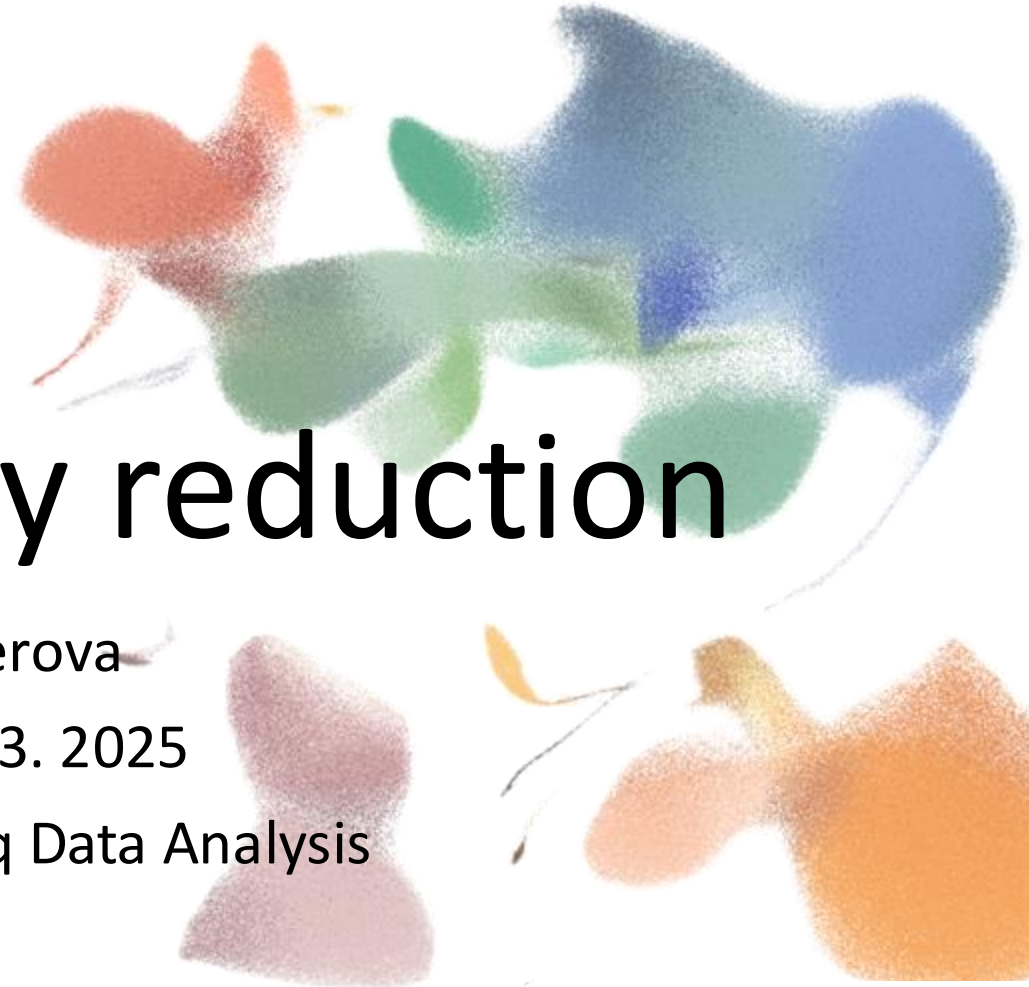


Dimensionality reduction

Lucie Pfeiferova

December 1.-3. 2025

Course on scRNA-seq Data Analysis

elixir CZECH REPUBLIC

IMG

- What is it
- How to choose features
- Which type – linear, graph based

Idea behind

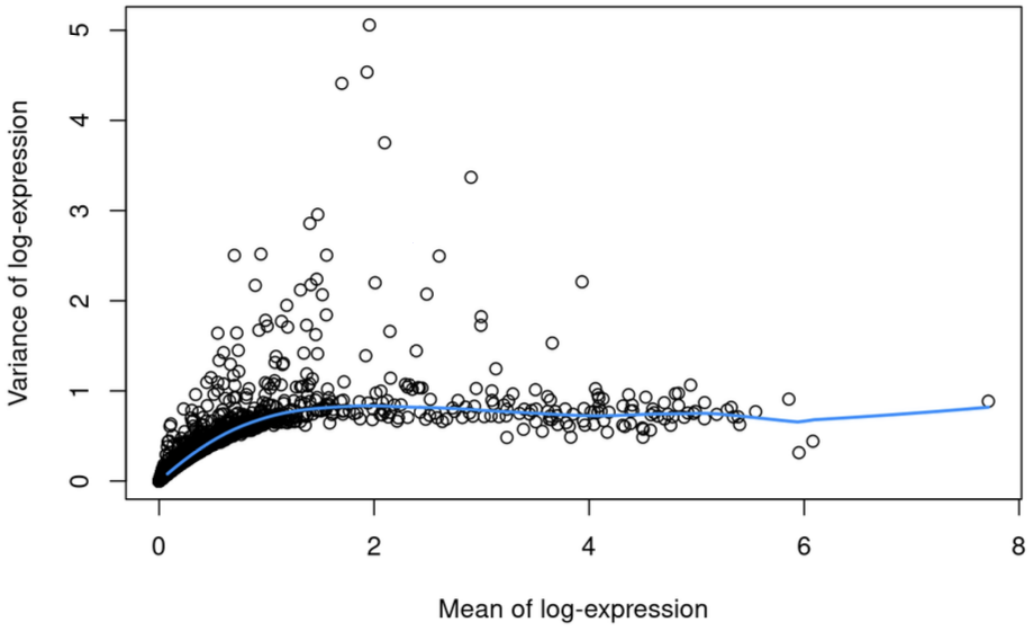
- In single-cell data we typically have thousands of genes across thousands (or millions!) of cells.
- Interpretation/visualisation beyond 2D is hard.
- As we increase the number of dimensions, our data becomes more sparse.
- High computational burden for downstream analysis (such as cell clustering)
- Solution: collapse the number of dimensions to a more manageable number, while preserving

PCA	linear	Matrix Factorization		
ICA	linear	Matrix Factorization		
MDS	non-linear	Matrix Factorization		
Sparce NNMF	non-linear	Matrix Factorization	2010	https://pdfs.semanticscholar.org/664d/40258f12ad28ed0b7d4c272935ad72a150db.pdf
cPCA	non-linear	Matrix Factorization	2018	https://doi.org/10.1038/s41467-018-04608-8
ZIFA	non-linear	Matrix Factorization	2015	https://doi.org/10.1186/s13059-015-0805-z
ZINB-WaVE	non-linear	Matrix Factorization	2018	https://doi.org/10.1038/s41467-017-02554-5
Diffusion maps	non-linear	graph-based	2005	https://doi.org/10.1073/pnas.0500334102
Isomap	non-linear	graph-based	2000	10.1126/science.290.5500.2319
t-SNE	non-linear	graph-based	2008	https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf
- BH t-SNE	non-linear	graph-based	2014	https://lvdmaaten.github.io/publications/papers/JMLR_2014.pdf
- Flt-SNE	non-linear	graph-based	2017	arXiv:1712.09005
LargeVis	non-linear	graph-based	2018	arXiv:1602.00370
UMAP	non-linear	graph-based	2018	arXiv:1802.03426
PHATE	non-linear	graph-based	2017	https://www.biorxiv.org/content/biorxiv/early/2018/06/28/120378.full.pdf
scvis	non-linear	Autoencoder (MF)	2018	https://doi.org/10.1038/s41467-018-04368-5
VASC	non-linear	Autoencoder (MF)	2018	https://doi.org/10.1016/j.gpb.2018.08.003

