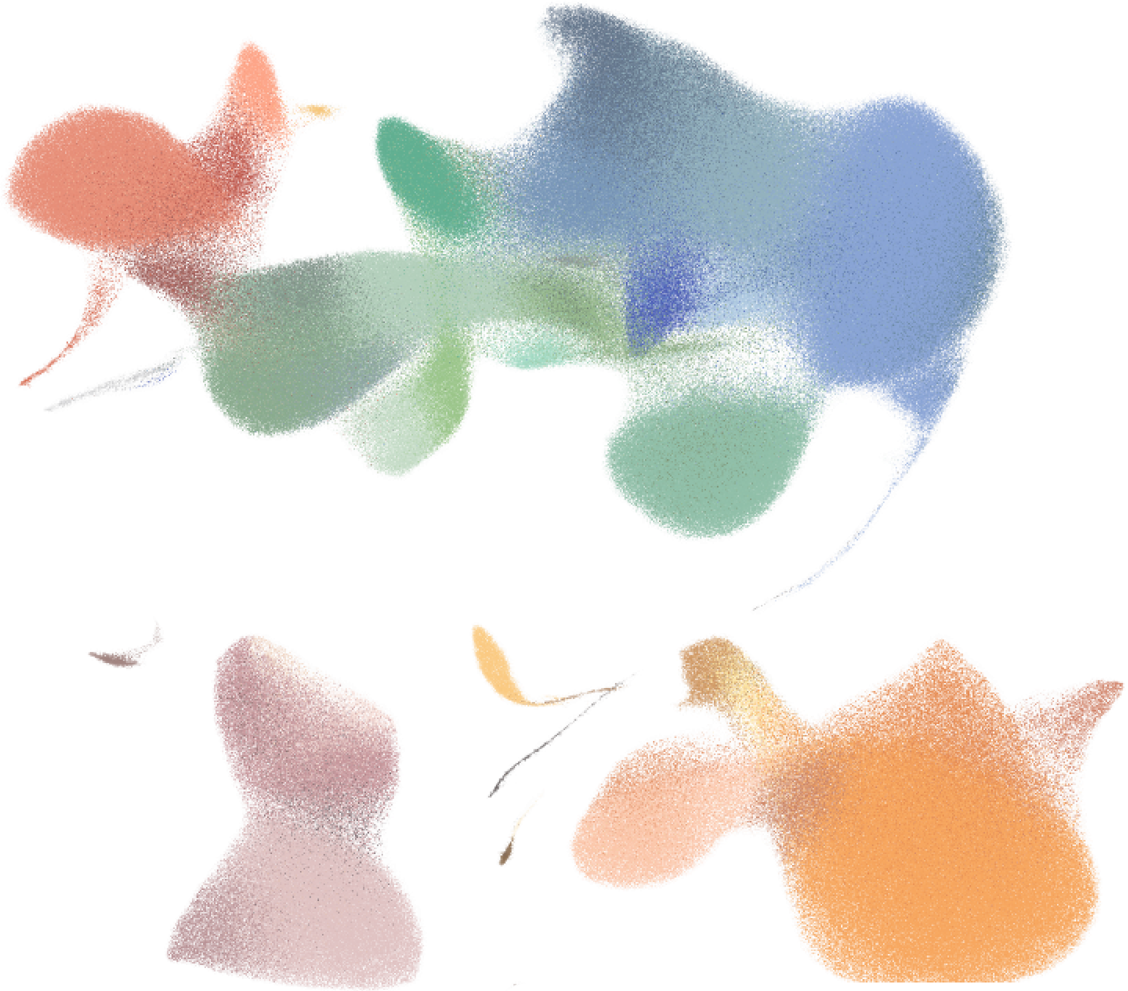
Clustering

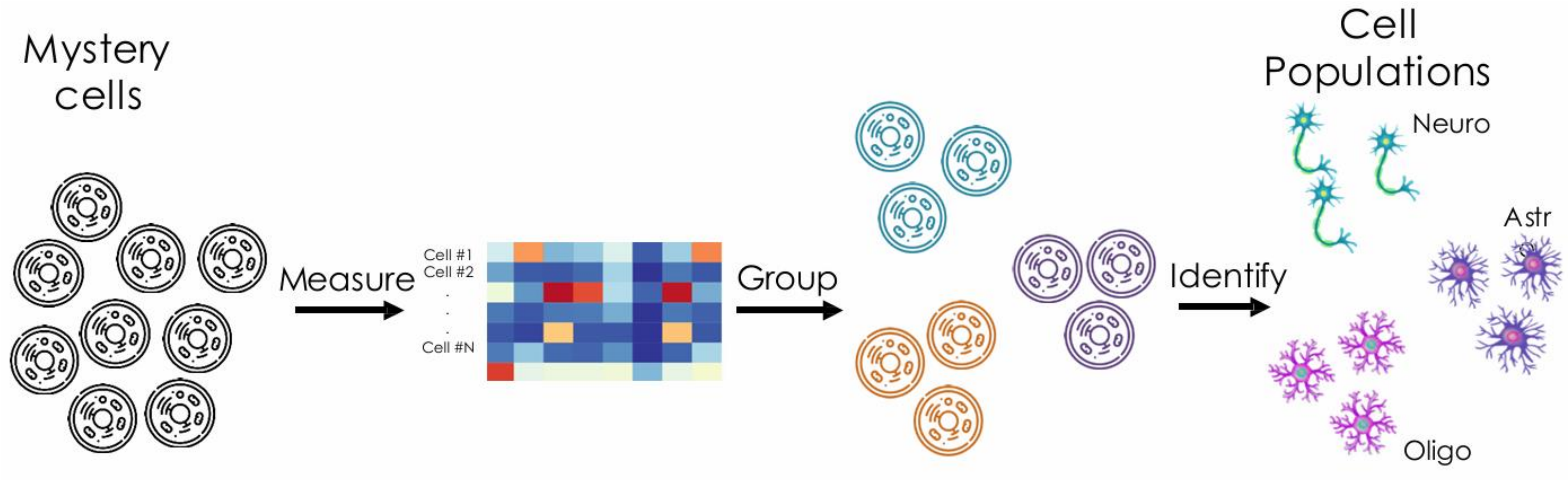
Yusuf Caglar Odabasi

December 1.-3. 2025

Course on scRNA-seq Data Analysis



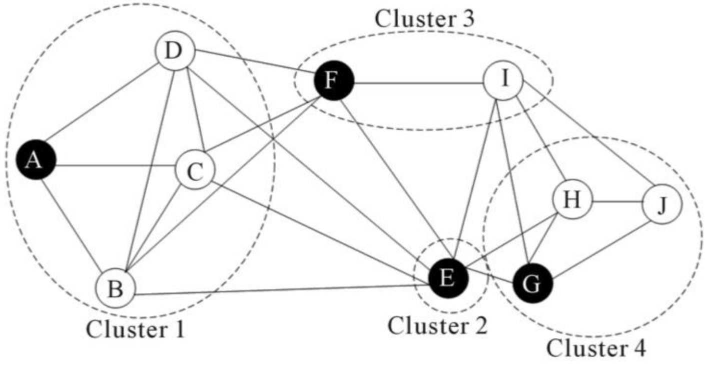
Clustering

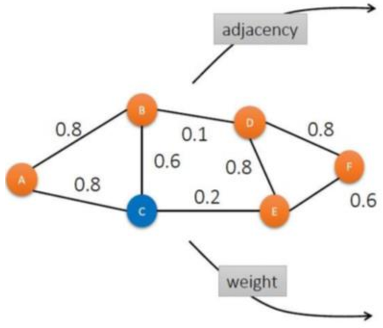


The goal of clustering is to group cells with similar gene expression profiles

