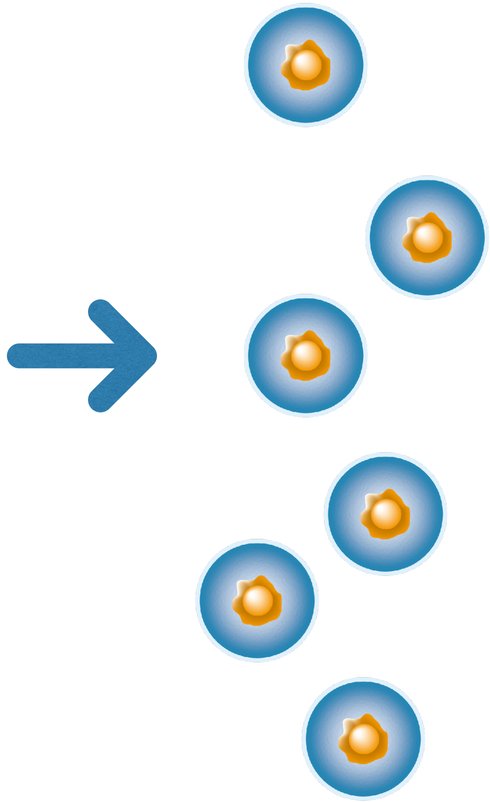
演習C

プラナリアの１細胞RNA-seqデータ

生物

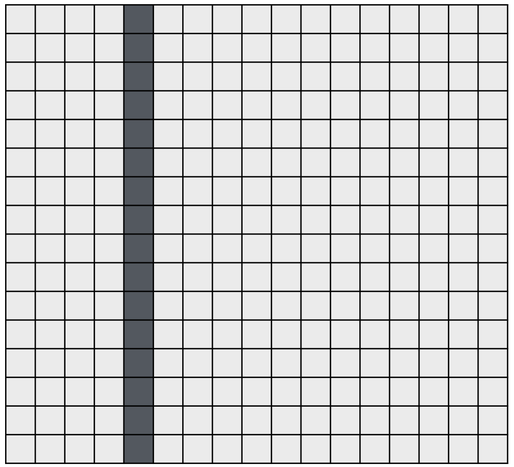
バラバラにされた
細胞たち



１細胞RNA-seqデータ

細胞数: *N*

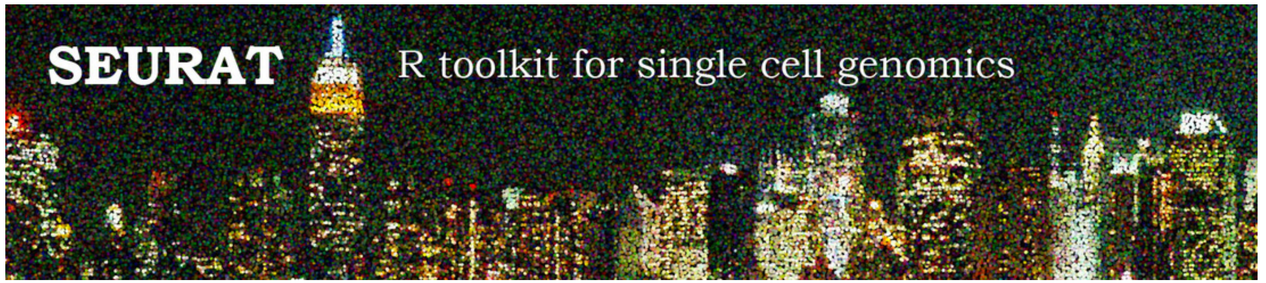
