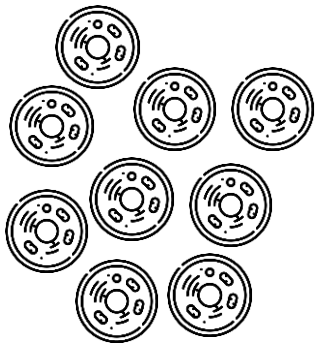


Clustering: Cell Identity

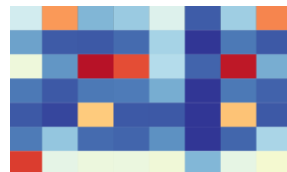
Mystery
cells



Measure



Cell #1
Cell #2
⋮
Cell #N



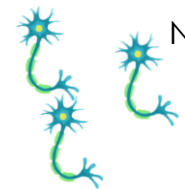
Group



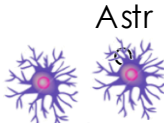
Identify



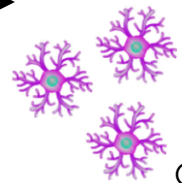
Cell
Populations



Neuro

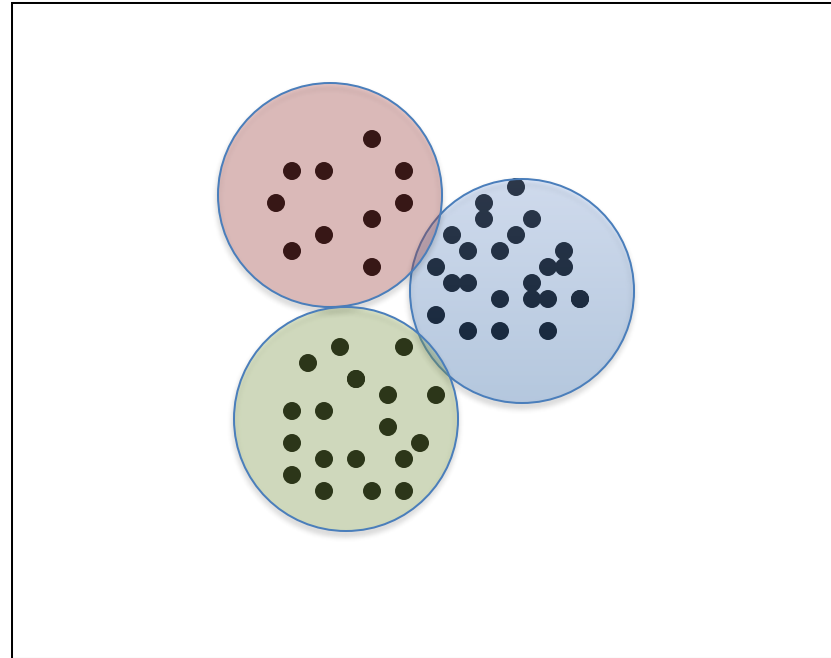


Astr



Oligo

