Gene expression analysis

Ulf Leser and Karin Zimmermann

Last lecture

What are **microarrays**? - Biomolecular devices measuring the transcriptome of a cell of interest.

Workflow of a **microarray experiment** - RNA extraction, cDNA rewriting, labeling, hybridization to microarray, scanning, spot detection, spot intensity to numeric values, normalization, *analysis* (today)

Normalization – Assumption, that the vast majority of genes is not differentially expressed between the two classes. Remove technical bias to detect the biological differences.

This lecture

Differential expression
Clustering
Standards in the gene expression data management
Databases

Differential Expression - Motivation

Why find genes that behave differently in two classes (e.g. normal and tumor)?

Better understanding of the genetic circumstances that cause the difference (disease) hopefully leads to better therapy.

Detection of marker-genes enables the early recognition of diseases as well as the recognition of subtypes of diseases.

Once a cause is identified therapy can become more specific, more effective and reduce side-effects.

Differential Expression

We **have**:

$N_1,...,N_m$: normale samples

$T_1,...,T_n$: tumor samples

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

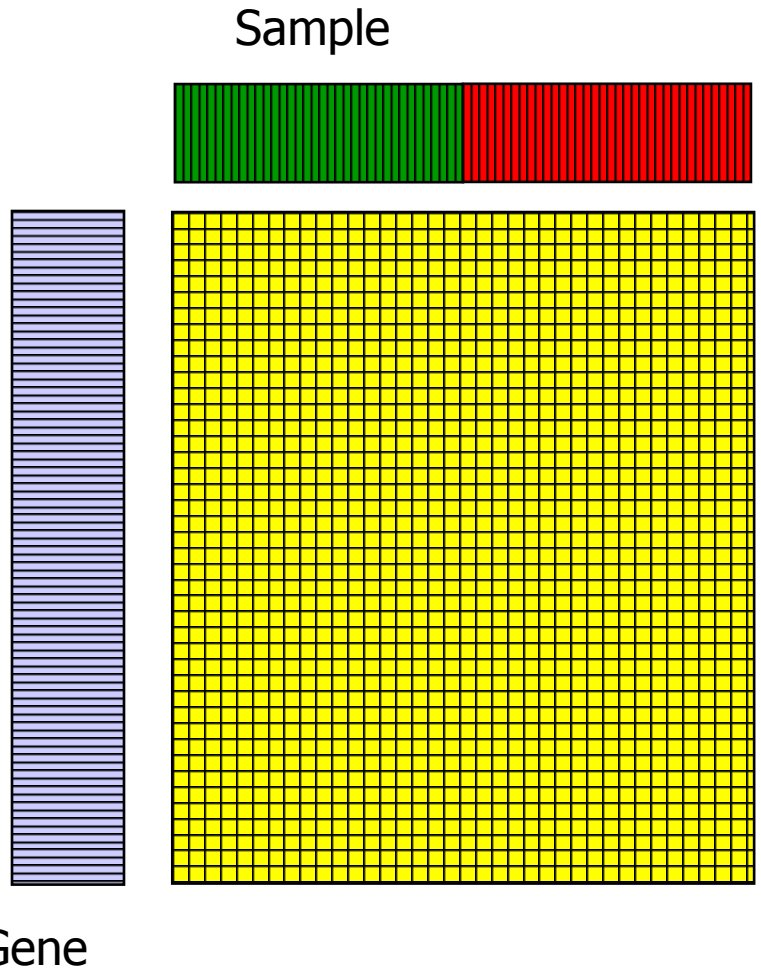
Compare values of gene X from group N with those of group T

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

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

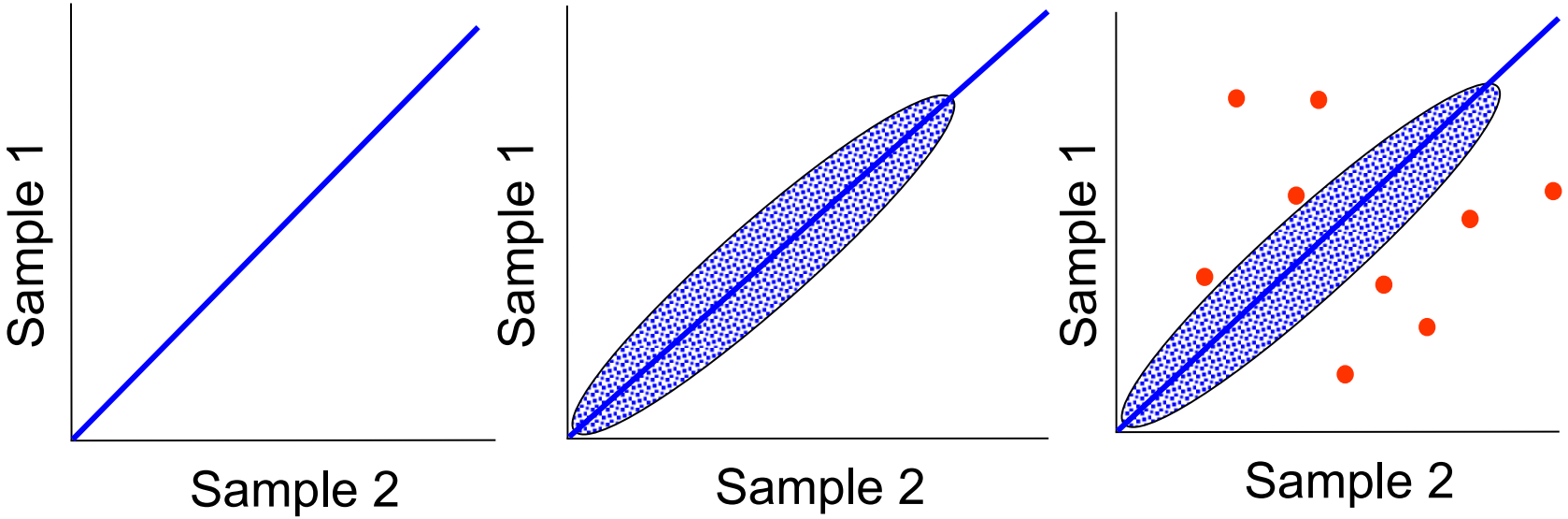
many methods, here:

- Fold change
- t-test



Visualization - Scatterplot

one point = one gene



totally identical distribution

distribution of intensity differences

outlier: interesting genes

Fold Change

Definition **Fold Change (FC)**:

$$2^{\left|\log 2\left(\frac{avg(T)}{avg(N)}\right)\right|}$$

Significance of result is determined by **threshold** fc:

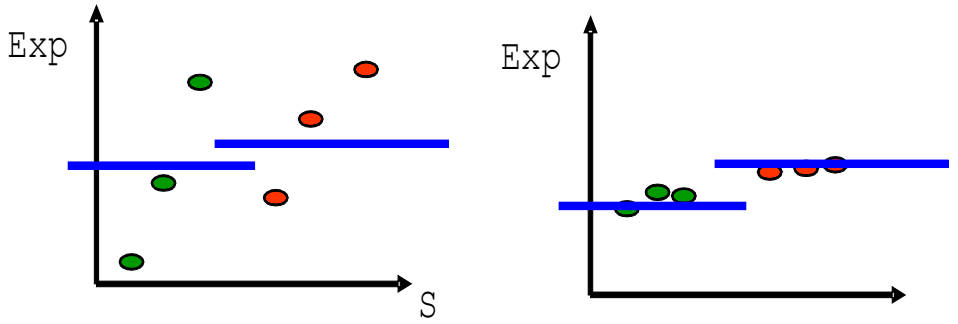
fc < 2 not interesting
2 < fc < 4 interesting
fc > 4 very interesting

Why log2 ?

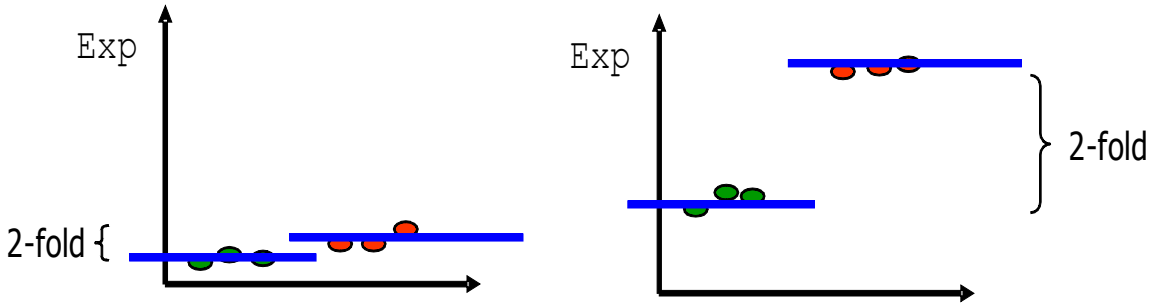
	mean(tumor)	mean(normal)	mean(t) / mean(n)	FC
gene x	16	1	16	16
gene y	0.0624	1	1/16	16

Fold Change– Advantages / Disadvantages

+ intuitive measure
- independent of scatter



- independent of absolut values



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

T-test – Hypothesis testing

Hypothesis

H0 Null hypothesis (the one we want to reject)
H1 Alternative hypothesis (logical opposite of H0)

Test statistic

Function of the sample that summarizes the characteristics of the latter into one number with a known distribution.

