scVerse 2024 workshop

National Health Data Science Sandbox

Samuele Soraggi, PhD

Data scientist - special consultant

Bioinformatics Research Center (BiRC), Aarhus University

samuele@birc.au.dk

Wednesday, 2024.09.11

BiRC

scVerse 2024 workshop

Projections, interactive plots and interactive online docs of your scRNA project

Samuele Soraggi, PhD

Data scientist - special consultant

Bioinformatics Research Center (BiRC), Aarhus University

samuele@birc.au.dk

Wednesday, 2024.09.11

scVerse 2024 workshop

https://shorturl.at/bj6QS

Or

https://hds-sandbox.github.io/scverse-2024-workshop

Samuele Soraggi, PhD

Data scientist - special consultant

Bioinformatics Research Center (BiRC), Aarhus University

samuele@birc.au.dk

Wednesday, 2024.09.11

Outline of the workshop

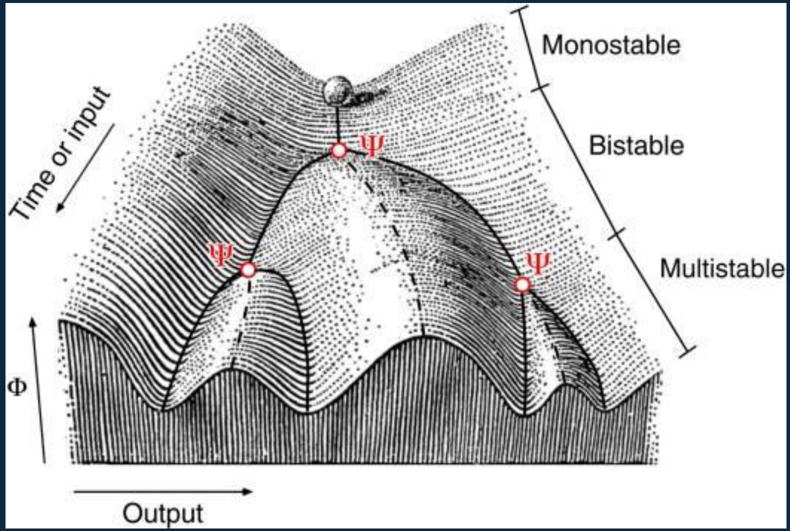
- **Dimensionality reduction (DR)**
  - DR and the Differentiation landscape
  - DR approaches in single cell data
  - Suggested workflow
  - Recap, pros&cons
- **Documentation**
  - Github and GH-pages
  - Quarto (or other) docs tool
  - Interactive (static) plots on notebooks
- **Tutorial**
  - Methods comparison
  - Diagnostics
  - Interactive plots for GH-pages

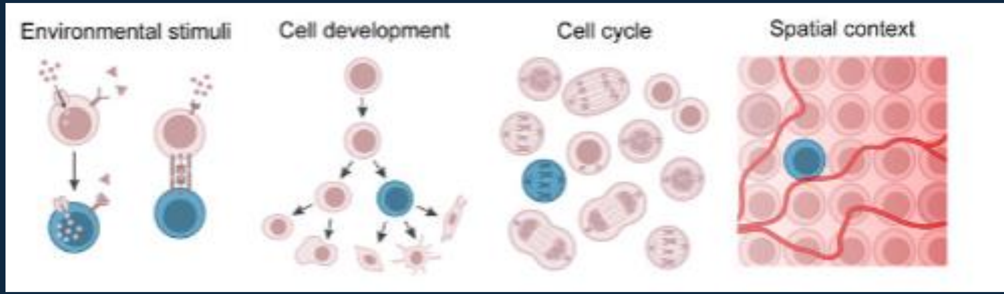
15 min

35-45 min

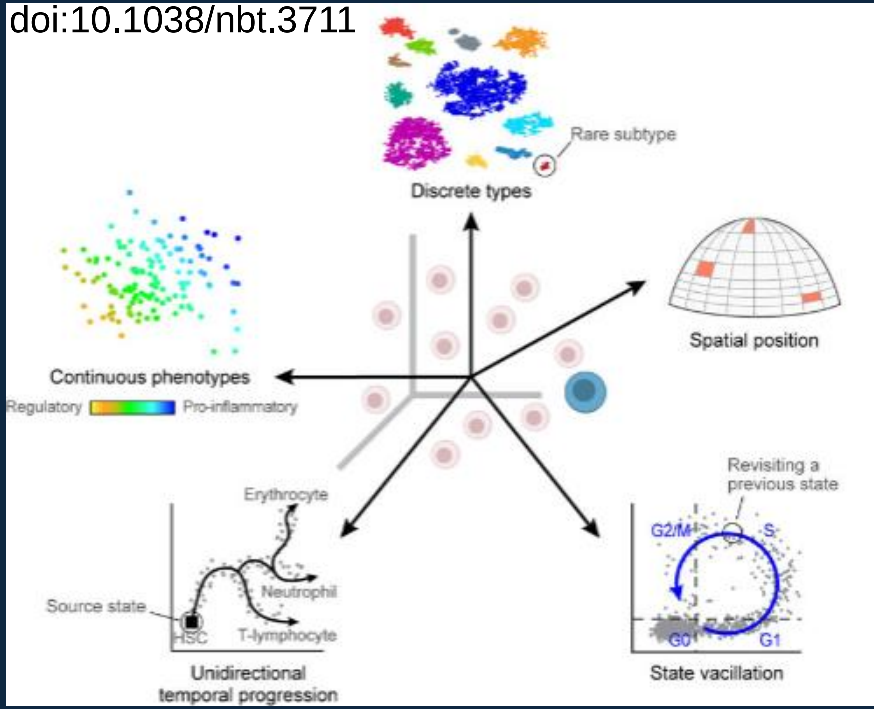
DR and the differentiation landscape

Complex differentiation landscape
Modeled with differential equations





Influence of other factors beyond on differentiation



Q: Could we get basis extracting essential underlying manifold of the differentiation and factors acting on it?

DR and the differentiation landscape

Yes! Dimensionality Reduction!

TSNE

Diff Maps

LBO

PCA

PaCMap

UMAP

ISOMAP

TriMap

DR and the differentiation landscape

Who wins? You decide!

TSNE

Diff Maps

LBO

Topometry

PCA

ISOMAP

UMAP

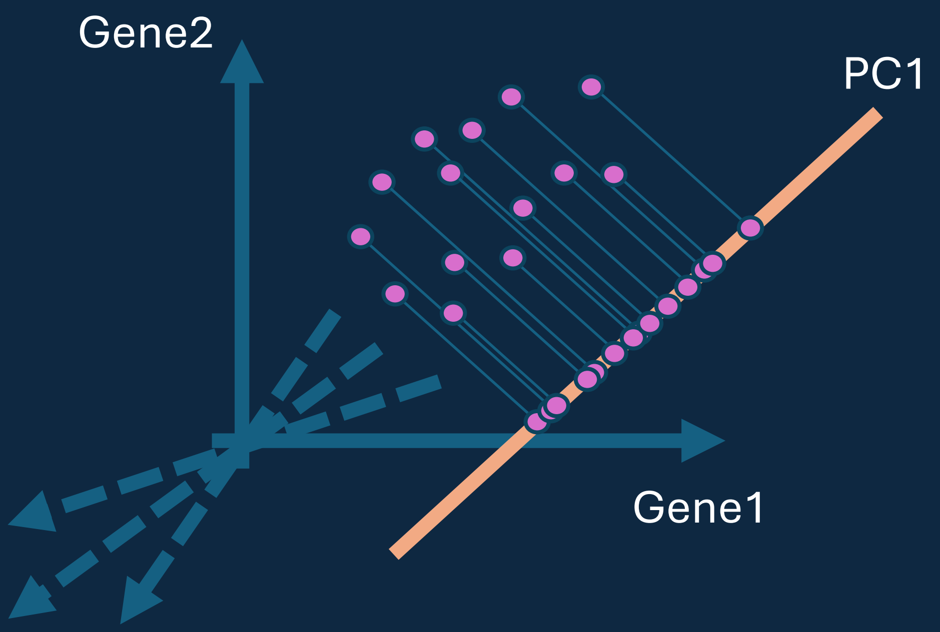
PaCMap

TriMap

DR approaches in single cell data

- Matrix decomposition

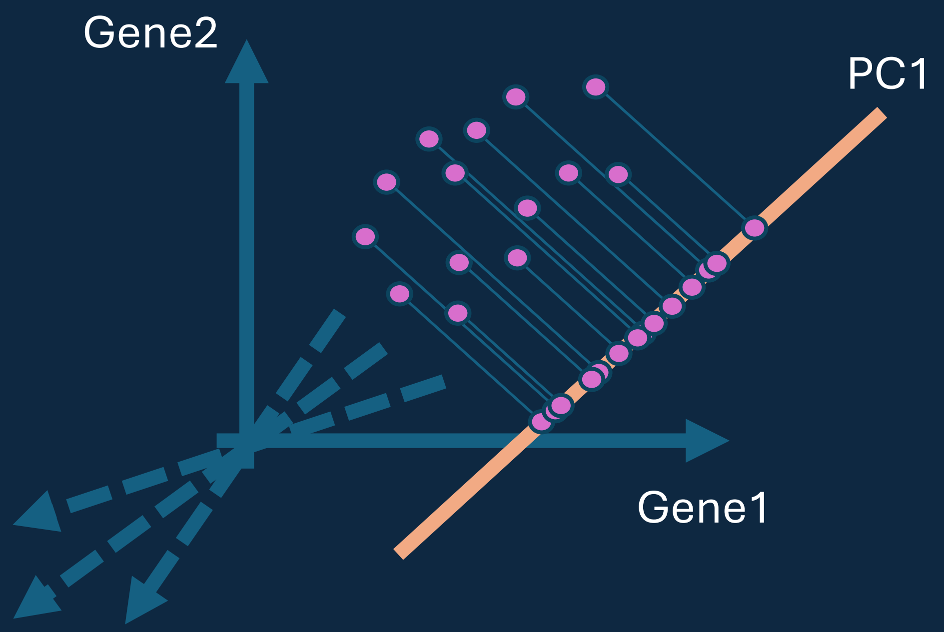
  PCA used as a base for most other methods in sc analysis



DR approaches in single cell data

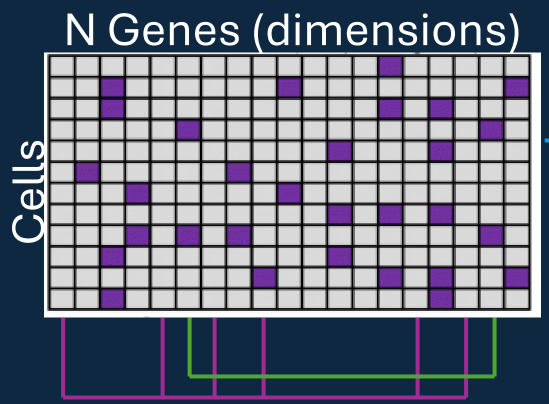
- Graph optimization methods

  tSNE, UMAP, triMAP, PaCMap → Usually run on the data's PCA



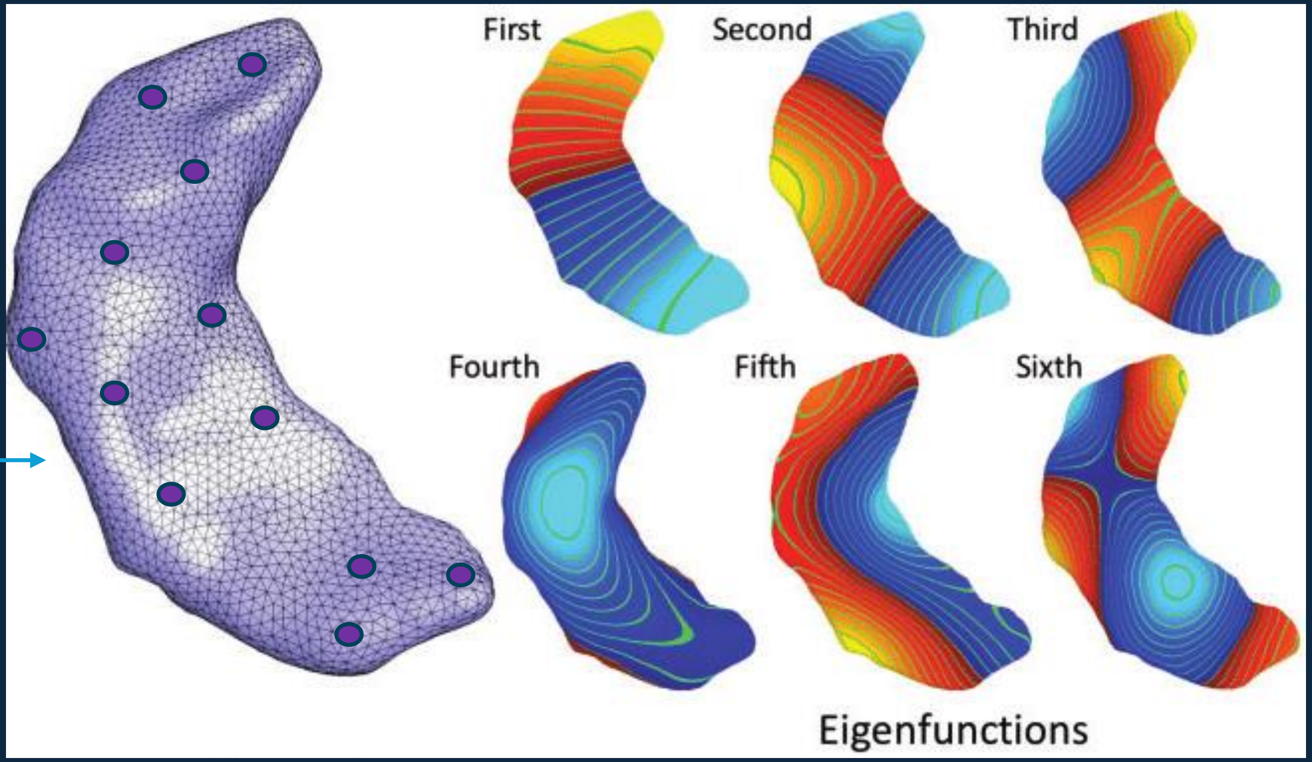
DR approaches in single cell data

- Spectral methods (topoMetry)
  - Data sampled from ”smooth surface” M
  - M has actually less dimensions than the data (e.g. genes acting in modules)
  - M pieced together as a series of basis functions (eigenfunctions)



Data sampled from S of dimens<N (intrinsic dimensionality)

- Highly correlated dimensions
- Not covering the entire possible space in N dimensions

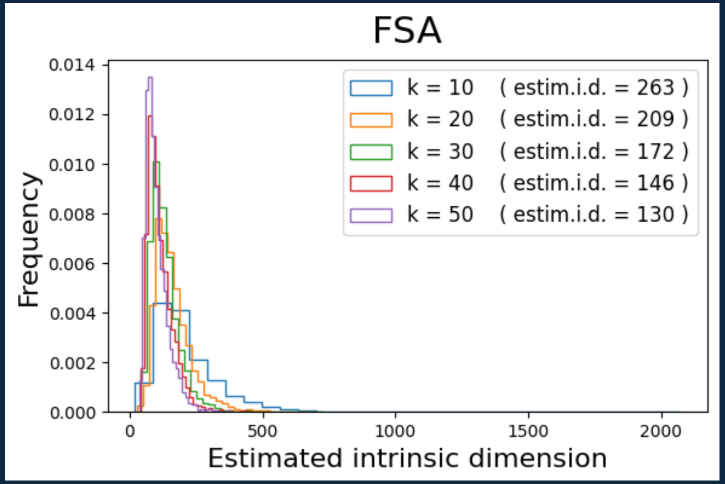


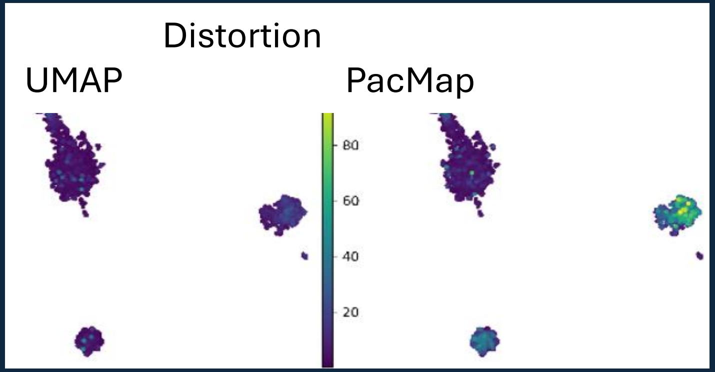
Data sampled from S of dimens<N (intrinsic dimensionality)

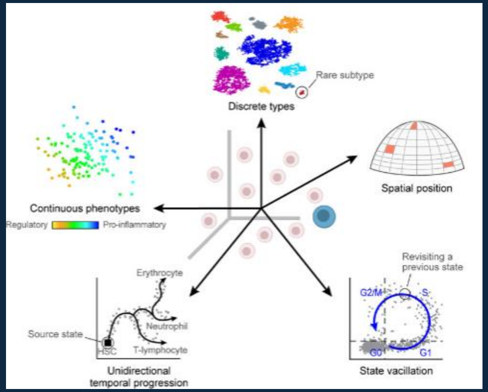
Suggested workflow

- Spectral methods

  - Find n<N eigenbasis (builds a matrix E)
  - Do projection from E (e.g. UMAP, PacMap)
    - Evaluate distortions to choose the best
  - E useful also for clustering
  - Explore eigenbasis on projections to see what each intrinsic dimension represents







Recap, pros/cons

Matrix decomposition & Graph optimization
(PCA, UMAP, triMAP, tSNE, PacMap)

In theory:

- **Linear**: hardly any hyperplane can capture the data variations (PCA)
- All graph-based methods use **PCA** as denoised data, missing non-linearity aspects and **distorting distances**
- All above methods are based on a loss function, aiming at preserving distances. **Curse of dimensionality!!!**

In practice:

- **Missing or false clusters**
- Loss of many **complex relationships** across dimensions
- Creation of **false relationships** through linearization

Recap, pros/cons

Matrix decomposition & Graph optimization (PCA, UMAP, triMAP, tSNE, PacMap)

In theory:

- **Linear**: hardly any hyperplane can capture the data variations (PCA)
- All graph-based methods use **PCA** as denoised data, missing non-linearity aspects and **distorting distances**
- All above methods are based on a loss function, aiming at preserving distances. **Curse of dimensionality!!!**

In practice:

- **Missing or false clusters**
- Loss of many **complex relationships** across dimensions
- Creation of **false relationships** through linearization

Spectral decomposition

In theory:

- Only based on the geometry of ”data surface”
- No assumption and previous projections to be based on
- Provides a basis of intrinsical dimensionality describing the effects dominating the data

In practice:

- **Rigorous decomposition** of the data
- Components can have **biological-technical meaning**
- Geometric distortion to **evaluate projections** of the eigenvectors
- **Clustering** of the data based only on geometrical information

Documentation

Tutorial