Distance and Clustering

XDAS 2019

Kris Gunsalus

Outline

- Review
  - Normalization
  - Differential expression
  - FDR
  - Distance measures
- Clustering
  - Hierarchical
  - K-means

⇒*ALSO SEE LECTURE NOTES ON CLUSTERING!*

Genome-wide expression analysis

- Goal: to measure RNA levels of all genes in a genome under various experimental conditions
- RNA levels vary with:
  - Cell type
  - Developmental stage
  - External stimuli
  - Disease state
- Time and location of expression provide information on genes' function and interactions, and can be useful for many purposes, including disease diagnostics and medical applications.

Normalizing the data

Sensitivity of RNA-seq is a function of

- Number of reads per transcript
- Transcript length
- Size of the library (total # of reads)

FPKM (Fragments per Kilobase of exons per million reads)

$$FPKM = \frac{R}{NT}$$

TPM (Transcripts per million reads)

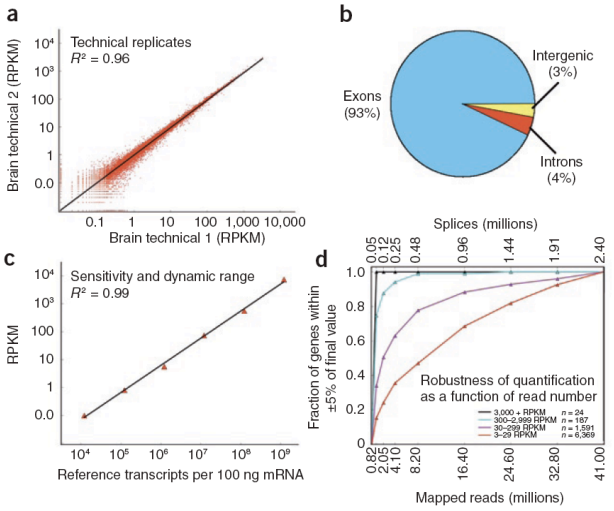
$$RPK = R/N$$

$$TPM = \frac{RPK}{\sum RPK / 1M}$$

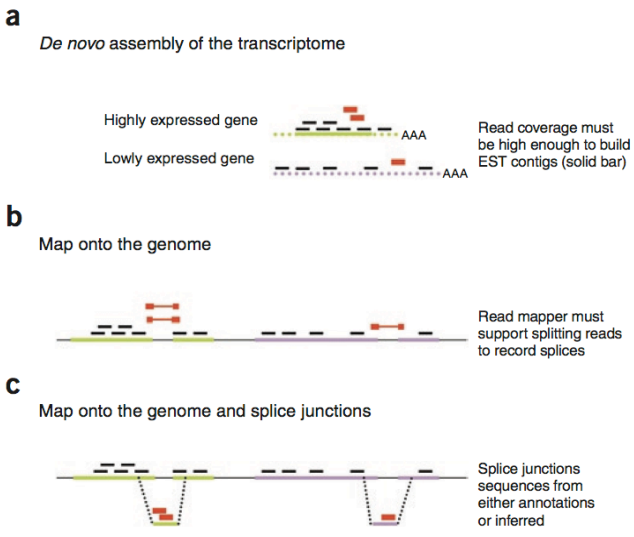
$$TPM = \frac{RPK}{\sum RPK} * 1M$$

R = # of unique reads for the gene
L = length of the gene (sum of exons in nt)
N = Size of the gene in kb (sum of exons / 1000)
T = total number of reads in the library mapped to the genome / 1,000,000

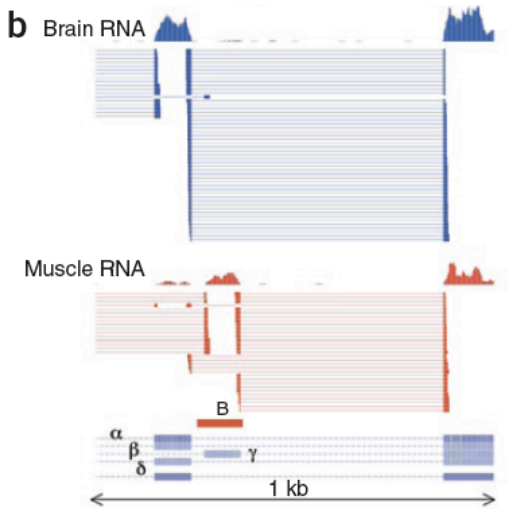
Reproducibility, linearity and sensitivity



RNA-seq provides more information than just expression level



Candidate new and revised exons



NATURE METHODS | VOL.5 NO.7 | JULY 2008 | 621

Evaluating expression differences: Which Looks Better ?

