

EB62104 「バイオインフォマティクス」 2日目

- 2日目（1月22日（日））のリアルタイムオンライン講義、午後は 13:30から開始です
- 暇な方はクイズをどうぞ
 - 長さが3塩基のDNA配列は何種類あるでしょうか？ただし、逆相補的配列が互いに一致する場合は1種類と数える（例：AATとATTは1種類）。
 - 長さがN塩基の場合、何種類あるでしょうか？ただし、（以下略）

この画面が見えているということは、ちゃんと接続できています

演習C

(資料 + リアルタイムオンライン)

録画を開始します

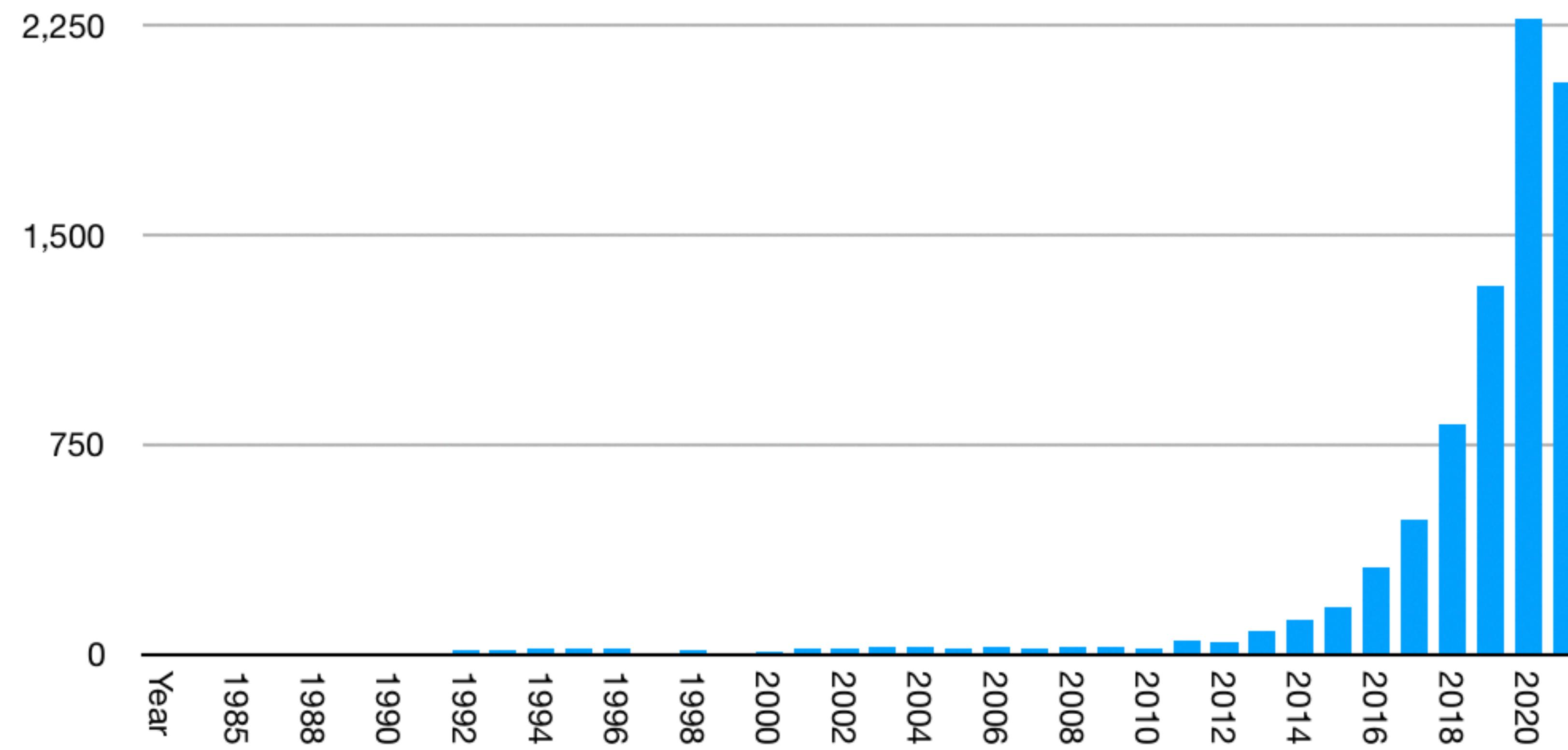
- リアルタイムオンラインだと通信環境によって聞き取れないことが起こります
- そのため、録画してアップロードするという措置を取っています
- 筑波大内部向けのアップロードであり、インターネットで誰でもみられるようになるという意味ではないのでご安心ください

シングルセルRNA-seq解析

「シングルセルRNAシークエンシング」

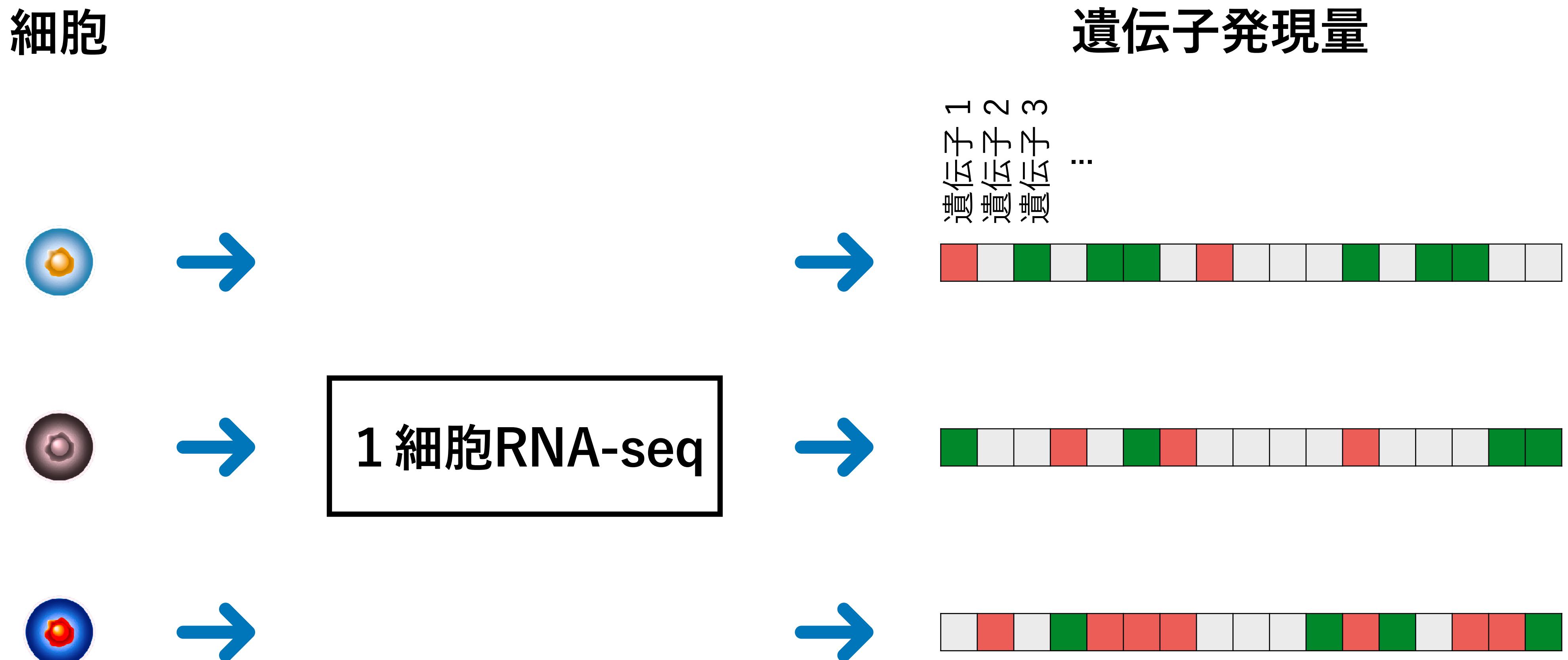
- 1 細胞RNA-seq
- シングルセルRNA-seq
 - アールエヌエーセック (アールエヌエーシーク)
- Single-cell RNA sequencing
- Single-cell RNA-seq
- scRNA-seq

