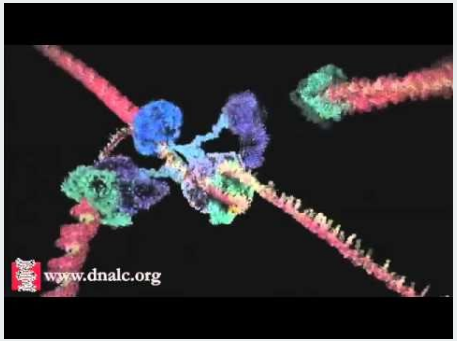
Measuring gene expression

Grundlagen der Bioinformatik

07.06.2022

Raik Otto - raik.otto@hu-berlin.de



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

Agenda

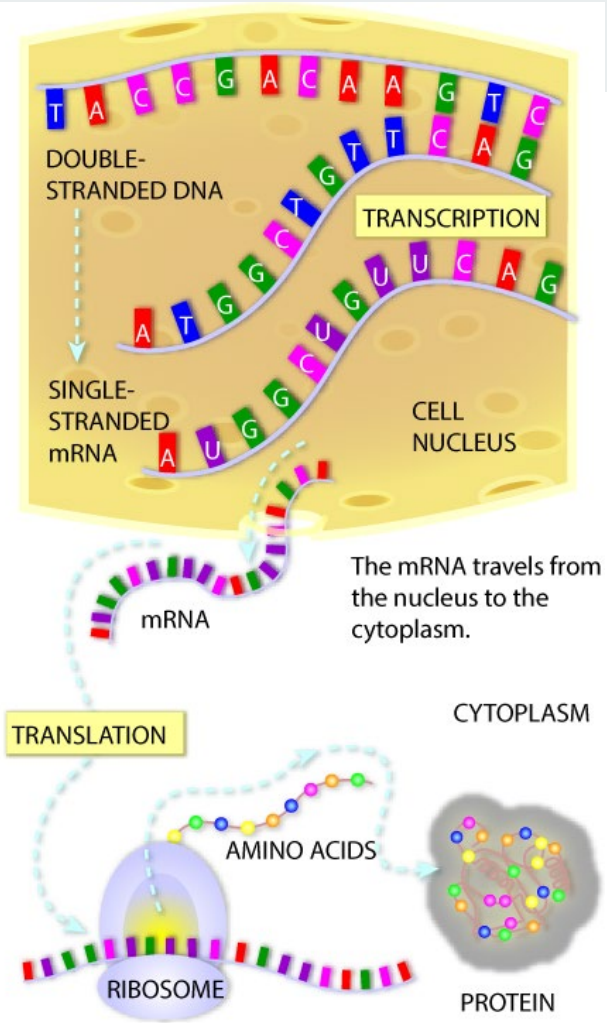
- Biology
  - Gene expression
- Technologies
  - FISH
  - Microarrays
  - RNA-seq
- Visualization Methods

Biology - Gene Expression

Your thoughts?

