Analysis of gene expression data

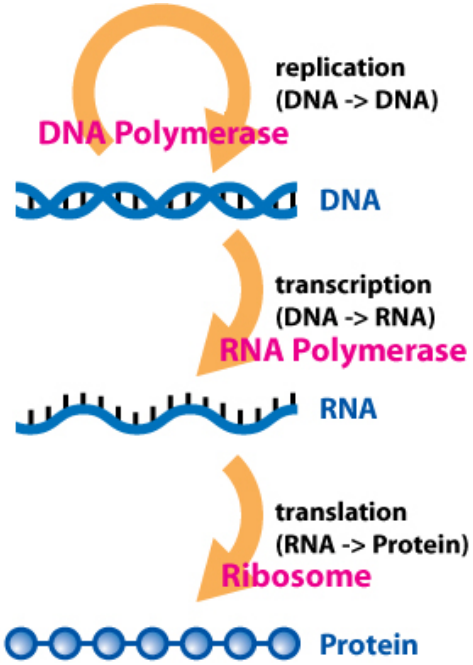
Ulf Leser and Philippe Thomas

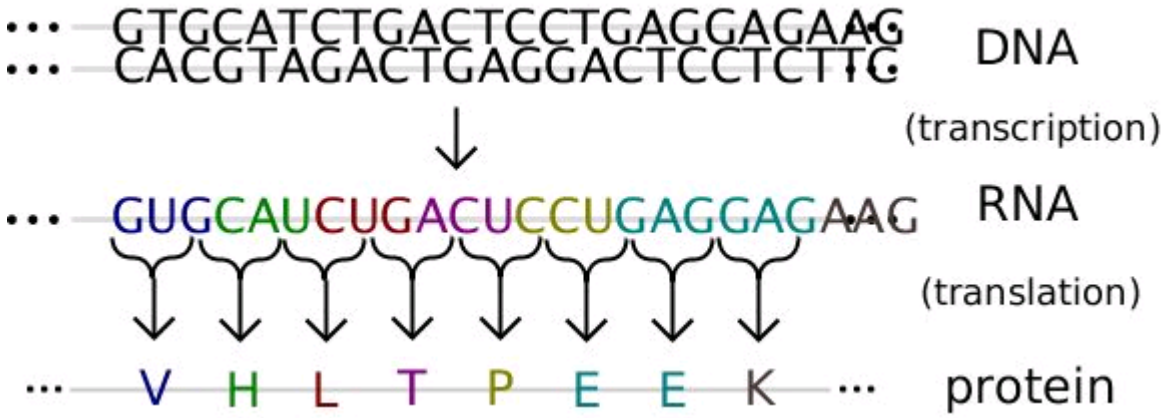
This Lecture

- Protein synthesis
- Microarray
  - Idea
  - Technologies
  - Applications
  - Problems
- Quality control
- Normalization
- Analysis next week!

Protein synthesis

- Gene expression has 2 phases:
  - Transcription (DNA -> mRNA) (40 nt /second)
  - Translation (mRNA -> protein) (40 aa/second)





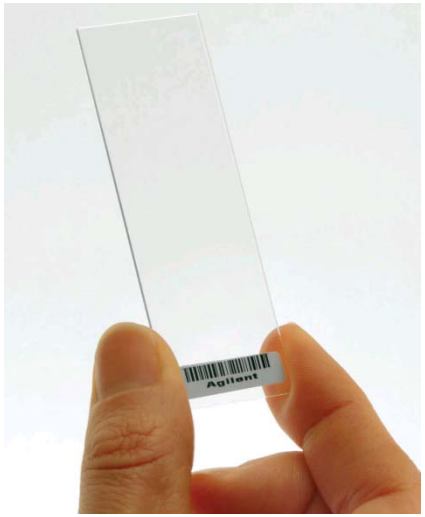
Wikipedia

AG:Proteomics Algorithms and Simulations

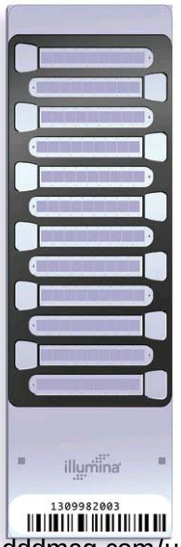
Tübingen

mRNA quantification

- Reporter Gene (GFP)
- Northern blotting
- Real time PCR
- Microarray
  - High throughput (multiple genes with one experiment)



http://www.agilent.com/about/newsroom/lsca/imagelibrary/images/lsca_160_Array_Slide.jpg



http://www.dddmag.com/uploadedImages/rticles/2007_10/dd7odisnp5.jpg

This Lecture

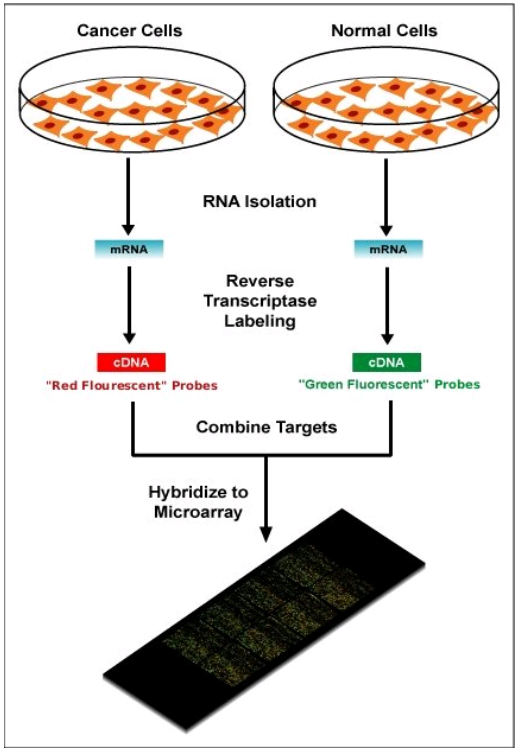
- Protein synthesis
- Microarray
  - Idea
  - Technologies
  - Applications
  - Problems
- Quality control
- Normalization
- Analysis next week!

Microarray Experiment I

- Comparative
  - Between experiment not between gene
- Measure mRNA expression
  - For all genes
  - At a specific time point
  - One sample
- Assumes that mRNA expression correlates with protein synthesis
  - Not entirely true
- Trend towards next generation sequencing
  - But: Microarrays are an important and prominent technology

