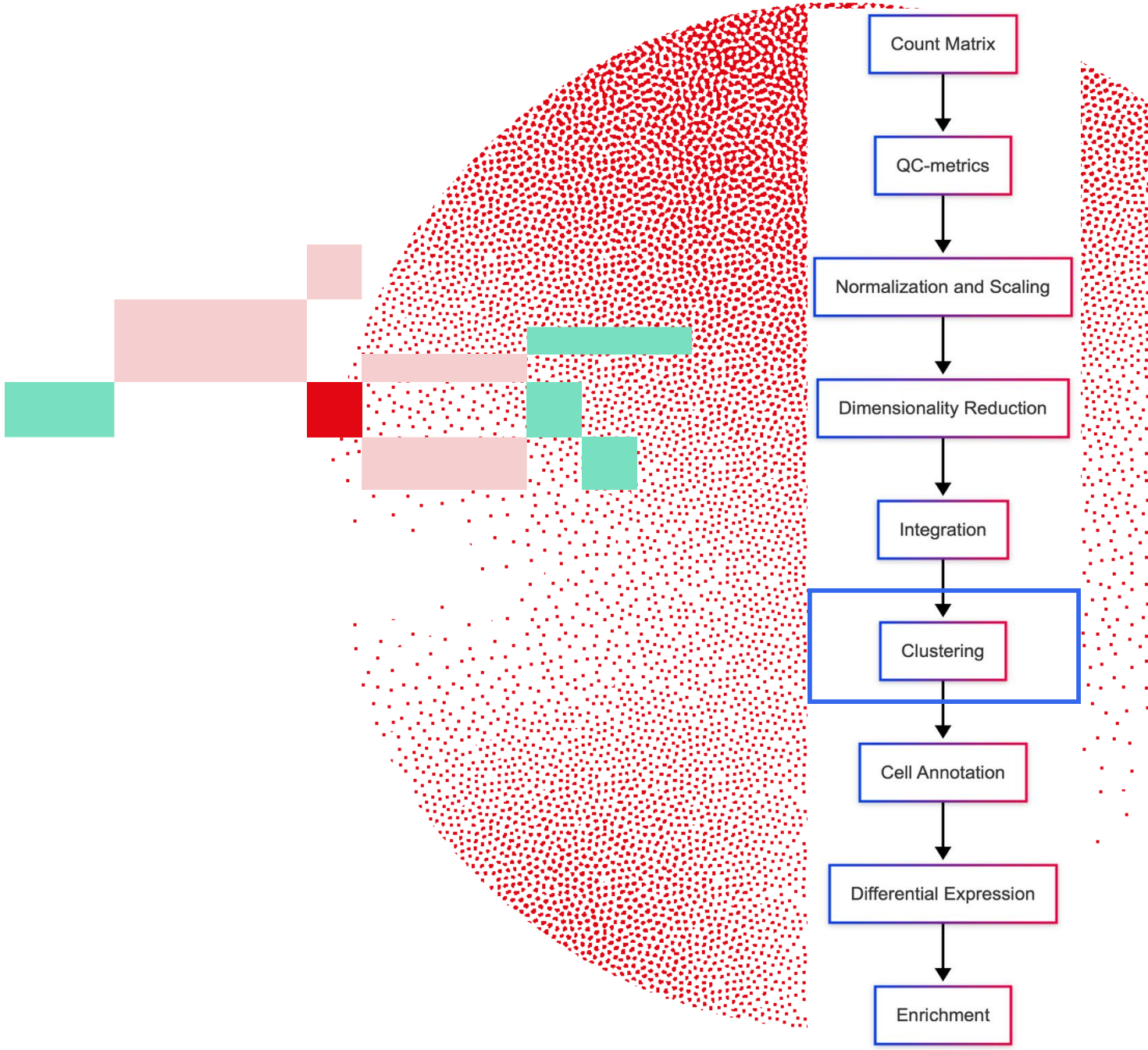
Swiss Institute of Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

Clustering



Joana Carlevaro Fita

July 2-4, 2025

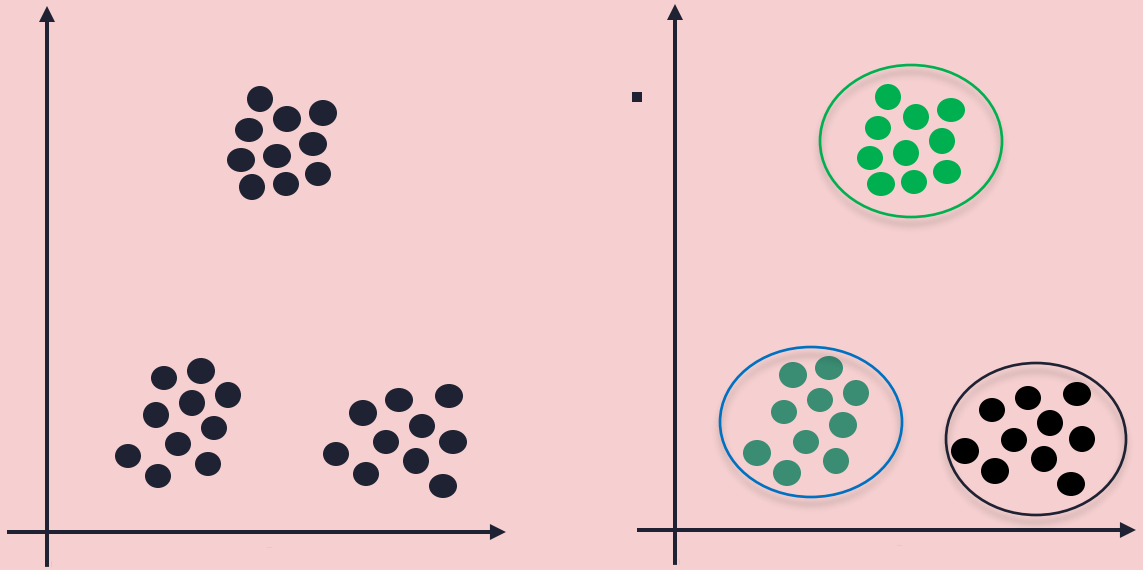
Adapted from previous year courses

elixir SWITZERLAND

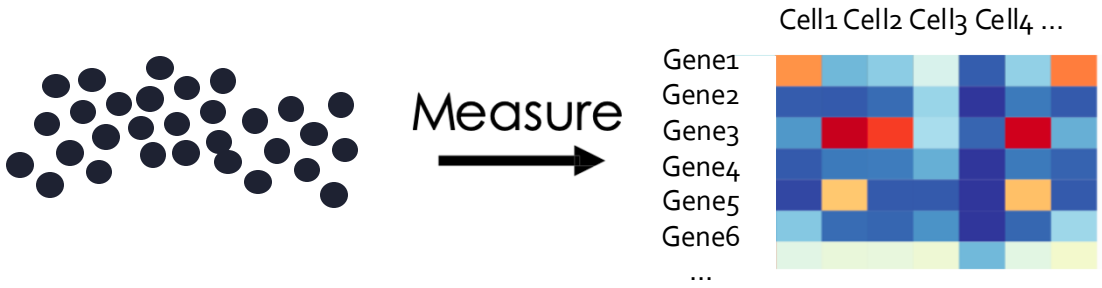
Learning objectives

- Understand the importance of clustering in single-cell RNA-seq data analysis
- Apply graph-based clustering techniques, including the Louvain algorithm, to define cell populations
- Understand the main steps and challenges during clustering workflow

