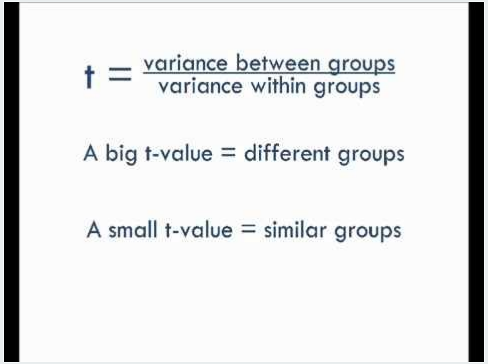
Biostatistics

Grundlagen der Bioinformatik SS2019



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

Agenda

- Normalization
- Differential expression
    - Fold Change
    - P-value
    - t-test
- Clustering

Experimental Design

$N_{1},...,N_{m}$: **control** samples

$T_{1},...,T_{n}$: **case** samples

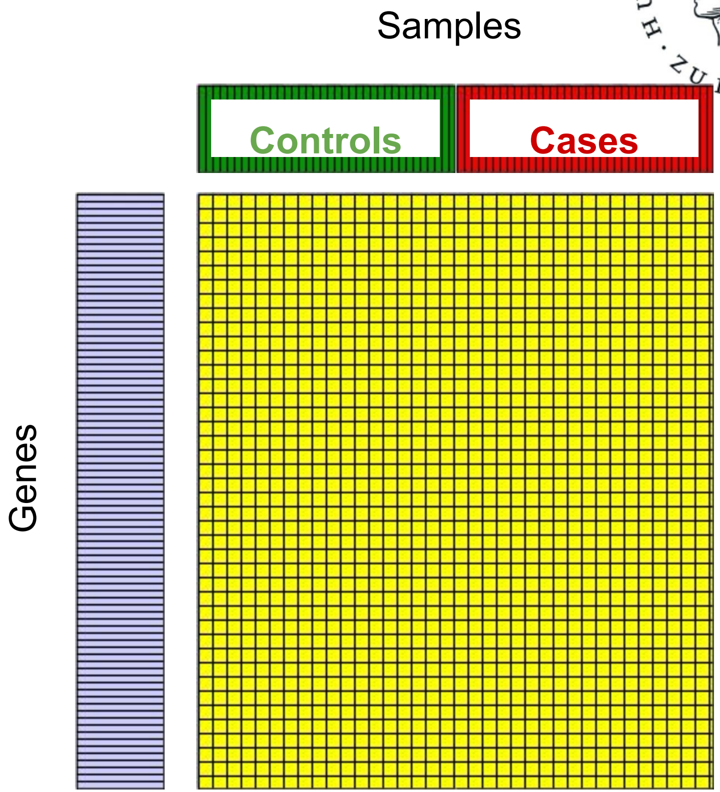
We **look for**:

Genes with significant differences between N and T

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

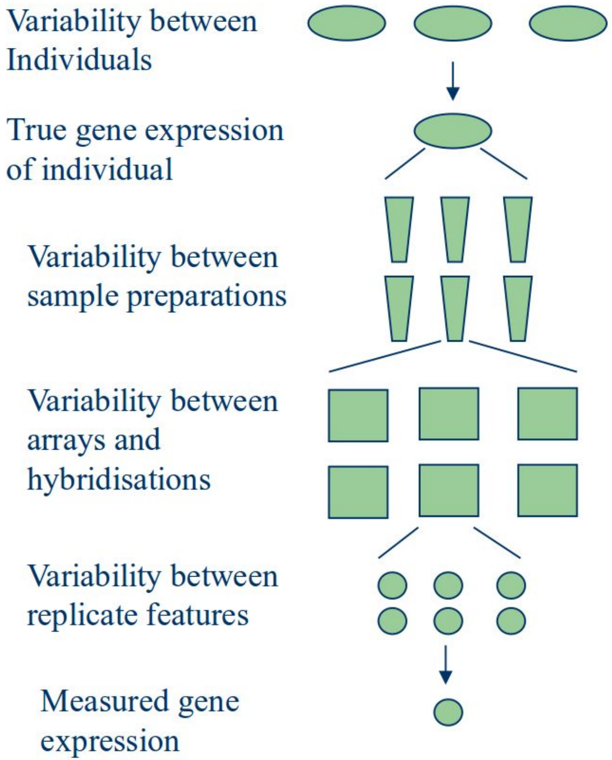
$N = \{n_{1},...,n_{m}\}$ $T = \{t_{1},...,t_{n}\}$

Many methods exist, here: Fold change t-test



Motivation normalization

- Interested in: true biological difference of mRNA expression
- What we measure: Mixture of (unwanted) technical and biological noise
- Correct undesired noise!



