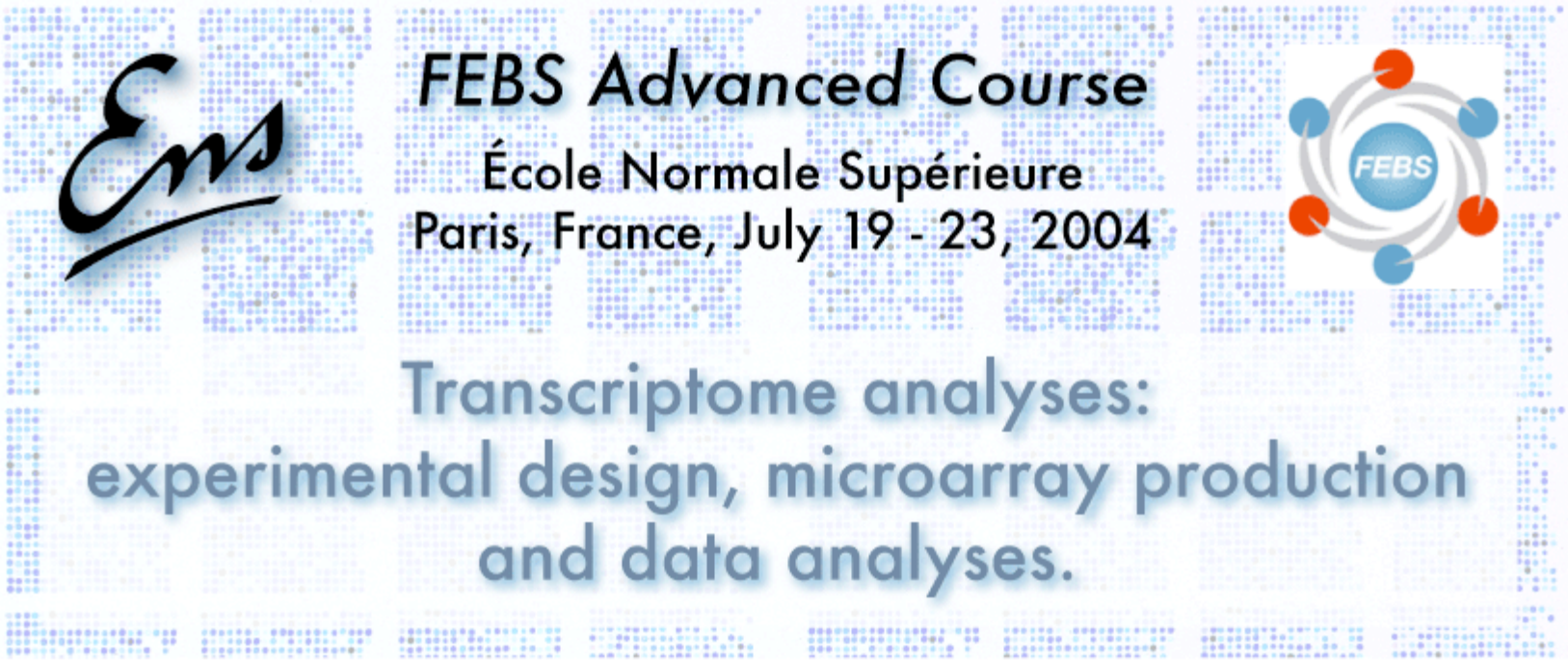
Lecture 3
Image analysis and normalisation

Stéphane LE CROM
lecrom@biologie.ens.fr



Large scale work revolutions

- *1900 - industrial revolution*

- *2000 - biology revolution?*

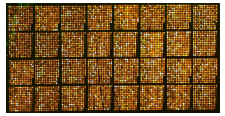
The various steps of a DNA microarray experiment

Experimental steps

Data analysis

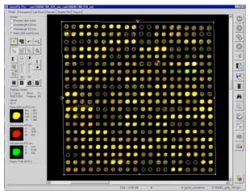
Experimental design set up

- **Chips on catalog**
- **Home made chips**



Hybridisation

Image analysis



Raw data treatment

- Normalisation
- Statistical analysis
- Storage

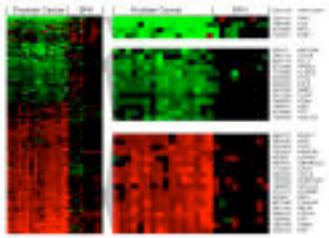
Data treatment

Available databases

Data mining

Data representation

- Clustering



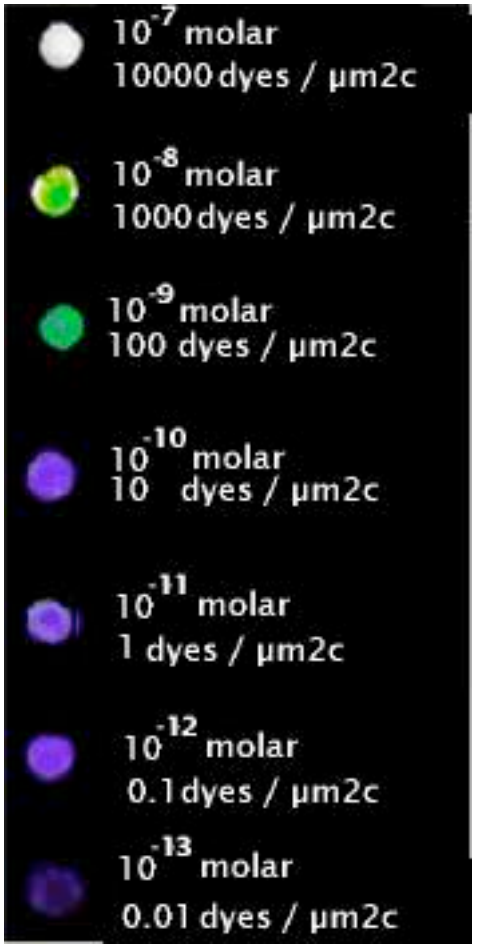
Data analysis

DNA microarray bioinformatic analysis

Image analysis

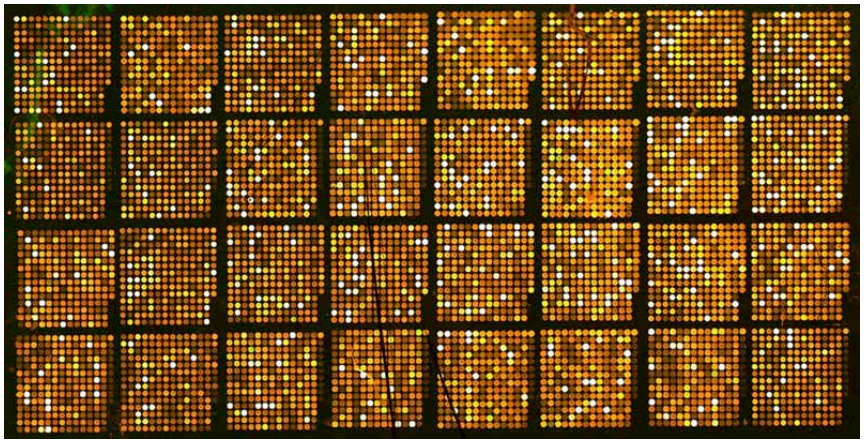
Various image type encountered

Color scale
=
Quantitative scale



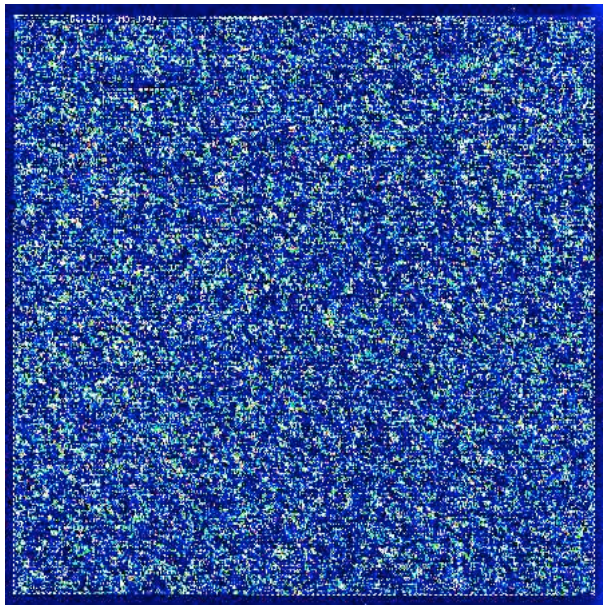
- cDNA/oligonucleotide arrays - Hitachi

- 2 channels
- Overlay
  (= Ratio)



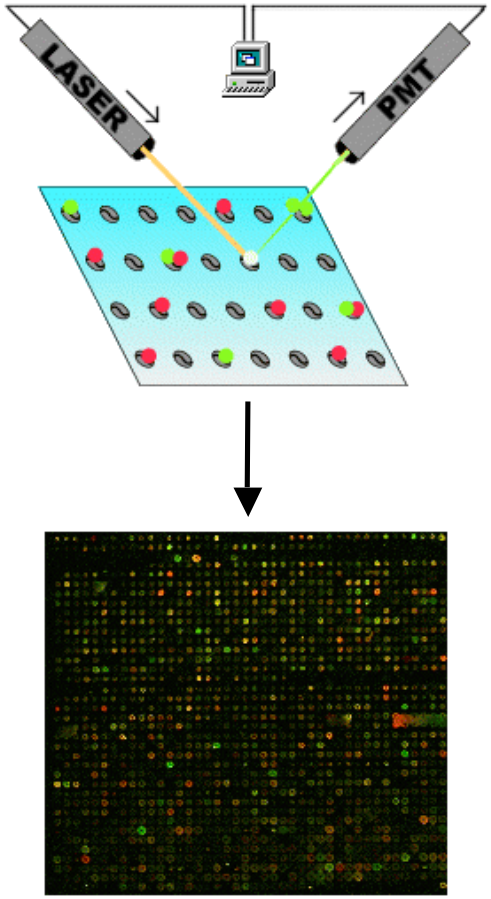
- Oligonucleotide chips (GeneChip) - Affymetrix

- 1 channel
- Intensity
  (= RNA quantity)

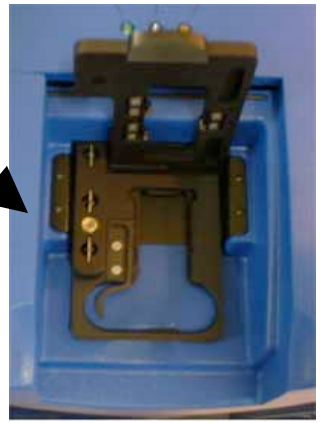


Slide scanning

- ***Fluorochrome excitation at a selected wavelength***

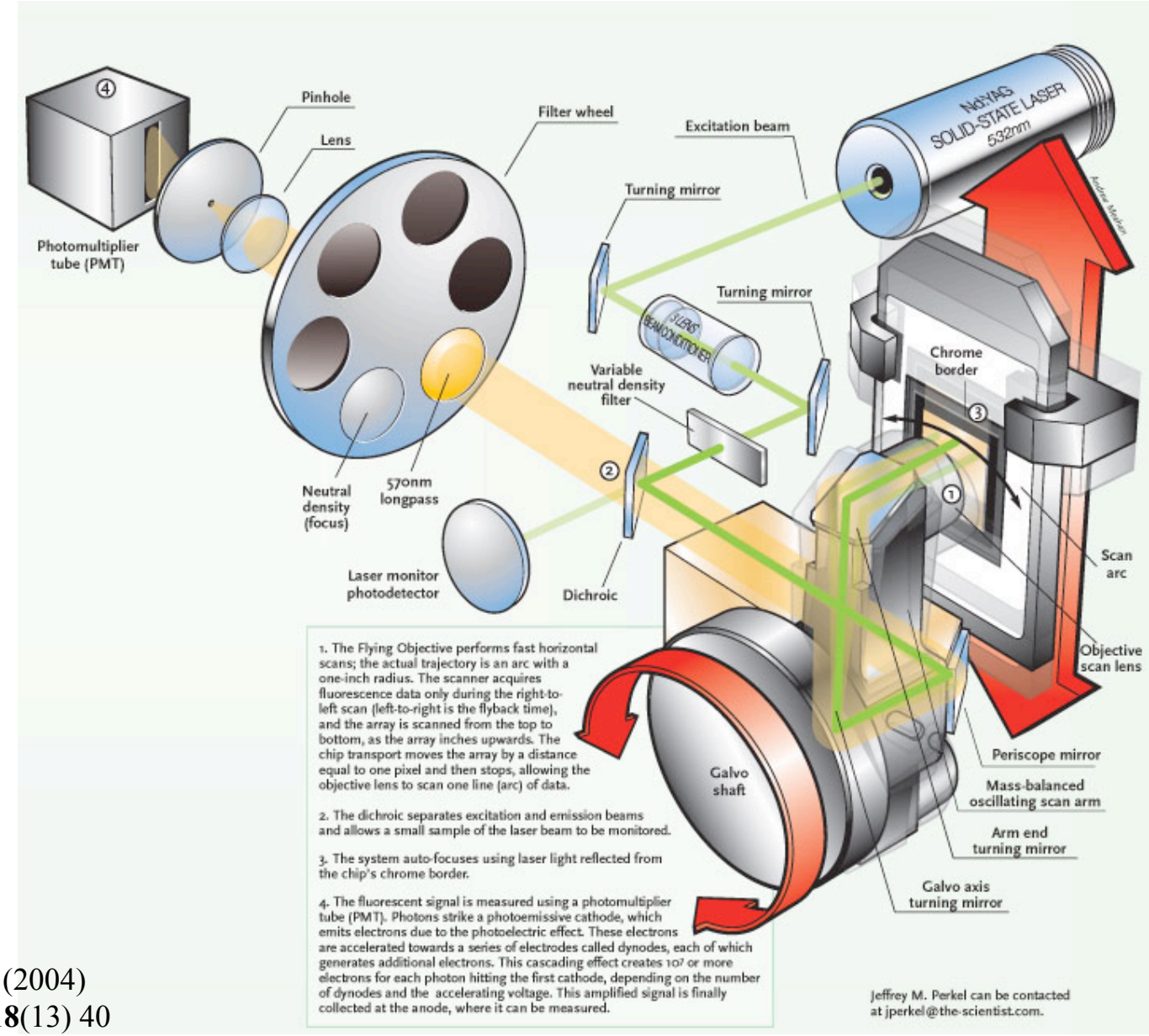


GenePix 4000 - Axon Instrument
(http://www.axon.com/GN_Genomics.html)



- ScanArray - Packard BioChip Technologies (http://www.packardbioscience.com)
- GeneMachine - Genomic Solutions (http://www.genomicsolutions.com)
- DNA Microarray Scanner - Agilent (http://www.chem.agilent.com)

How does a microarray scanner work ?