Concept of clustering

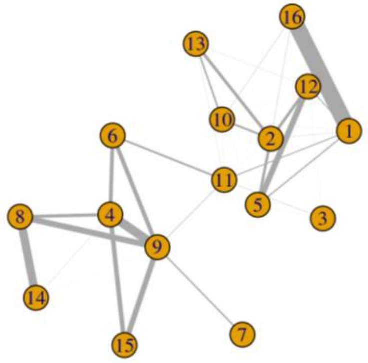
Graph-based clustering



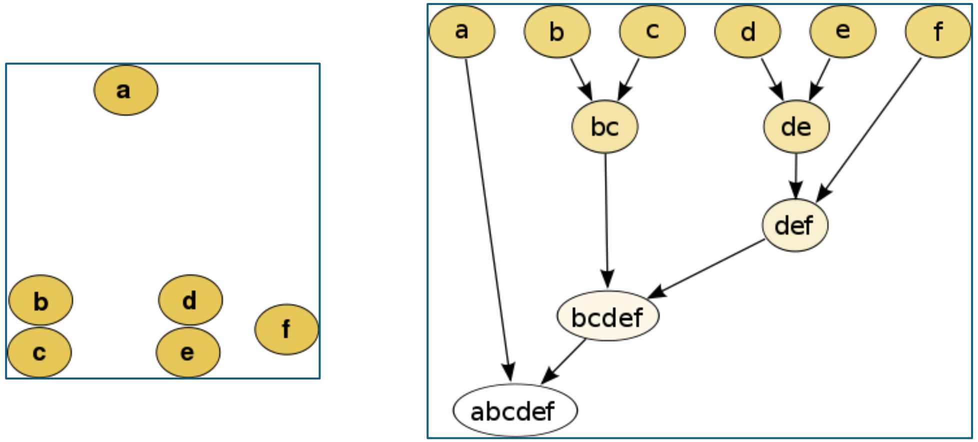


$$A = \begin{array}{c|cccccc} & A & B & C & D & E & F \\ \hline A & 0 & 1 & 1 & 0 & 0 & 0 \\ B & 1 & 0 & 1 & 1 & 0 & 0 \\ C & 1 & 1 & 0 & 0 & 1 & 0 \\ D & 0 & 1 & 0 & 0 & 1 & 1 \\ E & 0 & 0 & 1 & 1 & 0 & 1 \\ F & 0 & 0 & 0 & 1 & 1 & 0 \end{array}$$

