Institut de Recherche en Cancérologie
I R C Montpellier

HAU901I

Single-cell transcriptomics

Basic Algorithms

Pr Jacques Colinge

UM/Inserm - IRCM - ICM

Inserm
Institut national de la santé et de la recherche médicale

UNIVERSITÉ DE MONTPELLIER

FACULTÉ de MÉDECINE

ICM
Institut régional du Cancer Montpellier | Val d'Aurelle

Filtering data

+ (Chromium data loading will be seen at the exercises.)

+ Three main filtering steps (order can change):
  - Eliminate cells with too low a total UMI or too low a number of distinct genes
  - Eliminate likely dead cells, typically based on high mitochondrial content
  - Eliminate doublets

+ Optional: remove very high UMI/gene counts that may be due to amplification bias

+ As in bulk RNA-seq, reduce the count matrix to those genes that are at least expressed at a minimum level in some samples (criteria may vary, no single rule, but common sense should suffice)

Normalization

+ The total number of UMIs or gene counts varies significantly between cells even after low complexity filtering
  - Technical reason: different sequencing depth
  - Biological reason: different cell types contain more or less RNA or express more or less different genes

+ Elementary solution:
  - Divide each column of the matrix by its total and multiply by some arbitrary large number, say 10,000
  - Log-transform data to avoid excessive weight of highly expressed genes

+ More advanced statistical procedures have been proposed

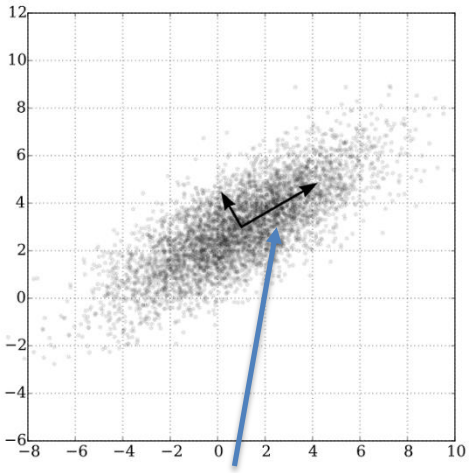
Clustering and 2D-projection

+ After a clean and normalized count matrix has been obtained, the usual next steps are to identify cell populations by clustering and to visualize data through a 2-dimensional projection

+ Although some clustering procedures could use the 2D-projection, it is usually not the case: clustering is performed independently
+ A common «exception» to this rule is to help the clustering algorithm by reducing data dimensionality before the latter is applied:
  - Reduce data dimension from >1000 to 50 or 30 for instance with principal component analysis (PCA)
  - Apply a clustering method on the reduced data
  - Compute a 2D-projection from the original data or the PCA-reduced data
  - Identify the clusters by colors in the 2D view

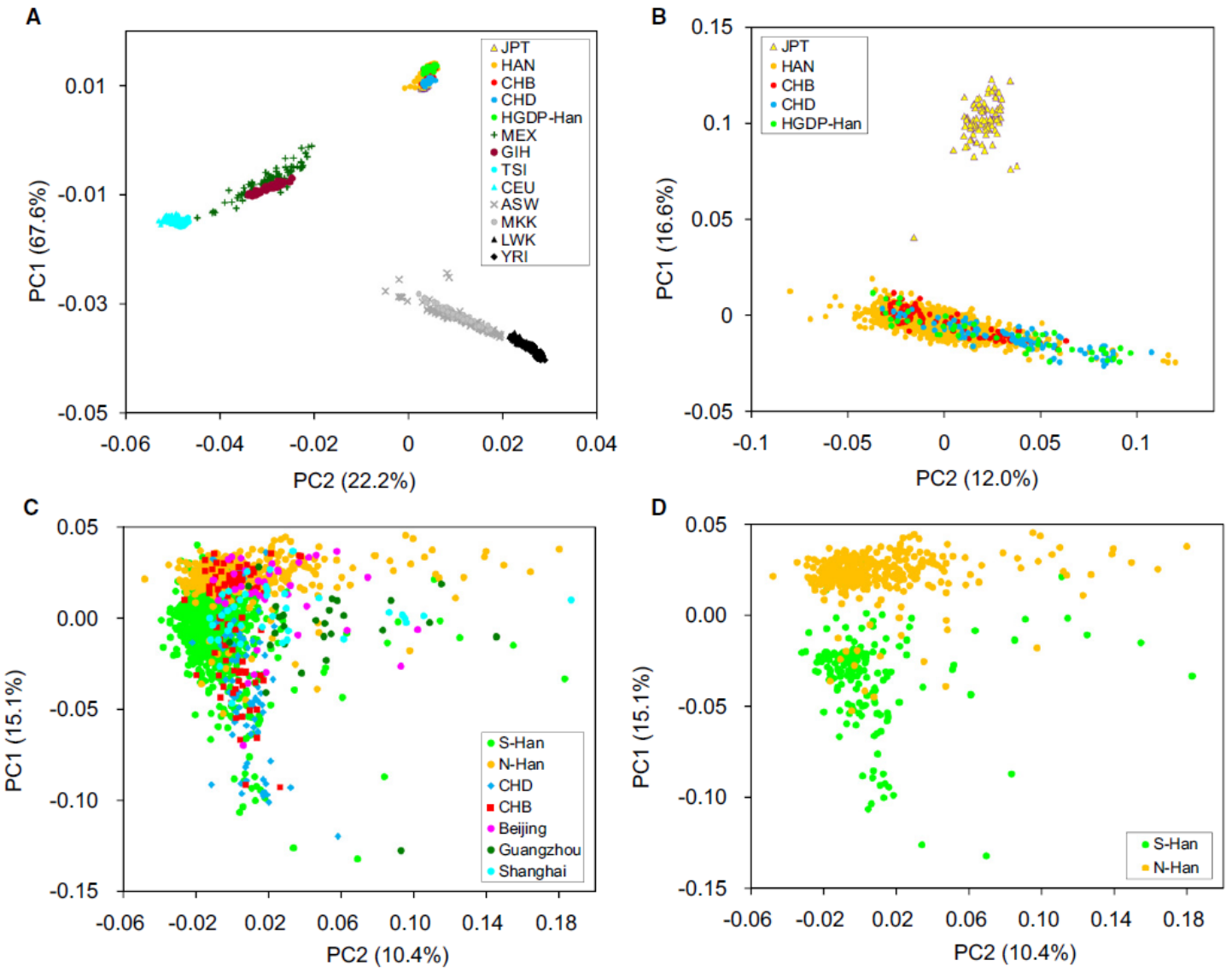
Principal component analysis (PCA)

+ A classical method is PCA that projects n-dimensional vectors on to a k-dimensional space (k<n) by preserving maximum variance in the data (limited loss of information in the projection)
+ A first direction in space (the first principal component) is found along which the projected data retain the largest variance
+ A second, orthogonal direction is then searched (second principal component)
+ Repeated until the k^{th} principal component
+ Well defined and understood mathematically
+ Very good to reduce dimension, not always to obtain a good 2D-projection



