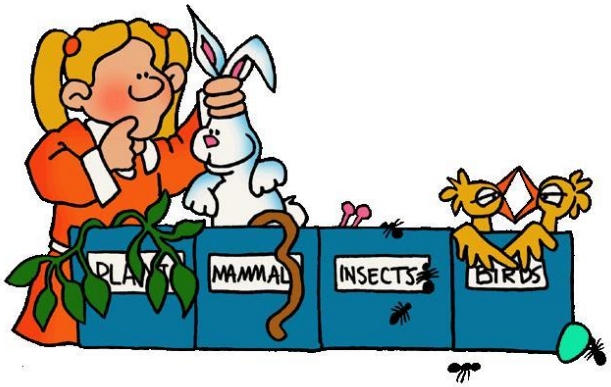
CLUSTERING AND EXPLORATORY ANALYSIS

Ester Feldmesser
Bioinformatics unit
December 8th 2019

Outline

- Classifications in high throughput experiments
  - Discriminant analysis versus cluster analysis
- What is clustering?
- Distance measures
- Clustering algorithms
  - Hierarchical clustering
  - K-Means Cluster
- Special considerations
- Summary
- Exploratory analysis



Supervised versus unsupervised analysis

- Discriminant analysis (supervised, i.e. ANOVA)

CLASSES KNOWN

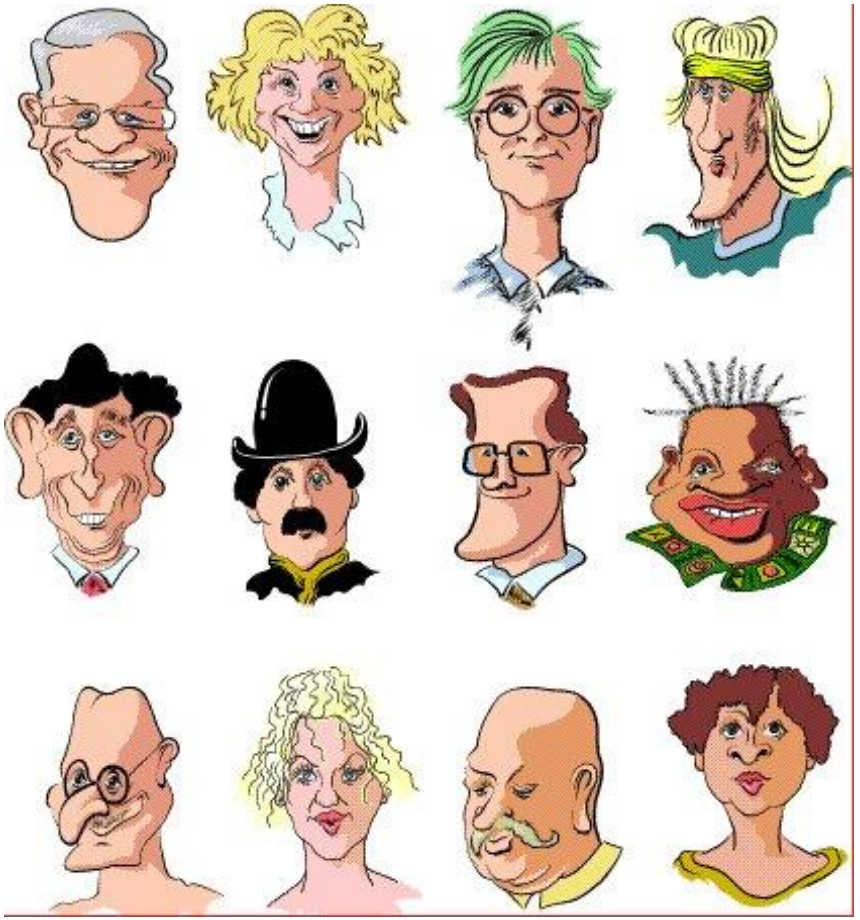
- Cluster analysis (unsupervised)

CLASSES NOT KNOWN

What is clustering?

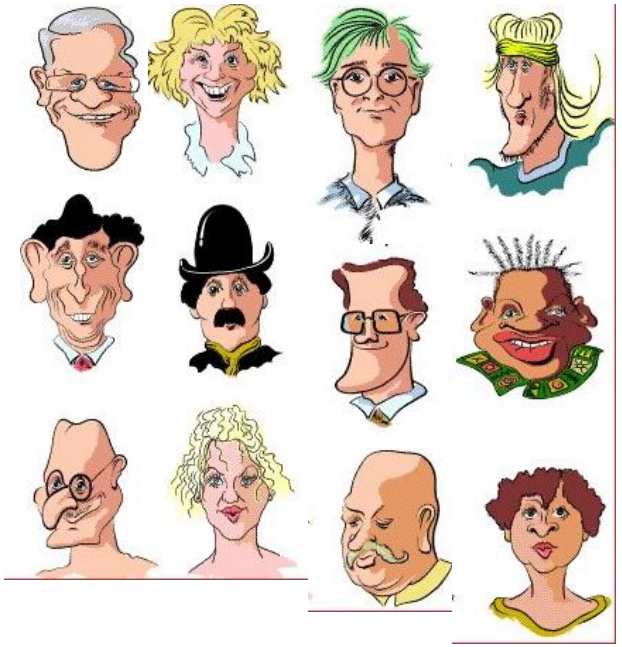
Clustering is a method to classify individual items which are placed into groups based on quantitative information on one or more characteristics inherent to the items

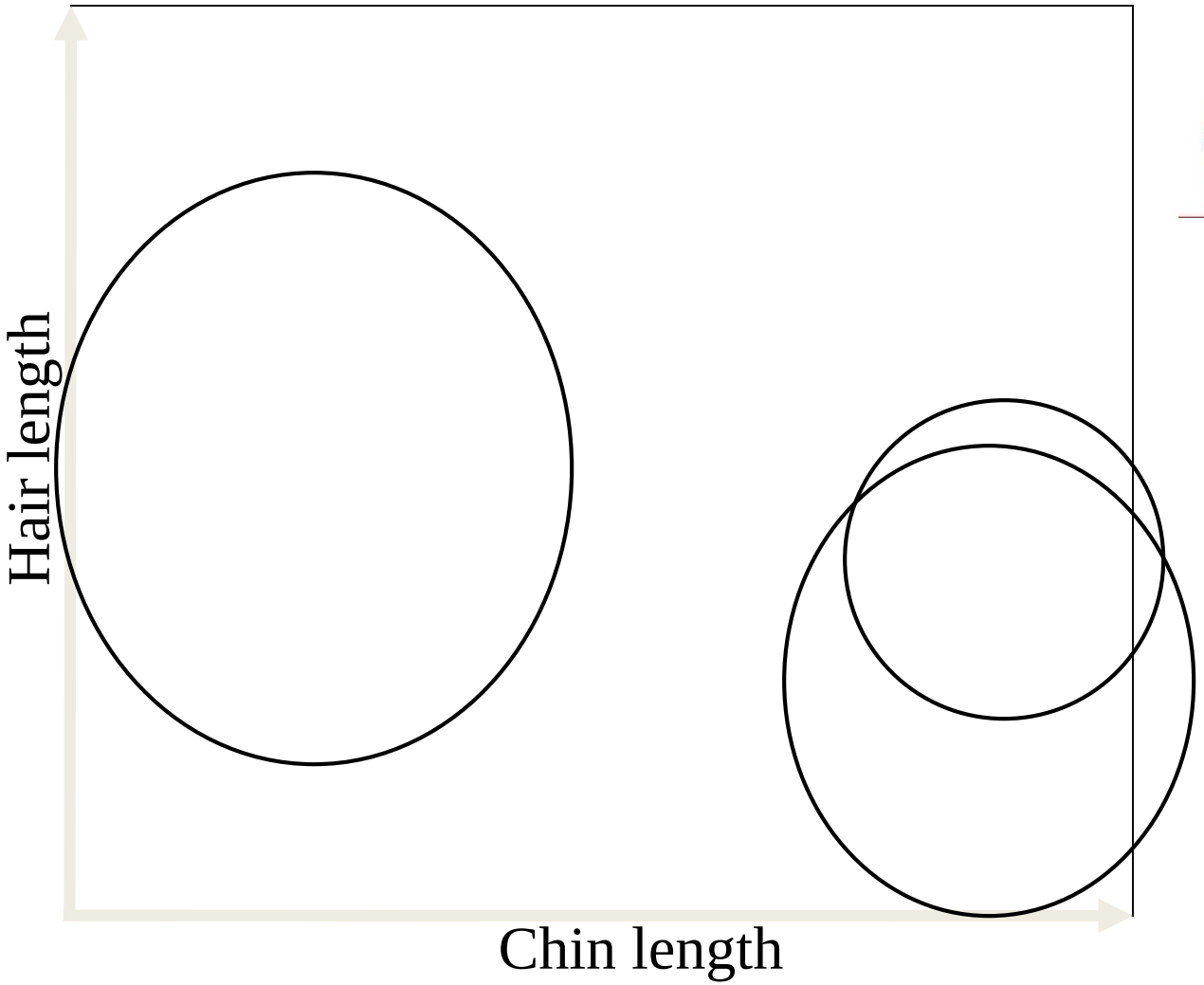
The objects within each cluster are more **closely related** to each other than to objects in other clusters



http://www.alanfielding.co.uk/multivar/ca.htm

Similarity Matrix





People in n-dimentional characteristics space

Characters

People

	Chin	Hair	Hat	Nose	Glasses	Neck
Person1	5	0	0	5	3	2
Person2						
Person3						
Person4						
Person5						
Person6						
Person7						
Person8						
Person9						

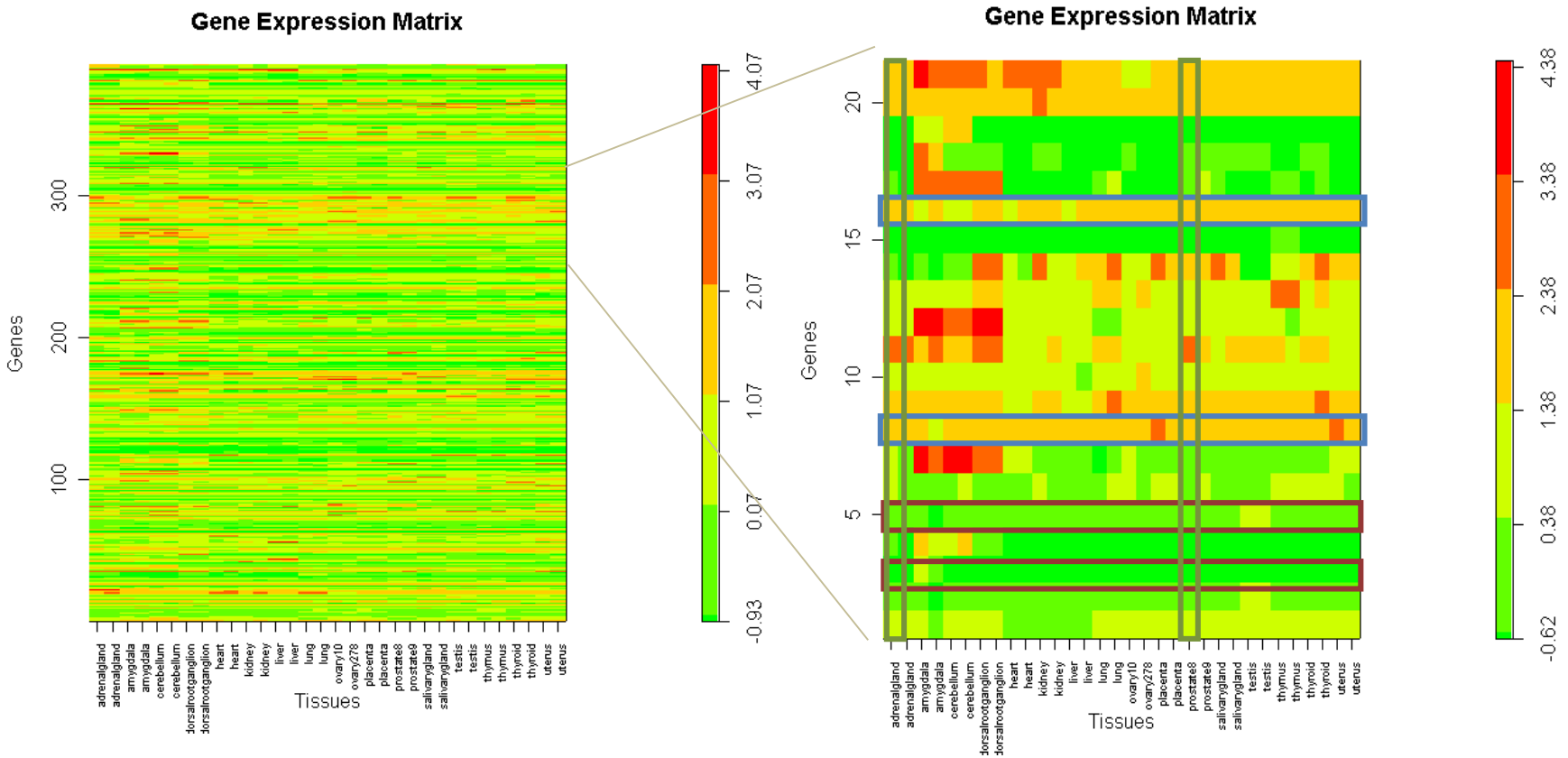
Genes in n-dimentional experimental conditions space