Microarray Experiment II

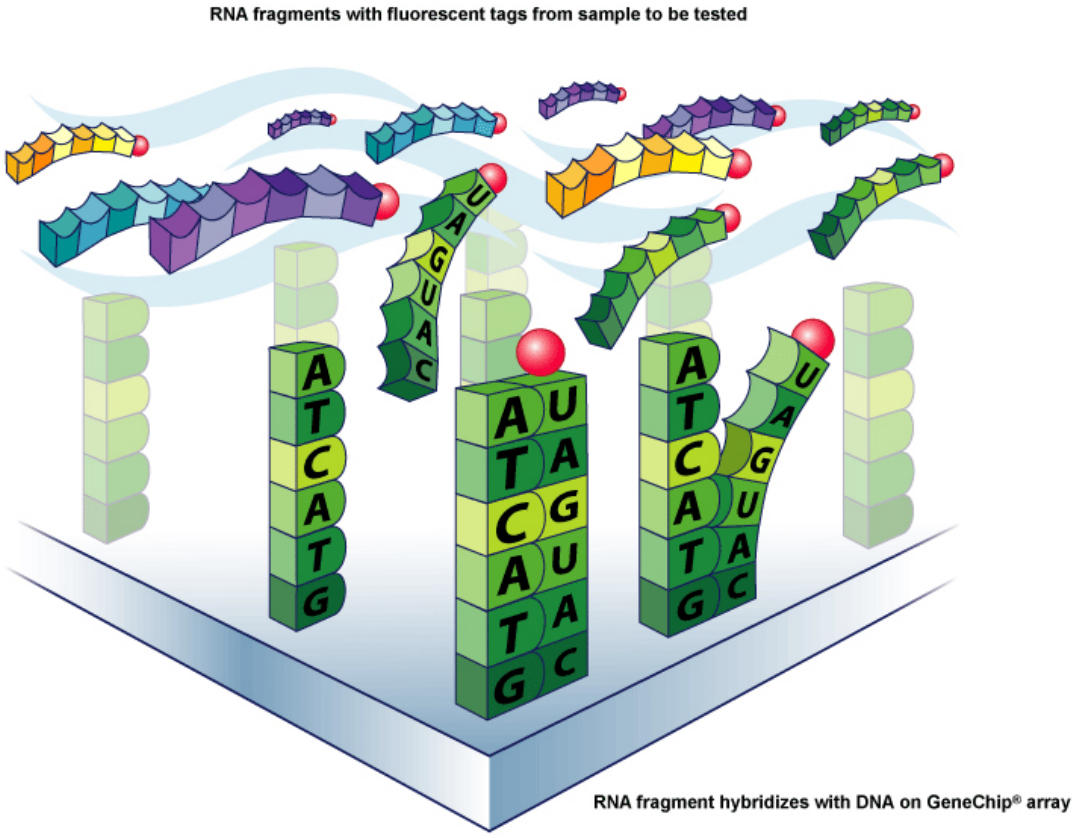
- Find differentially expressed genes between groups
  - Healthy vs. sick
  - Tissue /Cell types
  - Development state
    - Embryo, Child, Adolescent, Adult,
    - Cell development
  - Environment
    - Heat shock, Nutrition, Therapy
  - Disease subtypes
    - ALL vs. AML
    - 40% chemotherapy resistant in colon cancer
- Co-regulation of genes
  - Similar gene-pattern -> similar function?
  - Similar gene-pattern -> similar regulation?



Wikipedia

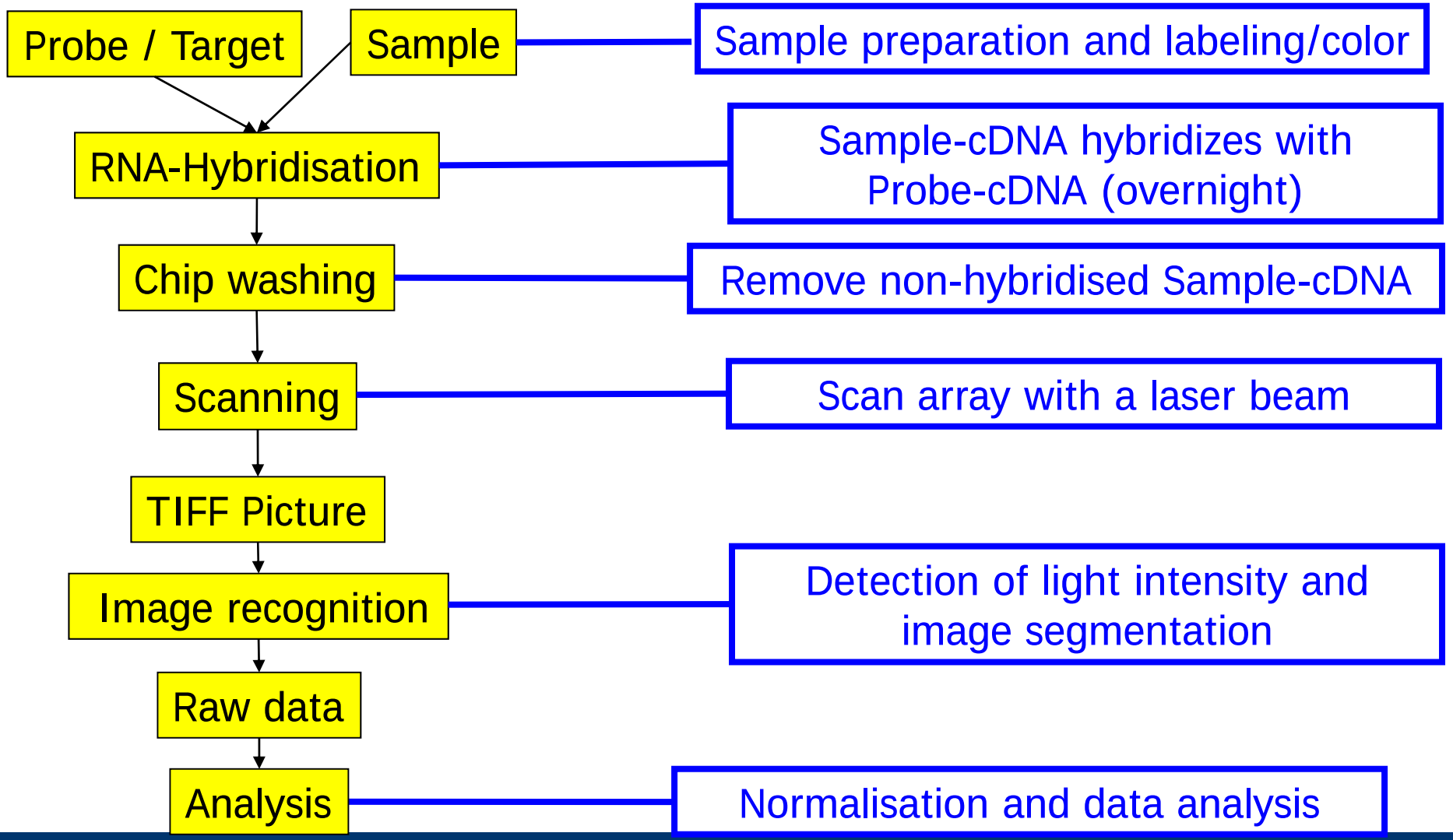
RNA-Hybridisation

Hybridization is the process of unifying two complementary single-stranded chains of DNA or RNA two one double-stranded molecule.



http://encyclopedia2.thefreedictionary.com/micro+array

Application flow



Probe preparation

- Replicate cDNA library (PCR)
- Spot on array
- Each spot represents a transcript/gene (idealized)
- Array-Layout: Redundancy, controls, maximum number of spots, ...

	S1	S2	S3	S4	S5	S6	...
Z1	G11	G11	G21	...	...	...	...
Z2	G12	...	...	...	...	...	...
Z3	G13	...	...	...	...	...	...
Z4	G14	...	...	...	...	...	...
Z5	G15	...	...	...	...	...	...
Z6	G16	...	...	...	...	...	...
...	...	...	...	...	...	...	...