Gene X

A

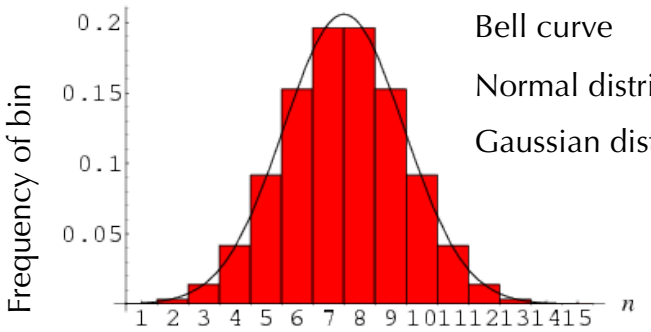
or

	e.g. control	e.g. treat
Rep1	20	30
Rep2	21	29
Rep3	19	31
Avg	20	30

B

	e.g. control	e.g. treat
Rep1	20	40
Rep2	30	30
Rep3	10	20
Avg	20	30

Multiple Measurements of the Same Genes Expression

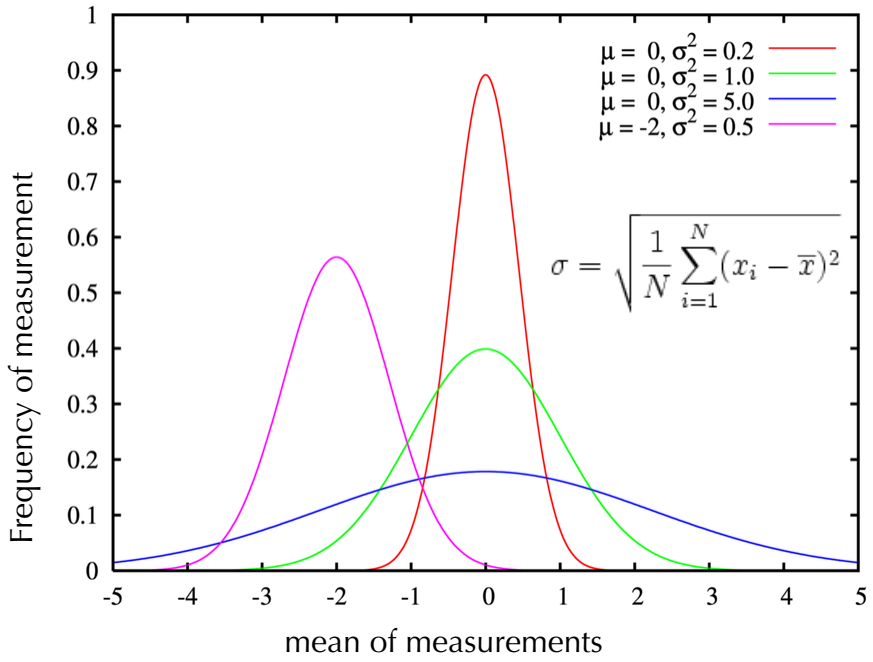


Bell curve

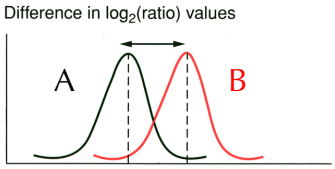
Normal distribution

Gaussian distribution

gene expression level bin

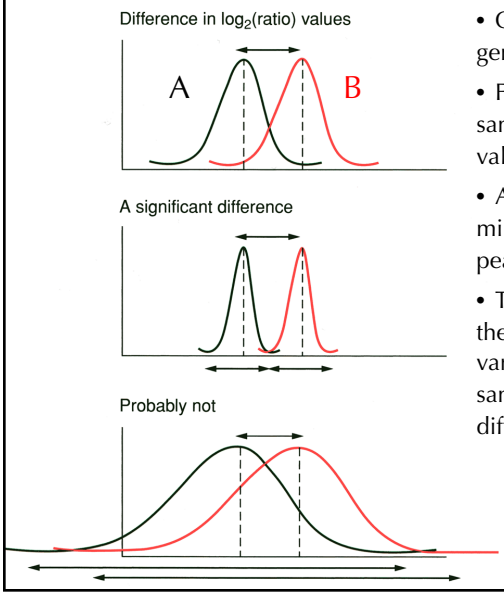


What is a "significant" difference in gene expression ?