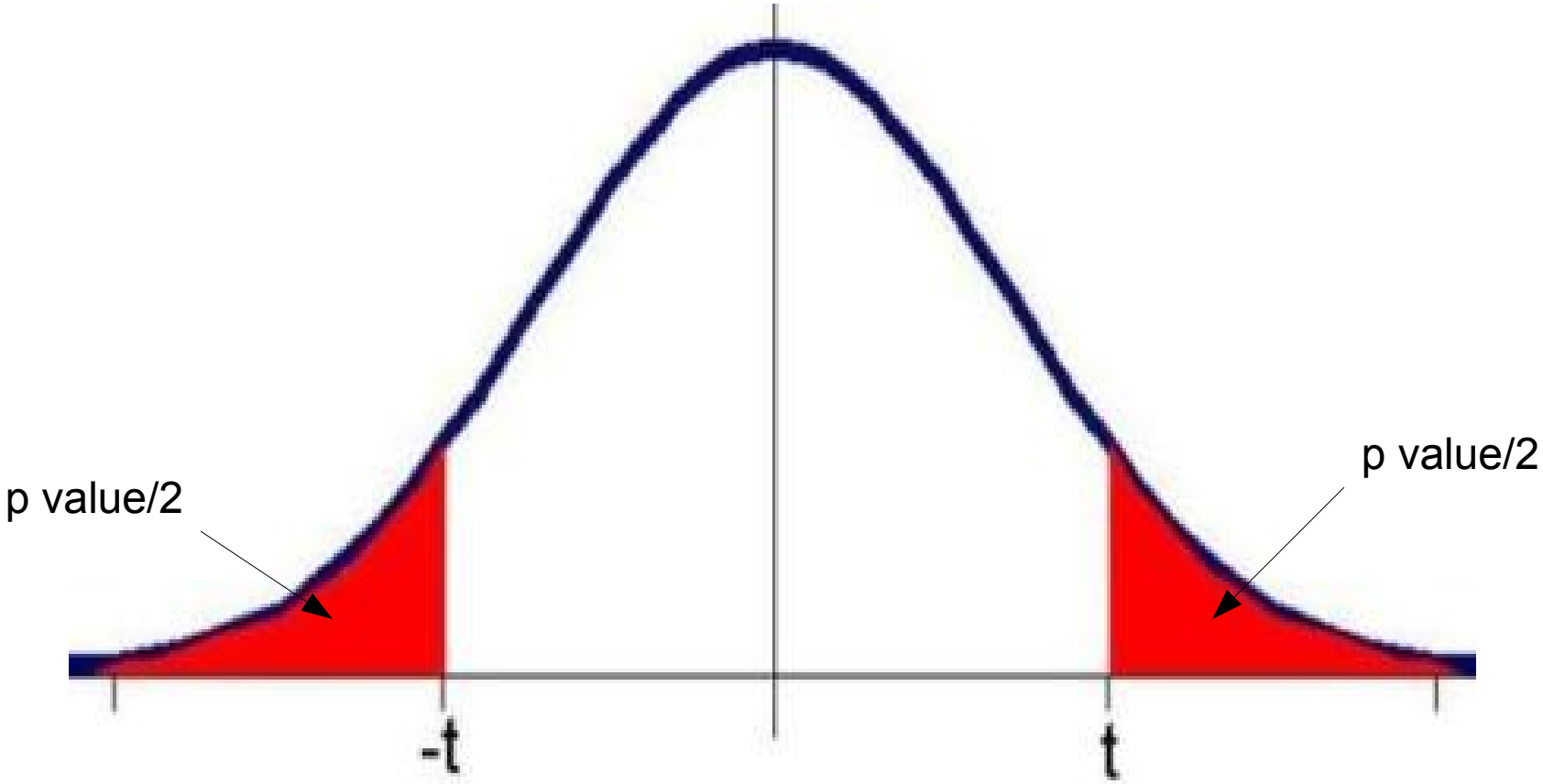
Significance level

Probability for a false positive outcome of the test,
the error of rejecting a null hypothesis when it is actually true

P-Value

Probability of obtaining the observed test-statistic or higher under the assumption, that the null hypothesis holds.

Hypothesis testing – p value



T-test (Welch-test)

Assumption: The values are normally distributed (note that for the normal t-test equal variances are assumed)

Teststatistik:

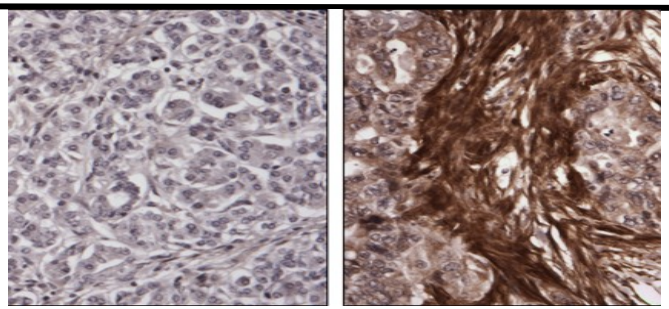
$$t = \frac{mean(N) - mean(T)}{\sqrt{\frac{sd(N)^2}{m} + \frac{sd(T)^2}{n}}}$$

the greater | t |, the greater the differential expression of gene X .

From t statistic to p value: t-value and significance level determine the p value (look-up tables)

Example

$N = \{5,7,6,9,5\}$



$T = \{2,4,3,5,3\}$

Hypothesis $H1: \mu_N - \mu_T \neq 0$ $H0: \mu_N - \mu_T = 0$

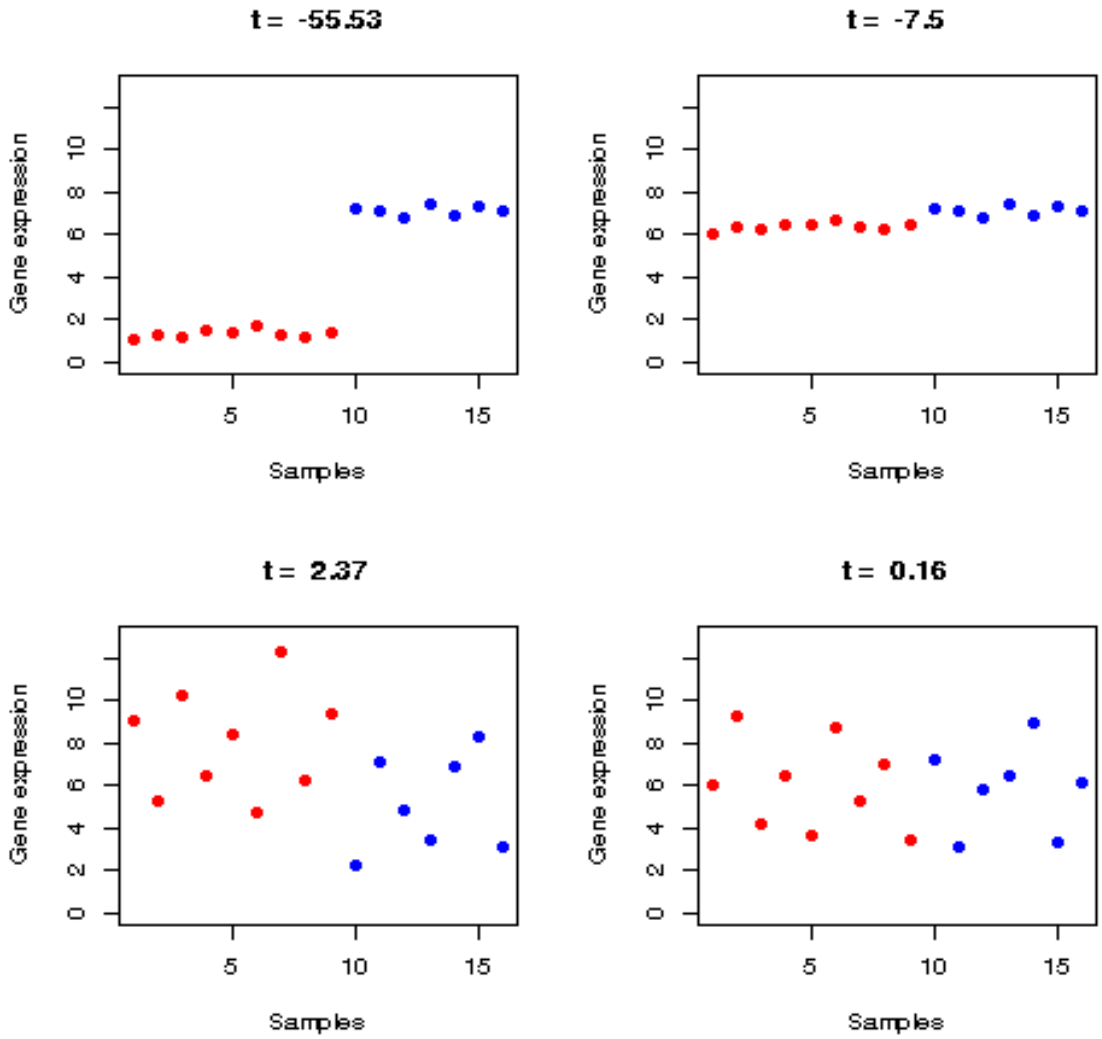
Significance level $\alpha = 0.05$

Test statistic

$$t = \frac{mean(N) - mean(T)}{\sqrt{\frac{sd(N)^2}{m} + \frac{sd(T)^2}{n}}} = 3.3129$$

P-Value $p - value = 0.0126$

Example



Further Methods

ANOVA – comparing more than one group as well as different factors.