RNA samples

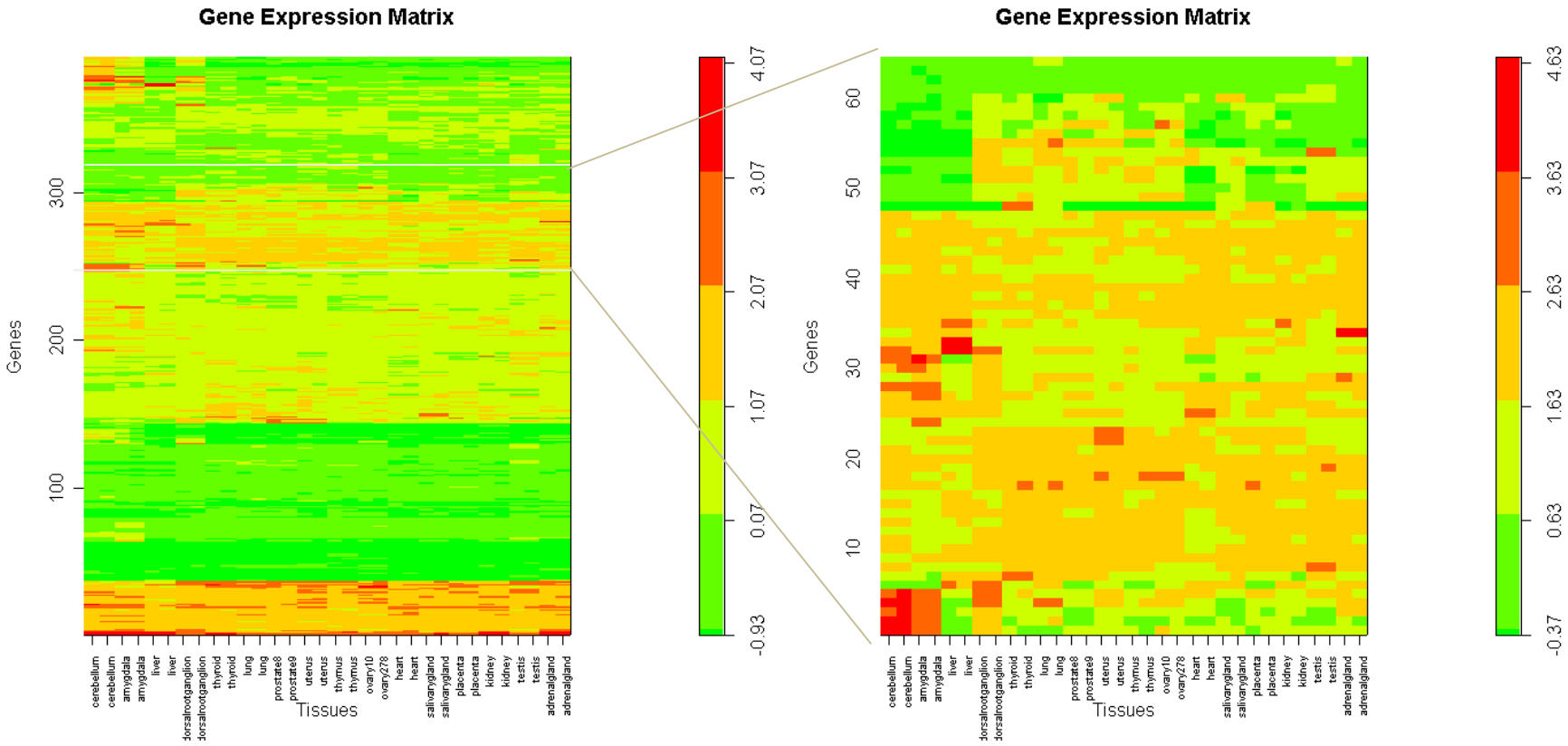
Genes

	Heart	Uterus	Liver	Kidney	Pancreas	Muscle
Gene1	5.72	9.36	4.12	4.85	4.75	5.51
Gene2						
Gene3						
Gene4						
Gene5						
Gene6						
Gene7						
Gene8						
Gene9						

Finding similar patterns in expression matrix

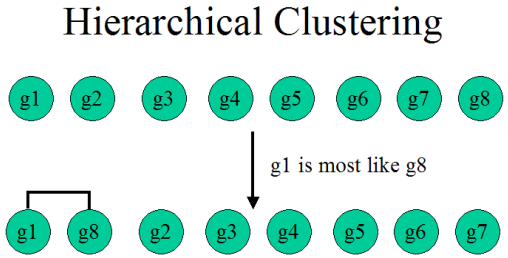


Reordered Gene Matrix



Hierarchical clustering

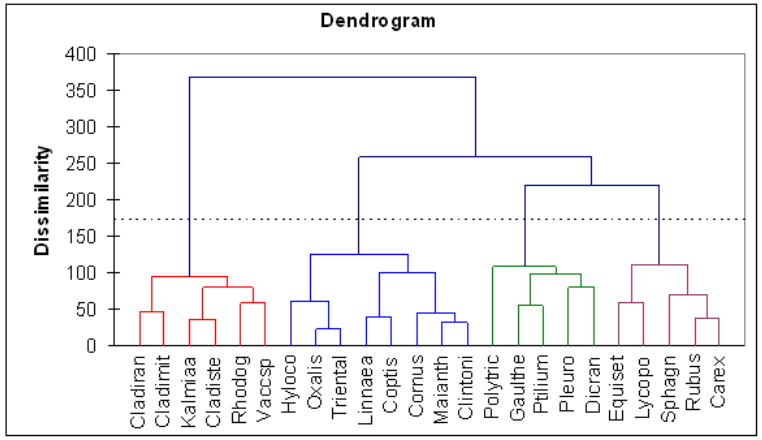
- **Hierarchical clustering** was first used in microarray research to cluster genes (Eisen et al, 1998)



- **Hierarchical clustering** uses
  - a measure of **distance** between pairs of observations
  - a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise **distances** of observations in the sets.
- Agglomerative – bottom to top
  - each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Hierarchical cluster algorithms

- Gives a dendogram (tree like structure)



- In each iteration, merge the two clusters with the minimal distance from each other - until you are left with a single cluster comprising all objects.

- Agglomerative clustering is used here, sometimes divisive clustering is used

- But what is the distance between two clusters?

http://compbio.utmem.edu/MSCI814/Module10.htm

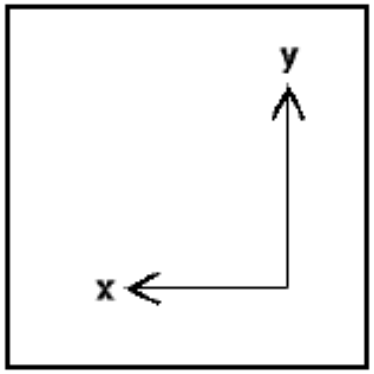
Distance measures

- **Distance** is a numerical description of how far apart objects are and it is how the ***similarity*** of two elements is calculated
- Distance measure
  - Euclidean
  - Manhattan
  - Pearson correlation

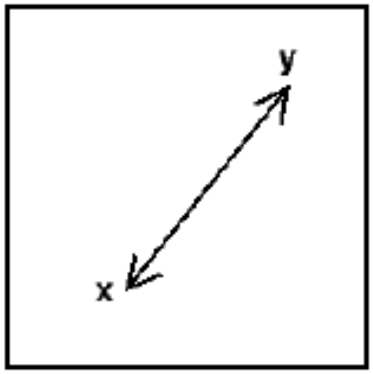
Difference between Manhattan distance and Euclidean distance

$$\sum |x - y|$$

6-0 + 6-0
12



Manhattan



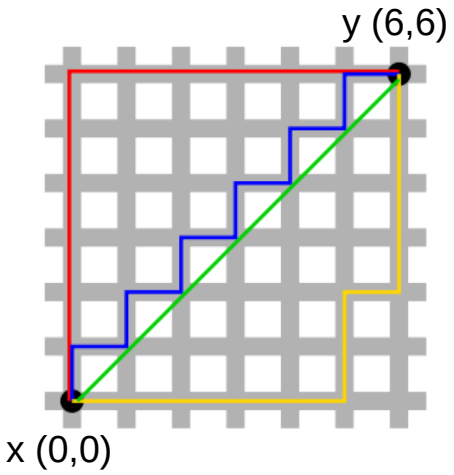
Euclidean

$$\sqrt{\sum (x_i - y_i)^2}$$

$\sqrt{(6-0)^2 + (6-0)^2}$
$6\sqrt{2} \approx 8.49$

Euclidean distance: measures true straight line distances. Note that the distance from Manchester to Chicago would not be Euclidean unless you bored a tunnel through the earth.

Manhattan distance: since it is not squared, it goes through a larger trajectory, like blocks in a city.



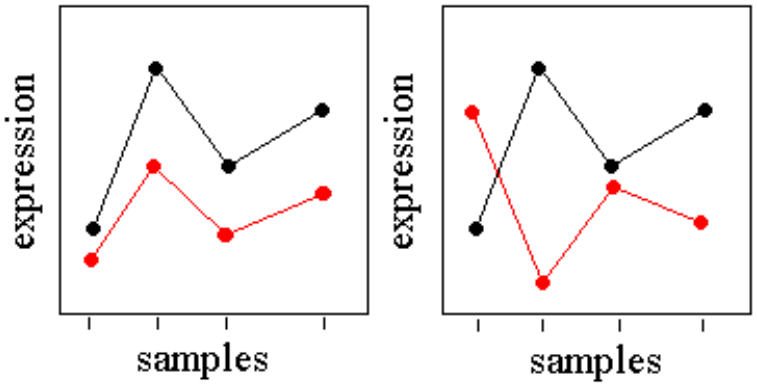
https://commons.wikimedia.org/w/index.php?curid=731390

Pearson correlation

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

- X and Y are different genes
- This correlation is the covariance (measure of how changes in one variable are associated with changes in a second variable)
- Divided by the sample standard deviations product of X and of Y (scaling)

It measures the similarity in shape between two profiles, but can also capture inverse relationships.



