1. Introduction
   - Gene regulation
   - Genomics and genome analyses
   - Hidden Markov models (HMM)

2. Gene regulation tools and methods
   - Regulatory sequences and motif discovery
   - TF binding sites, MicroRNA target prediction
   - Databases

3. Technologies
   - Microarrays
   - Deep sequencing
   - Single cell RNAseq spatial transcriptomics

4. Clustering
   - Unsupervized clustering (HCA,K-means, PCA, SOM)
   - Supervized clustering (classification)

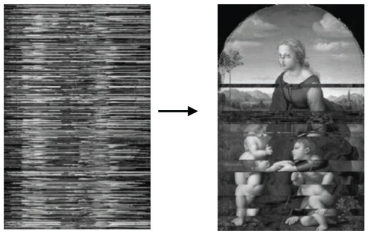
5. Gene ontology, Pathways, Enrichment analysis
   - Databases and tools
   - Gene set enrichment analysis

6. Biomolecular networks
   - Small world networks
   - Topology and parameter
   - Network motifs

Organize data

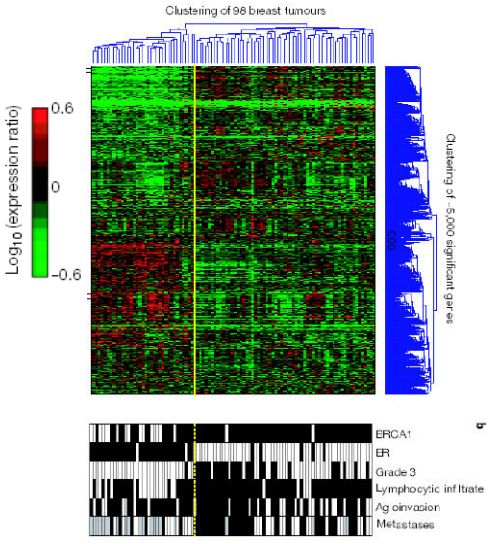
grayscale random order slices

clustering algorithms (hierarchical clustering)



Sherlock G, Kishan M, Narisamhan S

Gene expression profiling in breast cancer



Van t 'Veer et al. *Nature* 415:530-536, 2002

Clustering

- Unsupervized clustering
  - Hierarchichal Clustering
  - K-Means Clustering
  - Self organizing maps
  - Principal Component Analysis (PCA)
- Supervized clustering (Classification)
  - Support vector machines (SVM)
  - Logistic regression
  - ROC curve
  - Cross validation

Clustering

- Agglomerative

  Bottom up approach, whereby single expression profiles are successively joined to form nodes.

- Divisive

  Top down approach, each cluster is successively split in the same fashion, until each cluster consists of one single profile.

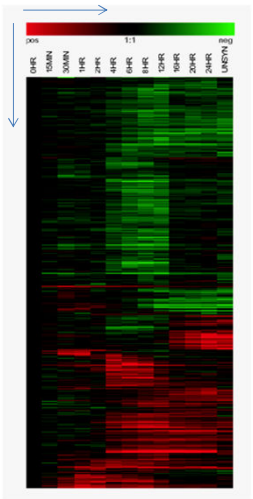
Methods for unsupervised clustering

- Hierarchical Clustering
- K-means
- Self Organizing Maps
- Principal component analyses (PCA)

Representation of gene expression

conditions (patients) (m)

genes (n)



heatmap

n x *m* matrix with *n* genes and *m* samples

- Representation as heatmap (e.g. *red* upregulated genes, green down regulated genes, *black* no change)

For experiments in reference design:

- log2-fold change (log2FC, log2(A/B), log2 ratio)

For patient samples and no reference:

- Mean (median) centered log2-levels for each gene
  log2-intensities for one-color arrays
  log2-RPKM for RNAseq
- z-score of log2-levels
  Z= (X-m)/s m…mean,
  s…standard deviation

How do we compare expression profiles?