SAM – Significance analysis of Microarrays. An 'improvement' of the t-test, as small variances can lead to very significant results without a considerable fold change.

Rank Produkt – sort genes by expression and determine Geometric mean of rank.

Multiple Testing Correction

Problem: Microarrays contain up to 20 000 genes, thus an α=0.05 leads to 20 000 * 0.05 = 1000 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 test, rarely used in practice.

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: $BH(p_i)_{i=1,...,m} = p_i * m/i$

4. BH (p) = significant if BH(p) ≤ α

Genes	p-value	rank	BH(p)	Significant? (α=0.05)
Gene A	0.00001	1	1000/1*0.00001=0.01	yes
Gene B	0.0004	2	1000/2*0.0004=0.02	yes
Gene C	0.01	3	1000/3*0.01=3.33	no

Clustering - Motivation

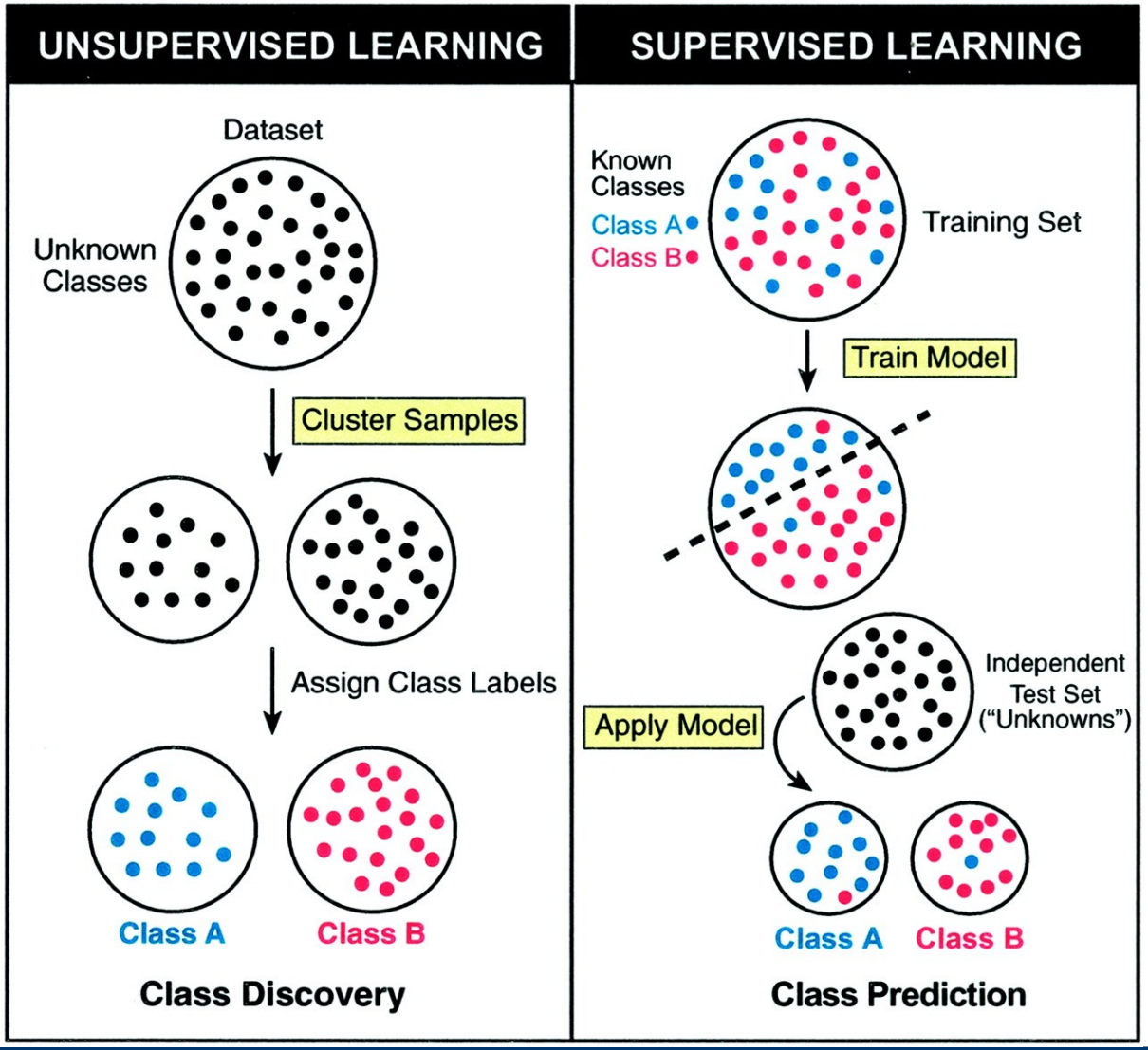
High dimensional data possibly containing all kinds of patterns and behavior of subgroups which might represent biolmedical phenomena. (explorative)

Clustering for **quality control.**

Expression patterns similar in spacial and temporal behavior → **co-regulated / expressed genes** (e.g. genes controlled by the same transkriptionfactor).

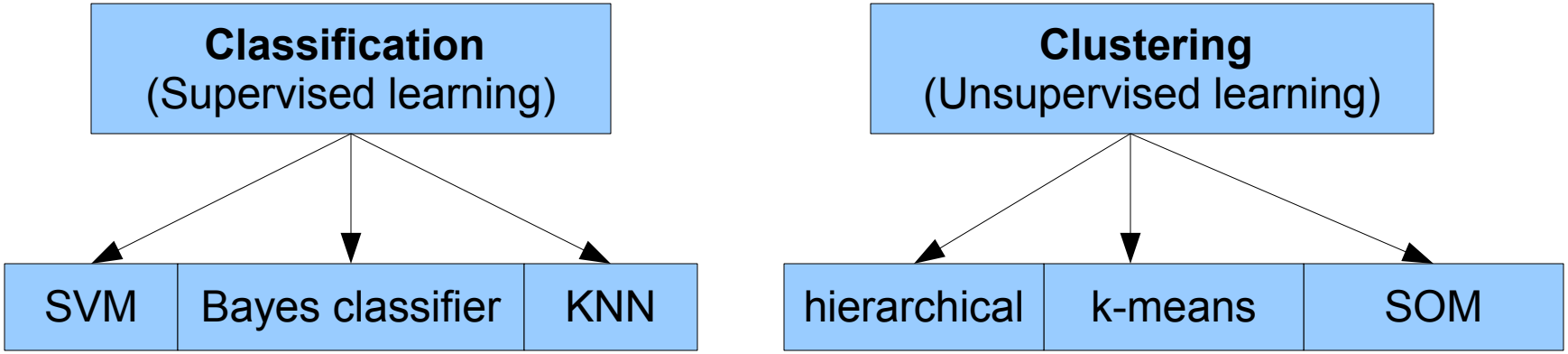
Discover new **disease subtypes** by clustering samples.