遺伝子数: *M*



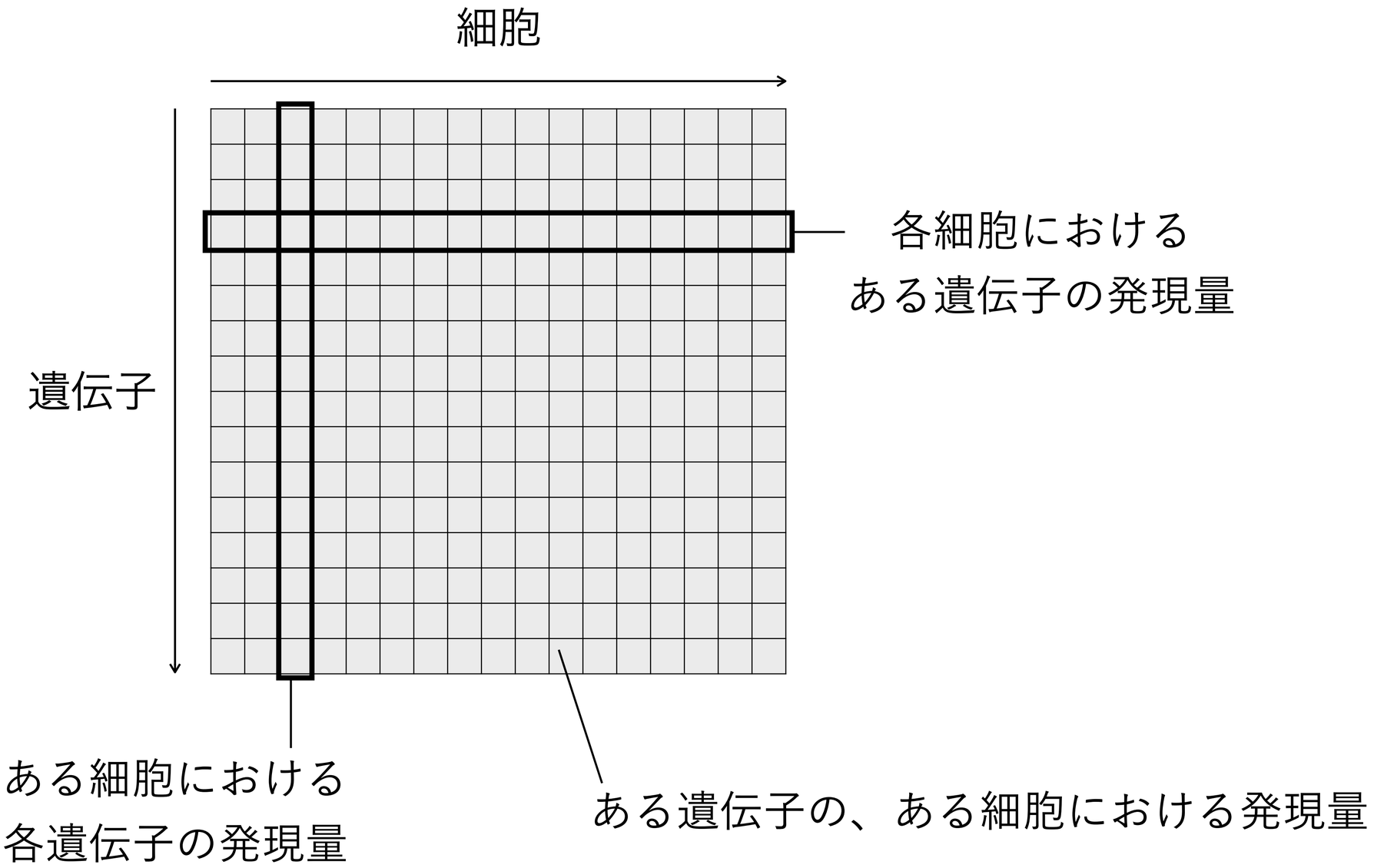
プラナリアの１細胞RNA-seqデータ

- 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 variabe 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

