Gene expression analysis

Ulf Leser

Last lecture

<u>mRNA-expression</u> arrays

Chips with probes that measure mRNA

Workflow mRNA arrays

RNA extraction, cDNA rewriting, labeling,
hybridization, scanning, spot detection, spot intensity to numeric values,
normalization, *analysis* (today)

Background-Correction & Normalization

Assume that no structural differences exist between samples
Homogenize measurements to render them comparable

This lecture

- Differential expression
  - Fold Change
  - t-Test
- Clustering
- Databases

Differential Expression - Motivation

Understand etiology

Identify early detection marker

Personalized medicine

Differential Expression

We **have**:

$N_1,\ldots,N_m$: normale samples

$T_1,\ldots,T_n$: tumor samples

We **look for**: genes with significant differences between N and T

Compare values of gene X from group N with those of group T

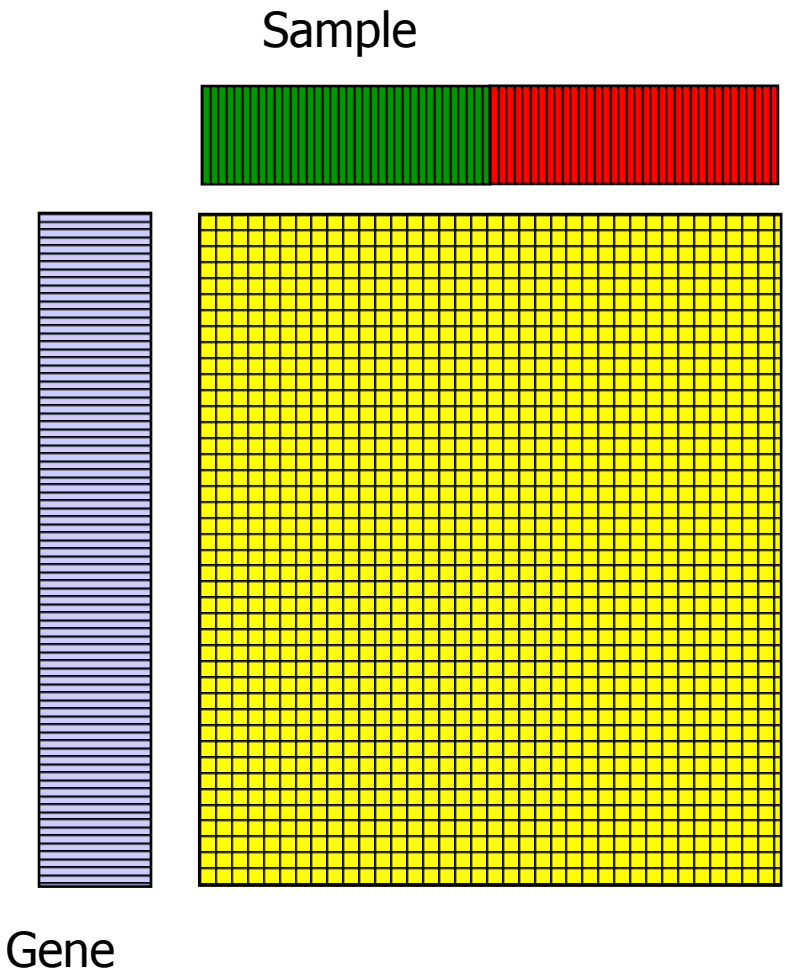
$N = \{n_1,\ldots,n_m\}$

$T = \{t_1,\ldots,t_n\}$

many methods, here:

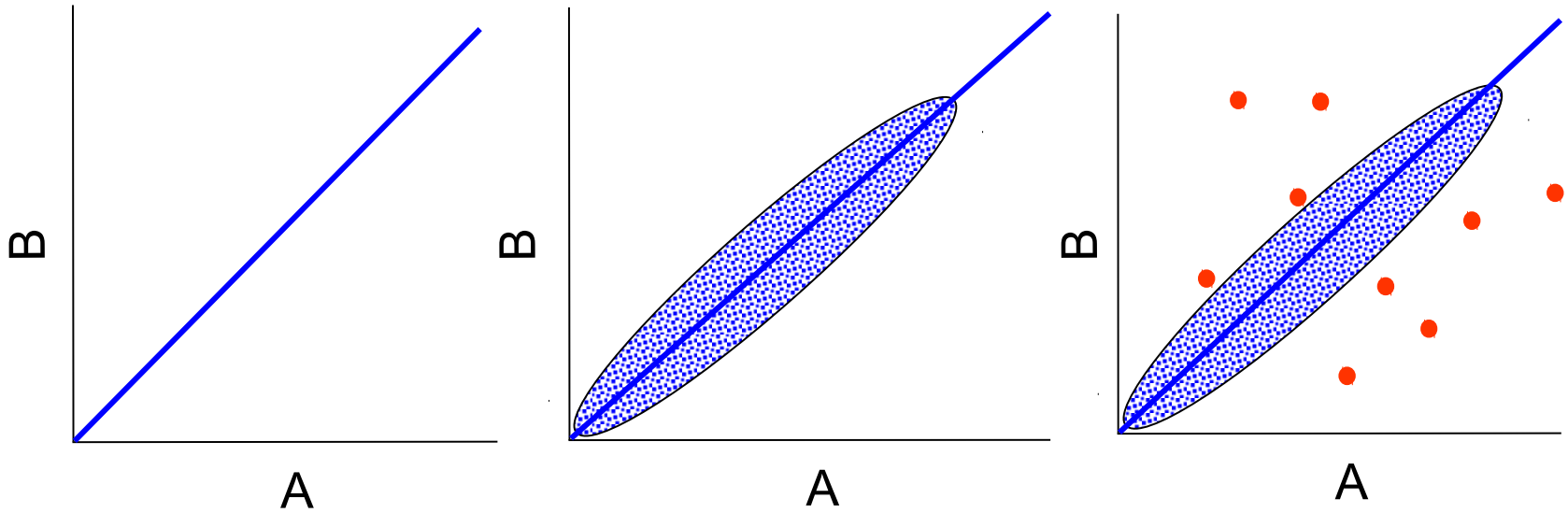
Fold change

t-test



What to look for - Scatterplot

one point = one gene



Identical

(Probably) Background-Noise

outlier: interesting genes

A : Gene expression Patient A

B : Gene expression Patient B

This lecture

- Differential expression
  - Fold Change
  - T-Test
- Clustering
- Databases

Fold Change

Fold Change (FC)

$$FC = \log_2\left(\frac{mean(T)}{mean(N)}\right) = \log_2(mean(T)) - \log_2(mean(N))$$

Thresholds (sort of because never really comparable)

|FC| < 1 not interesting
|FC| > 2 very interesting

Log2

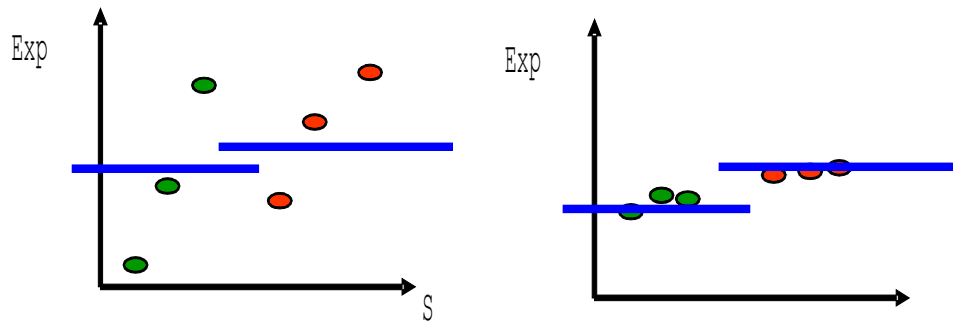
	mean(tumor)	mean(normal)	mean(t) / mean(n)	FC
gene a	16	1	16	4
gene b	0.0625	1	0.0625	-4
gene c	10	10	1	0
gene d	200	1	200	7.65

This lecture

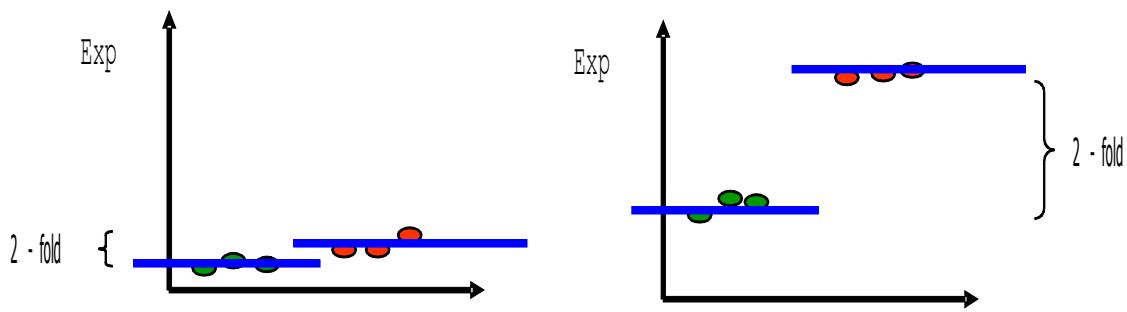
- Differential expression
  - Fold Change
  - T-Test
- Clustering
- Databases

