

Microarray data analysis

Yves Moreau and Janick Mathys

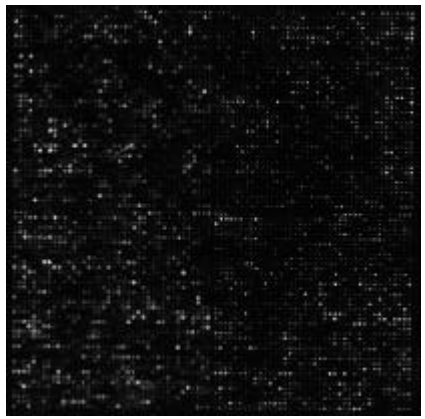
Master of Bioinformatics

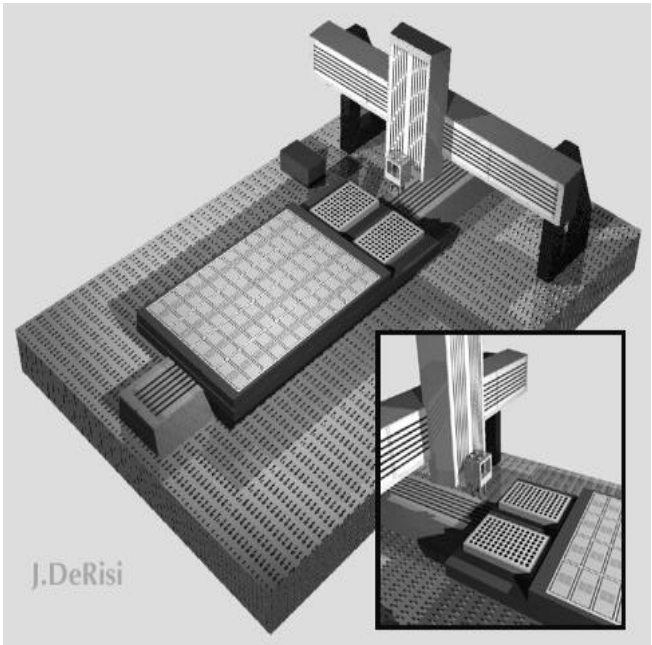
Katholieke Universiteit Leuven

2001-2002

Microarrays

- Genome-wide monitoring of gene activities by measurement of the levels of RNA transcripts
  - Massively parallel
  - Fully automated
  - Standardizable



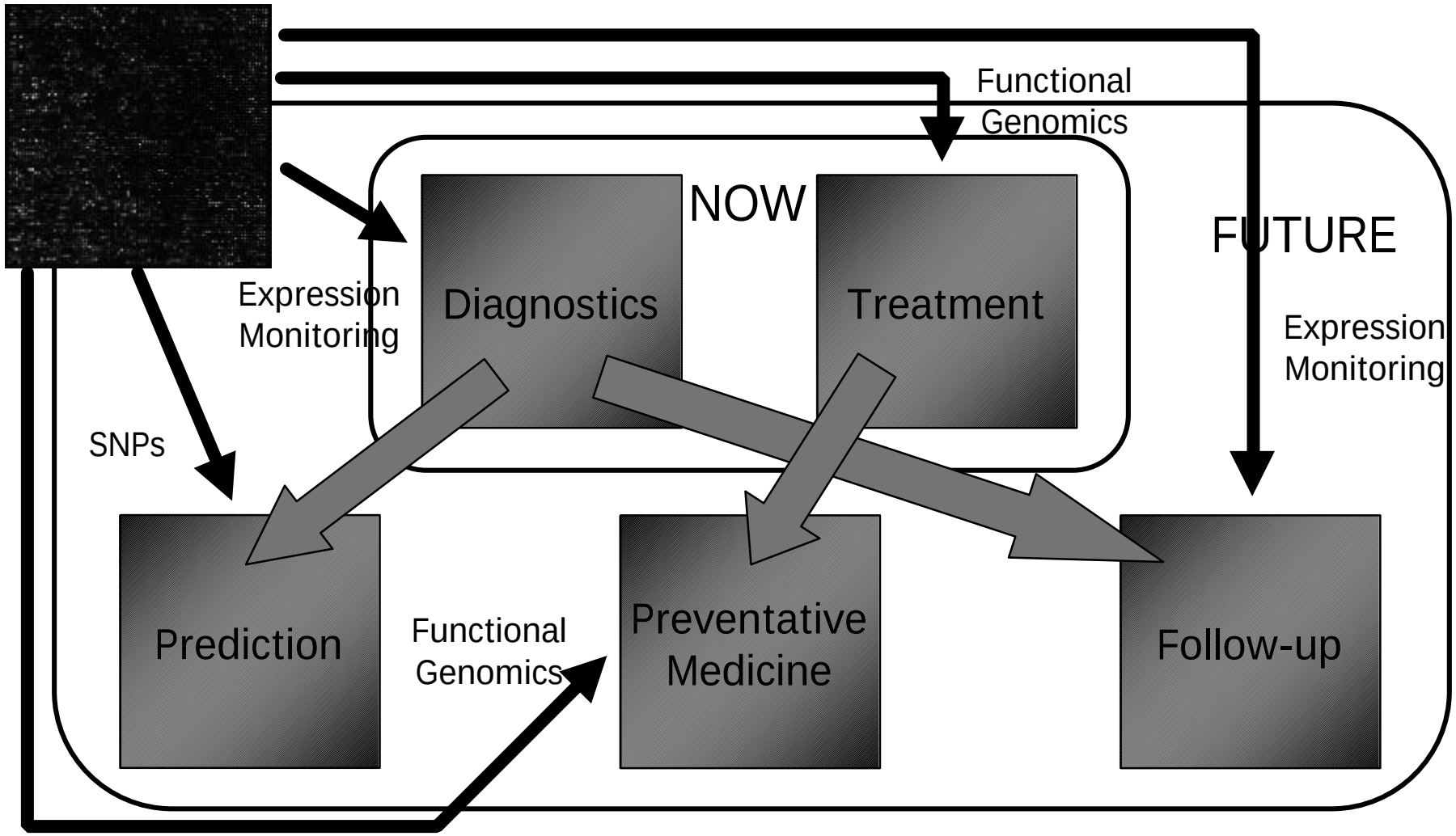


Applications

- Identification of regulatory mechanisms
- Modeling of genetic networks for function identification and transgenesis
- Drug development: study gene expression in
  - Disease
  - Model systems
  - Pathogens
  - Response to drug treatment
- Diagnosis

