Analysis of gene expression

Anne Kupczok

Center for Integrative Bioinformatics Vienna
Max F. Perutz Laboratories

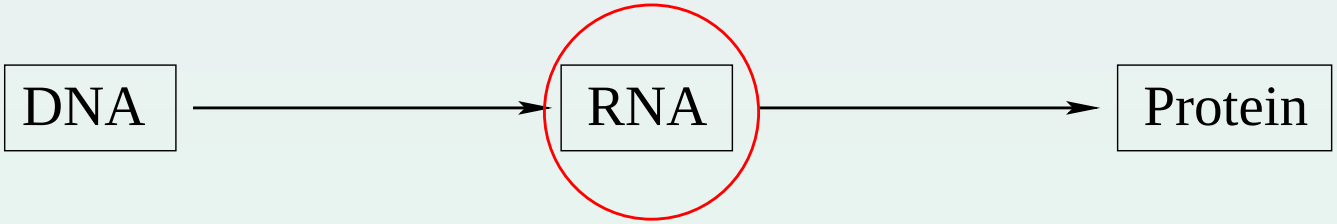
anne.kupczok@univie.ac.at

Contents

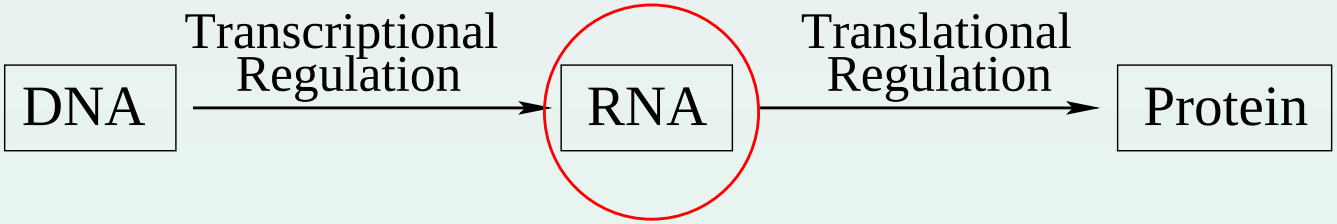
1. Motivation
2. Techniques
3. Tag data
   - Experiment
   - Design
   - Preprocessing
   - Differential expression
4. Microarray data
   - Experiment
   - Design
   - Preprocessing and Normalization
   - Differential expression
5. Functional analysis
6. Learning
   - Classification
   - Clustering
7. Literature

The central dogma of molecular biology

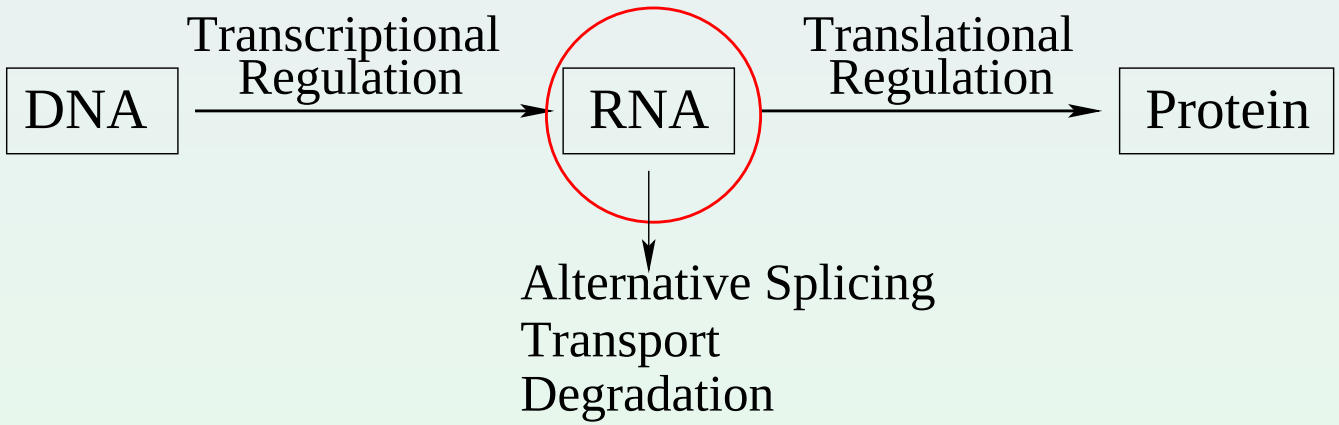
The central dogma of molecular biology



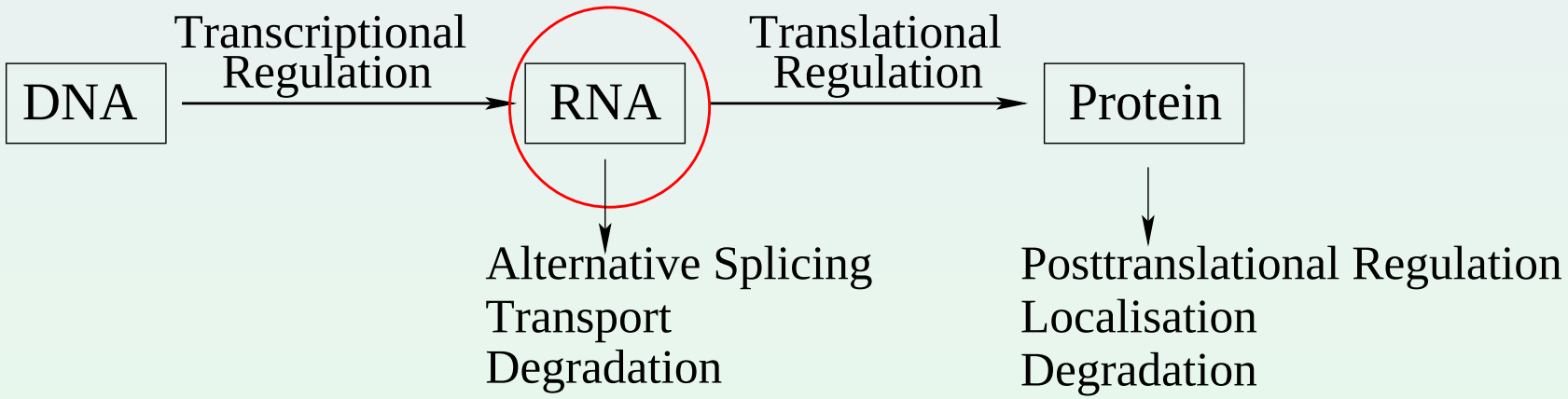
The central dogma of molecular biology



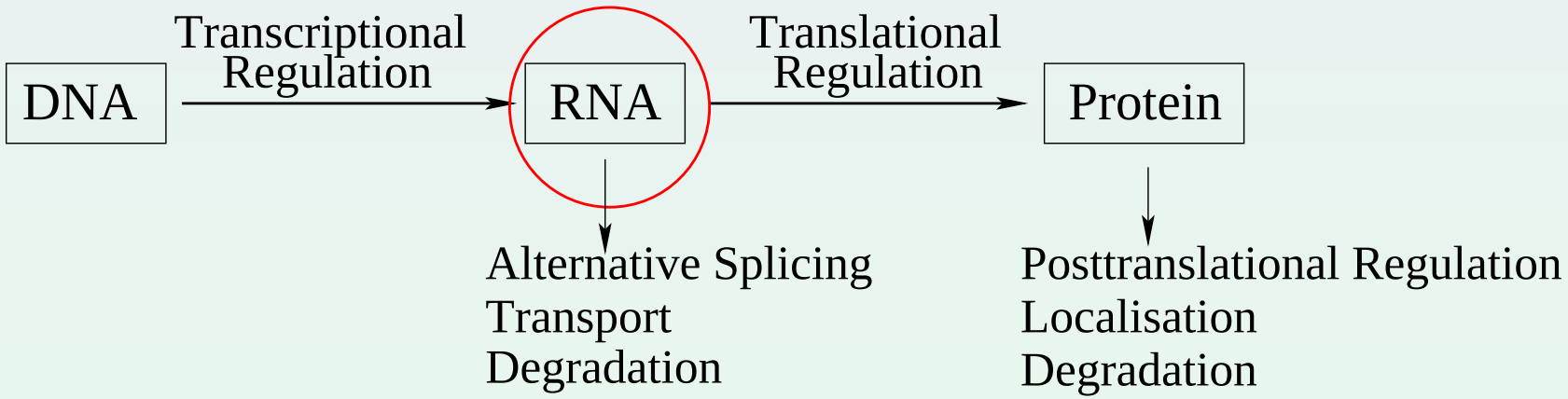
The central dogma of molecular biology



The central dogma of molecular biology



The central dogma of molecular biology



Analysis of 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, tissues)
- 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, tissues)
- Gene expression change during development
- Gene regulation networks

Simultaneous measurement of the expression of thousands of genes $\longrightarrow$ global view on gene expression

Differential expression: Is the expression of a special gene different in different treatments?
Learning: What accounts for the difference in different treatments?
Functional analysis: Which functional classes are different in different treatments?