Fold Change– Advantages / Disadvantages

+ intuitive measure
- independent of scatter



- independent of absolute values



→ score based only on the mean of the groups not optimal, **include variance!**

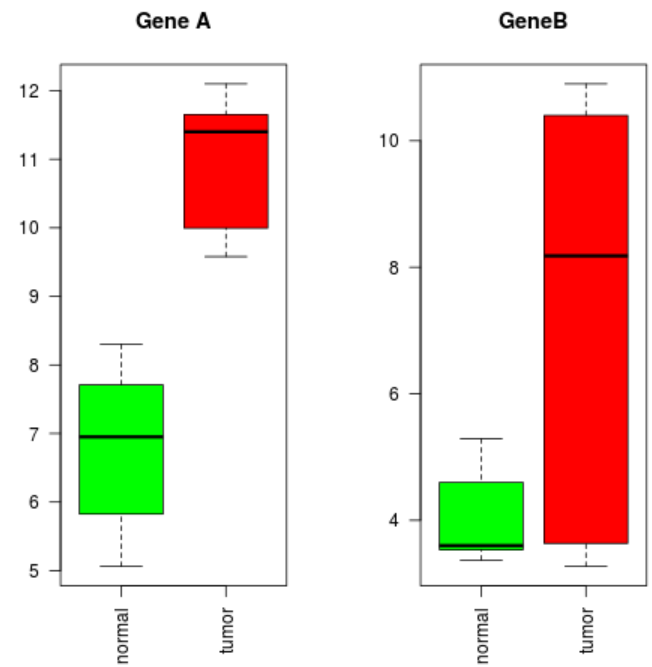
Hypothesis Testing – Comparing Two Samples

Gene expression matrix:

Gene	N1	N2	N3	N4	N5	N6	N7	T1	T2	T3	T4	T5	T6	T7	FC
A	5.06	5.22	8.3	8.03	6.95	6.43	7.39	10.1	9.89	11.7	11.6	11.4	9.58	12.1	**-4.14**
B	3.58	4.14	3.49	3.37	5.29	5.06	3.6	3.7	10.9	10.3	3.57	10.5	8.18	3.27	**-3.13**

High abs(FC) for Gene A and Gene B

But: variance very high in the tumor samples of Gene B



Evaluate 'randomness' of
→ T-test

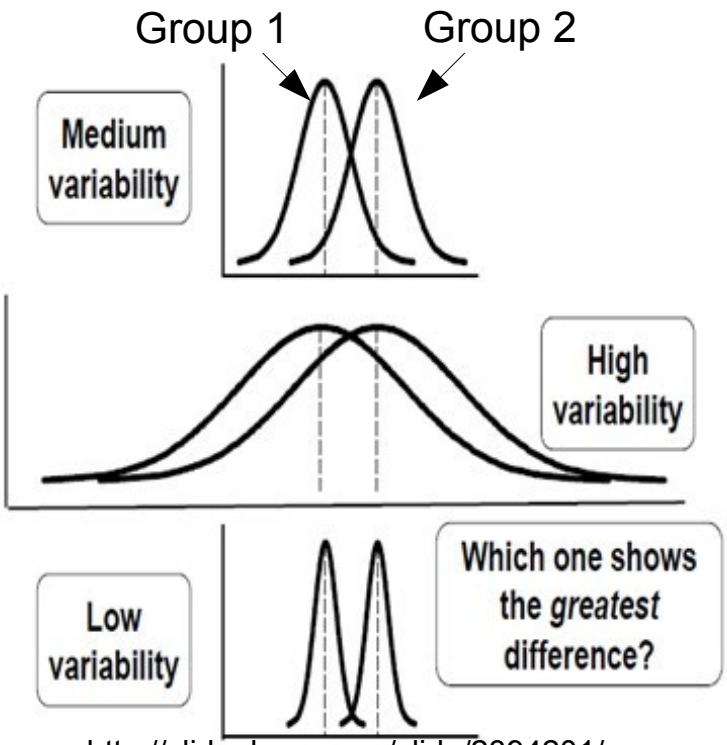
Hypothesis Testing

- Same Mean
  → Different variance

- Measure 'uncertainty' with standard deviation *sd*

- Combine both to likelihood for 'correctness'

- Assumption
  - Log-Normal distributions
  - Symmetric
  - Independent



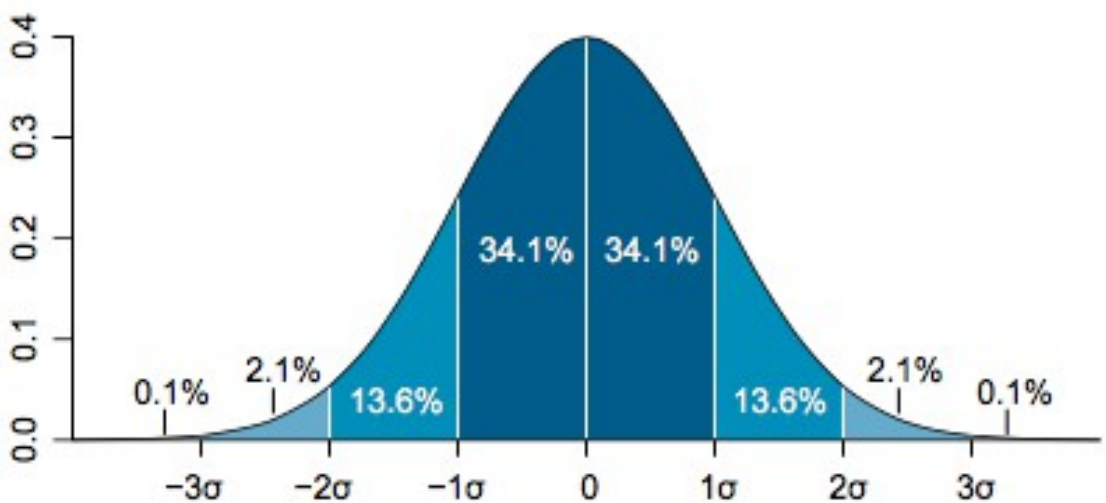
http://slideplayer.com/slide/2394201/

$$\sigma_X := \sqrt{\mathrm{Var}(X)}$$

$$\mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}(X))^2) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$$

Tschebyscheff-Inequation

$$P[|X - \mu| \geq k] \leq \frac{\sigma^2}{k^2}$$



Quite neat:
Z-transform your data
and see how likely a single value is

Hypothesis Testing

- **T-Test (unpaired two-sample)**
  compares the mean of two unpaired samples
- **Assumption:**
  - values normally distributed
    equal variances
- **Hypothesis:**
  H_0 (Null hypothesis): $\mu_1 = \mu_2$ vs. $\mu_1 \mathrel{!=} \mu_2$ (means are not equal)
- **Test statistic:** Function of the sample that summarizes the data set into one value that can be used for hypothesis testing

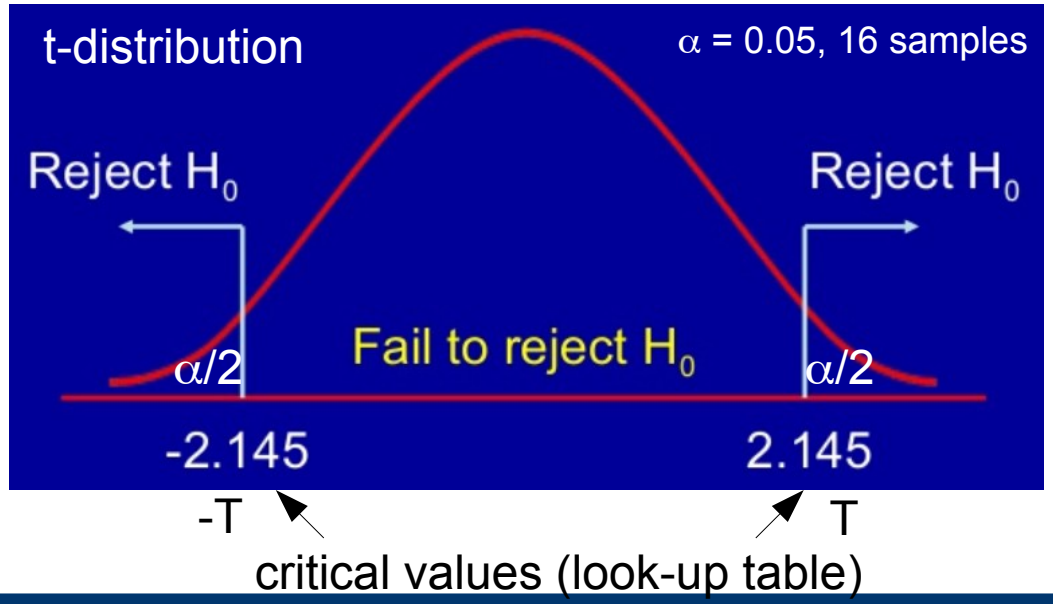
$$t = \frac{mean(T) - mean(N)}{\sqrt{\frac{sd(T)^2}{m} + \frac{sd(N)^2}{n}}}$$

