



Measuring gene expression (Microarrays)

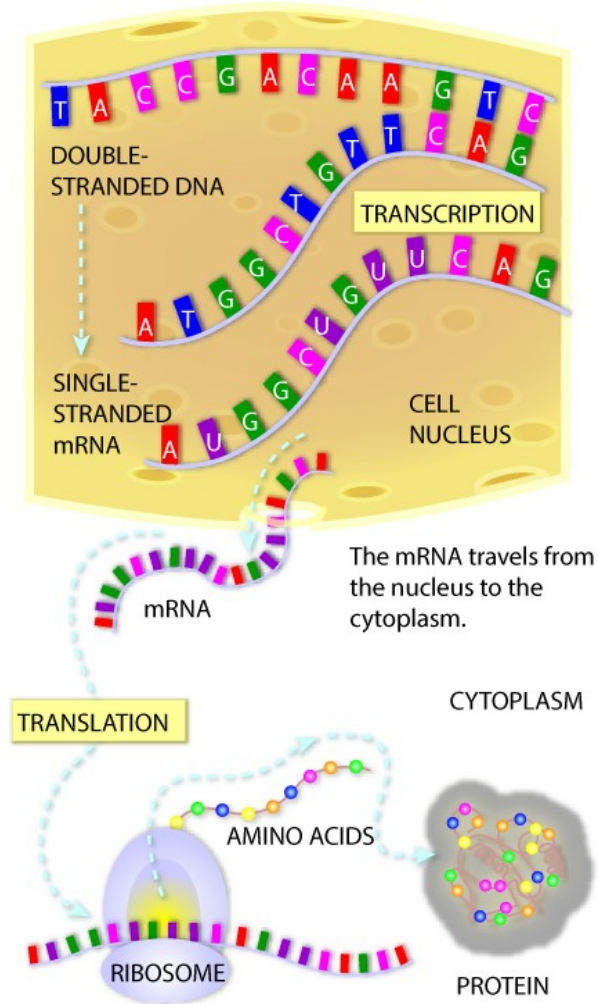
Ulf Leser

This Lecture

- Gene expression
- Microarrays
 - Idea
 - Technologies
 - Problems
- Quality control
- Normalization
- Analysis next week!

Recap – Gene expression (Protein Biosynthesis)

<http://learn.genetics.utah.edu/content/molecules/transcribe/>



- Gene expression has 2 phases:

- **Transcription**

DNA → mRNA

RNA polymerase

- **Translation**

mRNA → protein

(Ribosome)

alternatively mRNA (transcript) may encode for miRNA, rRNA, tRNA

Proteins have various functions:
antibodies, enzymes, hormones,
storage, transport, structure, regulation

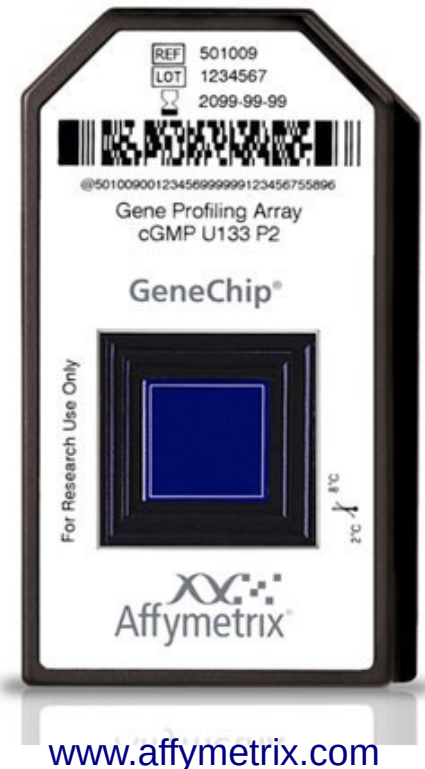
mRNA Quantification

Differences in gene expression correlate with drug response or disease risk!

- Measuring the relative amount of mRNA expressed in different experimental conditions
- Assumption: mRNA expression correlates with protein synthesis (not entirely true)

Techniques:

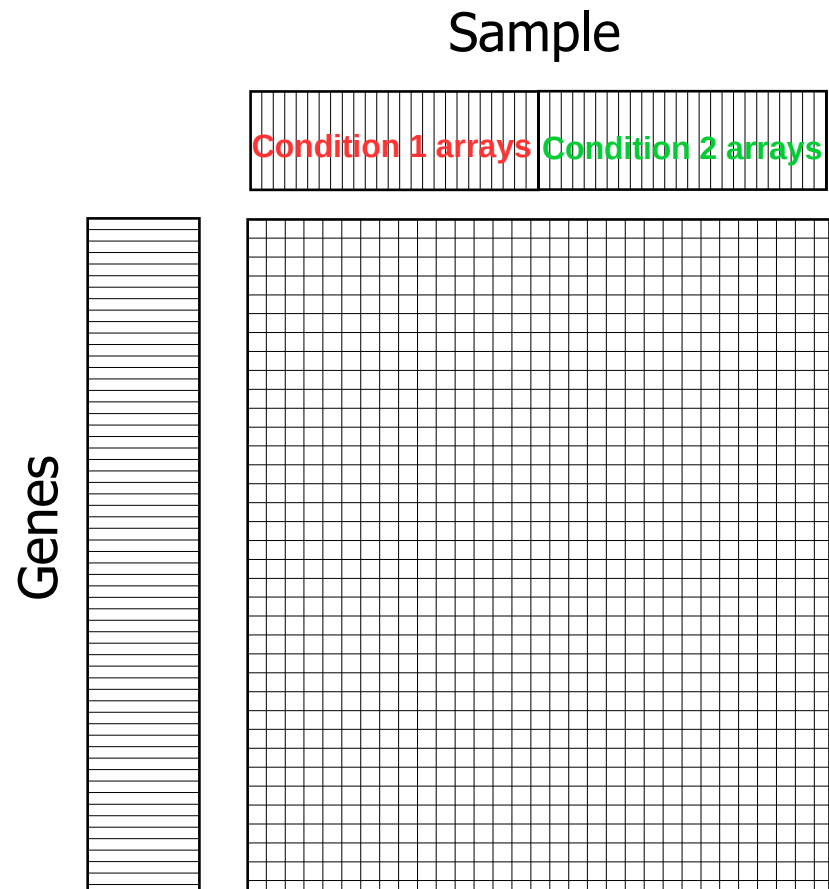
- Northern/Southern blotting
- Real time PCR
- High throughput analysis (multiple genes with one experiment):
 - **Microarrays** (since 1995)
 - Trend towards “next generation sequencing”, e.g. RNA-Seq (~since 2007)



Gene expression matrix

Birds-eye perspective:

- Data Visualization, Quality Control
 - (Background Correction)
 - Normalization
 - (log2 of data)
- gene expression matrix



Analysis next week!

This Lecture

- Gene expression
- Microarrays
 - Idea
 - Technologies
 - Problems
- Quality control
- Normalization
- Analysis next week!

