Analysis of microarray data

Sources

- From a lesson of Henrik Bengtsson
  - Bioinformatics, Bioinformatics Centre, University of Copenhagen
  - hb@maths.lth.se
- Robin Liechti, UNIL, Lausanne, CH
  - robin.liechti@ie-bpv.unil.ch

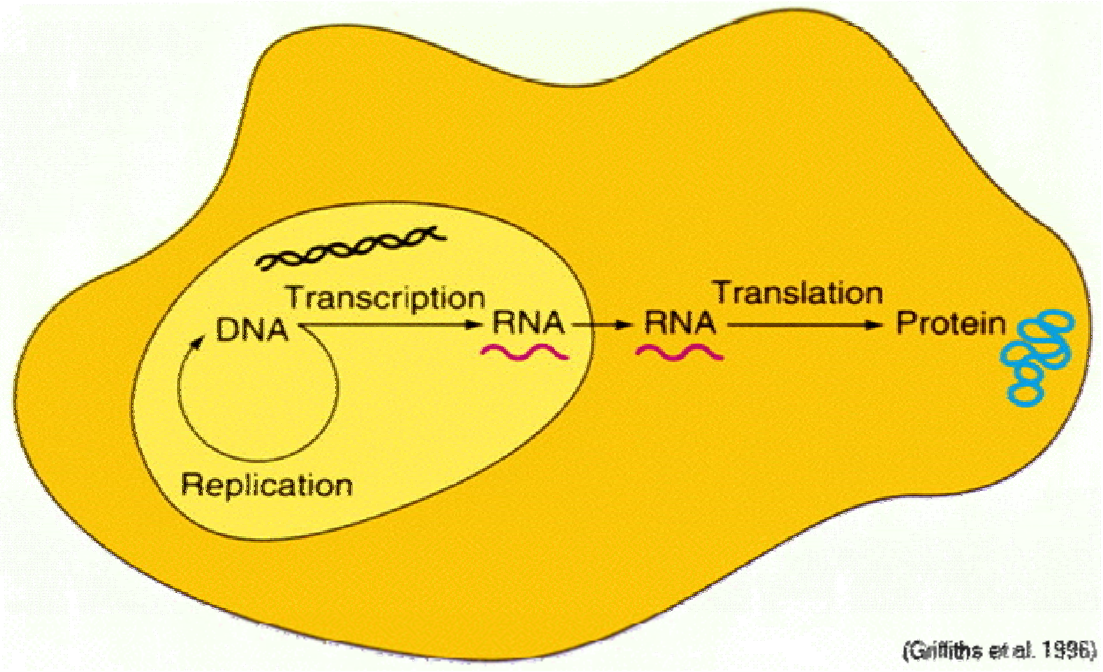
Outline

- Part I – (Very short) Background
  - Central Dogma of Biology
  - Idea behind the microarray technology
- Part II - Printing, Hybridization, Scanning & Image Analysis
  - From clone to slide
  - From samples to hybridization
  - From scanning to raw data
- Part III - Exploratory data analysis
  - The log-ratio log-intensity transform
  - Various graphs
- Part IV - Preprocessing of data
  - Background correction
  - Normalization
- Part V - Identifying differentially expressed genes
  - Cut-off by log-ratios values, the t-statistics and cut-off by T values
  - Multiple testing, adjusting the p-values
- Part VI – Normalization again...
  - Transformation of data
  - Linear and affine models
  - Affine normalization
  - Common normalization methods

The cDNA microarray technology-
PART I:
(Very short) Background

1. The Central Dogma of Biology
2. Idea behind the microarray technology

The Central Dogma of Biology



Idea of gene-expression techniques:

Measure the amount of mRNA to find genes that are expressed

The cDNA Microarray Technique

1. Put a large number of DNA sequences or synthetic DNA oligomers onto a glass slide
    1. - 5000-50000 gene expressions at the same time.
2. Measure amounts of cDNA (from mRNA) bound to each spot
3. Identify genes that behave differently in different cell populations

The cDNA microarray technology

-

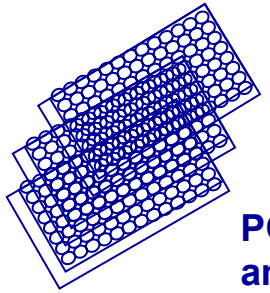
PART II:

Printing, Hybridization, Scanning & Image Analysis

1. From clone to slide
2. From samples to hybridization
3. From scans to raw data

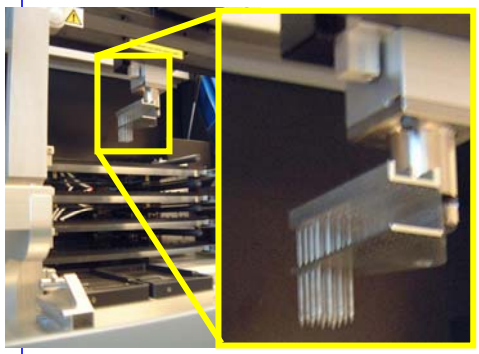
Overview

DNA probes



PCR product
amplification
purification

printing



Test sample

Reference sample

RNA

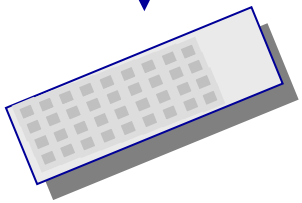
RNA

cDNA

cDNA

Hybridize

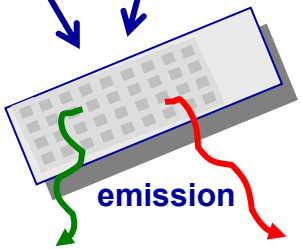
Production



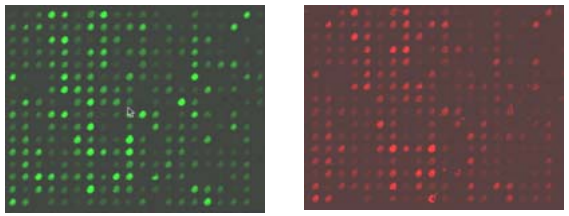
excitation

green laser

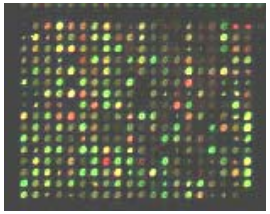
red laser



emission



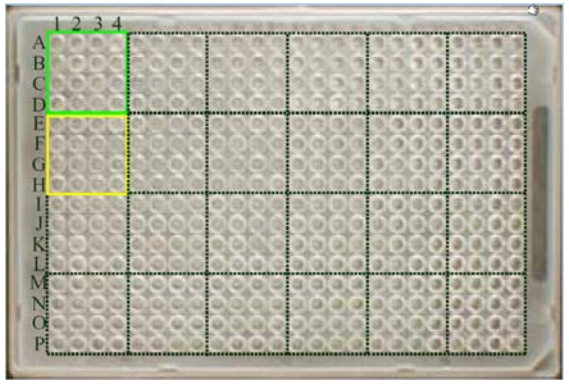
overlay images

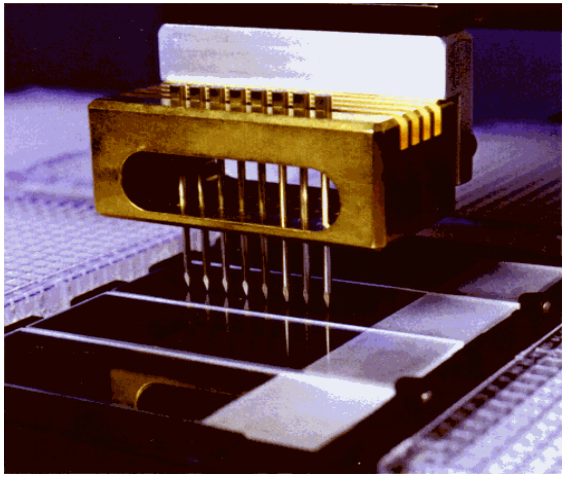


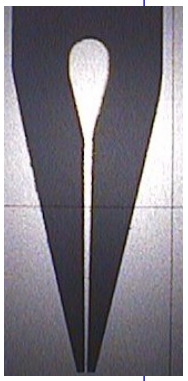
data: (R_{fg},G_{fg},R_{bg},G_{bg}, ...)

