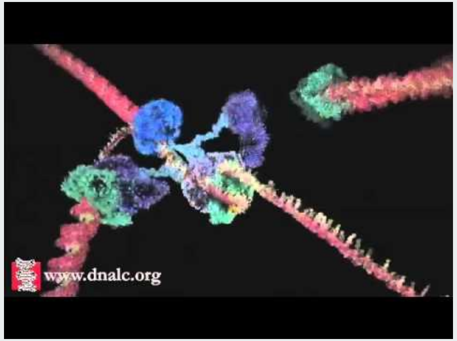
Measuring gene expression

Grundlagen der Bioinformatik SS2019



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

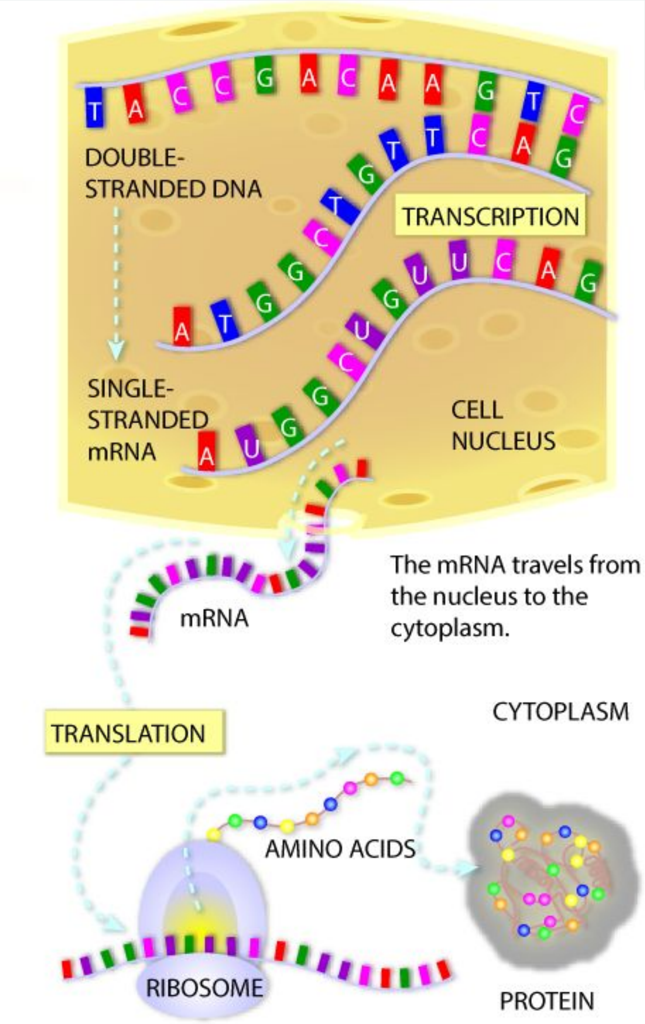
Agenda

- Gene expression
  - Biological background
- Technologies
  - FISH
  - Microarrays
  - RNA-seq
- How to detect technological biases
  - Visualization
  - Quality control
  - Normalization

Gene Expression - Background

Gene Expression - mRNA

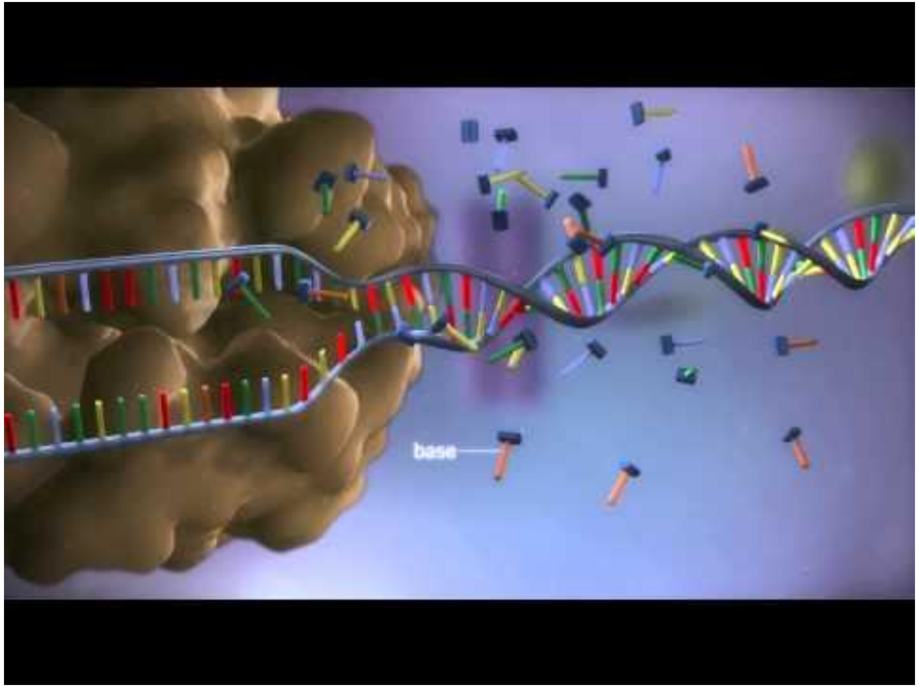
- mRNA expression ∝ gene activity
- Protein ~ active *form* of genes
- mRNA = messenger RiboNucleic Acid
- DNA->mRNA-> Protein



The mRNA travels from the nucleus to the cytoplasm.

