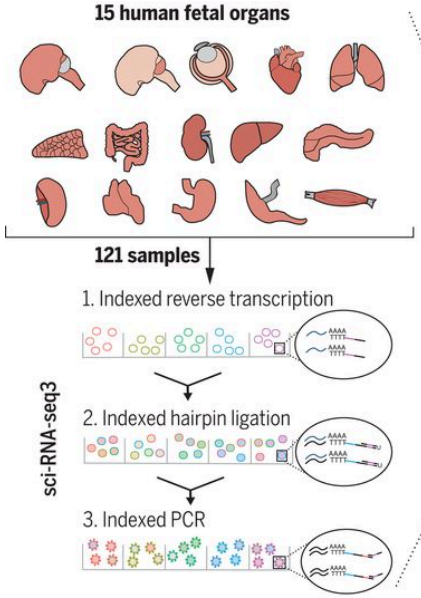
Bioinformatics Lab

Ivan Gesteira Costa & Mingbo Cheng & Martin Manolov & James Nagai & Mina Shaigan
Institute for Computational Genomics

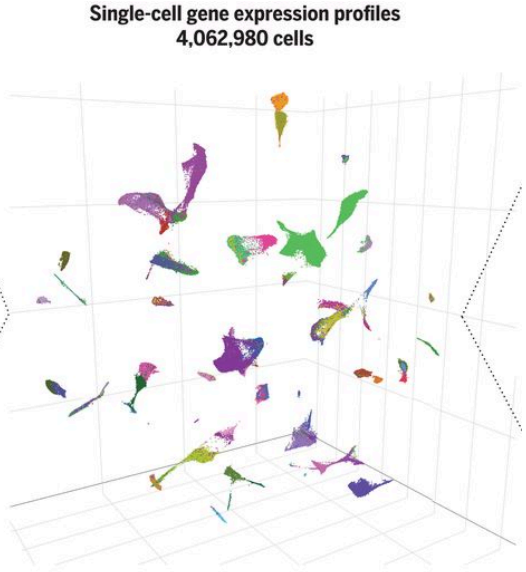
Institute for Computational Genomics
01011011010
10100100101
RWTH AACHEN UNIVERSITY

Problem Definition

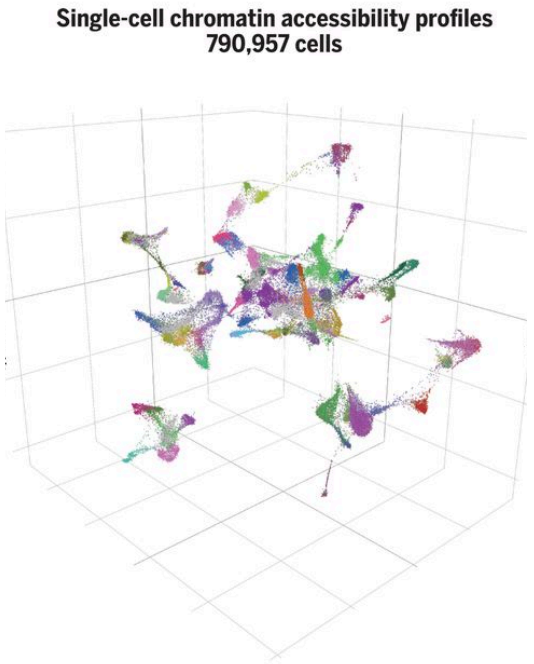
Clustering of cells / Human Fetal Cell Atlas