scanning

Printing / spotting

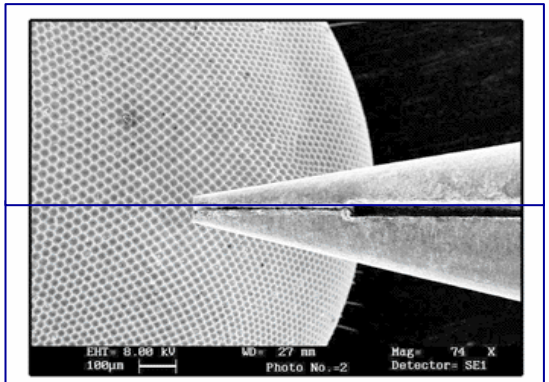




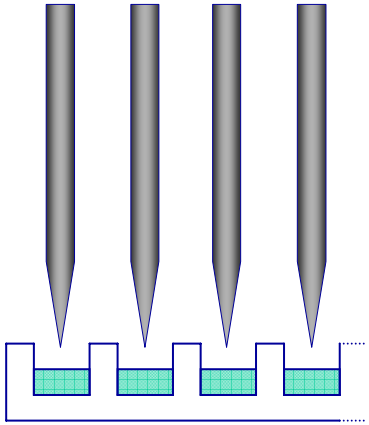


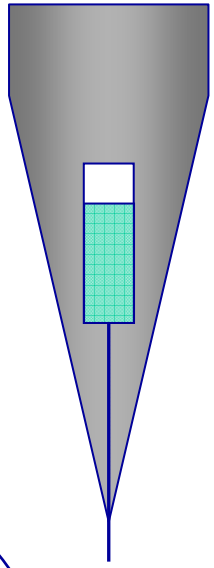
Arrayer (approx 100,000 EUR)

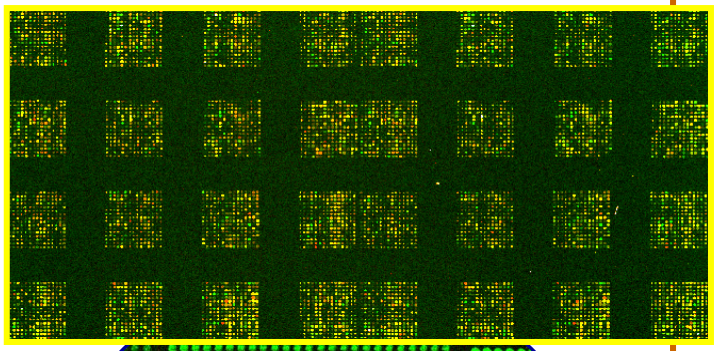
Microarray slide preparation

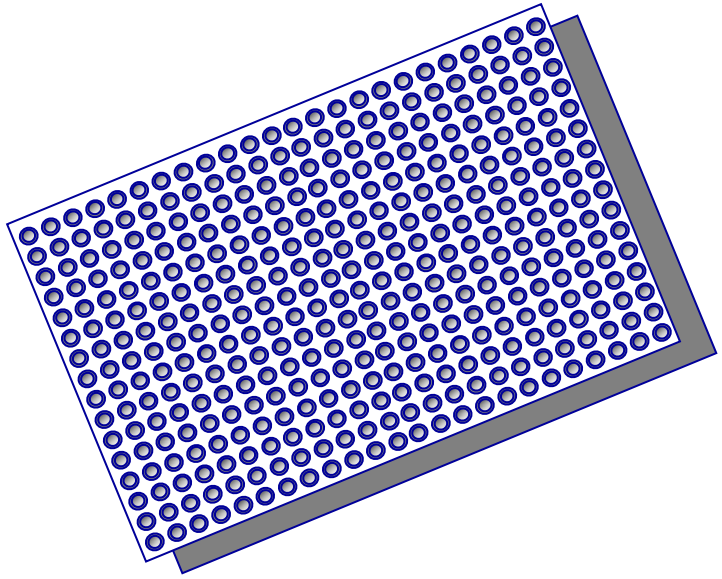


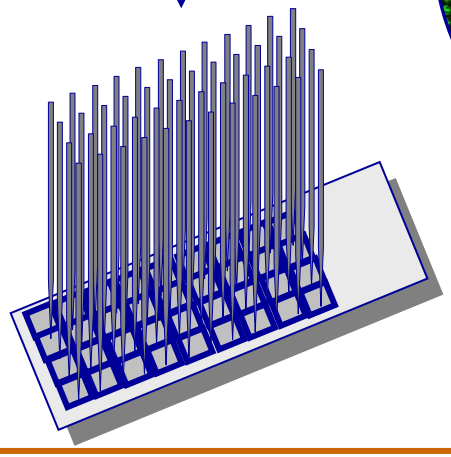
Electron micrograph of MicroSpot 2500 pin and the eye of a house fly

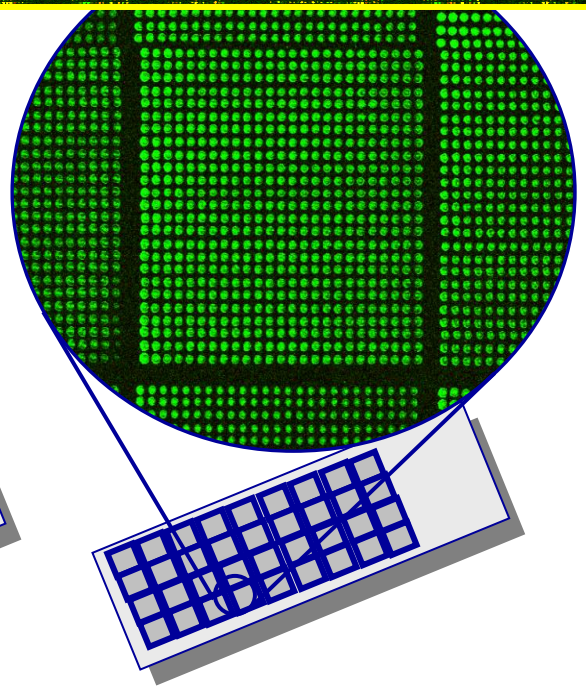








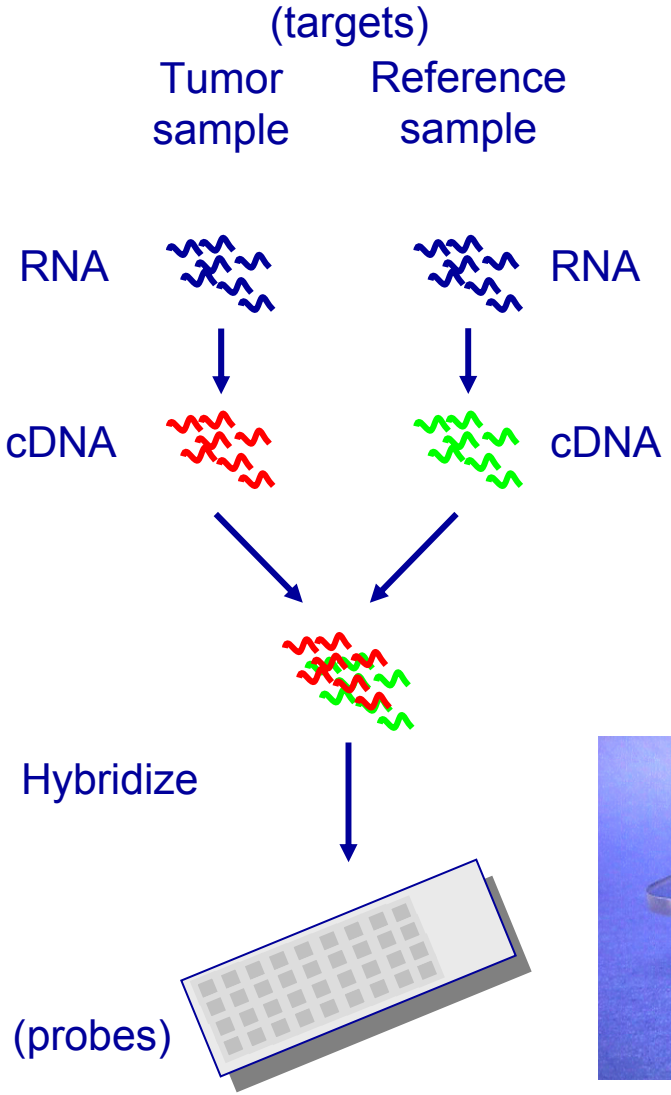




Terminology: probe and target

- As defined in Nature 1999:
  - The probes are the immobilized DNA sequences spotted on the array, i.e. spot, oligo, immobile substrate
  - The targets are the labeled cDNA sequences to be hybridized to the array, i.e. mobile substrate
  - The opposite usage can also be seen in some references. However, think of probes as the measuring device (which you can buy), and the targets (that you provide) as what you want to measure

RNA extraction & hybridization



1. **Extract mRNA from samples.**
2. **Reverse transcription of mRNA to cDNA.**
3. **Label with Cy3 and Cy5 fluorescent dyes.**
4. **Hybridize labeled cDNA cocktail to array.**
5. **Wash array.**

