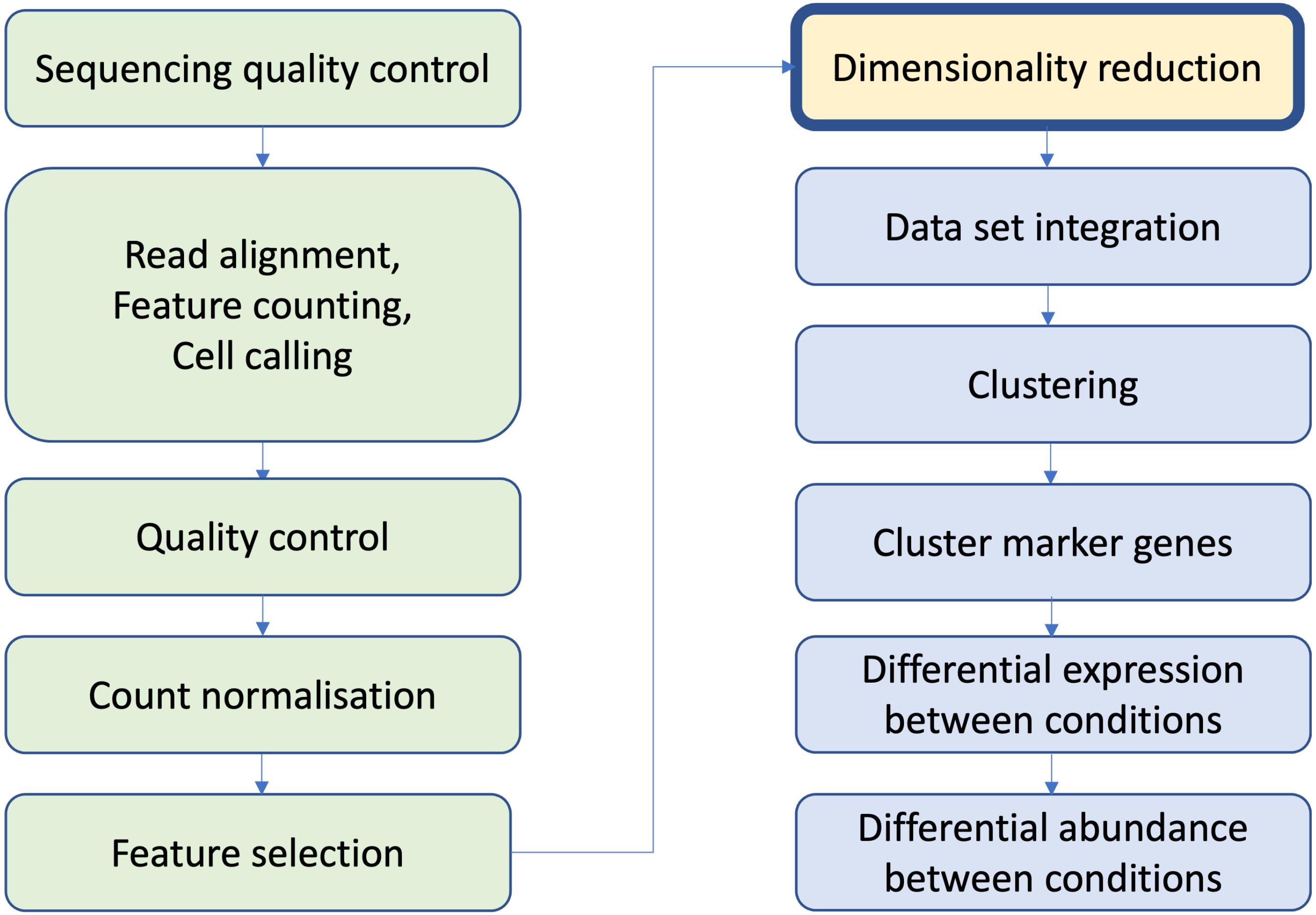
UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

