# Biostatistics

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

A big t-value = different groups

A small t-value = similar groups

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

Grundlagen der Bioinformatik

14.06.2022

Raik Otto
raik.otto@hu-berlin.de

What Bioinformaticians think of when hearing ‘Biostatistics’

What Computer Scientists think of when hearing ‘Biostatistics’

# Agenda

- Need for biostatistics
- Normalization
- Differential expression
  - Fold Change
  - P-value
  - t-test
- Clustering

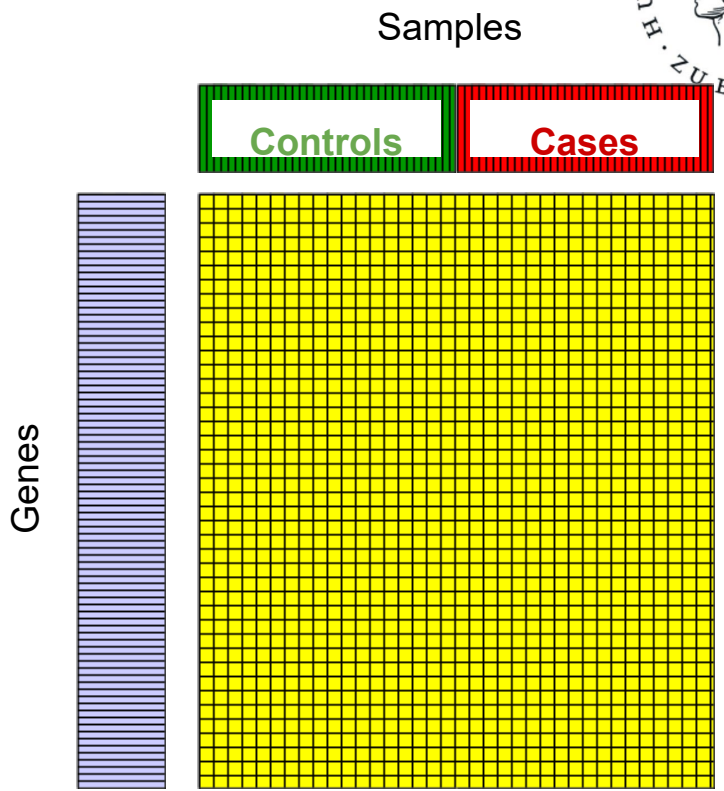
# Experimental Design

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

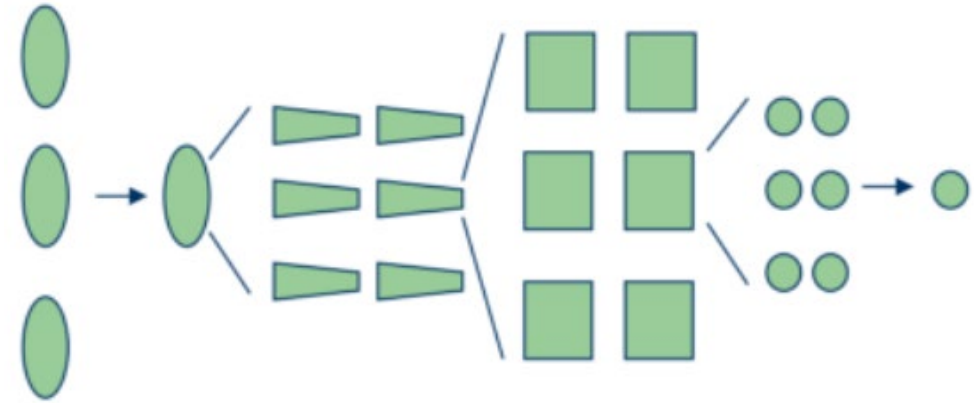
$T_1,...,T_n$: **case** samples

Identify genes with significant differences between **control** and **case**

Compare **control** gene X from group N with **case** gene X of group **control** = $\{n_1,...,n_m\}$ **case** = $\{t_1,...,t_n\}$