The greater | t |, the greater the difference between the means

Ways to get a larger t:
- Bigger difference in means
- Smaller standard deviation
- More samples

Hypothesis Testing – t-Test (Welch Test)

- **From T-statistic to p-value:**
  T-value, α and number of samples determine the p-value (look-up tables)
- **P-value:**
  - Probability of observing your data under the assumption that H_0 is true
  - Probability that you will be in error if rejecting H_0
- **Significance level (α):** Probability of a false positive outcome of the test, the error of rejecting H_0 when it is actually true



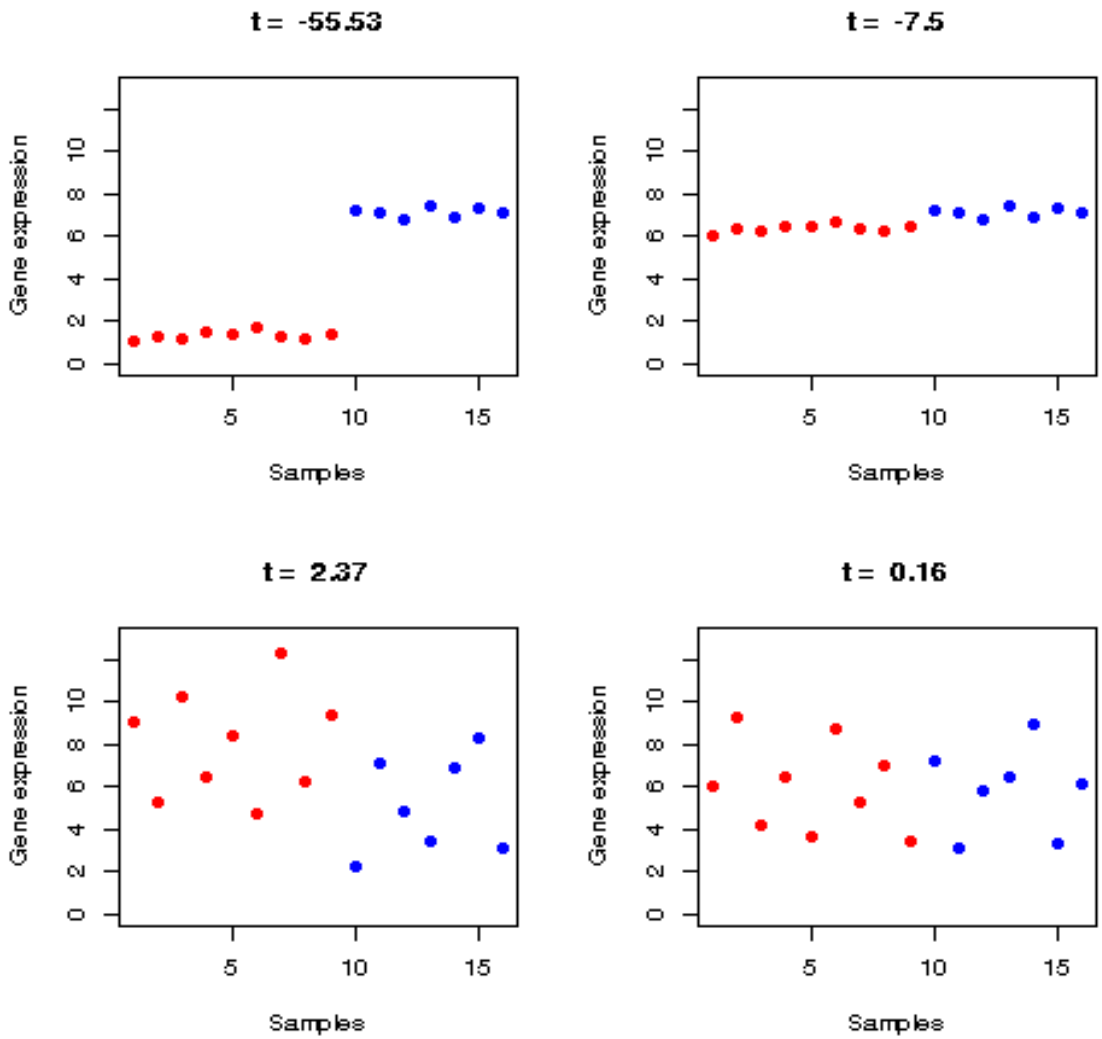
If $|t| > |T|$ we reject H_0

→ p-value is significant (p-value < α)

Steps of Hypothesis Testing

- Determine null and alternative hypothesis
- Select a significance level (alpha)
- Take a random sample from the population of interest
- Calculate a test statistic from the sample that provides information about the null hypothesis
- Decision

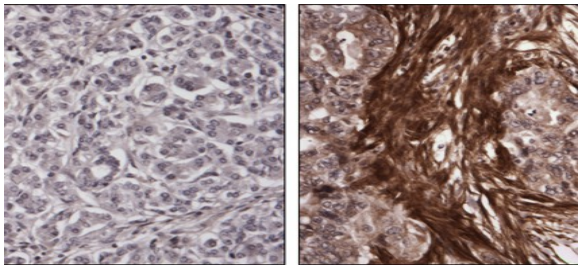
Example



Example

Example for Gene B from slide 11

N = {3.58, 4.14, 3.49, 3.37, 5.29, 5.06, 3.6}



T = {3.7, 10.9, 10.3, 3.57, 10.5, 8.18, 3.27}

Hypothesis H_0: $m_N - m_T = 0$ H_1: $m_N - m_T \mathrel{!=} 0$

Significance level $\alpha = 0.05$

Test statistic

$$t = \frac{mean(T) - mean(N)}{\sqrt{\frac{sd(T)^2}{m} + \frac{sd(N)^2}{n}}} = -2.27$$

(Critical value |T| = 2.45)

p-Value p-value = 0.06 → We cannot reject H_0, gene B ist not significantly differentially expressed!

Vulcano Plot

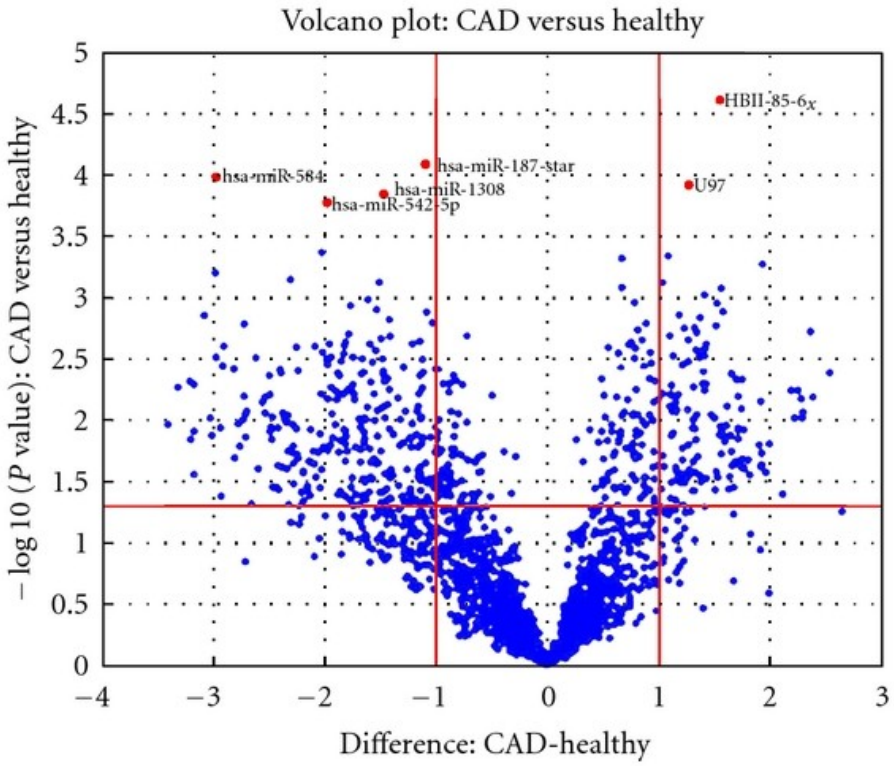
- Scatterplot significance versus fold change
  - Y-axis: Negative $\log_{10}$ of the p-value
  - X-axis: Fold-change

- Interested in the

  upper left

  upper right

  corner



Log2 FC

http://www.hindawi.com/journals/crp/2011/532915.fig.001.jpg

Multiple Testing Correction

Problem: Microarrays has 22k genes, thus an $\alpha=0.05$ leads to approximately 22 000 * 0.05 ~ 1100 FPs.