Differential expression: Is the expression of a special gene different in different treatments?
Learning: What accounts for the difference in different treatments?
Functional analysis: Which functional classes are different in different treatments?

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

Experimental techniques

Analog	**Digital**
Measurement	Counting
Microarrays	e.g. SAGE (Serial Analysis of Gene Expression)
Hybridization	Sequencing
Lots of statistics	Robust statistics
Need to design chip	Do not need sequence in advance

Sources of error

Biological noise:

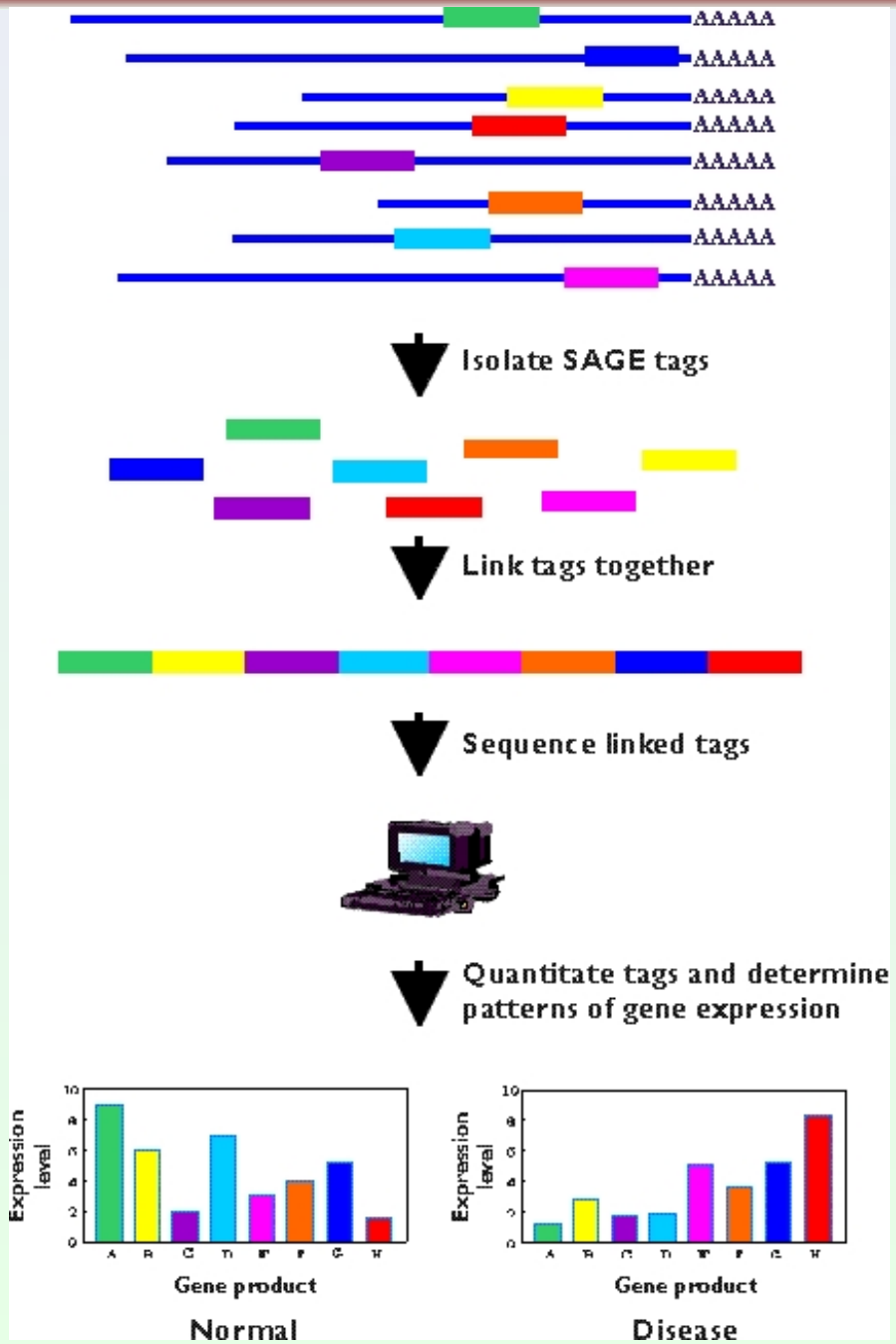
- Transcription is a stochastic process
- Posttranscriptional regulation
- Stability of the mRNA

Technical limitations:

- cDNA from mRNA
- Microarray:
  - Binding of the dye
  - Hybridization kinetics
  - cross-hybridization
  - Measurement of the signal
- SAGE:
  - Detection of tags
  - Tags not unique or not present
  - Sequencing errors

SAGE:

- extract short (10-20 bp) tags from cDNA
- cut with special restriction enzymes from 3' end
- if transcript is known → know 'virtual' transcript
- tags are concatenated, cloned and sequenced → get counts



Design issues - Probability of detecting a transcript

- The tags stem from randomly picked transcripts
- Due to experimental treatment, GC-bias has been observed

- Even in absence of any noise their frequencies are not a perfect representation of the frequencies in the cell but follow a binomial distribution: $P(k) = \binom{N}{k} p^k (1-p)^{N-k}$

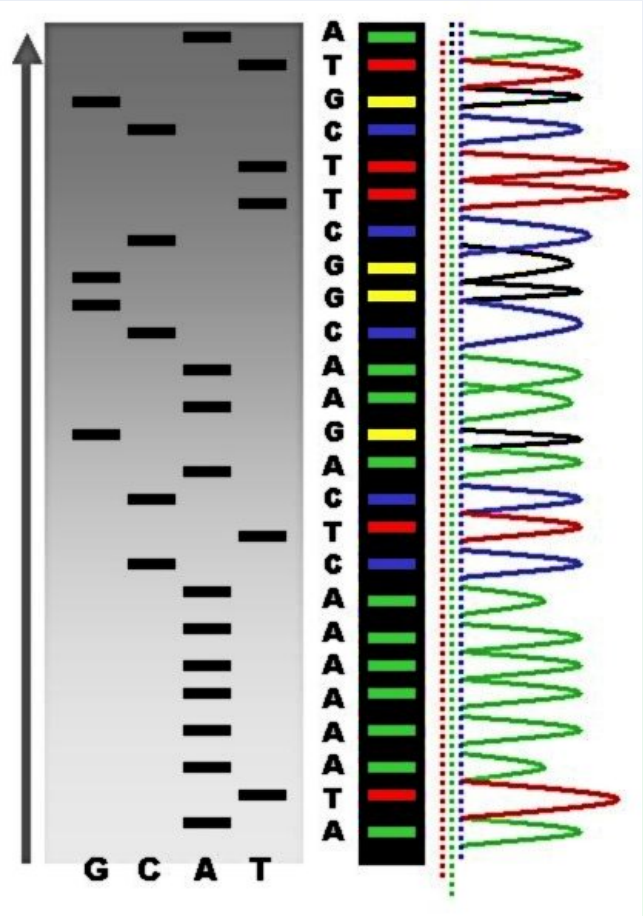
- N: Library size (total number of tags)
  k: count of tag x which occurs in cell with proportion p

Minimal count of a transcript in the cell to be detected with probability $> 95\%$ (total number of transcripts 300 000):

N	$k \geq 1$	$k \geq 2$
10 000	91	144
100 000	10	16
1 000 000	2	3

Minimal count of a transcript in the cell to be detected with probability $> 95\%$ (total number of transcripts 300 000):

N	$k \geq 1$	$k \geq 2$
10 000	91	144
100 000	10	16
1 000 000	2	3



As **preprocessing** single counts are usually excluded since they may be due to sequencing error. This reduces the detection probability of low abundance transcripts.