Clustering

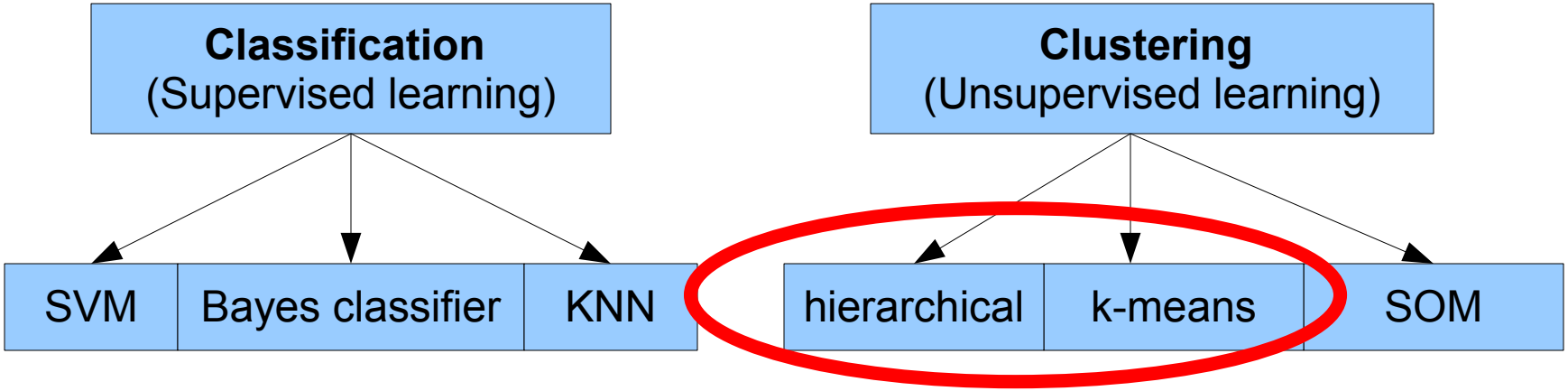


Ramaswamy & Golub 2002

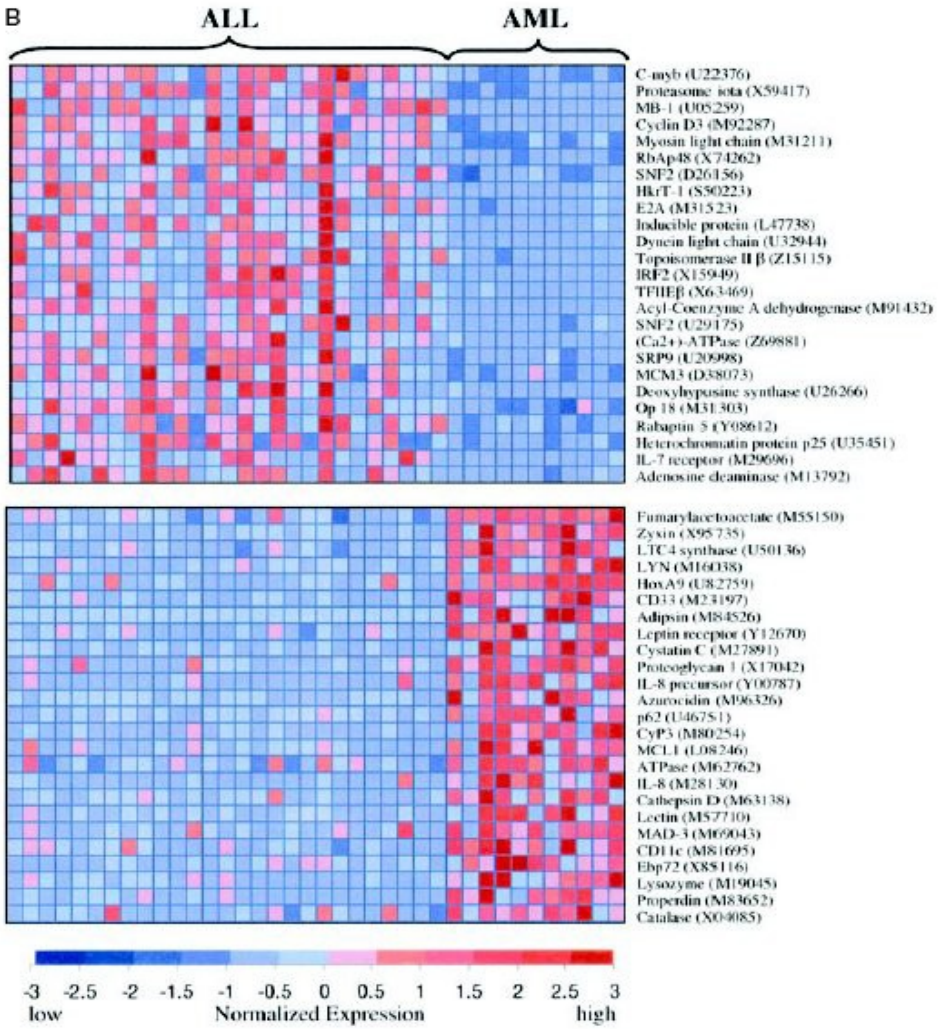
Clustering - Overwiev



Clustering - Overwiev



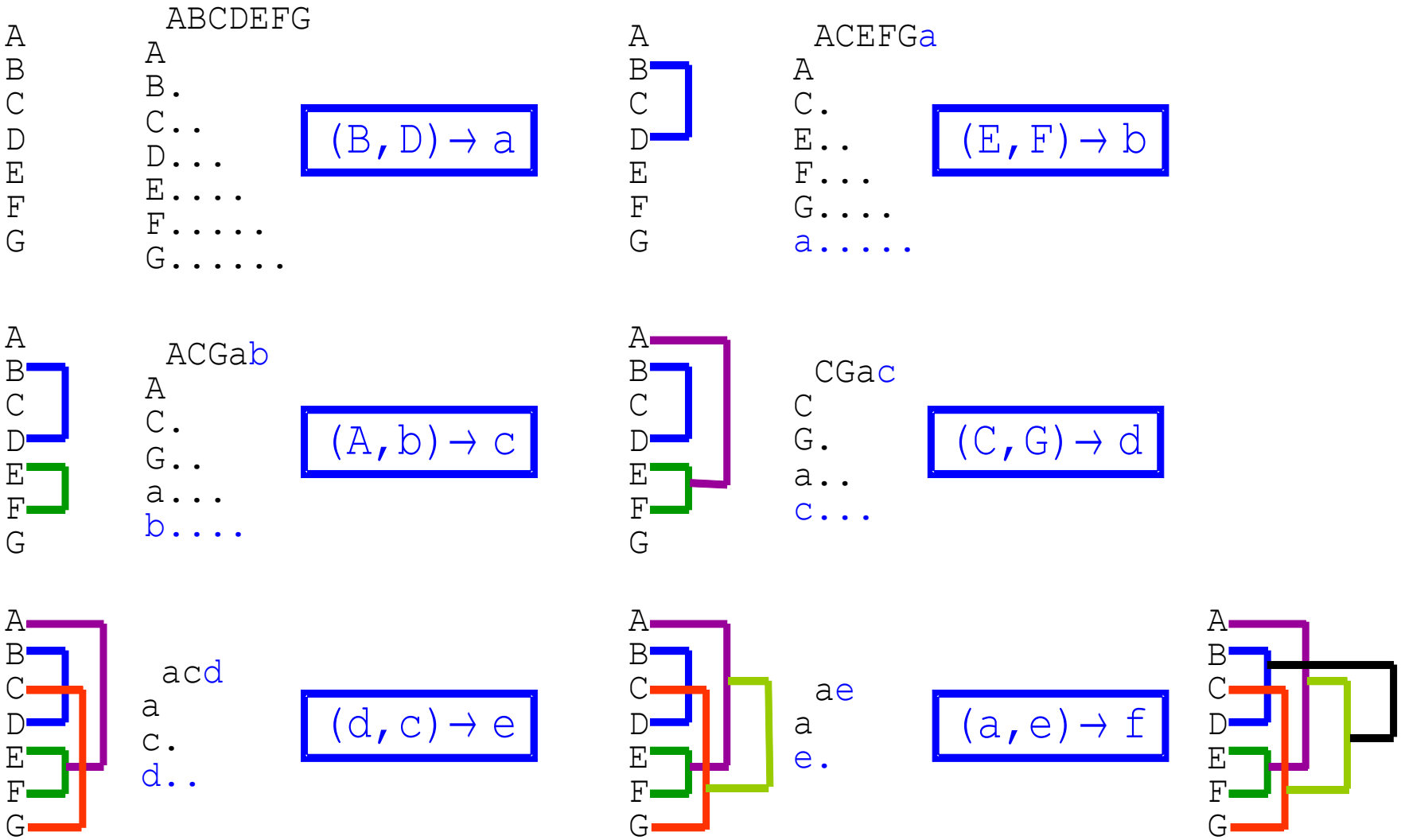
Clustering - Example



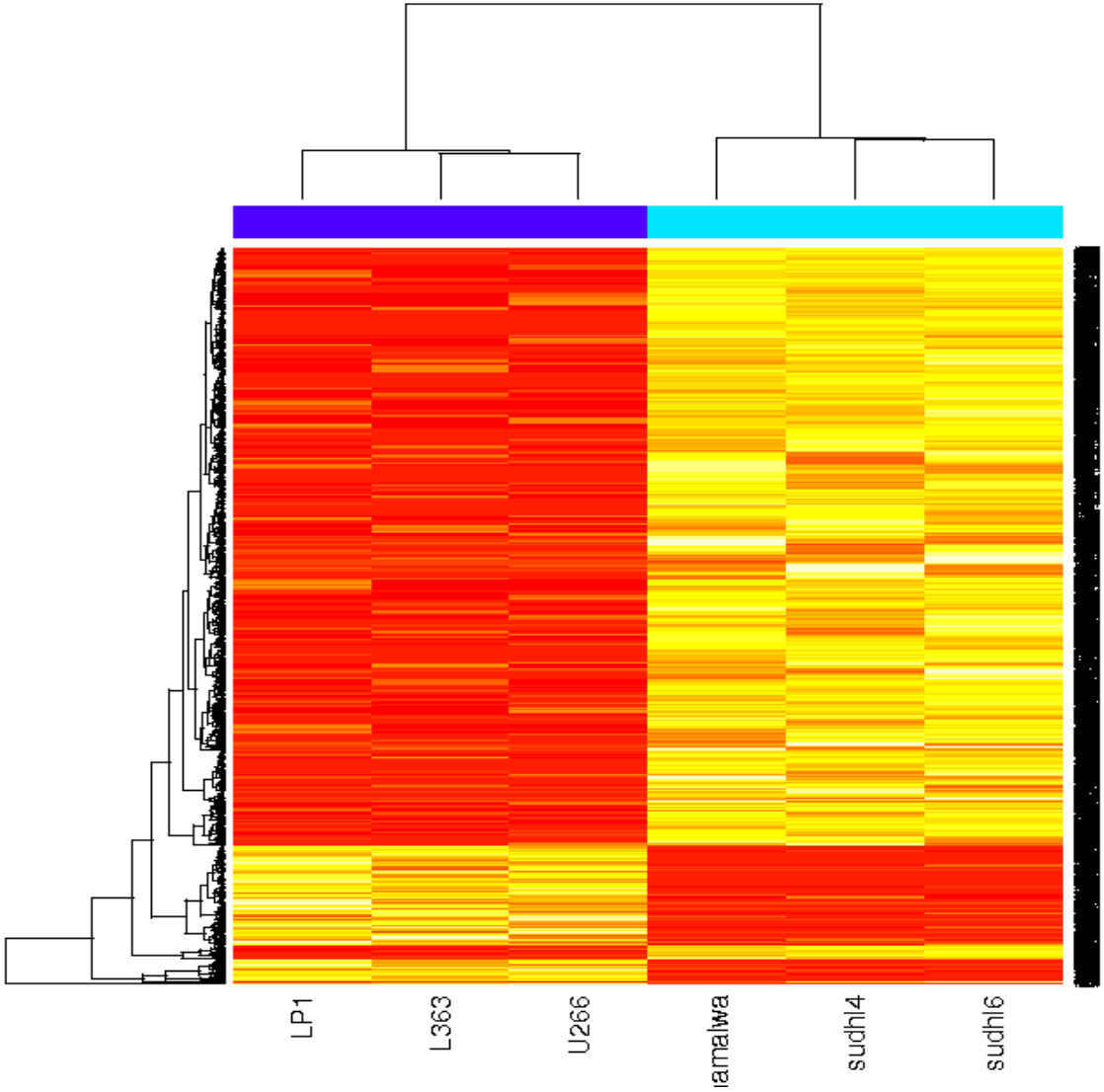
Hierarchical Clustering - Algorithm

1. choose a distance measure (e.g. euclidean, Pearson, etc.)
2. compute similarity matrix S
3. compute all pairwise distances in the matrix
4. while |S|>1
    5. determine pair (X,Y) with minimal distance
    6. compute new value Z = avg (X,Y), (single, average, or complete linkage)
    7. delete X and Y in S, insert Z in S
    8. compute new distances of Z to all elements in S
    9. visualize X and Y as pair

