Analysis of gene expression

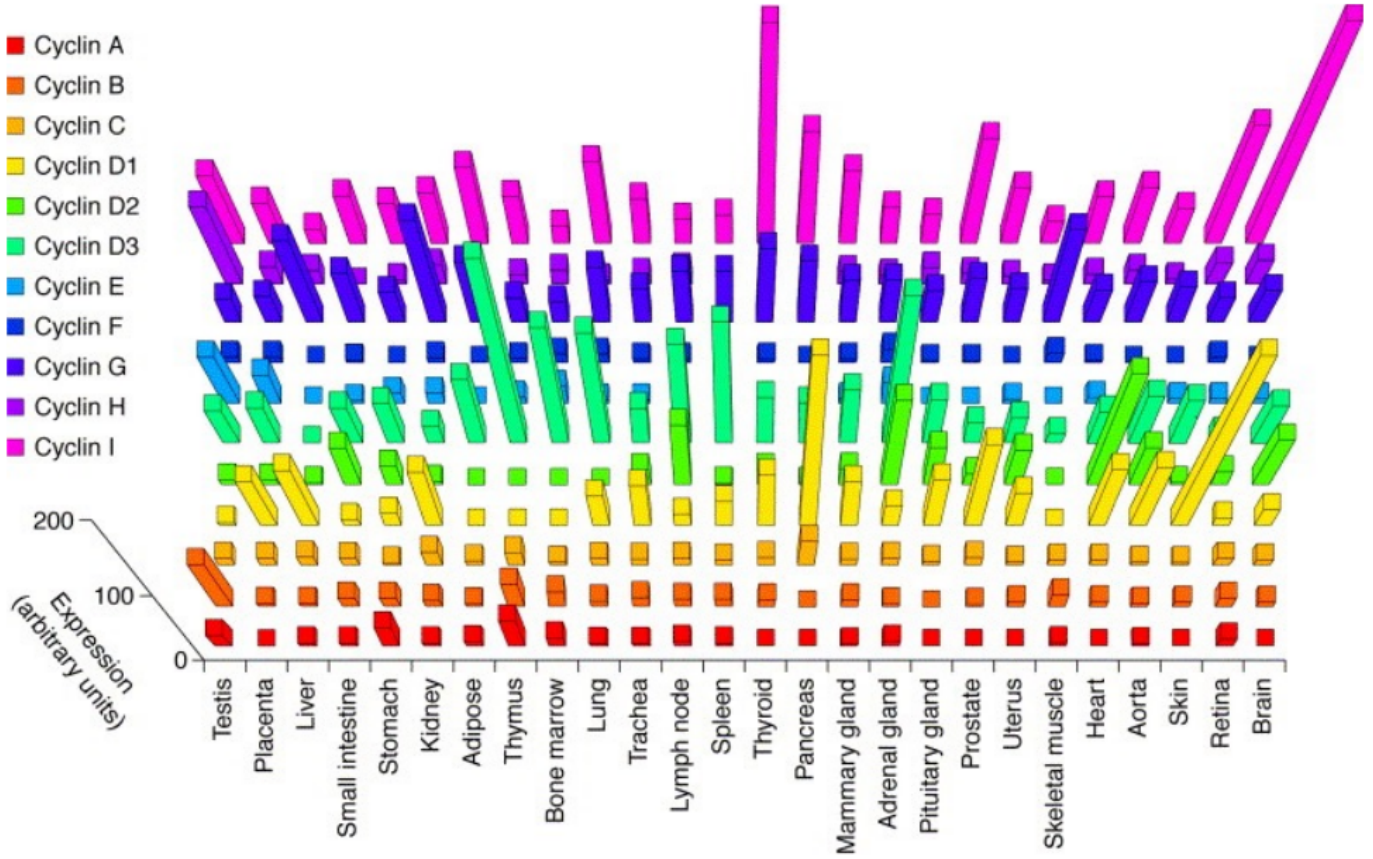
Alessandro Romanel

Bioinformatics Resources
2019-2020

Gene expression

- Expressed genes are those that have been transcribed
- A gene expression profile of a cell is the snapshot of which genes are expressed in that cell at the time the sample was taken
- Knowing which genes are expressed in a cell allows
  - identification of new genes or transcripts
  - comparison of expression profiles between samples
- Main motivations: disease, development, dynamic responses

Gene expression



- Gene expression varies from individual to individual, from tissue to tissue, from condition to condition, from cell to cell
- Variability is mainly due to alternative splicing and different regulation

Differential Gene Expression

- Expression profile of genes in one sample vs another
- Different cells, tissues, disease states, developmental stages, culture conditions, etc, can be compared
- Measure both, subtract the overlap, obtain the difference, interpret it.
- Pay due attention to controls, negative and positive
- Pay due attention to range of variability within samples
- High throughput

DGE analysis workflow

- Formulate the biological questions
- Experimental design
  - Which platform, which controls, how many replicates
- Run the experiment
- Image processing (by machine)
- Low-level analysis
  - Data preprocessing (normalization step)
- High-level analysis
  - Data analysis
- Reach biological conclusions
  - Interpretation of results

High throughput methods (pros)

- Fast
  - a lot of data produced quickly
- Comprehensive
  - entire genomes in one experiment
- Easy
  - submit RNA samples to a core facility
- Cost
  - getting cheaper still

High throughput methods (cons)

- Cost
  - Some researchers can’t afford to do appropriate numbers of controls, replicates
- RNA
  - The final product of gene expression is a protein
- Significance
  - How do you filter out non-coding RNA, or transcripts that are not translated?
- Quality
  - Artifacts with image analysis and data analysis
- Control
  - Not enough attention to experimental design
  - Need more collaboration with computational scientists

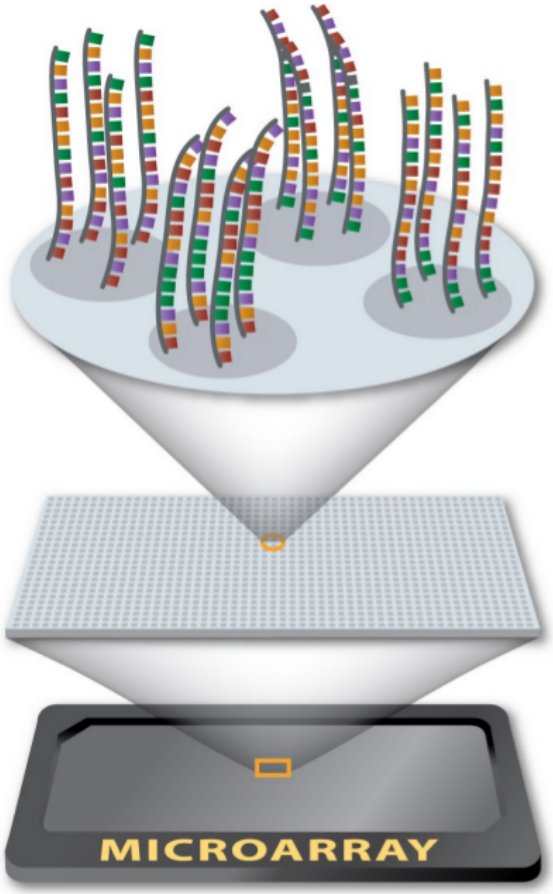
Main technologies

- Microarray technology
- High throughput RNA-seq technology
  - transcriptome sequencing
  - quantification of mRNA transcripts