Quantile normalization

✓ Differences between the separate values retained

✓ Identical distribution for each array

✘ Information lost

- Especially in the lower signals

1. Matrix X
   a. Columns = samples
   b. Row = transcripts
2. Sort each column of X -> X_{sort}
3. Calculate row-means and store in X'_{sort}
4. Obtain X_n by rearranging columns of X'_{sort} to have the same ordering as the corresponding input vector

What we want: comparability

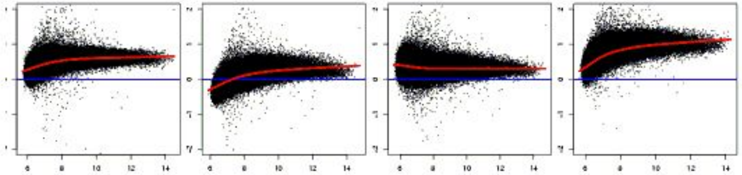


Figure 7A. Ratio Intensity Plot of all probes for four pairs of chips from GeneLogic spike-in experiment

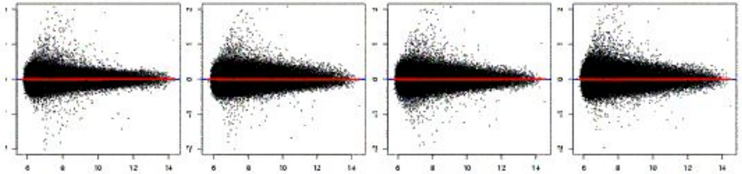
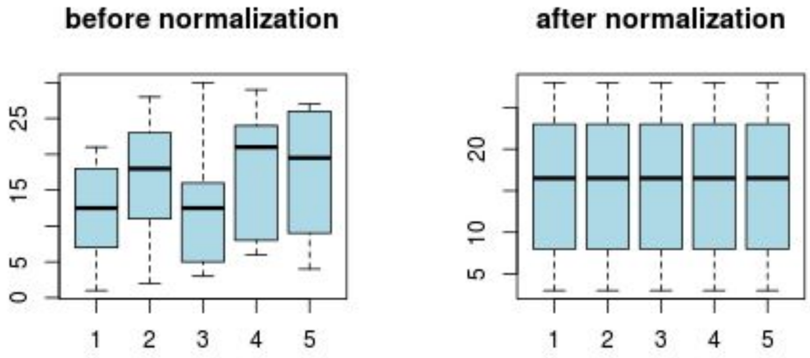


Figure 7B. As in A, after normalization by matching quantiles. Both figures courtesy of Terry Speed



Bolstad, Benjamin M., et al. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics 19.2 (2003): 185-193.

Important: normalization between samples, not within one sample

Example quantile normalization

Values

	E1	E2	E3	E4	E5
V1	1	11	13	29	26
V2	15	17	5	8	14
V3	21	2	12	20	25
	10	19	16	24	4
V4	18	28	3	22	27
V5	7	23	30	6	9

Indexes

1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6

Sort

E1	E2	E3	E4	E5
21	28	30	29	27
18	23	16	24	26
15	19	13	22	25
10	17	12	20	14
7	11	5	8	9
1	2	3	6	4

3	5	6	1	5
5	6	4	4	1
2	4	1	5	3
4	2	3	3	2
6	1	2	2	6
1	3	5	6	4

Replace

E1	E2	E3	E4	E5
28	28	28	28	28
23	23	23	23	23
19	19	19	19	19
14	14	14	14	14
8	8	8	8	8
3	3	3	3	3

3	5	6	1	5
5	6	4	4	1
2	4	1	5	3
4	2	3	3	2
6	1	2	2	6
1	3	5	6	4

Reorder

	E1	E2	E3	E4	E5
V1	3	8	19	28	23
V2	19	14	8	8	14
V3	28	3	14	14	19
	14	19	23	23	3
V4	23	28	3	19	28
V5	8	23	28	3	8

Differential expression

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

Z-score normalization

- Correct for different amount of mRNA per sample
- Z-score = scaling of counts
  - 0 = average
- Examples: 2, -1, 0.1

