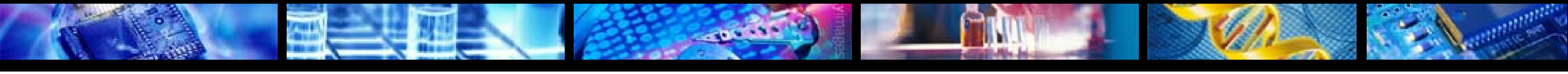


Computational Biology LU 2014

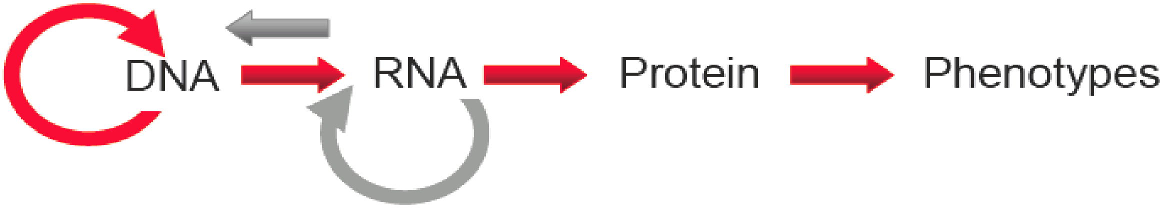
Exercise 2 - Microarray Analysis

Bioinformatics Group
Institute for Knowledge Discovery
Graz University of Technology
http://genome.tugraz.at
Petersgasse 14, A-8010 Graz

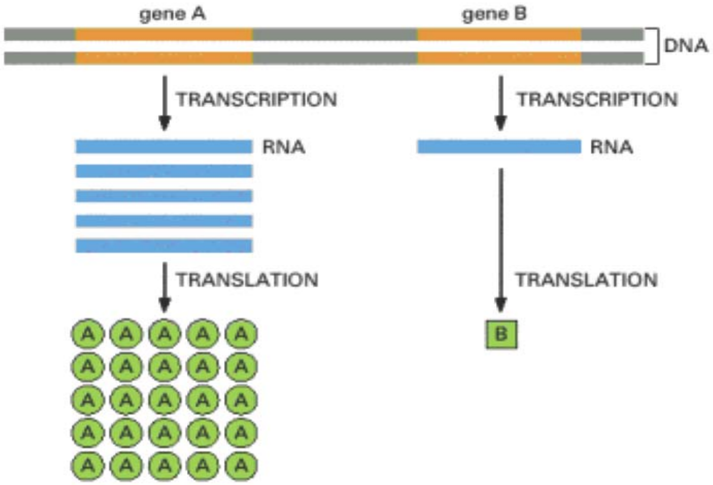
Outline

- Microarrays
  - Central Dogma of Molecular Biology
  - Omics
  - Two colour microarrays
  - Clustering
- Exercise 2 – help functions

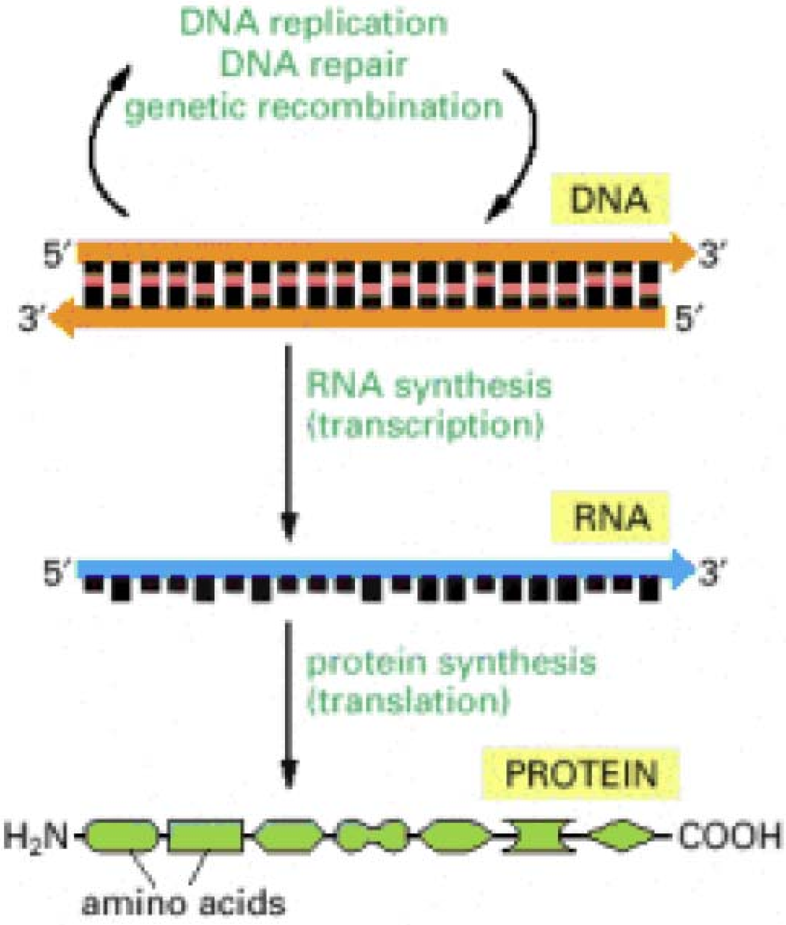
Central Dogma of Molecular Biology



DNA → RNA → Protein → Phenotypes



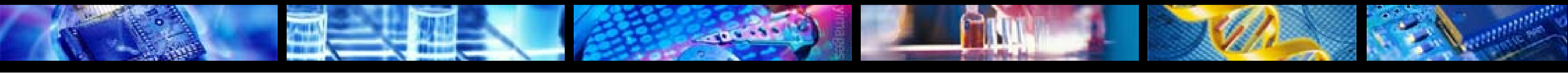
OMICS



→ GENOMICS

→ TRANSCRIPTOMICS

→ PROTEOMICS



Transcriptomics

- Transcriptomics = study of the transcriptome
(parts of the genome that are transcribed)

