

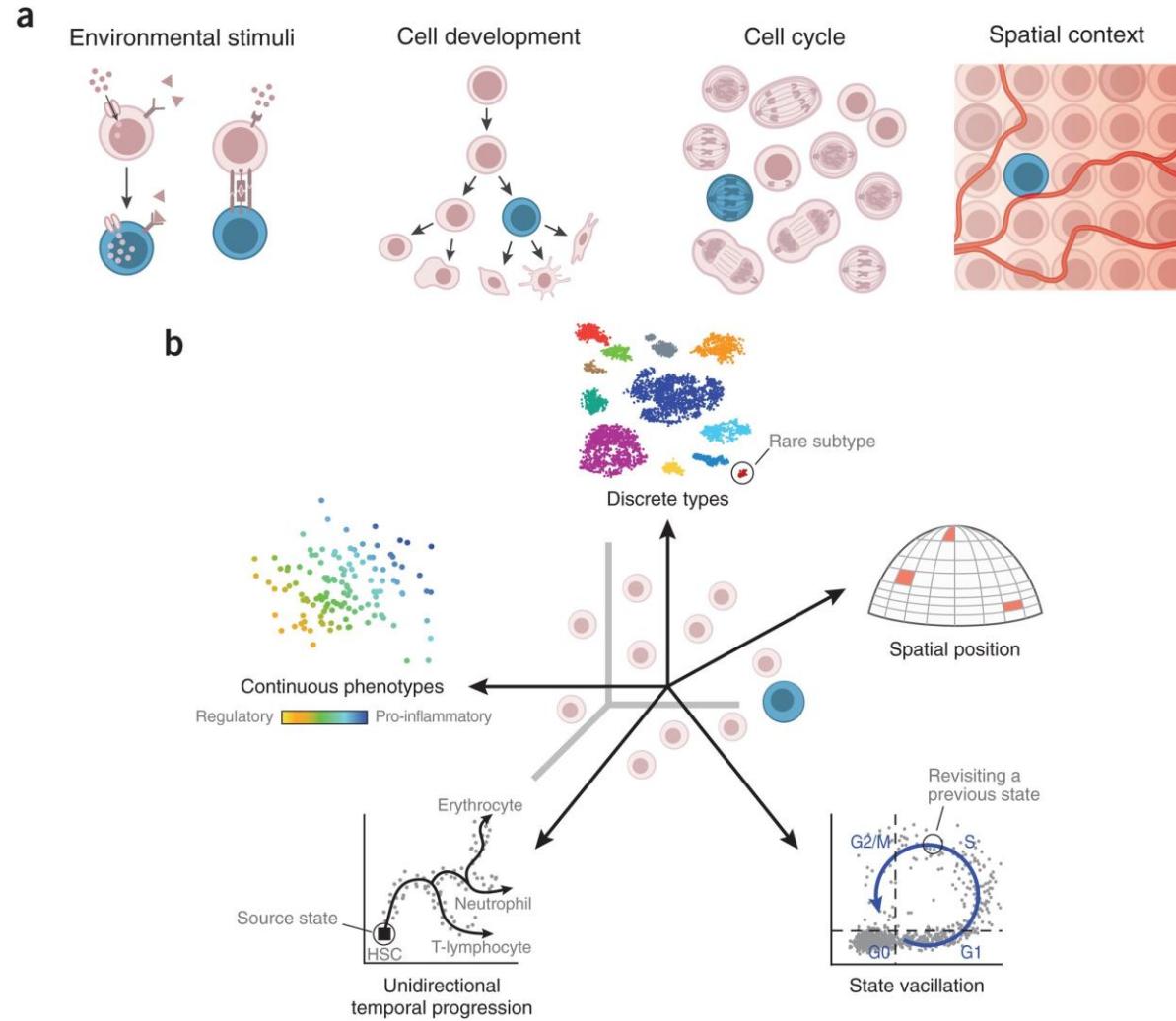
Clustering & cell annotation

Lieke Michielsen

Department of Human Genetics, LUMC
Delft Bioinformatics Lab, TU Delft

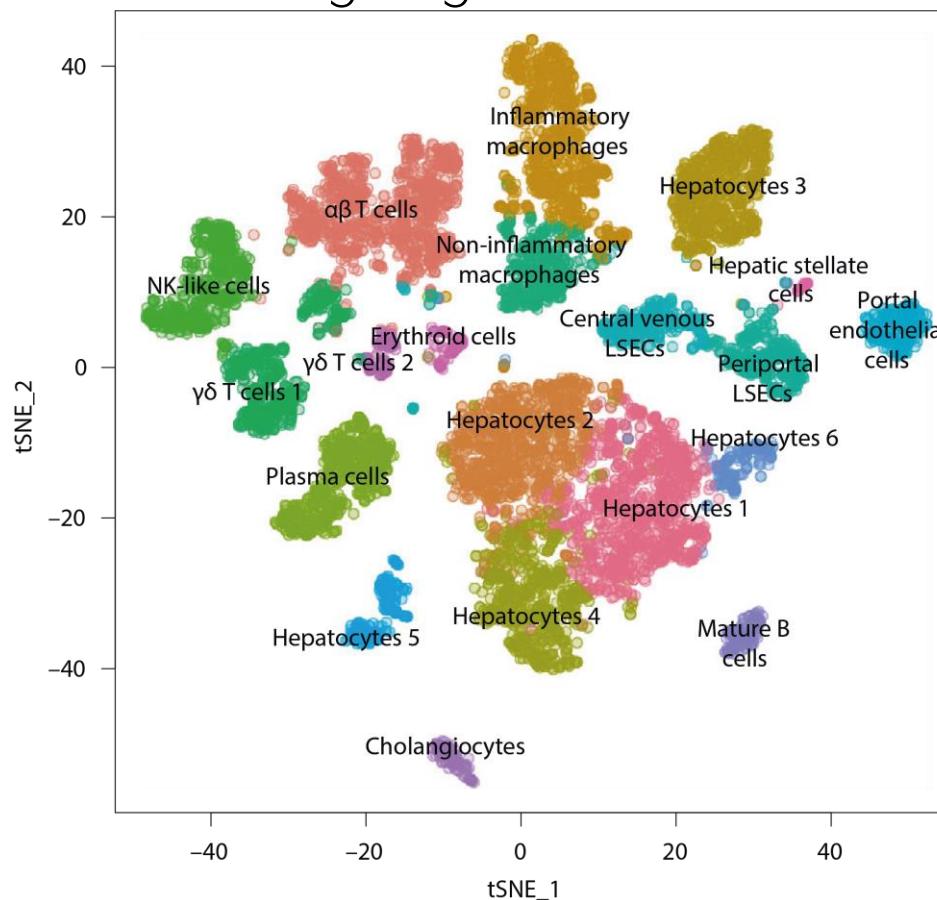
Cell identity

- **Cell type:** the more permanent aspects in a cell's identity (e.g., a hepatocyte typically cannot turn into a neuron).
 - Often organized in a hierarchical taxonomy
- **Cell state:** arise transiently
 - Unidirectional (e.g., during differentiation, stimulation, spatial patterning)
 - Vacillating (e.g., cell-cycle)



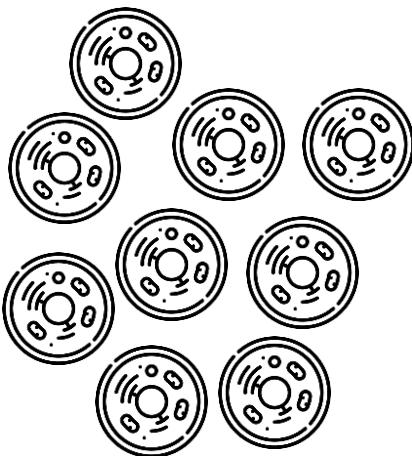
Cell annotation is a universal step

Annotation: giving each cell a label/name

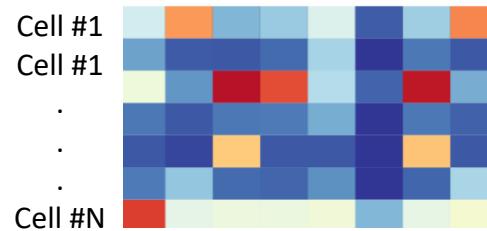


How can we identify cell populations?

Mystery cells



Measure

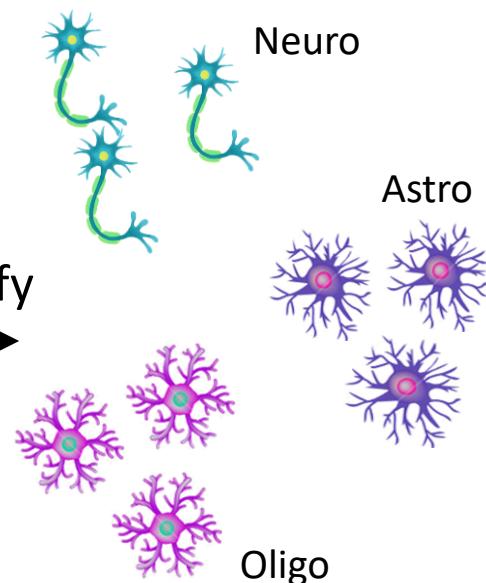


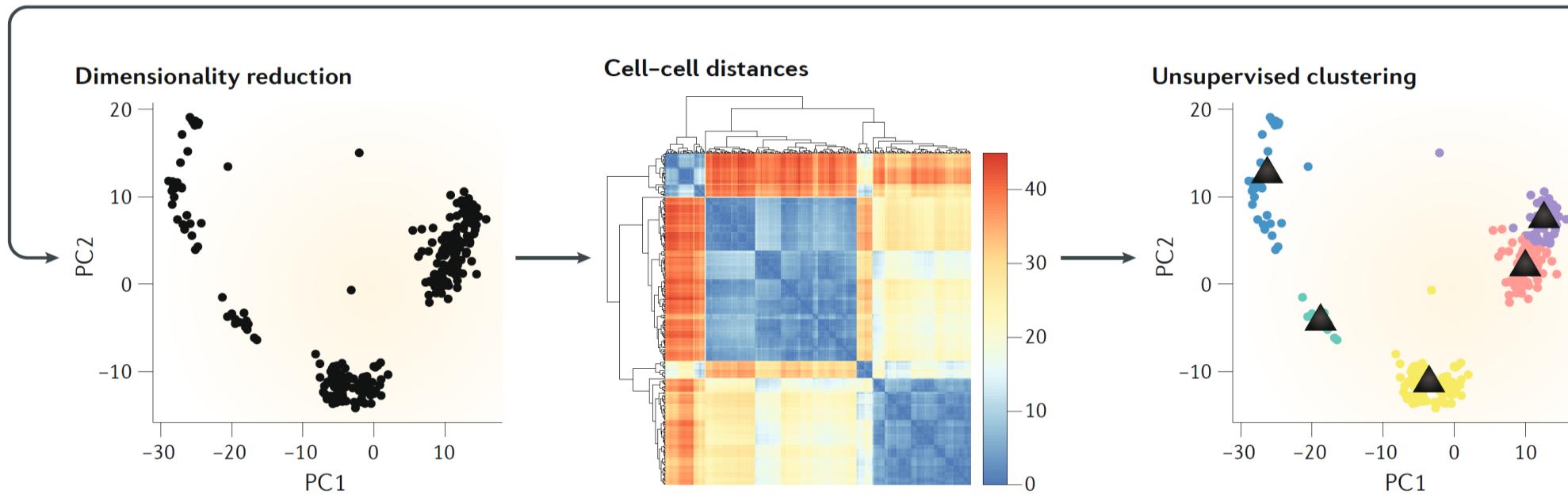
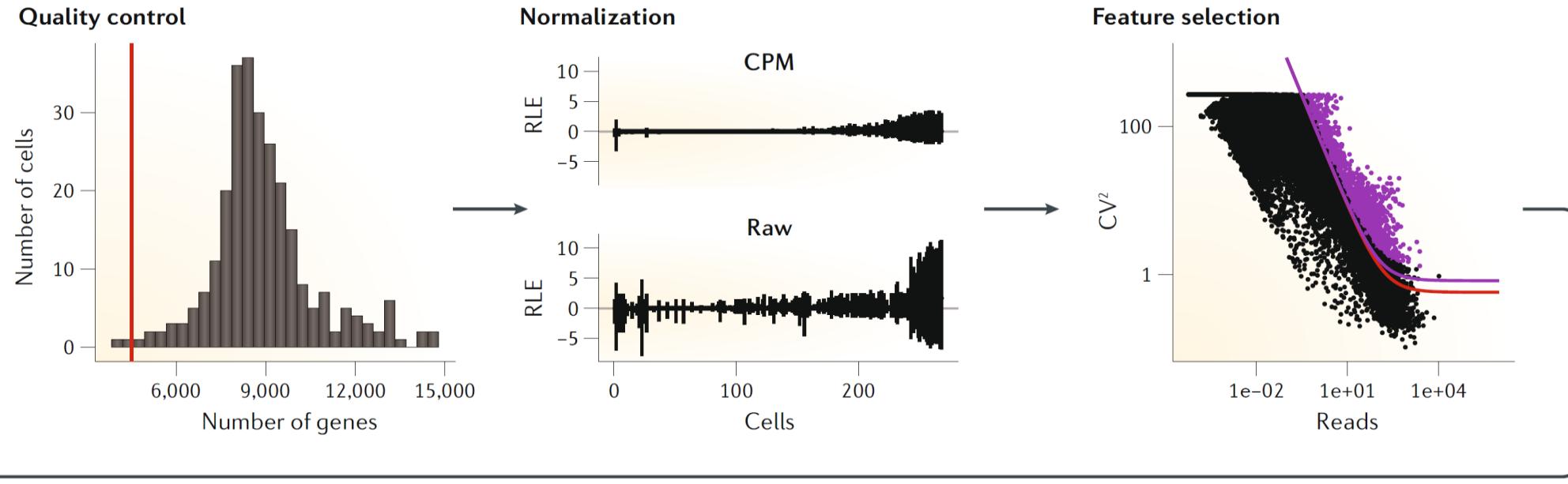
Group



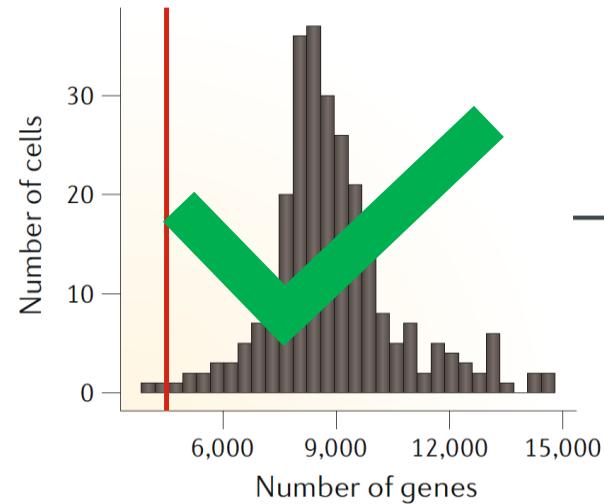
Identify

Cell populations

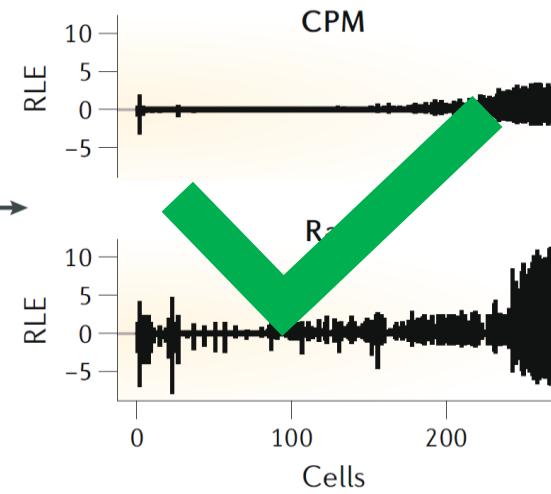




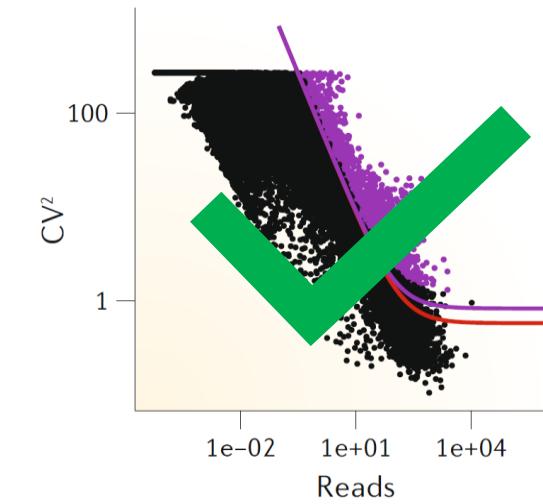
Quality control



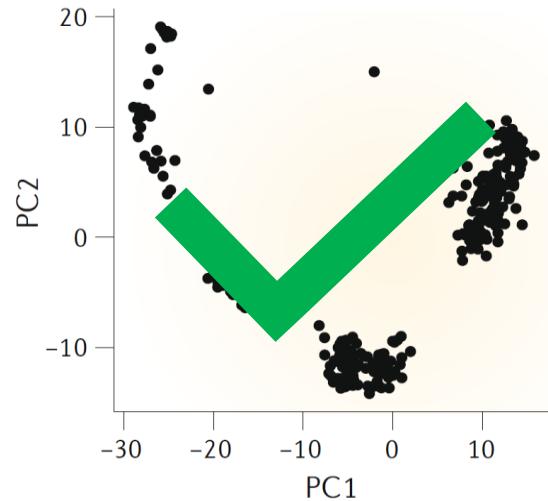
Normalization



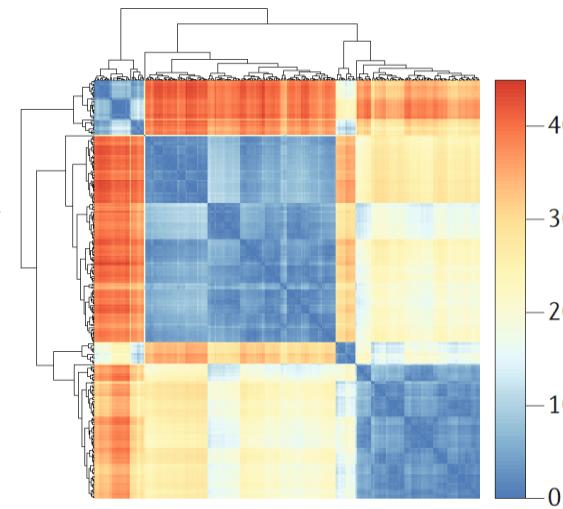
Feature selection



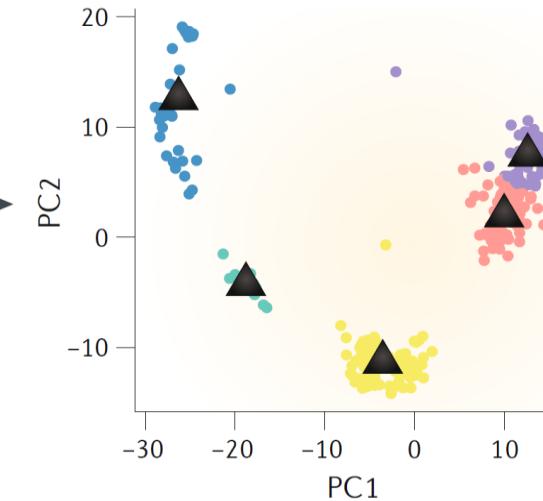
Dimensionality reduction



Cell-cell distances



Unsupervised clustering



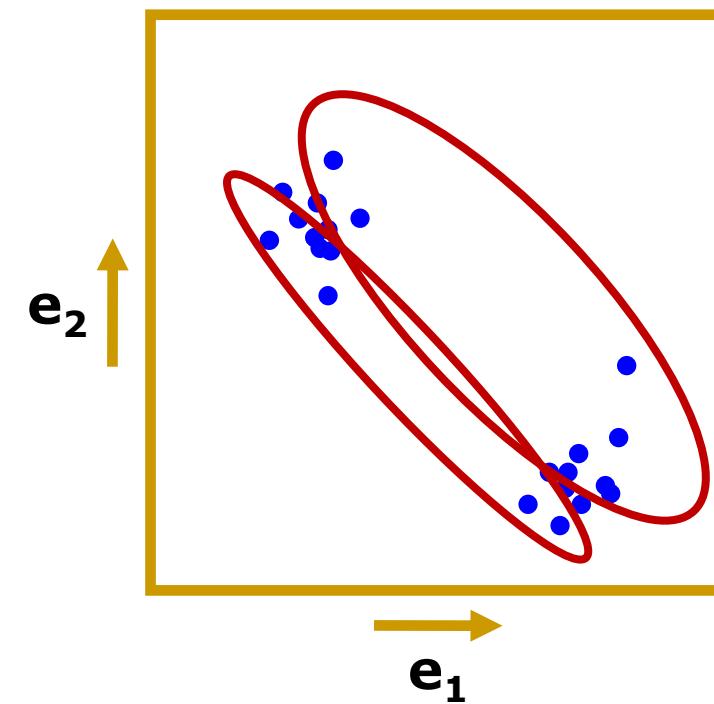
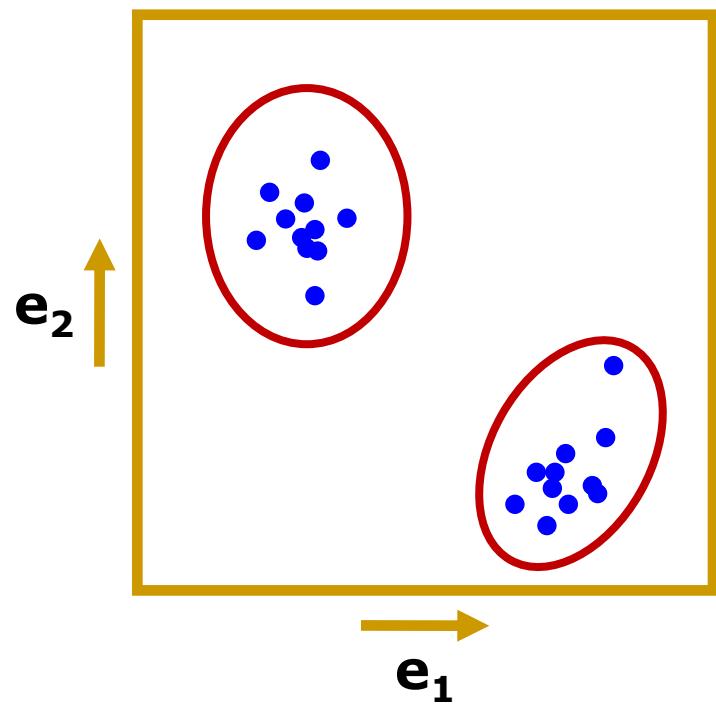
Outline