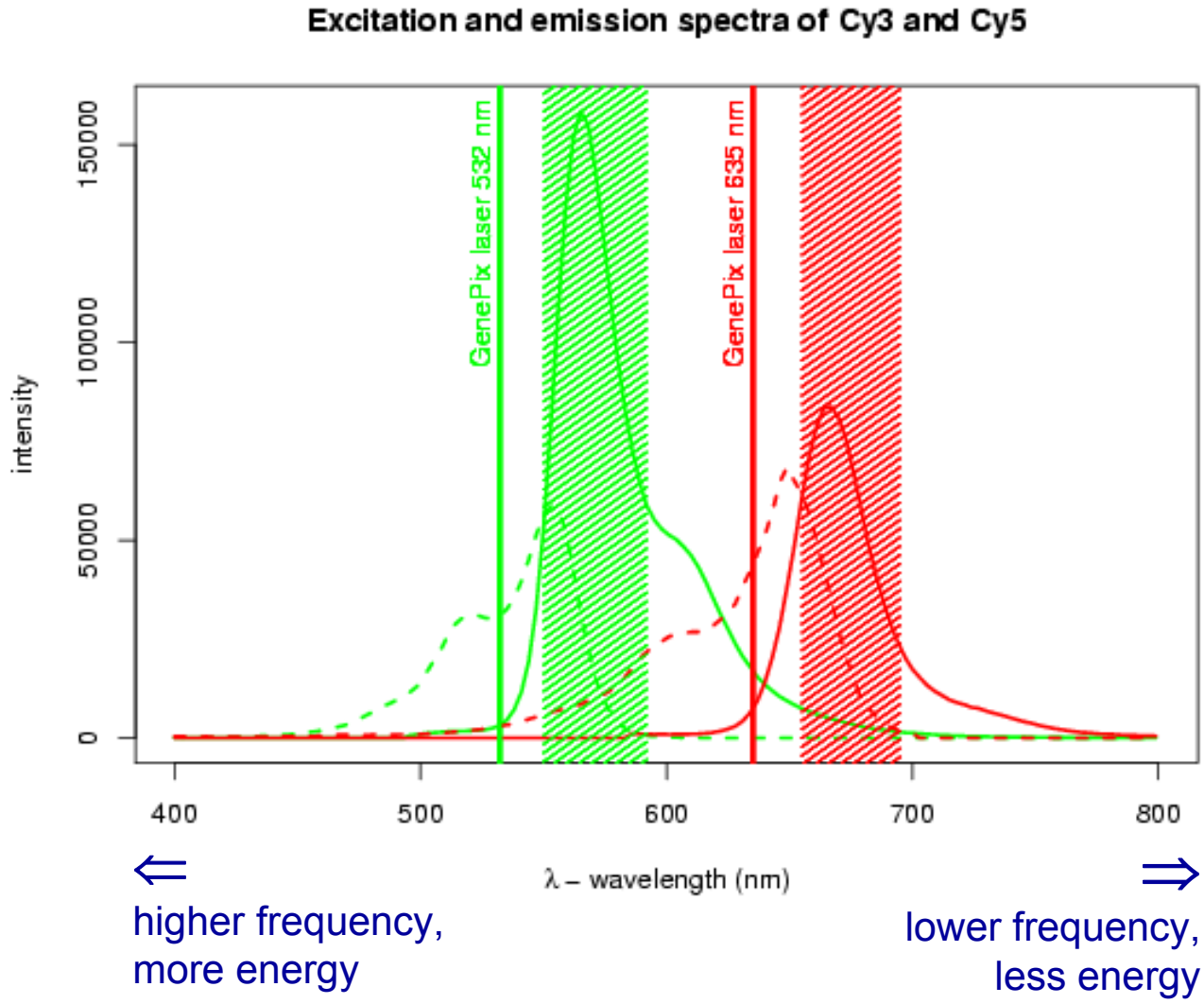
Figure: Hybridization chamber.

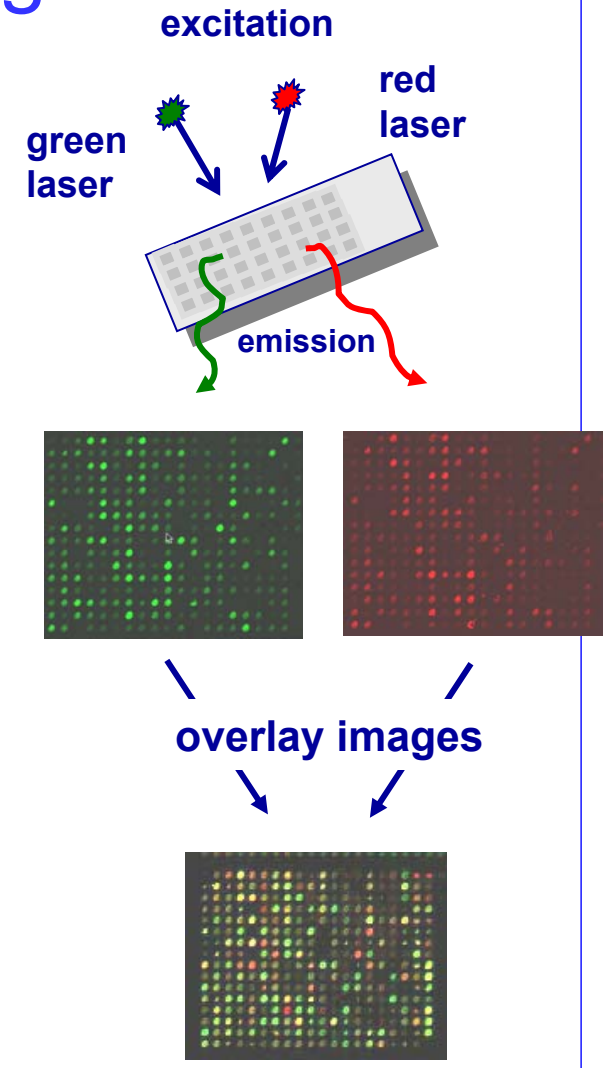
References

- Original cDNA microarray paper:
  - Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270(5235):467–470, October 1995.

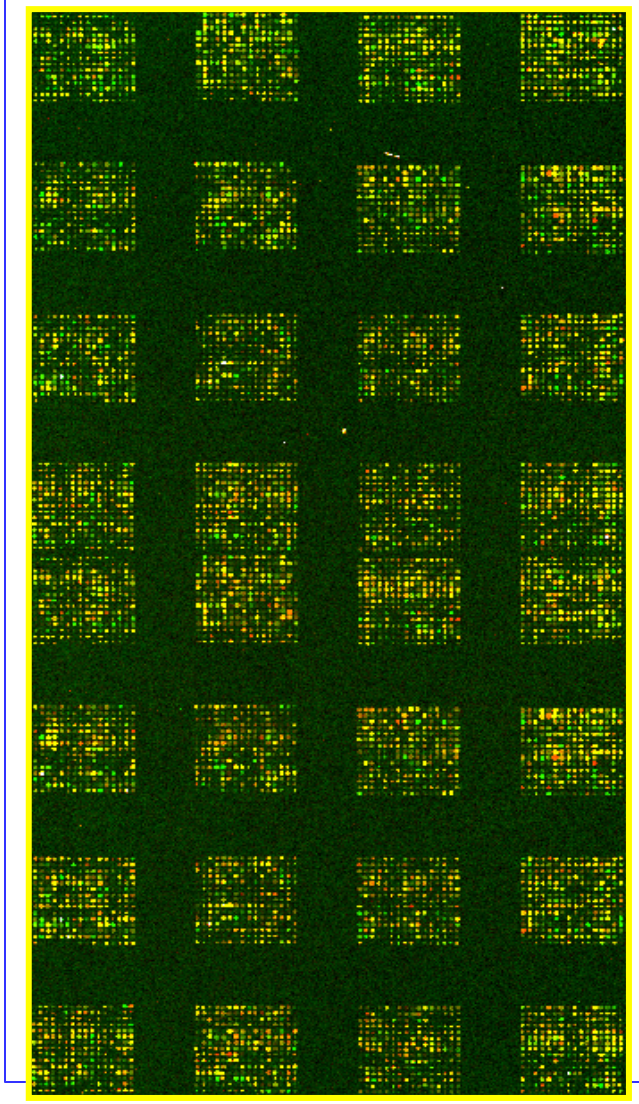
- General:
  - Mark Schena. Microarrays Analysis. John Wiley & Sons, Inc., Hoboken, New Jersey, 2003.
  - David J. Duggan, Michael Bittner, Yidong Chen, and Paul Meltzer & Jeffrey M. Trent. Expression profiling using cDNA microarrays. Nature Genetics, 21(1 Supplement):10–14, January 1999.