1 細胞RNA-seqを用いた論文が急増している



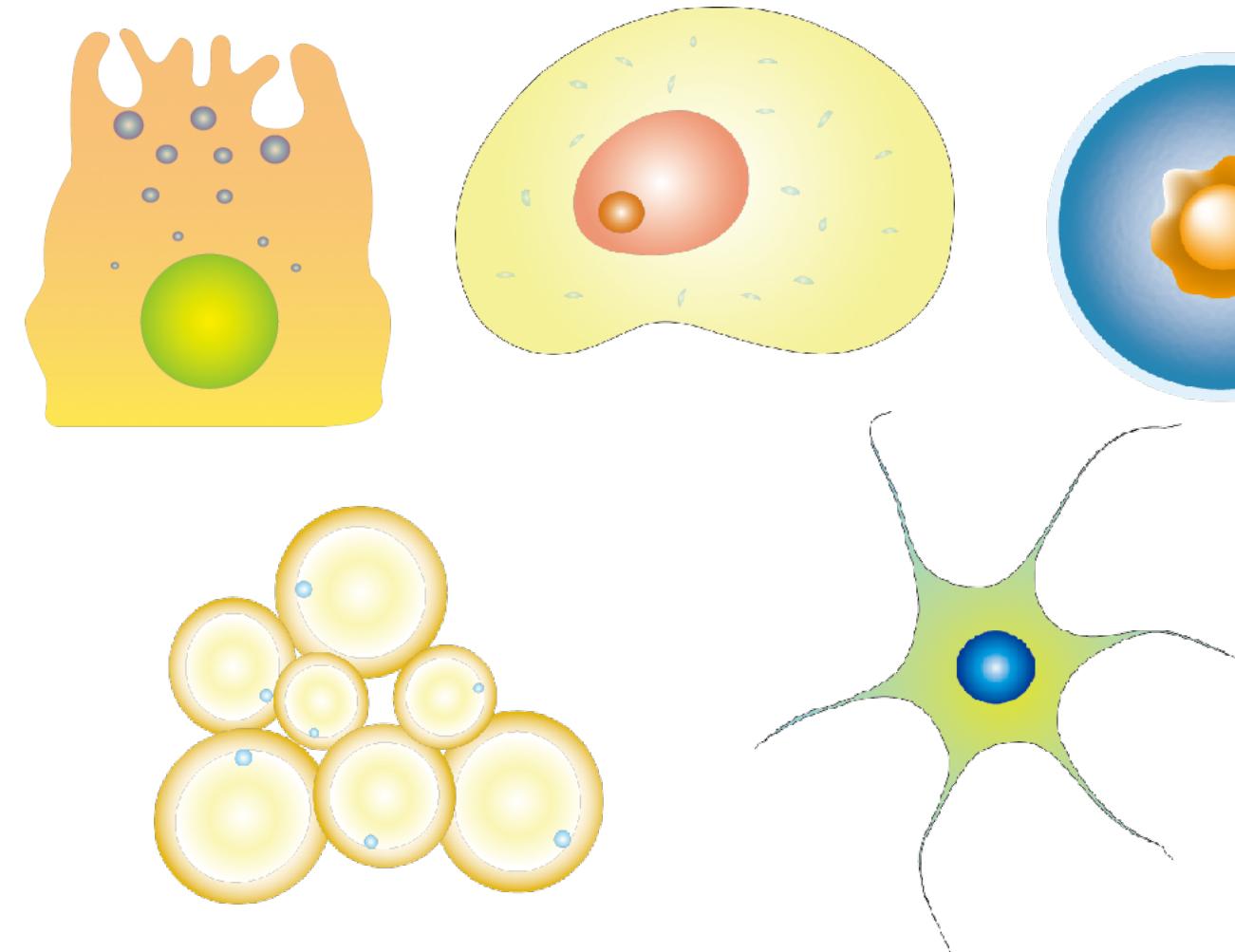
PubMed search on 2021/07/27

単一細胞の遺伝子発現量を計測する技術

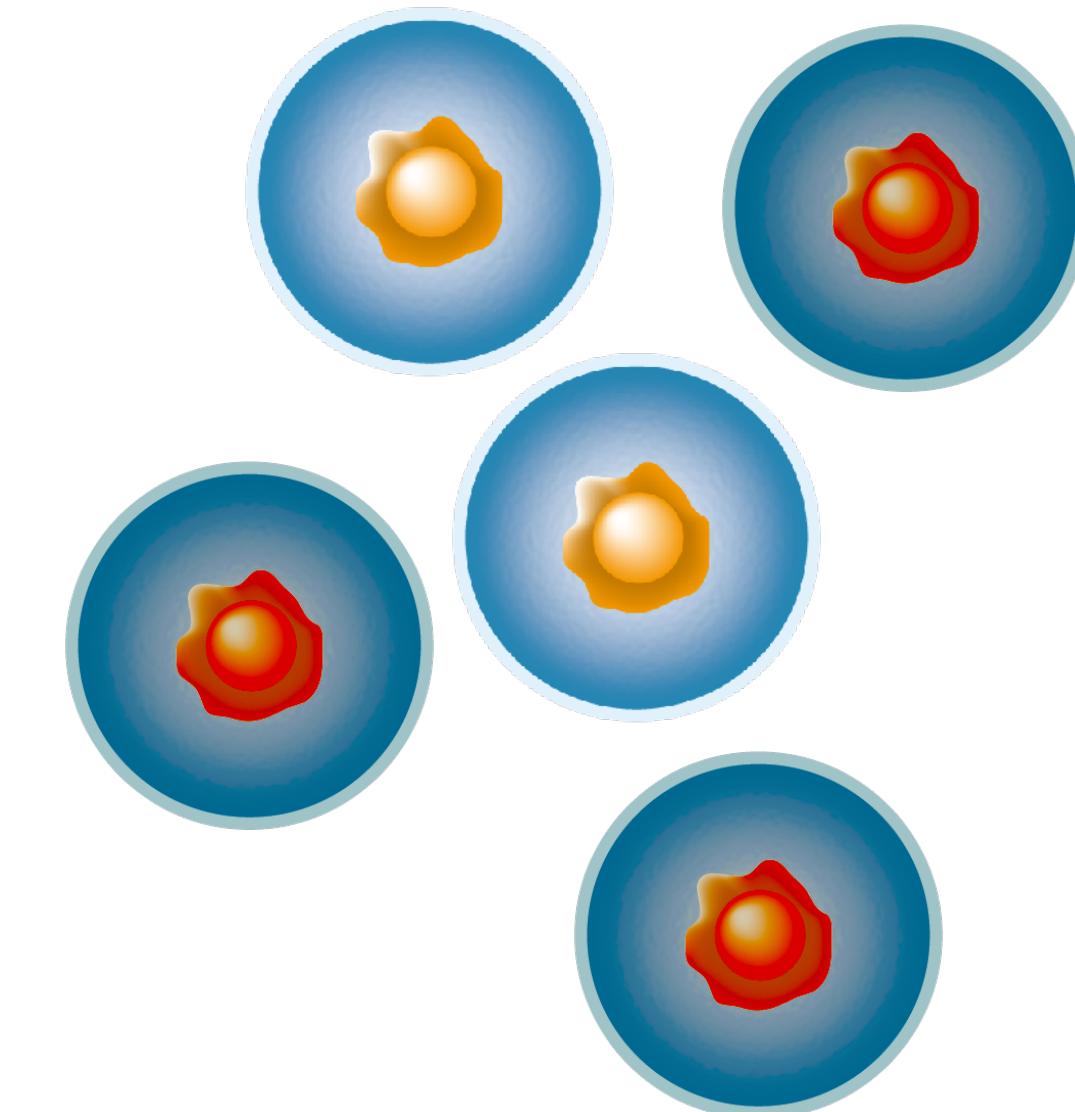


細胞は多様な形態や機能をもつ

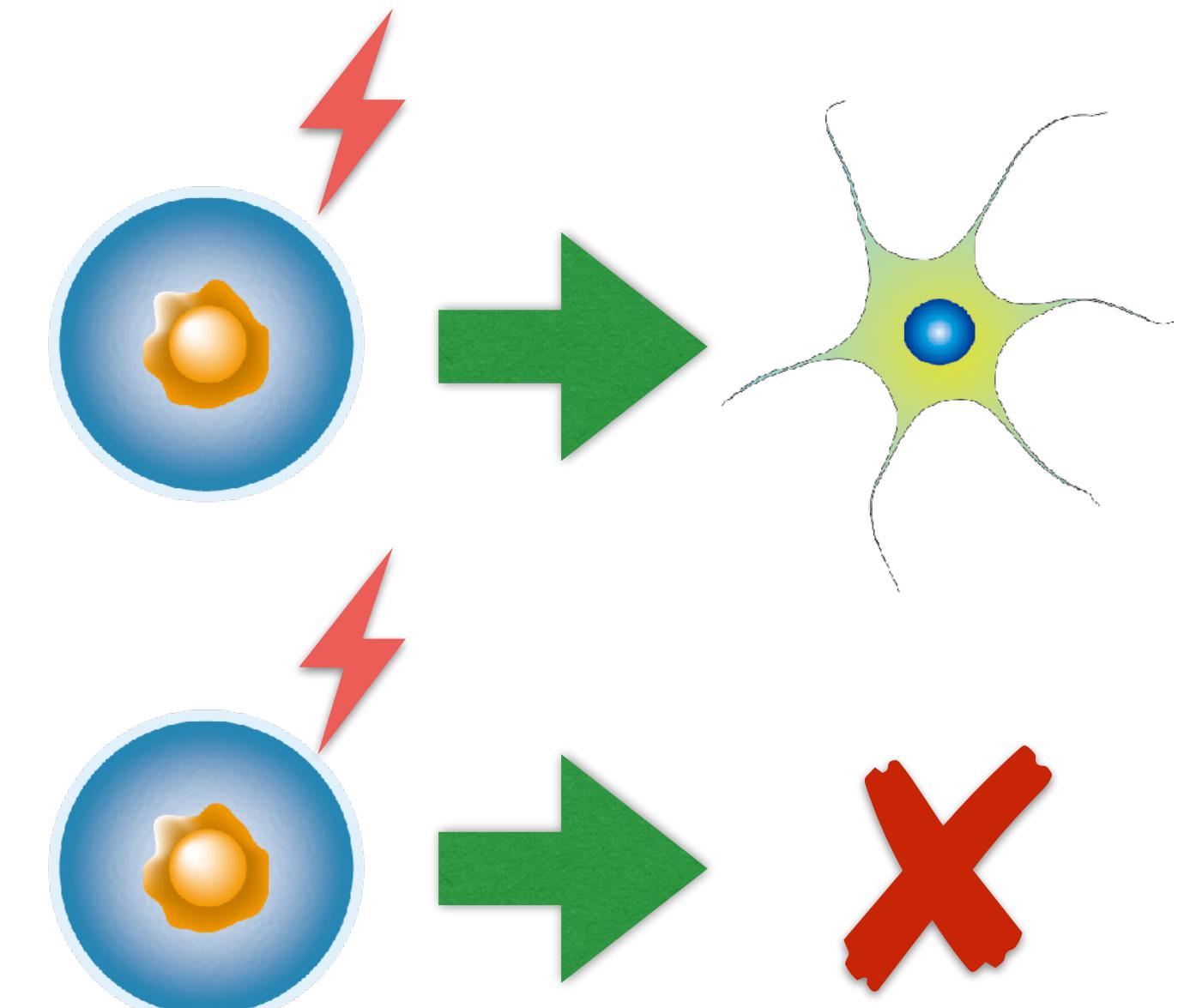
細胞型
Cell type



細胞状態
Cell state

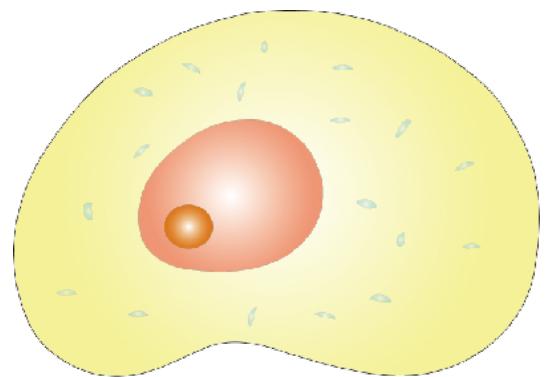


分化ポテンシャル
Differentiation potential

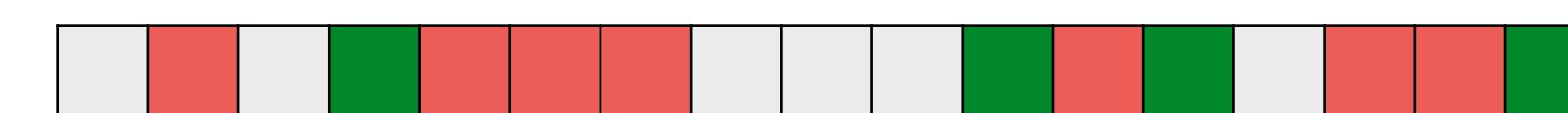
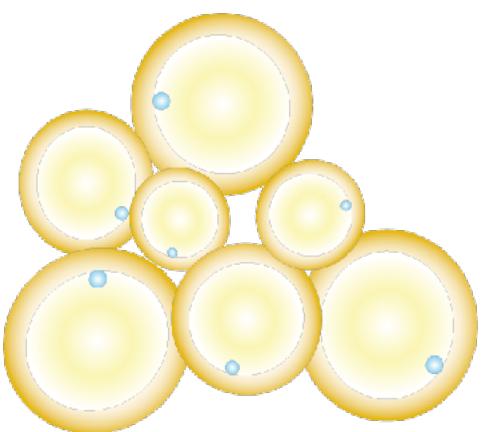
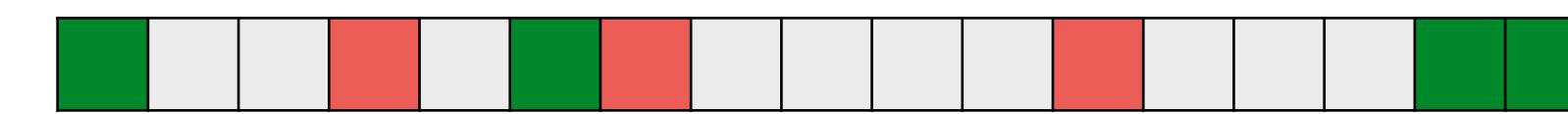
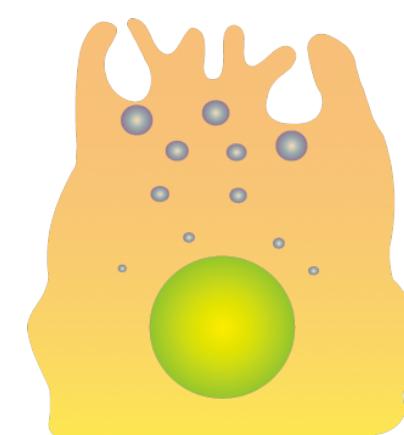
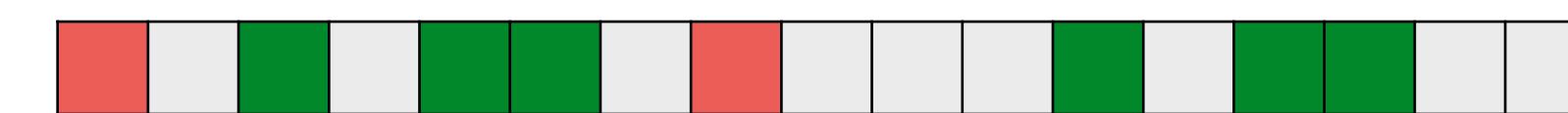


