EDA: Clustering Analysis

RNA-seq data analysis

Paulo Czarnewski
https://czarnewski.github.io/czarnewski/index.html

Why clustering?

The process of organizing objects into groups whose members are similar in some way

Typical clustering methods are:

- K-means clustering
- Hierarchical clustering
- Density based clustering
- Graph based clustering

Most clustering methods are divided in:

- **Step 1**: Calculate distances between data points
- **Step 2**: Partition | Group the data based on distances

Distance metrics

Most commonly used distances in RNA-seq:

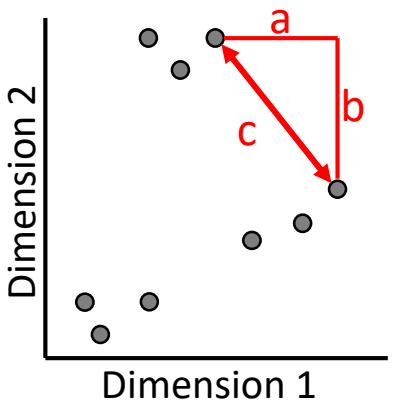
- Euclidean distance

  A.K.A., a straight line between 2 points

  Euclidean distance = c^2 = $b^2 + a^2$

- Inverted pair-wise correlations

  dist = - (correlation - 1)/2



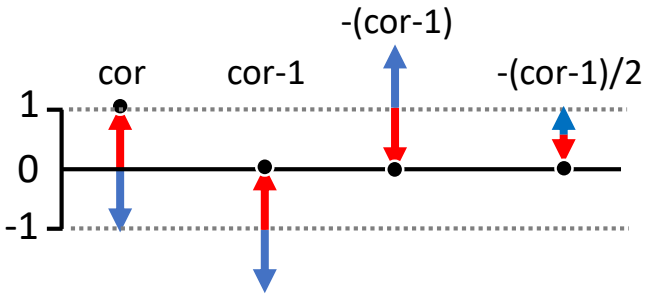
Others common distances are:

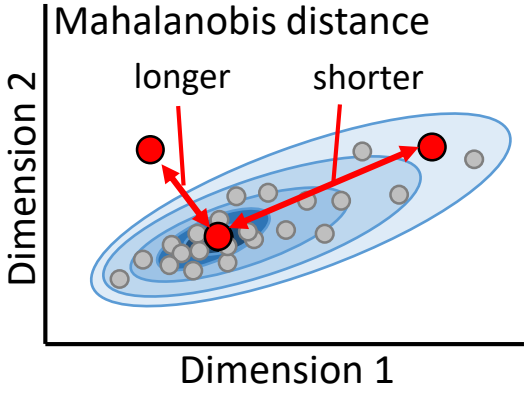
- Manhattan distance

  a + b

- Mahalanobis distance

  Takes data point distribution into consideration





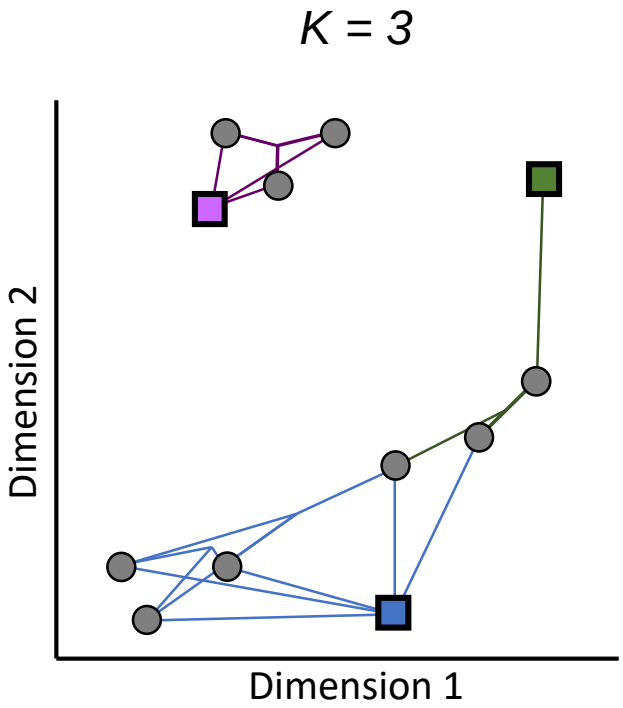
Distances can be measured in:

- In multidimensional space (raw data)
- In PCA reduced space (i.e. top PCs)