From Perkel J (2004)
The Scientist **18**(13) 40

Scanner setting verification

- <u>Scanner image acquisition :</u>

  - Confocal scanner

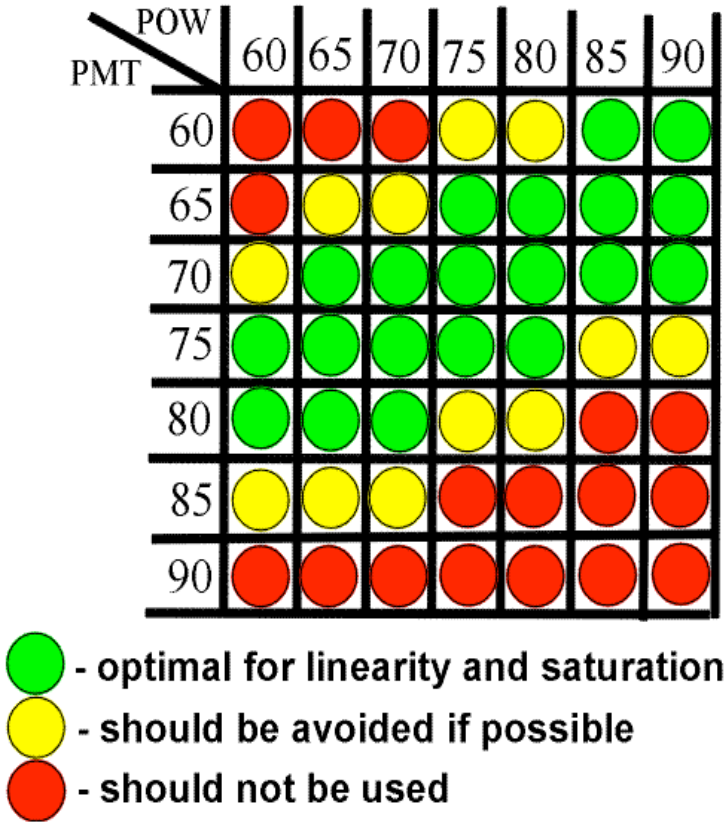
  - Two wavelength lecture

  - 16 bits TIFF image created (1 to 65536)

  - PMT power setting variable

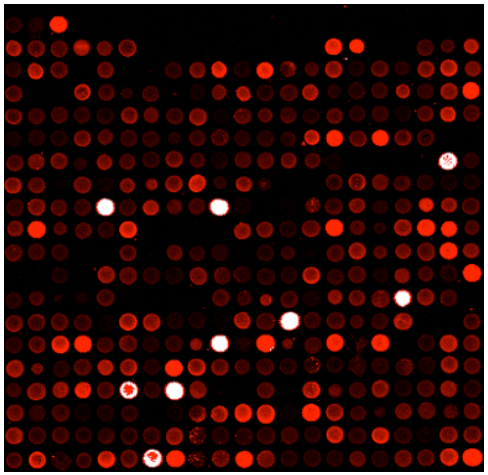
  - Linearity verification

Power and PMT settings for Cy3 and Cy5 lasers

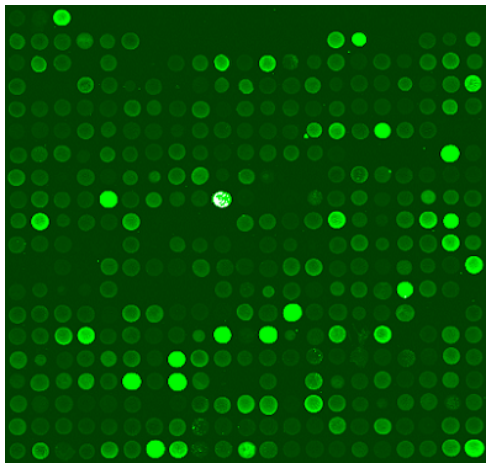


General Scanning 3000 scanner linearity

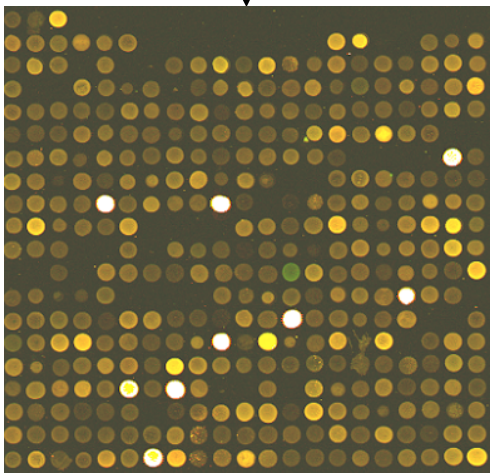
Image acquisition



Cy5 wavelength



Cy3 wavelength

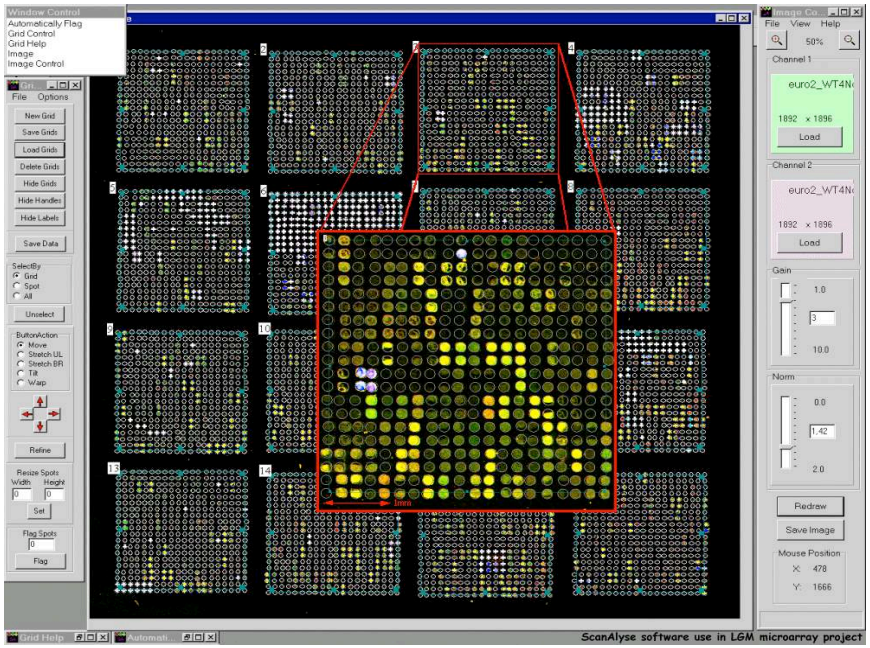


Final image

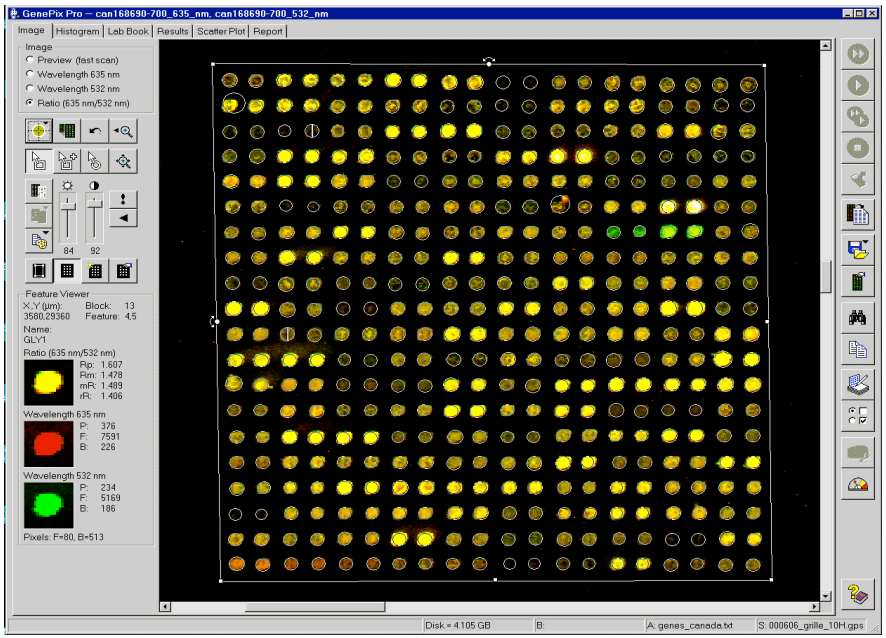
General principle of image analysis

- ***<u>Goal</u> : Convert the image pixels to digital values that quantify gene expression***

A lot of image analysis software are available



ScanAlyze
(M. Eisen Stanford University)



Genepix Pro
(Axon software)

Image analysis steps

1 — Define the spot localisation on the slide

For each spot:

2 — Find pixels belonging to the hybridisation zone

3 — Localise pixels to evaluate the background

4 — Calculate the global fluorescence intensity

On the whole slide:

5 — Identify the spots deformed by artefacts

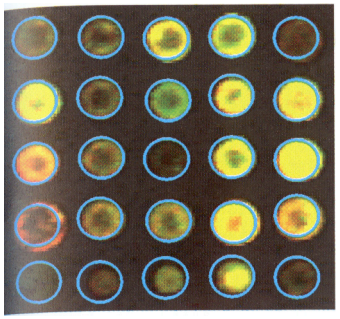
1- Spot localisation on the slide

It is crucial to assign each spot its correct gene identifier!