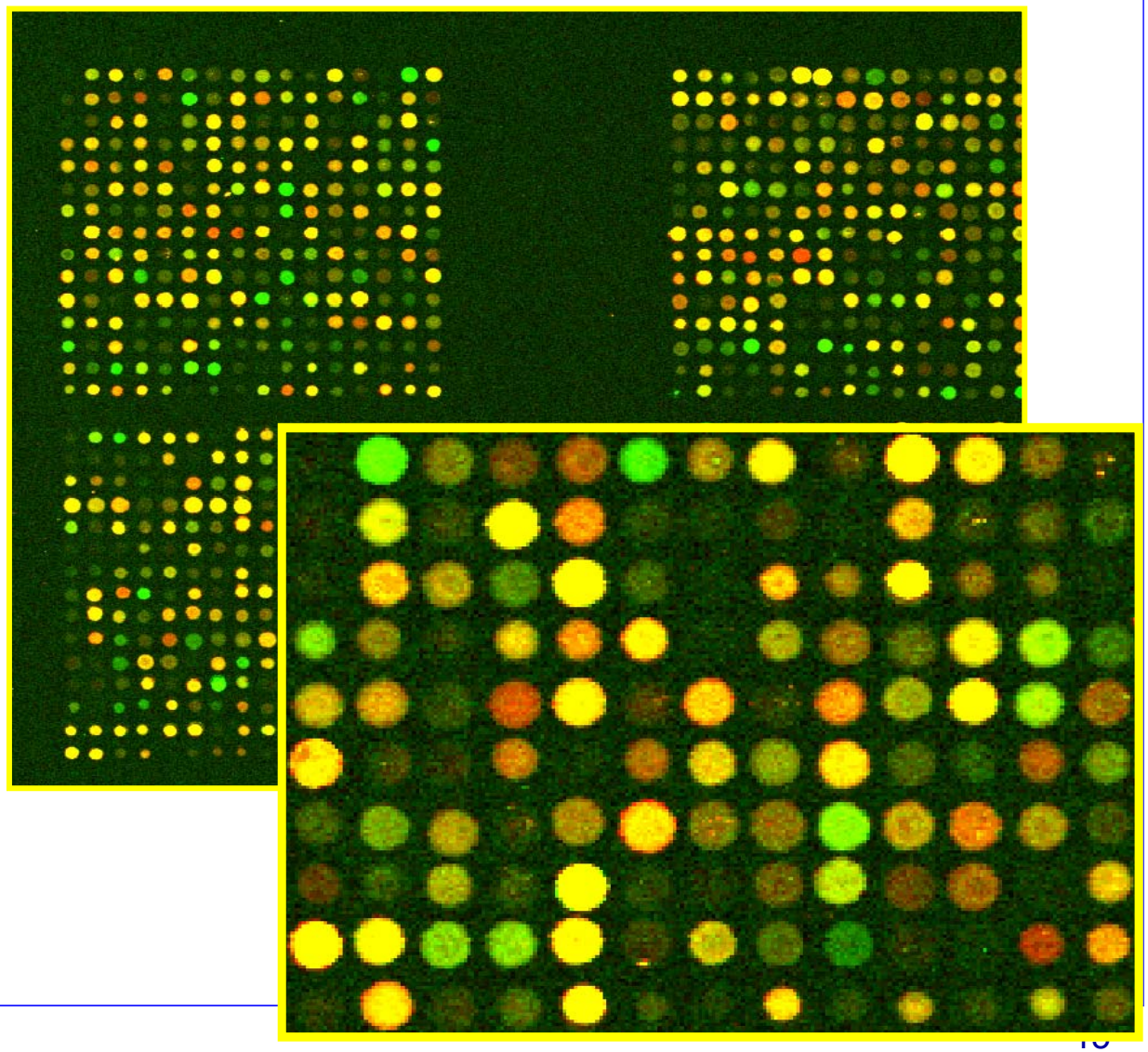
Two-channel scanning





Combined color image for visualization





Some scanners

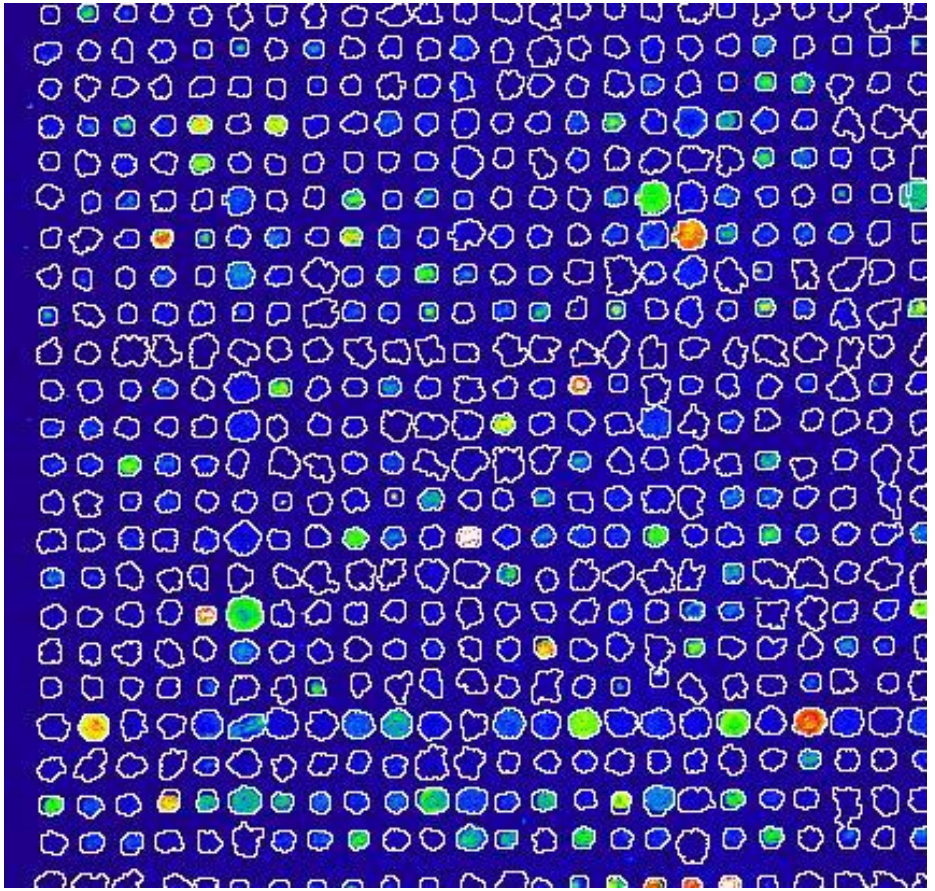
- Axon GenePix
- Agilent
- ScanArray
- ...

Signal quantification

1. Addressing
   - Locate spot centers.

2. Segmentation
   - Classification of pixels either as signal or background (using circles, seeded region growing or other).

3. Signal quantification
   a) foreground estimates
   b) background estimates
   c) ... (shape, size etc)



Robust signal estimates: hints

mean vs. median pixel signal

- Assume data with one outlier:

x = (8, 85, 7, 9, 5, 4, 13, 6, 8)

  – The *mean* of all x's, i.e. $(x_1+x_2+...+x_K)/K$, is affected by the outlier:

mean(x) = 16.11

  – The *median* of all x's, i.e. the middle value of $(x_1+x_2+...+x_K)$, is *not* (if < 50% values are outliers):

median(x) = 8.0

Use the median instead of the mean if you expect artifacts.

(If there are a lot of measurements and the errors are symmetrically distributed the median will give the same result as the mean without outliers.)

Some image analysis applications

Academic (free and non-free)

- Spot (CSIRO, Australia)
- ImaGene (BioDiscovery)
- ScanAlyze (Eisen Lab, US)
- Spotfinder (TIGR, US)
- ...

Commercial

- QuantArray (PerkinElmer Life Sciences)
- GenePix Pro (Axon)
- ...