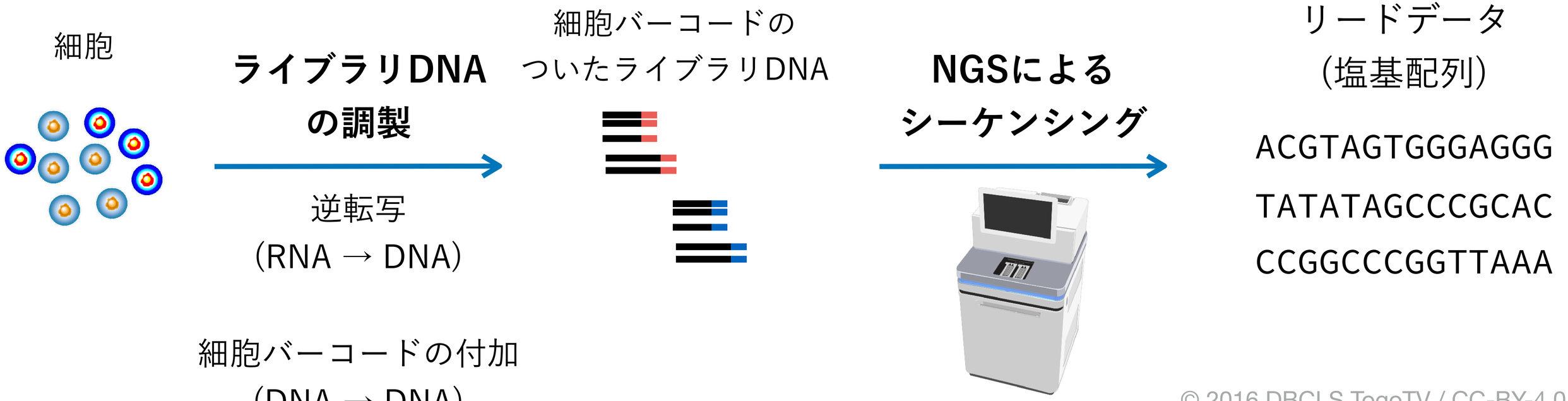


1. 遺伝子発現のカウント行列を読み込む



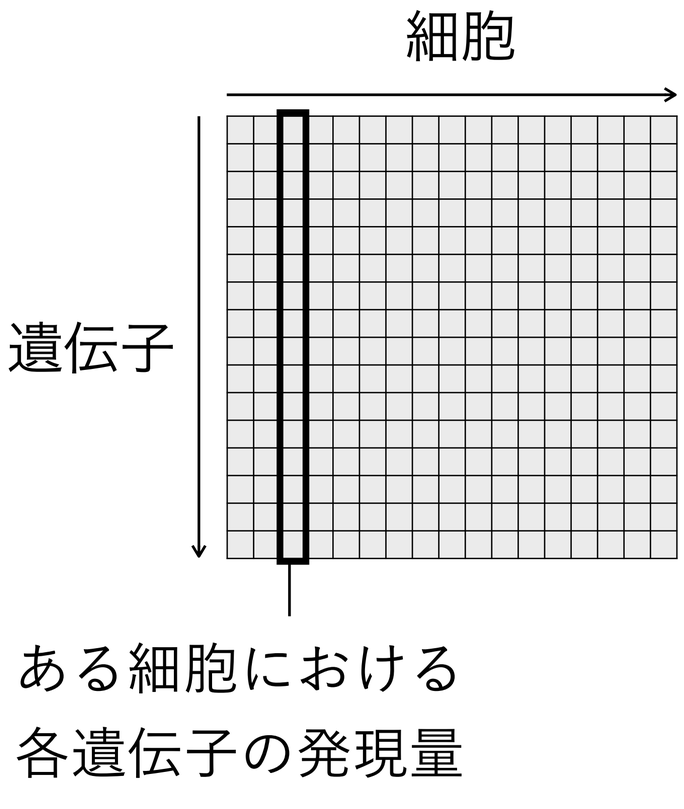
2. 品質の低い細胞をフィルターする

- 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



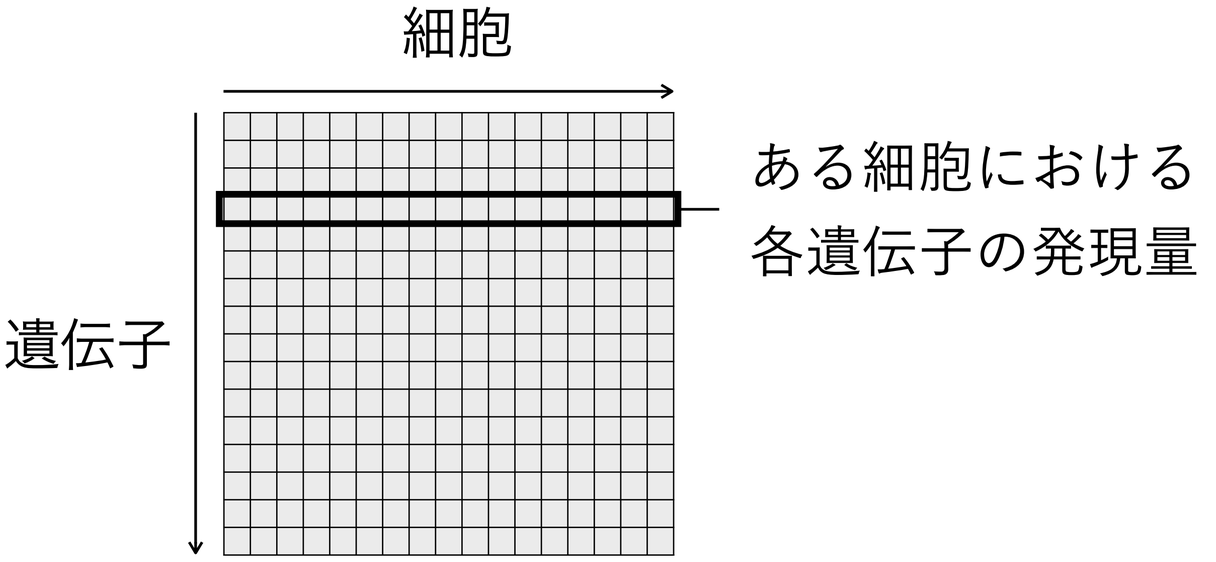
3. 発現量データを正規化する