- Introduction to clustering
 - Hierarchical clustering
 - k -Means clustering
 - Graph-based clustering
- scRNA-seq clustering
- Annotating clusters
- Cluster validation

Outline

- Introduction to clustering
 - Hierarchical clustering
 - k -Means clustering
 - Graph-based clustering
- scRNA-seq clustering
- Annotating clusters
- Cluster validation

Clustering



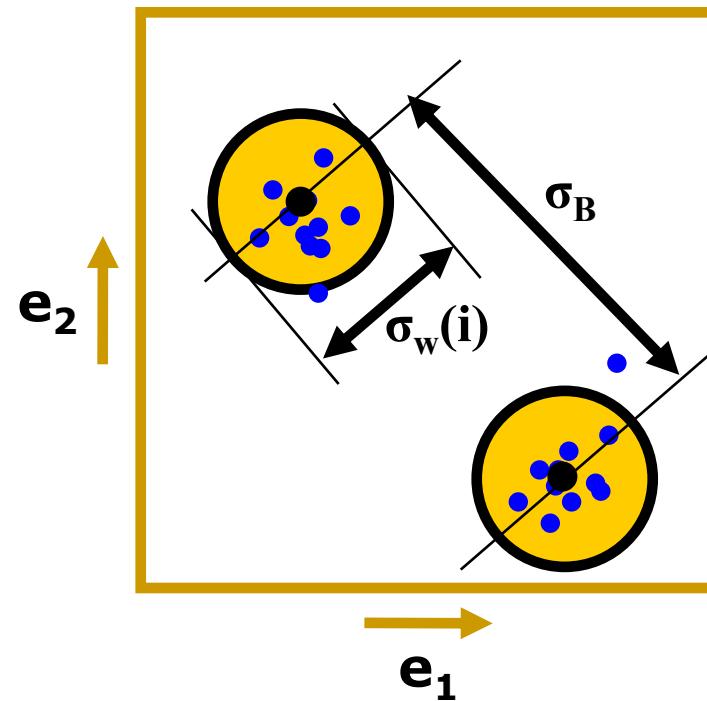
Clustering

Structure when:

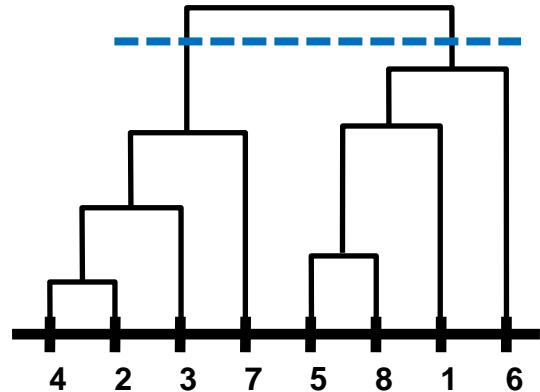
1. Samples within cluster resemble each other
(*small within variance, $\sigma_w(i)$*)
2. Clusters deviate from each other
(*large between variance, σ_B*)

Group samples such that:

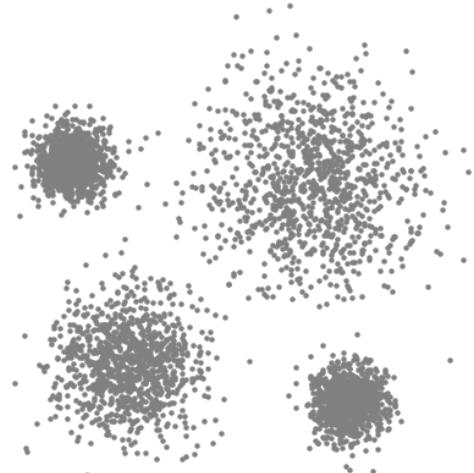
$$\min \left(\frac{\sum_{\text{clusters}} \sigma_w(i)}{\sigma_B} \right) \rightarrow \begin{matrix} \sigma_w: \text{small} & \& \\ \sigma_B: \text{large} & \end{matrix}$$



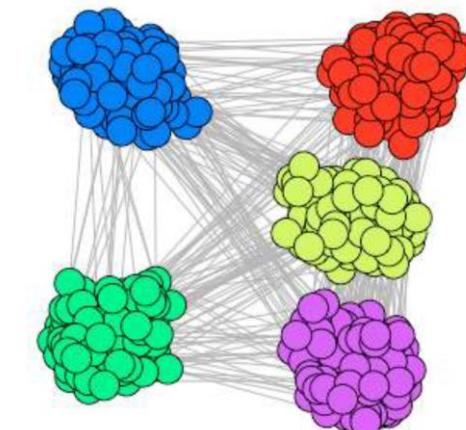
Many clustering approaches



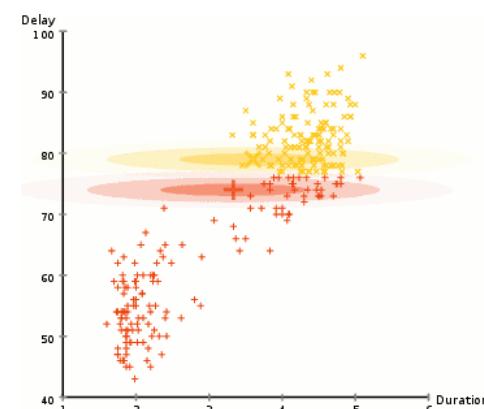
Hierarchical clustering



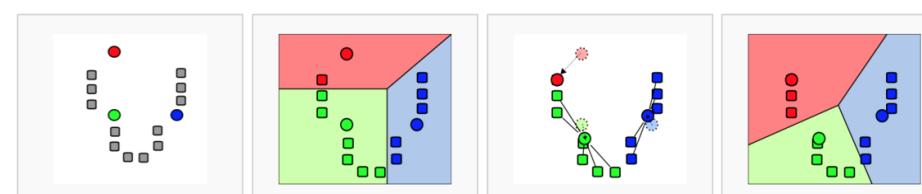
Mean shift clustering



Graph-based clustering

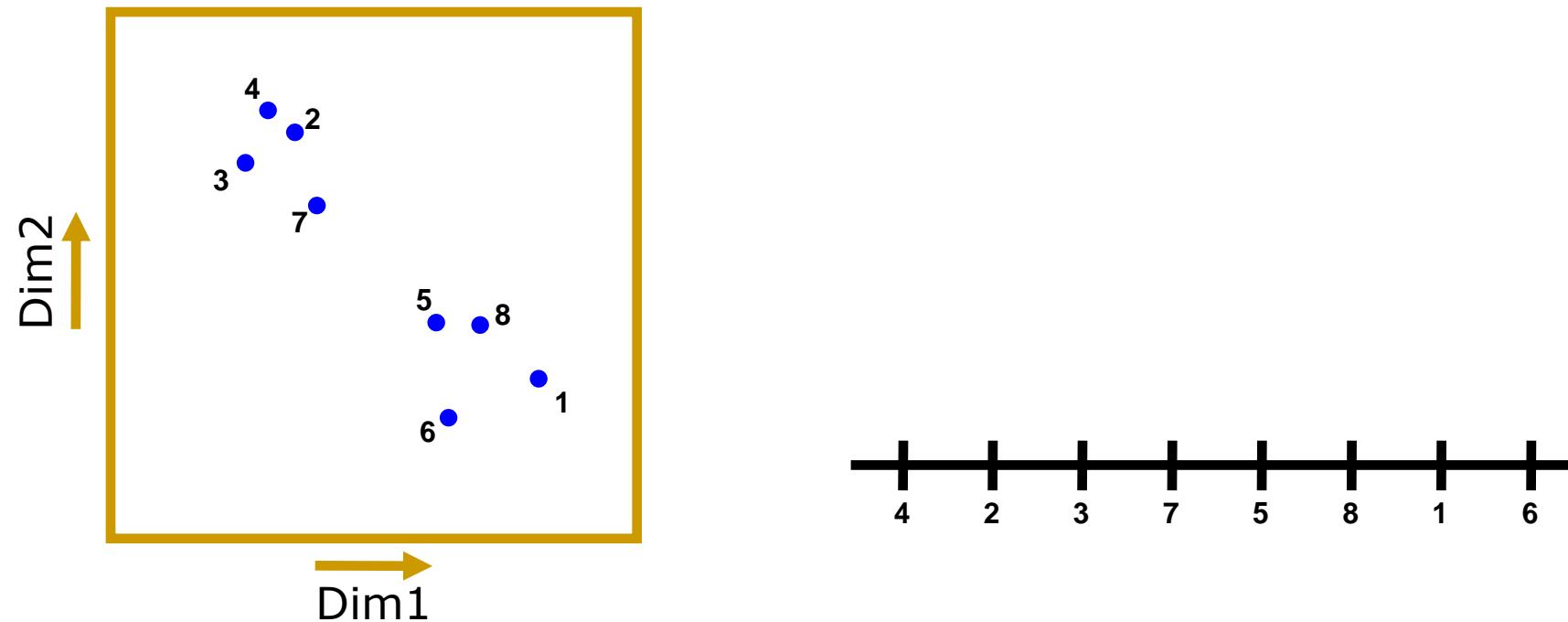


Gaussian mixture modeling



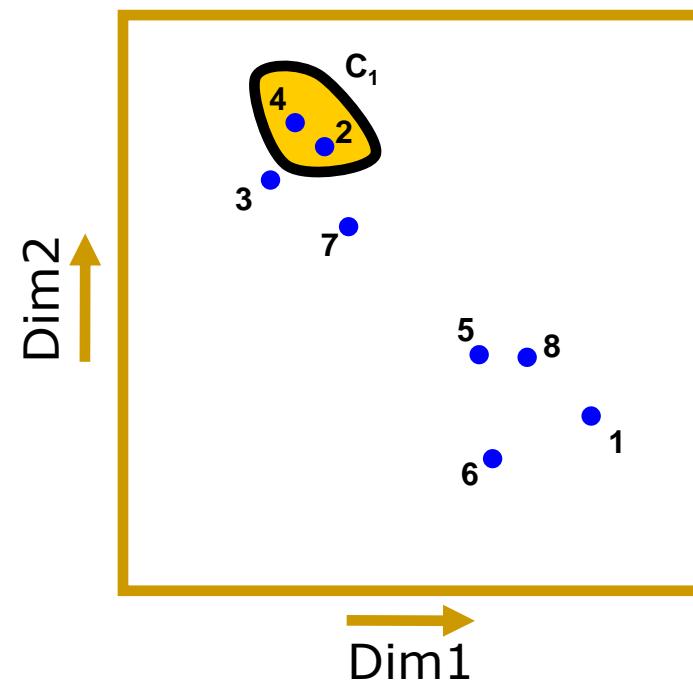
k -means clustering

Hierarchical clustering

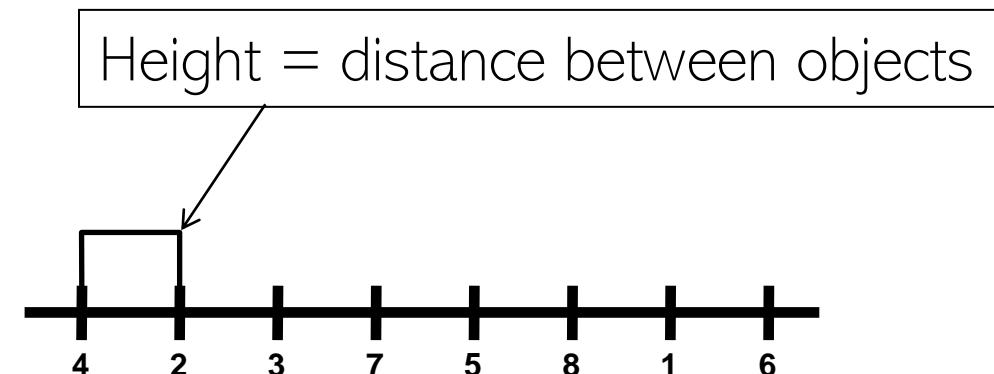


Find most similar objects (cells) and group them

Hierarchical clustering



Dendrogram

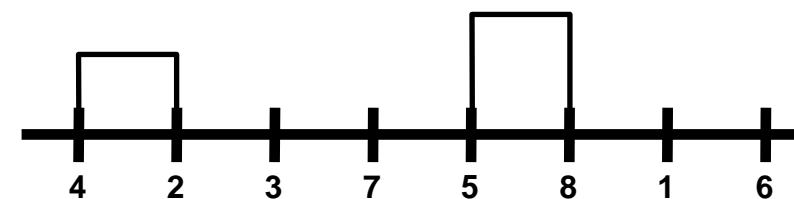
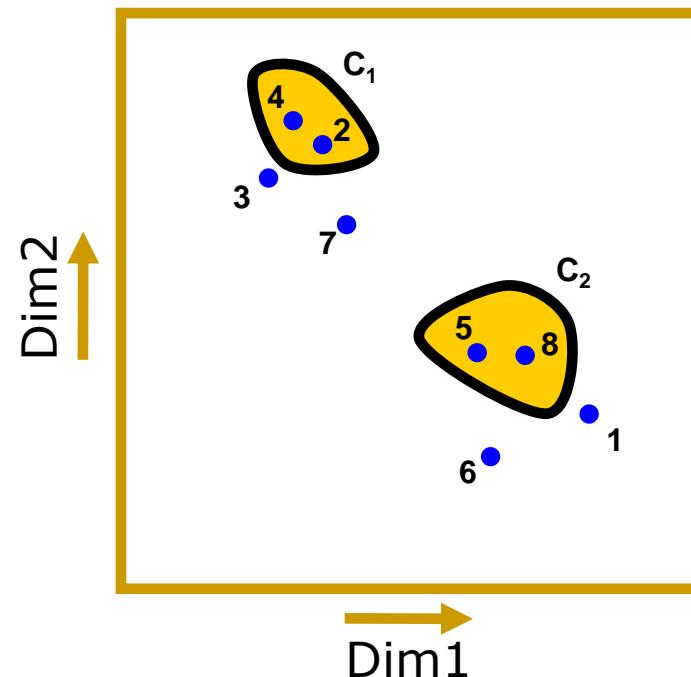


These are: objects 4 and 2

Again, find most similar objects (cells or clusters) and group them

Hierarchical clustering

Dendrogram

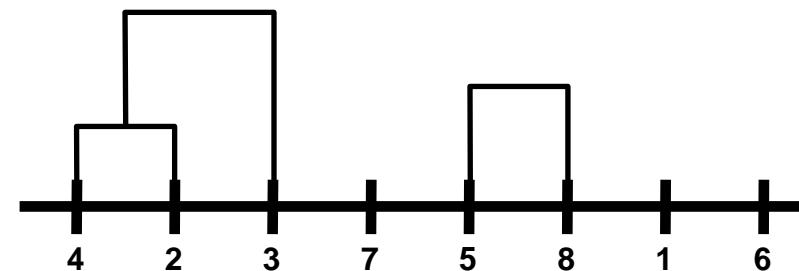
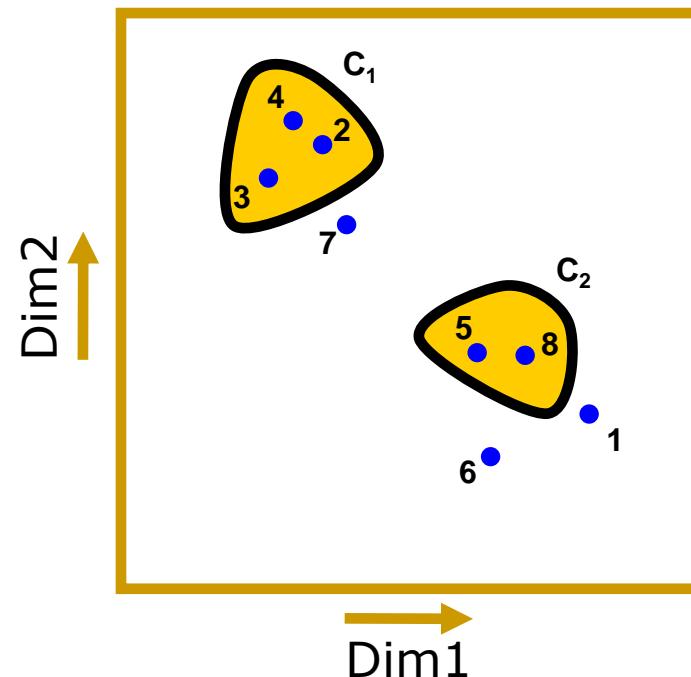