K-means

It was one the first and simplest clustering algorithms

It splits the data into *K* number of clusters



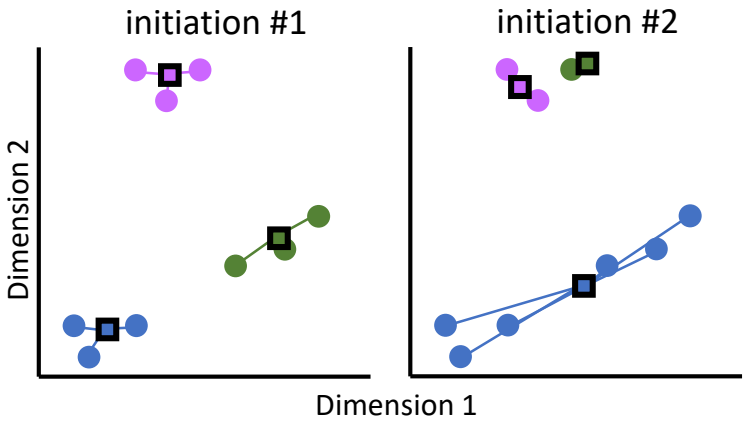
Steps:

1. Randomly initiate *K* cluster centroids
2. Assign points to the nearest centroid
   Using a distance metric (i.e. Euclidean)
3. Compute the average distance between points for every cluster
4. Update cluster centroid location
5. Repeat the same steps 2-4 until no more changes occur.

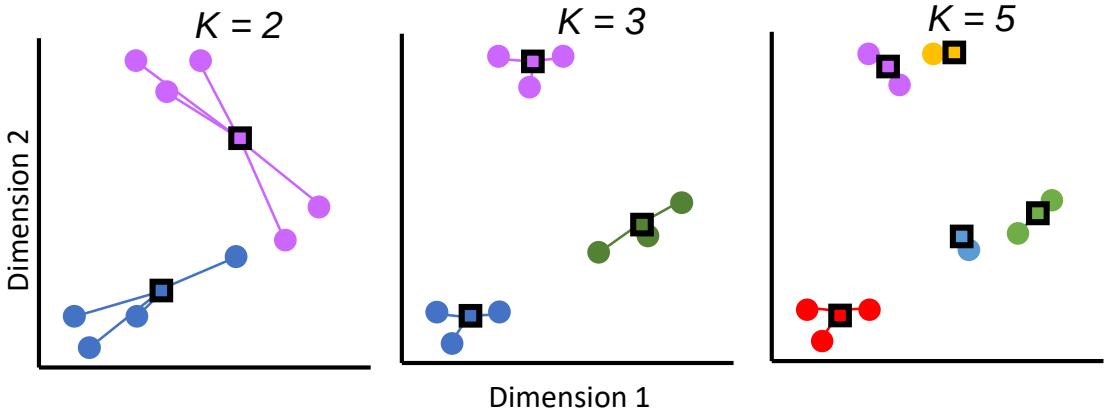
K-means

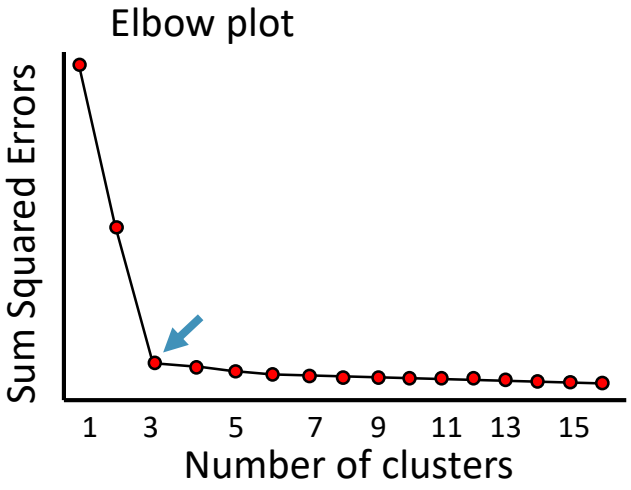
Problems:

1. The random initiation not always get the right clusters
   You need to repeat it ~1000 times and get the most frequent result.
   (this is included by default in most functions)



2. The **user** needs to define the number of clusters. Some methods might help you:
   - *Elbow method*
   - *Gap statistics method*
   - *Average Silhouette method*





Hierarchical Clustering

It is a clustering technique that performs a step-wise grouping of sample, thus seeking to build the hierarchy of clusters.

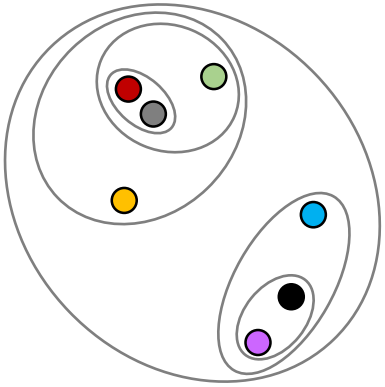
Final product is a dendrogram representing the order decisions at each merge/division of clusters.