- Survey of Single Nucleotide Polymorphisms
- Genotyping for drug tayloring

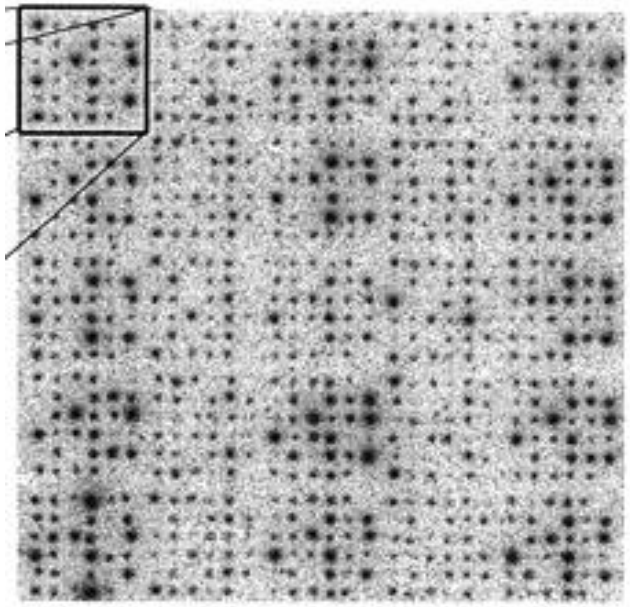
High-throughput genomics



Microarray technologies

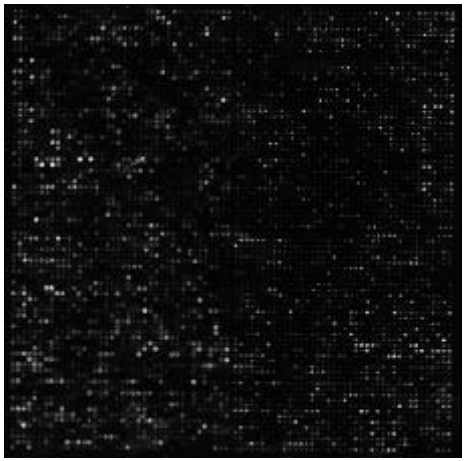
Filter macroarrays

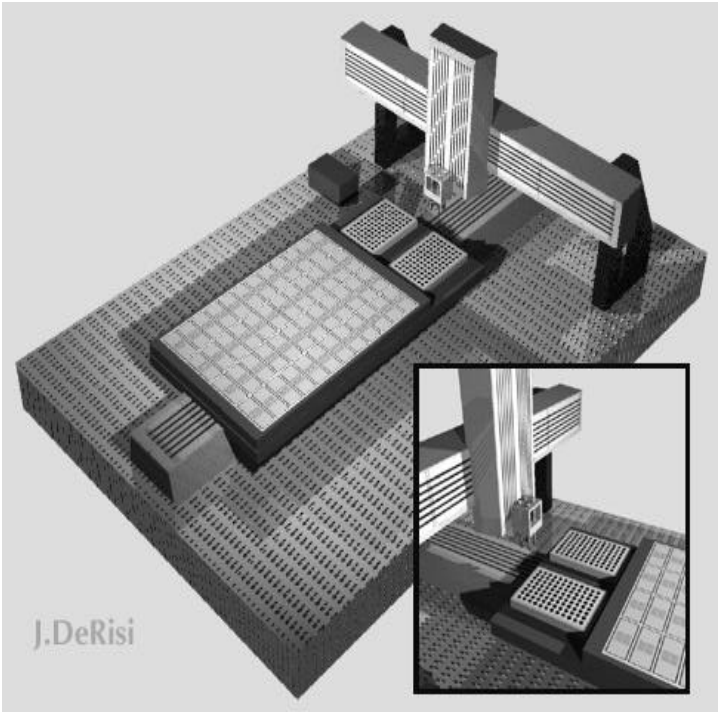
- Nylon membrane
- Spotted cDNA
- Sample is labeled by radioactivity (^{32}P or ^{33}P)



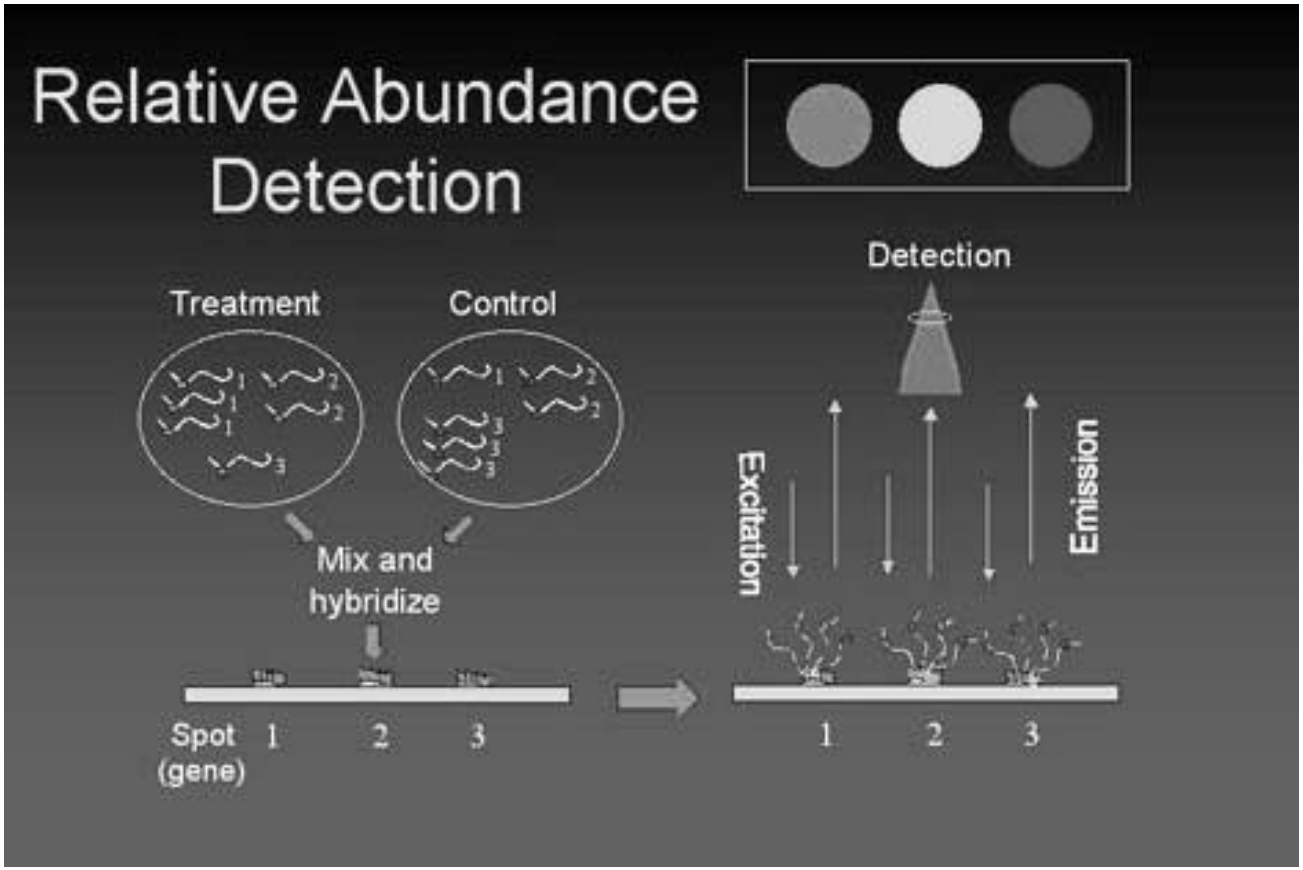
cDNA microarrays

- 5000 cDNAs (genes, ESTs) spotted on array
  - Glass slide
  - Cy3, Cy5 labeling of samples
  - Hybridization of test + control