- Consider the expression level of a single gene measured in two conditions (A vs B).
- For each gene, multiple replicates of the same condition measure a distribution of values centered on the mean.
- A fixed fold-change cutoff asks for a minimum separation between the average peaks in A and B.

What is a "significant" difference in gene expression ?



- Consider the expression level of a single gene measured in two conditions (A vs B).
- For each gene, multiple replicates of the same condition measure a distribution of values centered on the mean.
- A fixed fold-change cutoff asks for a minimum separation between the average peaks in A and B.
- The difference in $\log_2$(ratio) values between the means in both samples relative to the variability of measurements within each sample will determine whether the observed difference is significant.

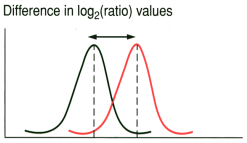
$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between groups}}{\text{variability with groups}}$$

What is a "significant" difference in gene expression ?

- A simple statistical test of significance is the Student's t test.
- The t test statistic can be used to asses the signal-to-noise ratio for an observed difference in expression of a particular gene in two experimental conditions.

Average $\log_2$(ratio) in condition A

Standard deviation of the mean

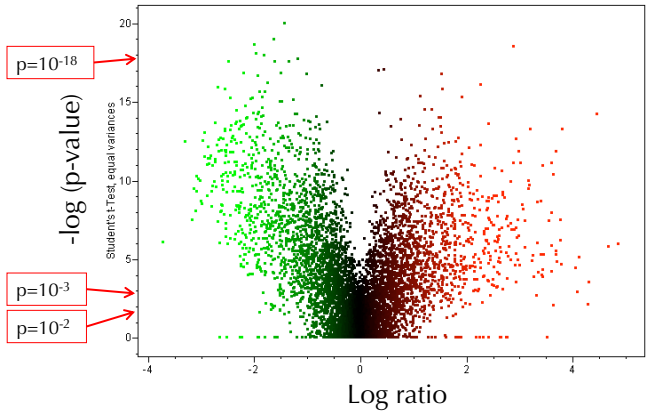


$$\bar{X}_A = \frac{1}{N_A}\sum_{i-1}^{N_A} X_{Ai} \qquad \sigma_A = \sqrt{\frac{\sum_{i-1}^{N_A} [X_{Ai} - \bar{X}_A]^2}{N_A}}$$

$$t = \frac{\text{signal}}{\text{noise}} = \frac{\text{difference between groups}}{\text{variability with groups}} = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}$$

Volcano plot

fold-change vs. significance



- A volcano plot is a scatter plot of -log (p-value) from a t-test or one-way ANOVA, versus log ratio. It allows you to visualize fold-change and statistical significance at the same time, so that one can find genes that are significant and have large fold change, or genes that are significant but have small fold change.

⇒ ***P-values should be corrected for multiple hypothesis testing, such as with FDR, to control for Type I errors***

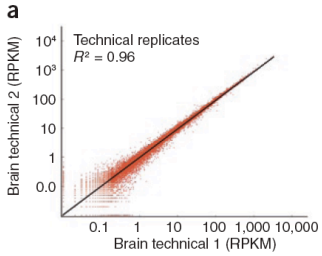
Common Analysis Tasks

Pattern Analysis

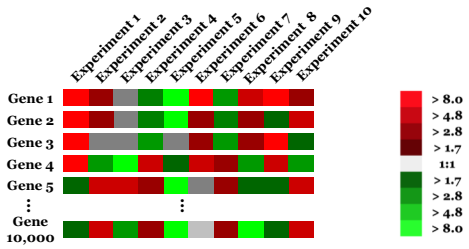
- Identify up- and down-regulated genes.
- Find groups of genes with similar expression profiles.
- Find groups of experiments (tissues) with similar expression profiles.
- Find genes that explain observed differences among tissues (feature selection).

