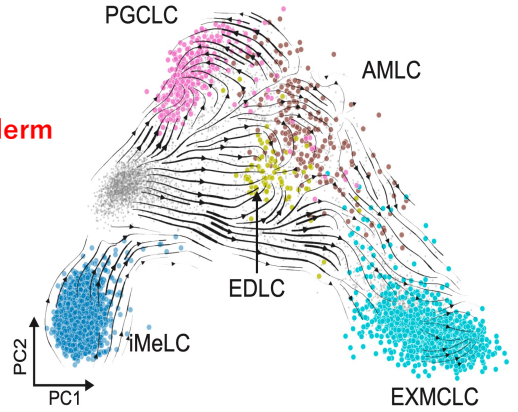
scEGOT



Cell state graph



Interpolation animation

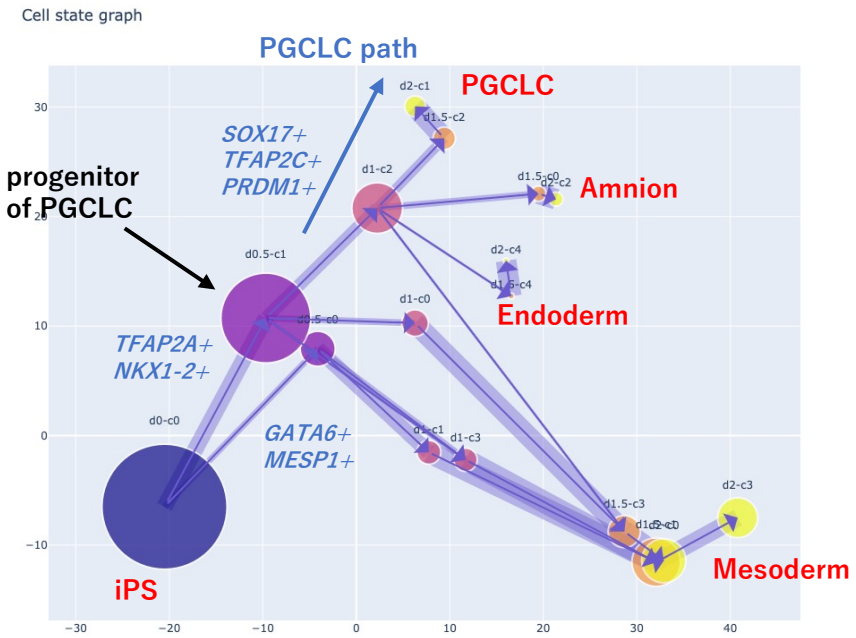


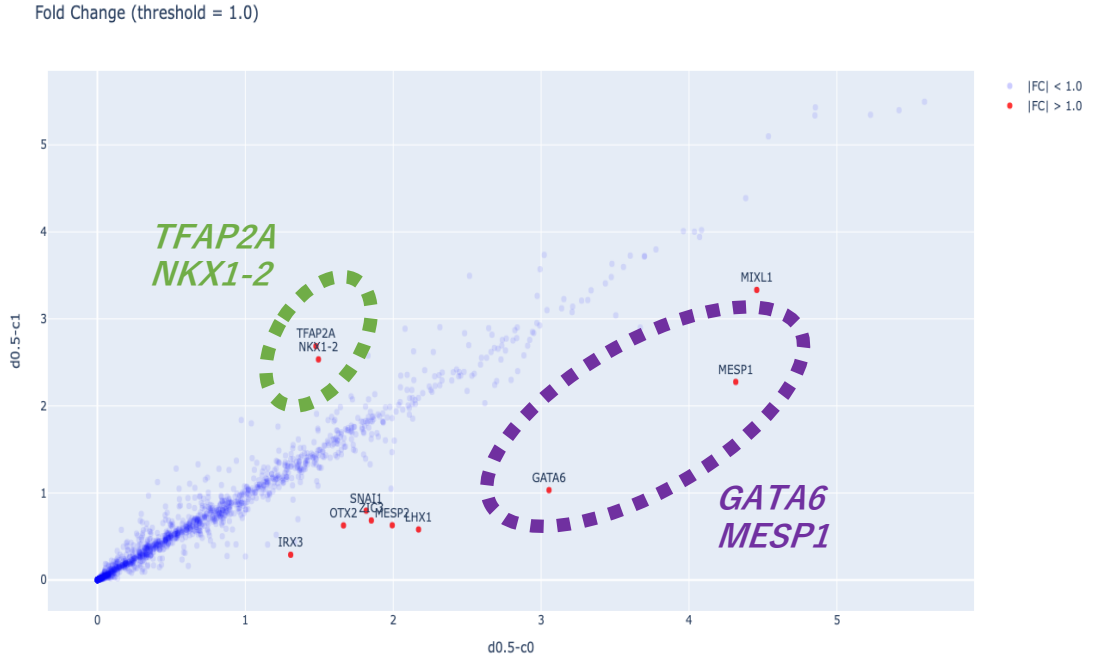
Cell velocity

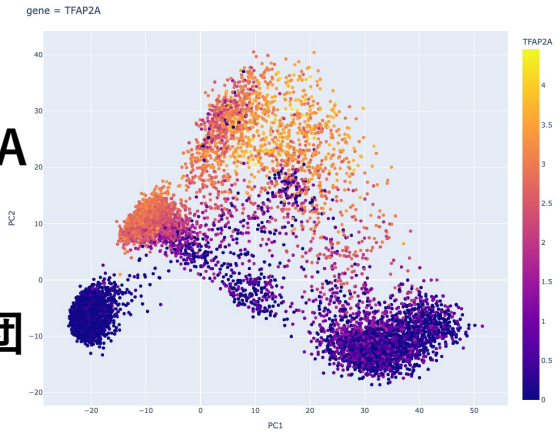
結果:

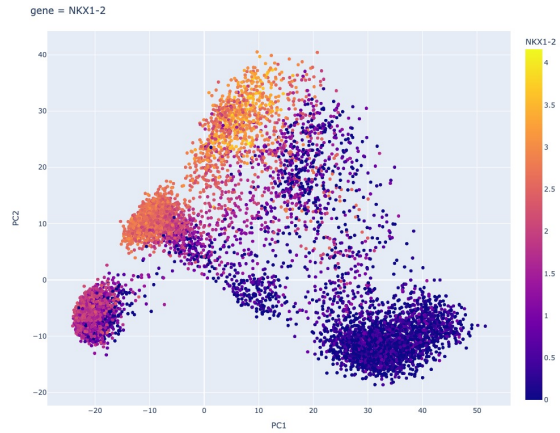