Sample preparation

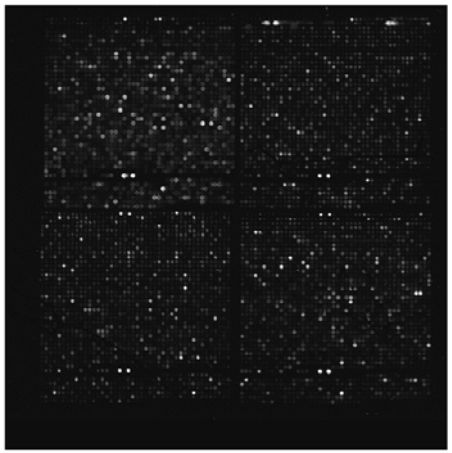
- Isolate cells in condition X
- RNA extraction
- Synthesize to cDNA
- cDNA labeling with colorized nucleotides
  - Labeled nucleotides
  - Biotinylated Oligo-dT
- RNA-Hybridisation

Differences in Technology – Spotted Array

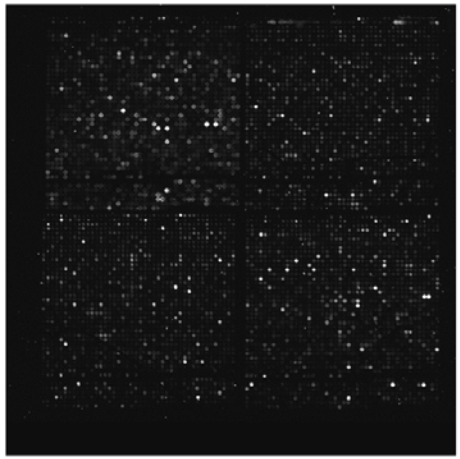
- Spotted on a microscope slide
- Spacing between two spots is ~120um
- 5ng of DNA/spot needed
- Many labs have required equipment
- Customization of probes (selection according to user need)
- Fewer Probes/Genes represented
- Probes are longer

Differences in Technology – Two Color Array I

- Two samples with one array
  - Two different colors Cy3 / Cy5
  - Laser uses two different wave length
- Usually spotted
- Cross hybridisation between the two samples possible

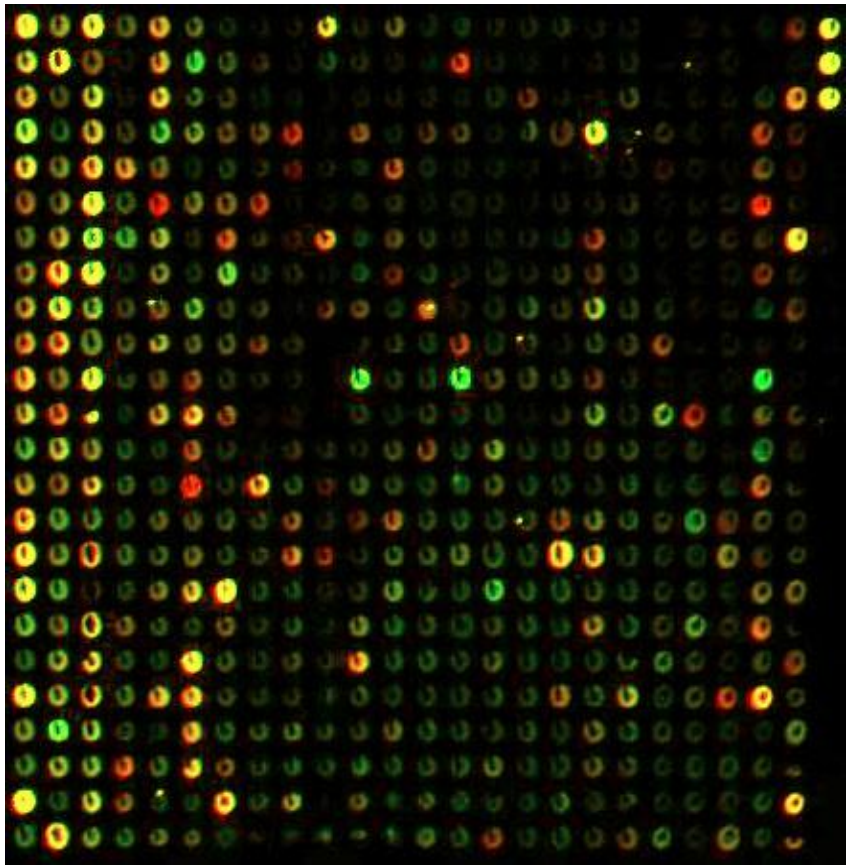


Sample A
(Red) Channel



Sample B
(Green) Channel

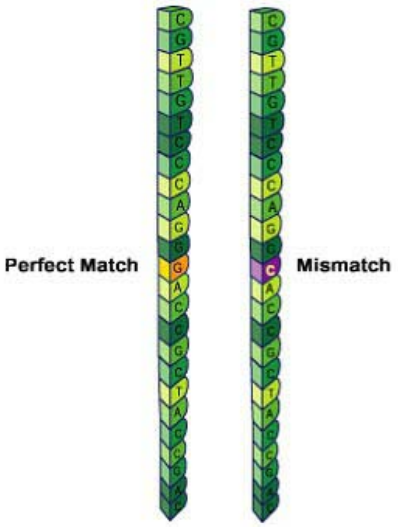
Differences in Technology – Two Color Array II



- Sample A: red
- Sample B: green
- Ratio Red/Green
  - **Black**: No signal in both samples
  - **Red**: High expression in A
  - **Green**: High expression in B
  - **Yellow**: Equal expression

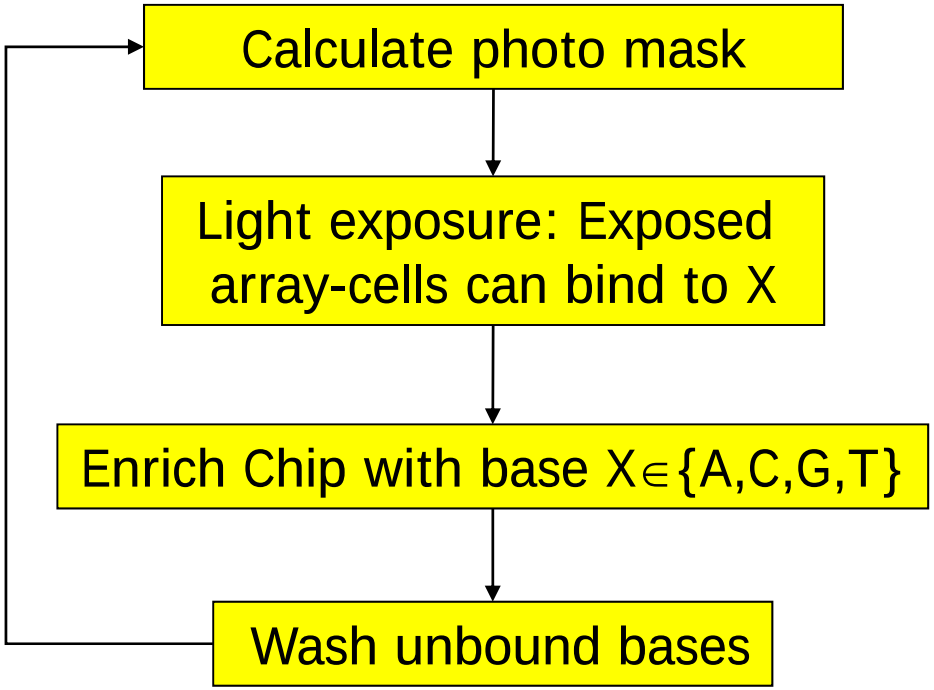
Differences in Technology – Oligo Microarray

