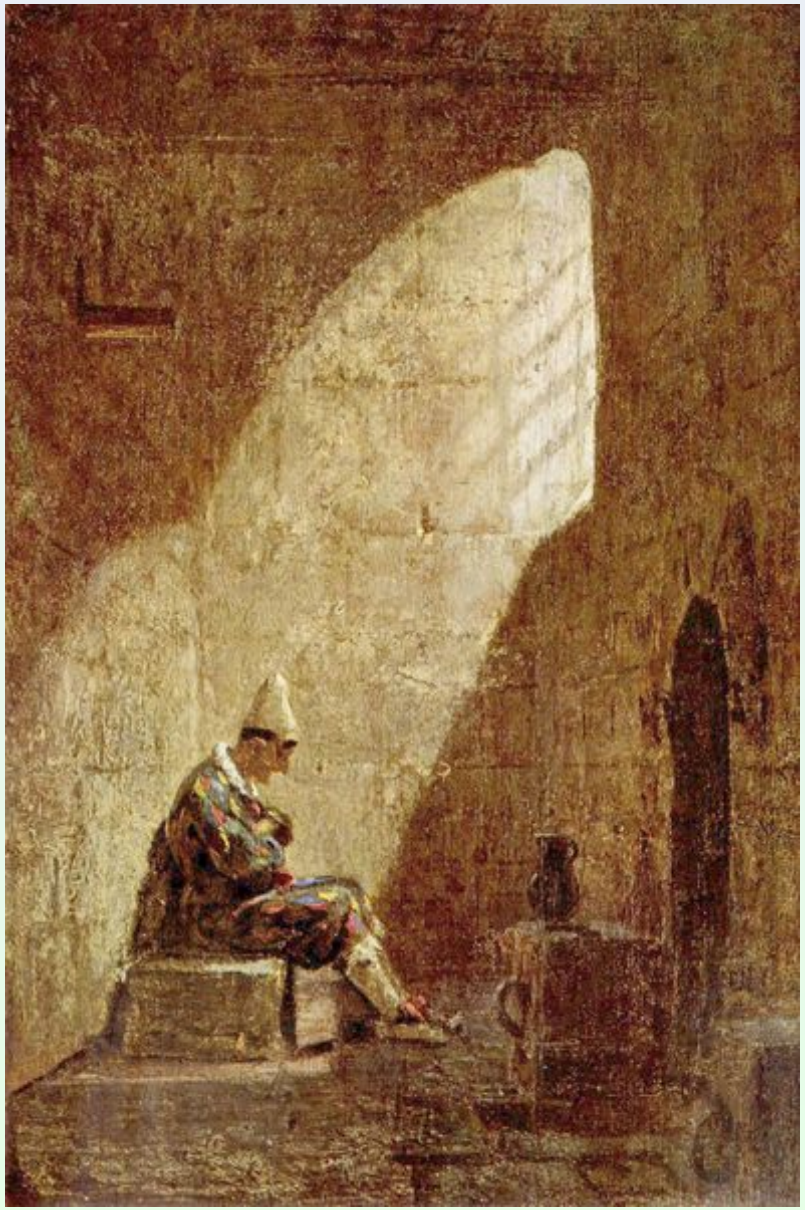
Microarray analysis

Anne Kupczok

Center for Integrative Bioinformatics Vienna
Max F. Perutz Laboratories

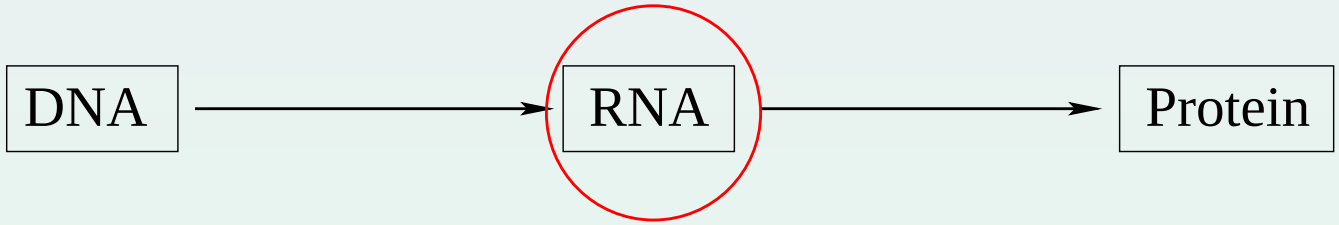
February 6th, 2008



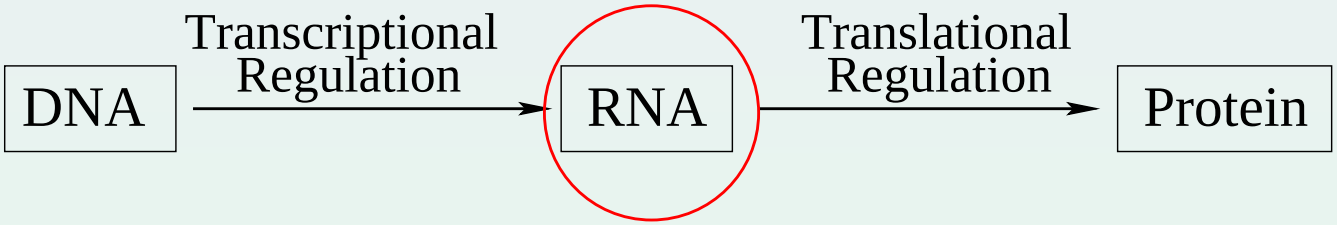
Contents

The central dogma of molecular biology

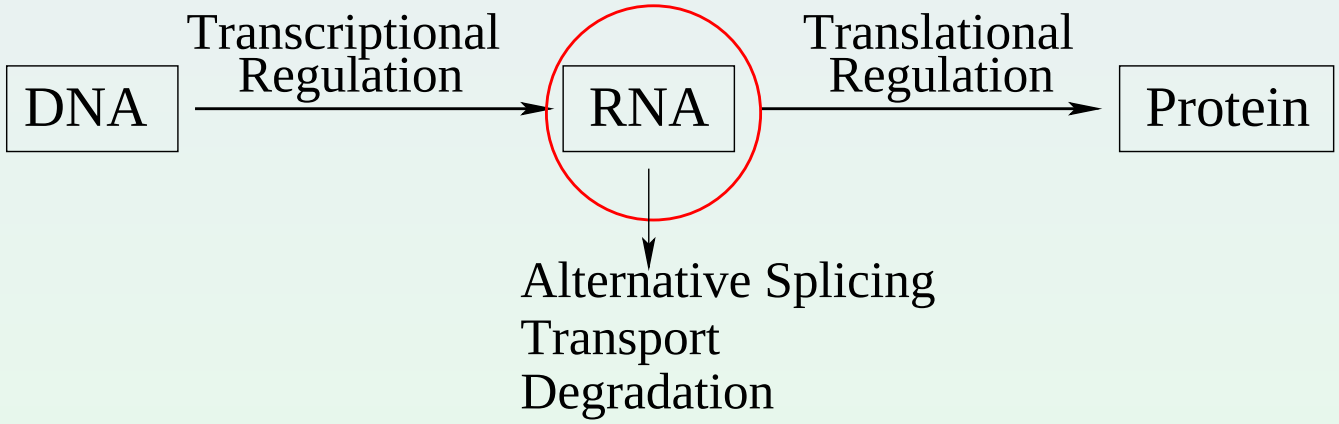
The central dogma of molecular biology



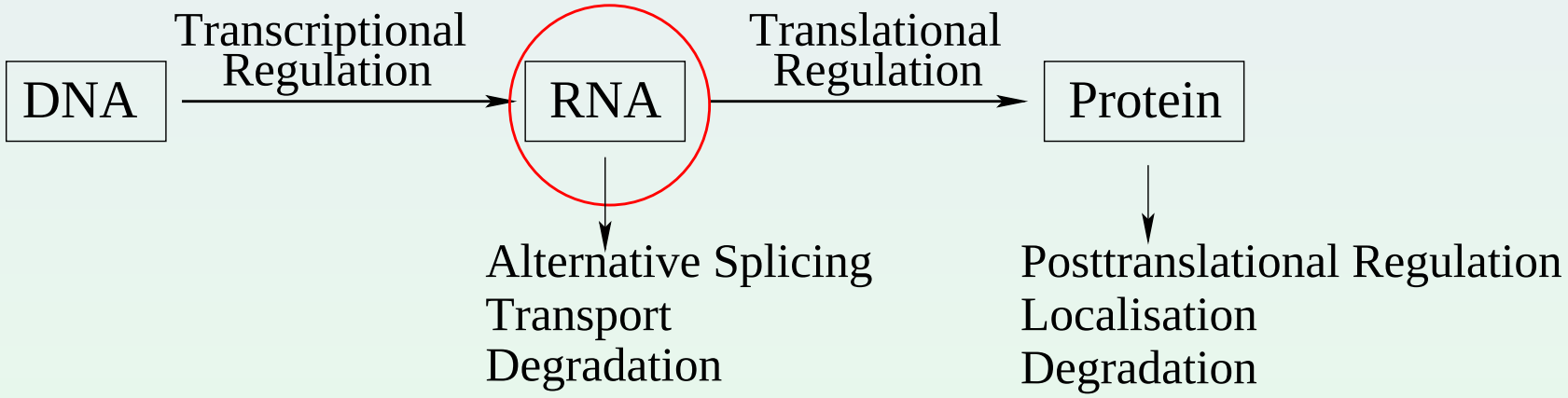
The central dogma of molecular biology



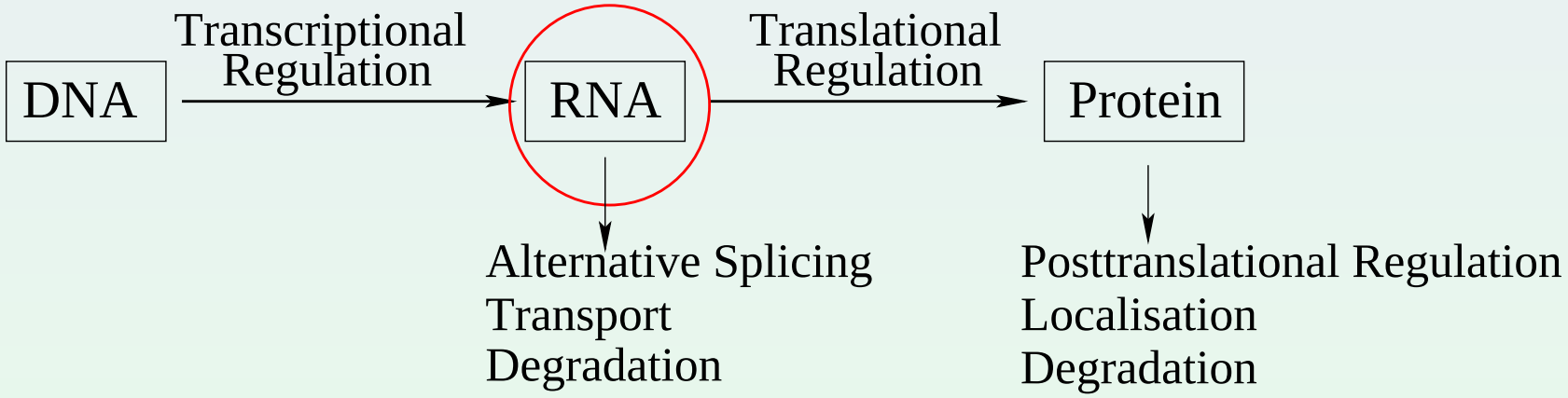
The central dogma of molecular biology



The central dogma of molecular biology



The central dogma of molecular biology



Microarrays analyse the gene expression by measuring the amount of mRNA in the cell at a special point in time.

Why expression analysis?

- Gene expression information is not available from the sequence alone
- Reaction of cells or organisms to different treatments
- Understand the difference between different entities (mutants)
- Gene expression change during development
- Gene regulation networks

Why expression analysis?