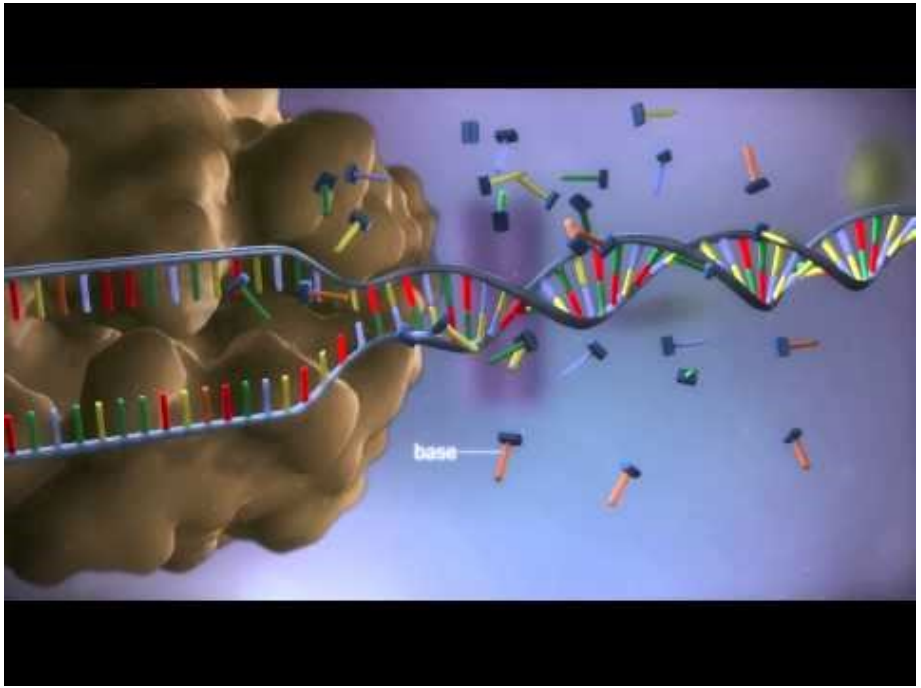
Gene Expression - mRNA

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



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

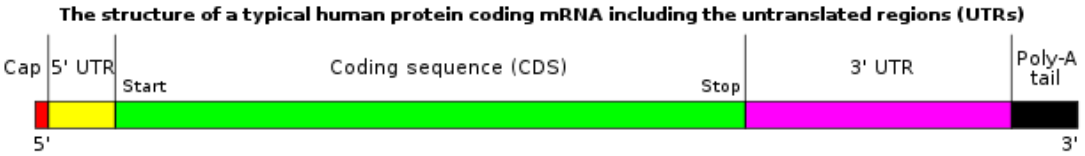
Video time



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

mRNA structure

- RNA copy of DNA gene
  - Modified copy -> not identical
- Sequence determines protein
- Cap and end
- Partially translated
- Aim: Detect mRNA expression

