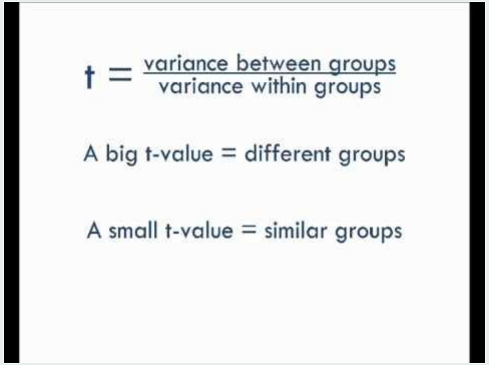
HUMBOLDT-UNIVERSITÄT · ZU BERLIN ·

Biostatistics

Grundlagen der Bioinformatik SS2018



$$t = \frac{\text{variance between groups}}{\text{variance within groups}}$$

A big t-value = different groups

A small t-value = similar groups

https://www.youtube.com/watch?v=0Pd3dc1GcHc

Agenda

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

Differential Expression

Motivation

- Etiology
- Biomarker
- Personalized medicine

Experimental Design

$N_1,...,N_m$: **control** samples

$T_1,...,T_n$: **case** samples

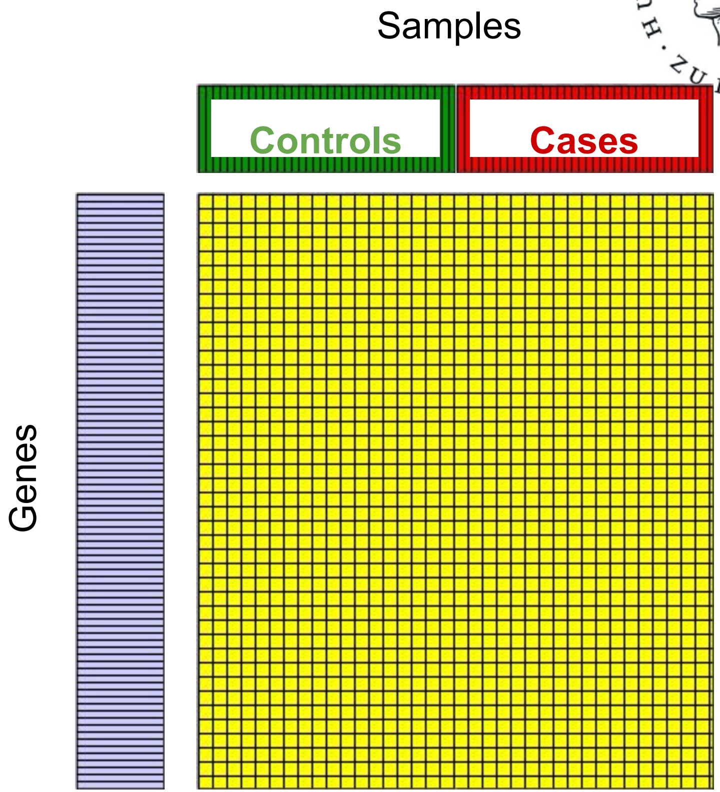
We **look for**:

Genes with significant differences between N and T

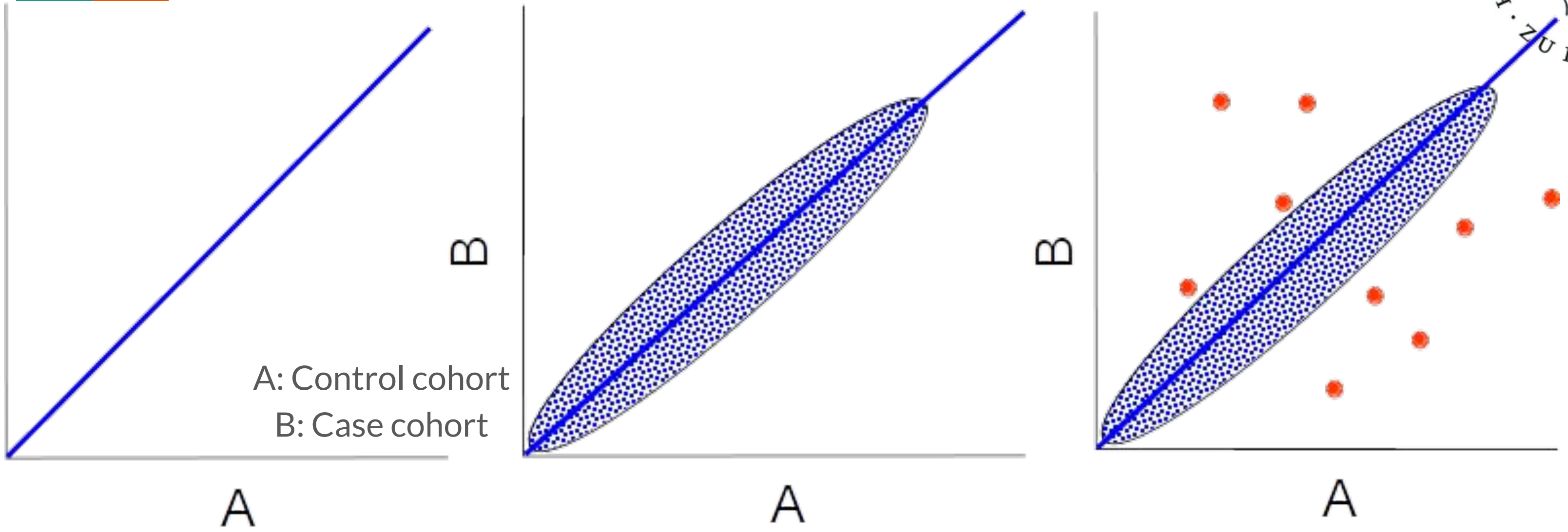
Compare gene X from group N with gene X of group T

$N = \{n_1,...,n_m\}$ $T = \{t_1,...,t_n\}$

Many methods exist, here: Fold change t-test



Scatterplot - Expression differences



A: Control cohort

B: Case cohort

Identical

Normal noise

Significant differences

Fold Change

$$FC = log_2(\frac{\overline{T}}{\overline{N}}) = log_2(\overline{T}) - log_2(\overline{N})$$