Solution: Multiple testing correction. Two basic approaches:

1. **Family wise error rate (FWER)** , the probability of having at least one false positive in the set of results considered as significant

2. **False discovery rate (FDR)**, the expected proportion of true null hypotheses rejected in the total number of rejections.(FDR measures the expected proportion of incorrectly rejected null hypotheses, i.e. type I errors)

Bonferoni (FWER)

Let N be the number of genes tested and p the p-value of a given probe, one computes an adjusted p-value using:

$$p_{adjusted} = p*N$$

Only if the adjusted p-value is smaller than the pre-chosen significance value, the probe is considered differentially expressed.

Very conservative (many failures to reject a false H_0), rarely used

Bonferoni assumes independence between the tests (usually wrong)

Appropriate when a **single false positive** in a set of tests would be a problem (e.g., drug development)

Benjamini – Hochberg (FDR)

1. choose a specific α (e.g. α=0.05)

2. rank all m p-values from smallest to largest

3. correct all p-values: $BH(p_i)_{i=1,\ldots,m} = p_i * m/i$

4. BH (p) = significant if BH(p) ≤ α

Genes	p-value	rank	BH(p)	Significant? (α=0.05)
Gene A	0.00001	1	0.00001*1000/1=0.01	yes
Gene B	0.0004	2	0.0004*1000/2=0.20	no
Gene C	0.01	3	0.01*1000/3=3.3 → 1.0	no

This lecture

- Differential expression
  - Fold Change
  - T-Test
- Clustering
- Databases

Clustering - Motivation

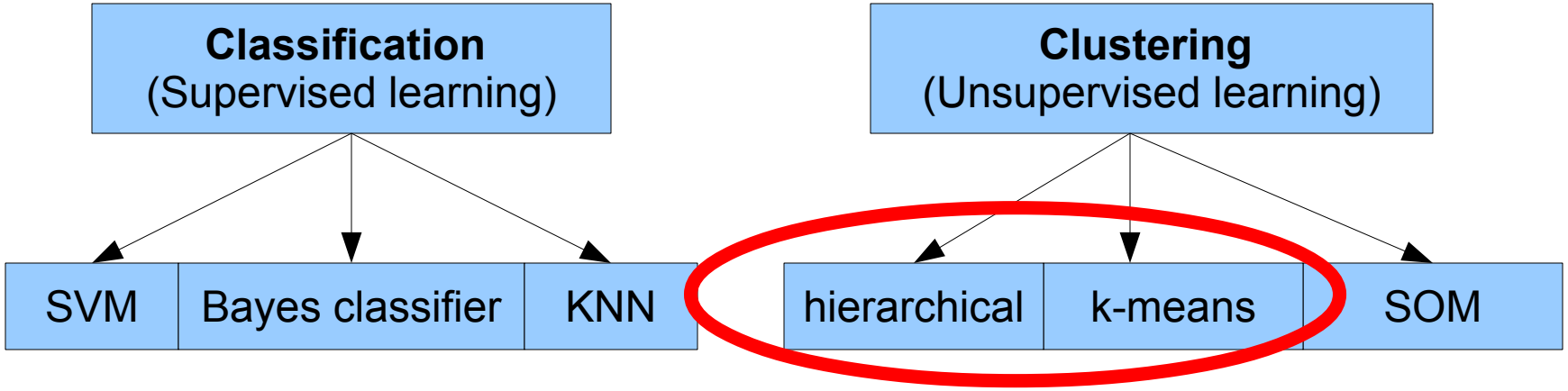
Subgroups detection

Quality control

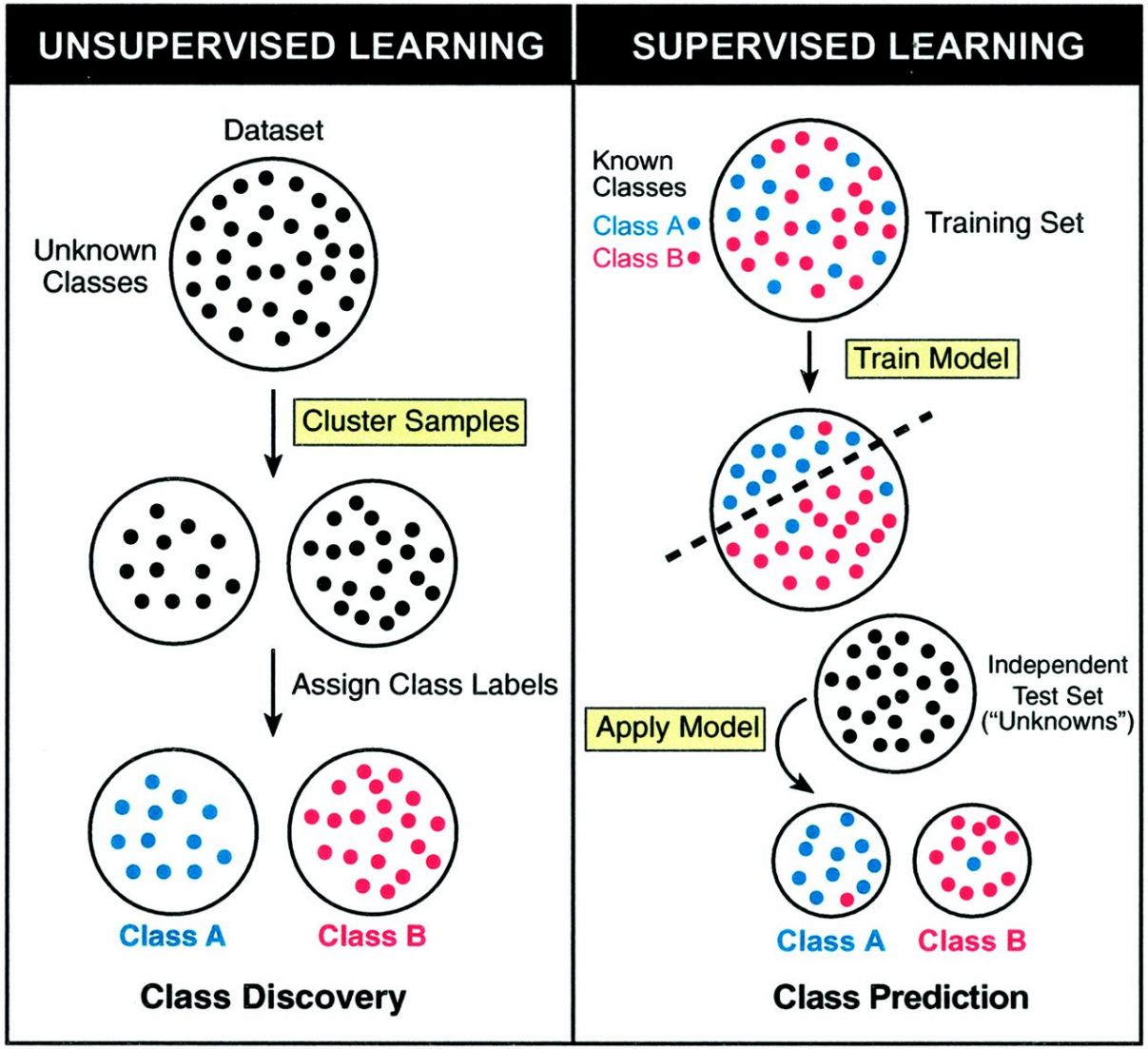
Similar-detection in spacial and temporal behavior
→ **co-regulated / expressed genes** (e.g. genes controlled by the same transcription-factor).