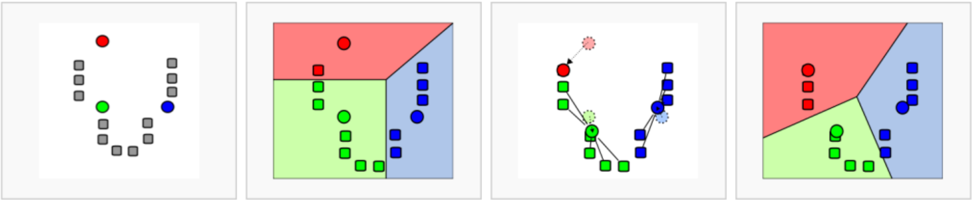
$$W = \begin{array}{c|cccccc} & A & B & C & D & E & F \\ \hline A & 0 & 0.8 & 0.8 & 0 & 0 & 0 \\ B & 0.8 & 0 & 0.6 & 0.1 & 0 & 0 \\ C & 0.8 & 0.6 & 0 & 0 & 0.2 & 0 \\ D & 0 & 0.1 & 0 & 0 & 0.8 & 0.8 \\ E & 0 & 0 & 0.2 & 0.8 & 0 & 0.6 \\ F & 0 & 0 & 0 & 0.8 & 0.6 & 0 \end{array}$$