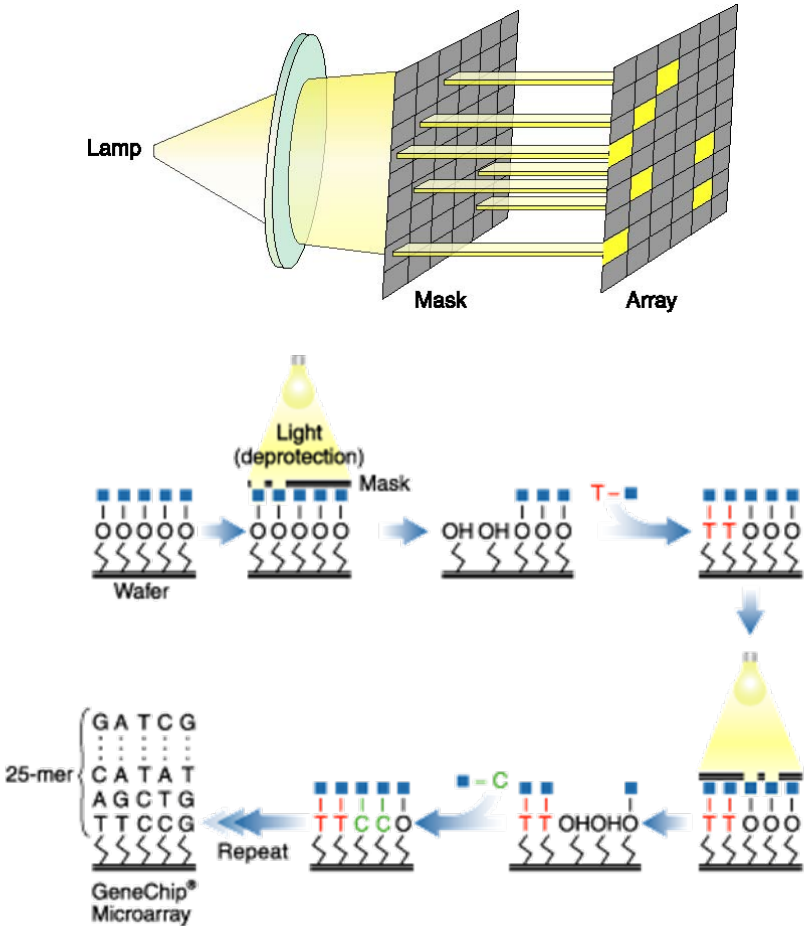
- Robustness between one array type
- Often one sample/chip
  - Twice as many chips as two-color
  - No cross hybridization between samples
- Short Probes (25nt – 80nt)
  - 11 - 20 Probes is one probeset
    - represent one gene transcript
    - scattered over array
  - Up to 4 Mio probes / Chip
  - Perfect match and mismatch probes



Affymetrix

Differences in Technology – Oligo Microarray

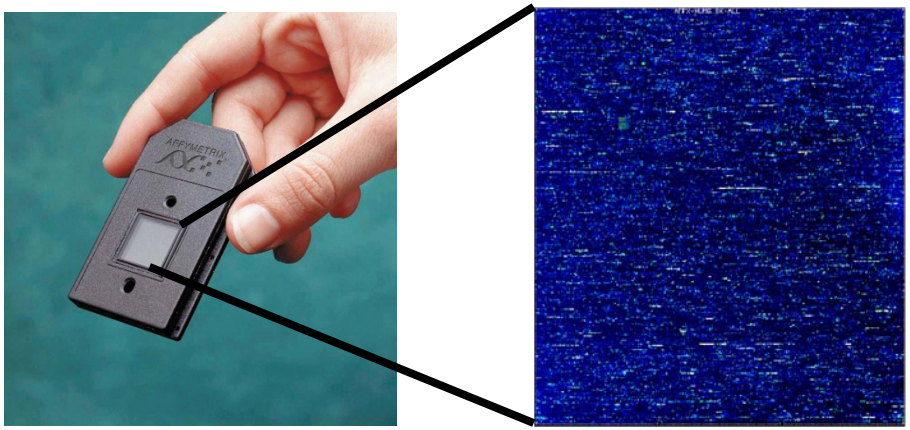




http://www.koreanbio.org/Biocourse/images/e/e0/Photolithography.gif

Differences in Technology – Oligo Microarray

- Good quality control
- No customization of probes
- Selection of good oligos is difficult
  - Probe self-hybridization
  - Probe should be unique for a transcript
  - Minimize the number of light expose cycles (reduce cost)



Comparison

- Oligo Chips
  - Densely packed
  - Companies sell „kits“ for every use case
  - Robust → reproducible results
  - Easy to handle in tools like R (probe annotation)
  - No customization of probes
  - Not available for all species
- cDNA Arrays
  - cDNA covers a longer mRNA fragment
  - Good customization
    - Very time intensive
  - Less spots / genes → limited redundancy
  - Error prone workflow
  - Results difficult to compare
  - Only available for species with EST (cDNA transcripts)

Replicates

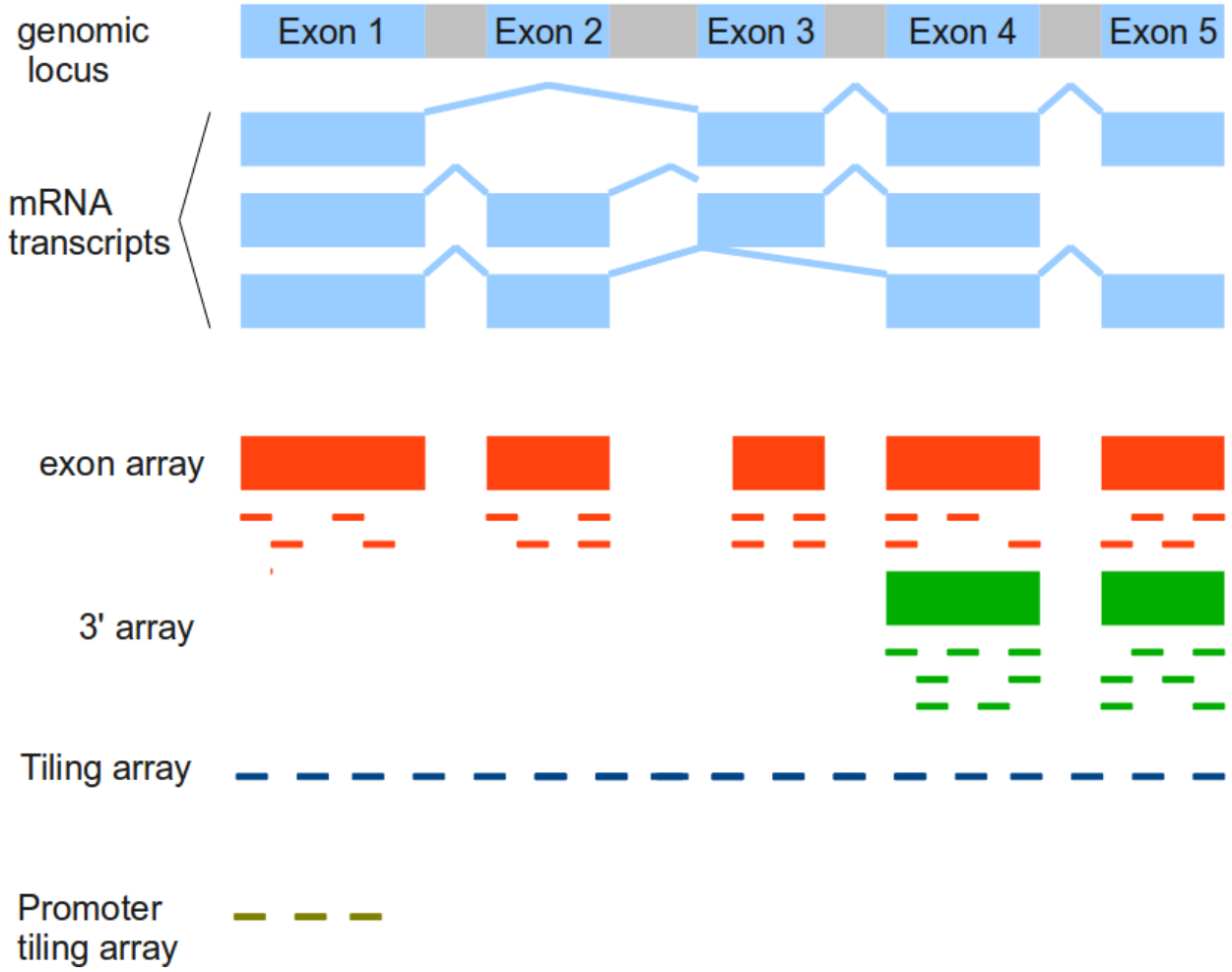
In order to exclude technical or biological bias, replicated measurements are exploited:

- Technical Replicates:
  - Same sample hybridized against several arrays
  - Statistical estimation of systematic effects
- Biological Replicates:
  - Different sample sources are used
  - They allow to estimate biological noise and reduce the randomness of the measurement.

Advances in technology

- Gene-Expression profiling
  - Usually referred to as Microarray or GeneChip
  - Measure expression level of all genes
- Exon array
  - Each exon of a gene is measured individually
- SNP array
  - Identifying single nucleotide polymorphism among alleles within or between populations.
- ChIP-on-chip
  - Detects DNA fragments specifically bound to a protein (e.g. transcription factor)

Advances in technology



Challenges

- Patient data has a high variance
  - Different genetic background
  - Mixture of cells from different tissues
  - Cells are in different stages (cell cycle; cell development)
- Gene representation
  - Very low mRNA level
  - Gene active during a short life time (embryonic stage, M-stage)
  - Genes not represented on Microarray
    - Annotation of a genome evolves over time; oligo-array is constant
- Environment has influence on hybridization quality
- Noise: Technical replicates never produce the same data

Challenges

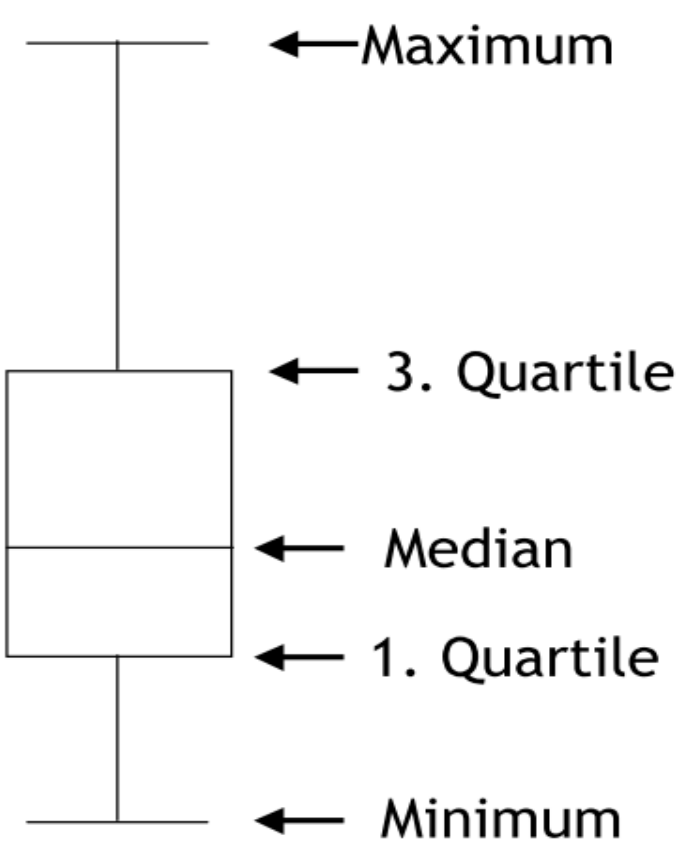
- Transient data
  - Select appropriate time point
  - Signaling might be very fast for some processes
  - Intermediate steps are lost
- Cause end effect
  - Tumors have high cell proliferation
- Biological interpretation difficult
- High number of transcripts
  - Multiple test correction
  - Choice of statistical test
- Time series results in day/night work and might result in a completely lost data set

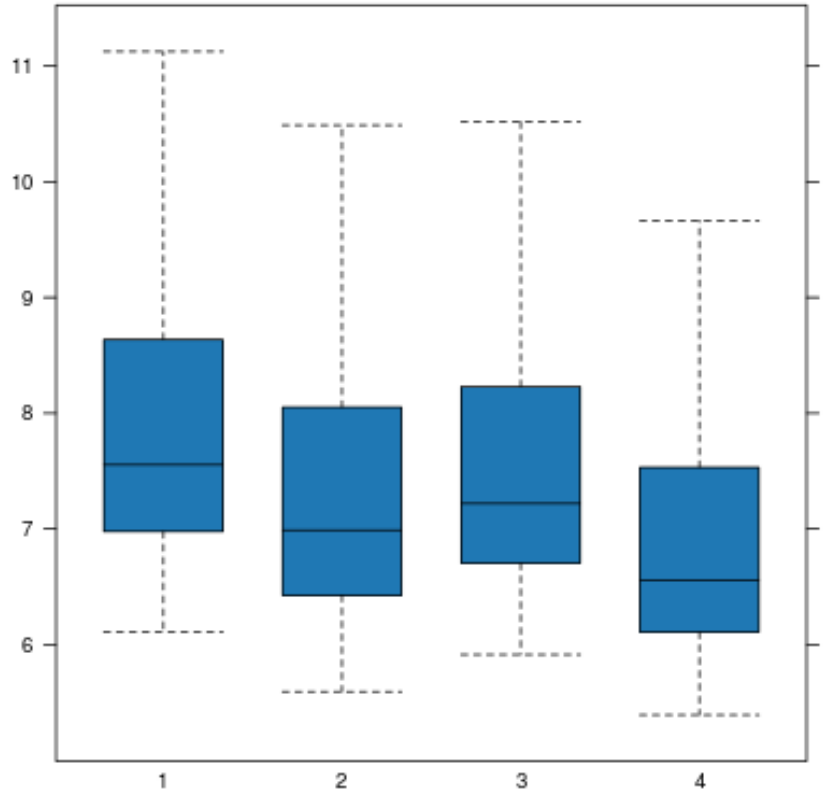
This Lecture

- Protein synthesis
- Microarray
  - Idea
  - Technologies
  - Applications
  - Problems
- Quality control
- Normalization
- Analysis next week!