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

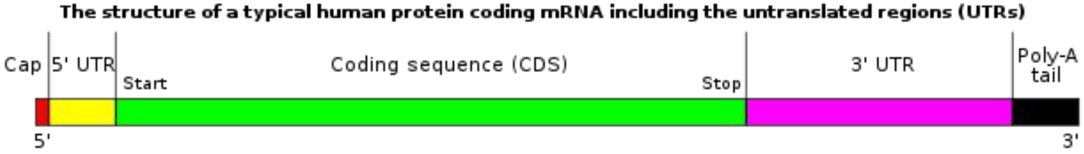
Video time



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


The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)
Cap
5' UTR
Start
Coding sequence (CDS)
Stop
3' UTR
Poly-A tail
5'
3'

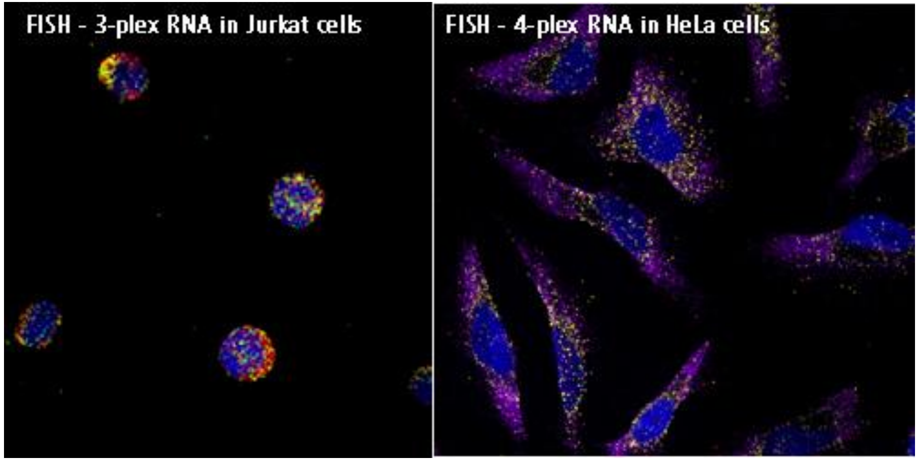
Simplified mRNA structure

Wikicommons

mRNA Quantification Technologies

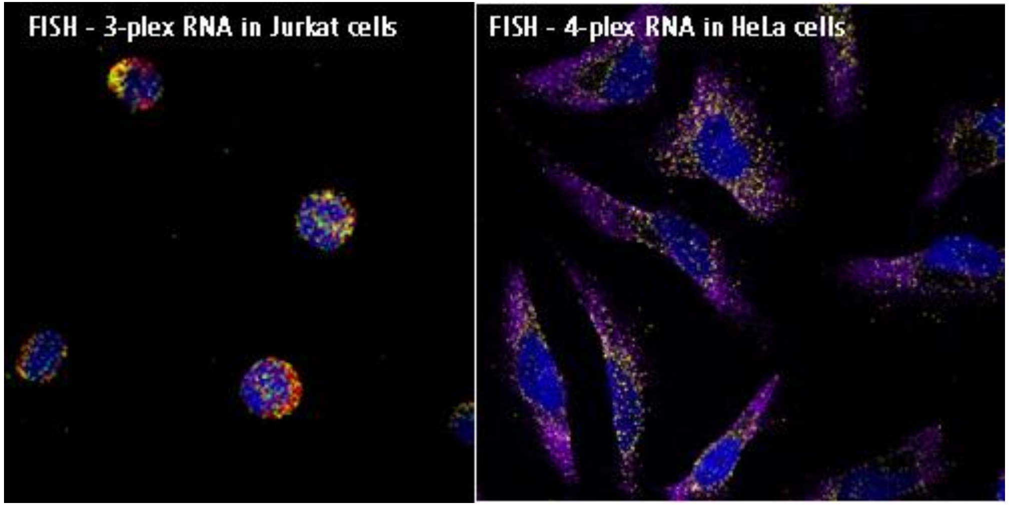
Fluorescence In Situ Hybridization

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


FISH - 3-plex RNA in Jurkat cells
FISH - 4-plex RNA in HeLa cells

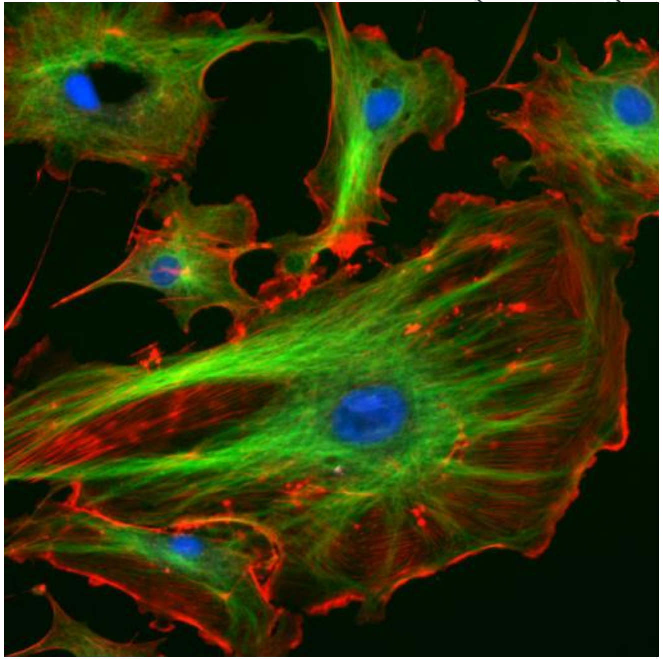