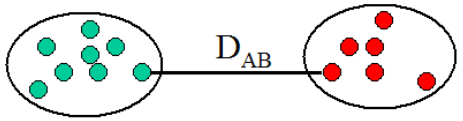
Hierarchical cluster algorithms

Single Linkage

Cluster-to-cluster distance is defined as the minimum distance between members of one cluster and members of the another cluster.

$$D_{AB} = \min (d(u_i, v_j))$$

where $u \in A$ and $v \in B$
for all $i = 1$ to N_A and $j = 1$ to N_B

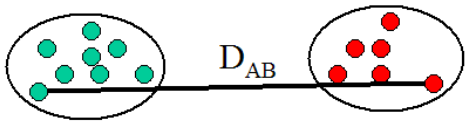


Complete Linkage

Cluster-to-cluster distance is defined as the maximum distance between members of one cluster and members of the another cluster.

$$D_{AB} = \max (d(u_i, v_j))$$

where $u \in A$ and $v \in B$
for all $i = 1$ to N_A and $j = 1$ to N_B



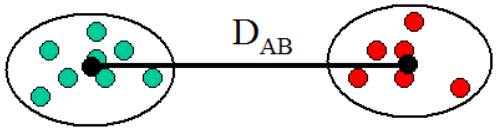
The distance D between two clusters A and B is based on the dissimilarity between objects from the two clusters

Average Linkage

Cluster-to-cluster distance is defined as the average distance between all members of one cluster and all members of another cluster.

$$D_{AB} = 1/(N_A N_B) \Sigma \Sigma (d(u_i, v_j))$$

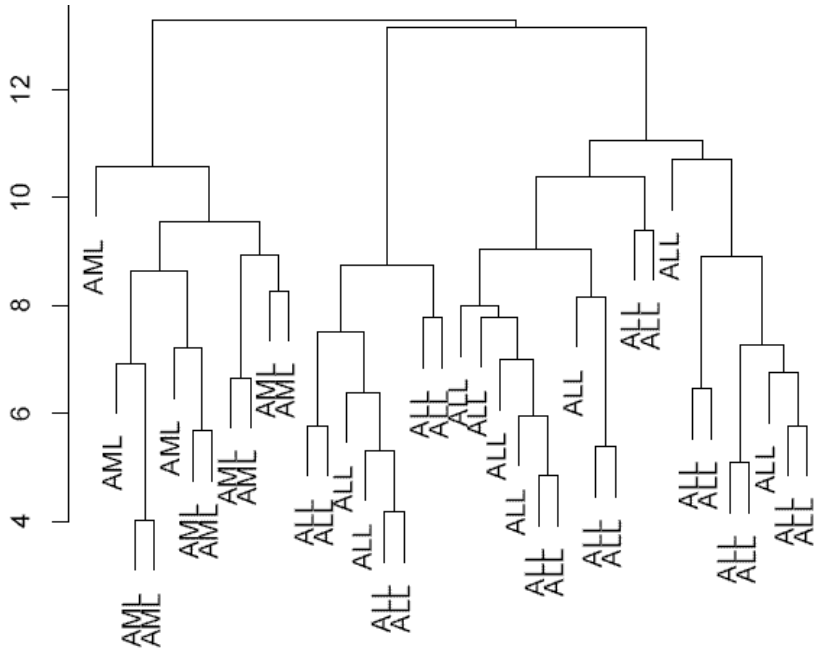
where $u \in A$ and $v \in B$
for all $i = 1$ to N_A and $j = 1$ to N_B



http://compbio.utmem.edu/MSCI814/Module10.htm

Hierarchical cluster visualization

- Similarity of objects is represented in a tree structure (**dendogram**).
- Advantage: no need to specify the number of clusters in advance. Nested clusters can be represented.
- The height of a node in the dendrogram represents the distance of the two children clusters.



A time series example

Time series example

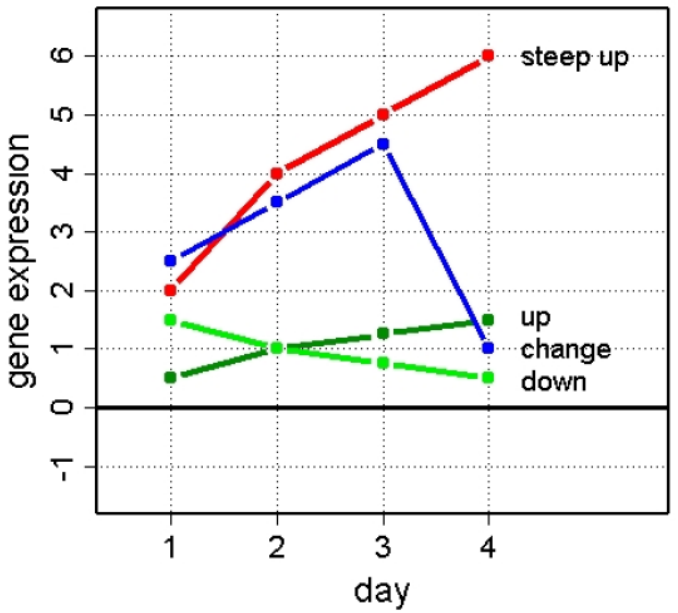
Biology

Measurements of gene expression on 4 (consecutive) days.

Statistics

Every gene is coded by a vector of length 4.

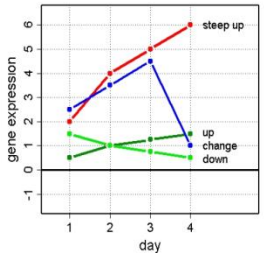
- steep up: $x_1 = (2, 4, 5, 6)$
- up: $x_2 = (2/4, 4/4, 5/4, 6/4)$
- down: $x_3 = (6/4, 4/4, 3/4, 2/4)$
- change: $x_4 = (2.5, 3.5, 4.5, 1)$



Euclidean distance

The distance between two vectors is the square root of the sum of the squared differences over all the coordinates

$$\sqrt{\sum(x_i - y_i)^2}$$



$$d_E(x_1, x_2) = \sqrt{(2\text{-}2/4)^2 + (4\text{-}4/4)^2 + (5\text{-}5/4)^2 + (6\text{-}6/4)^2} = 3\sqrt{3/4} \approx 2.598$$

- steep up: $x_1 = (2, 4, 5, 6)$
- up: $x_2 = (2/4, 4/4, 5/4, 6/4)$
- down: $x_3 = (6/4, 4/4, 3/4, 2/4)$
- change: $x_4 = (2.5, 3.5, 4.5, 1)$

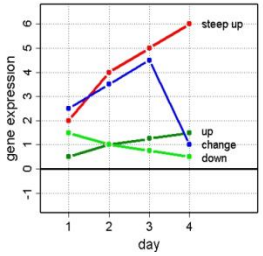
0	2.60	2.75	2.25
2.60	0	1.23	2.14
2.75	1.23	0	2.15
2.25	2.14	2.15	0

Matrix of pairwise distances

Manhattan distance

The distance between two vectors is the sum of the absolute differences over all the coordinates

$$\sum |x - y|$$



$$d_M(x_1, x_2) = |2\text{-}2/4| + |4\text{-}4/4| + |5\text{-}5/4| + |6\text{-}6/4| = 51/4 = 12.75$$

- steep up: $x_1 = (2, 4, 5, 6)$
- up: $x_2 = (2/4, 4/4, 5/4, 6/4)$
- down: $x_3 = (6/4, 4/4, 3/4, 2/4)$
- change: $x_4 = (2.5, 3.5, 4.5, 1)$

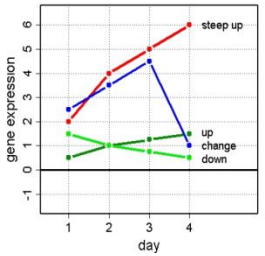
0	12.75	13.25	6.50
12.75	0	2.50	8.25
13.25	2.50	0	7.75
6.50	8.25	7.75	0

Matrix of pairwise distances

Pearson correlation

The distance between two vectors is 1 – r, where r is the Pearson correlation between the two vectors

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$



$$d_C(x_1, x_2) = 1 - \frac{(2-\frac{17}{4})(\frac{2}{4}-\frac{17}{16}) + (4-\frac{17}{4})(\frac{4}{4}-\frac{17}{16}) + (5-\frac{17}{4})(\frac{5}{4}-\frac{17}{16}) + (6-\frac{17}{4})(\frac{6}{4}-\frac{17}{16})}{\sqrt{(2-\frac{17}{4})^2 + (4-\frac{17}{4})^2 + (5-\frac{17}{4})^2 + (6-\frac{17}{4})^2}\sqrt{(\frac{2}{4}-\frac{17}{16})^2 + (\frac{4}{4}-\frac{17}{16})^2 + (\frac{5}{4}-\frac{17}{16})^2 + (\frac{6}{4}-\frac{17}{16})^2}}$$

- steep up: $x_1 = (2, 4, 5, 6)$
- up: $x_2 = (2/4, 4/4, 5/4, 6/4)$
- down: $x_3 = (6/4, 4/4, 3/4, 2/4)$
- change: $x_4 = (2.5, 3.5, 4.5, 1)$

0	0	2	1.18
0	0	2	1.18
2	2	0	0.82
1.18	1.18	0.82	0

Matrix of pairwise distances

Distance measures: summary

- Euclidean and Manhattan distance both measure average distance between vectors.
- May apply **standardization** to the genes:

  Subtract mean and divide by standard deviation:

$$x \mapsto \frac{x - \bar{x}}{\hat{\sigma}_x}$$

- After standardization, Euclidean and correlation distance are equivalent
- Correlation (Pearson dissimilarity) is good in detecting similar trends between vectors

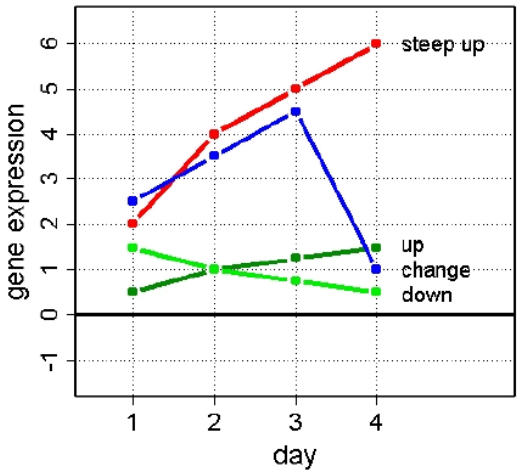
Distance measures: the idea

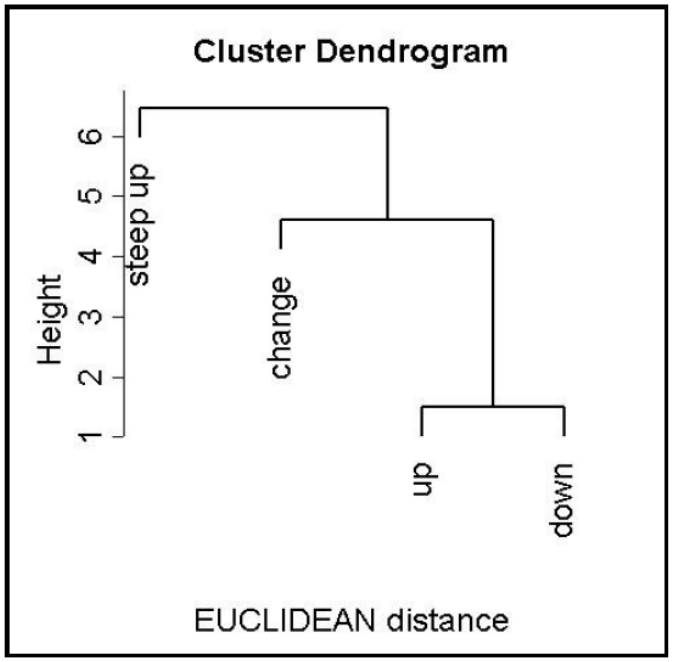
Choose the distance measure according to what you want you are interested in!