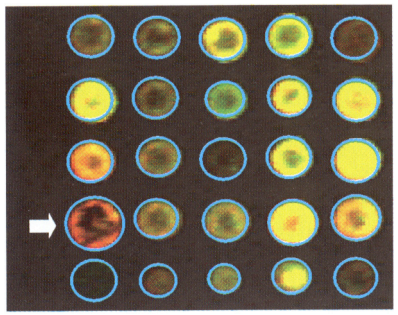
2 - Target detection (signal segmentation)

- ***Various methods are available:***

  - Fixed diameter circle
  - Variable diameter circle
  - Histograms
  - Adaptive shape



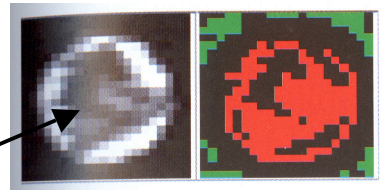
Fixed diameter circle



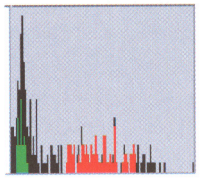
Variable diameter circle

- ***Several software are available:***

  - GenePix Pro
  - ScanAlyze
  - QuantArray
  - ImaGene
  - Dapple…



heterogeneous intensities



Histogram method

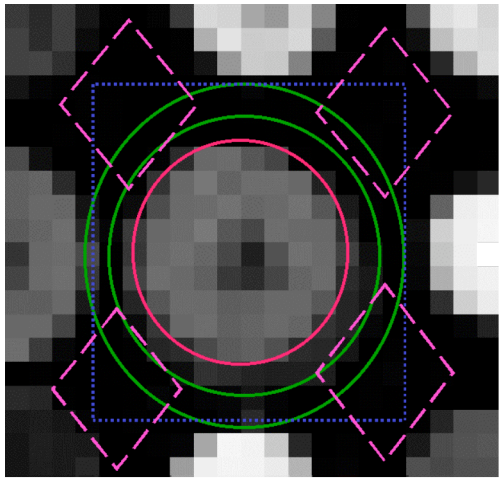
From "Microarray Bioinformatics", Dov Stekel

3 – Background evaluation

Observed signal have two components

Fluorescence coming from the specific hybridisation signal

Fluorescence coming from an unspecific hybridisation signal: **Background**

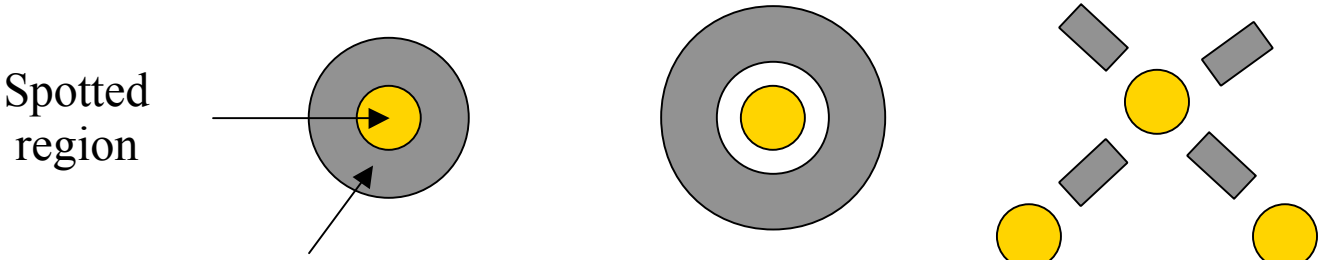


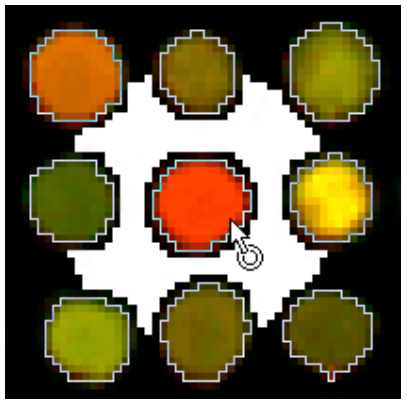
- ***Analysis of pixels localised closed to the spotting region:***

"ScanAlyze" method

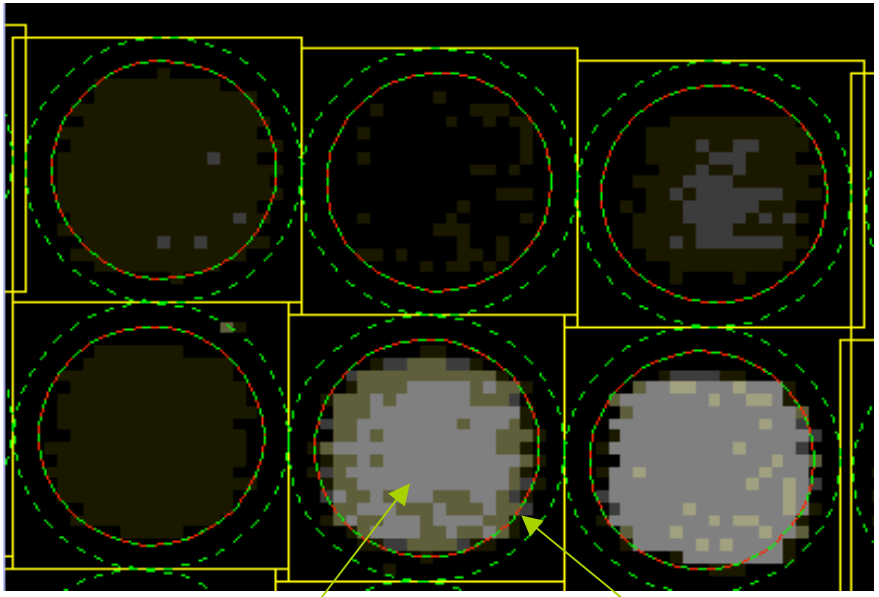
"ImaGene" method

"GenePix" method





4 – Calculation of the global fluorescence intensity



Raw signal
= intensity inside the spot

Background
measurement

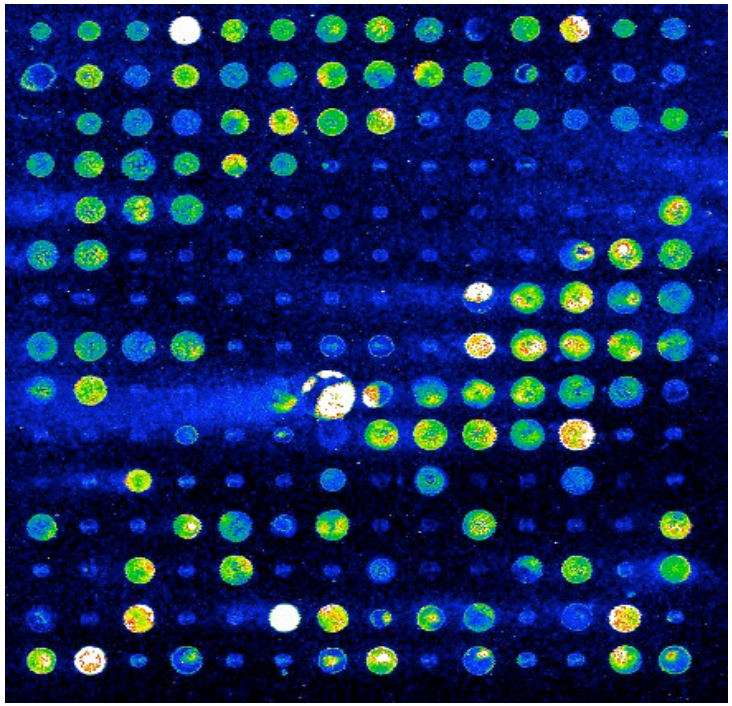
Net intensity
=
Raw intensity - Background

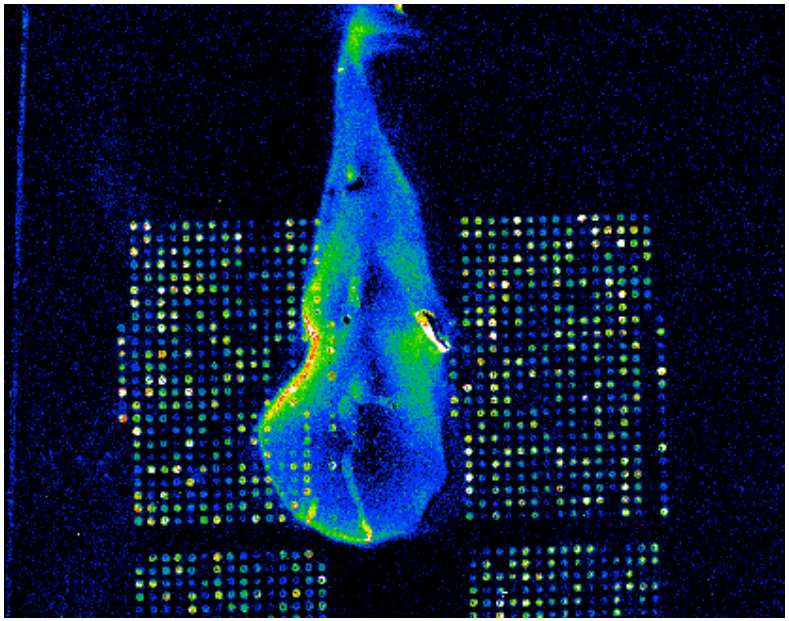
Intensity of each pixel → Mean or median? → Global intensity of the spot

Some quality criteria: the spot size, the standard deviation…

5 – Artefactual spot elimination

- ***Examples:***

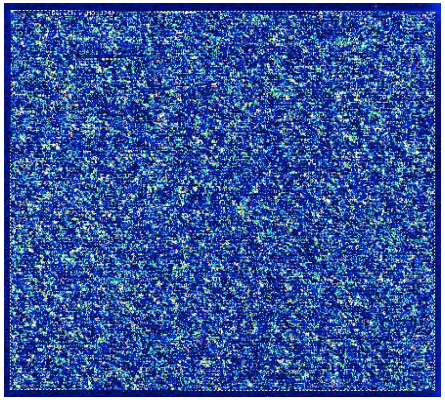




- ***Solutions:***

- Always look at the quality controls of the slide batch you ordered
- Refer to the troubleshooting guide (http://www.corning.com/lifesciences/technical_information/techdocs/troubleshootingUltraGAPS_ProntoReagents.asp)
- Apply manual or automatic flagging of artefactual spots

Affymetrix (GeneChip) arrays



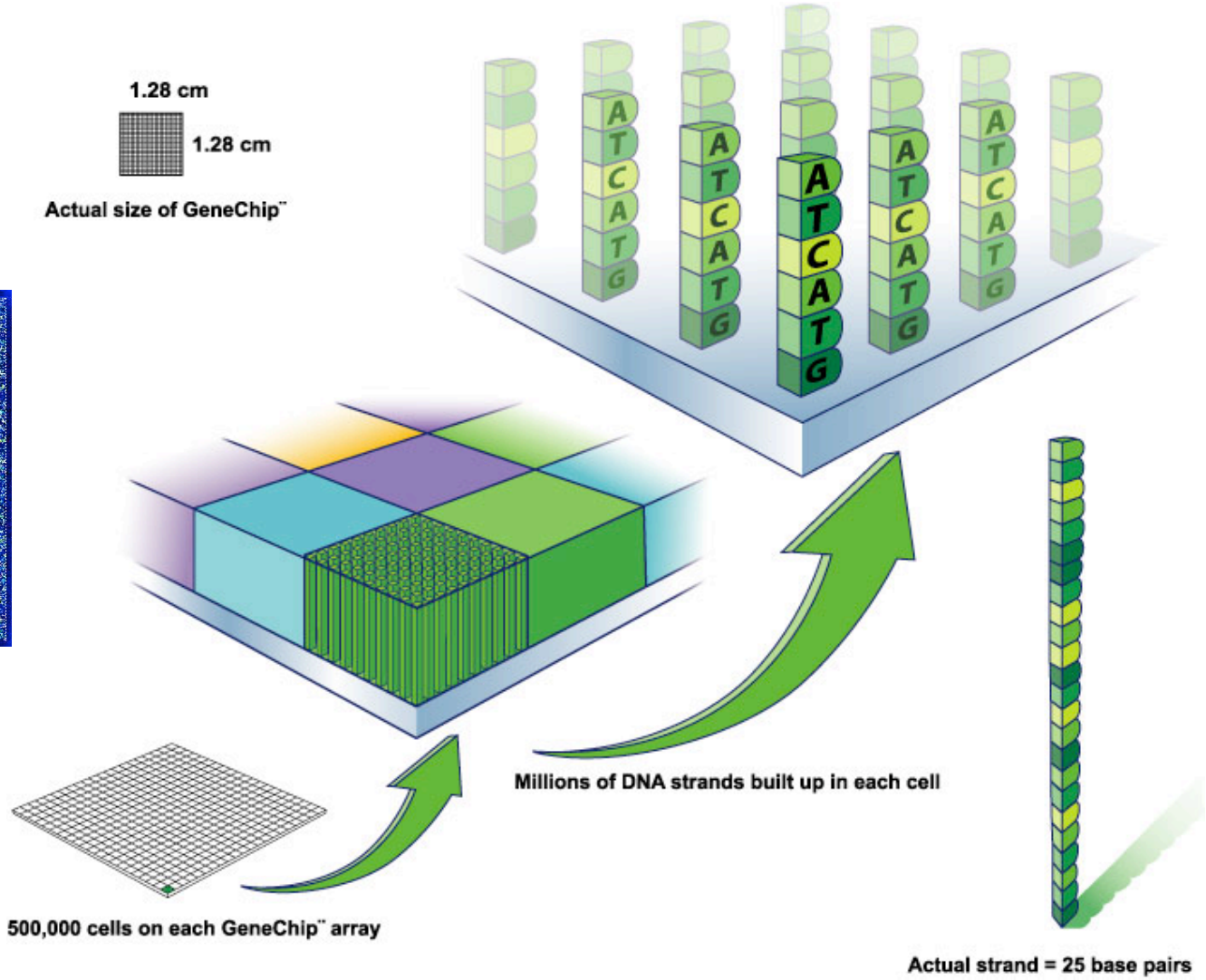
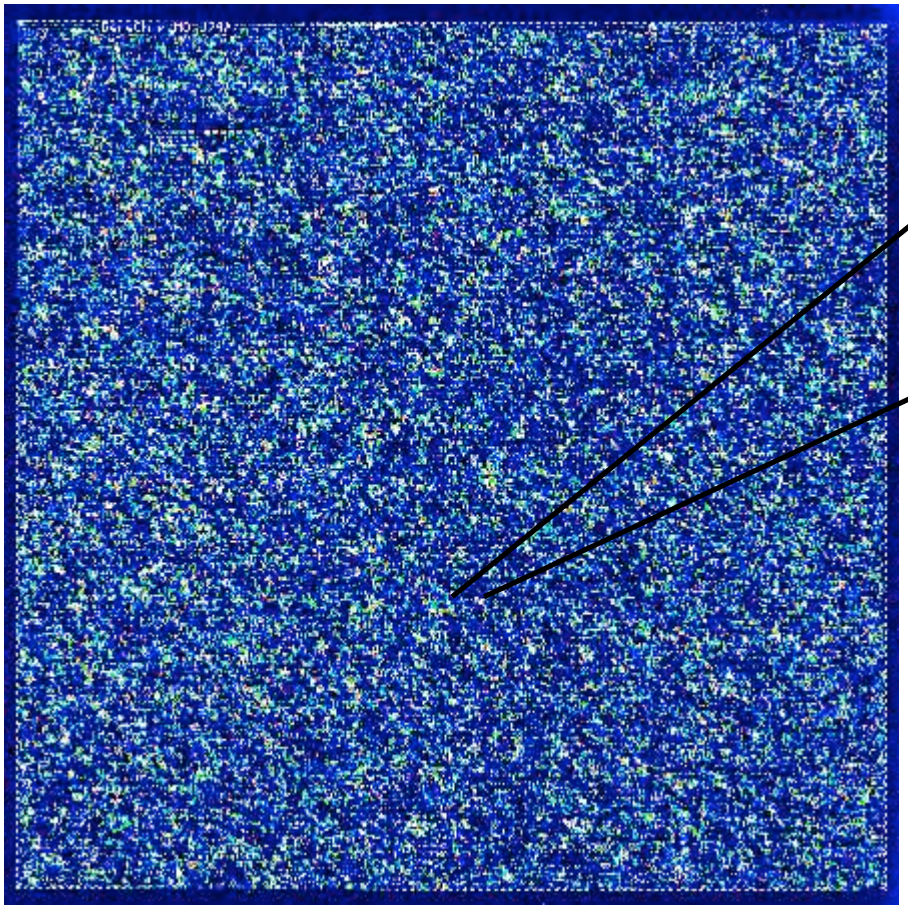
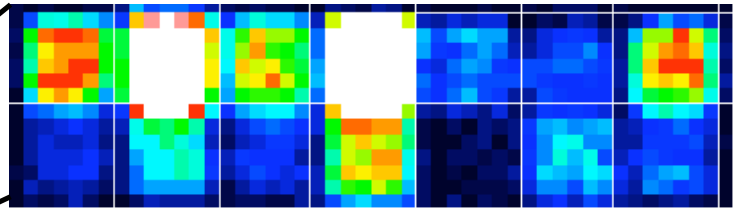