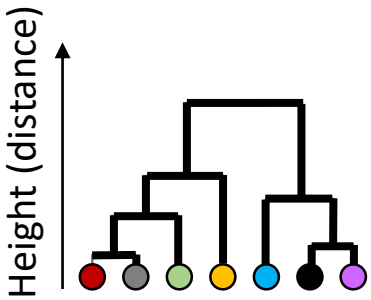
Agglomerative

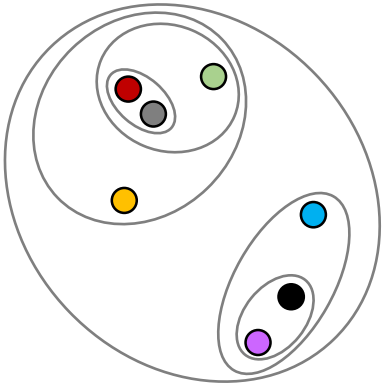
starts with all data points as individual clusters and joins the most similar ones in a bottom-up approach

Divisive

starts with all data points in one large cluster and splits it into 2 at each step. A top-down approach



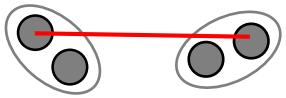




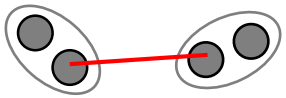
Linkage Methods

In order to merge the clusters, we need to define the position of the cluster.
The method to define this position is called Linkage

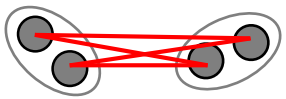
Complete Linkage
Maximum distance



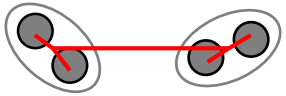
Single Linkage
Minimum distances



Average Linkage
Average of all points



Ward's linkage



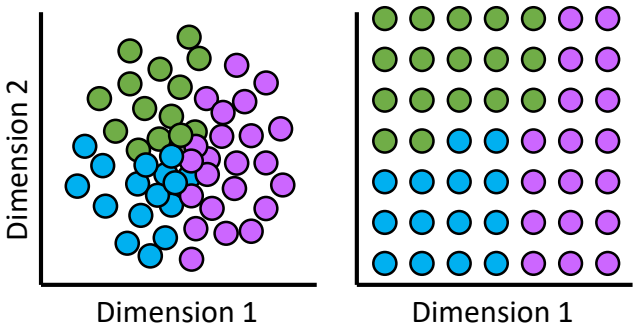
And many others …

Checking Clustering Robustness

A.k.a., Ensemble perturbations

Most clustering techniques will cluster random noise

They usually assume that clusters exists



One way of testing this is by clustering on parts of the data

i.e. re-cluster several times using 90% of the samples (or genes)

*This is often called as **clustering bootstrapping***

Table 3. Ensemble perturbations. Major perturbations applied to the data or to the clustering parameters in ensemble clustering and their intended purpose.

Perturbation	Reason behind perturbation	References
K	Identify the optimum number of clusters	(*46*, *60*, *63*)
Noise	Identify relationships within the data that are not affected by biological or experimental noise	(*45*, *47*, *64*)
Starting point (nondeterministic algorithms)	Identify those partitions that are independent of starting position or identify set of minima	(*23*, *48*)
Projections into lower dimensions	Increase robustness to clustering noise resulting from high dimensionality	(*30*, *63*, *65–67*)
Subsampling	Identify subsets of the data that cluster consistently	(*68–70*)
Parameters of clustering	Identify unpredicted biological information	(*32*)

Ronan et al (2016) *Science Signaling*

Linear Clustering Summary

It is possible to cluster on the raw data (i.e. Z-score) or on the top PCA coordinates
Recommendation: do on raw data or Z-score scaled data (general for bulk RNA-seq)
Cluster on PCA only if you have 10.000s of samples (i.e. single cell RNA-seq)

No clustering is fully automatic
There is always one parameter that needs to be set by the user

Most clustering methods will define clusters on random noise

K-means do not work well in unbalanced clusters
It will generate clusters with somewhat similar sizes

On hierarchical clustering, some distance metrics need to be used with a certain linkage method

Ronan et al (2016) *Science Signaling*

Want to know more ?