Hierarchical clustering



K-means clustering



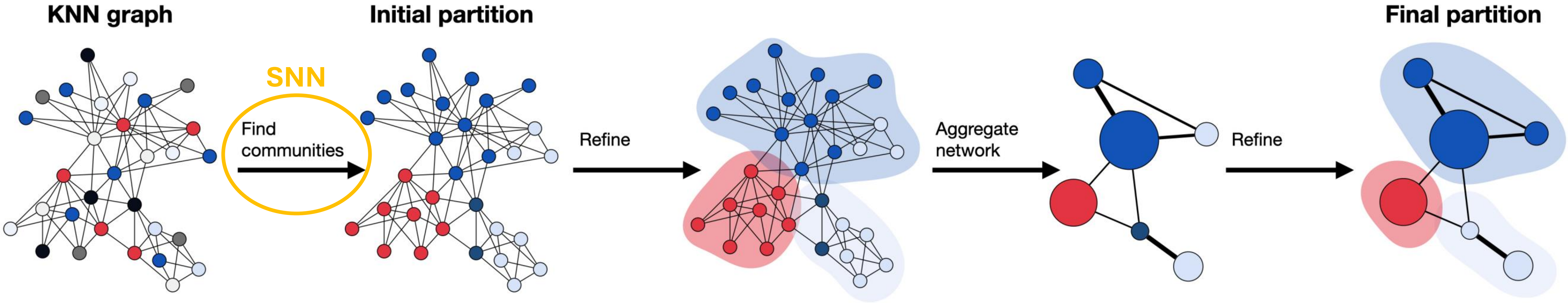
Graph-based clustering

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

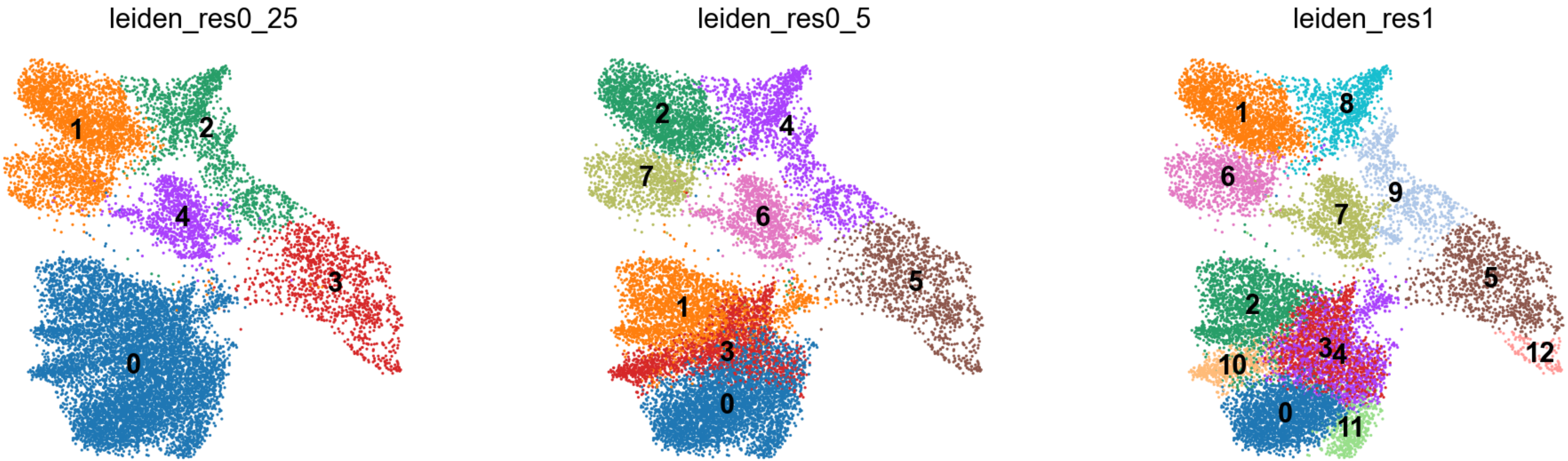
Two vertices p and q are connected by an edge, if the distance between p and q is among the k-th smallest distances from p to other nodes.