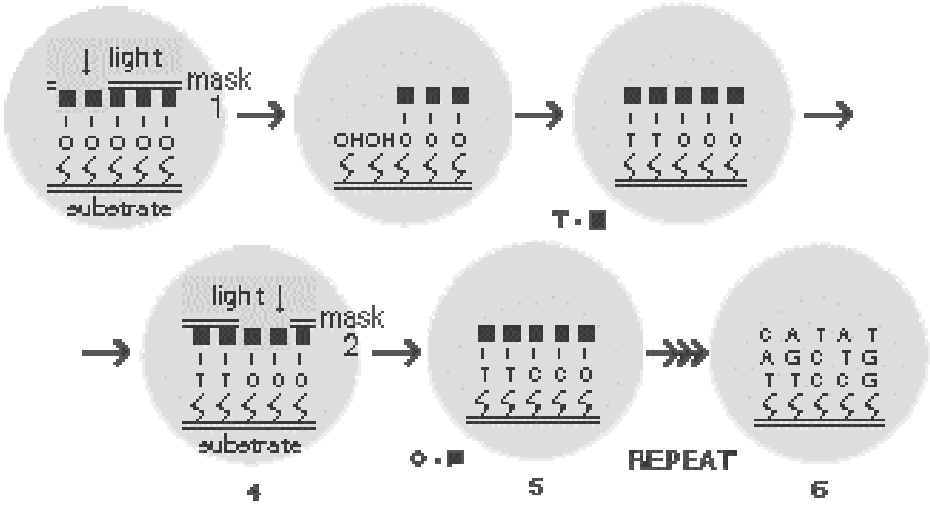
cDNA microarrays



DNA chips

- Silicon substrate
- In situ synthesis of oligonucleotides (25 bases) by photolithography
- Multiple probes per gene (match + mismatch)
- Ultra-high density possible





Inkjet arrays

- Glass slide
- Inkjet spotting
  - cDNA clones (no contact = better liquid handling)
  - OR in situ synthesis of long oligonucleotides (60 bases)
- Long oligos more specific than short oligos
- Possibility to study alternative splicing with long oligos

Relative vs. absolute expression

- Filters and oligonucleotide arrays
  - Single measurement per feature
  - Absolute expression level

- cDNA arrays
  - Two samples per experiment
    - Test
    - Control
  - Measurement is ratio of gene expression in test sample vs. control sample
  - Better reproducibility
  - More difficult to handle during analysis

Types of experiments

- Wild-type vs. mutant
  - Knock-out, conditional knock-out
  - Overexpression construct, inducible overexpression
  - Detection of the targets of a transcription factor
- Groups of patient samples
  - Different types of tumors
- Multiple conditions
  - Expression patterns in the presence of drugs or toxins
- Time-course experiment
  - Response to a signal
    - Stress
    - Development

Image analysis
in cDNA microarrays

Image analysis

- Raw images
  - Grey - scale images for red and green channels separately

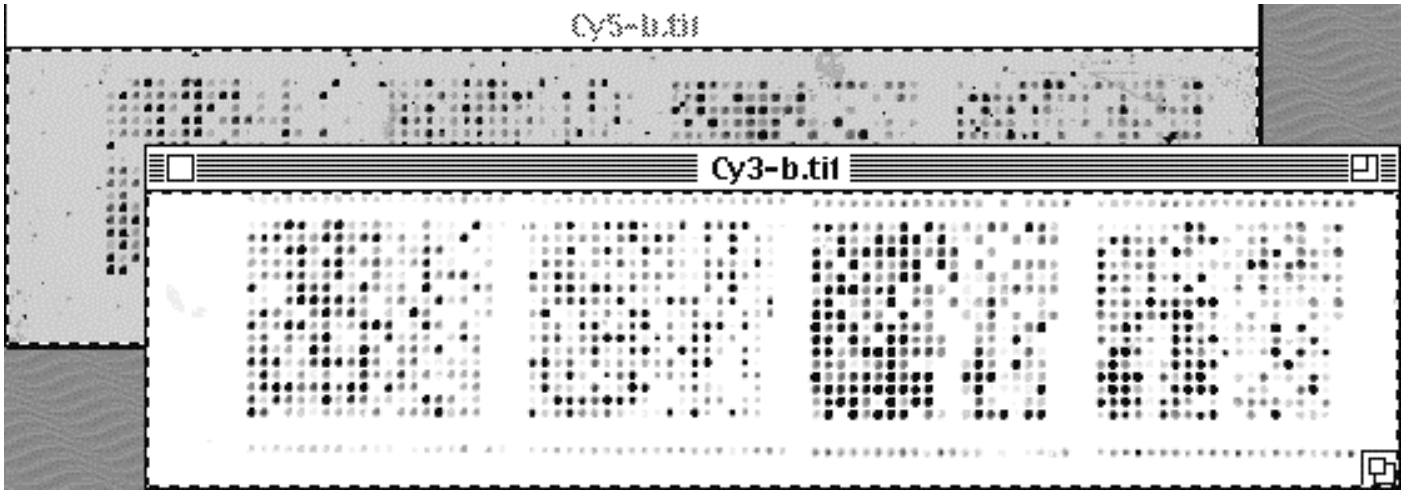


Image analysis

- Superimposed two channels to form color image
  - Red spot: gene was only expressed in test sample
  - Green: gene was only expressed in control sample
  - Yellow: gene was expressed both in test and in control sample
  - Black: gene was neither expressed in test nor in control sample