Microarrays

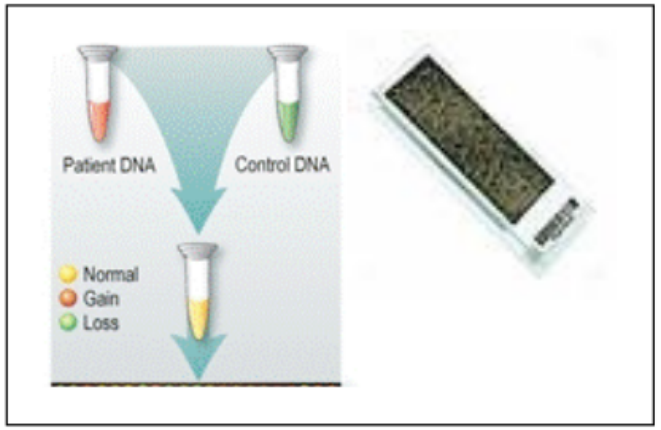
- Introduced at the beginning of ’00.
- First high throughput technology
- Useful to investigate genomic and transcriptomic profiles, methylome, DNA-protein interaction, microbiome, etc...
- Data interpretation is subject to specific computational analyses

Microarrays

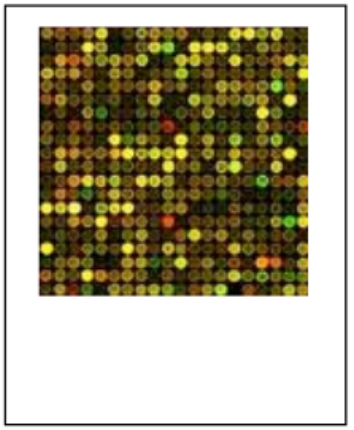


- Monitor thousand of genes in parallel
- Each spot contains multiple and identical DNA probles
- Thousand of spots disposed as a matrix on a solid support (usually glass)
- Microarrays have dimensions similar to that used in microscopy

Microarrays (workflow)

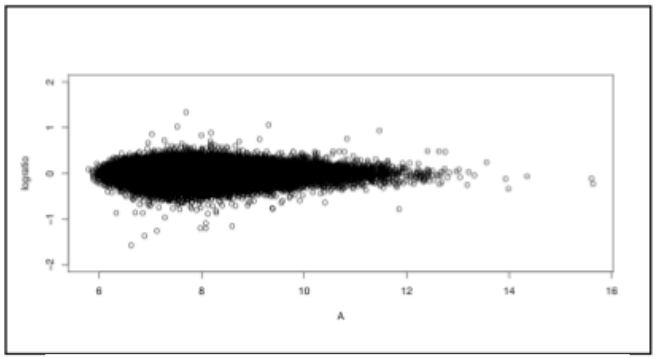


1. Experiment



2. Read the data

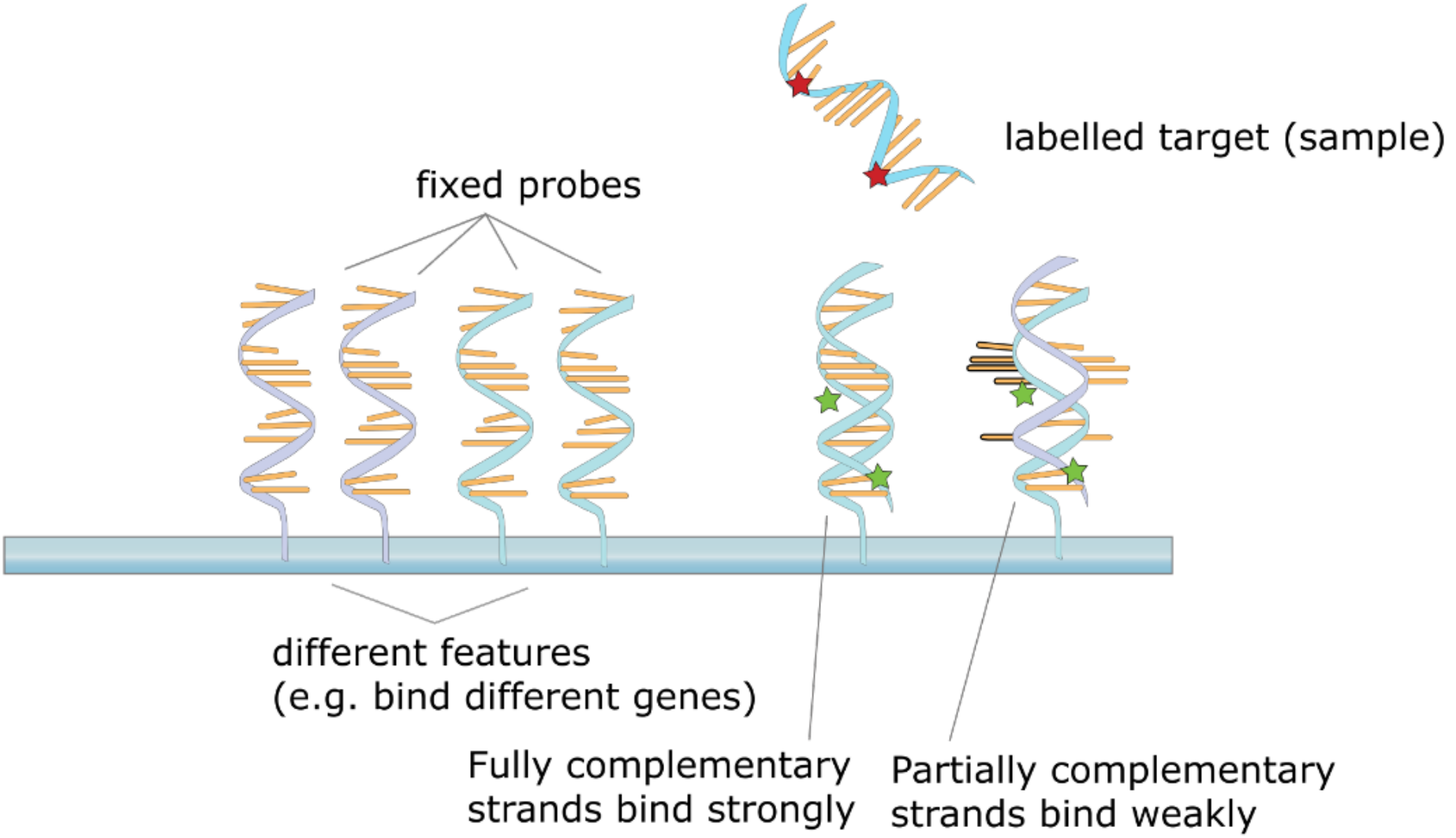
3. Image analysis



4. Data analysis

4. Data interpretation

Microarrays (ibridization)