scRNA-seq



scATAC-seq



https://descartes.brotmanbaty.org/

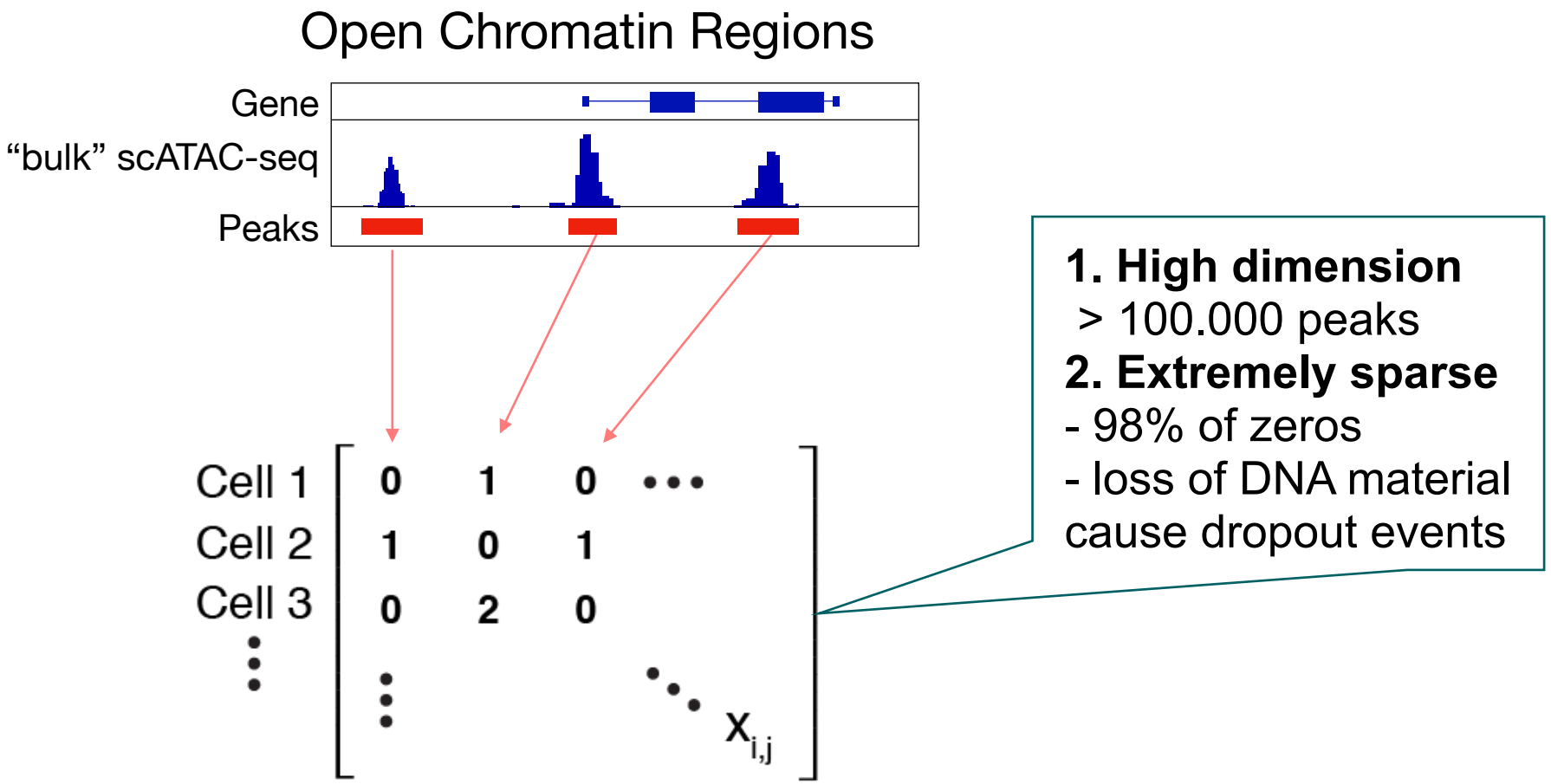
How to deal with large data sets (millions of cells)?

Adapted from Donke et al. 2020.

Single cell clustering / Project

- Finding groups of single cells require complex pipeline:
  - ~~Cell filtering~~
  - ~~Normalisation~~
  - ~~Artefact removal~~
  - **Dimension reduction**
  - ~~Integration~~
  - **Clustering**
  - ~~Cell annotation /~~ **visualisation**
- Open points:
  - How to deal with large data sets (millions of cells)?
  - How to deal with sparsity of single cell (scRNA-seq or scATAC-seq) data?