# Sources of biases which necessitate normalization



Human beings Experiments Tissues Similar genes

Sources of technical variance

Separate technical from biological signal

# What we want: comparability

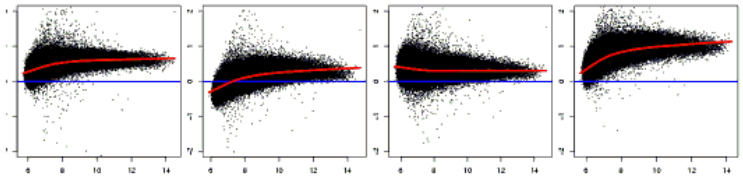


**Figure 7A. Ratio Intensity Plot of all probes for four pairs of chips from GeneLogic spike-in experiment**

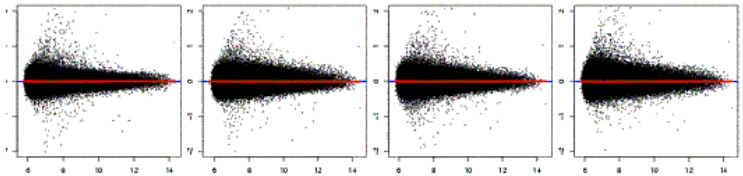
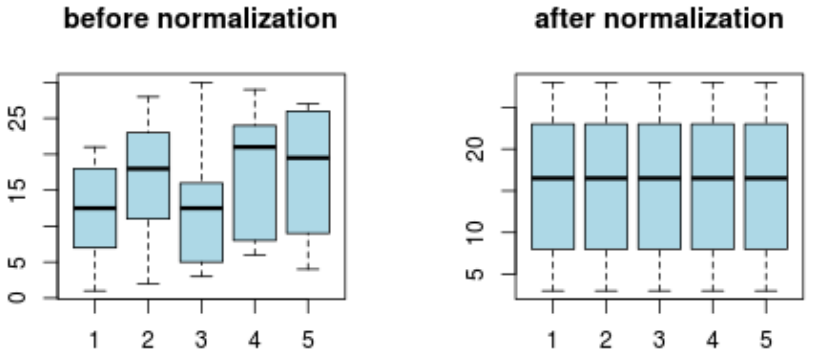


**Figure 7B. As in A, after normalization by matching quantiles. Both figures courtesy of Terry Speed**




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.