2 first principal components
Source: WIkipedia

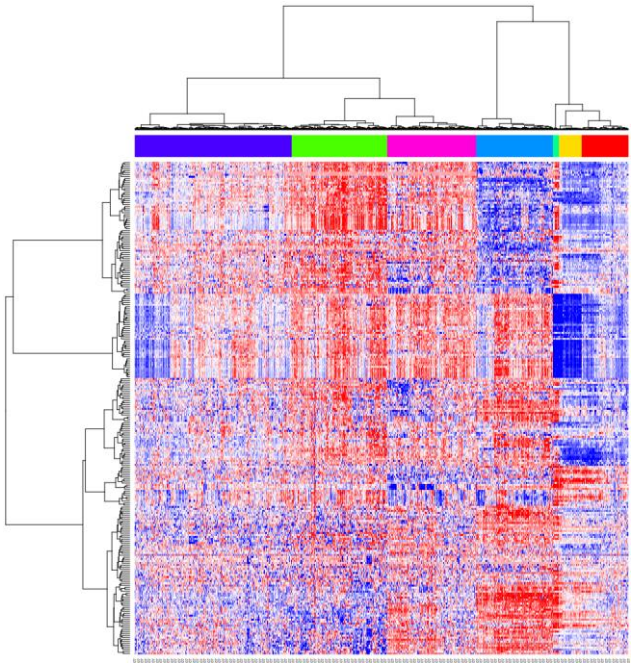
PCA on genotypes

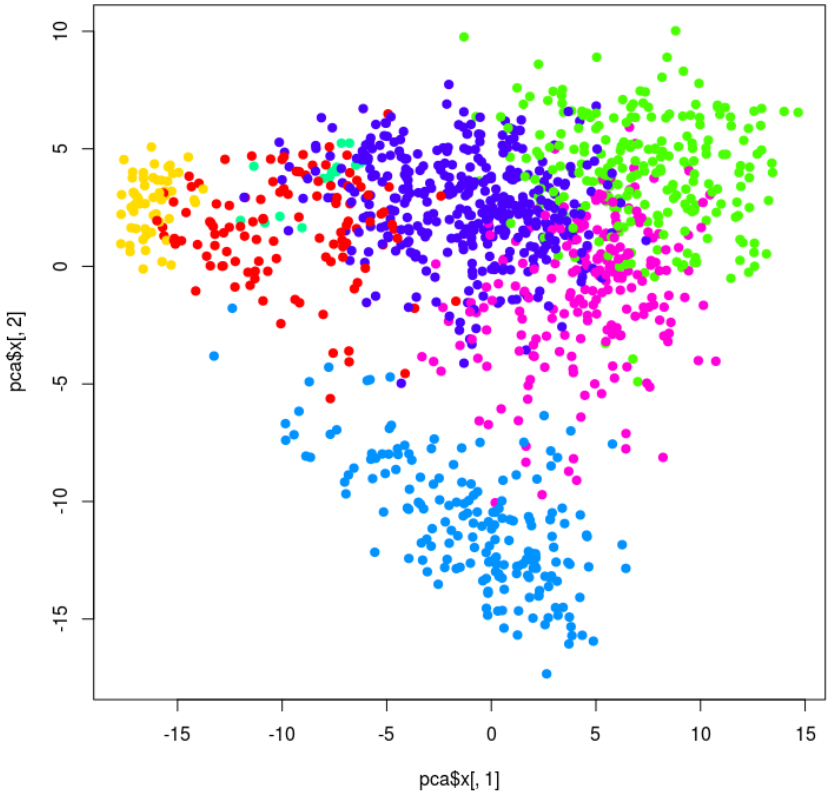


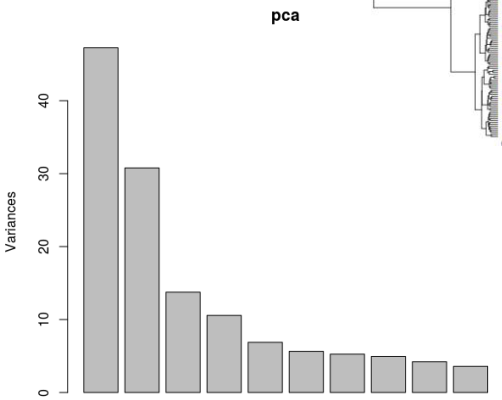
Source: Xu et al., AJHG, 2009

PCA on the BRCA transcriptomes

+ Transcriptomes of 113 normal adjacent and 1,094 cancer breast samples
+ Selection of the 246 most variable genes with clear expression levels



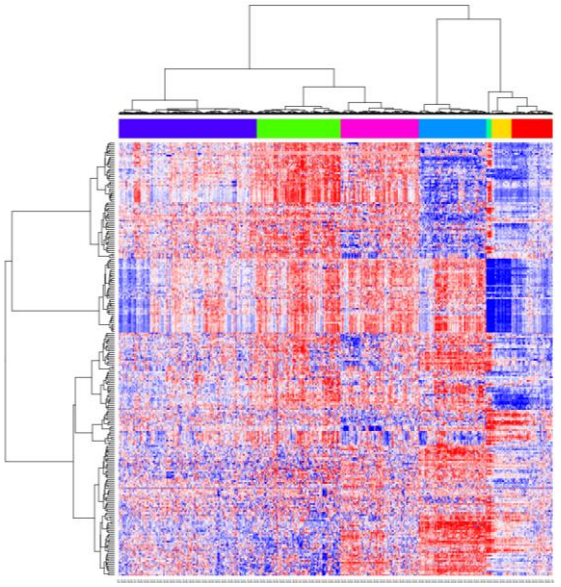


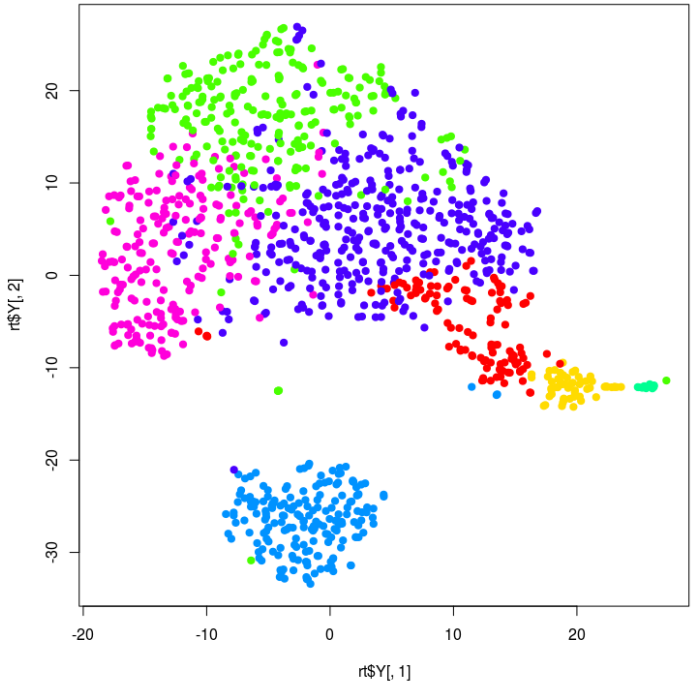


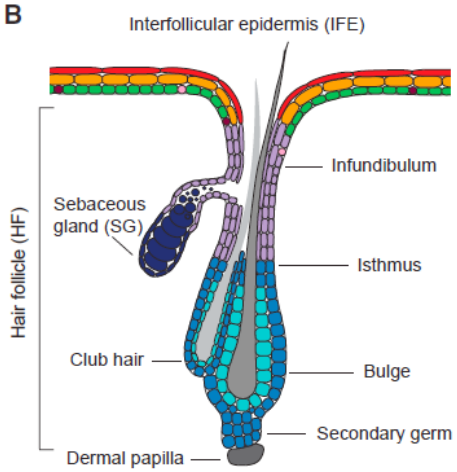
We see that projection onto the first two principal components captures an important part of the variance and preserves the clusters found with hierarchical clustering reasonably well

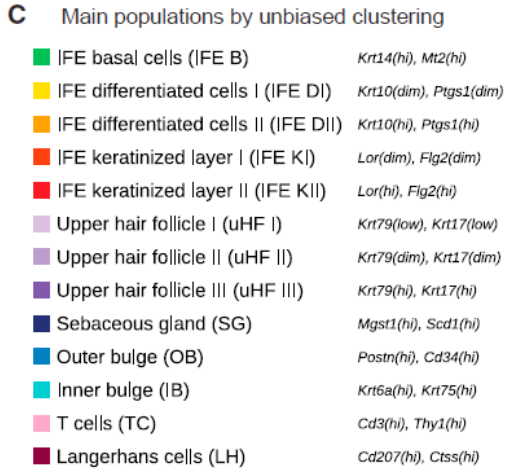
t-SNE

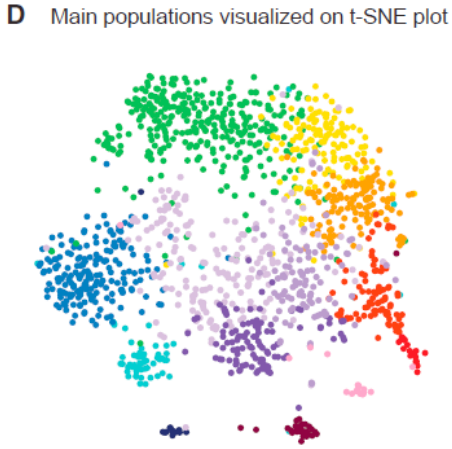
+ t-distributed stochastic neighbor embedding
+ t-SNE is a recent algorithm that outperforms PCA to generate a 2D view