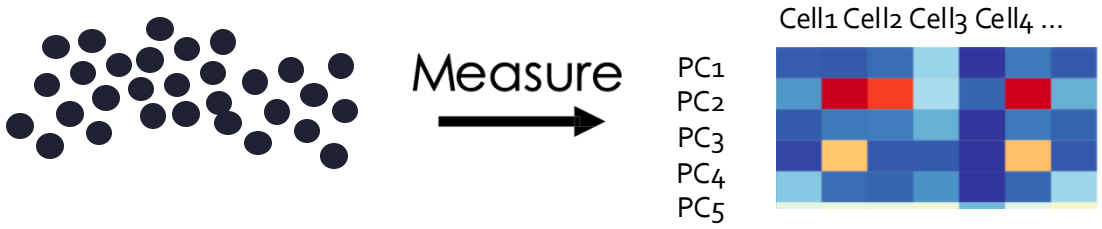
In single-cell RNA-seq, we use unsupervised clustering to empirically define groups of cells with similar expression profiles.



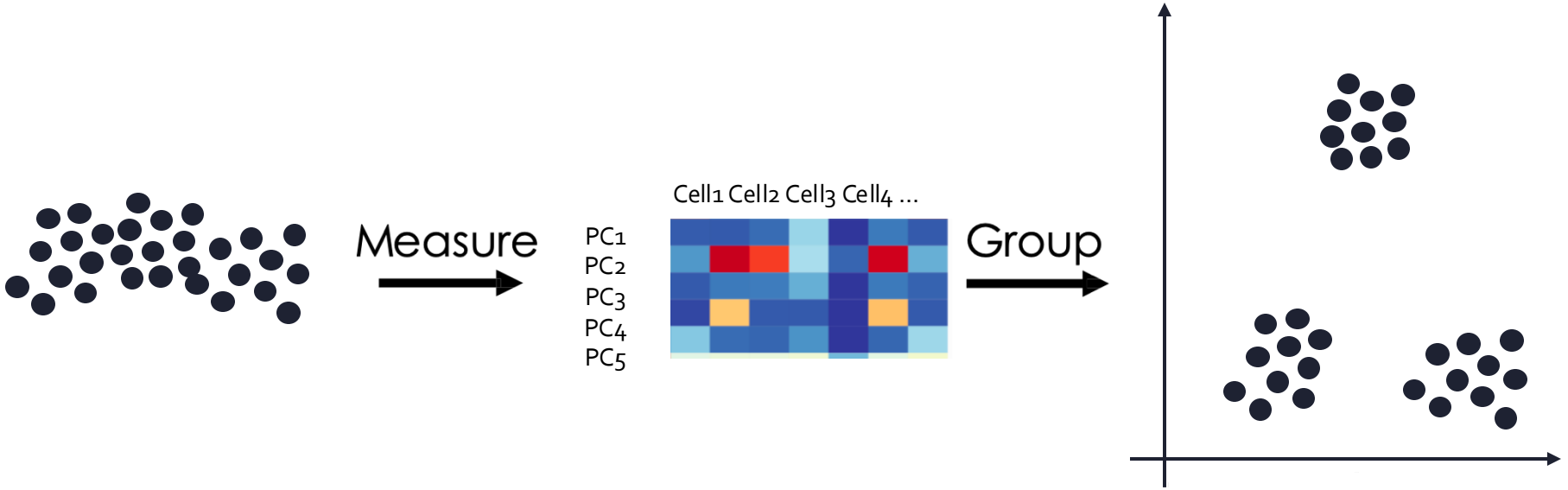
Why clustering?



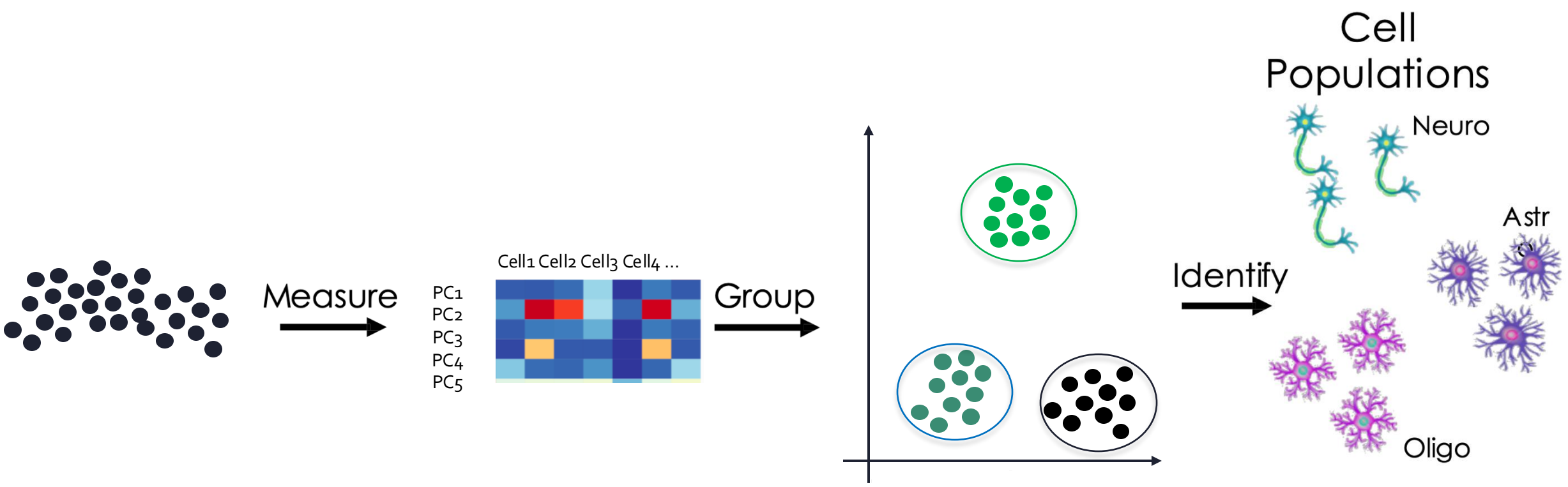
Why clustering?



Why clustering?