Hierarchical Clustering - graphical



Hierarchical Clustering – real data



HC

Result: binary tree, clusters have to be determined by the user.

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

The quality of the clustering can (then) be determined by the ratio of the mean distance in the cluster to the mean distance to points not in the cluster. Can be used as a measure for the cluster borders.

Dendrogram not unambiguous, 2^n possibilities. An $O(n^4)$ algorithm is known to optimize the dendrogram.

K means

1. choose k random cluster centers $\mu_1,...\mu_k$.
2. for all x in the dataset S compute nearest cluster center
3. for all Clusters C_i compute its cost:

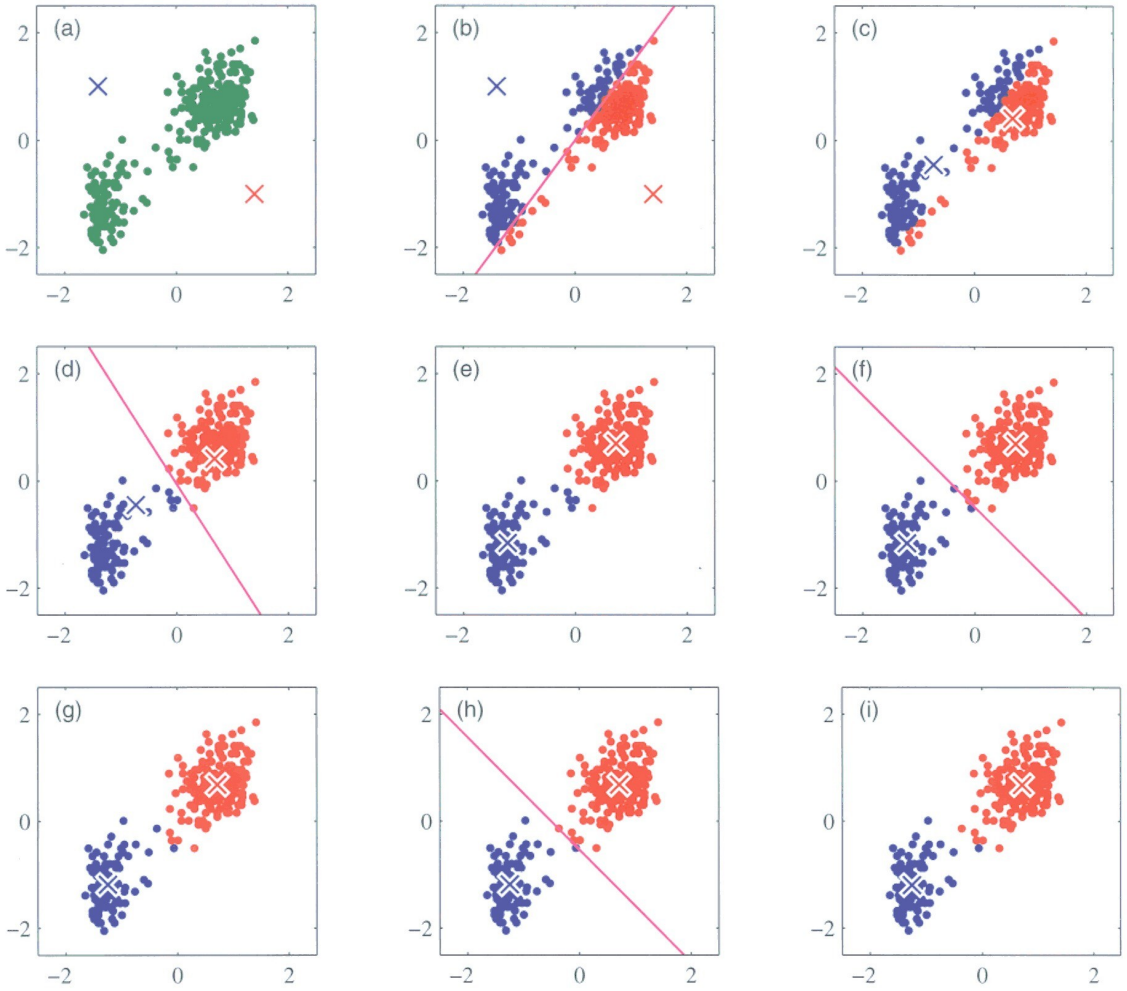
$$cost(Ci)=\sum_{r=1...|Ci|}(d(\mu_i,x_{r,i}))$$

4. compute a new center μ_i for every cluster C_i

$$c(Ci)=1/|Ci|\sum_{r=1}^{|Ci|}xri$$

5. 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 clustercenters for all clusters.

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

Standards

To determine the comparability of different experiments detailed information on the different steps is necessary.

RNA extraction,
cDNA rewriting,
labeling,
hybridization to microarray,
scanning,
spot detection,
spot intensity to numeric values,
normalization

MIAME

MIAME describes the **Minimum Information About a Microarray Experiment** that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment.

MIAME does **not** specify a particular **format** (→ use **MAGE-TAB** or **MAGE-ML**)

MIAME does **not** specify any particular **terminology** (use **MGED-ontology**)

MIAME Specification

1. **raw data** (.CEL, .gpr)

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

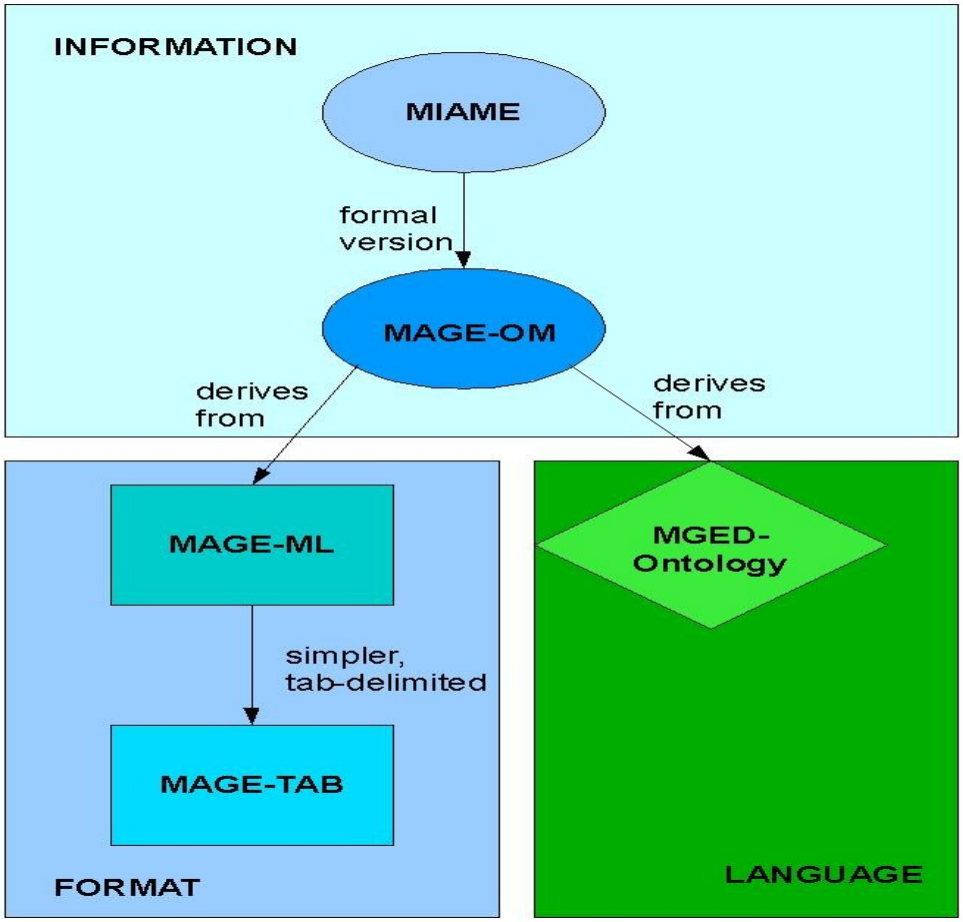
3. **sample annotation** (incl. Experimental factors and their values)

4. **experimental design** including sample data relationships (e.g.,hybridisations 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)

Standards - Overview



Standards - Overview

	DNA Microarray Data	High-throughput Sequencing Data	In Situ Hybridization and Immunohisto-chemistry Data	Tissue Microarray Data	Proteomics Data
Minimum Information Specification	MIAME	MINSEQE	MISFISHIE	???	MAIPE
Data Model	MAGE-OM	?	?	TMA-OM	PSI-OM
XML format	MAGE-ML	?	?	TMA-DES	PSI-ML
TAB-del. format	MAGE-TAB	?	?	TMA-TAB	?
Controlled vocabulary	MGED-ontology	?	?	?	?