Discover new **disease subtypes**

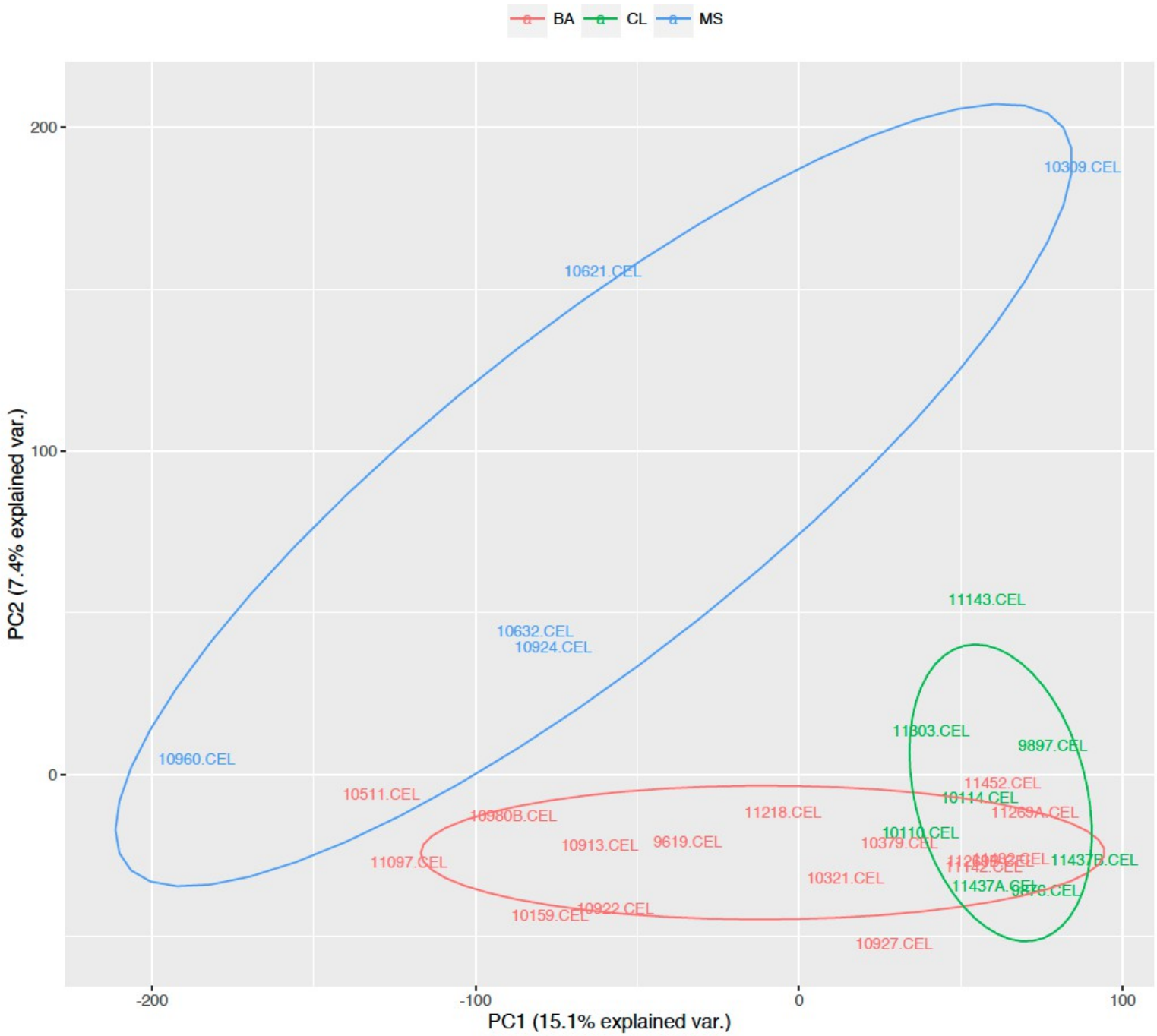
Overview (Un)Supervised Learning



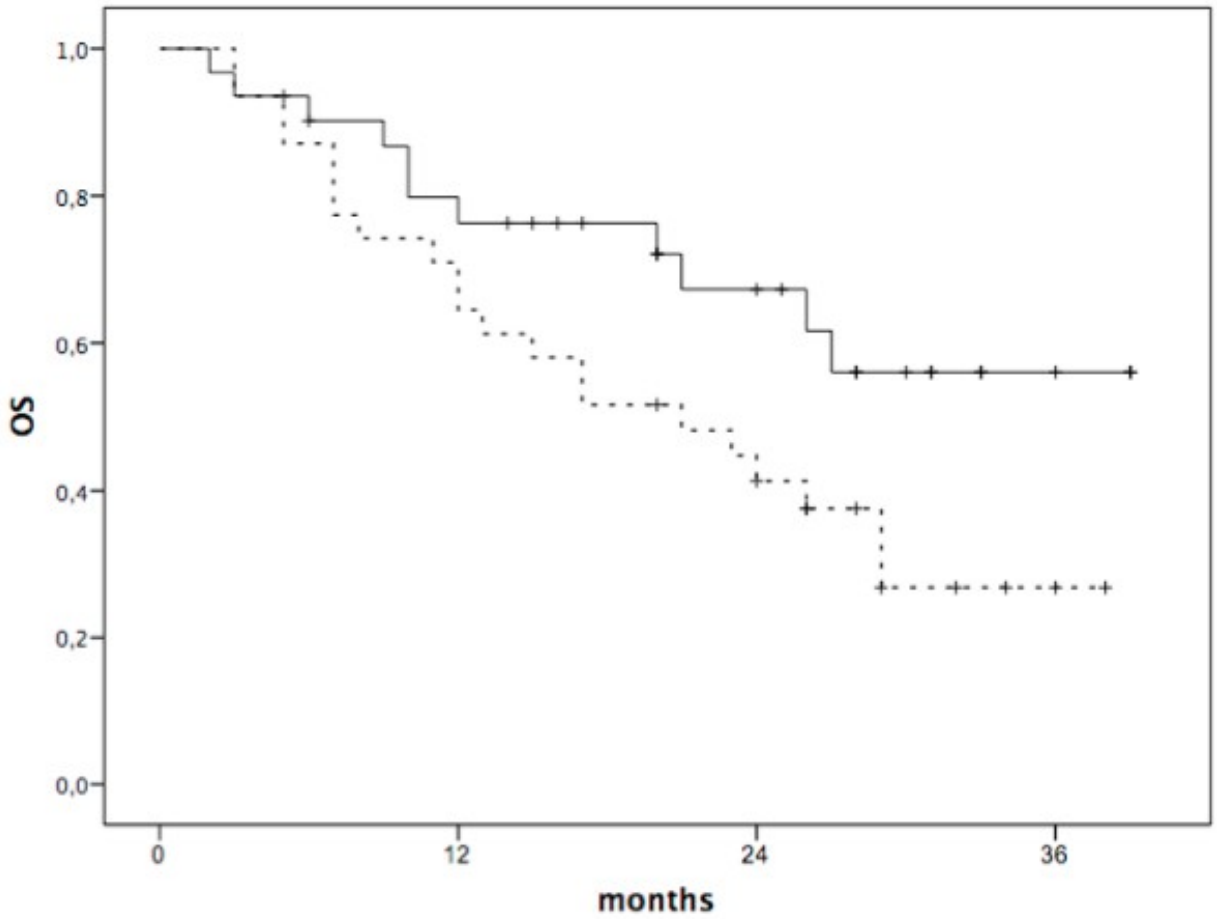
Clustering

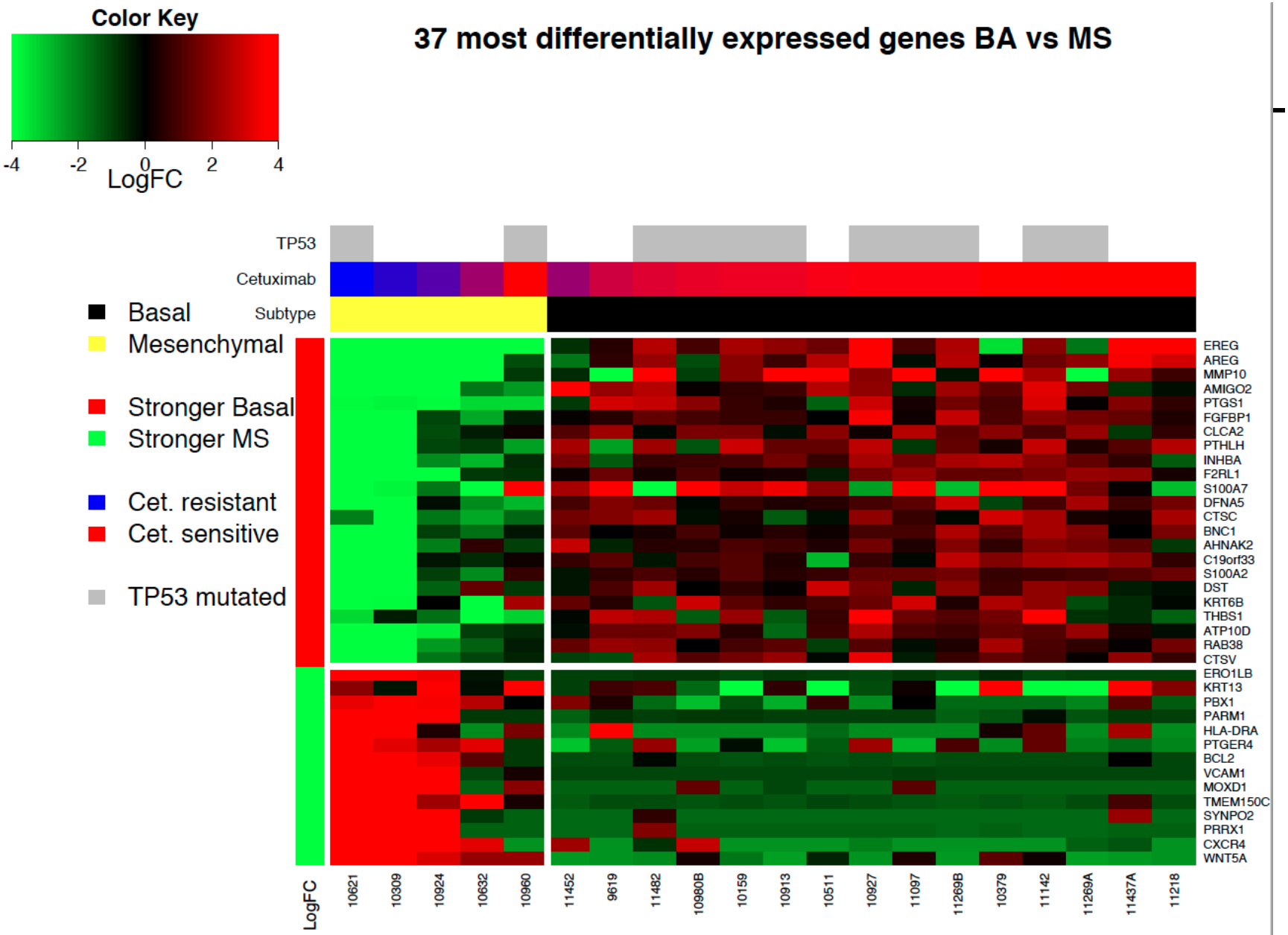


Ramaswamy & Golub 2002


BA
CL
MS
200
100
0
-200
-100
0
100
PC2 (7.4% explained var.)
PC1 (15.1% explained var.)
10309.CEL
10621.CEL
10632.CEL
10924.CEL
10960.CEL
11143.CEL
11303.CEL
9897.CEL
11452.CEL
10114.CEL
11269A.CEL
10511.CEL
10980B.CEL
11218.CEL
10110.CEL
10913.CEL
9619.CEL
10379.CEL
11097.CEL
11437B.CEL
10321.CEL
11437A.CEL
10159.CEL
10922.CEL
10927.CEL

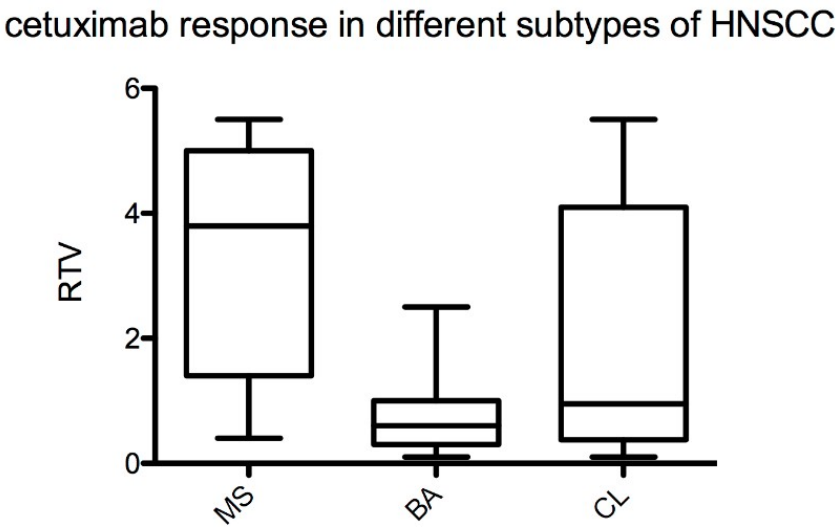
Kaplan-Meier Clusterplot




Color Key
-4
-2
0
2
4
LogFC
37 most differentially expressed genes BA vs MS
TP53
Cetuximab
Subtype
Basal
Mesenchymal
Stronger Basal
Stronger MS
Cet. resistant
Cet. sensitive
TP53 mutated
EREG
AREG
MMP10
AMIGO2
PTGS1
FGFBP1
CLCA2
PTHLH
INHBA
F2RL1
S100A7
DFNA5
CTSC
BNC1
AHNAK2
C19orf33
S100A2
DST
KRT6B
THBS1
ATP10D
RAB38
CTSV
ERO1LB
KRT13
PBX1
PARM1
HLA-DRA
PTGER4
BCL2
VCAM1
MOXD1