- Gene expression information is not available from the sequence alone
- Reaction of cells or organisms to different treatments
- Understand the difference between different entities (mutants)
- Gene expression change during development
- Gene regulation networks

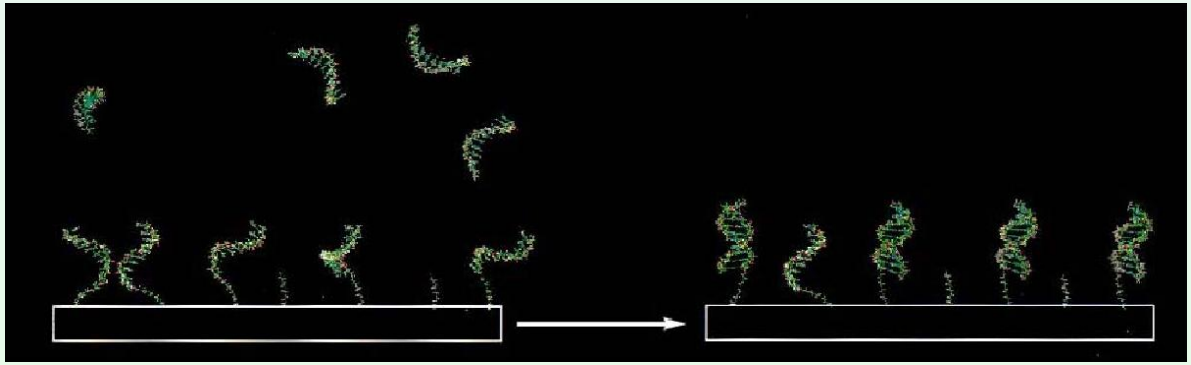
Microarrays simultaneously measure the expression of thousands of genes $\rightarrow$ global view on gene expression

Survey of one experiment

1. mRNA $\rightarrow$ cDNA (reverse transcription)
2. Amplification of the cDNA
3. Labelling of the cDNA with a dye
4. Hybridisation of the cDNA with DNAs on a slide

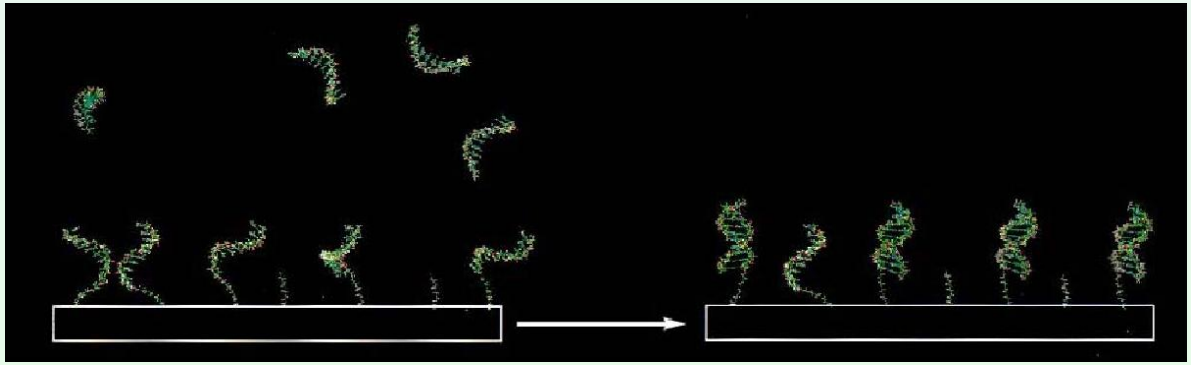
Survey of one experiment

1. mRNA → cDNA (reverse transcription)
2. Amplification of the cDNA
3. Labelling of the cDNA with a dye
4. Hybridisation of the cDNA with DNAs on a slide
   - On a slide a large number (10,000s) of DNA fragments (probes) are attached as a regular pattern



Survey of one experiment

1. mRNA → cDNA (reverse transcription)
2. Amplification of the cDNA
3. Labelling of the cDNA with a dye
4. Hybridisation of the cDNA with DNAs on a slide
   - On a slide a large number (10,000s) of DNA fragments (probes) are attached as a regular pattern



5. Hybridised DNA fragments are immobile → measure of the dye's intensities
6. Analysis of the measurement (here)

Sources of error

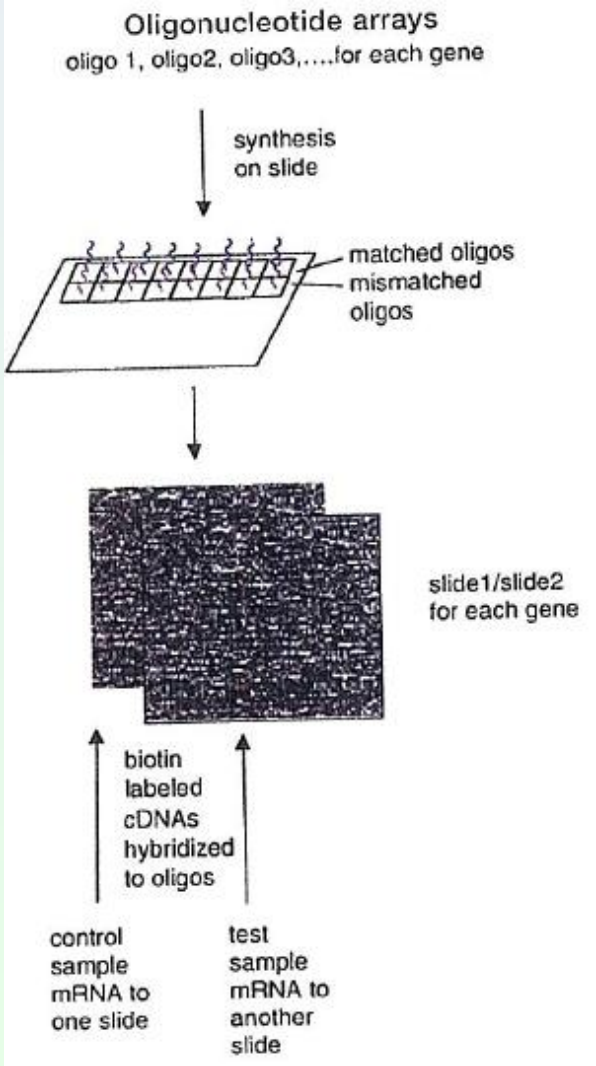
Biological noise:

- Transcription is a stochastic process
- Posttranscriptional regulation
- Stability of the mRNA

Technical noise:

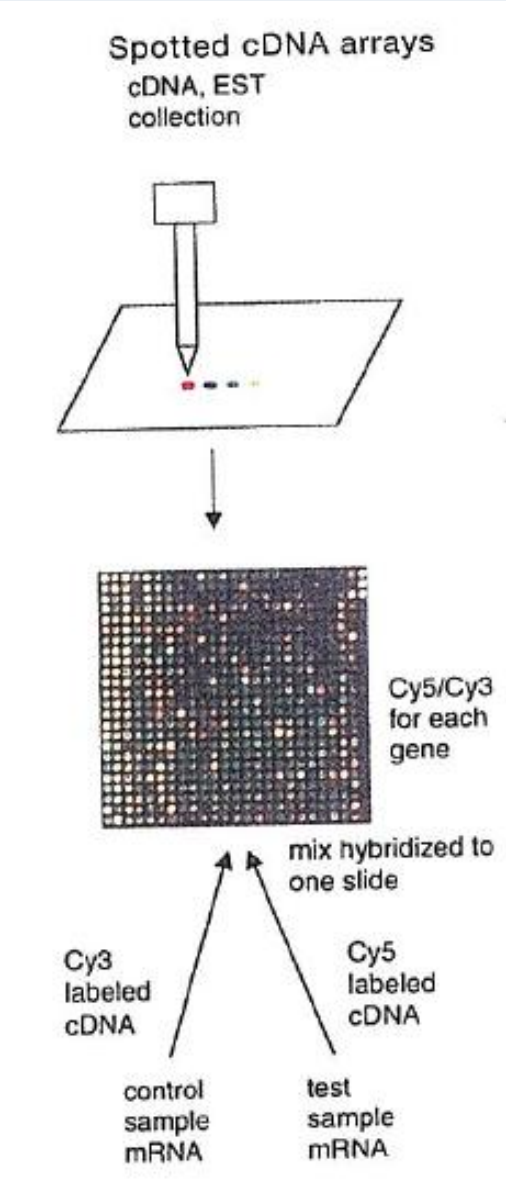
- cDNA from mRNA
- Binding of the dye
- Hybridisation
- Measurement of the signal

Oligonucleotide arrays



- Affymetrix arrays
- One biological sample per array (a new slide for every sample)
- cDNAs are labelled with biotin
- Oligonucleotides of length ≈ 25 on array
  - Perfect matching sequences
  - One or more mismatching nucleotides (control for non-specific binding)

cDNA arrays



- Two biological samples per array
- Each labelled with one of the fluorescent dyes Cy3 (green) or Cy5 (red)
- Mixture of labelled cDNAs on slide
- Intensities of the dyes measured $\rightarrow$ Ratio of the intensities provides information of the mRNA ratios in the original samples

Questions before the design

1. Scientific questions: Intention of the experiment
2. Logistic questions: Number of genes, number of measurements (probes) per gene, control genes (housekeeping genes)
3. Statistical questions: Control of the data quality, normalisation

Questions before the design

1. Scientific questions: Intention of the experiment
2. Logistic questions: Number of genes, number of measurements (probes) per gene, control genes (housekeeping genes)
3. Statistical questions: Control of the data quality, normalisation

Decisions about Blocking (distribution of the probes on the slides):

- Variables influencing the analysis on different blocks
- Reduction of block effects
- E.g. dye R or G

Replicates

Technical replicates:

- The same sample is spotted on different slides (but labelled independently)
- Measurements of errors in the procedure or in the technology

Replicates

Technical replicates:

- The same sample is spotted on different slides (but labelled independently)
- Measurements of errors in the procedure or in the technology

Biological replicates:

- Different samples spotted on different slides
- Inference of the underlying population
- Type I: different extracts of a cell line or a tissue
- Type II: the same tissue but different individuals (greater variability)

Replicates

Technical replicates:

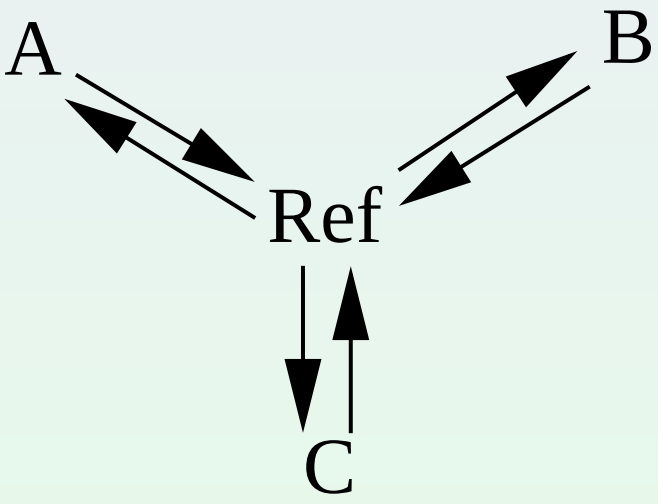
- The same sample is spotted on different slides (but labelled independently)
- Measurements of errors in the procedure or in the technology

Biological replicates:

- Different samples spotted on different slides
- Inference of the underlying population
- Type I: different extracts of a cell line or a tissue
- Type II: the same tissue but different individuals (greater variability)

The larger the number of replicates the better mean and variance can be estimated.