Characteristics of RNA-seq Data

- Extremely high dimensionality
  - Experiment = ($gene_1$, $gene_2$, …, $gene_N$)
  - Gene = ($experiment_1$, $experiment_2$, …, $experiment_M$)
  - N is often on the order of 10^4
  - M is often on the order of 10^1
- Noisy or missing data
  - Very lowly expressed genes are detected less reproducibly
  - Especially relevant to limited sample sizes, e.g. single cell analysis



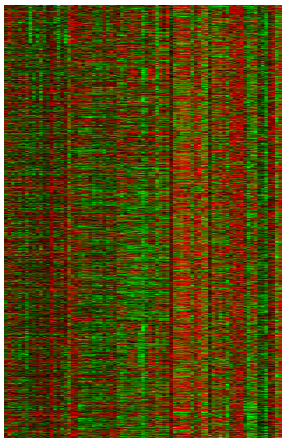
Gene expression profiling



How can we find patterns in the data?

Gene expression matrix

Experiments (*j*)



Genes (*i*)

The matrix entry at (*i, j*) is the expression level of gene *i* in experiment *j*.

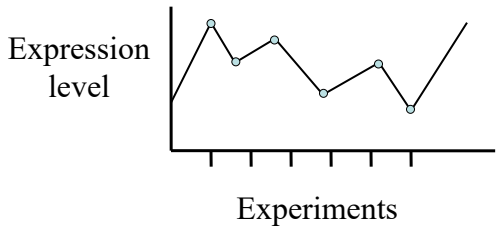
Experiments could be:

- Time series
- Different treatments
- Different tissues
- …

Types of analysis

- Unsupervised learning: learn from data only
  - visualization: find structure in data
  - clustering: find clusters/classes in data
- Supervised learning: learn from data plus prior knowledge
  - regression: predict a real value
  - classification: predict discrete classes
    - SVM, random forests, Bayes, KNN, neural networks

A series of experiments

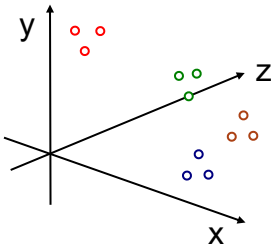


A 2-D plot of expression level for a single gene in many different conditions.

The data points are connected by lines just to help visualize the changes in level between conditions.

Gene expression in multiple dimensions

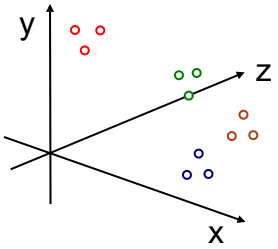
Consider 3 experiments: x, y, and z

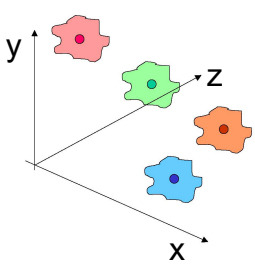


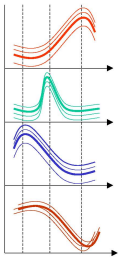
- The expression vector for each gene can be represented as a point in 3-dimensional space, in which each axis represents a different condition.
- Genes with similar expression patterns fall nearby one another in this multi-dimensional space.

Gene expression in multiple dimensions

Consider 3 experiments: x, y, and z







- The expression vector for each gene can be represented as a point in 3-dimensional space, in which each axis represents a different condition.
- Genes with similar expression patterns fall nearby one another in this multi-dimensional space.
- Genes with similar expression profiles are likely to have common or related functions, and possibly to be co-regulated.
- Similarly, conditions can be classified into different groups based on similarities in their expression profiles (all or subsets of genes).

Coordinated gene expression

Which genes are co-expressed?

- Hierarchal clustering
- K-means clustering
- Self-organizing maps
- Principal component analysis

Root of clustering approaches: a pairwise matrix of distances

	gene 1	gene2	gene 3
gene 1	1	0.5	0.8
gene 2	.	1	0.6
gene 3	.	.	1

This matrix describes all the pairwise relationships (distances) between the elements you are trying to group (genes in this case)

But how to define distance?

Calculating Distance

- Distance is the most natural method for numerical data
- Lower values indicate more similarity
- Distance metrics
  - Euclidean distance
  - Manhattan distance
  - Etc.
- Does not generalize well to non-numerical data
  - What is the distance between "male" and "female"?