Microarrays - Overview

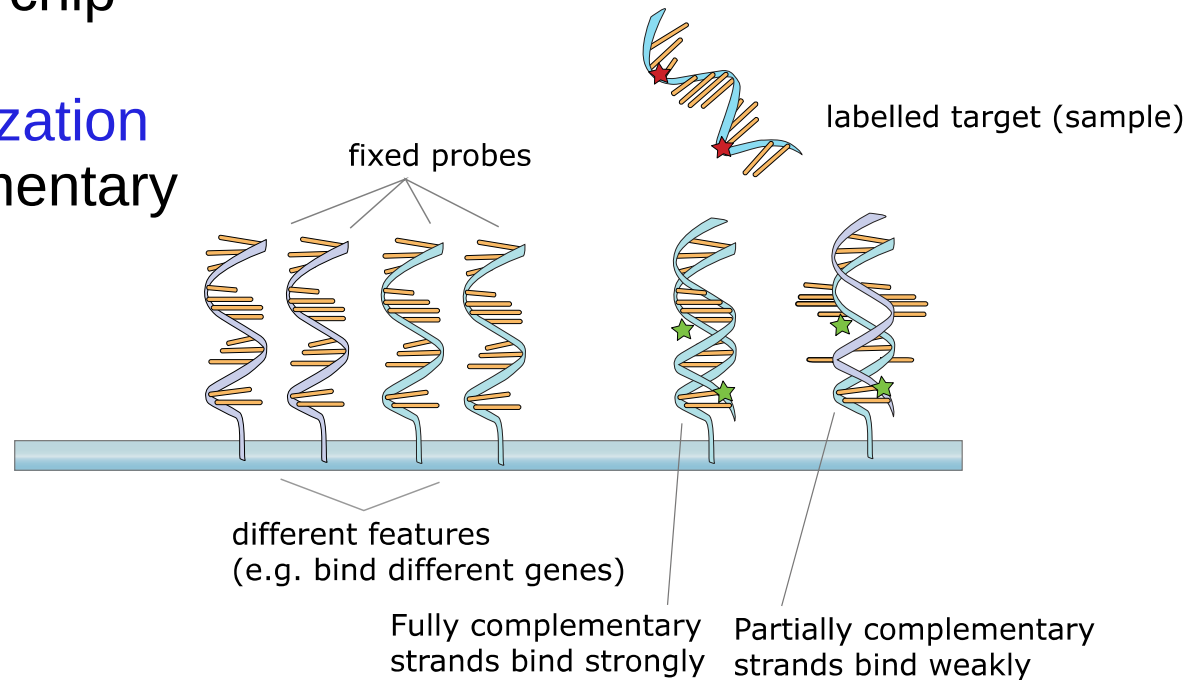
- Lab-on-a-chip: Measure mRNA expression for all genes at a specific time point in parallel
- Find differentially expressed genes between experiments/groups (not between genes):
 - × Healthy vs. sick
 - × Tissue/Cell types
 - × Development state
(embryo, adult, cell states, ...)
 - × Environment
(heat shock, nutrition, therapy)
 - × Disease subtypes (ALL vs. AML,
40% chemotherapy resistant in colon cancer)
- Co-regulation of genes:
similar gene-profile → similar function/regulation?
- Two types of microarrays:
cDNA microarrays and oligonucleotide microarrays

Microarray – Core Principle

Microarray: collection of single stranded DNA sequences (**probes**) attached to a solid surface (e.g. glass)

Sample mRNA is labeled with dyes and given onto the chip

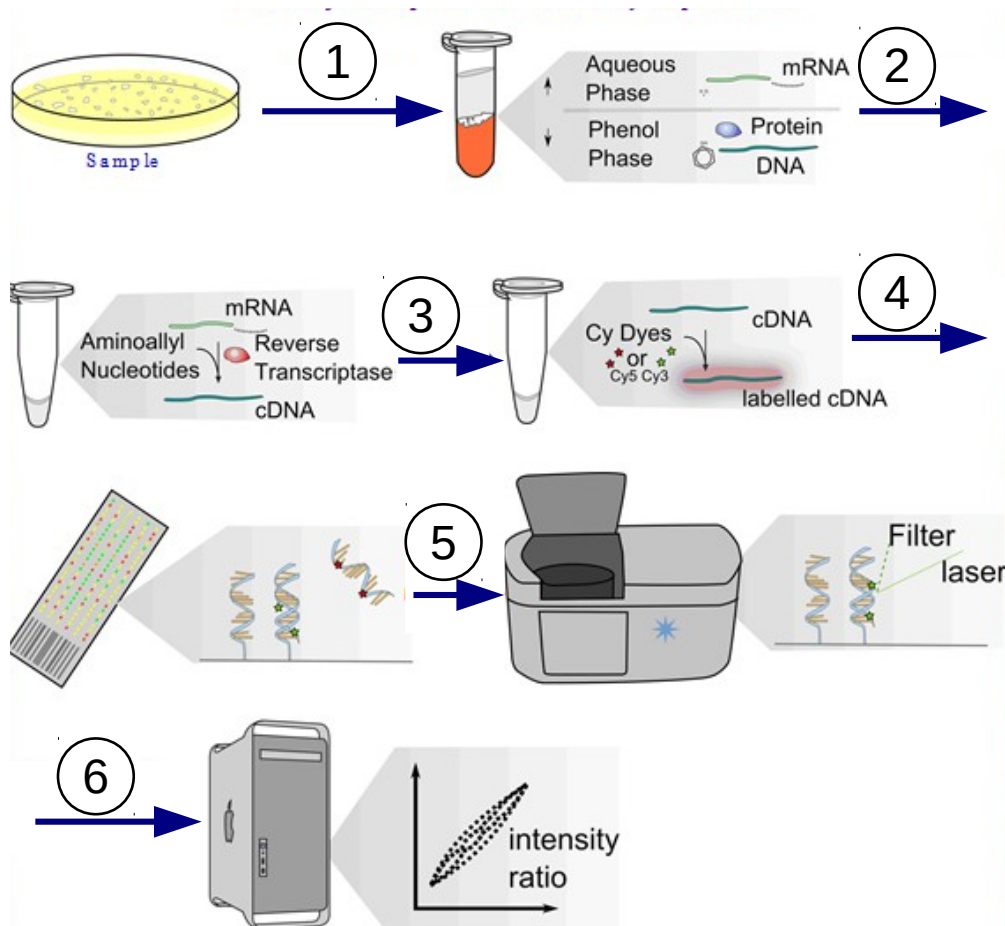
In the process of **hybridization** mRNA binds to complementary sequences on the array



Hybridisation: Process of unifying two complementary single- stranded chains of DNA or RNA to one double-stranded molecule.

www.wikipedia.com

Microarrays - Workflow

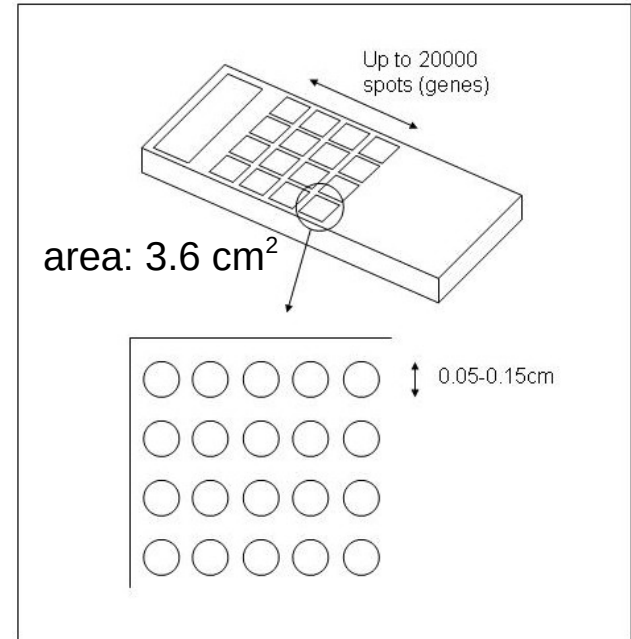


www.wikipedia.com

- 1) Sample preparation and purification, isolation of mRNA
- 2) Reverse transcription mRNA → cDNA (complementary DNA)
- 3) Labelling of sample cDNA with fluorescent dyes
- 4) Hybridization (overnight) and washing
- 5) Scanning of array with laser, detection of light intensities, image segmentation
- 6) Normalization of raw data and data analysis

cDNA Microarrays (Spotted Arrays)