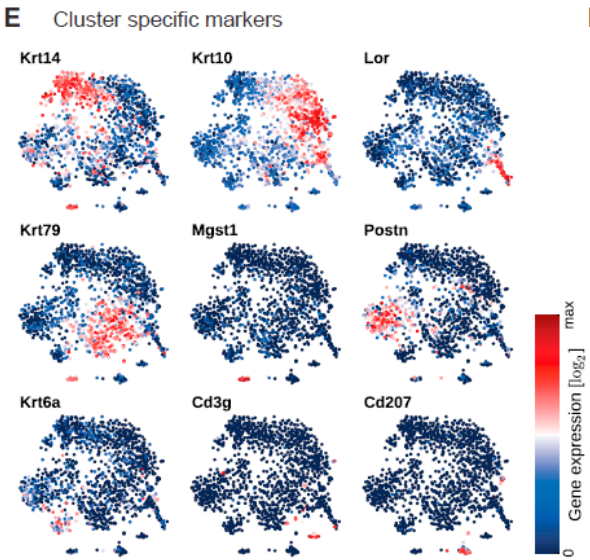


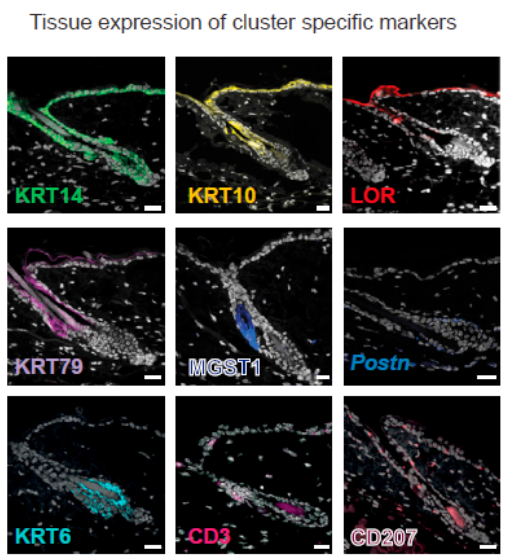


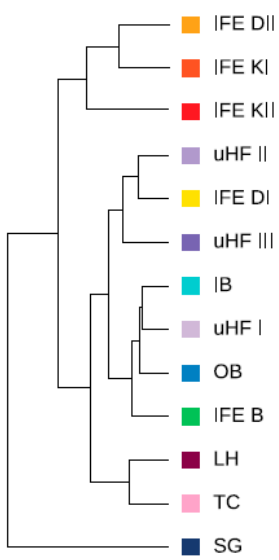












Joost, et al., Cell Rep, 2016

t-SNE principle

+ A parameter called perplexity (σ) is chosen by the user that controls how local/global the algorithm is
+ In the original space (high dimension), the similarity $p_{i,j}$ between each pair of data point x_i and x_j is computed
+ $p_{i,j}$ = similarity between data points i and j = Gaussian density at distance $\|x_i - x_j\|$ normalized by the total of such densities:

$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma^2\right)}{\sum_{k \neq l} \exp\left(-\|x_k - x_l\|^2/2\sigma^2\right)}$$

+ In the projection space (dimension 2 usually), projected data points are placed randomly and their similarity computed after the same principle but with a t distribution instead of a Gaussian:

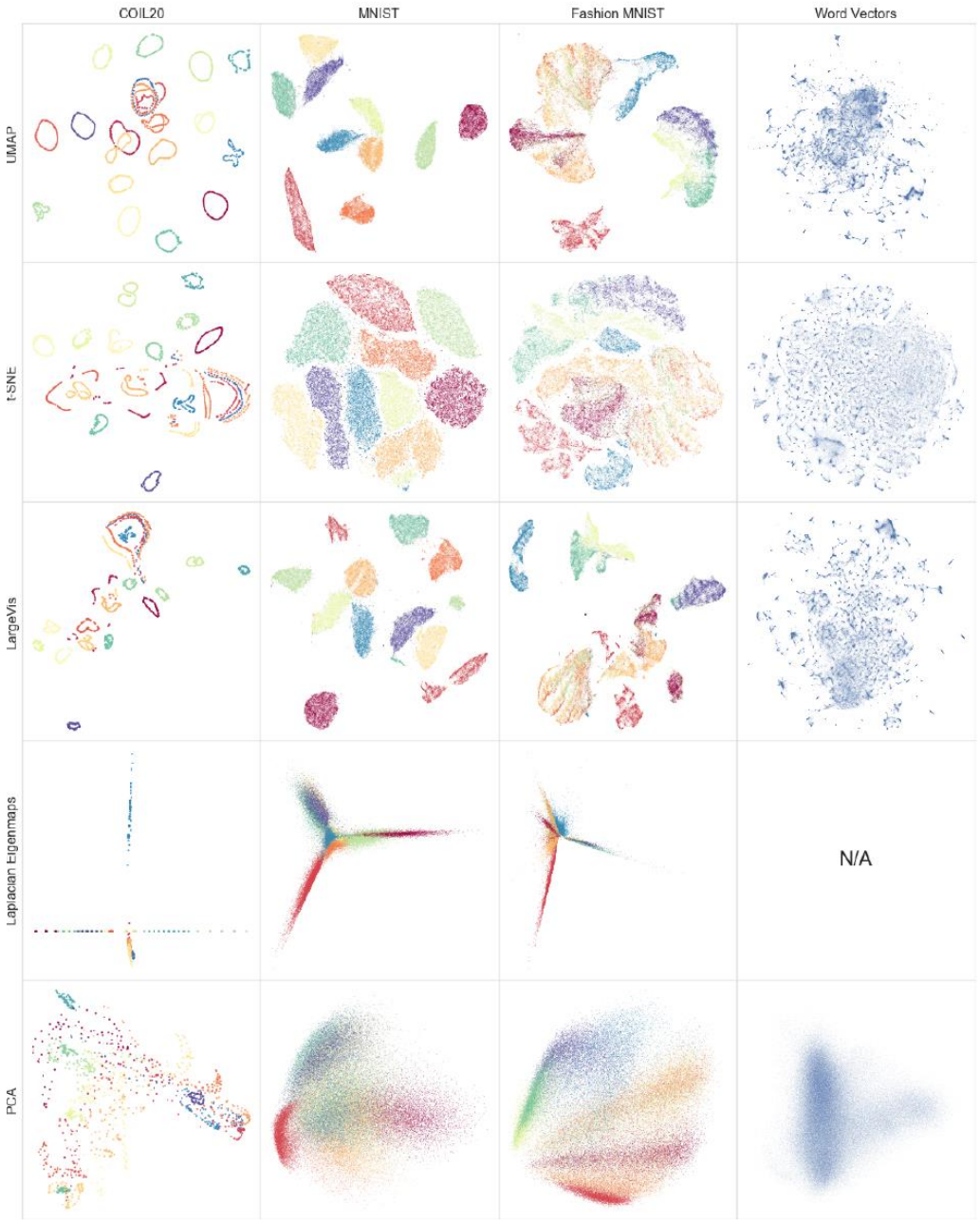
$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

+ The projected data points are moved by a gradient descent algorithm to minimize the Kullbak-Leibler divergence from the distribution $p_{i,j}$ to $q_{i,j}$