- 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



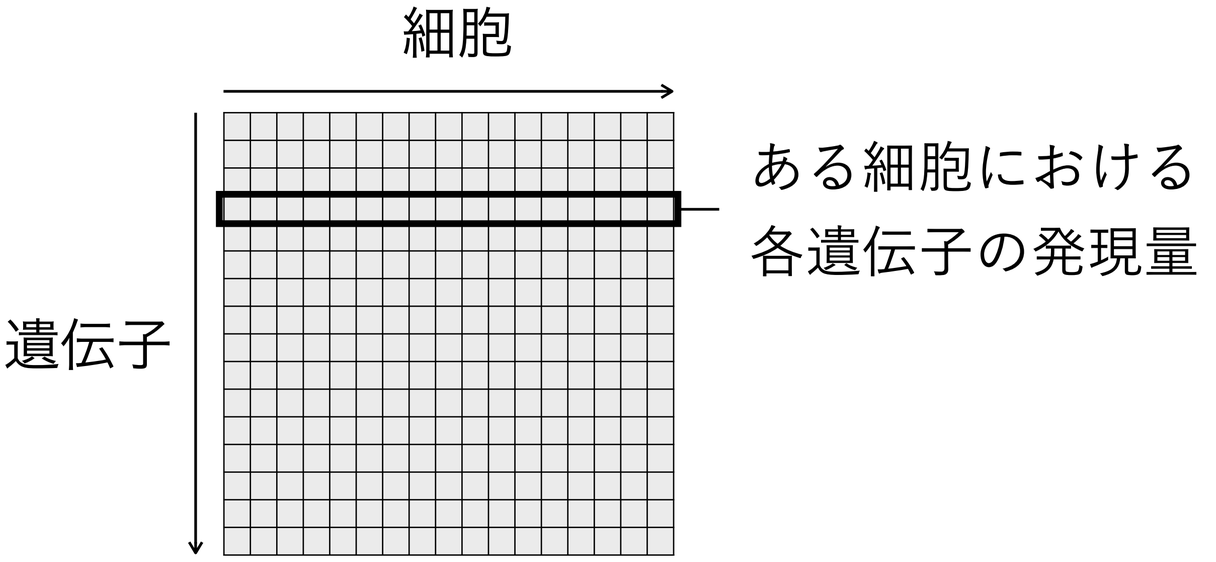
4. 高変動遺伝子（highly variabe genes）を抽出する

- 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」を見なければならない
- 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