Data visualization, quality control

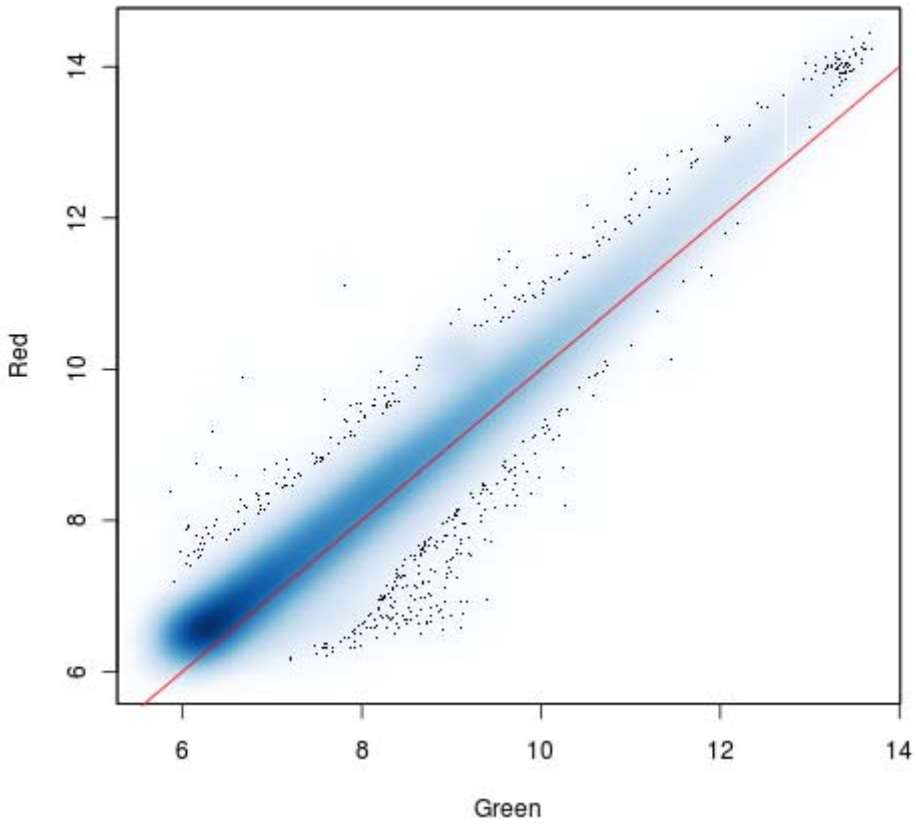
- Boxplot
  - Estimate the homogeneity of data





Data visualization, quality control

- Scatter plot
  - Each point represents one transcript in two experimental settings



MA-Plot I

- Fold Change or m-value
  - Fold change is the log2-ratio between two values
    - Log-values are symmetric
    - Visual interpretation (Difference between 4 to 16 vs. 0.25 to 0.0625)
    - Mathematical issues

$$FC(Value1/Value_2)=\log_2\left(\frac{Value_1}{Value_2}\right)$$

- For example:

$$FC(512/1024)=\log_2\left(\frac{512}{1024}\right)=\underline{-1}$$

$$FC(512/1024)=\left(\frac{512}{1024}\right)=\underline{0.5}$$

$$FC(123/123)=\log_2\left(\frac{123}{123}\right)=\underline{0}$$

$$FC(123/123)=\left(\frac{123}{123}\right)=\underline{1}$$

$$FC(512/256)=\log_2\left(\frac{512}{256}\right)=\underline{+1}$$

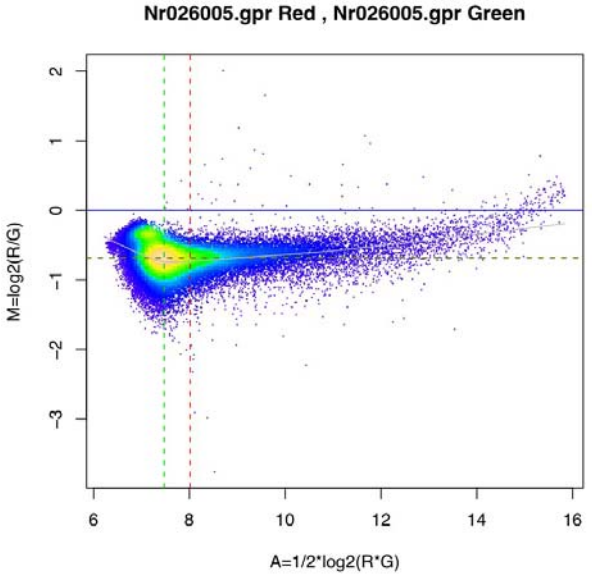
$$FC(512/256)=\left(\frac{512}{256}\right)=\underline{2}$$

MA-Plot II

- A-Value is the logged intensity mean value

$$A = \frac{1}{2} \times \left(\log_2(Value_1) + \log_2(Value_2) \right)$$

- Note that this scatter plot is a 45° rotated version with subsequent scaling of the normal scatter plot.



This Lecture

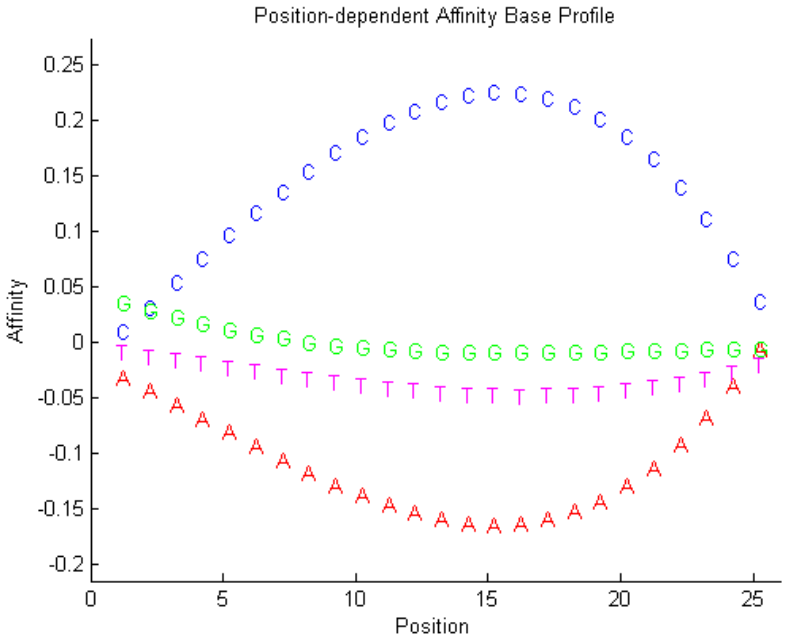
- Protein synthesis
- Microarray
  - Idea
  - Technologies
  - Applications
  - Problems
- Quality control
- Normalization
- Analysis next week!