1. scEGOTから，遺伝子発現空間における**細胞分化の状態グラフ(cell state graph)**や**細胞分化のダイナミクス(interpolation animation, cell velocity)**が得られる.
2. PGCLCの細胞分化経路と前駆細胞集団の特定.

scEGOT×単一細胞データ解析（始原生殖細胞データ）









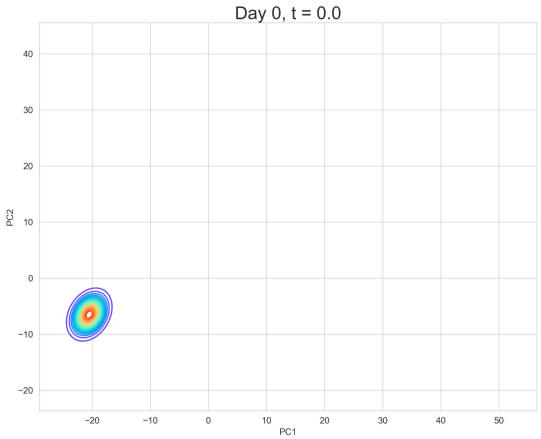
結果:

3. PGCLC分化誘導系において，**TFAP2A**や**NKX1-2**がマーカー遺伝子であることの発見

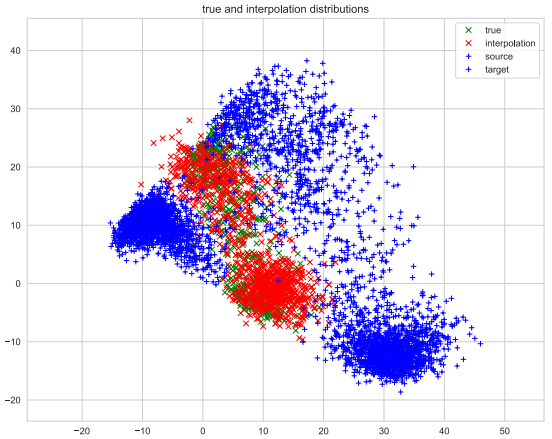
Human PGCはTFAP2A+の前駆細胞集団を経由して分化する.