- A gene is expressed when it is transcribed into RNA

- Genomes within and across species might be very similar
- The genes that are expressed is what makes the difference between individuals or between species

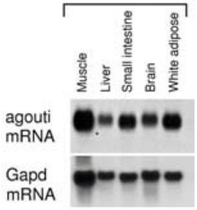
Transcriptomics

- The genome is static, the transcriptome is dynamic
- One gene has always the same sequence (except by mutations) but the same gene is differently expressed (at different rates) in different situations
- With transcriptomics we try to obtain a snap-shot of the cell transcriptional activity at a given time

RNA expression profiling

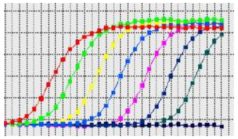
- Northern bloting

  - semi-quantitative
  - few genes



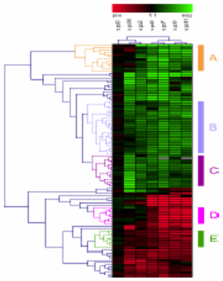
- Real time RT-PCR (qPCR)

  - medium throughput



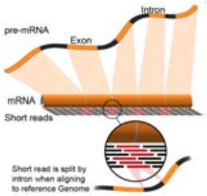
- Microarray analysis

  - high throughput
  - 10.000-500.000 elements per chip

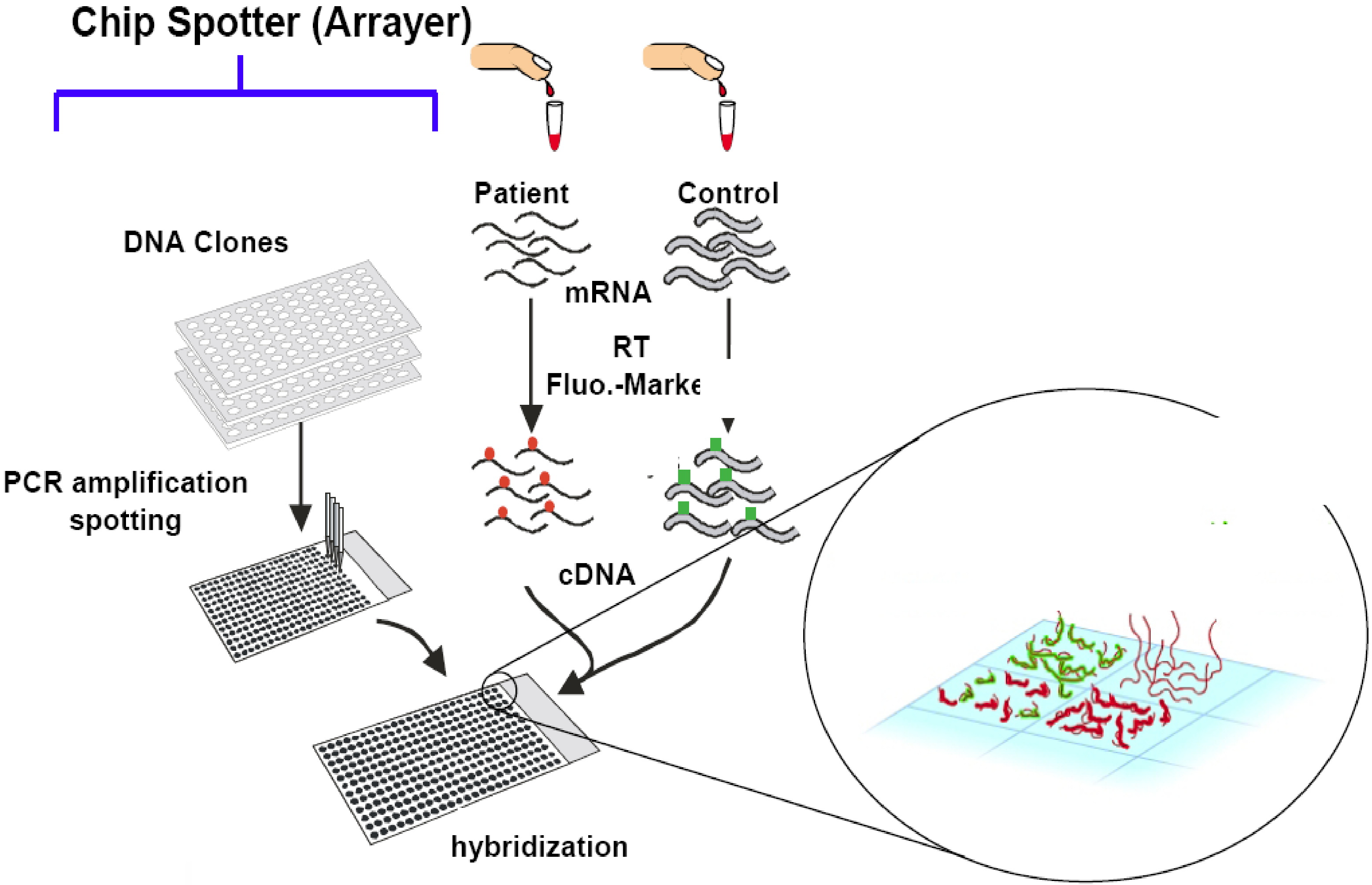


- RNA seq

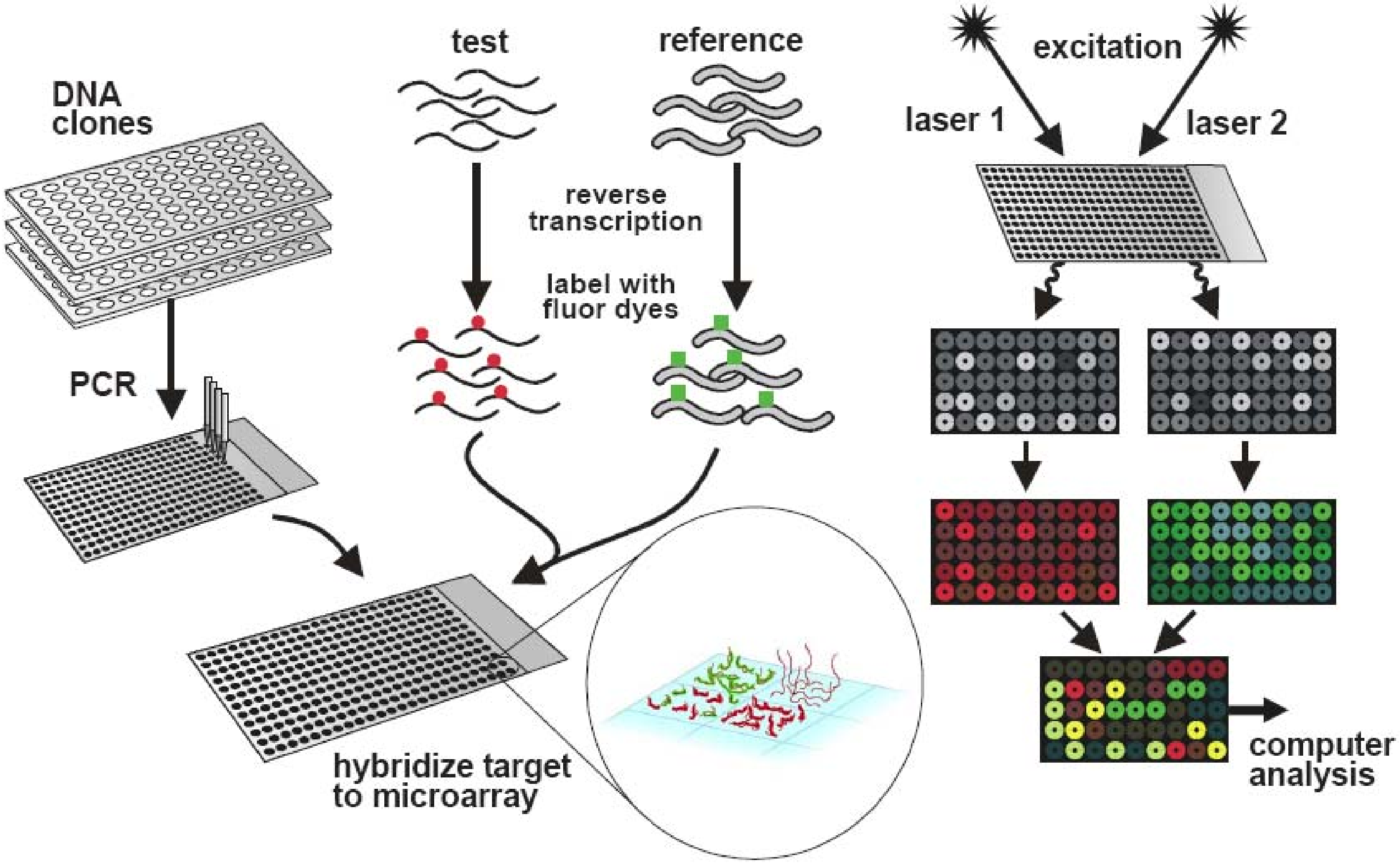
  - high throughput
  - deep sequencing (short reads 25bp)


Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

Two-color microarrays



Two-color microarrays



Analytical pipeline

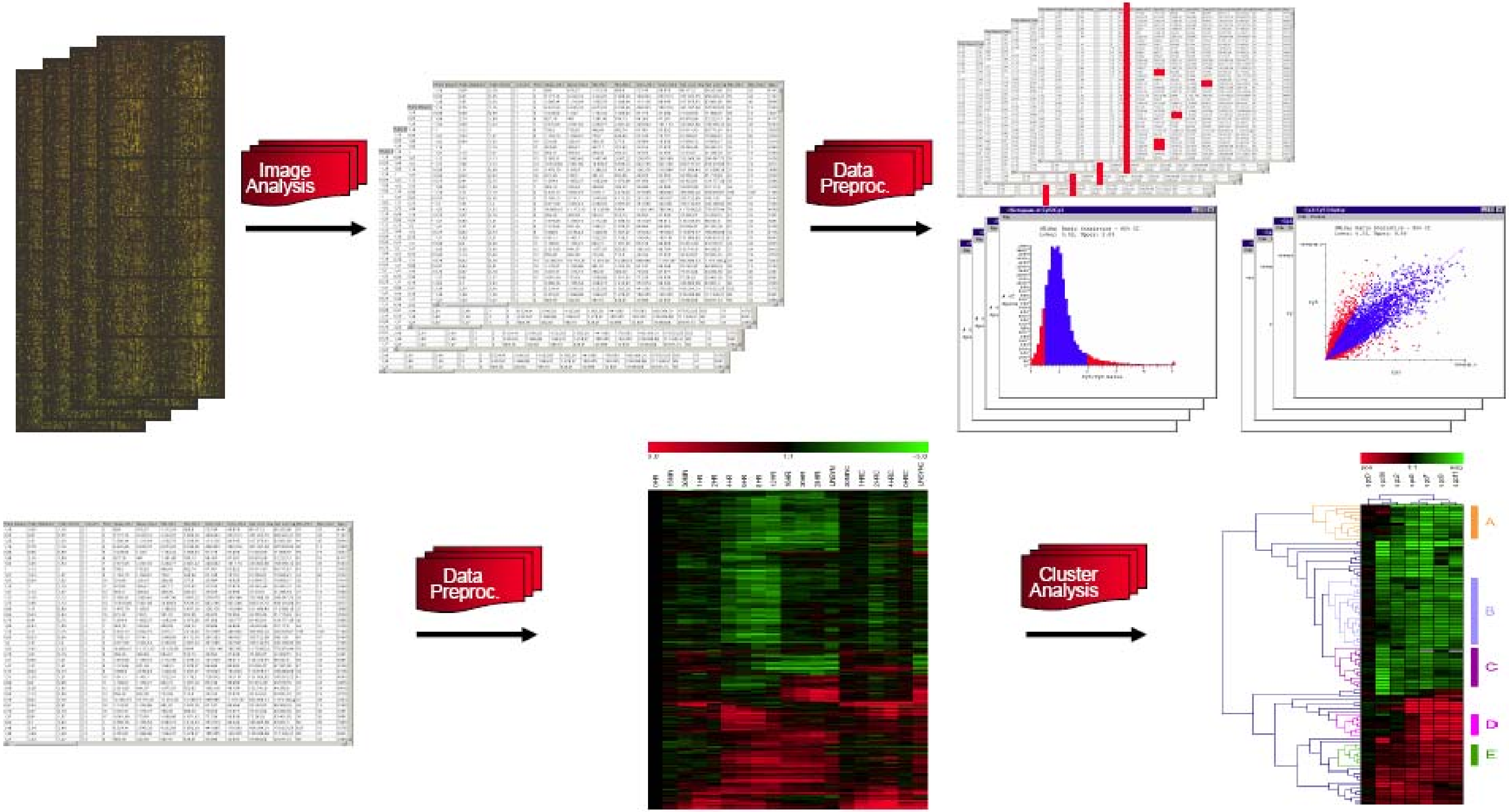
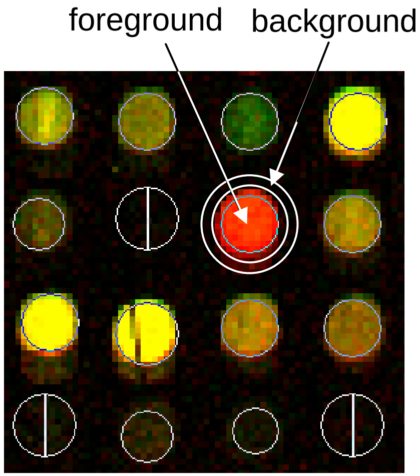


Image analysis and background correction

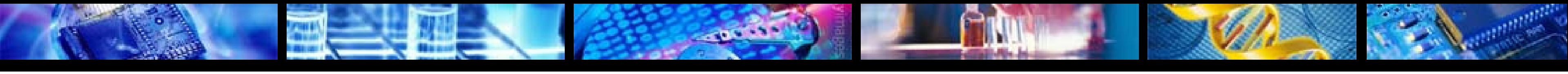
- Software available (GenePix, ImaGene, Agilent)
- Steps:
  - – Gridding, assigns coordinates and gene information to the different spots
  - – Segmentation: Foreground vs background
  - – Intensity extraction