Differential expression

Question: Given 2 Libraries S_1 and S_2, where tag x occures n_1 times and n_2 times, respecitvly, is x differentially expressed?

	S_1	S_2
x	n_1	n_2
others	$L_1 - n_1$	$L_2 - n_2$

Differential expression

Question: Given 2 Libraries S_1 and S_2, where tag x occures n_1 times and n_2 times, respecitvly, is x differentially expressed?

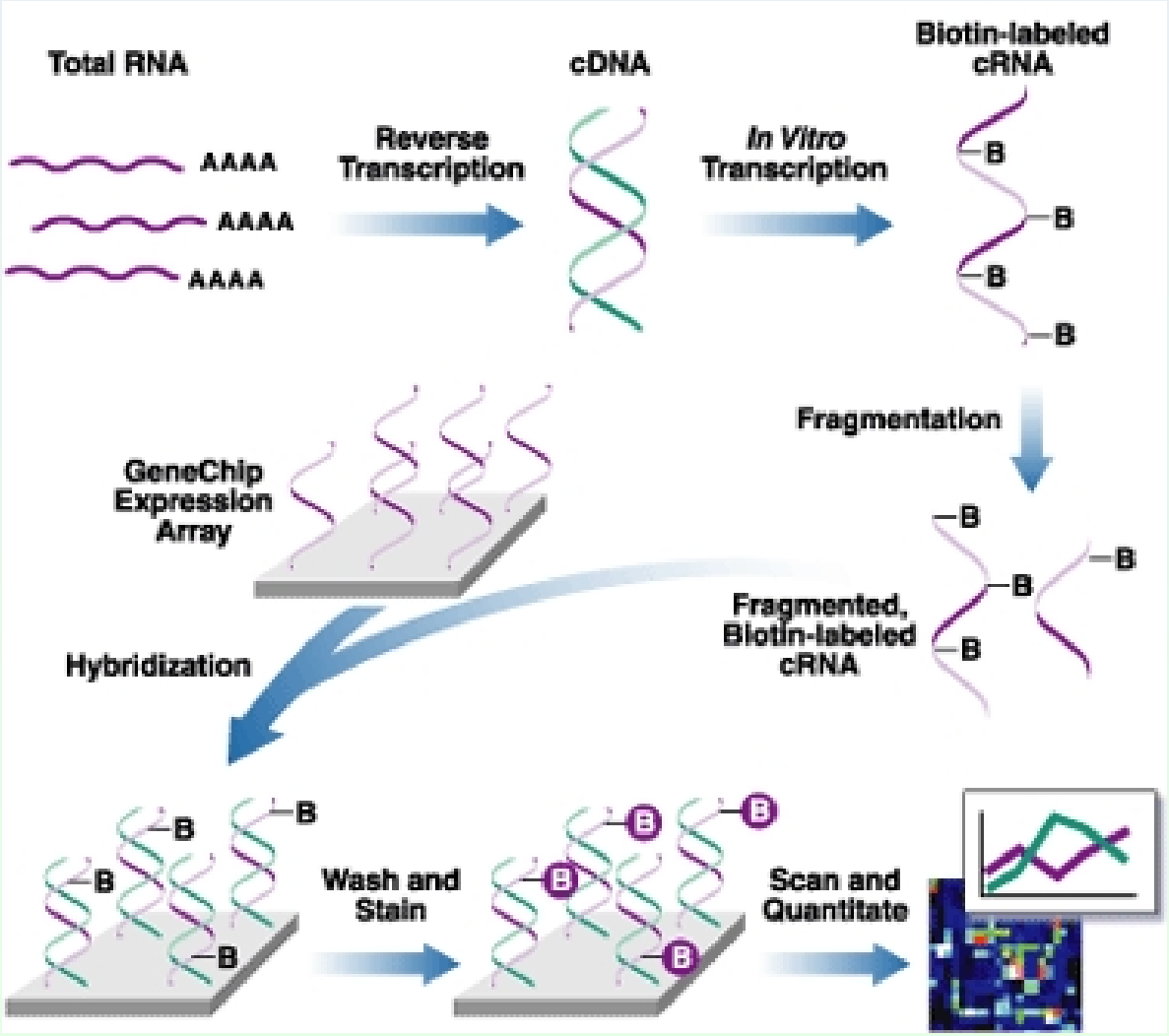
	S_1	S_2
x	n_1	n_2
others	$L_1 - n_1$	$L_2 - n_2$

Null Hypothesis:
The proportions are equal
Test Statistic:

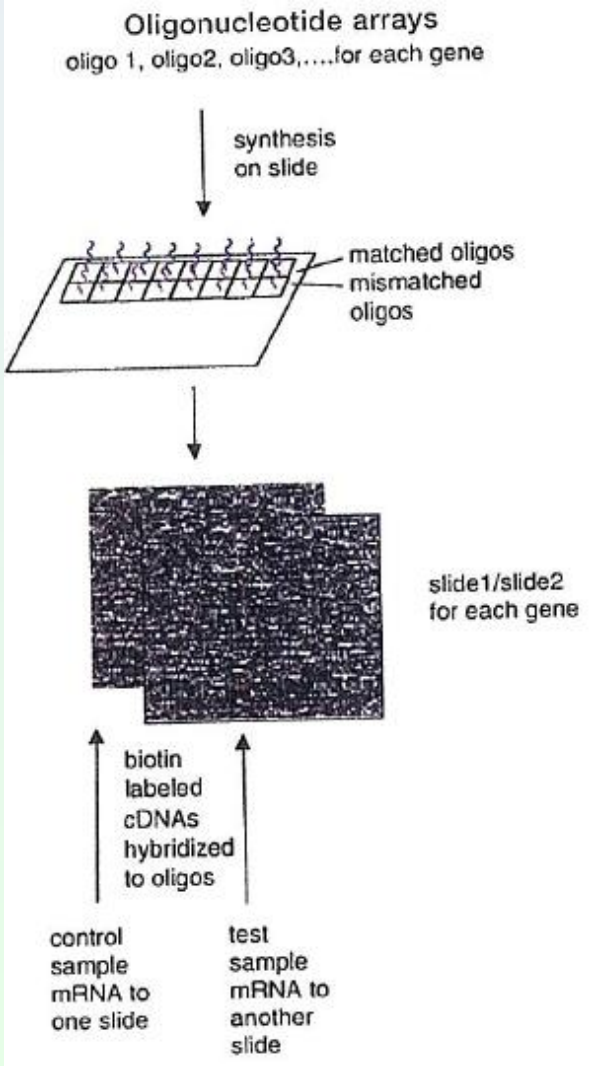
- χ^2
- Fisher's exact test

- Test proportions n_1/L_1 and n_1/L_1 for equality with z-Test

Survey of one microarray experiment

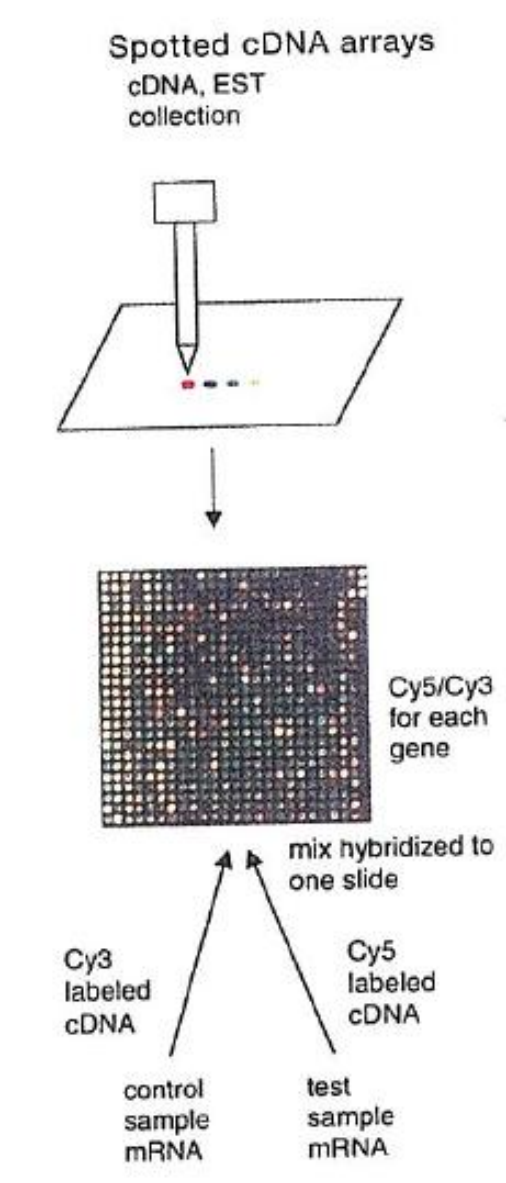


Oligonucleotide arrays



- e.g. Affymetrix arrays
- In situ synthesis is used to build probes bp by bp
- Oligonucleotides of length ≈ 25 on array
  - Perfect matching sequences
  - One or more mismatching nucleotides (control for non-specific binding)
- One biological sample per array (a new slide for every sample)
- cDNAs are labelled with biotin

cDNA arrays



- Spotting technology to attach probes to chip (e.g. cDNA library)
- 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

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.

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:

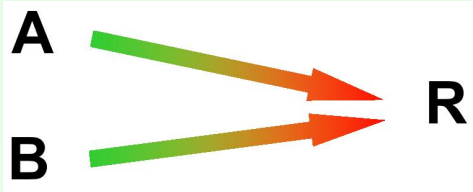
$$\frac{\text{Expression in sample B}}{\text{Expression in sample A}} \text{ (normally in } log_2\text{-scale)}$$

		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 via Refernce		$(logR - logA) - (logR - logB)$	$2\sigma^2$

Analysis of microarrays

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

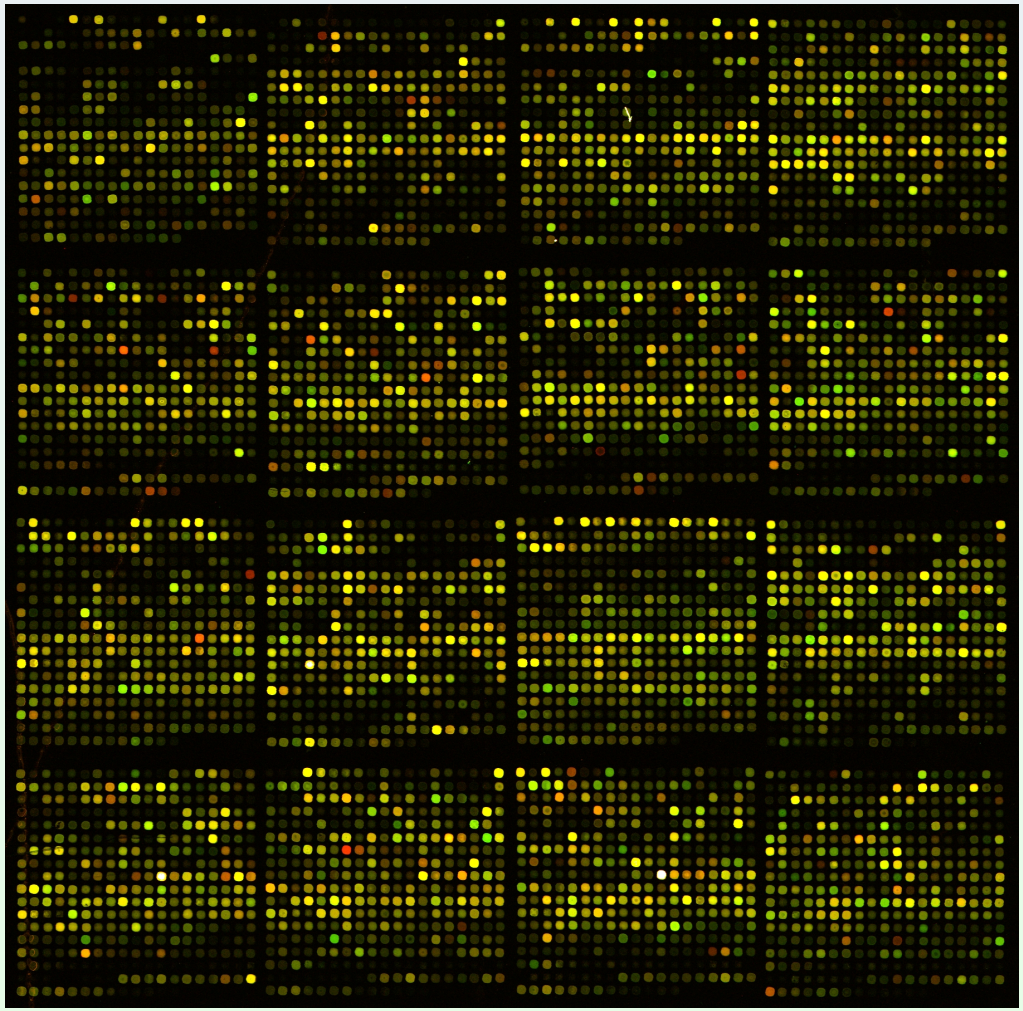
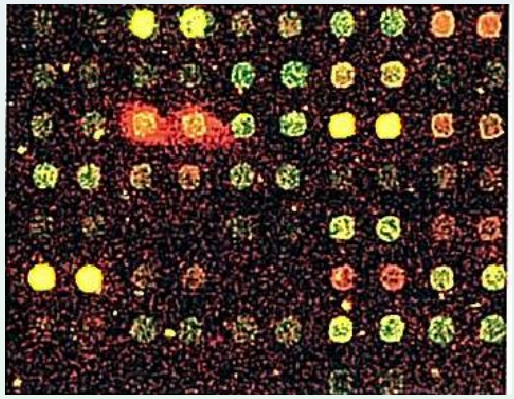