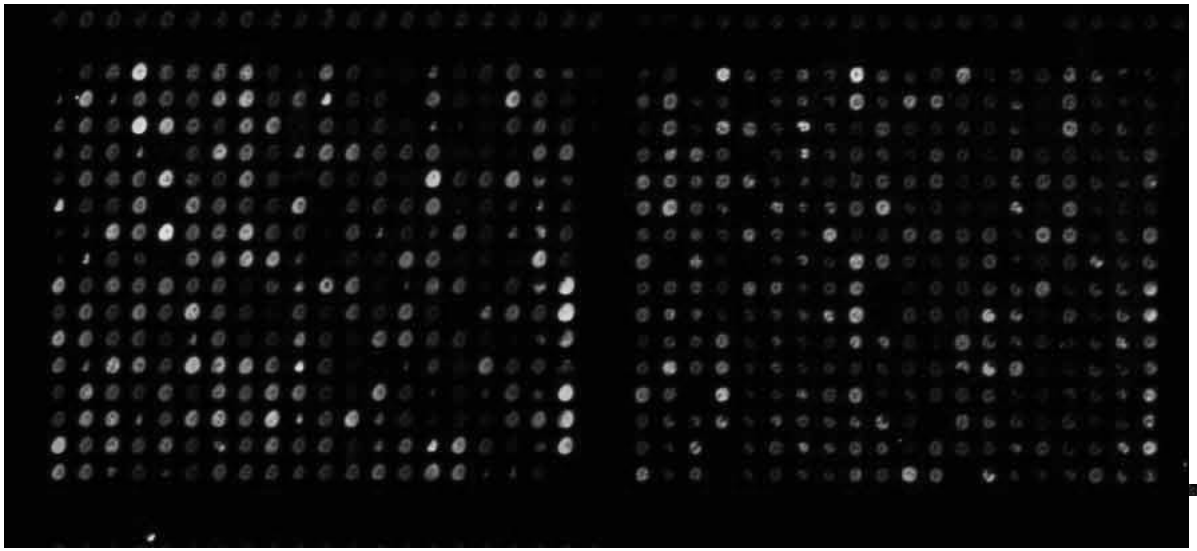
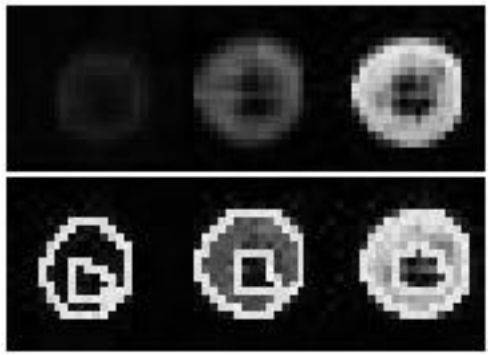


Image analysis



- Spot detection
  - Spots are not perfectly circular
  - Thresholding
    - Fixed threshold method
      - Threshold T is derived from local background mean intensity m and standard deviation σ by the relationship $T = m + 3\sigma$
      - Problems due to variability of background and target signal, particularly with weak signals (frequent finding in cDNA array experiments)
    - Man-Whitney method
      - Take sample pixels from background and perform rank-sum hypothesis test on target pixels
  - $\Rightarrow$ Treat red channel and green channel separately

Image analysis

- Background intensity extraction
  - Background is not uniform ⌫ extract local background intensities
  - Gray level histogram
    - Mean local background intensity
    - Standard deviation of the local background intensity
  - ⇒ Treat local red background intensity and local green background intensity separately

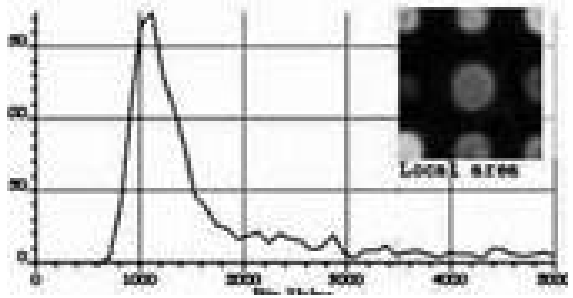
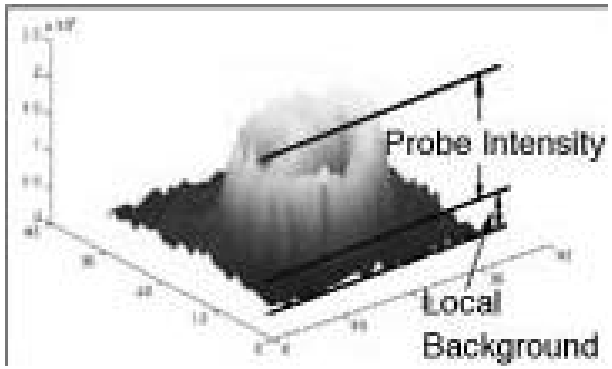


Image analysis

- Spot intensity extraction
  - Unite target regions detected for red and green channel
  - Probe intensity = average gray level within target region for R and G separately
  - Subtract local background values from reported probe intensities for R and G separately
    - R intensity = raw R intensity
      - local R background
    - G intensity = raw G intensity
      - local G background



- Fluorescent intensities = significant if average spot intensity is at least 2 standard deviations higher than corresponding average background intensity

Data analysis

Expression ratios

- Calculation of the relative (R/G) expression ratios
  - Mean or median R intensity / mean or median G intensity
  - Mean or median of ratios of R/G intensities from every pixel location
  - Linear regression slope of R-G gray-values from every pixel location

- !!! Ideally all genes should be expressed in the control sample (G)
  - Not true in practice
  - Intensity G = 0 $\Rightarrow$ ratio = intensity R / 0 (! missing value)
  - Reverse labeling can help here (adjust for variations between R and G channel)

Preprocessing

- Rescaling of different slides (experiments)
  - Different labeling quality between slides can result in systematic bias in the expression levels
  - Linear regression (or smoothing curve) of all genes present in both slides (or housekeeping genes)
  - Rescaling factor = 1 / slope of regression; new intensity = old intensity * rescaling factor
- Filtering of irrelevant genes
  - Retain genes with R/G ratios ≥ 2 or 3 in at least two experiments
  - Remove genes with missing values in x% of the experiments
  - Remove low variance / standard genes (housekeeping genes)
- Log_2 transformation of the R/G ratios: log scale is more intuitive

	2-fold induction	*constant expression*	*2-fold repression*	
R/G	2	1	1/2	...
log_2(R:G)	1	0	-1	...

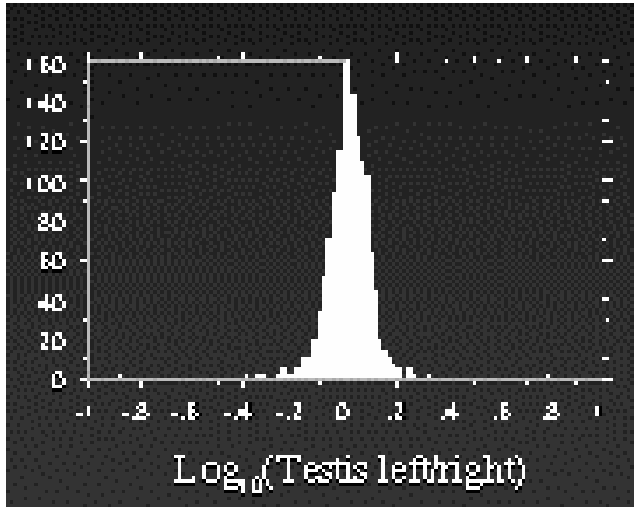
Analysis of expression data

- Single-slide analysis
  - Identify genes that are expressed in control and not in test sample
  - Identify genes that are expressed in test sample only
  - $\Rightarrow$ condition-specific genes