Consistent to Chen et al., 2019, Cell Reports, Venzor et al., Life Sci Alliance, 2023.

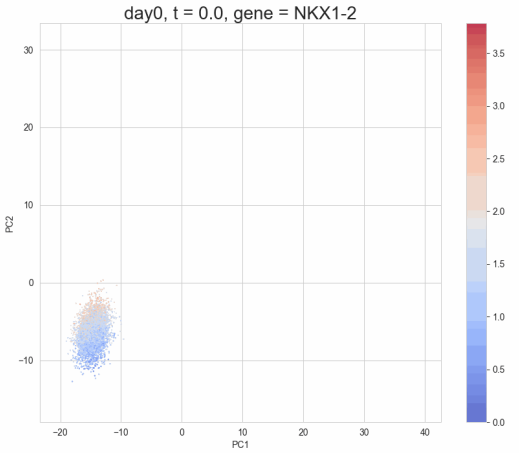
scEGOTのその他の機能（scRNA-seq dataの補間, Waddington landscapeの再構成, GRNの推定）



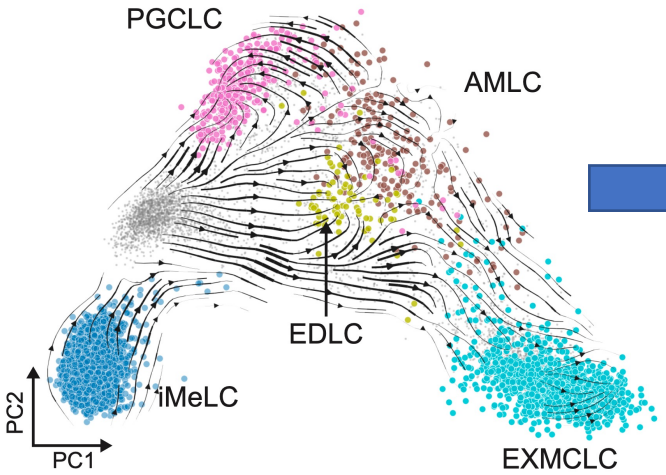
Interpolation animation



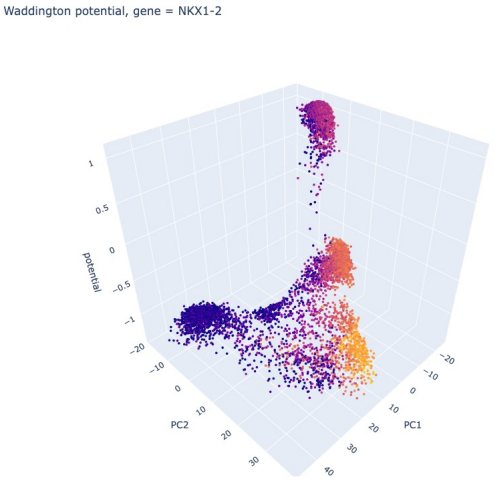
Interpolation of scRNA-seq data



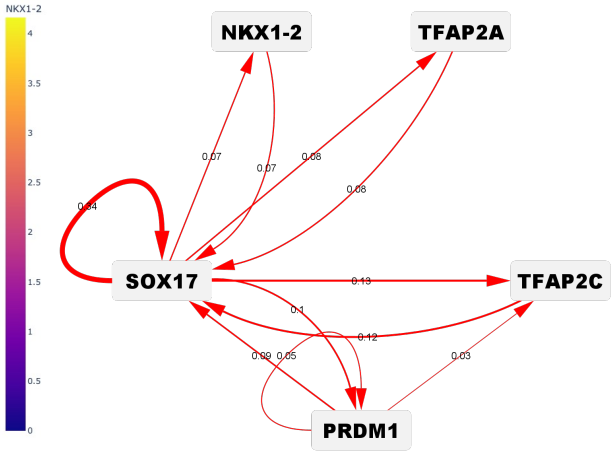
Interpolation animation of gene expression