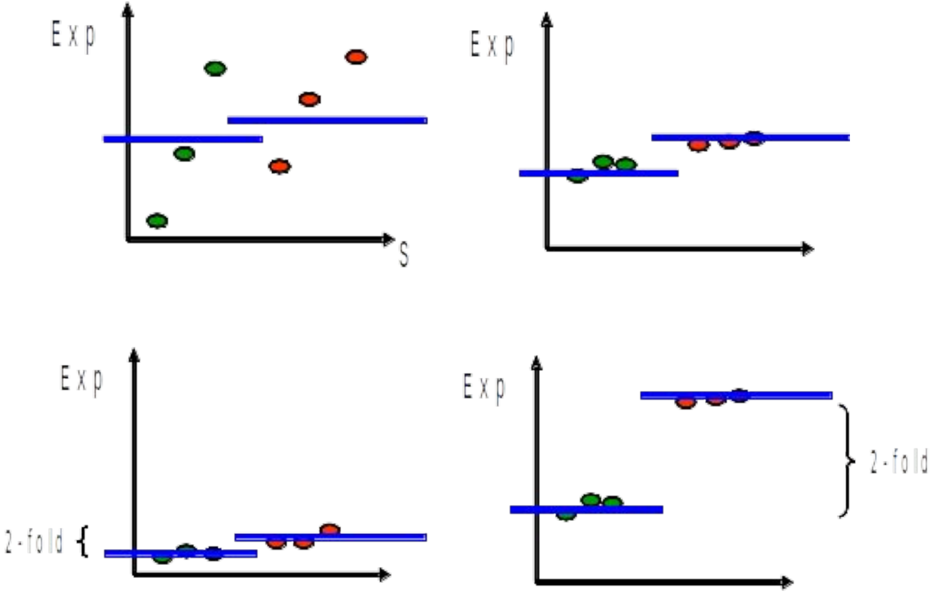
Thresholds (examples)

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

Genes	Mean Case	Mean Control	Mean Case / Control	FC
A	16	1	16	4
B	0.0625	1	0.0625	-4
C	10	10	1	0
D	200	1	200	7.65

Fold Change - Advantages / Disadvantages

- ✓ intuitive measure
- ✗ Independent of scatter
- ✗ Independent of absolute values
  - ○ Score only based on mean of groups
  - ○ **Spread** of data points essential

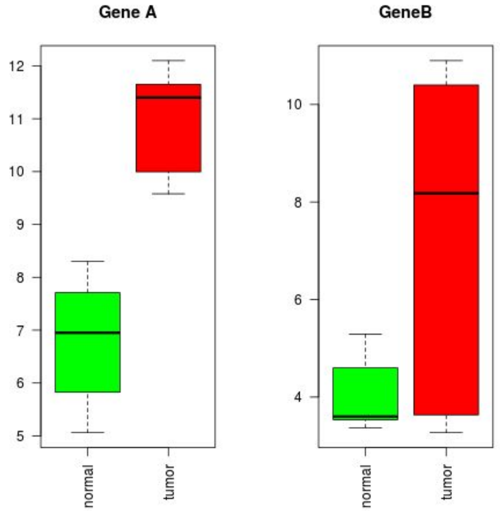


Variance essential

	N1	N2	N3	N4	N5	N6	N7	C1	C2	C3	C4	C5	C6	C7	FC
Gene A	5	5	8	8	7	6	7	10	10	12	12	11	10	12	-4
Gene B	3	4	3	3	5	5	4	4	11	10	4	11	8	3	-3

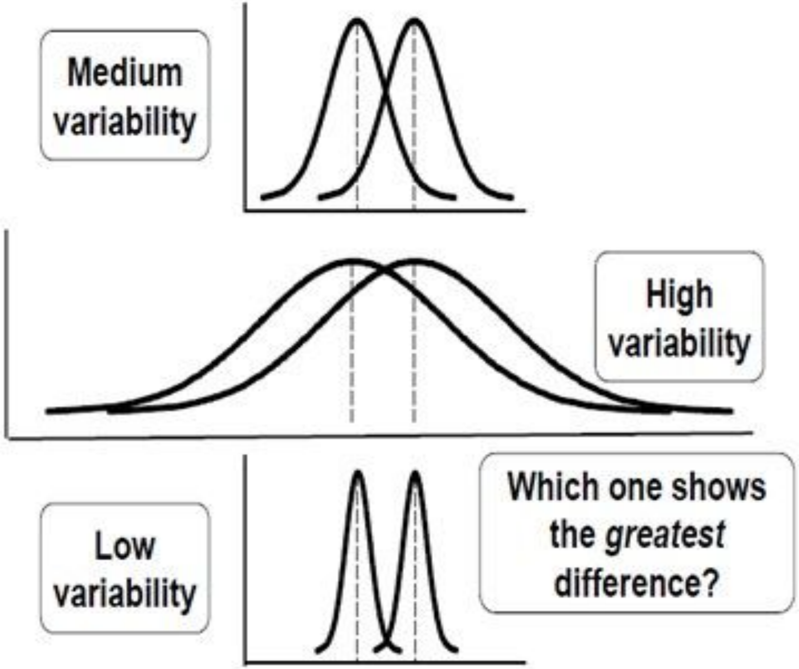
- High abs(FC) for Gene A and Gene B
- But: variance very high in the tumor samples of Gene B
- Find test for FC and variance

$$Var(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2$$



Hypothesis testing

- Same Mean
  - Different variance
- Measure 'uncertainty' with standard deviation *sd*
- Combine both to likelihood for 'correctness'
- Assumption
  - Log-Normal distributions
  - Symmetric
  - Independent



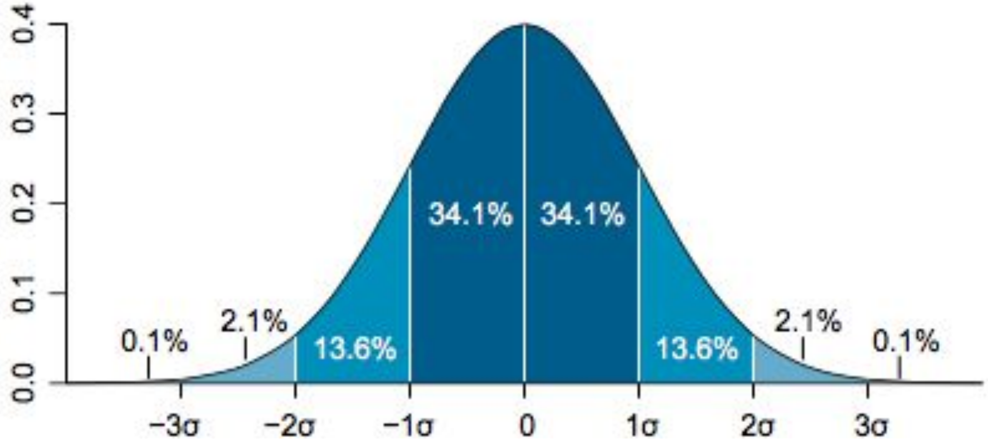
$$\sigma_X := \sqrt{\mathrm{Var}(X)}$$

$$Var(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2$$

Tschebyscheff-Inequation

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

- 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): $m_1 = m_2$ vs. $m_1 != m_2$ (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