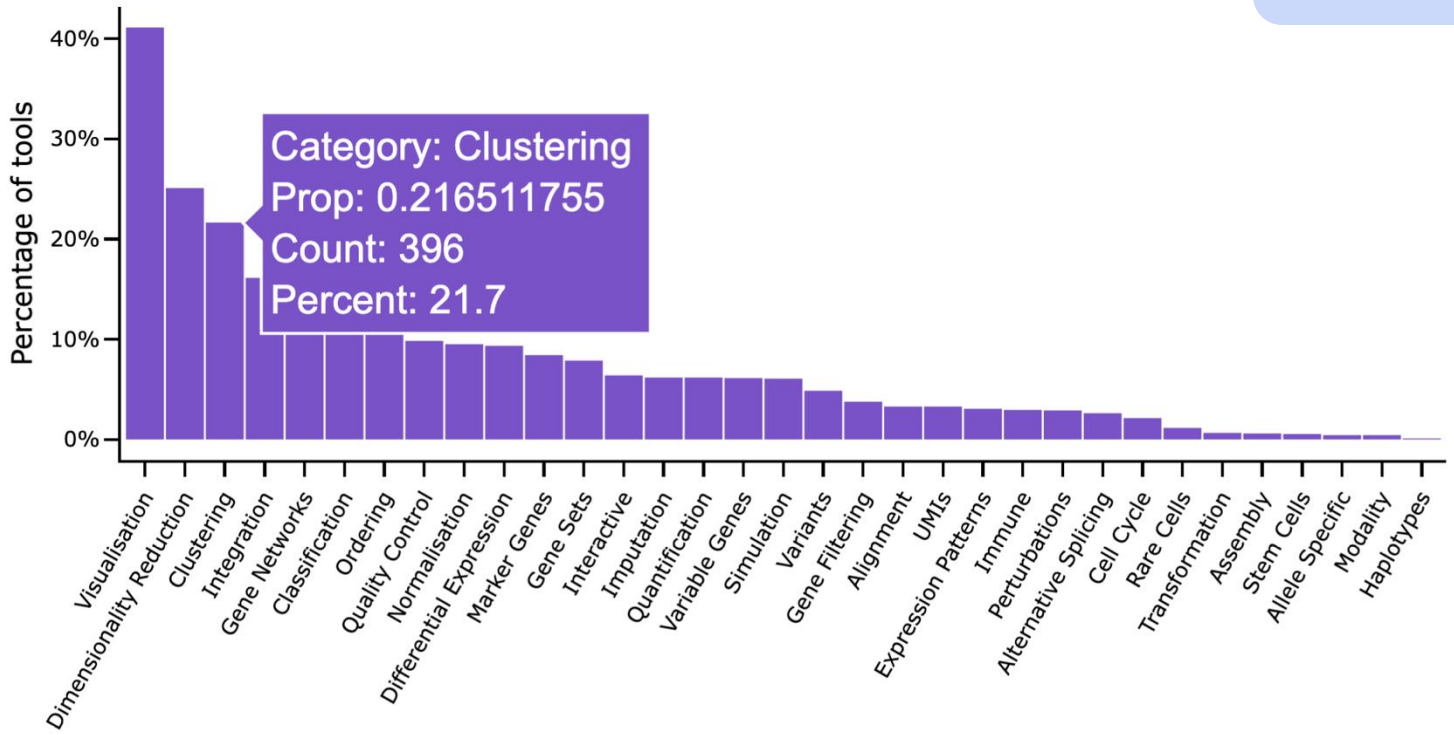
Why clustering?



scRNAseq clustering methods

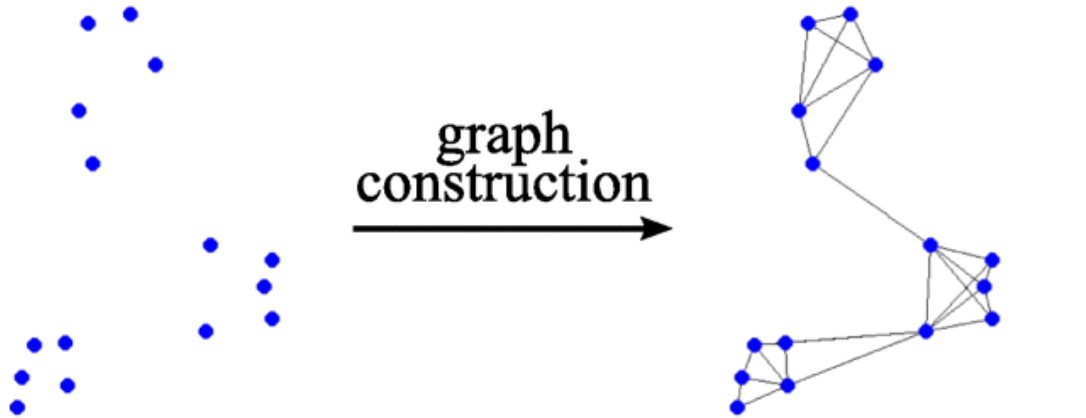
There are many possible methods: hierarchical clustering, k-means clustering, **graph-based clustering**…

Seurat uses graph-based clustering approach (KNN, SNN)



https://www.scrna-tools.org/

Graph-based clustering



Single cells after dimensionality reduction (like PCA)

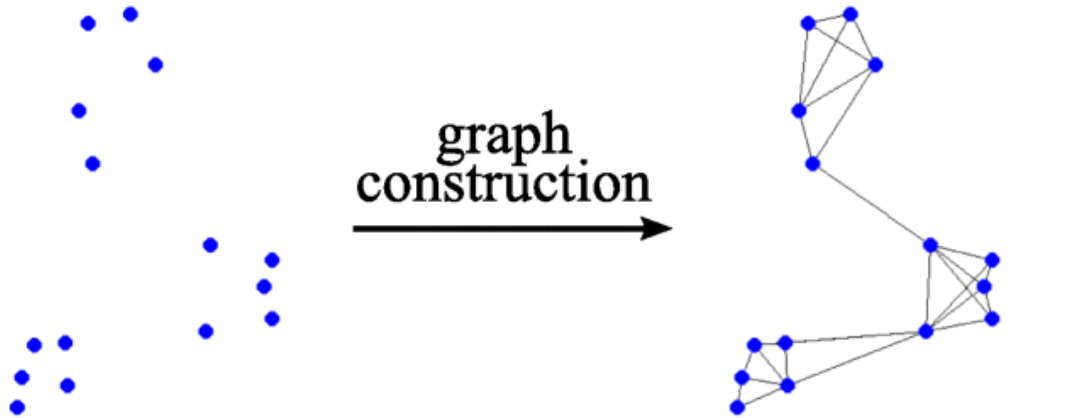
Nodes -> Cells

Edges → Similarity

Graph-based clustering

K-nearest neighbour (KNN) graph based on the euclidean distance in PCA space.

Two vertices *p* and *q* are connected by an edge, if the distance between *p* and *q* is among the *k*-th smallest distances from *p* to other nodes.