TMEM150C
SYNPO2
PRRX1
CXCR4
WNT5A
LogFC
10621
10309
10924
10632
10960
11452
9619
11482
10980B
10159
10913
10511
10927
11097
11269B
10379
11142
11269A
11437A
11218

Clustering

- Goal:
  - Partitioning
  - Biological interpretation of subtypes (clusters)
- Requires:
  - (useful) similarity measure
- Advantages:
  - Intuitive
  - Simple (you would think)



Hierarchical Clustering - Algorithm

1. Distance measure
   Euclidean
   Pearson, etc.

2. compute similarity matrix S

3. while |S|>1:
   4. determine pair (X,Y) with minimal distance
   5. compute new value Z = avg (X,Y), (single, average, or complete linkage)
   6. delete X and Y in S, insert Z in S
   7. compute new distances of Z to all elements in S
   8. visualize X and Y as pair

Hierarchical Clustering

Result: binary tree

Cutting the dendrogram at a particular height partitions the data into disjoint clusters

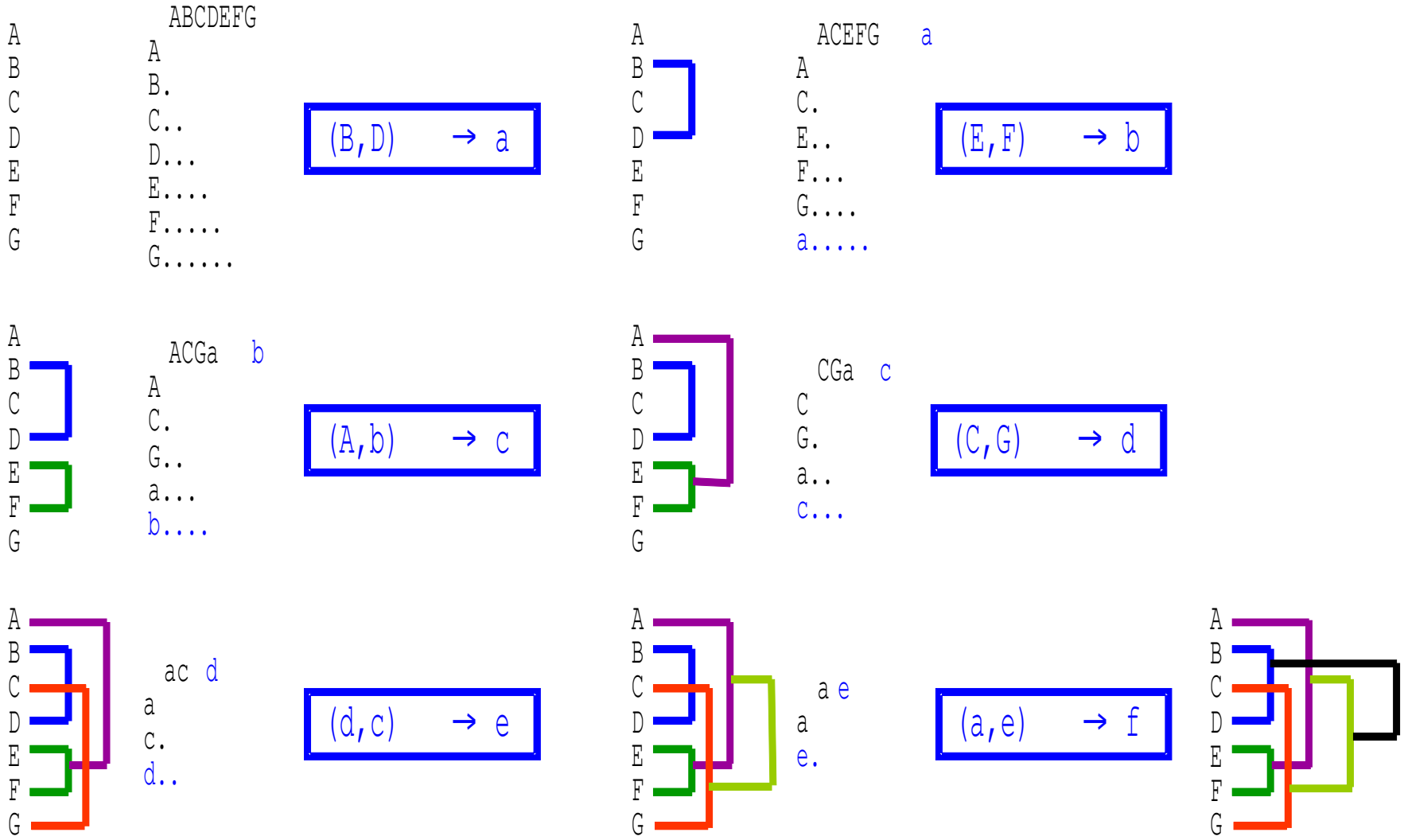
For an easier determination of clusters: length of branch is set in relation to the difference of the leafs.