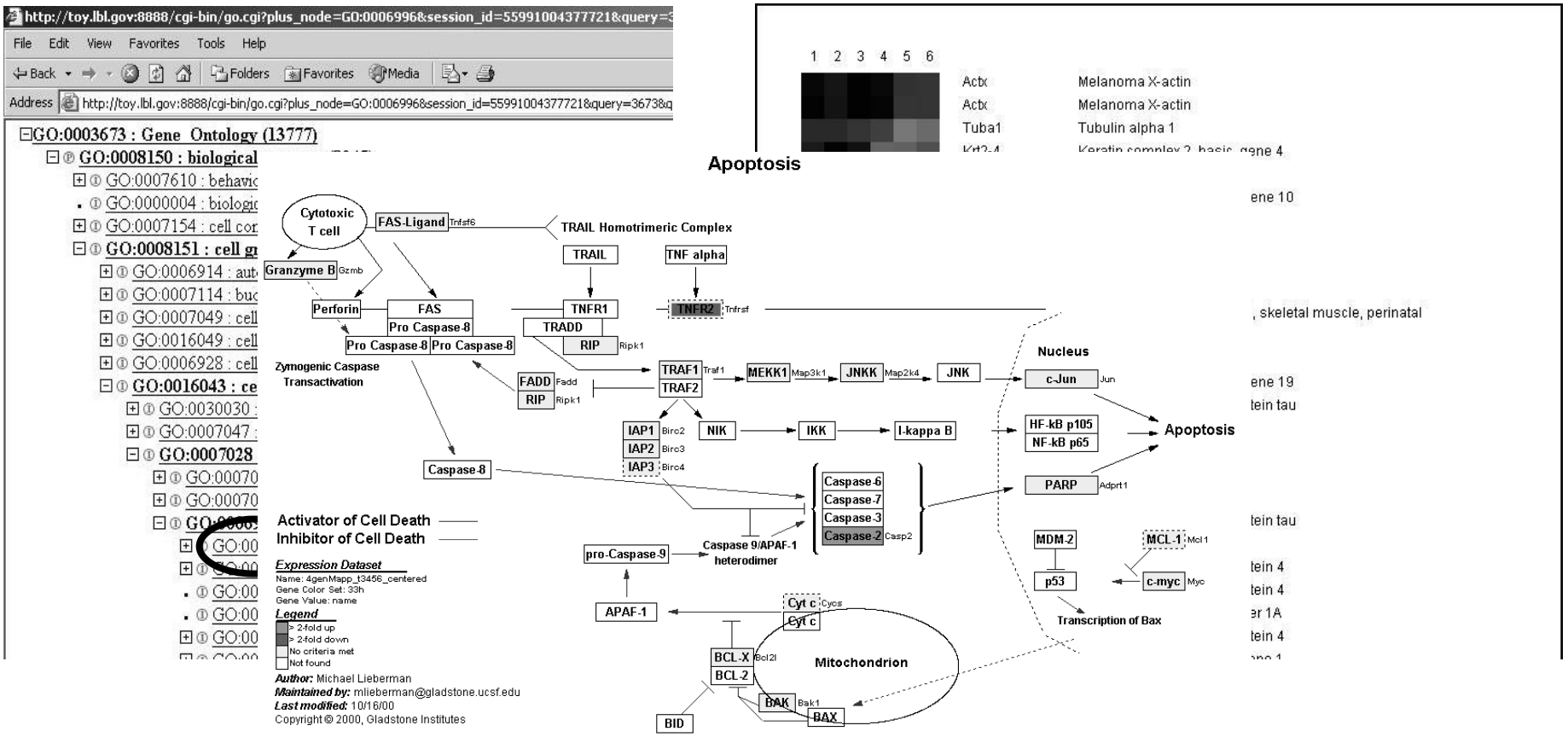
- Analysis of multiple slides
  - Cluster analysis of gene expression profiles
  - Partitions samples or genes into well-separated and homogeneous groups

Single-slide analysis

- Which genes are considered up- or downregulated ?
- VIB-MAF: mouse arrays, duplicate spots on the same slide
  - Frequency distribution: ratio of left spot against ratio of right spot for each gene
  - Average $m = 1$ as expected
  - Variation around m less than threefold
  - $\Rightarrow$ Less than threefold differential expression levels may not be statistically reliable