Hypothesis Testing – T-test (Welch Test)

From T-statistic to p-value

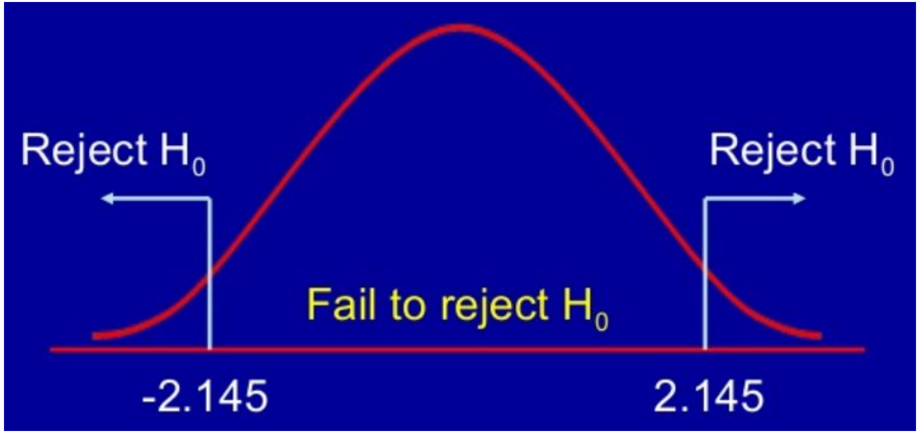
- T-value, a and number of samples determine the p-value (look-up tables)

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 (a)

- Probability of a false positive outcome of the test, the error of rejecting H_0 when it is actually true



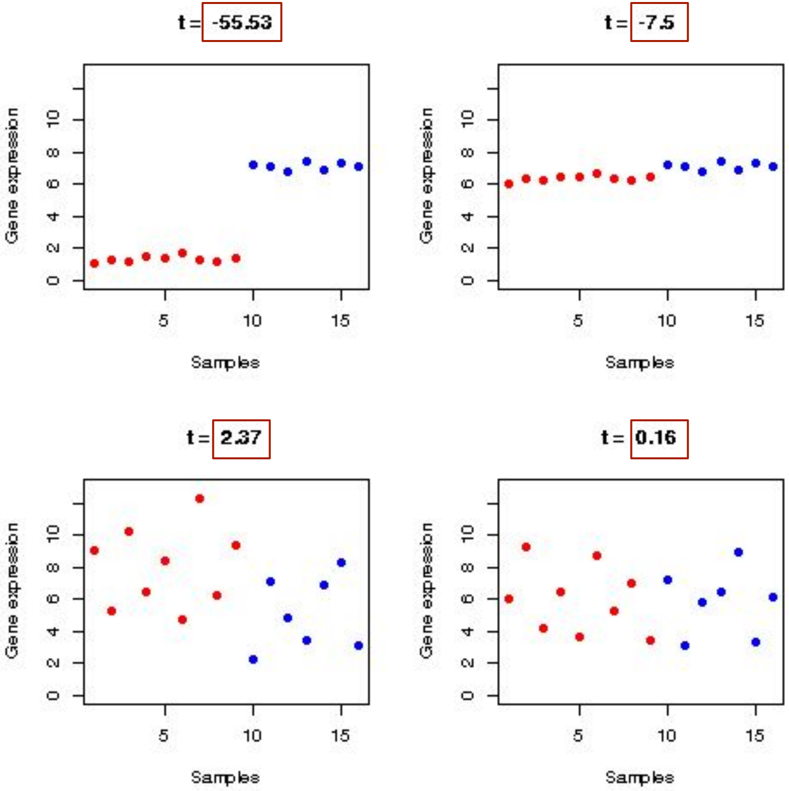
If |t| > |T| we reject H_0

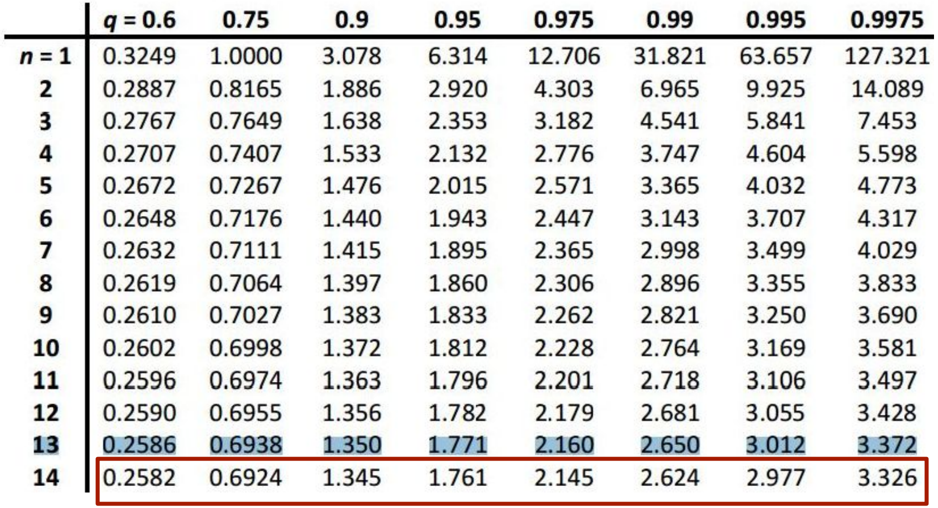
→ p-value is significant (p-value < a)