- Treat expression data for a gene as a multidimensional vector.
- Use a distance/correlation metric to compare the vectors.
- Similarity (distance) measures:
  - Pearson correlation
  - Spearman rank correlation
  - Euclidean distance
  - Manhattan distance
  - Mutual information
  - …

Similarity/distance measures

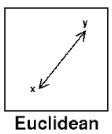
- Pearson correlation

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

$-1 \leq r \leq 1$

- Euclidian distance

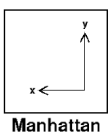
$$d_E = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



Euclidean

- Manhattan distance

$$d_M = (\sum_{i=1}^{n} |x_i - y_i|)$$

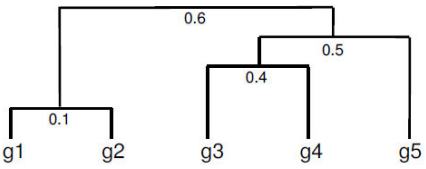


Manhattan

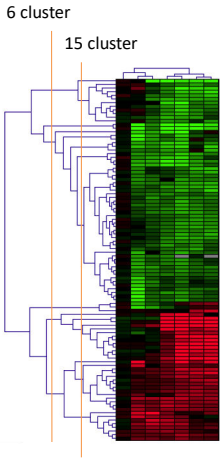
Hierarchical clustering

- Agglomerative (bottom up), unsupervized
- Cluster genes or samples (or both= biclustering)
- Distances are encoded in dendogram (tree)
- Cut tree to get clusters
- Pearson correlation (usually used)
- Computational intensive (correlation matrix)

1. Identify clusters (items) with closest distance
2. Join to new clusters
3. Compute distance between clusters (items) (see link
4. Return to step 1

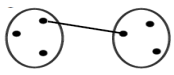


dendogram

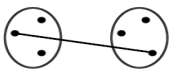


Linkage

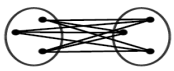
- Single-linkage clustering
  Minimal distance



- Complete-linkage clustering
  Maximal distance



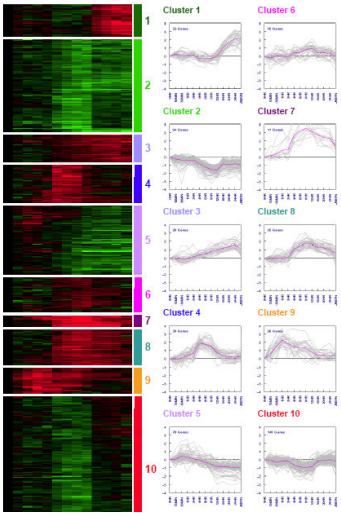
- Average-linkage clustering
  Calculated using average distance (UPGMA)
  Average from distances not! expression values



K-means

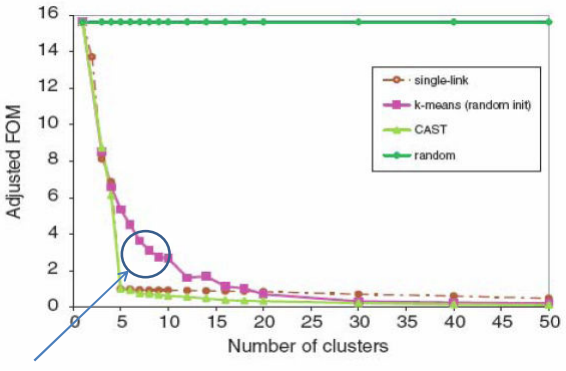
- partition *n* genes into *k* clusters, where *k* has to be predetermined
- k-means clustering minimizes the variability within and maximize between clusters
- Moderate memory and time consumption

1. Generate random points ("cluster centers") in n dimensions (results are depending on these seeds).
2. Compute distance of each data point to each of the cluster centers.
3. Assign each data point to the closest cluster center.
4. Compute new cluster center position as average of points assigned.
5. Loop to (2), stop when cluster centers do not move very much.