A time series example: clustering

Time Series Example

- **Euclidean distance**
  Similar values are clustered together

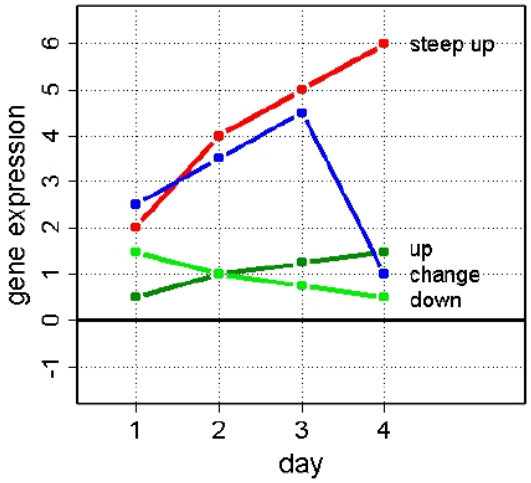


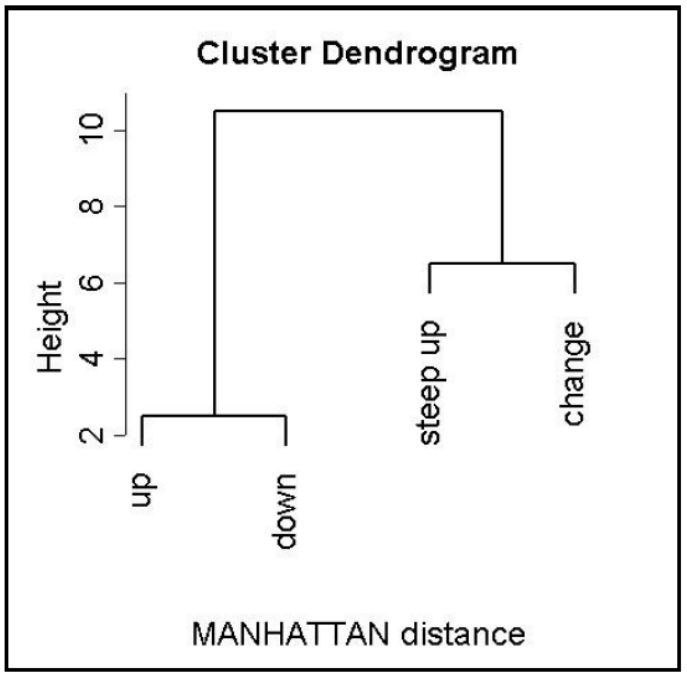


A time series example: clustering

Time Series Example

- **Manhattan distance**
  Similar values are clustered together (robust)

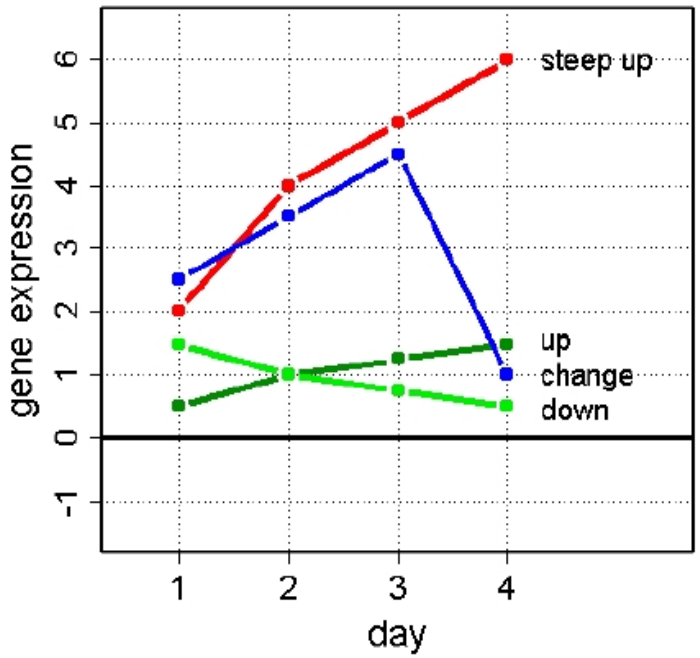


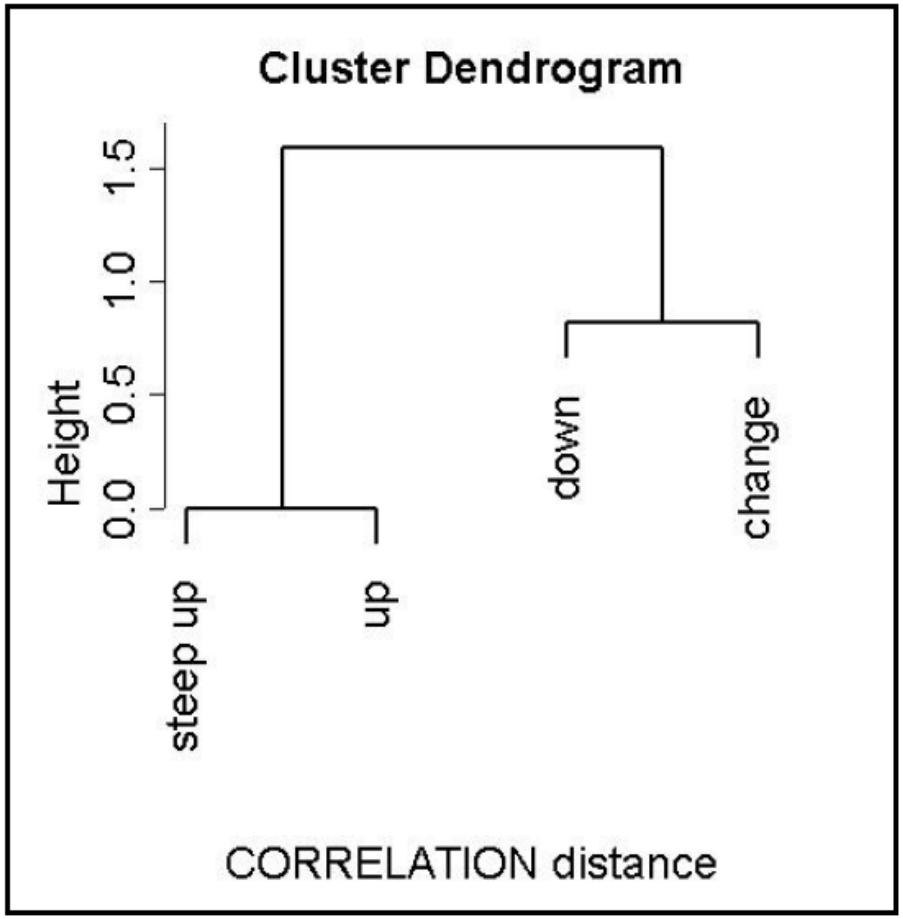


A time series example: clustering

- **Correlation distance**

  Similar trends are clustered together





Constructing a cluster (an example)

Data:

Six RNA samples from three duplicated tissues

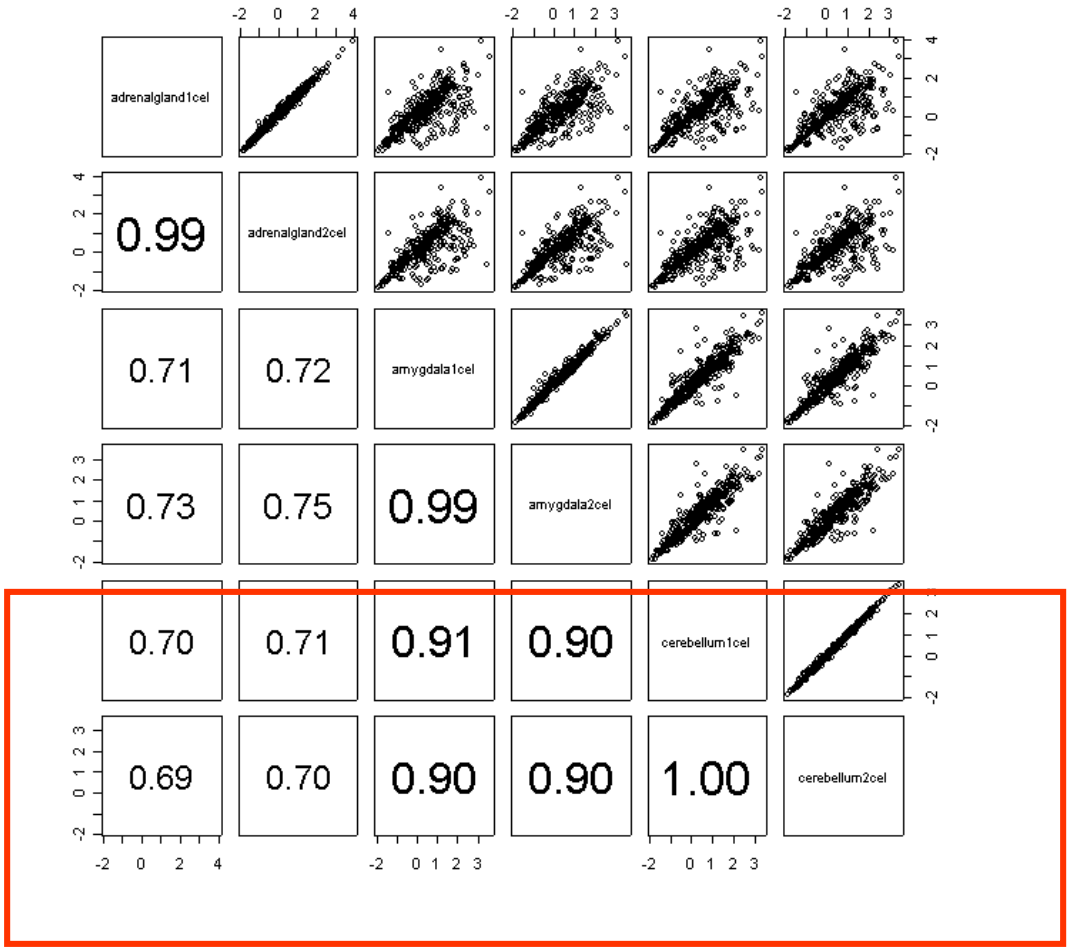
Method:

Agglomerative clustering

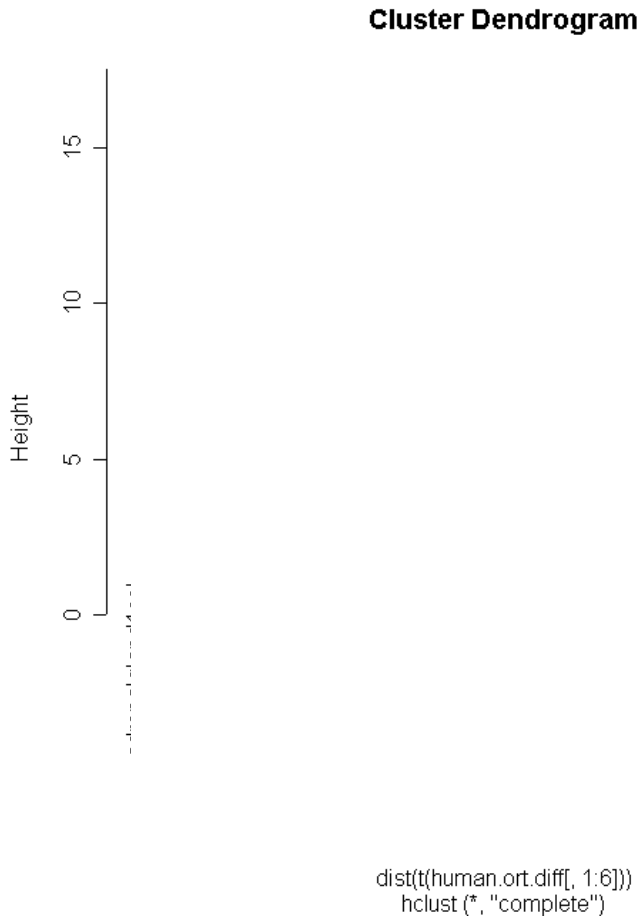
Steps:

- **Calculate all pairwise distances**
- **Define the relationship among samples**