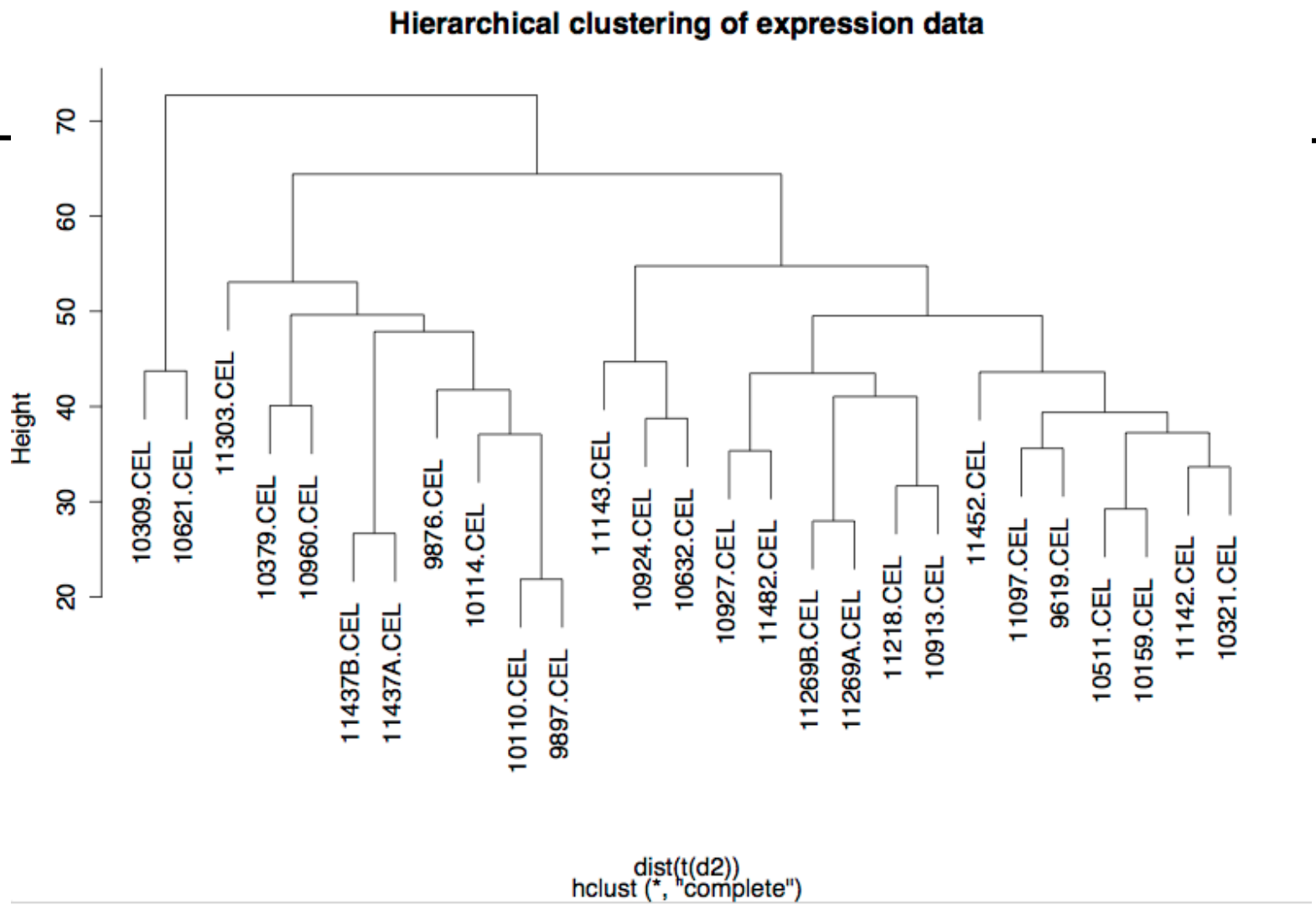
Linkage Rule essential

Hierarchical Clustering – Linkage

- Methods produce similar results for data with strong clustering tendency (each cluster is compact and separated)
- **Single** Linkage: $D(X,Y)=\min_{x\in X, y\in Y} d_{xy}$
  - Single smallest distance
  - Violates the compactness property (i.e., observations inside the same cluster should tend to be similar)
- **Complete** Linkage: $D(X,Y)=\max_{x\in X, y\in Y} d_{xy}$
  - Most distant elements
- **Average** Linkage: $D(X,Y)=\frac{1}{N_X N_Y}\sum_{x\in X}\sum_{y\in Y} d_{xy}$
  - Compromise

Hierarchical Clustering - graphical




Hierarchical clustering of expression data
Height
20
30
40
50
60
70
10309.CEL
10621.CEL
11303.CEL
10379.CEL
10960.CEL
11437B.CEL
11437A.CEL
9876.CEL
10114.CEL
10110.CEL
9897.CEL
11143.CEL
10924.CEL
10632.CEL
10927.CEL
11482.CEL
11269B.CEL
11269A.CEL
11218.CEL
10913.CEL
11452.CEL
11097.CEL
9619.CEL
10511.CEL
10159.CEL
11142.CEL
10321.CEL
dist(t(d2))
hclust (*, "complete")

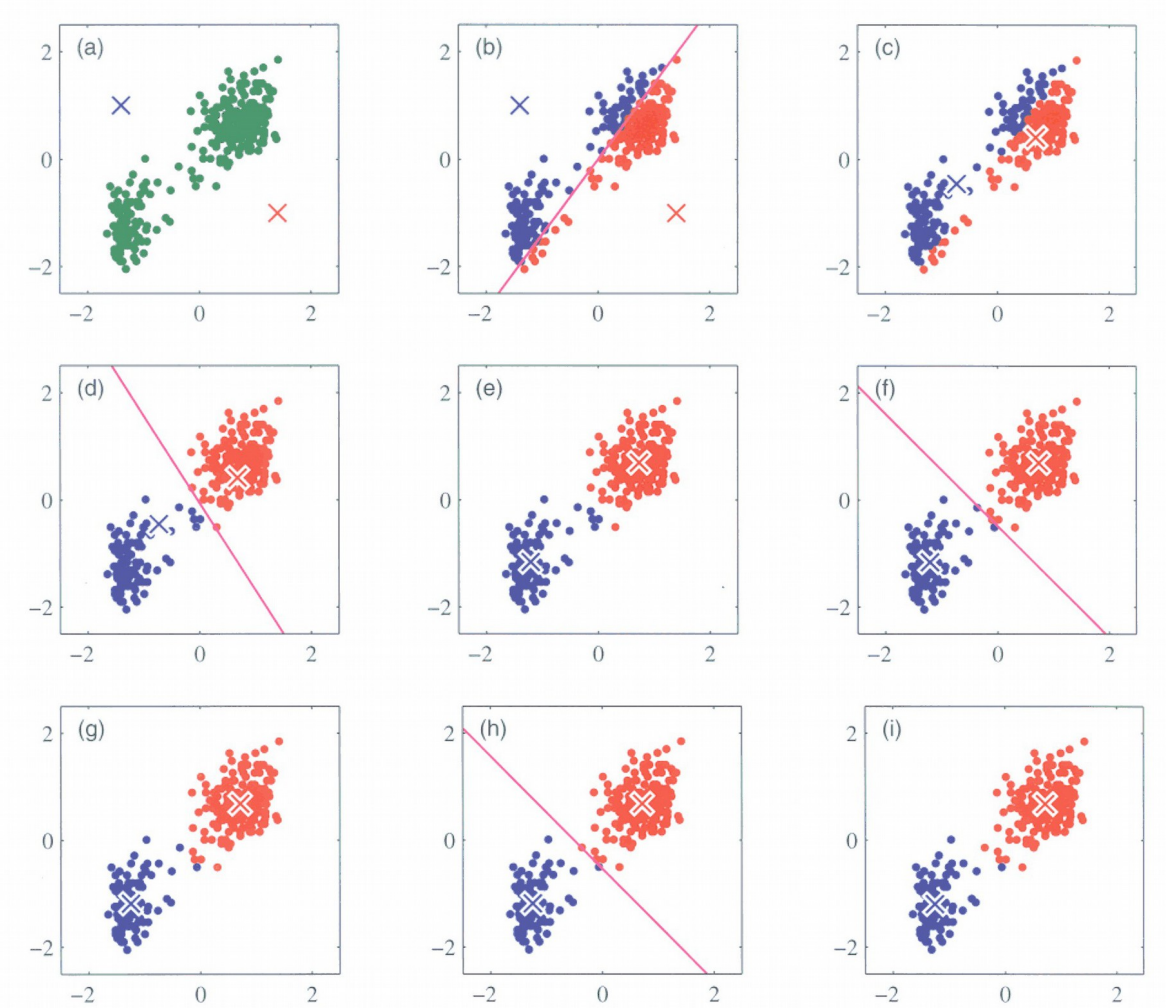
K means

K-means partitions the n observations into k clusters

Minimize the distance of the n data points from their respective cluster centres.