Workflow Hypothesis Testing

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

Examples





	q = 0.6	0.75	0.9	0.95	0.975	0.99	0.995	0.9975
n **= 1**	0.3249	1.0000	3.078	6.314	12.706	31.821	63.657	127.321
2	0.2887	0.8165	1.886	2.920	4.303	6.965	9.925	14.089
3	0.2767	0.7649	1.638	2.353	3.182	4.541	5.841	7.453
4	0.2707	0.7407	1.533	2.132	2.776	3.747	4.604	5.598
5	0.2672	0.7267	1.476	2.015	2.571	3.365	4.032	4.773
6	0.2648	0.7176	1.440	1.943	2.447	3.143	3.707	4.317
7	0.2632	0.7111	1.415	1.895	2.365	2.998	3.499	4.029
8	0.2619	0.7064	1.397	1.860	2.306	2.896	3.355	3.833
9	0.2610	0.7027	1.383	1.833	2.262	2.821	3.250	3.690
10	0.2602	0.6998	1.372	1.812	2.228	2.764	3.169	3.581
11	0.2596	0.6974	1.363	1.796	2.201	2.718	3.106	3.497
12	0.2590	0.6955	1.356	1.782	2.179	2.681	3.055	3.428
13	0.2586	0.6938	1.350	1.771	2.160	2.650	3.012	3.372
14	0.2582	0.6924	1.345	1.761	2.145	2.624	2.977	3.326

Degrees of freedom: |Samples| - 2,
Here 16 - 2 = 14

Example

Hypothesis	$H_0 : m_N - m_T = 0$ vs $H_1 : m_N - m_T != 0$
Significance level	0.05
Test statistic	
P-value	0.06

-> Not significant

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}

Data from slide 9

$$t = \frac{X_1 - X_2}{S_p \cdot \sqrt{\frac{1}{n_1} \cdot \frac{1}{n_2}}}$$

= - 2.27

Critical value = 2.45

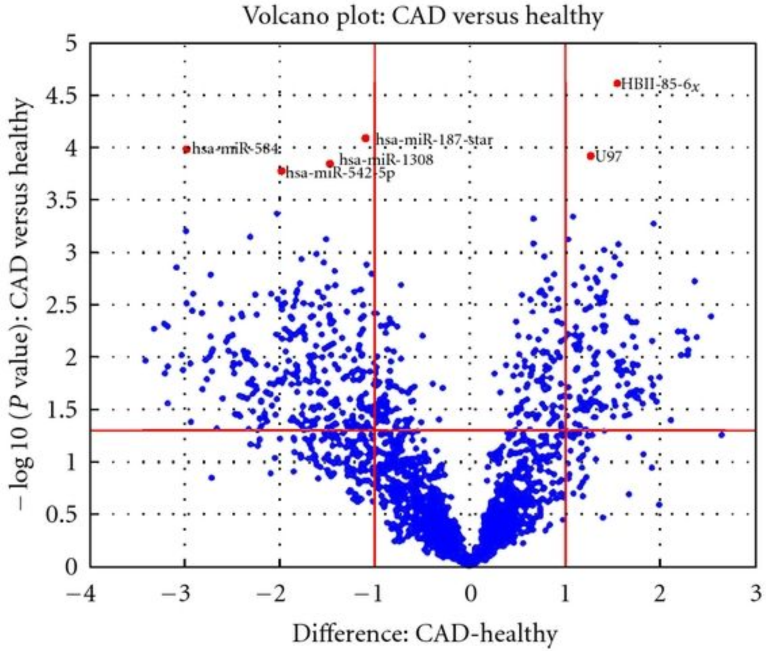
Volcano plot

Combine P-value and Log-FC

- Y-axis: Negative log10 of the p-value
- X-axis: Fold-change

Interested in

- Upper left
- Upper right corner



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

Log-FC

Multiple Testing Correction

Problem

Microarrays has 22k genes, thus an α=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 correction

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 H0), 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 correction

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(pi)i=1,...,m = p_i * m/i
4. BH (p) = significant if BH(p) ≤ α

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

Clustering - Motivation

Subgroups detection

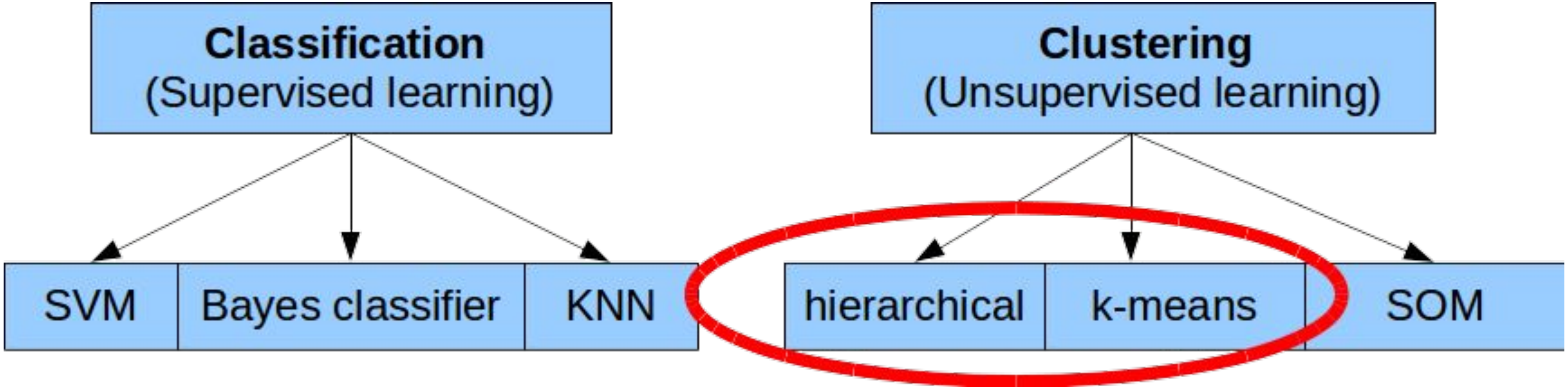
Quality control

Similarity-detection in spatial and temporal behavior

- Co-regulated / expressed genes
  - E.g. genes controlled by the same transcription-factor

Discovery of new disease subtypes

Overview unsupervised clustering



Classification
(Supervised learning)