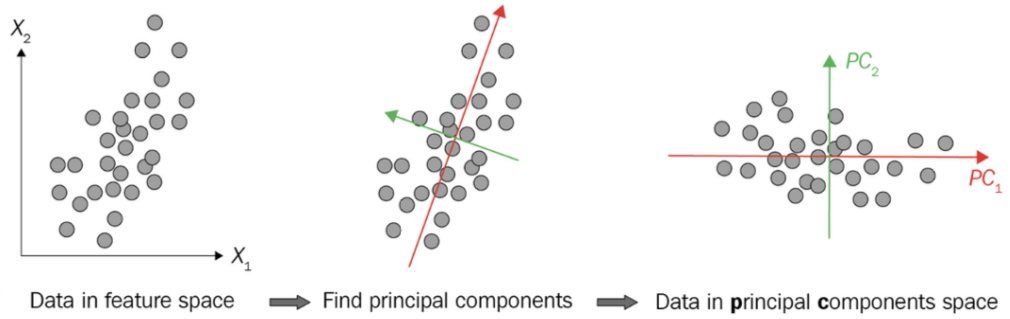
IMG

Feature selection

- Select genes which capture biologically-meaningful variation, while reducing the number of genes which
- only contribute to technical noise
- simplist the most variable genes based on their expression across the population
- quantifying per-gene variation is to compute the variance of the log-normalized expression values



PCA



- Principal Components Analysis (PCA) finds new axes (“principal components”) that capture the greatest variance in high-dimensional data.
- The **first principal component (PC1)** explains the largest amount of variation.
- Each subsequent PC is:
- **Orthogonal** to the previous ones.
- Chosen to capture the next-largest remaining variation.
- Result: PCA provides a compact, ordered set of components summarizing the major patterns in the data.