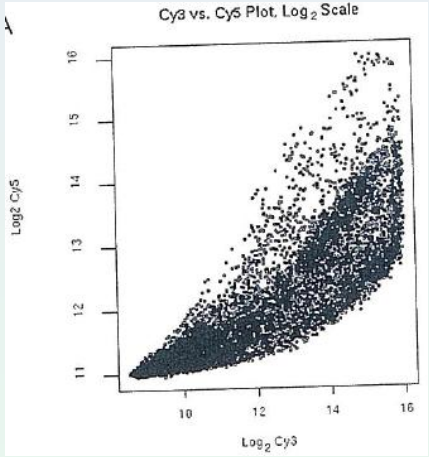
Image analysis



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

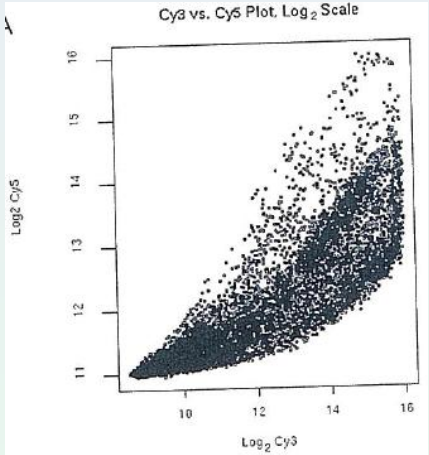
Normalization of cDNA arrays: M/A plot

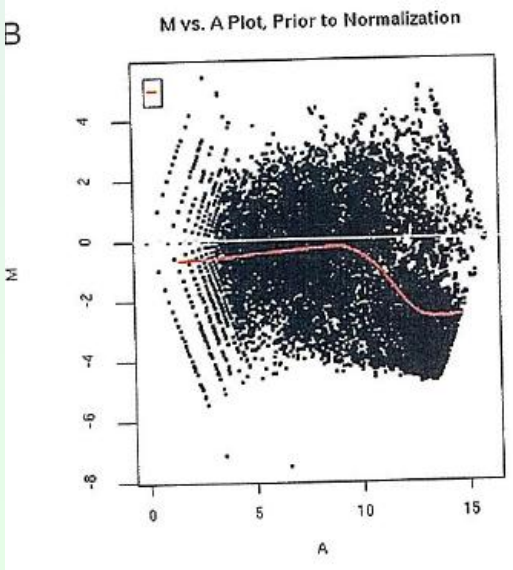
Assumption: Only a small part of the genes are differentially expressed, then the plot of R against G should be a line



Normalization of cDNA arrays: M/A plot

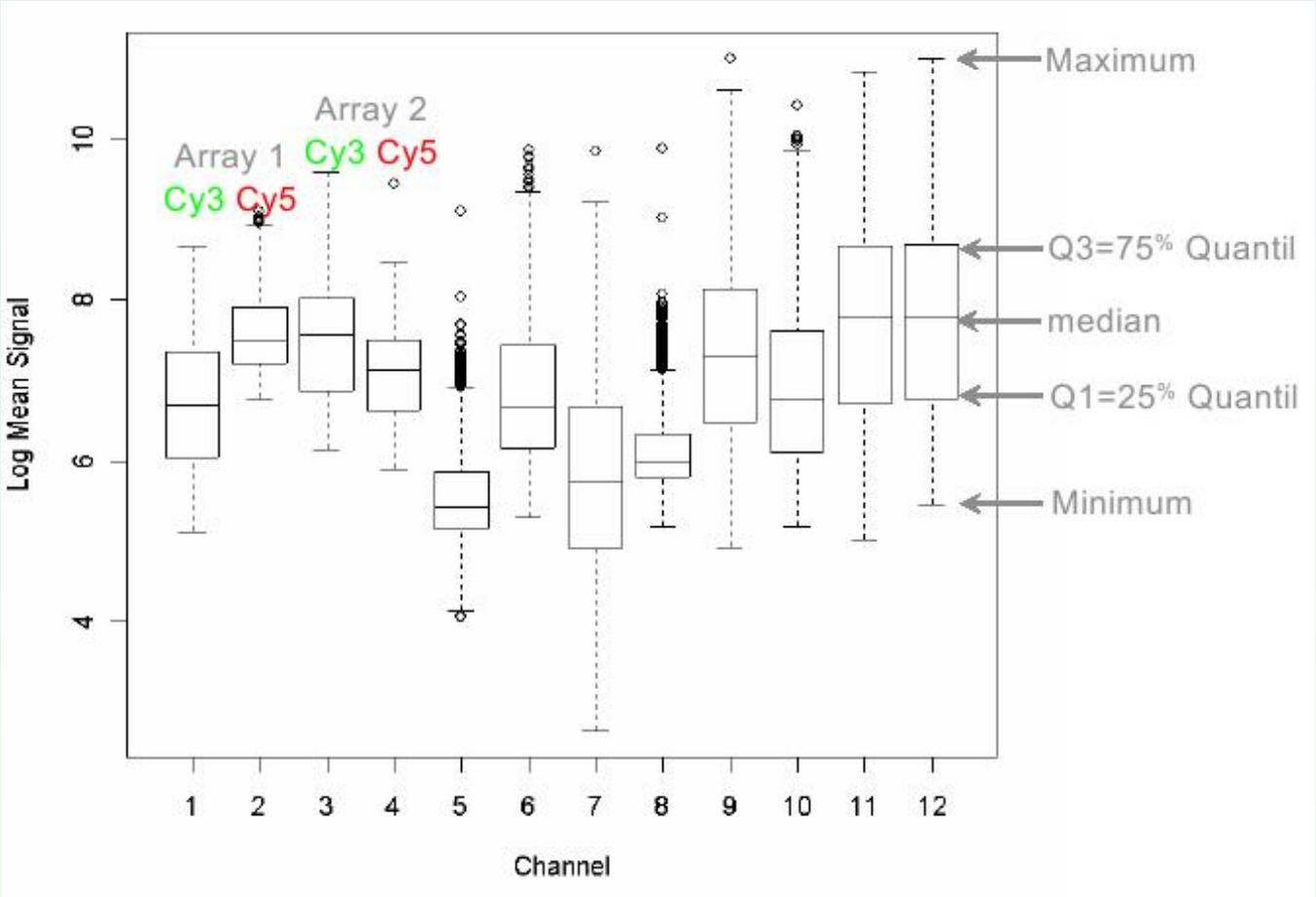
Assumption: Only a small part of the genes are differentially expressed, then the plot of R against G should be a line