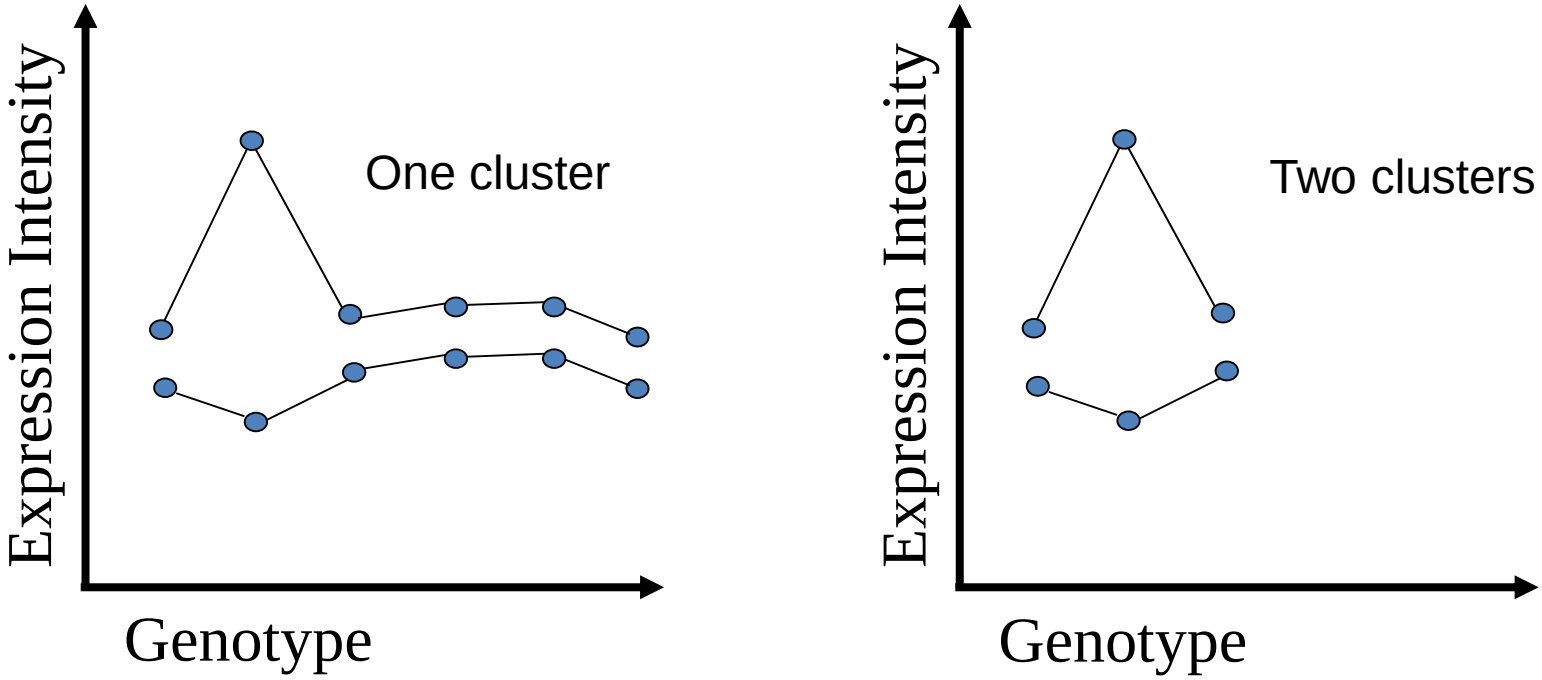
Distance matrix



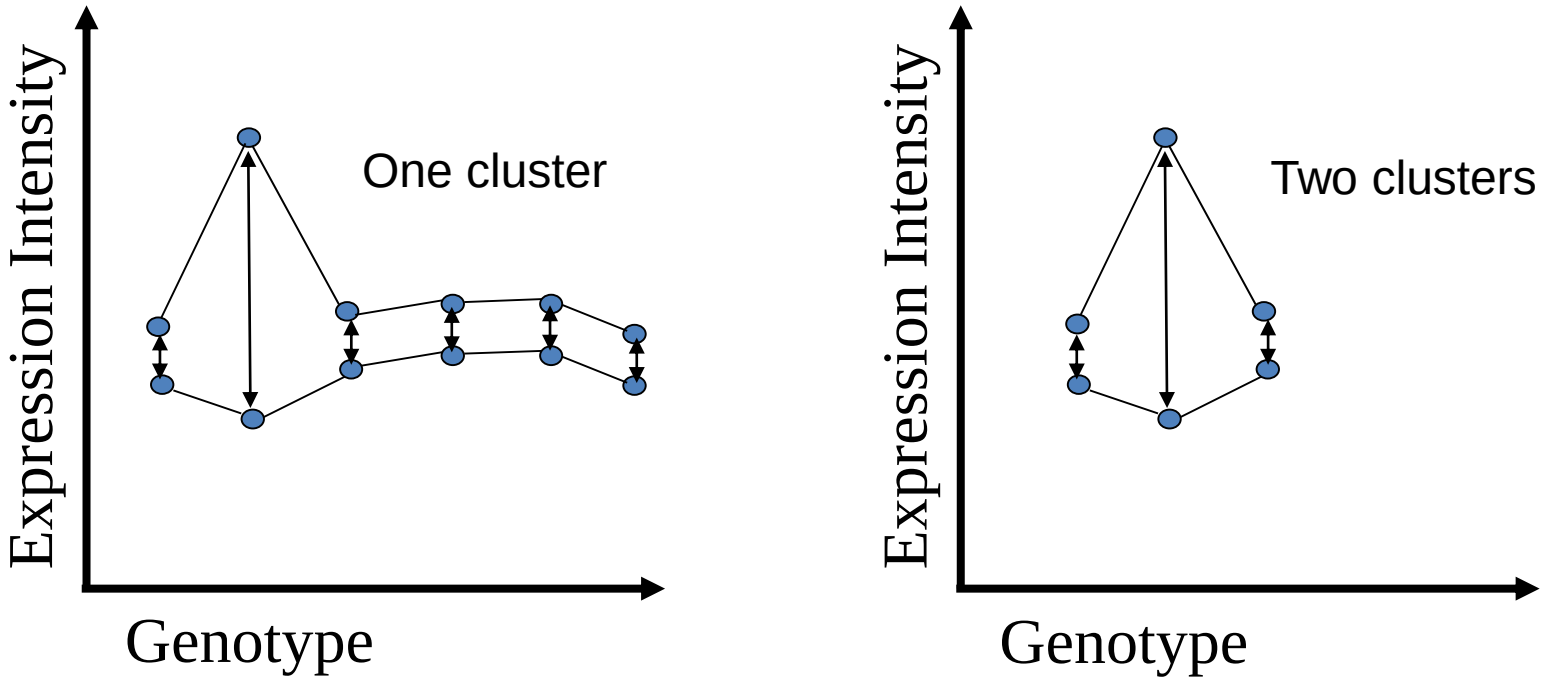
Create a Tissue Dendrogram



The impact of an object is dependent on the number of objects in the vector

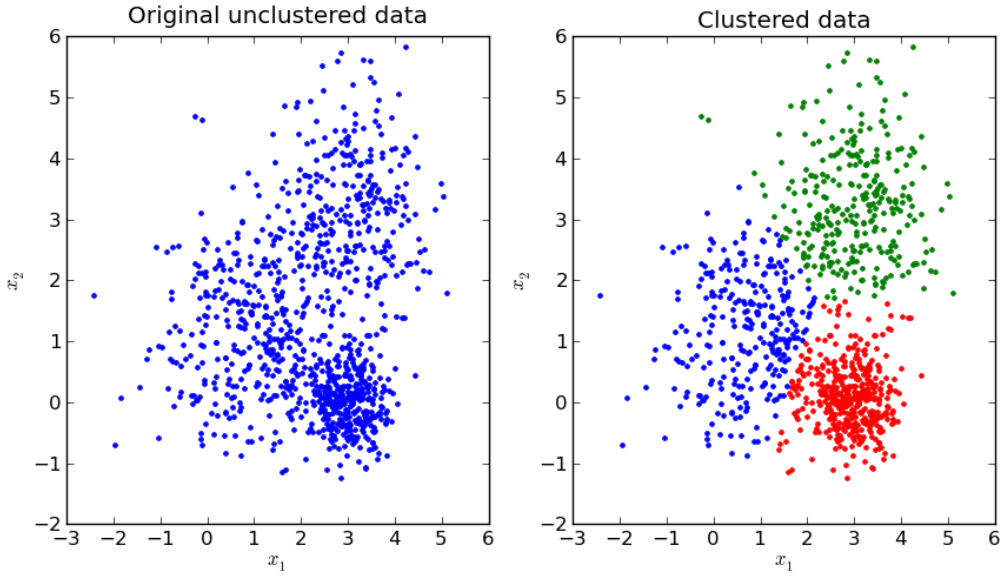


Can be implemented on non replicated samples

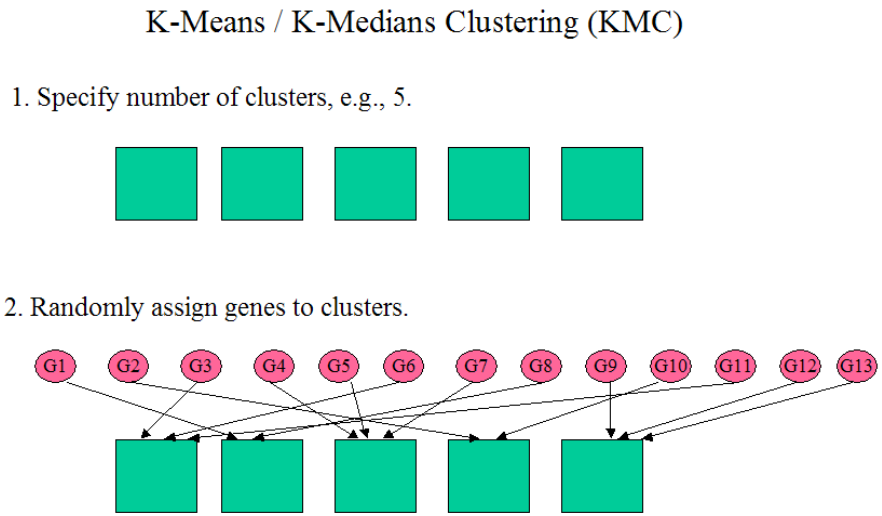


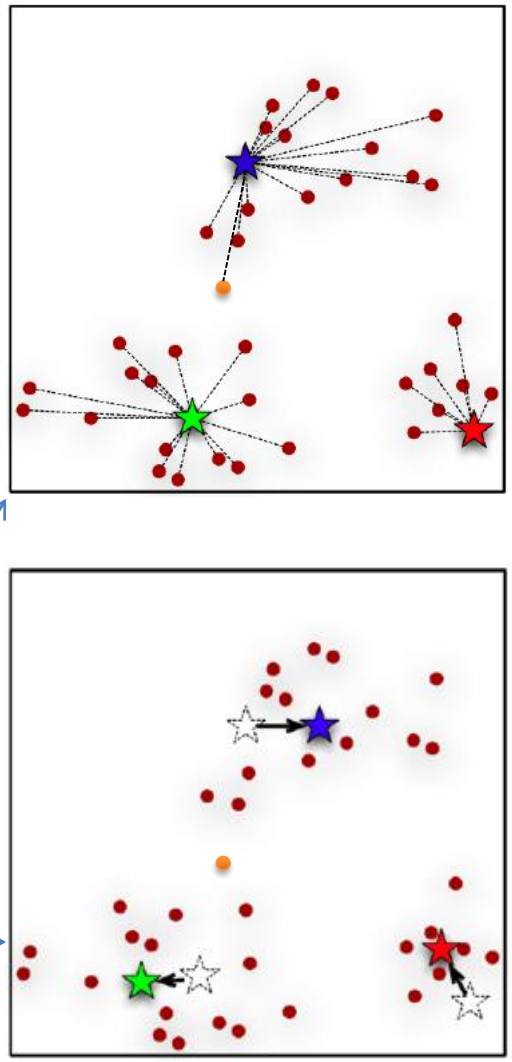
K-Means Clustering

- **k-means clustering** aims to partition n observations into **k clusters** in which each observation belongs to the **cluster** with the nearest **mean**, serving as a prototype of the **cluster**.