Distance Measures

- Euclidian distance metric

  Pythagorean theorem: $a^2 = b^2 + c^2$

  Euclidian distance in 3 dimensions between two points, $x=(x_1,x_2,x_3)$ and $y=(y_1,y_2,y_3)$:

  $$d_{12} = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + (x_3-y_3)^2}$$

  In n-dimensions:

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

- Pearson correlation and Pearson distance (semi-metric)

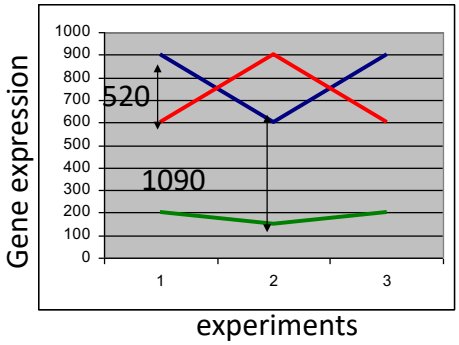
  $$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} \qquad -1 \le r \le 1$$

  $$d = 1 - r \qquad 0 \le d \le 2$$

  High degree of similarity implies a small distance and vice versa

Euclidean distance

$$dE(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



Implication for gene expression:
the ***magnitude*** of expression values will determine distances

Covariance and Correlation

Start with the concept of covariance:

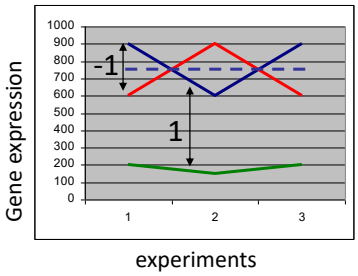
$$Cov_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}$$

But … covariance ranges from -∞ to +∞

Normalize the measure using the variance of two measurements, VarX and VarY

Pearson correlation coefficient

$$r = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}}{\sqrt{(VarX)(VarY)}}$$



Pearson correlation has the nice property of varying between -1 and 1

Implication for gene expression:
*the **shape** of gene expression responses will determine similarity*

Grouping Objects: Clustering

Given a collection of objects, put objects into groups based on similarity.

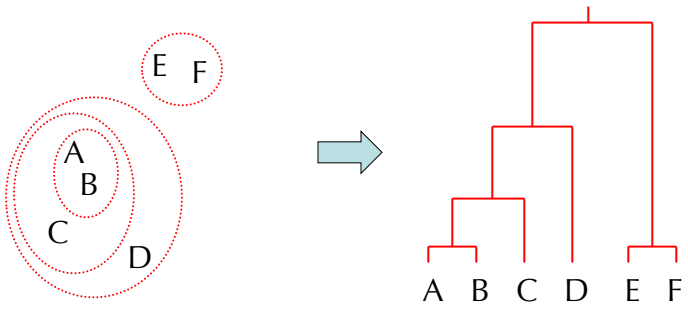
- Grouping complex entities such as expression data can be a fuzzy problem.
- Expression data are complex because each gene can have a value for many experiments ("high dimensionality")

Clustering approaches

- Agglomerative: hierarchical
- Divisive: partitioning methods

Hierarchical Clustering

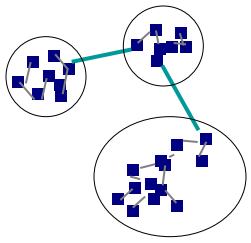
- Find the pair(s) with the highest pairwise similarity (***distance*** *measure*)
- ***Join*** these as a group and calculate an "average" profile (single, average, or complete linkage)
- Iteratively join groups until all are ***linked***



Linkage Methods

Single linkage:

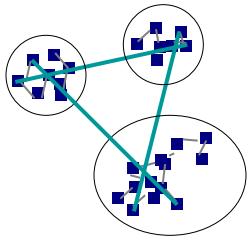
Use the distance between the closest two points between each pair of clusters



Linkage Methods

Complete linkage:

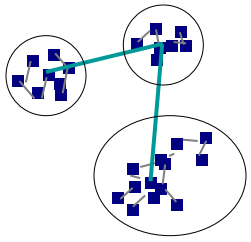
Use the distance between the furthest two points between each pair of clusters



Linkage Methods

Centroid linkage:

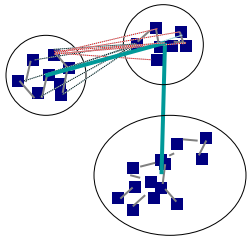
- Find the central point within each cluster based on all pairwise differences between them
- Use the distance between the centroids between each pair of clusters



Linkage Methods

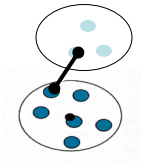
Average linkage:

Use the average distance between
each pair of points
between each pair of clusters



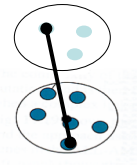
In phylogenetics, UPGMA (unweighted pair-group method with arithmetic means) uses average linking.