wikicommons

Impressions


FISH - 3-plex RNA in Jurkat cells
FISH - 4-plex RNA in HeLa cells

Ryan Jeffs

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

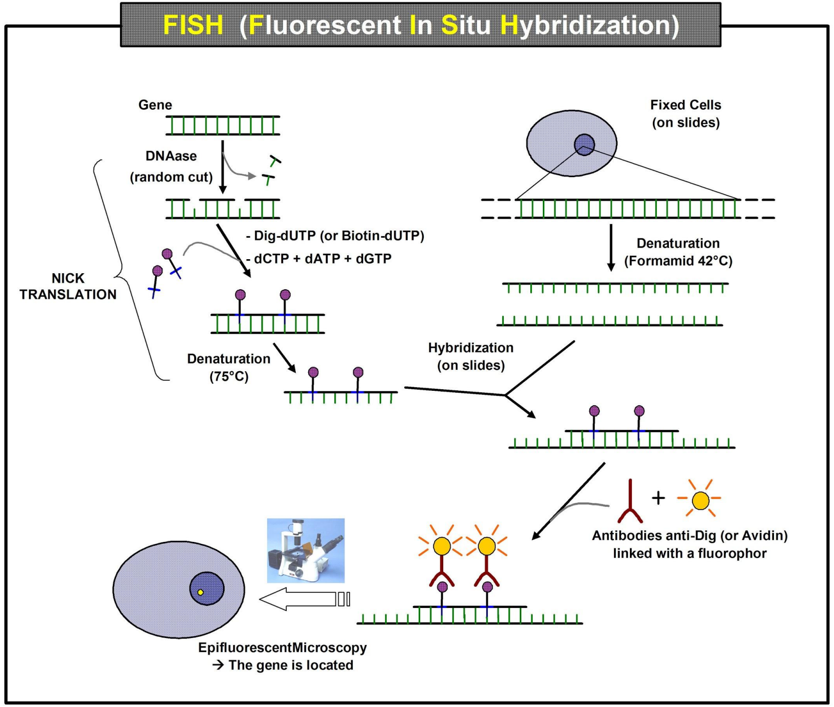


Nucleus, skeleton &
Cell-membrane

FISH method

- Here: shown for **DNA**

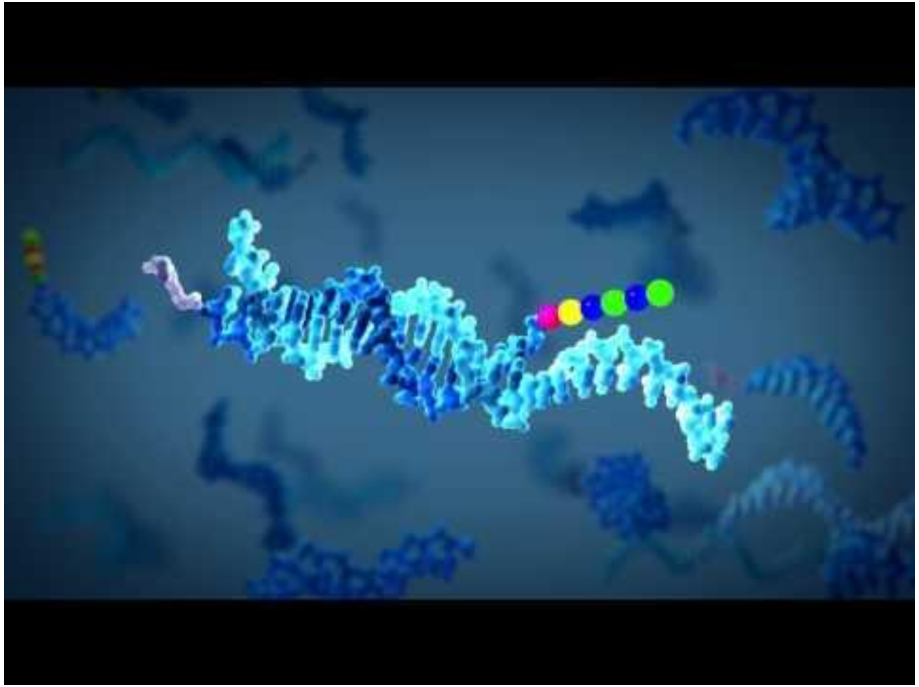
1. Cut DNA and paste anchor
2. Denature DNA
3. Hybridize
4. Attach antibody and shine



wikicommons

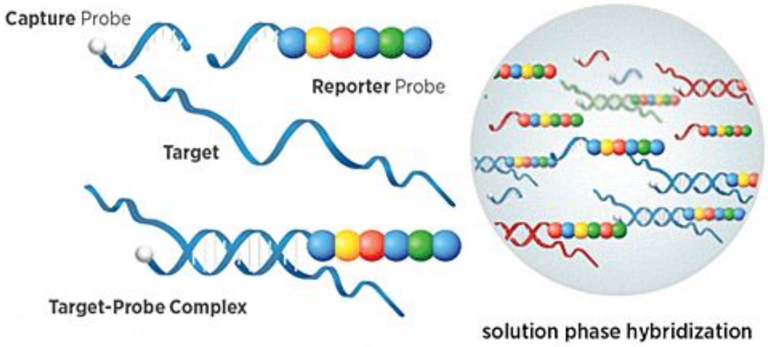
FISH / NanoString

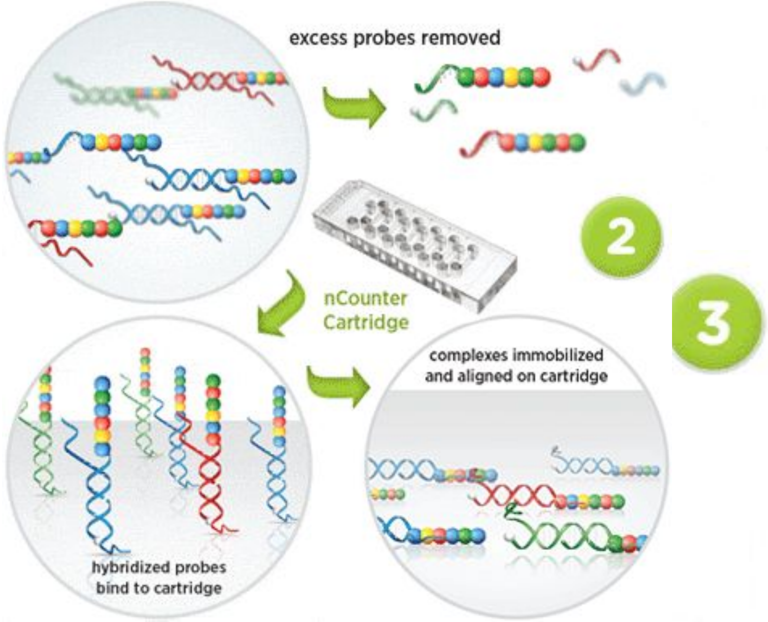
- Quantitative FISH-like -> counts available
- Separate capture
- Sequence matched
- Medium throughput