- Manufacturing steps :
 - Selection of genes
 - Preparation/purification of these genes
 - Amplification of corresponding cDNA via PCR*
 - micro spotting of cDNA sequences (probes) on a glass slide
 - 10000-20000 spots, each represents one gene
 - negative control spots (background correction)
- Known as „in-house“ printed microarrays:
 - Easy customization
 - No special laboratory equipment needed
 - Relatively low-costs

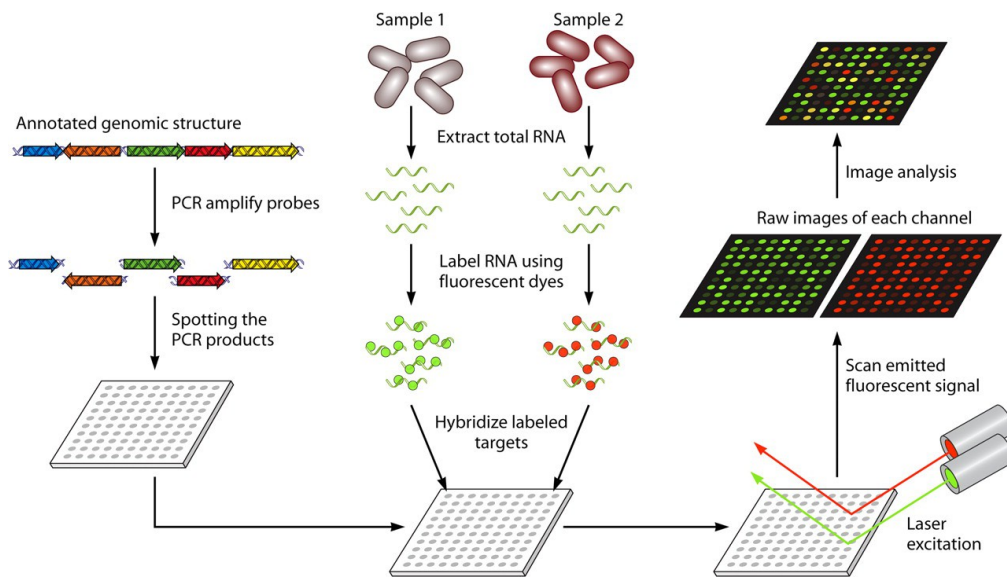


<http://grf.lshtm.ac.uk/microarrayoverview.htm>

*Polymerase Chain Reaction

cDNA Microarrays – Two Color Array

- cDNA microarrays are usually two color arrays
- Two samples with one array
 - Two different colors Cy3/Cy5
 - Laser uses two different wave length
→ Ratio Cy3/Cy5

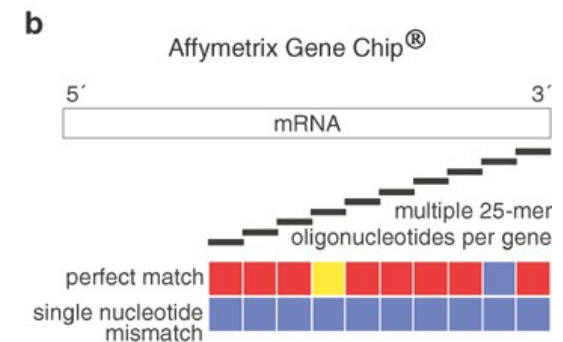
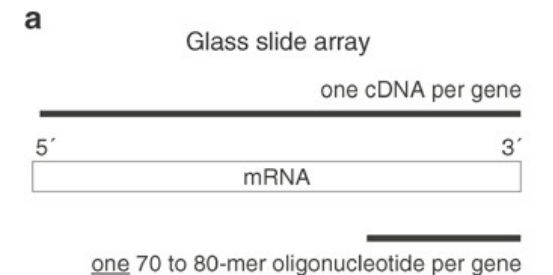


Green: high expression sample 1
Red: High expression sample 2
Black: no signal in both samples
Yellow: equal expression

<http://cmr.asm.org/content/22/4/611/F2.large.jpg>

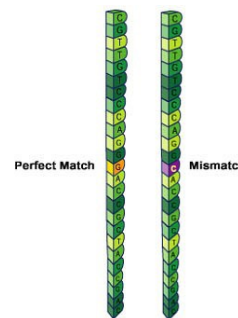
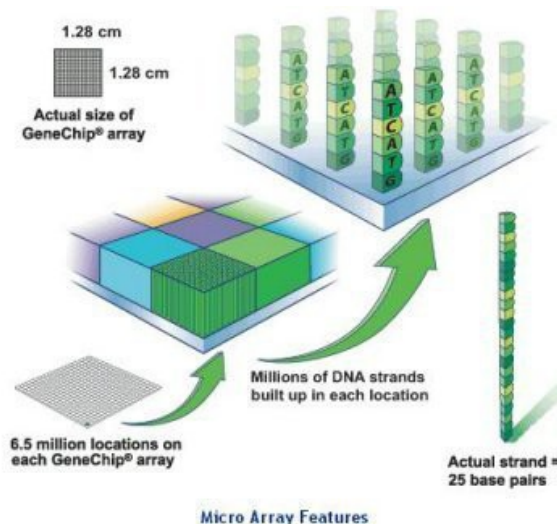
Oligonucleotide microarrays

- Short probes (25nt – 60nt) designed to **match parts of the sequence** of known genes
- 11 – 20 probes is one **probeset**
 - Represents one transcript
 - Scattered over array
- Perfect match (PM) and mismatch (MM) probes (local/global background)



Probes oligonucleotide
microarray vs. cDNA microarray

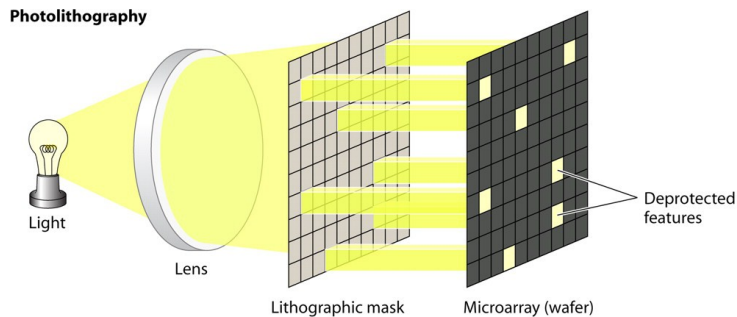
Staal, F. J. T., et al. "Leukemia 17.7 (2003): 1324-1332.