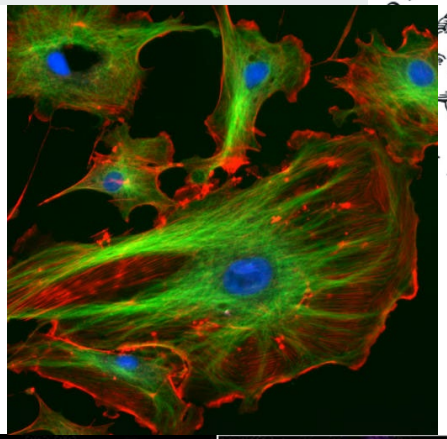


Simplified mRNA structure

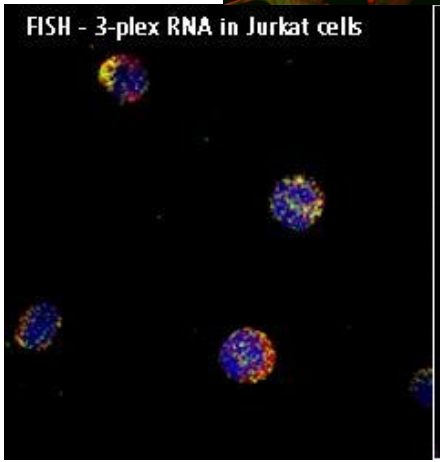
Wikicommons

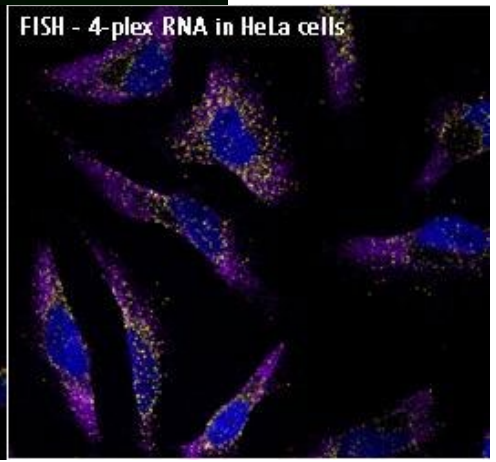
mRNA Quantification Technologies

Fluorescence In Situ Hybridization



- Fluorescence in situ hybridization = FISH
- Illumination
- Qualitative



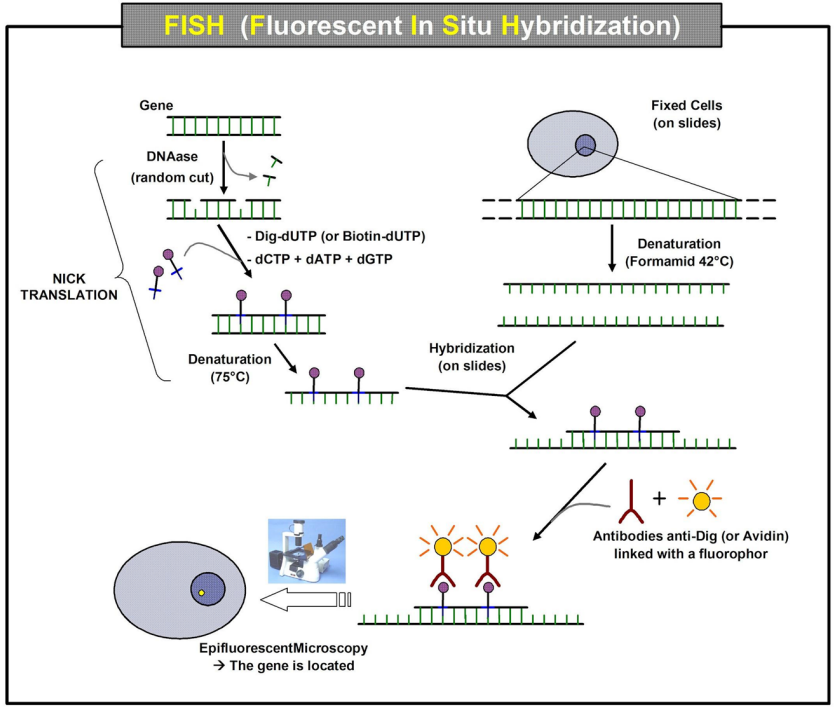


wikicommons

FISH method

- Here: shown for **DNA**

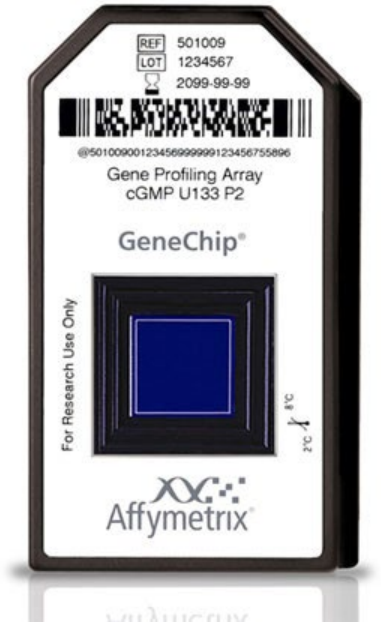
1. Cut DNA and paste anchor
2. Denature DNA
3. Hybridize
4. Attach antibody and shine