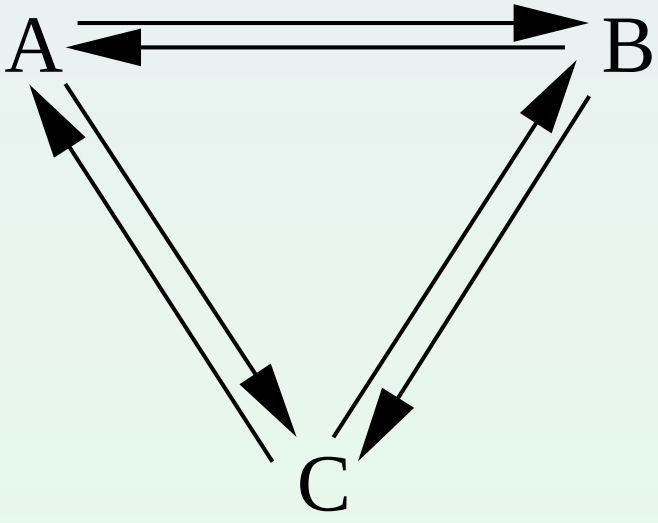
Reference design



Green $\longrightarrow$ Red

- One sample is the reference, everything else is hybridised to it
- With multiple mutants, all ratios can be computed
- A and B are compared **indirectly**
- May include a dye swap
- Advantage: Factorial experiment design, extendible
- Disadvantage: one-half of all hybridisations are the reference

Loop design



Green $\longrightarrow$ Red

- There is no reference
- On every array there is a different pair of samples (allows for biological replicates)
- With many variables, more ressources are needed compared to the reference design
- A and B are compared **directly**, i.e. on one slide
- Direct comparisons are more efficient, i.e. have a smaller variance

Fold change

The fold change (FC) is a measure for differential expression:

$$\frac{\text{Expression in sample B}}{\text{Expression in sample A}} \text{ (normally in } log_2\text{-scale)}$$

Fold change

The fold change (FC) is a measure for differential expression:

$$\frac{\text{Expression in sample B}}{\text{Expression in sample A}} \text{ (normally in } log_2\text{-scale)}$$

		log FC	variance
One cDNA array	A  B	$logB - logA$	σ^2

Fold change

The fold change (FC) is a measure for differential expression:

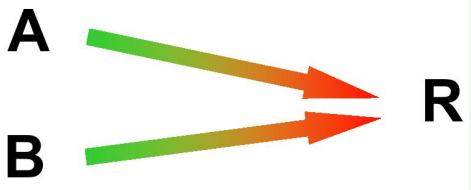
$$\frac{\text{Expression in sample B}}{\text{Expression in sample A}} \text{ (normally in } log_2\text{-scale)}$$

		log FC	variance
One cDNA array	A → B	$logB - logA$	σ^2
2 arrays, direct comparison	A ⇄ B	$[(logB - logA) - (logA - logB)]/2$	$\sigma^2/2$

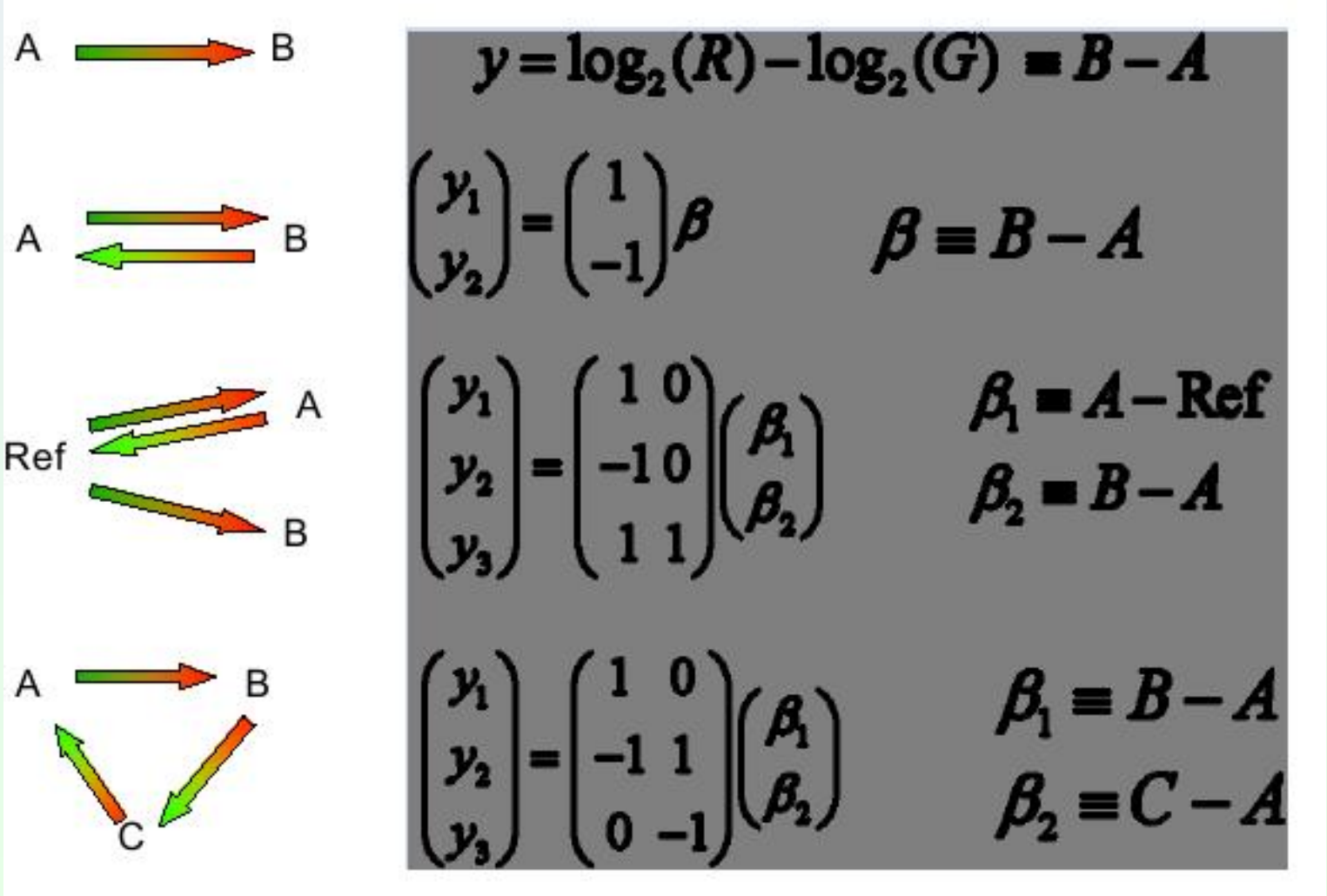
Fold change

The fold change (FC) is a measure for differential expression:

$$\frac{\text{Expression in sample B}}{\text{Expression in sample A}} \text{ (normally in } log_2\text{-scale)}$$

		log FC	variance
One cDNA array		$logB - logA$	σ^2
2 arrays, direct comparison		$[(logB - logA) - (logA - logB)]/2$	$\sigma^2/2$
2 arrays, indirect comparison		$(logR - logA) - (logR - logB)$	$2\sigma^2$

Simultanous computation of the expression values for a complex design



Analysis of microarrays

1. Image analysis
2. Normalisation (each slide separatly)
3. Differential gene expression (all slides, whole experiment)
4. Analysis of gene expression

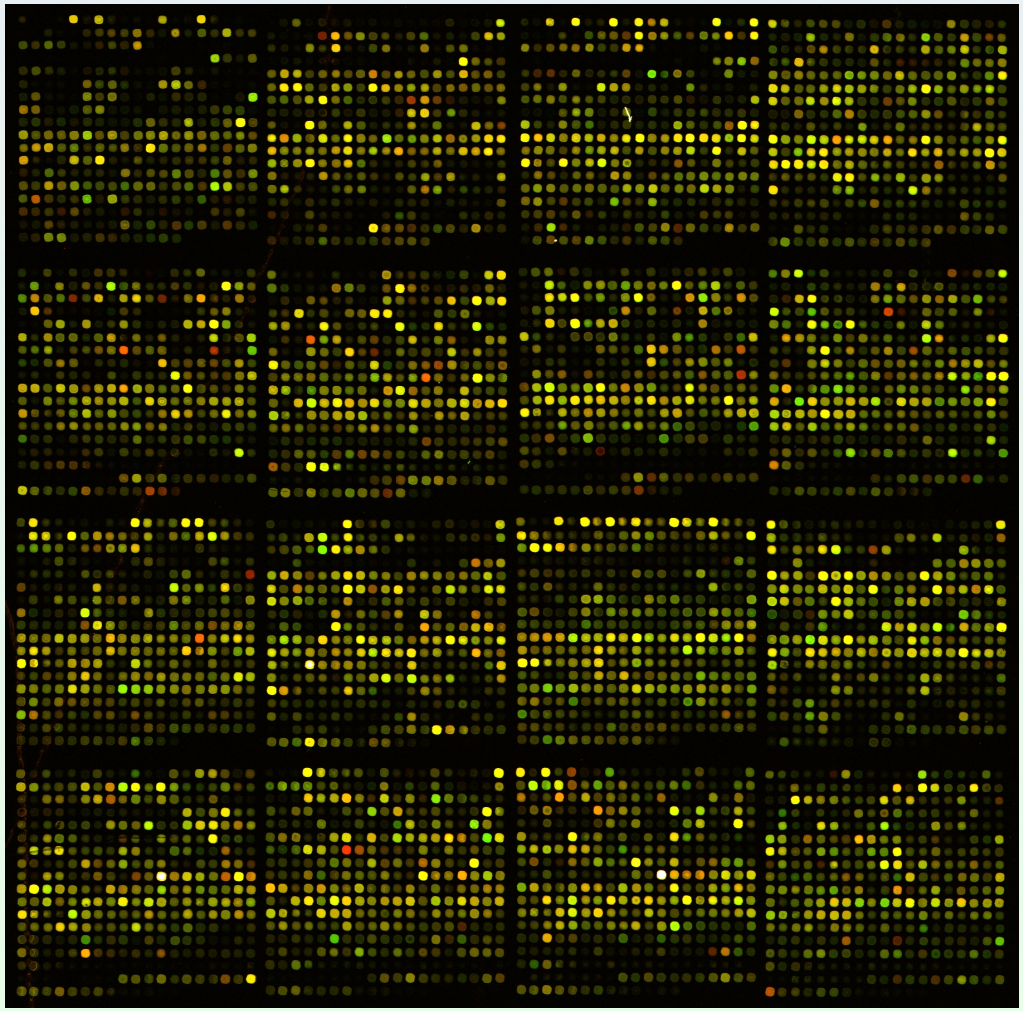
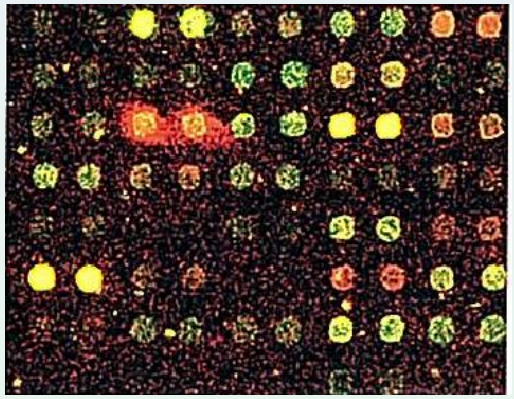
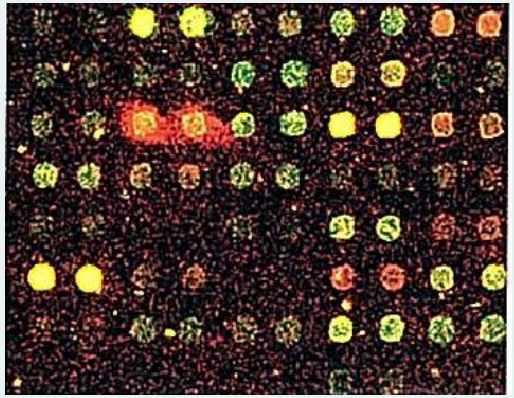


Image analysis



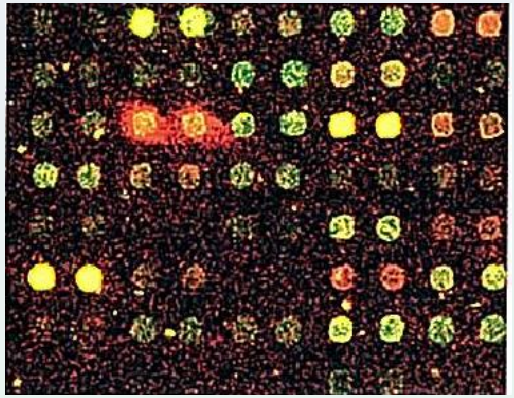
1. Localisation of the spots
2. Segmentation: Determination of the spot borders, partition in foreground and background
3. Computation of the intensities (next slide)
4. Filtering of low-quality spots