These are: objects 5 and 8

Repeat finding most similar objects (cells or clusters) and group them

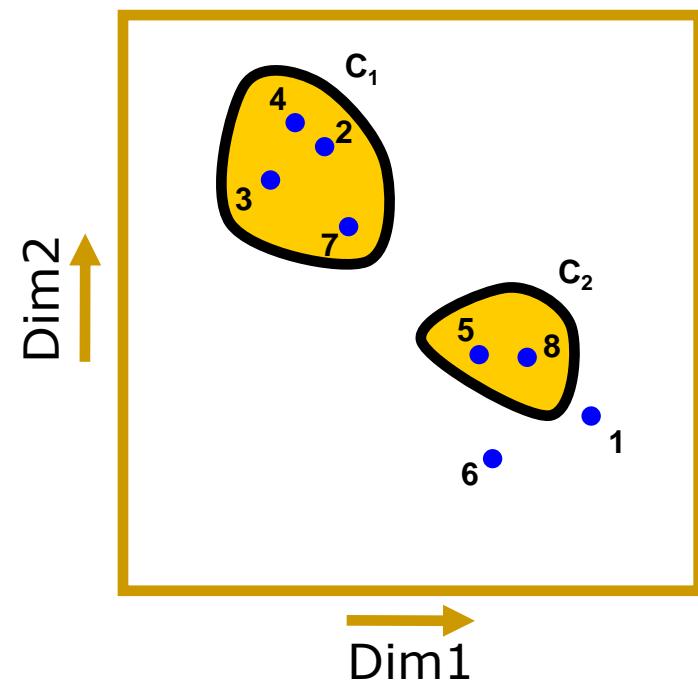
Hierarchical clustering

Dendrogram

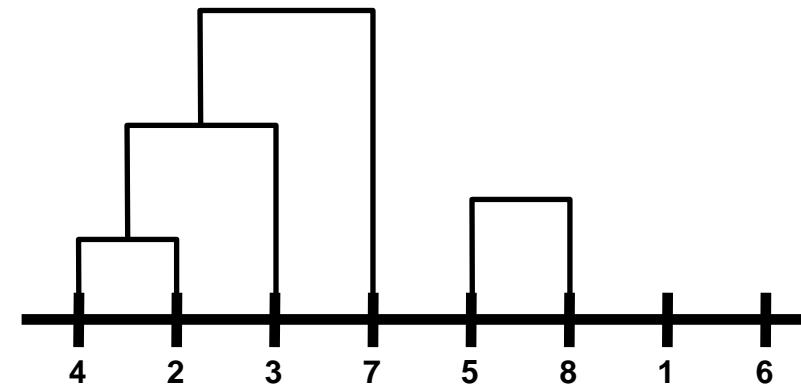


Join object 3 and cluster 1
Repeat process

Hierarchical clustering

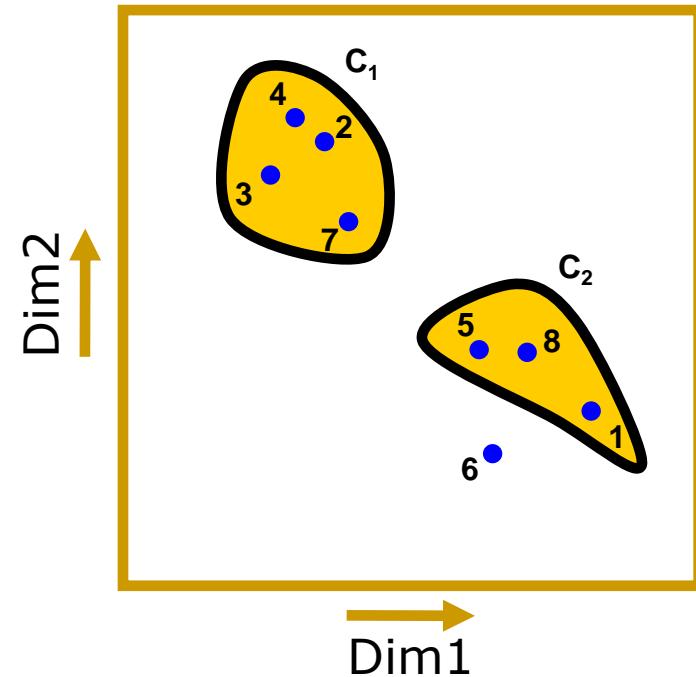


Dendrogram

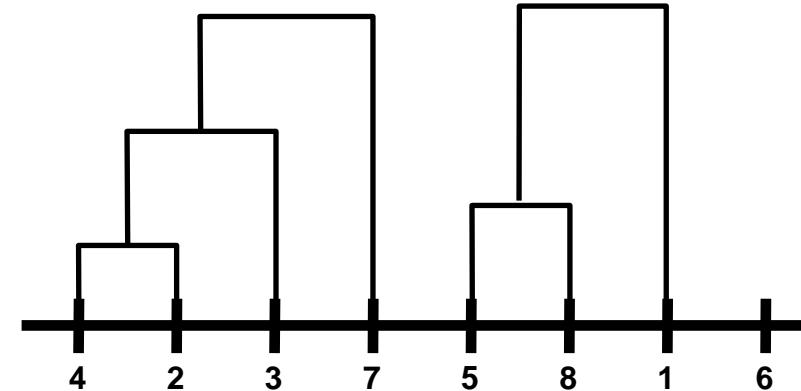


Join [object 7 and cluster 1] -> [cluster 1]
Repeat process

Hierarchical clustering

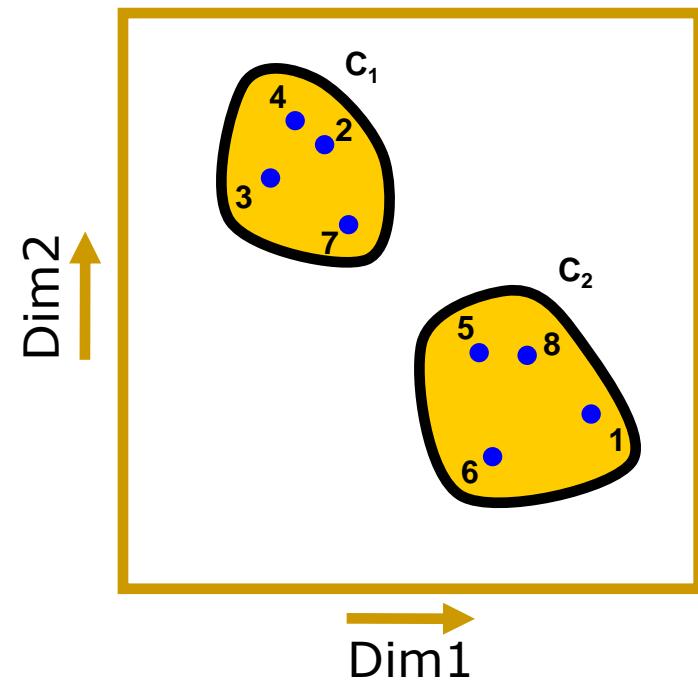


Dendrogram

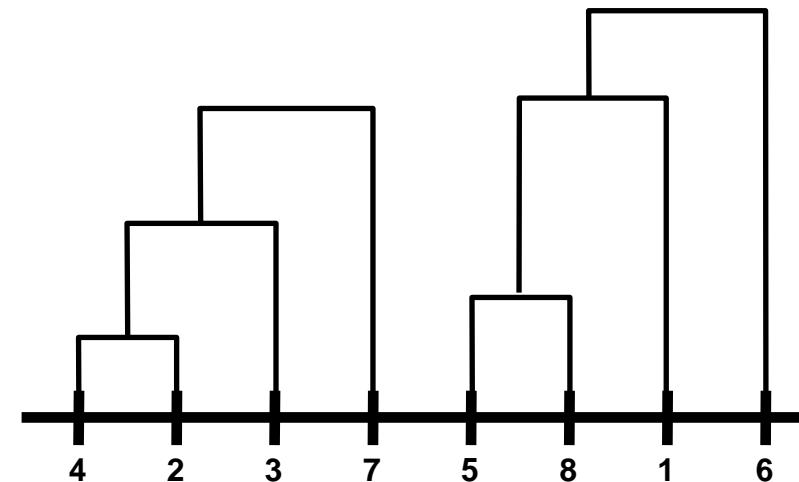


Join [object 1 and cluster 2] -> [cluster 2]
Repeat process

Hierarchical clustering

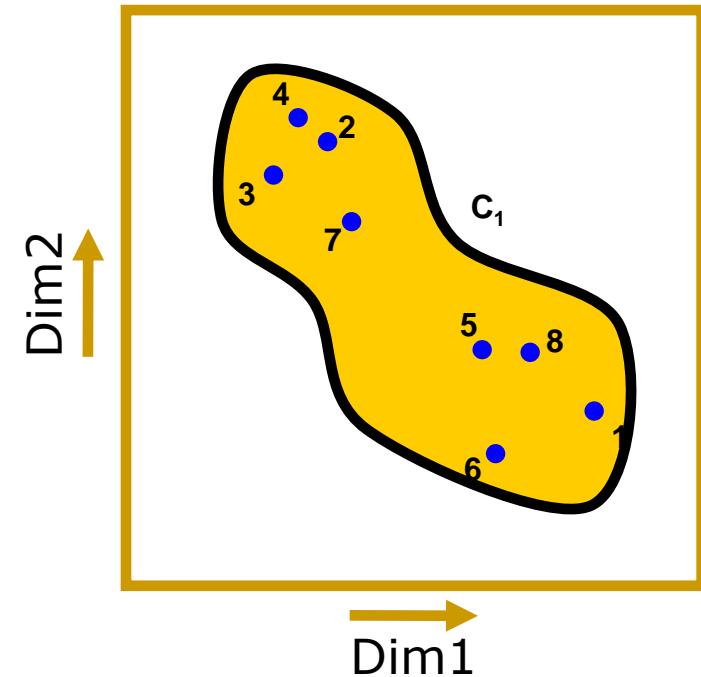


Dendrogram

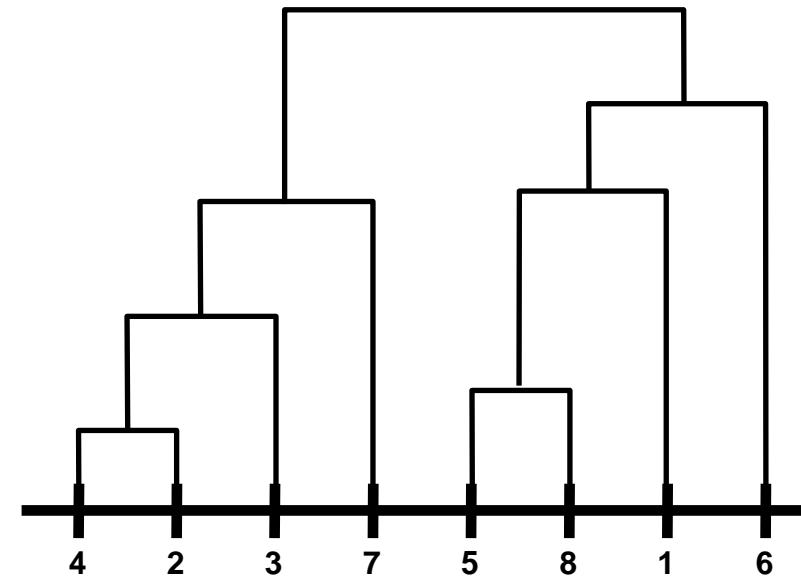


Join [object 6 and cluster 2] -> [cluster 2]
Repeat process

Hierarchical clustering

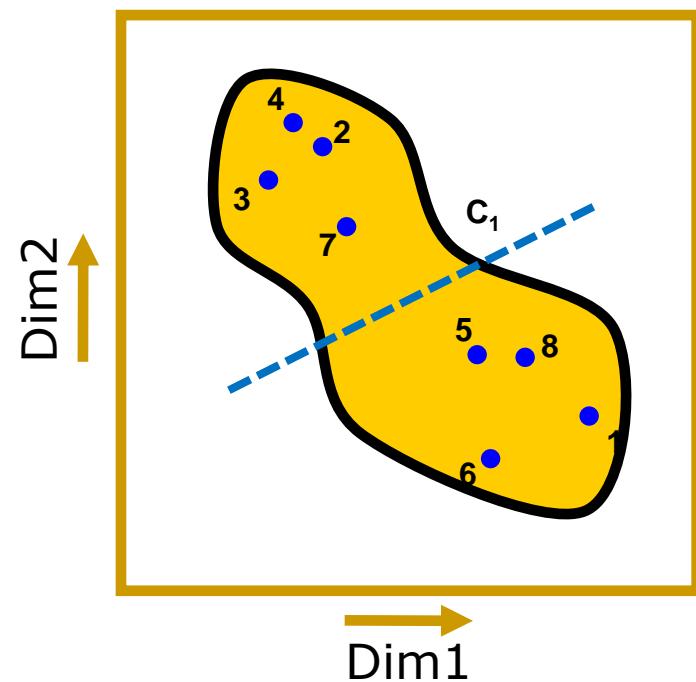


Dendrogram

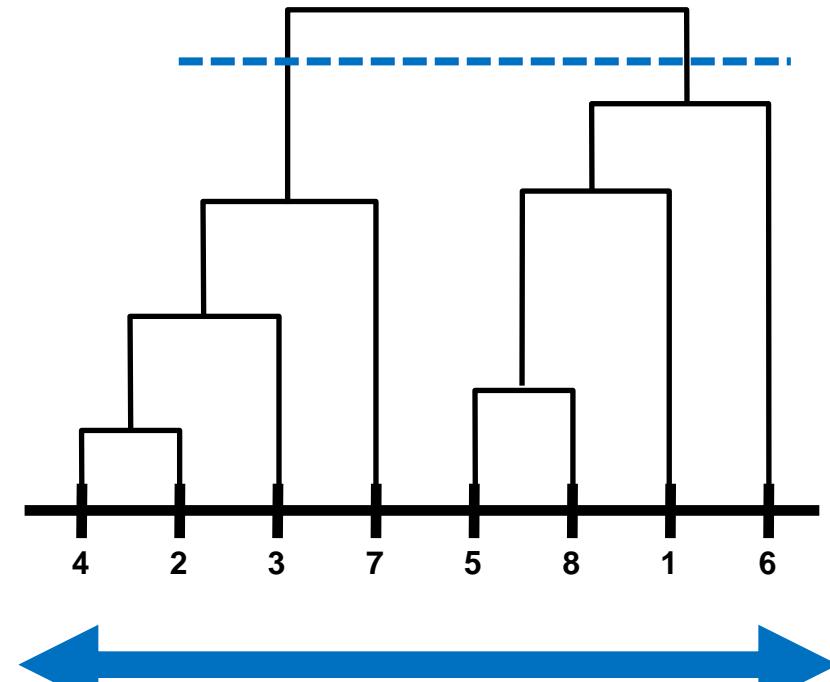


Join [cluster 1 and cluster 2] -> [cluster 1]
All in one cluster: FINISHED!

Hierarchical clustering



Dendrogram

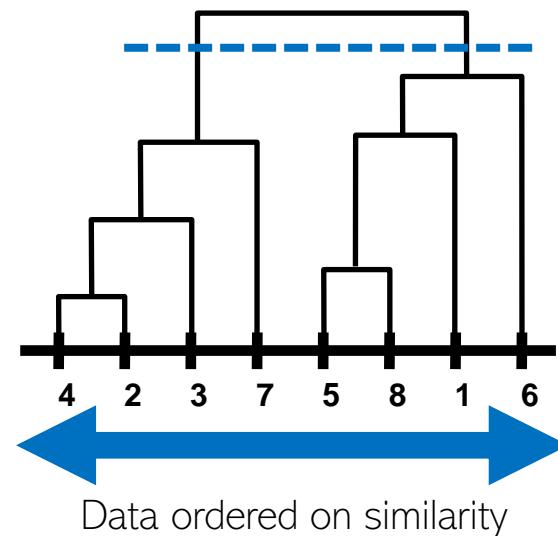
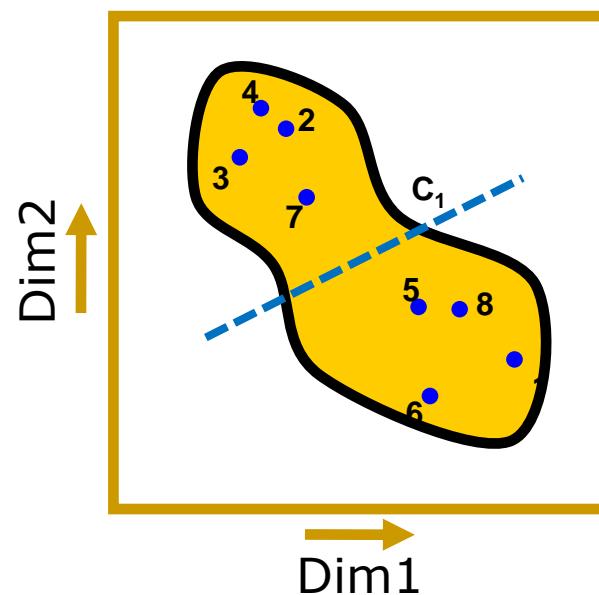


Data ordered on similarity

Hierarchical clustering

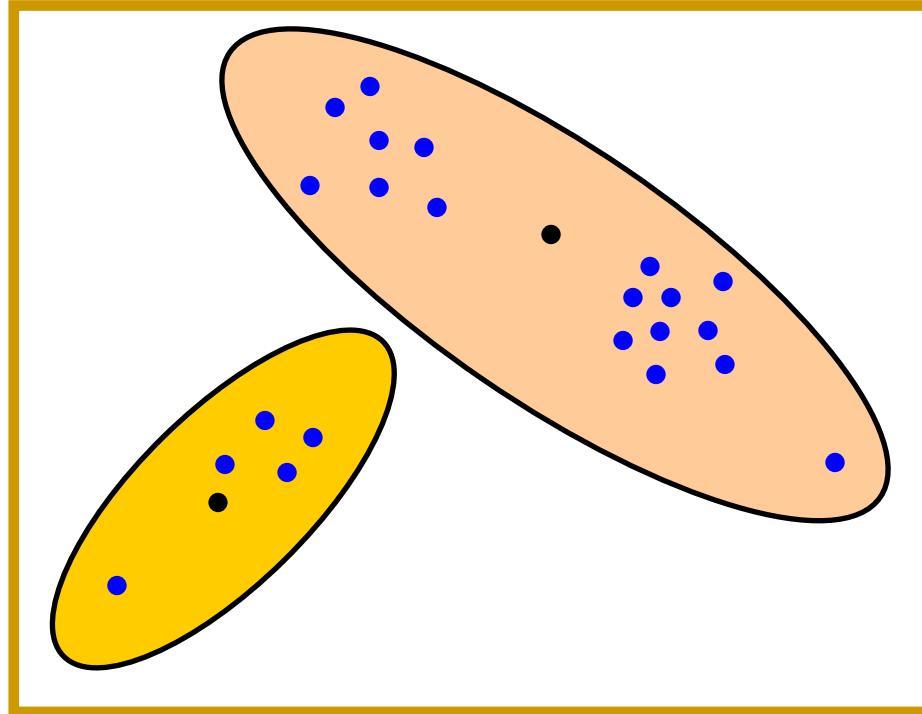
Need to know:

- Similarity between objects
- Similarity between clusters



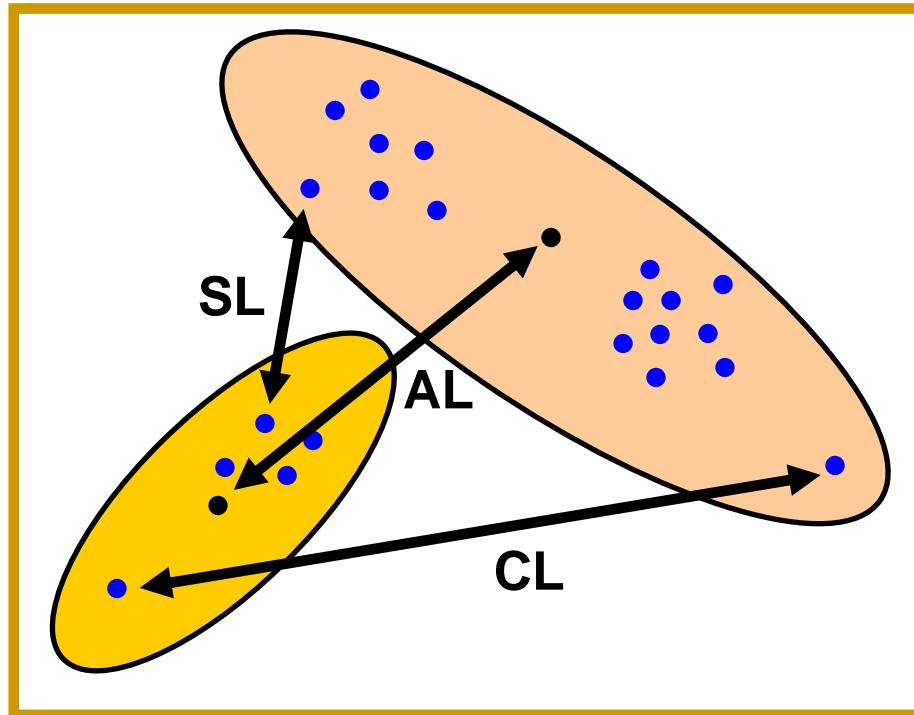
Hierarchical clustering

Similarity between clusters



Hierarchical clustering

Similarity between clusters



Single linkage:

Complete linkage:

Average linkage:

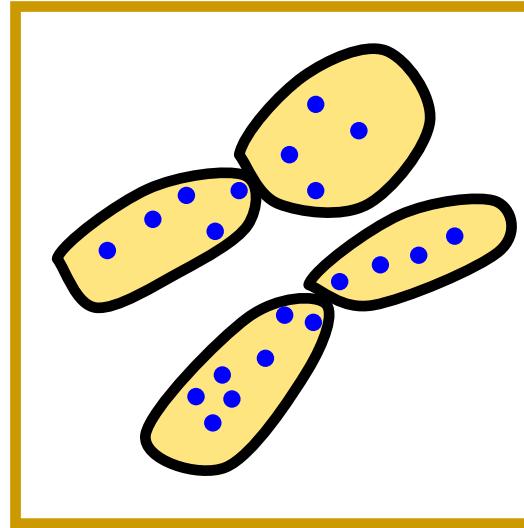
Closest objects

Furthest objects

Average similarity

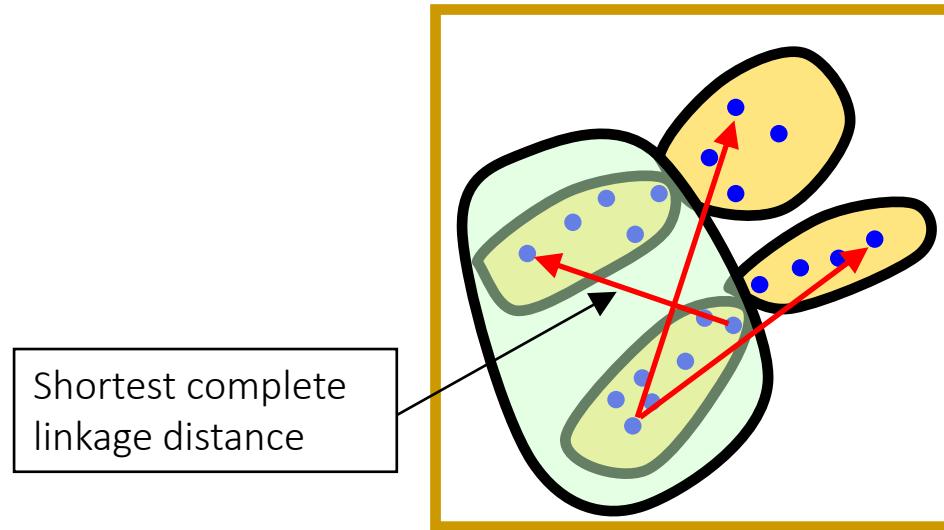
Hierarchical clustering

Similarity between clusters



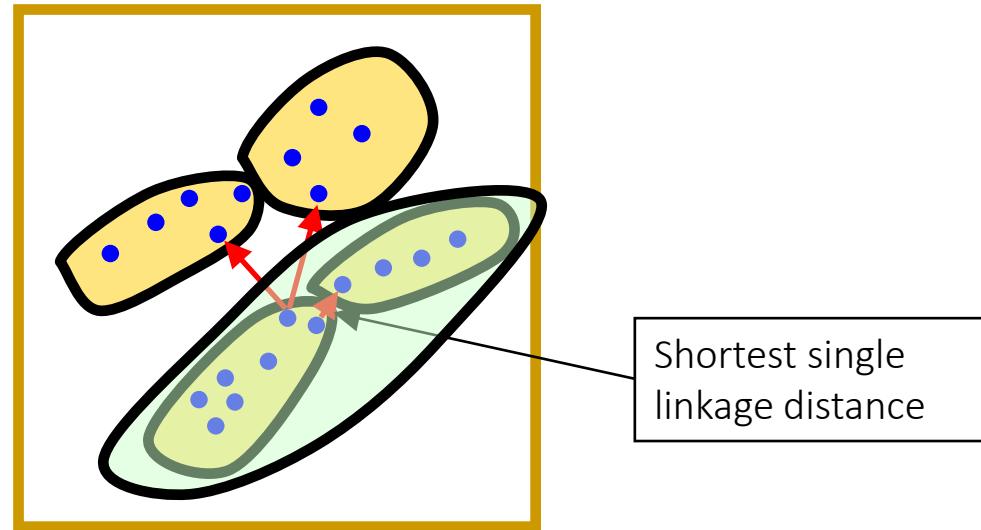
Hierarchical clustering

Similarity between clusters



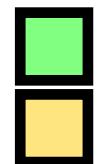
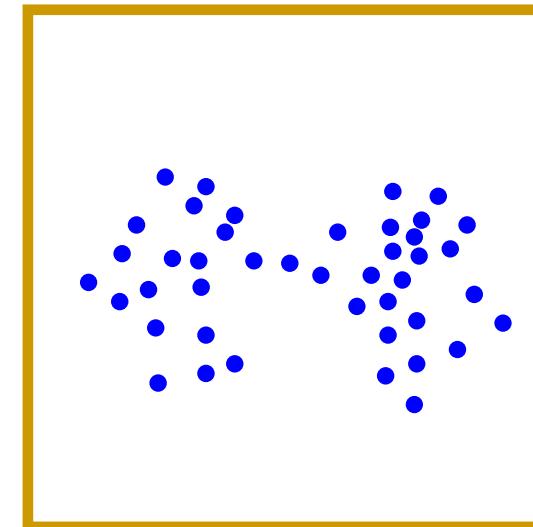
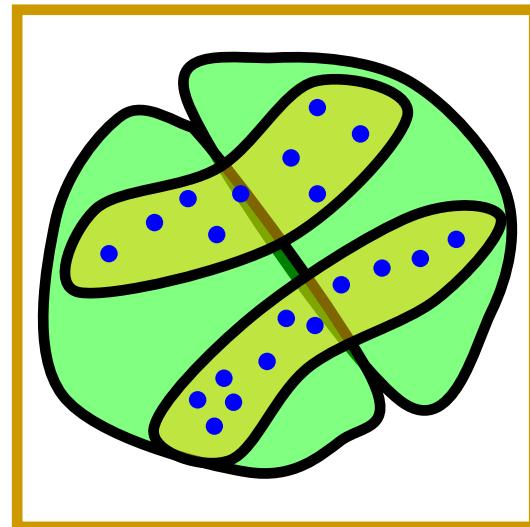
Hierarchical clustering