- Shared-nearest neighbour (SNN) graph

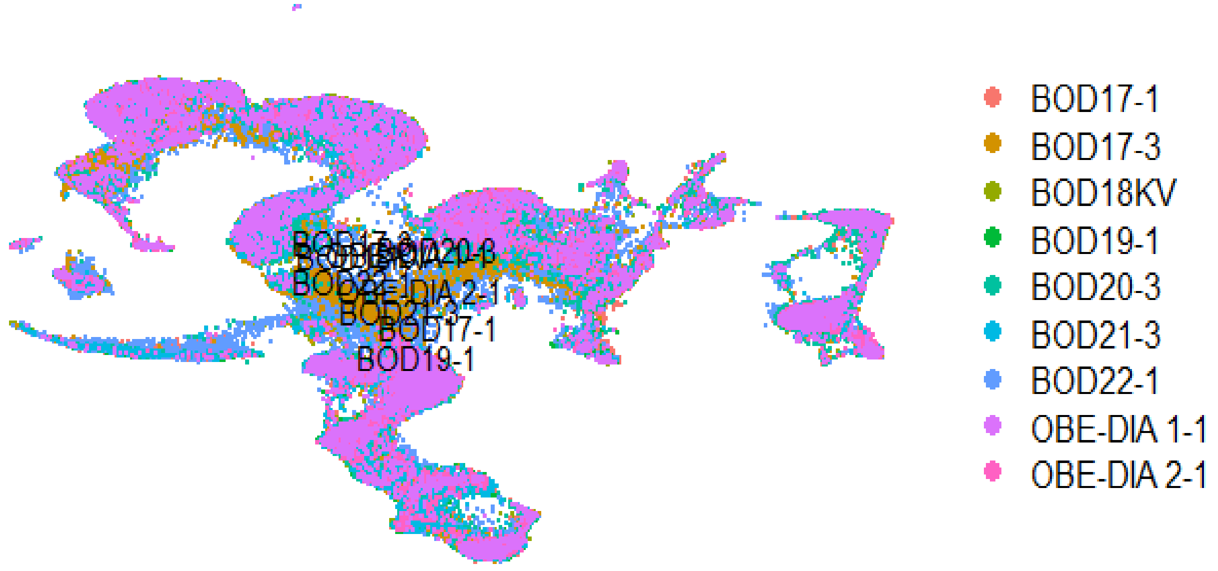
For each pair of cells (nodes), the number of shared neighbours is counted (according to the KNN graph). An edge is created between two cells if they share a sufficient number of nearest neighbours (above a certain threshold).

