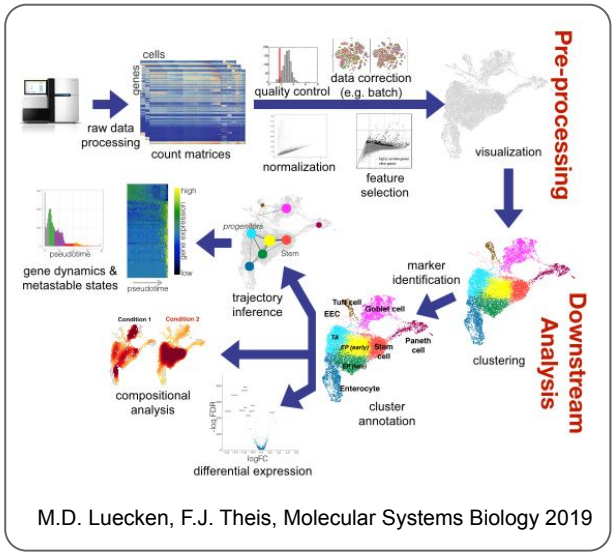
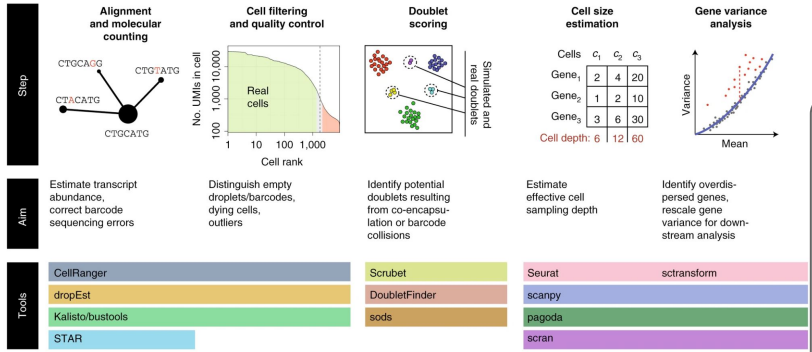
The Basic scRNAseq Analysis Workflow

Cell Discovery Network

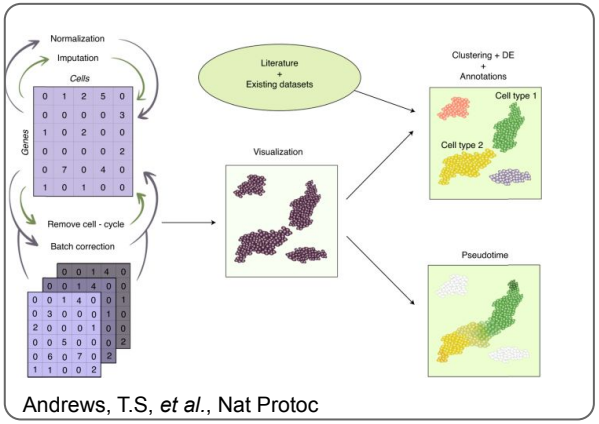
What does a standard workflow look like?



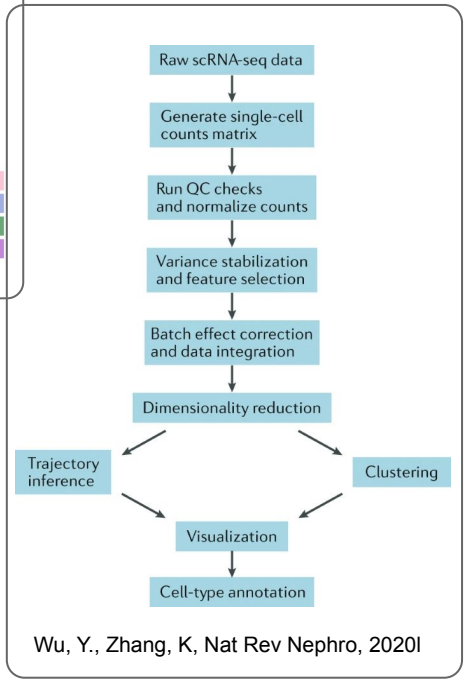
M.D. Luecken, F.J. Theis, Molecular Systems Biology 2019