Summary: Linkage Methods



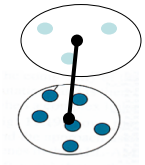
Single linkage

Minimum distance



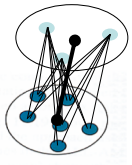
Complete linkage

Maximum distance



Centroid linkage

Mean distance



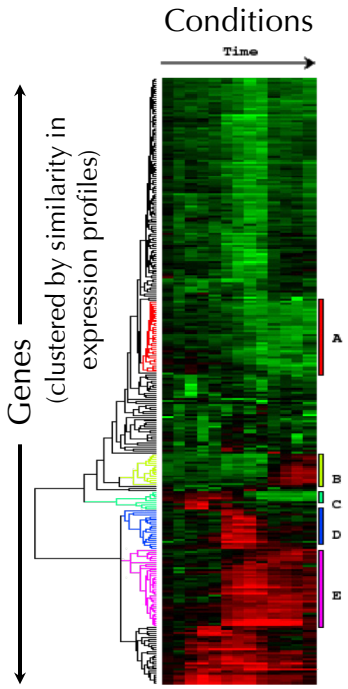
Average linkage

Average pair-wise distance

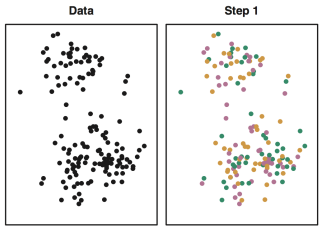
End Result

- Place genes with similar expression profiles into clusters.
- Similarity is defined by Pearson correlation.

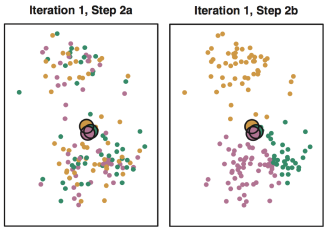
Genes are grouped according to similarities in their expression levels across a variety of conditions.



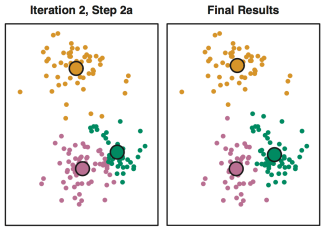
K-means: Example, k = 3



Step 1: Choose k and assign points randomly to different groups.

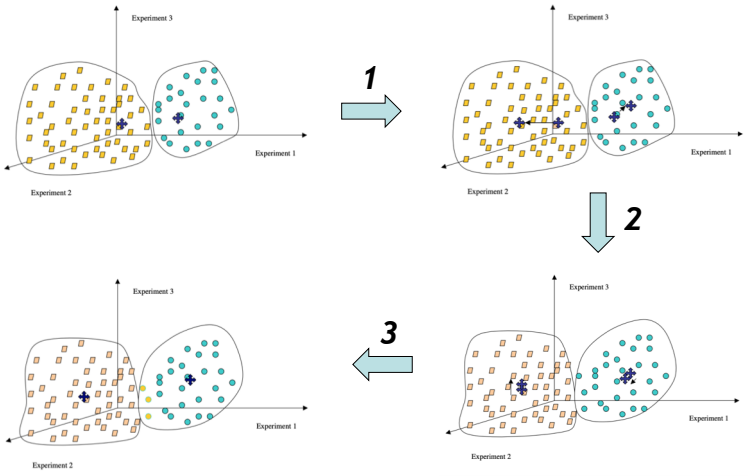


Step 2: Compute centroids (big dots) and reassign points to nearest centroids



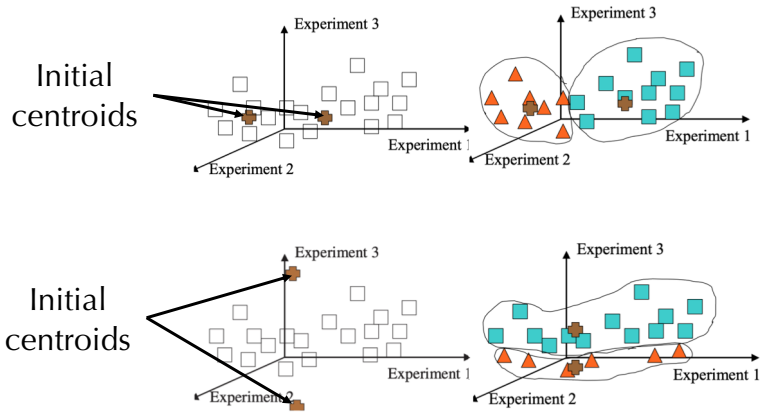
Step 3: Re-compute centroids, repeat until stable (right: after 10 iterations)