$Z = (X_i - mean_{est}) / sd_{est}$

X_i = expression gene i

$Mean_{est}$: (estimated) expr. average over all genes

Sd : (estimated) expr. standard deviation of all genes

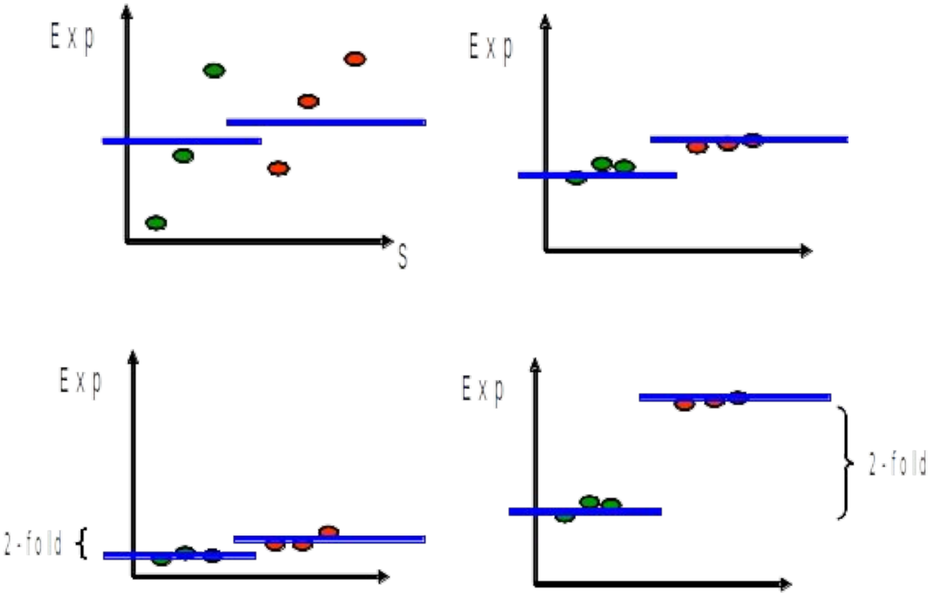
Fold Change - Advantages / Disadvantages