Single cell data and sparsity



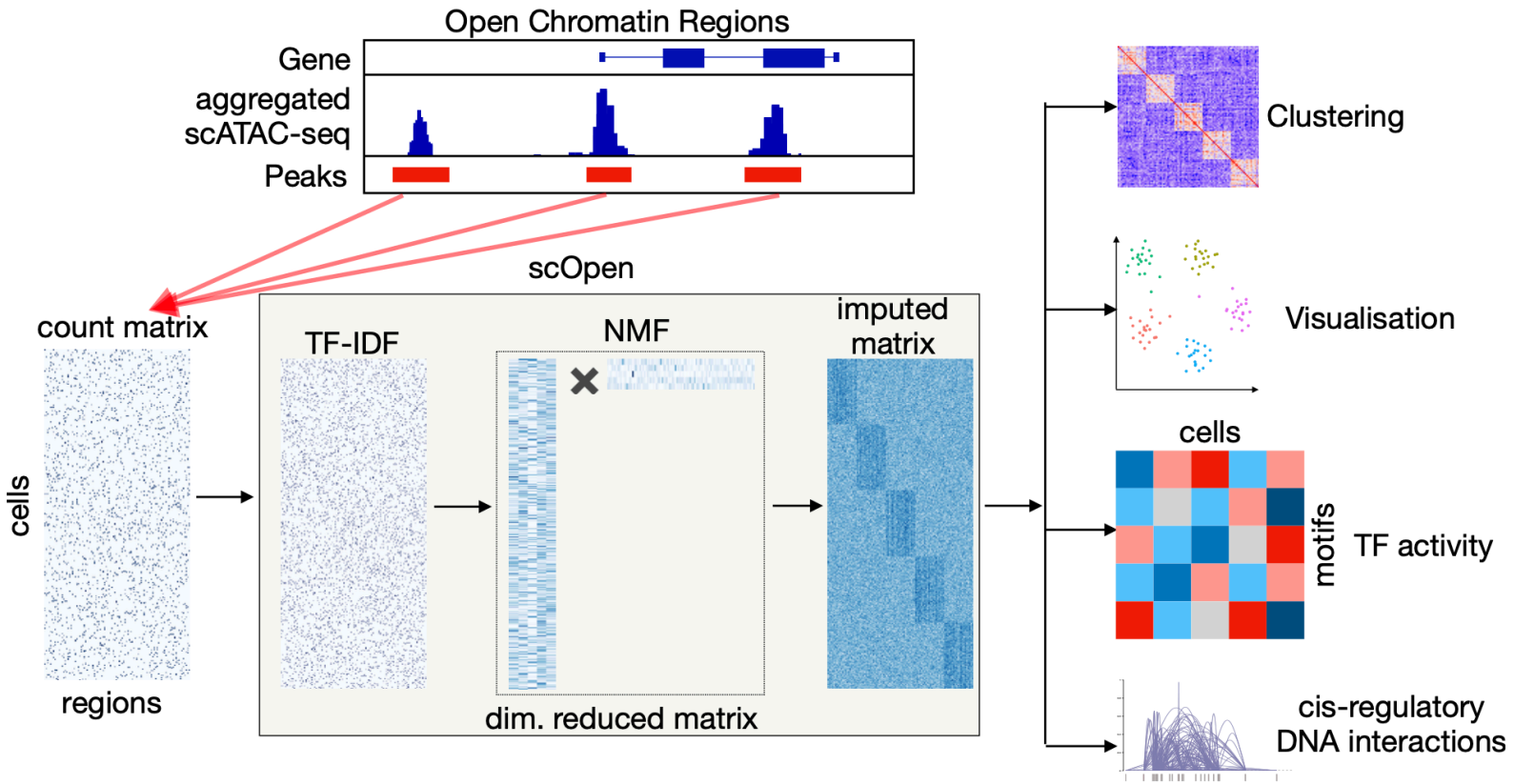
Single cell data and sparsity

- Sparsity example of scATAC and scRNA-seq data

Dataset	Type	Cells	Features	Non-zeros	Reads per cell
Cell lines	scATAC-seq	1,224	125,647	0.036	41,467.80
T cells	scATAC-seq	765	49,344	0,033	14,963.39
Hematopoiesis	scATAC-seq	2,210	109,418	0.039	34.656.15
Hematopoiesis	scRNA-seq	14,432	12,558	0.119	5.209,45
PBMC	scATAC-seq	10,032	106,935	0.067	13,486
UUO	scATAC-seq	31,129	150,593	0.042	13,933