G_k=F532 mean – B532

R_k=F635 mean – B635

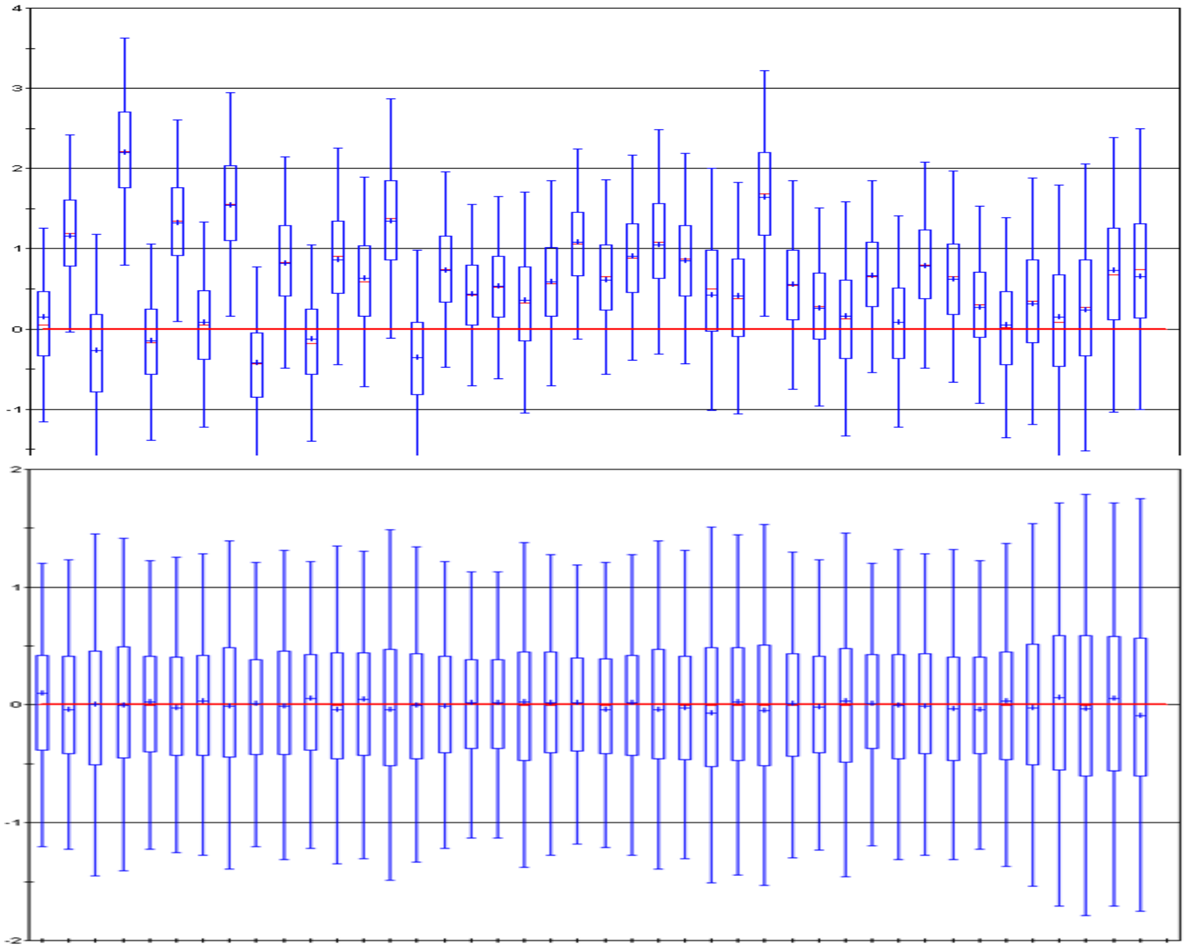
A lot of other parameter:
F635 % Sat., Flags, B532 SD,…



Normalization

- Removal of all sources of systematic non-biological variability and the reduction of the random errors.
- Basic assumption is that most of the genes are not changing their expression during the studied process