SVM

Bayes classifier

KNN

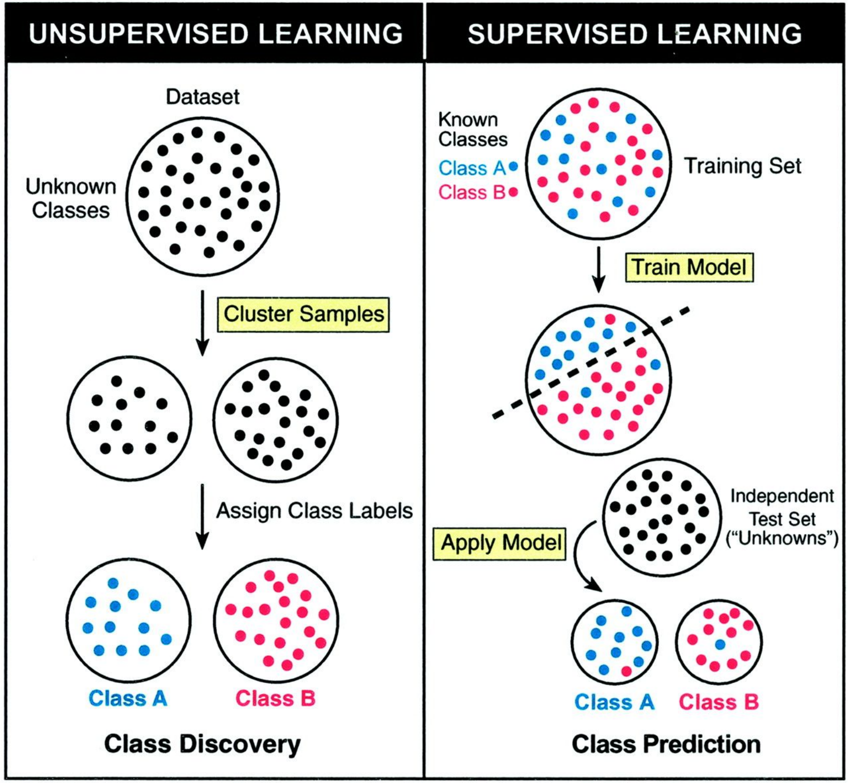
Clustering
(Unsupervised learning)