- scRNA-seq data has lower dimension (>20.000 features) and lower sparsity (20-40%).

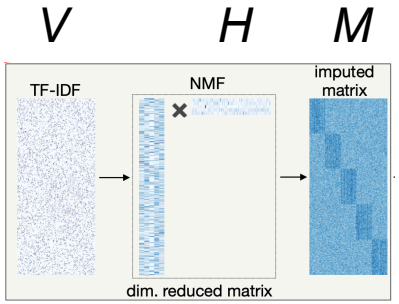
Imputation of sparse scATAC-seq matrices



Non-negative matrix factorisation

Li, Z., …., 2021. Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. Nat. Commun. 12, 6386. https://doi.org/10.1038/s41467-021-26530-2

Efficient Non-negative factorisation NMF for matrix completion



For an observed count matrix *V*, we want to infer *M* assuming it is low-rank

$$\hat{M} = argmin \sum_{i,j} (M_{ij} - V_{ij})^2 + \lambda ||M||_*, \quad s.t. \quad M_{ij} \geq 0$$

which is equivalent to:

$$\min_{W,H} f(W,H) = \sum_{ij} ((WH)_{ij} - V_{ij})^2 + \frac{\lambda}{2}||W||^2 + \frac{\lambda}{2}||H||^2, \quad s.t. \quad (WH)_{ij} \geq 0$$

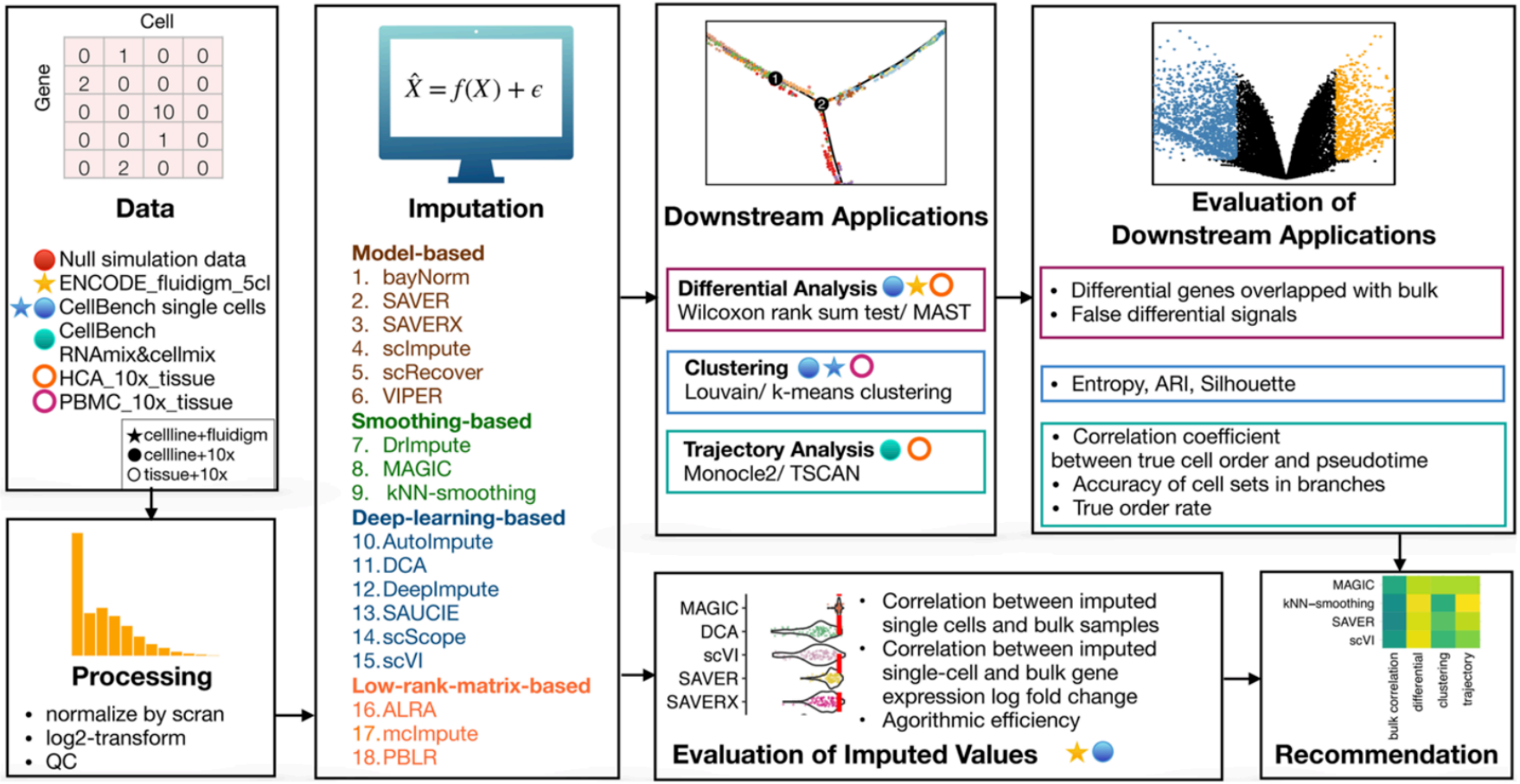
Above problem can be solved with cyclic coordinate descent method:

$$\min_{z} f(z) = \sum_{j=1}^{n} ((\sum_{t' \in k} w_{it'} h_{t'j} - w_{it} h_{tj}) + z h_{tj} - V_{ij})^2 + \frac{\lambda}{2} z^2, \quad s.t. \quad z \geq 0$$

Li, Z., …., 2021. Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. Nat. Commun. 12, 6386. https://doi.org/10.1038/s41467-021-26530-2

Examples of imputation algorithms

Benchmarking analysis of imputation methods:

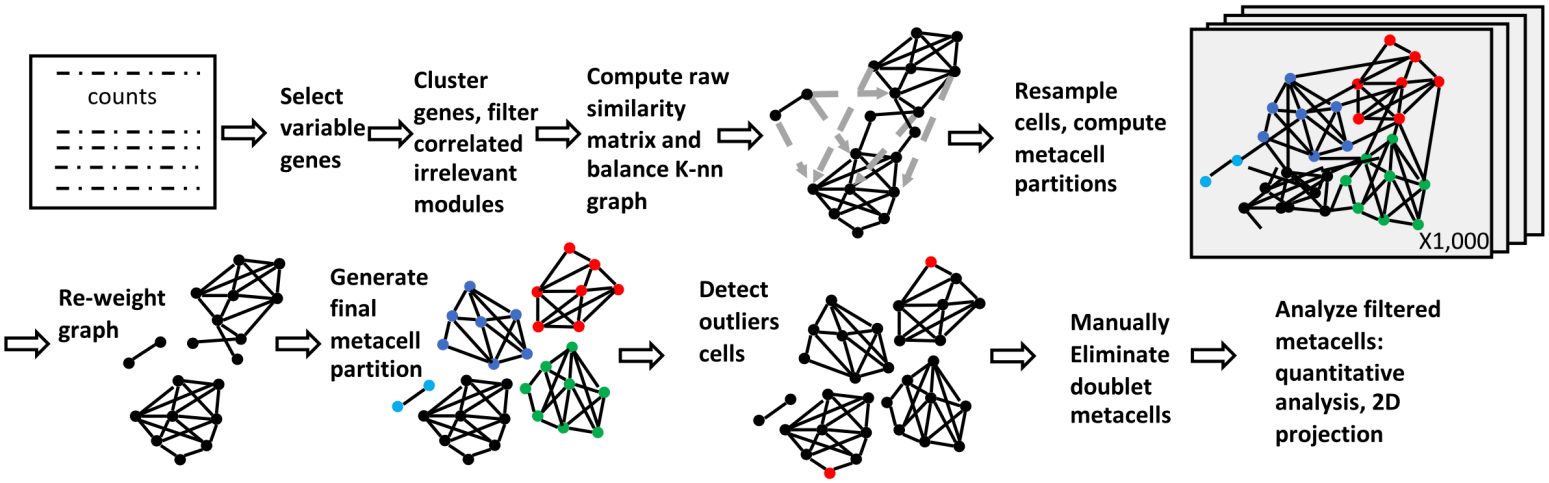