Single cells after dimensionality reduction (like PCA)

Nodes -> Cells

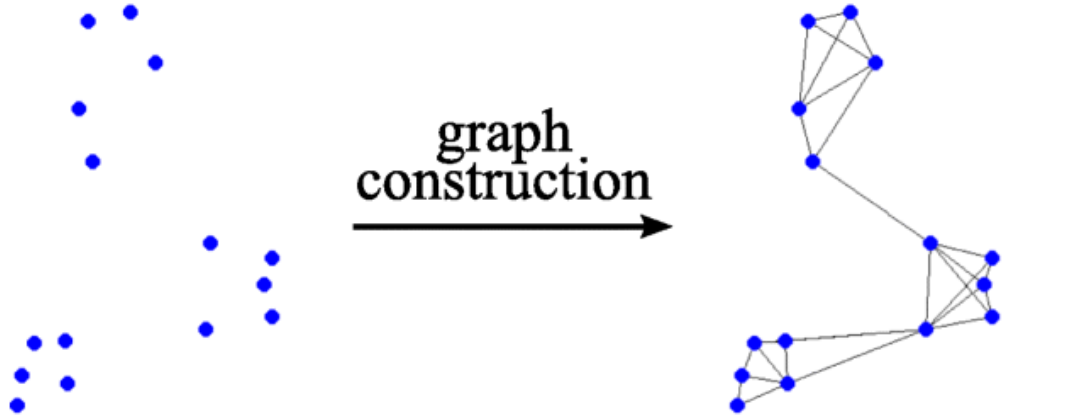
Edges → Similarity

```
Seurat::FindNeighbors(object, dims=1:10 ...)
```

Graph-based clustering

Shared-nearest neighbour (SNN) graph

For each pair of cells (nodes), the number of shared neighbours is counted (according to the KNN graph). An edge is created between two cells if they share a **sufficient number** of nearest neighbours (above a certain threshold).



Single cells after dimensionality reduction (like PCA)

Nodes -> Cells

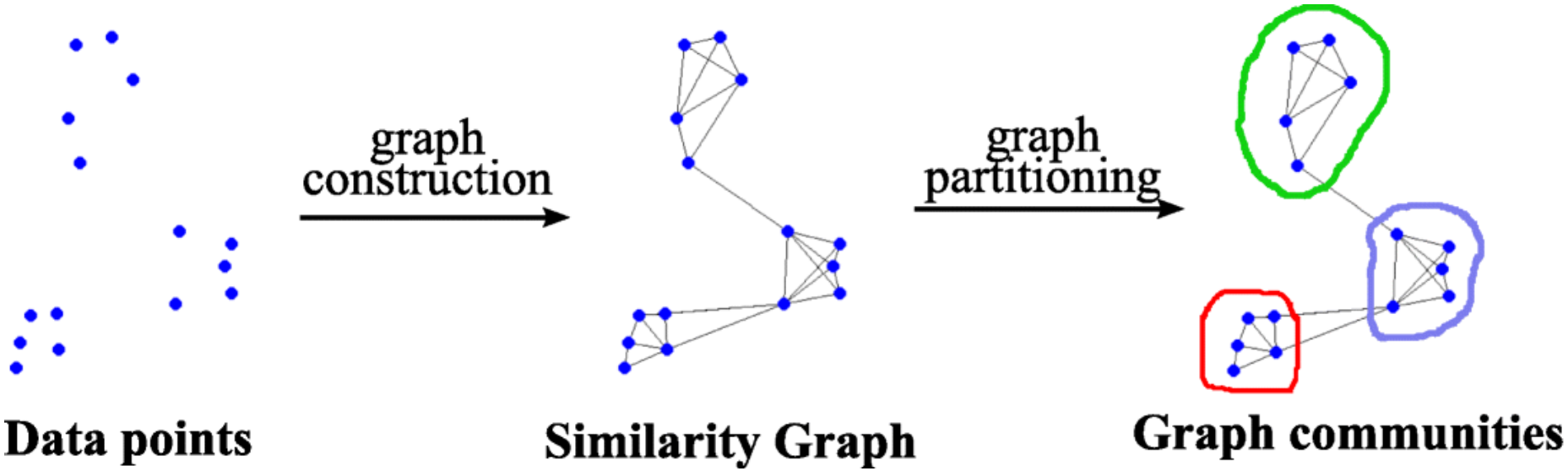
Edges → Similarity

```
Seurat::FindNeighbors(object, dims=1:10 ...)
```

Graph-based clustering

The graph is partitioned into communities (clusters) based on modularity optimization

A community hass more edges between the members of the community than edges linking nodes of the group with the rest of graph.

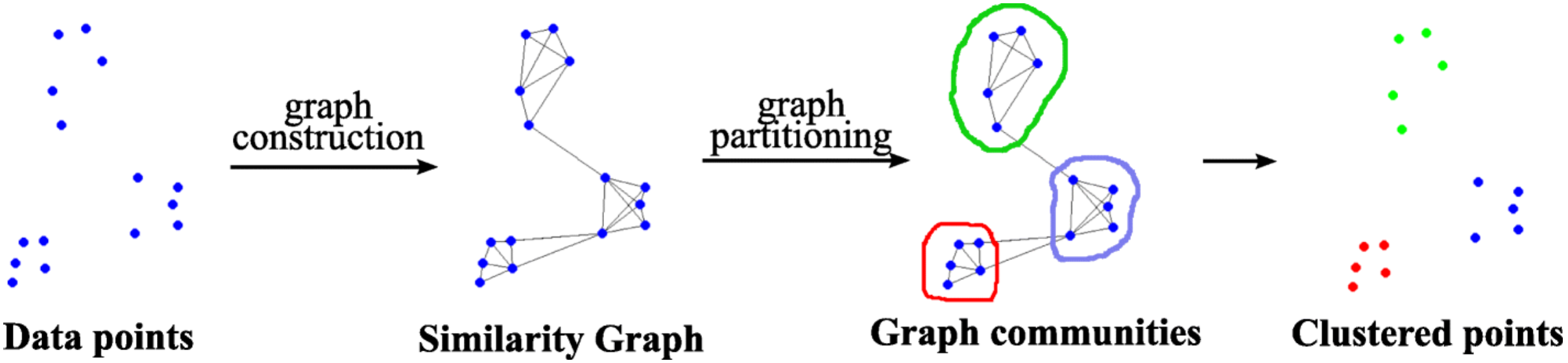


Seurat uses Louvain algorithm for community detection

```
Seurat::FindClusters(object, resolution = 0.1, ...)
```

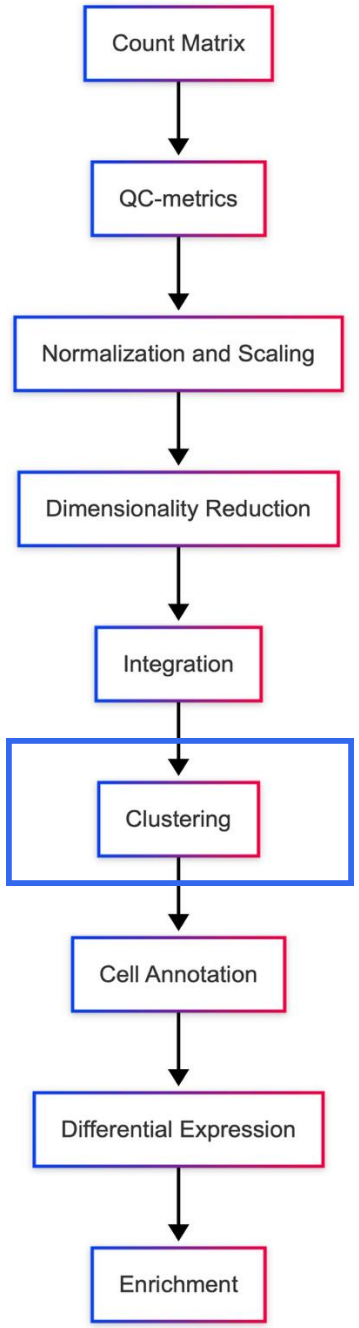
Graph-based clustering

Cells within the same community are assigned to the same cluster




```
Seurat::FindClusters(object, resolution = 0.1, ...)
```