✓ intuitive measure

✘ Independent of scatter

✘ Independent of absolute values

- Score only based on mean of groups
- Spread of data points essential

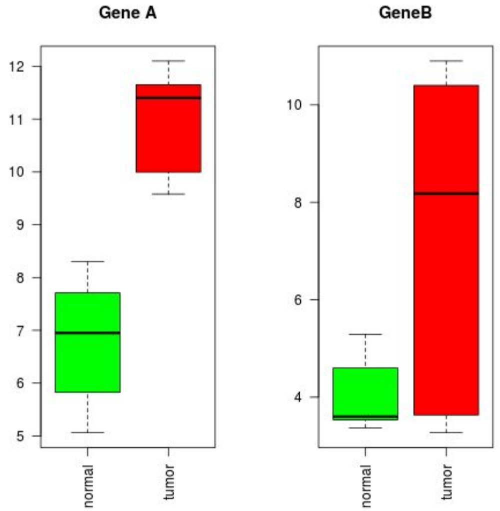


Variance essential

	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

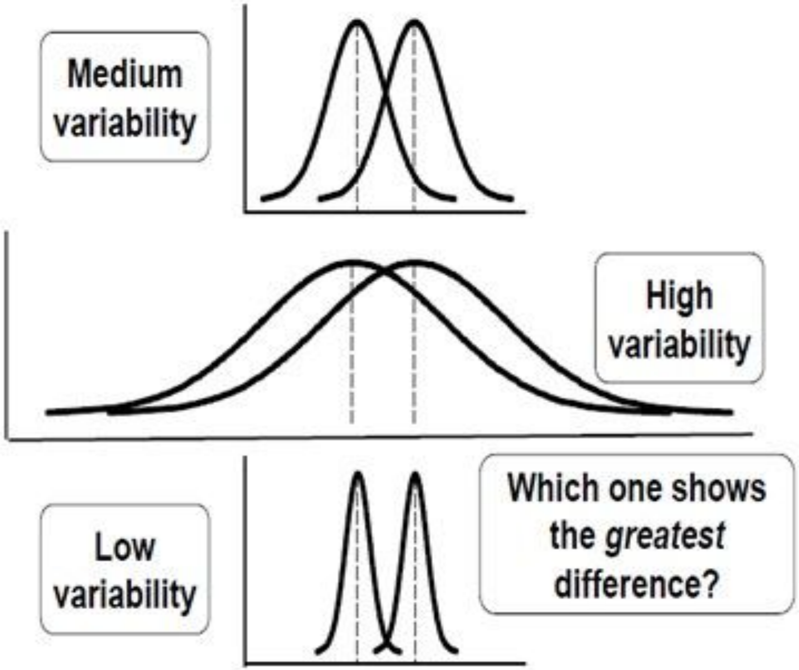
- High abs(FC) for Gene A and Gene B
- But: variance very high in the tumor samples of Gene B
- Find test for FC and variance

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



Hypothesis Testing 1

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



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

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

Hypothesis Testing 2

- **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 3

From T-statistic to p-value

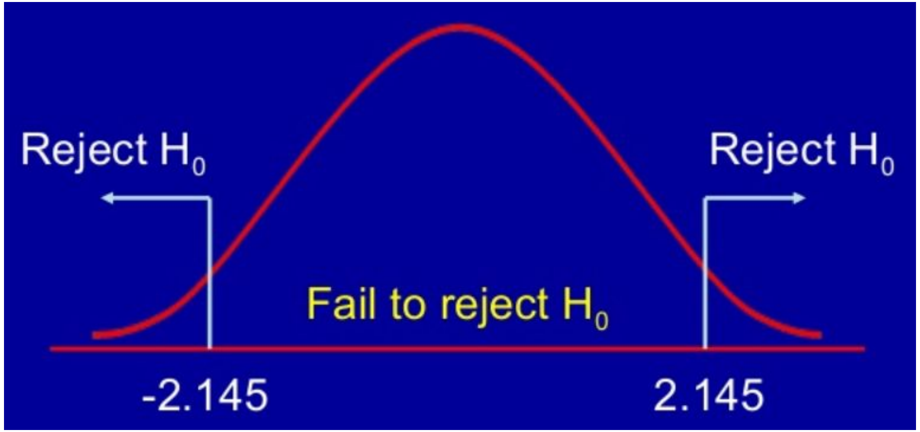
- T-value, a and number of samples determine the p-value (look-up tables)