References

Image analysis:

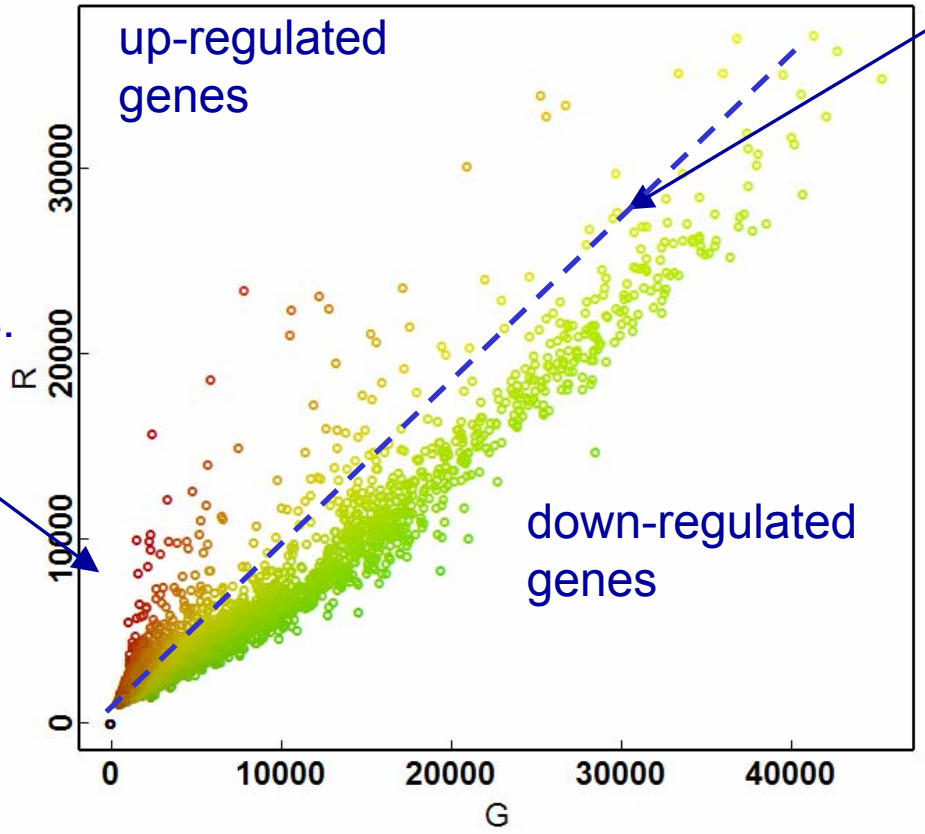
- **Yee Hwa Yang, Michael Buckley, Sandrine Dudoit, and Terry Speed. *Comparison of methods for image analysis on cDNA microarray data*. Technical Report 584, Department of Statistics, University of California at Berkeley, Nov 2000.**

- **Anders Bengtsson. *Microarray image analysis: Background estimation using region and filtering techniques*. Master's Theses in Mathematical Sciences, Mathematical Statistics, Centre for Mathematical Sciences, Lund Institute of Technology, Sweden, December 2003. 2003:E40.**

The cDNA microarray technology
-
PART III:
Exploratory Data Analysis

1. The log-ratio log-intensity transform
2. Various graphs

Scatter plot: R vs G



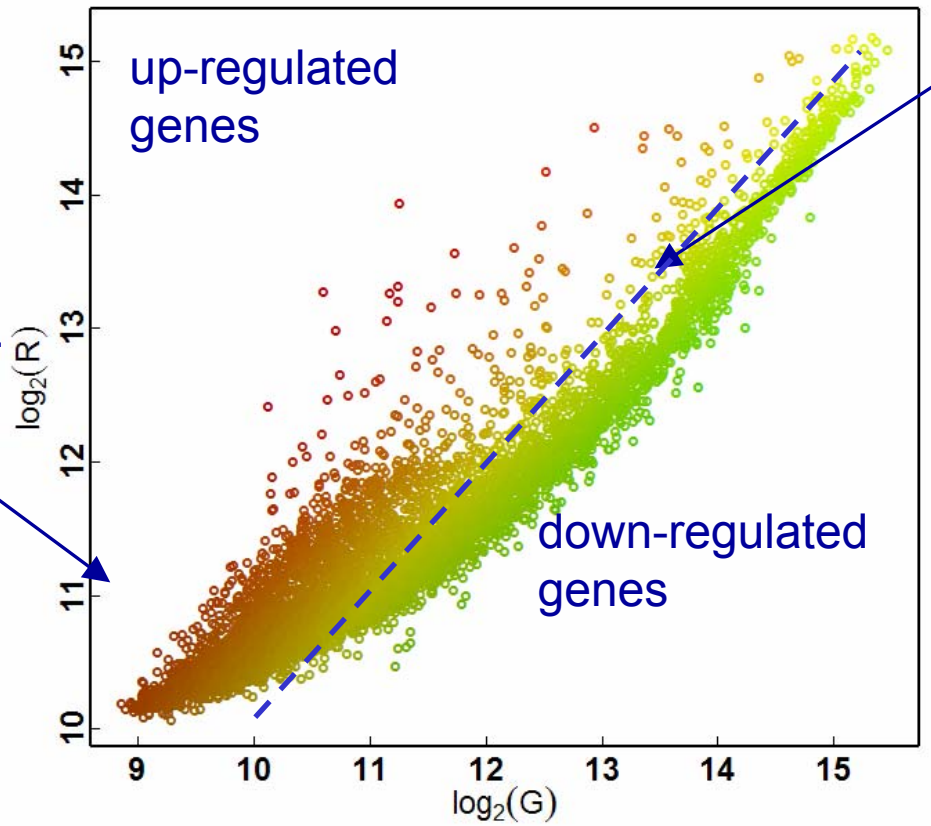
non-differentially expressed genes along the diagonal: R = G

Most genes have low gene expression levels. What happens here?

"Observed" data $\{(R,G)_i\}$:

R = signal in red channel, G = signal in green channel

Scatter plot: $\log_2 R$ *vs* $\log_2 G$



non-differentially expressed genes are still along the diagonal:

$\log_2 R = \log_2 G$

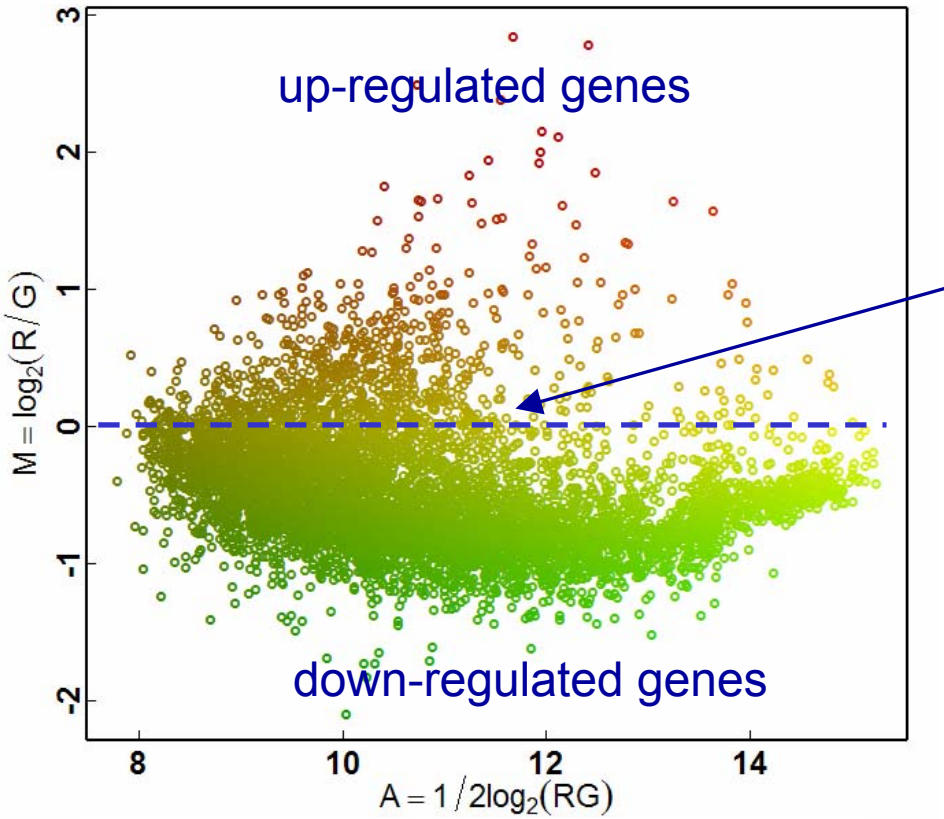
Low gene expression levels are "blown up".

"Observed" data $\{(\log_2 R, \log_2 G)_i\}$:

R = signal in red channel, G = signal in green channel

Scatter plot: M vs A (recommended)

Note: M vs A is basically a rotation of the $\log_2$R vs $\log_2$G scatter plot.



non-differentially expressed genes are now along the horizontal line:

$$M = 0 \Leftrightarrow \log_2 R - \log_2 G = 0 \Leftrightarrow R = G$$

Why: Now the quantity of interest, i.e. the *fold change*, is contained in *one variable*, namely M!

If M > 0, up-regulated.
If M < 0, down-regulated.