Databases

GEO (Gene Expression Omnibus)
Array Express

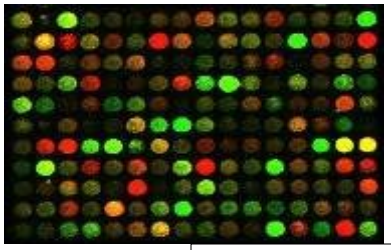
GEO – Gene Expression Omnibus

NCBI public repository
RDBMS schema

GPL
platform description

submitted by manufacturer

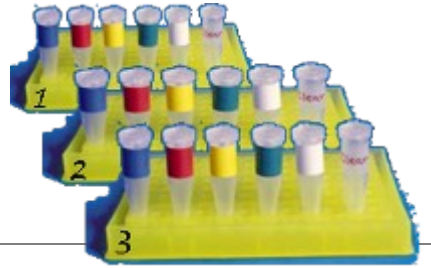
GSM
raw-processed intensities from a single or chip



GSE
grouping of chip data, a single experiment

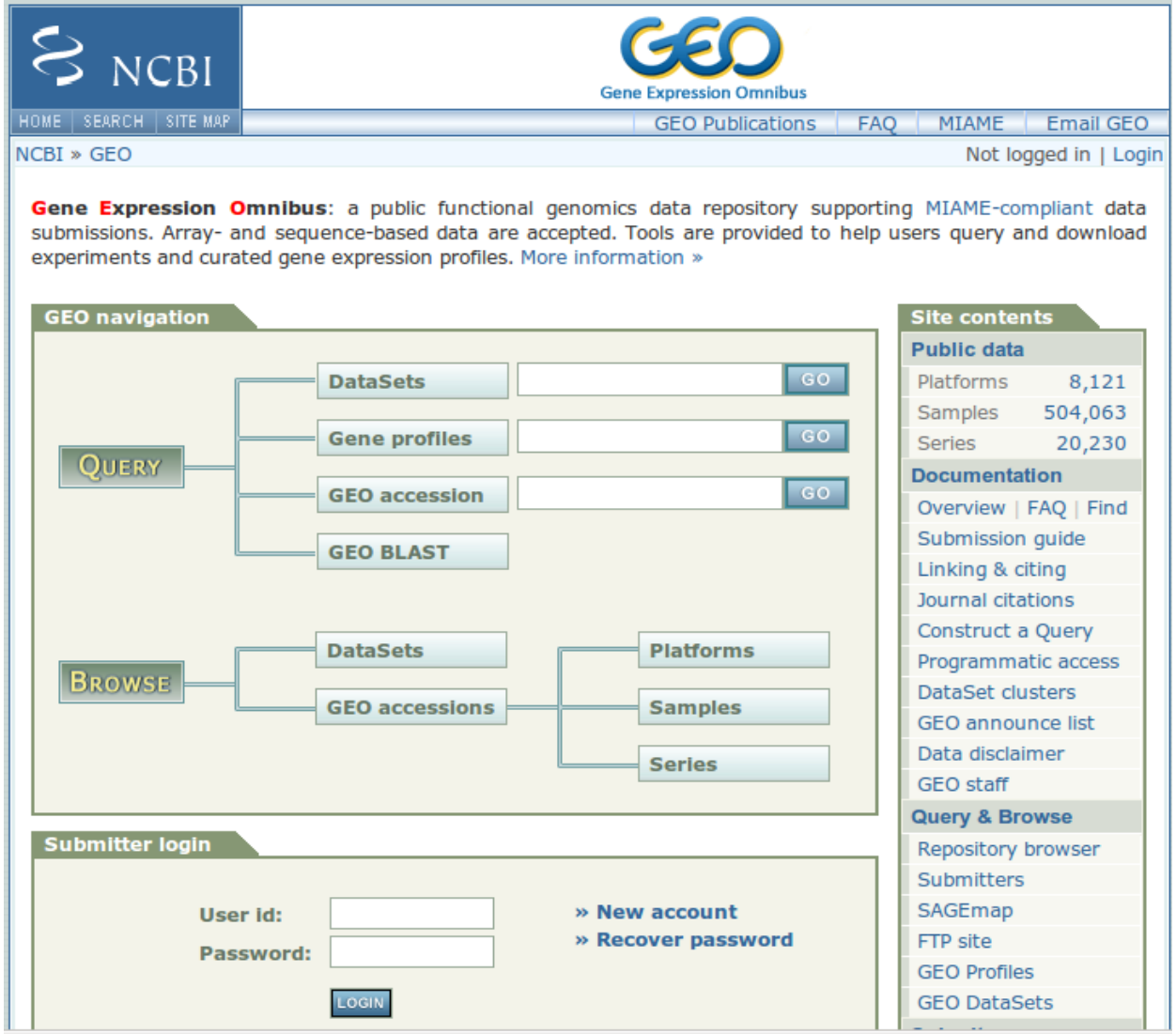
submitted by experimentalist

GDS
grouping of experiments

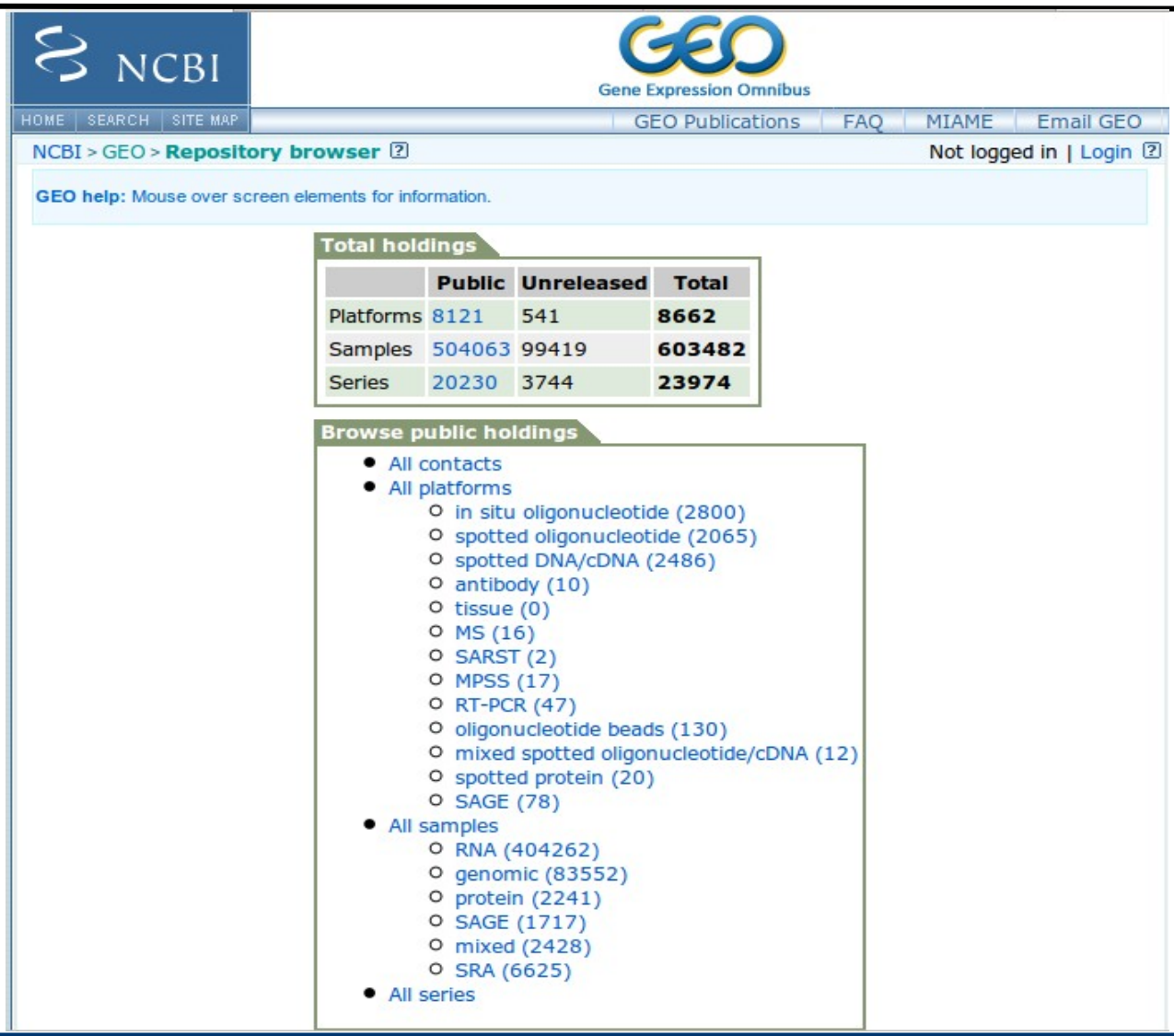


curated by NCBI

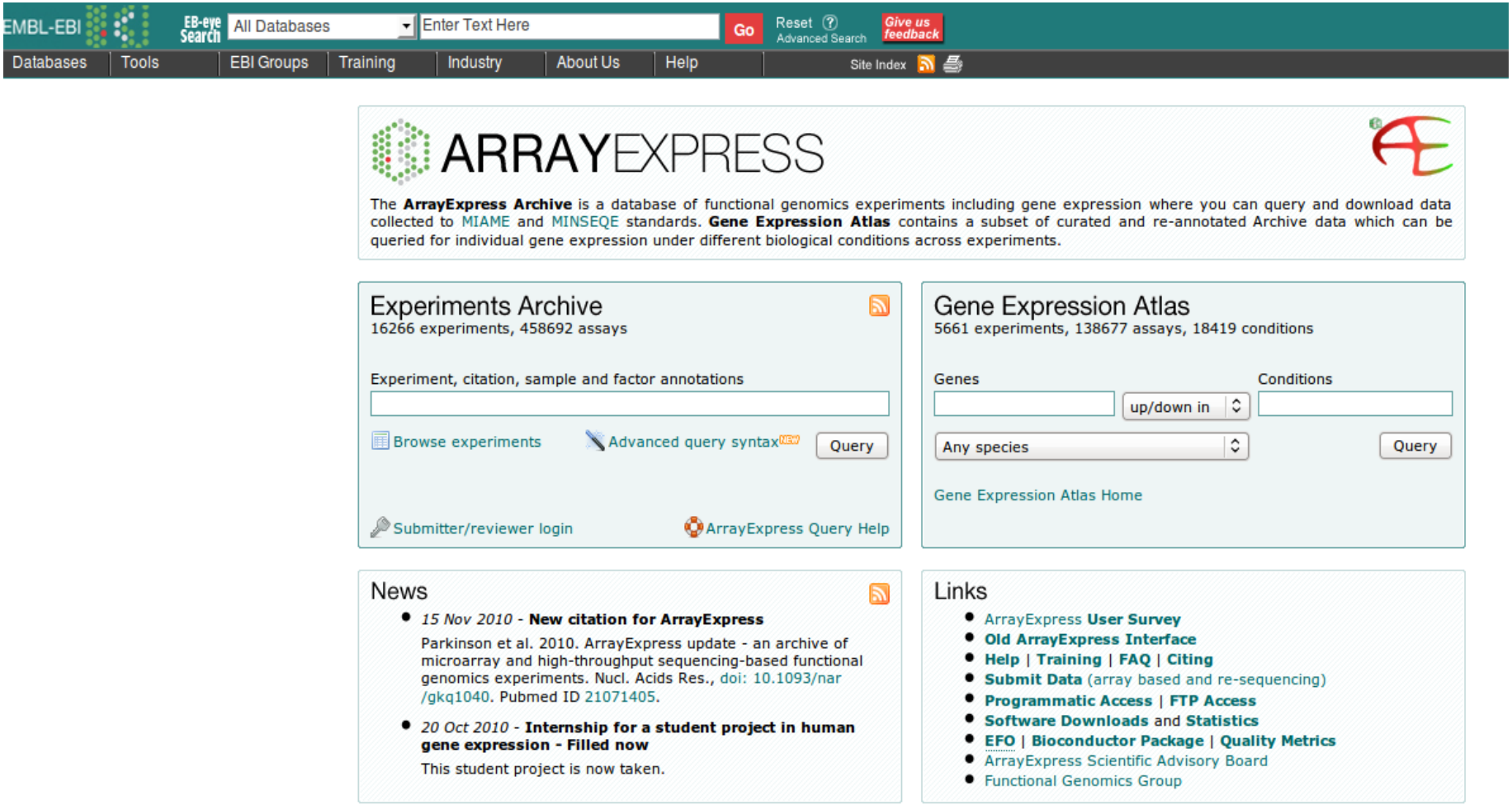
GEO


NCBI
HOME
SEARCH
SITE MAP
Gene Expression Omnibus
GEO Publications
FAQ
MIAME
Email GEO
NCBI » GEO
Not logged in | Login
Gene Expression Omnibus: 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. More information »
GEO navigation
QUERY
DataSets
Gene profiles
GEO accession
GEO BLAST
GO
BROWSE
DataSets
GEO accessions
Platforms
Samples
Series
Submitter login
User id:
Password:
LOGIN
» New account
» Recover password
Site contents
Public data
Platforms 8,121
Samples 504,063
Series 20,230
Documentation
Overview | FAQ | Find
Submission guide
Linking & citing
Journal citations
Construct a Query
Programmatic access