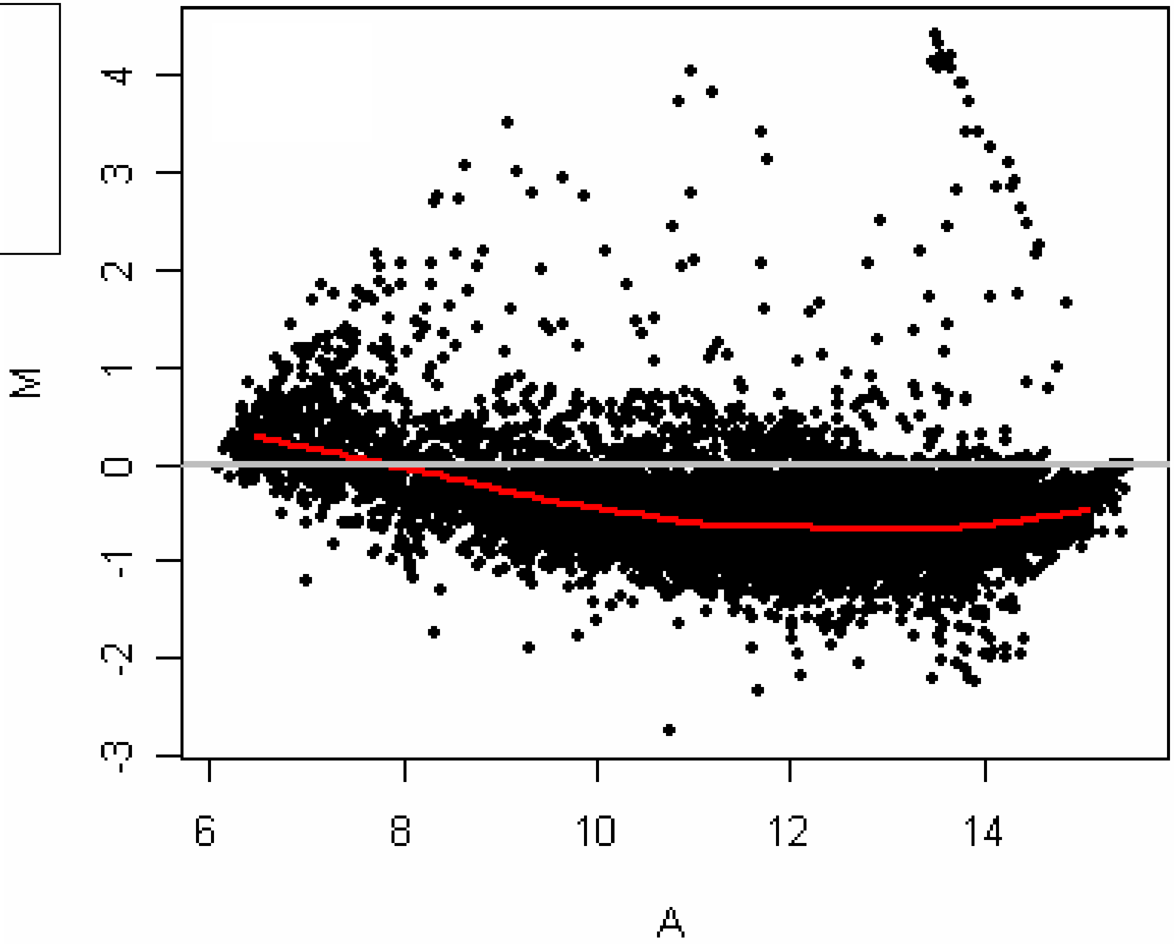
Box plot



MA plot

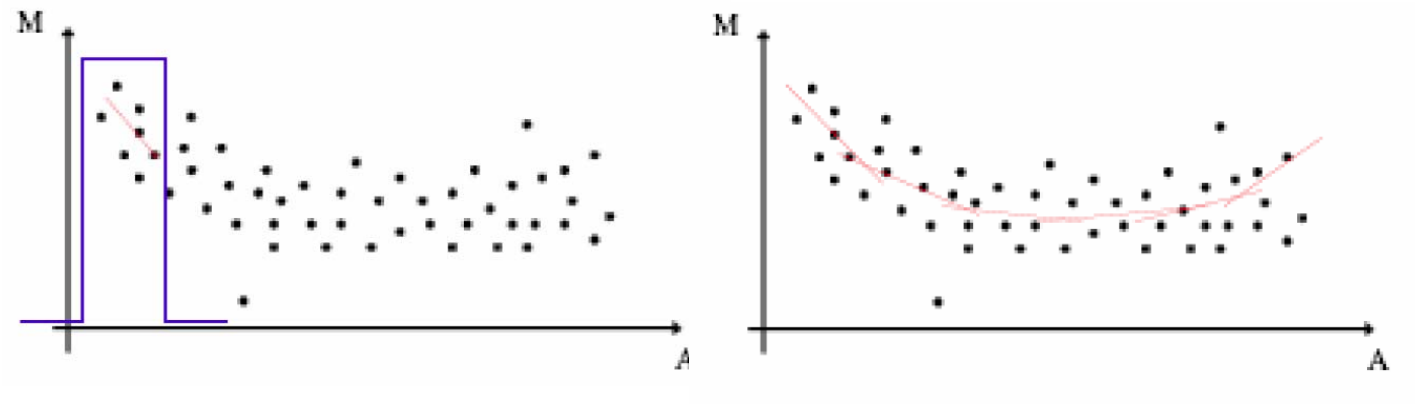
$M = \log_2(R/G)$

$A = \log_2(R*G)/2$



Intensity dependent normalization

- Apply a locally weighted polynomial regression for a fixed subset of genes in the neighborhood of every gene i (LOWESS).



- Weight function:

$$w(x_i) = \begin{cases} (1 - d(x, x_i)^3)^3 & : \quad d(x, x_i) < 1 \\ 0 & : \quad d(x, x_i) \geq 1 \end{cases}$$

Self normalization (dye-swap normalization)

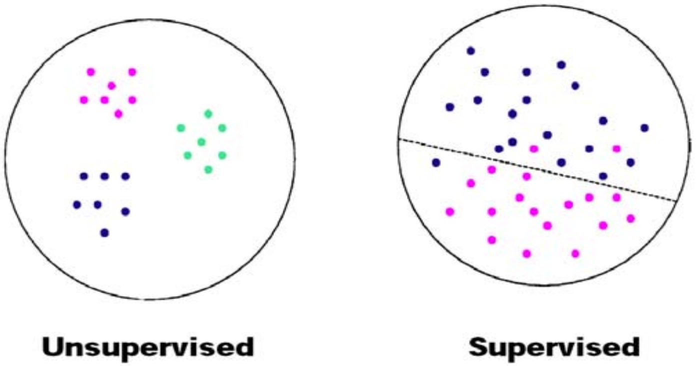
$M = \log_2 (R/G)$
$M' = \log_2 (R'/G')$

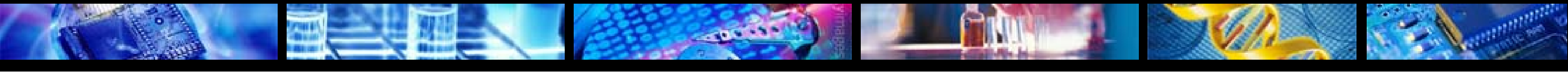
$Mn \approx [\ \log2\ (R/G)\ -\ \log2\ (R'/G')\ ]\ /\ 2$

$Mn = [\ \log2\ (RG'/GR')\ ]\ /\ 2$

Clustering

- Unsupervised or supervised (classification)





Why cluster gene expression profiles

- Functionally related genes are often co-expressed
- Relationship between co-expression and co-regulation
- If a gene has unknown function, but clusters with genes of known function, this is a way to assign its general function (‘guilt-by-association’)