Image acquisition with Affymetrix GeneChip





PM

MM

Oligonucleotide pairs have been created for each gene (8-10):

- "perfect match" PM
- "mismatch" MM

Mouse slide = 12000 genes

Correct spot selection

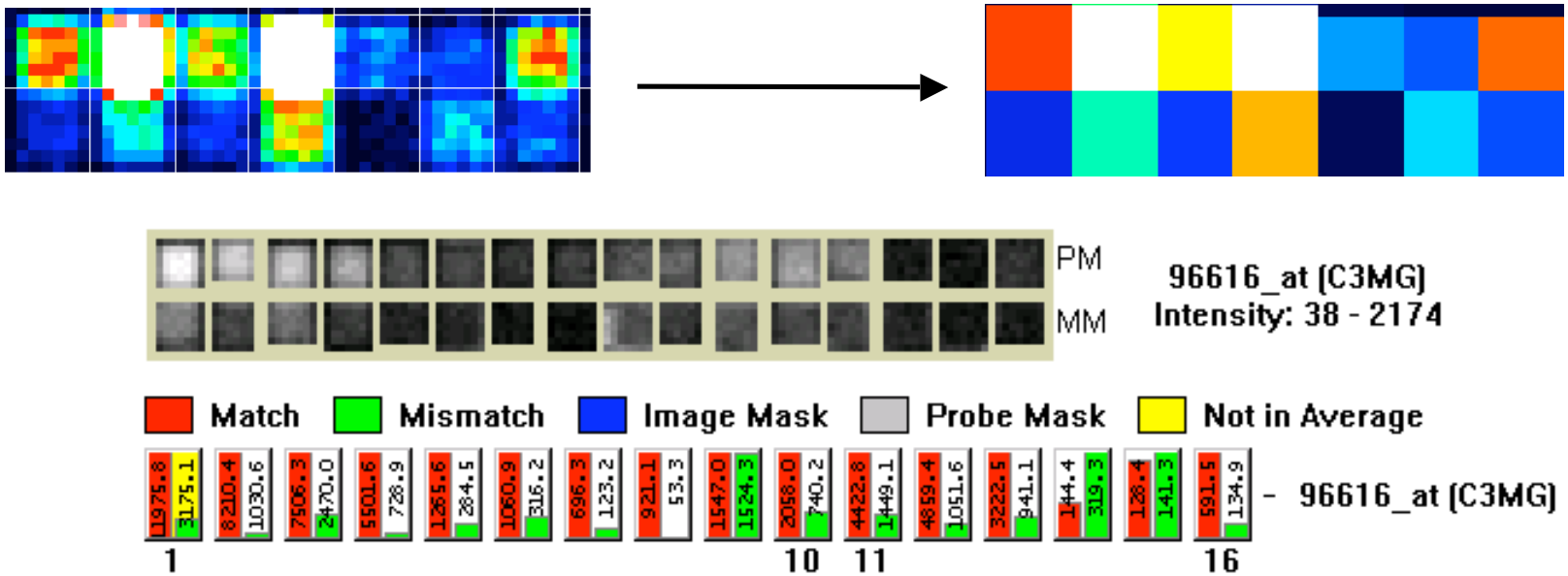
- ***Background calculation on the slide:***

  The measurement is based on the 2% cells with the lowest intensities in several blocs on the slide.

- ***Mean values estimation:***

Image obtained for one gene

Mean value calculated for each cell



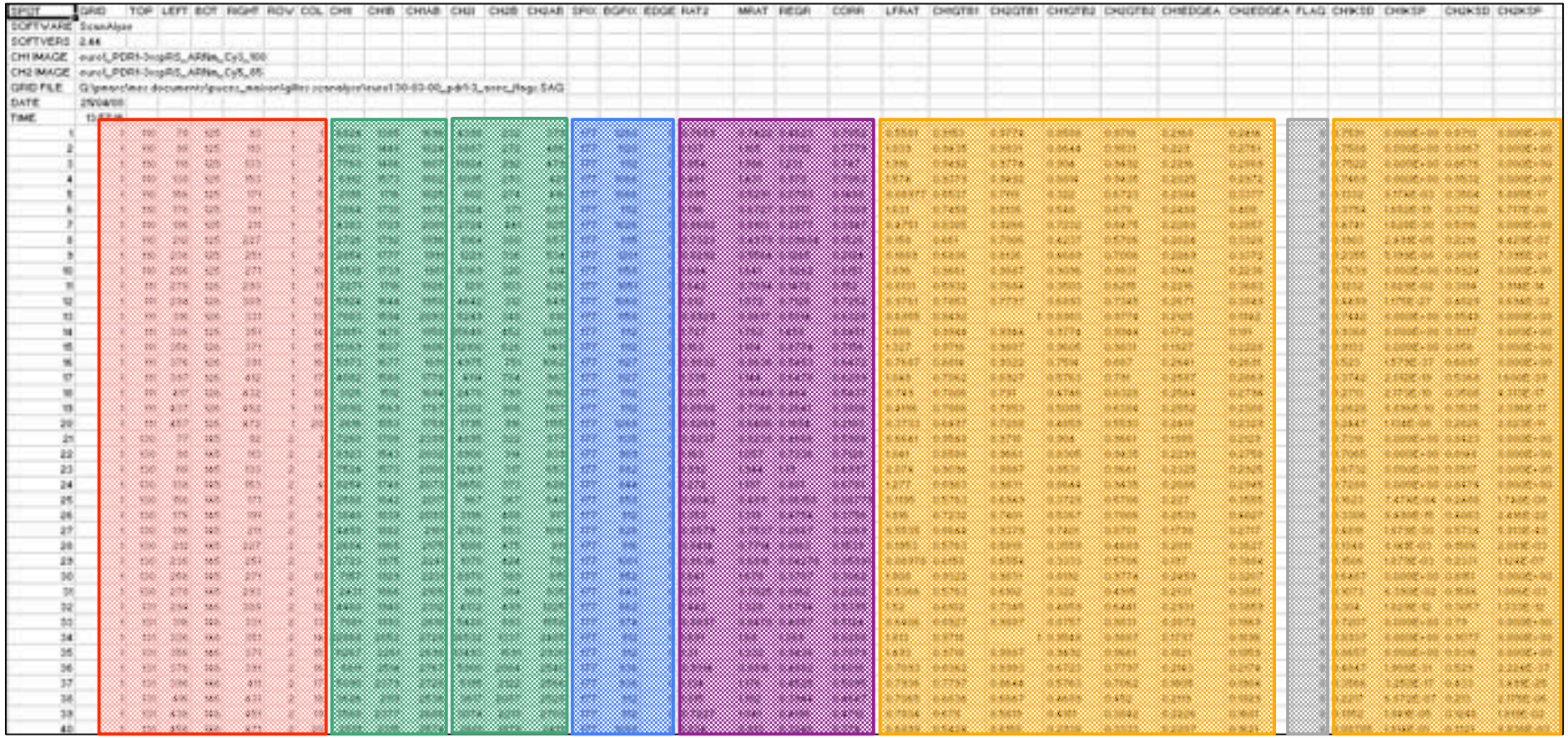
=> Calculation of the number of the positive and negative pairs (decision matrix)

=> Determination of the gene status (present, marginal or absent)

Image analysis results retrieval

- ***Example of an image analysis software output file:***

Position | Intensities (Raw, Bkg) | Size | Homogeneity | Statistics | Annotations



DNA microarray bioinformatic analysis

Within-array normalisation

Data variability sources with microarray

- DNA amount
- Efficiency of:
  - the RNA extraction
  - the reverse transcription
  - the labelling step
  - dyes incorporation

- the PCR yield
- the spotting quality
- the Unspecific cross-hybridisation effect

Systematic error

=> Same effects on various measurement
=> Corrections can be estimated from the data

Calibration

Stochastic error

=> Effects that crop up randomly and then can not be measured as noise

Error model

1st step: data cleaning and filtering

1— Elimination of the spots flagged as artefactual spots

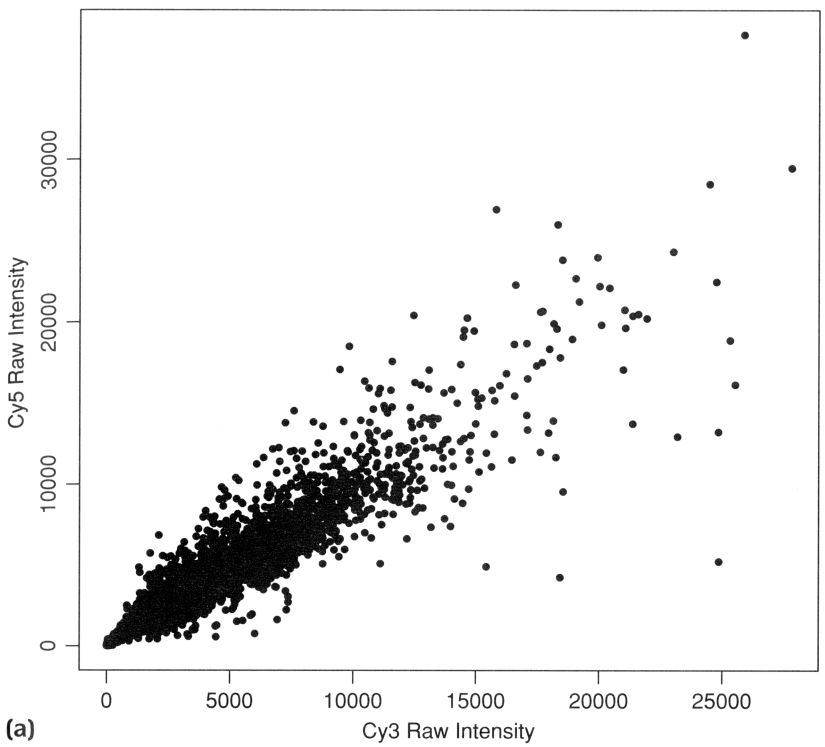
(Go back to original image if needed)

2— Intensity filtering:

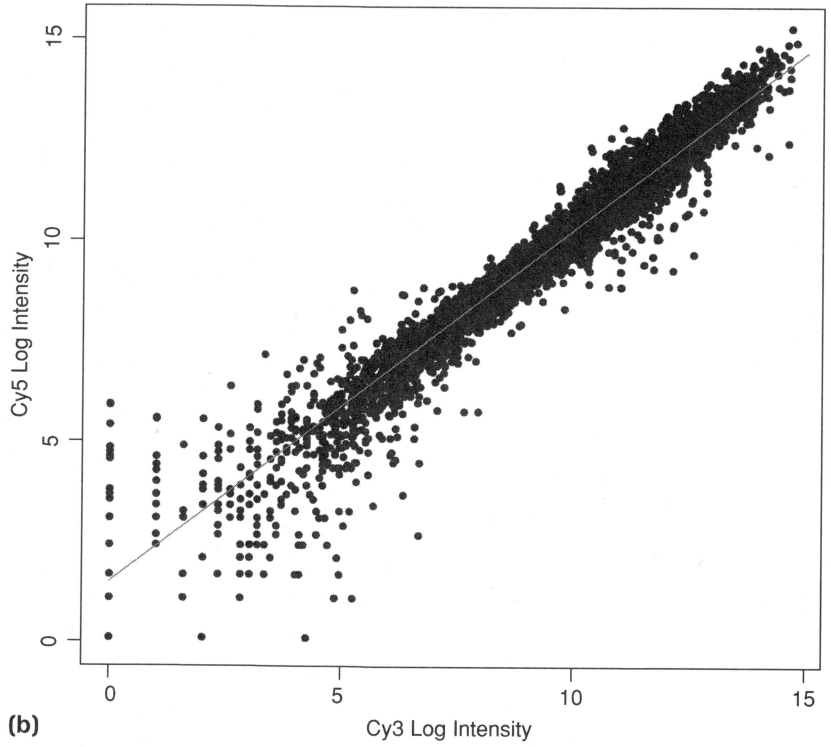
- Scanner saturated spots
- Spots where the difference between signal and background is too low

3—Then it is essential to apply some mathematical transformation on the raw data to help the analysis

Logarithmic transformation



There are more values toward the lowest intensities



Intensities are distributed in a uniform way

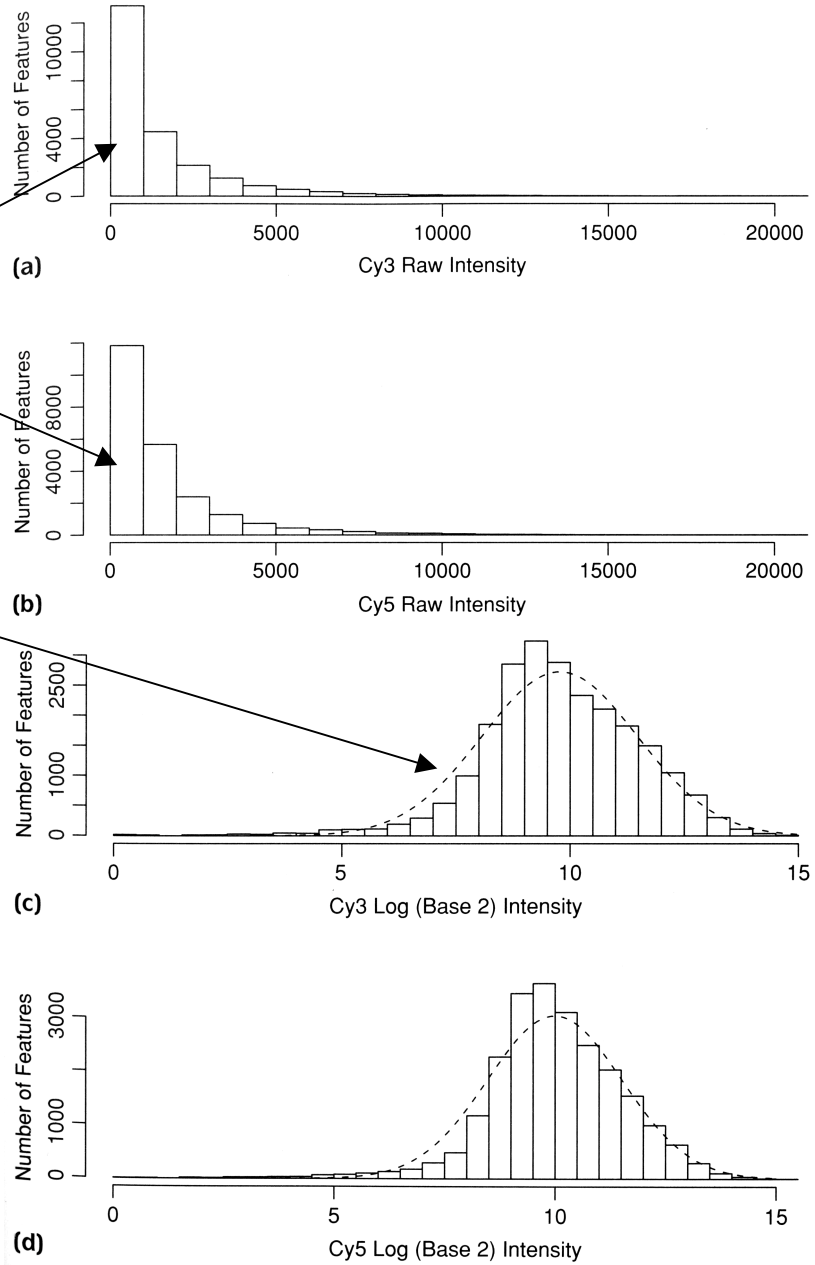
<u>Comments</u>: for microarray analysis, use the base 2 logarithm

Logarithmic transformation

- ***Effect on intensities distribution:***

Most of the measured intensities are weak

"Bell shape" distribution

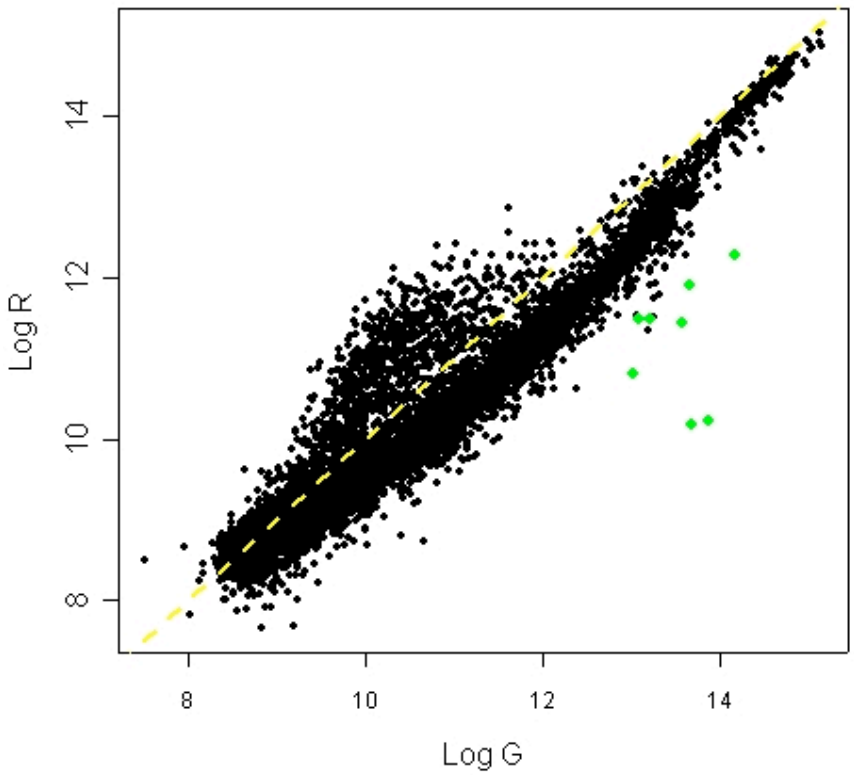


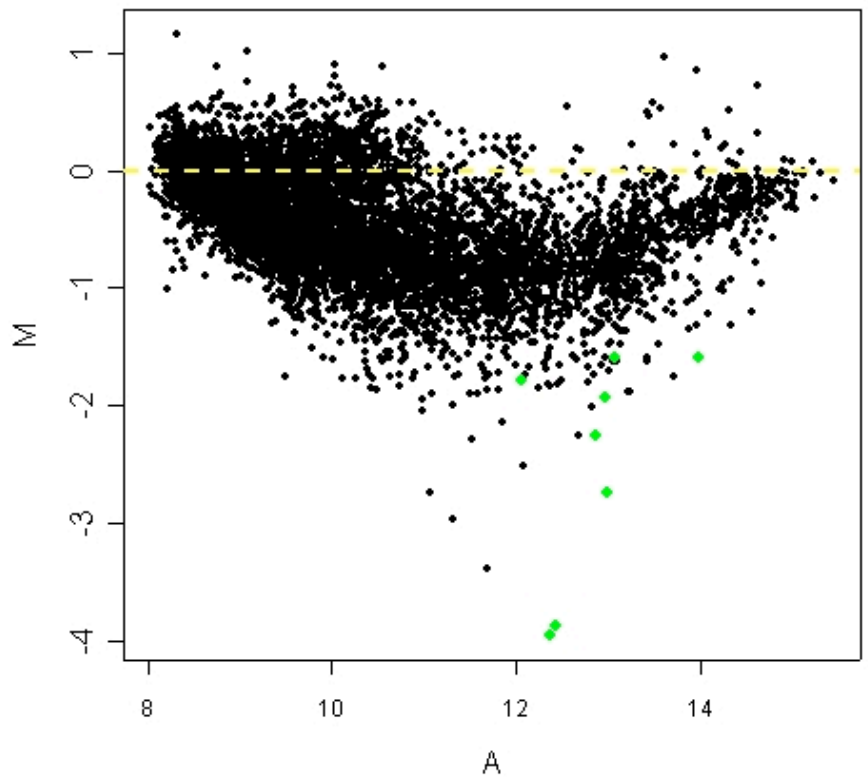
- Distribution centring:

It helps using mathematic and statistic transformation

From "Microarray Bioinformatics", Dov Stekel

Plot rotation





$\log_2 R$ vs $\log_2 G$

$M = \log_2 R/G$ vs $A = \log_2 \sqrt{RG}$

MA plot

- ***Difference between intensities:***

$$M = \log ratio = \log f/g = \log f - \log g$$

- ***VS intensity geometric mean***:

$$A = \log geometric\ mean = \log fg = [\log f + \log g]/2.$$

- ***In general***, f = Cy5 intensity for one spot on a cDNA glass slide and g = Cy3 intensity for the same spot on the same slide.

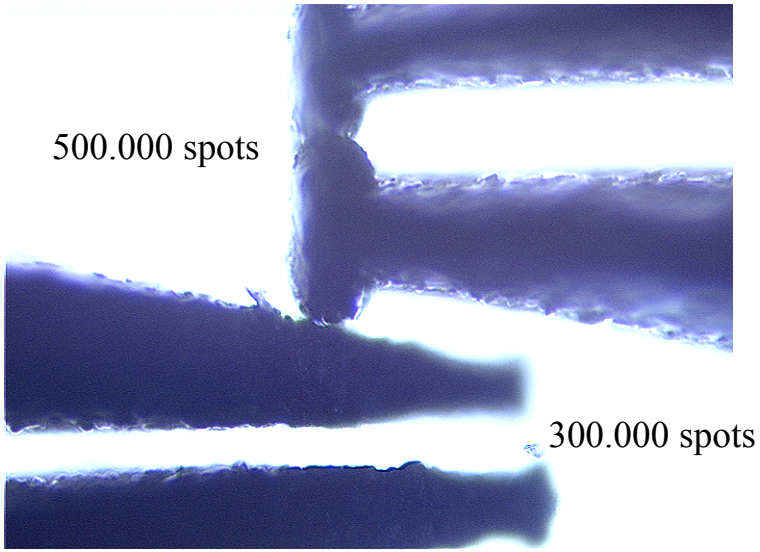
2nd step: normalisation

- ***What for?***

  Normalisation is used to correct systematic differences between samples on the same slide (within-array normalisation), which do not represent real biologic variations between samples.

- ***Why normalisation is necessary?***

  Replicates within an experiment or between experiments are supposed to be identical: there is no differential expression expected. Therefore, normalisation is necessary to discard the phenomenon that appears to be experimental bias.



Spotting pins end

We found bias depending on global spot intensity, spot localisation on the slide, dyes, plates used during the spotting process, spotting pins, microarray scanner, scanner parameters…

- Several methods are available to normalise data
- Normalisation calibrates systematic errors (and not stochastic ones)

2nd step: normalisation

The goal is to adjust the raw data to remove as much as possible systematic effects, still keeping in mind that:

- discrepancy between systematic effects and the others is far to be clear;
- it is usually difficult to prove that normalisation improve the results without attenuating the “true” signal;
- finding the perfect adjustment is almost impossible.