wikicommons

mRNA Micro-Arrays

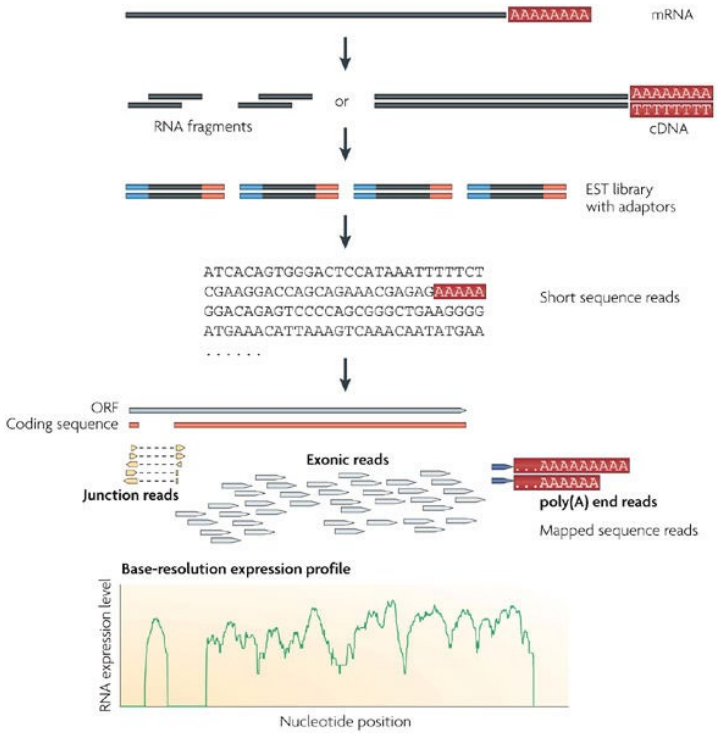
- Oligo-nucleotide arrays
- Array of pre-defined sequences
- Complementarily binding to mRNA
- mRNA illuminated
    - Expression measured as light-intensity



www.affymetrics.com

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

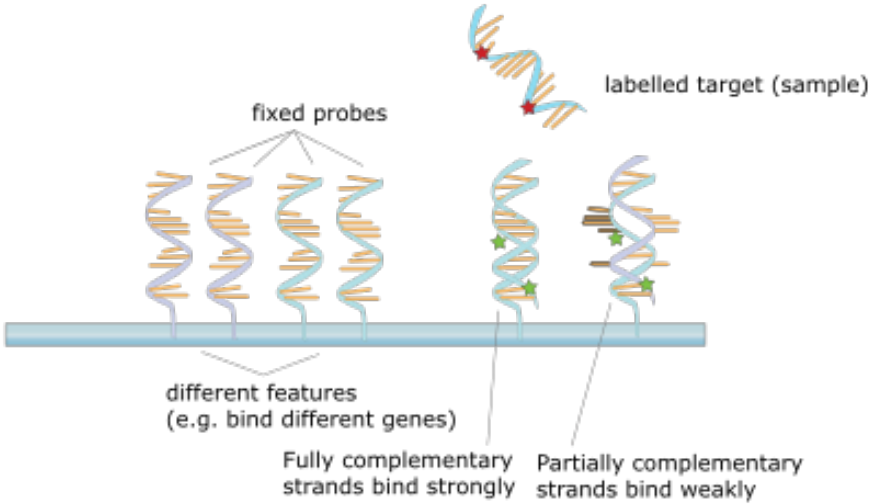


Wang et al., *Nature Reviews Genetics* volume10, pages57–63 (2009)
doi:10.1038/nrg2484

Nature Reviews | Genetics

Hybridization

- mRNA binds to sequences
- Sequences are labeled
- Binding intensity translates into illumination



wikicommons

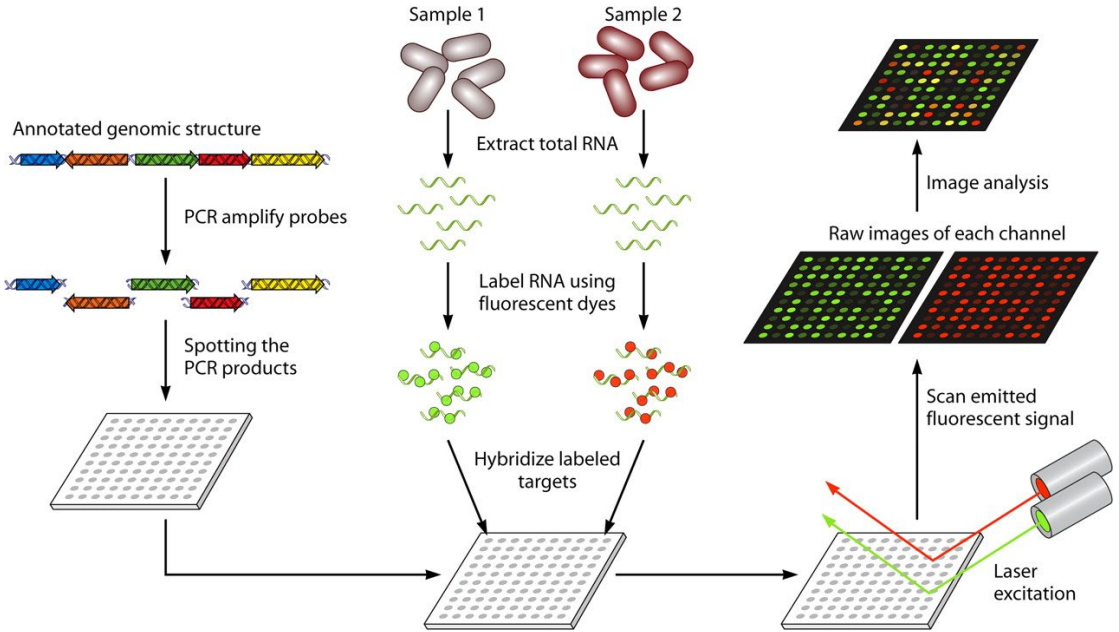
Probe sequence selection

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

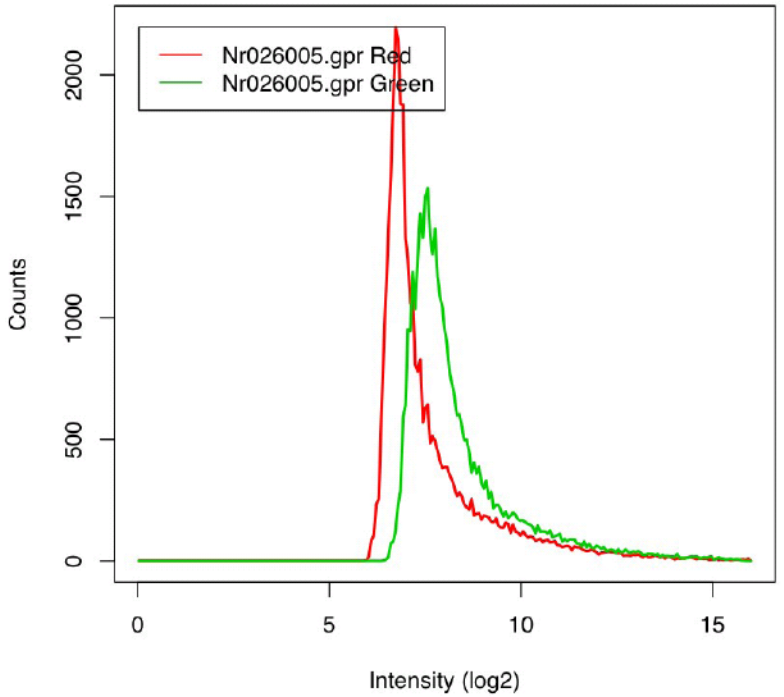
Two color array

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



Structural dye-bias two-color array

- Distortion of measurements
- Green brighter than red
- Intensity-dependent



Summary technologies

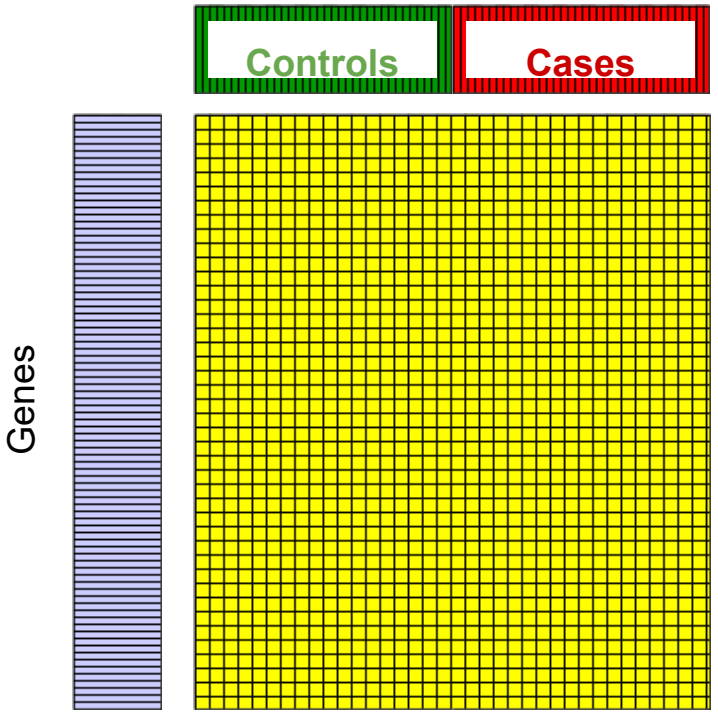
Technology	Type	Price	Amount genes	Supervised*