Clustering workflow

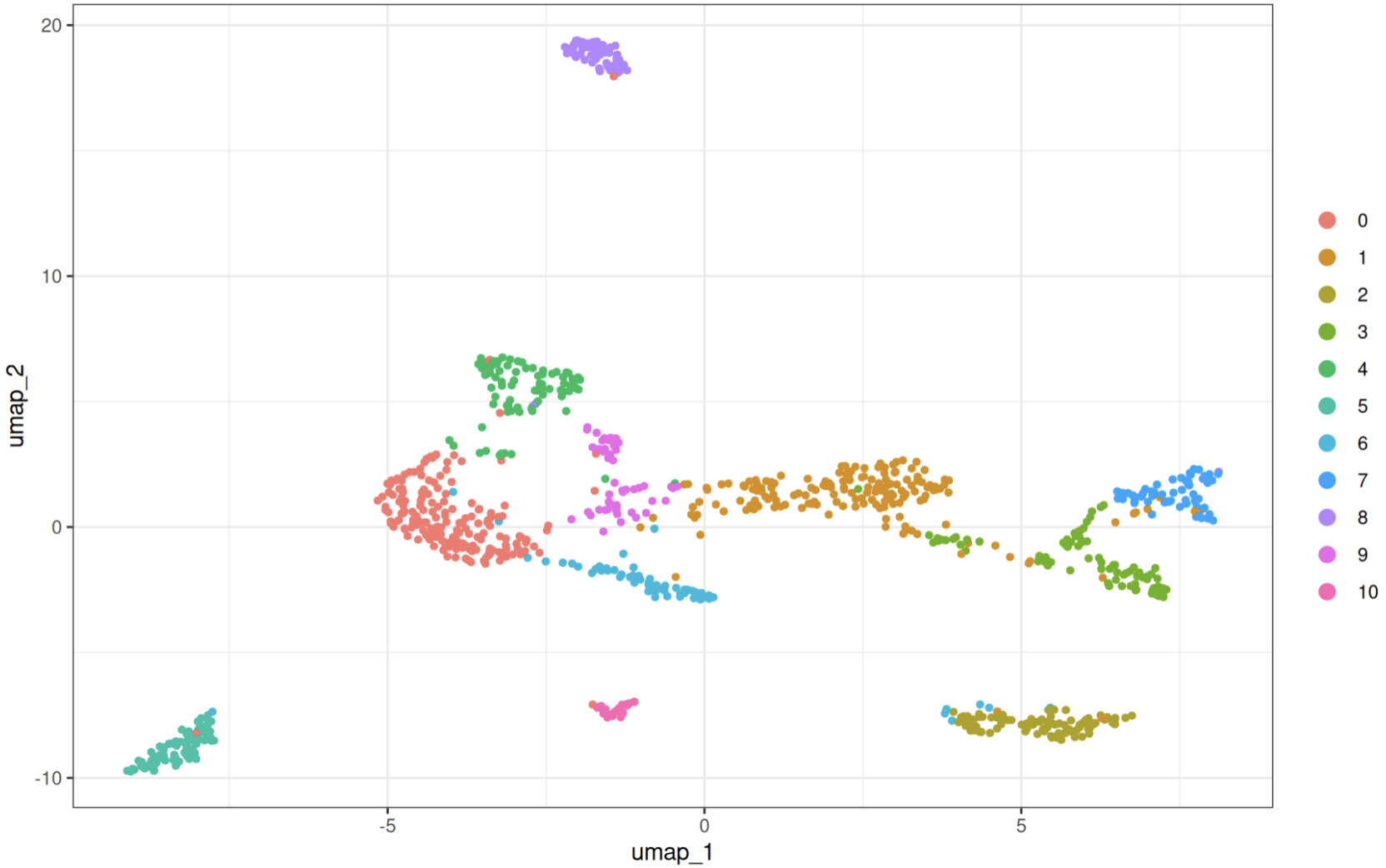


Clustering is not an isolated and single step

Clustering workflow

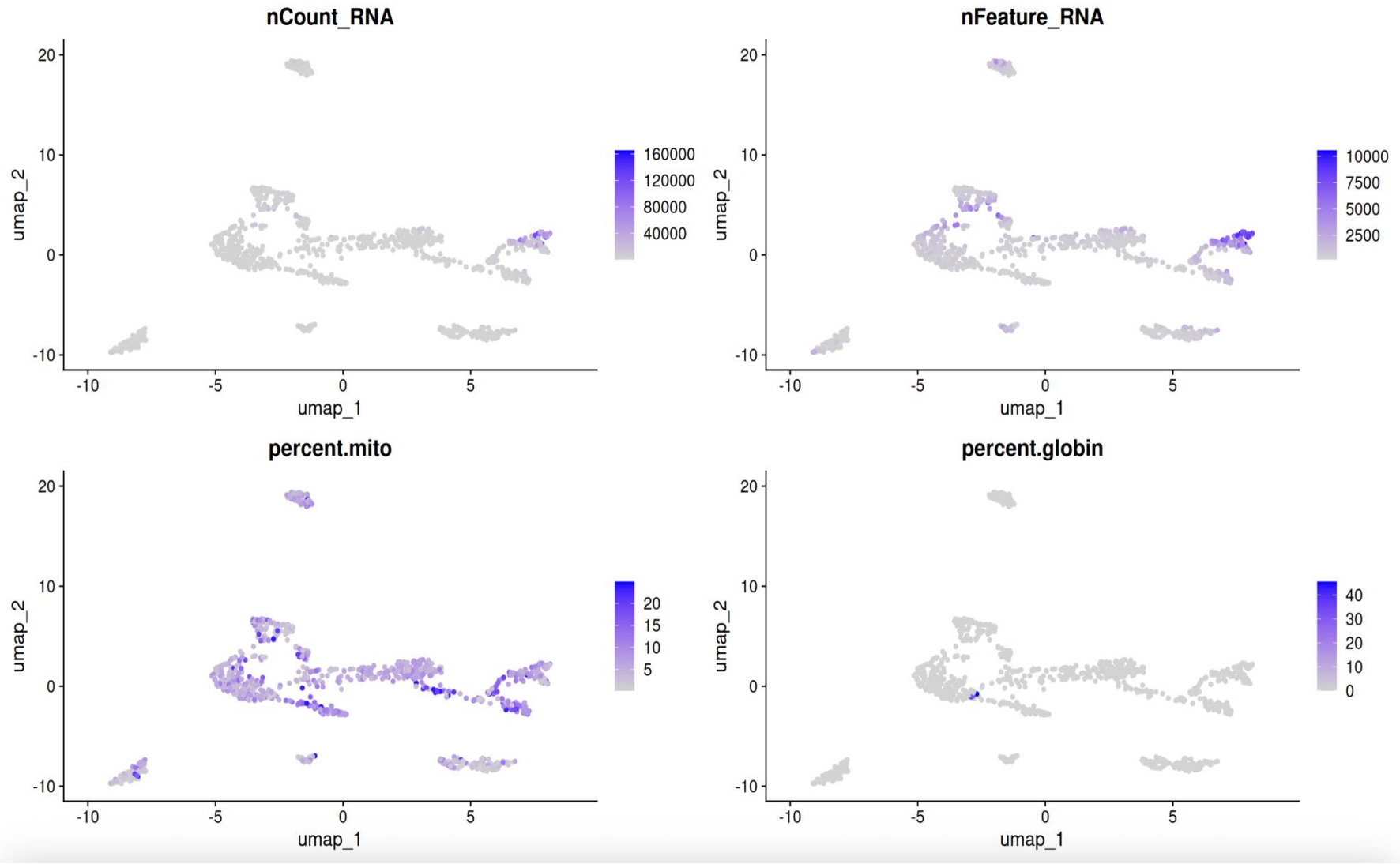
The goal of clustering is to group cells with similar gene expression profiles