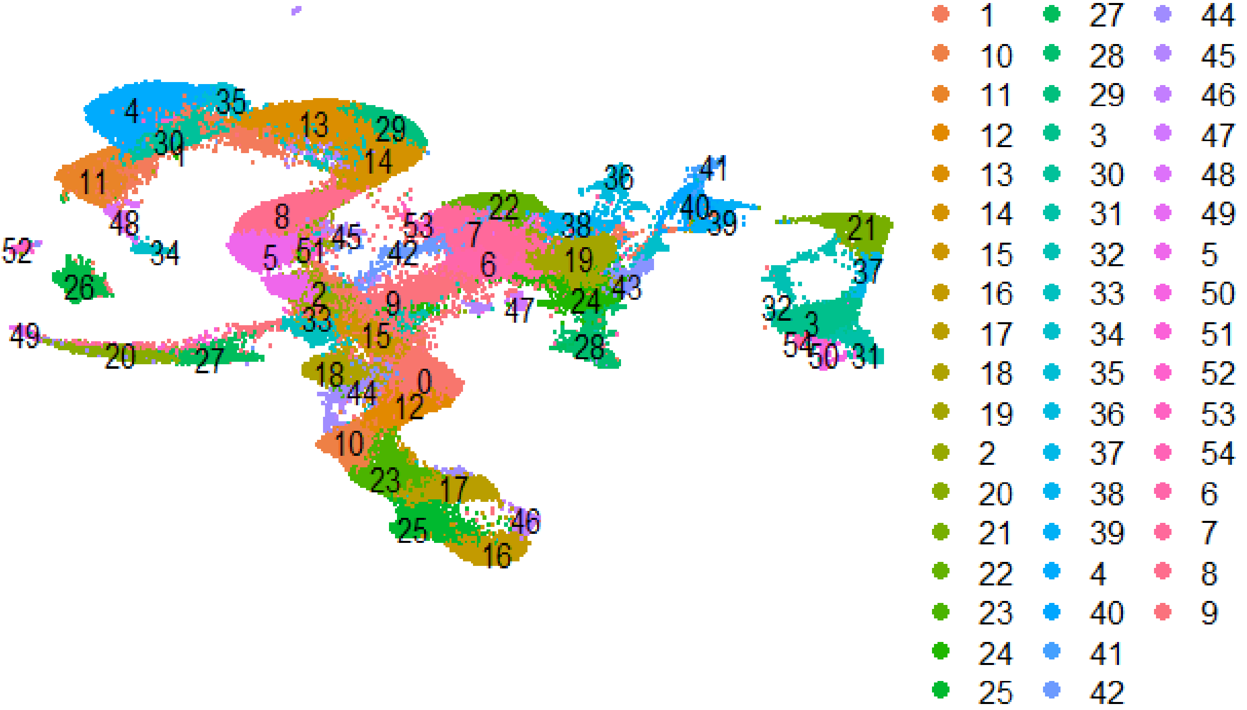


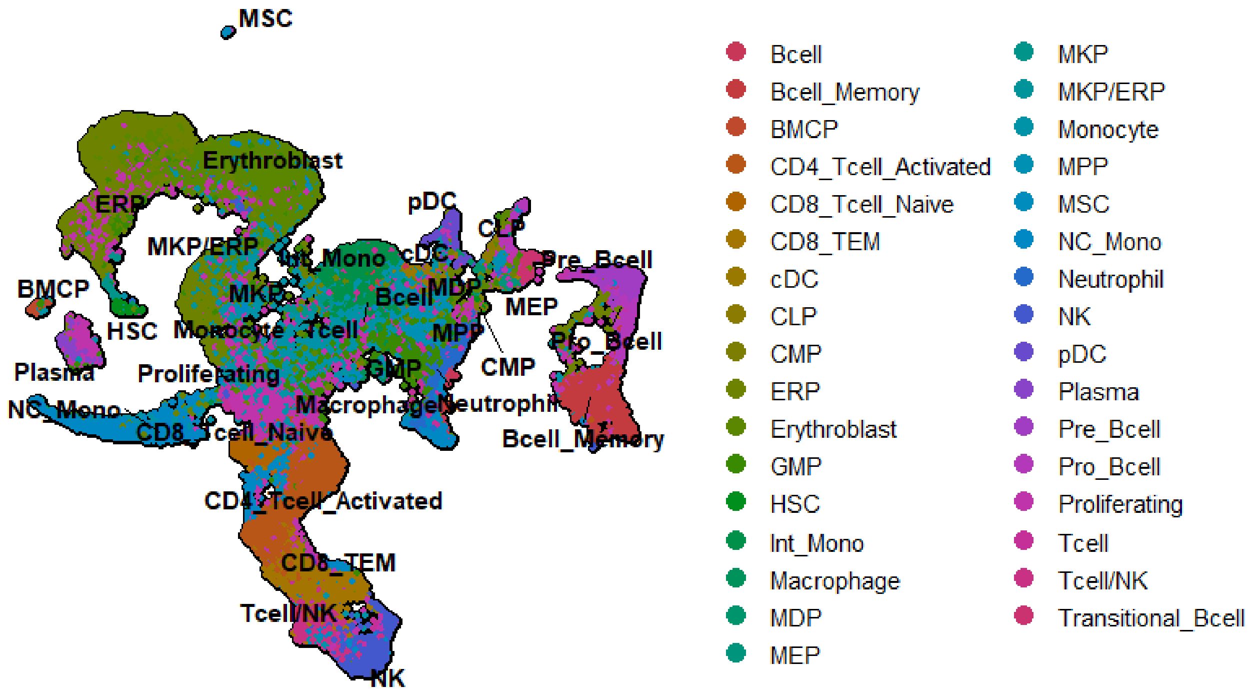
Leiden and Louvain

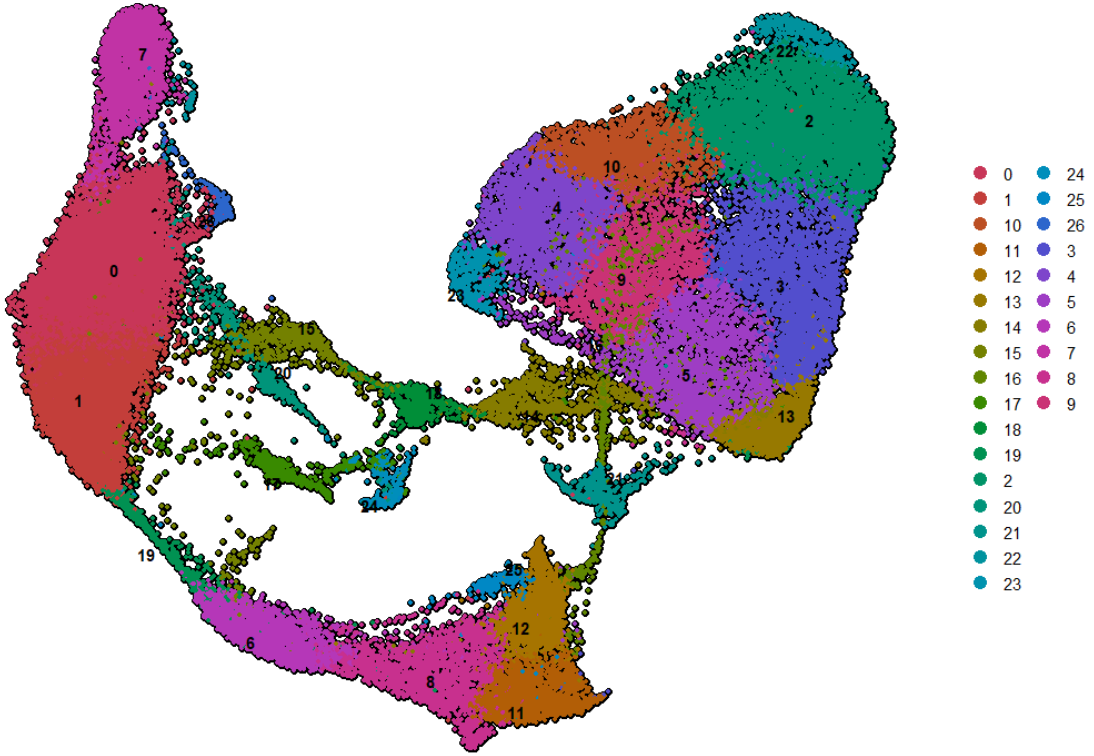