Peter Kharchenko, Nat Methods, 2021



Andrews, T.S, *et al.*, Nat Protoc



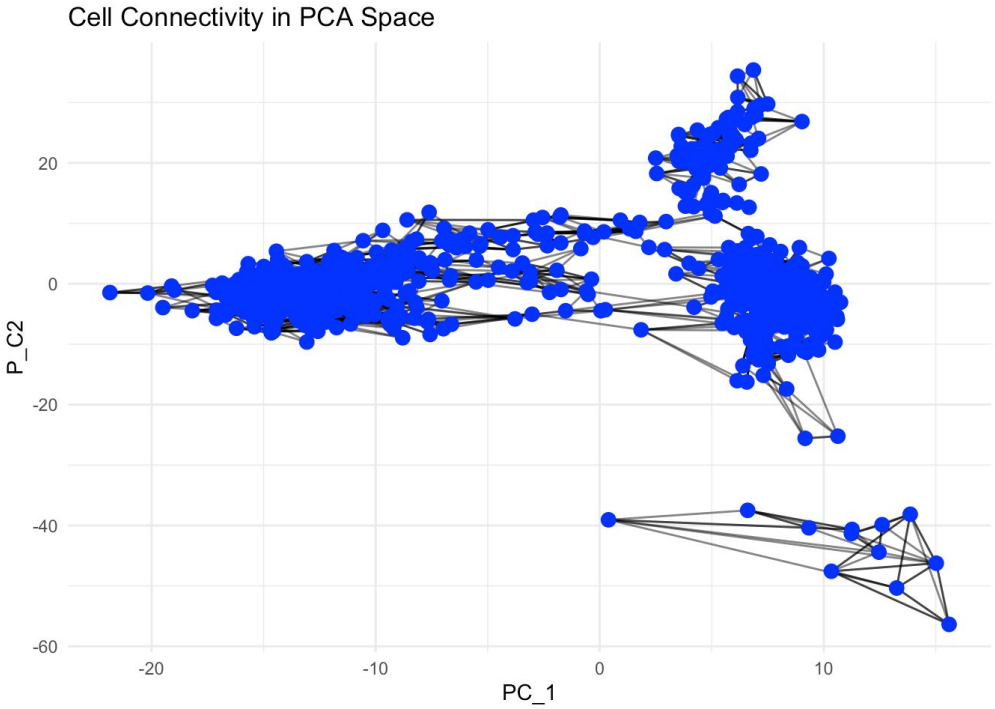
Wu, Y., Zhang, K, Nat Rev Nephro, 2020l

The Goal - KNN graph

Ultimately we want to build a KNN/SNN graph of our dataset.

Each node is a cell and each edge is the similarity between two cells.

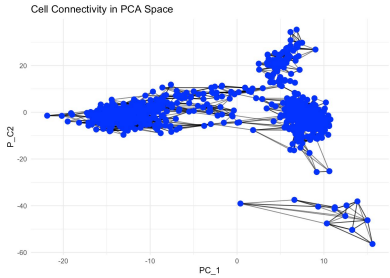
But… how do we measure this similarity?



Challenges in unsupervised clustering of single-cell RNA-seq data

Where we start - Count Matrix

		100K cells			
		Cell1	Cell2	...	CellN
30K genes	Gene1	3	2	.	13
	Gene2	2	3	.	1
	Gene3	1	14	.	18
	...	.	.	.	.
	...	.	.	.	.
	...	.	.	.	.
	GeneM	25	0	.	0



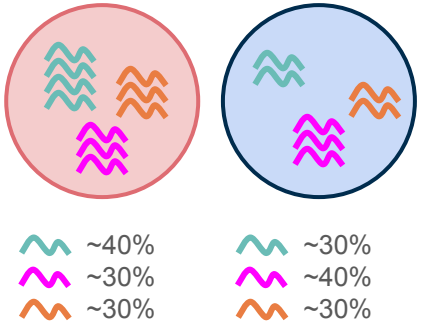
⛔ We can’t compute distances in this high dimensional space