How well does the clustering approximate the cell types/states of interest?



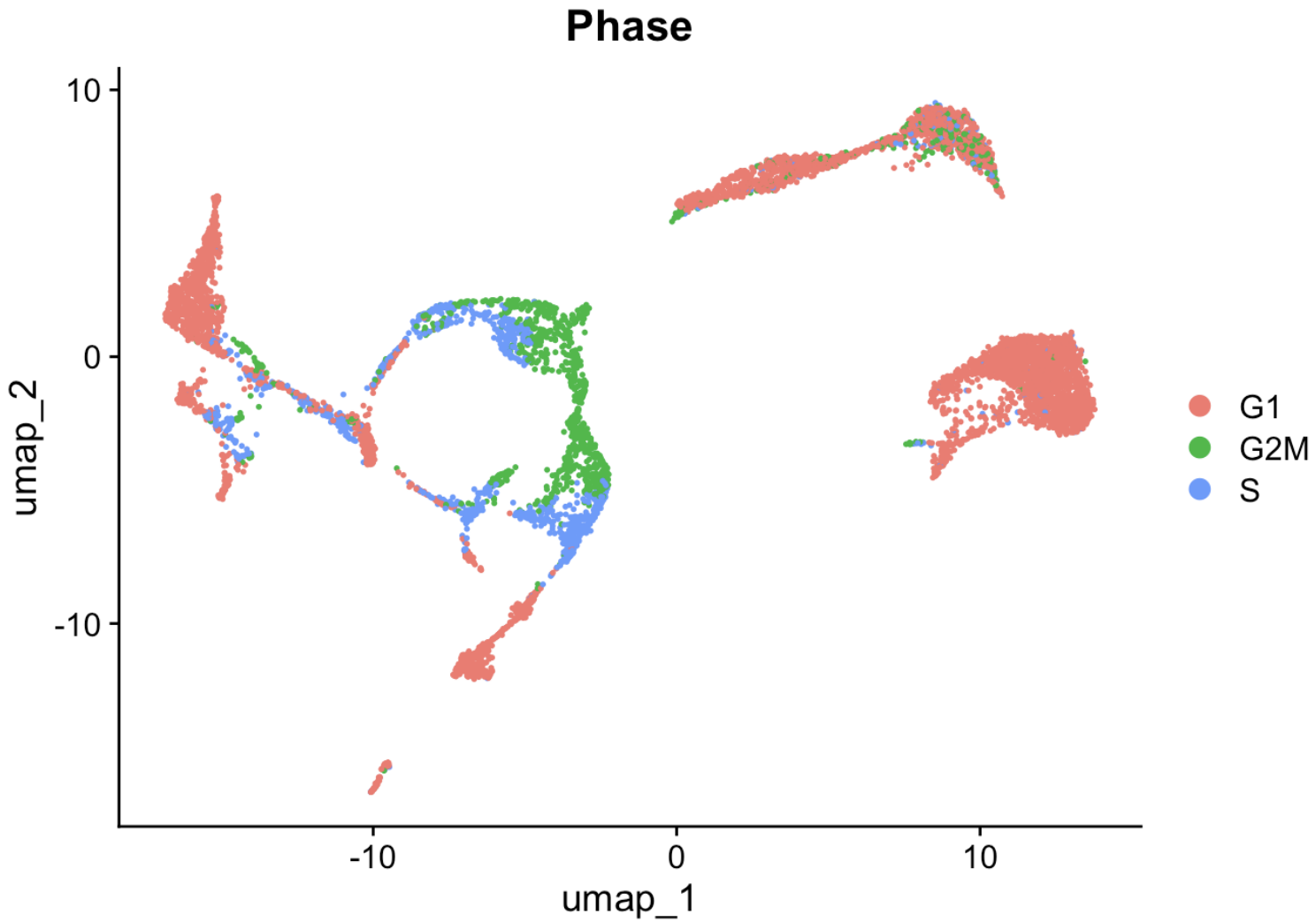
Clustering workflow

Control for quality measurements



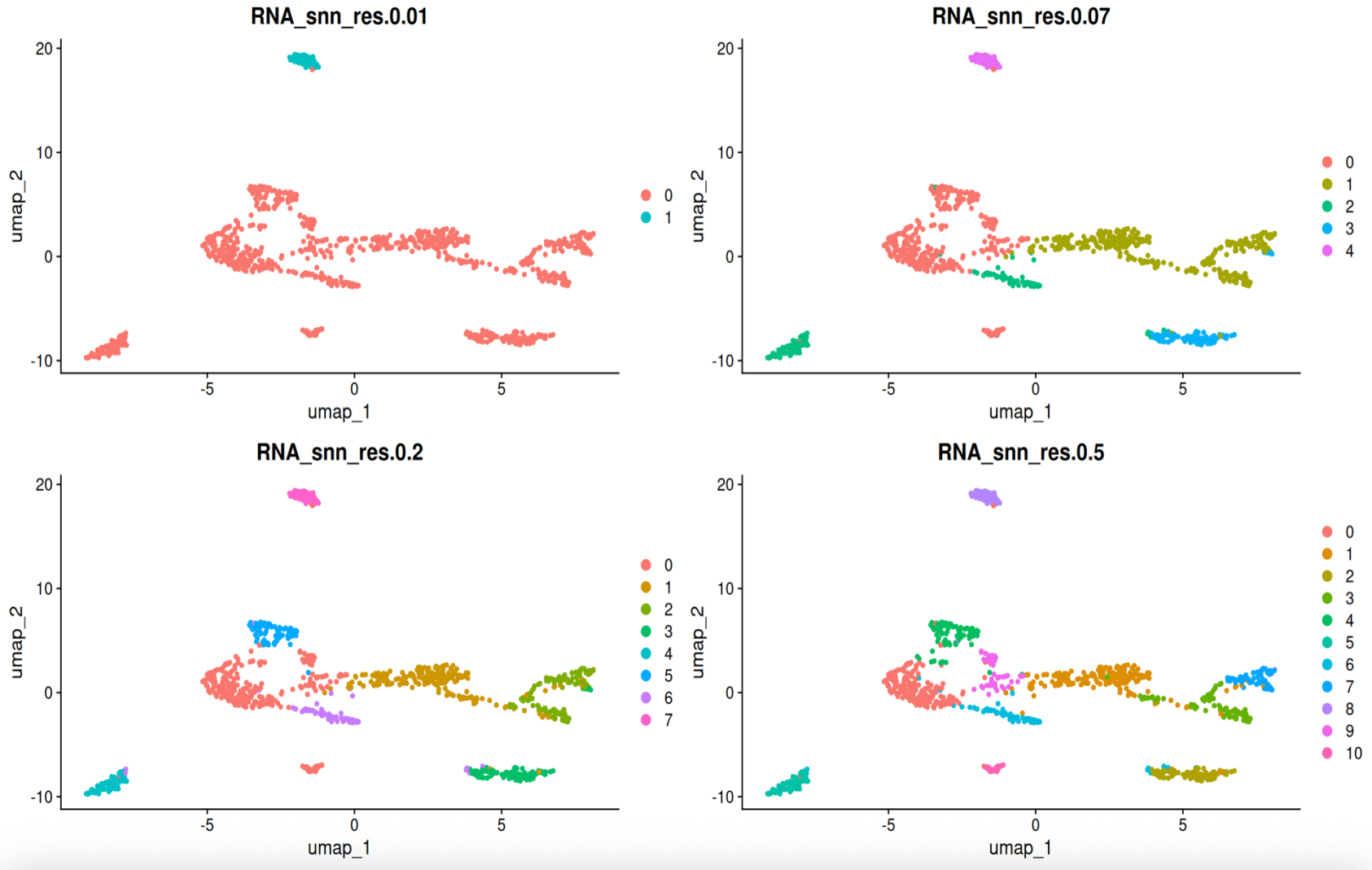
Clustering workflow

Control for cell cycle state



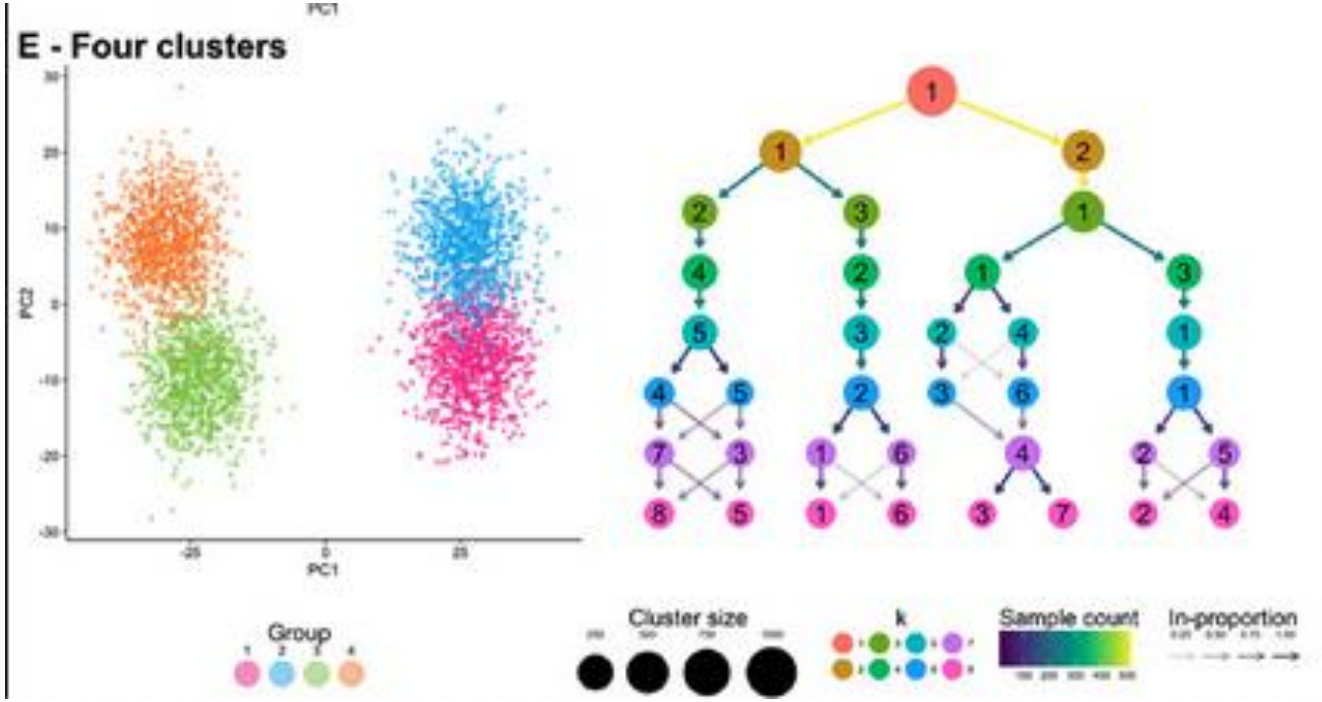
Clustering workflow

Try different resolutions



Clustering workflow

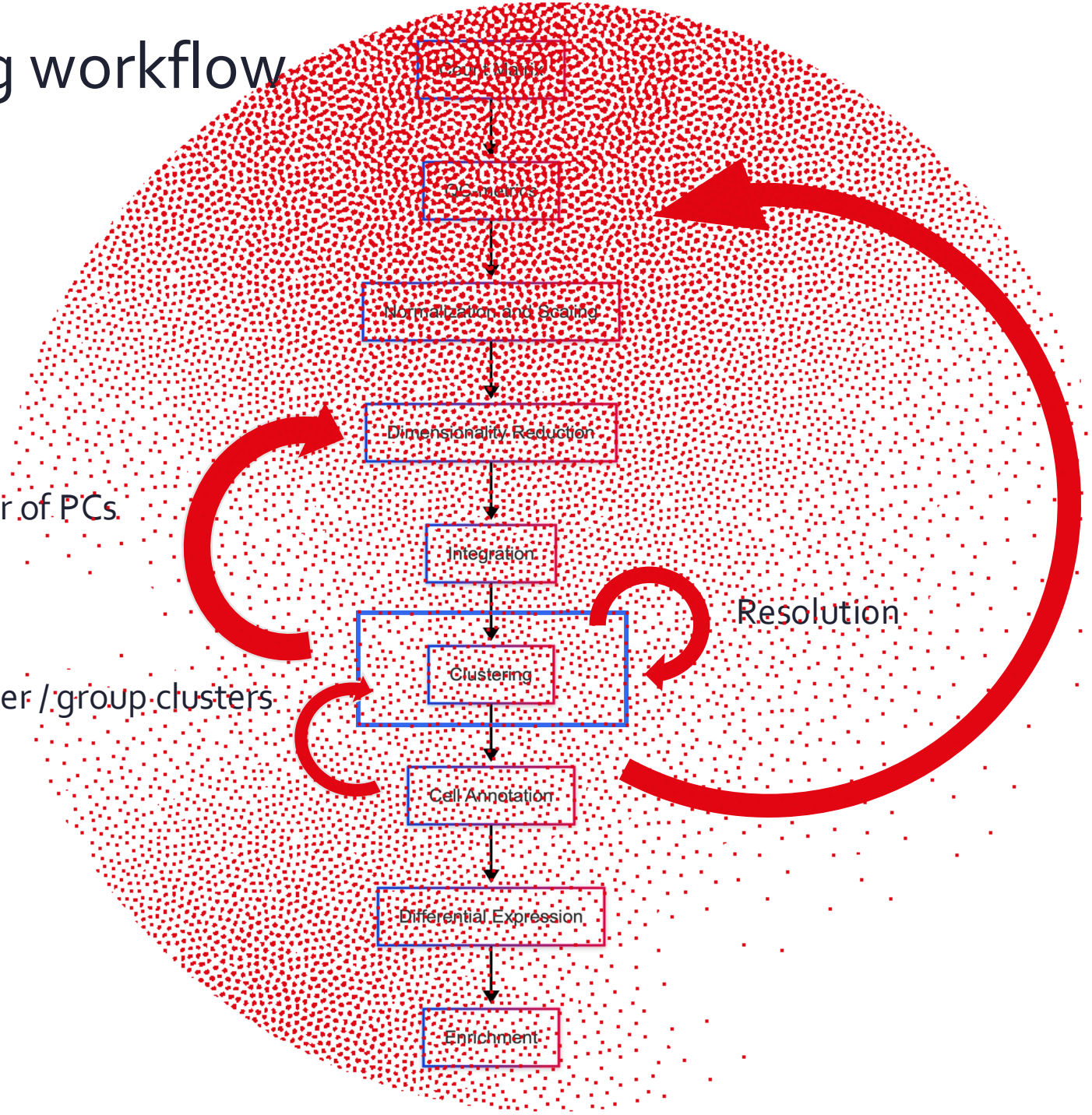
Clustree: "shows the relationships between clusters at multiple resolutions, allowing