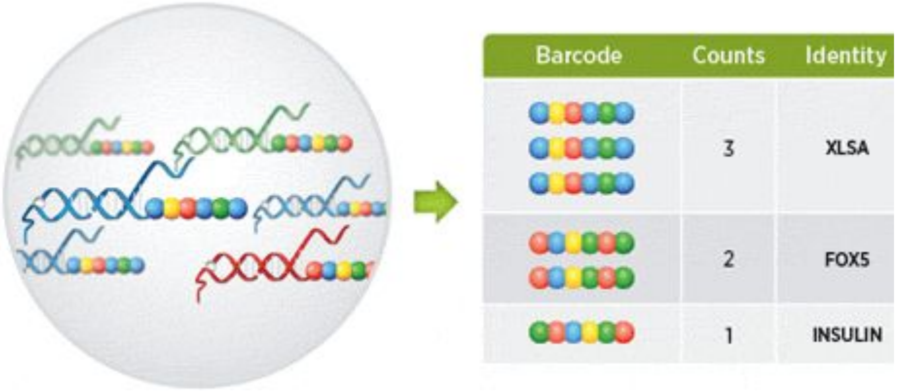


https://www.youtube.com/watch?v=XIVmmfujiro
>= 1 minute 22 seconds

Nanostring 1



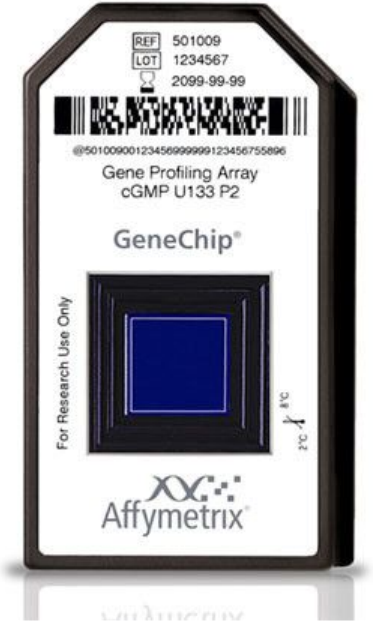




Barcode	Counts	Identity
●●●●●● ●●●●●● ●●●●●●	3	XLSA
●●●●●● ●●●●●●	2	FOX5
●●●●●●	1	INSULIN

mRNA Micro-Arrays

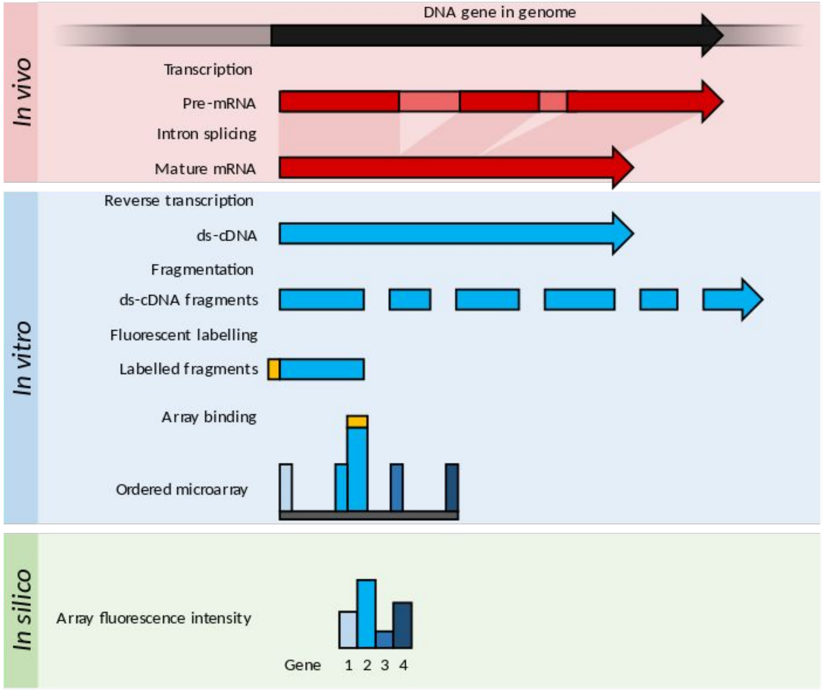
- Oligo-nucleotide arrays
- Array of pre-defined sequences
- Complementarily binding to mRNA
- mRNA illuminated
  - Expression measured as light-intensity



www.affymetrics.com

Workflow mRNA array

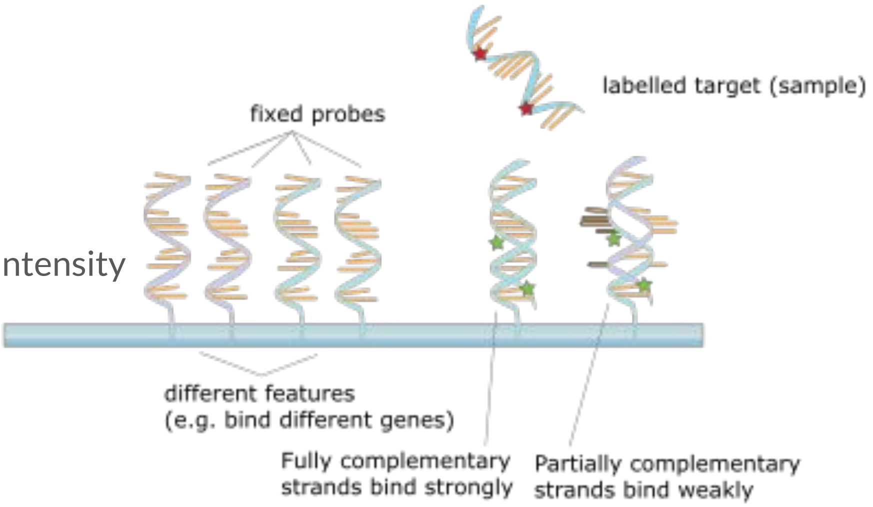
1. Isolation and purification
2. Reverse transcription
   a. cDNA == complementary DNA
3. Labelling fluorescent dye cDNA labeling
4. Hybridization
   a. Washing
5. Scanning
   a. Laser excitation
   b. detection of light intensities
   c. image segmentation
6. Normalization