Spot use in normalisation

- ***The positive control spots***

  - Are the spots containing housekeeping genes or genomic DNA: their expression is supposed to be well known

  - To be used in normalisation, they must be detectable, have a stable expression and their intensity has to be into the range detection of the scanner

Advantage: only a few number of genes is needed
Drawback: these genes undergo a lot of uncontrolled modifications in biological systems

- ***The global spot intensity measurement***

  - Is the spot intensity measured on the whole slide : It is supposed to show the gene majority have an invariant profile

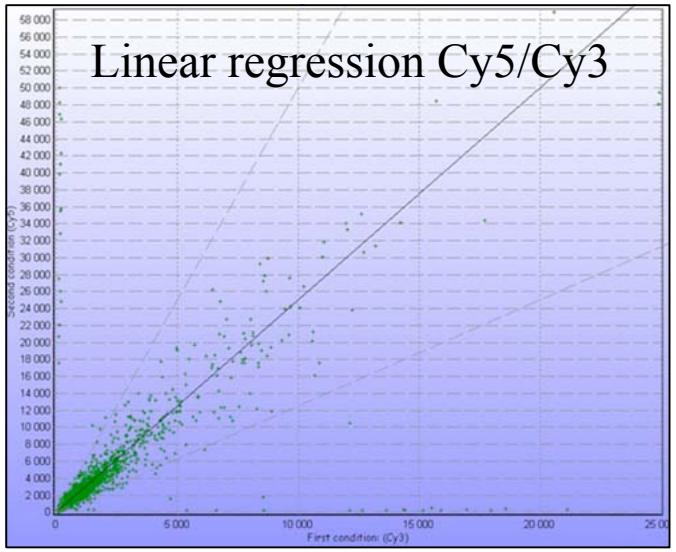
  - The global intensity measurement can not be done on a small set of spots and the spot intensities have to be homogenous

Advantage: efficient measurement on a large number of spots
Drawback: it is imperative that the majority of the analysed genes do not have a varying expression

Normalisation methods

1) A normalisation method based on global adjustment:

$$\log_2 R/G \;\square\; \log_2 R/G - c = \log_2 R/(kG)$$

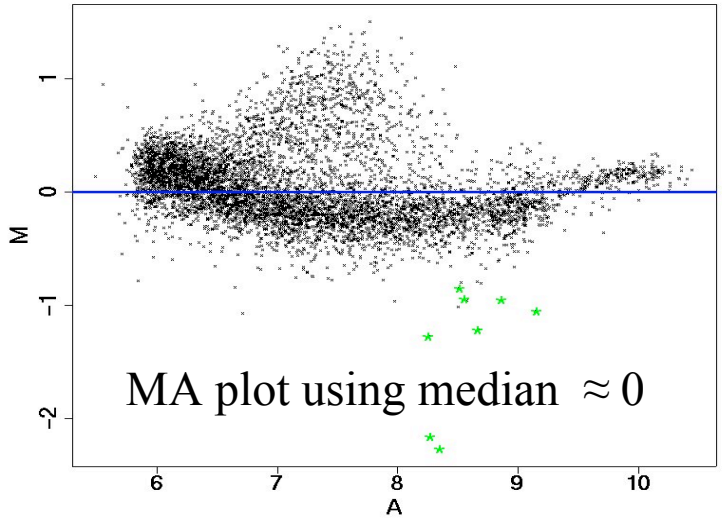


• The choice for k or $c = \log_2 k$ are of several types:

c = log ratio median or mean for a specific gene, or for a set of genes (such as housekeeping genes)

c = normalisation using global intensities where:

$$k = \sum R_i / \sum G_i.$$



Normalisation methods

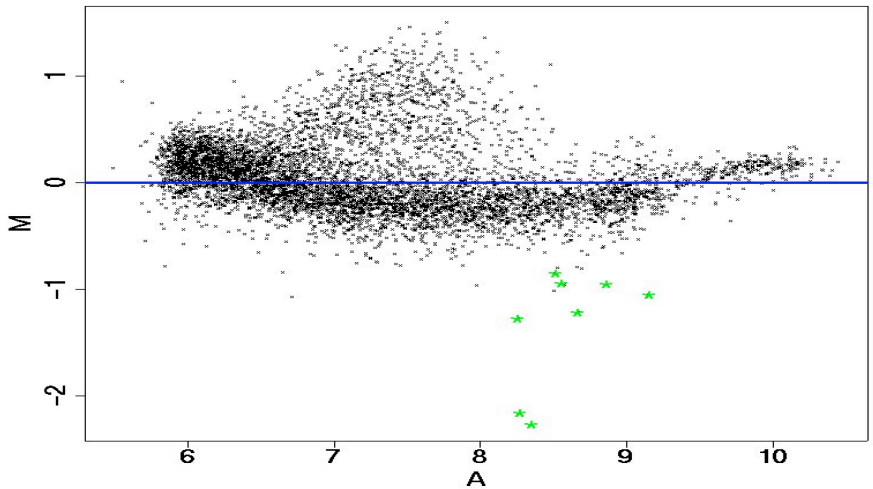
2) A normalisation method that takes into account intensities:

In this case we draw a line going through the centre of MA plot, modifying each M value in each (M,A) point using $c=c(A)$:

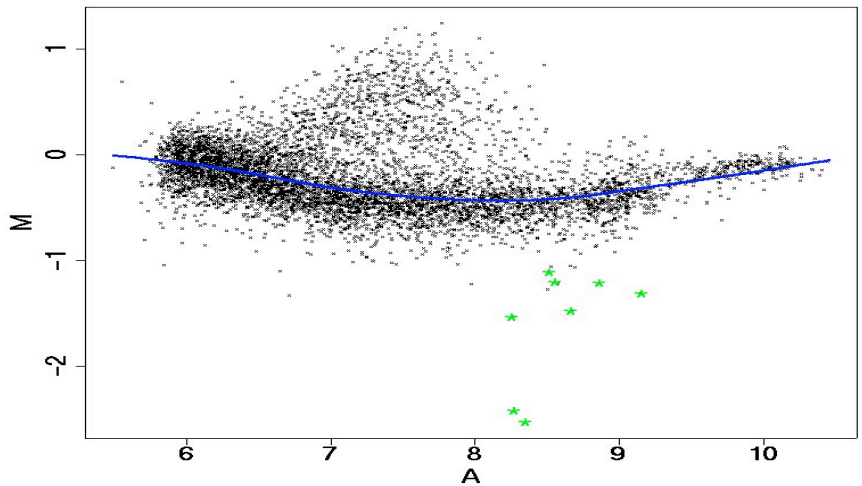
$$\log_2 R/G \;\square\; \log_2 R/G - c(A) = \log_2 R/(k(A)G).$$

An estimation of the $c(A)$ value is done using the Lowess (Loess) regression method from Cleveland (1979):