UMAP

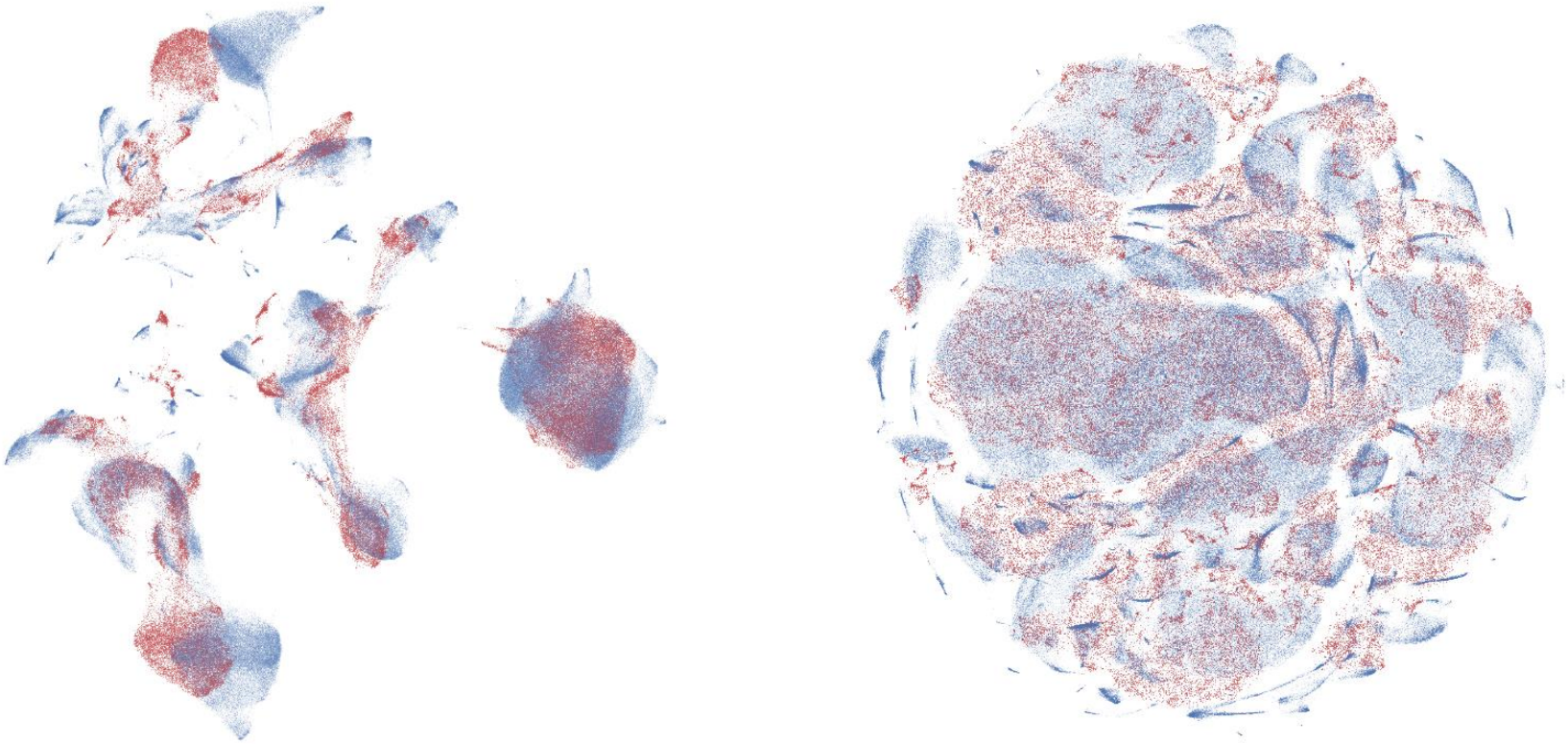
+ Uniform manifold approximation and projection
+ This complex algorithm relies on advanced geometry and topology to approximate a manifold by a low dimensional representation
+ A first step involves the construction of a graph based on simplices to connect data points
+ In a second step, this graph is projected to a lower-dimensional space
+ UMAP is usually faster than t-SNE and its strong theoretical background «guarantees» better preservation of data properties in the projection
+ Check https://pair-code.github.io/understanding-umap/ for a nice interactive presentation

Comparisons on artificial data sets



Source: McInnes, arXiv.org, 2020

A comparison on real data



(a) UMAP

(b) t-SNE

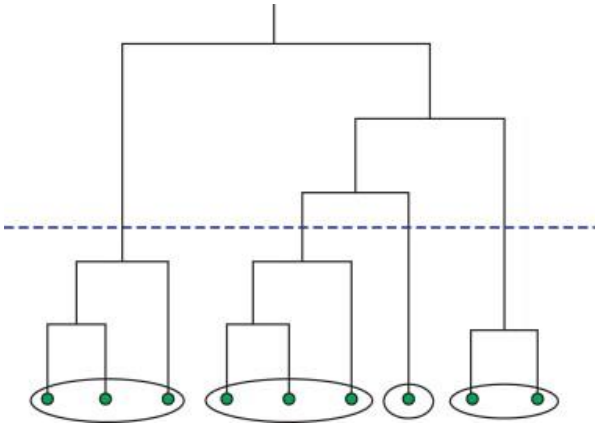
Flow cytometry data, red is 10% of the whole data set (in blue). UMAP can learn a reliable representation from 10% of the data only. Source: McInnes, arXiv.org, 2020.

Clustering

+ Clustering algorithms are meant to group similar objects together and to separate dissimilar objects

+ Obviously, how similar is defined has an impact on the obtained clusters
+ We usually use a notion of distance to measure similarity (close is similar)

+ Dendrograms are the graphical representation of a powerful way to cluster data called hierarchical clustering
+ Their topology depends on the distance used to compare objects and on the rule that decides when to link data points in the dendrogram



+ These notions are detailed in the next slides

Distances

+ Without loss of generality, each object is described by p features, which can be coded as real numbers
+ Each object is thus represented by a real-valued vector $x \in \mathbb{R}^p$
+ For instance, the objects can be biological samples and the features are the expression levels of the genes (transcriptomes)

+ There exists an infinite number of distances between $\mathbb{R}^p$ points
+ Common distances are provided by the Minkowski metric

$$d_k(x,y) = \sqrt[k]{\sum_{i=1}^{p} |x_i - y_i|^k},\ x,y \in \mathbb{R}^p, k \geq 1$$

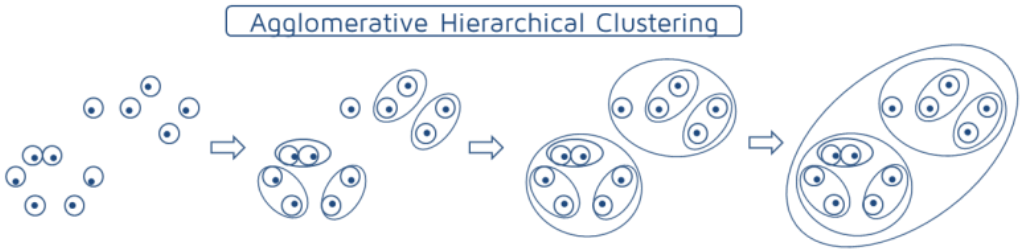
+ Familiar special cases are: Euclidean distance ($k = 2$), Manhattan distance ($k = 1$), and maximum distance ($k = \infty$)

+ An example of non-Minkowski distance is the Canberra distance

$$d(x,y) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

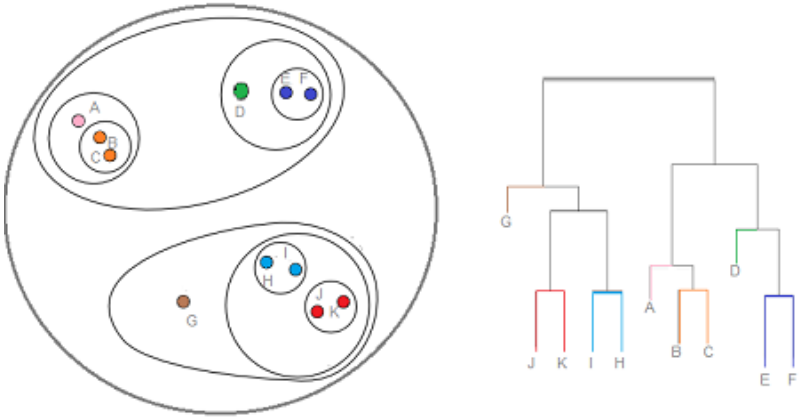
Hierarchical clustering

+ The principle of (agglomerative) hierarchical clustering is to assign each object to its own cluster and to proceed iteratively, merging the two closest clusters at each step
+ Let us assume that the feature vectors are 2-dimensional, then agglomerative hierarchical clustering is represented by the following schema:



Source: quantdare.com

+ The relationships between the embedded clusters can be represented by a tree (right) called a dendrogram



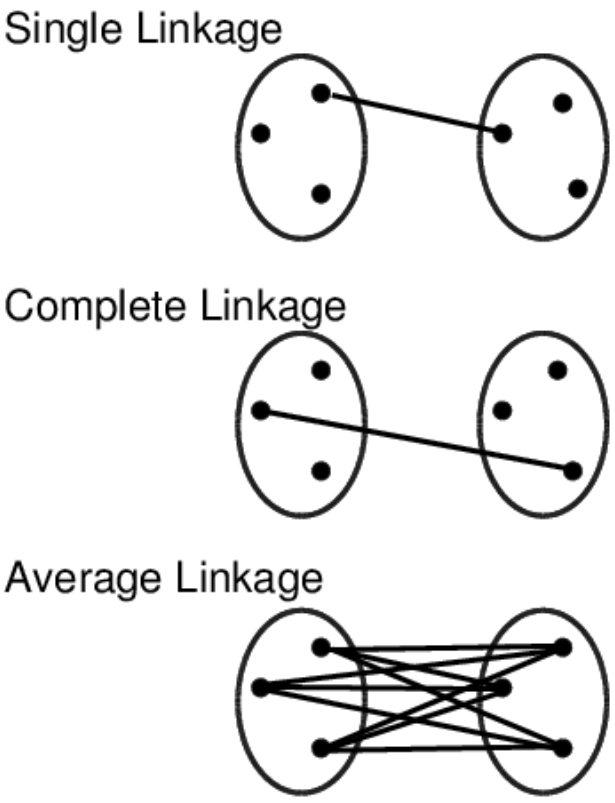
Source: Statistics How To

Rules to join/merge clusters

+ Single linkage defines the distance between two clusters as the smallest distance between all possible pairs
+ Complete linkage takes the largest distance
+ Average linkage (UPGMA) uses the average of all the distances between all the pairs