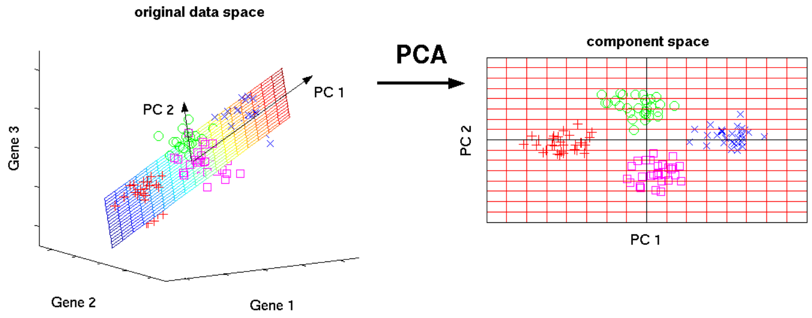
Feature Selection and Dimensionality Reduction

September 2022

Single Cell RNAseq Analysis Workflow

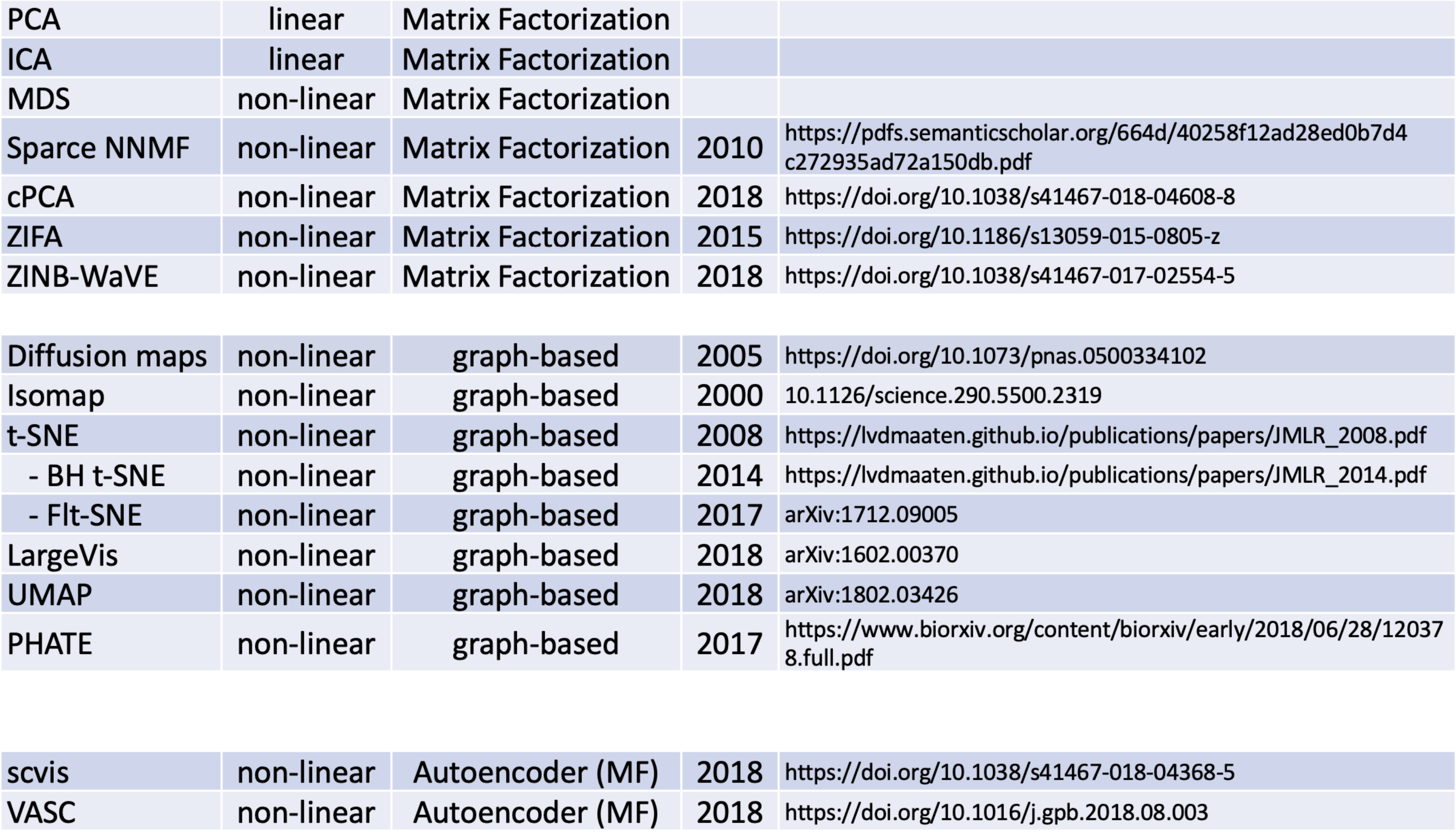

Sequencing quality control
Read alignment, Feature counting, Cell calling
Quality control
Count normalisation
Feature selection
Dimensionality reduction
Data set integration
Clustering
Cluster marker genes
Differential expression between conditions
Differential abundance between conditions

Why do high-dimensional data pose a problem?



- Single cell analysis is aimed at cluster the cells according to their gene expression
- Thousands of genes across the cells
- Problem with high-dimensional data
    - Human intuition and understanding is limited to a three dimensional world