Cell velocity

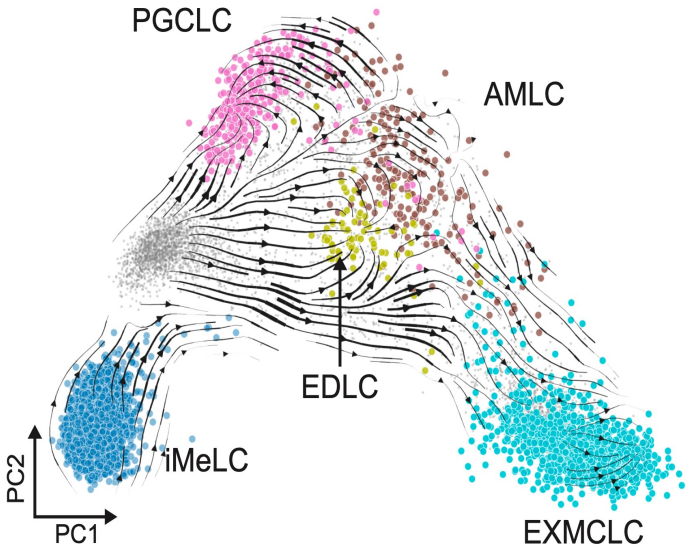


Waddington landscape construction

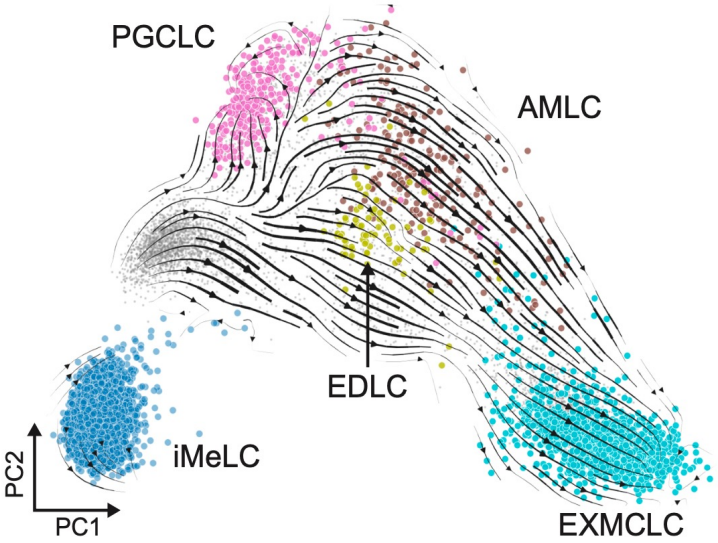


GRN estimation

RNA-velocity(従来手法)との比較

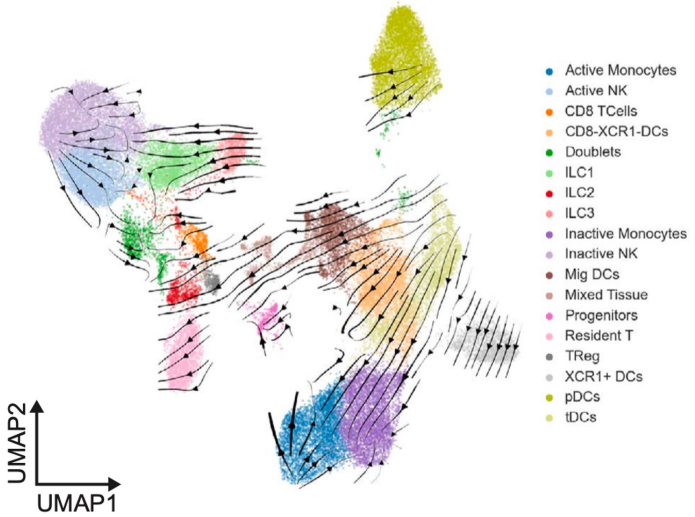


Cell-velocity (提案手法)



RNA-velocity(従来手法)