K-means in action:
tends to create round clouds



Source: Sorin, Drăghici. Statistics and Data Analysis for Microarrays Using R and Bioconductor, Second Edition by Chapman and Hall/CRC Series: Chapman & Hall/CRC Mathematical and Computational Biology, 2016

K-means: Weaknesses
Can give you a different result each time with exactly the same data



Source: Sorin, Drăghici. Statistics and Data Analysis for Microarrays Using R and Bioconductor, Second Edition by Chapman and Hall/CRC Series: Chapman & Hall/CRC Mathematical and Computational Biology, 2016

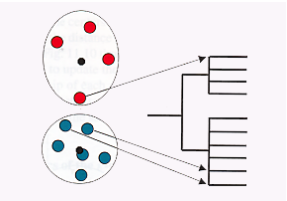
K-means: Weaknesses

- Must choose parameter k in advance, or try many values.
- Data must be numerical and must be compared via Euclidean distance (there is a variant called the k-medians algorithm to address these concerns)
- The algorithm works best on data which contains spherical clusters; clusters with other geometry may not be found.
- The algorithm is sensitive to outliers -- points which do not belong in any cluster. These can distort the centroid positions and ruin the clustering.

Clustering has no one answer

- Given a collection of objects, put objects into groups based on similarity.
- It really depends on how you measure similarity/dissimilarity

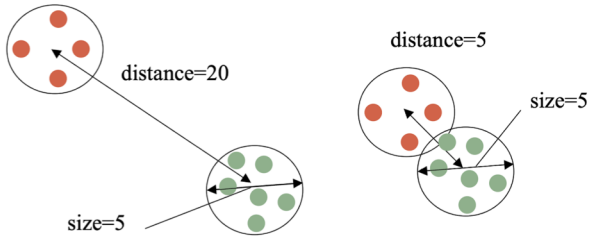
Problem: Sometimes genes with pretty similar expression can end up in different clusters!



Judging Clustering Quality: Silhouette width

Ideally, we want well separated, distinct groups

- Maximize ***between***-cluster distance
- Minimize ***within***-cluster distance



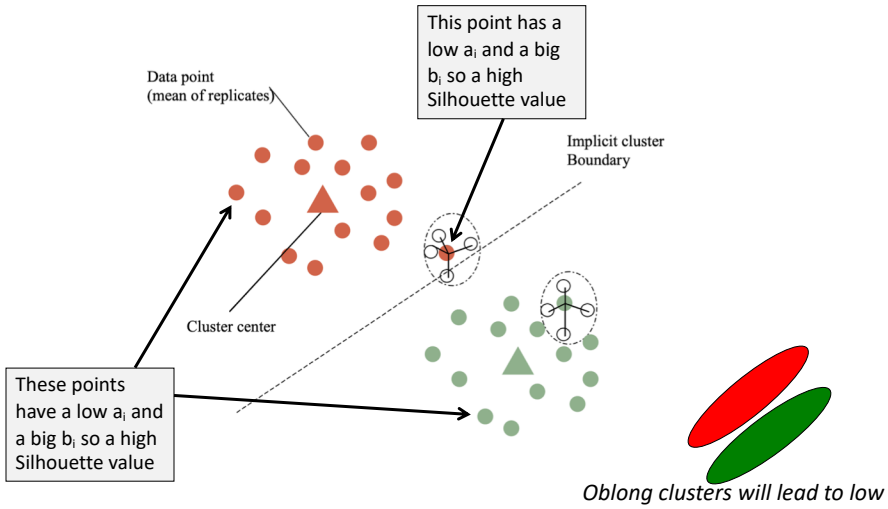
$$\mathbf{Sil}_i = (b_i - a_i)/\max(a_i, b_i)$$

a_i: average within cluster distance with respect to gene i

b_i: average between cluster distance with respect to gene i

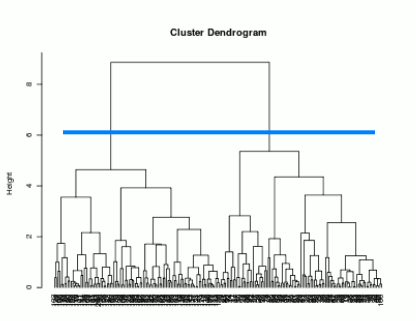
Source: Sorin, Drăghici. Statistics and Data Analysis for Microarrays Using R and Bioconductor, Second Edition by Chapman and Hall/CRC Series: Chapman & Hall/CRC Mathematical and Computational Biology, 2016