Similarity between clusters



Hierarchical clustering

Similarity between clusters

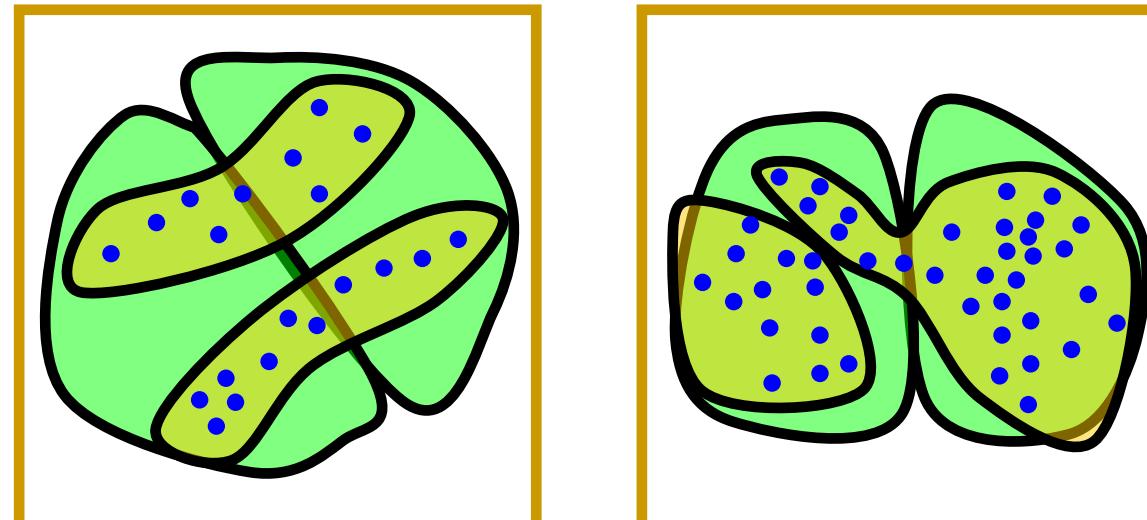


complete linkage
single linkage

Hierarchical clustering

Similarity between clusters

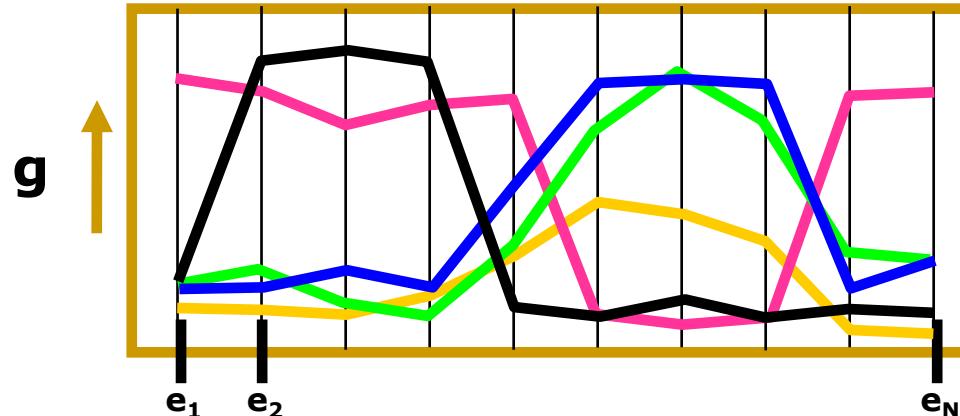
- Single linkage → long and “loose” clusters
- Complete linkage → compact clusters



complete linkage
single linkage

Hierarchical clustering

Similarity between objects



Euclidean distance

$$d(g_i, g_j) = \sqrt{(\sum((x_i - x_j)^2))}$$

$d(\bullet, \bullet) < d(\bullet, \circ)$
 $d(\bullet, \bullet) << d(\bullet, \circ)$
 $d(\bullet, \bullet) << d(\bullet, \bullet)$

Match exact shape

Pearson correlation

$$1 - \rho_{ij}$$

$d(\bullet, \bullet) \approx d(\bullet, \circ)$
 $d(\bullet, \bullet) << d(\bullet, \circ)$
 $d(\bullet, \bullet) << d(\bullet, \bullet)$

Ignore amplitude

Mixed Pearson correlation

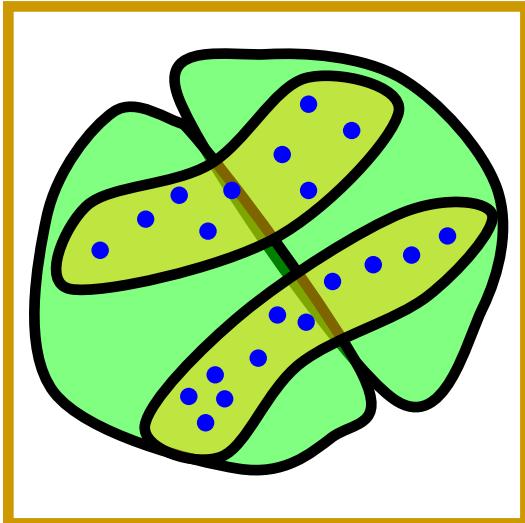
$$1 - |\rho_{ij}|$$

$d(\bullet, \bullet) \approx d(\bullet, \circ)$
 $d(\bullet, \bullet) \approx d(\bullet, \circ)$
 $d(\bullet, \bullet) << d(\bullet, \bullet)$

Ignore amplitude and sign

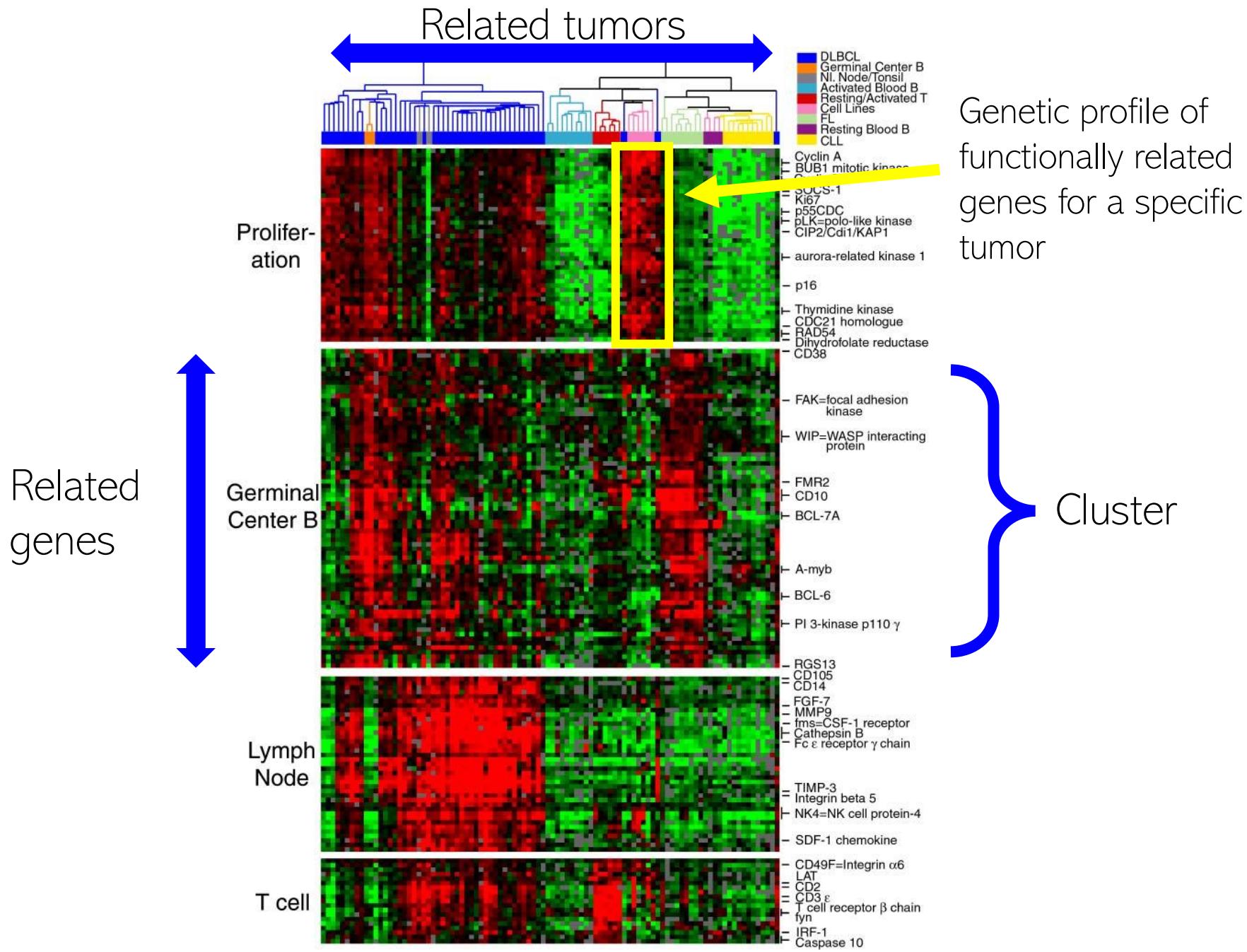
Hierarchical clustering

Overview



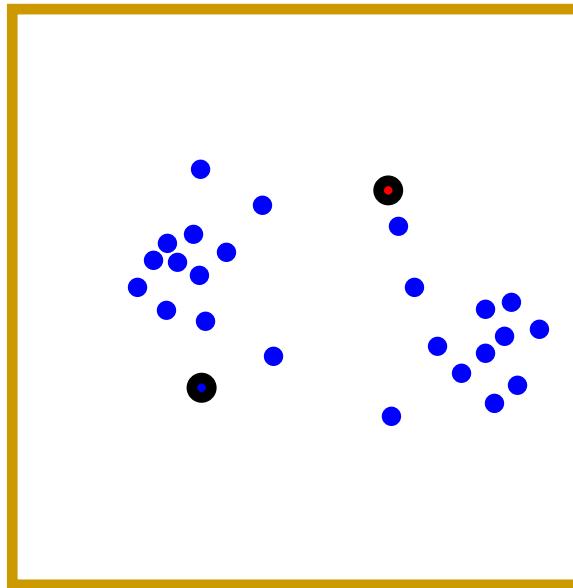
complete linkage
single linkage

- Hierarchical clustering
 - Choice of distance measure
 - Choice of linkage type
- Distance measure
 - Euclidean
 - Correlation
- Linkage
 - Single
 - Average
 - Complete
- Number of clusters
 - Validate clustering!



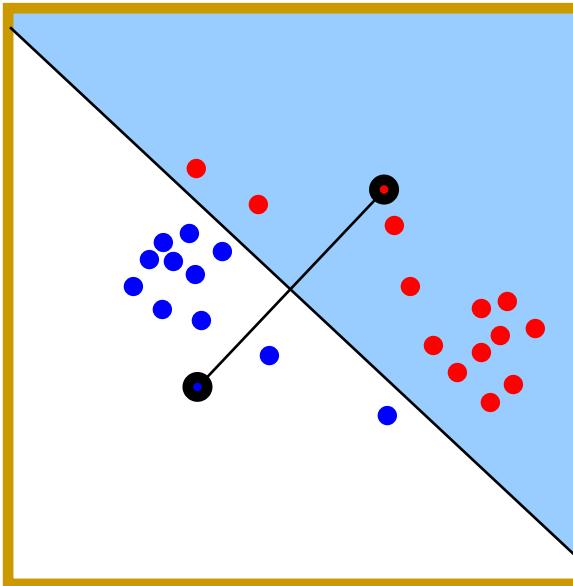
Genetic profile of functionally related genes for a specific tumor

k -Means clustering



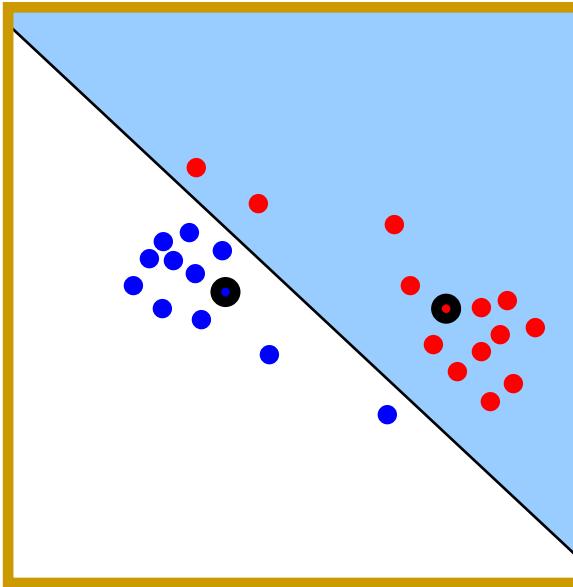
Choose randomly 2 prototypes

k -Means clustering



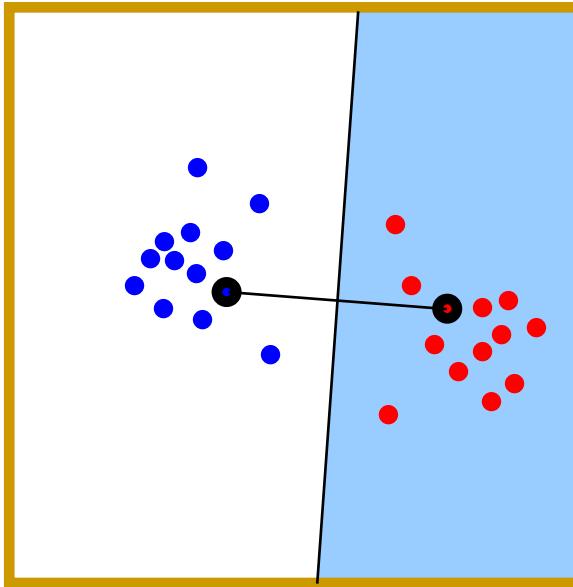
Assign objects to closest prototype
Blue area: cluster 1
White area: cluster 2

k -Means clustering



Calculate new cluster prototypes
By averaging objects

k -Means clustering

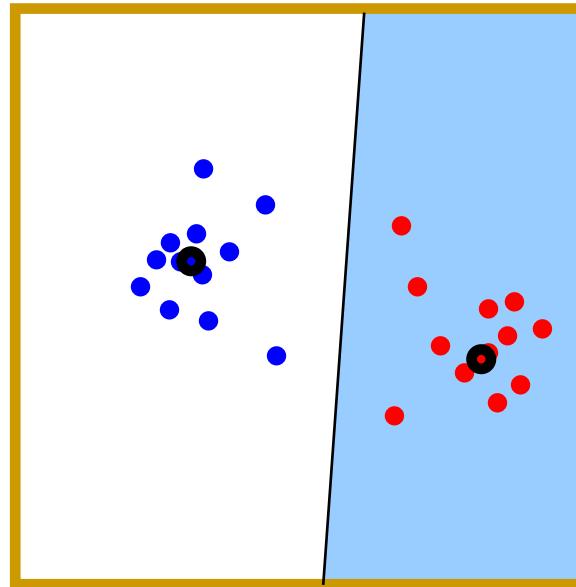


Re-assign objects to closest prototype

Blue area: cluster 1

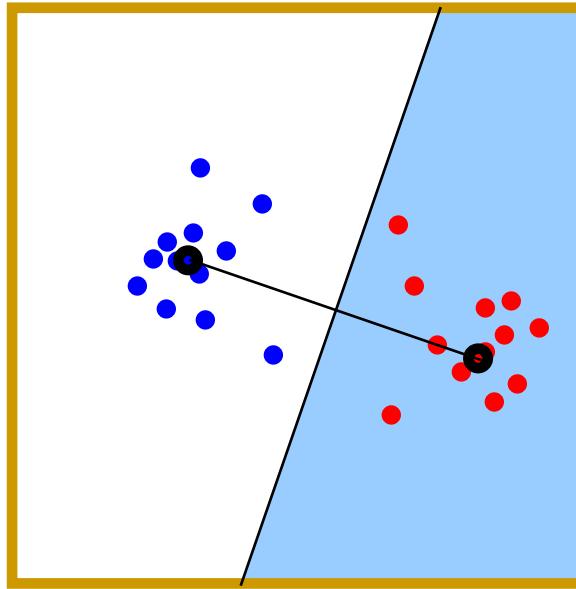
White area: cluster 2

k -Means clustering



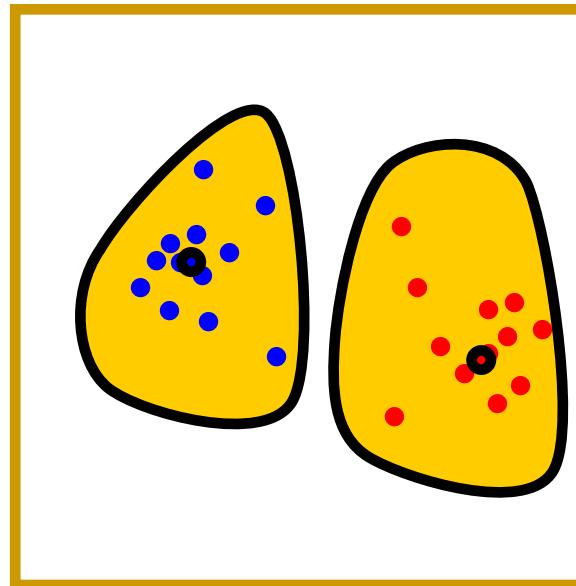
Re-calculate new cluster prototypes

k -Means clustering



Re-assign objects to closest prototype
If no objects change cluster, then finished

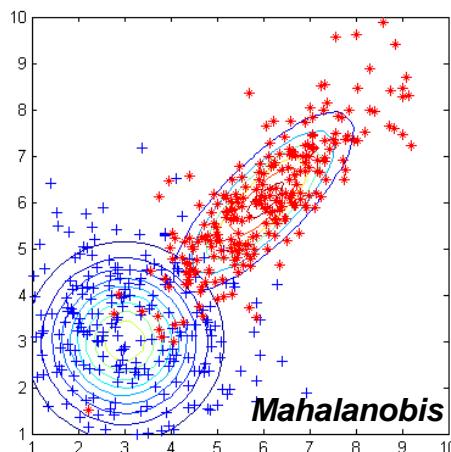
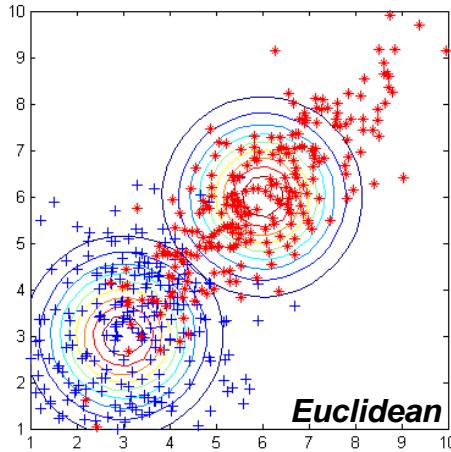
k -Means clustering



Establish clusters

k -Means clustering

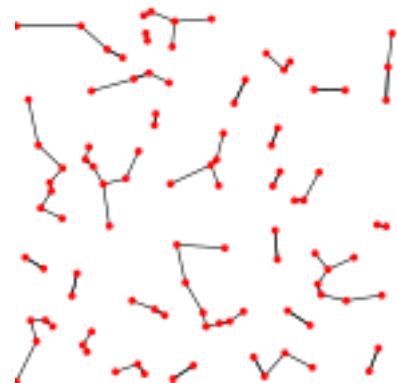
Overview



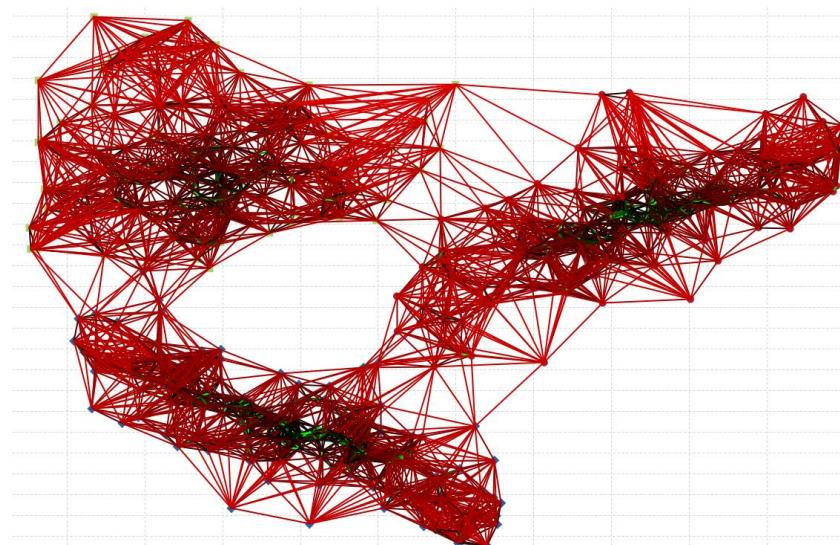
- k -Means
 - Fixed number of clusters (need to know a priori)
 - Choice of distance measure
 - Prototype choice
- Distance measure
 - Euclidean: Round clusters
 - Mahalanobis: Elongated clusters
- Prototype choice
 - Point
 - Line etc.
- Number of clusters
 - Validate clustering!