hierarchical

k-means

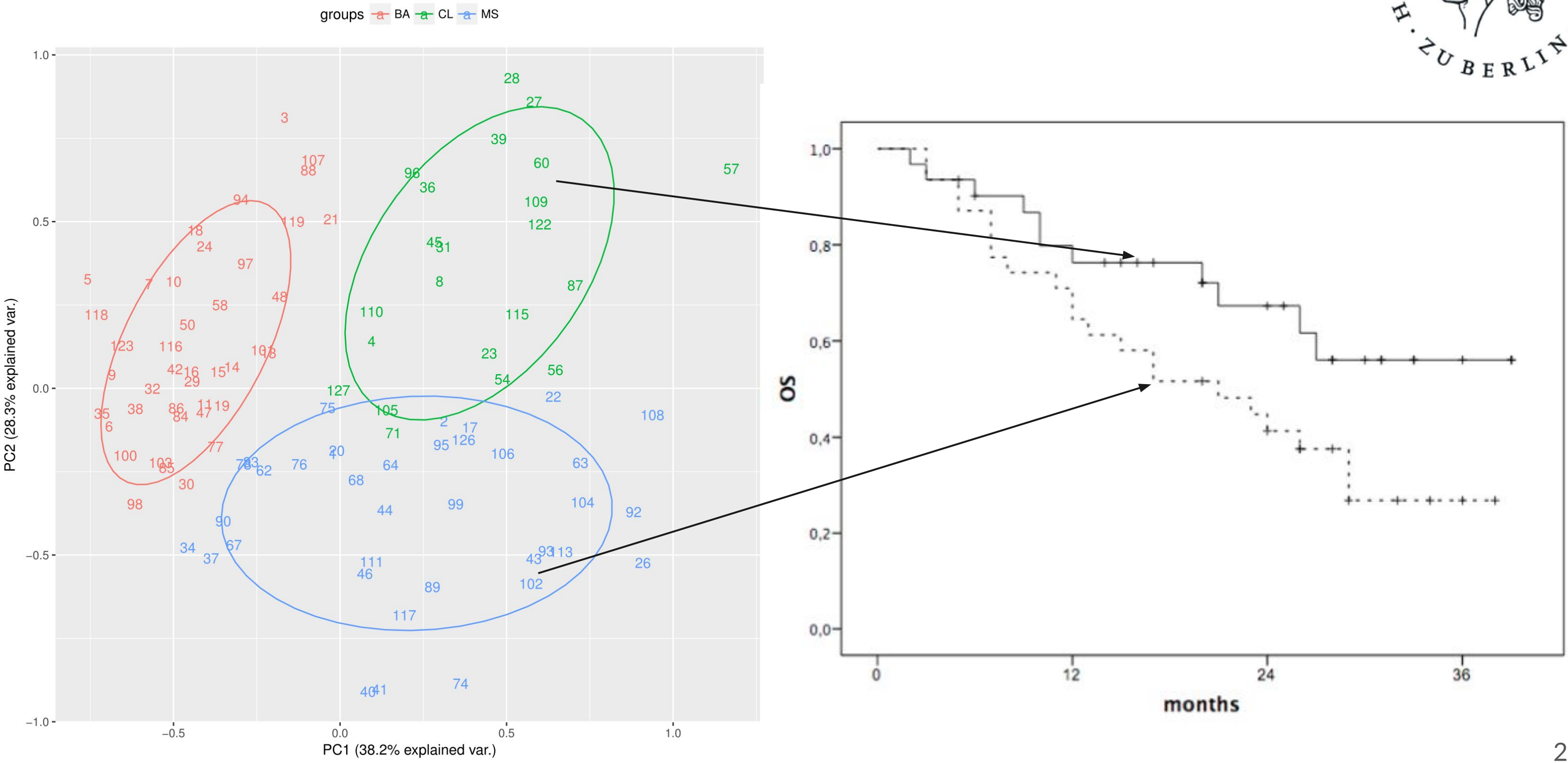
SOM

Clustering



Ramaswamy & Golub 2002

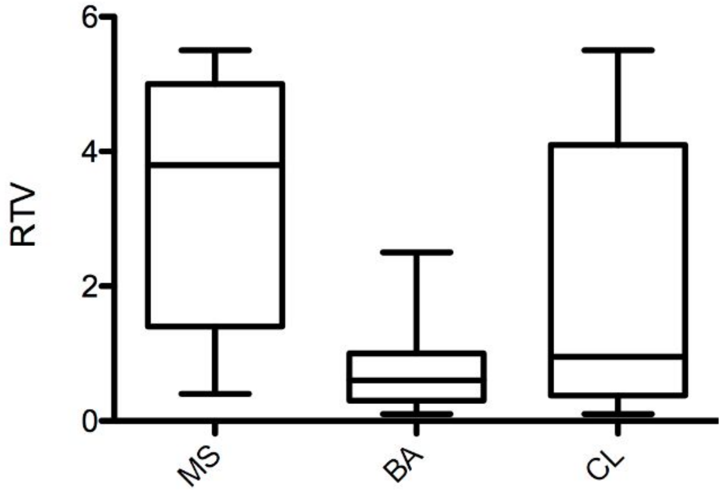
Example


groups
BA
CL
MS
PC2 (28.3% explained var.)
PC1 (38.2% explained var.)
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
28
27
39
60
57
96
36
109
122
45
31
8
87
110
115
4
23
56
54
127
105
71
3
107
88
94
119
21
18
24
97
5
7
10
48
58
118
50
123
116
103
9
42
16
15
14
29
32
11
119
35
38
86
84
47
6
77
100
102
85
30
98
22
108
75
2
17
95
126
106
20
78
83
62
76
64
63
68
44
99
104
92
90
34
67
37
111
46
43
93
113
26
102
89
117
40
41
74
OS
months
1,0
0,8
0,6
0,4
0,2
0,0
0
12
24
36

Clustering

- **Goal**
  - ○ Partitioning Biological interpretation of subtypes (clusters)
- **Requires**
  - ○ (Useful) similarity measure
- **Advantages**
  - ○ Intuitive Simple (you would think)

cetuximab response in different subtypes of HNSCC



Hierarchical Clustering - algorithm

1. Distance measure
    a. Euclidean
    b. Pearson, etc.

2. Compute similarity matrix S

3. While |S|>1:
    a. Determine pair (X,Y) with minimal distance
    b. Compute new value Z = avg (X,Y), (single, average, or complete linkage)
    c. Delete X and Y in S, insert Z in S
    d. Compute new distances of Z to all elements in S
    e. Visualize X and Y as pair

Hierarchical Clustering

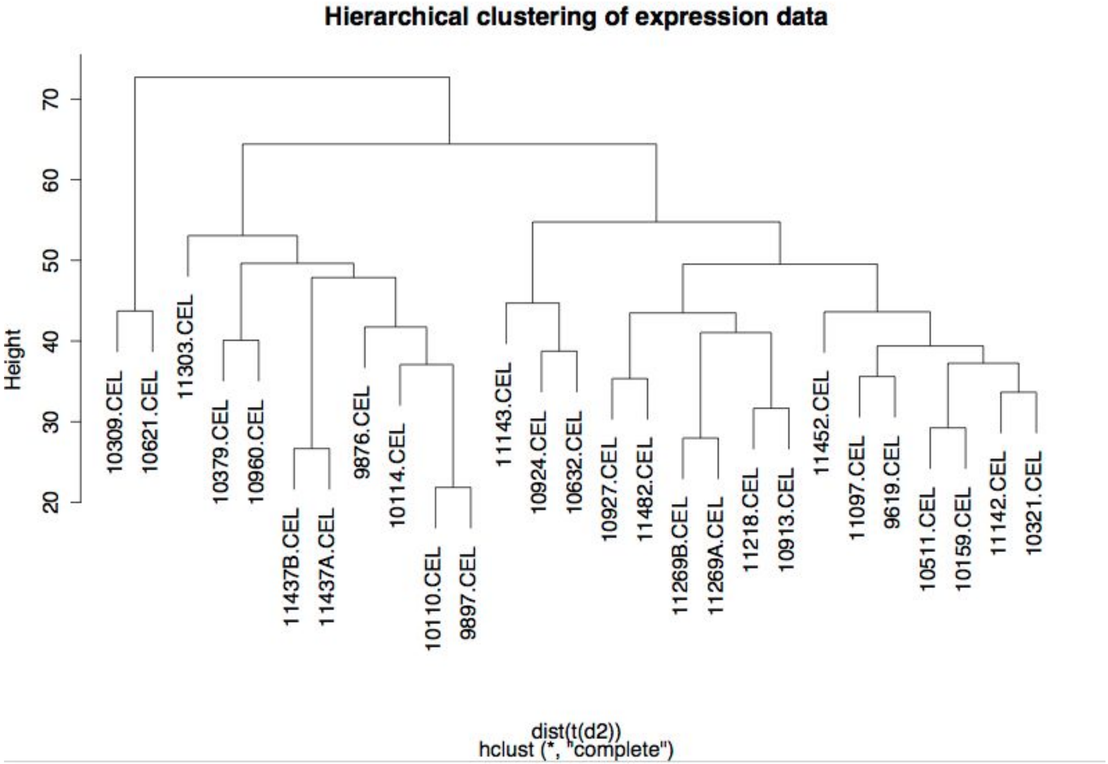
- Binary tree
- Cutting the dendrogram at a particular height partitions the data into disjoint clusters
- For an easier determination of clusters
  - Length of branch is set in relation to the difference of the leafs.

Linkage Rule essential

Hierarchical Clustering – Linkage

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

Hierarchical Clustering



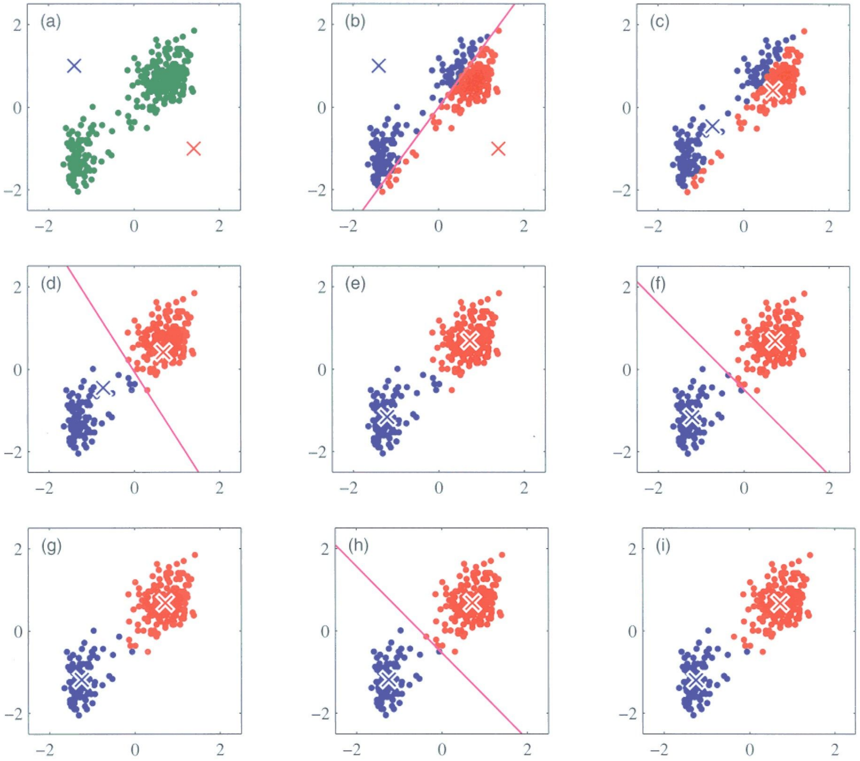
K-means

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 µ1,...µk
2. Assign for each point x in dataset S the closest cluster center
3. Compute a new center µi for every cluster Ci
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 not assured
- Cluster quality can be computed by determining the mean distance of a gene to its cluster-center
- 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.