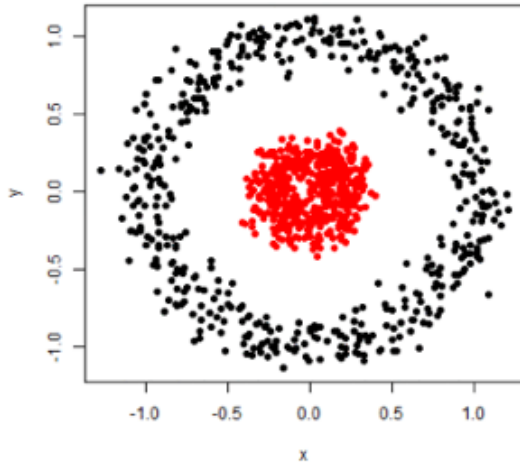
Clustering: Cell Identity



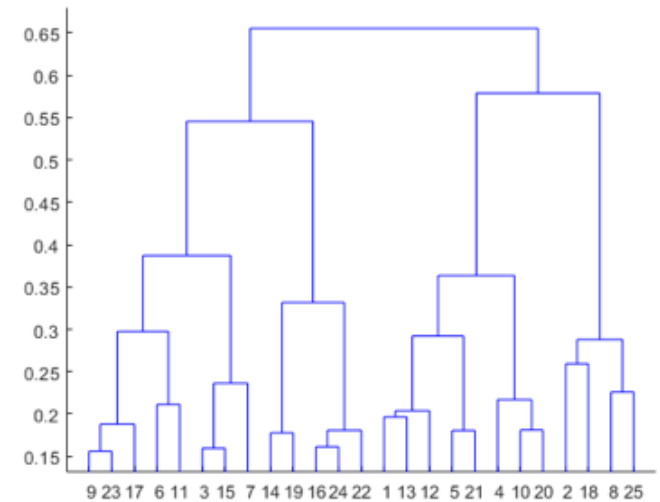
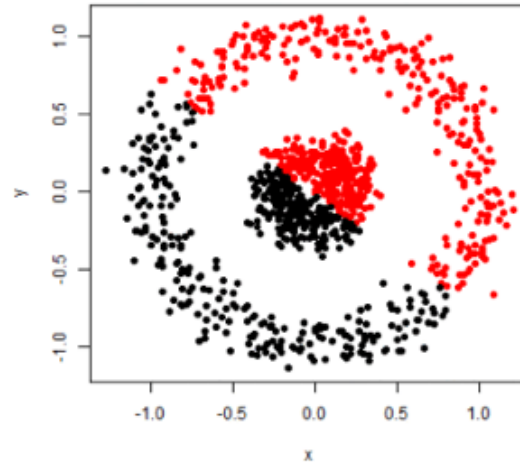
Clustering

Clustering method are divided into two categories* :

Partitioning clustering



Hierarchical clustering



*Handbook of cluster analysis, Hennig C. et al.

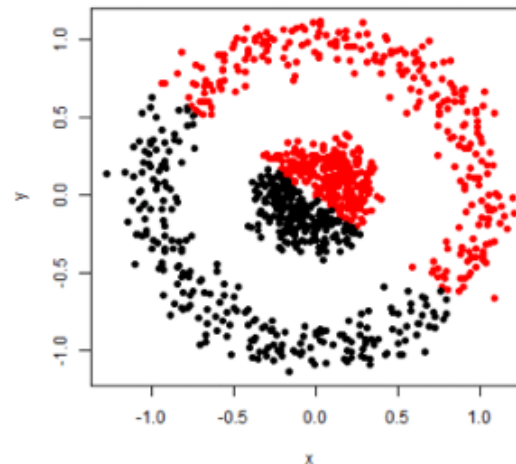
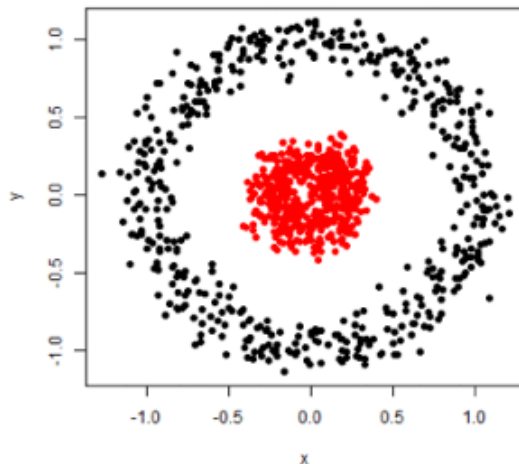
Partitioning clustering