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

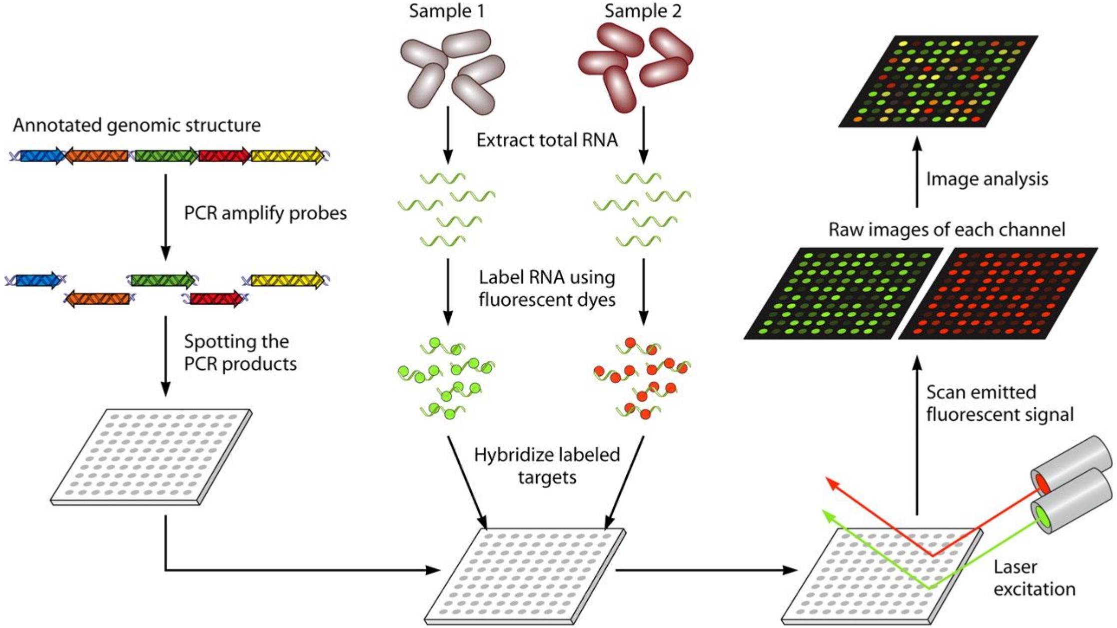
- Sensitive sequence may not be specific
    - E.g. cap or poly-A tail sequences
- Sensitivity := TP / (TP + FN)
- Specificity := TN / (TN + FP)
- Interesting optimization problem

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

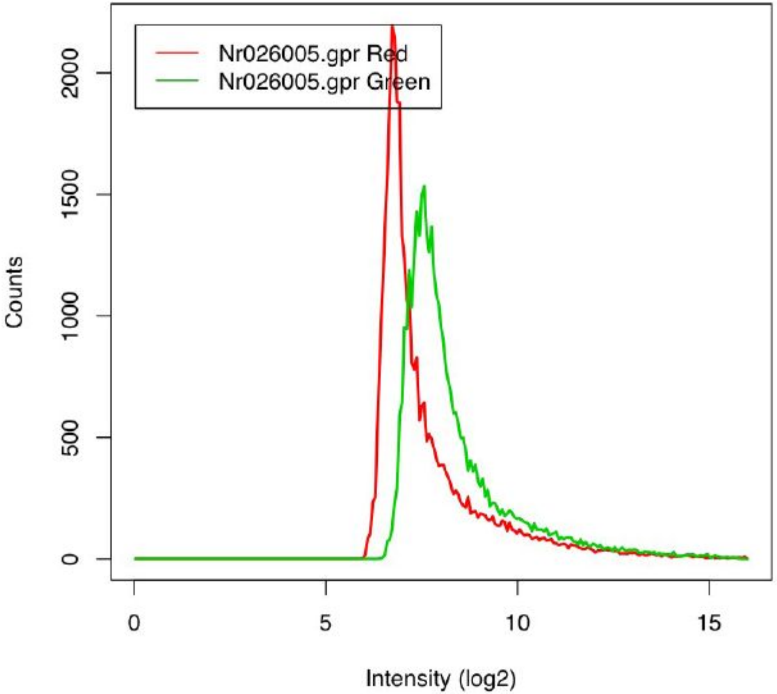
Two color array

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


Sample 1
Sample 2
Annotated genomic structure
Extract total RNA
PCR amplify probes
Label RNA using fluorescent dyes
Spotting the PCR products
Hybridize labeled targets
Image analysis
Raw images of each channel
Scan emitted fluorescent signal
Laser excitation

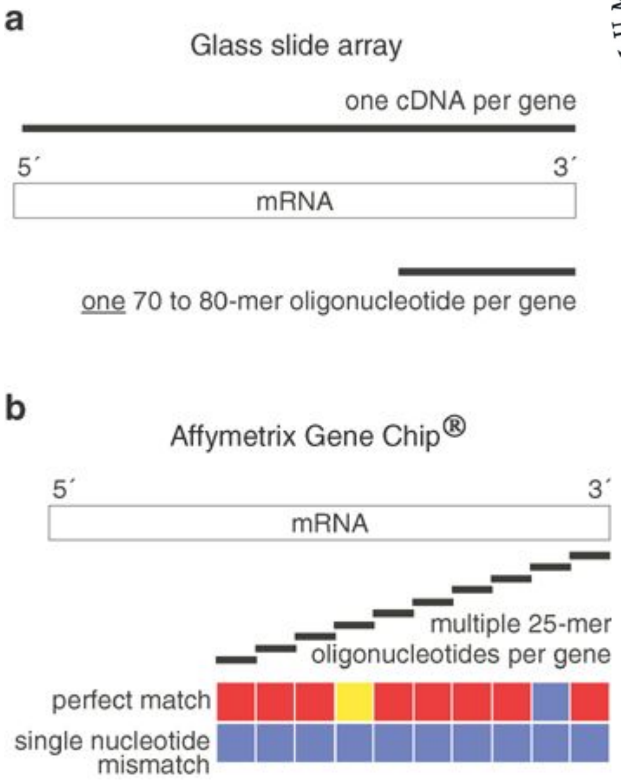
Structural dye-bias two-color array

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



One color array

- 25nt - 60nt
- Probe-seq matches known genes
- 20 probes := 1 probeset = 1 sequence
  - But: can target many genes
- Ratio match-mismatch critical