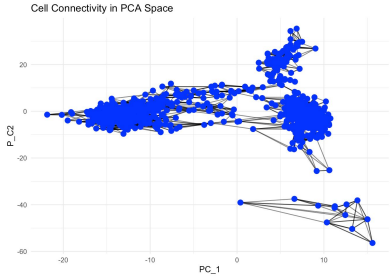
⛔ Not memory efficient or computationally feasible

✅ We need to reduce the dimensions

Intermediate steps - Data Normalization

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0



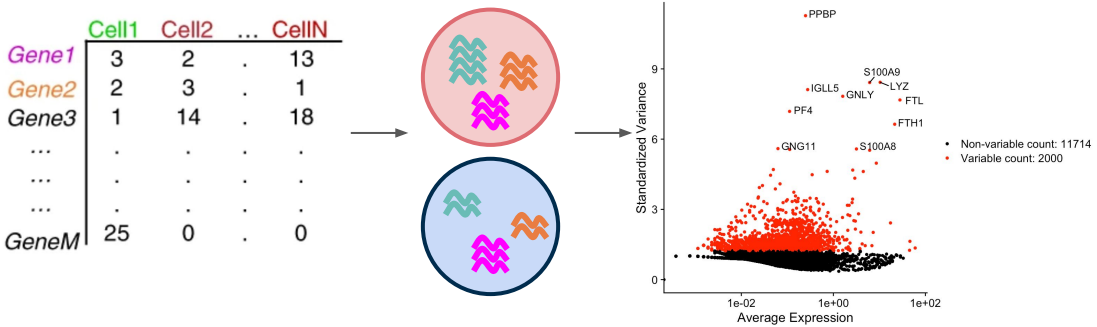


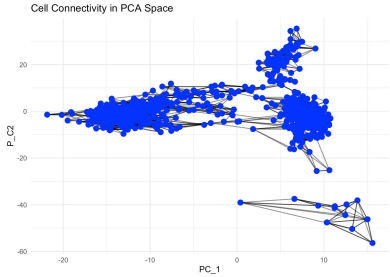
⛔ We can’t compare raw counts

⛔ scRNAseq has a lot of technical noise

✅ We need to normalize by library size and log

Intermediate steps - Feature Selection



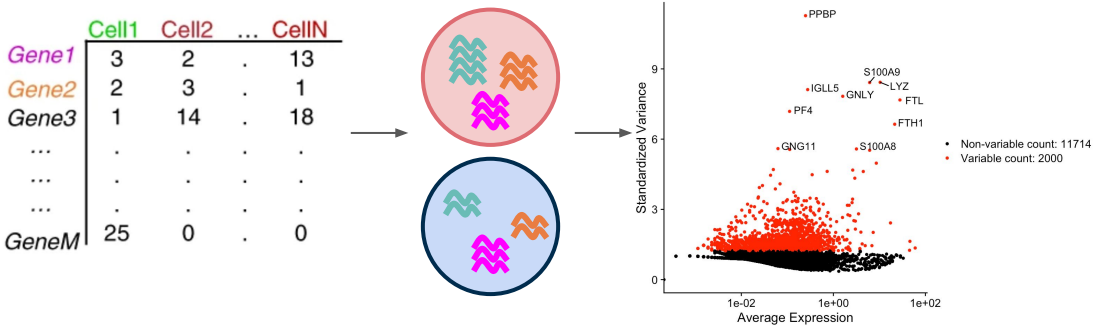


⛔ We can’t do PCA using all genes

✅ We aim to to subset our genes to the most biologically relevant