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



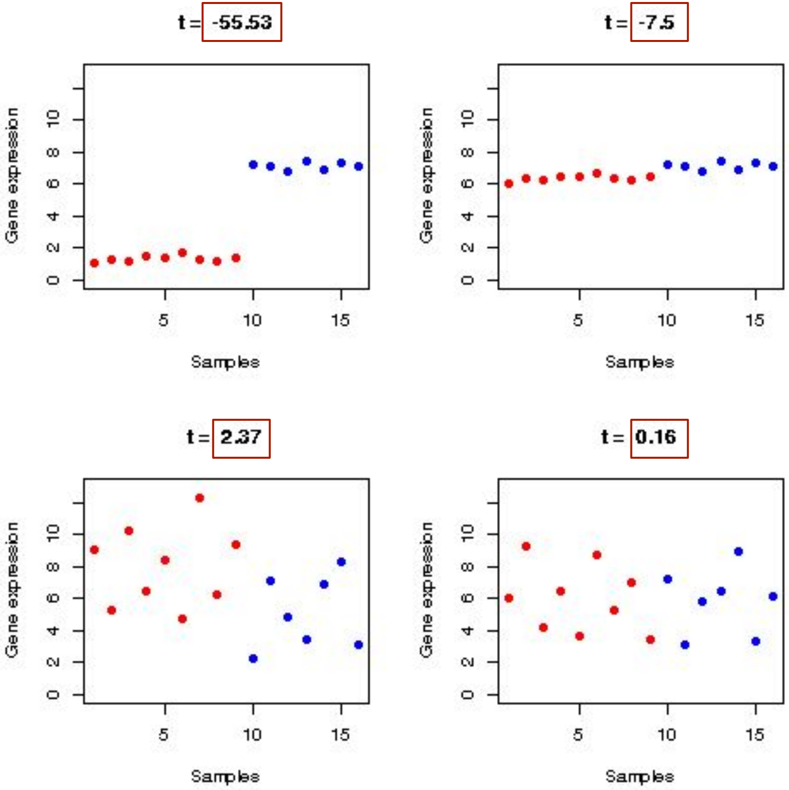
If $|t| > |T|$ we reject H_0

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

Hypothesis Testing - Workflow

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

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

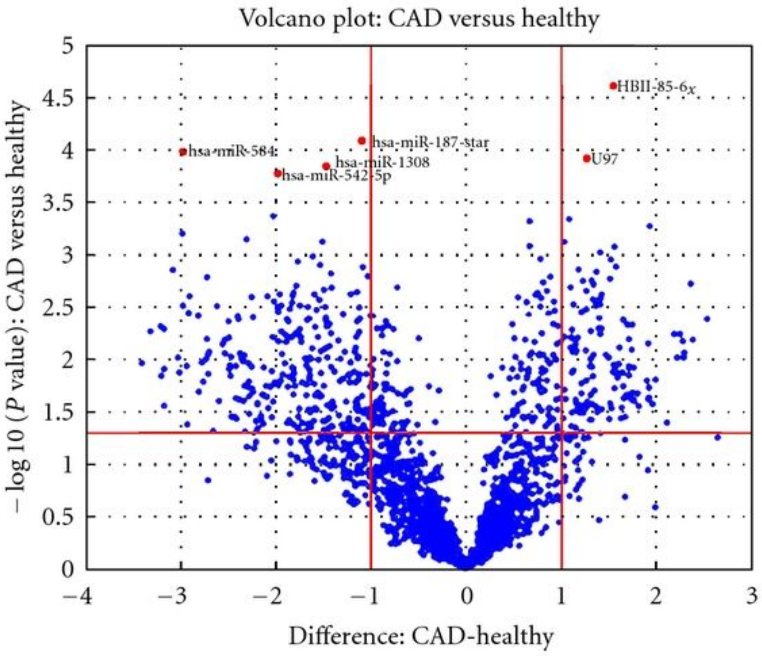
Volcano plot

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)

Bonferroni 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

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

E.g. case of two p-values (multiply by 2)

1. 0.001 -> 0.002
2. 0.03 -> 0.06

- Iff the adjusted p-value is smaller than the *alpha*, the probe is considered differentially expressed.
- Bonferroni 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