Transformed data $\{(M,A)_i\}$:

$M = \log_2(R) - \log_2(G)$ (minus)

$A = \frac{1}{2} \cdot [\log_2(R) + \log_2(G)]$ (add)

Details on M vs A

***Log-ratios*:**

$$M = \log_2(R) - \log_2(G) = \text{[logarithmic rules]} = \mathbf{\log_2(R/G)}$$

Average *log-intensities*:

$$A = \tfrac{1}{2}\cdot[\log_2(R) + \log_2(G)] = \text{[logarithmic rules]} = \mathbf{\tfrac{1}{2}\cdot\log_2(R\cdot G)}$$

There is a one-to-one relationship between (M,A) and (R,G):

$$R=(2^{2A+M})^{1/2},\quad G=(2^{2A-M})^{1/2}$$

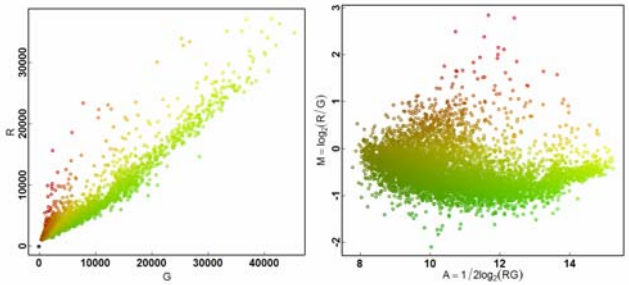
More on why log and why M vs A?

- It makes the distribution symmetric around zero ⇒

- Logs stretch out region we are most interested in and makes the distribution more normal. ⇓

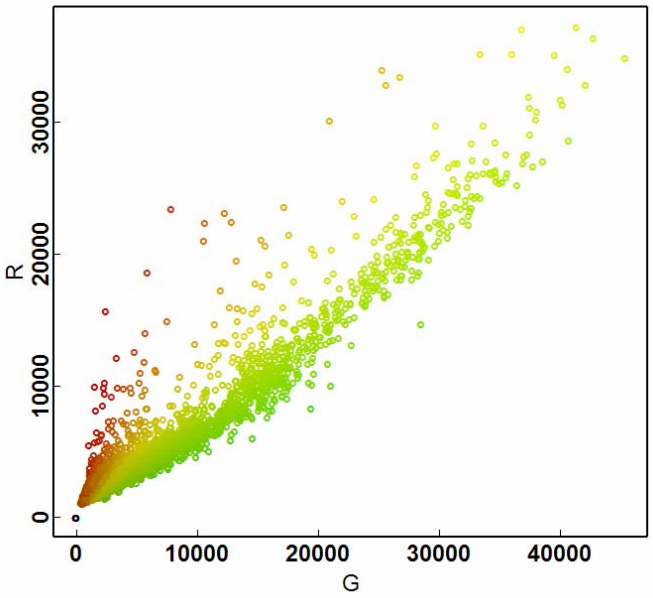
R:G	M=log(R/G)	R:G	M=log(R/G)	comment
1:1	0			**$2^0=1$**
2:1	+1	1:2	-1	**$2^1=2$, $2^{-1}=1/2$**
4:1	+2	1:4	-2	**$2^2=4$, $2^{-2}=1/4$**
8:1	+3	1:8	-3	**$2^3=8$, $2^{-3}=1/8$**
16:1	+4	1:16	-4	**$2^4=16$, $2^{-4}=1/16$**
32:1	+5	1:32	-5	**$2^5=32$, $2^{-5}=1/32$**

- Easier to see artifacts of the data, .e.g. intensity dependent variation and dye-bias. ⇓

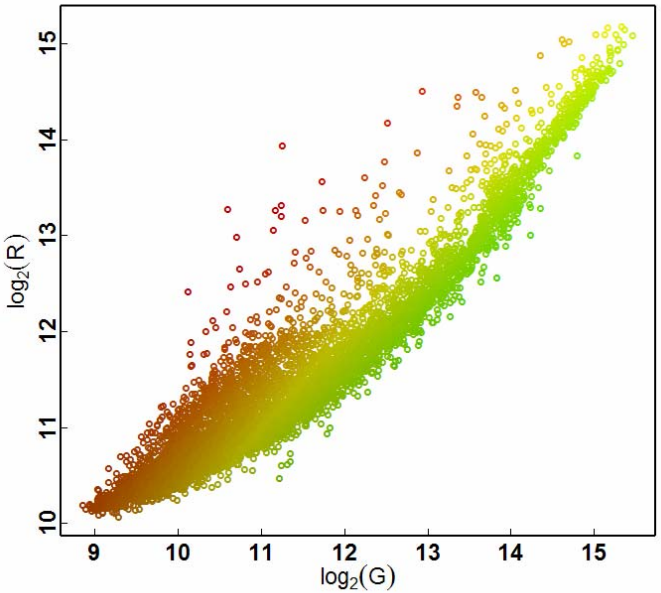


- **Log base 2 because the raw data is binary data (max intensity is 2^{16}-1 = 65535). It is also naturally to think of 2-, 4-, 8-fold etc up and down regulated genes. For the actual analysis, any log-base will do.** [

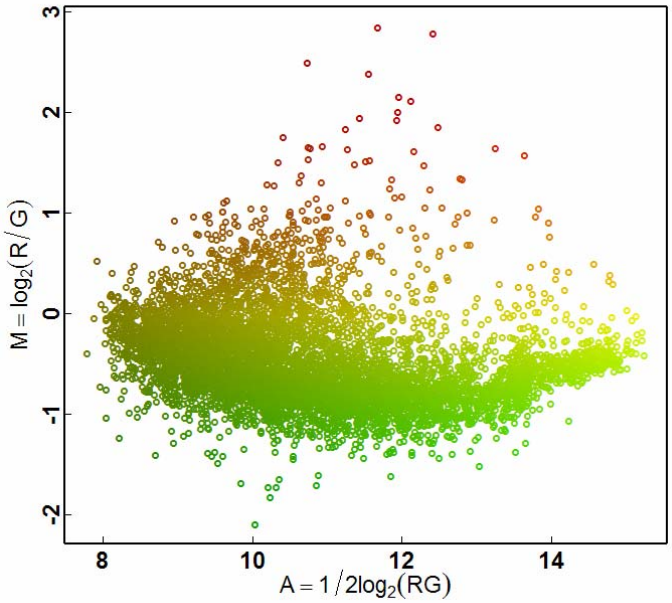
Summary of (R,G) ↔ ($\log_2$R,$\log_2$G) ↔ (M,A)



R vs G



log(R) vs log(G)



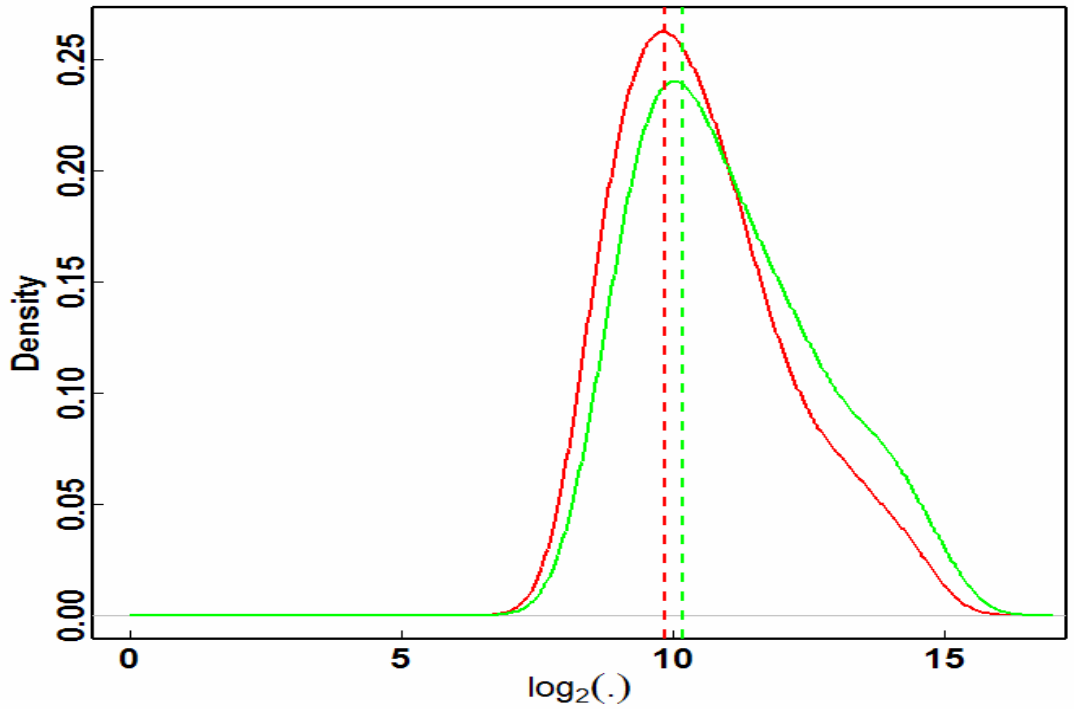
M vs A

R = red channel signal
G = green channel signal

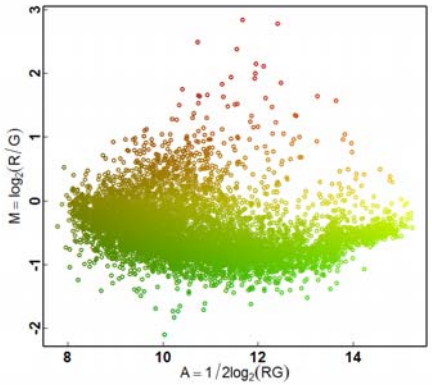
$M = \log_2(R/G)$ (log-ratio),
$A = \frac{1}{2}\cdot\log_2(R\cdot G)$ (log-intensity)