Background normalisation



- Background signal (noise) varies across slide
- For every spot: means over intensities in the neighborhood are substracted from foreground intensities

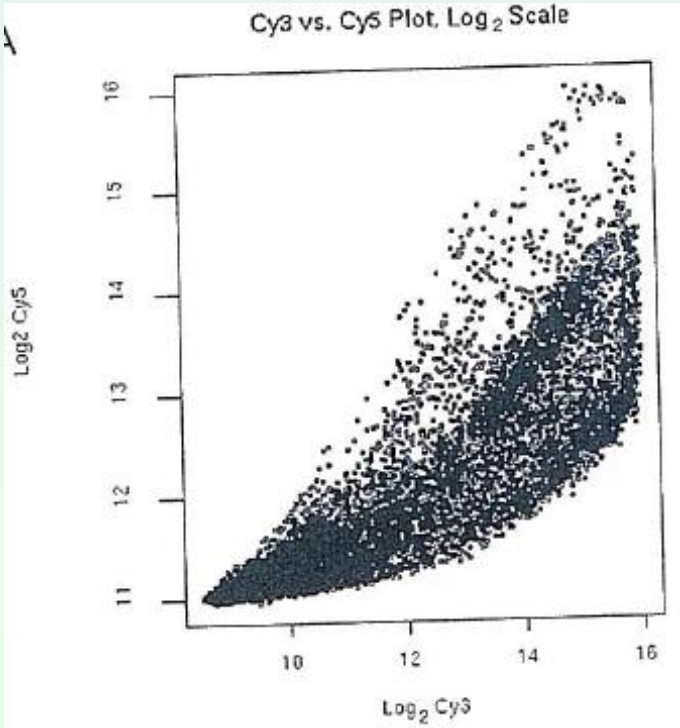
Background normalisation



- Background signal (noise) varies across slide
- For every spot: means over intensities in the neighborhood are substracted from foreground intensities
- Background values can be greater then corresponding foreground values
  1. Removing of genes with negative intensities
  2. Replacement by the minimal value of the array (problem: decreases the variance)
  3. Statistical approach based on the assumption that foreground is always larger than background

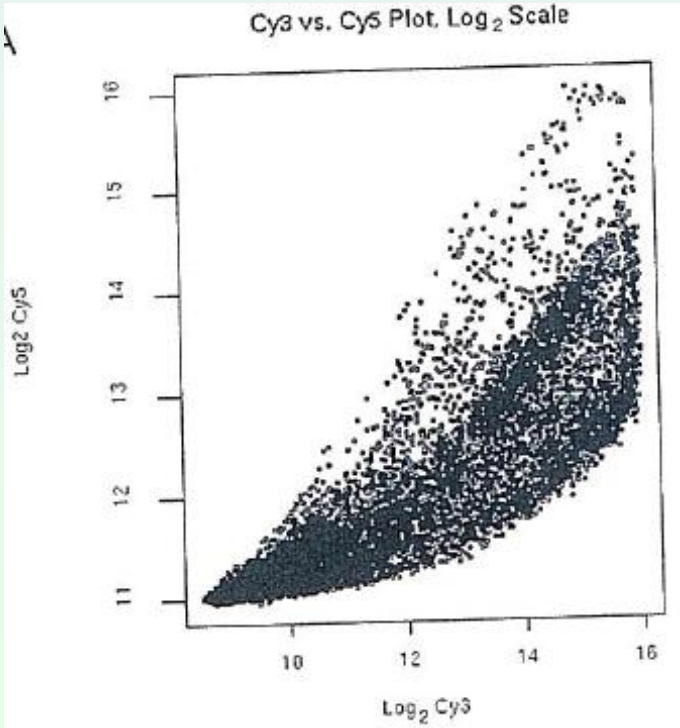
M/A plot

Assumption: Only a small part of the genes are differentially expressed, then the plot of R against G should be a line

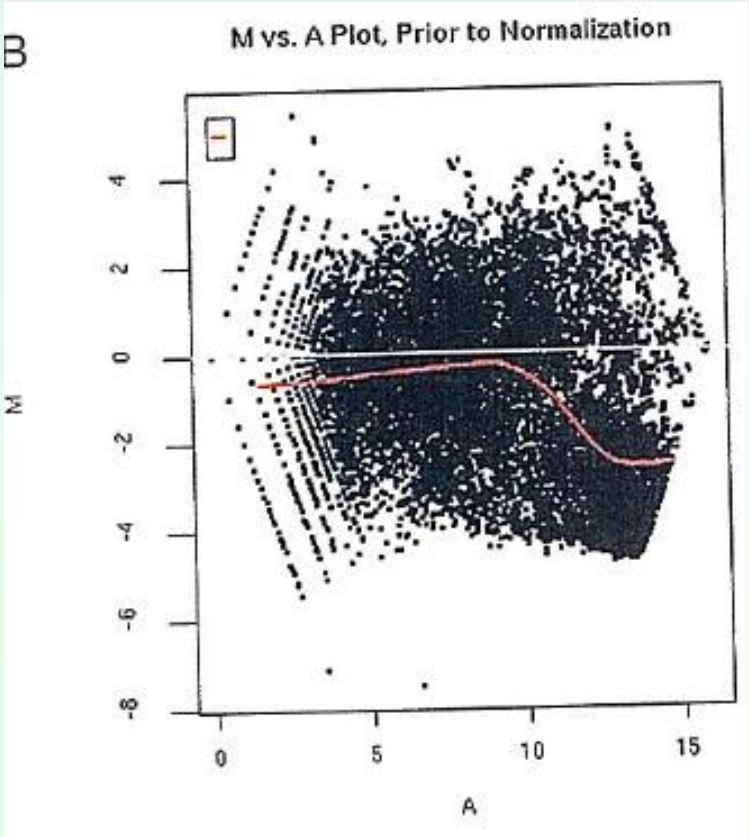


M/A plot

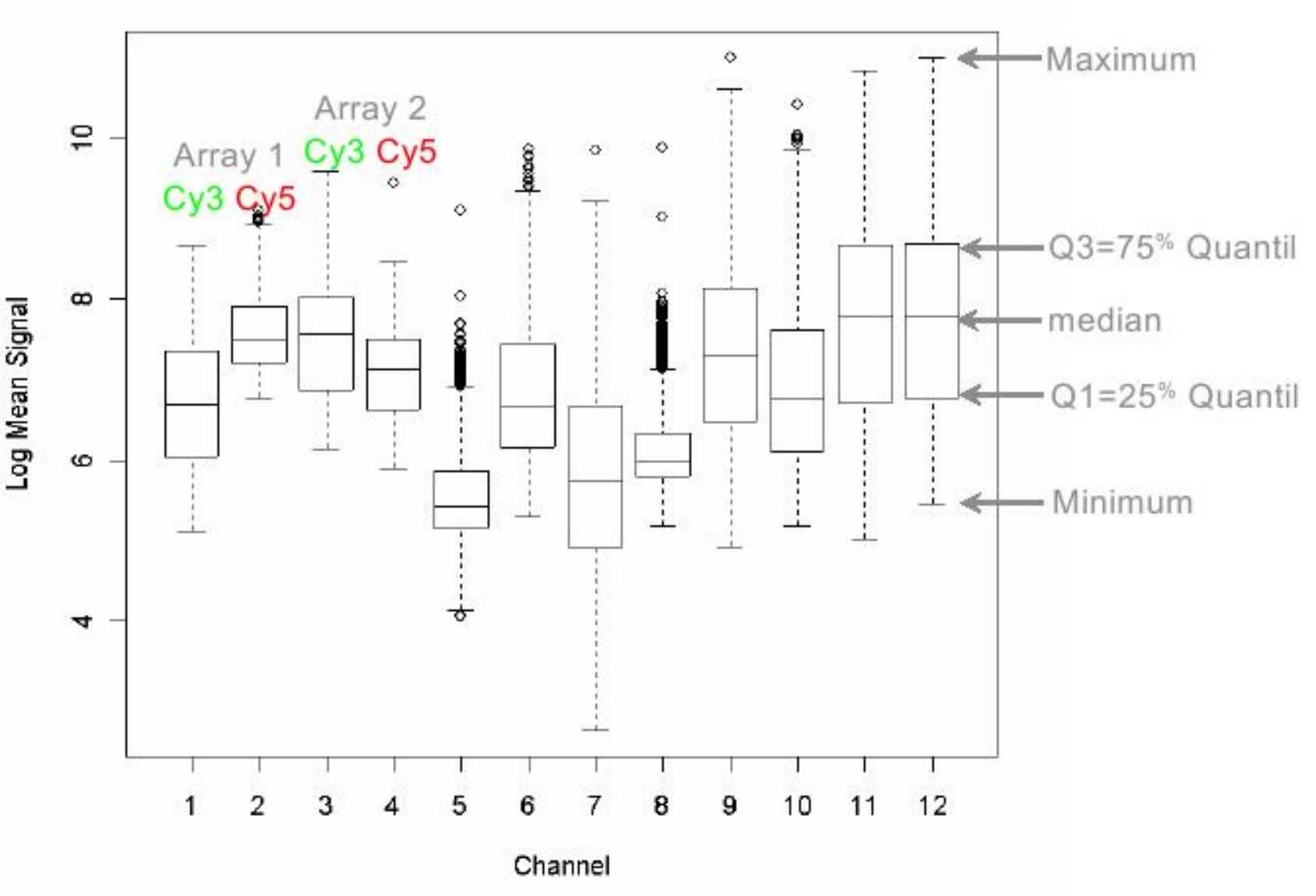
Assumption: Only a small part of the genes are differentially expressed, then the plot of R against G should be a line



- $A = (log_2(R) + log_2(G))/2$
  (**A**ddition, mean intensity)
- $M = log_2(R) - log_2(G)$
  (**M**inus, differential expression)

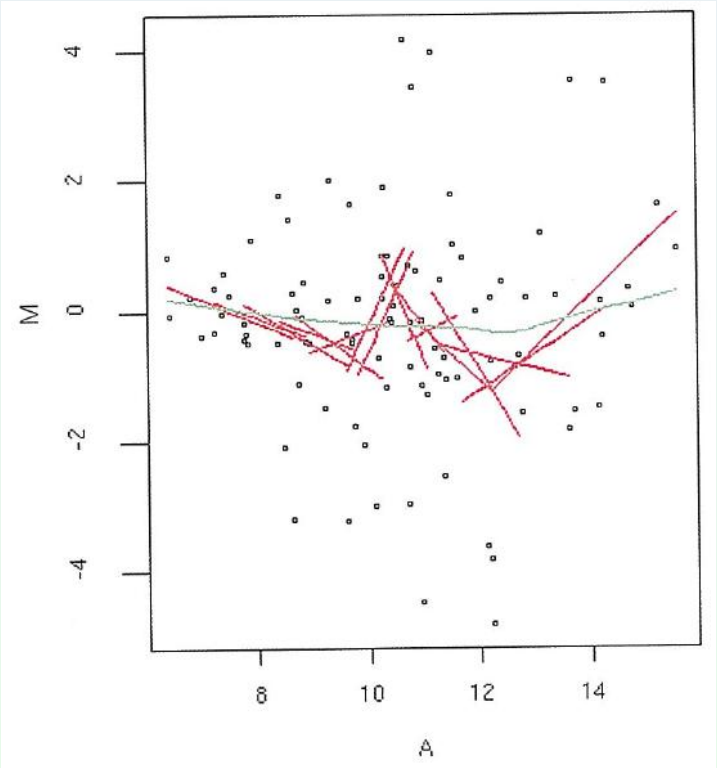


Boxplot of the intensities



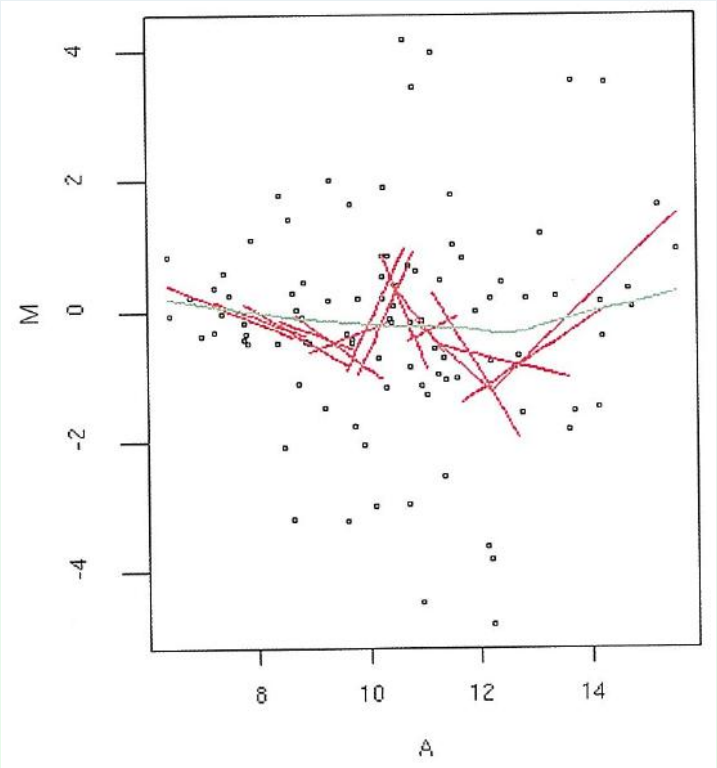
Lowess normalisation

- Normalisation within one array
- Data within a small window are fitted to a straight line
- Straight segments are averaged $\rightarrow$ non-linear fit
- Normalisation: $M_{new} = M_{old} - c(A)$
- Risk of overfitting