    - As we increase the number of Dimensions, our data becomes more sparse. The average distance in between two points of our data set is increased and invariant.

There are many dimensionality reduction algorithms



➡ 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

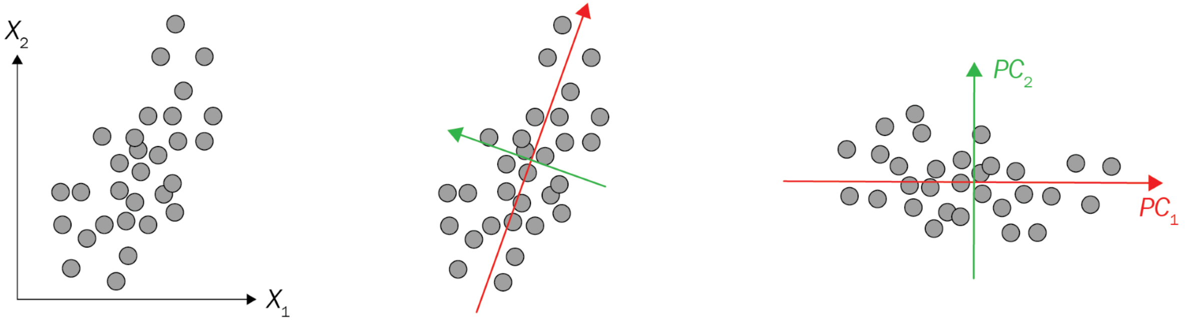
Diffusion maps	non-linear	graph-based	2005	https://doi.org/10.1073/pnas.0500334102
Isomap	non-linear	graph-based	2000	10.1126/science.290.5500.2319
➡ t-SNE	non-linear	graph-based	2008	https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf
- BH t-SNE	non-linear	graph-based	2014	https://lvdmaaten.github.io/publications/papers/JMLR_2014.pdf
- 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

Which genes should we use for downstream analysis?

- We want to select genes that contain biologically meaningful variation, while reducing the number of genes which only contribute with technical noise
- We can model the gene-variance relationship across all genes to define a data-driven "technical variation threshold"
- From this we can select **highly variable genes** (HVGs) for downstream analysis (e.g. PCA and clustering)

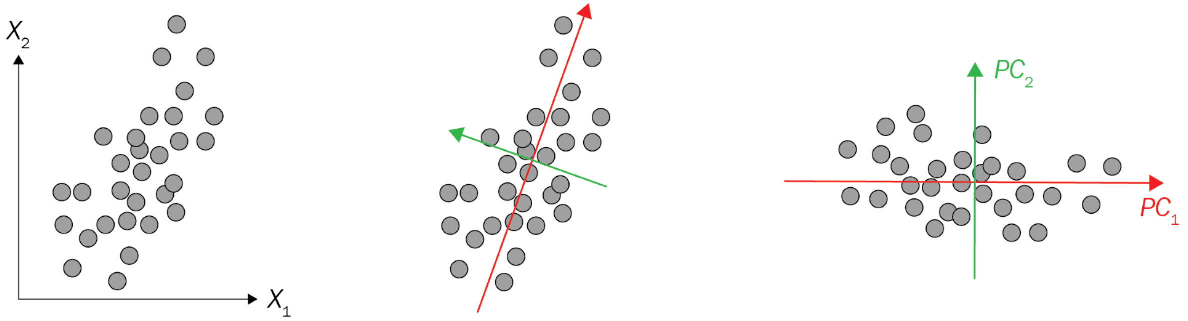
Principal Components Analysis (PCA)



Data in feature space → Find principal components → Data in **p**rincipal **c**omponents space

- It’s a linear algebraic method of dimensionality reduction
- Finds principal components (PCs) of the data
  - Directions where the data is most spread out (highest variance)
  - PC1 explains most of the variance in the data, then PC2, PC3, etc.
  - PCA is primarily a dimension reduction technique, but it is also useful for visuvalization
  - A good separation of dissimilar objects is provided
  - Preserves the global data structure

Principal Components Analysis (PCA)

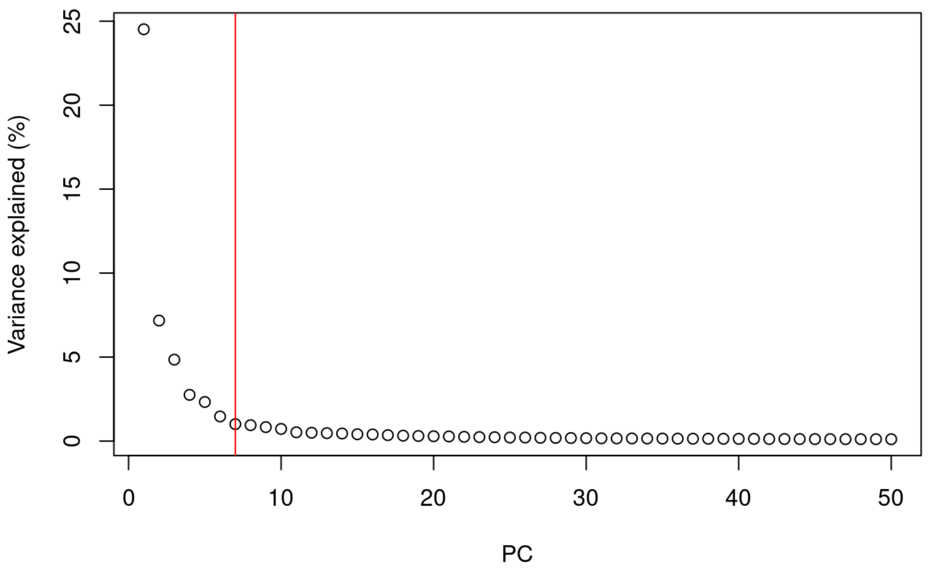


Data in feature space ⟹ Find principal components ⟹ Data in **p**rincipal **c**omponents space

- When data is very highly-dimensional, we can select the most important PCs only, and use them for downstream analysis (e.g. clustering cells)
    - This reduces the dimensionality of the data from ~20,000 genes to maybe 10-20 PCs
    - Each PC represents a robust ‘metagene’ that combines information across a correlated gene set
- Prior to PCA we scale the data so that genes have equal weight in downstream analysis and highly expressed genes don’t dominate

How many principal components for downstream analysis?

After performing PCA we are still left with as many dimensions in our data as we started



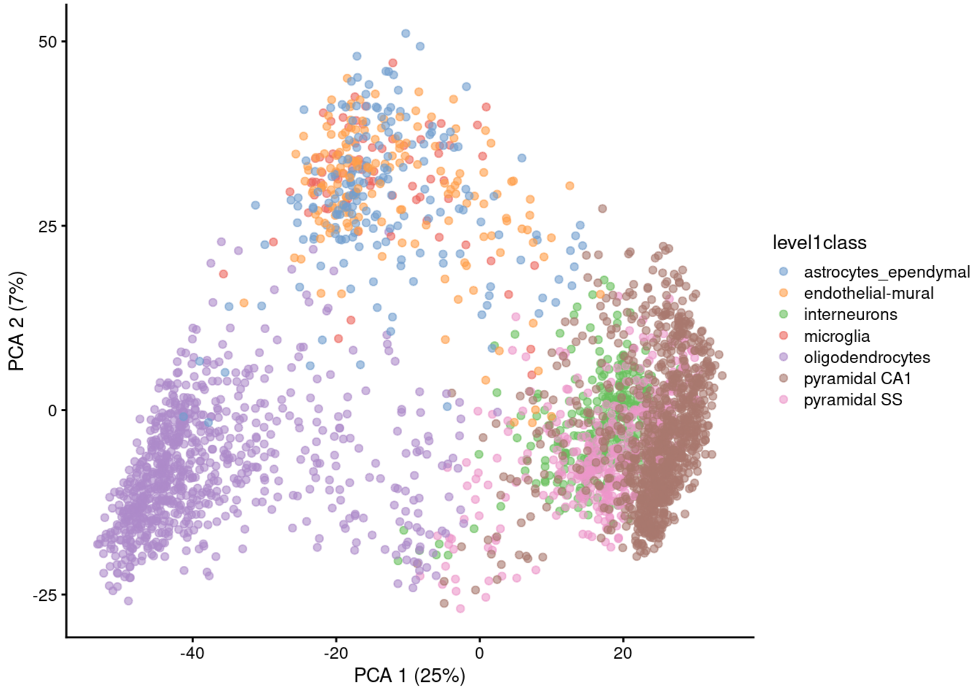
But our principal components progressively capture less variation in the data

How do we select the number of PCs to retain for downstream analysis?

- Using the “Elbow” method on the scree plot
- Using the model of technical noise (shown earlier)
- Trying downstream analysis with different number of PCs (10, 20, or even 50)

Visualizing PCA results: PC scores

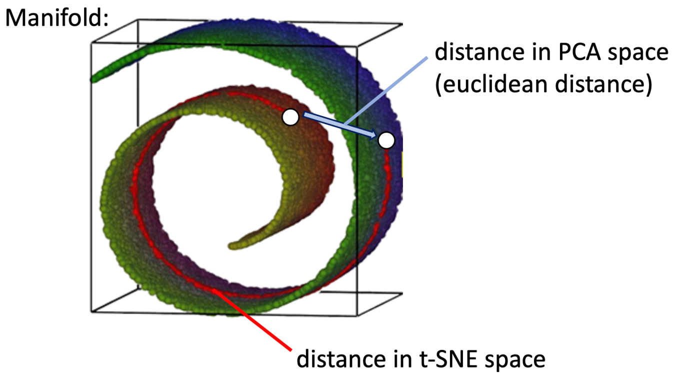
Because PC1 and PC2 capture most of the variance of the data, it is common to visualise the data projected onto those two new dimensions.



Gene expression patterns will be captured by PCs -> PCA can separate cell types

Note that PCA can also capture other things, like sequencing depth or cell heterogeneity/complexity!

Other dimensionality reduction methods

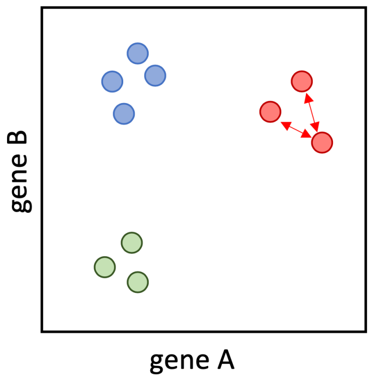


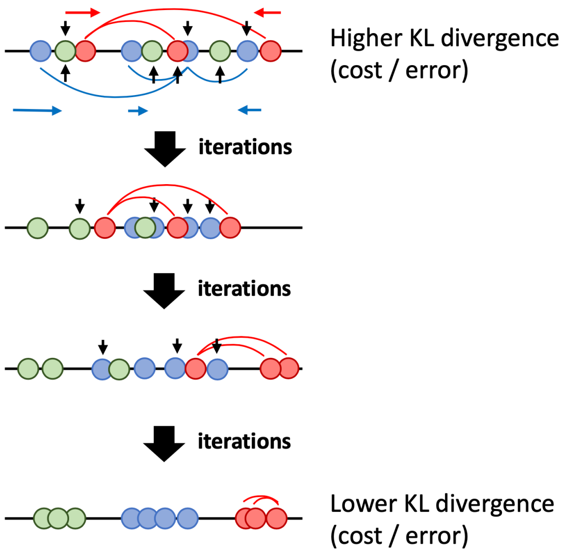
Graph-based, non-linear methods: **UMAP** and **t-SNE**

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

t-Distributed Stochastic Neighbor Embedding (t-SNE)





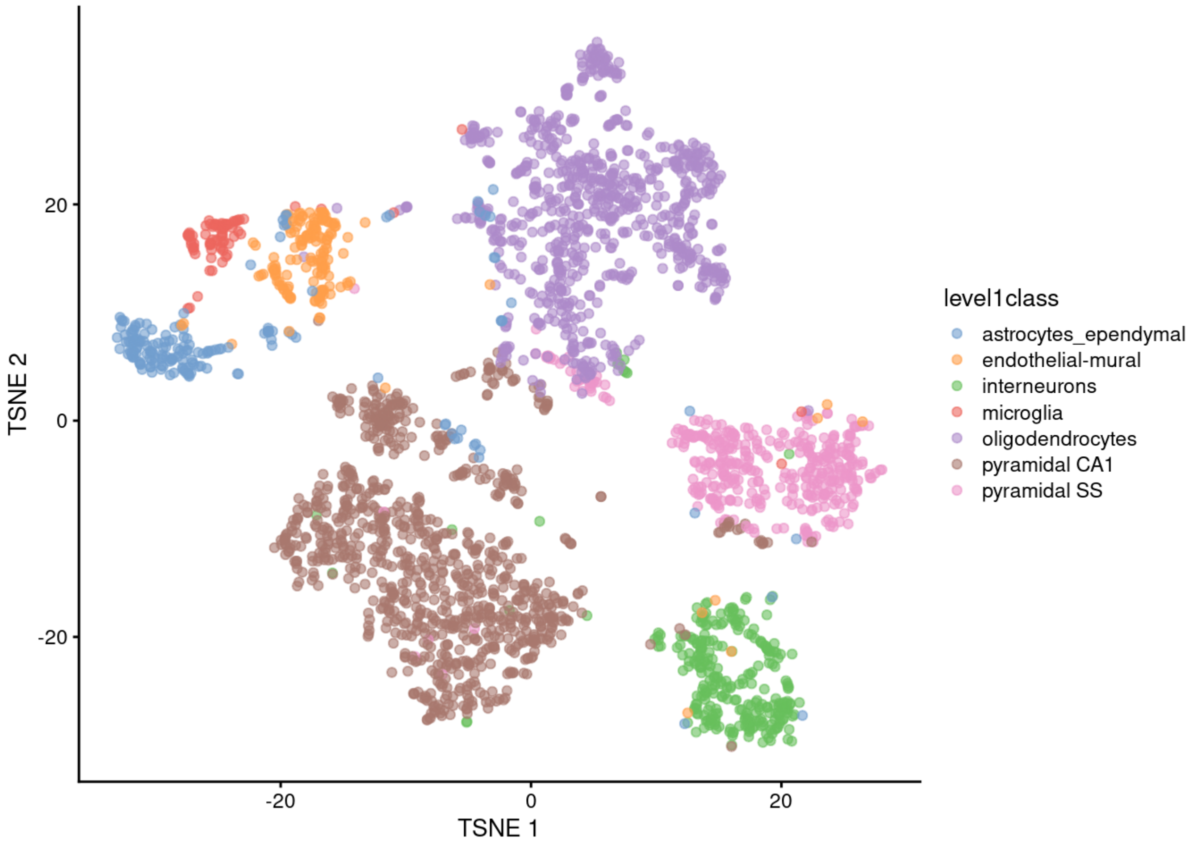
It has a stochastic step (results vary every time you run it)

Only local distances are preserved, while distances between groups are not always meaningful

Some parameters dramatically affect the resulting projection (in particular “perplexity”)

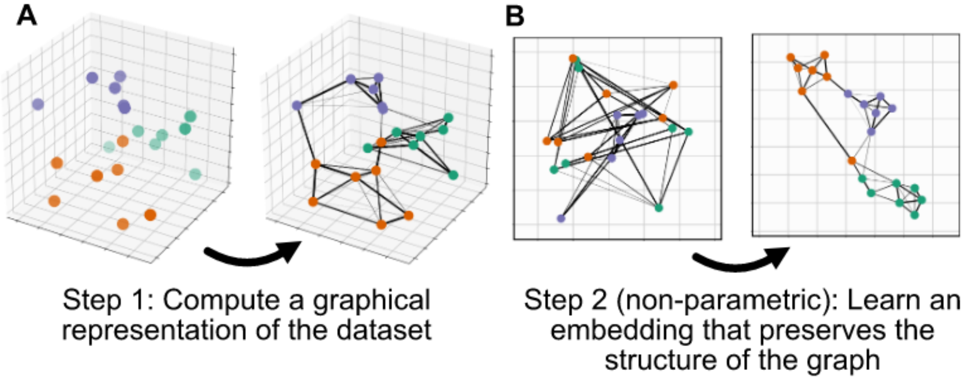
Learn more about how t-SNE works from this video: [StatQuest: t-SNE, Clearly Explained]

t-SNE


TSNE 2
TSNE 1
20
0
-20
-20
0
20
level1class
astrocytes_ependymal
endothelial-mural
interneurons
microglia
oligodendrocytes
pyramidal CA1
pyramidal SS

UMAP

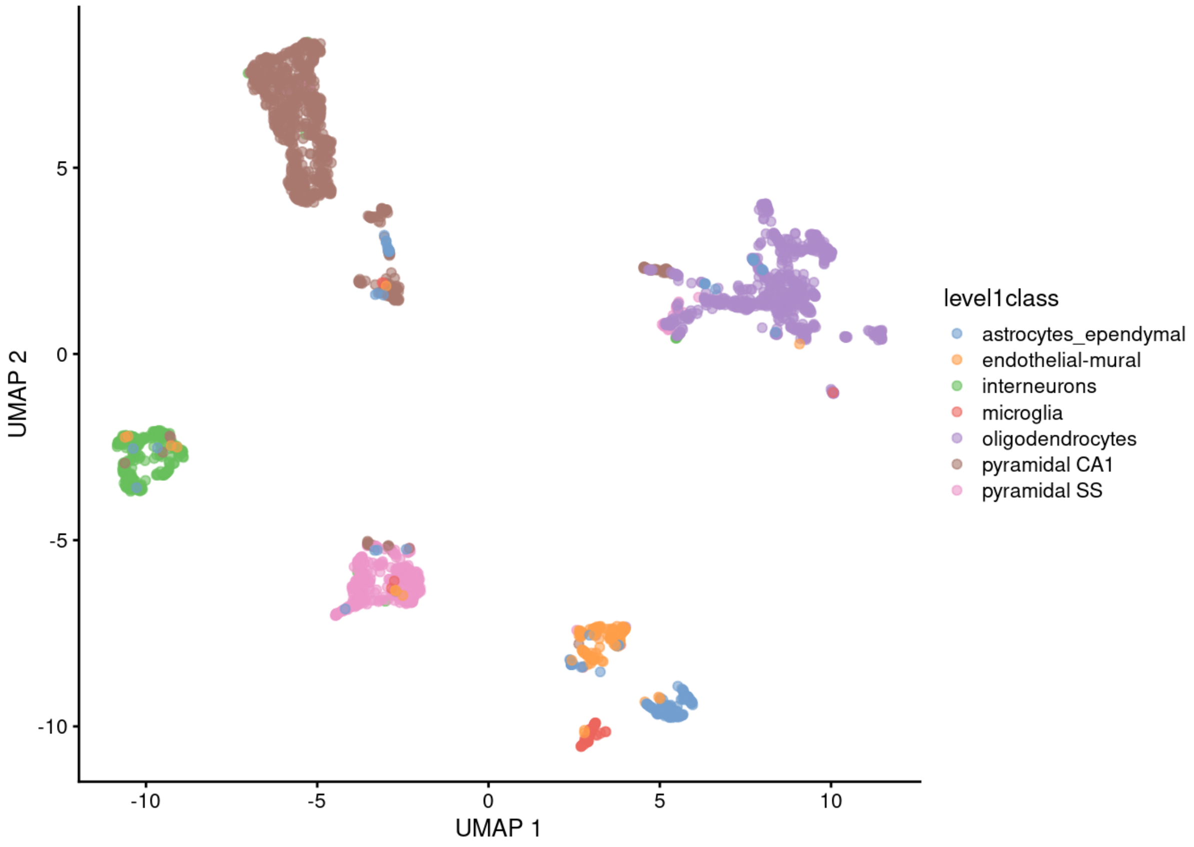
- Non-linear graph-based dimension reduction method like t-SNE
- Newer & efficient = fast
- Runs on top of PCs
- Based on topological structures in multidimensional space
- Unlike tSNE, you can compute the structure once (no randomization)
  - faster
  - you could add data points without starting over
- Preserves the global structure better than t-SNE



Step 1: Compute a graphical representation of the dataset

Step 2 (non-parametric): Learn an embedding that preserves the structure of the graph

UMAP



UMAP 2
UMAP 1
5
0
-5
-10
-10
-5
0
5
10
level1class
astrocytes_ependymal
endothelial-mural
interneurons
microglia
oligodendrocytes
pyramidal CA1
pyramidal SS

Key Points

- Dimensionality reduction methods allow us to represent high-dimensional data in lower dimensions, while retaining biological signal.
- The most common methods used in scRNA-seq analysis are PCA, t-SNE and UMAP.
- PCA uses a linear transformation of the data, which aims at defining new dimensions (axis) that capture most of the variance observed in the original data. This allows to reduce the dimension of our data from thousands of genes to 10-20 principal components.
- The results of PCA can be used in downstream analysis such as cell clustering, trajectory analysis and even as input to non-linear dimensionality reduction methods such as t-SNE and UMAP.
- t-SNE and UMAP are both non-linear methods of dimensionality reduction. They aim at keeping similar cells together and dissimilar clusters of cells apart from each other.
- Because these methods are non-linear, they should only be used for data visualisation, and not for downstream analysis.

Acknowledgments

Slides are adapted from Paulo Czarnewski and Zeynep Kalender-Atak

References (image sources):

- Orchestrating Single-Cell Analysis with Bioconductor
- Parametric UMAP embeddings for representation and semi-supervised learning