# Quantile normalization

✓ Biological differences retained

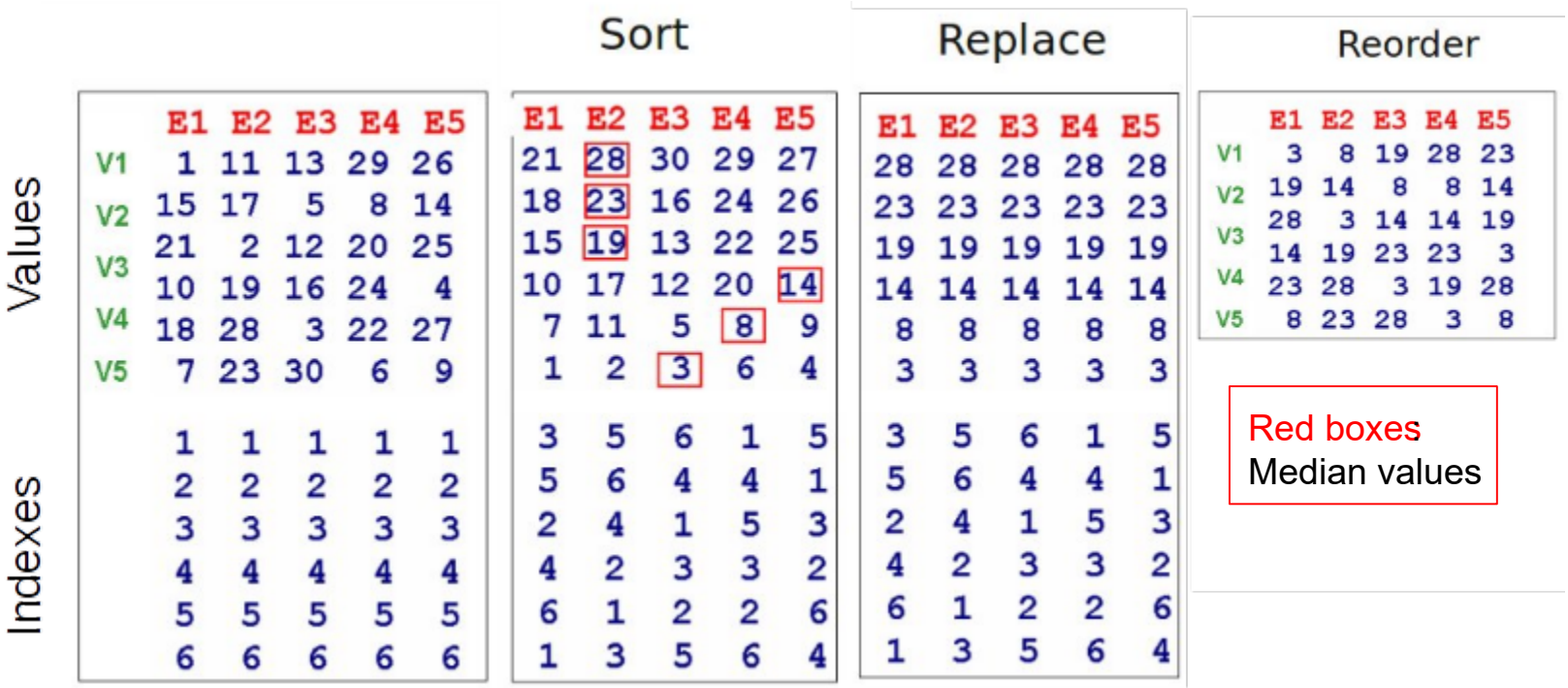
X Biological differences reduced

X: Expression matrix

C: columns $\{c_1, .., c_m\}$

1. **Sort columns: $X_c$ -> $X_{c\ sorted}$**
2. **Replace all rows in $X_{c\ sorted}$ with row medians**
3. **Unsort replaced $X_c$ -> $X_{c\ sorted}$**

# Example quantile normalization



# Differential expression

# 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

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

**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 |

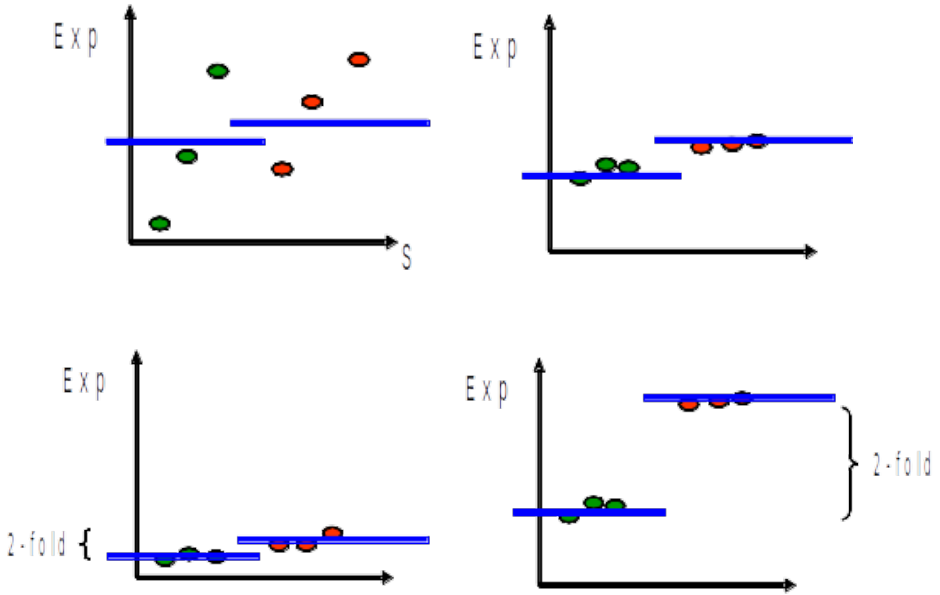
# Fold Change - Advantages / Disadvantages