細胞の性質は遺伝子発現に反映されるはず

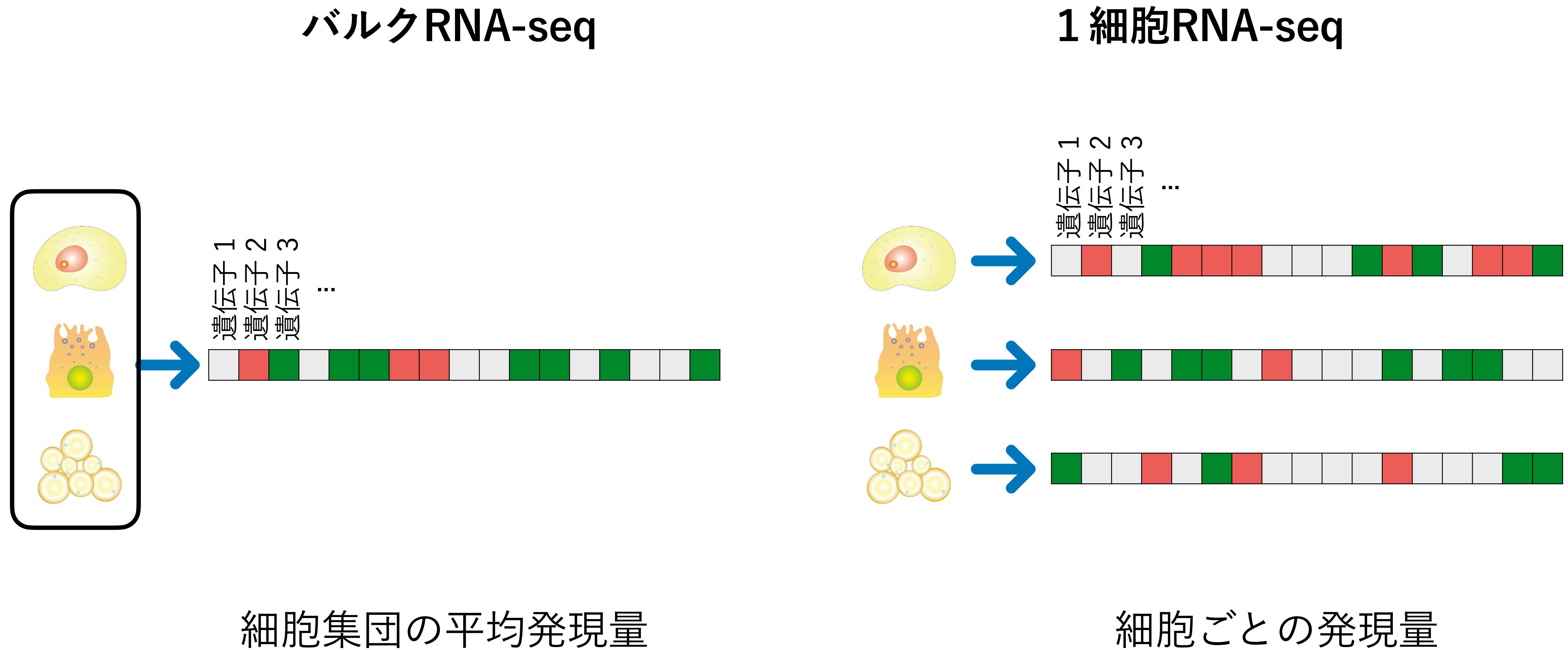
ゲノムは同じ



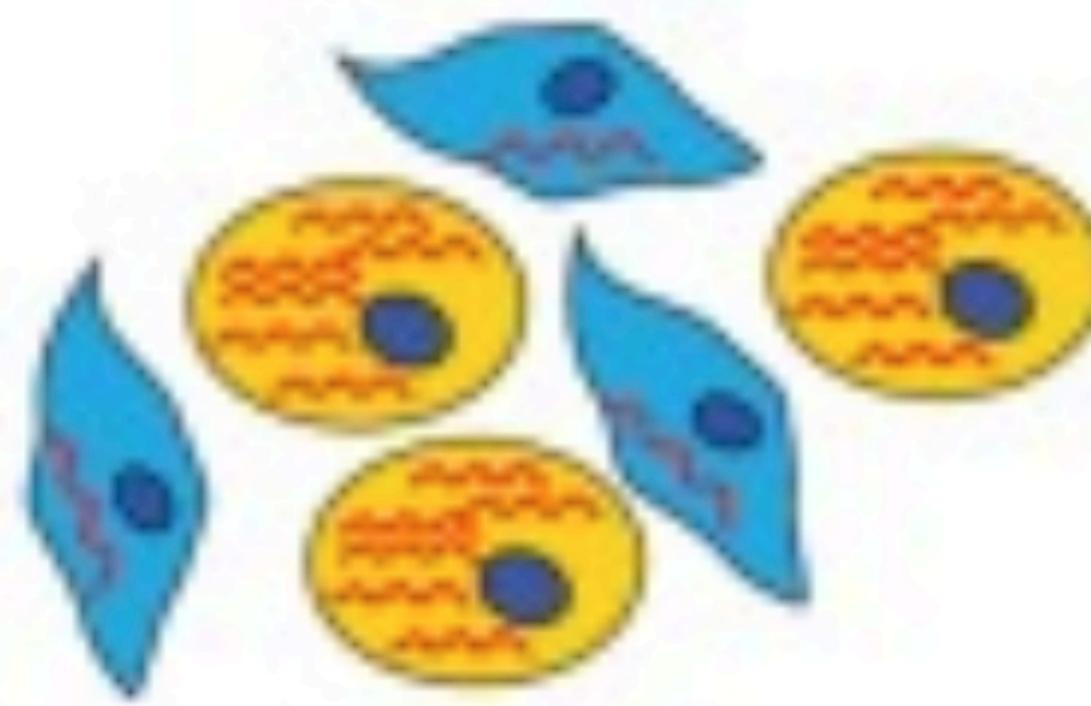
遺伝子発現量は異なる



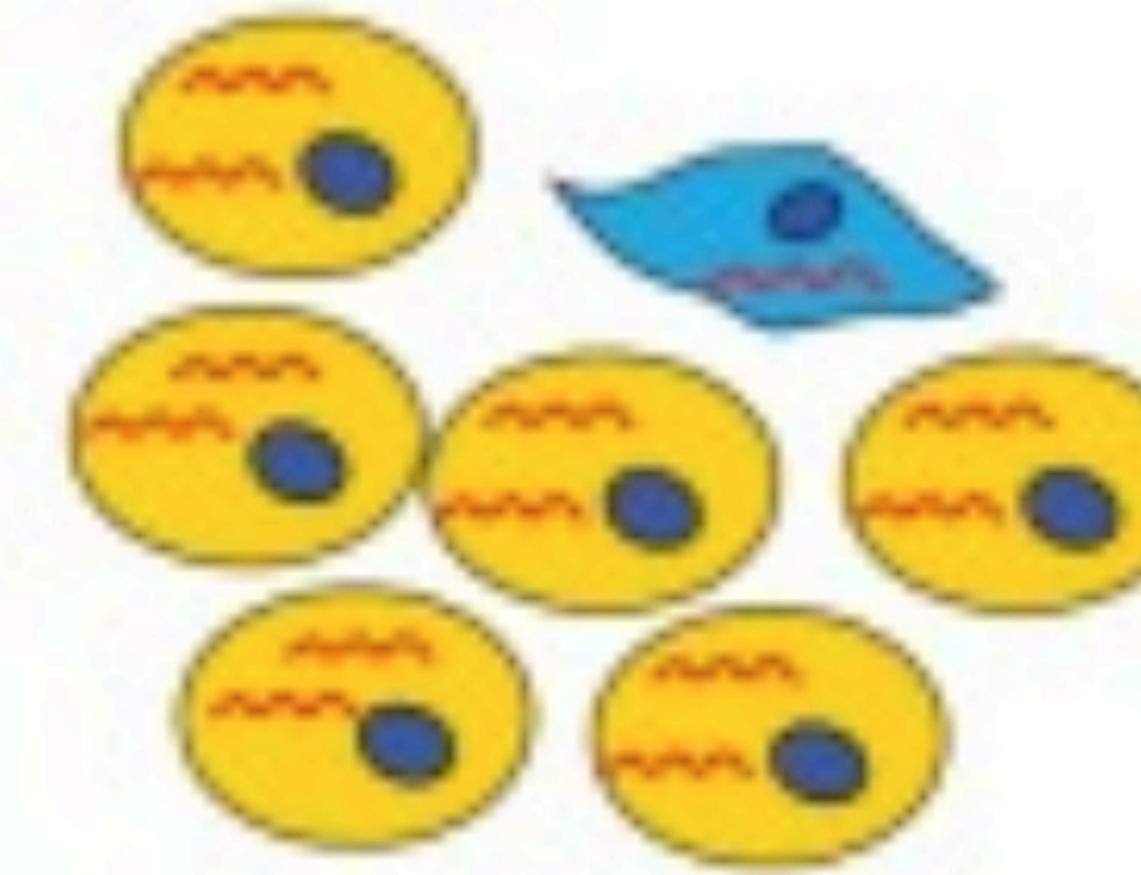
1 細胞解像度 Single-cell resolution



細胞集団は不均一な集団

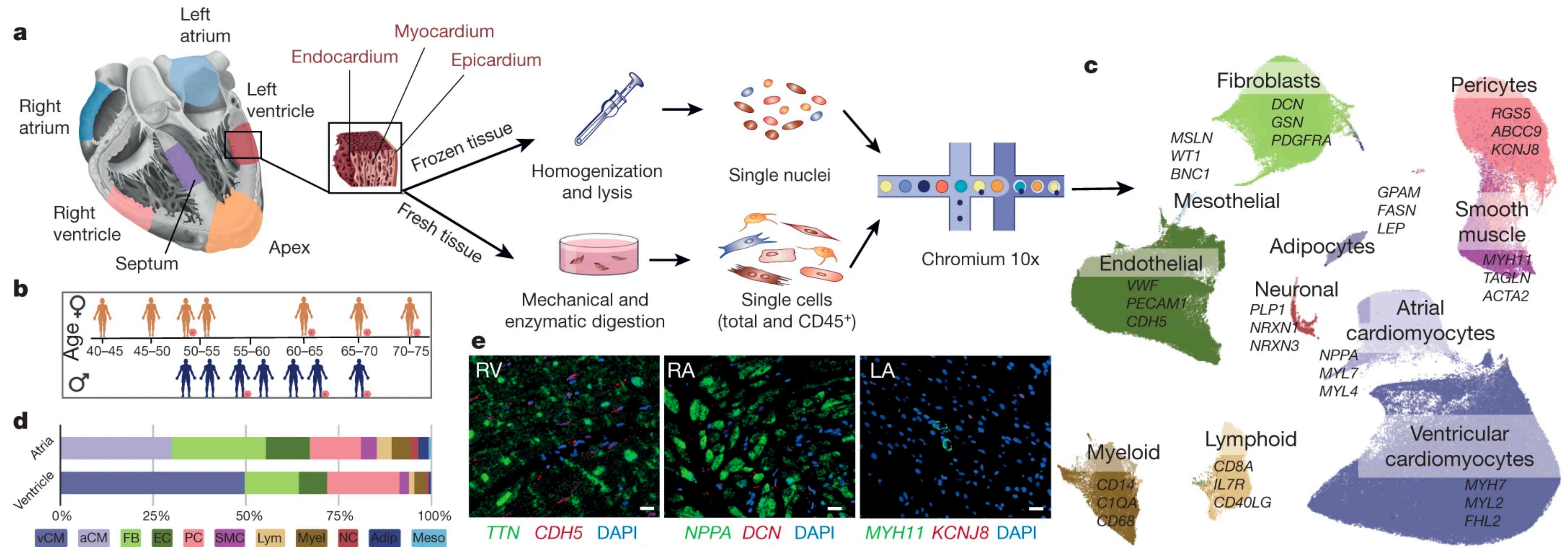


Change in
regulation



Change in
composition

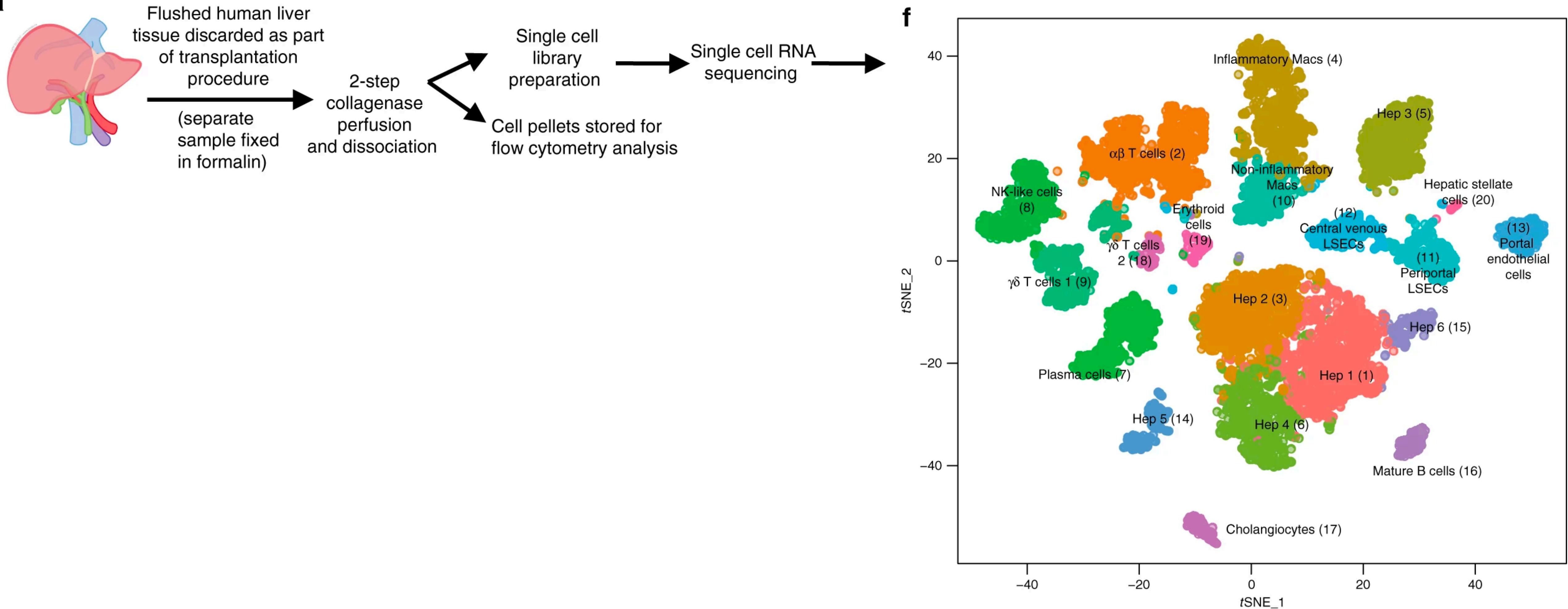
臓器は不均一な細胞からできている（心臓）



"Cells of the adult human heart"

DOI: 10.1038/s41586-020-2797-4

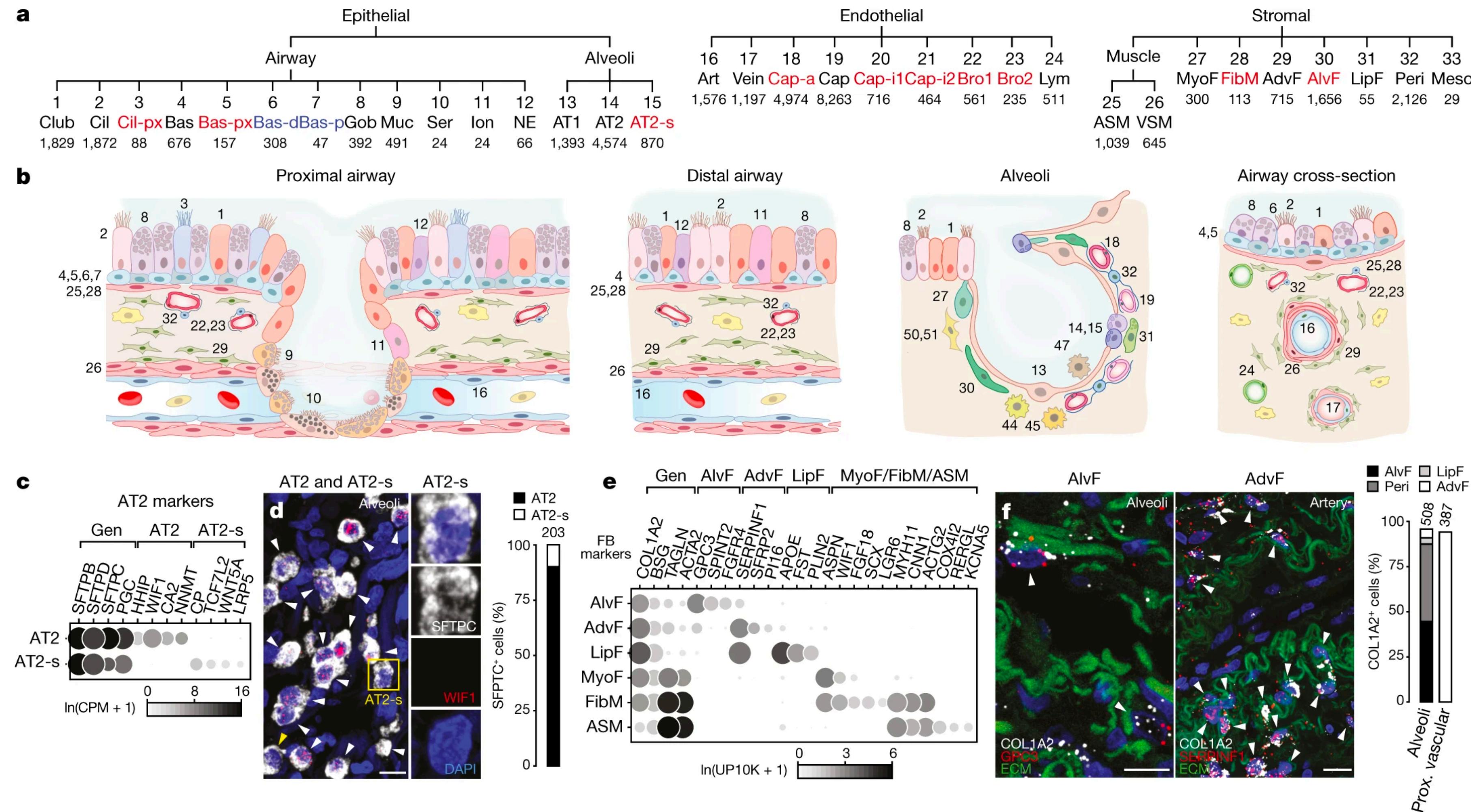
臓器は不均一な細胞からできている（肝臓）



"Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations"

DOI: 10.1038/s41467-018-06318-7

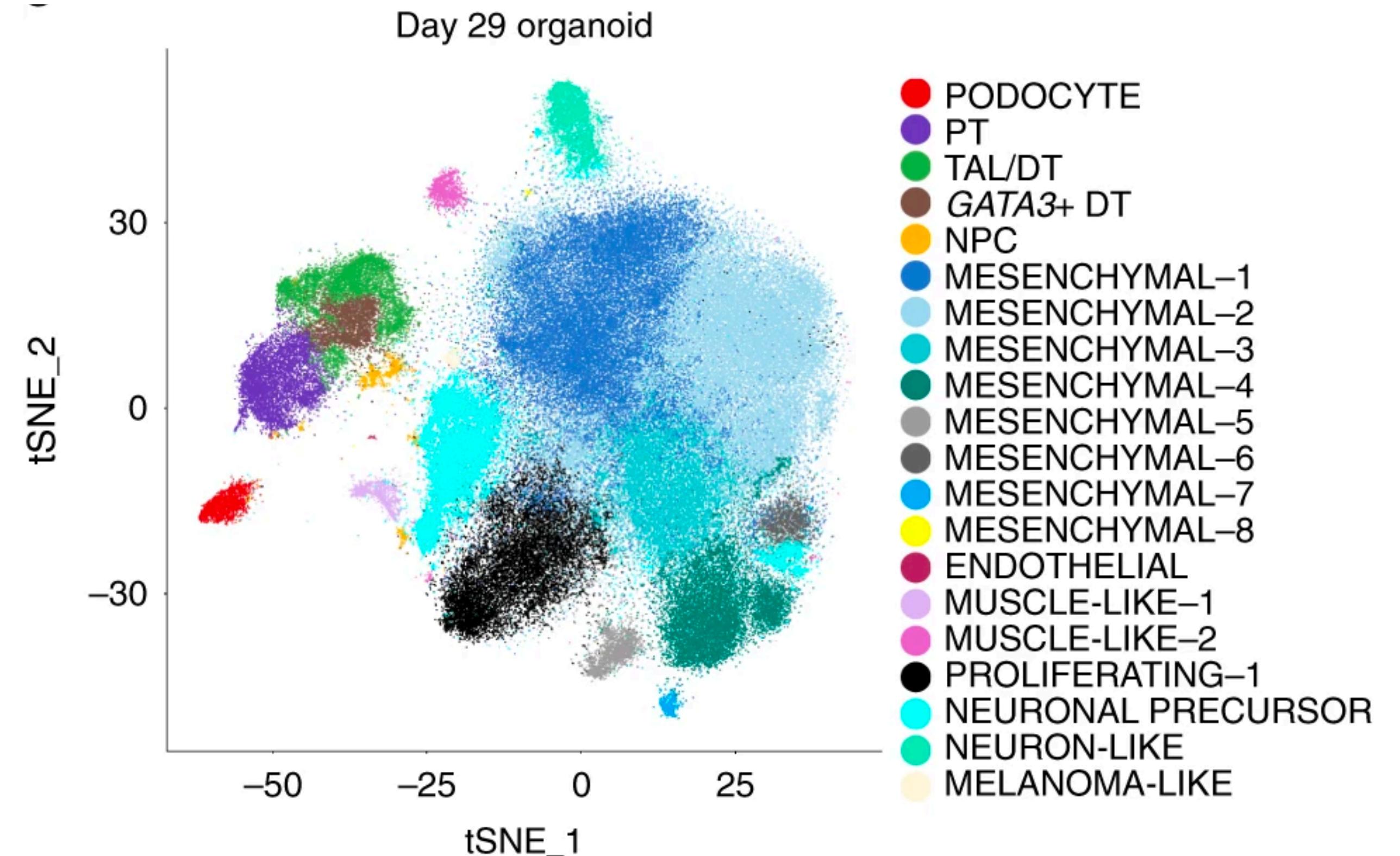
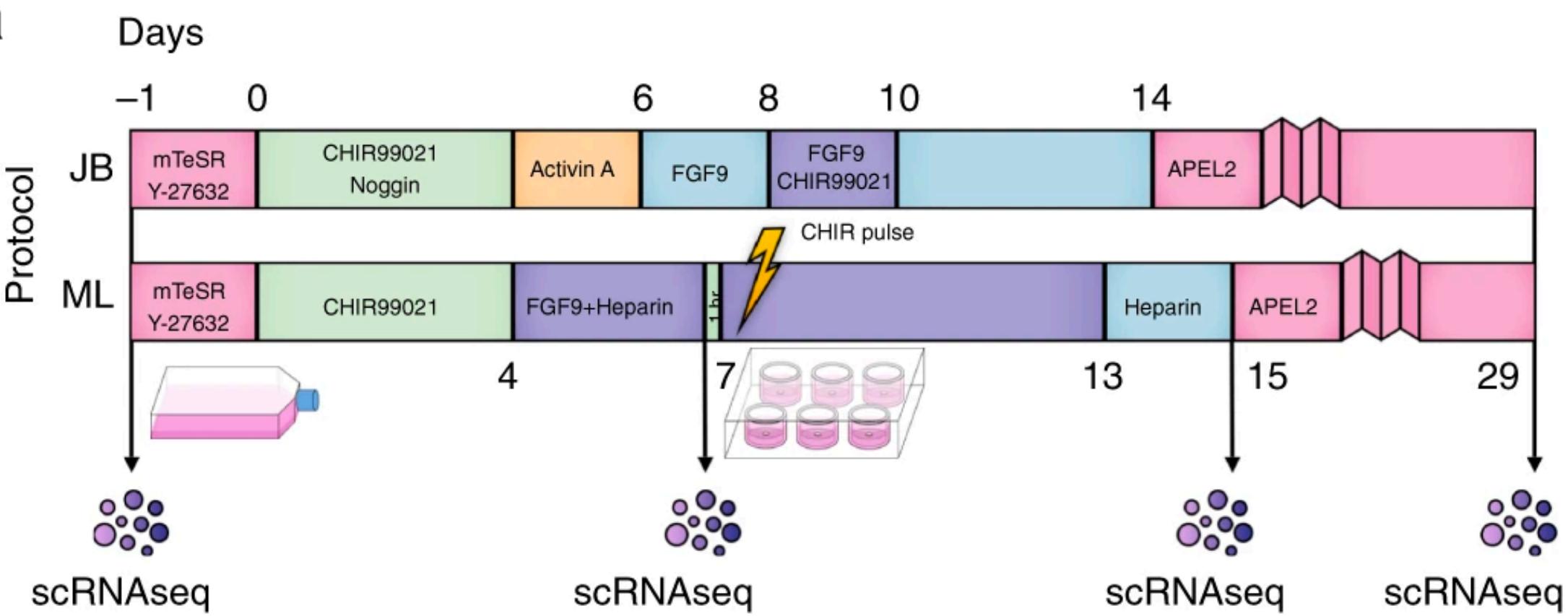
組織幹細胞も見つかる（肺）



"A molecular cell atlas of the human lung from single-cell RNA sequencing"

DOI: 10.1038/s41586-020-2922-4

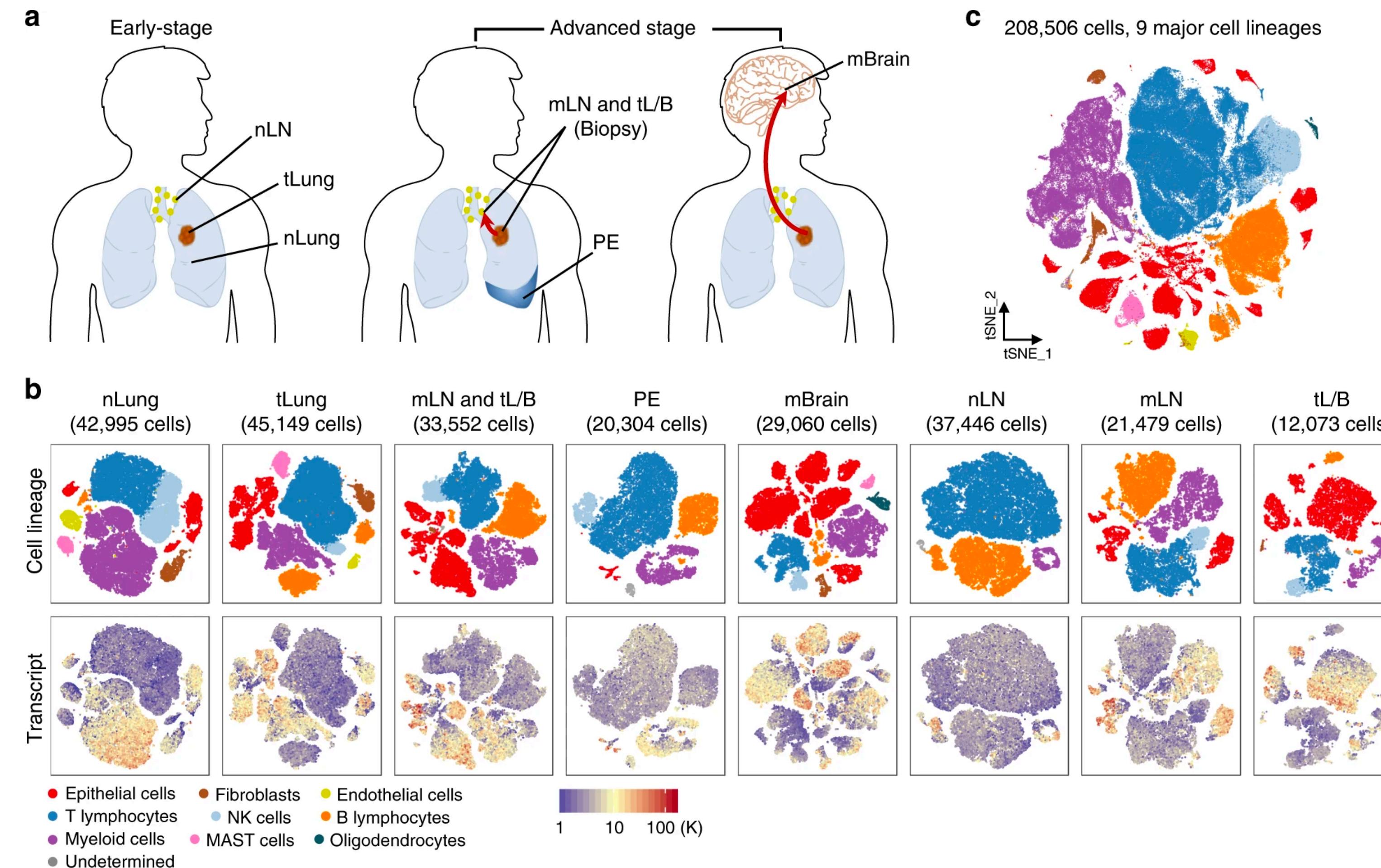
オルガノイド（ミニ臓器）

a

"Single cell census of human kidney organoids shows reproducibility and diminished off-target cells after transplantation"