Normalization

- Microarrays are comparative experiments
  - Distinguish between biological from technical variation
- Measurements between two experiments are not directly comparable
- Minimize the influence systematic errors on the experiment
  - Sample preparation
  - Different quantities of RNA
  - Probe affinity
  - Fluorescence detection non-linear
  - self fluorescence of microarray surface
  - Experimentator variability
- No method to handle bad RNA quality

Normalization

- G and C bindings have a higher energy than A and T bindings.
- Furthermore, the binding specificity depends on the nucleotide positions in the probe.



mathworks.com (based on Naef and Magnasco, Phys. Rev., 2004)

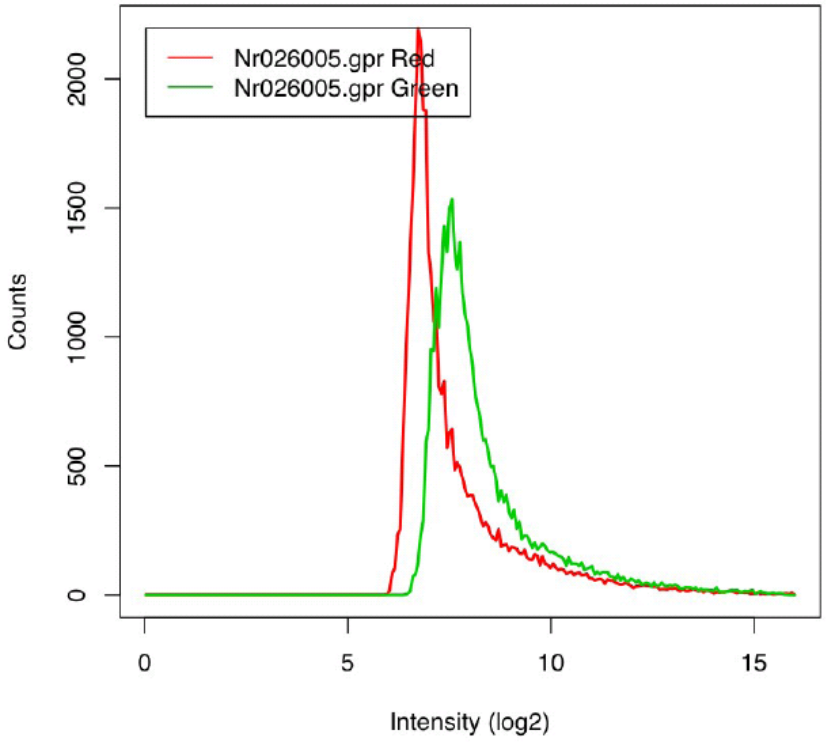
Normalization – Global scaling

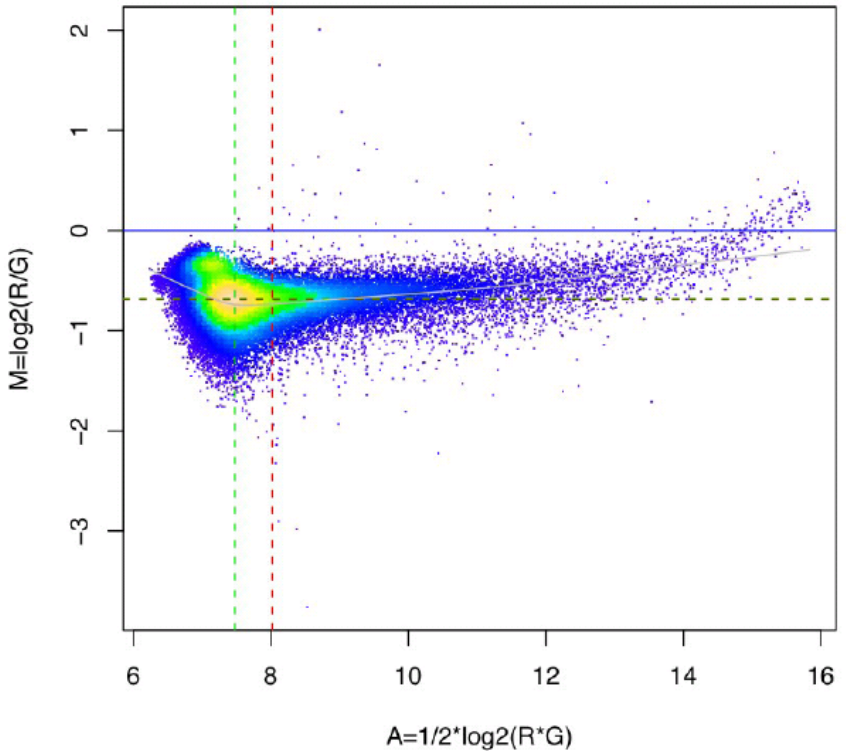
- mRNA in a sample
  - Assumption: „cells contain same proportion of RNA“
  - Measure total mRNA
  - Divide intensities by this value
- Reference gene
  - Assumption: „These genes are similar expressed across tissue“
  - Selection of „Housekeeping“ genes
  - Divide intensities by this value

Non linear methods – Lowess I

- Dye-bias in two color array
  - Green channel appears consistently brighter then red channel
  - Intensity based
- Fit simple models to localized subsets
  - Needs no global function of any form to fit a model to the data
  - It requires large, densely sampled data sets in order to produce good models

Non linear methods – Lowess II





Quantile normalization I

- Impose same empirical distribution of entities to each array
- Each hybridization is thus the transformation of an underlying common distribution
- Usually outperforms linear methods
  - Sophisticated methods like RMA use quantile normalization

Quantile normalization II

1. Given a matrix X where p x n where each array is a column and each transcript is a row
2. Sort each column of X separately to give X_{sort}
3. Take the mean, across rows, of X_{sort} and create X'_{sort}
4. Get X_n by rearranging each column of X'_{sort} to have the same ordering as the corresponding input vector

Quantile normalization III

1. Given a matrix X where p x n where each array is a column and each transcript is a row

Sort by column

	Array 1	Array 2	Array 3
Gene1	1	6	8
Gene2	2	5	9
Gene 3	3	4	7

	Array 1	Array 2	Array 3
Gene1	1	4	7
Gene2	2	5	8
Gene 3	3	6	9

Row wise mean

	Array 1	Array 2	Array 3
Gene1	4	4	4
Gene2	5	5	5
Gene 3	6	6	6

Rearrange

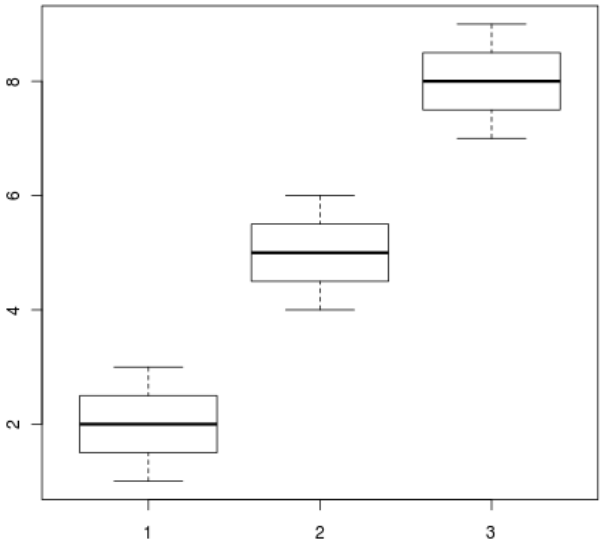
	Array 1	Array 2	Array 3
Gene1	4	6	5
Gene2	5	5	6
Gene 3	6	4	4