Convex partitioning. Example: K-means

Density based approaches. Example: DBSCAN

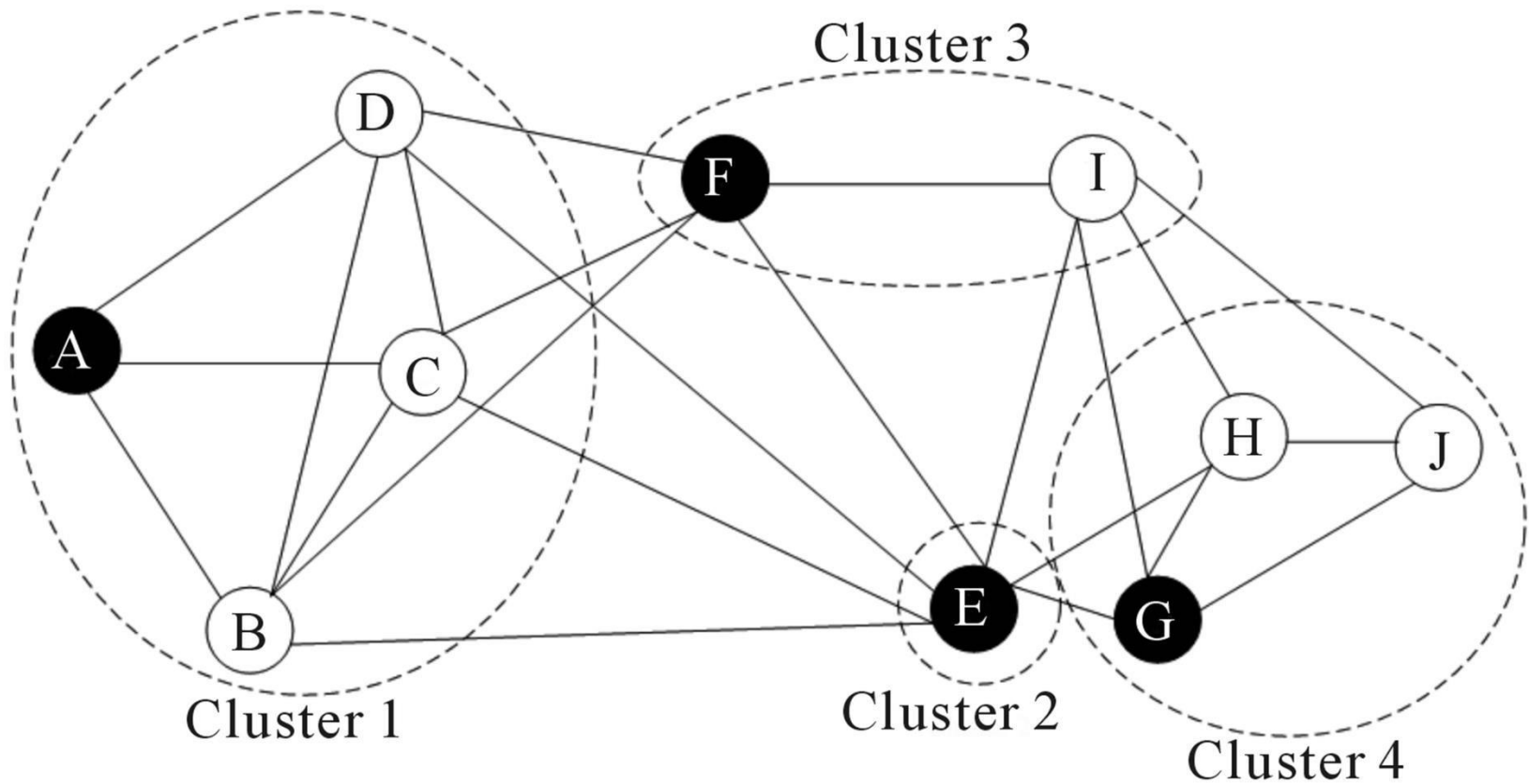
Model-based approaches. Example: Mclust

Graph based approaches : Example to follow



Graph-based

- Nodes -> cells
- Edges -> similarity ()