a
Glass slide array
one cDNA per gene
5´
3´
mRNA
one 70 to 80-mer oligonucleotide per gene
b
Affymetrix Gene Chip®
5´
3´
mRNA
multiple 25-mer oligonucleotides per gene
perfect match
single nucleotide mismatch

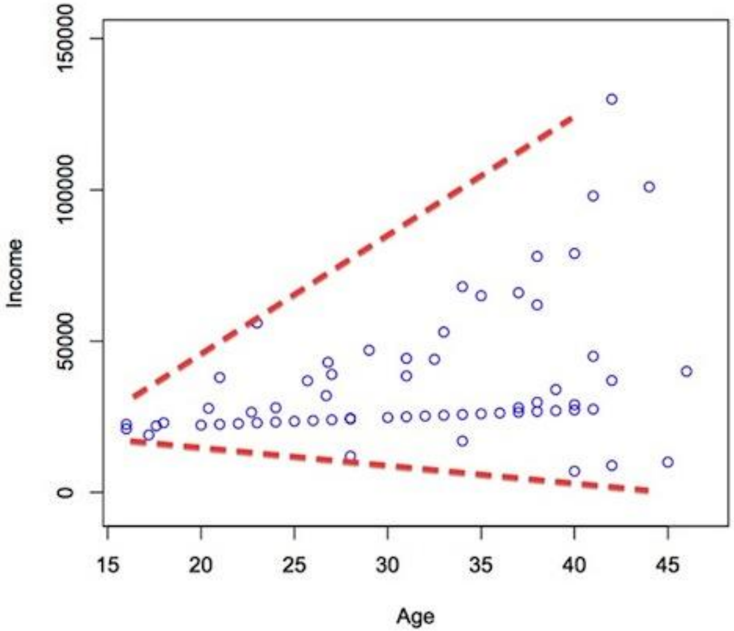
Staal, F. J. T., et al. ." Leukemia 17.7 (2003): 1324-1332.

Example bias one-color array: Heteroscedasticity

Heteroscedasticity when comparing two

independent one-colory array measurements

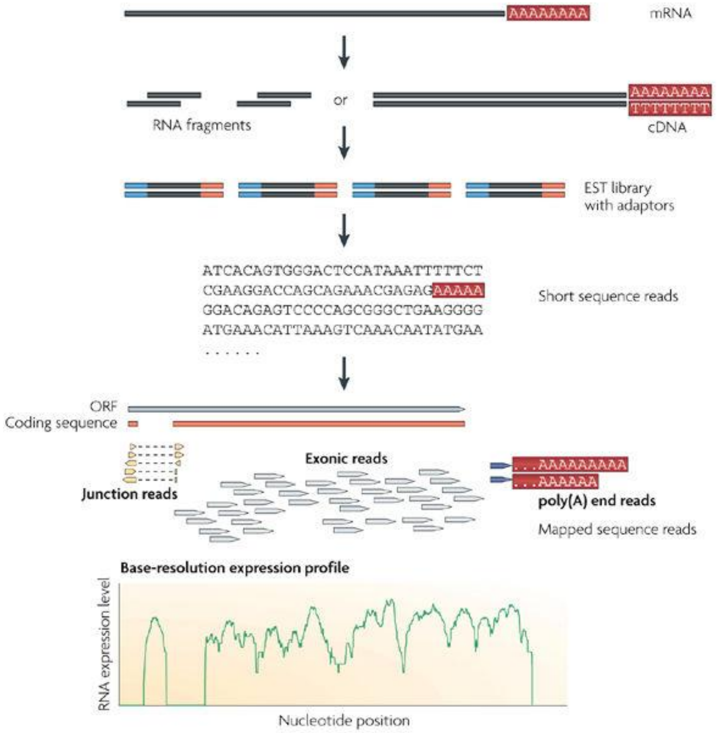
- Variance expression dependent
- Differential expression detection fails
- Normalization required



Analogous example for income

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

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

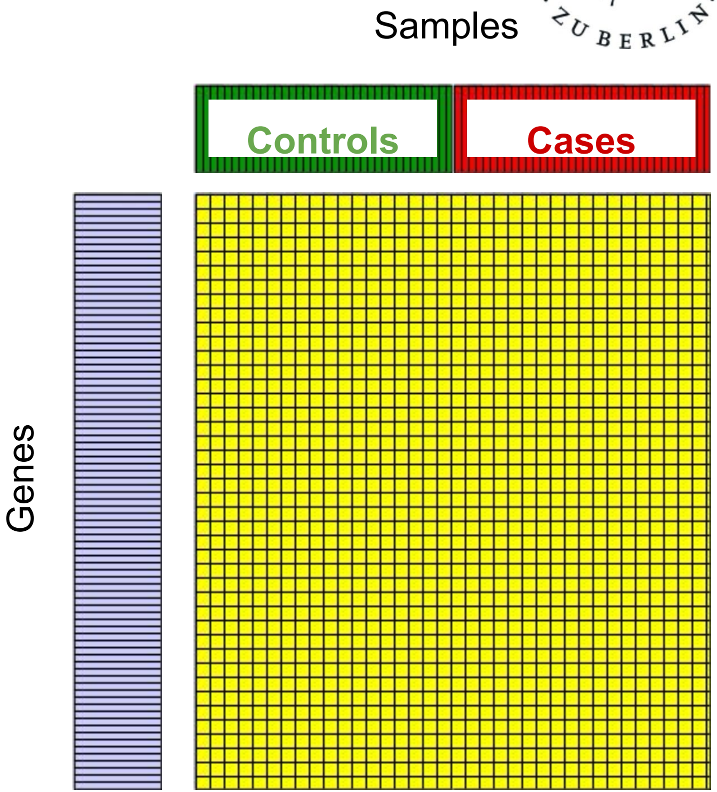
Technology	Type	Price	Amount genes	Supervised*
FISH	Qualitative	Low	Small	Yes
mRNA-Array	Qualitative/ Quantitative	Low	Large	Yes
RNA-seq	Quantitative	High	Very large	No