Methods for unsupervised clustering

- Hierarchical Clustering
- K-means
- …

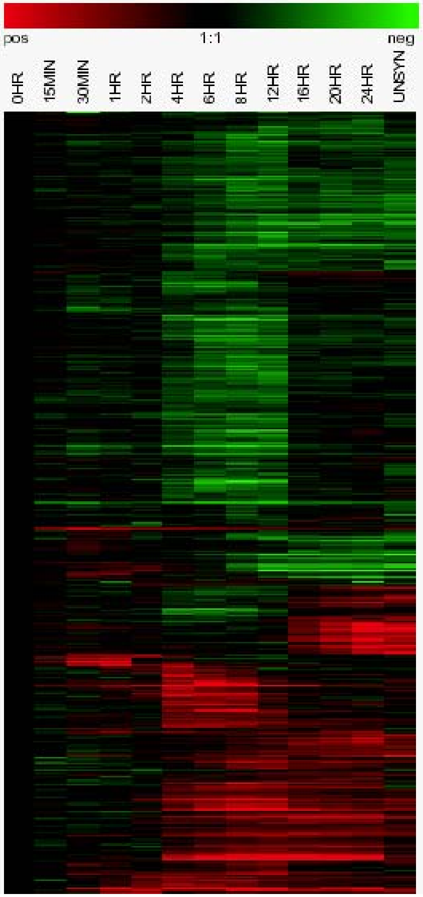
Data format

(n x m) Matrix of n Genes and
m Experiments

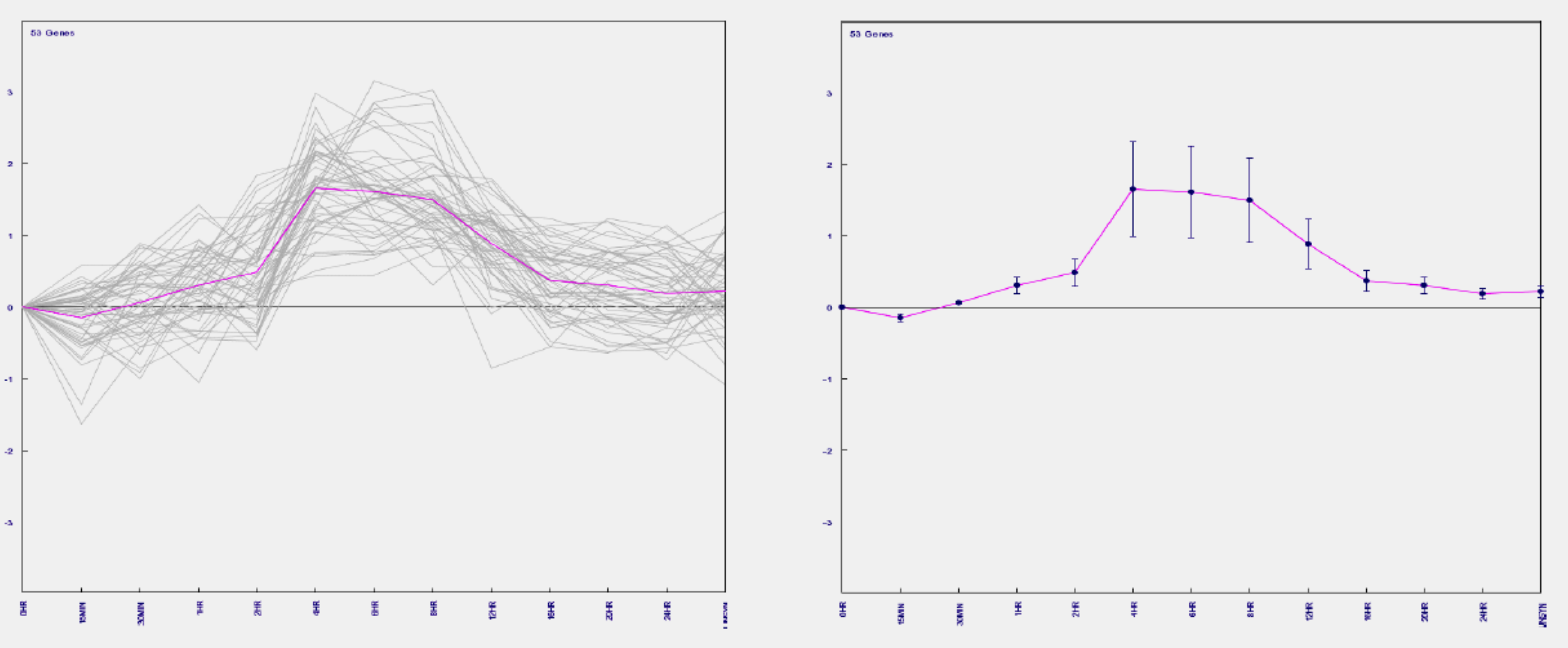
$$x_{ij} = \log_2 \frac{C5_{ij}}{C3_{ij}}$$