✓ Intuitive

✗ Independent of scatter

✗ Independent of absolute values

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

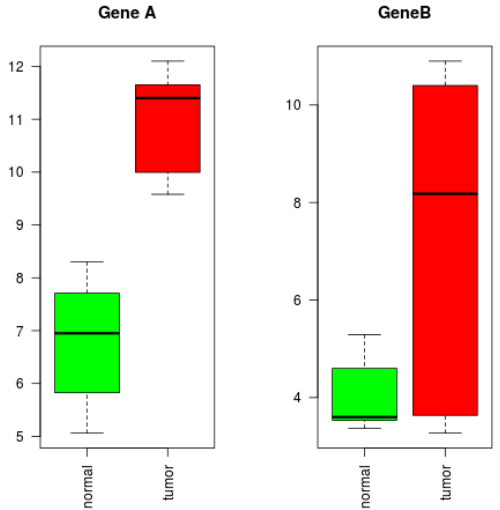


# Impact of variance

| | 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

1. Determine null H0 and alternative hypothesis H1
    - H0 is opposite of what you want to show (H1)
2. Select a significance level (alpha) -< accepted risk
3. Take a random sample from the population of interest
4. Calculate a test statistic
5. Decision

# Hypothesis Testing 2 - tests on risk

## T-test (unpaired two -sample)

- Compares the mean of two unpaired samples

## Assumption

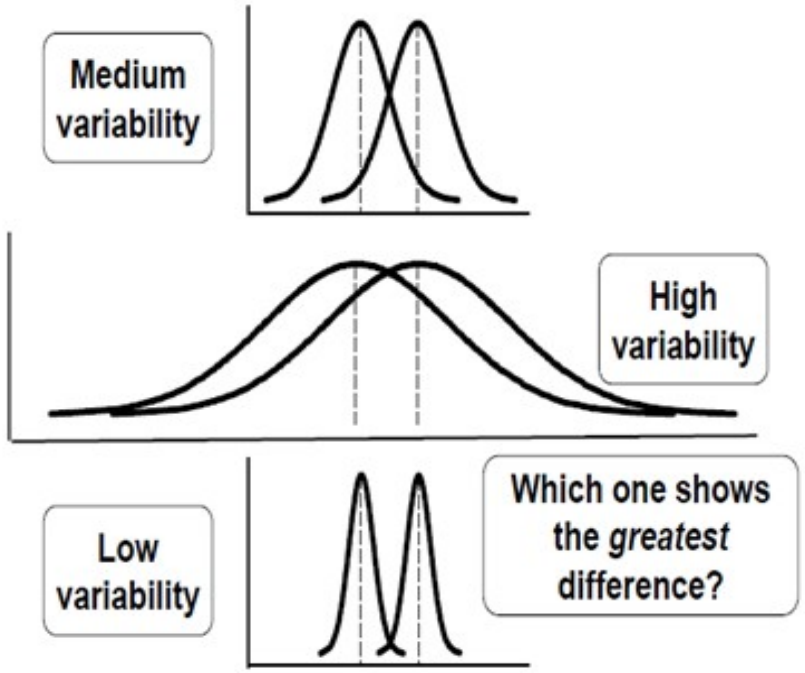
- Values normally distributed
- Equal variances

## 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




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

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

# Hypothesis Testing 3

**From T-statistic to p -value**

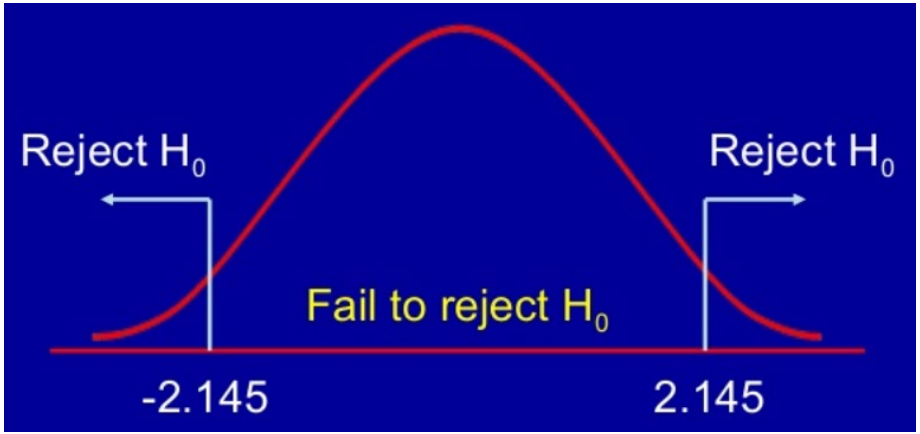
- T-value, a and number of samples determine the p-value (look-up tables)