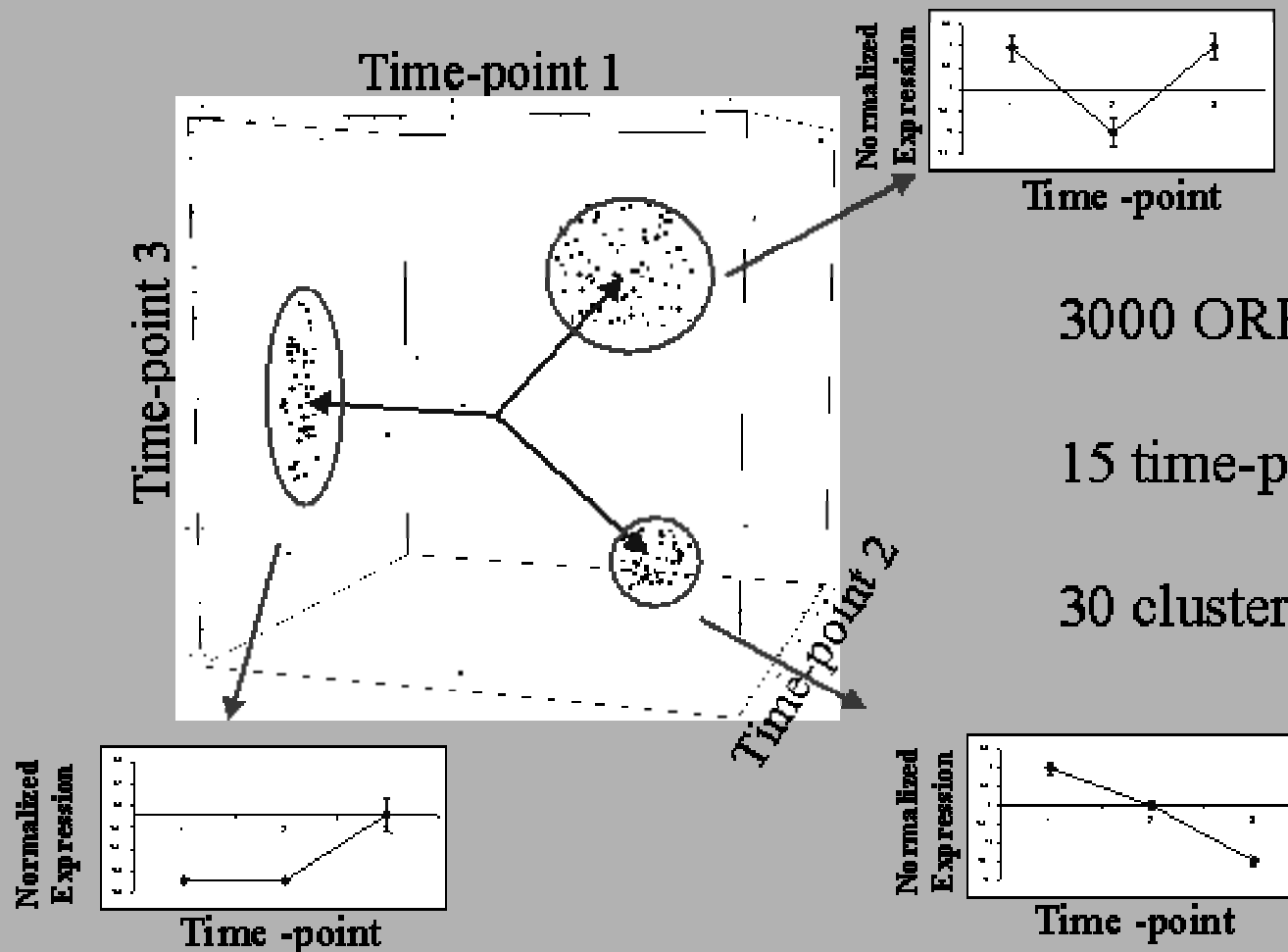


Data exploration



Clustering

Identifying prevalent expression patterns (clusters)



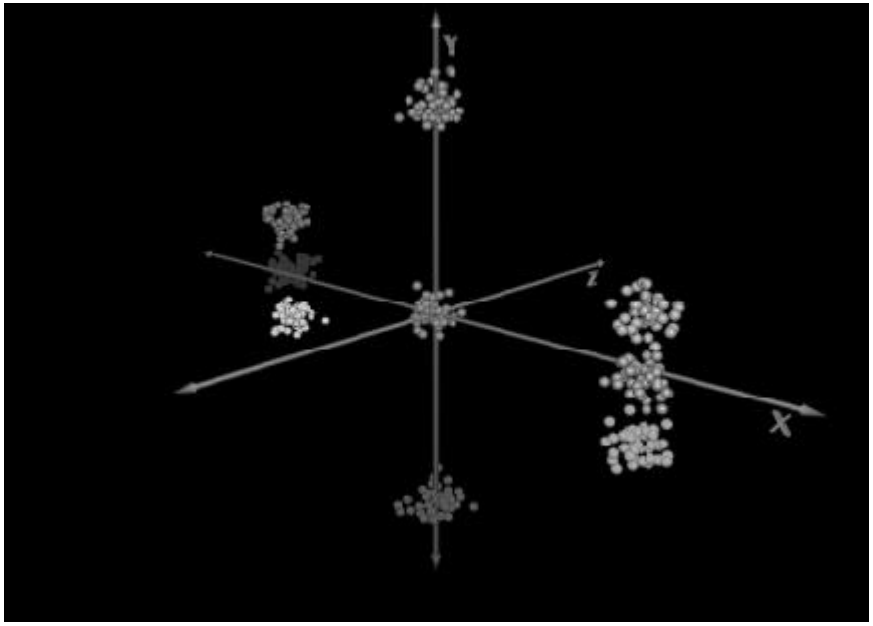
3000 ORFs

15 time-points

30 clusters

Principal component analysis

- PCA detects the directions that capture the most information about the data



Principal component analysis

- PCA is performed by a linear transformation of the data set based on the Singular Value Decomposition (SVD)
- PCA can be applied to genes as well as samples
  - “Eigengenes”
  - “Eigensamples”
- Successive principal component capture less and less information about the data
- We can truncate the representation of the data to a limited number of principle components = dimensionality reduction

Cluster analysis

- Preprocessing of the data
  - Mean (or median) centering: mean (or median) of ratios = 0
  - Normalization : standard deviation of ratios = 1
  - Reduces noise caused by differences in RNA yield, labeling efficiency and image analysis
- Cluster algorithms
  - Hierarchical clustering
  - K-means clustering
  - Self-Organizing Maps
  - Adaptive quality-based clustering

Hierarchical clustering

- Evaluate pairwise similarities between genes
- For 2 genes X $(x_1,x_2,...,x_m)$ and Y $(y_1,y_2,...,y_m)$, the Pearson correlation coefficient (= similarity) is

$$r = \frac{1}{N} \sum_{i=1}^{N} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{s_X \quad s_Y}$$

- Euclidian distance = 1 - similarity
- $\Rightarrow$ Distance matrix :

	gene 1	*gene 2*	*gene 3*	*...*	*gene n*
gene 1	0			...	
gene 2	Dist(gene1,gene2)	0		...	
gene 3	Dist(gene1,gene3)	Dist(gene2,gene3)	0	...	
...	...	...	...	...	
gene n	Dist(gene1,gene n)	Dist(gene2,gene n)	Dist(gene3,gene n)	...	0

Hierarchical clustering

- Find pair of genes with shortest distance (e.g., gene 2 and gene 3)
- Substitute ratios for gene 2 and gene 3 in data matrix by average ratios for gene 2 and gene 3 combined
- $\Rightarrow$ Data matrix:

	exp1	***exp2***	***...***	***expm***
gene 1	$\log_2(R_1/G_1)$	$\log_2(R_1/G_1)$	...	$\log_2(R_1/G_1)$
gene 2 + 3	$\frac{\log_2(R_2/G_2) + \log_2(R_3/G_3)}{2}$	$\frac{\log_2(R_2/G_2) + \log_2(R_3/G_3)}{2}$	...	$\frac{\log_2(R_2/G_2) + \log_2(R_3/G_3)}{2}$
...	...	...	...	...
gene n	$\log_2(R_n/G_n)$	$\log_2(R_n/G_n)$	...	$\log_2(R_n/G_n)$

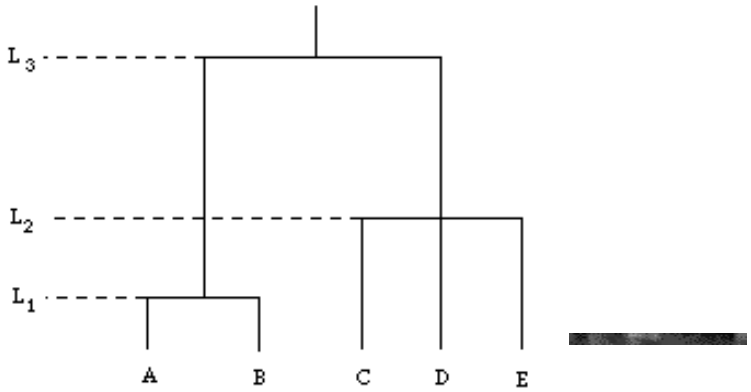
- Recalculate pairwise similarities

Hierarchical clustering

- ⇒ Distance matrix:

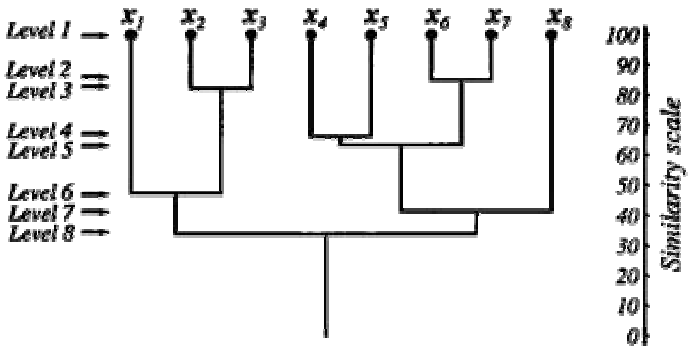
	gene 1	*gene 2 + 3*	*...*	*gene n*
gene 1	0		...	
gene 2 + 3	Dist(gene1,gene2+3)	0	...	
...	...	...	...	
gene n	Dist(gene1,gene n)	Dist(gene2+3,gene n)	...	0

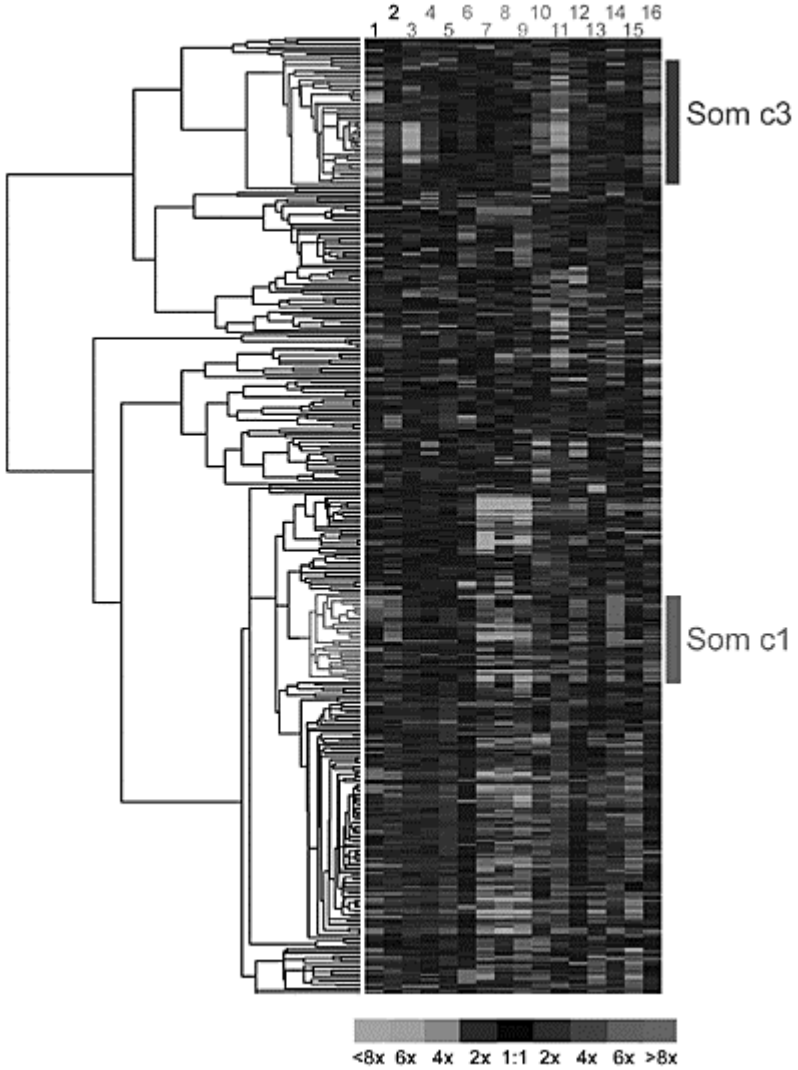
- Repeat until distance matrix contains 1 element
- Visualization: dendrogram (= tree)
- Length of each branch is indicative of distance between clusters
  - Average distance between centers
  - Maximal distance between clusters
  - Minimal distance between clusters