$C5_{ij}$ …Cye-5 of gene *i* in microarray experiment *j*

$C3_{ij}$ …Cye-3 of gene *i* in microarray experiment *j*

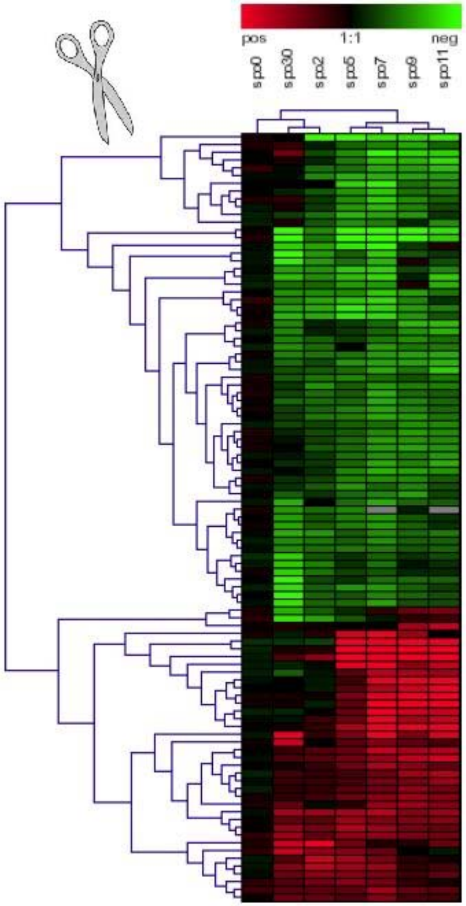


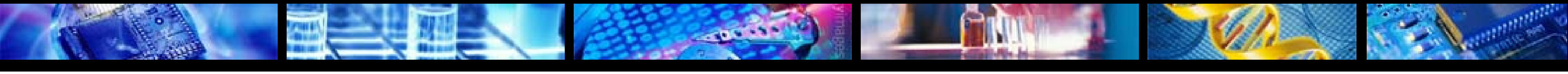
Graphical presentation



Hierarchical clustering

- Reorders the vectors regarding similarity
- Distances are encoded in dendrogram (tree)
- Unsupervised
- Very computational intensive
- Clusters of genes and experiments (bi-clustering)





Hierarchical clustering

Nodes = genes or groups of genes. Initially all nodes are rows of data matrix

1) Compute matrix of all distances (correlation coefficient)

2) Find two closest nodes.

3) Merge them by averaging measurements (weighted)

4) Compute distances from merged node to all others

5) Repeat until all nodes merged into a single node

Linking within hierarchical clustering

- Single-linkage clustering
  The distance between two clusters, *i* and *j,* is calculated as the minimum distance between a member of cluster *i* and a member of cluster *j*.

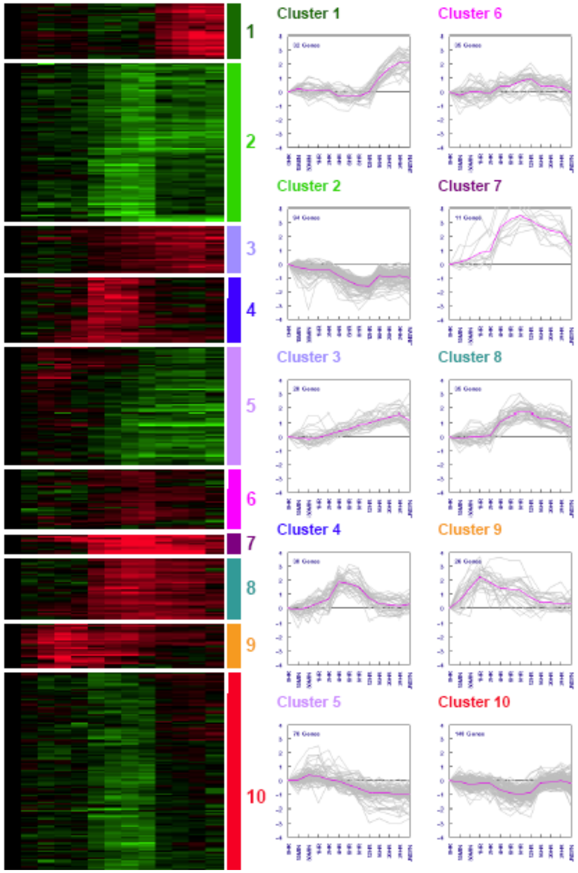
- Complete-linkage clustering
  Here the maximum distance is used

- Average-linkage clustering
  Calculated using average values (UPGMA)

- Weighted pair-group average
  Like UPGMA but weighted according cluster size

K-means

- It partitions *n* genes into *k* clusters, where *k* has to be predetermined
- k-means clustering minimizes the variability within each cluster
- Tries to maximize the distance between clusters
- Moderate memory and time consumption



K-means

0) Choose number of clusters

1) Generate random points (“cluster centers”) in n dimensions (results are depending on these seeds)

2) Compute distance of each data point to each of the cluster centers.

3) Assign each data point to the closest cluster center.

4) Compute new cluster center position as average of points assigned.

5) Loop to (2), stop when cluster centers do not move very much.

Exercise

- 2 Parts:
  - Two colour microarrays
    - Microarray background correction, normalization, dye swaps, M values.
  - Clustering of interesting genes
    - K-means, hierarchical

Exercise Microarrays

- We will use Bioconductor – open source software for bioinformatics; R based
- We will use the following packages:
  - marray, limma, Biobase, OLIN, gplots
- Installing packages
  - From Bioconductor
    - source(”http://bioconductor.org/biocLite.R”)
    - biocLite(“marray”);
  - From R
    - install.packages(“gplots”)
- Load packages
  - library(“marray”)

Exercise Microarrays

- Useful functions
  - getwd()/setwd() – get/set working directory
  - par(mfrow = c(2,2), mar = c(2,2,2,2))

  - read.marrayInfo()/read.Galfile()/read.GenePix()
  - backgroundCorrect2()/maNorm()
  - maPlot()/maBoxplot()/plot()

  - read.csv()/cor()/heatmap()/image()
  - hclust()/kmeans()