researchers to see how samples move as the number of clusters increases"



Luke Zappia, Alicia Oshlack; 2018

QUESTION 11

Clustering workflow



Different number of PCs

Subcluster / group clusters

Resolution

Filter out cells
Remove doublets
Regress out unwanted variability (cell cycle)

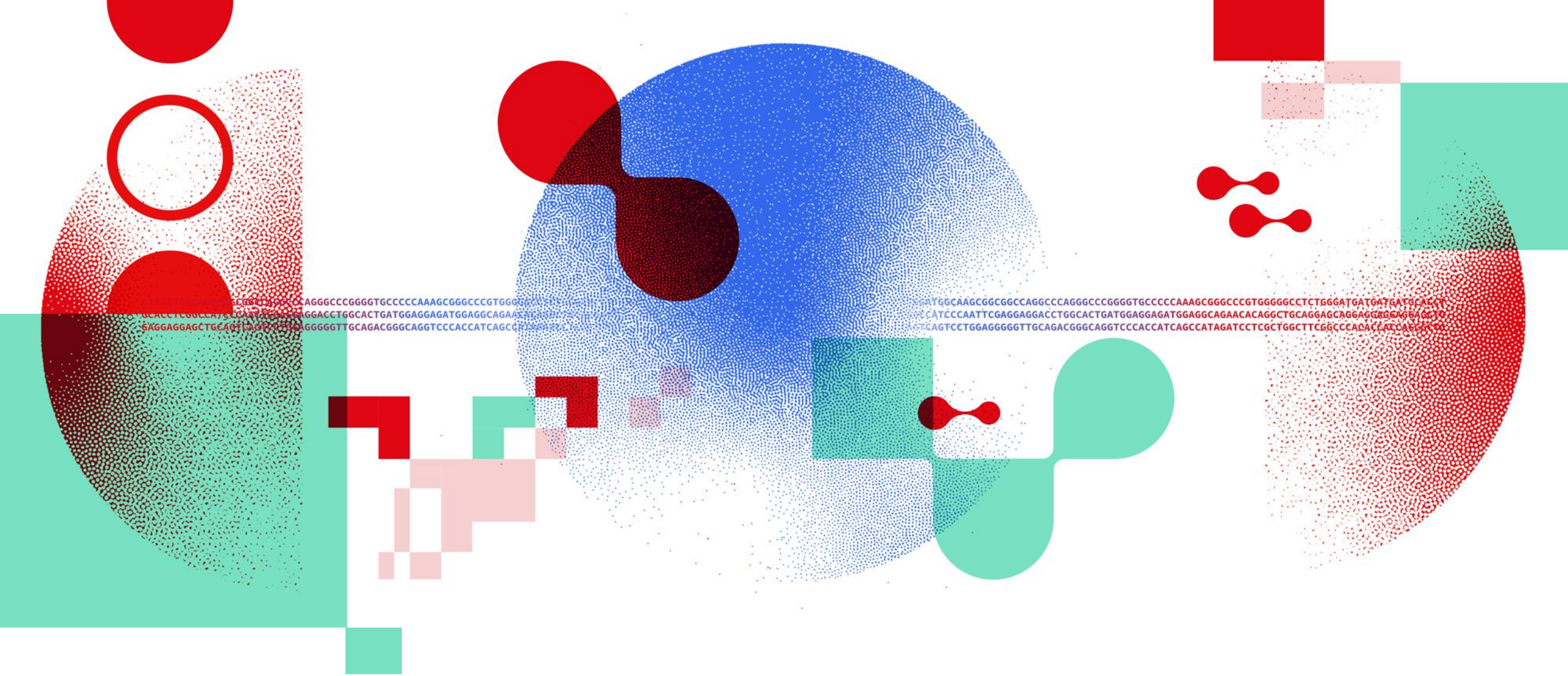
Summary

In scRNAseq analysis, clustering is an unsupervised learning procedure that is used to group cells together based on their expression profile similarities

There are many clustering methods, a popular one in scRNAseq analysis is graph based

There is not a correct number of clusters, it will depend on the context and biological question

- Assess whether cells cluster based on unwanted sources of variation (quality, cell cycle phase…)
- Having a good idea about expected heterogeneity/cell types present in the dataset may help
- You can test the effect of changing parameters (number of PCs, K, resolution..)



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss