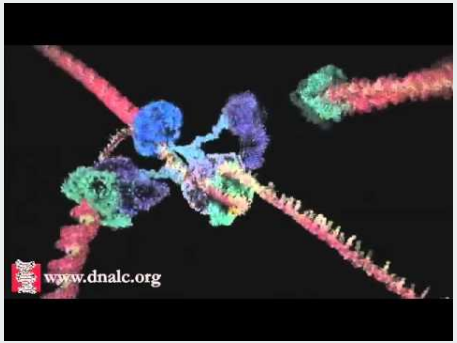
Measuring gene expression

Grundlagen der Bioinformatik SS2018



https://www.youtube.com/watch?v=v8gH404a3Gg

Agenda

- Organization
- Gene expression
  - Background
- Technologies
  - FISH
  - Nanostring
  - Microarrays
  - RNA-seq
- How to detect technological biases
  - Visualization
  - Quality control
  - Normalization

Shift date of next lecture & thursday's exercise

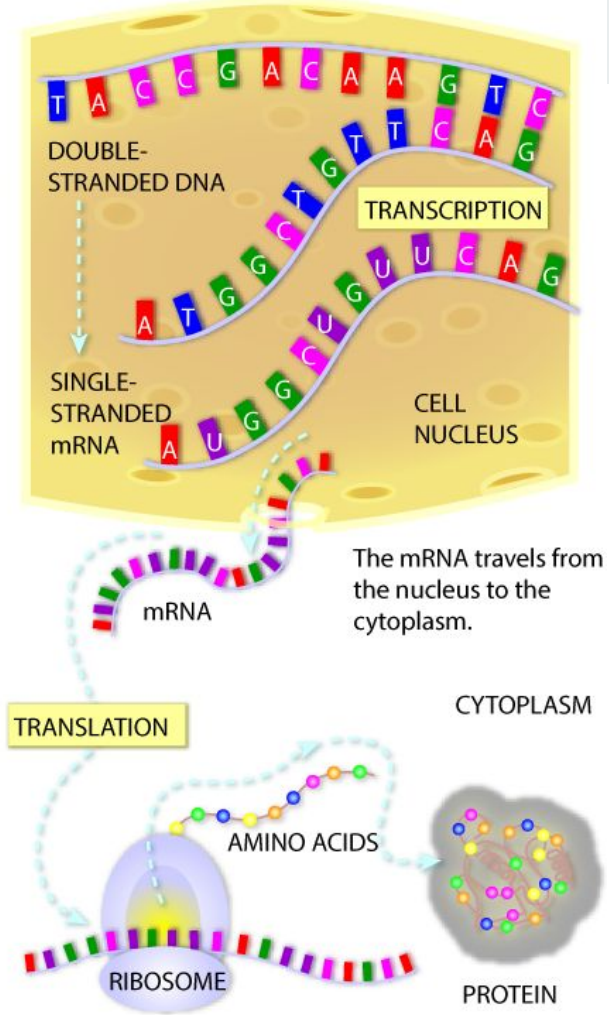
Move next Thursday's lecture and excersise to Friday the 15th

- Lecture 9 a.m.
- Exercise 1 p.m. (11 a.m. - 1 p.m. = Friday's exercise)

Gene Expression - Background

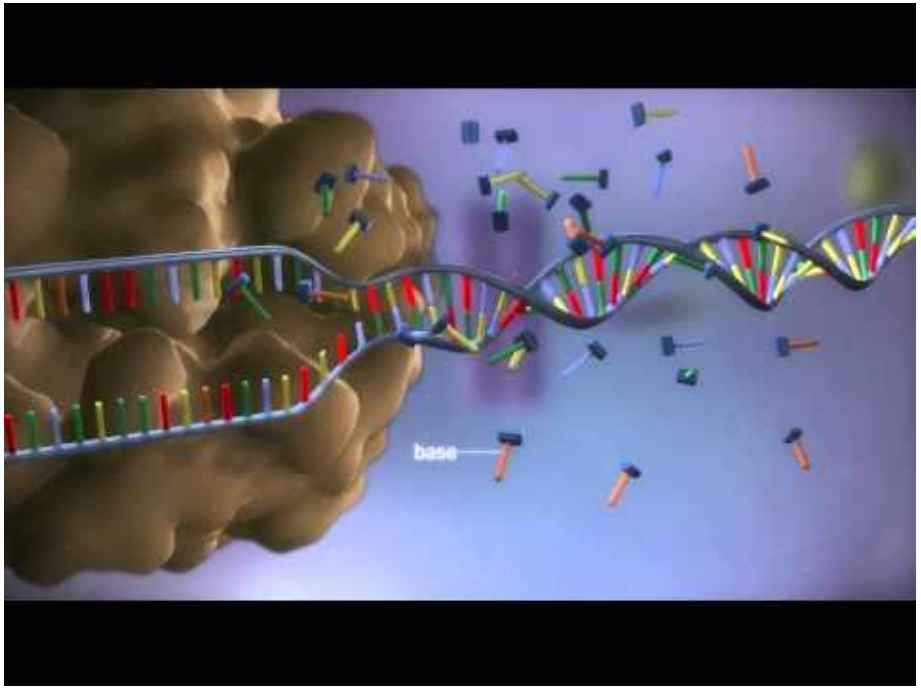
Gene Expression - mRNA

- mRNA expression ∝ gene activity
- Protein ~ active *form* of genes
- mRNA = messenger RiboNucleic Acid
- DNA->mRNA-> Protein



http://learn.genetics.utah.edu/content/basics/

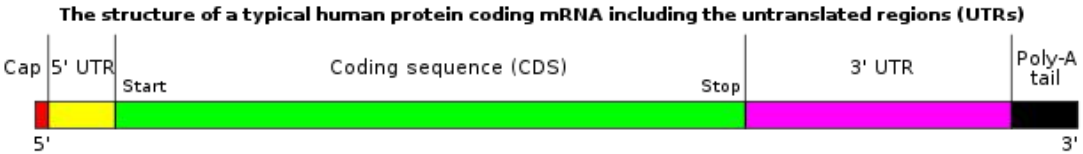
Video time



https://www.youtube.com/watch?v=gG7uCskUOrA

mRNA structure

- RNA copy of DNA gene
  - Modified copy -> not identical
- Has specific sequence of bases that determine proteine
- Has additional cap and end
  - E.g. Poly-A tail
- Only parts are translated
- Aim: Detect mRNA expression



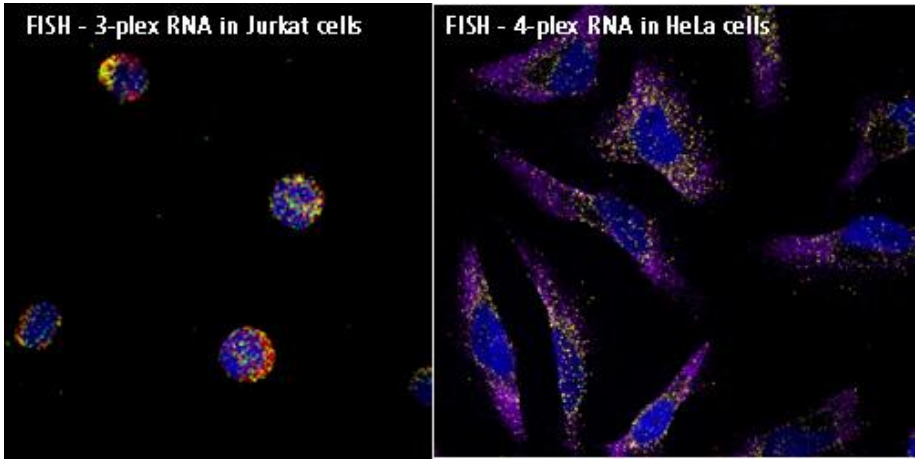
Simplified mRNA structure

Wikicommons

Technologies

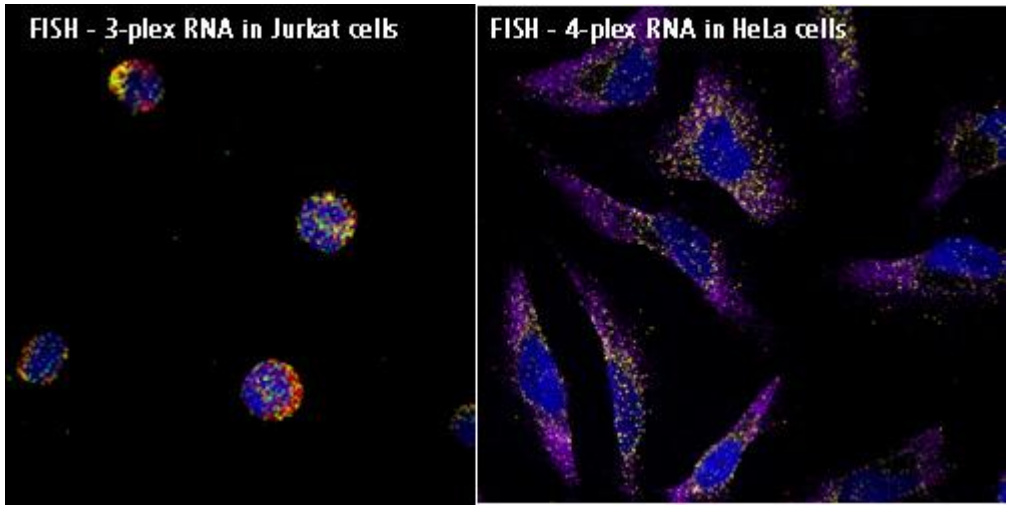
Fluorescence In Situ Hybridization

- Fluorescence in situ hybridization = FISH
- Illuminate mRNA
- Qualitative -> no count information
- Match sequence
- Low throughput



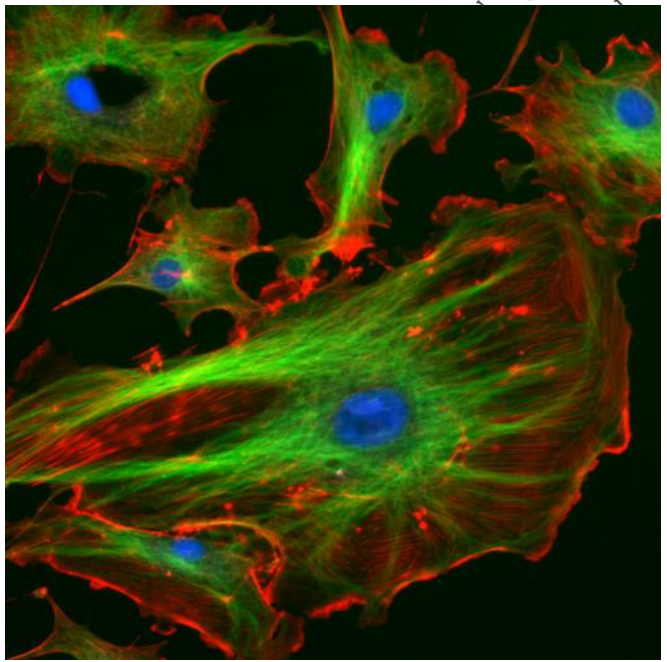
wikicommons

FISHy impressions



Ryan Jeffs

Illumination of RNA via FISH
Colors specific for mRNA
-> location detection

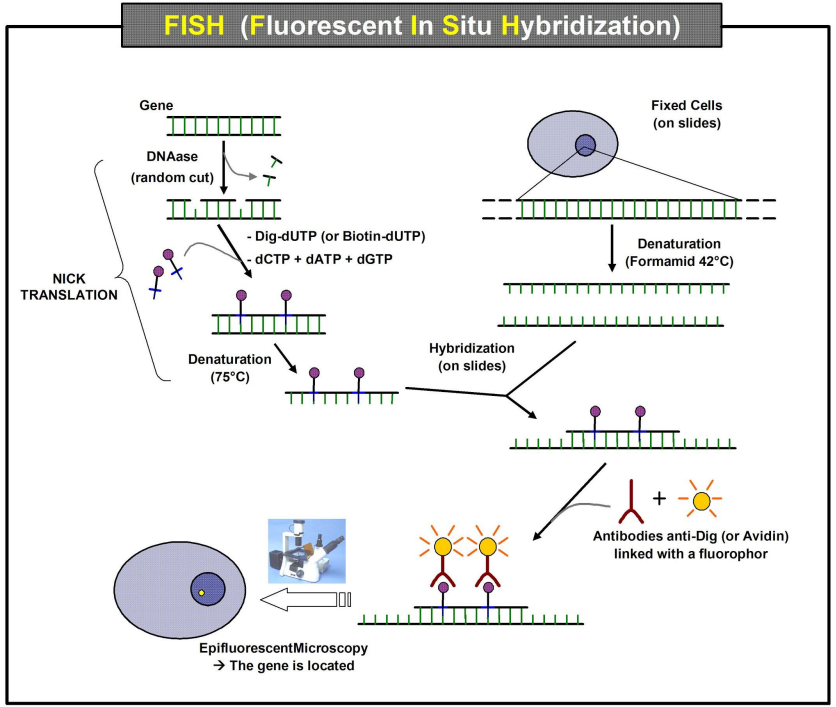


Nucleus, skeleton &
Cell-membrane

FISH method

- Here: shown for **DNA**

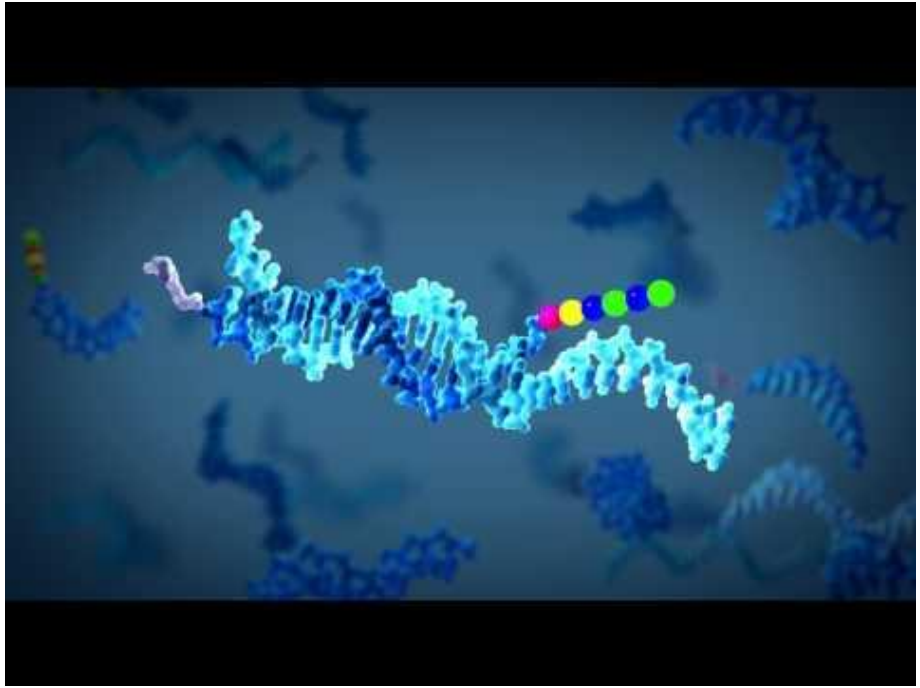
1. Cut DNA and paste anchor
2. Denature DNA
3. Hybridize
4. Attach antibody and shine



wikicommons

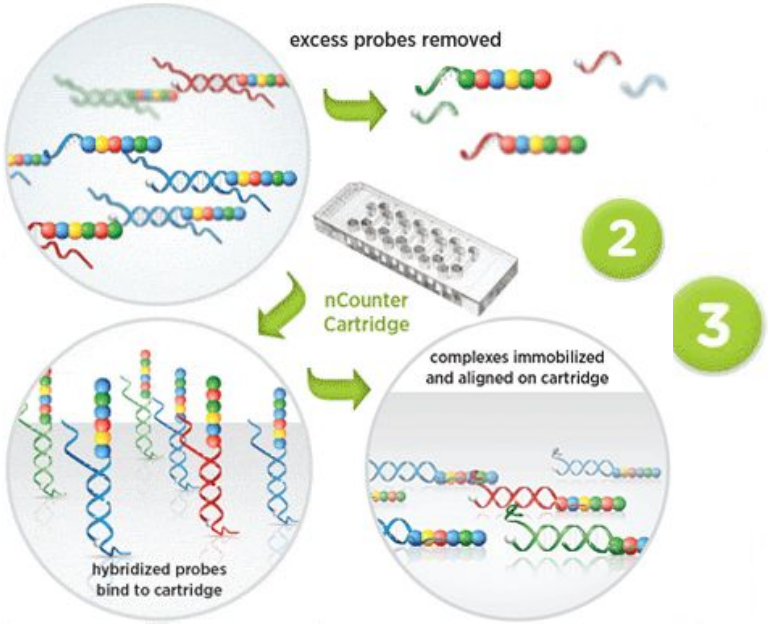
FISH / NanoString

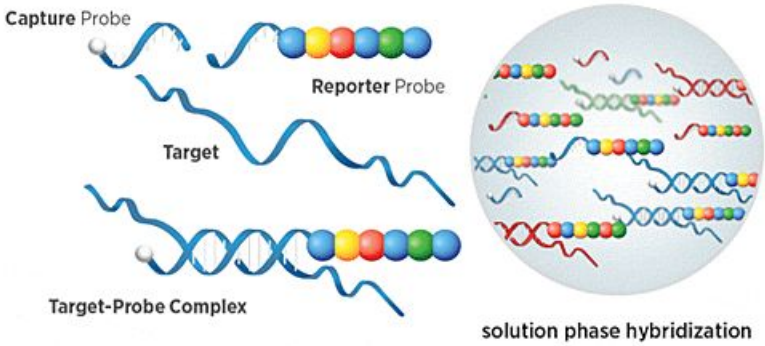
- Quantitative FISH-like -> counts available
- Separate capture
- Sequence matched
- Medium throughput

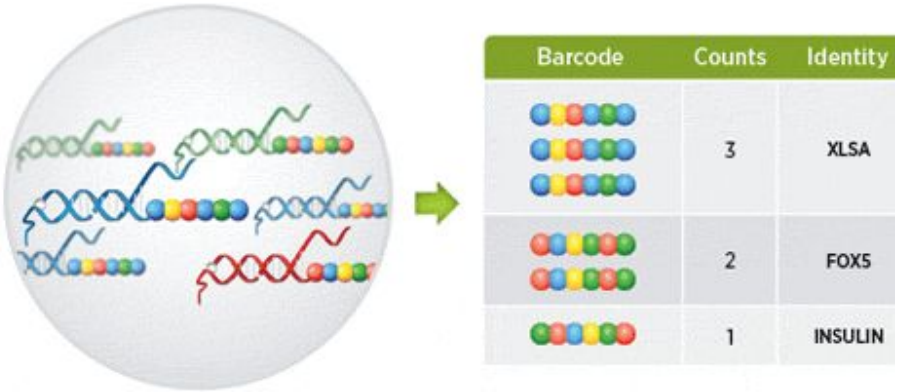


https://www.youtube.com/watch?v=XIVmmfujiro
>= 1 minute 22 seconds

Nanostring

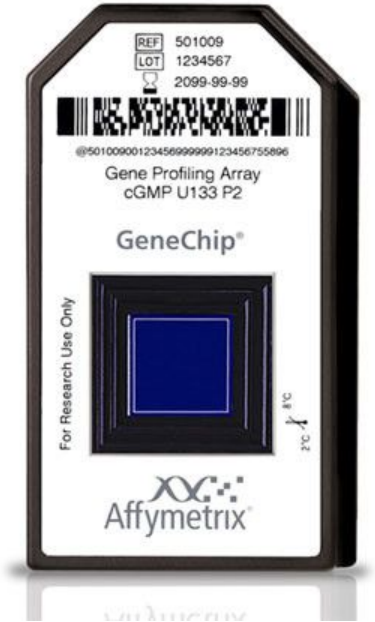

excess probes removed
nCounter Cartridge
complexes immobilized and aligned on cartridge
hybridized probes bind to cartridge


Capture Probe
Reporter Probe
Target
Target-Probe Complex
solution phase hybridization


Barcode
Counts
Identity
3
XLSA
2
FOX5
1
INSULIN

mRNA Micro-Arrays

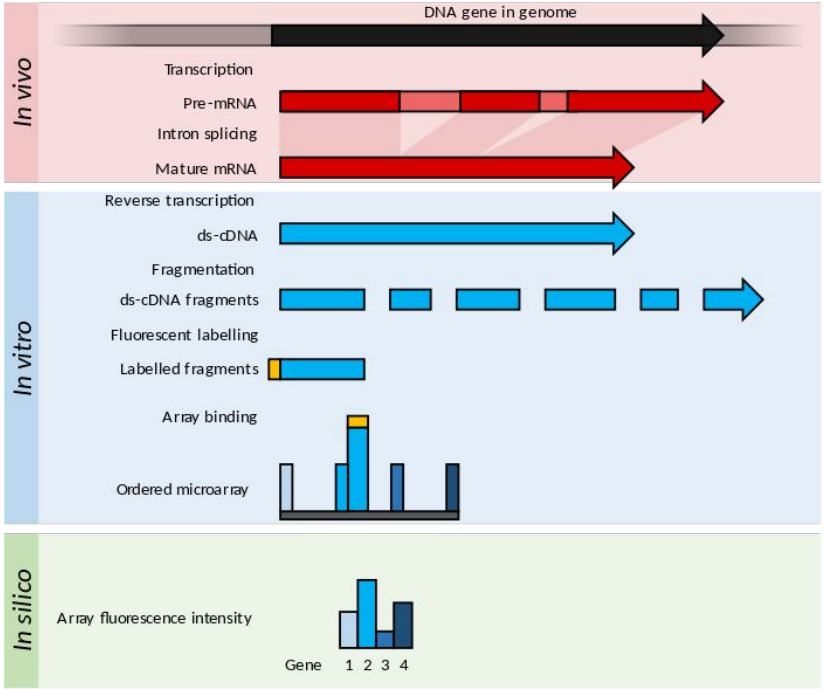
- Oligo-nucleotide arrays
- Array of pre-defined sequences
- Complementarily binding to mRNA
- mRNA illuminated
    - Expression measured as light-intensity



www.affymetrics.com

Workflow mRNA array

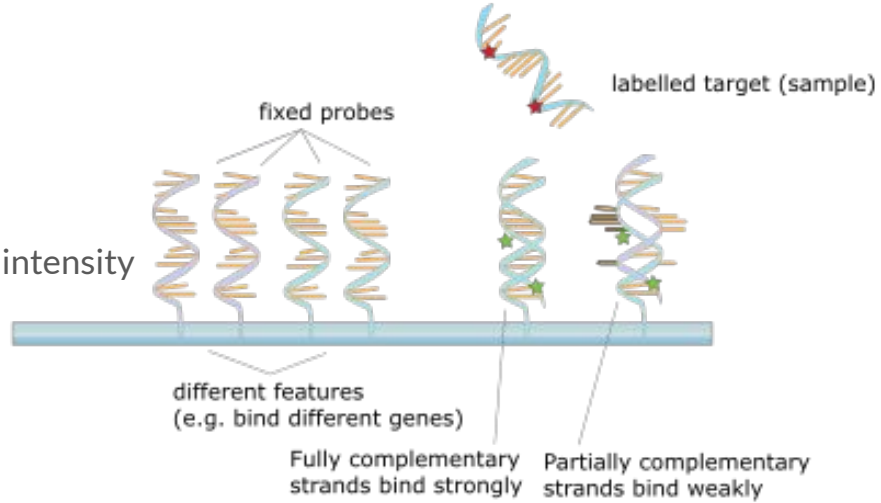
1. Isolation and purification
2. Reverse transcription
    a. cDNA == complementary DNA
3. Labelling fluorescent dye cDNA labeling
4. Hybridization
    a. Washing
5. Scanning
    a. Laser excitation
    b. detection of light intensities
    c. image segmentation
6. Normalization



wikicommons

Hybridization

- Binding of free mRNA by pre-defined probe sequences
- Targets mRNA sequences labeled
- Amount matches / mismatches determines illumination intensity



wikicommons

Probe sequence selection

Trade-off Sensitivity versus Specificity

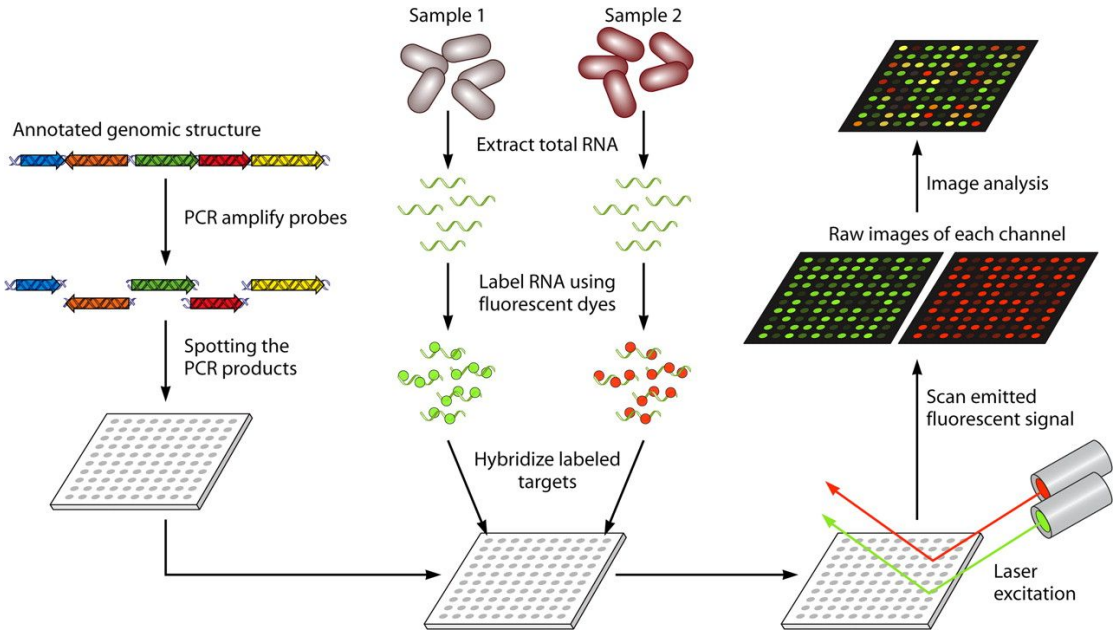
- Sensitive sequence may not be specific
  - E.g. cap or poly-A tail sequences
- Sensitivity := TP / (TP + FN)
- Specificity := TN / (TN + FP)
- Interesting optimization problem

Probe-hybridization subject to plethora of factors

- Probe length
- GC content
- Secondary structure
- Amount matches over all transcripts
- Probe self or cross hybridisation
- Position of probe in the transcript
- Probe uniqueness
  - Sensitivity vs. specificity

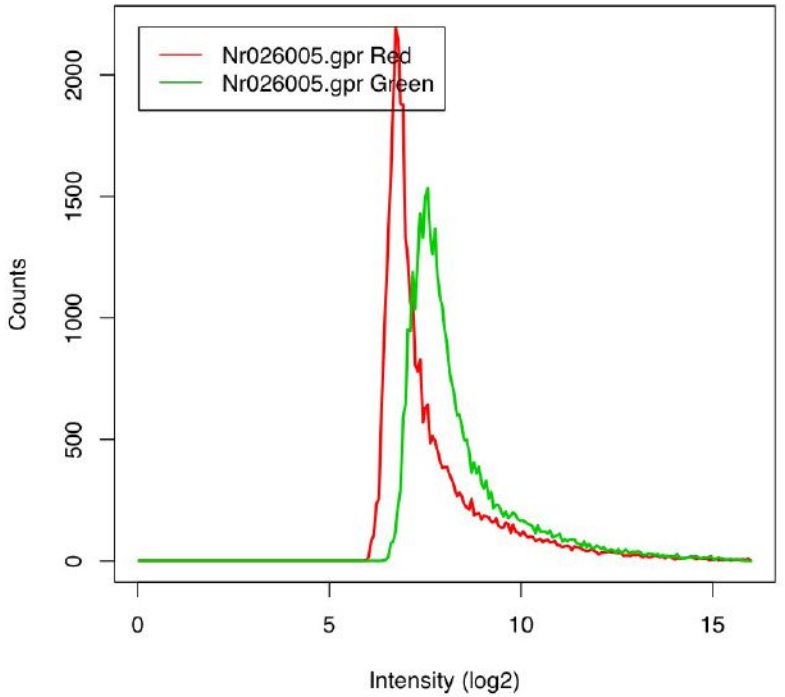
Two color array

- Expressed in sample 1
- Expressed in sample 2
- Expressed in samples 1 & 2
- Not expressed in samples 1 & 2



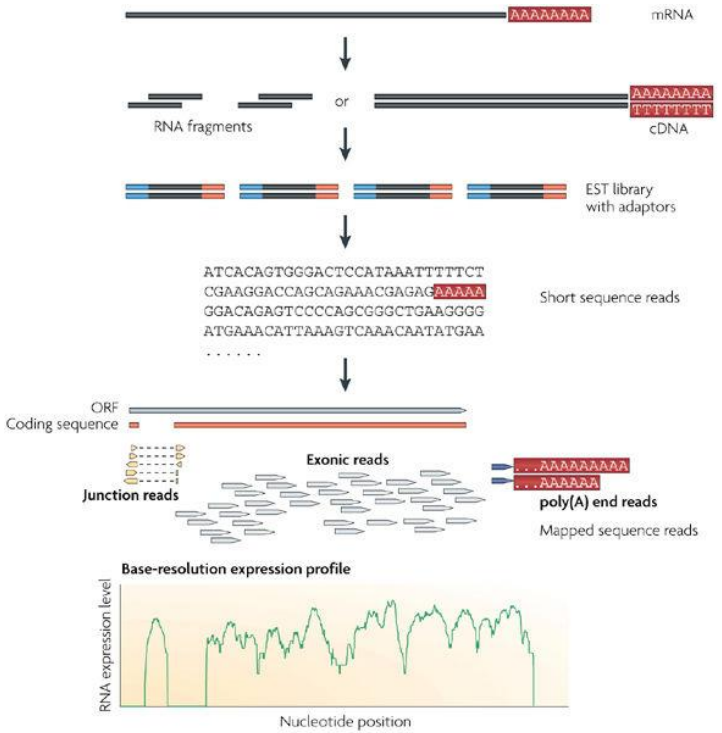
Structural dye-bias two-color array

- Distortion of expression measurement
- Green channel consistently brighter than red channel
- Intensity-dependent



RNA-seq

1. mRNA library preparation
    a. Shotgun-sequencing or
    b. cDNA-sequencing
2. Amplification fragments (PCR)
3. Map reads to genome
4. Count reads per gene



Wang et al., *Nature Reviews Genetics* volume10, pages57–63 (2009)
doi:10.1038/nrg2484

Nature Reviews | Genetics

Comparison Arrays vs. RNA-seq

Arrays		RNA-seq	
✓	Cheap	✘	Expensive
✓	Standardized	✘	Non-standardized
✓	Well understood	✘	Still subject to active research
✘	Limited to know genes	✓	Detects all genes
✘	Limited detection range	✓	Dynamic range
✘	Non-specific hybridization	✓	Specific detection

Summary technologies

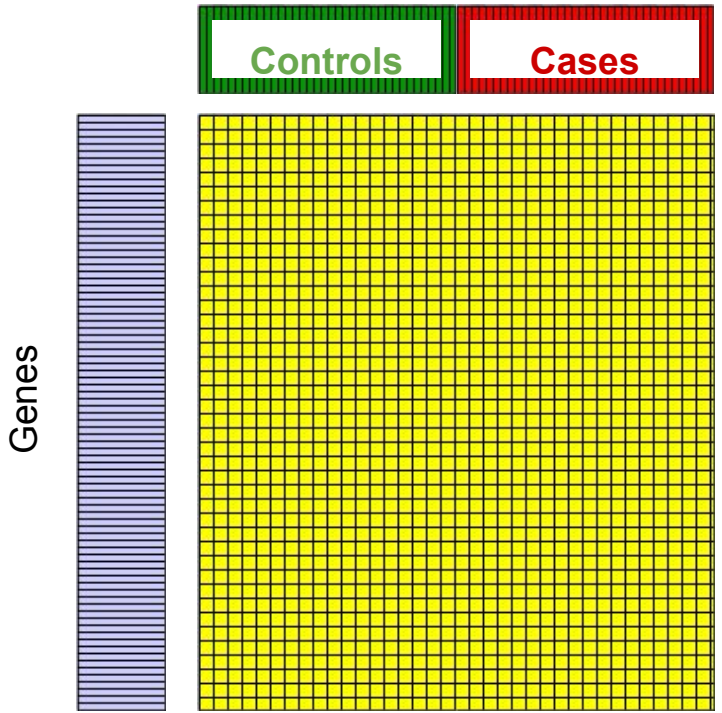
Technology	Type	Price	Amount genes	Supervised*
FISH	Qualitative	Low	Small	Yes
mRNA-Array	Qualitative/ Quantitative	Low	Large	Yes
NanoString	Quantitative	Medium	Medium	Yes
RNA-seq	Quantitative	High	Very large	No

*Supervised := Can only detect what we actively look for
Unsupervised := Can detect novel mRNA transcripts

Methods

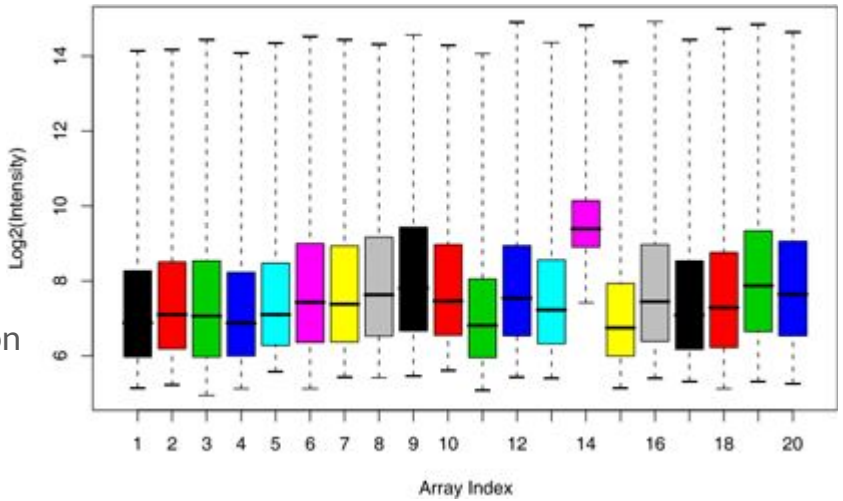
mRNA experiment design

- Two or more groups (called cohorts)
  - Control
  - Case
- Identify aggregated expression within cohorts
- Identify differences between aggregated expressions
- Ensure that measurements are comparable



Visualization - Boxplot

- Data overview
- Outlier identification
- Homogeneity-estimation



OUTLIER Greater than 3/2 times the upper quartile

MAXIMUM Greatest value, outliers not included

UPPER QUARTILE 25% data greater than this value

MEDIAN Middle of the dataset

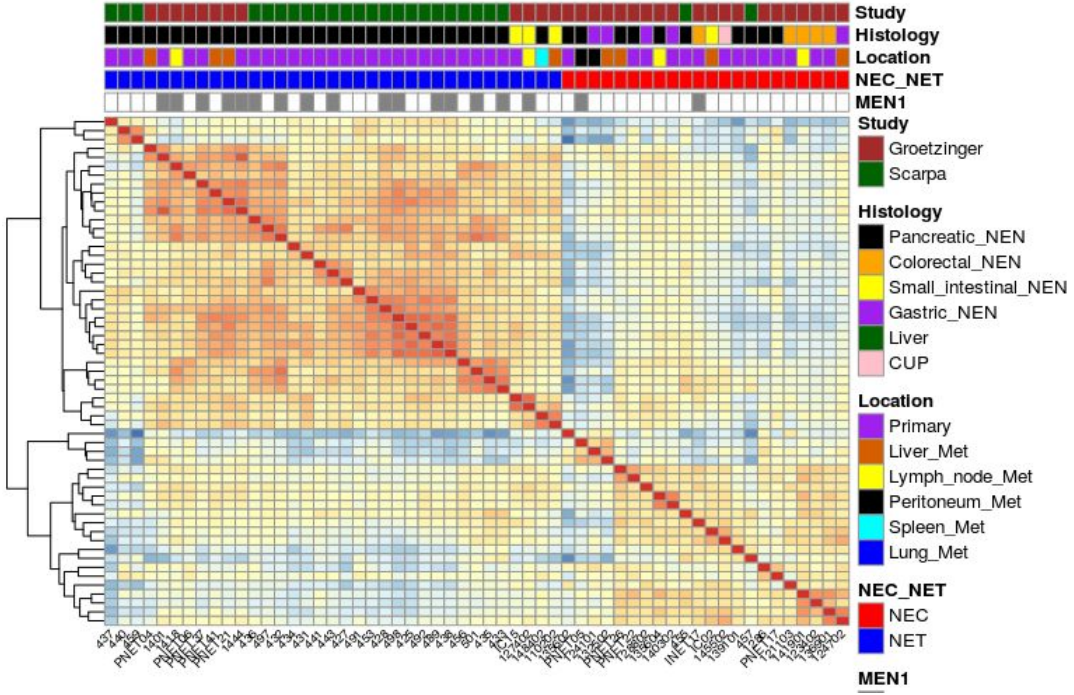
LOWER QUARTILE 25% data less than this value

MINIMUM Least value, outliers not included

OUTLIER Less than 3/2 times the upper quartile

Visualization - Correlation heatmap

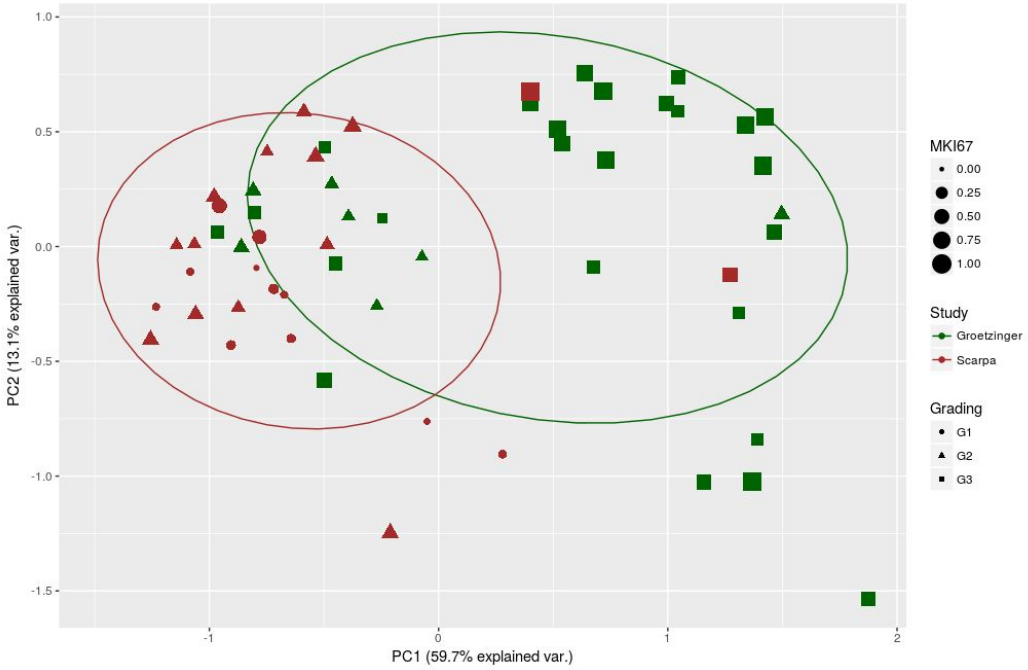
- *Pairwise-similarity* of samples
- Clustering informative
  - Bad: clustering based on study
  - Good: clustering based on cancer-type
    - NEC (Carcinoma) vs NET (Tumor)



Real-world heatmap

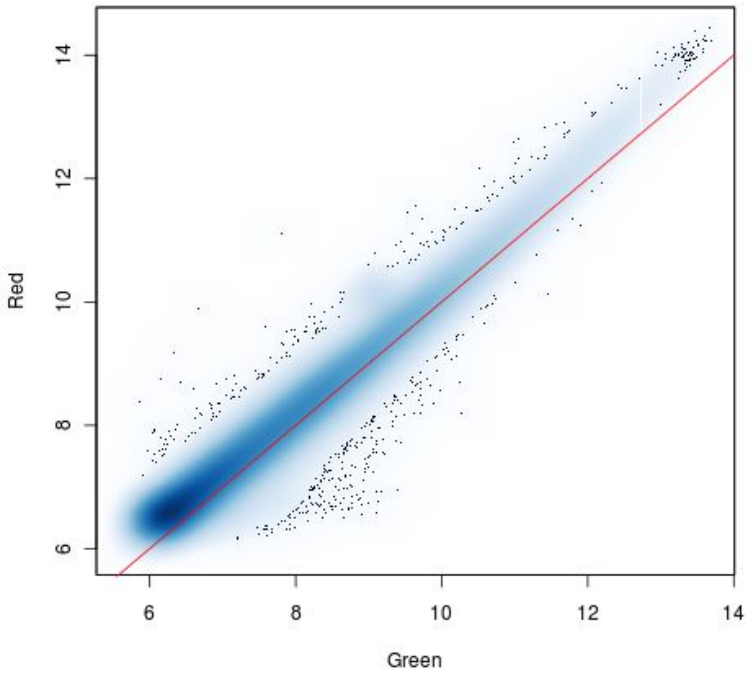
Principal component analysis (PCA)

- Two-dimensional similarity of samples
- Clustering
- Principal effects on data shown in
  - PC1 (greatest effect)
  - PC2 (second greatest effect)



Scatter plot

- Dot := one transcript in two experimental settings
- Points should appear around the horizontal line
  - only a few genes are expressed at different levels
- Higher variation with low intensities

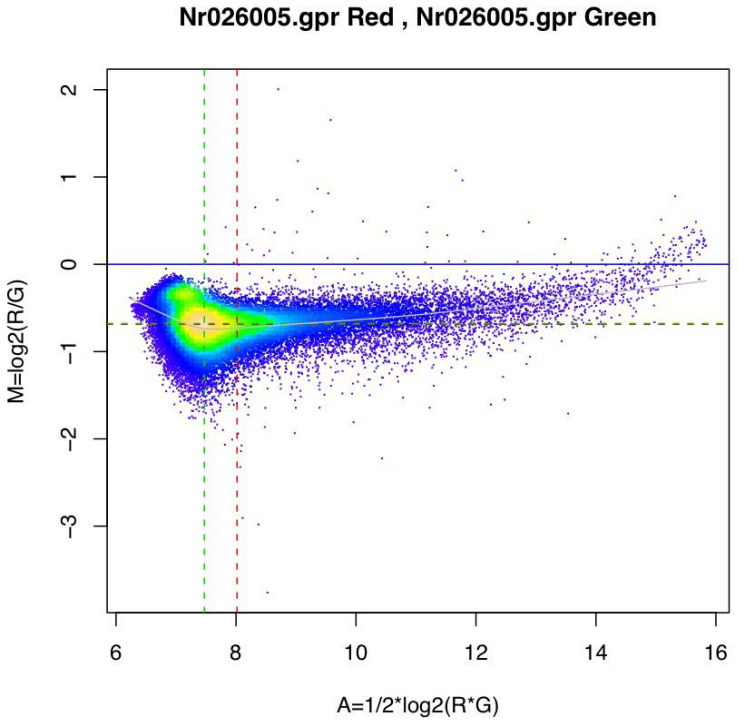


Mean-average (MA)-plot

- Visualization relationship mRNA expression vs.

  Log_2 expression difference

- Bias-correction two-color array
  - Banana-shape indicates bias
  - Shift signal to zero -> bias-correction
- Modified scatter plot
  - 45° rotated
  - Scaled



M & A calculation

M := Log_2 fold change (difference)

FC($Value_1$ / $Value_2$) := log_2 ($Value_1$ / $Value_2$)

FC(512 / 1024) := log_2 (512/1024) = -1

FC(123 / 123) := log_2 (123/123) = 0

FC(512/ 256) := log_2 (512/256) = 1

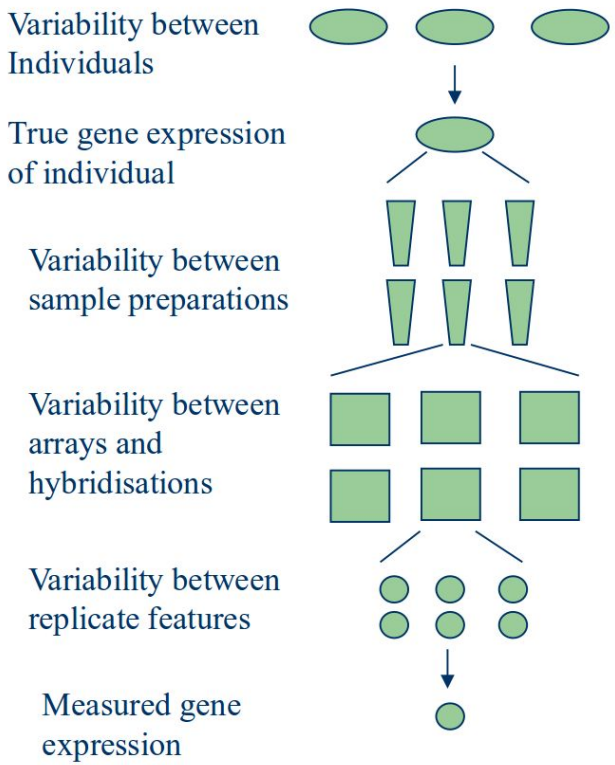
A := logarithm of mean expression intensity

A := 0.5 * ($log_2 Value_1$) + log($Value_2$))

A := 0.5 * (log_2 4) + $log_2$2) == 1.5

Motivation normalization

- Interested in: true biological difference of mRNA expression
- What we measure: Mixture of (unwanted) technical and biological noise
- Correct undesired noise!



Z-score normalization

- Correct for different amount of mRNA per sample
- Z-score = scaling of counts
  - 0 = average
- Examples: 2, -1, 0.1

$Z = (X_i - mean_{est}) / sd_{est}$

X_i = expression gene i

$Mean_{est}$: (estimated) expr. average over all genes

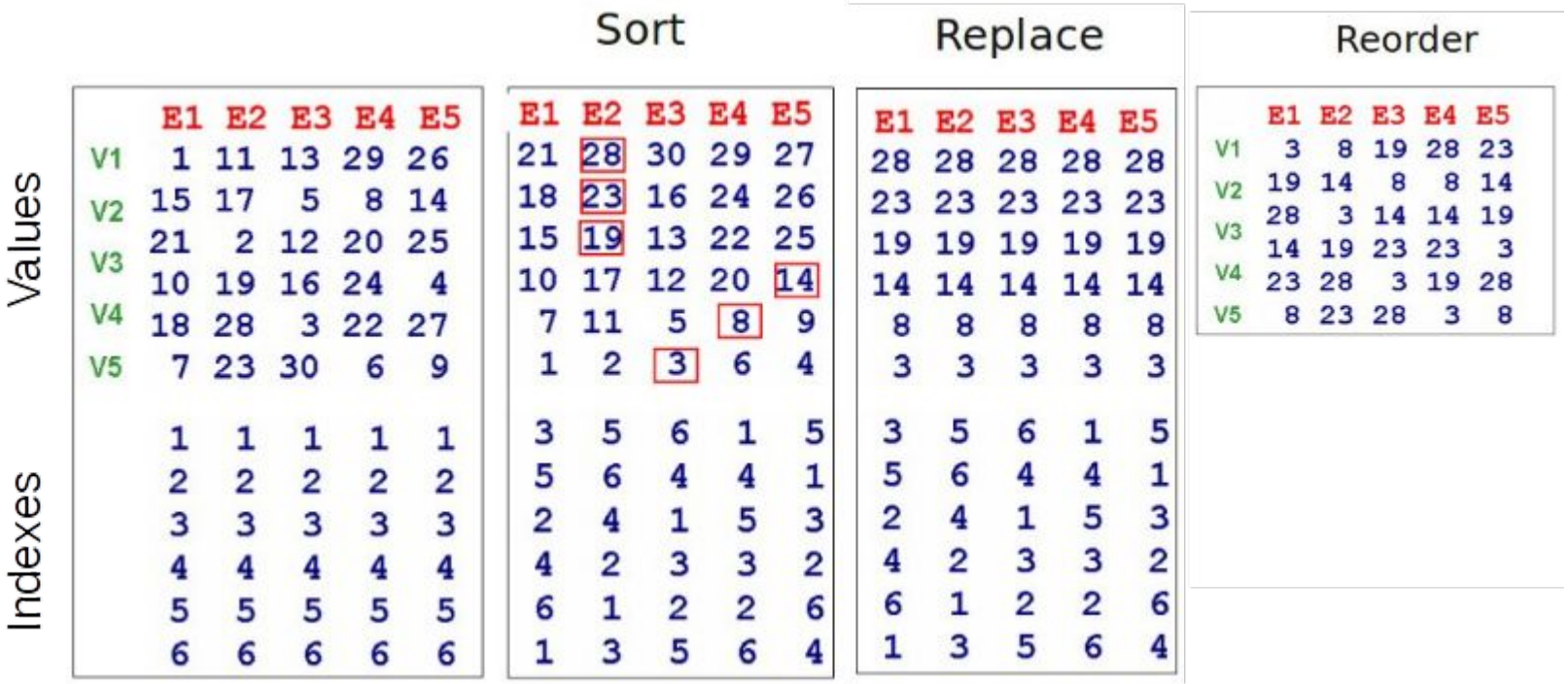
Sd : (estimated) expr. standard deviation of all genes

Quantile normalization

- ✓ Differences between the separate values retained
- ✓ Identical distribution for each array
- ✘ Information lost
  - Especially in the lower signals

1. Matrix X
   a. Columns = samples
   b. Row = transcripts
2. Sort each column of X -> X_{sort}
3. Calculate row-means and store in X'_{sort}
4. Obtain X_n by rearranging columns of X'_{sort} to have the same ordering as the corresponding input vector

Example quantile normalization



Example effect quantile normalization

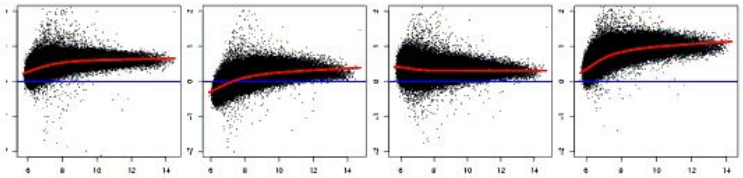


Figure 7A. Ratio Intensity Plot of all probes for four pairs of chips from GeneLogic spike-in experiment

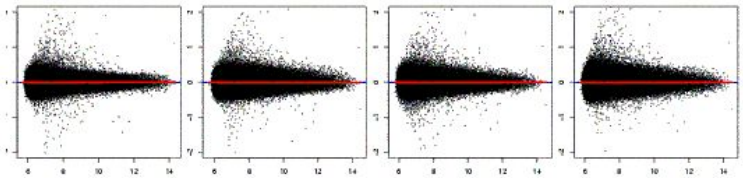
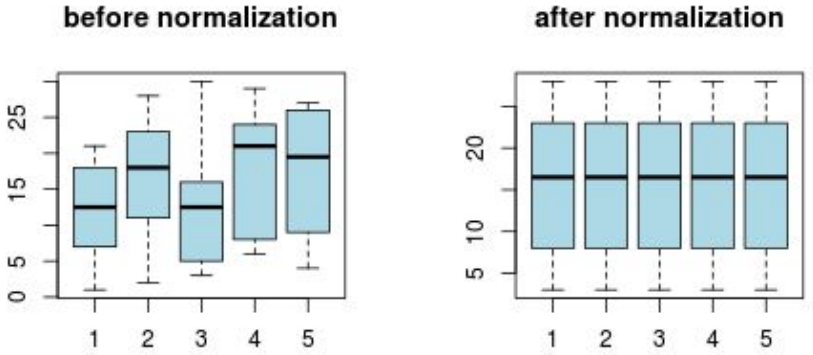


Figure 7B. As in A, after normalization by matching quantiles. Both figures courtesy of Terry Speed



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.

Important: normalization between samples, not within one sample