Lowess normalisation

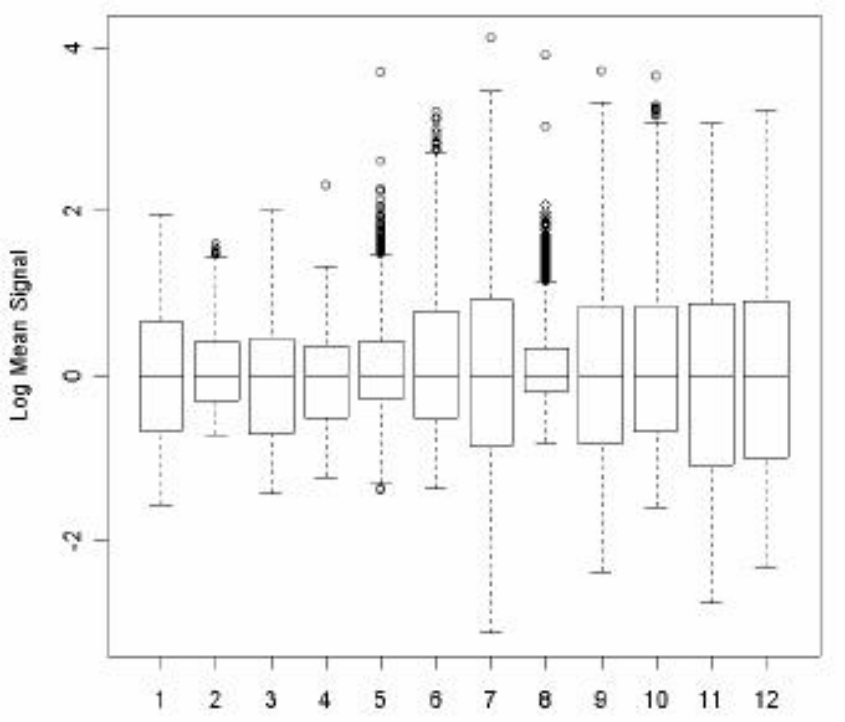
- Normalisation within one array
- Data within a small window are fitted to a straight line
- Straight segments are averaged $\rightarrow$ non-linear fit
- Normalisation: $M_{new} = M_{old} - c(A)$
- Risk of overfitting



<u>Loess normalisation</u>: Similar to Lowess normalisation, but instead of a straight line, a complex polynomial function (e.g. quadratic or cubic) is fitted.

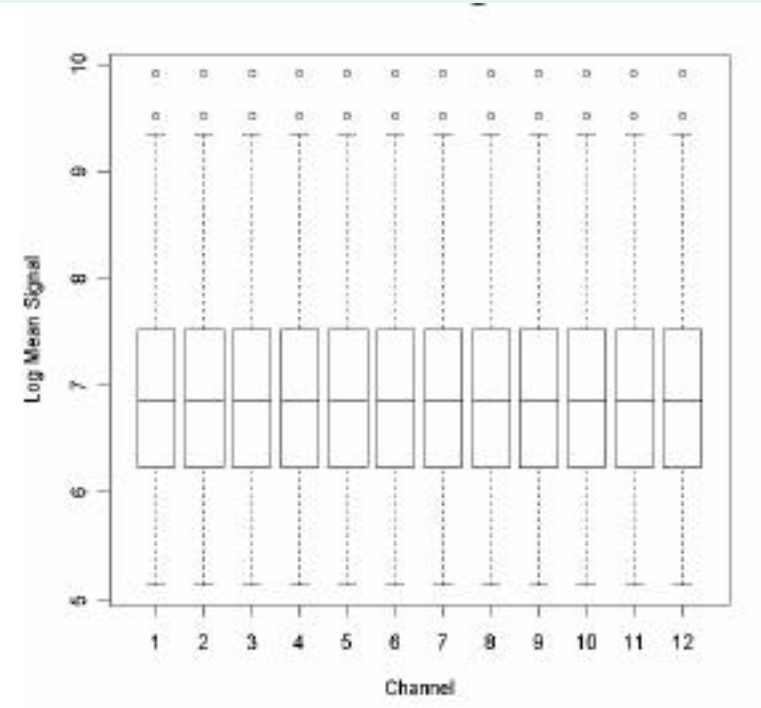
Median centering

- Microarrays, combined in an experiment, have different statistical distributions
- From all expression values of one array the median is substracted and they are divided by the standard deviation
- Global method: Normalisation between arrays after normalisation within arrays



Quantile normalisation

- Ranking the genes by their intensity values
- One array is the masterarray, its intensities are copied to the other arrays for the genes of the same rank
- The intensity distributions are then identical
- Can also be applied to the two dyes of one slide only



Types of experiments

Question: Is the expression of a special gene different in different treatments?

1. One factor
   - Two samples
   - Multiple samples
2. Time courses
3. Factorial experiments

One factor, two samples - no replicates

1. Absolute value of $M = log_2(R) - log_2(G)$

$M < 0$ Gene over-expressed in green-labelled sample compared to red-labelled sample

$M = 0$ Gene equally expressed in both samples

$M > 0$ Gene over-expressed in red-labelled sample compared to green-labelled sample

One factor, two samples - no replicates

1. Absolute value of $M = log_2(R) - log_2(G)$

$M < 0$ Gene over-expressed in green-labelled sample compared to red-labelled sample

$M = 0$ Gene equally expressed in both samples

$M > 0$ Gene over-expressed in red-labelled sample compared to green-labelled sample

2. Statistical modelling of (R, G)-pairs

- But: very error-prone

One factor, two samples - no replicates

1. Absolute value of $M = log_2(R) - log_2(G)$

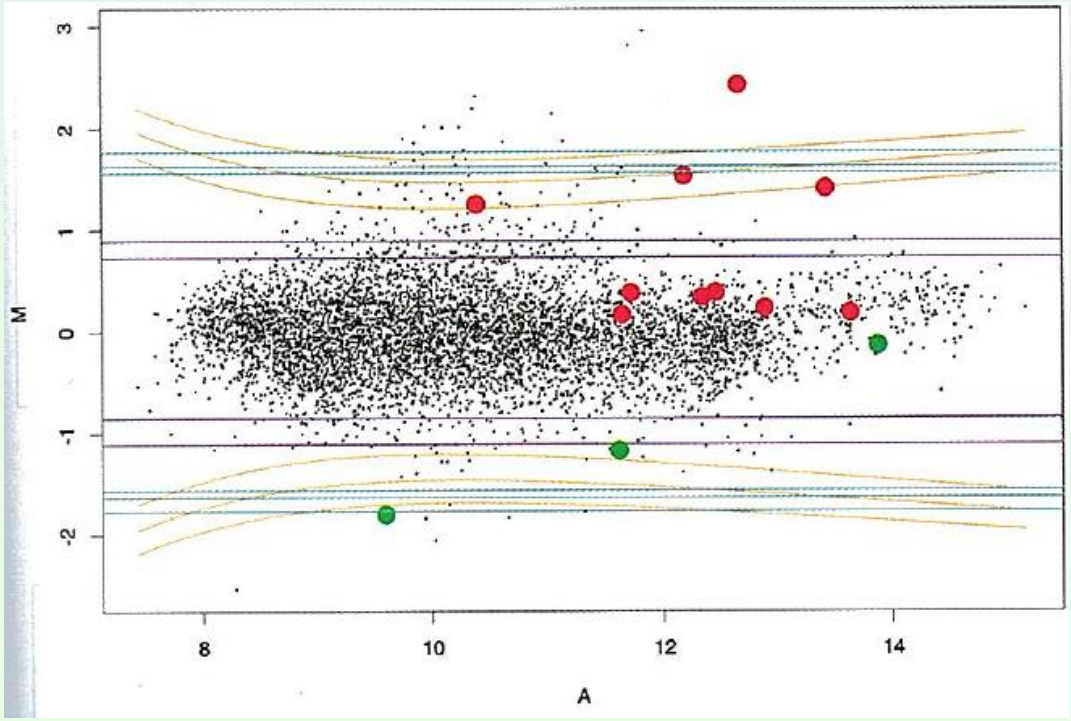
$M < 0$ Gene over-expressed in green-labelled sample compared to red-labelled sample

$M = 0$ Gene equally expressed in both samples

$M > 0$ Gene over-expressed in red-labelled sample compared to green-labelled sample

2. Statistical modelling of (R, G)-pairs

- But: very error-prone



Ranking the genes - $|\overline{M}|$

There are different statistical methods to rank the genes by differential expression when having m replicates:

$|\overline{M}|$ Mean intensities: $\overline{M} = \frac{1}{m} \sum_{i=1}^{m} M_i$

- Problem: Variance of the M-values not considered

Ranking the genes - $|T|$

T-test Null hypothesis: two distributions show the same mean

- here: Does the distribution of M values deviate from mean 0?