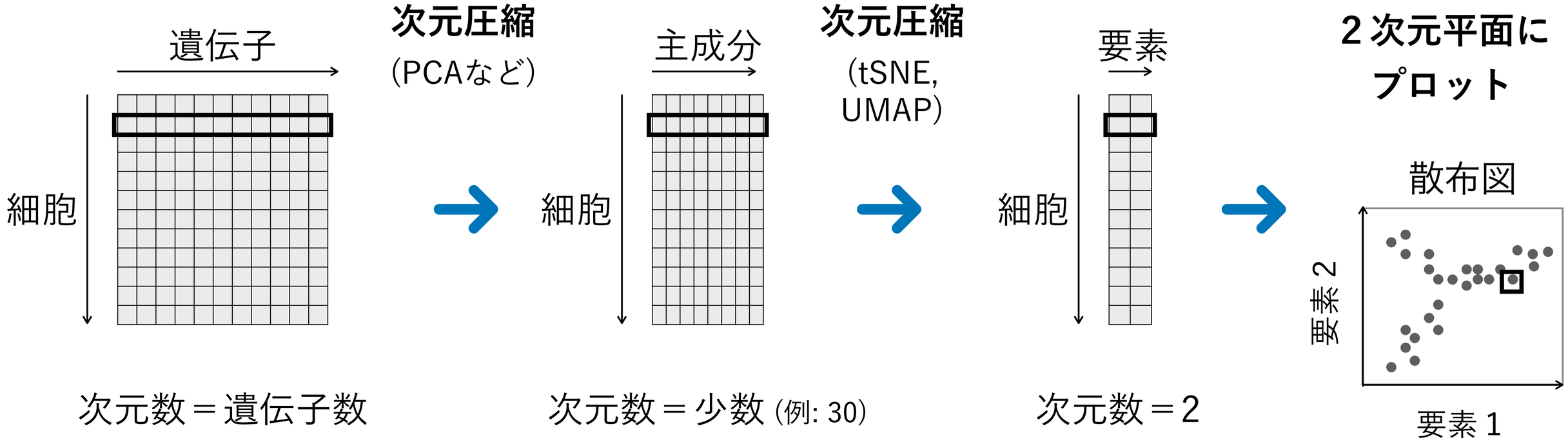
5. 発現量データをスケーリングする

- 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

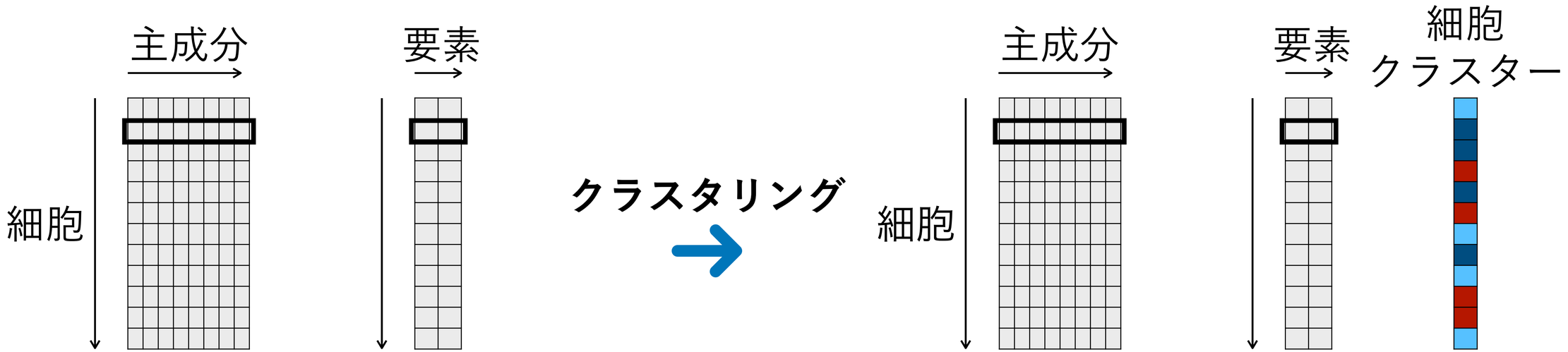


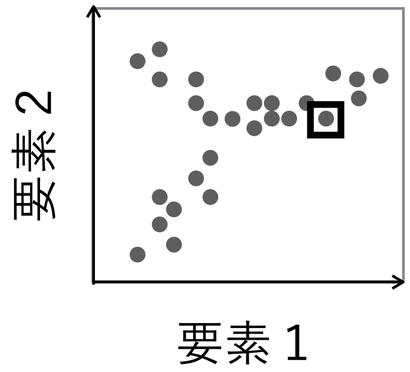
6. PCA（主成分分析）を用いて次元削減を行う

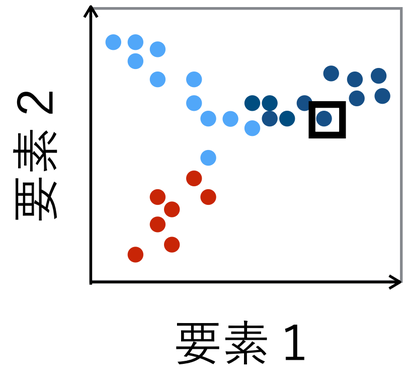
- クラスタリングの前に次元圧縮をすることで、データの多様性をなるべく損ねずに効率的にクラスタリングができる
  - PCA (Principal component analysis) は情報の損失少なく次元圧縮できる