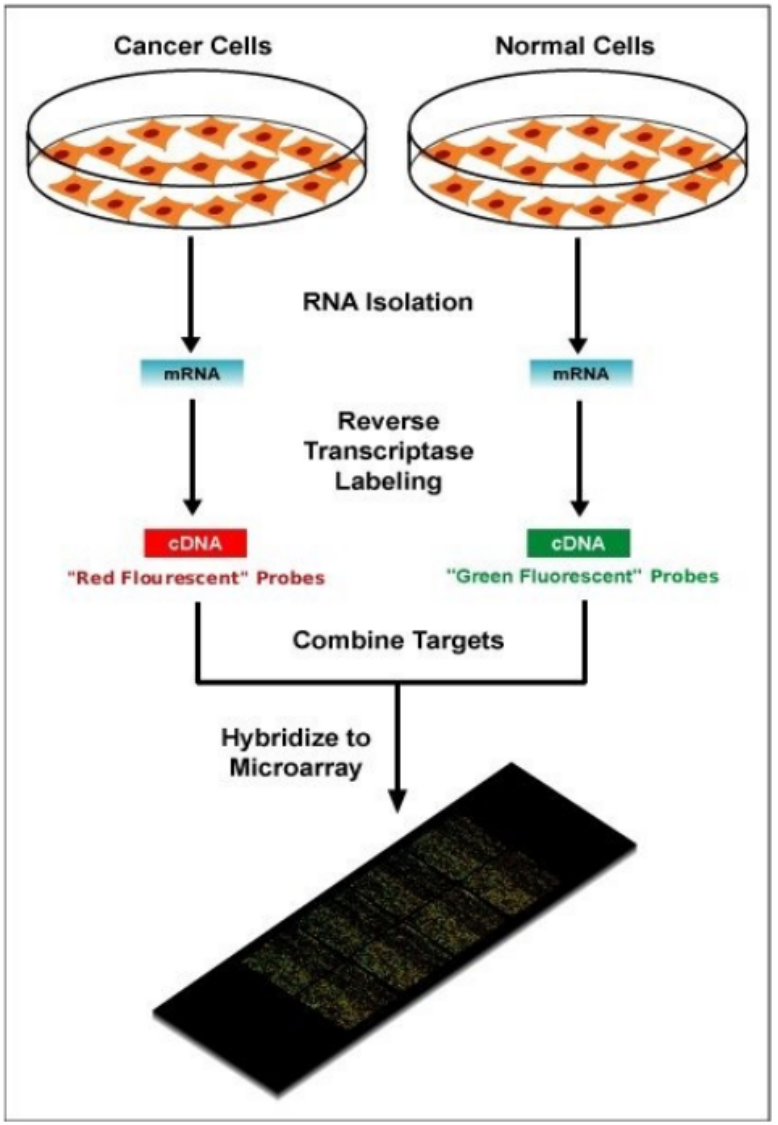


https://en.wikipedia.org/wiki/DNA_microarray

Microarrays

- Microarrays can be fabricated using different technologies
- Probes can be oligonucleotides, cDNA, small PCR fragments related to specific mRNA
- The probes are synthesized and then placed on the support
- Probes can have different length (25-60nt).

Microarrays (types)



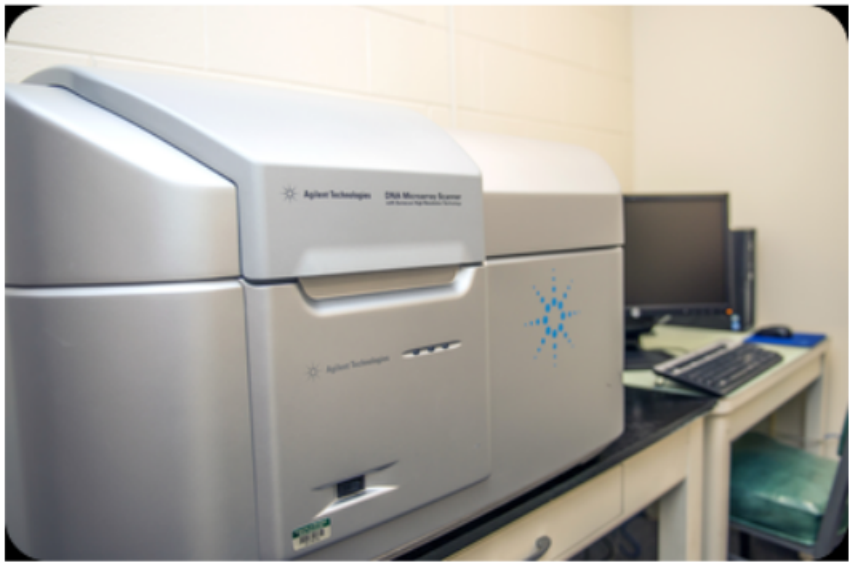
2 channels

- test vs control samples
- labelling with different fluorophores
- comparative technique

1 channel

- One sample per time
- More samples are needed

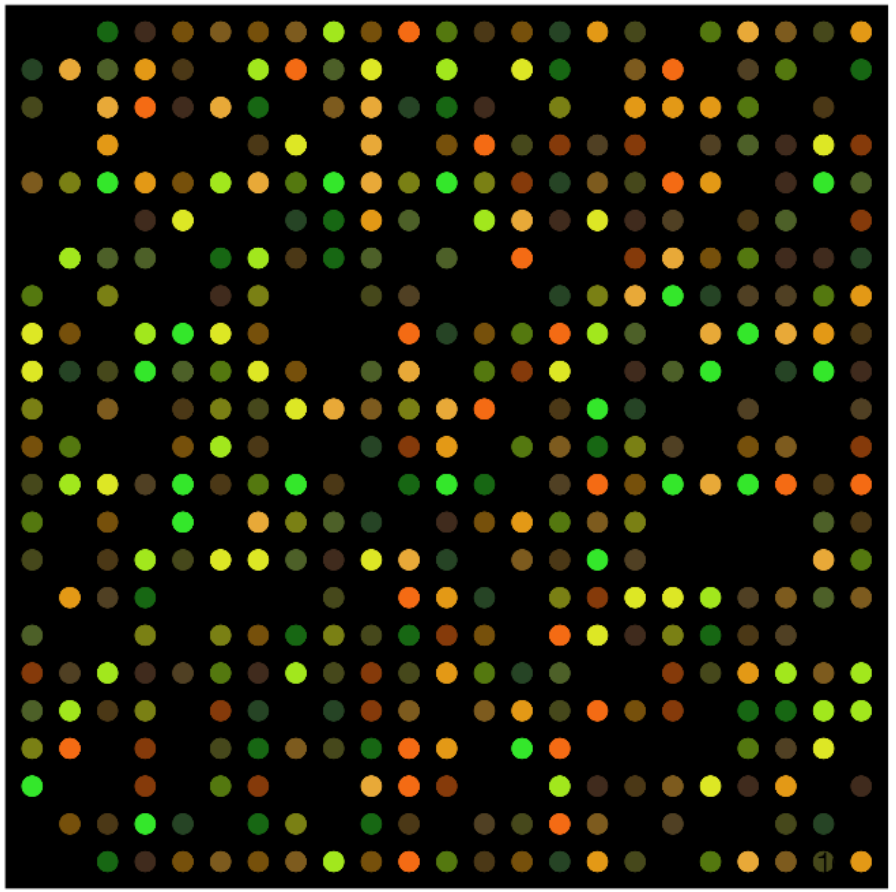
Reading system



- A “scanner” allows to read the fluorescence light emitted by the fluorophores
- The information is store in 2 images (2 channels) at 16 bit resolution

Image analysis

- The image is in grey scale but is usually represented in a red/green scale that represents the light emitted by the two fluorophores



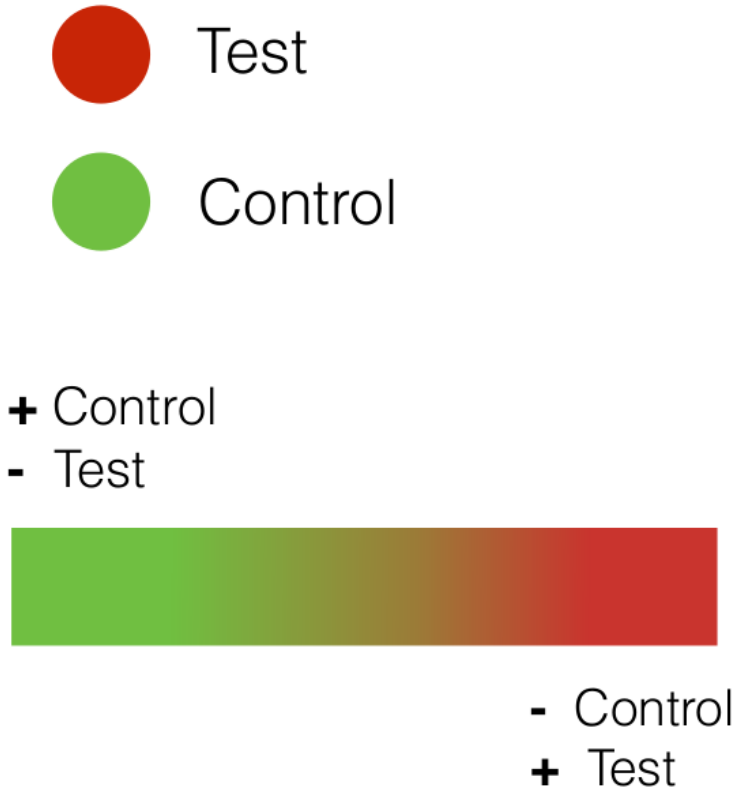
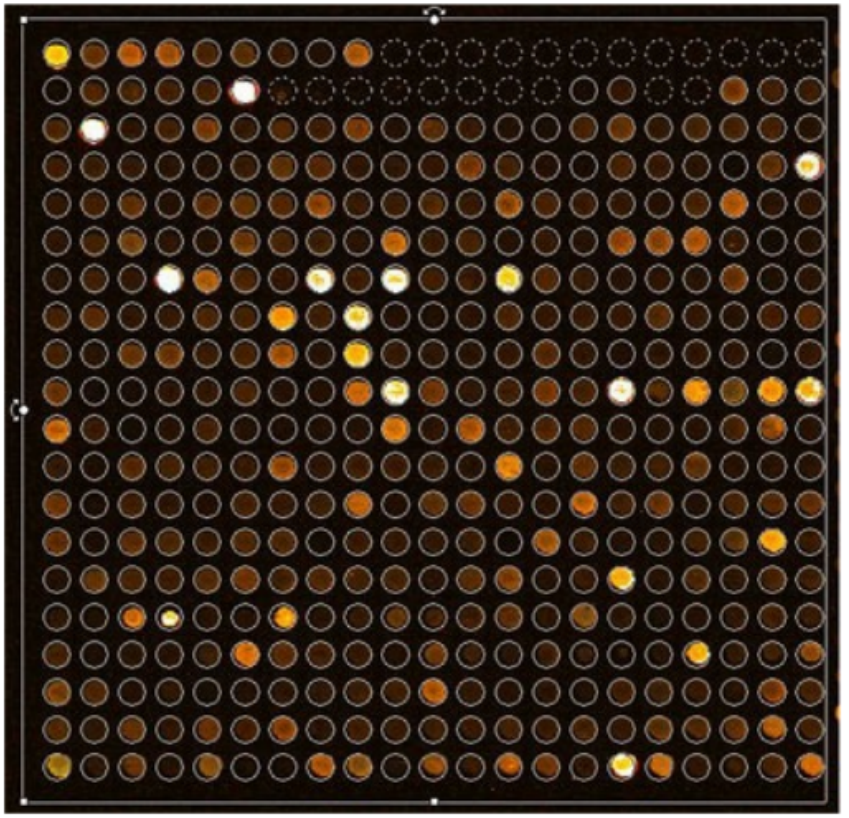


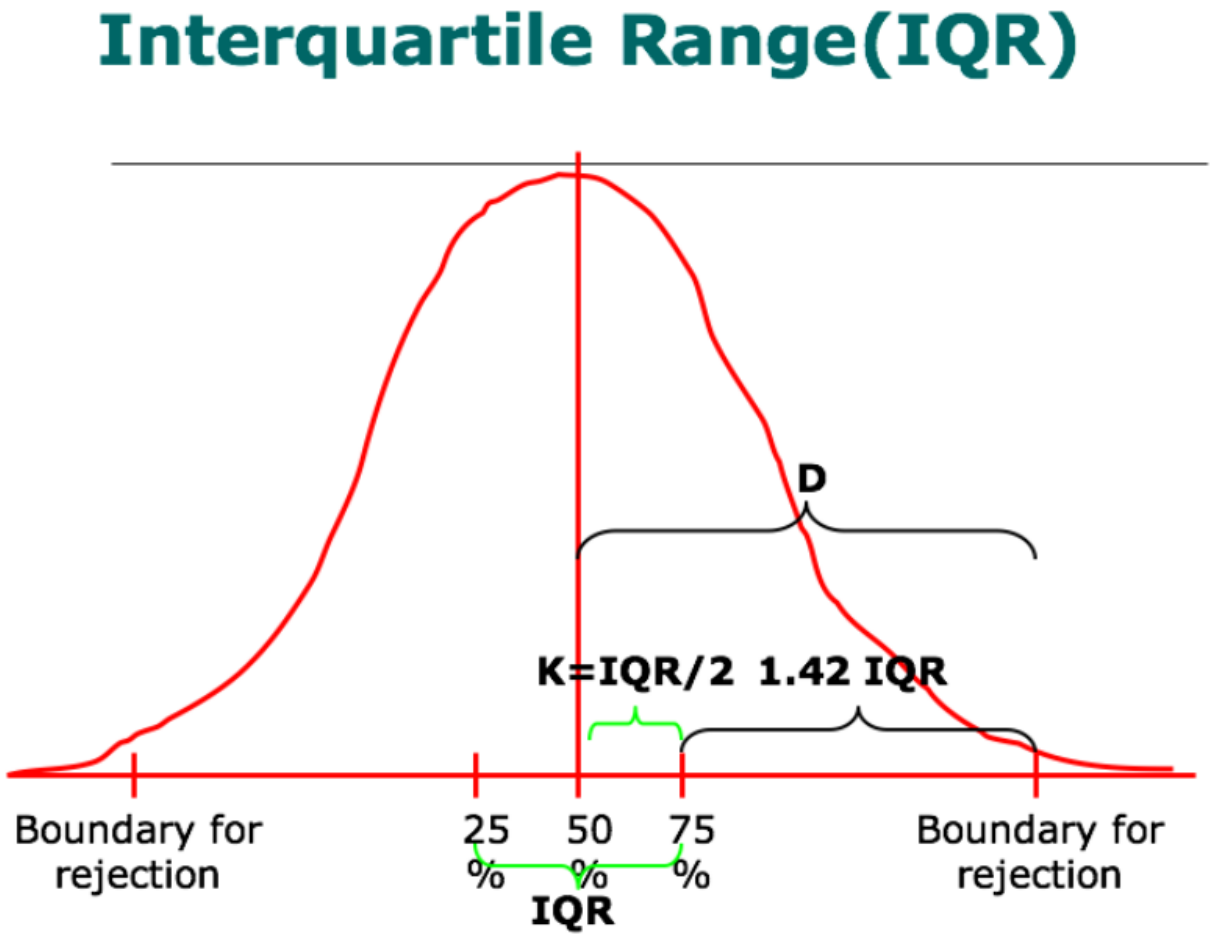
Image analysis

- Positioning of the grid
- Segmentation
- Estimation of the fluorescence for each spot
  - Information extraction
- Results writing



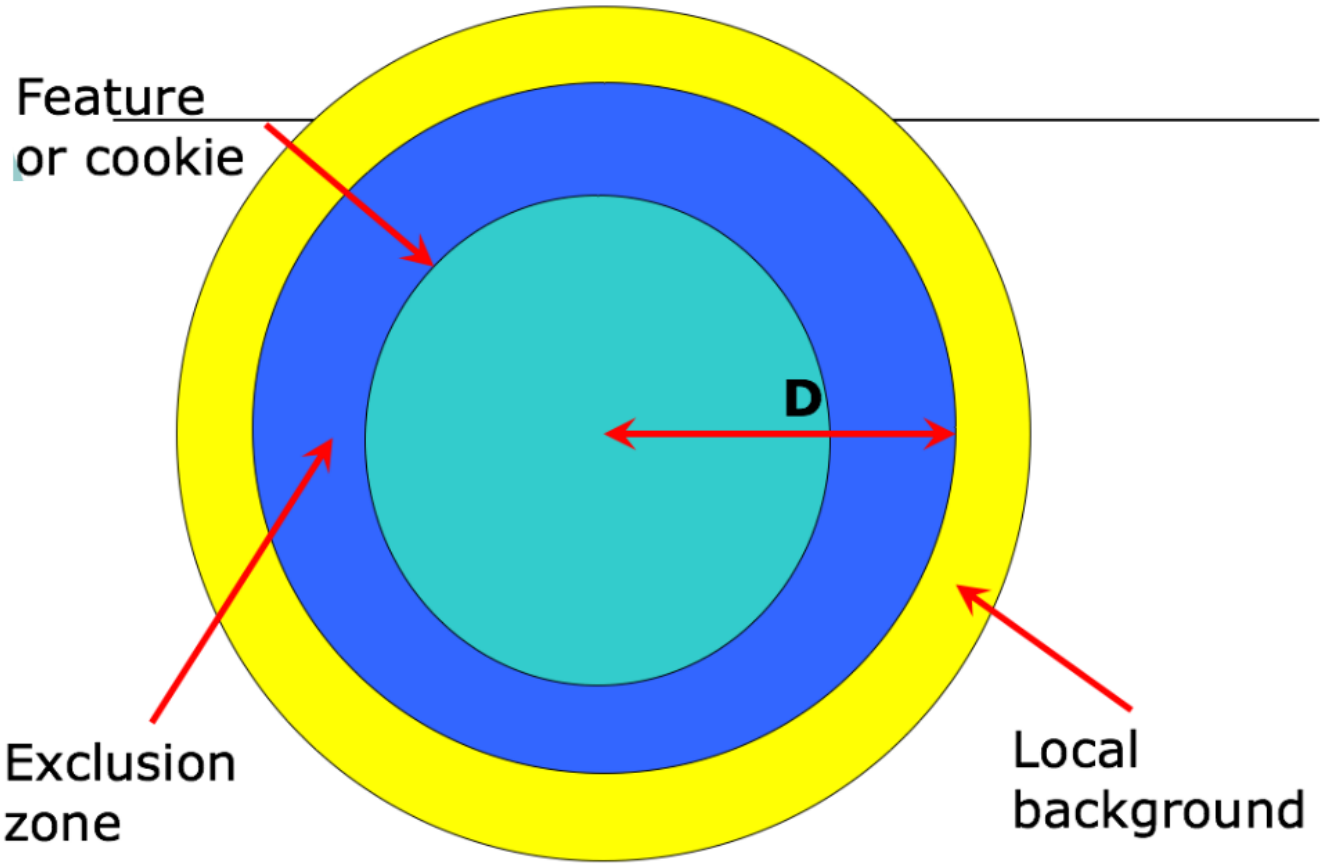
Segmentation

Background and *foreground* at each spot



Segmentation

Background and *foreground* at each spot



Segmentation

Quality control

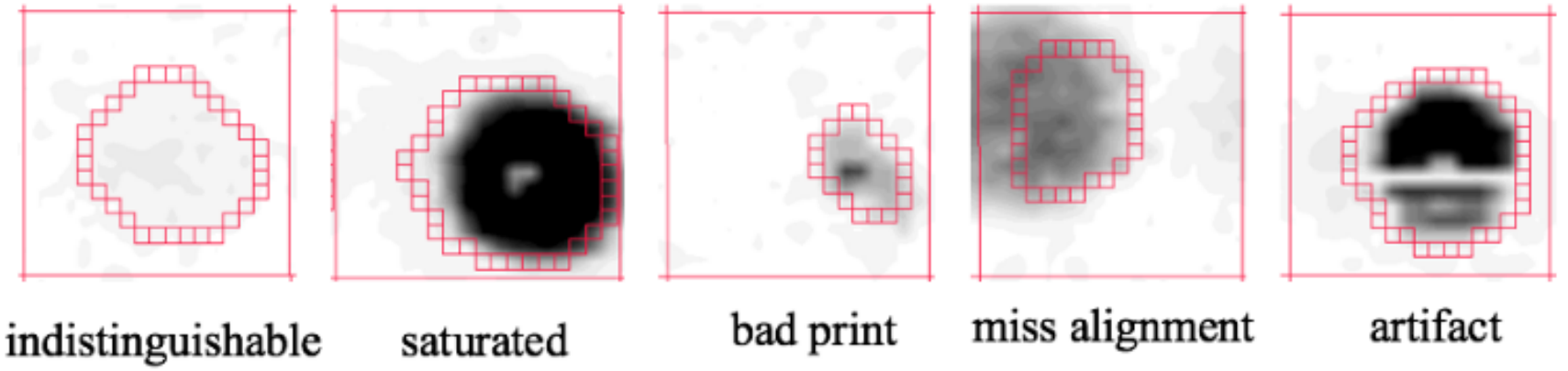


Image analysis

- Each microarray is provided with a specific design file

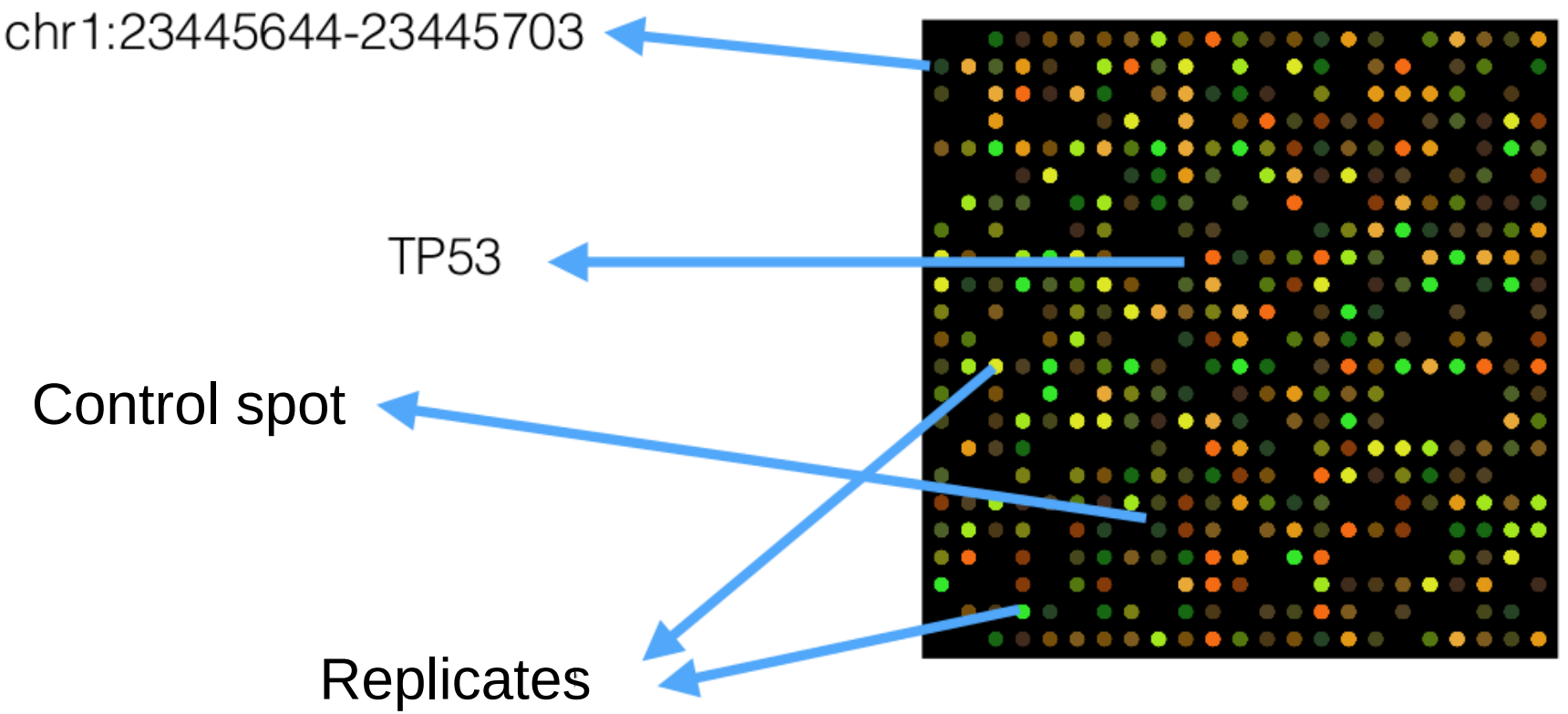


Image analysis

Green intensity

Green background

Green b.g.-corrected

Red intensity

Red b.g.

Red b.g.-corrected

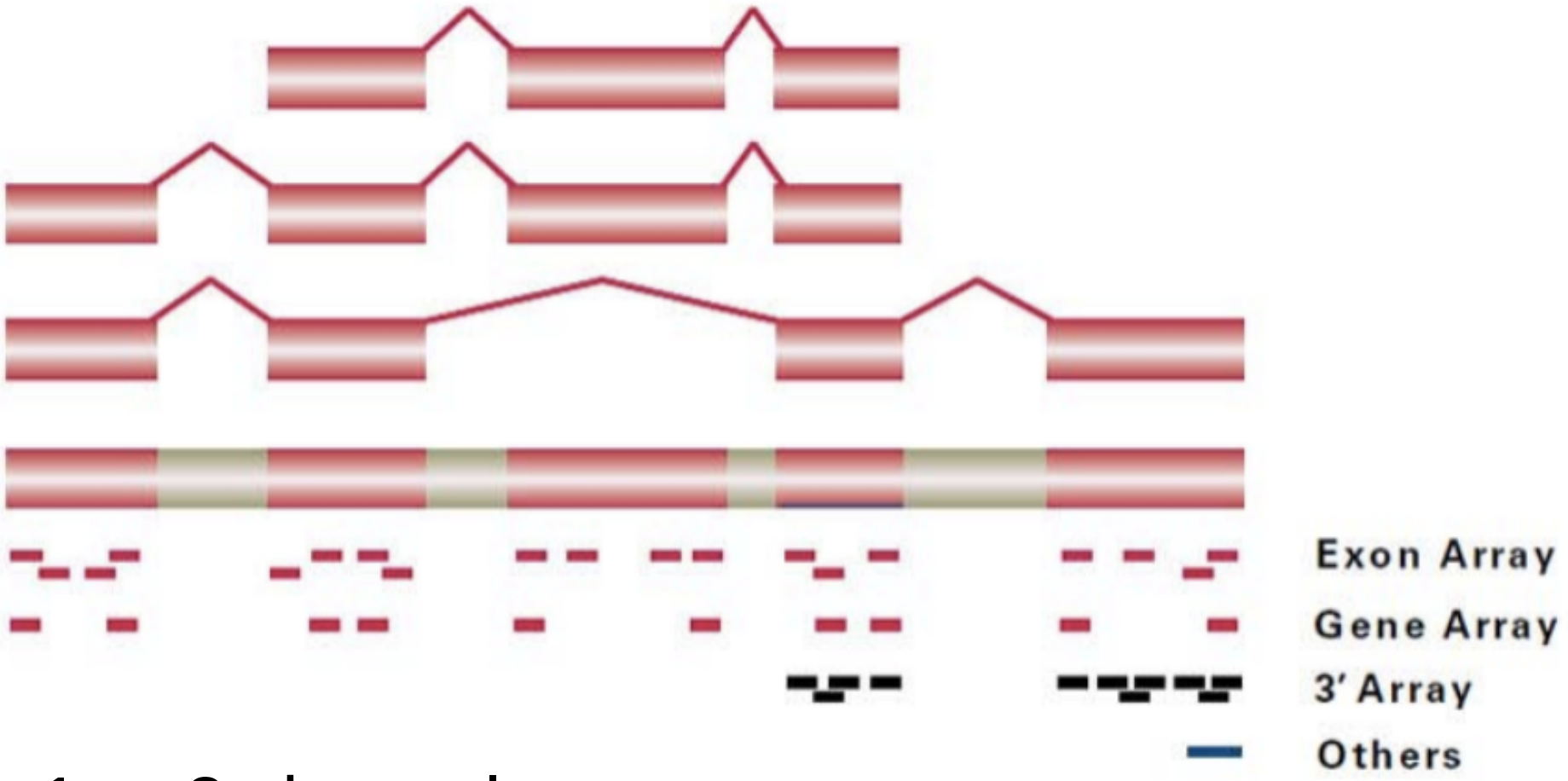
(R. b.g.-c)/(G. b.g.-c)

Systematic name

Gene function

	Ctrl D x A – PS	Ctrl Bkgd	Ctrl sDxA	Data D x A – PS	Data Bkgd	Data sDxA	Ratio (sDxA): Data / Ctrl		
A_1_1	59358.75	512.92	58845.83	50953.13	1779.913	49173.22	0.835628	YAL003W	translation elongation factor eef1beta
A_1_2	1209.19	512.92	696.271	2522.345	1779.913	742.4323	1.066298	YAR053W	hypothetical protein
A_1_3	1948.2	512.92	1435.28	3100.152	1779.913	1320.239	0.919848	YBL078C	essential for autophagy
A_1_4	4940.806	512.92	4427.886	6670.604	1779.913	4890.691	1.104521	YAL008W	protein of unknown function
A_1_5	1485.59	512.92	972.671	2916.086	1779.913	1136.173	1.168096	YAR062W	putative pseudogene
A_1_6	32642.03	512.92	32129.11	42304.13	1779.913	40524.22	1.261293	YBL087C	60s large subunit ribosomal protein l23.e
A_1_7	6919.441	512.92	6406.521	8540.246	1779.913	6760.333	1.055227	YAL014C	
A_1_8	2698.301	512.92	2185.382	4314.47	1779.913	2534.557	1.159778	YAR068W	strong similarity to hypothetical protein yhr21
A_1_9	7167.958	512.92	6655.038	7379.286	1779.913	5599.373	0.841374	YBL100C	questionable orf
A_1_10	5470.062	512.92	4957.142	6953.799	1779.913	5173.886	1.043724	YAL025C	nuclear viral propagation protein
A_1_11	27879.49	512.92	27366.57	33746.9	1779.913	31966.99	1.168103	YBL002W	histone h2b.2
A_1_12	2589.613	512.92	2076.693	4385.568	1779.913	2605.655	1.254713	YBL107C	hypothetical protein
A_1_13	6196.245	512.92	5683.326	8840.475	1779.913	7060.562	1.242329	YDR044W	coproporphyrinogen iii oxidase
A_1_14	34737.1	512.92	34224.18	36129.62	1779.913	34349.7	1.003668	YDR134C	strong similarity to flo1p, flo5p, flo9p and ylr1
A_1_15	34035.35	512.92	33522.43	27128.53	1779.913	25348.62	0.756169	YDR233C	similarity to hypothetical protein ydl204w
A_1_16	1638.381	512.92	1125.461	2988.042	1779.913	1208.129	1.073453	YDR048C	questionable orf
A_1_17	3873.718	512.92	3360.799	4955.141	1779.913	3175.228	0.944784	YDR139C	ubiquitin-like protein
A_1_18	2433.625	512.92	1920.706	3502.406	1779.913	1722.493	0.896802	YDR252W	strong similarity to egd1p and to human btf3
A_1_19	1800.736	512.92	1287.816	3011.855	1779.913	1231.942	0.956613	YDR053W	questionable orf
A_1_20	1296.689	512.92	783.77	2636.549	1779.913	856.6356	1.092968	YDR149C	questionable orf
A_1_21	3453.24	512.92	2940.32	4968.026	1779.913	3188.113	1.084274	YDR260C	hypothetical protein
A_1_22	10731.55	512.92	10218.63	9307.246	1779.913	7527.333	0.736629	YDR056C	hypothetical protein
A_1_23	6191.309	512.92	5678.39	8808.398	1779.913	7028.485	1.23776	YDR152W	weak similarity to c.elegans hypothetical prot
A_1_24	3589.998	512.92	3077.078	4420.744	1779.913	2640.831	0.858227	YDR269C	questionable orf
A_1_25	27568.34	512.92	27055.42	20856.2	1779.913	19076.29	0.705082	YGL189C	40s small subunit ribosomal protein s26e.c7
A_1_26	1956.182	512.92	1443.262	3150.716	1779.913	1370.803	0.949795	YGL261C	strong similarity to members of the srp1/tip1