K-Means Clustering 2



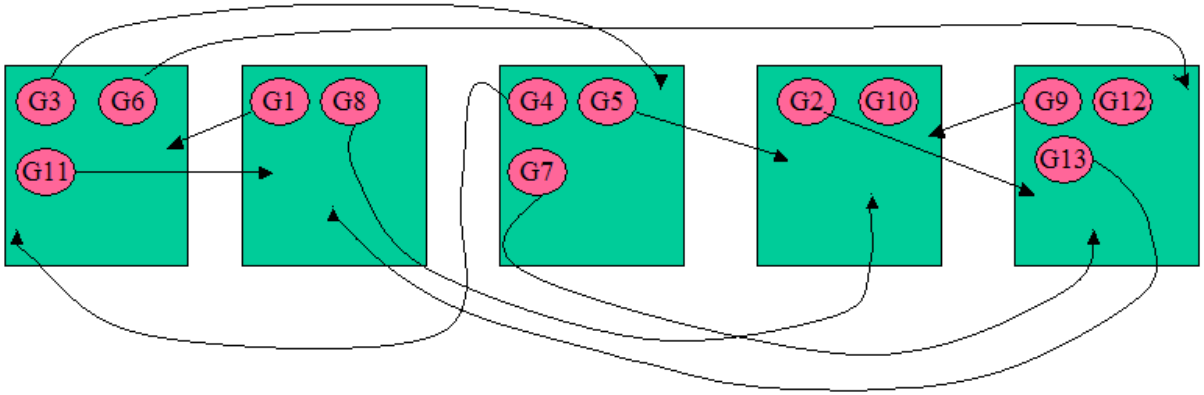


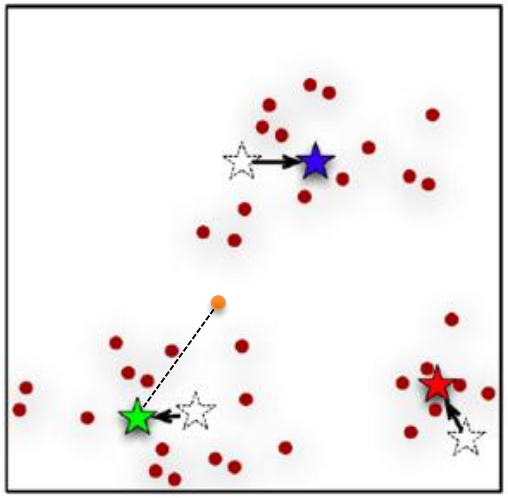
- Initialize *k* cluster centers. Iterate until convergence
- Assign each object to the cluster with the closest center (Euclidean distance)
- Calculate the mean of each cluster
- The centroids/mean vectors of the obtained clusters are taken as new cluster centers.

http://compbio.utmem.edu/MSCI814/Module10.htm
http://bigdata-madesimple.com/possibly-the-simplest-way-to-explain-k-means-algorithm/

K-Means Clustering 3

4. Shuffle genes among clusters such that each gene is now in the cluster whose mean / median expression profile (calculated in step 3) is the closest to that gene's expression profile.





5. Repeat steps 3 and 4 until genes cannot be shuffled around any more, OR a user-specified number of iterations has been reached.

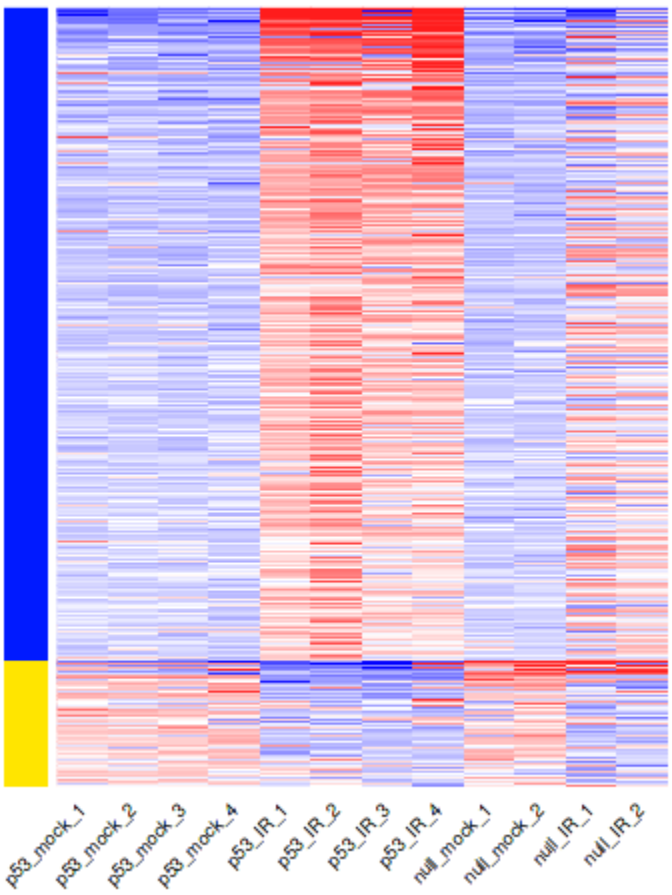
K-Means / K-Medians is most useful when the user has an a-priori hypothesis about the number of clusters the genes should group into.

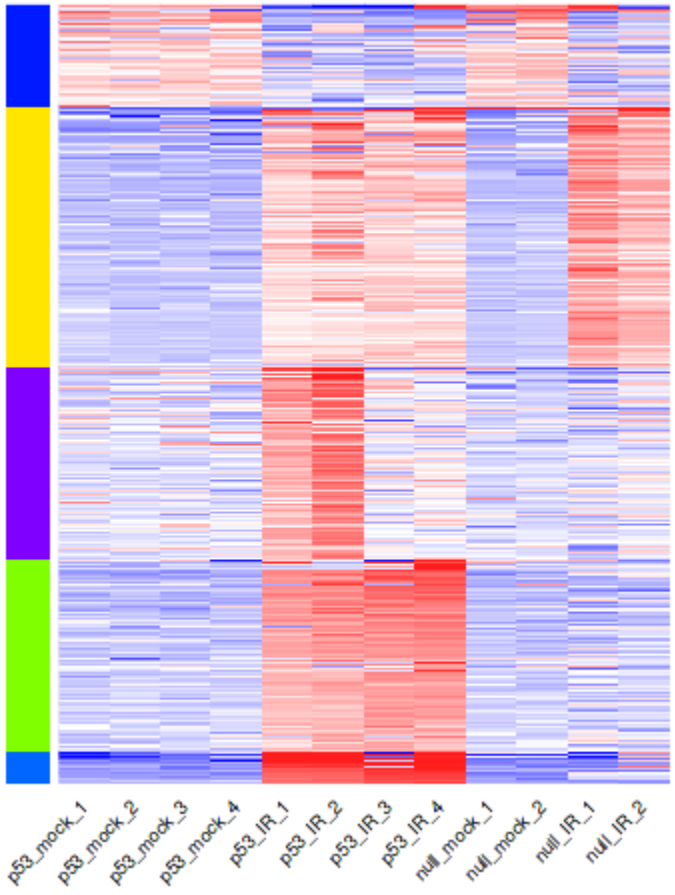
http://compbio.utmem.edu/MSCI814/Module10.htm
http://bigdata-madesimple.com/possibly-the-simplest-way-to-explain-k-means-algorithm/

Calculate K-means

- Example
- https://www.youtube.com/watch?v=wE8H-MEHSKs

How many clusters

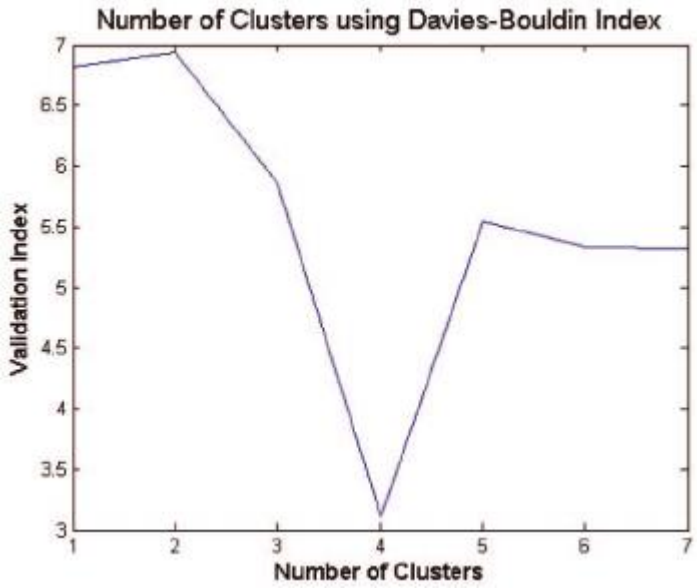




- Cluster homogeneity or within cluster variation
- Cluster separation or variance explained as a function of the number of clusters

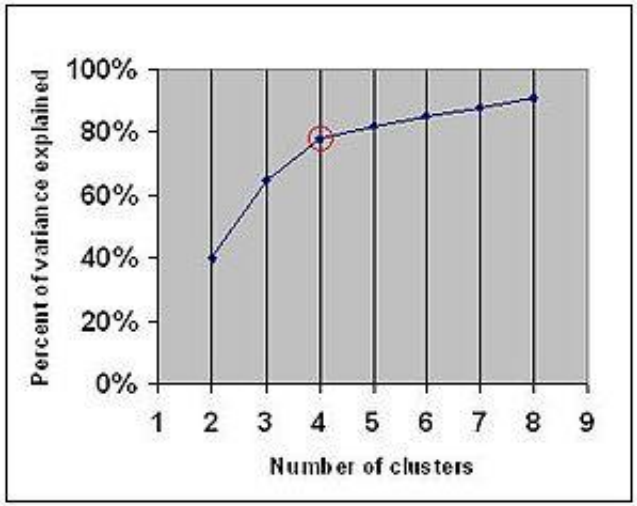
Methods to determine the cluster number in K-Means

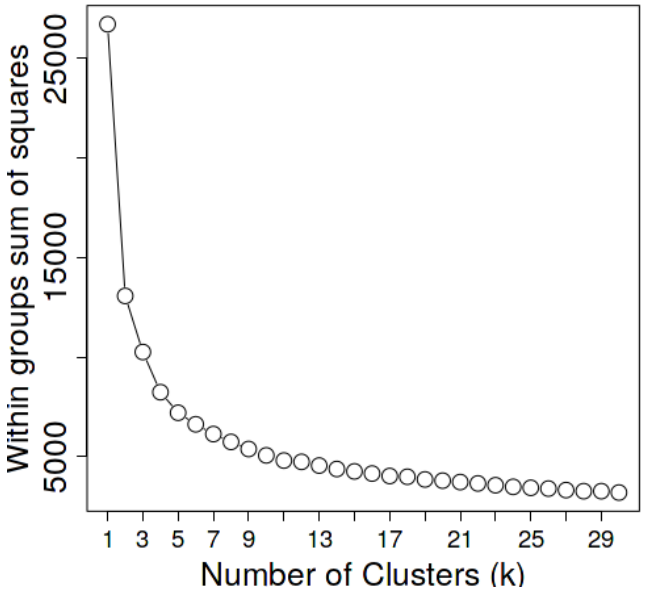
- Davies-Bouldin index – a function of the ratio of the within cluster variation, to the between cluster separation, a lower value will mean that the clustering is better.



Methods to determine the cluster number in K-Means (2)

- Elbow method - this method looks at the percentage of variance explained as a function of the number of clusters, at some point the marginal gain will drop, giving an angle in the graph: "elbow criterion". A slight variation of this method plots the curvature of the within group variance (within groups sum of squares).





Standardization

- **Standardization** rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance).
- For K-means, standardize before clustering.
- For hierarchical clustering, standardize after clustering.

Special considerations

r log transformation

- Transforms the count data to the log2 scale in a way which minimizes differences between samples for rows with small counts, and which normalizes with respect to library size (rld)

- Better suits for calculating Pearson correlation

About differentially expressed genes and clustering

• Sometimes, people first select genes that appear to be differentially expressed between groups of samples. Then they cluster the samples based on the expression levels of these genes. Is it remarkable if the samples then cluster into the two groups?

• No, this doesn’t prove anything, because the genes were selected with respect to the two groups! Such effects can even be obtained with a matrix of random numbers.

BE AWARE, IN THIS CASE YOU ARE NOT CLASSIFYING YOUR DATA, JUST VISUALIZING THE SUPERVISED ANALYSIS

Summary

- Discriminant analysis: Classes **known**

  Cluster analysis: Classes **not known**
- Choose the **distance measure** according to your **purpose**
- For hierarchical cluster algorithms it is not necessary to know the model (how many clusters)
- K-means is most useful when there is prior knowledge about the number of clusters or when defined clusters are required for further analysis
- It is difficult to assess the validity/significance of clustering. Even “random” data with no structure can yield clusters or exhibit interesting looking patterns.