7. 細胞をクラスタリングする (1/2)





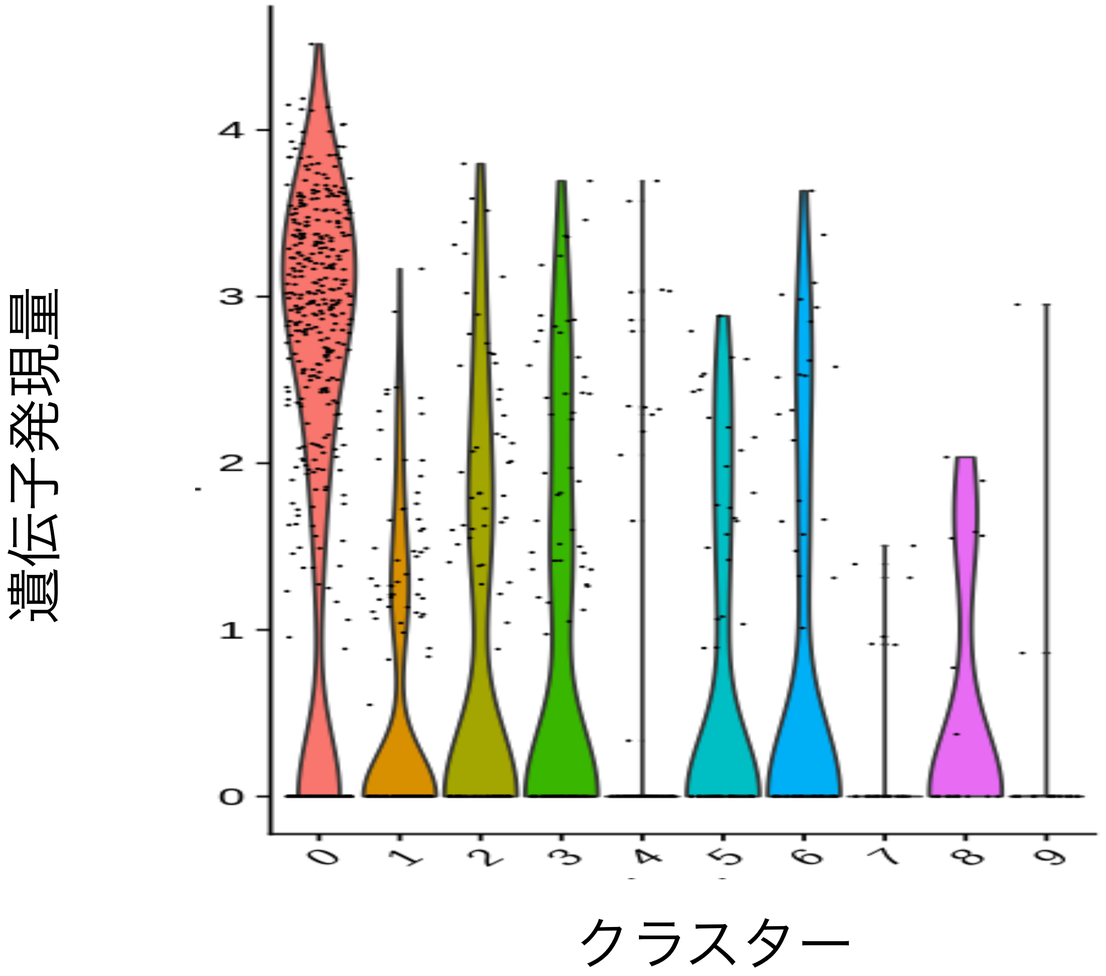


7. 細胞をクラスタリングする (2/2)

- 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さる
- 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