Graph-based: types

- k-Nearest Neighbor (**kNN**) graph

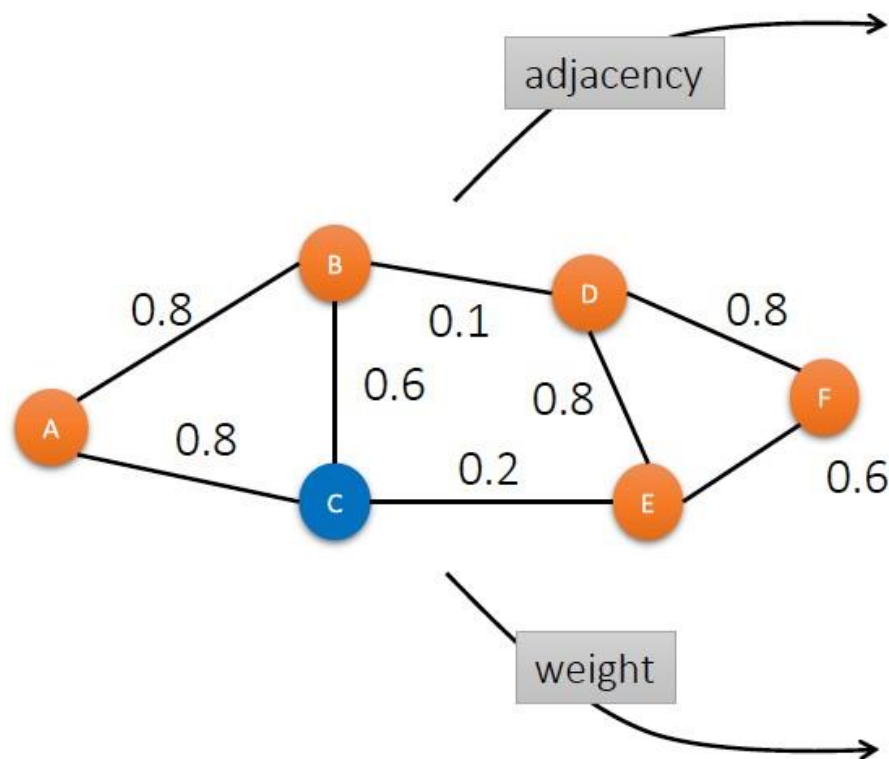
A graph in which two vertices p and q are connected by an edge, if the distance between p and q is among the k -th smallest distances from p to other objects from P .

- Shared Nearest Neighbor (**SNN**) graph

A graph in which weights define proximity, or similarity between two nodes in terms of the number of neighbors (i.e., directly connected nodes) they have in common.

```
obj <- FindNeighbors(obj)
```

Graph-based: types

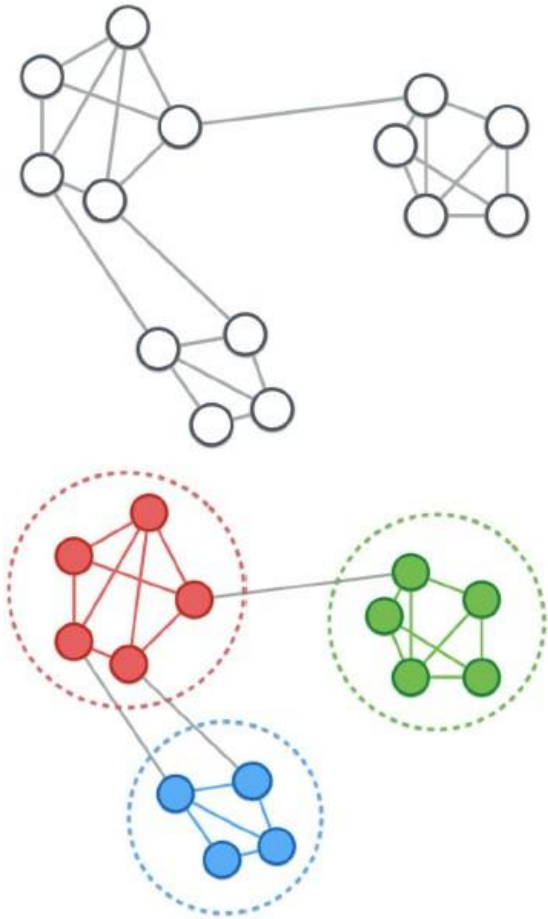


$$A = \begin{matrix} & \begin{matrix} A & B & C & D & E & F \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix} & \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \end{matrix}$$