Hou, W., Ji, Z., Ji, H., Hicks, S.C., 2020. A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biol. 21, 1–30. https://doi.org/10.1186/s13059-020-02132-x

Meta-cell analysis

General idea: group cells into small groups (10-20 cells)

Metacell: k-NN based metacells



Advantages:

- reduction of sample size
- denoising (meta cells have their expression sum up)

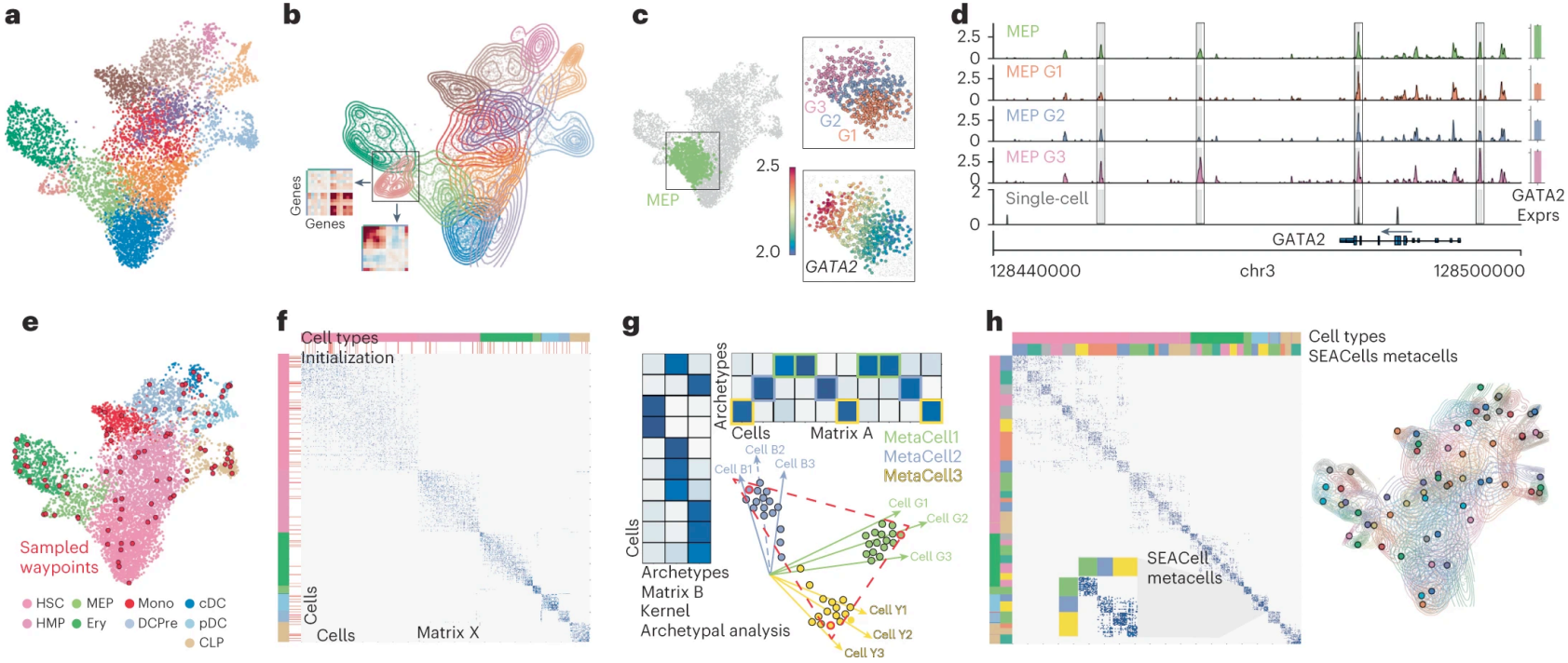
Baran, Y., …., 2019. MetaCell: Analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. 20, 1–19. https://doi.org/10.1186/s13059-019-1812-2.