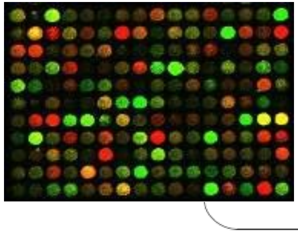
Databases - GEO – Gene Expression Omnibus

- 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

submitted by
manufacturer

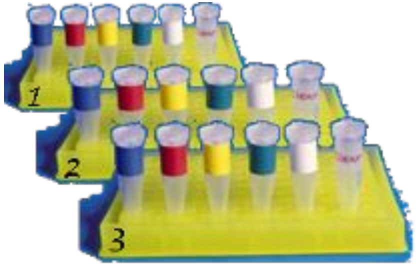
GSM
(GEO sample)
raw-processed
intensities from a
single or chip



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

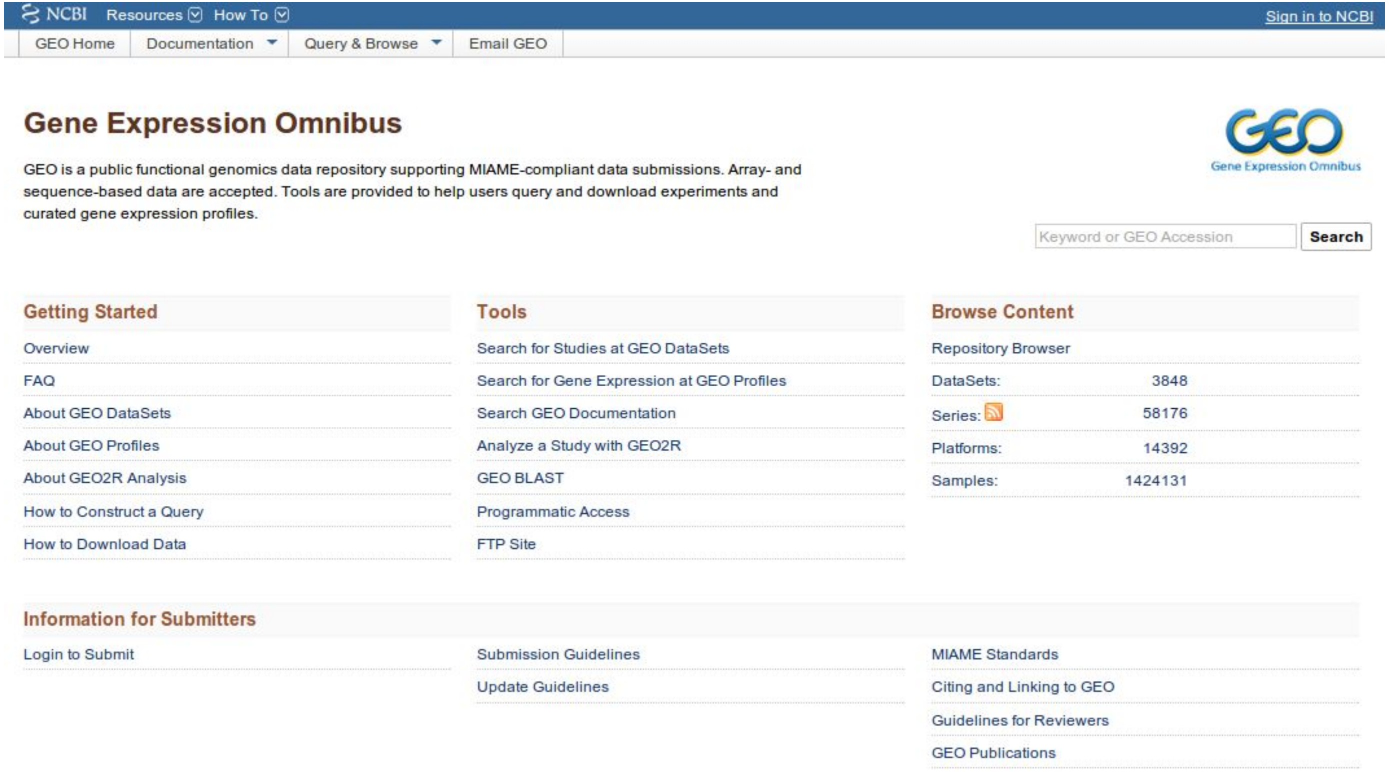
submitted by
experimentalist

GDS
(GEO dataset) grouping
of experiments



curated by NCBI

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.