How to choose k

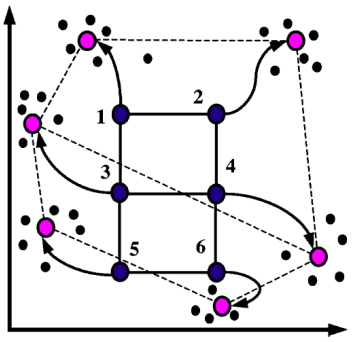
Figure of Merit (FOM)



choose k here (e.g. k=8)

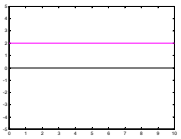
Self organizing maps (SOM)

- Neural network approach
- Usually one or 2D map
- Hexagonal or rectangular net topology
- Moderate memory and time consumption
- Number of clusters has to be specified!

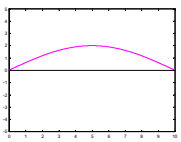


Principal Component Analysis (PCA)

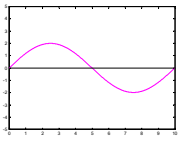
Is it possible to represent each profile by overlay of few patterns?



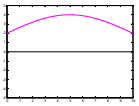
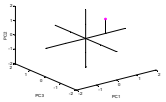
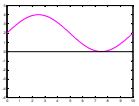
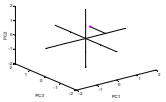
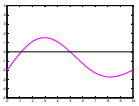
PC1



PC2



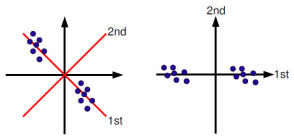
PC3

Description	Gene Expression	Point in PC Space
PC1 + PC2		
PC1 + PC3		
-PC1 + PC2 +PC3		

Principal component analysis (PCA)

PCA is a data reduction technique that allows to simplify multidimensional data sets into smaller number of dimensions (r<n).

Variables are summarized by a linear combination to the principal components. The origin of coordinate system is centered to the center of the data (mean centering) . The coordinate system is then rotated to a maximum of the variance in the first axis.

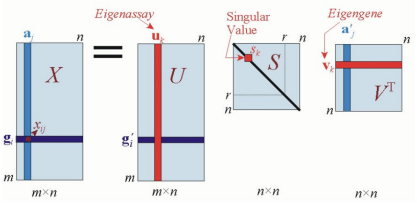


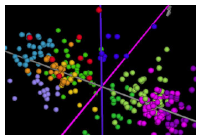
Subsequent principal components are orthogonal to the 1st PC. With the first 2 PCs usually 80-90% of the variance can already be explained.

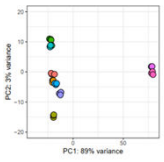
This analysis can be done by a matrix decomposition (singular value decomposition SVD).

Singular value decomposition (SVD)

$$X = USV^T \text{ with } UU^T = V^TV = VV^T = I$$







For mean centered data the Covariance matrix C can be calculated by XX^T. U are eigenvectors of XX^T and the eigenvalues are in the diagonal of S defined by the characteristic equation $|C - \lambda I| = 0$.

Transformation of the input vectors into the principal component space can be described by $Y = XU$ where the projection of sample i along the axis is defined by the j-th PC:

$$y_{ij} = \sum_{t=1}^{m} x_{it} u_{tj}$$

Other dimension reduction methods

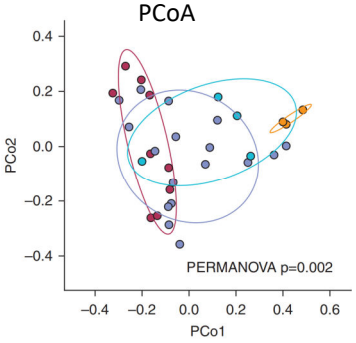
Metric multidimensional scaling (MDS)
Principal Coordinate Analysis (PCoA)