NBIS SciLifeLab

REVIEW

COMPUTATIONAL BIOLOGY

Avoiding common pitfalls when clustering biological data

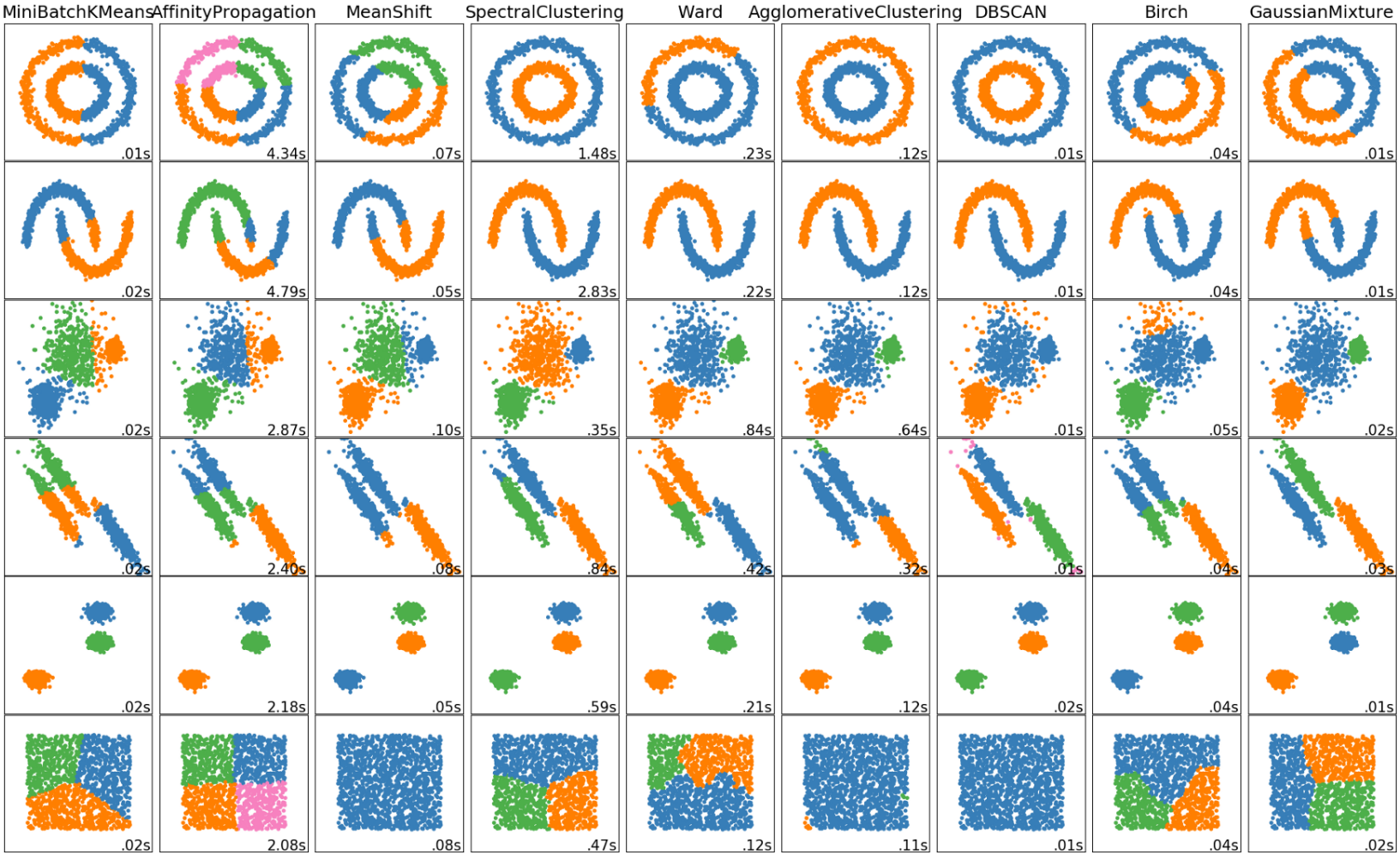
Tom Ronan, Zhijie Qi, Kristen M. Naegle*

Table 4. Summary of in-depth review articles. A collection of reviews for more in-depth coverage of each topic.

In-depth review area	References
Specific clustering algorithms, their trade-offs, and how they function	(*2*, *3*, *71*–*73*)
Analysis of the effects of different distance metrics on clustering gene expression data	(*74*, *75*)
Practical and mathematical implications of high dimensionality on clustering	(*15*, *16*)
A thorough review of validation metrics	(*35*, *38*, *62*, *76*)
The most common multiple hypothesis correction procedures including Bonferroni correction and FDR correction.	(*55*, *58*)
The effects of specific distances on data clustering for lower-dimensional spaces	(*77*)
The effects of specific distances on data clustering for high-dimensional spaces	(*15*, *78*)
Ensembles of some algorithms incompatible with high-dimensional data can be useful on higher-dimensional data, even when a single clustering solution is uninformative.	(*23*, *24*)
A more in-depth analysis of ensembles, including evaluating the results of multiple clustering runs and determining consensus	(*60*, *63*)

Ronan et al (2016) *Science Signaling*

Clustering method comparison



https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

Thank you. Questions?

Paulo Czarnewski