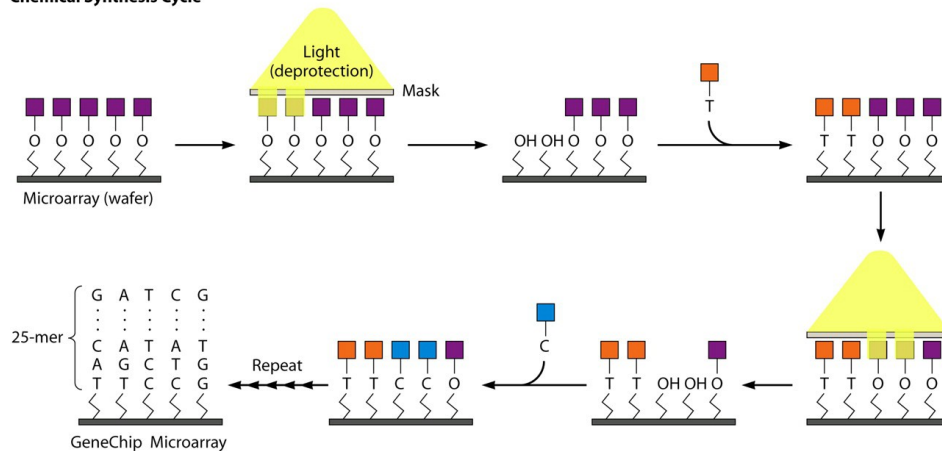
Affymetrix.com

Oligonucleotide microarray - Photolithography

Manufacturing via photolithographic synthesis



Chemical Synthesis Cycle

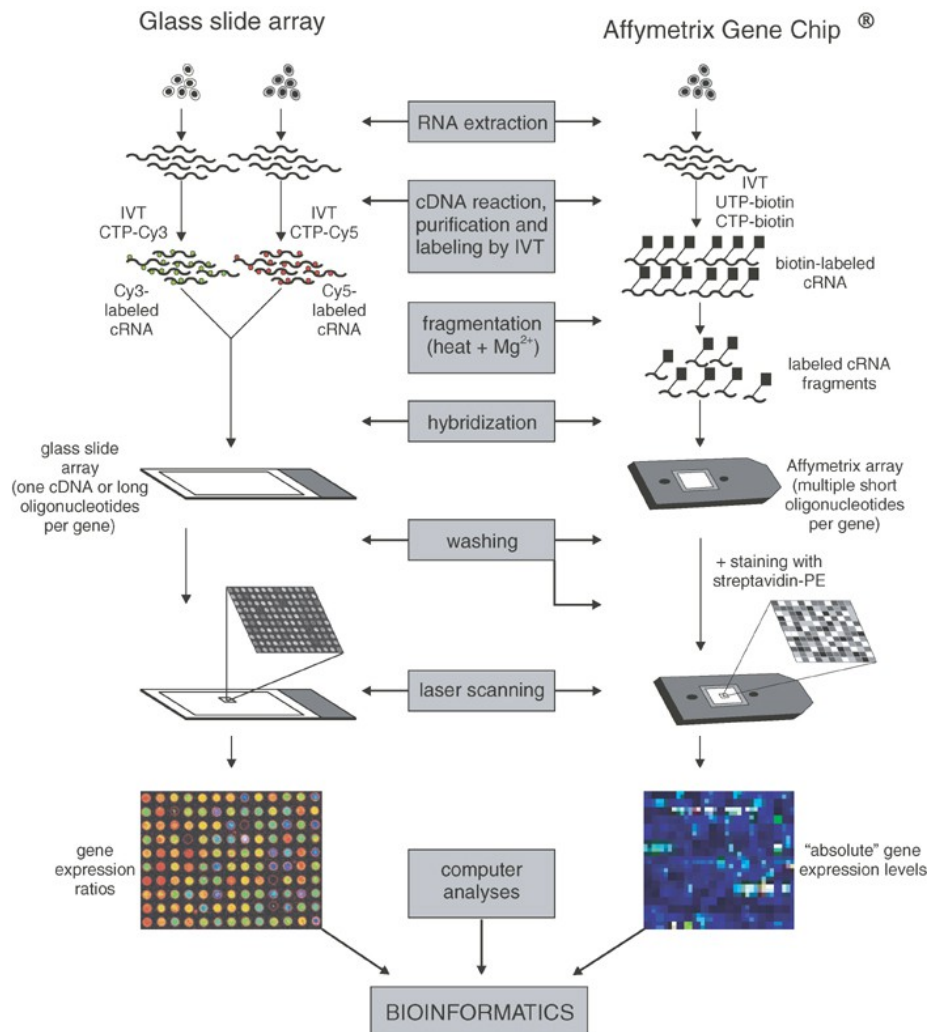


- UV light is passed through mask
→ either transmits or blocks the light from array surface
- UV light removes protecting groups
- Array is flooded with one kind of nucleotide (A, C, G or T)
- One nucleotide is added to each deprotected position
- Unbound nucleotides are washed away
- Process is repeated ~70 times

Miller MB, Tang Y-W. Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology. Clinical Microbiology Reviews. 2009;22(4):611-633. doi:10.1128/CMR.00019-09.

Reading the signal - comparison

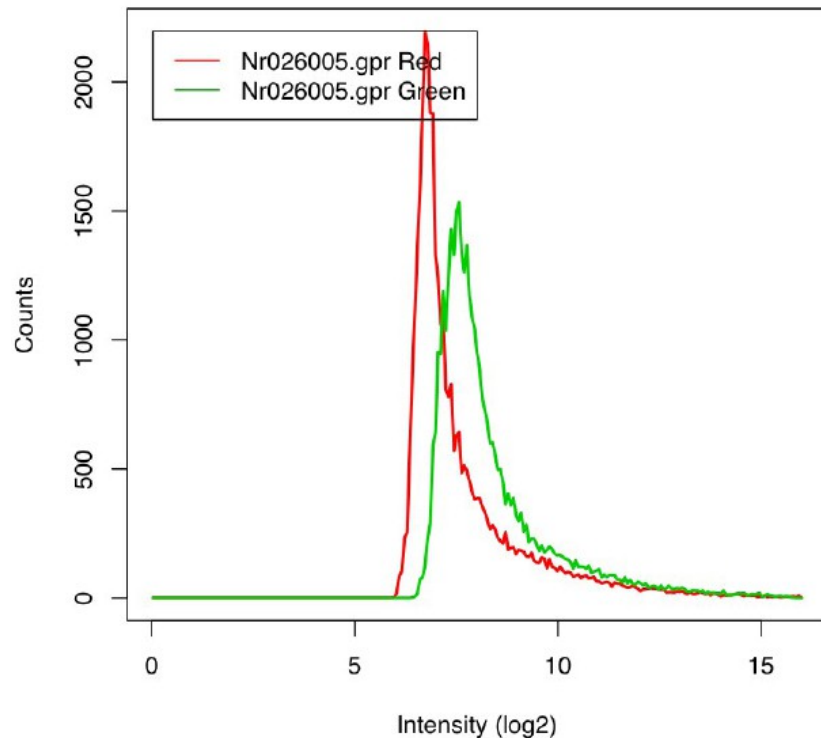
cDNA-Microarray vs. Oligonucleotide Microarray



Staal, F. J. T., et al. "DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers." *Leukemia* 17.7 (2003): 1324-1332.