Graph-based clustering

- k-NN graph: connect every node to its k -nearest neighbors
- Find densely connected components (communities)



$k=1$



$k=20$

Graph-based clustering

- Maximize modularity score
 - Dense connections between nodes within clusters
 - Sparse connections between nodes in different clusters

Observed edges
in cluster c

Expected edges
in cluster c

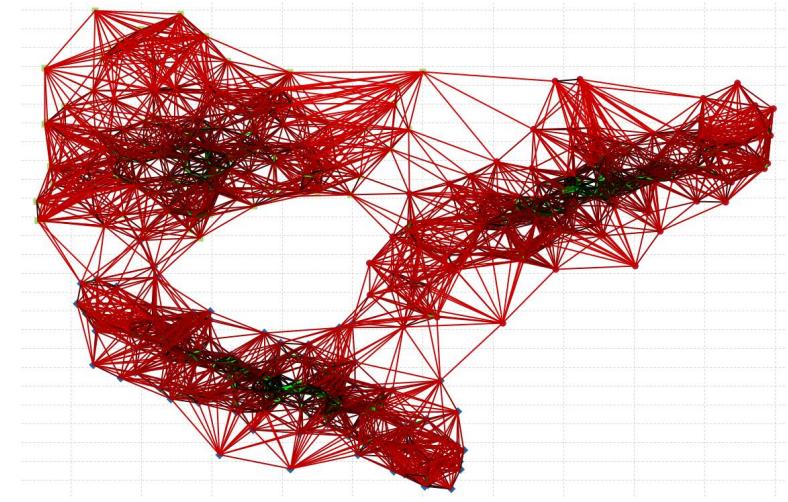
$$H = \frac{1}{2m} \sum_c (e_c - \gamma \frac{K_c^2}{2m})$$

m = edges in the graph

e_c = edges in cluster c

K_c = $\sum_{n \in c}$ degree(n)

γ = resolution

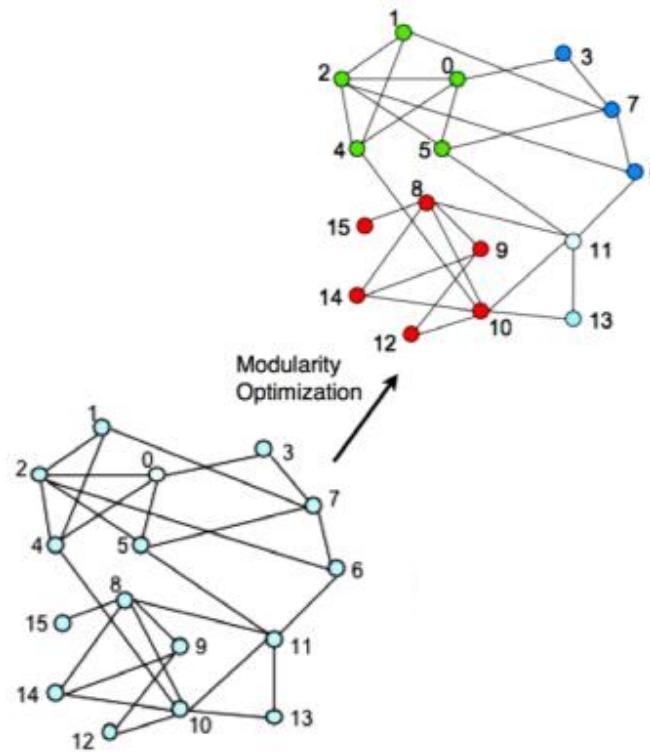


Graph-based clustering

Louvain algorithm

Two steps

1. Local moving of nodes: move node i to community of neighbor j , if this increases H
2. Aggregate nodes



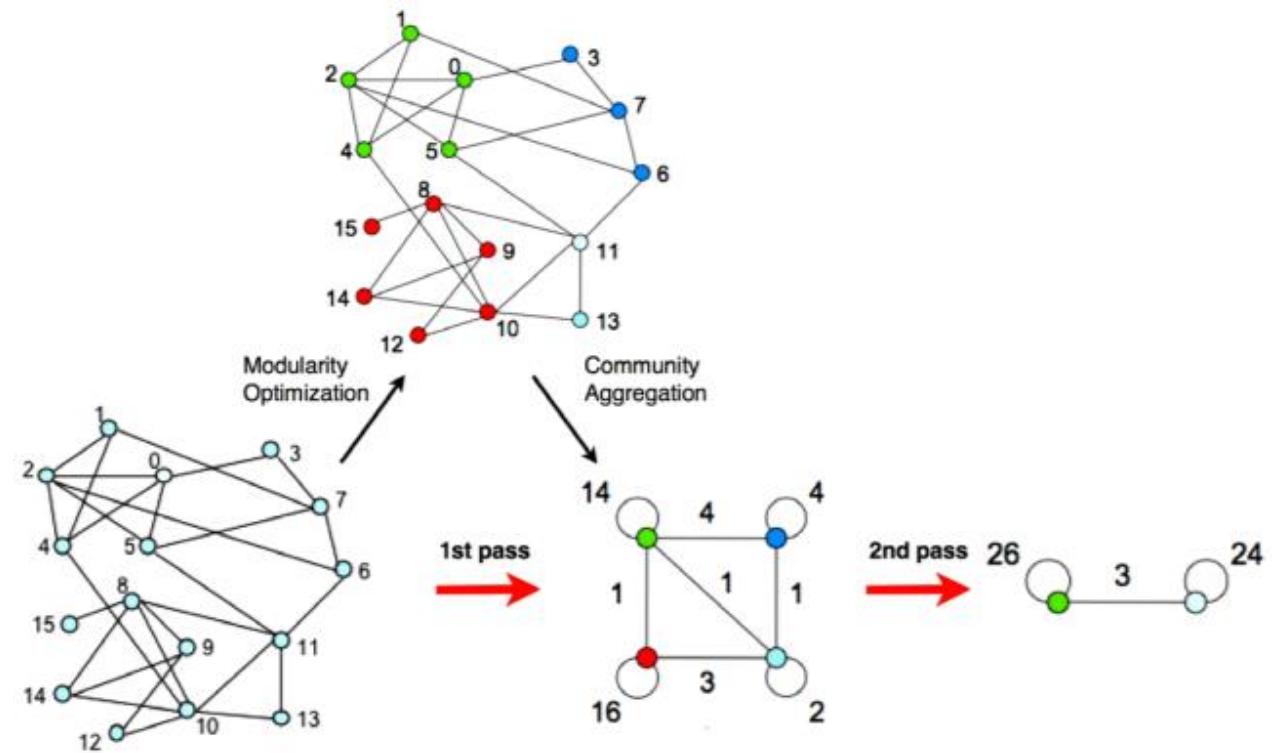
Graph-based clustering

Louvain algorithm

Two steps

1. Local moving of nodes: move node i to community of neighbor j , if this increases H
2. Aggregate nodes

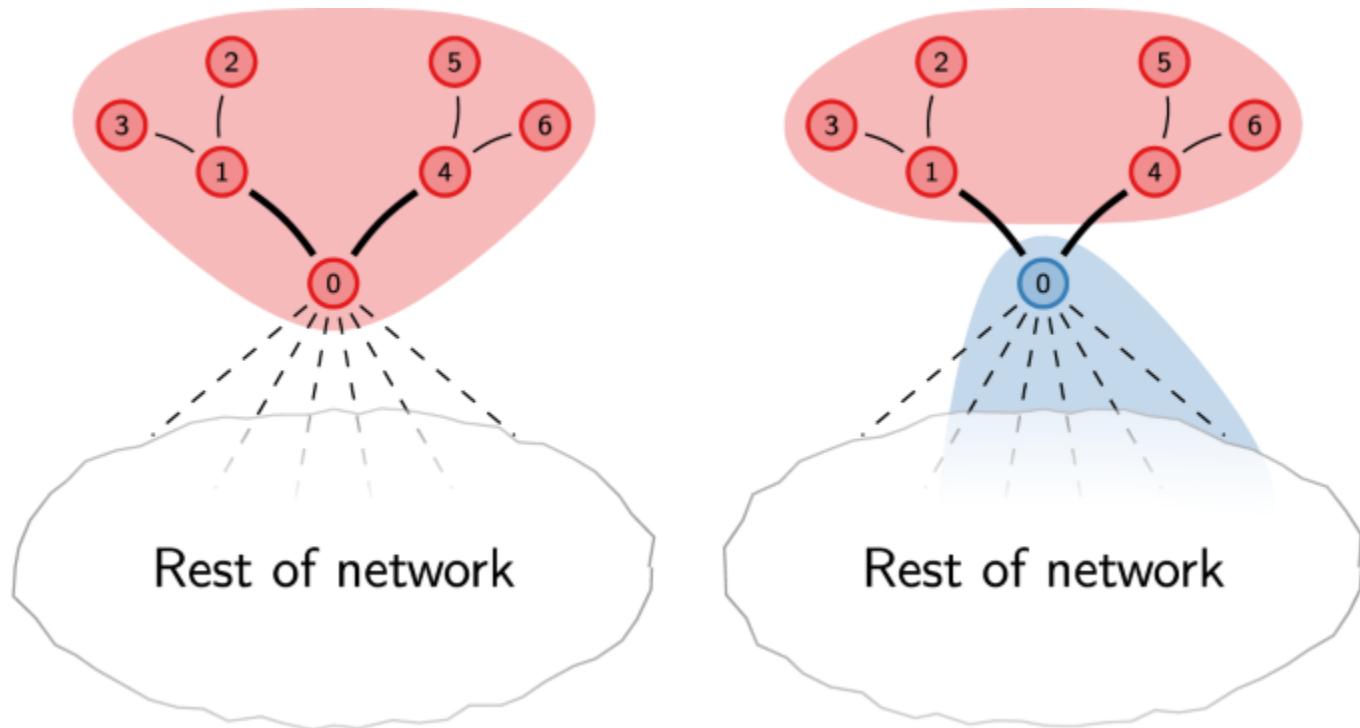
Iterate until no more changes



Graph-based clustering

Louvain algorithm

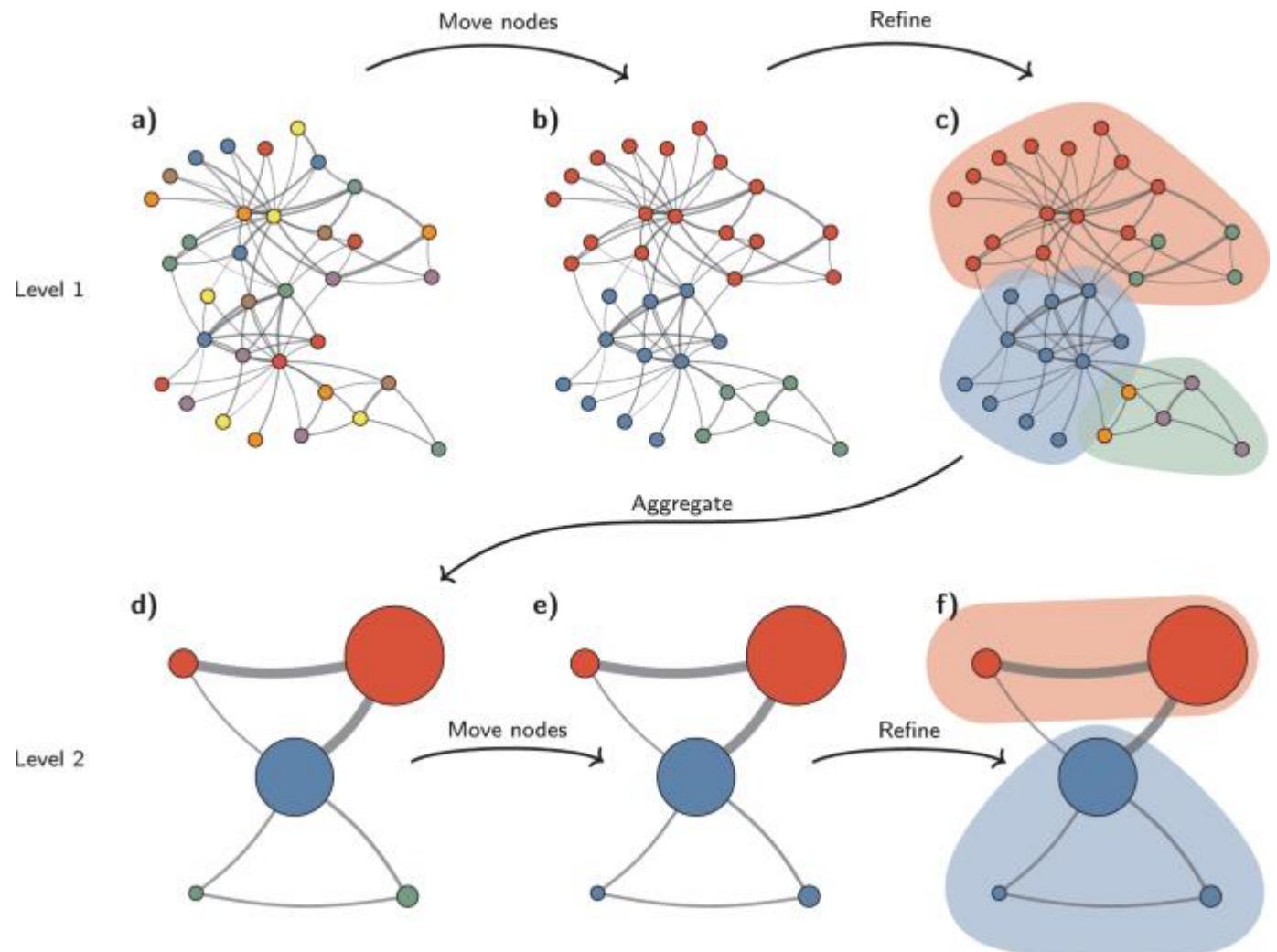
- During the ‘moving step’, nodes can become internally disconnected
- Nodes 1-6 still locally optimal assigned



Graph-based clustering

Leiden algorithm

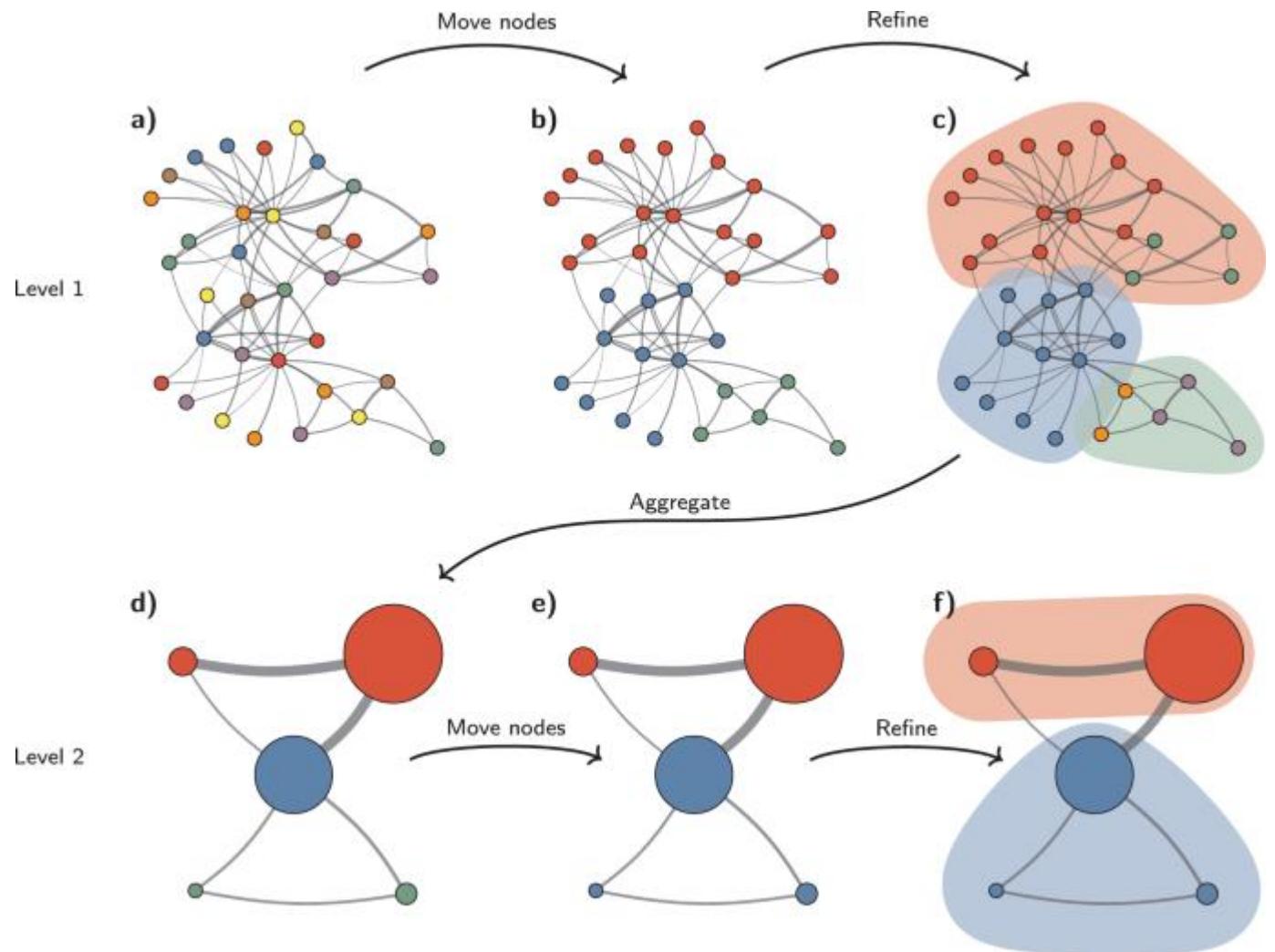
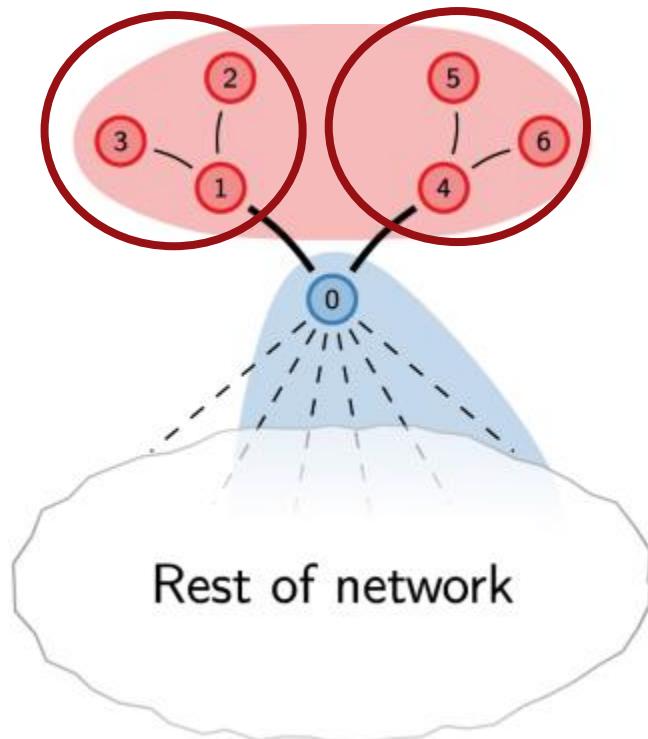
- Solution: add refinement step



Graph-based clustering

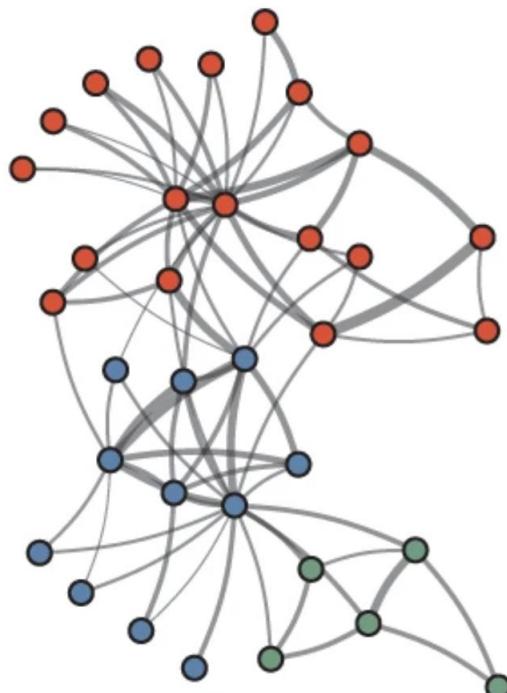
Leiden algorithm

- Solution: add refinement step



Graph-based clustering

Overview

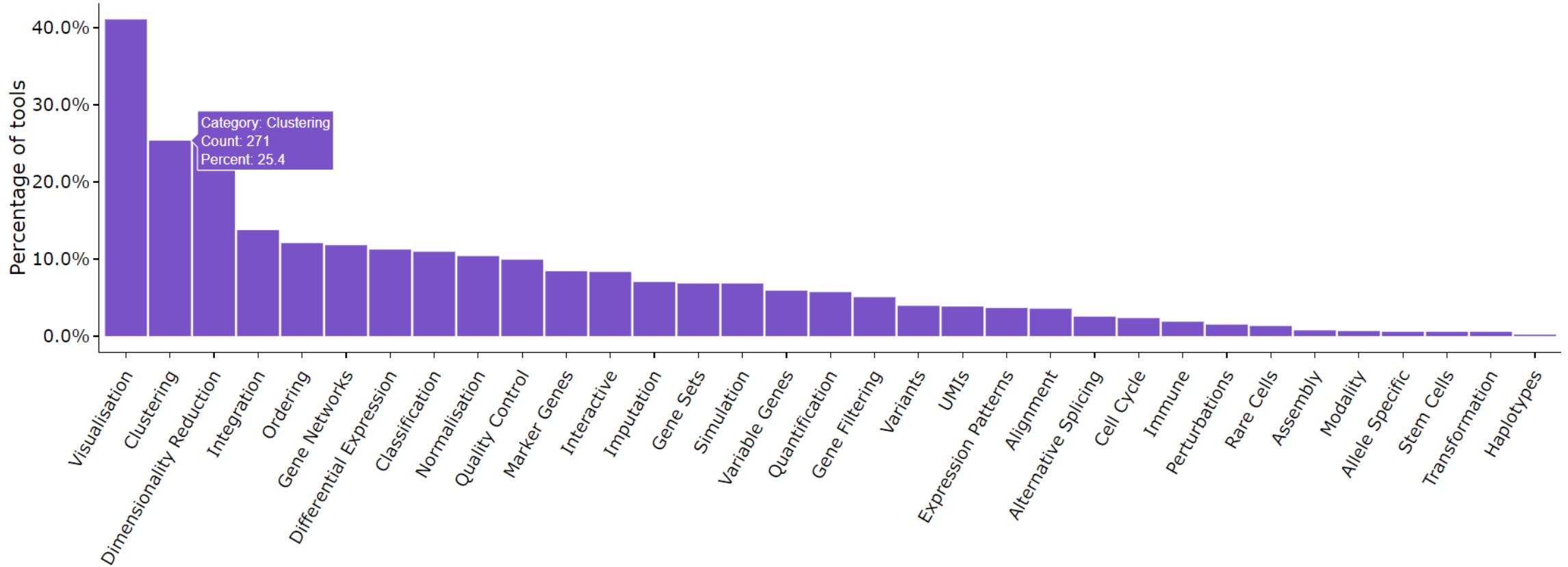


- Graph-based clustering
 - Number of neighbors when constructing the graph
 - Resolution parameters
- Resolution
 - High → less clusters
 - Low → more clusters
- Number of clusters
 - Determined using resolution parameter
 - Validate clustering!