$$W = \begin{matrix} & \begin{matrix} A & B & C & D & E & F \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix} & \begin{pmatrix} 0 & 0.8 & 0.8 & 0 & 0 & 0 \\ 0.8 & 0 & 0.6 & 0.1 & 0 & 0 \\ 0.8 & 0.6 & 0 & 0 & 0.2 & 0 \\ 0 & 0.1 & 0 & 0 & 0.8 & 0.8 \\ 0 & 0 & 0.2 & 0.8 & 0 & 0.6 \\ 0 & 0 & 0 & 0.8 & 0.6 & 0 \end{pmatrix} \end{matrix}$$

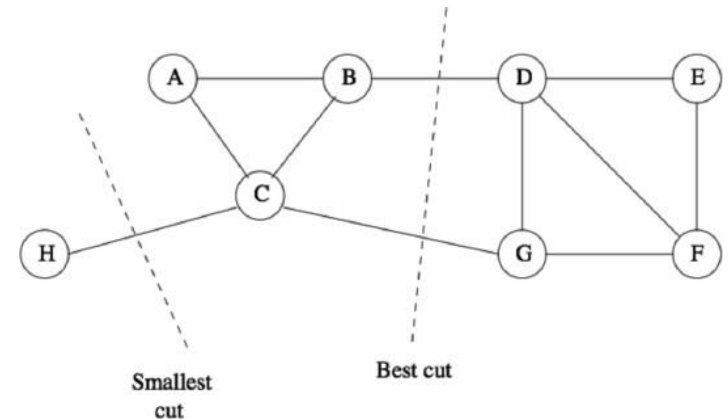
Graph-based: communities

- Communities (clusters):
 - groups of nodes **with higher probability of being connected** to each other than to members of other groups
- Community detection:
 - find a group (community) of nodes with **more edges inside** the group than edges linking nodes of the group with the rest of the graph.