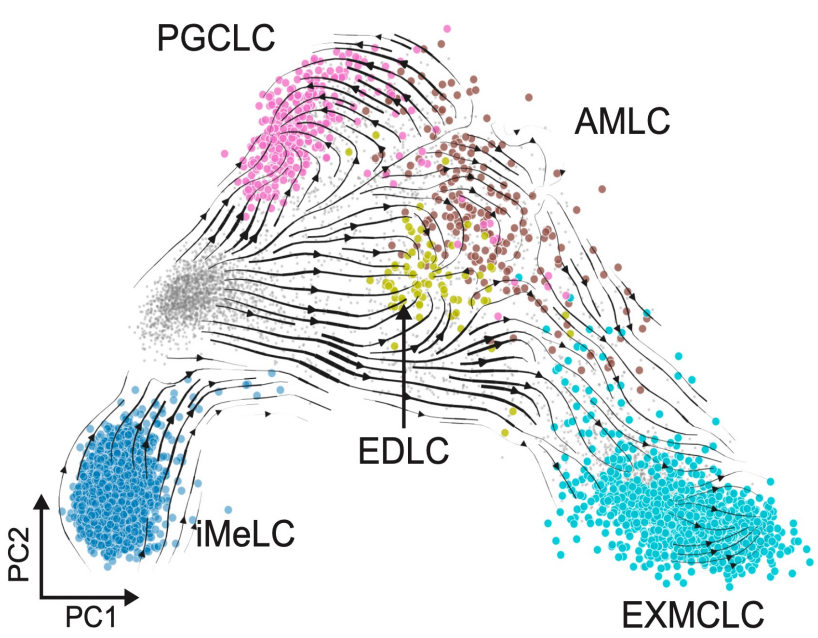
- Cell-velocityはspliced/unspliced RNAの情報を必要としない(**データ駆動型手法**).
- Cell-velocityは**他の単一細胞データ**(e.g. scATAC-seq)**に適用可能**.

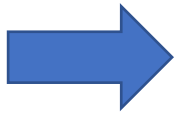


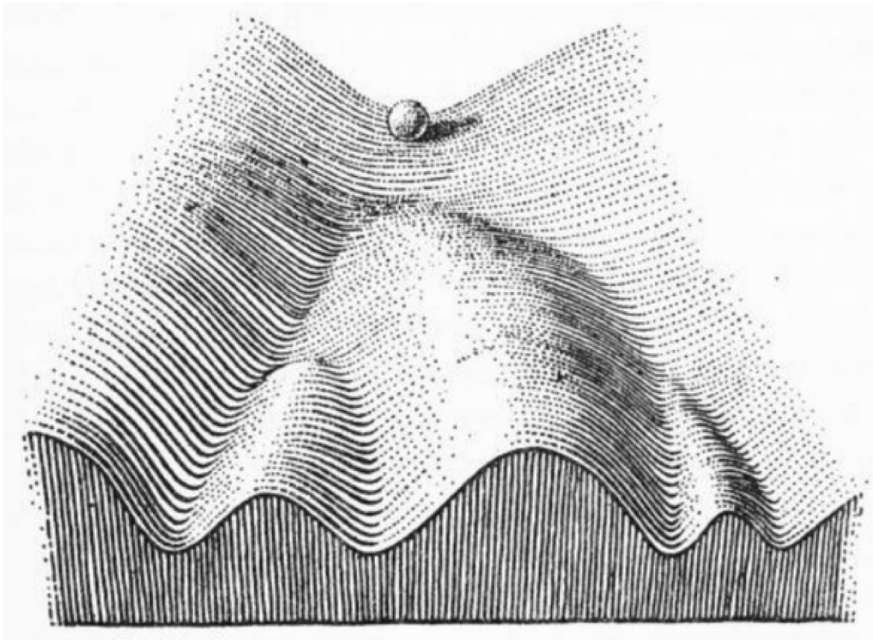
Cell velocity for scATAC-seq data of mouse innate immune cells at three-time points (days 0, 1, and 28)

Waddington landscapeの再構成

cell-velocityからWaddington landscapeを再構成する.





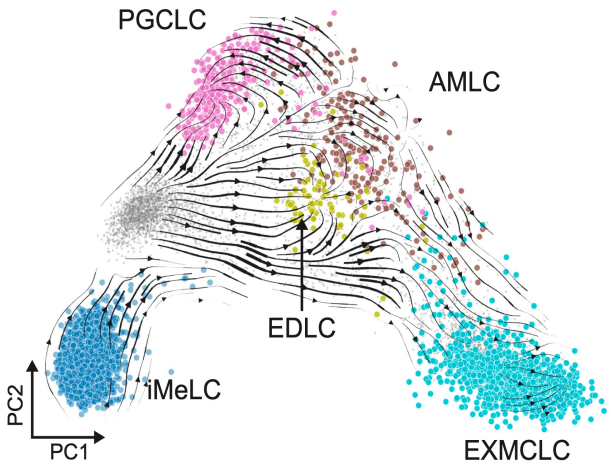


Helmholtz-Hodge分解から，以下のように書ける.

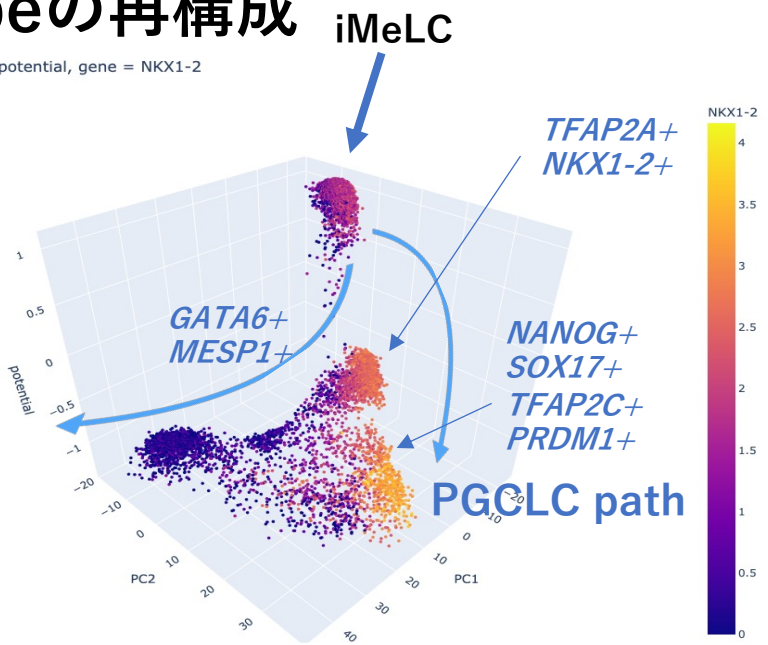
$$v = -\nabla\varphi + q, \quad \operatorname{div} q = 0$$

Gradient potential　Divergence free part

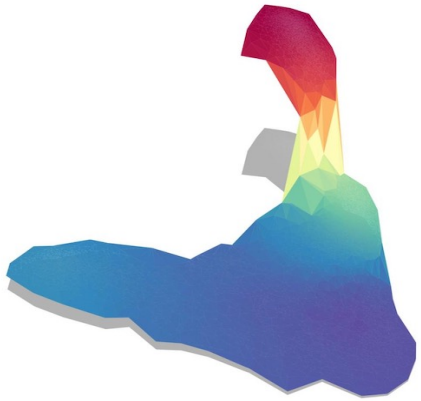
Waddington landscapeの再構成



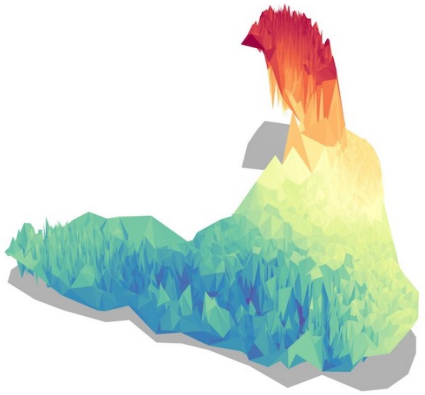
$$-\Delta\varphi = \operatorname{div} v_{\mathrm{cell}}$$



scEGOTにより，**Waddington's landscape**のような勾配ポテンシャルが書ける．



scEGOTによるポテンシャル推定



CytoTRACE 2 (深層学習フレームワークによるポテンシャル推定)

Mapping single-cell developmental potential in health and disease with interpretable deep learning, bioRxiv, 2024

GRNの推定

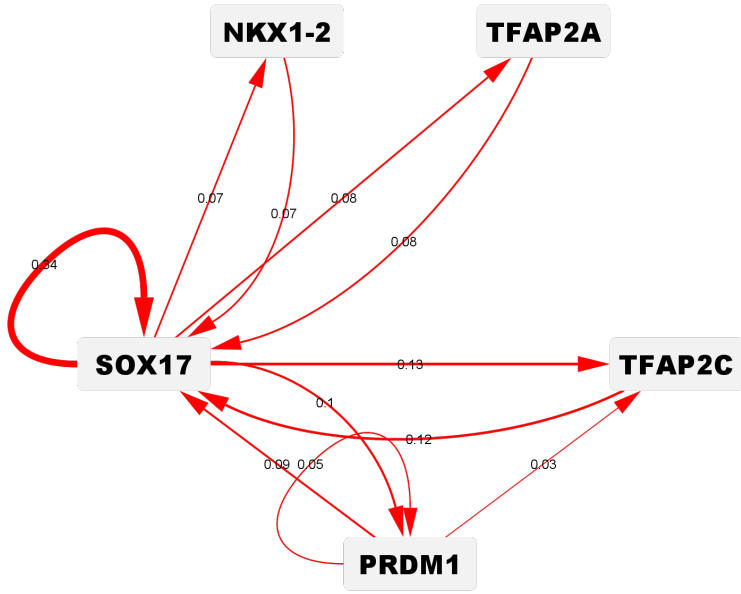
Algorithm:

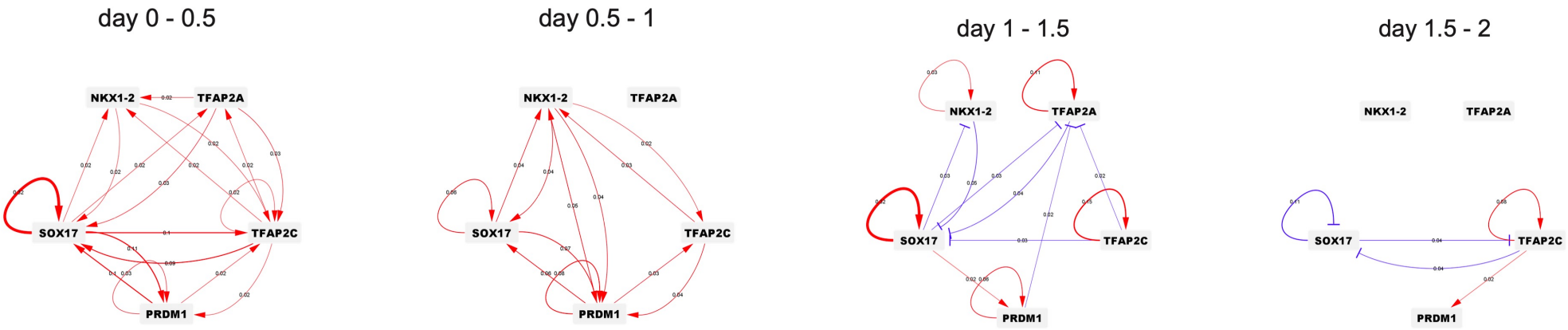
1. 線形ダイナミクスを仮定 $\frac{dx}{dt} = Ax$

2. Cell velocityから, $\frac{dx}{dt} = v_{\text{cell}}$

3. ridge回帰を解いてGRNを推定

$$\min_{A \in \mathbb{R}^{G \times G}} \|v_{\text{cell}} - Ax\|_2^2 + \lambda \|A\|_2^2$$

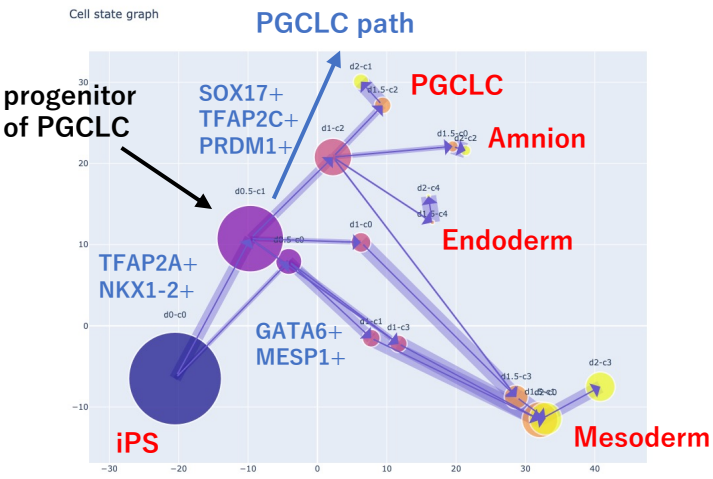




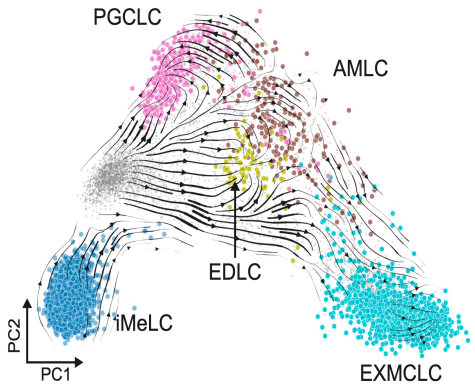
PGC経路に沿った動的GRN

- scEGOTによりWaddington landscapeを形成する背景のGRNを推定することができる.
- 経路に沿ったGRN推定も可能.
- SOX17はPGC誘導に必須のマーカー遺伝子であり，この遺伝子をKOするとPGCが誘導されないことが知られている.

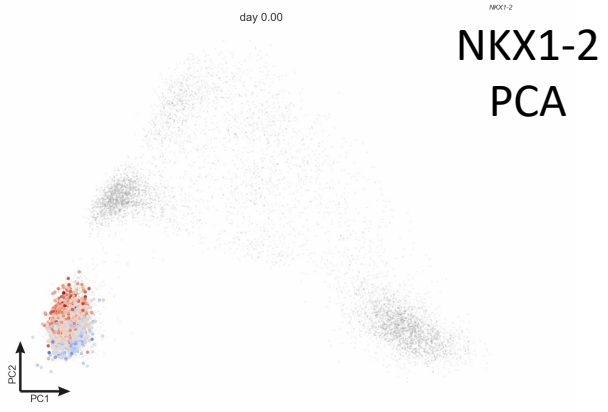
まとめ

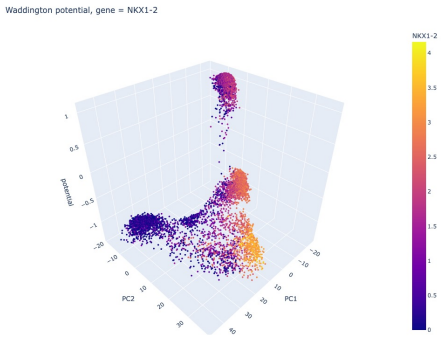


Cell state graph

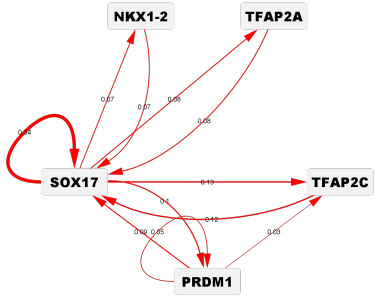


Cell velocity

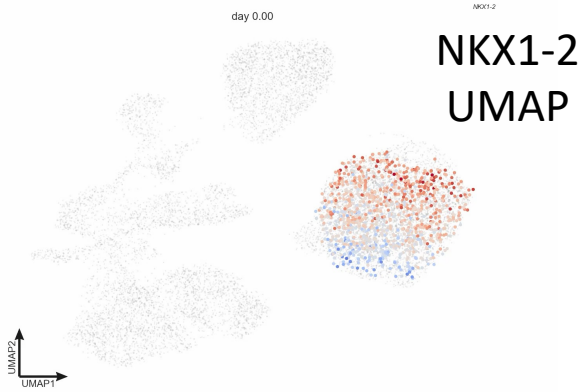




Landscape reconstruction



GRN estimation



Gene expression animation

- scEGOTは細胞分化の経路だけでなく，細胞分化のダイナミクス（遺伝子発現アニメーションやcell velocity）も推定できる.
- さらに，scEGOTは遺伝子発現空間におけるWaddingtonのランドスケープの再構築や背景のGRNを推定することができる.
- scEGOTを始原生殖細胞の誘導系に関する時系列scRNA-seqデータに用いることにより, PGC誘導に関連する新規遺伝子(e.g. **TFAP2A, NKX1-2, GATA6, MESP1**)を発見した.