+ At every step, the two closest clusters are merged

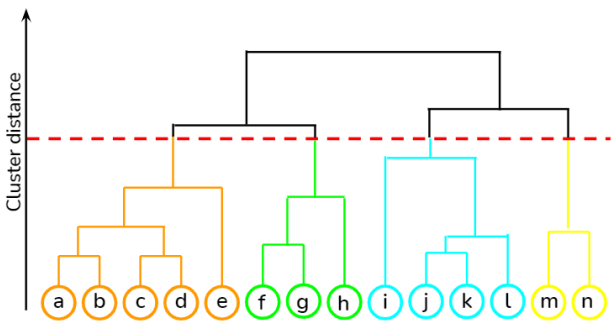
+ Ward's method tries to minimize the new clusters internal variance

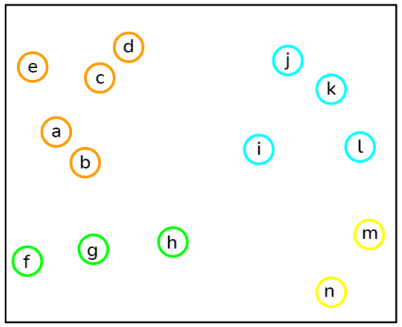


Source: Girke.bioinformatics.ucr.edu

Distances and cuts in a dendrogram

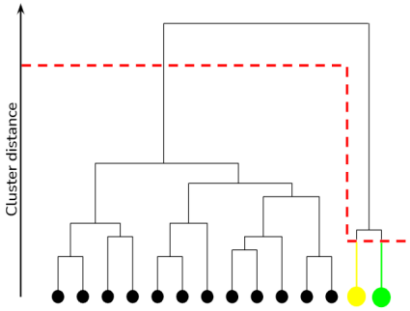
+ A dendrogram plot retains distance information on its vertical axis

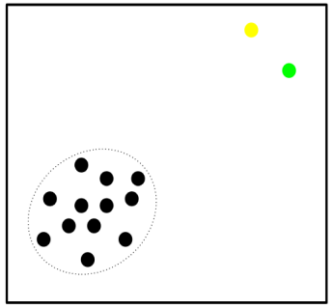




+ The sum of vertical distances in the dendrogram is equal to the distance between two objects

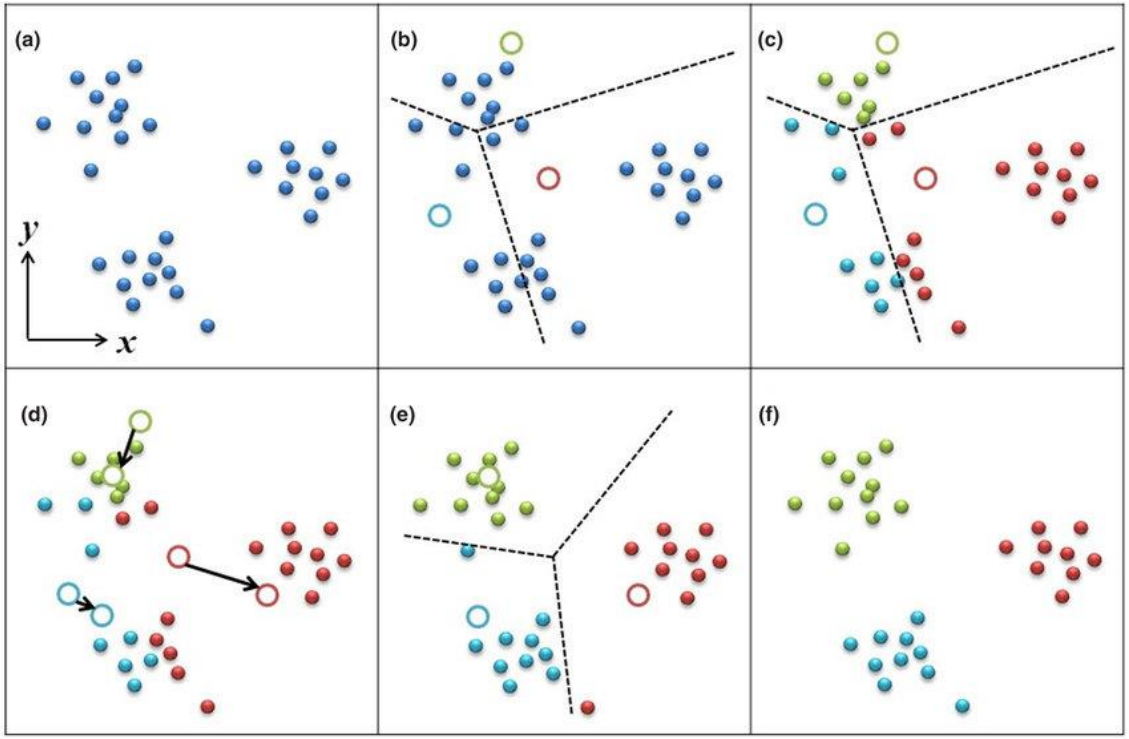
+ Cutting at a certain height defines clusters
+ There are many rules (silhouette plots, gap statistics, etc.) to choose the best height(s), none is universally reliable!





+ Check that cluster contents make sense and cluster are robust. For instance, compute clusters on 90% of the data and check 2 given objects cluster together most of the time or similar concepts