Hierarchical clustering

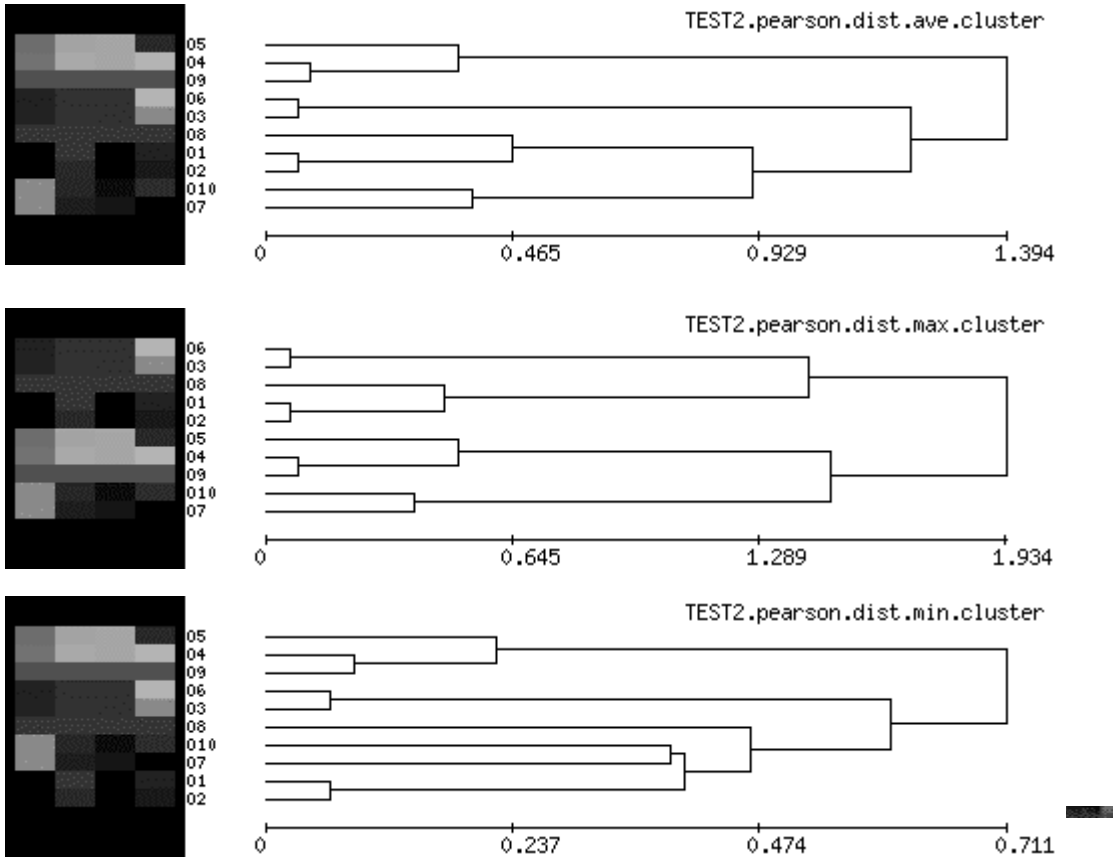
- Visual determination of clusters



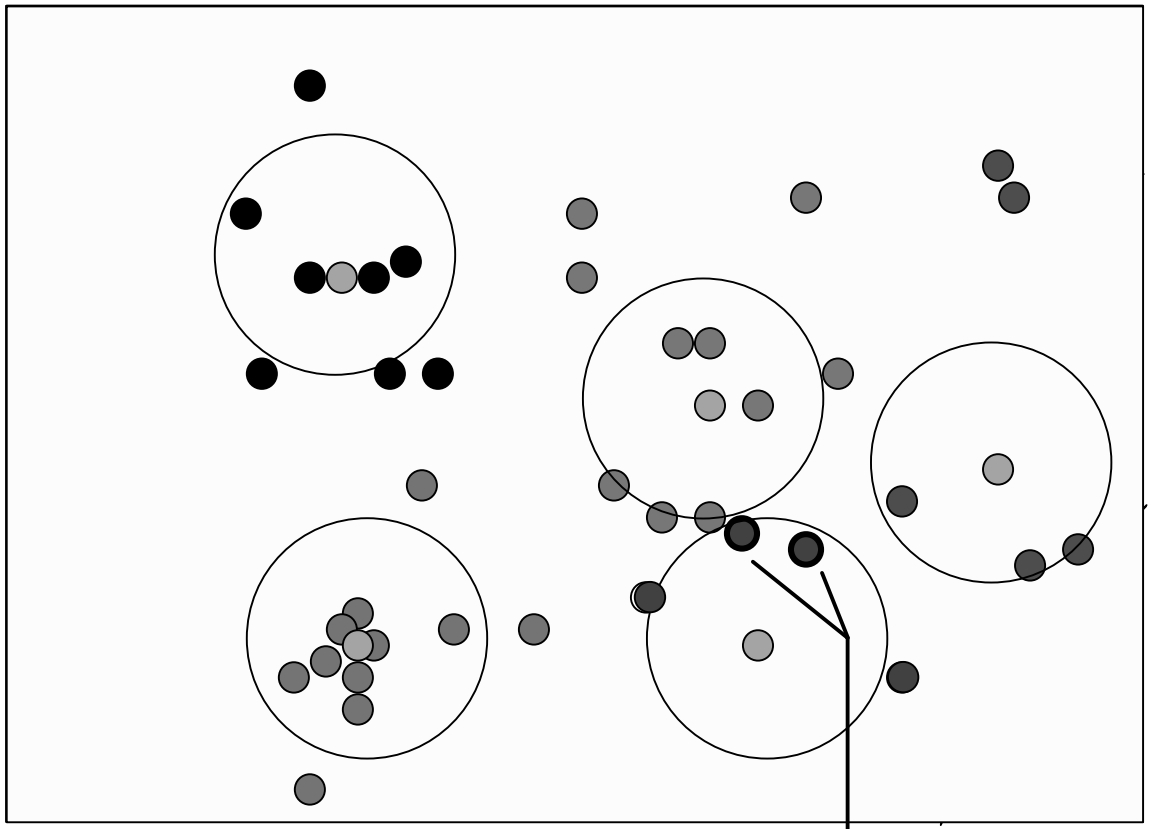


Hierarchical clustering

- Distance measure influences outcome of clustering
  - Lack of robustness
  - Nonunique



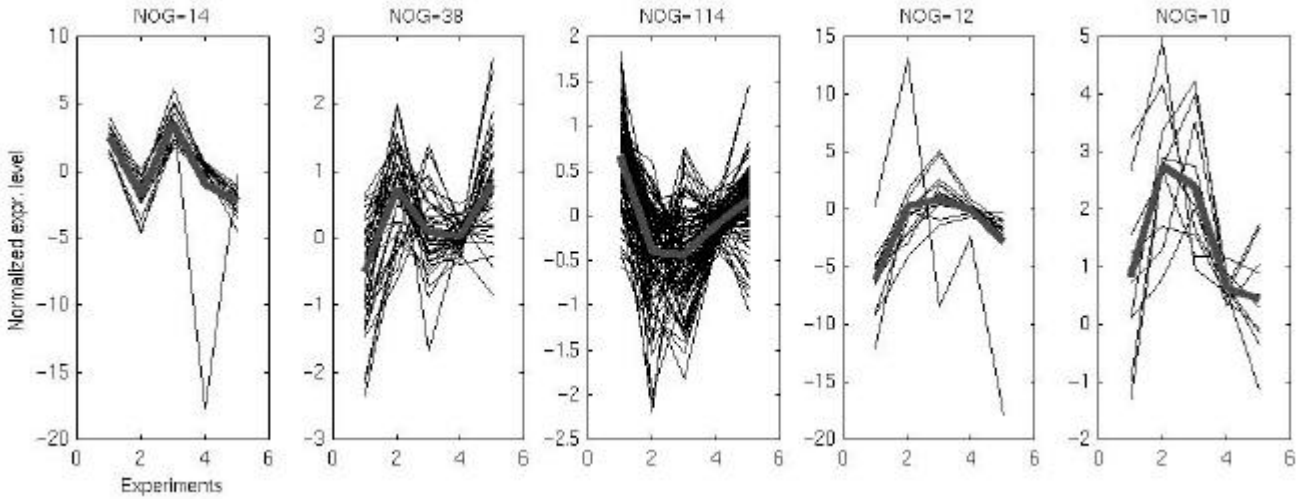
K-means clustering



1. Predefined number of clusters = 5; initialisation: randomly choose cluster centers (red points)
2. Attribute each point (gene) to cluster with closest center
3. Recalculate cluster centers = mean expression profile of genes in cluster
4. Repeat the whole process until centers remain stationary

K-means clustering

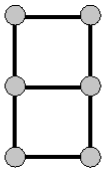
- Visualization: gene expression profiles per cluster



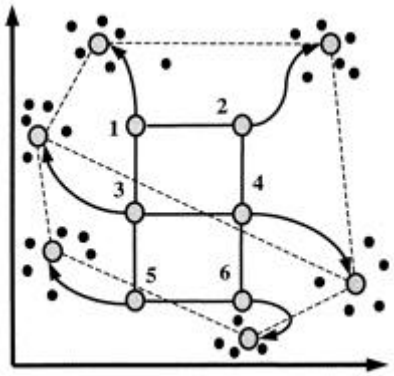
- Drawbacks
  - Number of clusters is predefined
  - All genes are assigned to a cluster
  - Lack of robustness
  - Nonunique

Self-organizing maps

- Choose a geometry of nodes (e.g., 3 x 2 rectangular grid)



- Nodes are mapped into m-dimensional space
- A data point p is selected randomly and the nodes are moved in the direction of p
- Node closest to p (Np) is moved most
- Other nodes are moved by smaller amounts ~ distance to Np in initial grid
- 20.000 - 50.000 iterations
- Nodes = cluster centers and neighbouring nodes define related clusters



Self-organizing maps

- Visualization: gene expression profiles per cluster