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

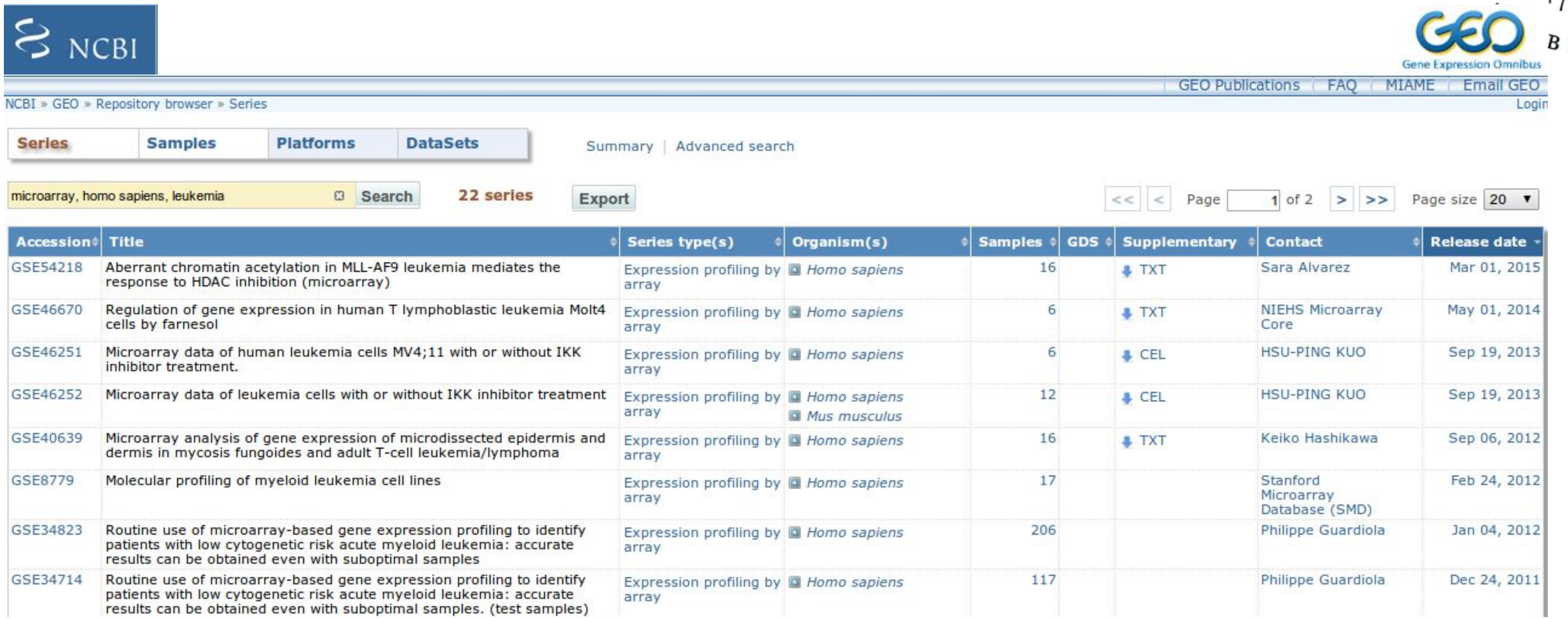
Repository Browser

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

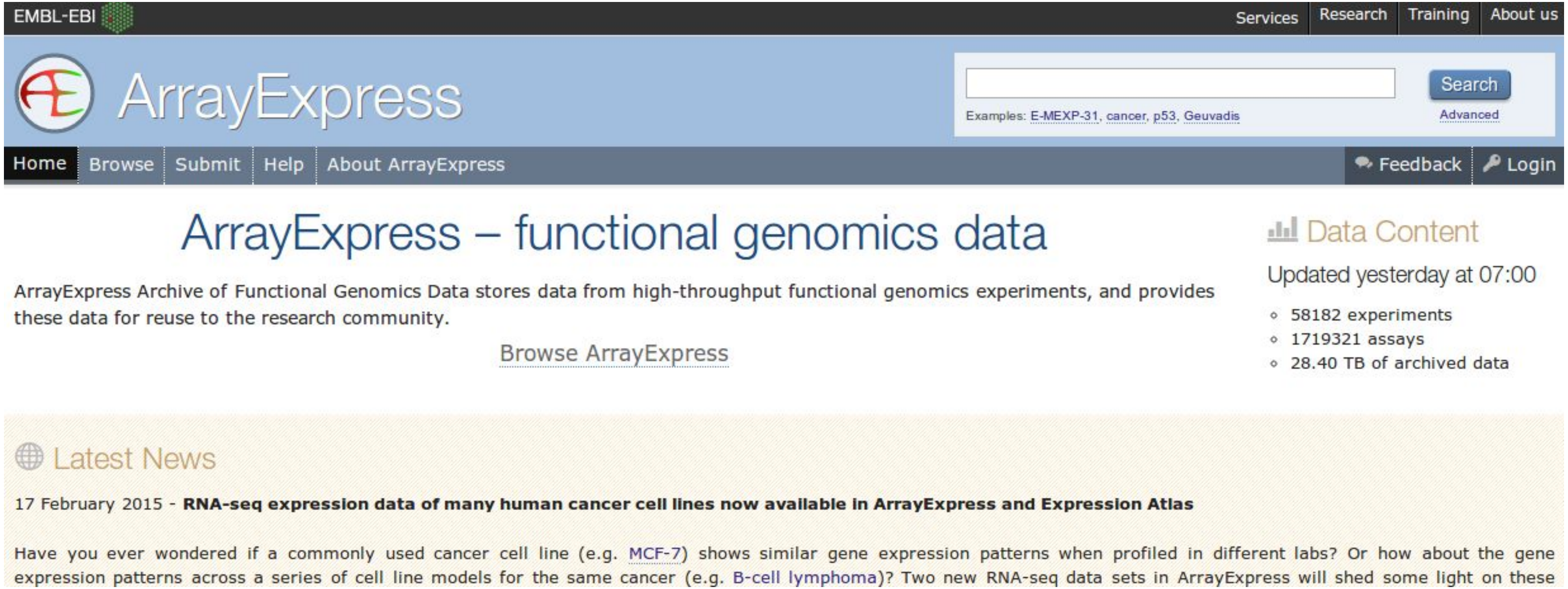
Accession	Title	Series type(s)	Organism(s)	Samples	GDS	Supplementary	Contact	Release date
GSE54218	Aberrant chromatin acetylation in MLL-AF9 leukemia mediates the response to HDAC inhibition (microarray)	Expression profiling by array	*Homo sapiens*	16		TXT	Sara Alvarez	Mar 01, 2015
GSE46670	Regulation 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

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)



ArrayExpress – functional genomics data

ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.

Data Content

Updated yesterday at 07:00

- 58182 experiments
- 1719321 assays
- 28.40 TB of archived data

Latest News

17 February 2015 - **RNA-seq expression data of many human cancer cell lines now available in ArrayExpress and Expression Atlas**

Have you ever wondered if a commonly used cancer cell line (e.g. MCF-7) shows similar gene expression patterns when profiled in different labs? Or how about the gene expression patterns across a series of cell line models for the same cancer (e.g. B-cell lymphoma)? Two new RNA-seq data sets in ArrayExpress will shed some light on these

All ArrayExpress submissions follow the MIAME checklist

GEO vs. ArrayExpress

- Both encompass MIAME compliance

- Both provide a good possibility for making data publicly available 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.