Ranking the genes - $|T|$

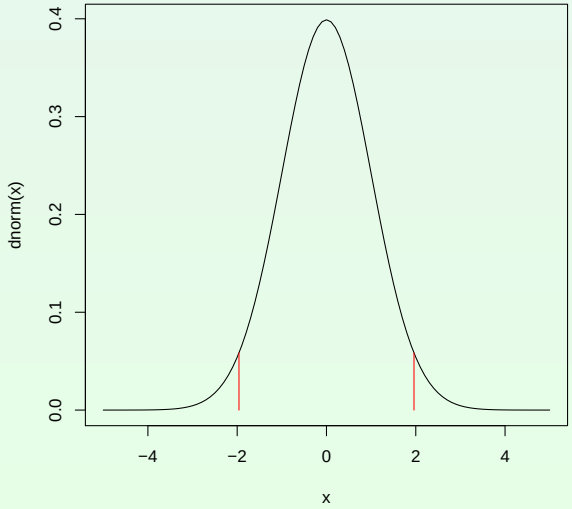
T-test Null hypothesis: two distributions show the same mean

- here: Does the distribution of M values deviate from mean 0?
- $T = \frac{\overline{M}}{\sigma/\sqrt{m}}$ (Standard deviation $\sigma = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(M_i - \overline{M})^2}$)
  - Problem: Large T value can also be caused by a low standard deviation
  - With small sample size σ cannot be well estimated $\rightarrow$ moderated T-statistic (variances are borrowed from other genes)

Ranking the genes - P-value of the T-test

P-value probability that a $|T|$ is larger or equal to the observed $|T|$, while the null hypothesis is true

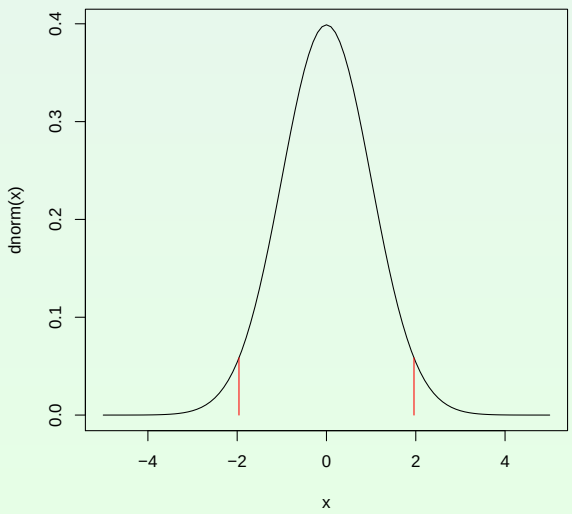
- If P is smaller than a prior chosen cutoff the null hypothesis is rejected
- E.g. 10000 genes on a chip and a cutoff of 0.05, $10000 \times 0.05 = 500$ significant results (false-positives) are expected under the null hypothesis



Ranking the genes - P-value of the T-test

P-value probability that a $|T|$ is larger or equal to the observed $|T|$, while the null hypothesis is true

- If P is smaller than a prior chosen cutoff the null hypothesis is rejected
- E.g. 10000 genes on a chip and a cutoff of 0.05, $10000 \times 0.05 = 500$ significant results (false-positives) are expected under the null hypothesis



	# non-rejected hypo.	# rejected hypo.	
# true null hypo. (non-diff.)	U	V (FP)	n_0
# false null hypo. (diff.)	T (FN)	S	$n - n_0$
	$n - R$	R	n

Ranking the genes - P-value with multiple tests

Solution When testing multiple times, the P-values must be **adjusted**

FWER Family-wise error rate: Probability of at least one false-positive - $Pr(V > 0)$

- Bonferroni correction: multiply all P-values with the number of tests

Ranking the genes - P-value with multiple tests

Solution When testing multiple times, the P-values must be **adjusted**

FWER Family-wise error rate: Probability of at least one false-positive - $Pr(V > 0)$

- Bonferroni correction: multiply all P-values with the number of tests

FDR False discovery rate: Expected proportion of true null hypotheses on all rejected hypotheses - $E[V/R]$

Ranking the genes - P-value with multiple tests

Solution When testing multiple times, the P-values must be **adjusted**

FWER Family-wise error rate: Probability of at least one false-positive - $Pr(V > 0)$

- Bonferroni correction: multiply all P-values with the number of tests

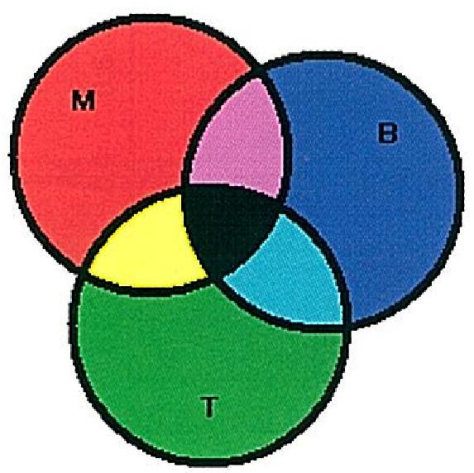
FDR False discovery rate: Expected proportion of true null hypotheses on all rejected hypotheses - $E[V/R]$

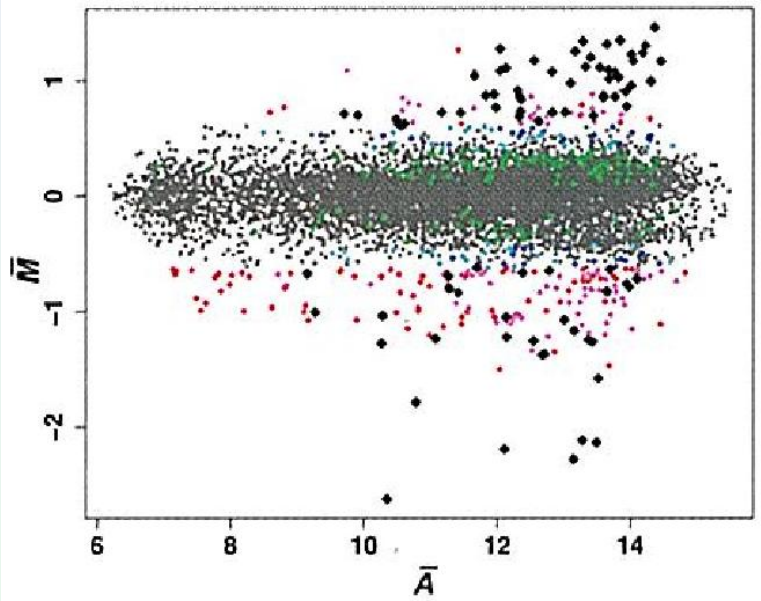
The significance level 5 % now controls FWER or FDR.

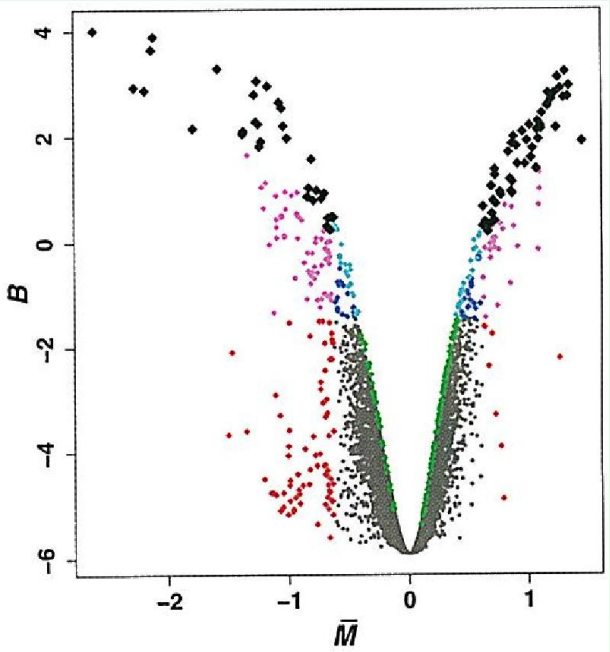
Ranking the genes - B

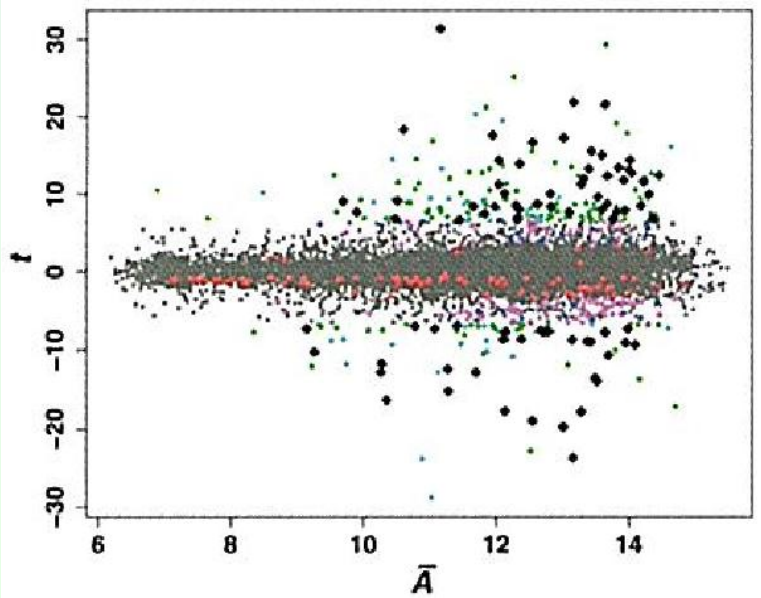
- Empirical Bayes method estimates posterior probabilities for differential expression
- Need prior assumtions about distribution of differentially expressed genes
- B-values are posterior log odds for differential expression
- Estimated variables are used for other statistics:
  - Moderated T
  - Moderated F: Combination of T-statistics to an overall test for significance for that gene

Example



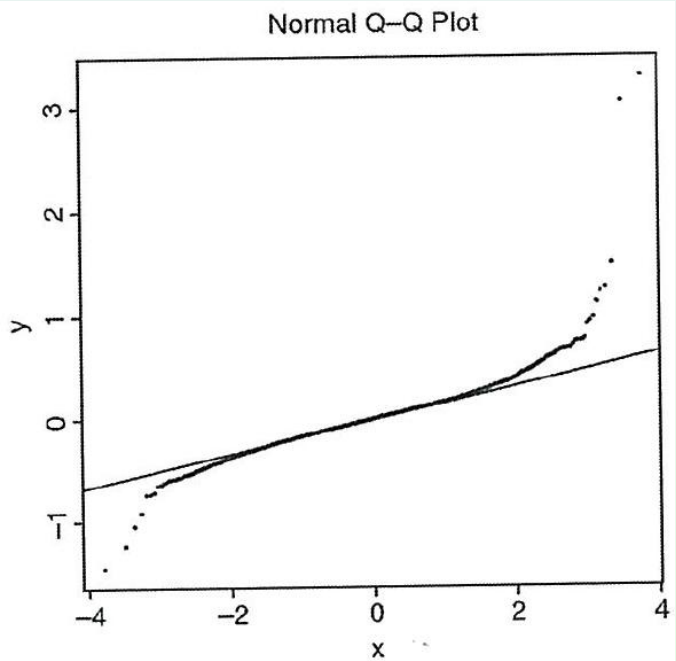






Graphical representations

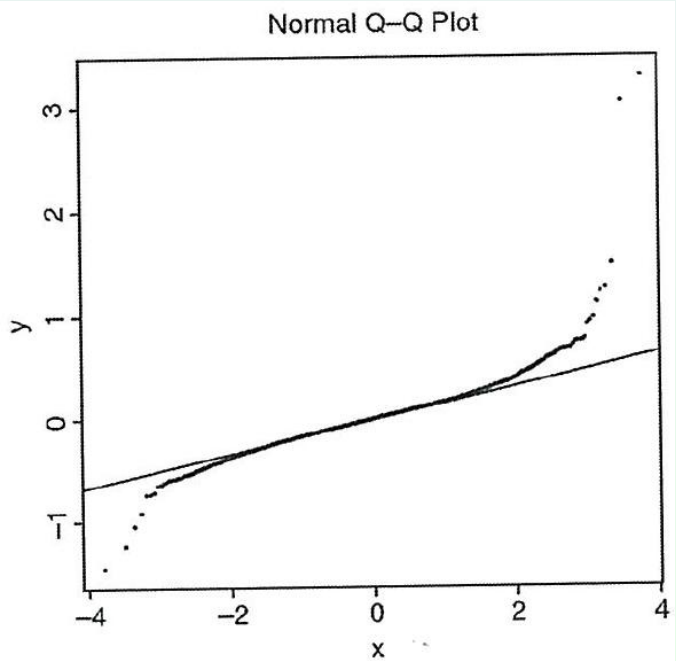
Quantile-quantile plot:
M or T against the normal distribution



Graphical representations

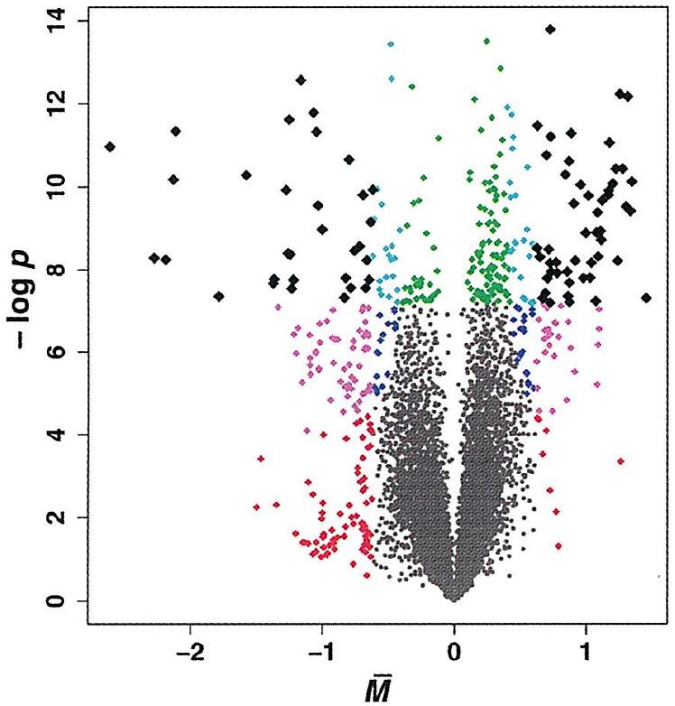
Quantile-quantile plot:

M or T against the normal distribution



Volcano plot:

Plot of the P-values against M



Gene sets

Use the functional information (meta-data, annotations) available for the genes on the array to define gene sets:

GO Gene Ontology: Molecular function, biological process and cellular component

- Annotations arranged in a directed acyclic graph

Pathways KEGG, BioCarta, GenMapp

Loc Chromosomal Localisation $\rightarrow$ clusters of co-regulated genes

TFBS Transcription factor binding sites

...