1. choose *k* random cluster centers $\mu_1,...\mu_k$
2. Assign for each point x in dataset S the closest cluster center
3. compute a new center μ_i for every cluster C_i
4. repeat 2-3. until cluster centers do not change

K means



http://www.itee.uq.edu.au/~comp4702/lectures/k-means_bis_1.jpg

K means

Convergence is not assured.

Cluster quality can be computed by determining the mean distance of a gene to its clustercenter.

Number of clusters has to be chosen in advance.

The initialization of the cluster centers has a great impact on the clustering quality, compute more than one initial constellation.

This lecture

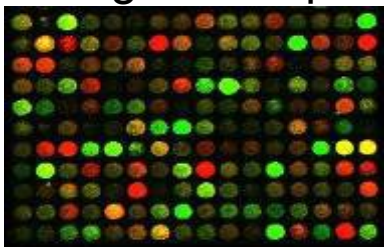
- Differential expression
  - Fold Change
  - T-Test
- Clustering
- Databases

GEO – Gene Expression Omnibus

NCBI public repository http://www.ncbi.nlm.nih.gov/geo/ archives microarray, NGS, and other high-throughput genomics data submitted by the research community

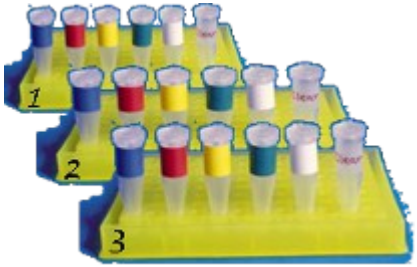
GPL
(GEO platform)
platform description

GSM
(GEO sample)
raw-processed intensities from a single or chip



GSE
(GEO series)
grouping of chip data, a single experiment

GDS
(GEO dataset)
grouping of experiments



submitted by manufacturer

submitted by experimentalist

curated by NCBI

GEO

NCBI Resources How To Sign in to NCBI

GEO Home | Documentation | Query & Browse | Email GEO

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Gene Expression Omnibus

Keyword or GEO Accession **Search**

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- GEO BLAST
- Programmatic Access
- FTP Site

Browse Content

- Repository Browser
- DataSets: 3848
- Series: 58176
- Platforms: 14392
- Samples: 1424131

Information for Submitters

- Login to Submit
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications

GEO

NCBI

GEO Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

NCBI » GEO » Repository browser » Series

Login

Series | Samples | Platforms | DataSets

Summary | Advanced search

microarray, homo sapiens, leukemia Search

22 series

Export

<< < Page 1 of 2 > >> Page size 20

Accession	Title	Series type(s)	Organism(s)	Samples	GDS	Supplementary	Contact	Release date
GSE54218	Aberrant chromatin acetylation in MLL-AF9 leukemia mediates the response to HDAC inhibition (microarray)	Expression profiling by array	*Homo sapiens*	16		TXT	Sara Alvarez	Mar 01, 2015
GSE46670	Regulation of gene expression in human T lymphoblastic leukemia Molt4 cells by farnesol	Expression profiling by array	*Homo sapiens*	6		TXT	NIEHS Microarray Core	May 01, 2014
GSE46251	Microarray data of human leukemia cells MV4;11 with or without IKK inhibitor treatment.	Expression profiling by array	*Homo sapiens*	6		CEL	HSU-PING KUO	Sep 19, 2013
GSE46252	Microarray data of leukemia cells with or without IKK inhibitor treatment	Expression profiling by array	*Homo sapiens* *Mus musculus*	12		CEL	HSU-PING KUO	Sep 19, 2013
GSE40639	Microarray analysis of gene expression of microdissected epidermis and dermis in mycosis fungoides and adult T-cell leukemia/lymphoma	Expression profiling by array	*Homo sapiens*	16		TXT	Keiko Hashikawa	Sep 06, 2012
GSE8779	Molecular profiling of myeloid leukemia cell lines	Expression profiling by array	*Homo sapiens*	17			Stanford Microarray Database (SMD)	Feb 24, 2012
GSE34823	Routine use of microarray-based gene expression profiling to identify patients with low cytogenetic risk acute myeloid leukemia: accurate results can be obtained even with suboptimal samples	Expression profiling by array	*Homo sapiens*	206			Philippe Guardiola	Jan 04, 2012
GSE34714	Routine use of microarray-based gene expression profiling to identify patients with low cytogenetic risk acute myeloid leukemia: accurate results can be obtained even with suboptimal samples. (test samples)	Expression profiling by array	*Homo sapiens*	117			Philippe Guardiola	Dec 24, 2011

All GEO submissions follow the MIAME checklist

MIAME (Minimum Information about a Microarray Experiment)

1. raw data (e.g. .CEL, .txt)

2. final **processed** (normalized) **data**

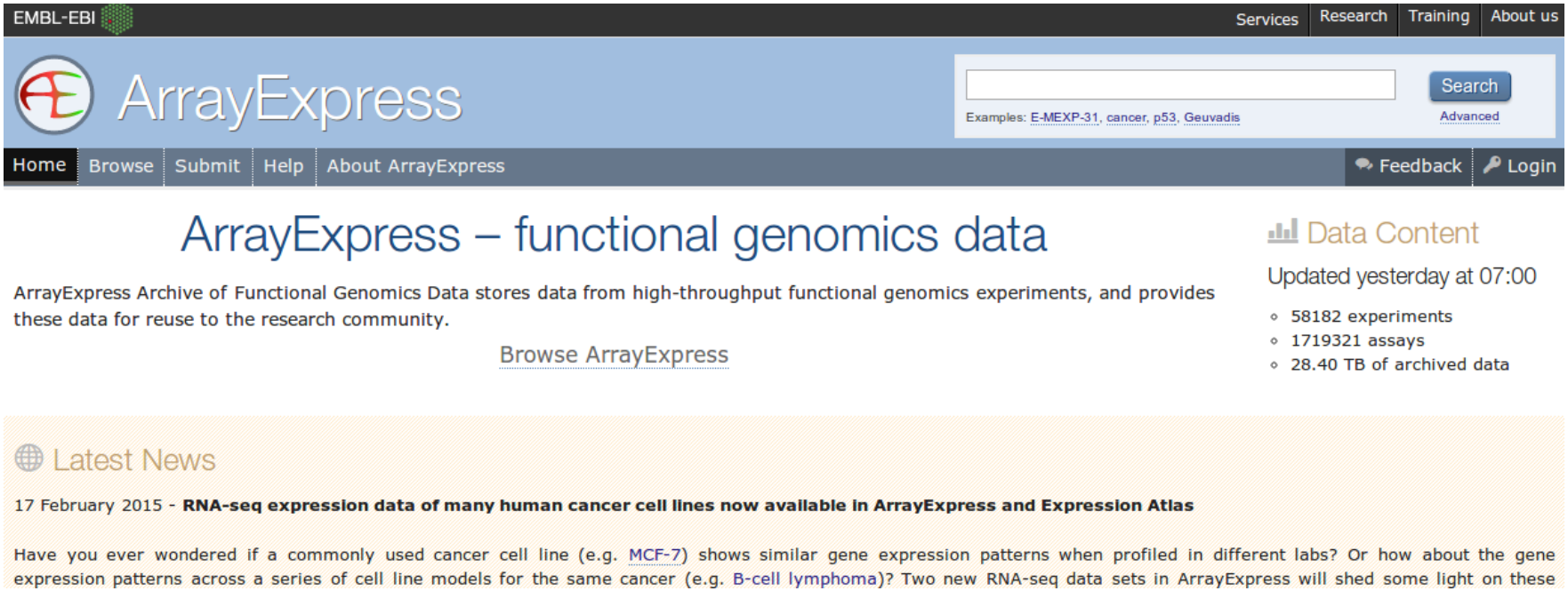
3. sample annotation (incl. Experimental factors and their values, scan protocol,e.g. drug, dosage)

4. experimental design including sample data relationships (e.g., overall design; technical or biological replicates)

5. annotation of the **array** (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences)

6. laboratory and **data processing protocols** (e.g., what normalisation method)

ArrayExpress (EMBL-EBI)



All ArrayExpress submissions follow the MIAME checklist

GEO vs. ArrayExpress

- both encompass MIAME compliance

- both provide a good possibility for making data publicly availabe as often requested by journals

- ArrayExpress provides analysis tools

Summary

Combine T-test and fold change for optimal detection of differential expression (Volcano plot)

More explorative analyses like clustering can detect patterns inherent in the expression data like co-regulated genes or new disease subtypes.

Public repositories like GEO and ArrayExpress offer a rich fundus of data.