Outline

- Introduction to clustering
 - Hierarchical clustering
 - k -Means clustering
 - Graph-based clustering
- scRNA-seq clustering
- Annotating clusters
- Cluster validation

scRNA-seq tools



scRNA-seq clustering methods

Name	Year	Method type	Strengths	Limitations
scanpy ⁴	2018	PCA+graph-based	Very scalable	May not be accurate for small data sets
Seurat (latest) ³	2016			
PhenoGraph ³²	2015			
SC3 (REF. ²²)	2017	PCA+k-means	High accuracy through consensus, provides estimation of k	High complexity, not scalable
SIMLR ²⁴	2017	Data-driven dimensionality reduction+k-means	Concurrent training of the distance metric improves sensitivity in noisy data sets	Adjusting the distance metric to make cells fit the clusters may artificially inflate quality measures
CIDR ²⁵	2017	PCA+hierarchical	Implicitly imputes dropouts when calculating distances	
GiniClust ⁷⁵	2016	DBSCAN	Sensitive to rare cell types	Not effective for the detection of large clusters
pcaReduce ²⁷	2016	PCA+k-means+hierarchical	Provides hierarchy of solutions	Very stochastic, does not provide a stable result
Tasic et al. ²⁸	2016	PCA+hierarchical	Cross validation used to perform fuzzy clustering	High complexity, no software package available
TSCAN ⁴¹	2016	PCA+Gaussian mixture model	Combines clustering and pseudotime analysis	Assumes clusters follow multivariate normal distribution
mpath ⁴⁵	2016	Hierarchical	Combines clustering and pseudotime analysis	Uses empirically defined thresholds and a priori knowledge
BackSPIN ²⁶	2015	Biclustering (hierarchical)	Multiple rounds of feature selection improve clustering resolution	Tends to over-partition the data
RaceID ²³ , RaceID2 (REF. ¹¹⁵), RaceID3	2015	k-Means	Detects rare cell types, provides estimation of k	Performs poorly when there are no rare cell types
SINCERA ⁵	2015	Hierarchical	Method is intuitively easy to understand	Simple hierarchical clustering is used, may not be appropriate for very noisy data
SNN-Clip ⁸⁰	2015	Graph-based	Provides estimation of k	High complexity, not scalable

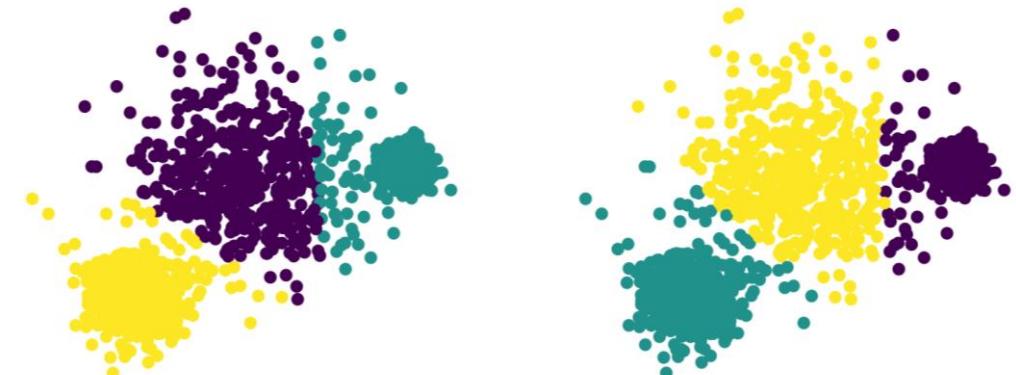
Comparing different cluster labels

Adjusted Rand Index (ARI)

Measure of the similarity between two data clusterings

Given a set S of n elements, and two groupings or partitions of these elements

$$X = \{X_1, X_2, \dots, X_r\}, \quad Y = \{Y_1, Y_2, \dots, Y_s\}$$

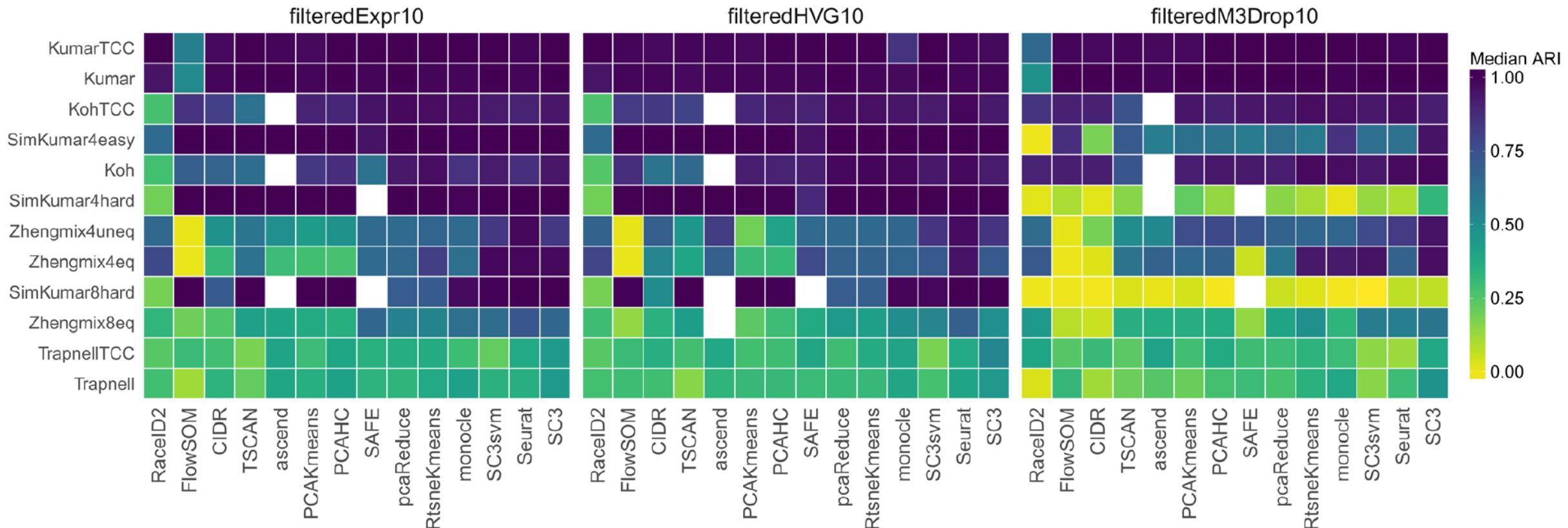


$X \setminus Y$	Y_1	Y_2	\dots	Y_s	Sums
X_1	n_{11}	n_{12}	\dots	n_{1s}	a_1
X_2	n_{21}	n_{22}	\dots	n_{2s}	a_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
X_r	n_{r1}	n_{r2}	\dots	n_{rs}	a_r
Sums	b_1	b_2	\dots	b_s	

Adjusted Index
 \widehat{ARI}

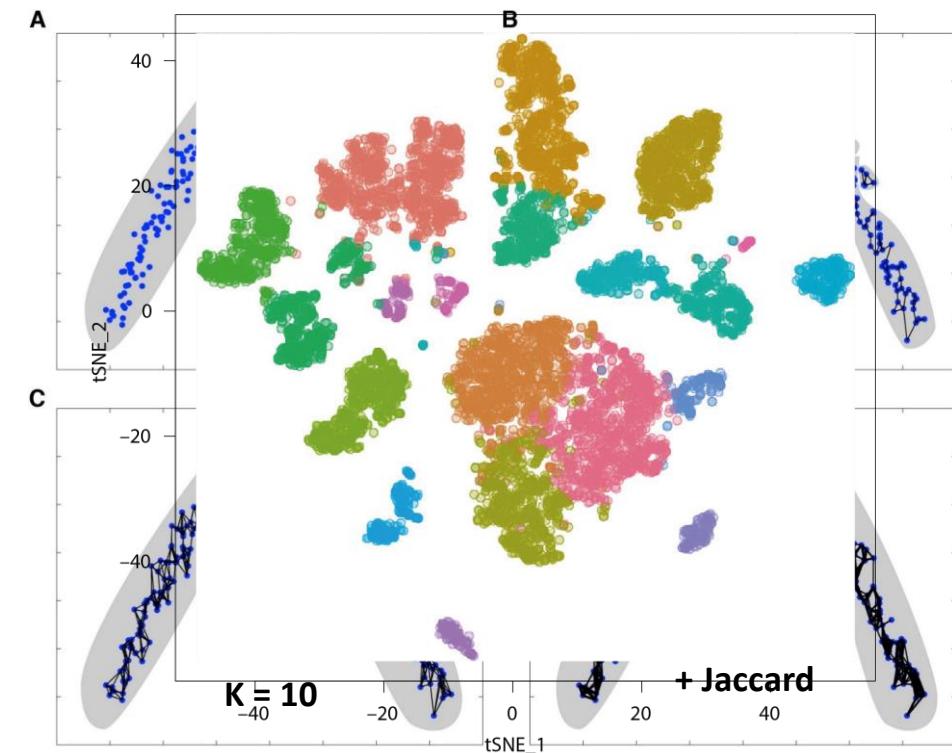
$$\widehat{ARI} = \frac{\overbrace{\sum_{ij} \binom{n_{ij}}{2}}^{\text{Index}} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}{\underbrace{\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}]}_{\text{Max Index}} - \underbrace{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}_{\text{Expected Index}}}$$

Benchmarking scRNA-seq clustering methods



Standard clustering approach

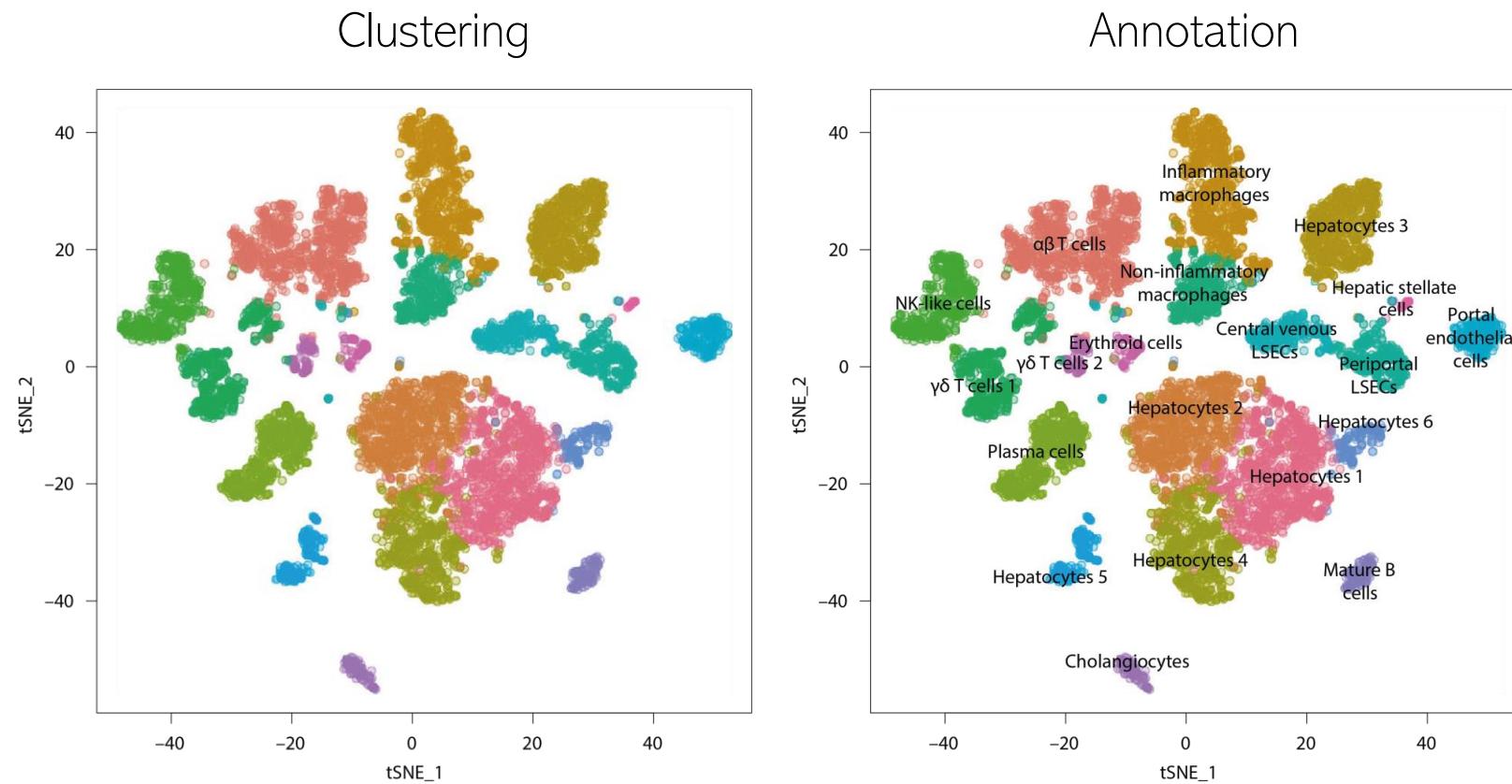
1. Select highly variable genes (~1000-5000 genes)
2. Reduce dimensions using PCA (~30-50 dimensions)
3. Construct kNN graph (~15-20 neighbors)
4. Louvain/Leiden community detection



Outline

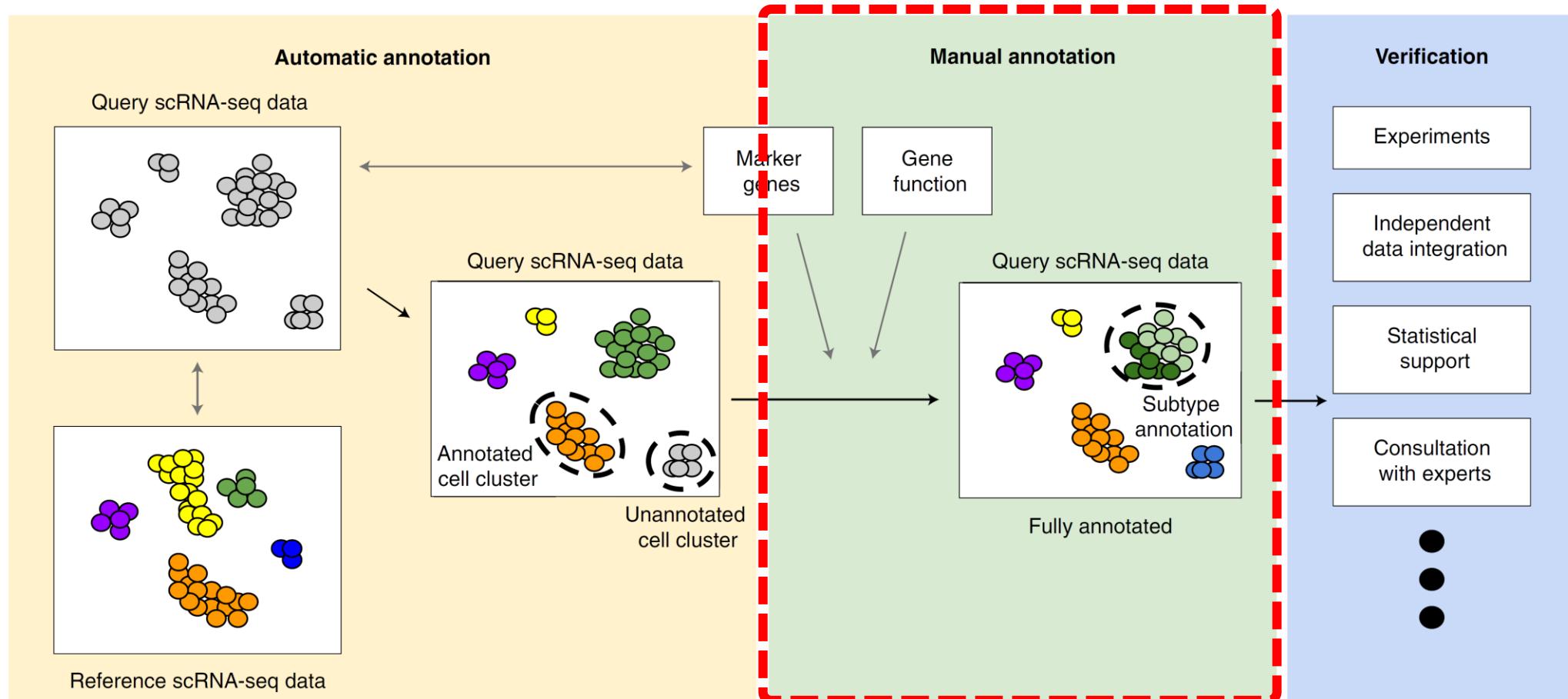
- Introduction to clustering
 - Hierarchical clustering
 - k -Means clustering
 - Graph-based clustering
- scRNA-seq clustering
- **Annotating clusters**
- Cluster validation