✅ We select top X most informative (variable) genes

Intermediate steps - Feature Selection

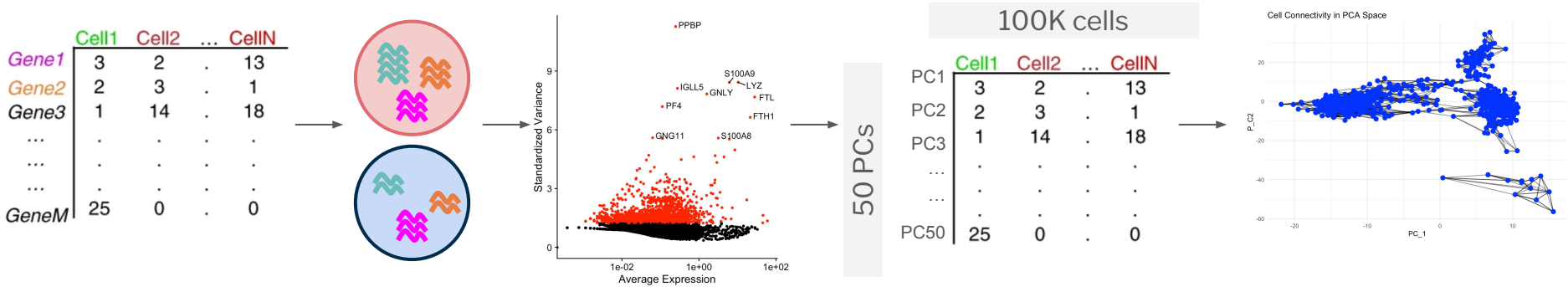


⛔ We can't do PCA using all genes

✅ We aim to to subset our genes to the most biologically relevant

✅ We select top X most informative (variable) genes

Intermediate steps - PCA

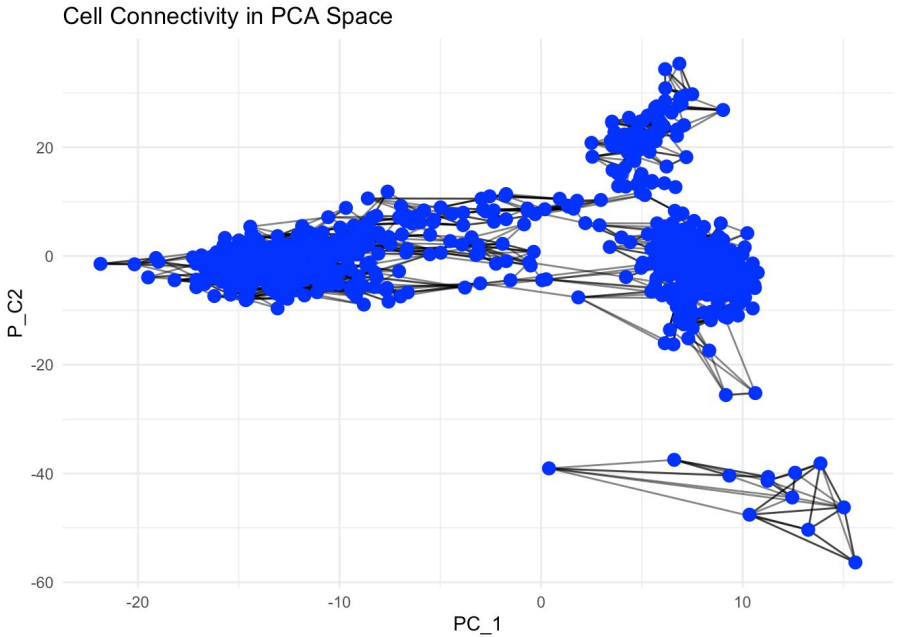


⛔ We can’t compute distances in this high dimensional space

✅ PCA reduces the dimensions to <50 learning orthogonal information

✅ We can also use NMF, scVI, SVD…

Downstream steps - Clustering

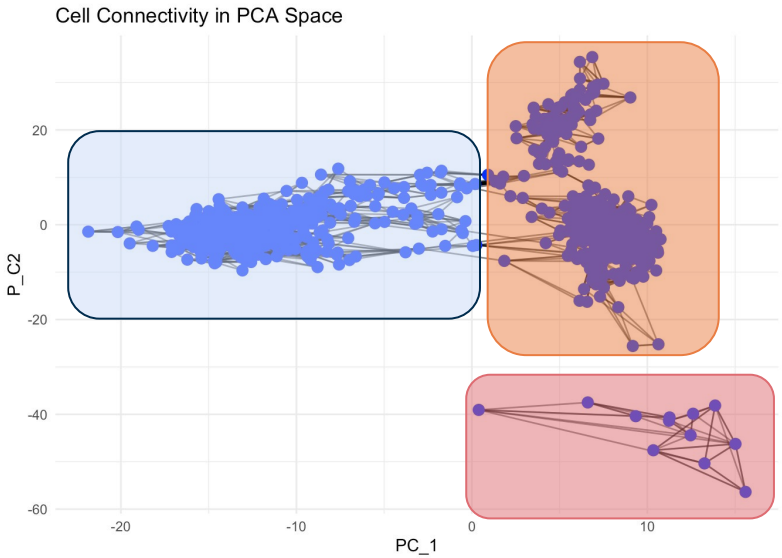