Expression microarrays



1 or 2 channels

Data pre-processing

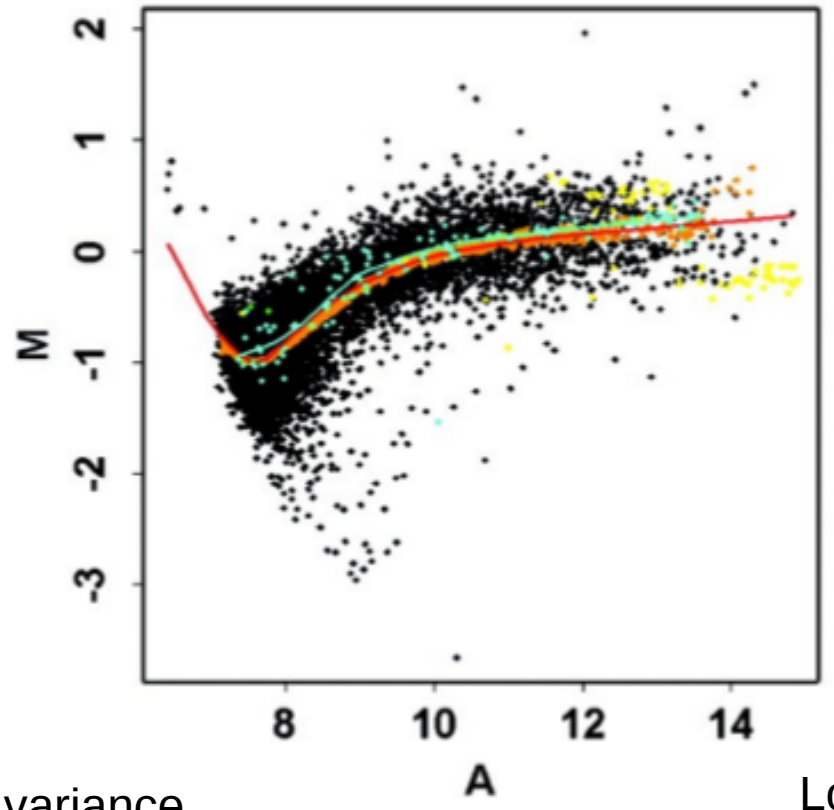
- You choose the parameters, software does the work
- Background subtraction:
  - Eliminates background noise
- Normalization:
  - This step takes care of
    - Unequal quantity of starting sample
    - Difference in labeling efficiency
    - Difference in detection efficiency
    - System biases, etc.
  - Brings all samples into a similar range of distribution
- Summarization
  - Summary of information from several spots into a single measure for each gene
- Statistical QC
  - Removes low quality samples and probesets

2 channels array

- Background correction
  - Separate Signal (Rs, Gs) and Background (Rb, Gb) estimates
  - Background corrected estimates (Rc, Gc)
    - Rc = Rs – Rb, Gc = Rs – Rb, OR (better)
    - Rc = max(Rs -Rb, 0), Gc = max(Gs -Gb, 0)
- Summarization & Transforms: log-Ratios
  - Estimate relative expression as log(Rc/Gc)

MA normalization

- Useful to identify systematic intensity-dependent bias in the data



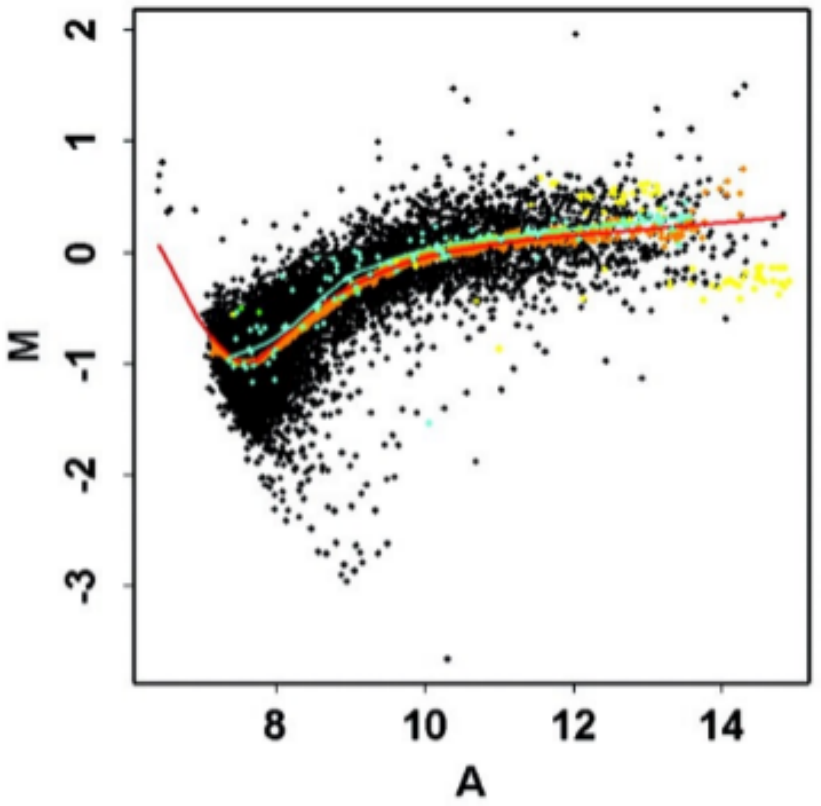
$$M = \log2(R/G)$$

$$A = 0.5*\log2(R*G)$$

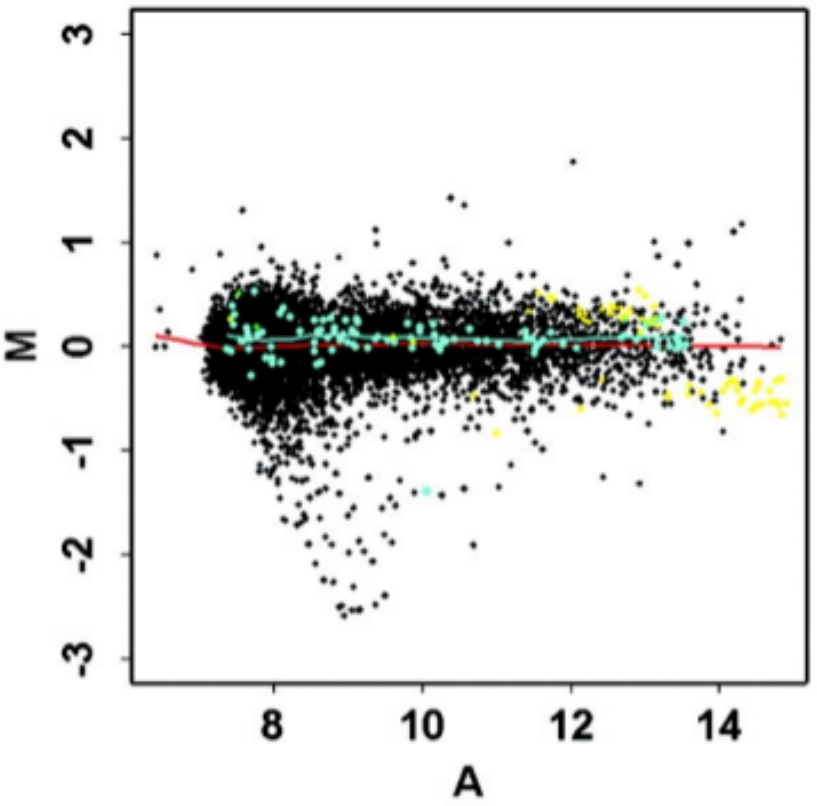
High variance

Low variance

MA normalization



For the majority of the probes M should equal 0



Loess (polynomial regression) normalization

y = f (A)
M’ = M - y

1 channel array

- Many methods have been developed to preprocess affymetrix arrays.
  - Advanced methods : GCRMA, PLIER
  - Popular methods: RMA and MAS5
  - Rudimentary methods: MAS4, LOESS