From clusters to annotations



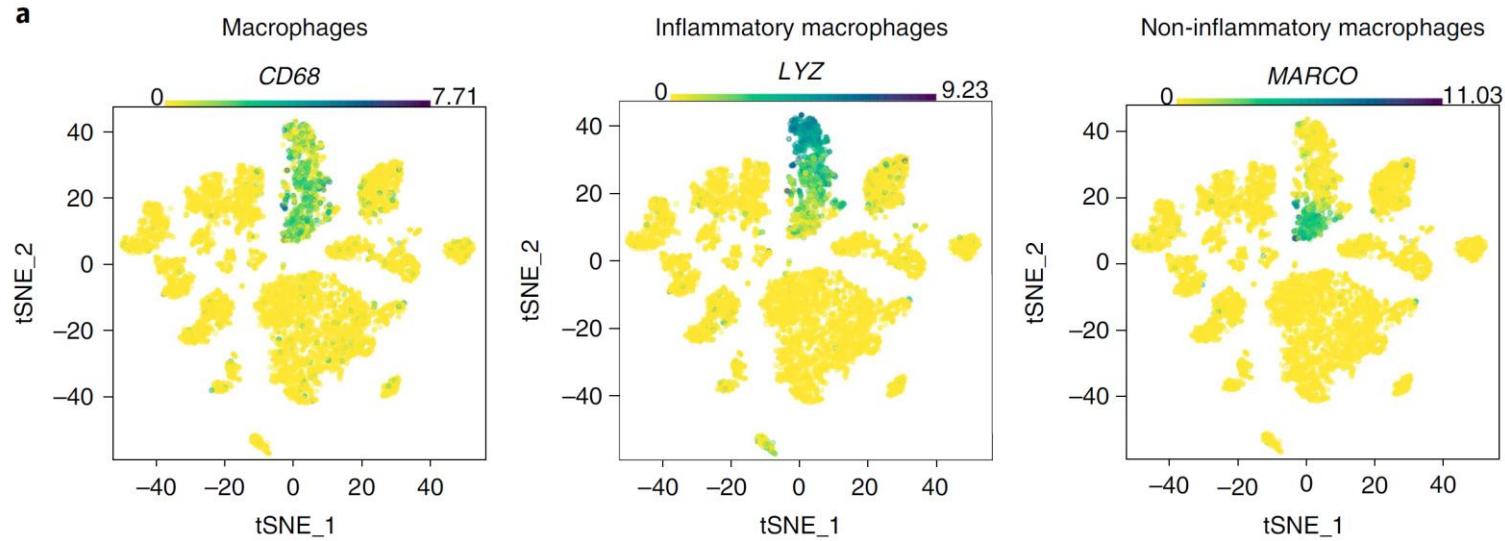


Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods

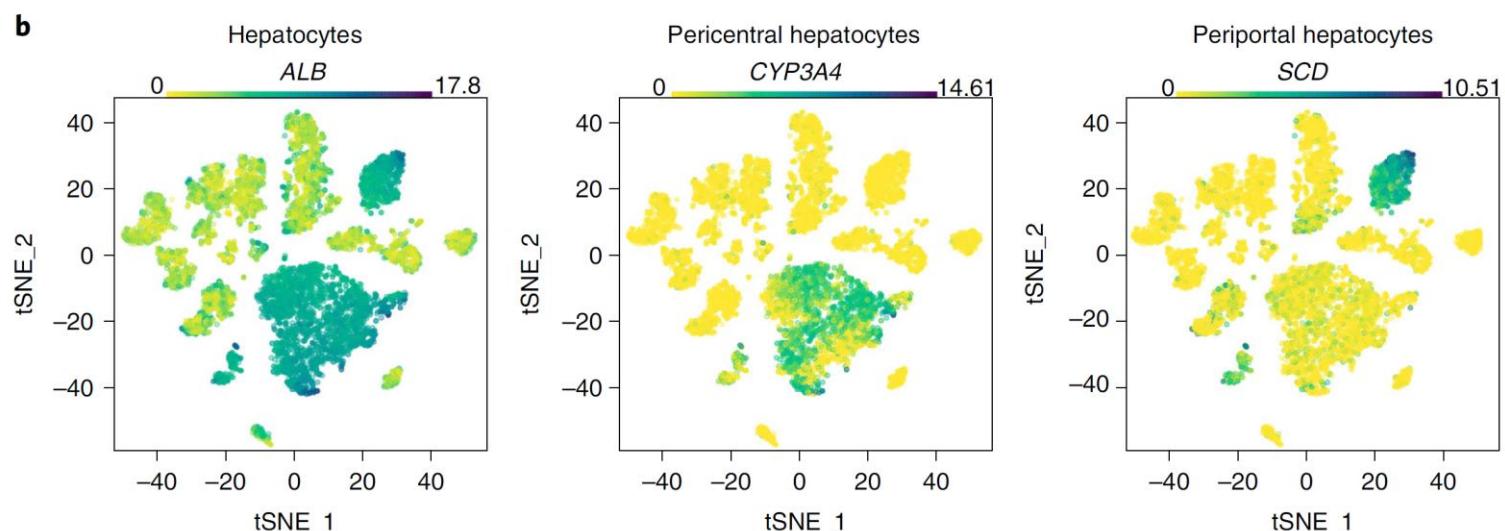


Gene expression overlay

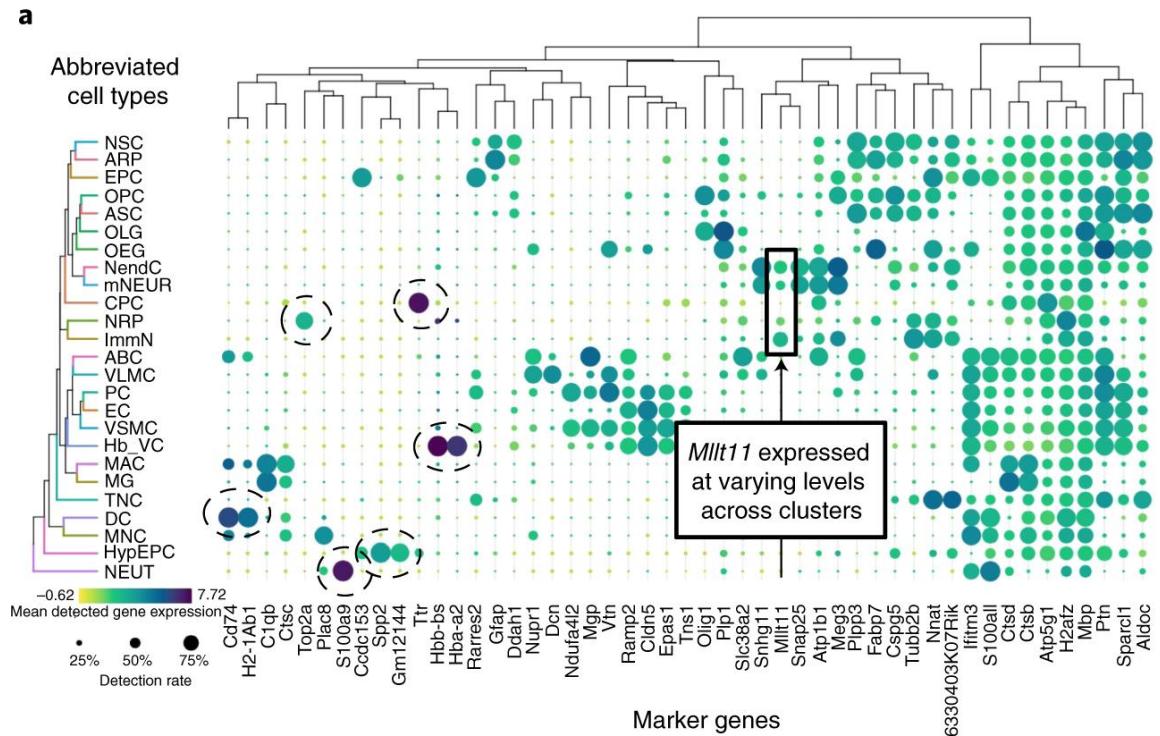
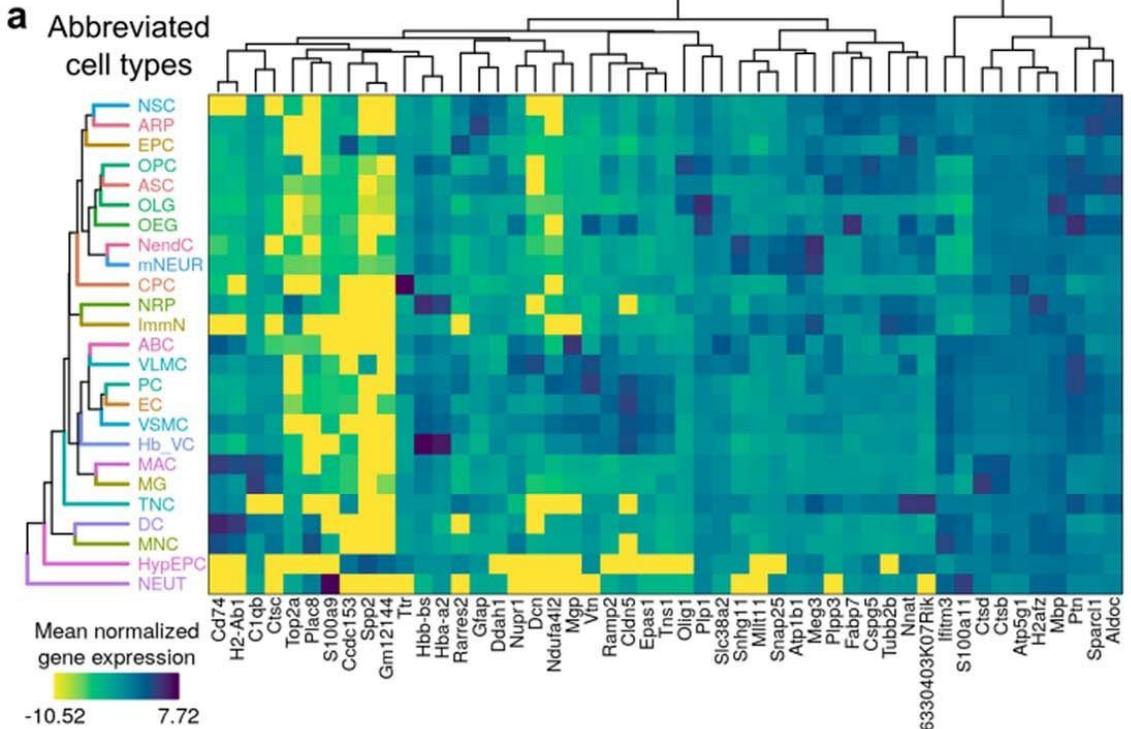
Easy



Challenging

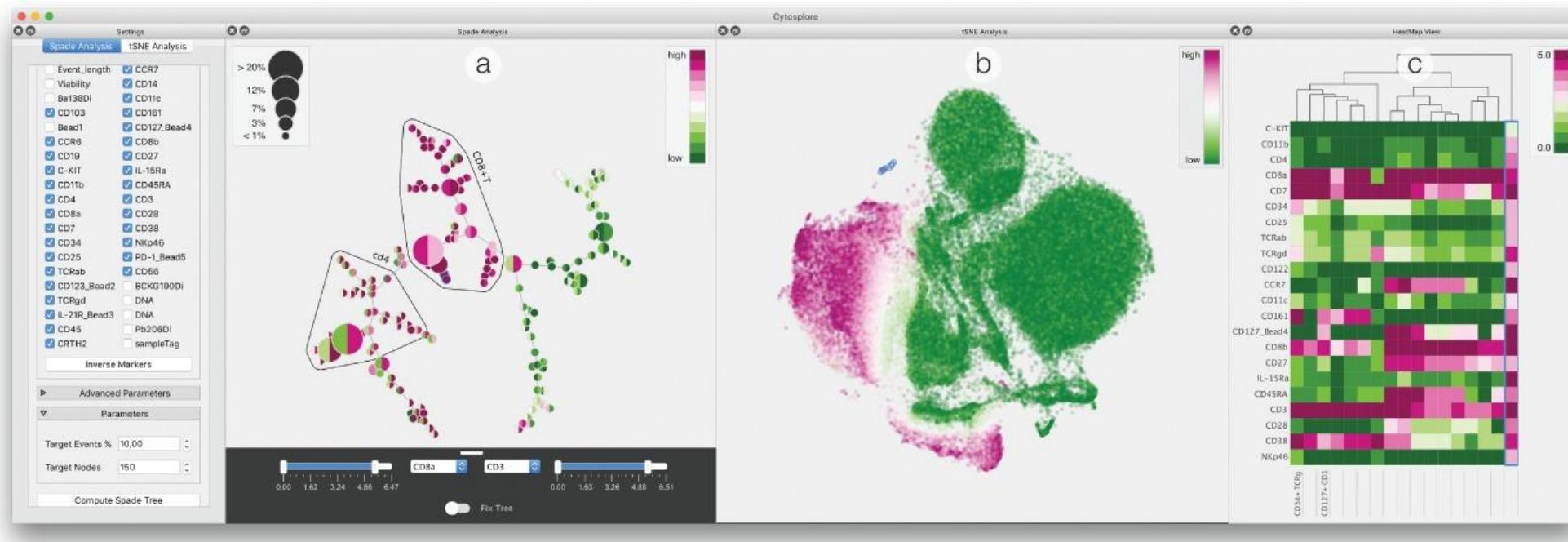


Alternatively: heatmaps & dot plots



Interactive visualization is important

- *Interactive* tools: Cytosplore, Loupe, cellxgene, ...
- *Iterative* visualization: Seurat, scanpy, ...



Where do we get these marker genes?

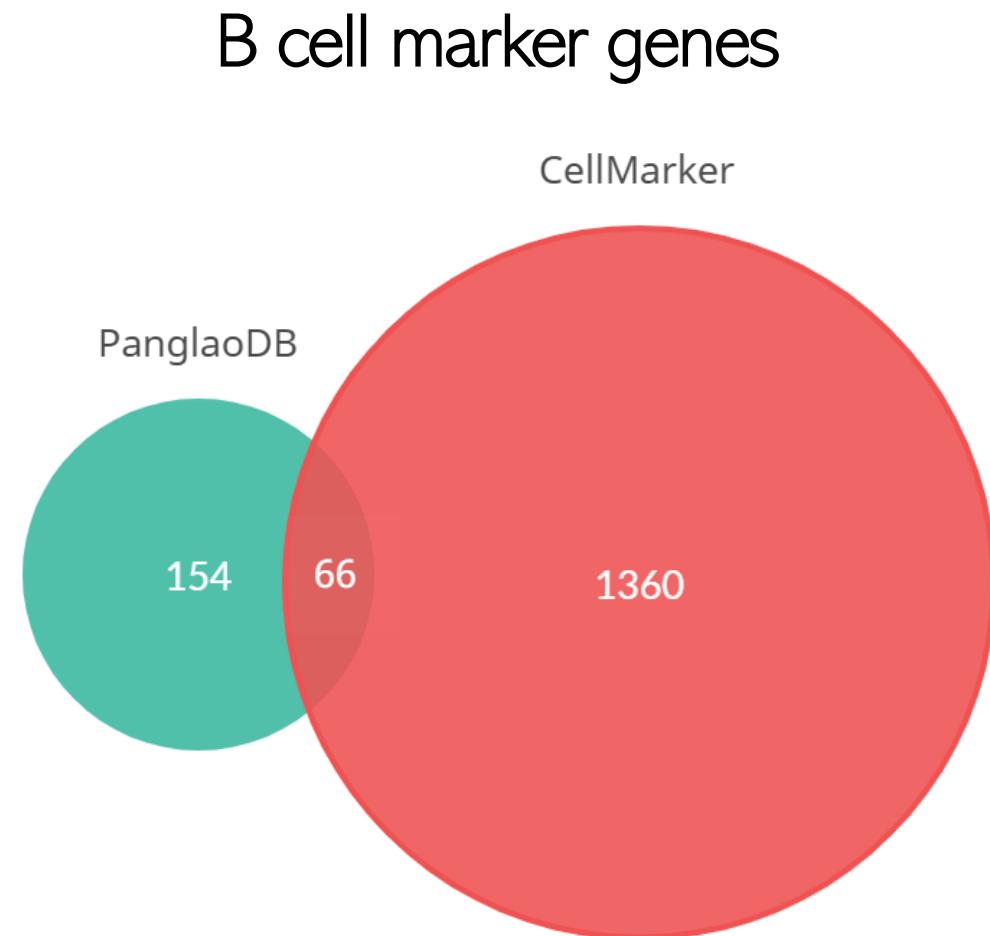
Ideally: from a single cell atlas from a relevant organism, organ and disease context

- “Expert knowledge”
- Literature
- Other scRNA-seq data
- Marker databases: PangaloDB, CellMarker,...

Challenges:

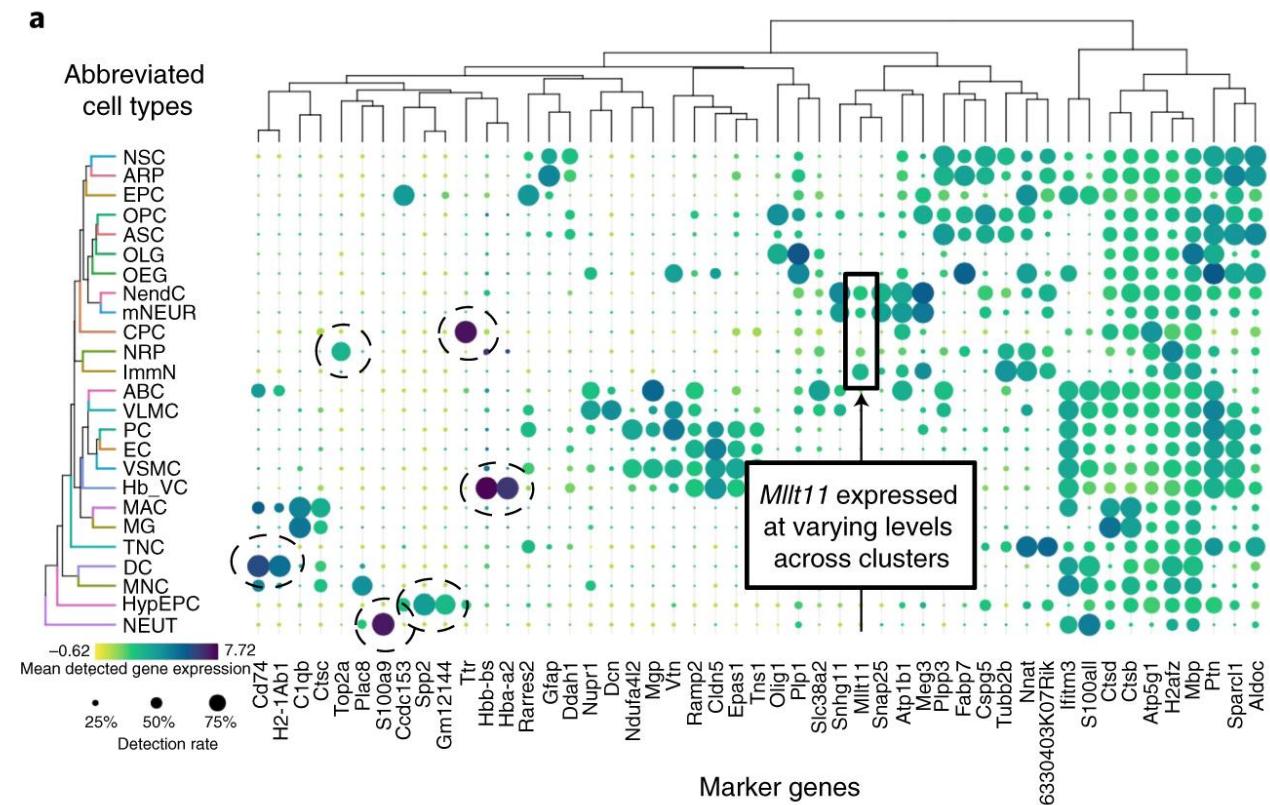
- Few well-known markers
- Some well-known markers may not be as specific as expected

Where do we get these marker genes?



What if I don't have that many markers

- Identify “novel” markers by computing differential expression between a cluster and all other cells or between pairs of clusters
- Manually research markers to find functional information that may help identify the cell type



Complicating factors

1. Clusters that express markers of more than one cell type
 - Doublets?
 - Likely small, higher than average genes and UMIs per cell
 - Doublet detection tools: Scrublet, DoubletFinder, scds
2. Ambient RNA
 - RNA derived from one or more cell types that are sensitive to tissue dissociation
 - Markers of the contaminating cell types may be spread to all other cell clusters
 - Ambient RNA correction tools: SoupX, CellBlender

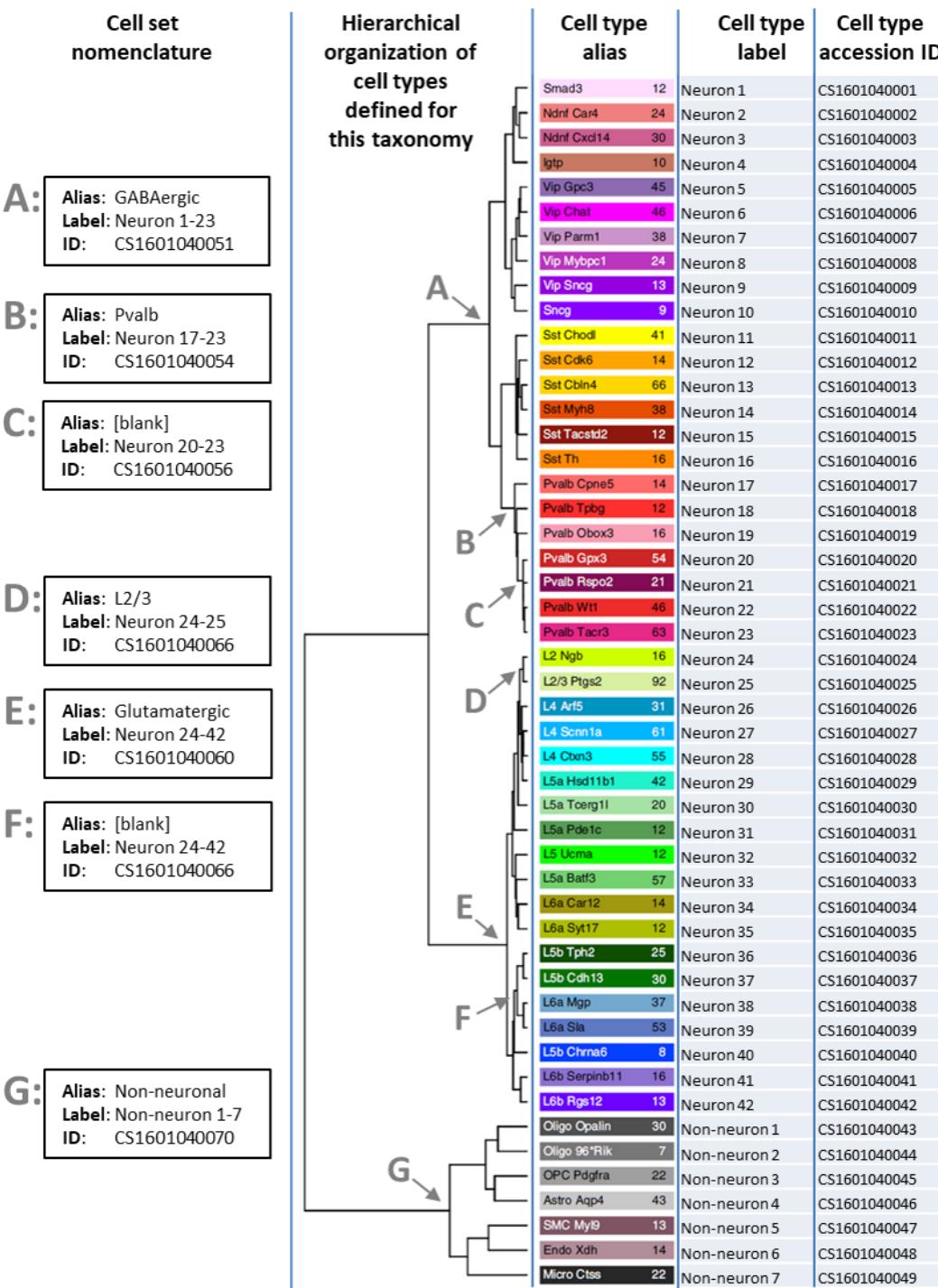
Watch out not to remove rare “interesting” cells!