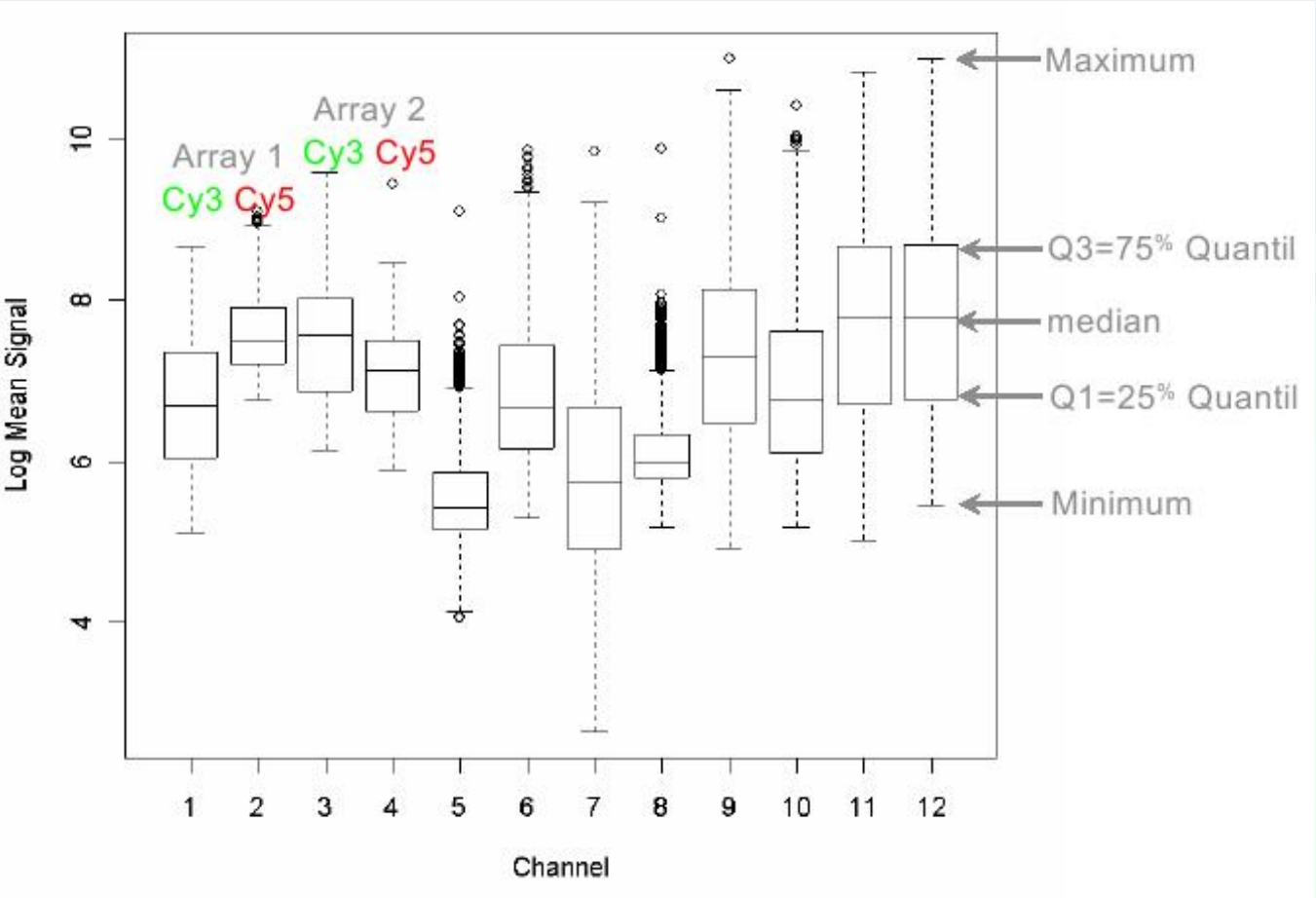


- $A = (log_2(R) + log_2(G))/2$ (**A**ddition, mean intensity)
- $M = log_2(R) - log_2(G)$ (**M**inus, differential expression, log fold change)
- Fit curve by Lowess or Loess normalization.

cDNA array: Intensity Boxplot



cDNA array: Intensity Boxplot



Distributions adjusted by median centering or quantile normalization.

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

- $M = log_2(R) - log_2(G)$

$M < 0$ 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

- Absolute value of of M is indicator for differential expression

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

- $M = log_2(R) - log_2(G)$

$M < 0$ 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

- Absolute value of of M is indicator for differential expression

- m replicates: Mean intensity $\overline{M} = \frac{1}{m} \sum_{i=1}^{m} M_i$
- Problem: Variance of the M-values not considered

Ranking the genes - $|T|$, p and B

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

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

Ranking the genes - $|T|$, p and B

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 σ
  - Problem: Large T value can also be caused by low σ
  - With small sample size σ cannot be well estimated $\rightarrow$ **moderated T-statistic** (variances are borrowed from other genes)

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

- Must be adjusted for multiple testing

Ranking the genes - $|T|$, p and B

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 σ
  - Problem: Large T value can also be caused by low σ
  - With small sample size σ cannot be well estimated $\rightarrow$ **moderated T-statistic** (variances are borrowed from other genes)

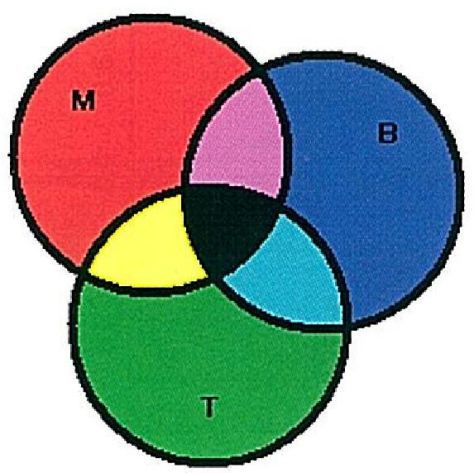
P-value probability that a $|T|$ is larger or equal to the observed $|T|$, while the null hypothesis is true

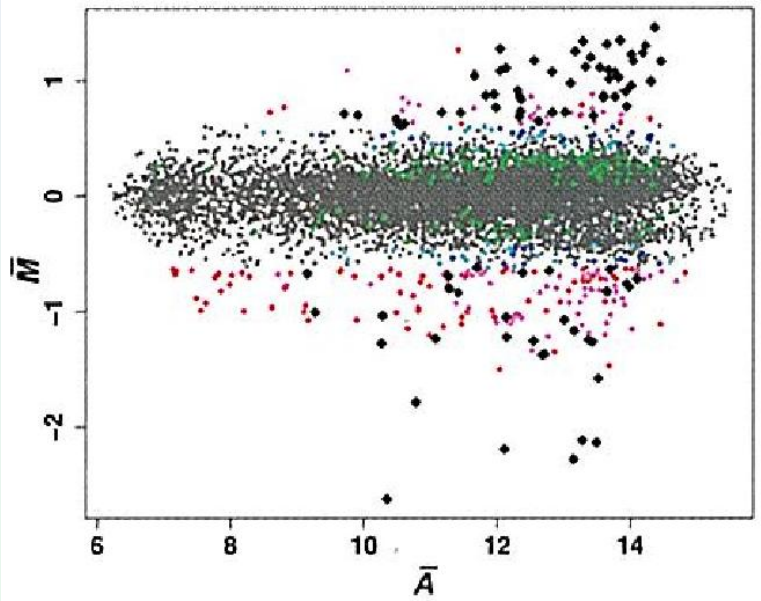
- Must be adjusted for multiple testing

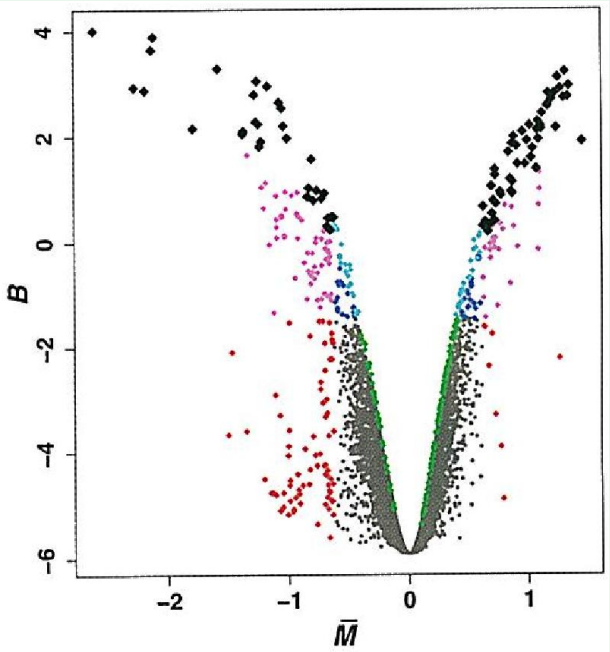
BEB Bayes empirical Bayes: Posterior probabilities for differential expression (log odds)

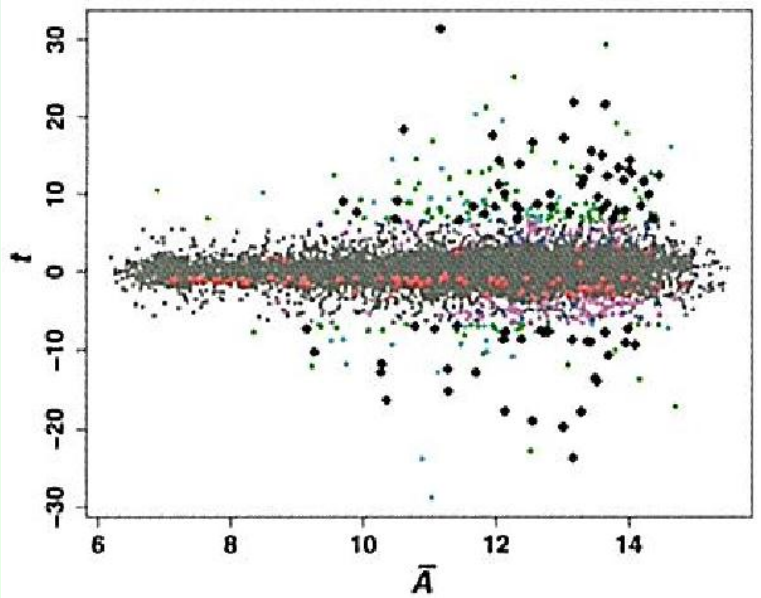
- Estimated variables are used for moderated T-statistic