FISH	Qualitative	Low	Small	Yes
mRNA-Array	Qualitative/ Quantitative	Low	Large	Yes
RNA-seq	Quantitative	High	Very large	No

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

Visualization Methods

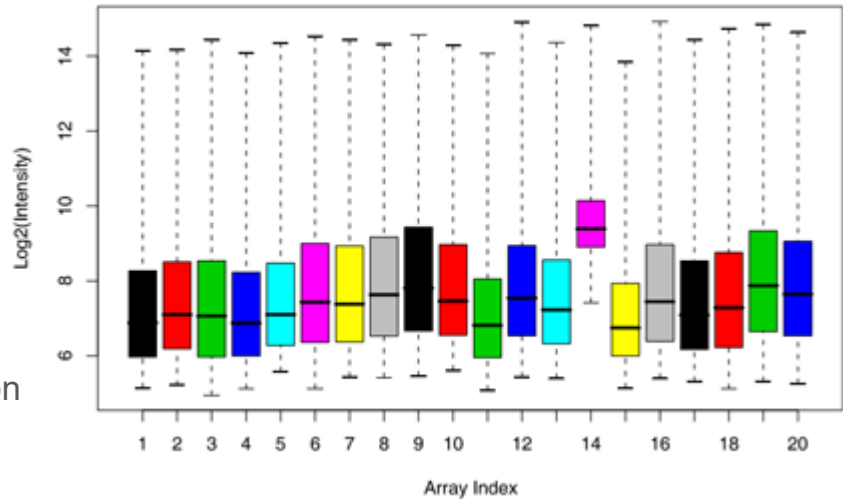
mRNA experiment design

- Two or more cohorts
  - Control
  - Case
- Measure aggregated intra-cohort expression
- Render measurements comparable
- Identify differences between aggregated expressions



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

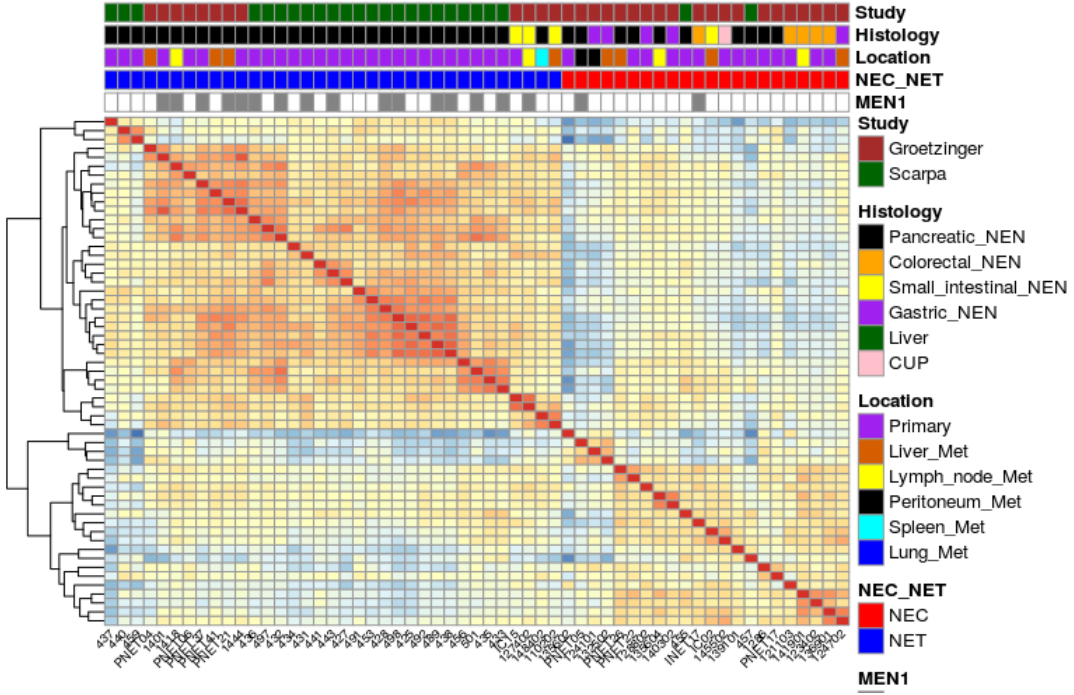
LOWER QUARTILE 25% data less than this value

MINIMUM Least value, outliers not included

OUTLIER Less than 3/2 times the upper quartile

Visualization - Correlation heatmap

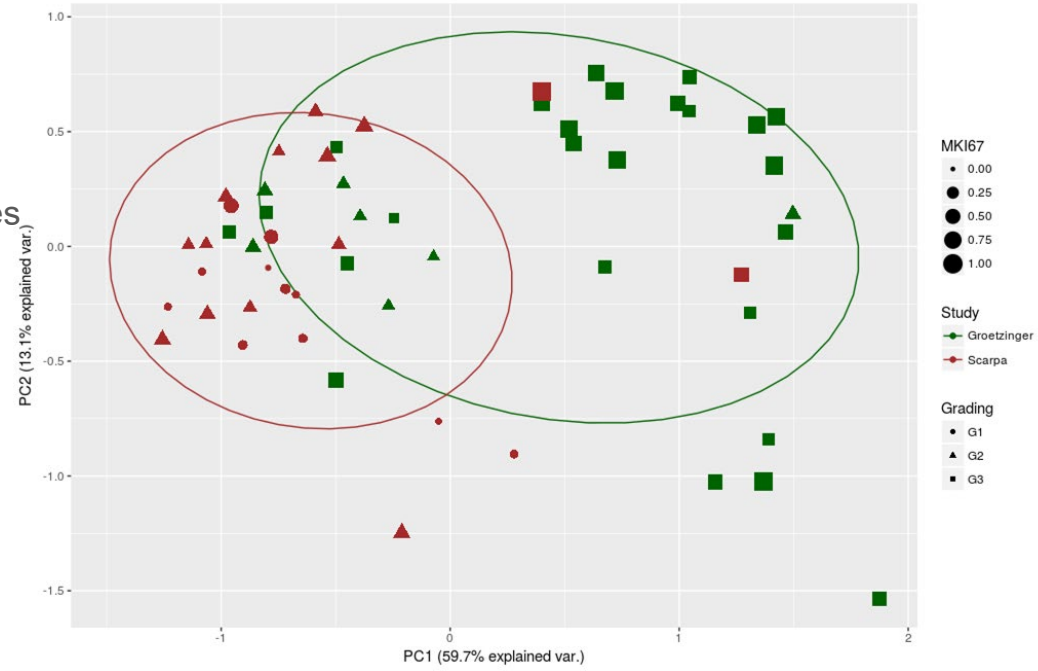
- *Pairwise-similarity* of samples
- Clustering informative
  - Bad: clustering based on study
  - Good: clustering based on cancer-type
    - NEC (Carcinoma) vs NET (Tumor)