DataSet clusters
GEO announce list
Data disclaimer
GEO staff
Query & Browse
Repository browser
Submitters
SAGEmap
FTP site
GEO Profiles
GEO DataSets

GEO


NCBI
Gene Expression Omnibus
HOME
SEARCH
SITE MAP
GEO Publications
FAQ
MIAME
Email GEO
NCBI > GEO > Repository browser
Not logged in | Login
GEO help: Mouse over screen elements for information.
Total holdings
Public
Unreleased
Total
Platforms 8121 541 8662
Samples 504063 99419 603482
Series 20230 3744 23974
Browse public holdings
All contacts
All platforms
in situ oligonucleotide (2800)
spotted oligonucleotide (2065)
spotted DNA/cDNA (2486)
antibody (10)
tissue (0)
MS (16)
SARST (2)
MPSS (17)
RT-PCR (47)
oligonucleotide beads (130)
mixed spotted oligonucleotide/cDNA (12)
spotted protein (20)
SAGE (78)
All samples
RNA (404262)
genomic (83552)
protein (2241)
SAGE (1717)
mixed (2428)
SRA (6625)
All series

ArrayExpress (EMBL-EBI)



GEO vs. ArrayExpress

- both encompass MIAME compliance

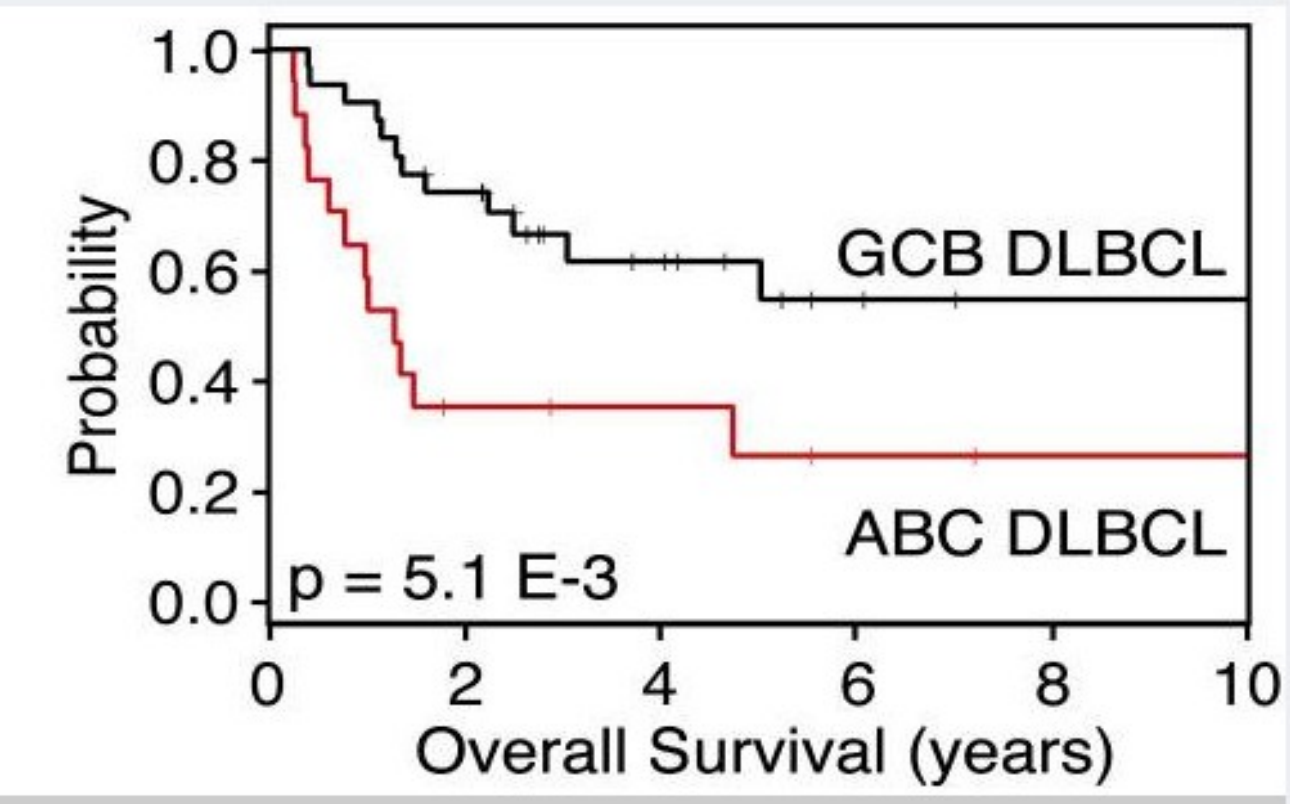
- both provide a good possibility for making data publicly availabe as often requested by journals

- GEO contains more data

- ArrayExpress provides analysis tools (and seq data?)