Intensity related variation

density of all signals



Example: Signals in the red channel seem to be slightly weaker than the signals in the green channel.
Compare to M vs A plot:

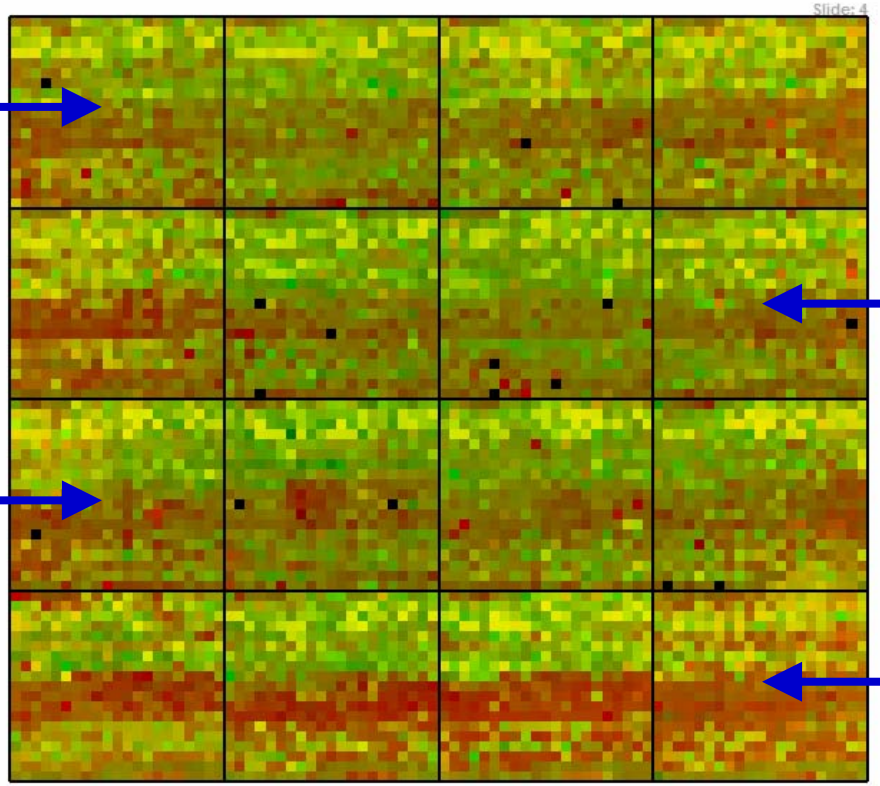


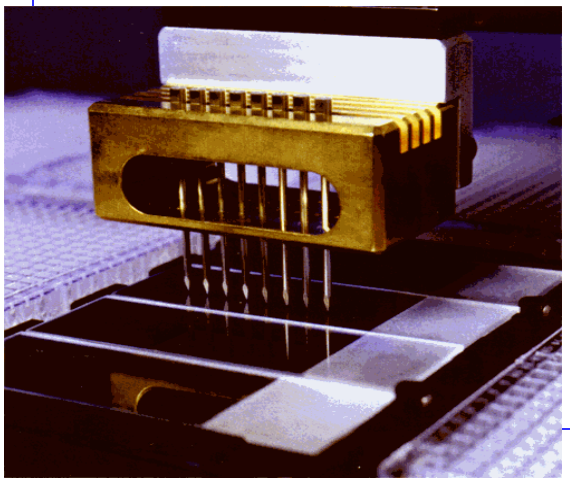
...more green.

$\log_2$R = red channel signal
$\log_2$G = green channel signal

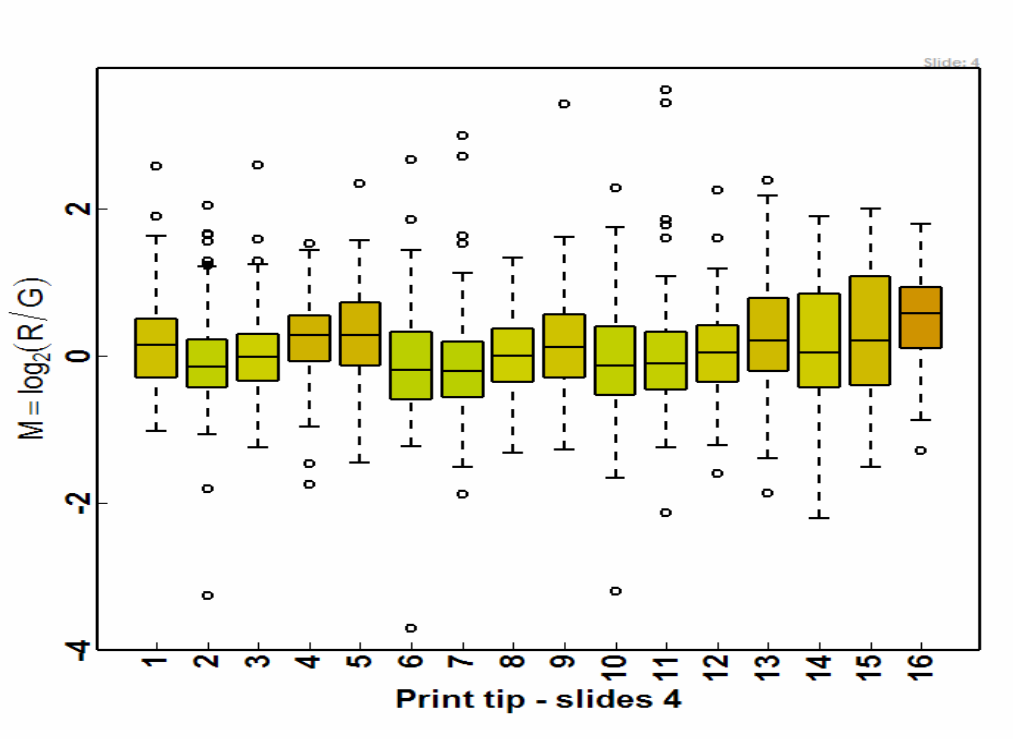
Spatial variation

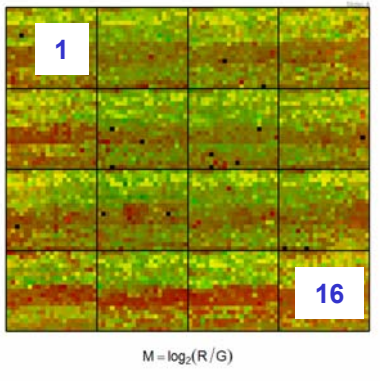
Spatial plot of log-ratios (M values)





Print-tip box plot of log-ratios





Spatial variations

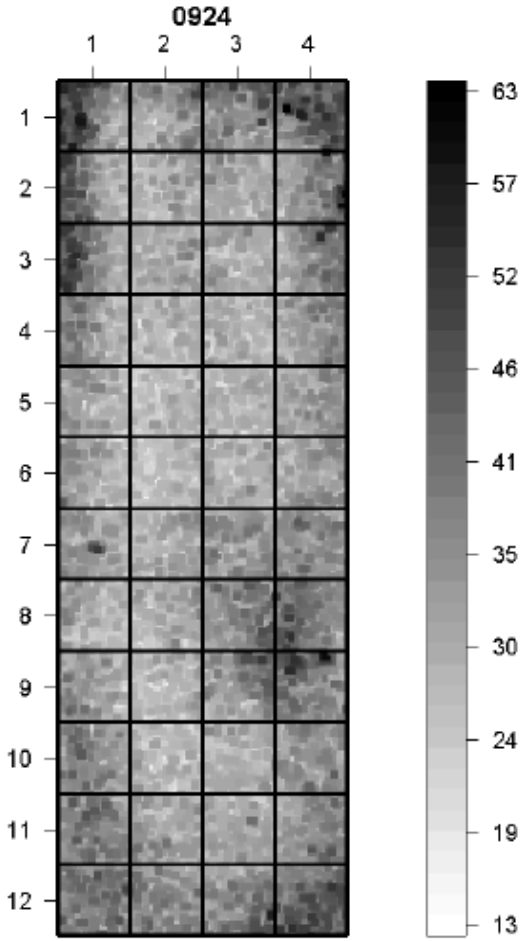


Figure 2. Spatial plot of green background values. The array was printed using a 12 x 4 pattern of print tips. The image shows that the array tends to be more green around the edges in the four corners. There is also a green patch in tip rows 8 and 9 and columns 3 and 4.