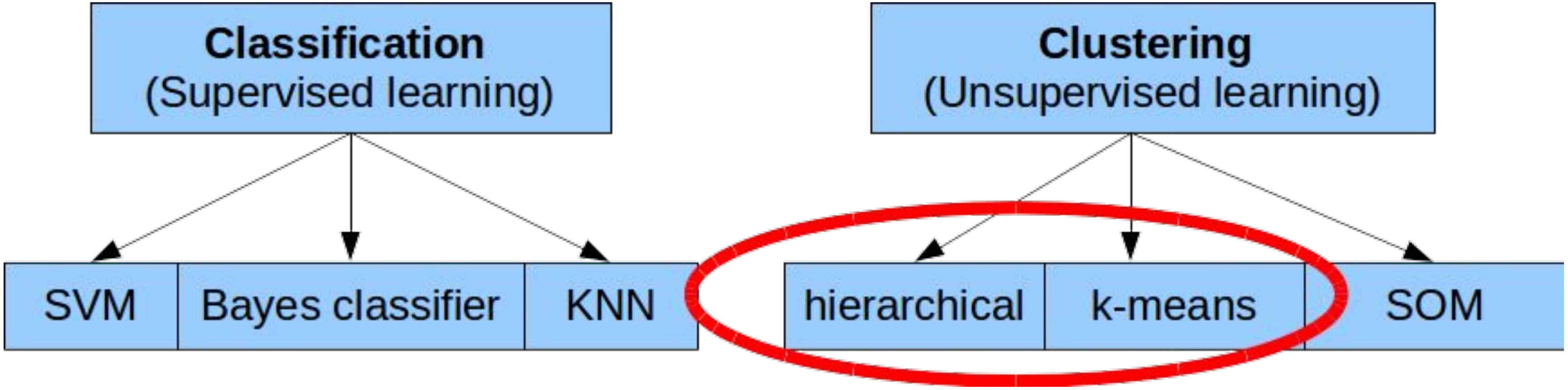
Quality control

Similarity-detection in spatial and temporal behavior

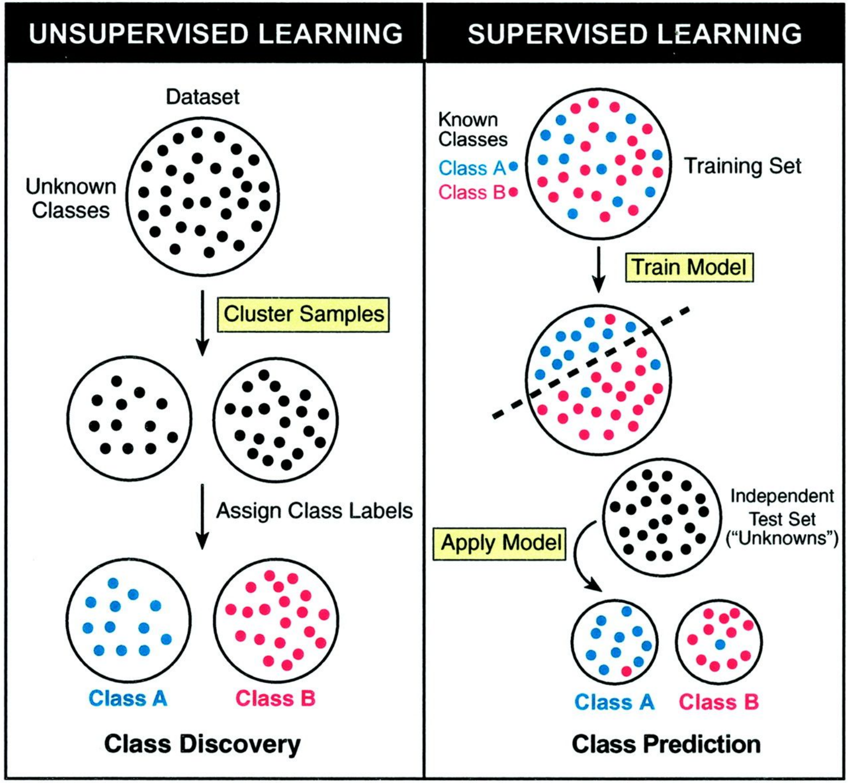
- Co-regulated / expressed genes
  - E.g. genes controlled by the same transcription-factor