Example









Annotation

- Previous analyses are done on the level of probes or tags
- Now: include function information
- First step: Find corresponding genes

Microarray	SAGE
Probe-to-gene-mapping	Tag-to-gene-mapping
Annotation data packages for specific platforms (e.g. Affymetrix) in Bioconductor, e.g.`annotate`	Mapping by 'virtual' transcript, e.g. SageGenie `http://cgap.nci.nih.gov/SAGE/AnatomicViewer`

Gene sets

Use the functional information (meta-data, annotations) available for the genes 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

$\longrightarrow$ 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

- Gene set Enrichment Analysis (GSEA)
- Genes ranked by a measure for differential expression (e.g. fold change, $|T|$, B), but no cutoff needed

Rank-based Gene-Class Testing

- Gene set Enrichment Analysis (GSEA)
- Genes ranked by a measure for differential expression (e.g. fold change, $|T|$, 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:

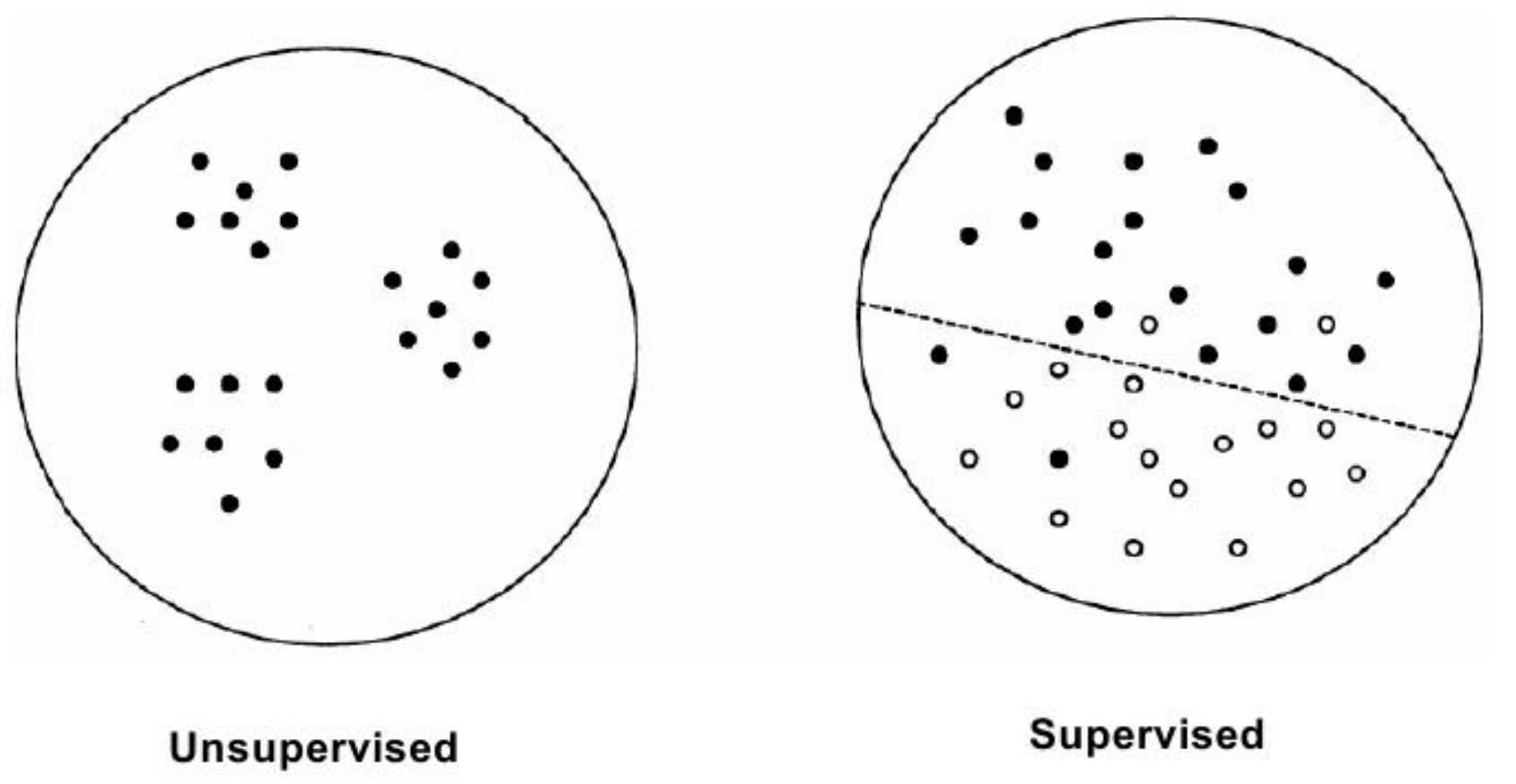
Gene	Sample 1	...	m
1	Expres-		
⋮	sion		
n	values		

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 $\longrightarrow$ 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:

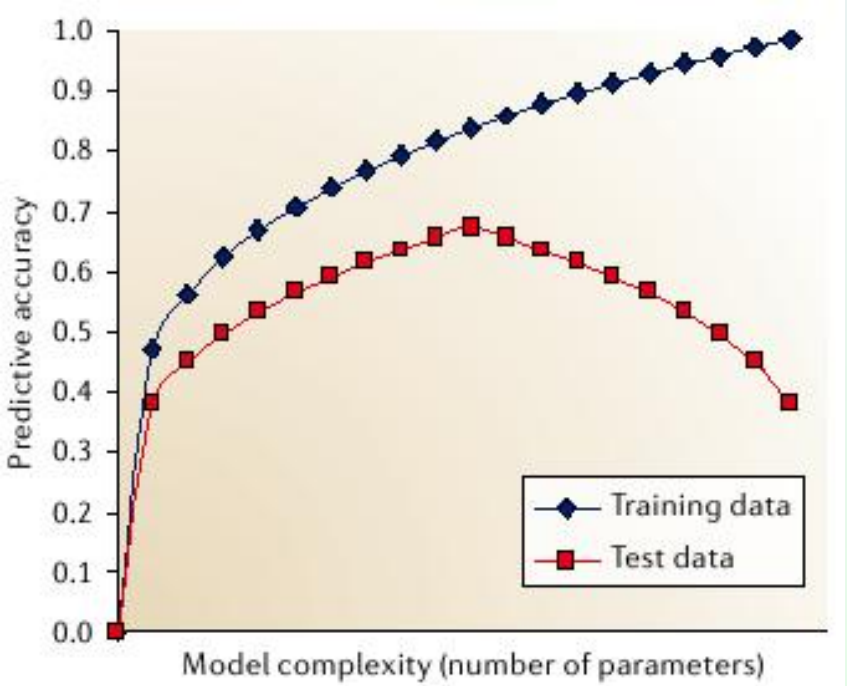
- With a mathematical expression, the objects are transfered in a space where they can be separated with a straight line