Bray-Curtis dissimilarity index
in microbiome analyses

Non-linear transformation
Keeps local and global structures

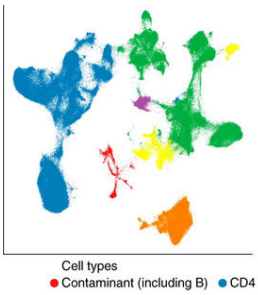
Single cell analysis

PCoA

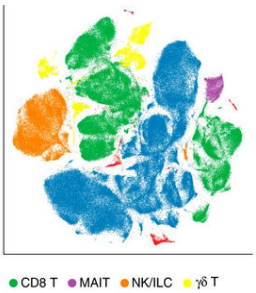


Moosbruger-Martinz et al. J Invest Derm 2020

UMAP



t-SNE



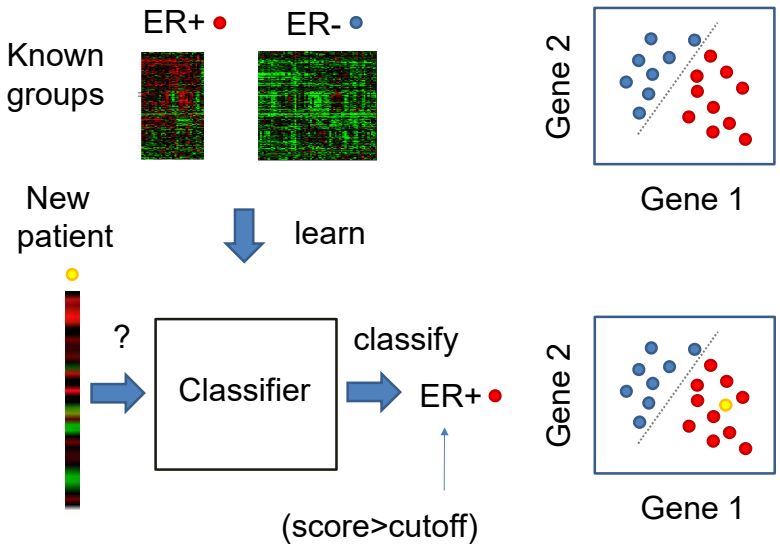
Becht E et al. Nat Biotechnol 2018

Clustering vs. classification

Clustering uses the primary data to group together measurements, with **no information** from other sources (*unsupervized machine learning*)

Classification uses **known groups** of interest (from other sources) to learn the features associated with these groups in the primary date, and create rules for associating the data with the group of interest (supervized machine learning)

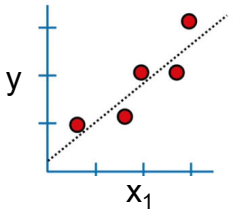
Classification



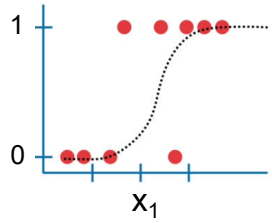
Methods for classification

- K-nearest neighbors
- Linear Models
- Discriminant analysis
- Logistic Regression
- Naïve Bayes
- Decision Trees
- Random Forests
- Support Vector Machines

Linear and logistic regression



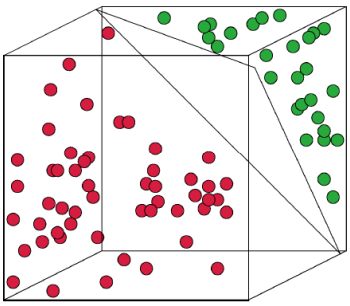
$y=b_0+b_1*x_1+b_2*x_2+...$

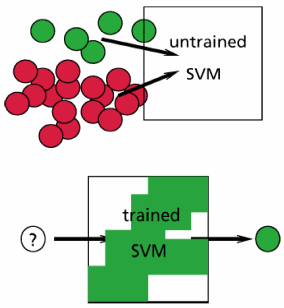


$\ln (P/(1-P))=b_0+b_1*x_1+b_2*x_2+...$

$$L_{LASSO}(\beta) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i|^2 + \lambda\sum_{j=1}^{J}|\beta_j|.$$

Support vector machines (SVM)





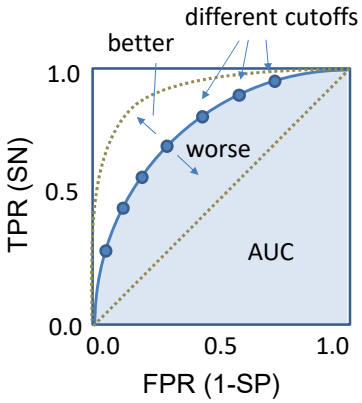
A SVM tries to find an optimal hyperplane that separates all training samples correctly and maximaizes the margins. If this is not possible in the input space (e.g. 2 dimensions) a hyperplane can be found in the higher dimensional features space (e.g. 3 dimensions).