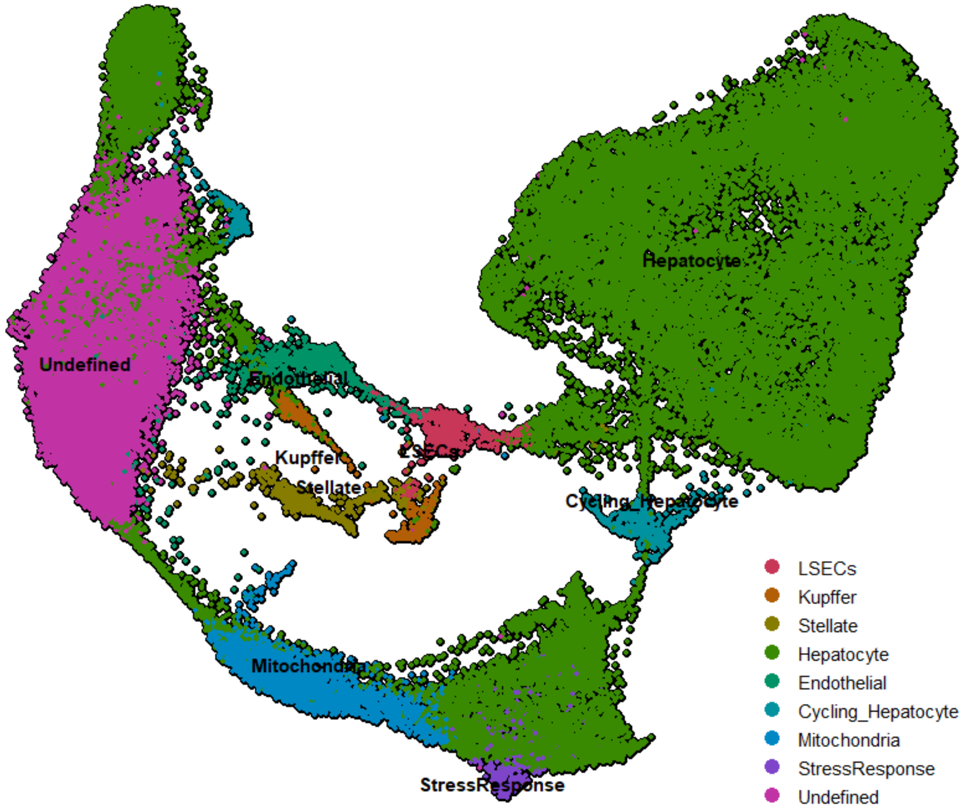
There is not a correct number of clusters, it will depend on the context and biological question