LOcally WEighted Scatterplot Smoothing

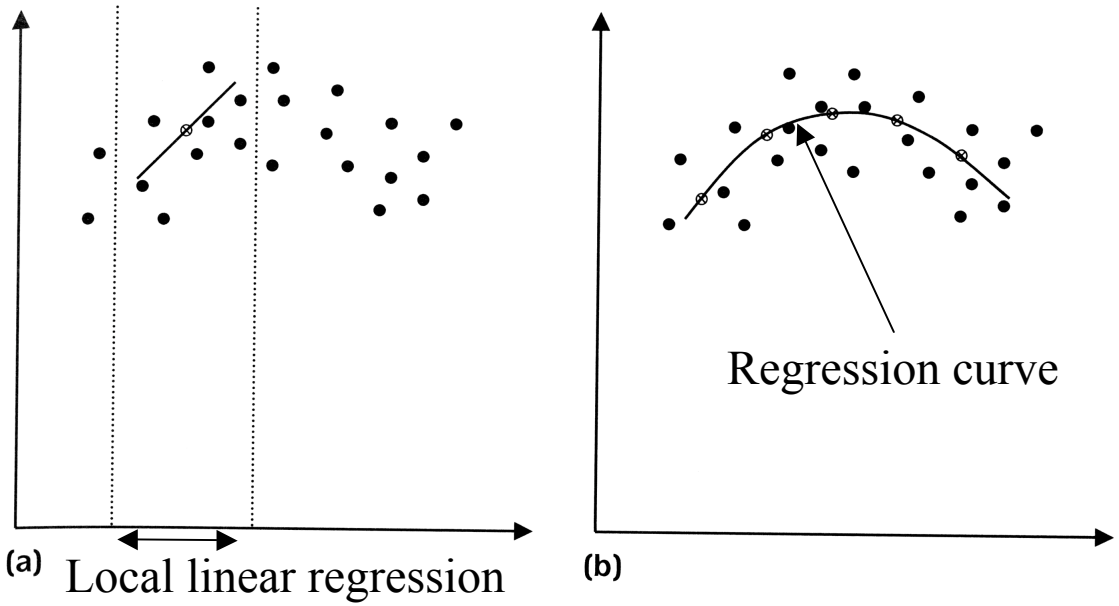


Global intensity normalisation



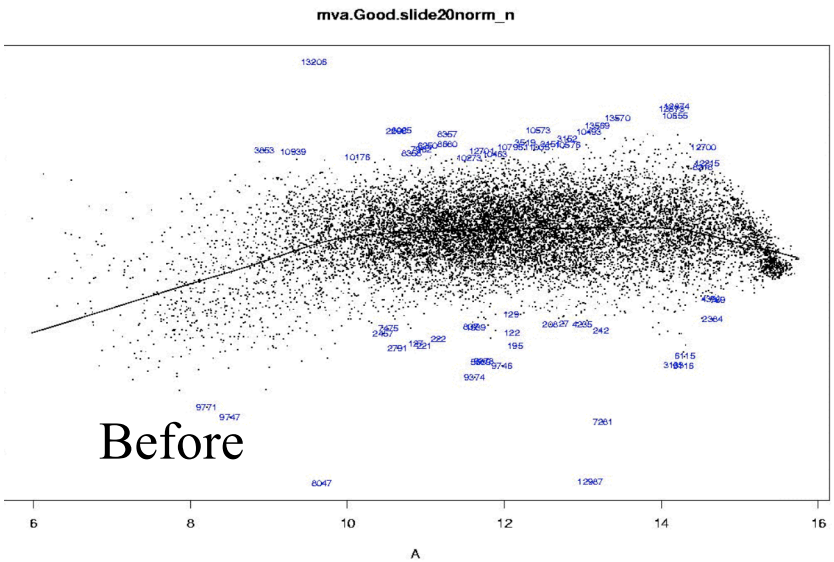
Intensity based normalisation

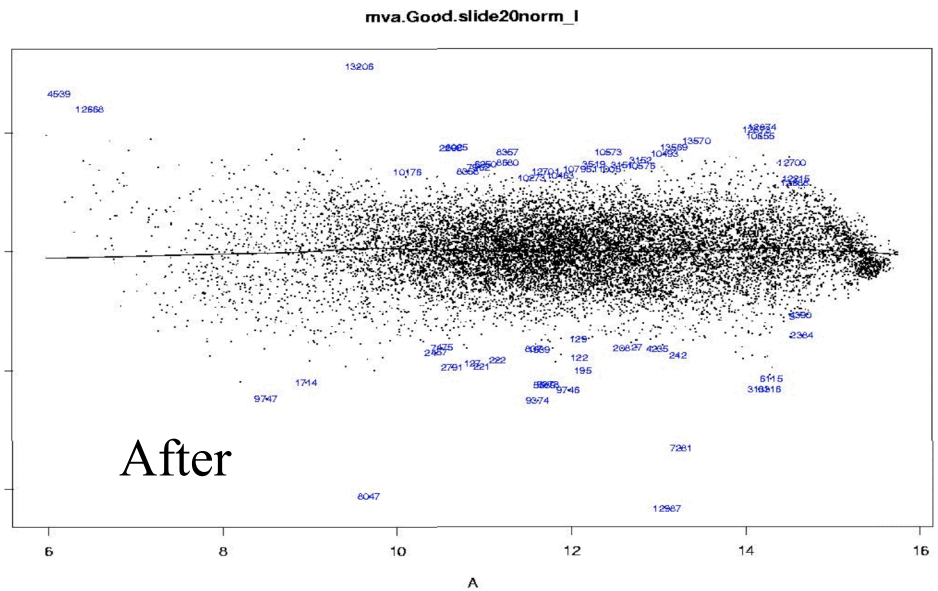
The Loess normalisation



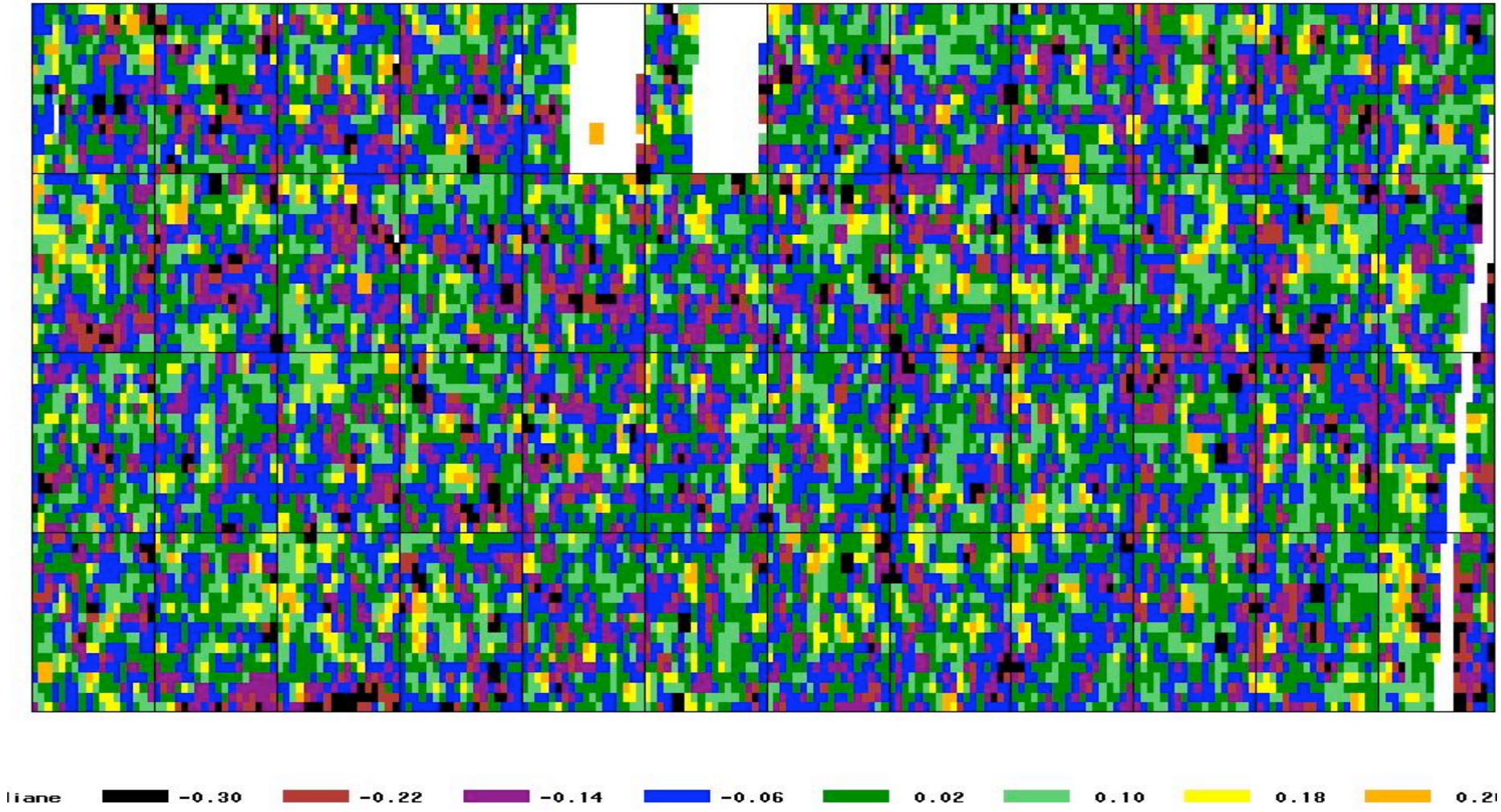
Two parameters to set up:

- The windows size
- The windows overlap





How to fix spatial effects ?



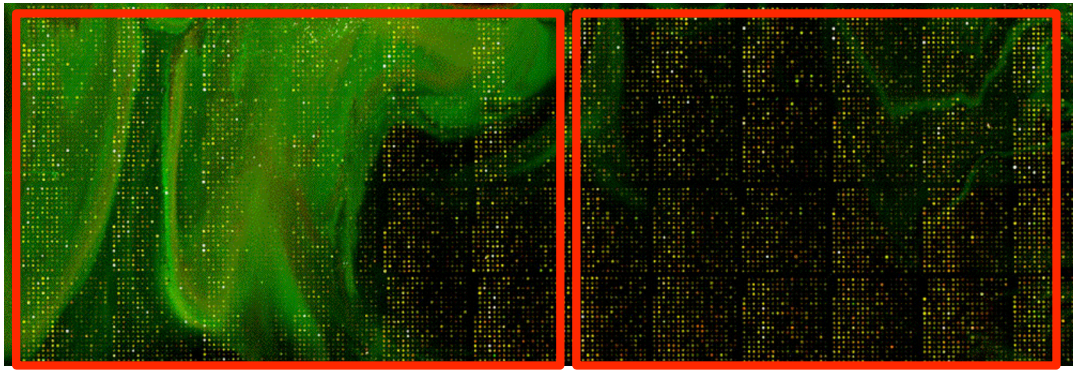
Block or print-tip effect

Normalisation methods

3) A normalisation method that takes into account spotting pins

In addition to intensity dependent variations, spatial bias can also be an important source of systematic error.

Only a few normalisation methods can fix spatial effects like hybridisation artefacts, print-tip differences or collection plate discrepancy during the slide spotting.

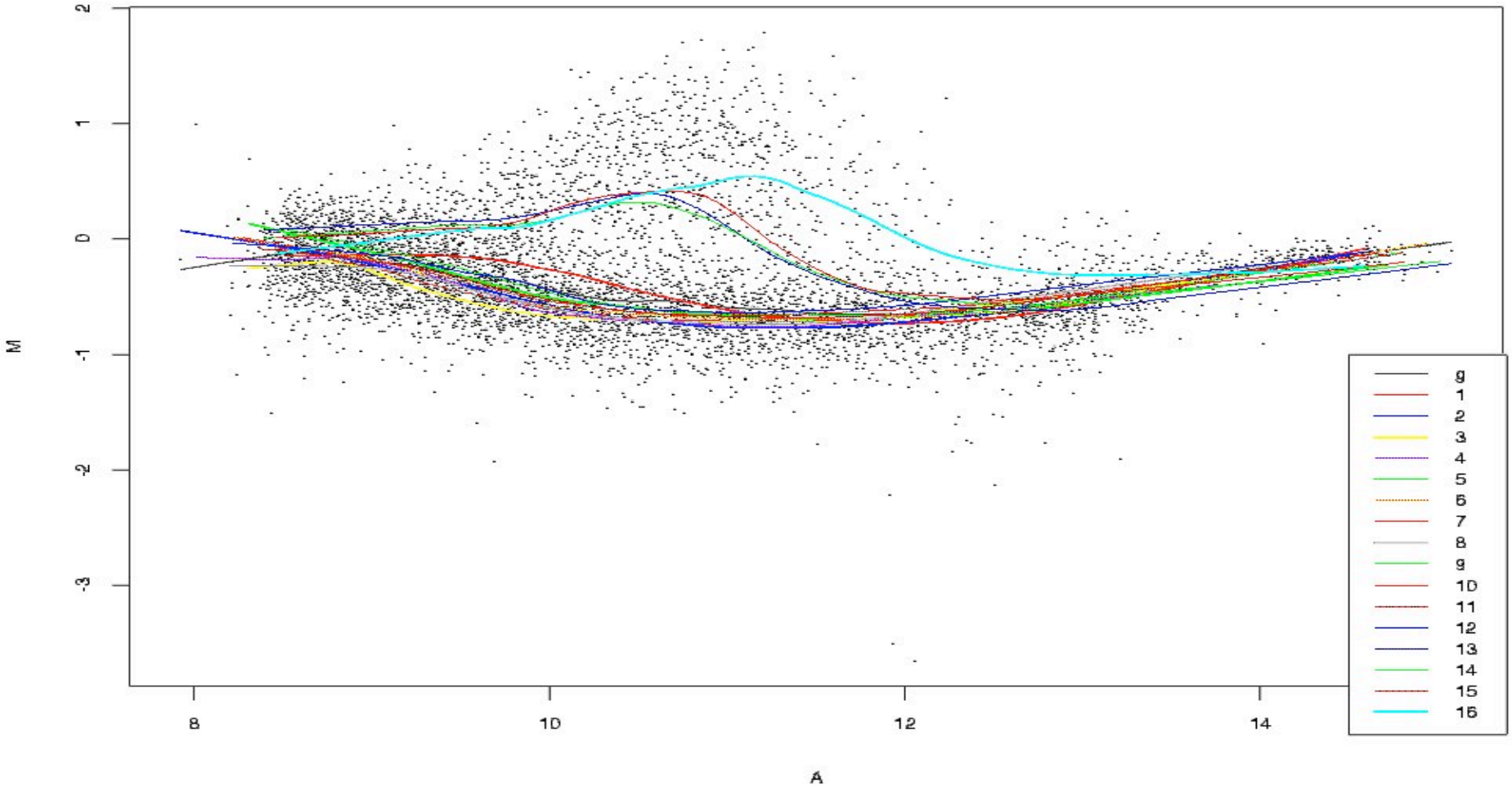


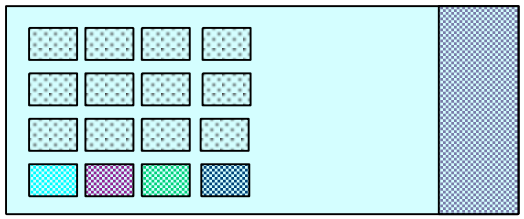
It is possible to fix simultaneously the bias due to intensities and print-tips using a Loess regression on the data in each group of spotting pins:

$$\log_2 R/G \rightarrow \log_2 R/G - c_i(A) = \log_2 R/(k_i(A)G),$$

where $c_i(A)$ is the Loess regression coefficient on the MA plot for the i grid.

Print-tip Loess



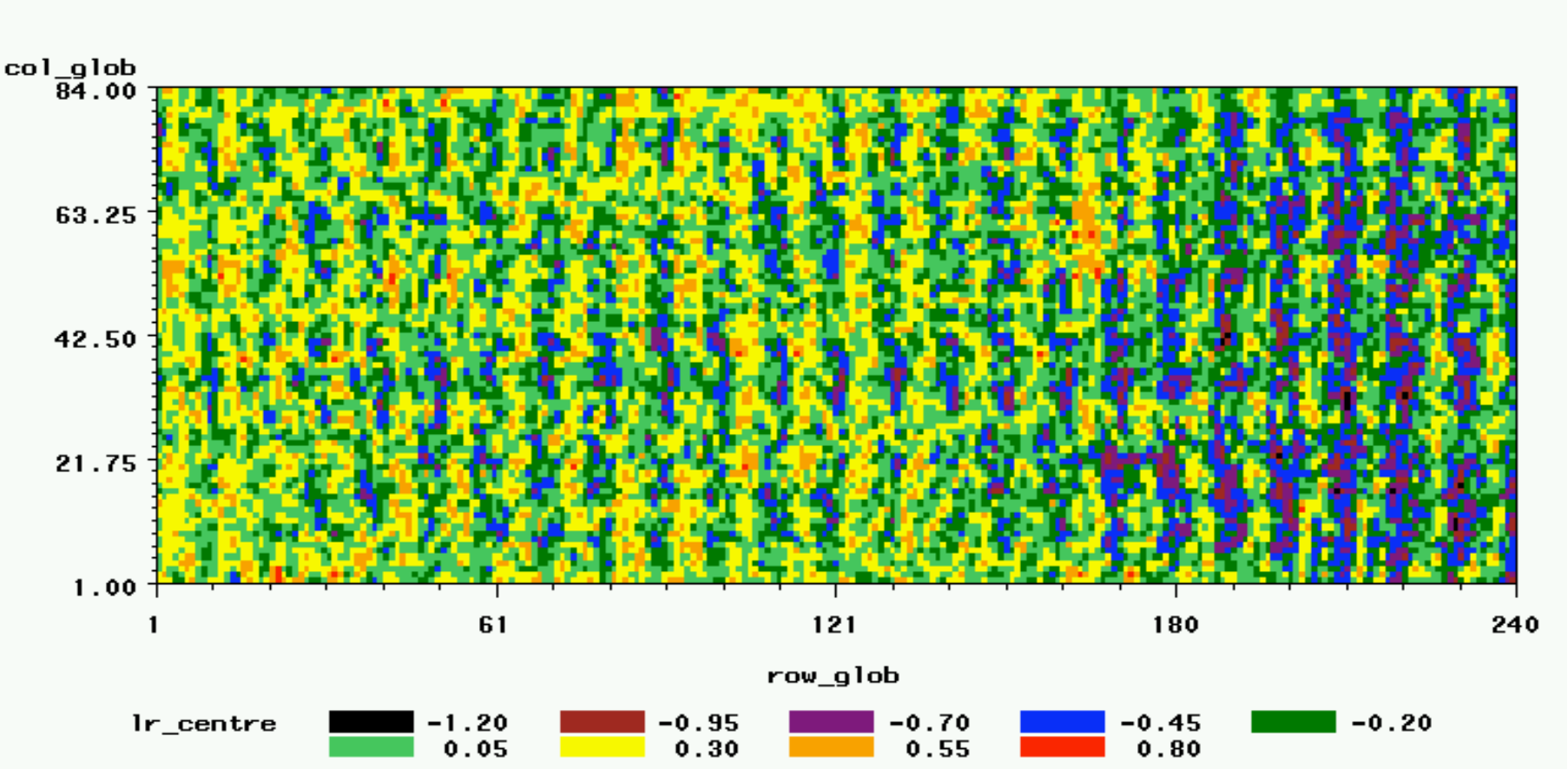


Microarray

cDNA glass slide

4x4 blocs = 16 print-tip groups

Other type of spatial effects



The gene collection plate effects :

- Differences in the plate preparation
- Differences in the gene localisation within the plate

From Mary-Huard et al (2004) *BMC Bioinformatics* **5**: 63

Normalisation methods

Advantages	Drawbacks
----- Linear regression Cy5 vs Cy3 -----	
• Very simple method • Good data overview	• Cy5/Cy3 asymmetry • Low efficiency on deformed data distribution
----- Linear regression "MA plot" -----	
• Very simple method • Horizontally aligned	• Low efficiency on deformed data distribution
----- Loess -----	
• Works well on deformed distribution • Cy5 and Cy3 intensities correction	• Needs the use of a statistic software • Pay attention to the regression parameters

Conclusions on normalisation

- ***There are some hypotheses to keep in mind:***

  - Using the global Loess method assumes at the level of mRNA abundance that:
    - only a minority of genes is differentially expressed
    - there is an equal number of induced or repressed differentially expressed genes

  - For the print-tip specific method, it is necessary that all previous conditions are respected for each bloc. From a statistical point of view, the number of spots concerned by the method cannot be too small.