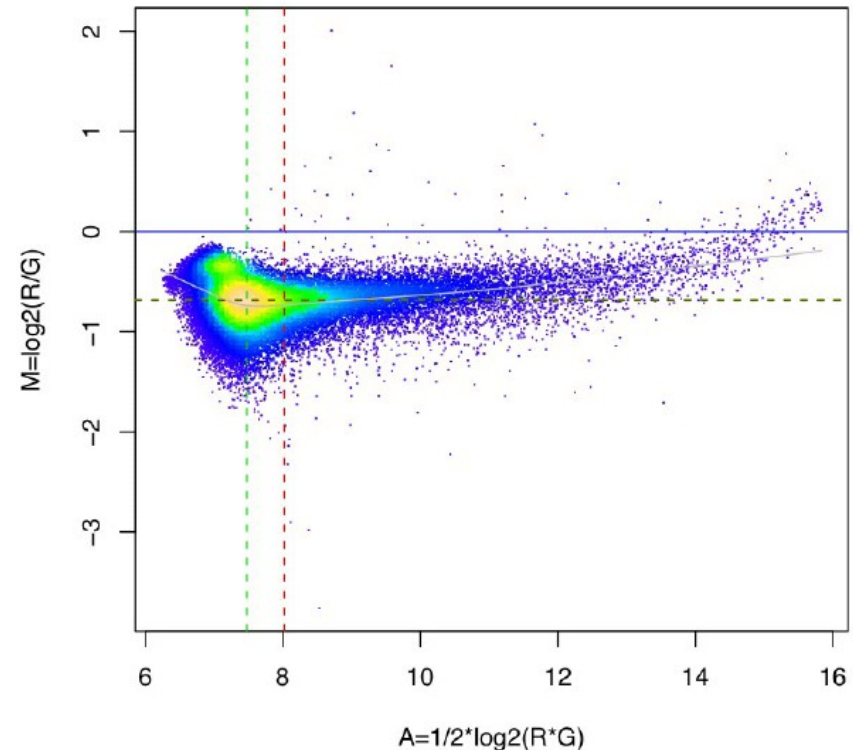
Example - Real world problem

- Dye-bias in two color array
 - Green channel appears consistently brighter than red channel
 - Intensity based



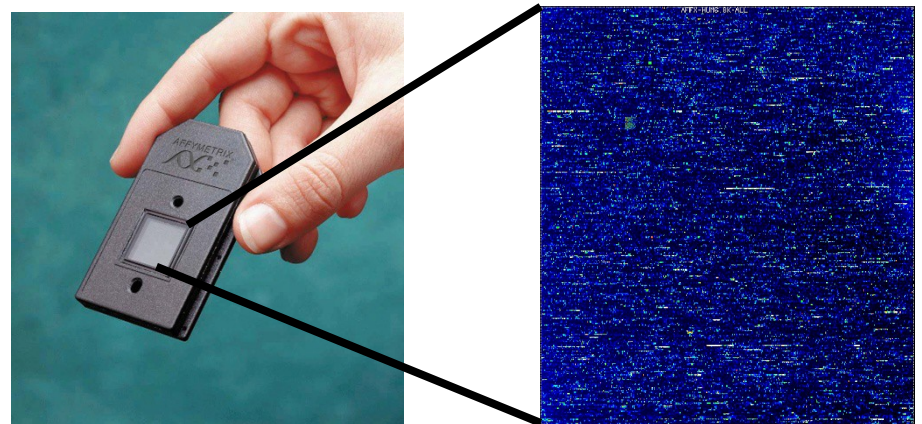
Non linear methods – Lowess

- 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



Oligonucleotide microarrays

- Industrial manufacturing
 - Prominent example: Affymetrix Microarrays
 - No customization of probes possible
 - Robustness between one array type
 - Good quality control
 - More expensive than cDNA microarrays
- Selection of good oligos is difficult



Probe selection

Requirements:

- High sensitivity (SN): strong signal if complementary target sequence is in sample solution
- High specificity (SP): Weak signal if complementary target sequence is not in solution (probe uniqueness)

$$SN = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

Criteria/binding depends on:

probe length, GC content, secondary structure, number of matches on all transcripts, probe self or cross hybridisation, position of probe in the transcript, number of matches on all transcripts, probe uniqueness (sensitivity vs. specificity), ...

<http://grf.lshtm.ac.uk/microarrayoverview.htm>

Comparison: cDNA vs Oligonucleotide Microarrays

cDNA microarrays (also: spotted arrays)

- No special laboratory equipment needed
- Good customization
- Error prone workflow
- Less spots/genes (limited redundancy)
- One cDNA or long oligonucleotides per gene
- High sensitivity: long probe sequences
- Lower specificity: cross hybridisations more probable
- Lower costs

Oligonucleotide microarrays

- Industrial manufacturing
- Densely packed
- Companies sell „kits“
- No customization possible
- Higher reproducibility (e.g. no PCR amplification needed)
- Better Quality Control
- Higher specificity
- Sensitivity: lower but multiple probes per gene as compensation
- Expensive

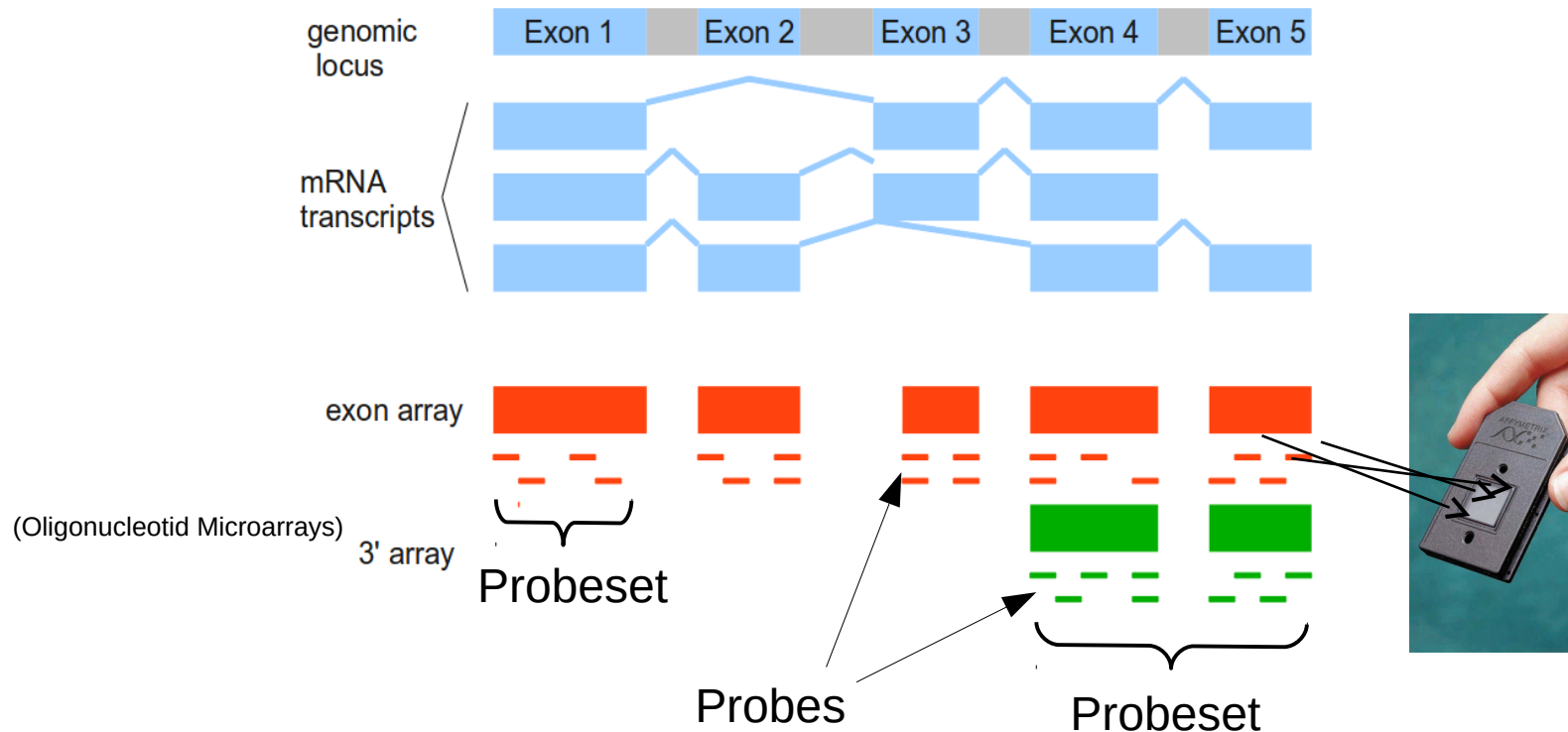
Experimental Design - Replicates

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 – Exon Array

- Gene-Expression profiling with microarrays
- Exon arrays: each exon of a gene is measured individually
- SNP arrays, ChIP-on-Chip Arrays, ...