RMA (Robust Multi-array Average)

- Background correction
  - Remove local artifacts and “noise”
    - so measurements aren’t so affected by neighboring measurements
- Normalization
  - Remove array effects
    - so measurements from different arrays are comparable
- Summarization
  - Combine probe intensities across arrays
    - so final measurement represents gene expression level

RMA (background correction)

- Assumes PM (Probe Measure) data is combination of background and signal

  PM = Signal + Background, where

  - Signal: $S \sim \exp(\lambda)$ and
  - Background: $B \sim N(\mu,\sigma 2)$

- By assuming strictly positive distribution for signal background corrected signal is also positively distributed
- Background correction performed on each array seperately
- Estimate μ, σ, and a separately in each chip using the observed distribution of PMs
- By introducing them in the above formula we obtain an estimate of E(S|PM) for each PM value

RMA (normalization)

- Normalizes across all arrays to make all distributions the same
- ‘Quantile Normalization’ used to correct for array biases
- Compares expression levels between arrays for various quantiles
- Can view this on quantile-quantile plot
- Protects against outliers

RMA (summarization)

- Combine intensity values from the probes in the probe set to get a single intensity value for each gene (probeset)
- Uses ‘Median Polishing’
  - Each chip normalized to its median
  - Each gene normalized to its median
  - Repeated until medians converge
  - Maximum of 5 iterations to prevent infinite loops.

Gene expression microarray

- An expression microarray experiment is used to test differences in gene expression between two (or more) conditions
  - Conditions: cancer vs normal, treatment A vs B, ...
- Each condition can be represented by one or more samples

Gene expression microarray

- The null hypothesis to test is that *there exist no difference between the gene expression in the two conditions*
- The comparison between the samples is done using the ratio between the test and the control samples
- The ratio should not differ in case of null hypothesis validity
- These ratios are also defined as *fold changes*

Fold change

Ratio	FC
1	0
2	2
3	3
1/2	-2
1/3	-3

FC = Ratio if Ratio > 1
FC = -1/Ratio if Ratio <1

Fold change

- Ratios are not symmetric with respect to 1
- This complicates statistics
- We should then use the logarithm
- The log-ratio of the null hypothesis should be 0

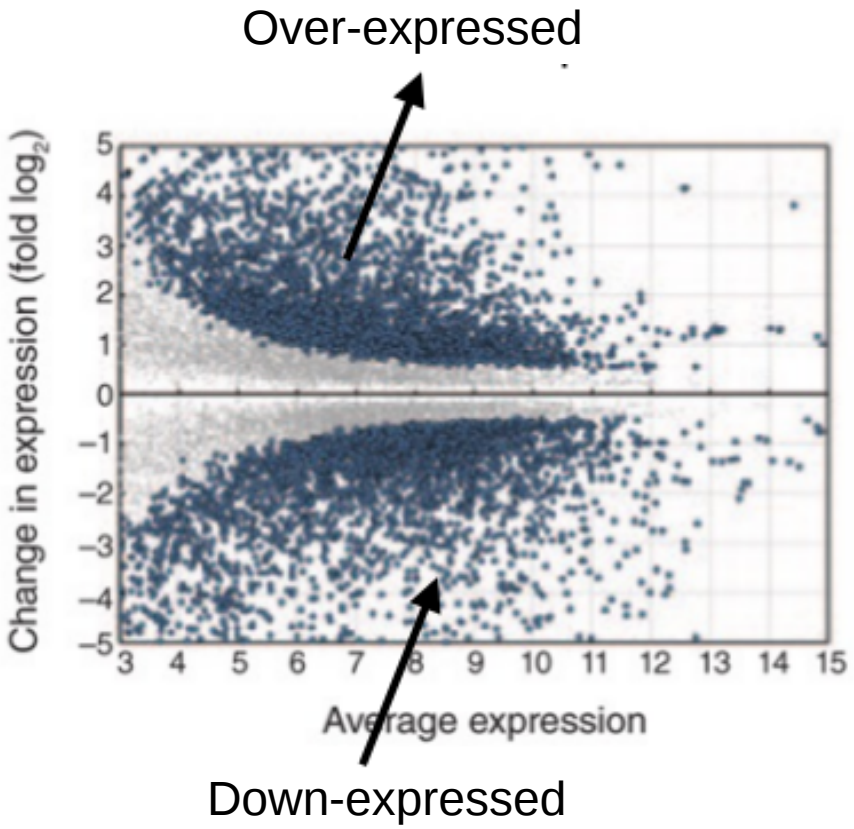
Replicates

- Replicates are needed considering the noise of microarray data
- Technical replicates:
  - Experiments on more RNA samples obtained from the same biological source
- Biological replicates:
  - Experiments on more biological sources belonging to the same condition
- Ideally each condition should be represented by more biological replicates in order to perform a statistical test
- Replicates can also be summarized as mean
  - for each gene

Gene expression table

	Condizione A			Condizione B		
	Sample1	Sample 2	..	Sample m	Sample m+1	...
Gene1	1050	1010	..	1022	999	..
Gene2	30222	35156	..	20111	19053	..
...	..	..	..	..		..
Gene m	10134	50222	..	12222	15560	..

MA plot



MA correlation can be exploited to identify differentially expressed genes.

A gene il called differentially expressed through a Z statistics.

$$Z = \frac{X - \mu}{\sigma}$$

Calculated at different bins of A

Statistical tests

- To find differentially expressed gene we should use a test statistics for each gene
- Test statistics provide p-values
- A low p-value is interpreted as evidence that the null hypothesis can be false
  - A gene is differentially expressed

Statistical tests

- T-test
  - Parametric test to check the difference between the mean of two groups
  - Assumes variance of the two groups is the same

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

- Walch t-test
  - Considers different variance between the two groups
  - Default implementation for R t.test() function

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{\sqrt{s_X^2/m + s_Y^2/n}}.$$

Statistical tests

- Wilcoxon test
  - Non parametric test to check equality of two distributions
- Permutation test
  - Generate a null distribution on an observation of interest by changing the group label
  - Compare the value observed in data versus value in the generated null distribution

$$p = \#\{b : |T_b| \geq |T_{obs}|\}/B$$

Multiple hypothesis correction

- Say you have a set of hypotheses that you wish to test simultaneously. Let's, consider a case where you have 20 hypotheses to test, and a significance level of a = 0.05. What's the probability of observing at least one significant result just due to chance?
  - P(at least one significant result) = 1 − P(no significant results)

    = 1−(1−0.05)^20

    ≈ 0.64
  - We have a 64% CHANCE to find one significant result randomly
- Correction methods:
  - Bonferroni (very conservative): significance threshold is a/N
  - FDR (False Discovery Rate): check if the k_th ordered p-value is larger than (k × a)/N

ANOVA

- ANOVA analysis
  - control vs treatment 1 vs treatment 2
  - ANalysis Of VAriances: Allows to test the null hypothesis that the differences within and between at least 3 groups are the same on average
- 2-way ANOVA (eg 2 cell lines, 2 treatments)
  - compares the mean differences between groups that have been split on two independent variables called factors
  - understand if there is an interaction between the two independent variables on the dependent variable

ROC curve

- ROC: Receiver operating characteristic
- Find genes that better discriminate between the two conditions