BOD19-1
BOD17-1
BOD17-1
BOD17-3
BOD18KV
BOD19-1
BOD20-3
BOD21-3
BOD22-1
OBE-DIA 1-1
OBE-DIA 2-1


4
35
30
13
29
14
11
8
48
34
52
26
5
51
45
53
42
7
22
6
38
19
36
41
40
39
21
37
24
43
47
2
33
9
15
32
3
54
50
31
28
49
20
27
18
44
0
12
10
23
17
46
25
16
0
1
10
11
12
13
14
15
16
17
18
19
2
20
21
22
23
24
25
26
27
28
29
3
30
31
32
33
34
35
36
37
38
39
4
40
41
42
43
44
45
46
47
48
49
5
50
51
52
53
54
6
7
8
9


MSC
Erythroblast
ERP
MKP/ERP
BMCP
HSC
Plasma
NC_Mono
Int_Mono
MKP
Monocyte
Tcell
Prolifer ating
CD8_Tcell_Naive
pDC
cDC
CLP
Pre_Bcell
Bcell
MDP
MEP
MPP
Pro_Bcell
GMP
CMP
Macrophage
Neutrophil
Bcell_Memory
CD4_Tcell_Activated
CD8_TEM
Tcell/NK
NK
Bcell
Bcell_Memory
BMCP
CD4_Tcell_Activated
CD8_Tcell_Naive
CD8_TEM
cDC
CLP
CMP
ERP
Erythroblast
GMP
HSC
Int_Mono
Macrophage
MDP
MEP
MKP
MKP/ERP
Monocyte
MPP
MSC
NC_Mono
Neutrophil
NK
pDC
Plasma
Pre_Bcell
Pro_Bcell
Proliferating
Tcell
Tcell/NK
Transitional_Bcell


7
22
2
10
4
9
3
0
23
15
20
18
14
13
5
1
17
24
21
19
25
12
6
8
11
0
1
10
11
12
13
14
15
16
17
18
19
2
20
21
22
23
24
25
26
3
4
5
6
7
8
9


Hepatocyte
Undefined
Endothelial
Kupffer
Stellate
LSECs
Cycling_Hepatocyte
Mitochondria
StressResponse
LSECs
Kupffer
Stellate
Hepatocyte
Endothelial
Cycling_Hepatocyte
Mitochondria
StressResponse
Undefined

- https://satijalab.org/seurat/articles/integration_introduction
- https://www.sc-best-practices.org/cellular_structure/integration.html
- https://www.singlecellcourse.org/biological-analysis.html#clustering-introduction
- https://bioconductor.org/books/3.12/OSCA/clustering.html#k-means-clustering
- https://github.com/quadbio/scRNAseq_analysis_vignette/blob/master/Tutorial.md#step-2-3-data-integration-using-liger