DOI: <https://doi.org/10.1038/s41467-019-13382-0>

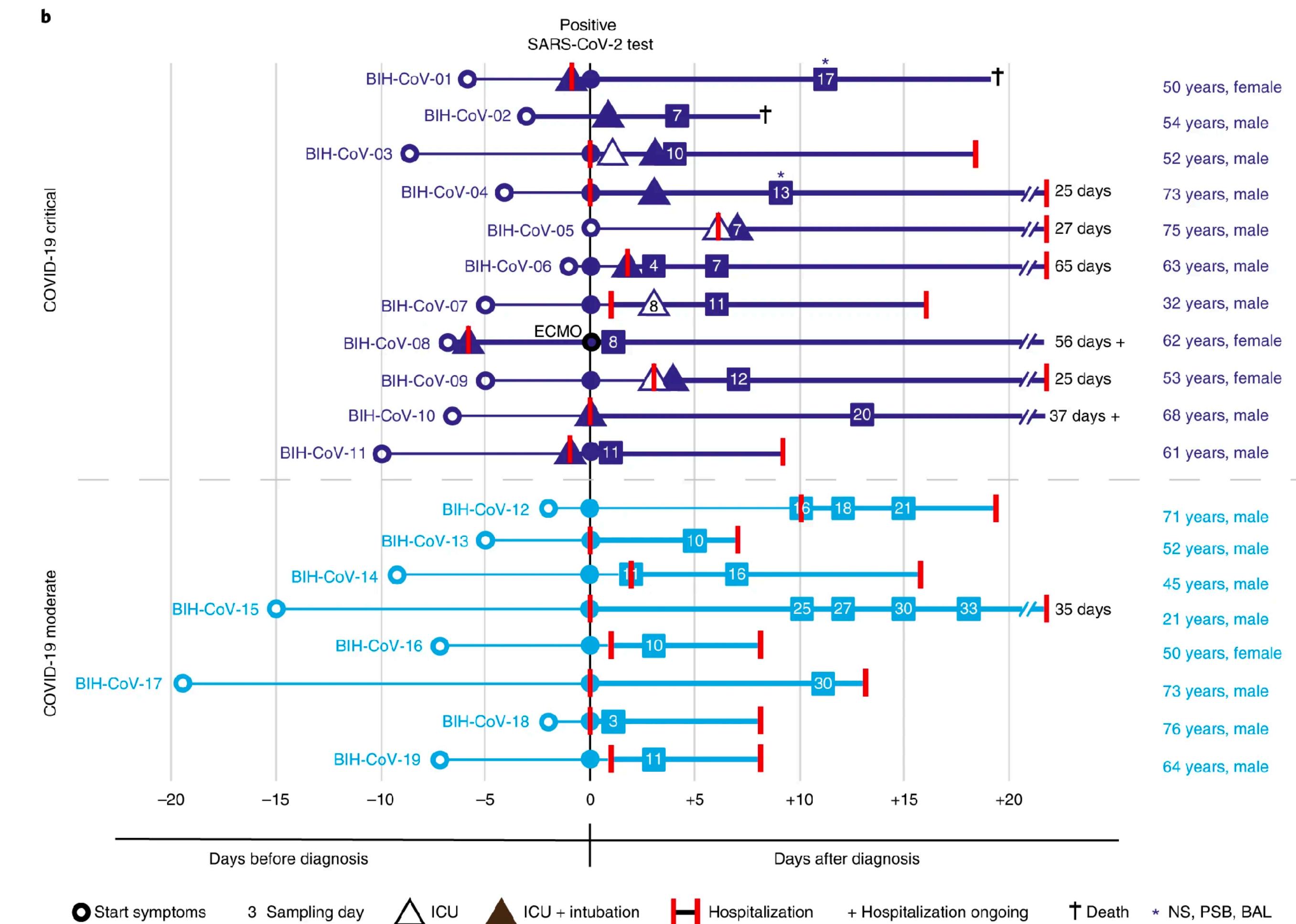
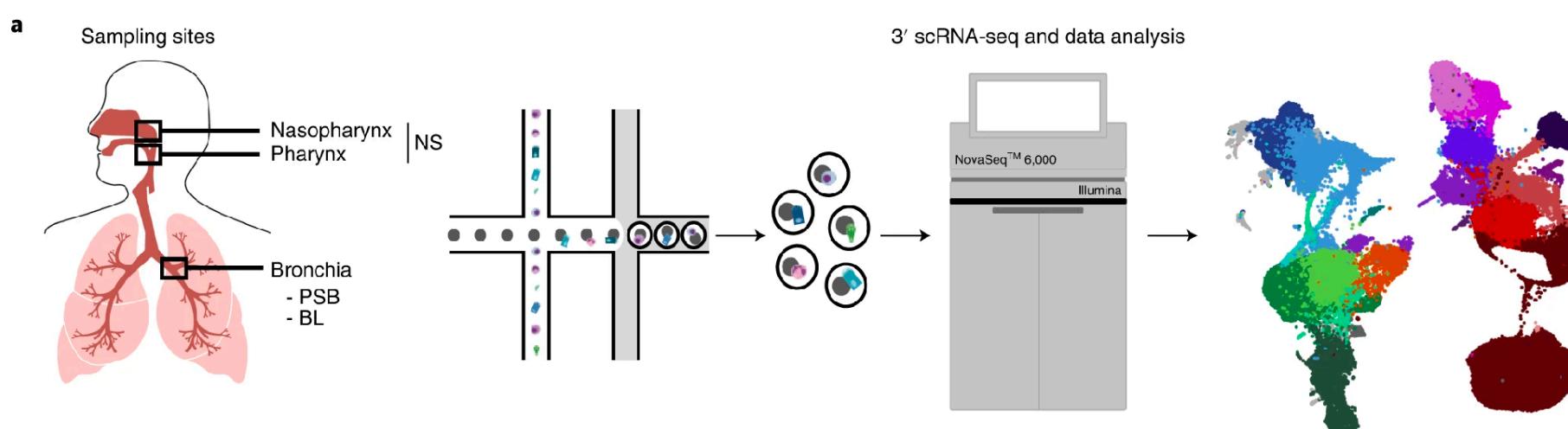
がん組織



"Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma"

DOI: 10.1038/s41467-020-16164-1

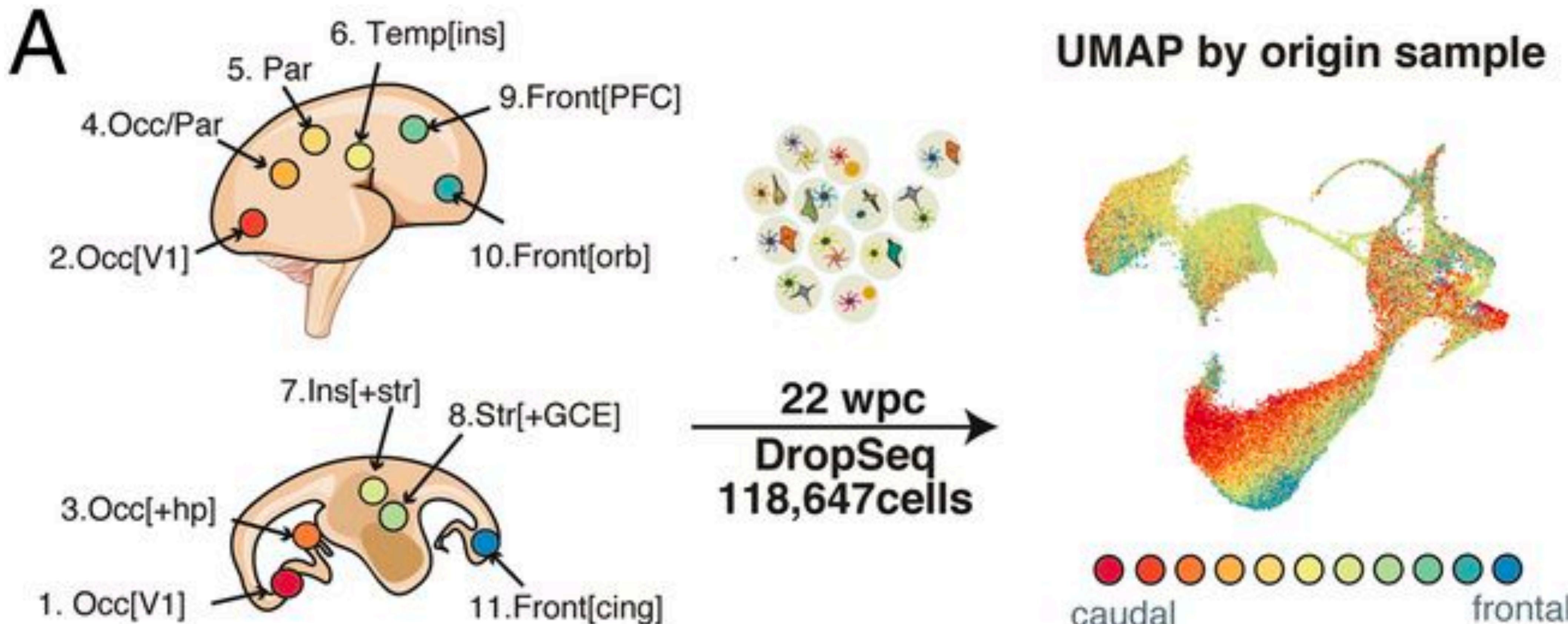
COVID-19重症度と鼻咽頭・気管支の細胞集団



"COVID-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis"

DOI: 10.1038/s41467-020-16164-1

脳領域の細胞



"Early role for a Na⁺,K⁺-ATPase (ATP1A3) in brain development"

DOI: 10.1073/pnas.2023333118

国際コンソーシアムでも1細胞RNA-seqを活用

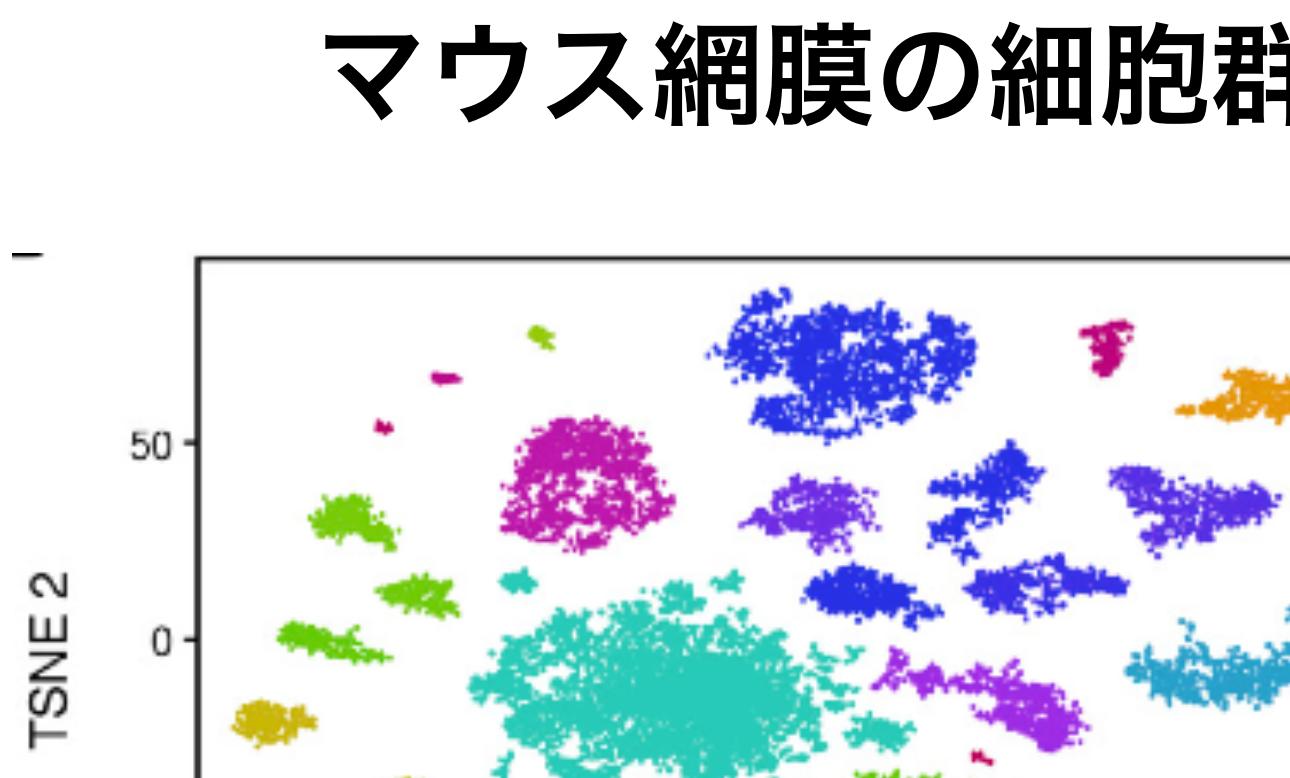
Human Cell Atlas
(HCA)

Human Tumor Atlas Network
(HTAN)

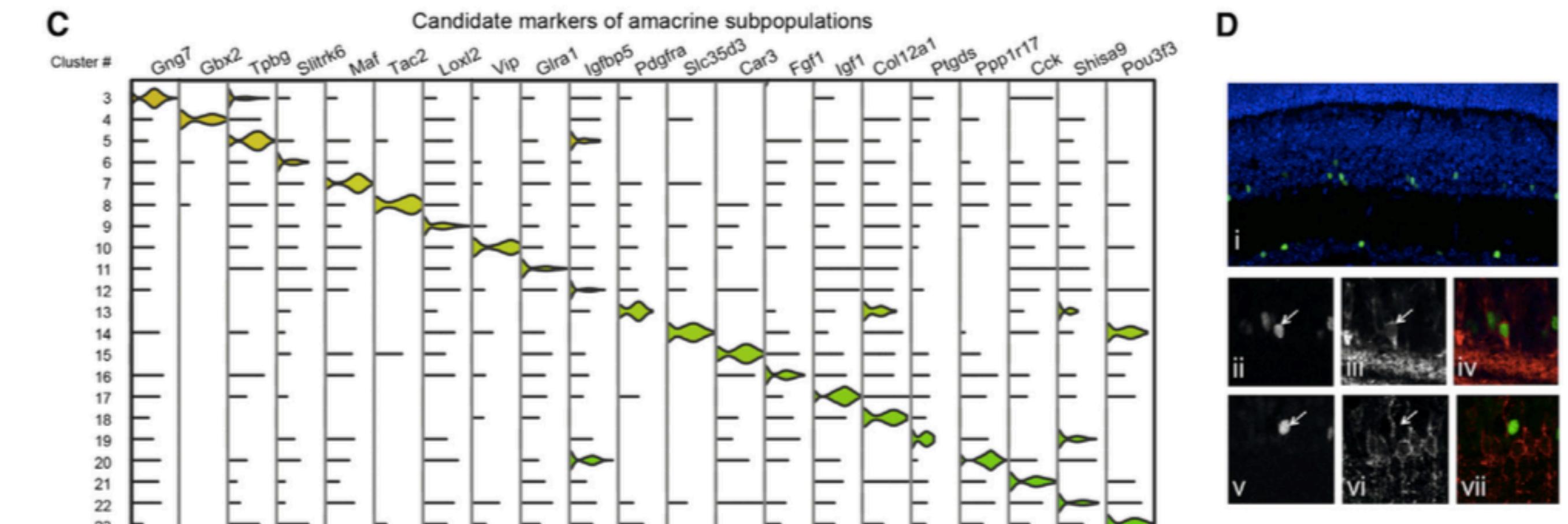
sc-eQTLGen consortium

不均一な細胞集団構成の理解

- 事前知識・マーカーがなくても細胞集団の組成（細胞型・サブタイプ）を明らかにできる
- e.g., 不均一な細胞が混合した組織、オルガノイド

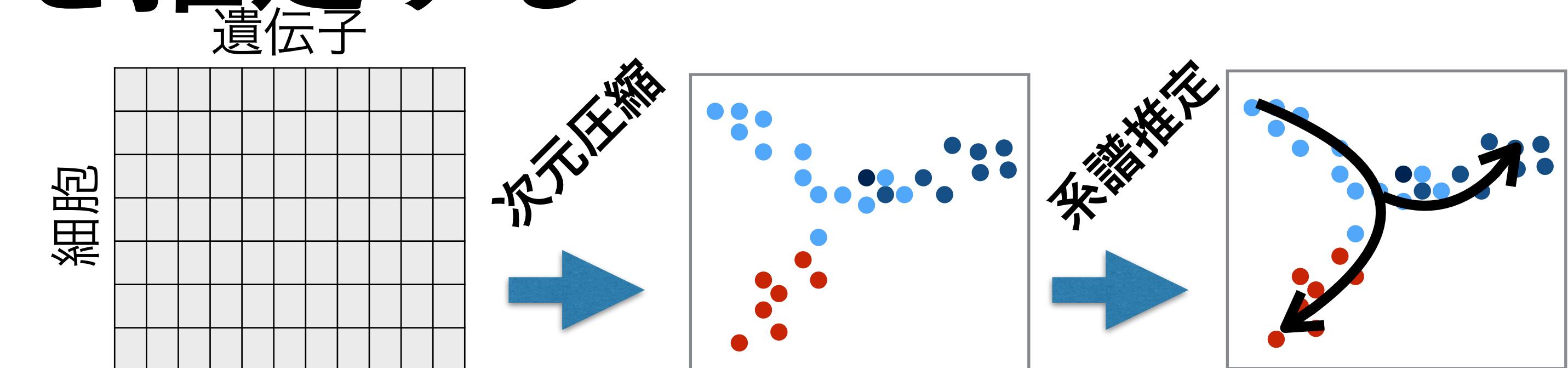


新規サブタイプの発見・マーカーの同定

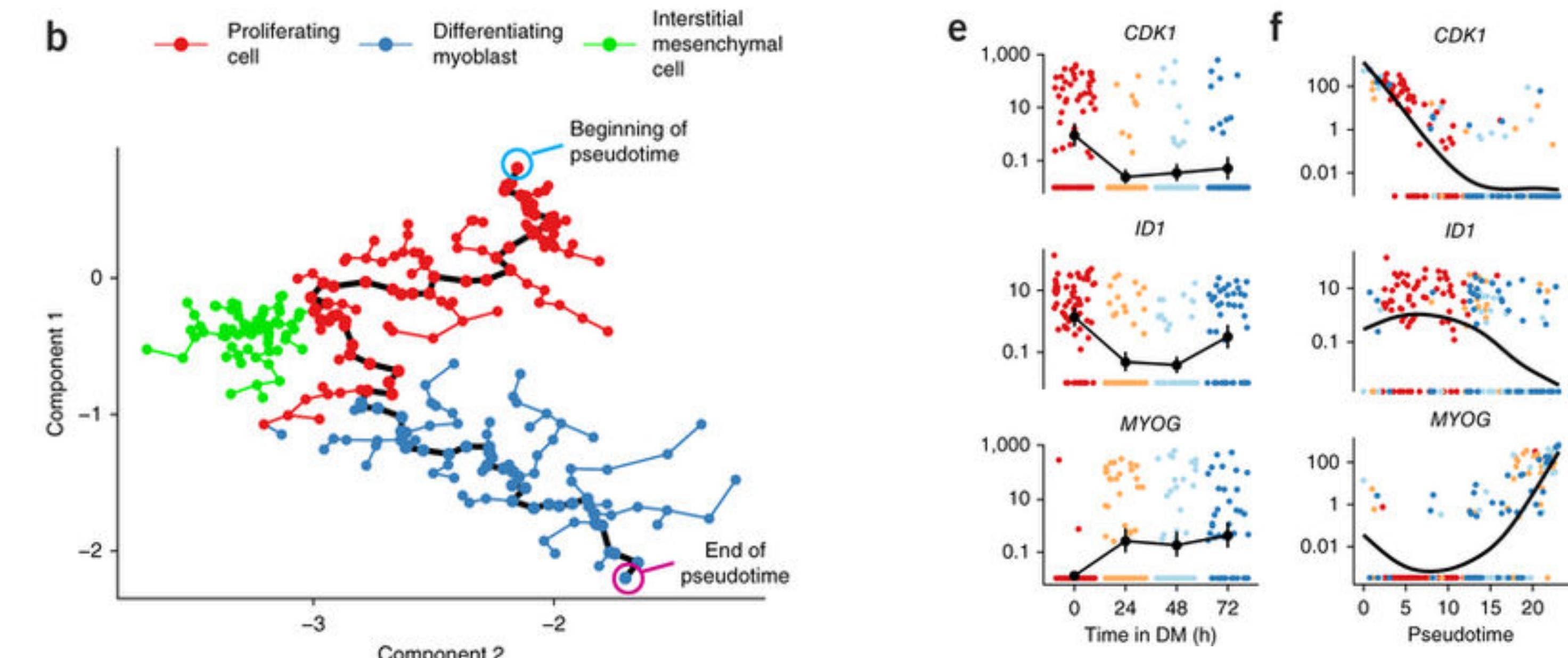


主観的時間・系譜を推定する

- 細胞の主観的時間（擬時間）や系譜を、時間の情報なしに再構築する
- 不均一な細胞が混在する系で細胞状態の時間発展を再現できる
 - e.g. 細胞分化、刺激応答

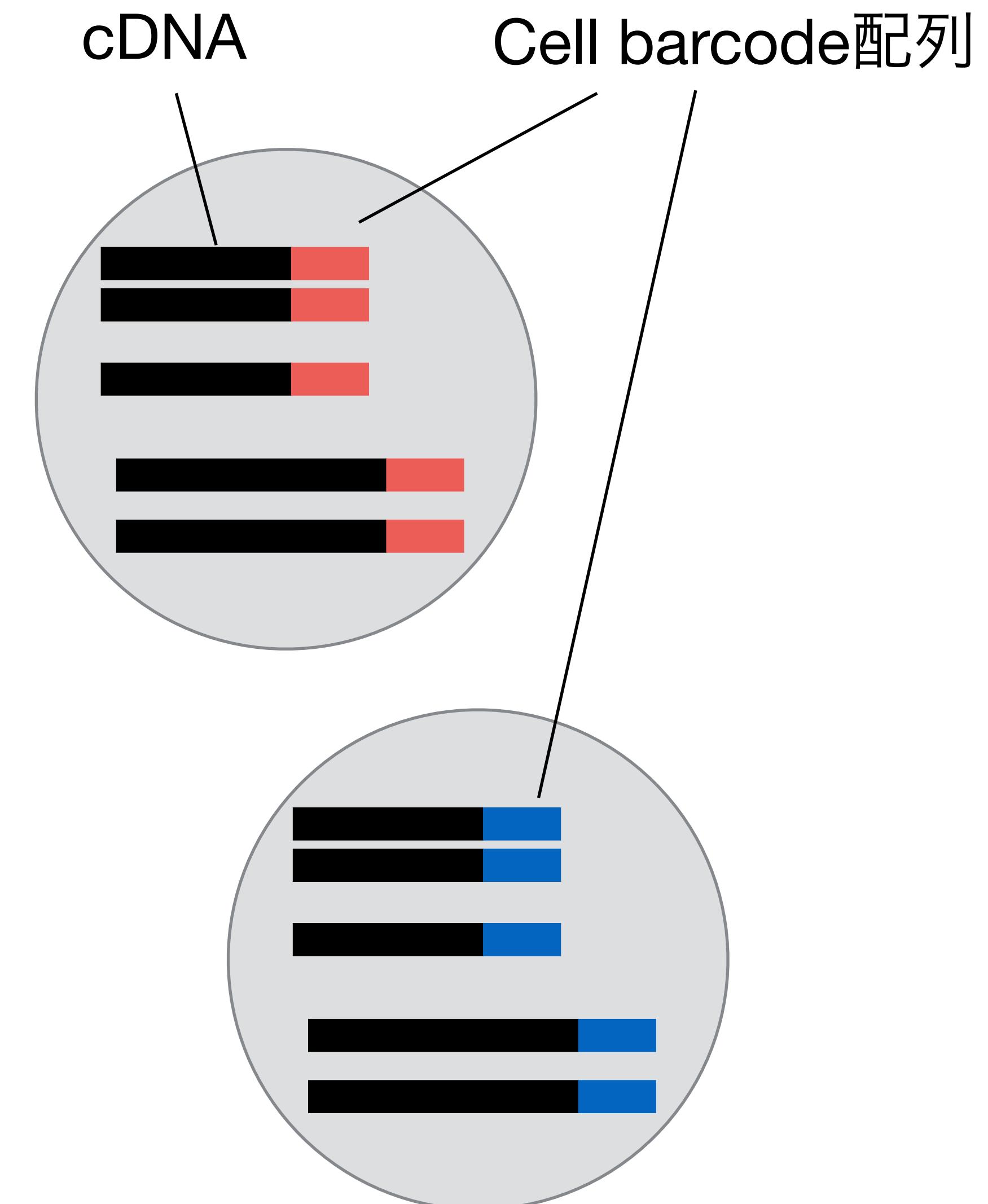


細胞を分化進行度に応じて並べ替える
(線維芽細胞から筋芽細胞への分化誘導)

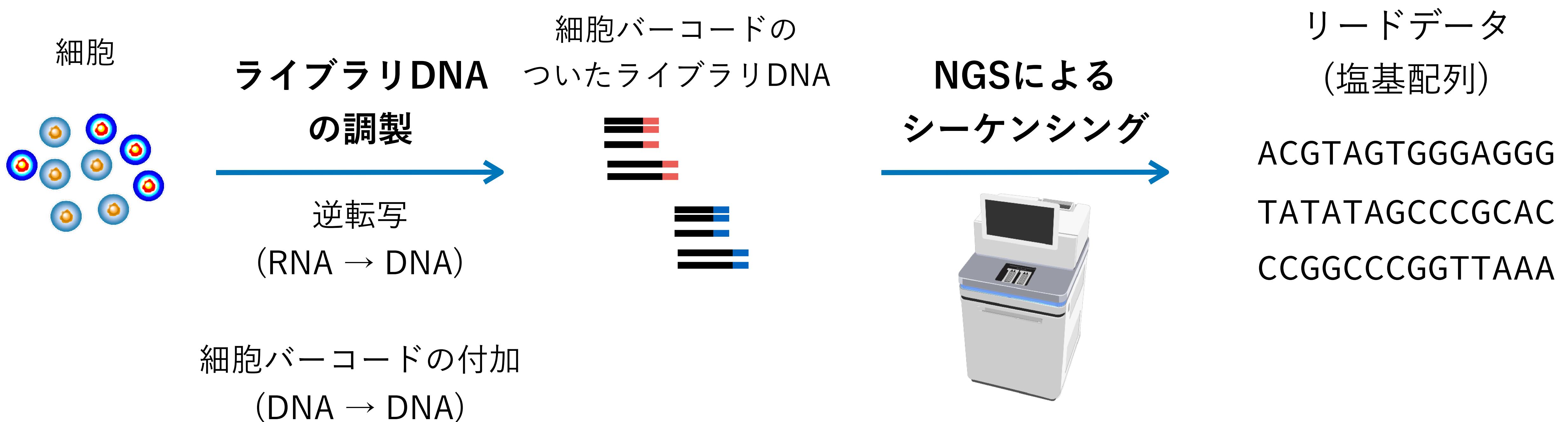


scRNA-seqの仕組み (Cell barcode)

- 細胞単位での超マルチプレックス化
- 実験: マルチプレックスの前に細胞ごとに異なる目印 (Cell barcode配列) が付与
- データ解析: Cell barcode配列に基づいてNGSのリードを細胞に振り分ける

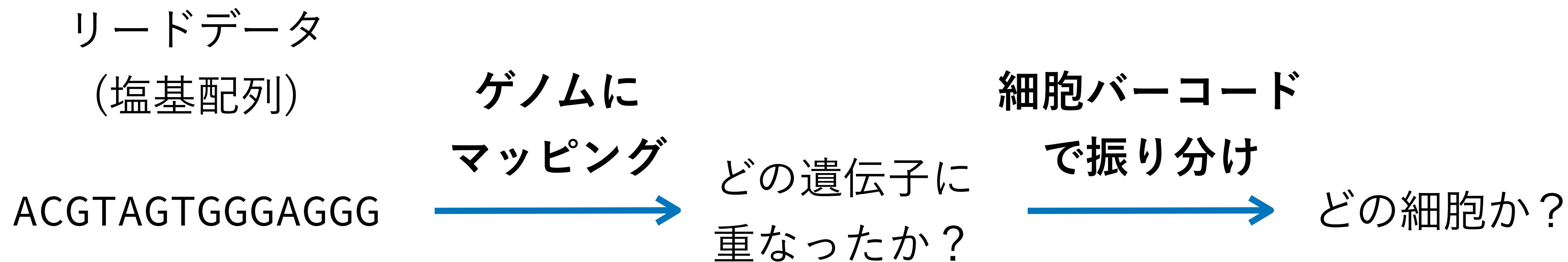


細胞内のRNAが塩基配列データになるまで

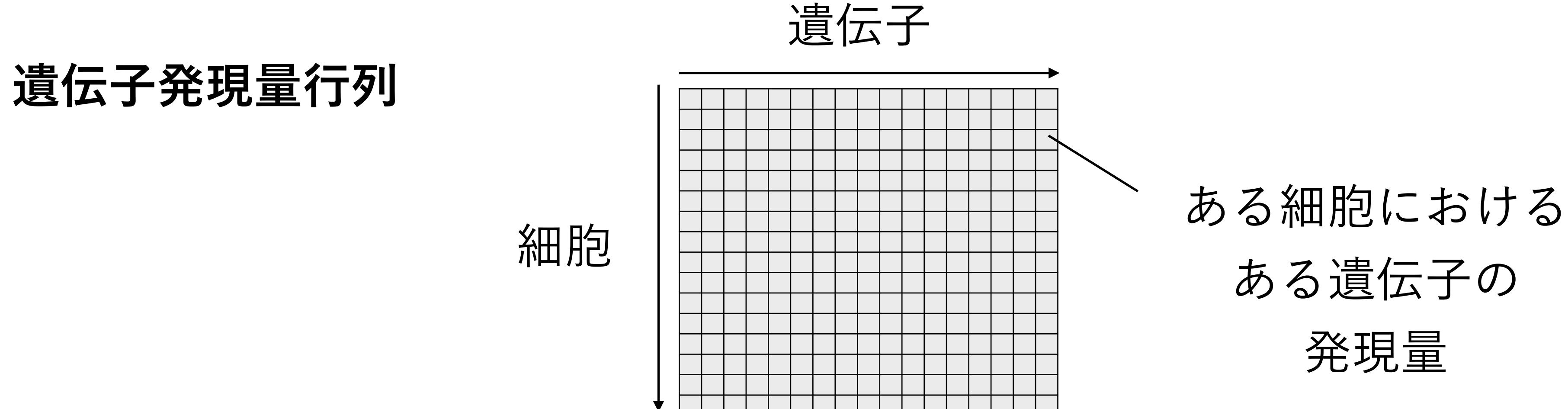


塩基配列データが遺伝子発現量行列になるまで

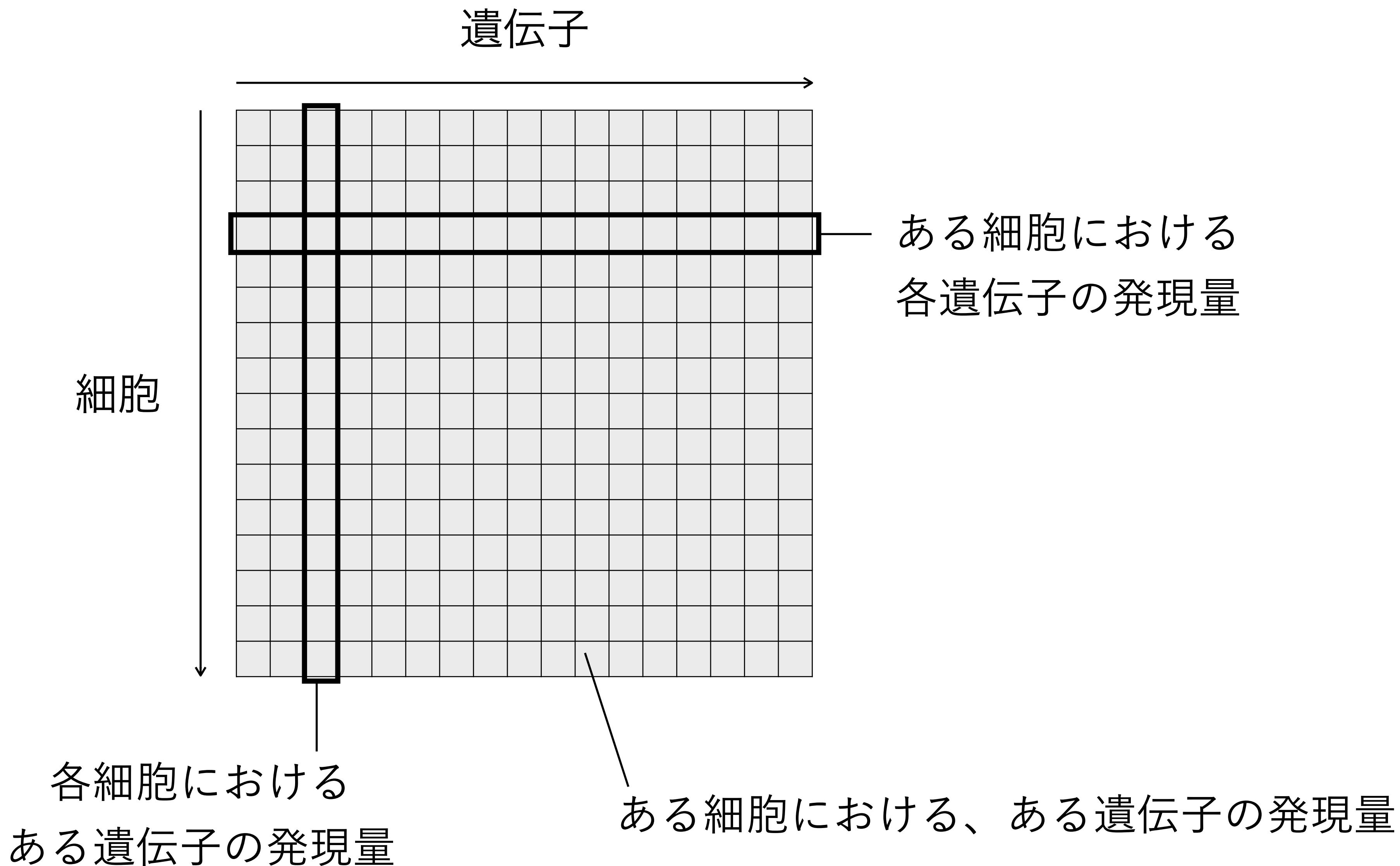
- 個々のリードがどの細胞・どの遺伝子の由来かを当てる



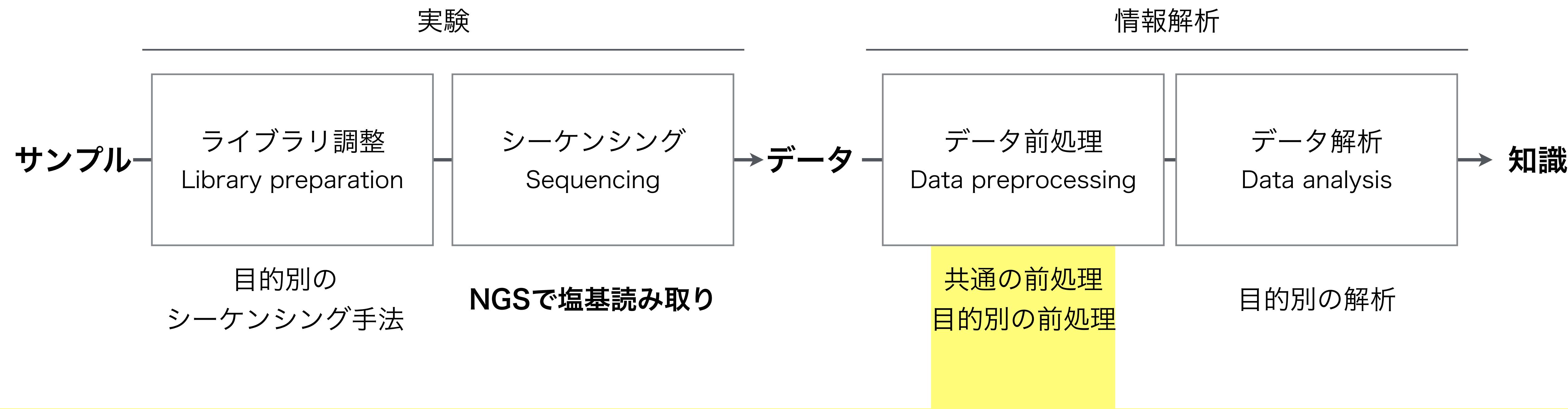
- 細胞ごとに各遺伝子に割り当てられたリードの数をカウントする



1 細胞RNA-seqのデータは遺伝子発現量行列



シングルセル解析の流れ



不要な細胞の除去

低品質・empty (検出遺伝子数が低い)

ダブルレット (検出遺伝子数が高い)

正規化

細胞間でのリード数・
UMI数の差を補正する
(細胞ごとに合計値で割
る、log変換するなど)

スケーリング

遺伝子間での発現量の値
の範囲を揃える
(次元圧縮、クラスタリ
ング、ヒートマップなど
で高発現遺伝子の影響を

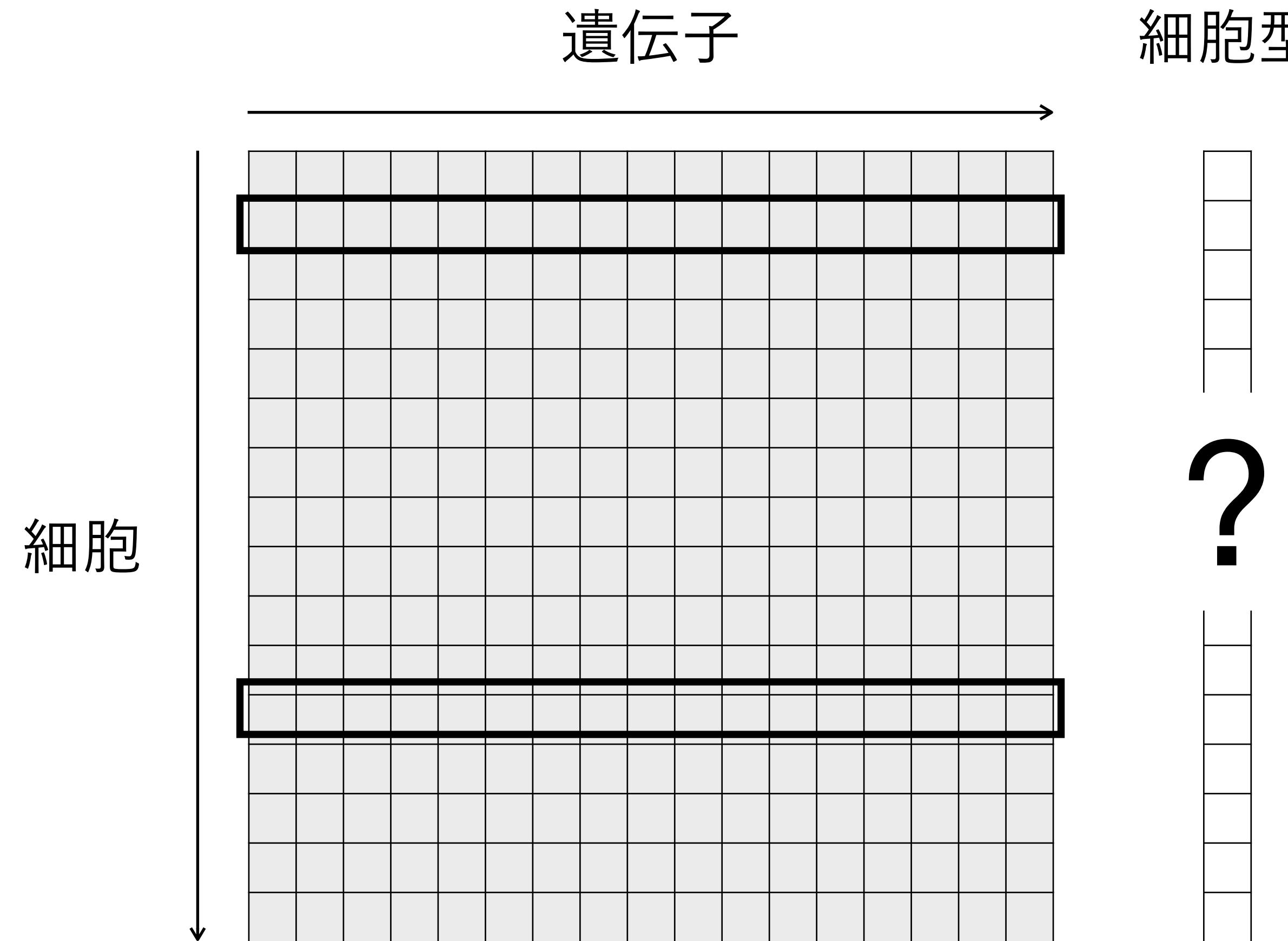
遺伝子の選択

後の解析に都合の良い遺伝子を選びたい
(高変動遺伝子、高Gini係数遺伝子など)

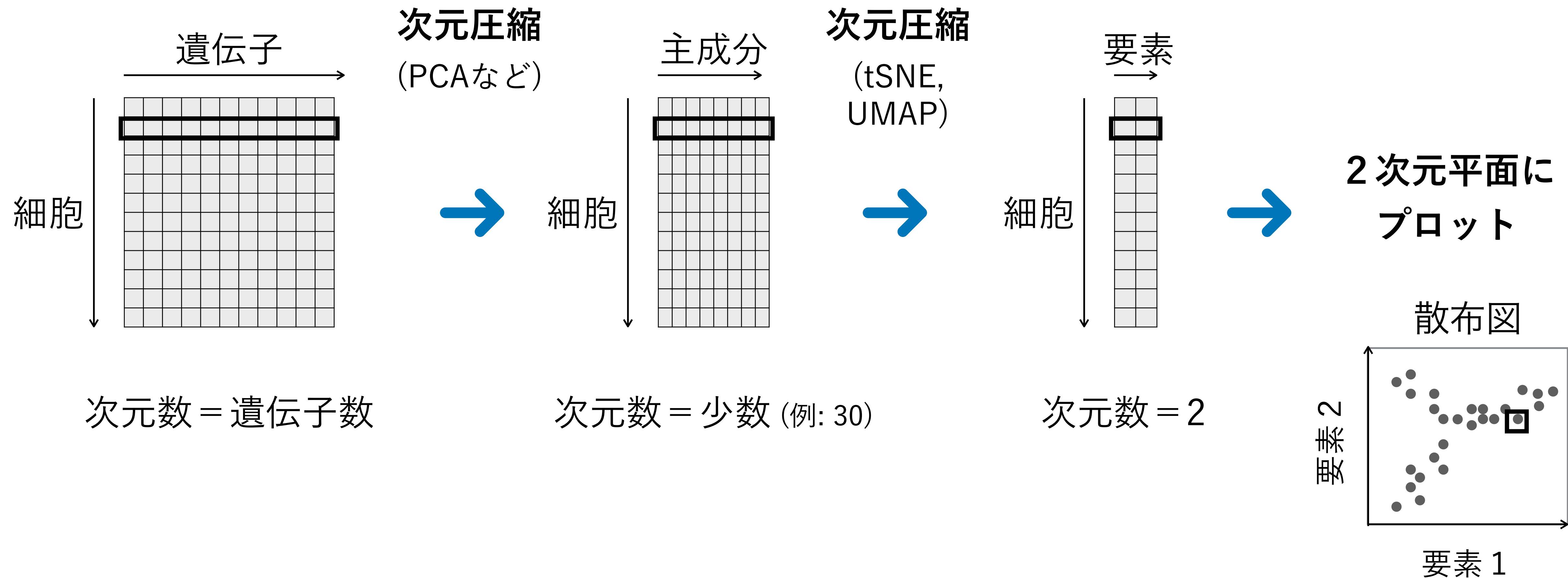
シングルセル解析の流れ



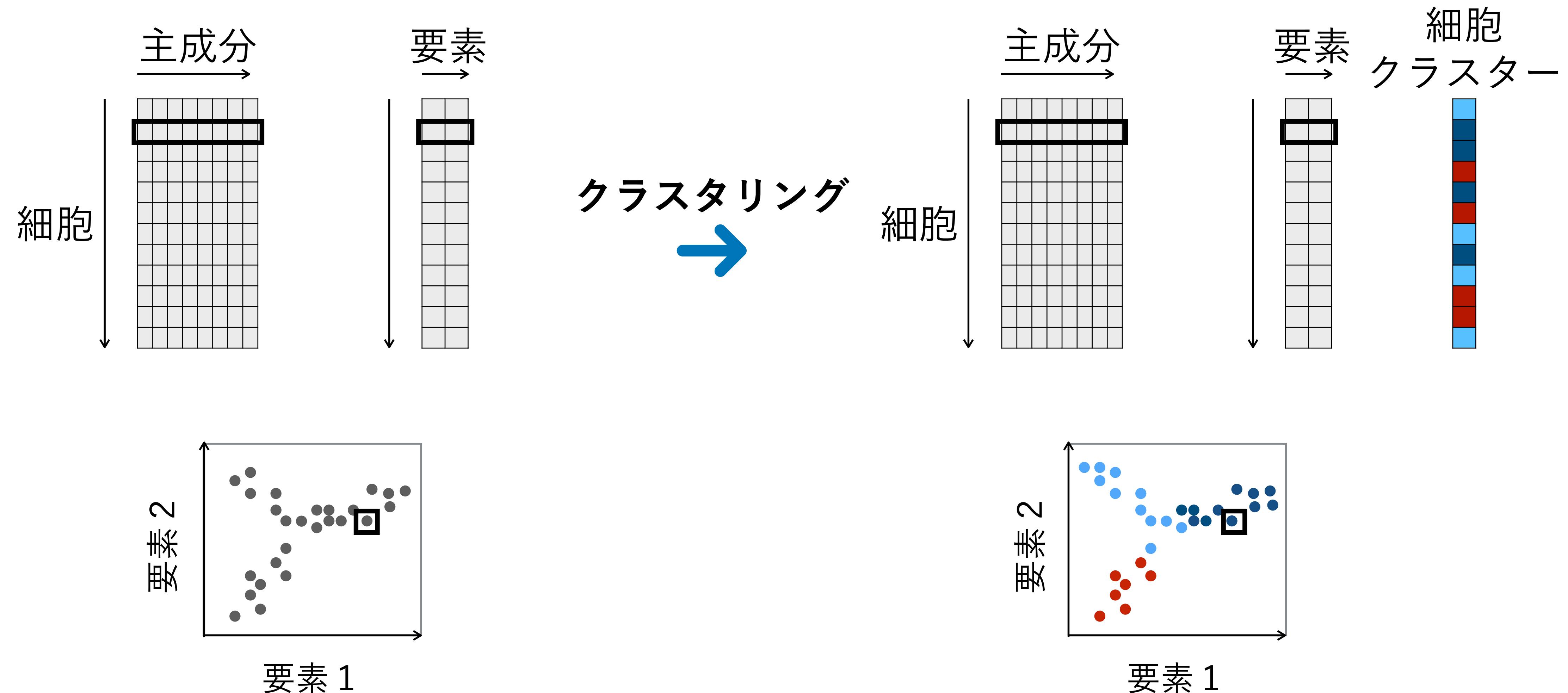
細胞型がわからない



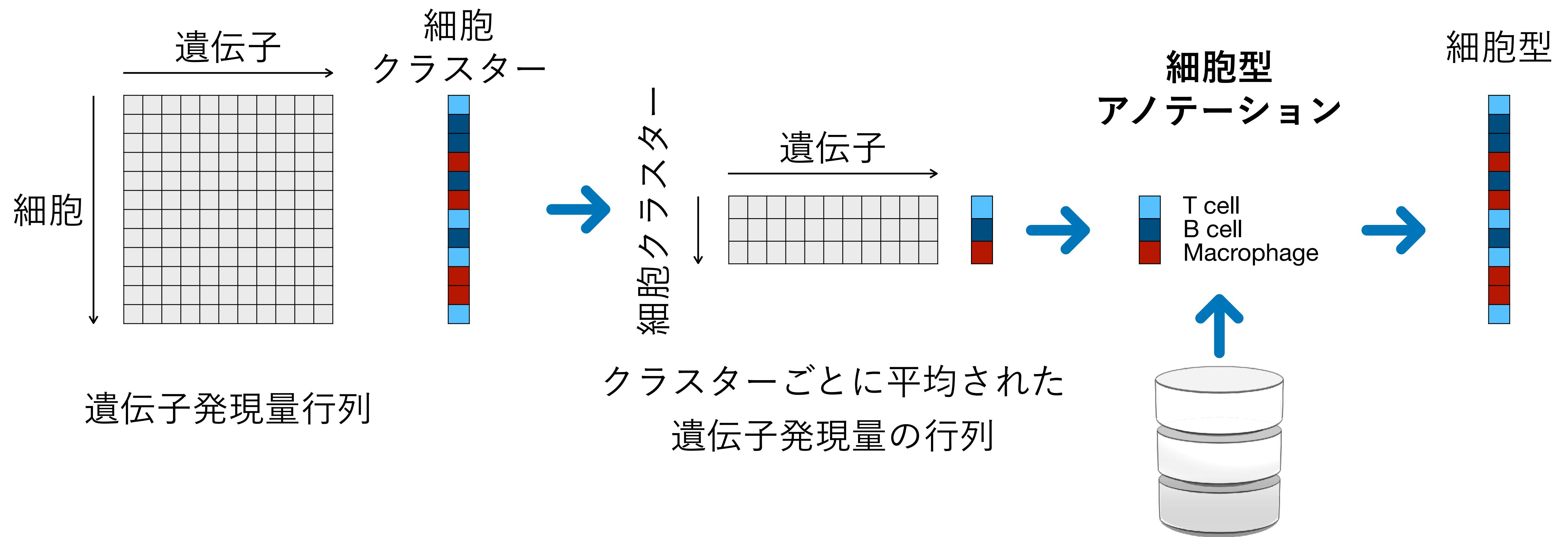
細胞の次元圧縮：細胞の類似関係がわかる



細胞のクラスタリング：細胞をグループ分け



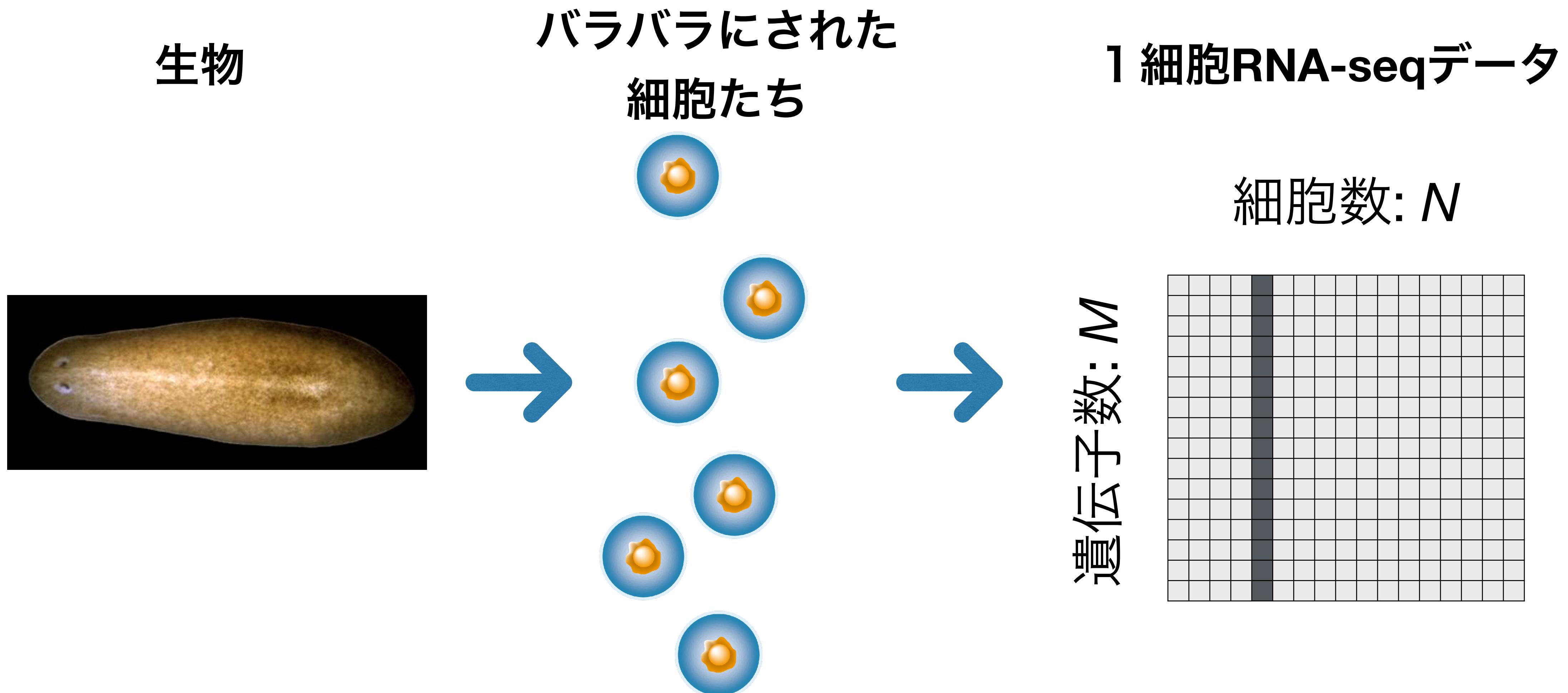
細胞型アノテーション：クラスターに細胞型をラベリング



演習C

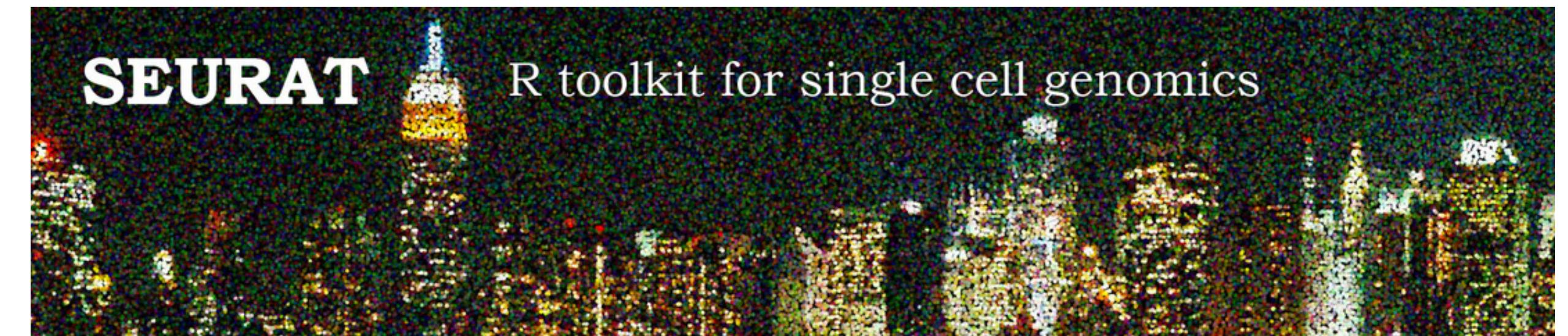


プラナリアの1細胞RNA-seqデータ

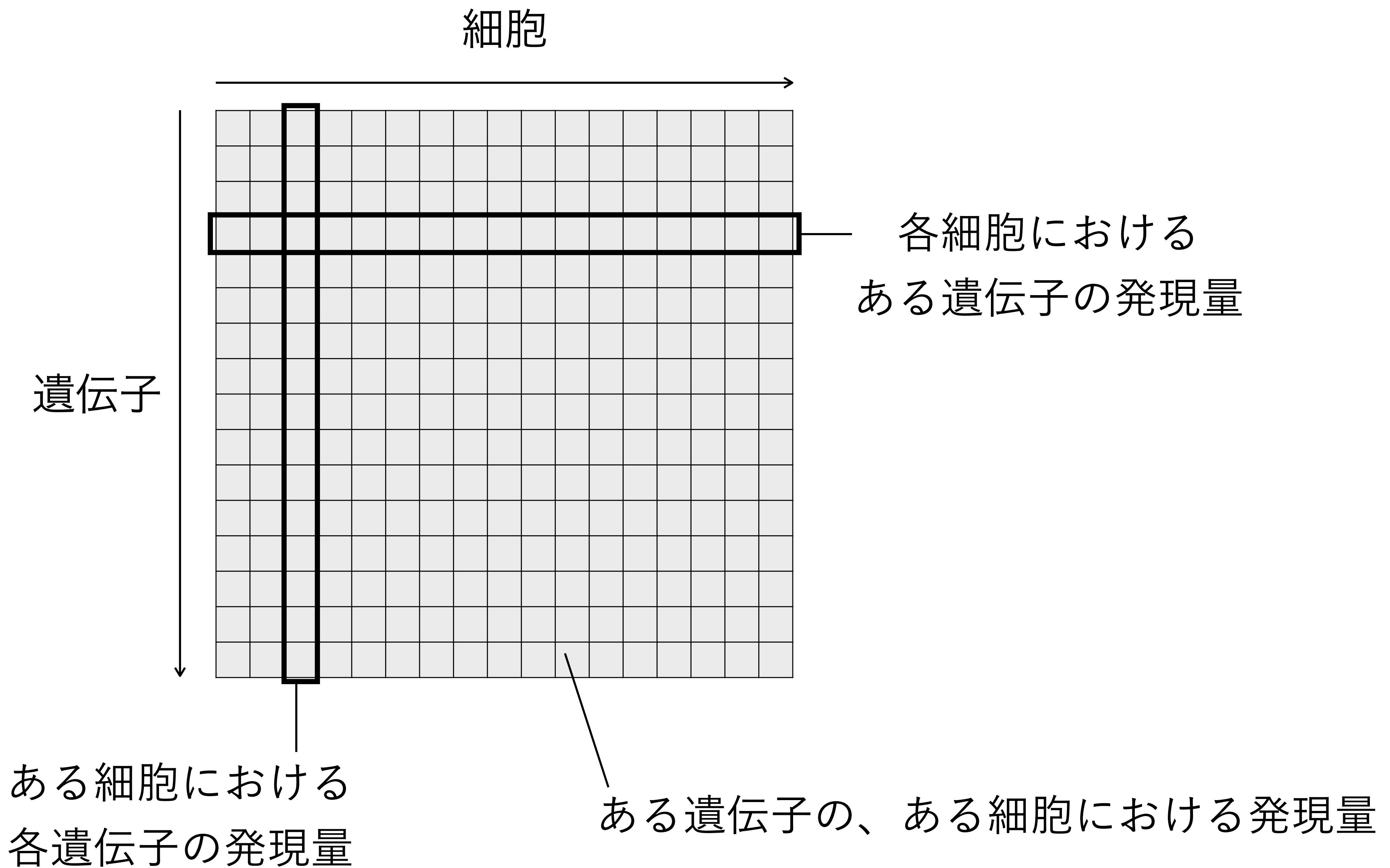


プラナリアの1細胞RNA-seqデータ

- 1細胞RNA-seq解析でよく使われる Seurat（すーら）というパッケージを使用します
- 1細胞RNA-seq解析の基本的な流れを学びます
 1. 遺伝子発現のカウント行列を読み込む
 2. 品質の低い細胞をフィルターする
 3. 発現量データを正規化する
 4. 高変動遺伝子 (highly variable genes) を抽出する
 5. 発現量データをスケーリングする
 6. PCA (主成分分析) を用いて次元削減を行う
 7. 細胞をクラスタリングする
 8. 各クラスターに特徴的な遺伝子群を探す
 9. 各クラスターがどんな細胞型かを類推する

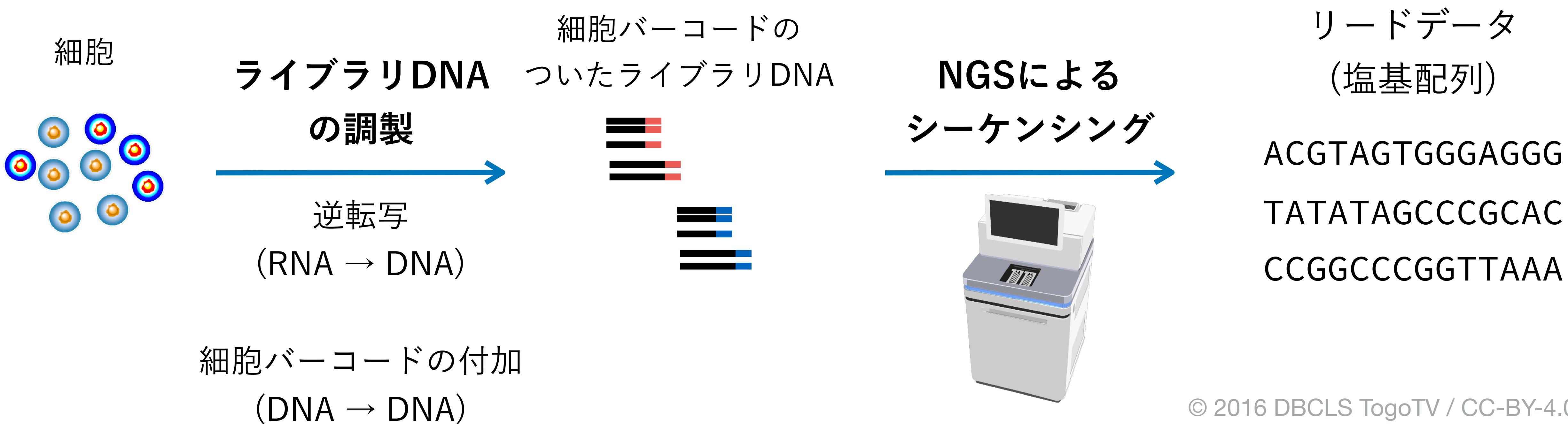


1. 遺伝子発現のカウント行列を読み込む



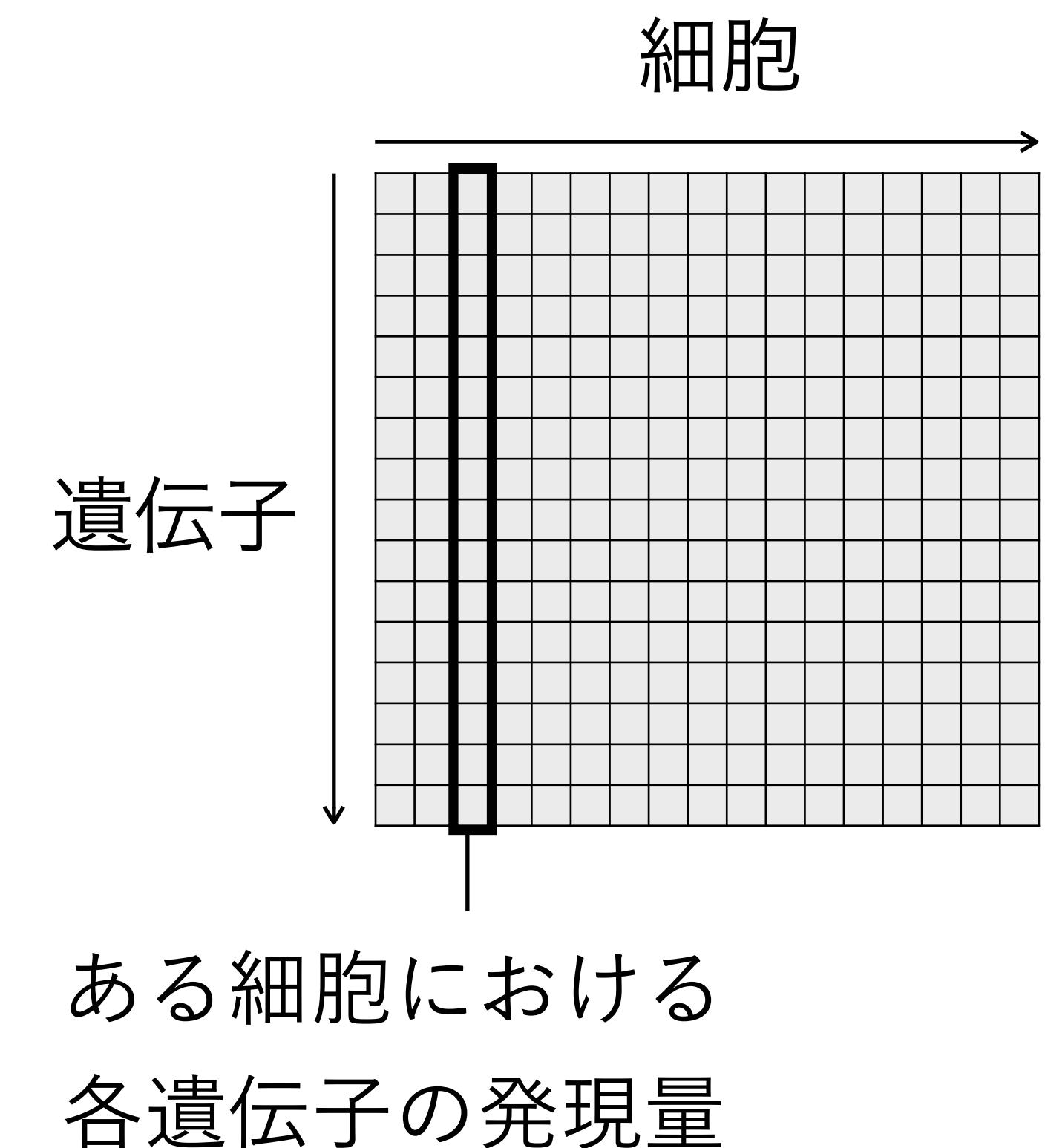
2. 品質の低い細胞をフィルターする

- 細胞によってはmRNAのリードが少ないとことがある
 - 化学反応なので、収率は100%ではない
 - 元の細胞の状態が悪かったり、死細胞が混ざっていることもある



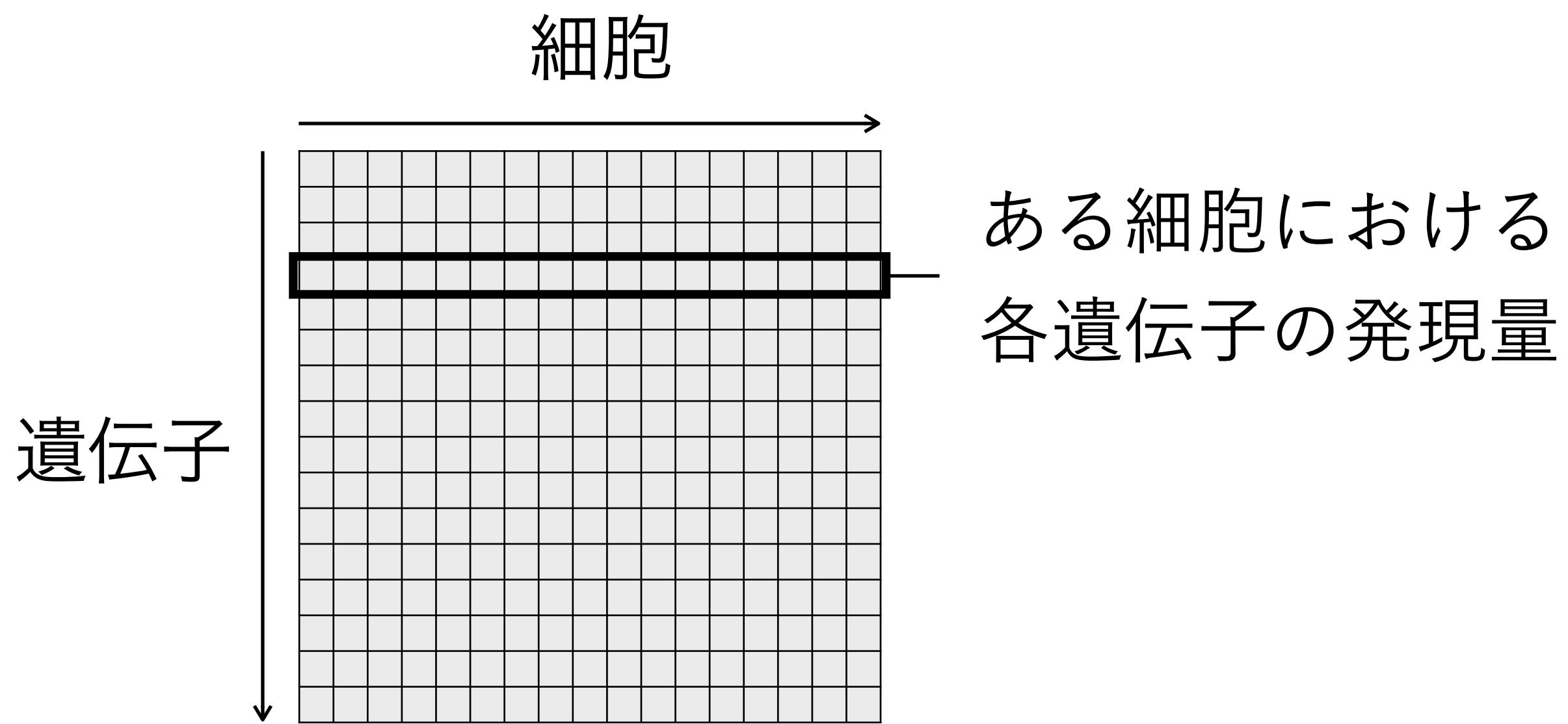
3. 発現量データを正規化する

- 細胞ごとにリードカウントの合計値が違う場合、元のカウントを細胞間で比較しても意味がない
 - 遺伝子発現量は「割合」に近いイメージ
- そこで、細胞間で遺伝子発現量を比較できるように、カウントデータを正規化する
 - 列ごとに、合計値で値を割る



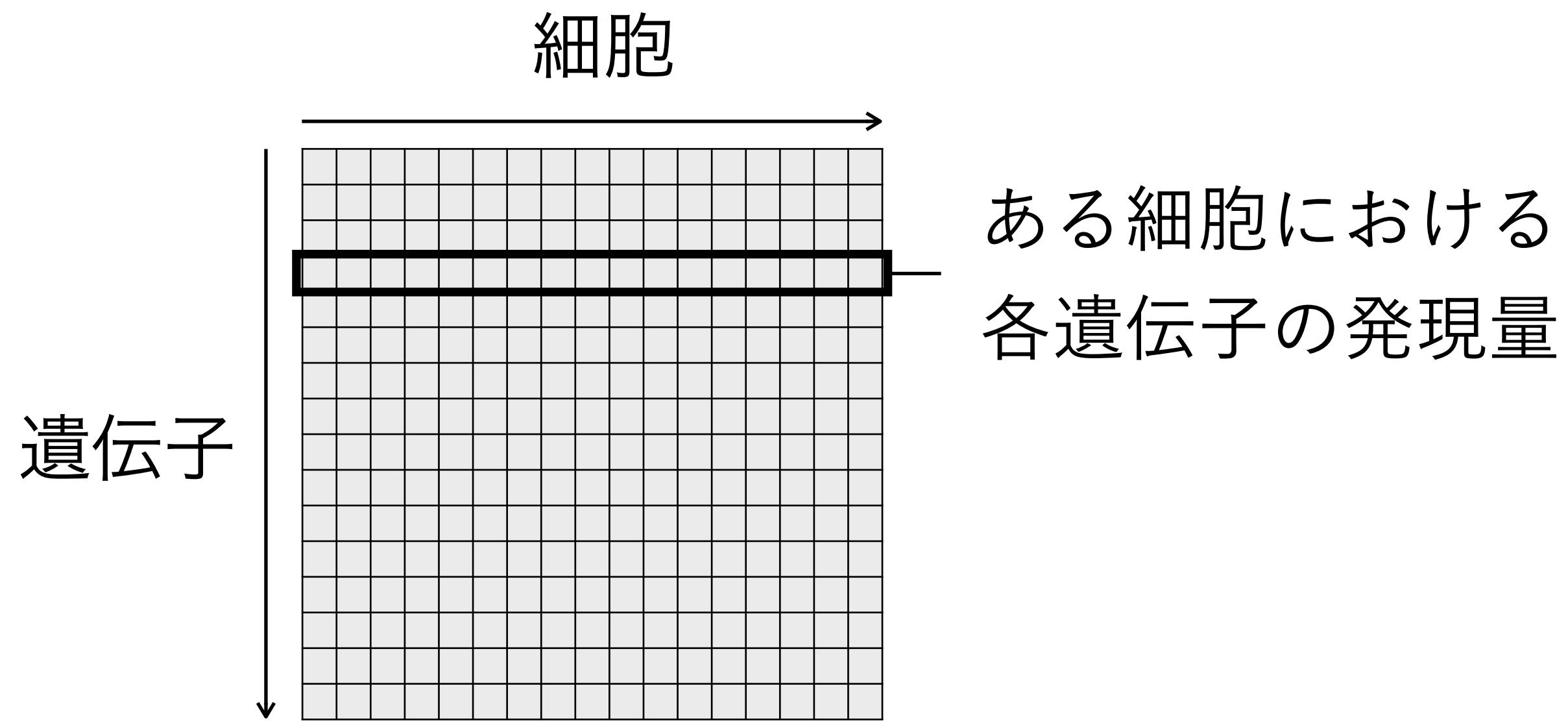
4. 高変動遺伝子 (highly variable genes) を抽出する

- 全ての遺伝子の発現量が重要なわけではない
 - 「個々の細胞の細胞型の違い・多様性」を見分けるためには、「細胞間で発現量が異なる遺伝子」を見なければならぬ
- そこで「細胞間で発現量が大きく変動している遺伝子」を抽出する
 - この際、遺伝子によっては「ノイズ」のように変動するものもあるため、統計学的に有意に高い変動を示す遺伝子を抽出することが重要



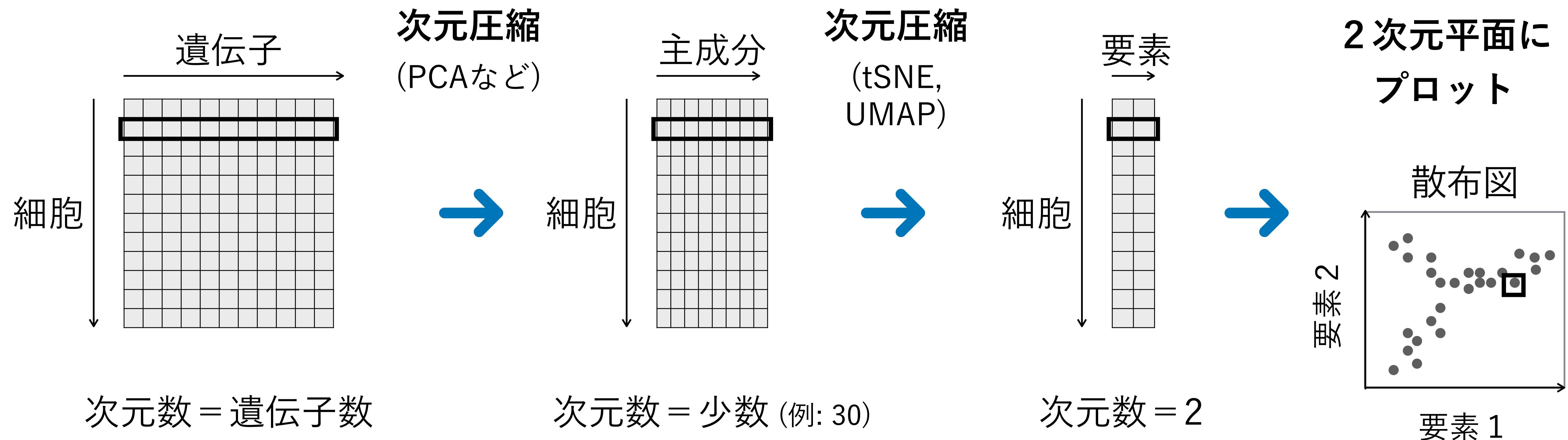
5. 発現量データをスケーリングする

- 計算処理の高速化や計測ノイズをならす意味がある
- 細胞のクラスタリングには、遺伝子たちを変数として使用する
- この際、このままでクラスタリングすると、発現量が大きい遺伝子の影響が大きくなる
- そのため、遺伝子間での発現量のスケールを揃える（スケーリング）ことが必要となる

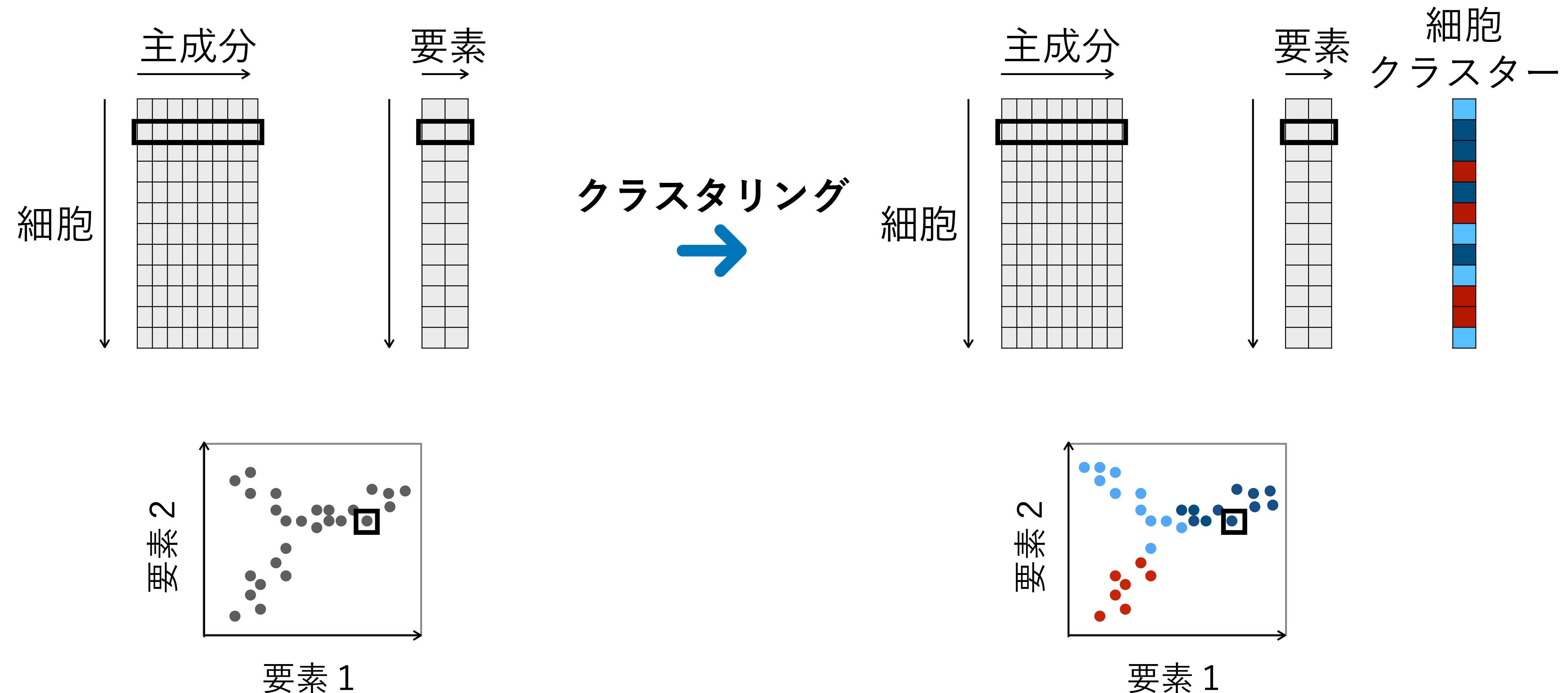


6. PCA（主成分分析）を用いて次元削減を行う

- ・クラスタリングの前に次元圧縮することで、データの多様性をなるべく損ねずに効率的にクラスタリングができる
- ・PCA (Principal component analysis) は情報の損失少なく次元圧縮できる



7. 細胞をクラスタリングする (1/2)

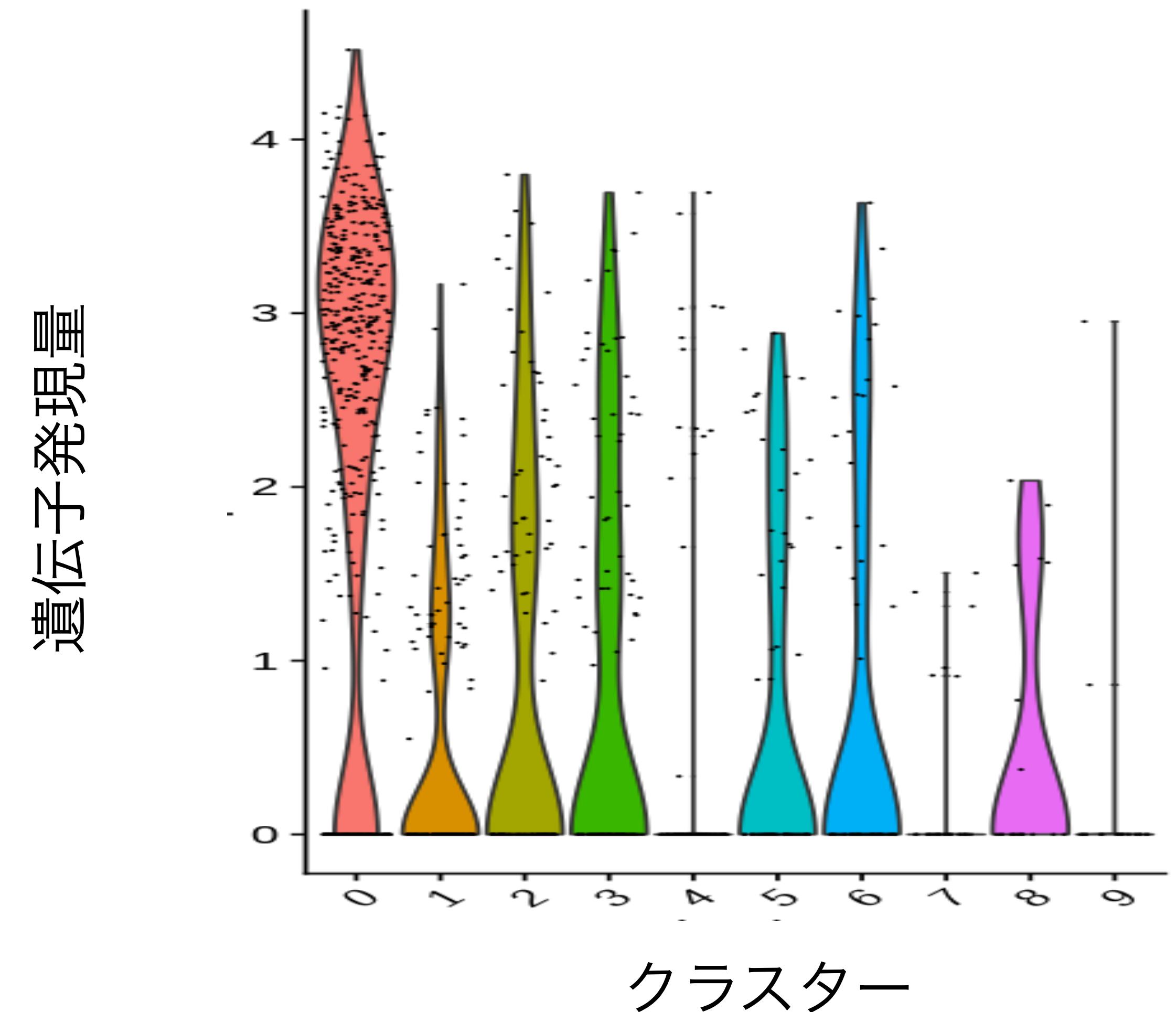


7. 細胞をクラスタリングする (2/2)

- クラスタリングとは、たくさんのサンプルを、データの値の類似性に基づいて、いくつかのグループに分けることである
- 1細胞RNA-seqの場合は、たくさんの細胞を、遺伝子発現量の類似性に基づいて、いくつかのグループに分けることである
- クラスタリングにより見つかったグループをクラスターと呼ぶ

8. 各クラスターに特徴的な遺伝子群を探す

- あるクラスターについて、他のクラスターに比べて発現量が高い遺伝子は、そのクラスターの細胞の特徴を反映している可能性が高い
- マーカー遺伝子 (marker genes)とも呼ばれる



9 各クラスターがどんな細胞型かを類推する

- ・遺伝子機能の知識が不足している場合は、オーソログの情報を使うと良い

9 各クラスターがどんな細胞型かを類推する

- ・遺伝子機能の知識が不足している場合は、オーソログの情報を使うと良い

演習Cを始めましょう

- <https://github.com/bioinfo-tsukuba/FY2022-EB62104-Bioinformatics/tree/main/%E6%BC%94%E7%BF%92C>

課題

- 基本課題C-1
 - `planarian_single_cell.ipynb` を Jupyter Hub 上で開き、上から 1 つずつセルを実行せよ。
- 発展課題C-1
 - 各クラスターに特異的な遺伝子群がどのような機能を持つ遺伝子かを調べ、レポートとしてまとめよ。
- 発展課題C-2
 - ヒトの13の組織において ACE2など SARS-COV2 の感染に関連する受容体の発現を調査している論文で使われているデータをダウンロードし、データ前処理・解析を行い、ACE2遺伝子の発現量が高い細胞があるかを調べよ