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)
- Environment has influence on hybridization quality
- Noise: Technical replicates never produce the same data
- Transient data
 - Select appropriate time point
 - Signaling might be very fast for some processes
 - Intermediate steps are lost

Challenges

- High number of transcripts
 - Multiple testing correction
 - Choice of statistical test
- Time series results in day/night work
- Cause and effect
 - Tumors have high cell proliferation
 - EGF, p53 likely highly expressed
- Biological interpretation difficult

This Lecture

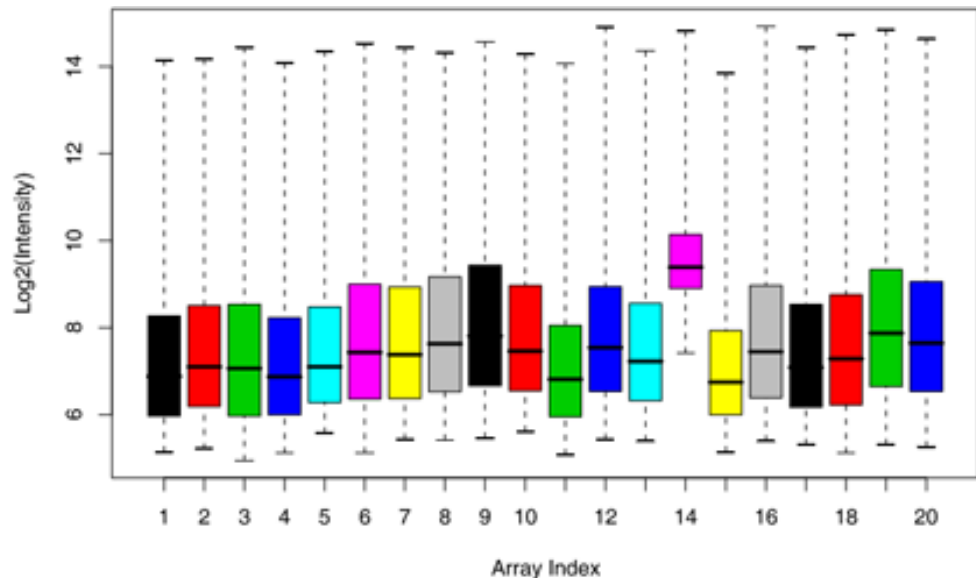
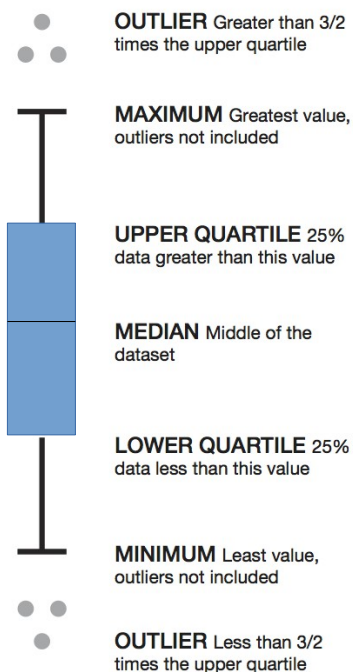
- Protein synthesis
- Microarray
 - Idea
 - Technologies
 - Problems
- Quality control
- Normalization
- Analysis next week!

Data Visualization, Quality Control

- Detect arrays with poor quality (outliers)
- Identify arrays behaving different than others