Annotation verification

1. Using *independent* data (e.g. fluorescence in situ hybridization)
2. Multi-modal single cell data
 - SNVs & CNVs
 - TCR/BCR
 - scRNA-seq+scATAC (mRNA + accessibility)
 - CITE-seq (surface proteins + mRNA)

Nomenclature

- How should we name cells?



Summary

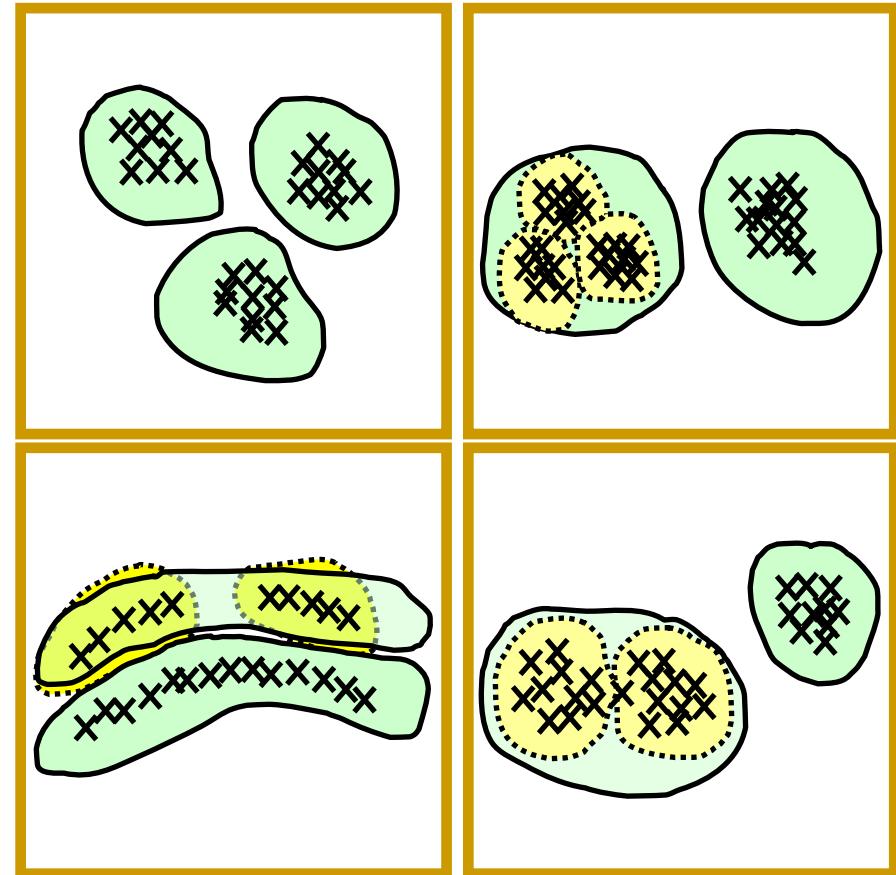
- Start by identifying major well-known cell types (clearly defined, discrete cell clusters)
- Sometimes, it is useful to split the data into broad subsets (e.g., immune, endothelial and tumor) and analyze each separately
- Cell subtypes or poorly defined clusters are challenging
- Manual annotations heavily relies on marker genes and expert knowledge

Outline

- Introduction to clustering
 - Hierarchical clustering
 - k -Means clustering
 - Graph-based clustering
- scRNA-seq clustering
- Annotating clusters
- Cluster validation

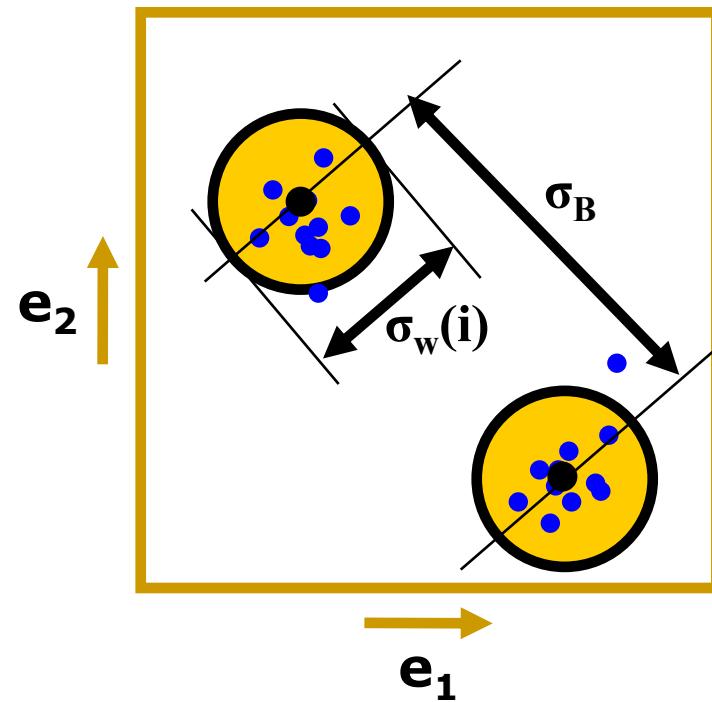
Clustering is subjective!

- Principle choices
 - Similarity measure
 - Algorithm
- Different choice leads to different results
 - Subjectivity becomes reality
- Cluster process
 - Validate, interpret (generate hypothesis), repeat steps



Cluster criteria

- Silhouette score
 - Goal: optimize cohesion within a cluster and separation between clusters
 - Seek: clustering that maximizes SI



Silhouette score

1. Mean distance between i and all other points in cluster C_i

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

2. Mean nearest cluster distance of i

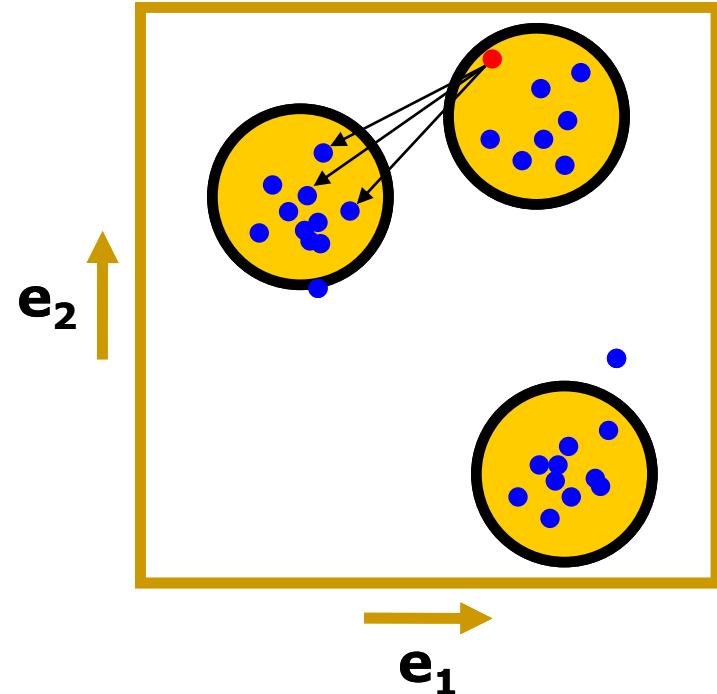
$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

3. Silhouette score for i

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

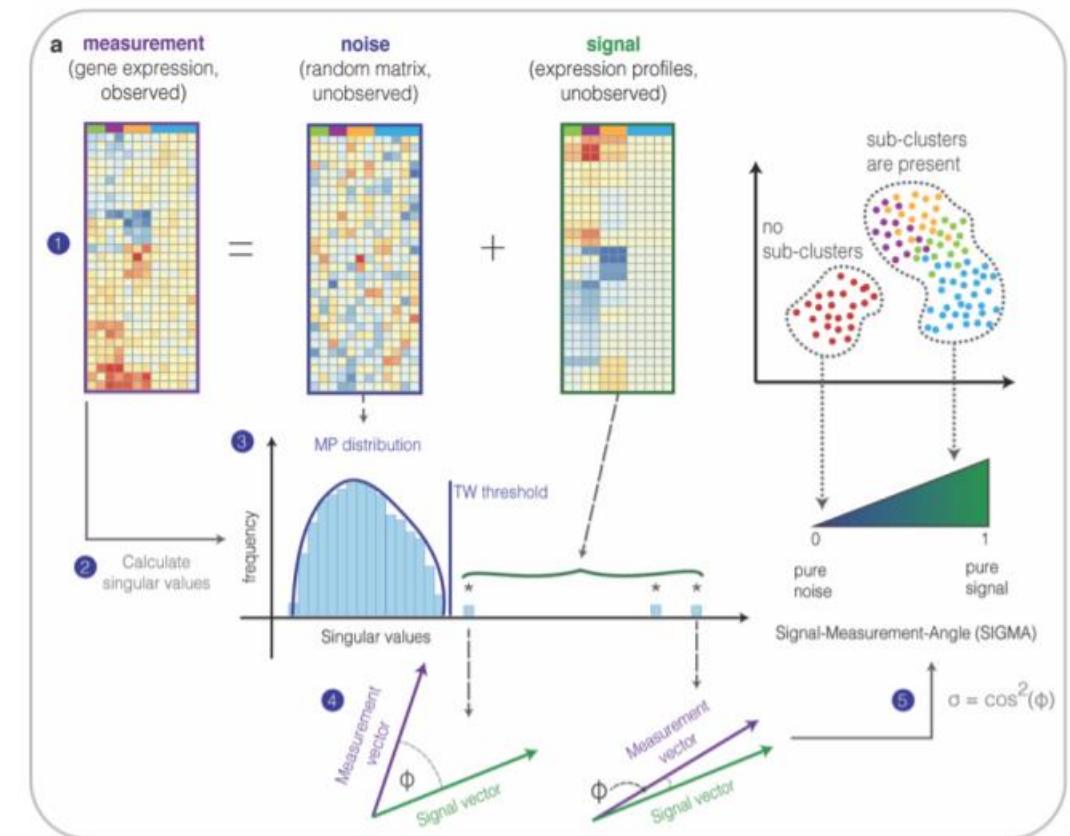
4. Total silhouette score

$$SI = \frac{1}{N} \sum s(i)$$

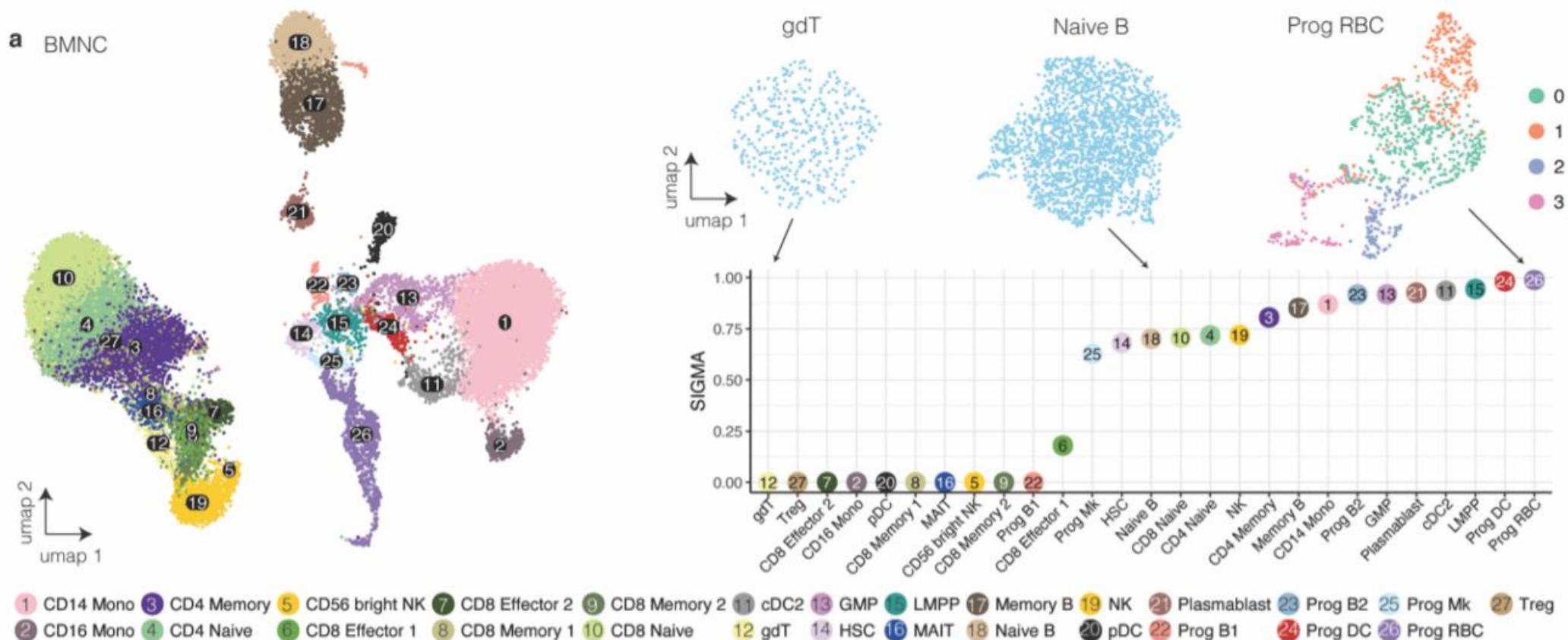


SIGMA: a clusterability measure

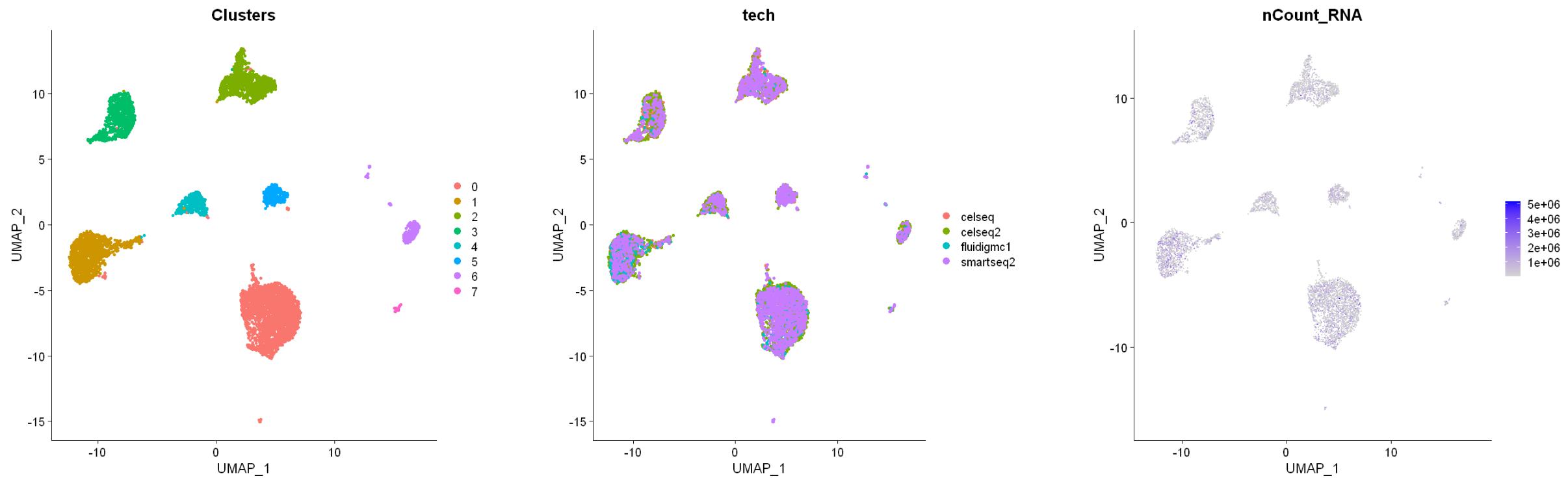
- Variability within a cluster due to
 - Presence of subpopulations
 - Or random noise?



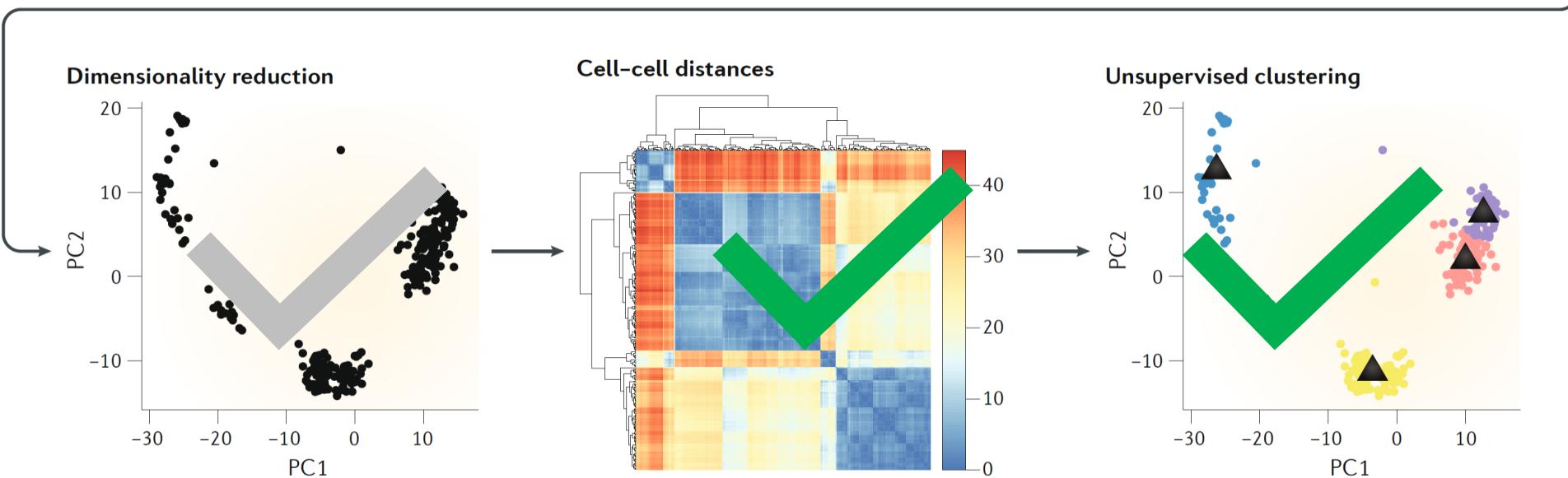
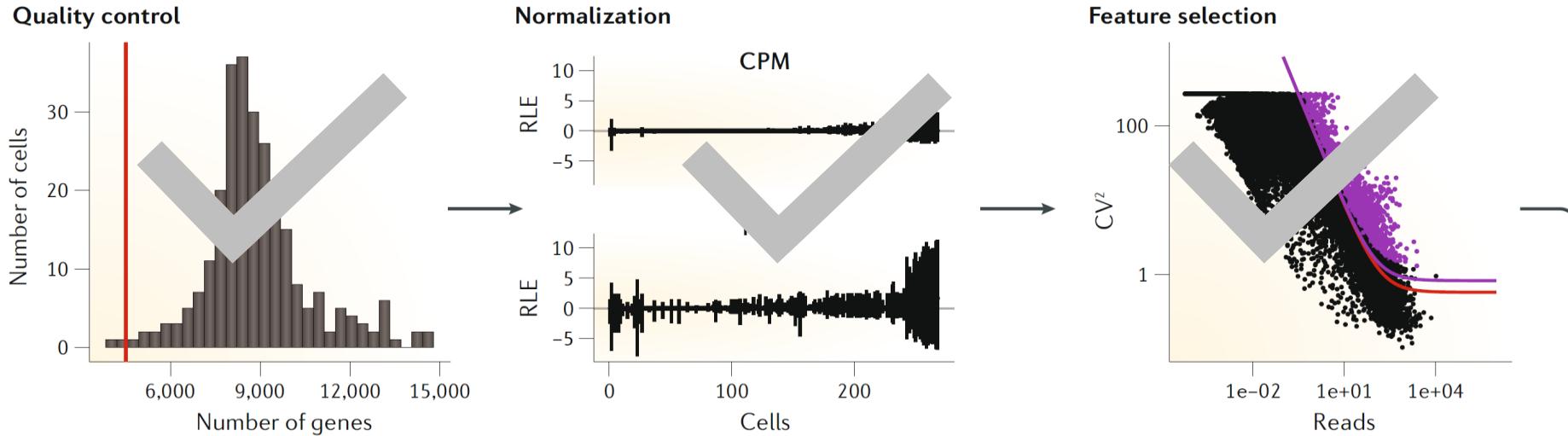
SIGMA: a clusterability measure



Always check QC data



Summary



Summary & challenges

- **Subjectivity:** what is a cell type
 - *Different parameters will give different results*
 - *Validation is important*
- **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 $\sim 10^6$
 - *Computational efficiency*
 - *Visual exploration, crowding problem*

Clustering practical

- Hierarchical clustering: distances and linkage methods
 - k -Means
 - Graph-based clustering
-
- Annotating clusters

Resources

- Kiselev et al. "Challenges in unsupervised clustering of single-cell RNA-seq data"
<https://doi.org/10.1038/s41576-018-0088-9>
- Duò et al. "A systematic performance evaluation of clustering methods for single-cell RNA-seq data"
<https://doi.org/10.12688/f1000research.15666.2>
- Orchestrating Single-Cell Analysis with Bioconductor
<https://osca.bioconductor.org/>
- Hemberg single cell course: Analysis of single cell RNA-seq data
<https://scrnaseq-course.cog.sanger.ac.uk/website/index.html>
- Slides Åsa Björklund (NBIS, SciLifeLab)
<https://github.com/NBISweden/workshop-scRNAseq/tree/master/slides2019>
- Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods
<https://doi.org/10.1038/s41596-021-00534-0>