**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)**

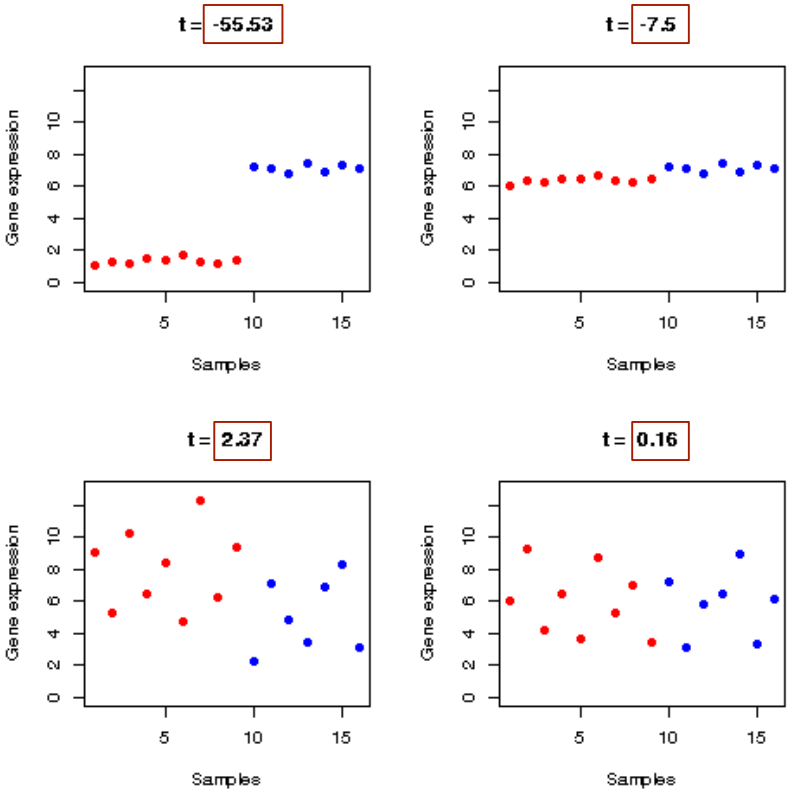
- Probability of a false positive outcome of the test, the error of rejecting $H_0$ when it is actually true




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

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

# 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 t -value

Null-Hypothesis $H_0 : m_N - m_T = 0$ vs $H_1 :$

$m_N - m_T \neq 0$

Significance level (alpha or evalue in slide before): 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