Receiver operator characteristics (ROC)

Classified (> cutoff)	truely ER+	truely ER-
ER+	TP	FP
ER-	FN	TN

Sensitivity
SN=TP/(TP+FN)

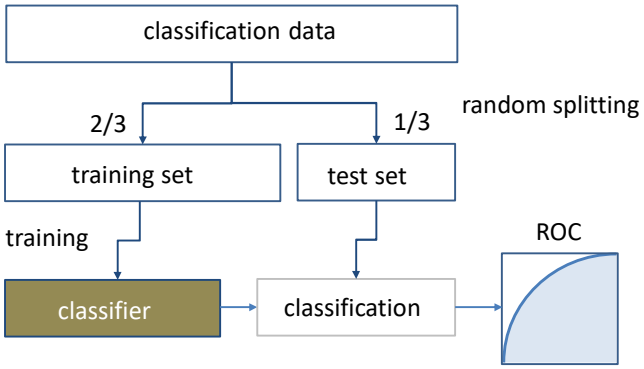
Specificty
SN=TN/(TN+FP)

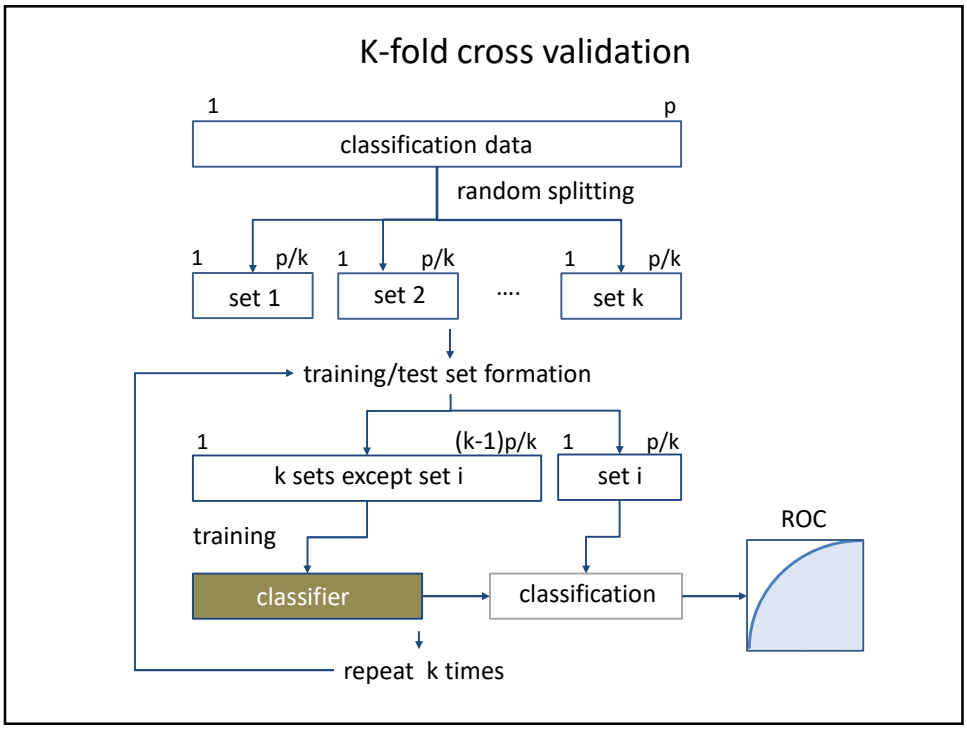


Area under curve (AUC)
AUC=1.0 optimal
AUC=0.5 random

Holdback cross validation

To avoid overfitting data should be splitted into training and test set




K-fold cross validation
1
p
classification data
random splitting
1
p/k
1
p/k
1
p/k
set 1
set 2
....
set k
training/test set formation
1
(k-1)p/k
1
p/k
k sets except set i
set i
training
ROC
classifier
classification
repeat k times