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

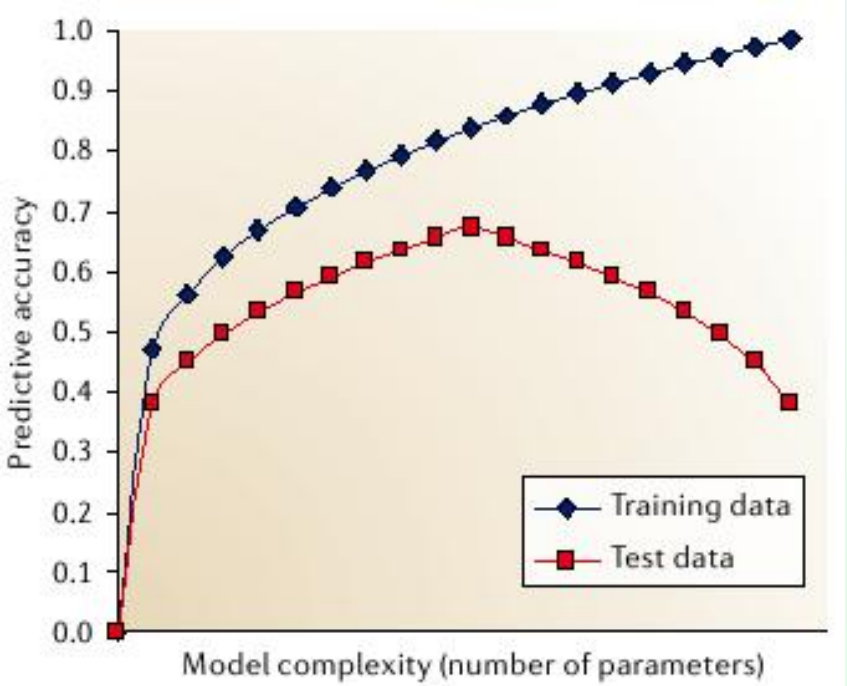


Validation

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 $\rightarrow$ 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

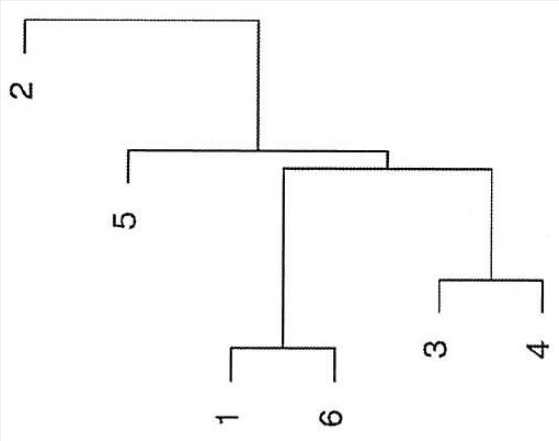
Also: Clustering of samples or two-sided clustering

Problems:

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

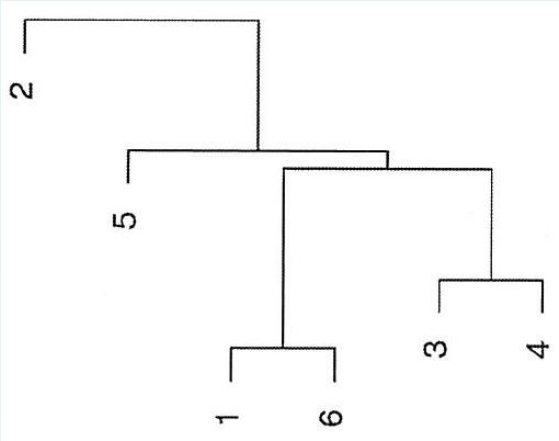
Clustering algorithms

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

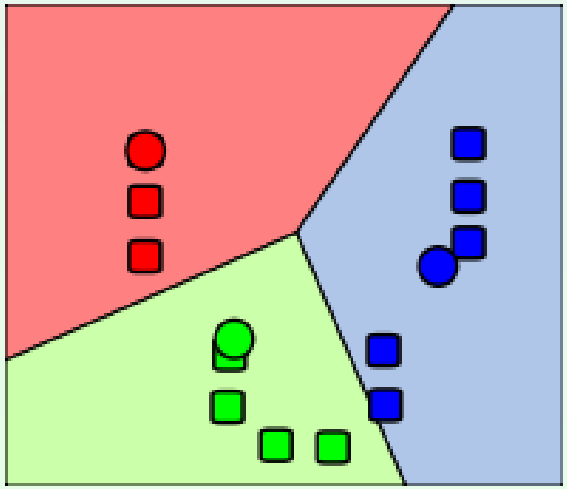


Clustering algorithms

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



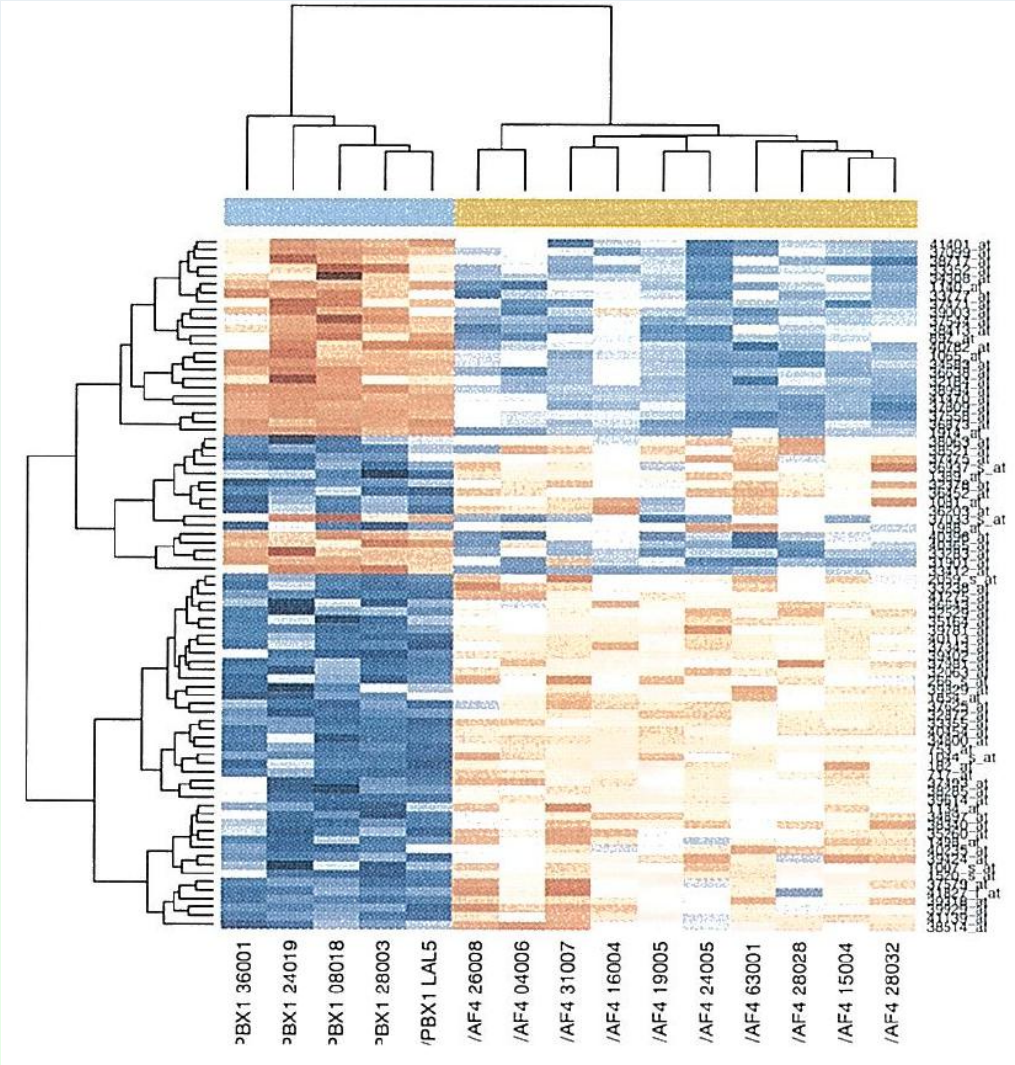
- **Partition clustering**
- k-means: k classes $\rightarrow$ class means $\rightarrow$ classification according to smallest distance $\rightarrow$ new classes $\rightarrow$ ...
- The classes are recomputed in every step



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
- San Ming Wang. 2006. Understanding SAGE data. *Trends in Genetics* 23: 42-50
- Robert Gentleman et. al. 2005 *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Springer
- Florian Hahne et. al. 2008 *Bioconductor Case Studies.* Springer