Rearrange

Quantile normalization IV

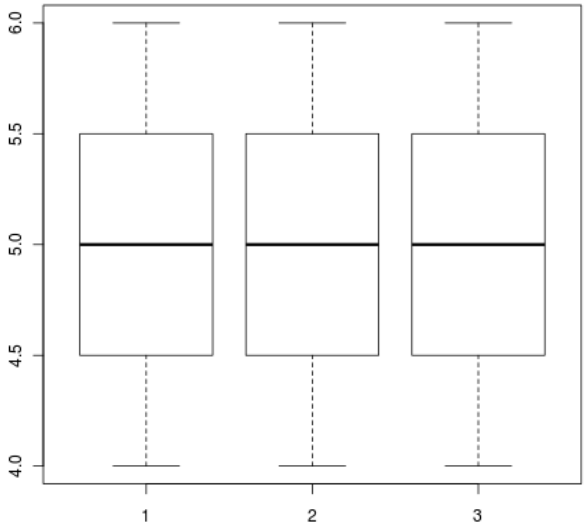
Before

	Array 1	Array 2	Array 3
Gene1	1	6	8
Gene2	2	5	9
Gene 3	3	4	7



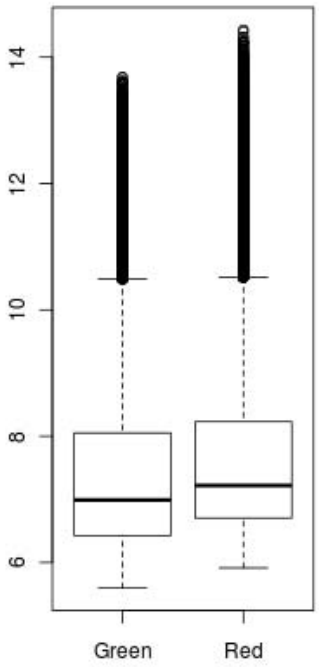
After

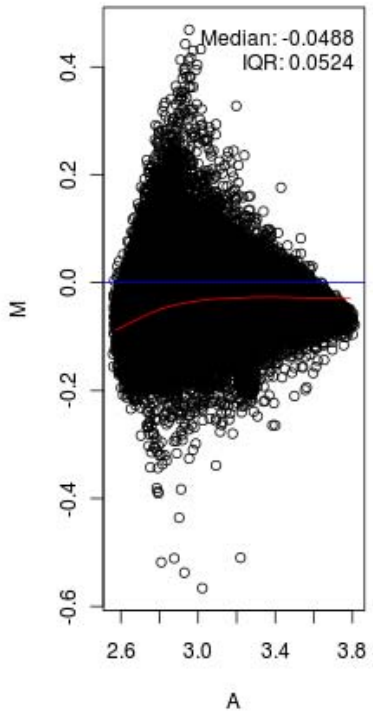
	Array 1	Array 2	Array 3
Gene1	4	6	5
Gene2	5	5	6
Gene 3	6	4	4



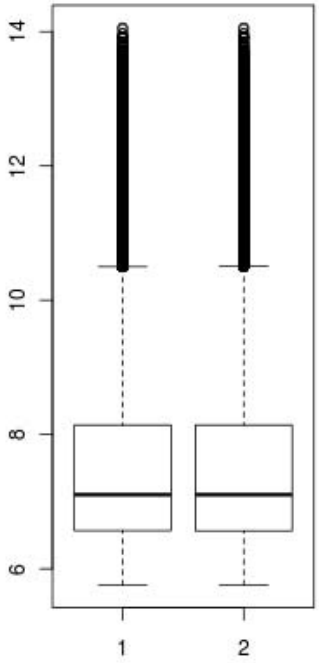
Quantile normalization V

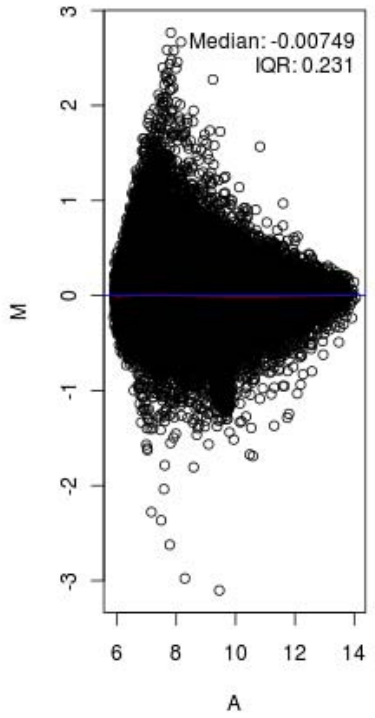
Before



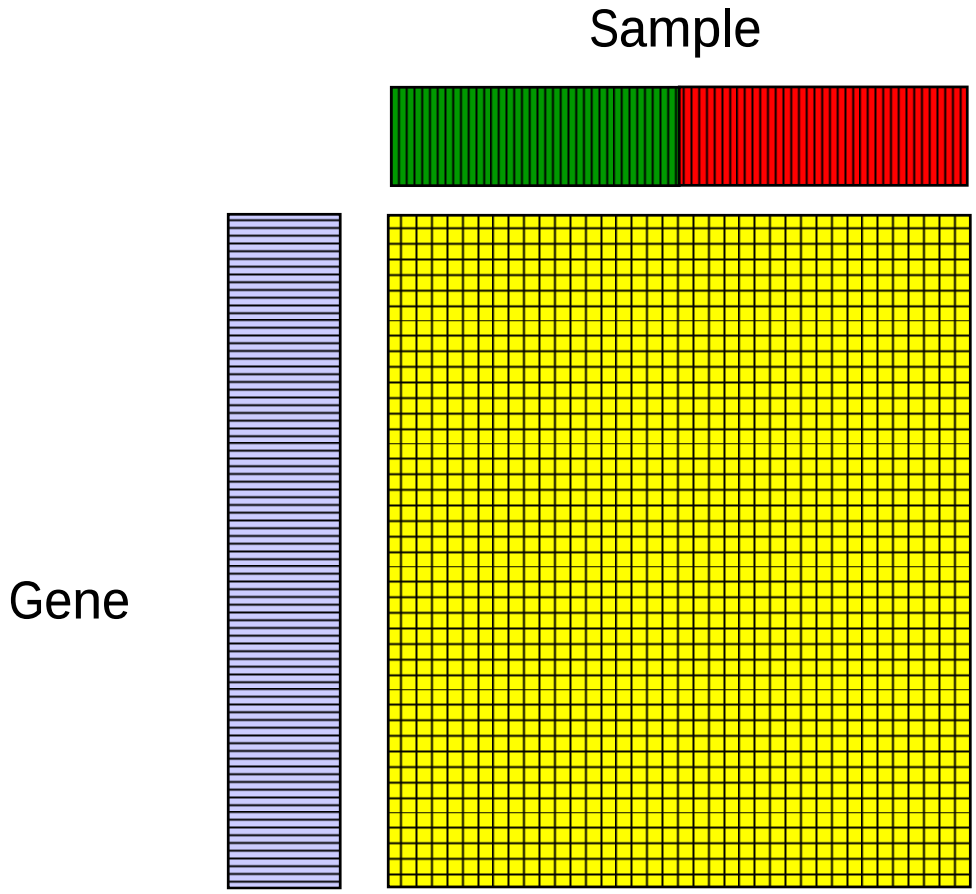


After





Gene expression matrix



Conclusion

- Different techniques
  - cDNA: cDNA library
  - Oligo: artificial Oligos
- Problems
  - Image recognition
  - Normalization
- Comparative tool
- Findings about interplay of genes in pathways
- Often used

This Lecture

- Protein synthesis
- Microarray
  - Idea
  - Technologies
  - Applications
  - Problems
- Quality control
- Normalization
- Analysis next week!