Gen-Class Testing (differentially expressed genes)

Guess: List of differentially expressed genes are functionally related

Problem: Find functional group(s) which are related to the differentially expressed genes

Procedure: Choose gene sets of known function and test every set whether it is overrepresented in the set of differentially expressed genes

Gen-Class Testing (differentially expressed genes)

Guess: List of differentially expressed genes are functionally related

Problem: Find functional group(s) which are related to the differentially expressed genes

Procedure: Choose gene sets of known function and test every set whether it is overrepresented in the set of differentially expressed genes

Test Null hypothesis: The amount of a category K is equally distributed among differentially and non-differentially expressed genes

2×2 Contingency table:

	diff	nd
K	a	b
not K	c	d

$\rightarrow$ Fisher-Test (hypergeometric distribution)

Gen-Class Testing (differentially expressed genes)

Guess: List of differentially expressed genes are functionally related

Problem: Find functional group(s) which are related to the differentially expressed genes

Procedure: Choose gene sets of known function and test every set whether it is overrepresented in the set of differentially expressed genes

Test Null hypothesis: The amount of a category K is equally distributed among differentially and non-differentially expressed genes

2×2 Contingency table:

	diff	nd
K	a	b
not K	c	d

$\longrightarrow$ Fisher-Test (hypergeometric distribution)

Attention: Multiple tests and complex dependencies

Rank-based Gene-Class Testing

- Genes ranked by a measure for differential expression (e.g. $|T|$, B), but no cutoff needed

Rank-based Gene-Class Testing

- Genes ranked by a measure for differential expression (e.g. $|T|$, B), but no cutoff needed

KS Kolmogorov-Smirnov-Test: Does the genes of category K occur more frequently in the beginning of the list?

- Null distribution estimated by permutation

Distance functions

Data matrix E:

Gene	Sample		
	1	...	m
1	Expres-		
⋮	sion		
n	values		

Distance functions

Data matrix E:

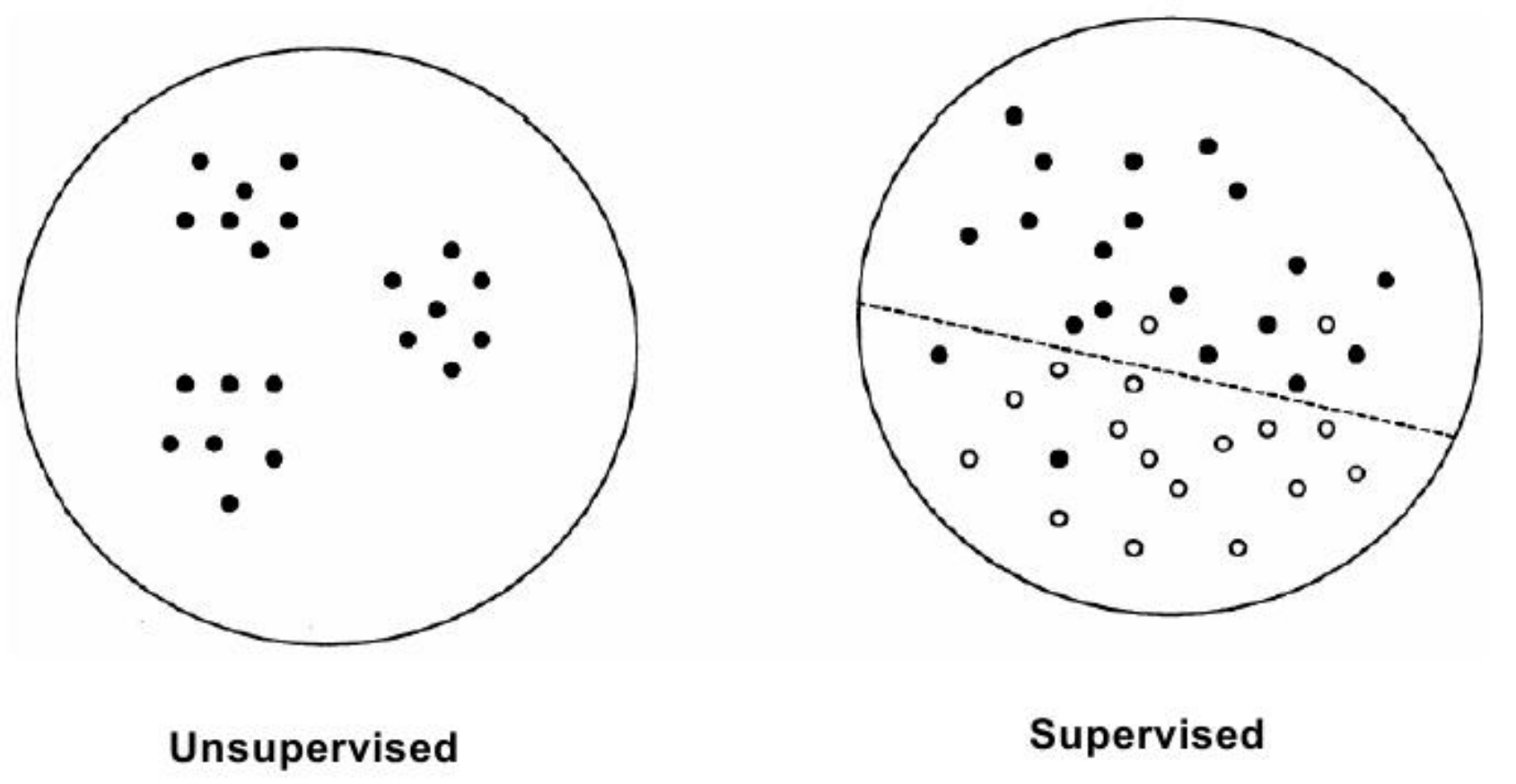
Gene	Sample 1	...	m
1	Expres-		
⋮	sion		
n	values		

Application of distance funtions to the n-dimensional column vectors:

1. Euclidean distance:
$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$
2. $1 - r(x,y)$ with correlation coefficient r
3. $1 - |r(x,y)|$

Analogous for the m-dimensional row vectors

Types of learning



Classification

Classification is a form of unsupervised learning $\rightarrow$ external information is used.

Question: Classification of patients by their expression profiles (learn with healthy and ill persons)

Classification

Classification is a form of unsupervised learning $\longrightarrow$ external information is used.

Question: Classification of patients by their expression profiles (learn with healthy and ill persons)

Multilevel process:

1. Feature selection: Select informative components
2. Learn a classifier with labelled samples
3. Classify an unlabelled sample with the classifier

Feature selection (Gene filtering)

- A Classification with the complete n-dimensional data is often problematic
- Improvement: extract N genes, that distinguish best between the classes and learn the classifier only with the reduced N-dimensional data

Feature selection (Gene filtering)

- A Classification with the complete n-dimensional data is often problematic
- Improvement: extract N genes, that distinguish best between the classes and learn the classifier only with the reduced N-dimensional data

- m_1 data sets for class 1 and m_2 data sets for class 2
  1. T-Test for every gene, whether two classes have the same mean expression value
  2. Wilcoxon-Test whether two classes have the same median (non-parametric test)

  - Only thake the N most significant genes

Classification algorithms

k-NN k nearest neighbors:

- Majority decision of the k objects with the smallest distance to the classified object

Classification algorithms

k-NN k nearest neighbors:

- Majority decision of the k objects with the smallest distance to the classified object

LDA Linear discriminant analysis

- For every class, a "feature vector" is learned which represents the class

Classification algorithms

k-NN k nearest neighbors:

- Majority decision of the k objects with the smallest distance to the classified object

LDA Linear discriminant analysis

- For every class, a "feature vector" is learned which represents the class

CART Classification and regression trees:

- Decision trees: Partitioning with respect to a component (gene expression value) on every inner node, class labels on the leaves

Classification algorithms

k-NN *k* nearest neighbors:

- Majority decision of the *k* objects with the smallest distance to the classified object

LDA Linear discriminant analysis

- For every class, a "feature vector" is learned which represents the class

CART Classification and regression trees:

- Decision trees: Partitioning with respect to a component (gene expression value) on every inner node, class labels on the leaves

SVM Support vector machines:

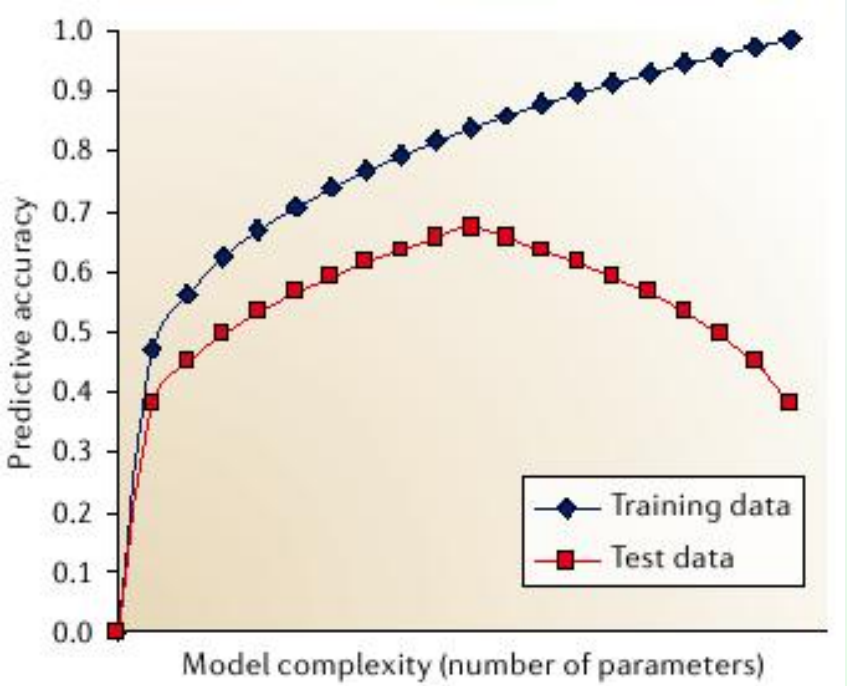
- With a mathematical expression, the objects are transfered in a space where they can be separated with a straight line

Validation

To protect the classifier against overfitting, a test data set is neccessary.