Boxplot

Estimate the homogeneity of data

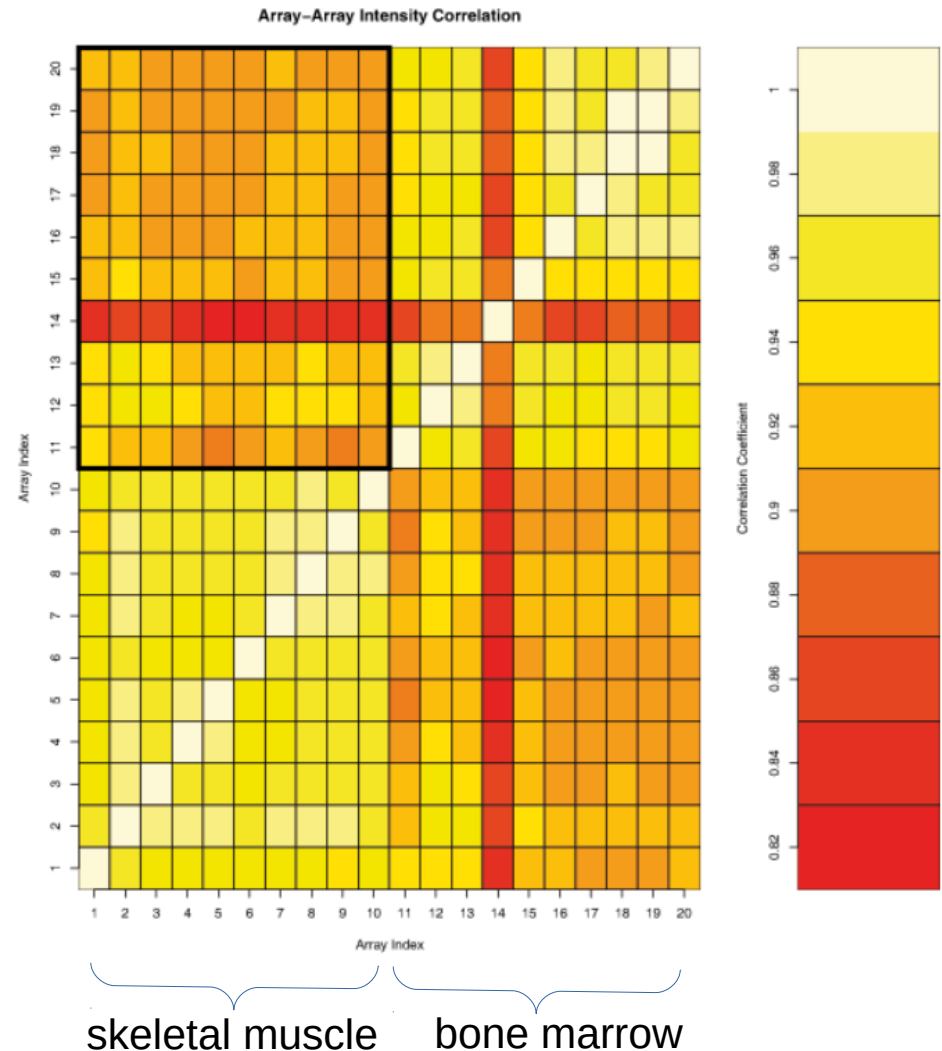


<https://genevestigator.com>

Array 14: overall higher signal intensity

Array-Array Correlation Plot

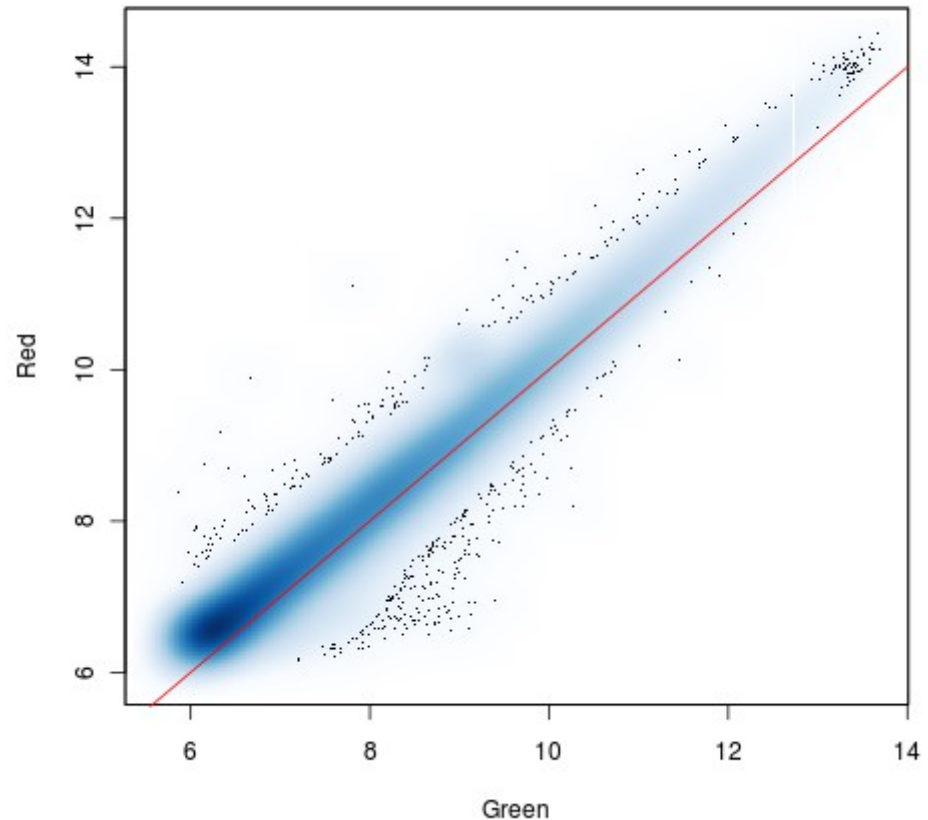
- correlate all arrays from an experiment with each other
→ replicates should show high correlation
- Example: Array 14 poor correlation
→ possible reasons:
higher noise in the data,
stronger background, ...



Data visualization, Quality control

Scatter Plot:

- Each point represents one transcript in two experimental settings
- Most points should appear around the horizontal line (only a few genes are expressed at different levels)
- Higher variation with low intensities



MA-Plot

- 45° rotated version with subsequent scaling of the scatter plot
- Log2 Fold Change or M-value
 - Is the \log_2 ratio between two values
 - Log-values are symmetric
 - Visual interpretation (Difference between 4 to 16 vs. 0.25 to 0.0625)

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

- For example:

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

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

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

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

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

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

MA-Plot

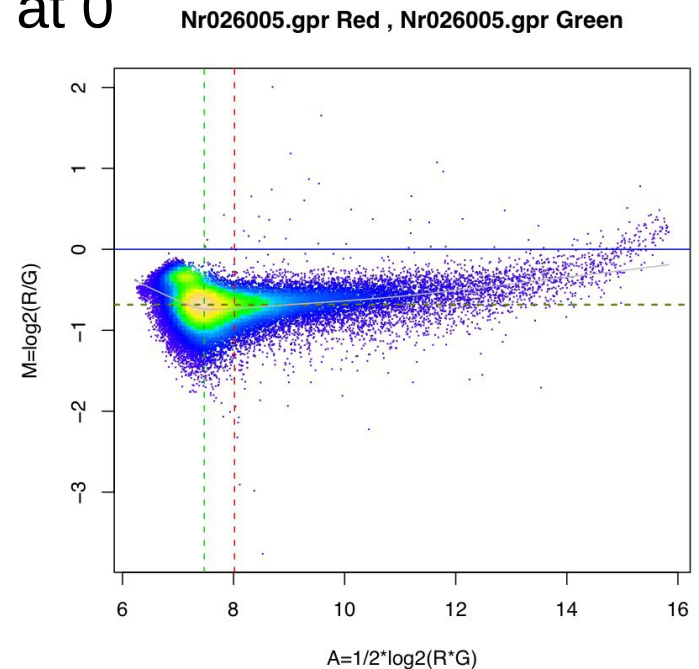
- A-Value is the **logarithm of the intensity mean value**

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

- Points should be located on the y-axis at 0
- Bana shape in two-color arrays:
green brighter than red

Further quality control possibilities:

- Image analysis
- RNA degradation plots
- Residual plots
- PCA
- ...



This Lecture

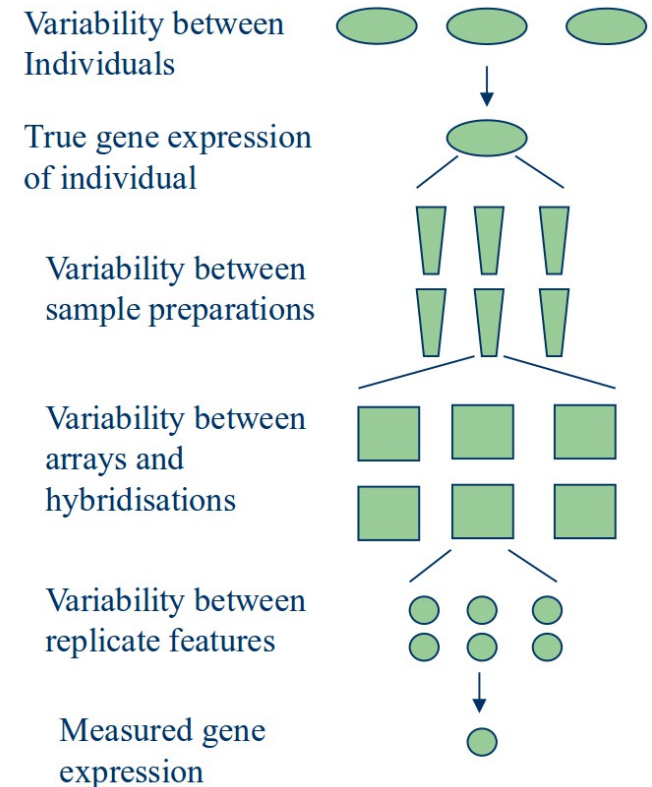
- Protein synthesis
- Microarray
 - Idea
 - Technologies
 - Problems
- Quality control
- Normalization
- Analysis next week!