K-means

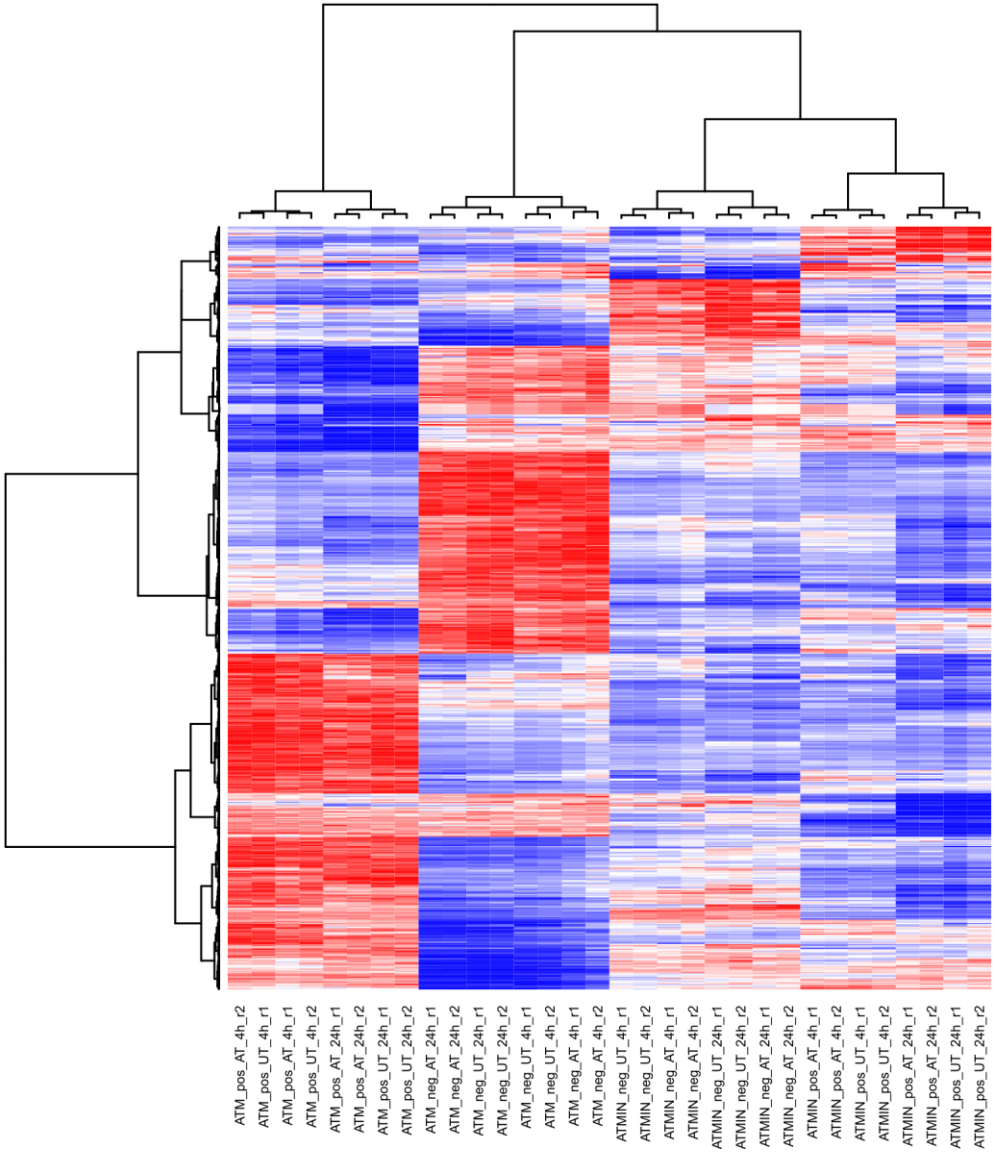


+ K-means is a old and widely used algorithm
+ The principle is to construct k prototypical data points, *i.e.*, proteomes or transcriptomes for us, and to cluster all the data points around these representative prototypes
+ The number of clusters k is decided in advance (there are criteria to compare different k values)
+ Algorithm: (a) N data points should be clustered; (b) k random prototypes are chosen; (c) each data point is assigned to the closest prototype; (d) prototypes are updated to be equal to the average of the data points in the clusters they represent; (e) new assignments to the closest prototype, etc. until nothing changes. (f) Final clustering
+ Very simple and effective algorithm

Figure source: Chen & Lai, Phys Rev E, 2018

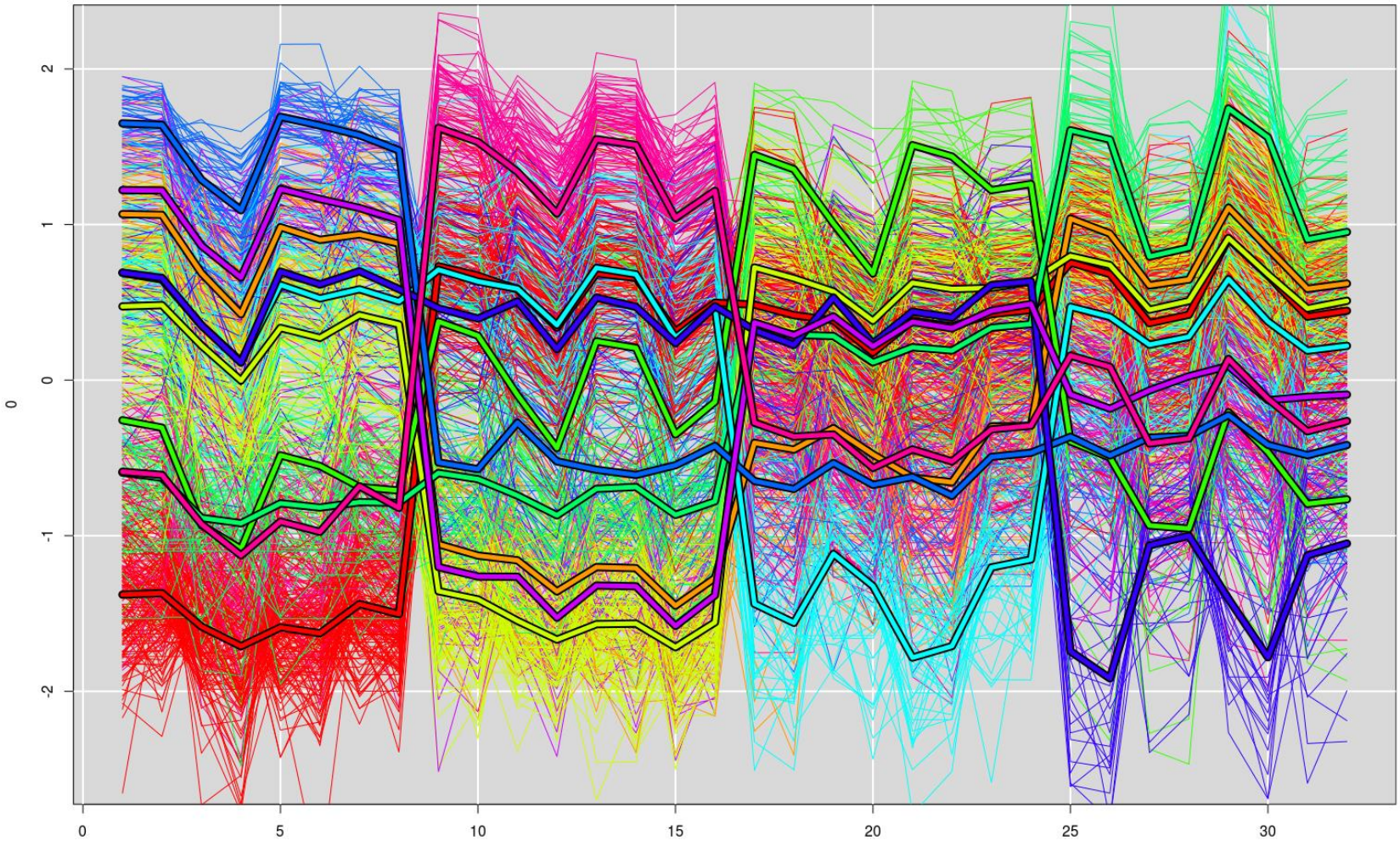
K-means

+ Murine cells were treated with DNA damage-inducing reagent
+ Atm and Atmin (involved in DNA repair) were KO
+ Selection of regulated genes with edgeR identifies genes associated with the experimental conditions
+ Hierarchical clustering unravels clear patterns in gene expression data



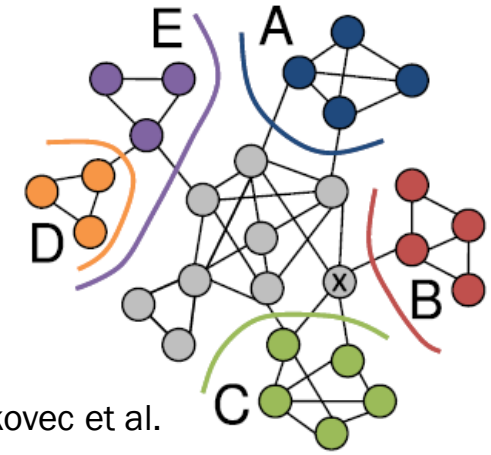
K-means

+ The K-means algorithm identifies typical gene expression patterns (thick lines) that represent clusters of individual gene expression profiles (thin lines)
+ k=10 in this example ; R script in `example-K-means.R`