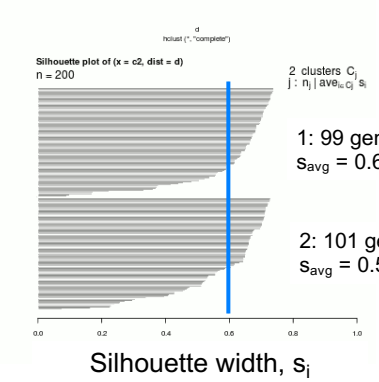
Measuring the Quality of Clusters



Oblong clusters will lead to low or negative Silhouette widths

Silhouette plots



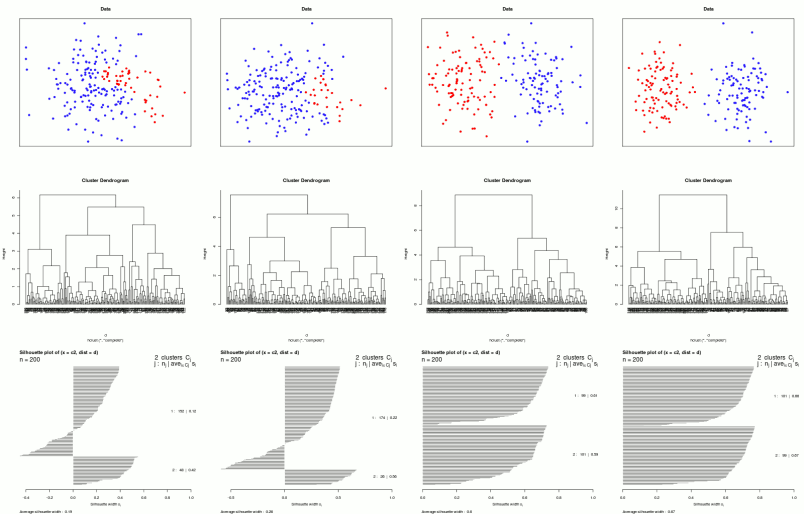


Silhouette width, s_i

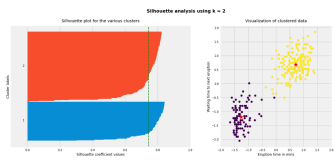
Average silhouette width: 0.6

Silhouette plots

Four different datasets:

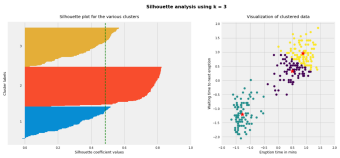


Another example



k = 2

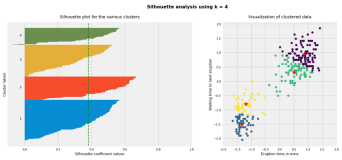
The partitioning with k = 2 has the highest average silhouette width, and thus provides the most distinct clusters.



k =3

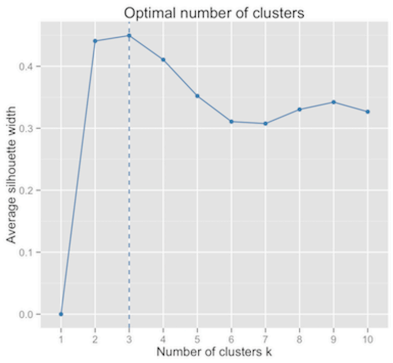
You may have additional data, however, suggesting that there really are more than 2 groups

(e.g. single-cell data in which the yellow and purple clusters can be distinguished based on coherent expression of cell-type-specific markers / gene sets)

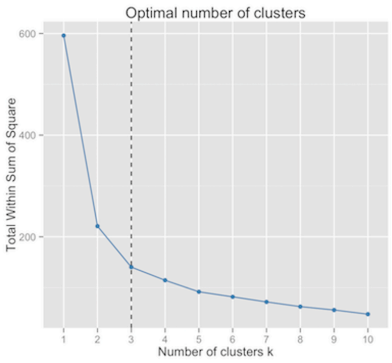


k = 4

Choosing the right number of clusters



Maximum average silhouette width



Elbow method