⛔ Classical methods like k-means or hierarchical clustering don’t scale

✅ We can use community detection algorithms on the SNN -

Challenges in unsupervised clustering of single-cell RNA-seq data

Downstream steps - Clustering

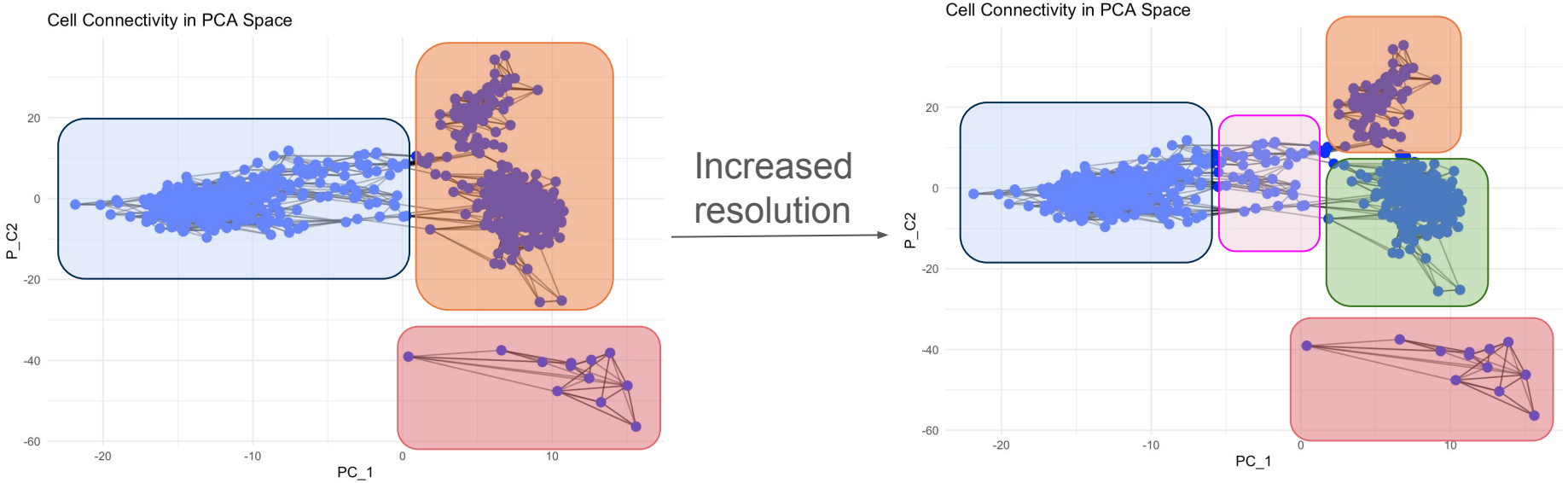


⛔ Classical methods like k-means or hierarchical clustering don’t scale

✅ We can use community detection algorithms on the SNN -

4min explanation here & Leiden paper

Downstream steps - Clustering

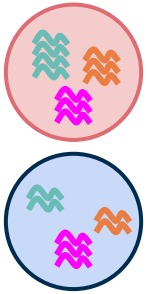


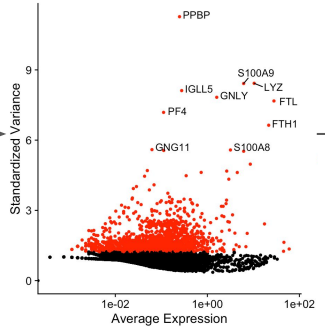
✅ Community detection algorithms have a resolution parameter