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

Methods

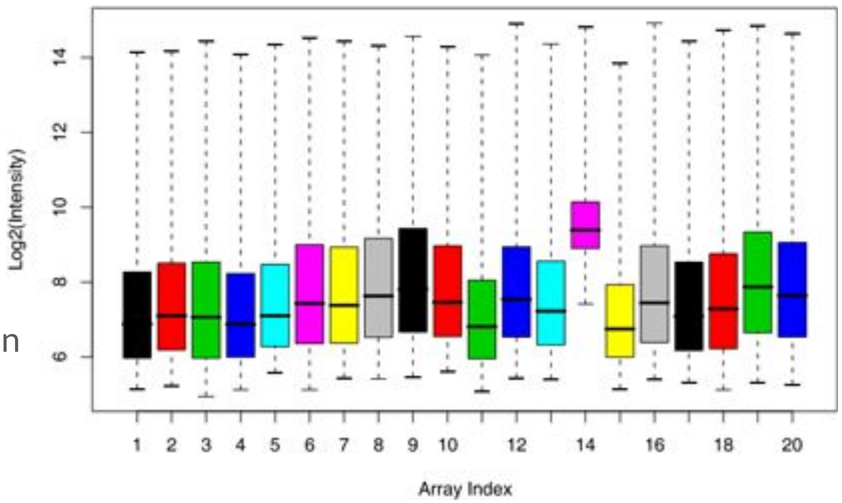
mRNA experiment design

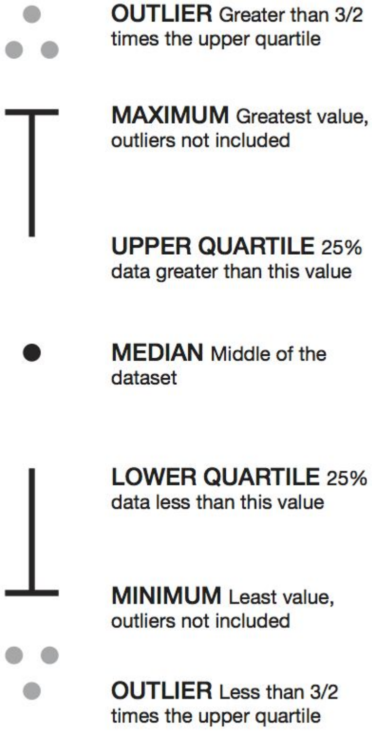
- Two or more groups (called cohorts)
  - Control
  - Case
- Identify aggregated expression within cohorts
- Identify differences between aggregated expressions
- Ensure that measurements are comparable



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

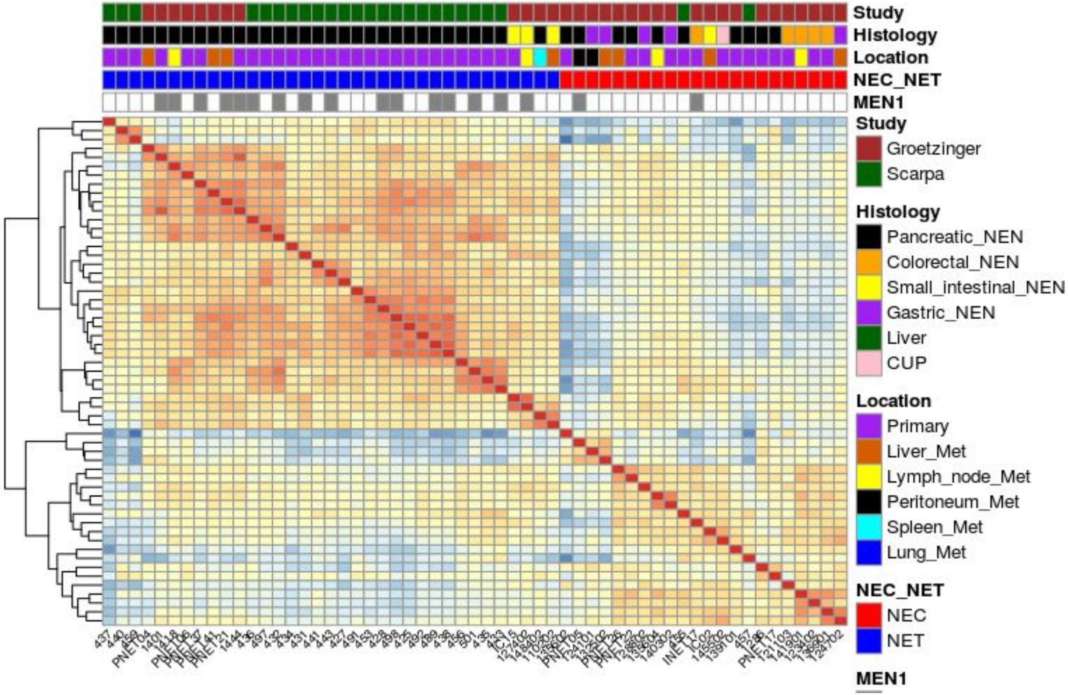
LOWER QUARTILE 25% data less than this value