  - Using a subset of specific genes for normalisation (control, housekeeping genes) imply similar hypotheses.

- ***There are improvements to bring:***

  - The use of an adaptive normalisation method allowing the selection of a data adapted method.
  - It is important not to crush variations and not to create false positive genes.

Conclusions on normalisation

- ***We recommend:***

  - To use log2 ratios (MA plot)
  - To use the Loess normalisation to fix dye bias.
  - To use print-tip normalisation (Loess or median) to take into account spatial bias
  - To be aware of the controversial effects such as the background subtraction
  - To keep in mind the bias correction changes the raw data: it is necessary to adapt the normalisation method to the data and to be aware of these problems
  - To guaranty the same technical conditions for all the slides you use in one experiment (same bench scientist, same scanner, same slide batch…)
  - Not to hesitate on the time you spend with quality controls

- ***In principle:***

  - The bad spots should be eliminated using replicates

=> The most difficult thing is to remove technical bias without changing anything to the signal you study

DNA microarray bioinformatic analysis

Between-array normalisation

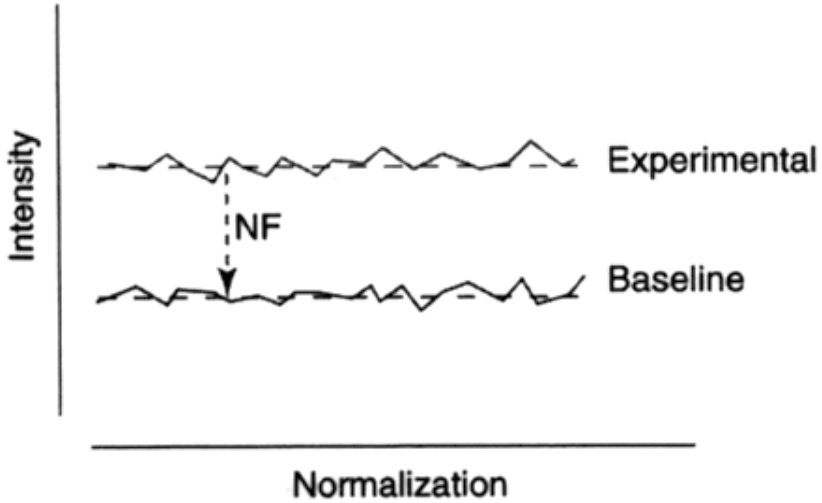
Affymetrix chips normalisation

- ***Normalisation***

Comparison between two experiments:

- Each condition is hybridised on one array.

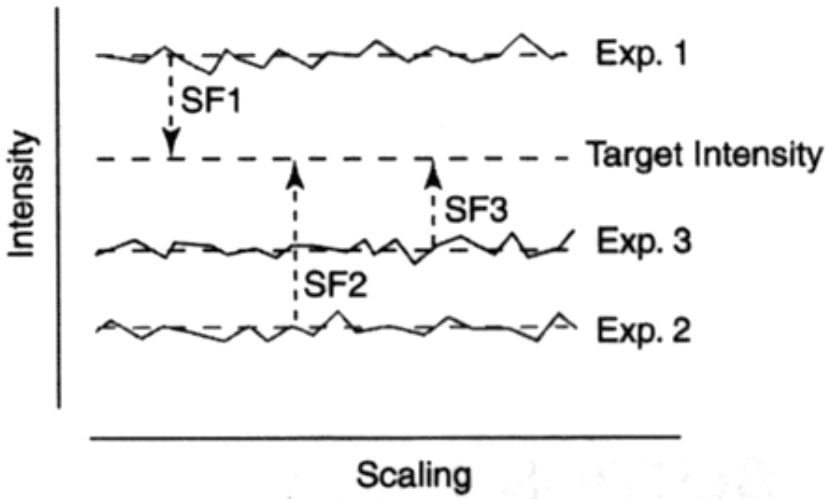
- Using global intensities allow all value normalisation and then comparison between experiments.



- ***Standardisation***

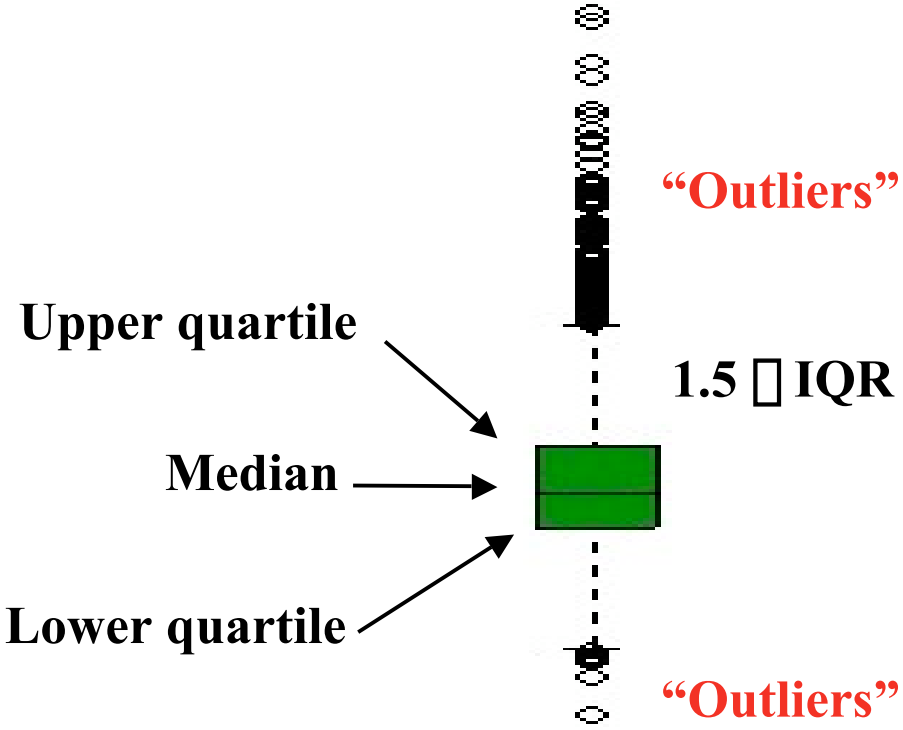
Comparison between several experiments:

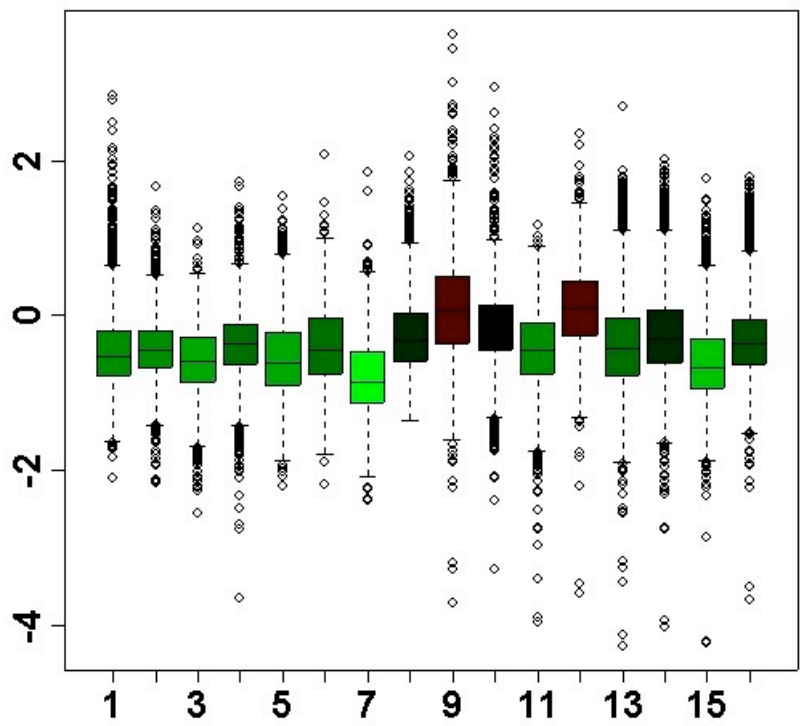
- Each experiments must be compared to a control condition.
- It is necessary to use a target intensity, arbitrarily fixed or obtained in an absolute way (housekeeping genes).



Box plots

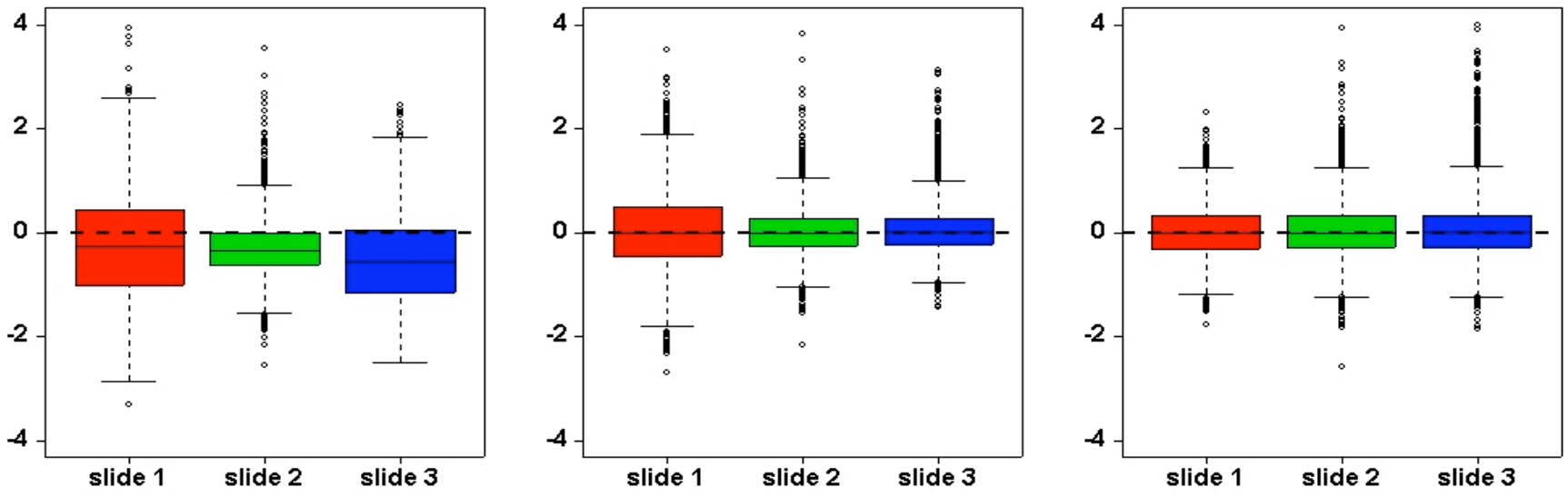
- Each experiment distribution is illustrated as a box
- We can directly visualise the global shape of one distribution (median, standard deviation) and then quickly and easily compare $\log_2$(ratios)





Between-array normalisation

<u>Hypothesis:</u> The variations of the distribution observed are not real biological changes



Box plot distribution of log2(ratios) for 3 identical hybridisations (replicates):

- Left: without any normalisation
- Centre: after a print-tip Loess normalisation (centring)
- Right: after between-array normalisation (scaling)

Some bibliography about normalisation

- Quackenbush J. Microarray data normalization and transformation. *Nat Genet.* 2002 Dec;**32** Suppl:496-501.

- Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 2002 Feb 15;**30**(4):e15.

- Leung YF, Cavalieri D. Fundamentals of cDNA microarray data analysis. *Trends Genet.* 2003 Nov;**19**(11):649-59.

Transcriptome bioinformatic at the ENS

Stéphane LE CROM - Gaëlle LELANDAIS - Sophie LEMOINE - Laurent JOURDREN

http://transcriptome.ens.fr