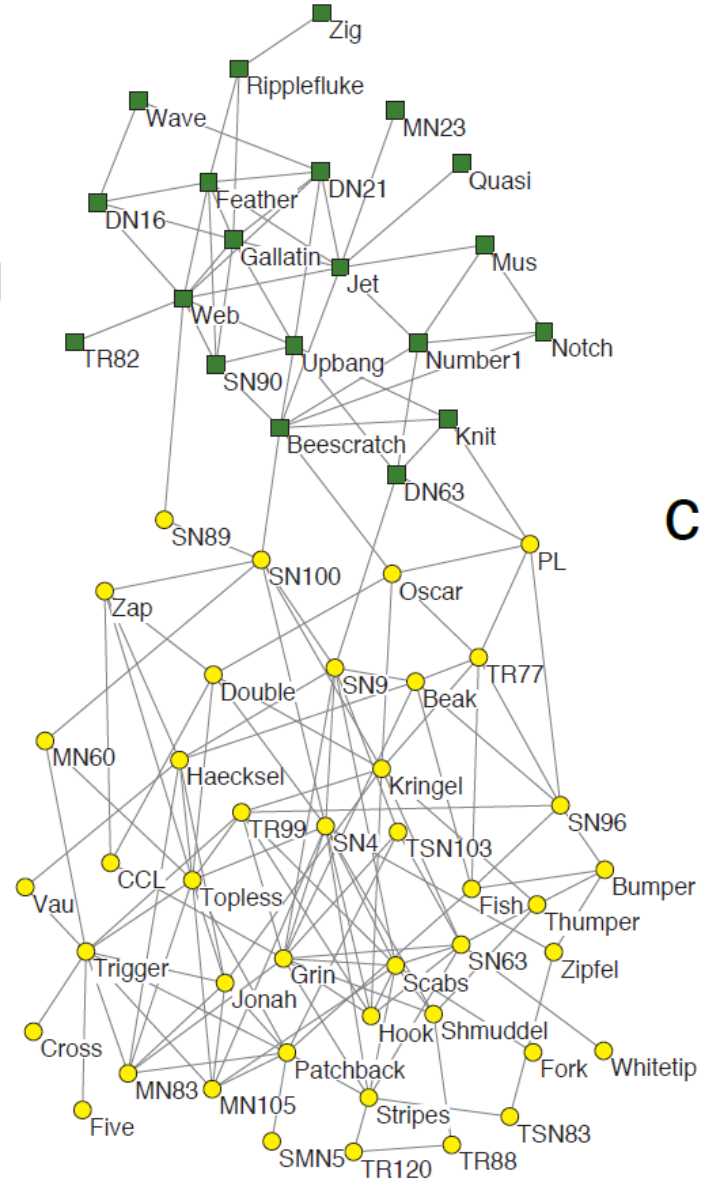


Graphs and communities

+ Graphs are a very versatile mathematical framework to model relationships between objects : gene or protein interactions, signaling pathways, social or commercial networks, airlines, roads, etc.
+ One common problem in graph theory is the detection of communities, *i.e.* subsets of graph nodes that are more connected with eachother than with the rest of the network : genes involved in a same biological function, friends, commercial partners, etc.



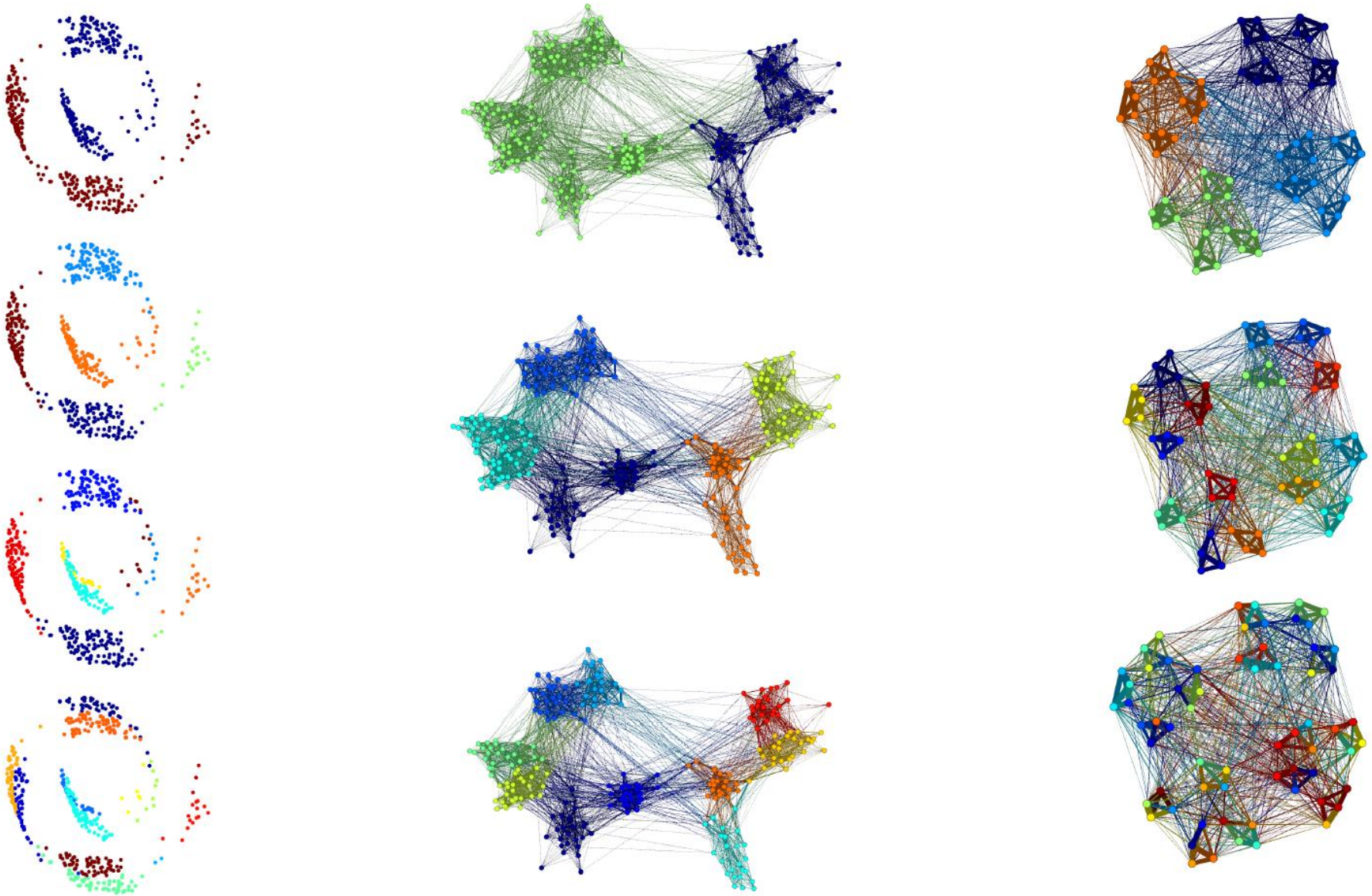
Source : Leskovec et al.