Preliminary (exploratory) analysis

1. Hierarchical clustering

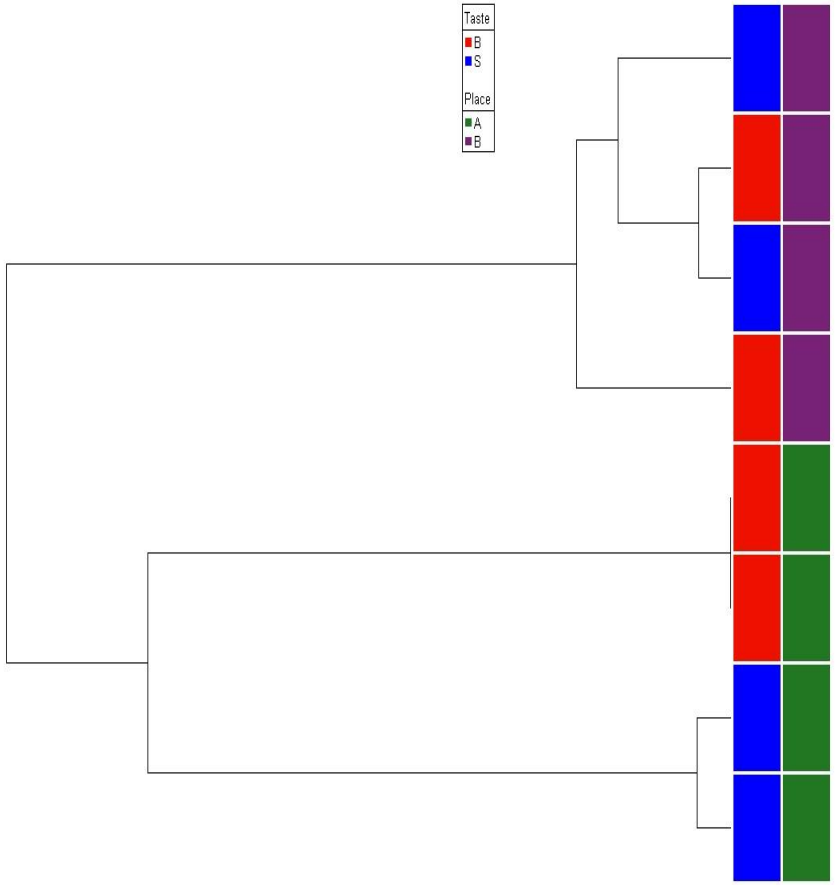
2. Principal component analysis (PCA)

Goals

- Exploratory analysis aims to find patterns in the data that aren't predicted by the experimenter's current knowledge or pre-conceptions.
- Identify groups of samples that are closely related or to find unknown subgroups among samples.
- Identify bad replicates or outliers
- To address these questions, researchers have turned to methods such as cluster analysis and principal components analysis.

Hierarchical clustering

Grouping a collection of objects into subsets or "clusters", such that those within each cluster are more closely related to one another.
The partition is performed according to the degree of similarity (or dissimilarity) between the individual objects being clustered.

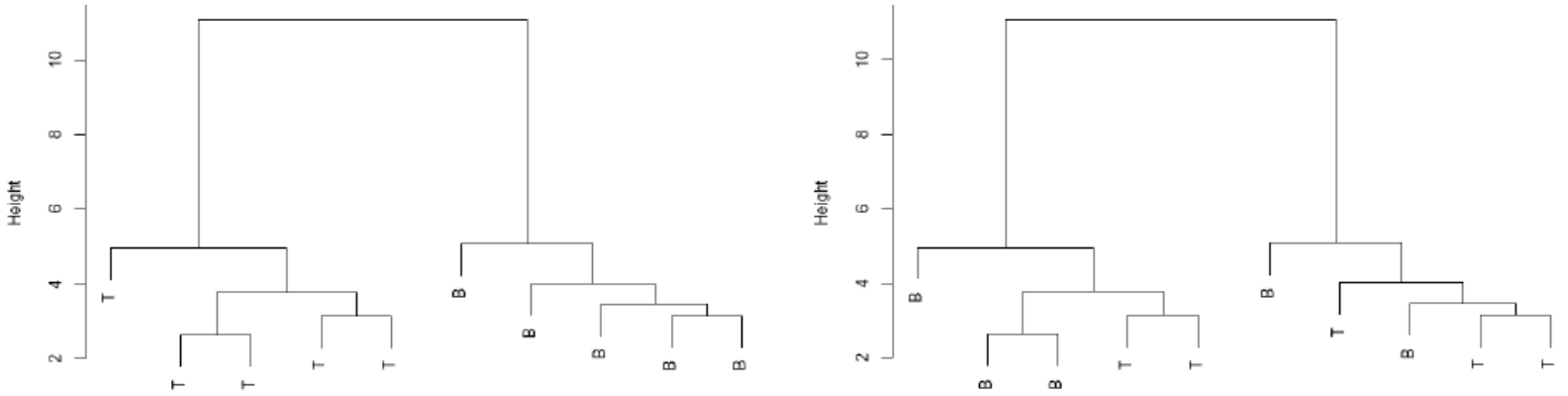


Sample clustering in exploratory analysis

- Clustering after supervised feature selection

  NO! Do not first select genes based on the outcome of some covariable (e.g. tumor type) and then look at the clustering.

  You will ALWAYS find difference w.r.t. your covariable, since this is how you selected the genes!



Left dendrogram obtained by

1. Random assignment of sample labels
2. Selection of best discriminating genes
3. Clustering with selected genes

Right plot shows original labels

If the data is very noisy, a practical proceeding is too choose the most variable genes

How to use clustering?

- Exploratory analysis

  Do not select genes based on a factor as genotype or treatment.

  If the data is noisy, you can select the most variable genes.

- Visualization of DE genes or finding common expression patterns for DE genes.

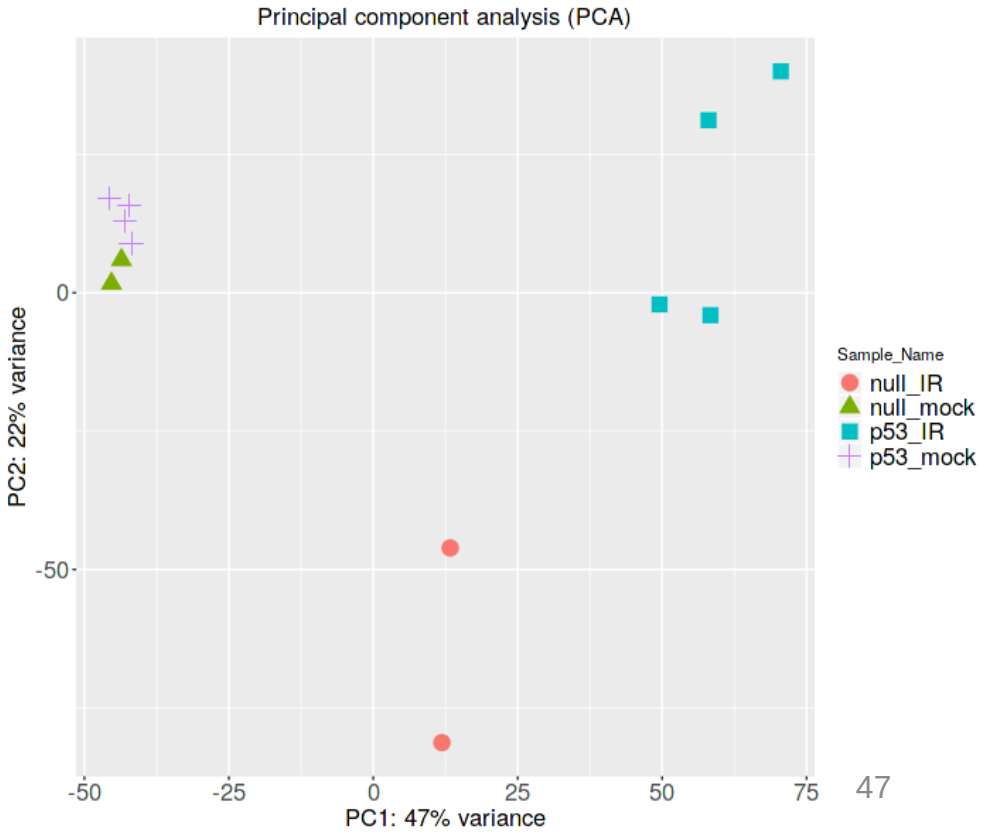
  Select genes. This is not a formal analysis.

  Do not be surprised if the samples separate according to the DE genes that were chosen.

Principal component analysis (PCA)

It is a technique used to reduce multidimensional data sets to lower dimensions for analysis. Iteratively, the direction with largest variance is selected as i-th principal component, in such a way that the maximum variability is visible.

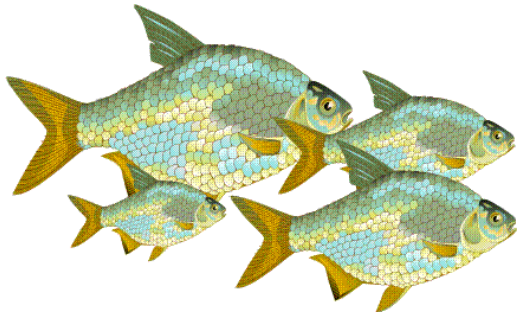
- Each shape represent a sample.
- The closest the samples are the more similar.
- The variance explained is shown



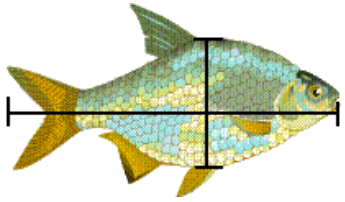
PCA – A graphical explanation

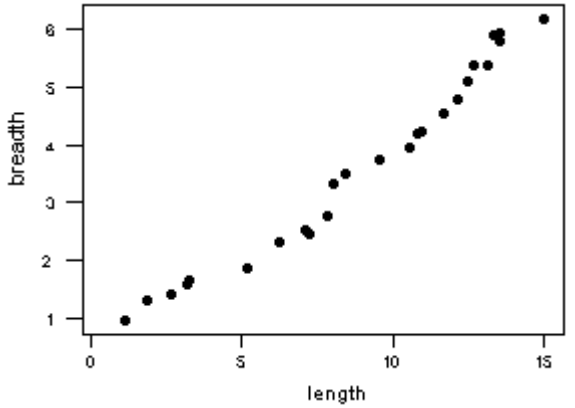
PCA is dimension reduction technique in that points in multidimensional space are projected onto a space of fewer dimensions

The orientation of the projection will aid our understanding of any relationships between the points



We could measure, for each fish, its length and breadth.

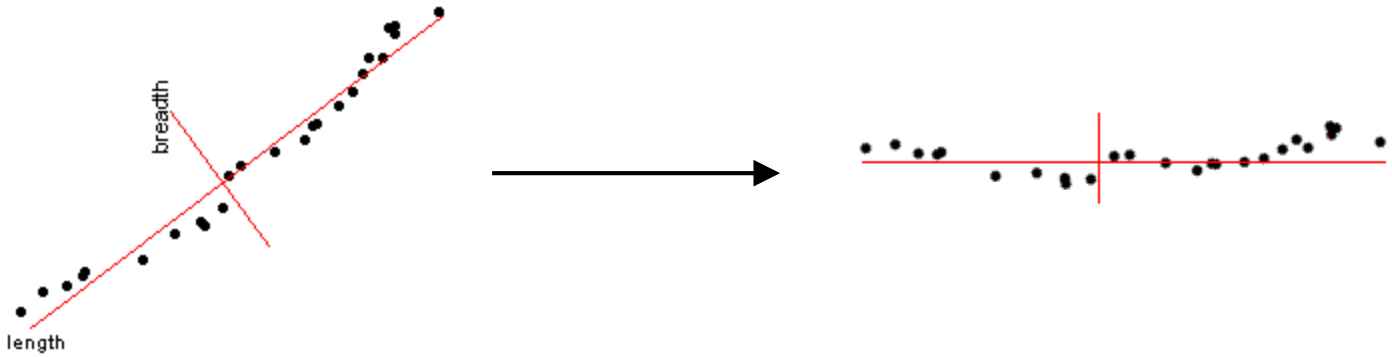




http://www.doe-mbi.ucla.edu/~parag/multivar/pca_graf.htm

PCA – A graphical explanation 2

- We can move and rotate the axes
- Moving the axes does not change the underlying pattern in the data



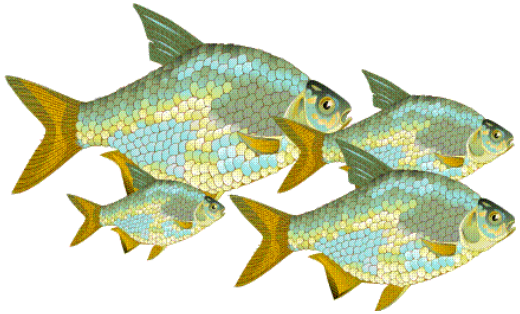
The new major axis: size = 0.75 x length + 0.25 x breadth

What about the second axis of the ellipse?