Real-world heatmap

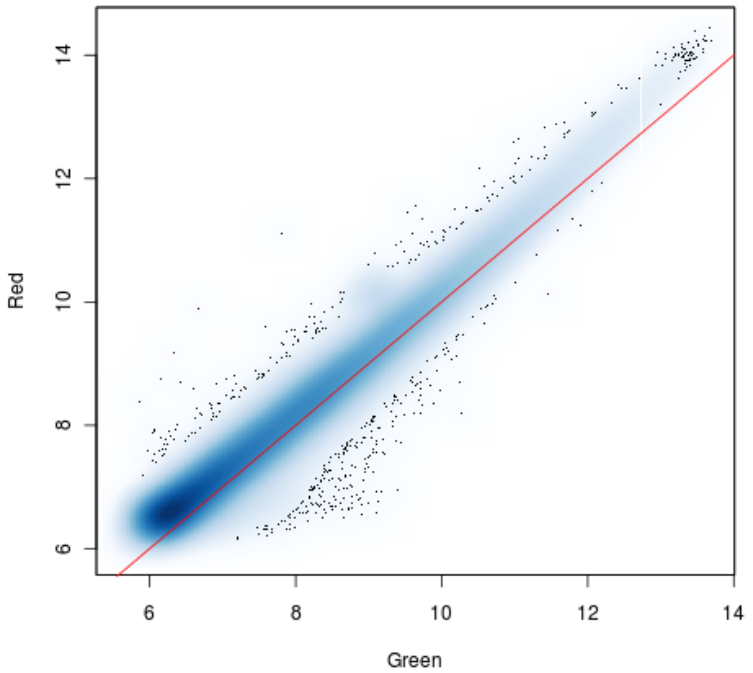
Principal component analysis (PCA)

- 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

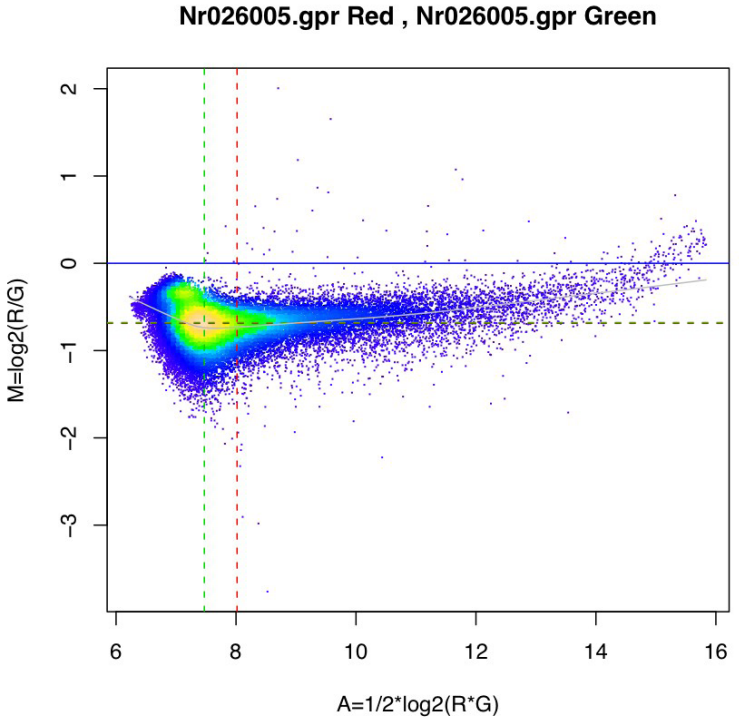


Mean-average (MA)-plot

- 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