C

Source : S Fortunato, 2010

Communautés multi-échelles



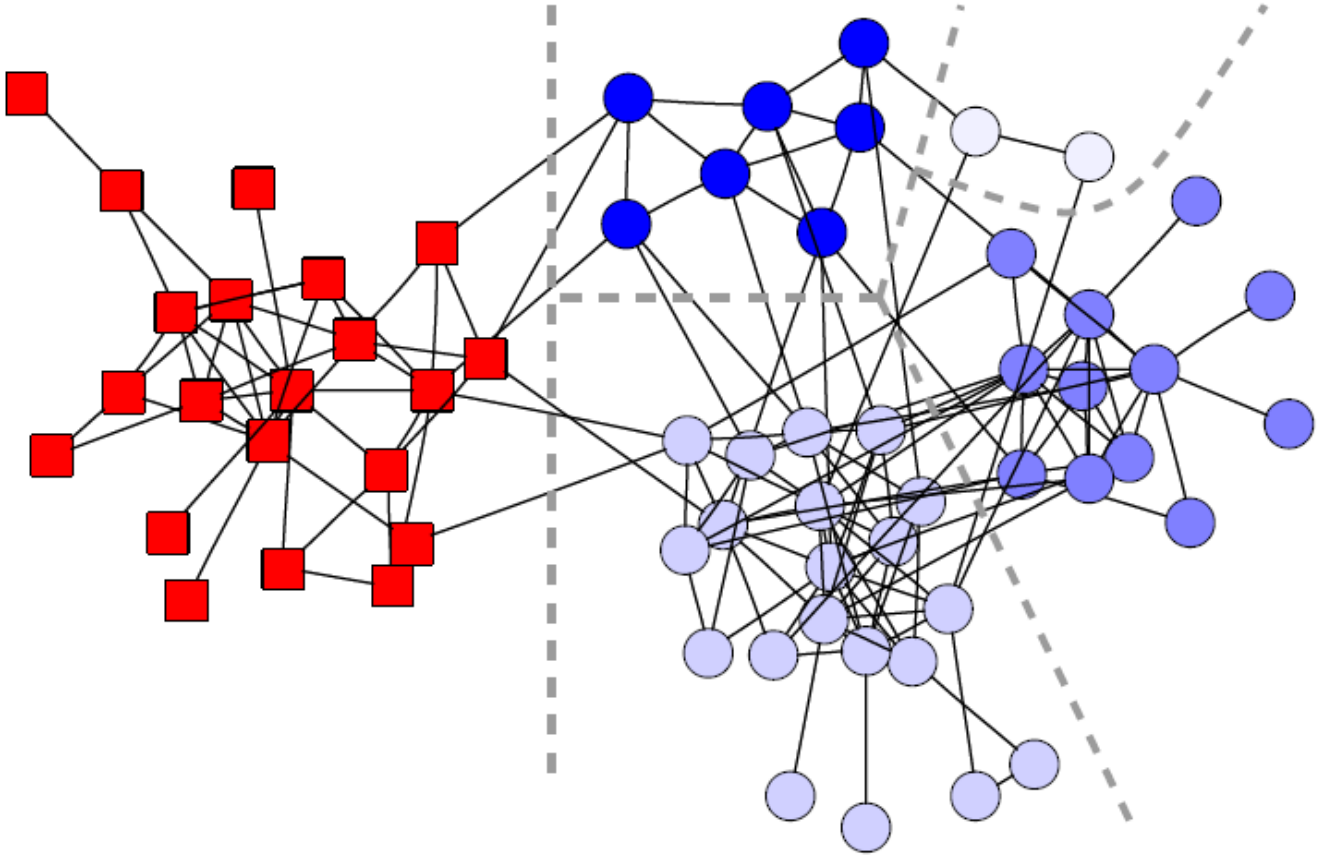
Source : Nicolas Tremblay, ENS, Lyon

Méthodes par suppression d'arêtes

+ Une approche de la recherche de communautés consiste à supprimer des arêtes afin de décomposer le graphe en composantes connexes
+ Girvan et Newman (2002) on proposé l'algorithme suivant:
  - Calculer le betweenness des arêtes (concept similaire au betweenness des nœuds ci-dessus)
  - Supprimer l'arête de betweenness le plus élevé
  - Répéter
+ On obtient une décomposition toujours plus fine

+ Pour décider où s'arrêter, on peut appliquer un critère de qualité aux communautés, par exemple la modularité (= #arêtes internes/#arêtes externes)

Exemple Girvan & Newman

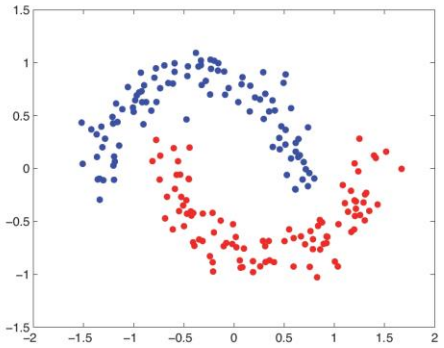


Carrés et disques dénotent le premier découpage. Les disques sont ensuite découpés plus finement par l'algorithme.

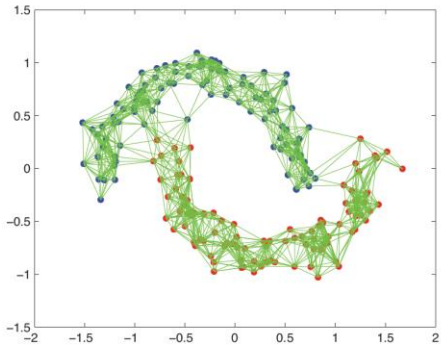
Source : MEJ Newman

Graph-based clustering

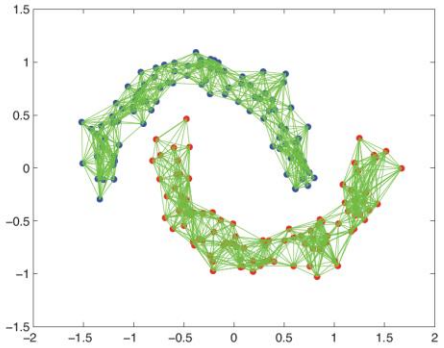
+ Broad family of clustering methods
+ The general principle is to build a graph by linking data points that are close enough in the original space
+ A graph is thus obtained, which is submitted to a community detection algorithm
+ The communities are the clusters



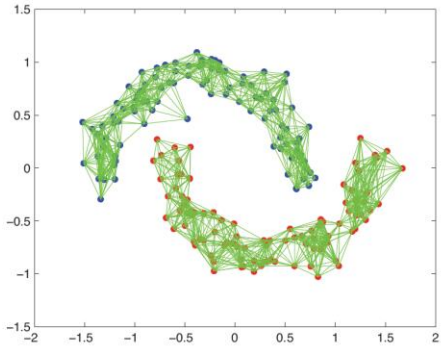
(a) Original Points



(b) Original Connected Graph



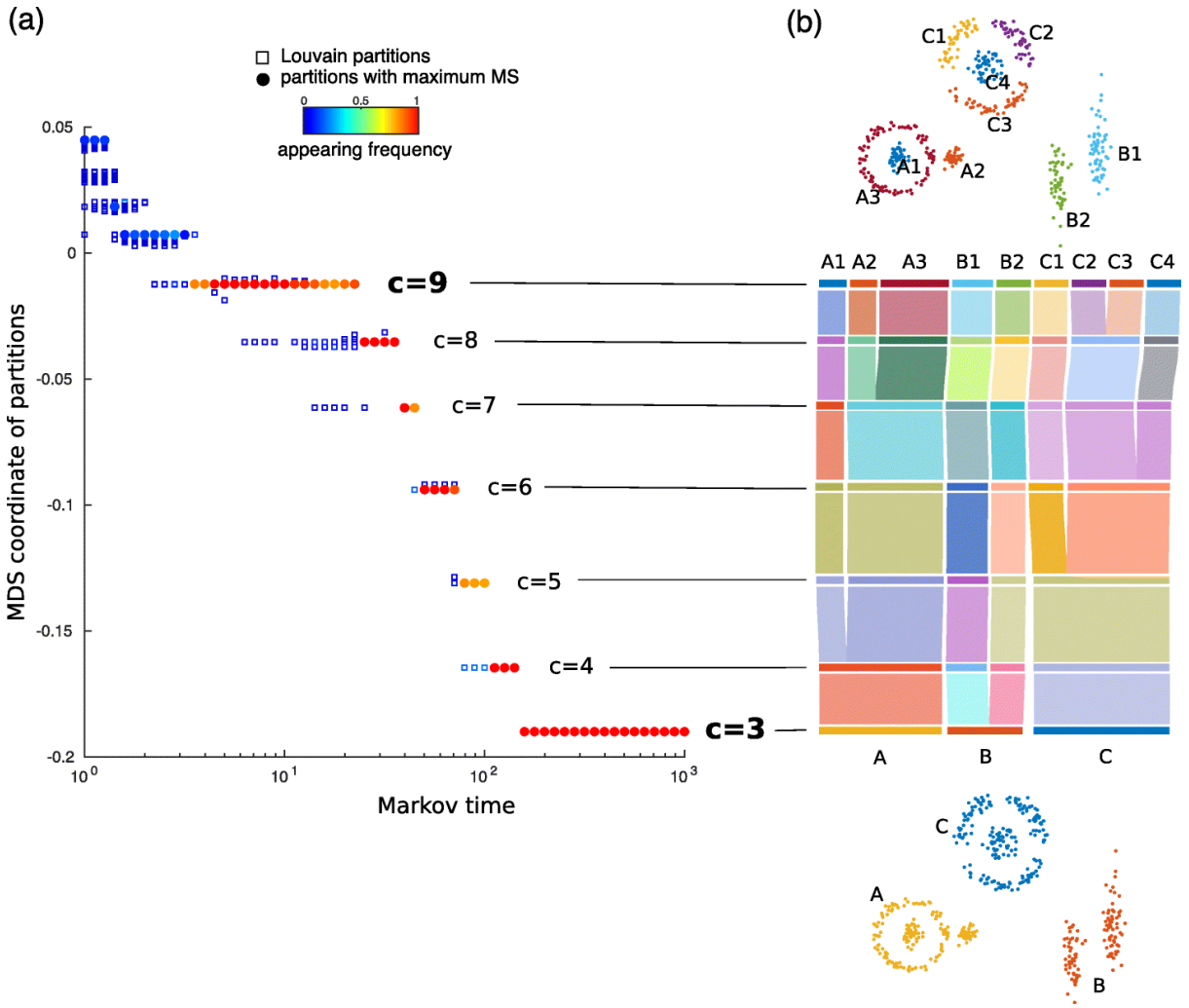
(c) CLR_L1 Result



(d) CLR_L2 Result

Source: Nie et al, Proceedings of the 13th AAAI Conf on AI, 2016

Louvain algorithm (used in Seurat and common in sc studies)



Liu & Barahona, Appl Network Sci, 2020

A heatmap also provides an excellent 2D-projection

+ The plot is indeed in 2 dimensions
+ It does the clustering as well
+ A glimpse of how and why the different clusters differ is provided
+ PCA, t-SNE, or UMAP 2D-projections do not reveal the nature of the differences between clusters