4min explanation here & Leiden paper

Downstream steps - Clustering

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

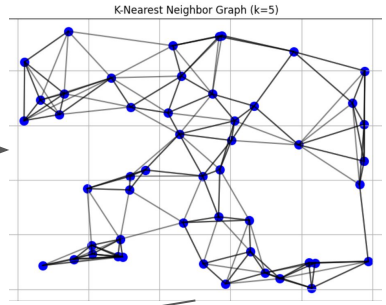


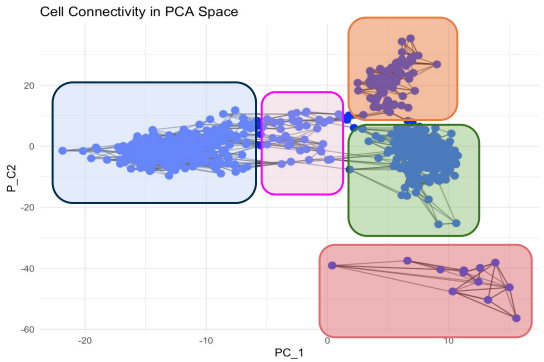


50 PCs

100K cells

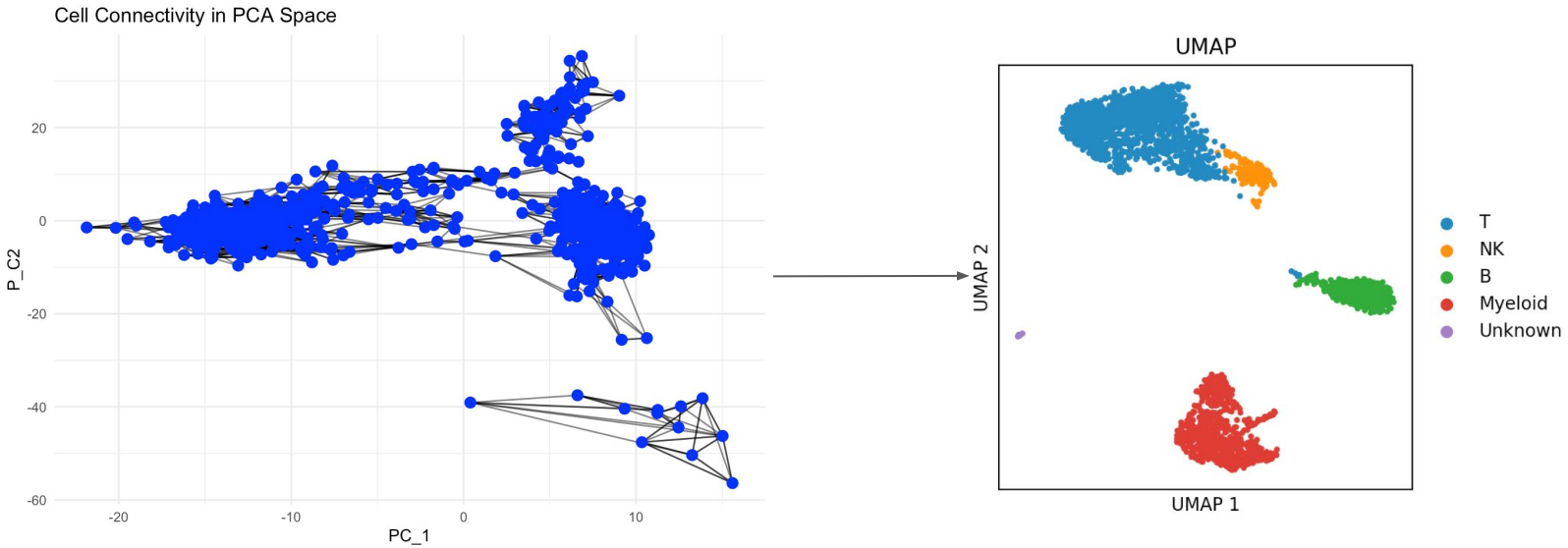
	Cell1	Cell2	...	CellN
PC1	3	2	.	13
PC2	2	3	.	1
PC3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
PC50	25	0	.	0