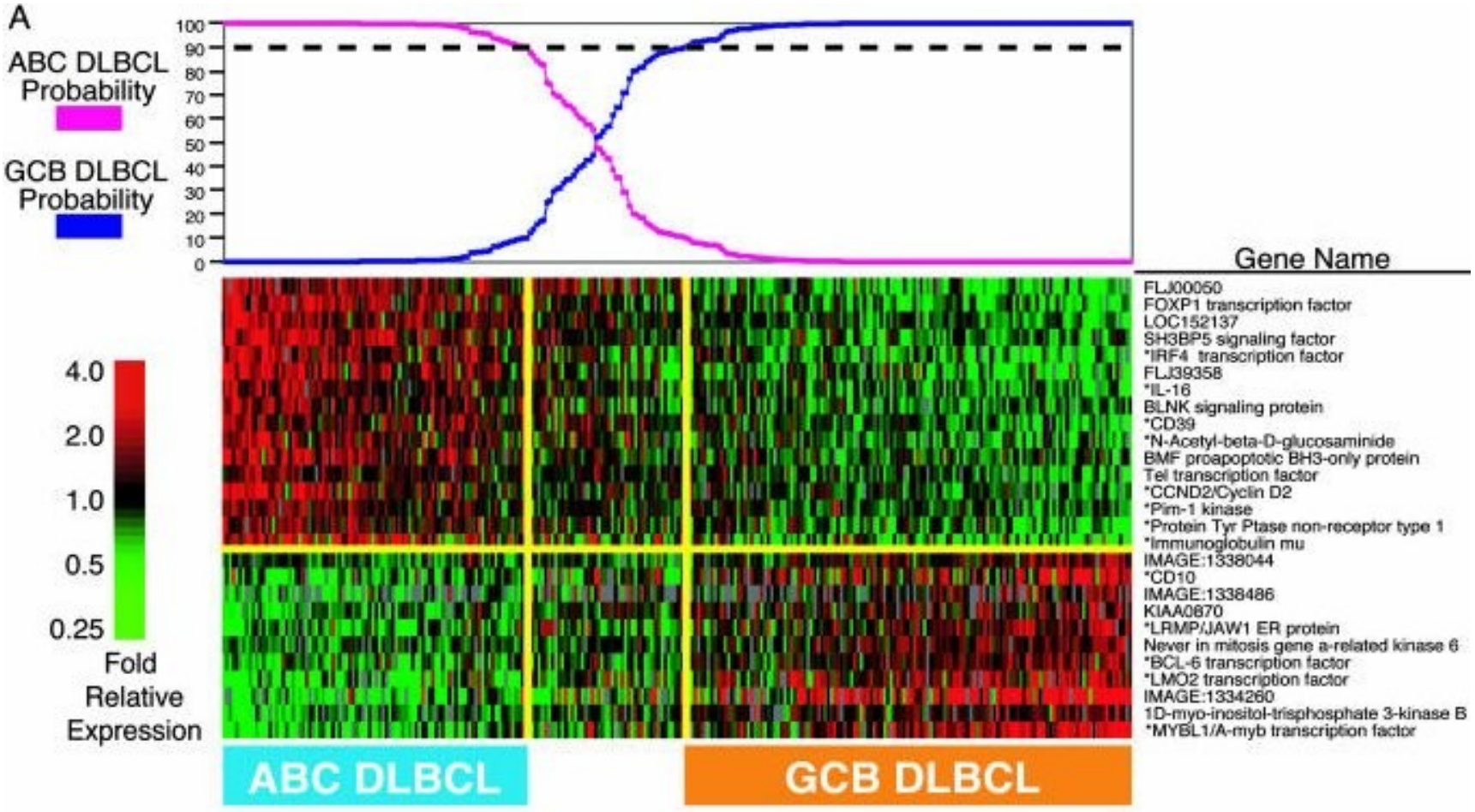
DLBCL Subtypes

germinal center B-cell-like (GCB), activated B-cell-like (ABC) with 5-year survival rates of 59% and 30%



Wright 2003

DLBCL Subtypes



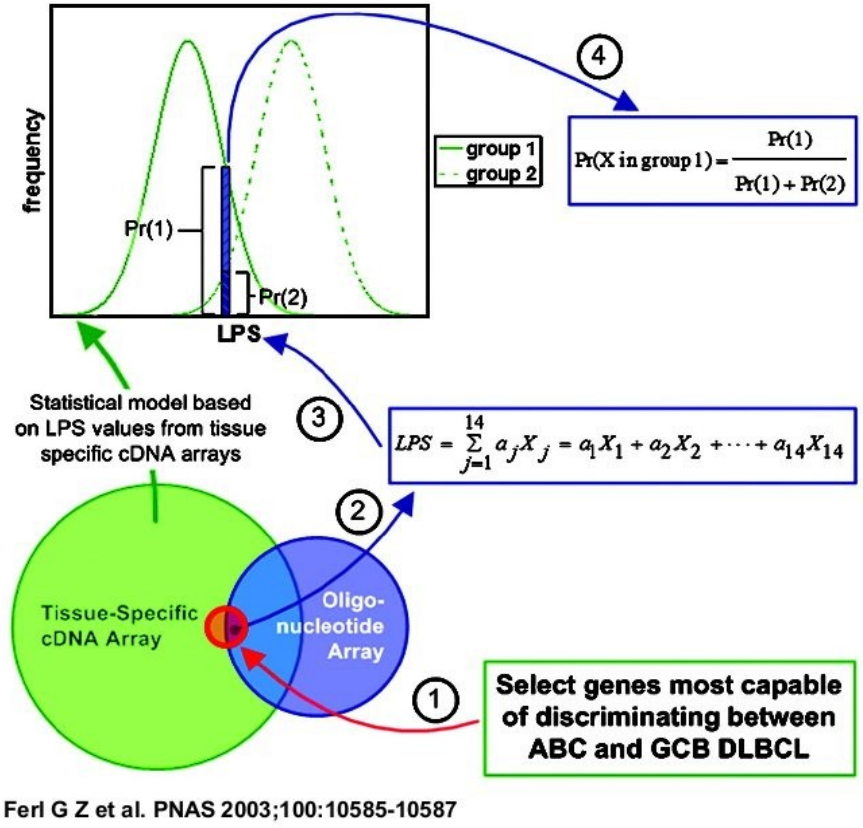
Wright 2003

DLBCL Subtypes

40 Exon arrays of DLBCL patients, subtype unknown.
Do we see the division in subgroups with a different technology and different probes?

DLBCL Subtypes

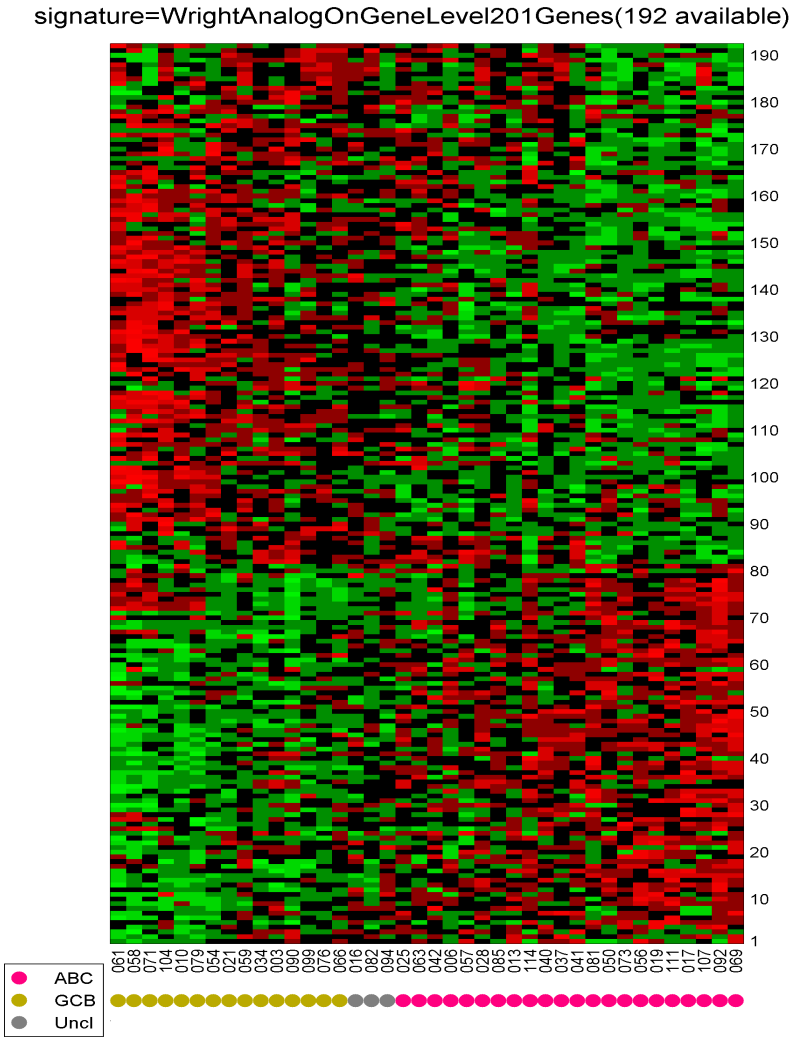
Schematic representation of how gene expression results can be compared across microarray platforms.



Ferl G Z et al. PNAS 2003;100:10585-10587

PNAS

DLBCL Subtypes



Summary

Combine t-test and fold change for optimal detection of differential expression.

More explorative analysis 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.