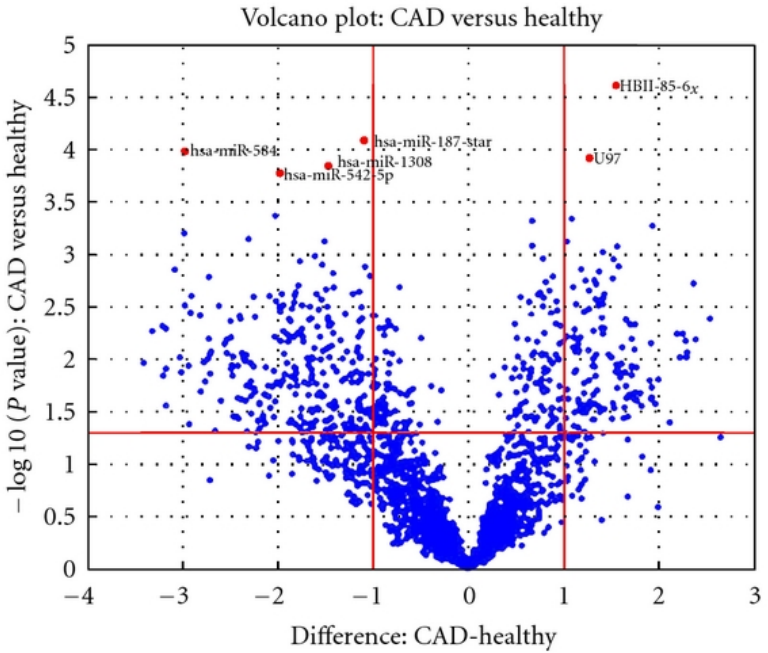
# Volcano plot

Combine P-value and Log-FC

- Y-axis: Negative log10 of the pvalue
- 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 $\alpha$=0.05 leads to approximately $22\,000 * 0.05 \sim 1100$ FPs.

**Solution:** Multiple testing correction

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

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

# Bonferroni correction

*N* genes
*p* p-value

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

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

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

- Iff $p_{adjusted}$ smaller than the *alpha* test significant
- Independence between the tests assumed
  - Possibly not the case
- 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 specifiα (e.g. α=0.05)
1. Rank all m p-values from smallest to largest
1. Correct all p-values: BH(pi)i=1,...,m = $p_i$ * m/i
1. 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