Cross validation:

- The labelled data is partitioned several times in training data and test data
- The classifier is learned with the training data and the test data is classified

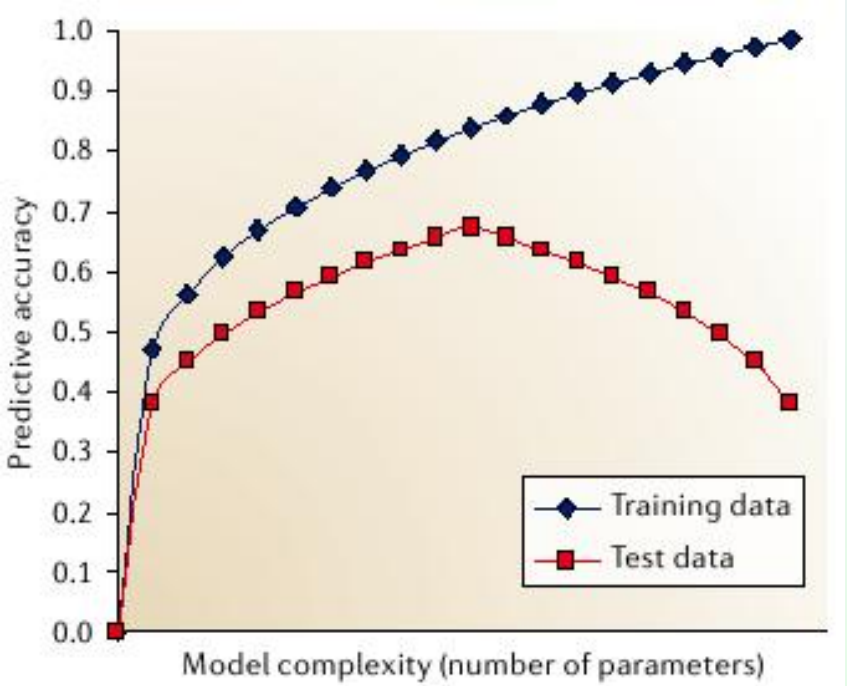


Validation

To protect the classifier against overfitting, a test data set is neccessary.

Cross validation:

- The labelled data is partitioned several times in training data and test data
- The classifier is learned with the training data and the test data is classified



The gene selection can also be validated (avoids overfitting to the selected genes)

Clustering

Clustering is a form of unsupervised learning → no external information is used.

Input: Distances computed between the genes from a microarray experiment

Output: Assignment of classes to the genes

Clustering

Clustering is a form of unsupervised learning → no external information is used.

Input: Distances computed between the genes from a microarray experiment

Output: Assignment of classes to the genes

Also: Clustering of samples or two-sided clustering

Clustering

Clustering is a form of unsupervised learning → no external information is used.

Input: Distances computed between the genes from a microarray experiment

Output: Assignment of classes to the genes

Also: Clustering of samples or two-sided clustering

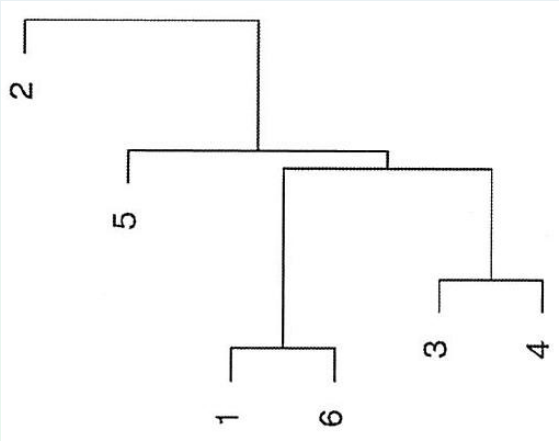
Problems:

- Few known about reliability and problems of clustering methods
- Hard to reproduce
- Does not answer biological question for differential expression

Clustering algorithms

HC Hierarchical clustering

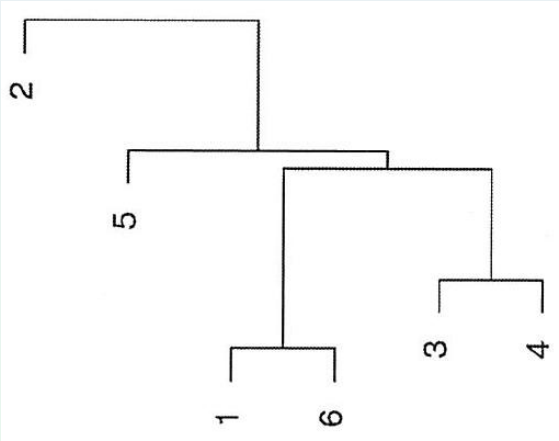
- Genes with the smallest distance are merged
- New distances computed to inner node
- Tree (dendogram) is produced
- Mistakes cannot be taken back



Clustering algorithms

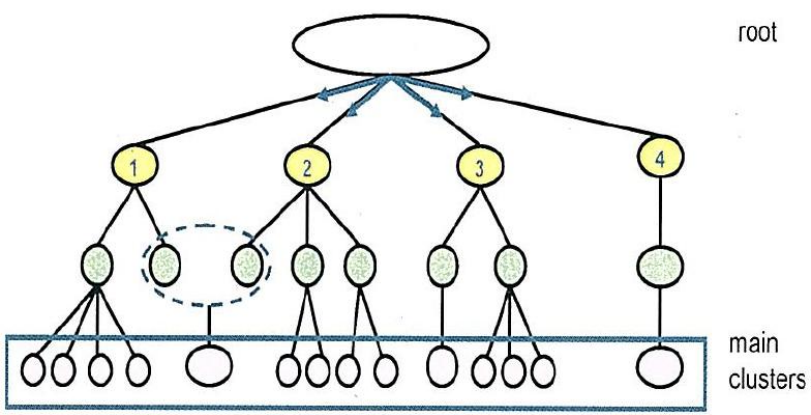
HC Hierarchical clustering

- Genes with the smallest distance are merged
- New distances computed to inner node
- Tree (dendogram) is produced
- Mistakes cannot be taken back



Hopach Hierarchical ordered partitioning and collapsing hybrid

- Partitioning und merging steps



Clustering algorithms

k-means Partition clustering

- k classes $\rightarrow$ class means $\rightarrow$ classification according to smallest distance $\rightarrow$ new classes $\rightarrow \ldots$
- The classes are recomputed in every step
- Initialised randomly, must not converge always

Clustering algorithms

k-means Partition clustering

- k classes $\rightarrow$ class means $\rightarrow$ classification according to smallest distance $\rightarrow$ new classes $\rightarrow \ldots$
- The classes are recomputed in every step
- Initialised randomly, must not converge always

PCA Principal component analysis

- Dimension reduction: Extraction of the components, that explain the largest portion of the variation (can see clusters)

Clustering algorithms

k-means Partition clustering

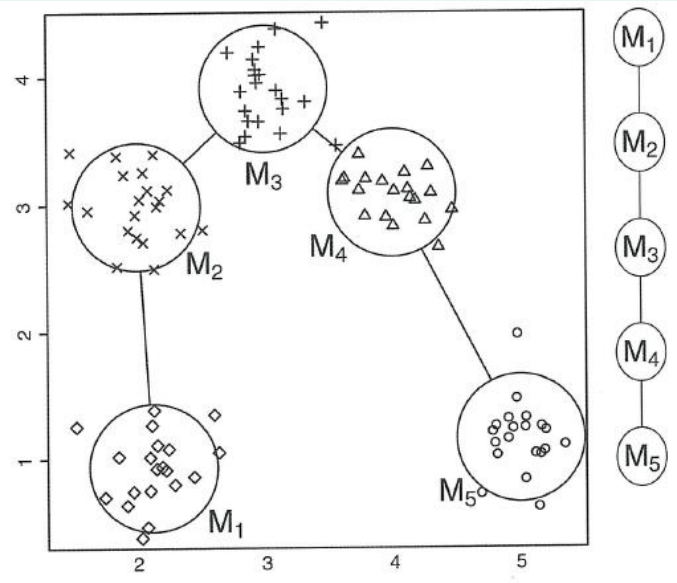
- k classes $\rightarrow$ class means $\rightarrow$ classification according to smallest distance $\rightarrow$ new classes $\rightarrow$ ...
- The classes are recomputed in every step
- Initialised randomly, must not converge always

PCA Principal component analysis

- Dimension reduction: Extraction of the components, that explain the largest portion of the variation (can see clusters)

SOM Self-organising maps

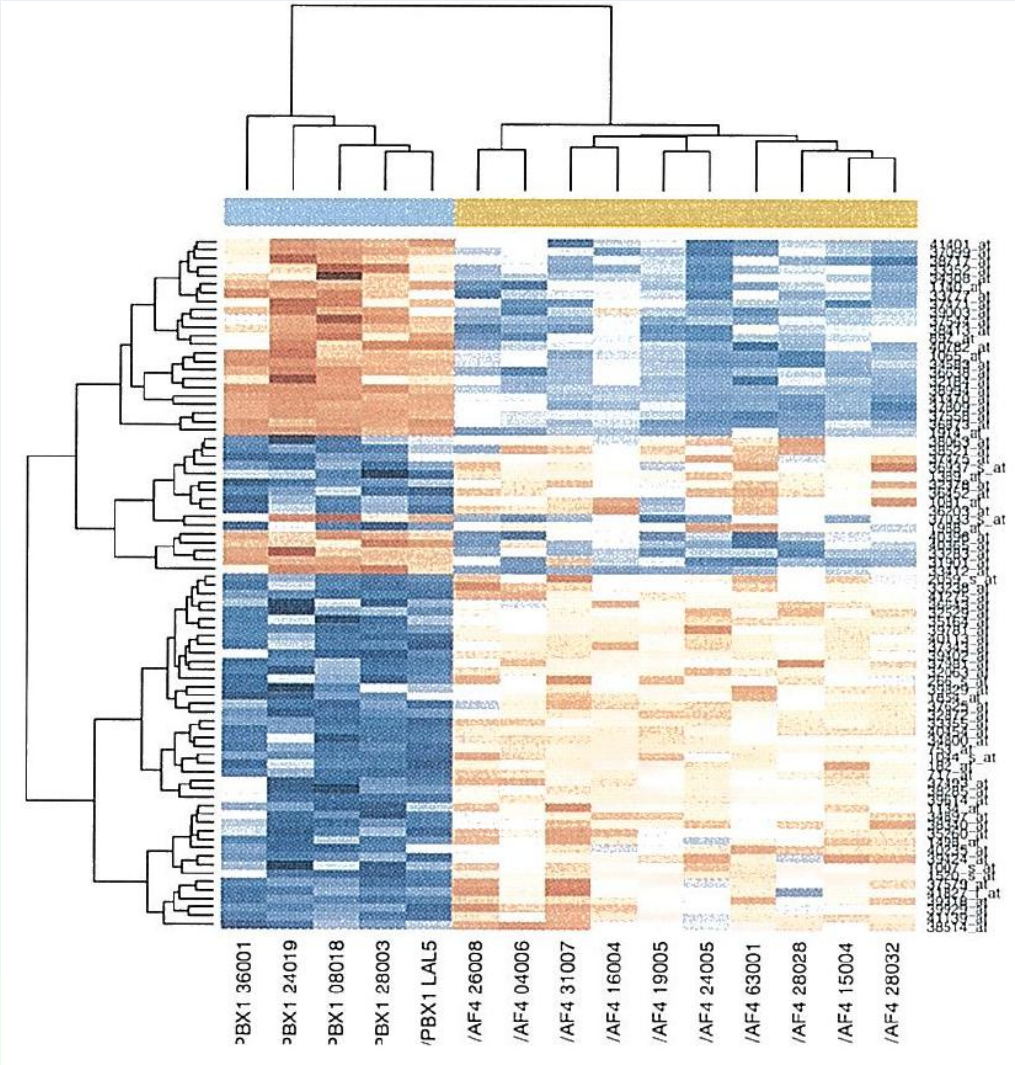
- Dimension reduction: high-dimensional data is represented by a lower dimensional grid



Clustering as a visualisation tool

Heatmap:

- Color-coding of the expression level
- Two-sided hierarchical clustering
- Rearrangement of rows and columns such that similar rows (columns) are placed next to each other



Literature

- David W. Mount. 2005. *Bioinformatics: Sequence and Genome Analysis. Second edition.* (Chapter 13) CSHL Press
- Terry Speed. 2003 *Statistical Analysis of Gene Expression Microarray Data.* Chapman & Hall
- David B. Allison et. al. 2005. Microarray data analysis: from disarray to consolidation and consensus. *Nature reviews genetics* 7: 55-65
- Robert Gentleman et. al. 2005 *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Springer
- Limma's Usersguide: `http://bioconductor.org/packages/2.0/bioc/vignettes/limma/inst/doc/usersguide.pdf`