Ben-Kiki, O., …, 2022. Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. Genome Biol. 23, 1–18. https://doi.org/10.1186/s13059-022-02667-1

Meta-cell analysis

General idea: group cells into small groups (10-20 cells)

SeaCell: matrix factorisation based metacells



Persad, S., et al., 2023. SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. Nat. Biotechnol. 2022.04.02.486748. https://doi.org/10.1038/s41587-023-01716-9

Challenges: Denoising vs. Metacells

1. **Several methods denoising and meta-cell methods.**

2. **Meta-cells can denoise data at an expense of lower capture of rare cells**

Evaluation: trade off between performance (time & memory) and accuracy (clustering)

Are this similar once considering scRNA vs. scATAC-seq?

Relevant work:

Baran, Y., Bercovich, A., Sebe-Pedros, A. *et al.* MetaCell: analysis of single-cell RNA-seq data using *K*-nn graph partitions. *Genome Biol* **20,** 206 (2019).

Ben-Kiki Tanay, A., 2022. Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. Genome Biol. 23, 1–18.

Hou, W., Ji, Z., Ji, H., Hicks, S.C., 2020. A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biol. 21, 1–30. https://doi.org/10.1186/s13059-020-02132-x

Li, Z., …., 2021. Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. Nat. Commun. 12, 6386. https://doi.org/10.1038/s41467-021-26530-2.

Persad, S., et al., 2023. SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. Nat. Biotechnol. 2022.04.02.486748. https://doi.org/10.1038/s41587-023-01716-9

Overall Design / Basic Approach

Perform dimension reduction and clustering to find groups of cells

1. Dimension reduction
2. Clustering

Data sets:

- Use quality checked and pre-labeled data from Human cell fetal atlas
  - Either scRNA-seq or scATAC-seq
- Perform integration with Harmony

Evaluation:

- Use adjusted Rand index (and similar indices) to evaluate clustering accuracy compared to labels
  - https://en.wikipedia.org/wiki/Rand_index
- Benchmark both time/memory requirements
- Evaluate scalability with increase in sample size
- Evaluate the impact of parameters (size of meta-cells) in performance (accuracy vs. Time)

Project Proposal

- **Groups: 3-4 participants each**

Challenges

1. Denoising vs. Meta-cells with scRNA or scATAC-seq data

- **Projects code should be deposited in gitlab**
  - **git.rwth-aachen.de**
- **Groups should discuss in a discord room**
  - **channel for group: https://discord.gg/q7rQz9eX83**

Calendar

15.05.2023 – 3.7.2023 – Project development

10.07.2023 – Project Presentation

Links

- **Machine learning libraries:**
  - **python - scikit - https://scikit-learn.org/stable/**
  - **python & gpu - https://keras.io/**
  - **Python - scanpy & epyscanpy**
    - **https://scanpy.readthedocs.io/en/stable/**
    - **https://episcanpy.readthedocs.io/en/latest/**

- **Data**

  Relevant data will be provided at the RWTH Cluster

  /hpcwork/lect0094/

  https://descartes.brotmanbaty.org/

  Data matrices for distinct organs + meta-data (cell label - true label)

- **Computing**

  - can be done at HPC as described earlier today

Thank you!