MINIMUM Least value, outliers not included

OUTLIER Less than 3/2 times the upper quartile

Visualization - Correlation heatmap

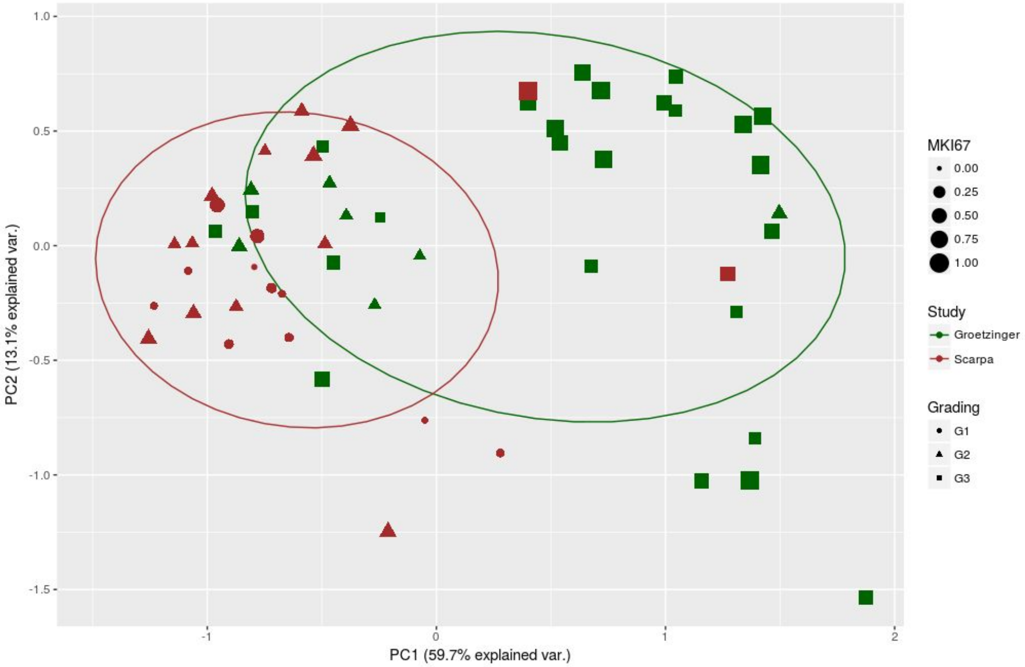
- *Pairwise-similarity* of samples
- Clustering informative
  - Bad: clustering based on study
  - Good: clustering based on cancer-type
    - NEC (Carcinoma) vs NET (Tumor)



Real-world heatmap

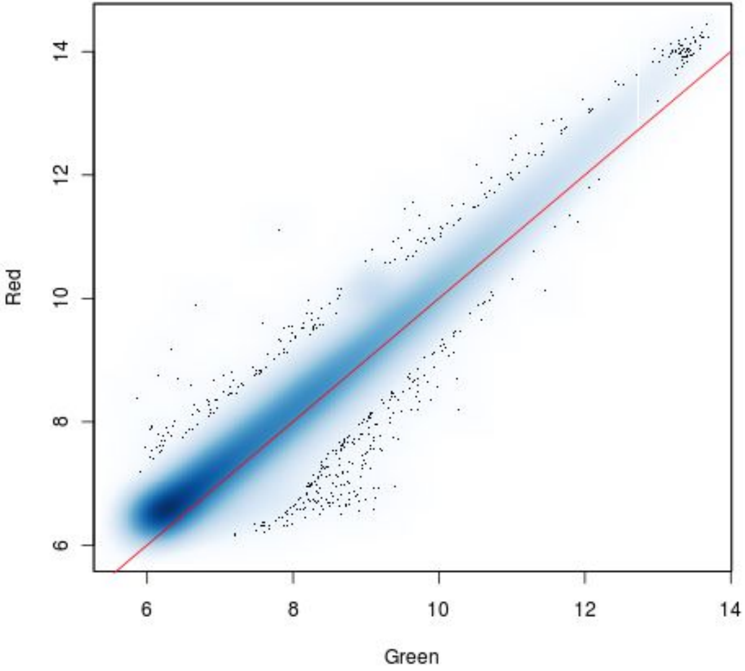
Principal component analysis (PCA)

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



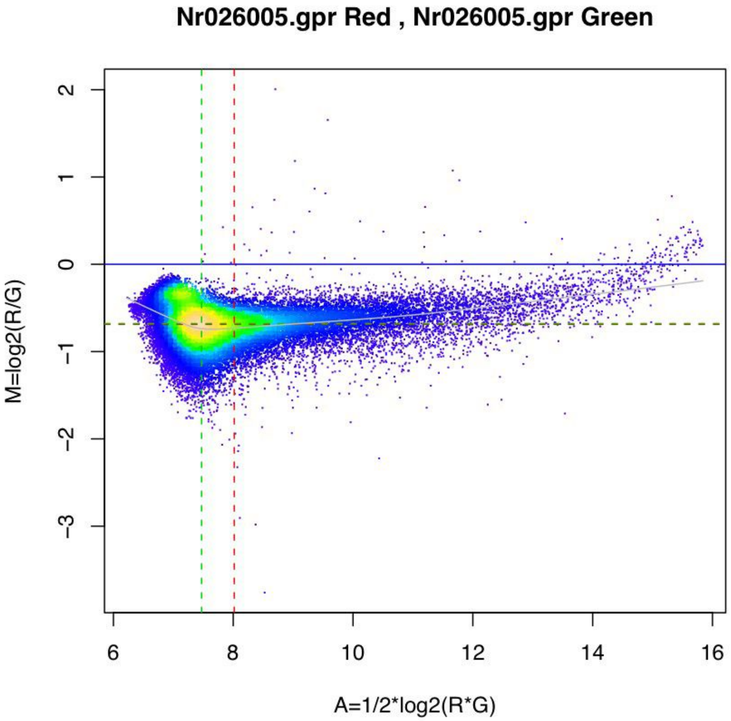
Scatter plot

- 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 * ($\log_2 \text{Value}_1$) + log(Value_2))

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