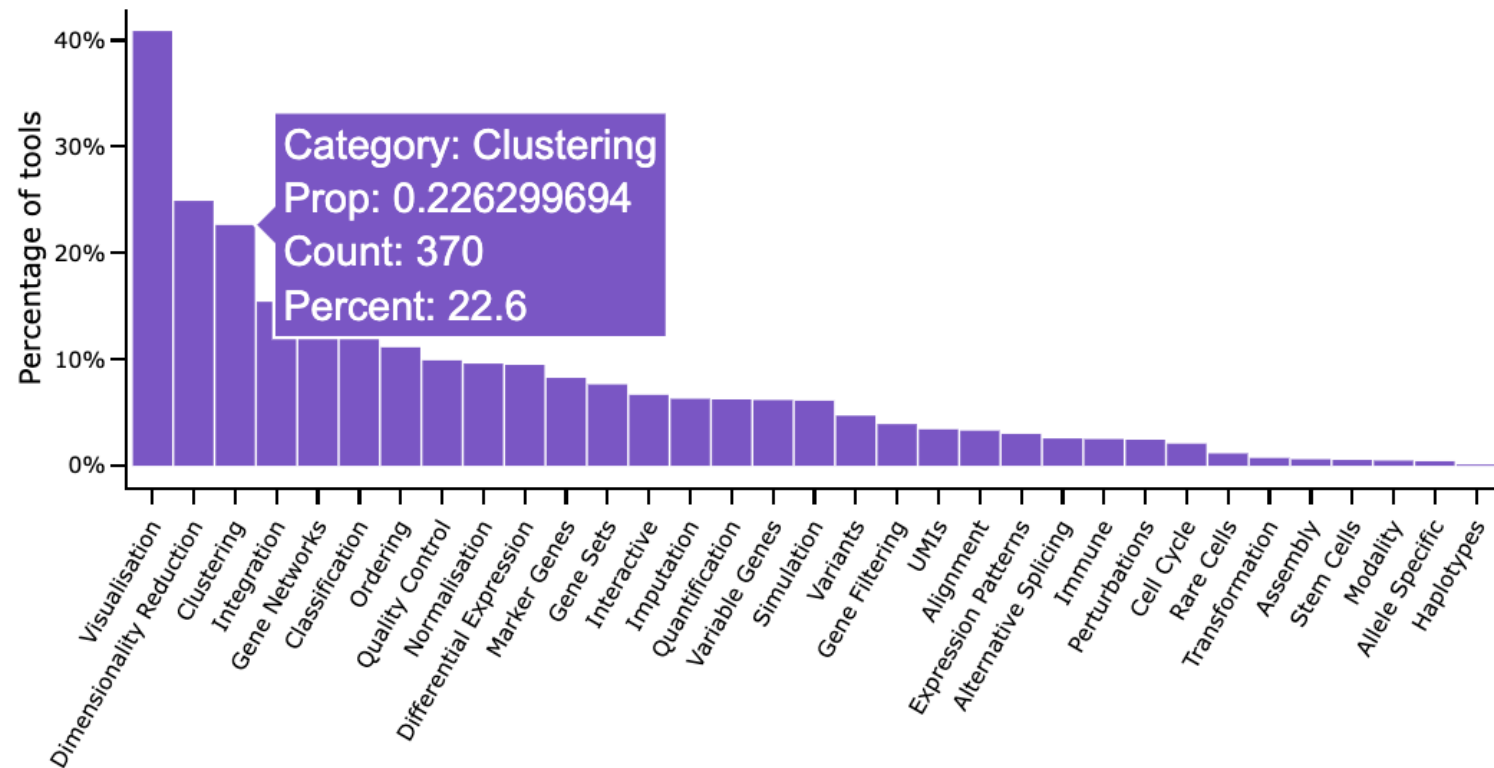


Graph-based: Cuts

- Graph cut partitions a graph into subgraphs
- Cut size is the number of cut edges
- Clustering by graph cuts: find the smallest cut that bi-partitions the graph
- The smallest cut is not always the best cut
- NP-hard
 - Heuristic methods applied e.g. Louvain