Clustering

Downstream steps - UMAP visualization

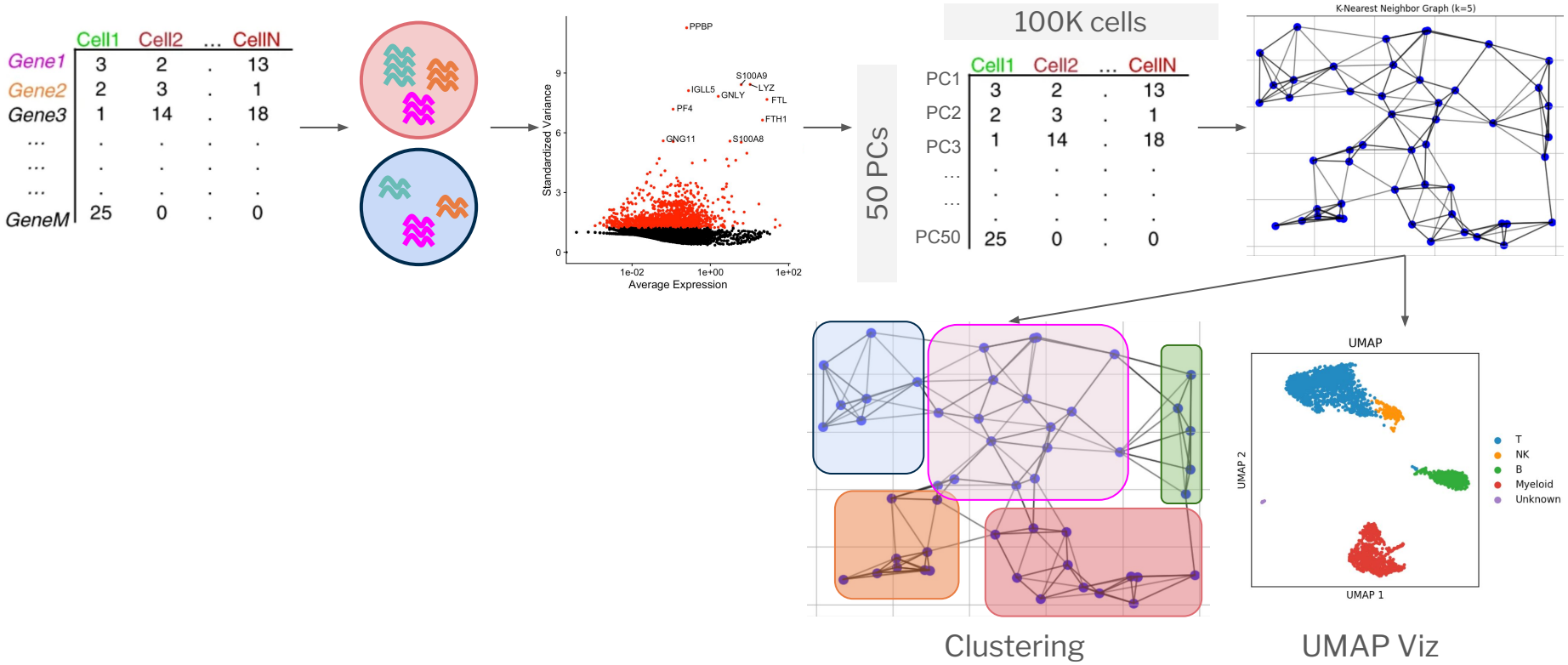


⛔ Classical methods like KNN or hierarchical clustering don't scale

✅ We can use community detection algorithms on the SNN -

15min explanation

Downstream Analysis



Downstream Analysis