Discovery of new disease subtypes

Overview unsupervised clustering

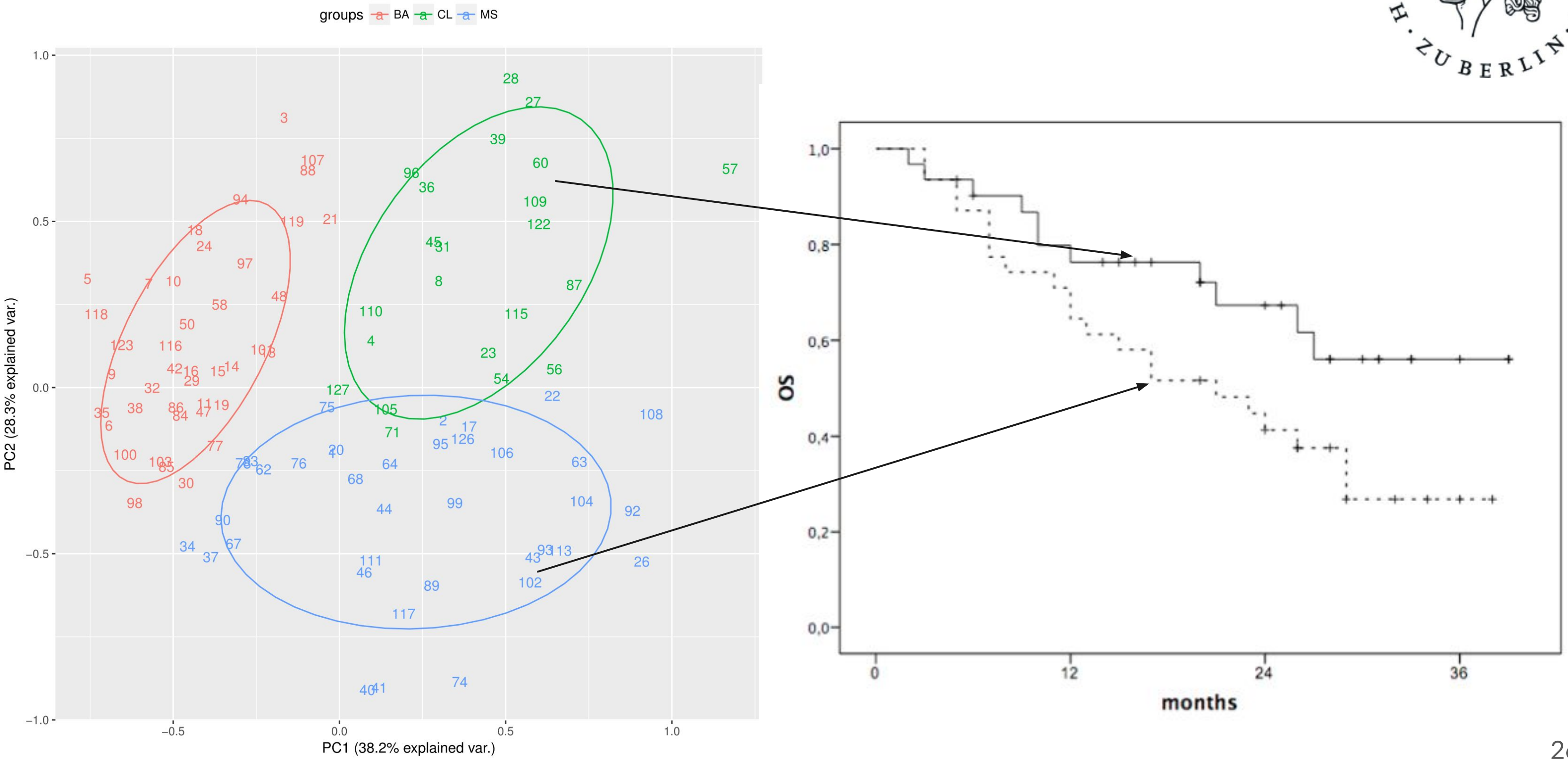


Clustering



Ramaswamy & Golub 2002

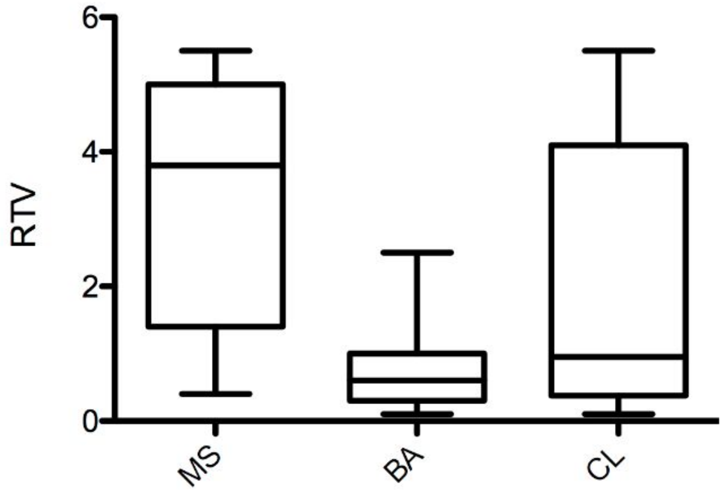
Example


groups BA CL MS
PC1 (38.2% explained var.)
PC2 (28.3% explained var.)
OS
months

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