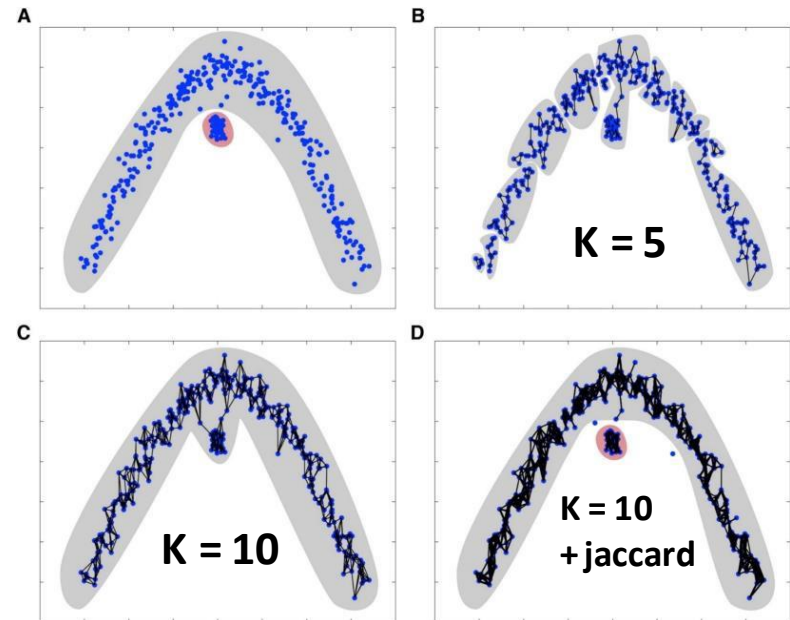
scRNA-seq clustering methods



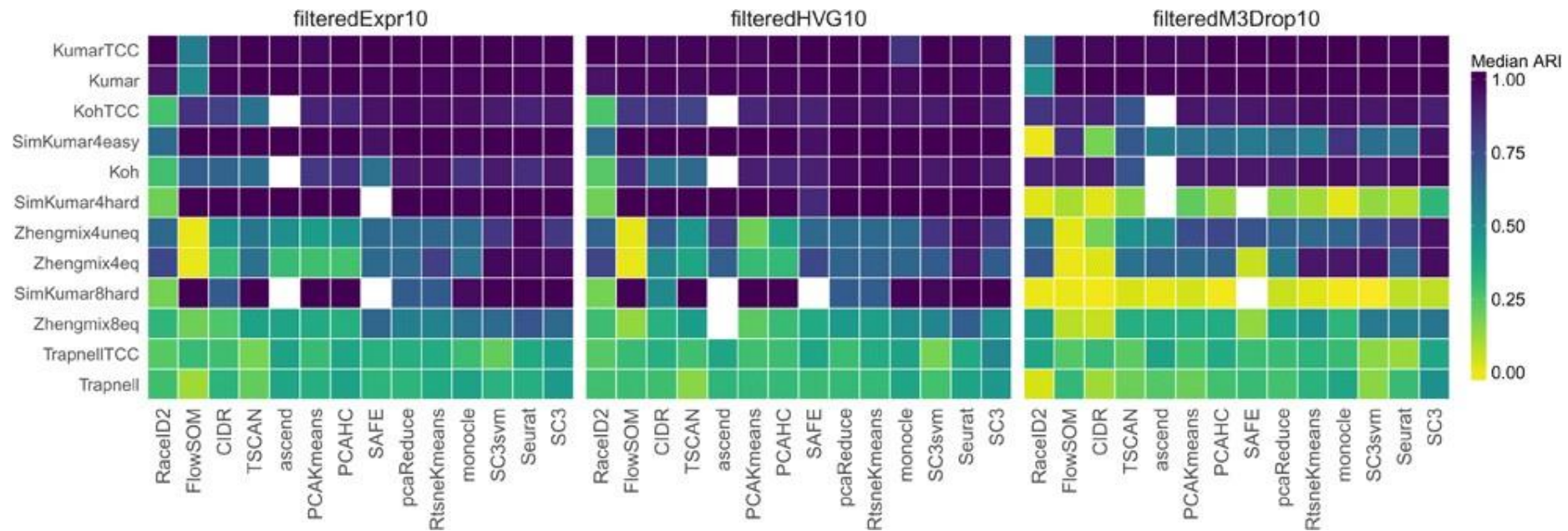
Seurat

1. Construct SNN graph based on the Euclidean distance in PCA space.
(Default, but could be also kNN)
2. Refine the edge weights between any two cells based on the shared overlap in their local neighborhoods (Jaccard index).
3. Cluster cells by optimizing for modularity (cuts) (Louvain algorithm)
(Modularity is a cost function, resolution is a parameter used to calculate the modularity)

```
obj <- FindClusters(obj)
```



Benchmarking



Clustering: Challenges

- What is a cell type?
 - What is the number of clusters k ?
 - Check QC after clustering to see if no biases are constituting your clusters
 - Clustering is subjective – No ground truth
 - How stable are the clusters
 - How dependent are the clusters on the surrounding cells
-
- Scalability: in the last few years the number of cells in scRNA-seq experiments has grown by several orders of magnitude from $\sim 10^2$ to $\sim 10^6$