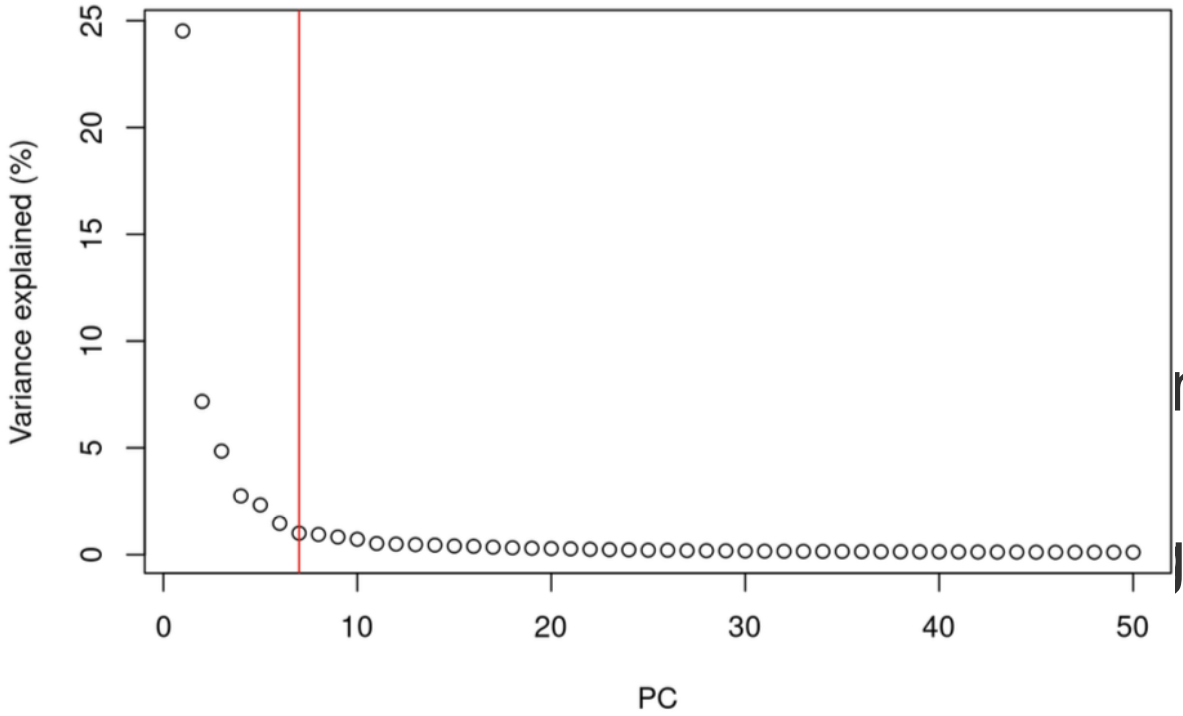
Idea behind PCA

- Biological processes affect *many genes together*, producing correlated patterns.
- Correlated variation is captured in **early PCs**, representing major biological structure.
- Random noise affects genes independently → cannot be captured well by any single PC.
- Noise tends to fall into **later PCs**, so using early PCs:
  - Enhances biological signal
  - Reduces noise
  - Speeds downstream analysis

Component selection

- After PCA we are still left with as many dimensions in our data as we started
- “reasonable” but arbitrary value, typically ranging from 10 to 50
- Elbow Plot as Visual Decision Method
- Look for the **“elbow”**—the point where adding more PCs yields little additional variance.
- PCs **before** the elbow capture the most meaningful structure.
- Reduces the risk of:
  - Keeping noise-dominated PCs
  - Using arbitrary thresholds

Compone



- After PCA w ns in our data as we starte
- “reasonable’ from 10 to 50
- Elbow Plot as
- Look for the **“elbow”**—the point where adding more PCs yields little additional variance.
- PCs **before** the elbow capture the most meaningful structure.
- Reduces the risk of:
  - Keeping noise-dominated PCs
  - Using arbitrary thresholds

IMG

Non-linear dimensionality reduction methods

- Graph-based, non-linear methods: t-SNE (t-stochastic neighbor embedding) and UMAP (Uniform manifold approximation and projection)

- These methods can run on the output of the PCA, which speeds their computation and can make the results more robust to noise
- t-SNE and UMAP should only be used for visualisation, not as input for downstream analysis

- *PCA is used before other dimensionality reduction methods to UMAP and t-SNE to remove noise, reduce dimensionality, emphasize biological structure, and produce faster, more stable, and more reliable low-dimensional embeddings.*

t-SNE

- t-SNE (**t-Distributed Stochastic Neighbor Embedding**) is a nonlinear dimensionality reduction method used mainly for **visualizing high-dimensional data** in 2D or 3D.
- It focuses on preserving **local neighborhoods**, meaning it keeps similar cells close together in the plot.
- Distances between clusters, cluster sizes, and global layout **are not meaningful** due to strong geometric distortion.
- Highly effective for exploring cluster structure but **not suitable** for downstream analysis (e.g., clustering, trajectories).
- Sensitive to random initialization; different runs can produce different embeddings.

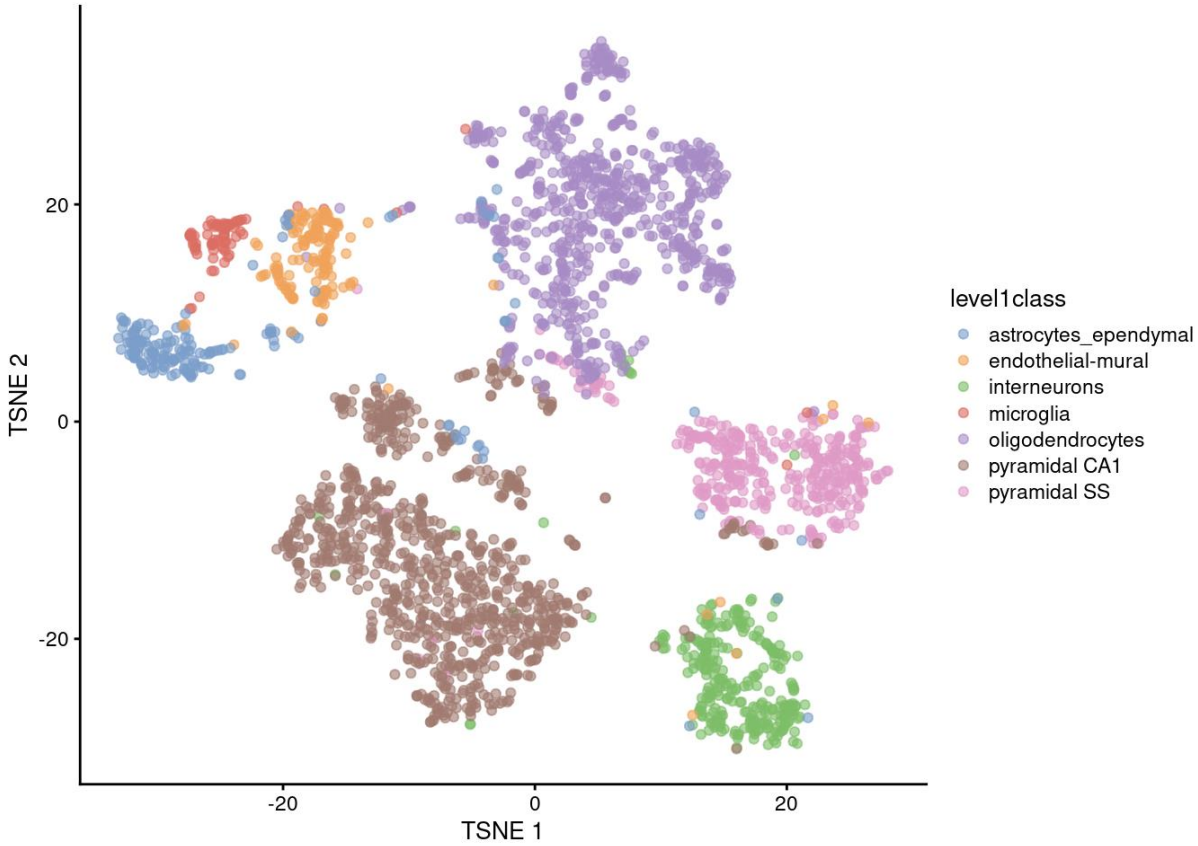
UMAP

- UMAP (**Uniform Manifold Approximation and Projection**) is a visualization method that creates a low-dimensional representation while preserving **local structure** and some **global structure**.
- Usually faster, more stable, and better at maintaining overall organization than t-SNE.
- Still performs heavy non-linear transformations, so **distances and positions are not quantitatively accurate**.
- Very useful for visually identifying patterns and relationships between cell populations.
- Like t-SNE, UMAP should be used for **visualization only**, not as input for clustering or trajectory analysis.

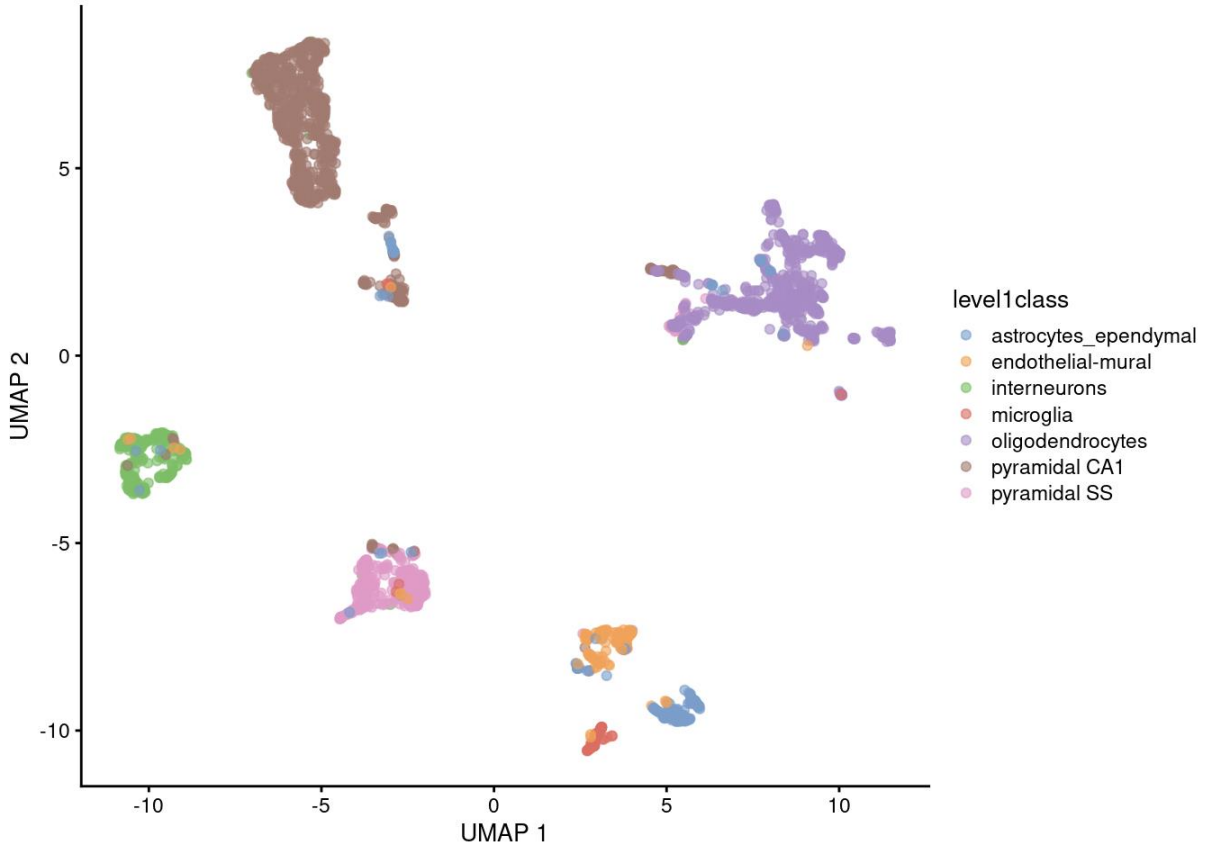
Differences between t-SNE and UMAP

Feature	t-SNE	UMAP
Structure	Preserves **local** neighborhoods only	Preserves **local + more global** structure
Stability	Sensitive to randomness; may vary between runs	More stable and reproducible
Speed	Slower; less scalable for large datasets	Faster; scales well to large datasets
Cluster appearance	Clusters often exaggerated/separated	Clusters smoother and more realistic
Interpretability	Good for tight clusters; global layout unreliable	Easier to interpret cluster relationships
Primary use	Visualization only	Visualization only, but more informative

IMG



t-SNE



UMAP

IMG

Summary

- Dimensionality reduction methods simplify high-dimensional data while preserving biological signal.
- Common methods in scRNA-seq analysis include PCA, t-SNE, and UMAP.
- PCA transforms the data linearly to capture the main variance and reduce the dimensionality from thousands of genes to a few principal components.
- PCA results can be utilized for downstream analysis like cell clustering and trajectory analysis, and as input for non-linear methods such as t-SNE and UMAP.
- t-SNE and UMAP are non-linear methods that group similar cells and separate dissimilar cell clusters.
- These non-linear methods are primarily for data visualization, not for downstream analysis.

IMG

- https://sib-swiss.github.io/single-cell-r-training/
- https://bioconductor.org/books/3.14/OSCA.basic/
- https://bioconductor.org/books/release/OSCA/