Normalization

Microarrays are **comparative experiments**, **BUT** measurements between two experiments are **not directly comparable**

Several levels of variability in measured gene expressions:

- Highest level: biological variability in the population from where the sample derives
- Experimental level: variability between
 - preparations and labelling of samples (different amounts of RNA or dye, experimenter variability, day/night work ...)
 - hybridisations
 - The signal on replicate features on the same array (probe affinity)
- Further sources: different scanner settings, ..



<http://slideplayer.com/slide/2394201/>

Normalization

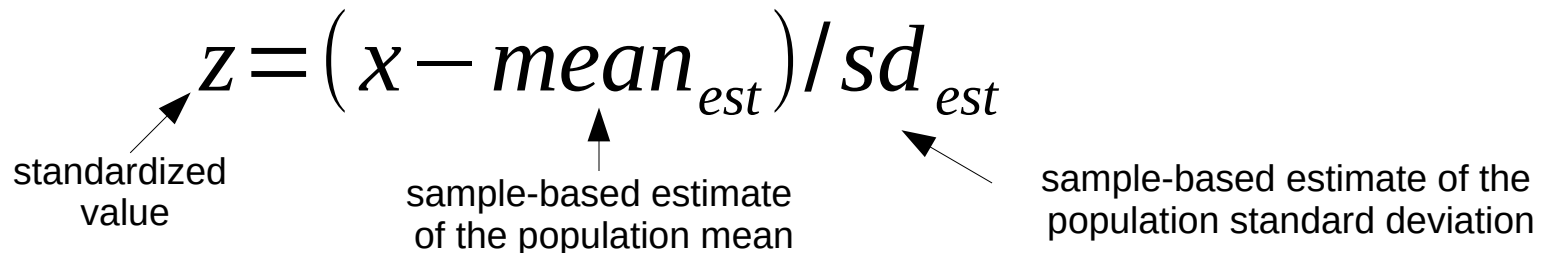
- Aim: identification of the real biological differences among samples (compensation for minor technical divergence)
- **Assumptions**
 - **overall number of mRNA molecules changes not „*much*“ (*linear*) between samples**
 - **only a few genes expressed differently between cohorts**
- (not entirely true for comparing highly transformed cancer cells with normal cells)
- Two-color arrays: normally within array normalization followed by across array normalization

Normalization

- mRNA in a sample
 - Assumption: „cells contain same relative composition of RNA“
 - Measure total mRNA, divide
- Z-Score (mean)
 - Standardization: set mean to 0 and standard deviation to 1:
 - Centering: subtract the mean from each value
 - Scaling: divide the centered value by the standard deviation

$$z = (x - mean_{est}) / sd_{est}$$

standardized value sample-based estimate of the population mean sample-based estimate of the population standard deviation

The diagram shows the Z-score formula $z = (x - mean_{est}) / sd_{est}$. Three arrows point from descriptive text below to the variables in the formula: an arrow from 'standardized value' points to 'z', an arrow from 'sample-based estimate of the population mean' points to 'mean_{est}', and an arrow from 'sample-based estimate of the population standard deviation' points to 'sd_{est}'.

- Z-Score (median)

Quantile normalization

- Normalize so that the quantiles of each array are equal
 - Distribution of expression values on each microarray is made identical
- Simple and fast algorithm
- Usually outperforms linear methods

Steps:

- Given a matrix X with $p \times n$ where each array is a column and each transcript is a row
- Sort each column of X separately to give X_{sort}
- Take the mean across rows of X_{sort} and create X'_{sort}
- Get X_n by rearranging each column of X'_{sort} to have the same ordering as the corresponding input vector

Quantile normalization

		Sort					Replace					Reorder				
		E1	E2	E3	E4	E5	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
Values	V1	1	11	13	29	26	21	28	30	29	27	28	28	28	28	28
	V2	15	17	5	8	14	18	23	16	24	26	23	23	23	23	23
	V3	21	2	12	20	25	15	19	13	22	25	19	19	19	19	19
	V4	10	19	16	24	4	10	17	12	20	14	14	14	14	14	14
	V5	18	28	3	22	27	7	11	5	8	9	8	8	8	8	8
Indexes		7	23	30	6	9	1	2	3	6	4	3	3	3	3	3
		1	1	1	1	1	3	5	6	1	5	3	5	6	1	5
		2	2	2	2	2	5	6	4	4	1	5	6	4	4	1
		3	3	3	3	3	2	4	1	5	3	2	4	1	5	3
		4	4	4	4	4	4	2	3	3	2	4	2	3	3	2
		5	5	5	5	5	6	1	2	2	6	6	1	2	2	6
		6	6	6	6	6	1	3	5	6	4	1	3	5	6	4

- + differences between the separate values are retained
- + identical distribution for each array
- some data can be lost, especially in the lower signals

Quantile normalization

Boxplot for raw data (left) and normalized (right) data from previous slide

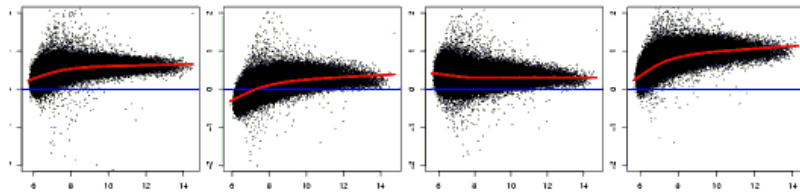
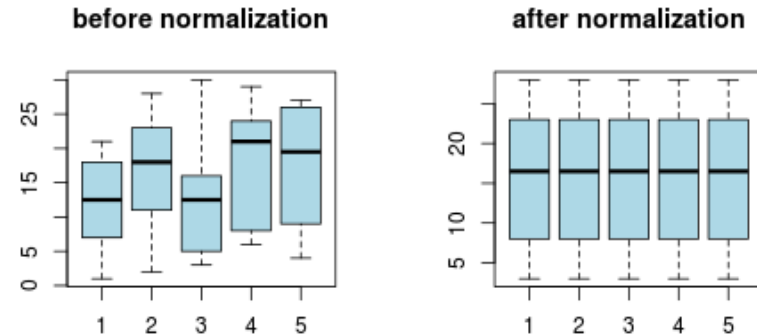


Figure 7A. Ratio Intensity Plot of all probes for four pairs of chips from GeneLogic spike-in experiment

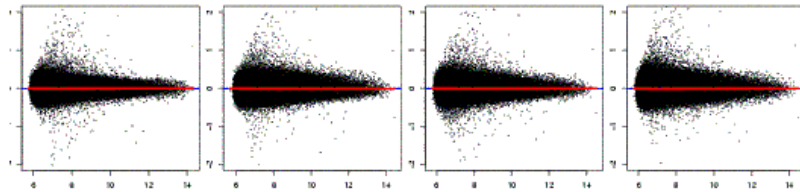


Figure 7B. As in A, after normalization by matching quantiles. Both figures courtesy of Terry Speed

MA Plots before (top) and after (bottom) quantile normalization (red line: lowess line)

Bolstad, Benjamin M., et al. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." *Bioinformatics* 19.2 (2003): 185-193.

RMA – Robust Multichip Average

- Used for Affymetrix microarray normalization
- Three steps:
 - 1) Background correction (for each array separately)
 - 2) Quantile normalization (across all arrays)
 - 3) Summary of probesets (median-polish)

→ returns normalized data in log2 scale

RMA does not use the Mismatch probe intensity information (MM signal often higher than PM signal)

Computing RMA: affy package for R
(Exercise next week!)

Further normalization possibilities:

- Non linear methods like Lowess (two-color, non-linear, within array)
- Statistical approaches like MAS5 or VSN
- ...