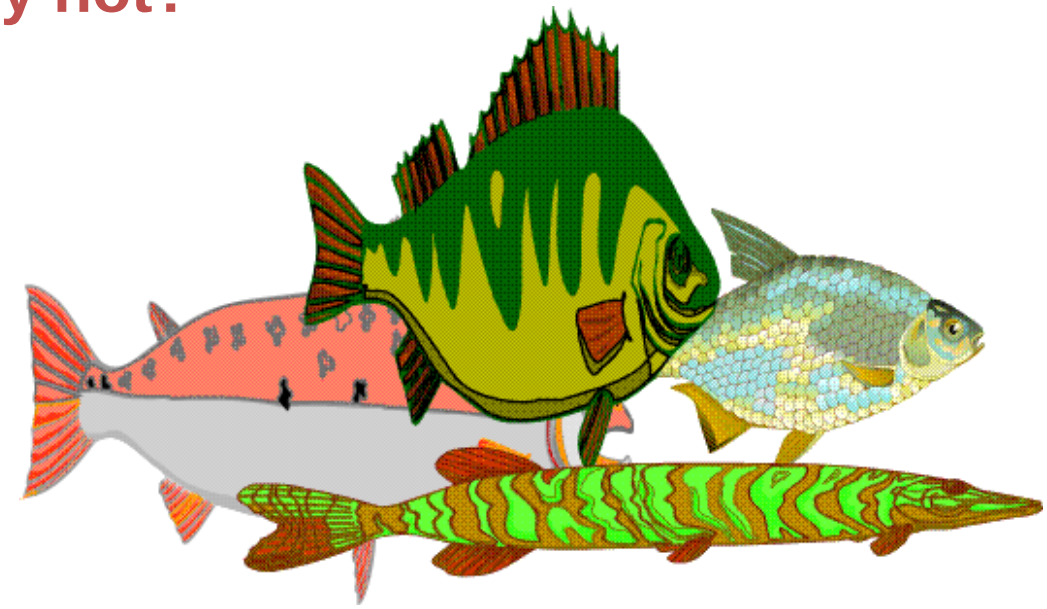
- It should account for as much of the remaining variation as possible
- It must also be uncorrelated (orthogonal) with the first

CONCLUSION

Since length and breadth are highly correlated, we can reduce the dimensionality of the data from two (length and breadth) to one (size), with little loss of information



Could you represent these fish by a size variable alone? If not, why not?



PCA in expression data

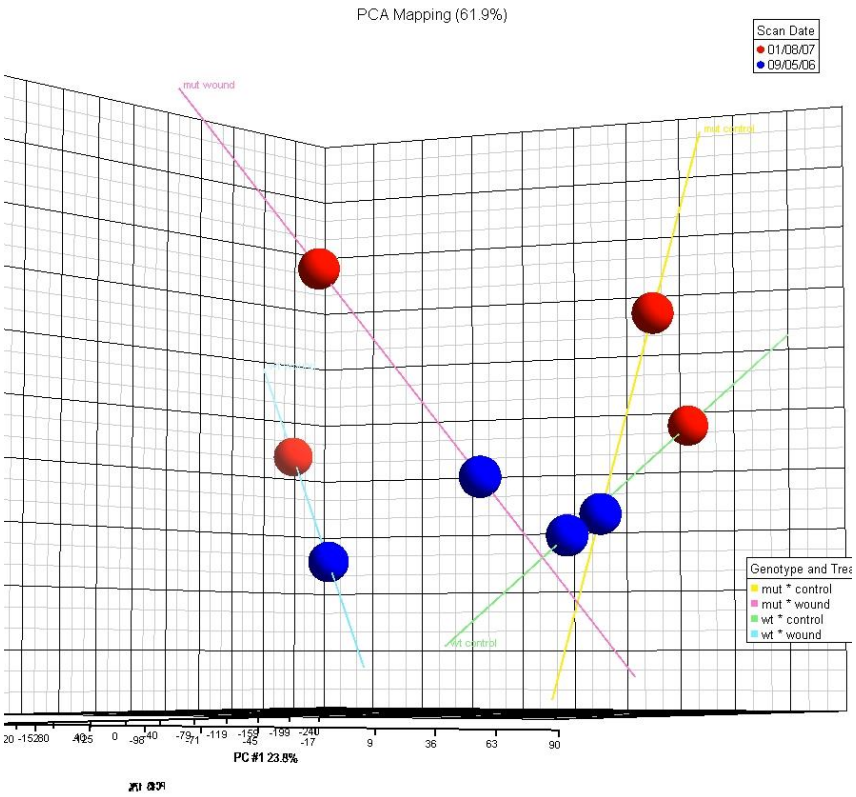
- Original data – a set of (possibly) partially correlated variables.
  - 20.000 gene expression values in several samples, some of them correlated.
- Output – a new set of uncorrelated variables, each of which is a linear combination of the original variables.
- The first new variable (or principal component) accounts for as much of the variation in the original data amongst all linear combinations of the original variables.
- $PC1 = a_1x_1 + a_2x_2 + a_3x_3 + \ldots$
- $PC2 = b_1x_1 + b_2x_2 + b_3x_3 + \ldots$

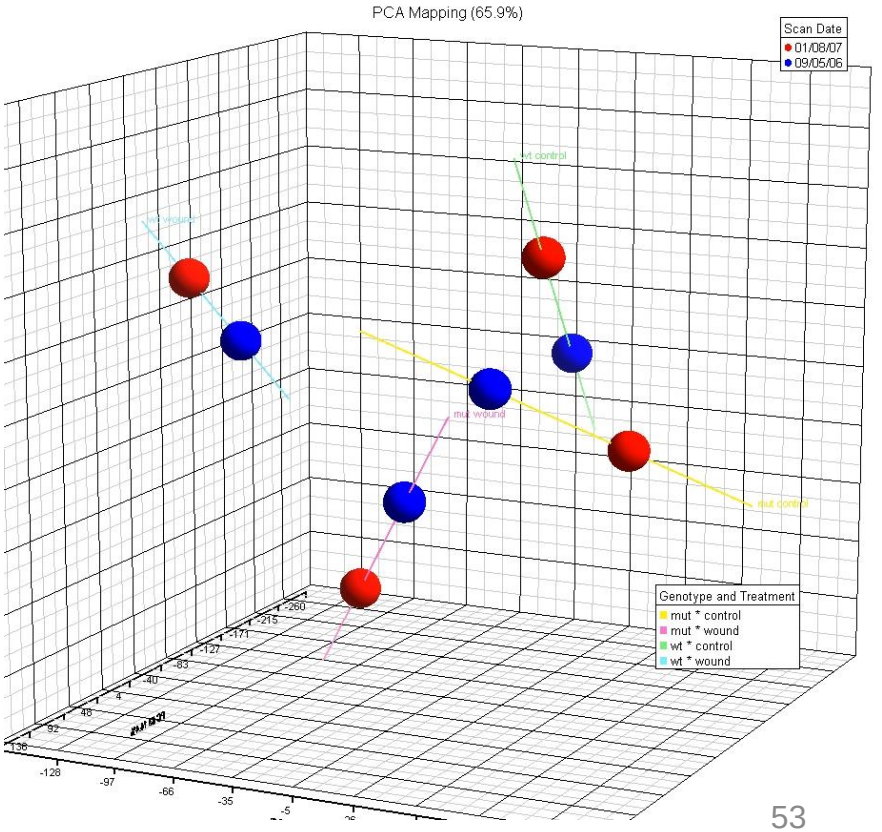
PCA is unsupervised

- PCA doesn't care about groups, and does not try to differentiate between treatments! It only cares about variance.

Batch effect

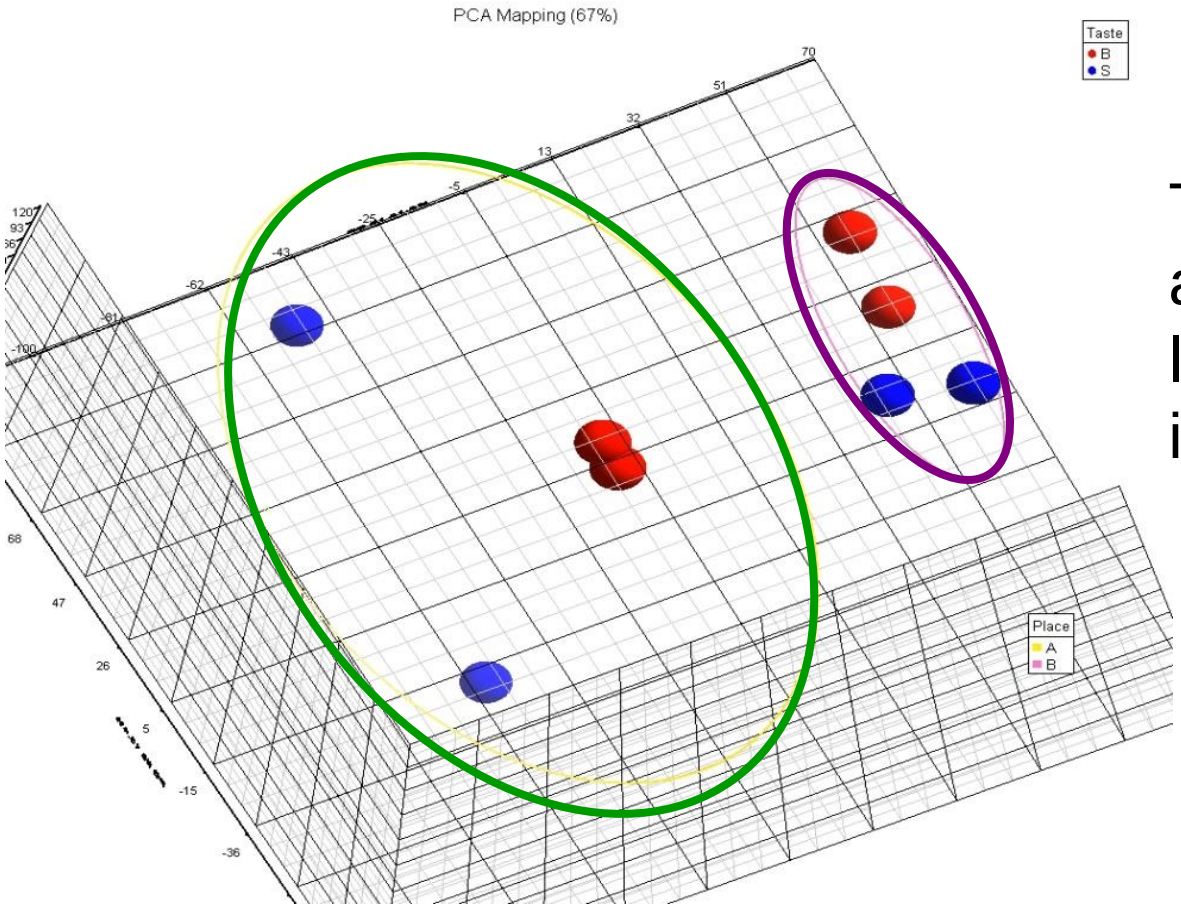
Non-biological experimental variation or *batch effects* are commonly observed across multiple batches of experiments.





PCA with 2 factors

Experiment: Influence of cultivation location and tomato strain on gene expression



The separation is according to the location, therefore its influence is stronger.


THE END


Thank you
Gracias
Wa'-do
Shukran
Komapsumnida
Spasibo
Cheets
Sipas
Xie xie
Thenk ye
Danke
Cám ón
Merci