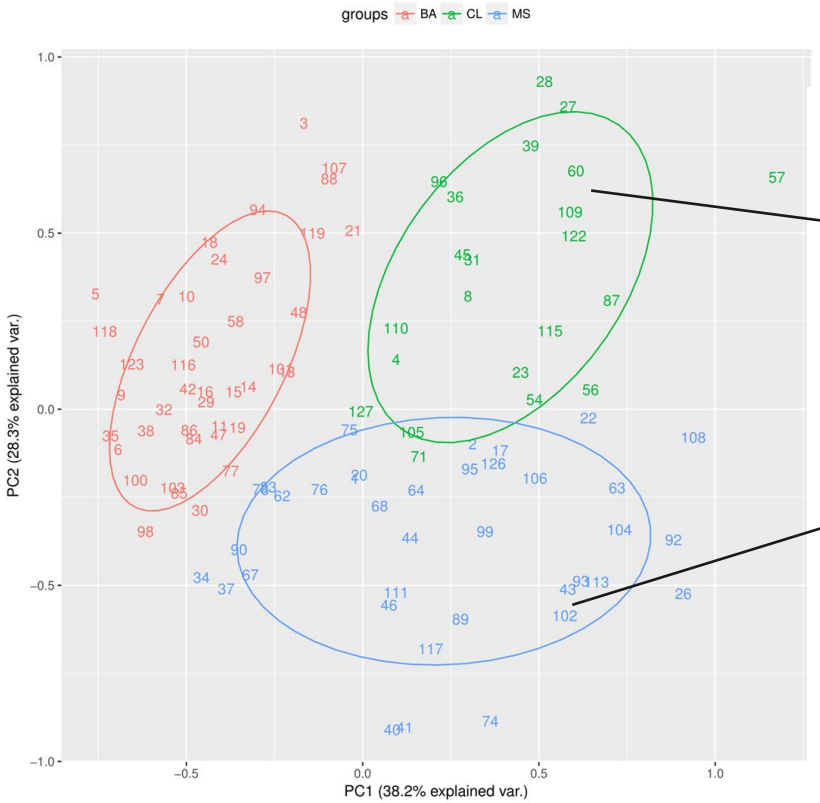
Quality control

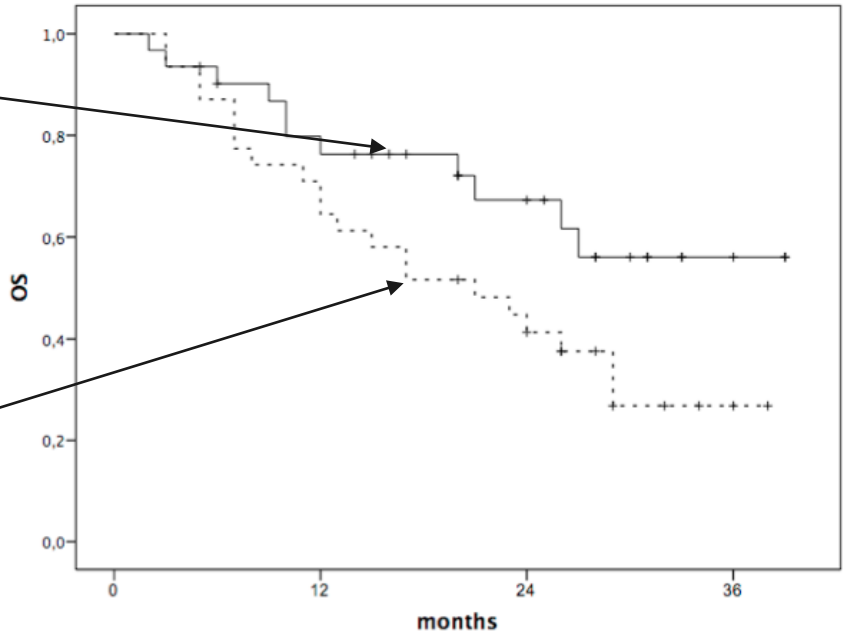
Similarity -detection in spatial and temporal behavior

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

Discovery of new disease subtypes

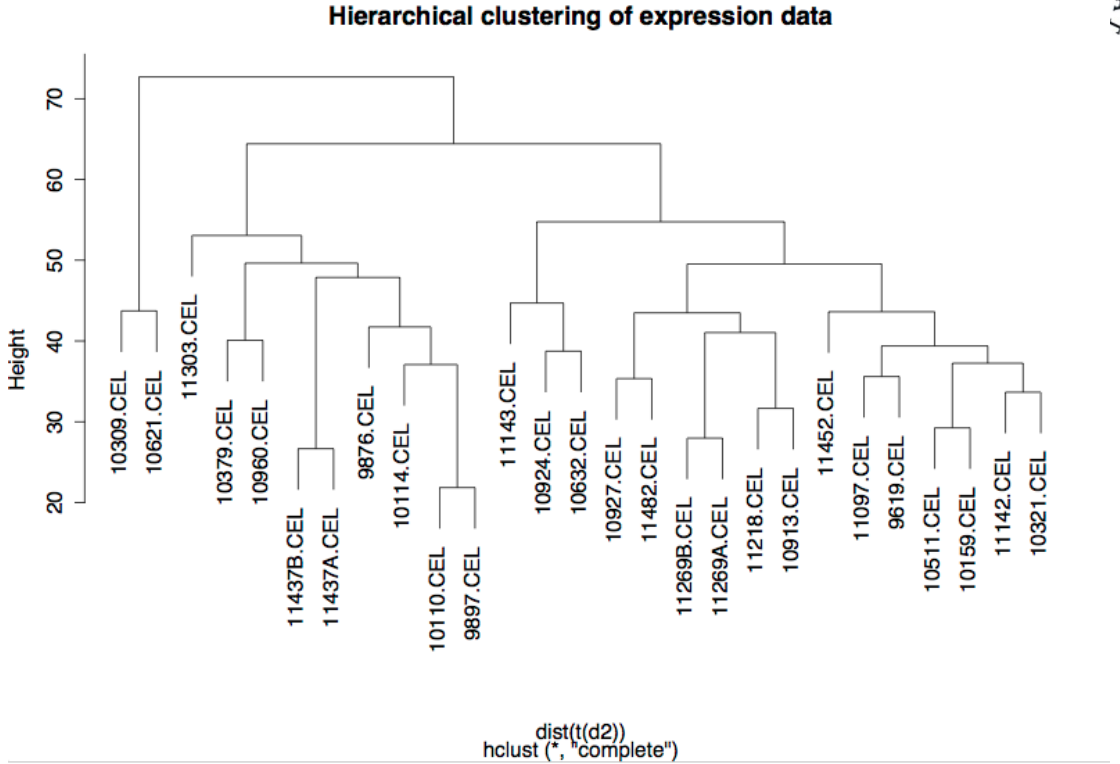
# Example


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


OS
months

# Clustering

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



# Hierarchical Clustering - algorithm

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

1. Compute similarity matrix S

1. 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 – 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 $D(X,Y) = \frac{1}{N_X N_Y} \sum_{x \in X} \sum_{y \in Y} d_{xy}$