Spatial variations

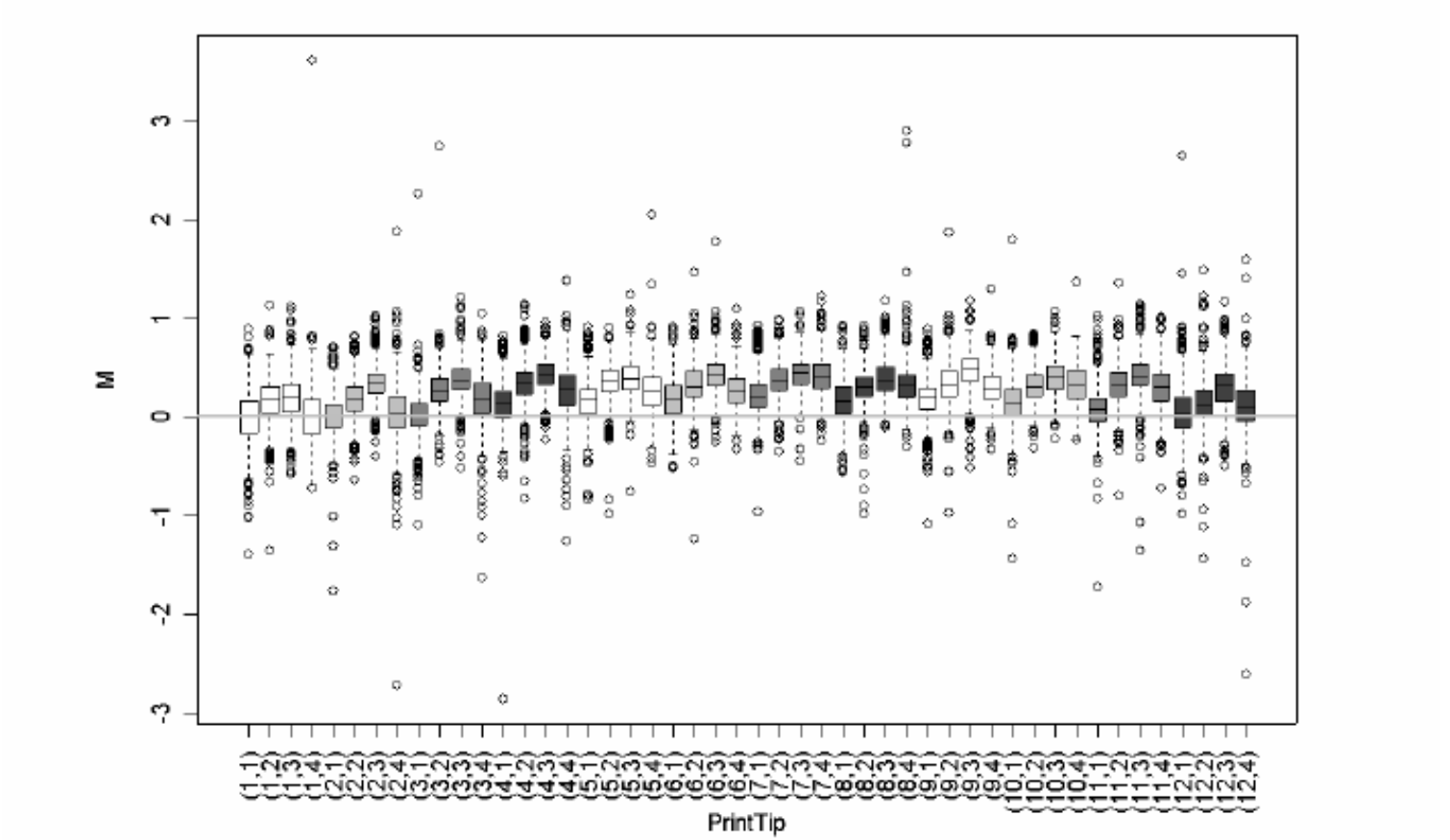
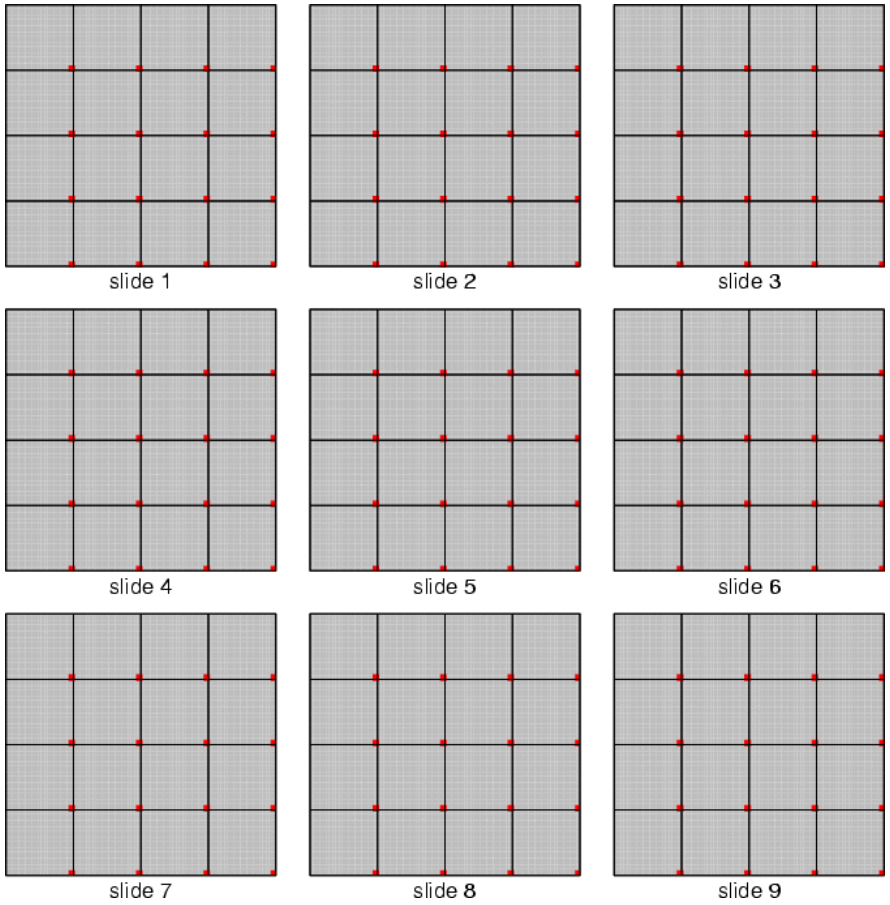


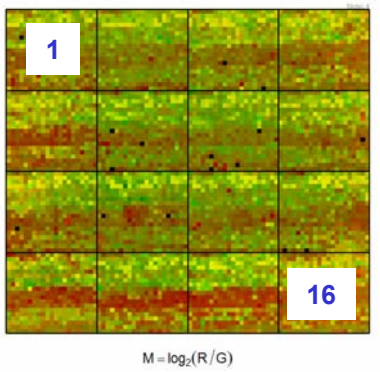
Figure 3. Boxplots of the M-values for each print tip group. The M-values are higher (more red) in the middle of each sequence of four and in the middle of the overall sequence.

Printing order of spots

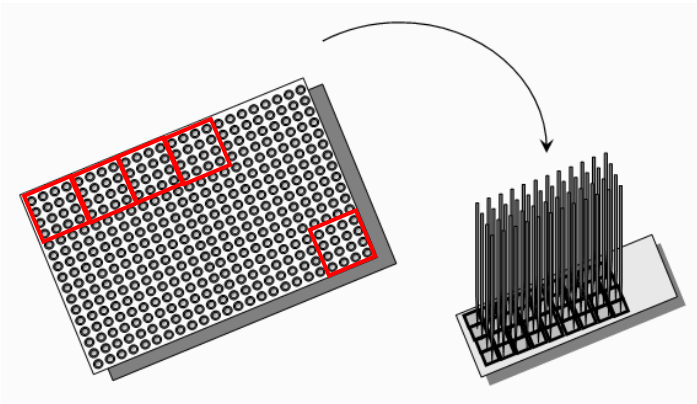


Above: 9 arrays

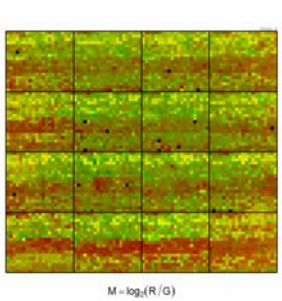
Hmm... why the horizontal stripes?)



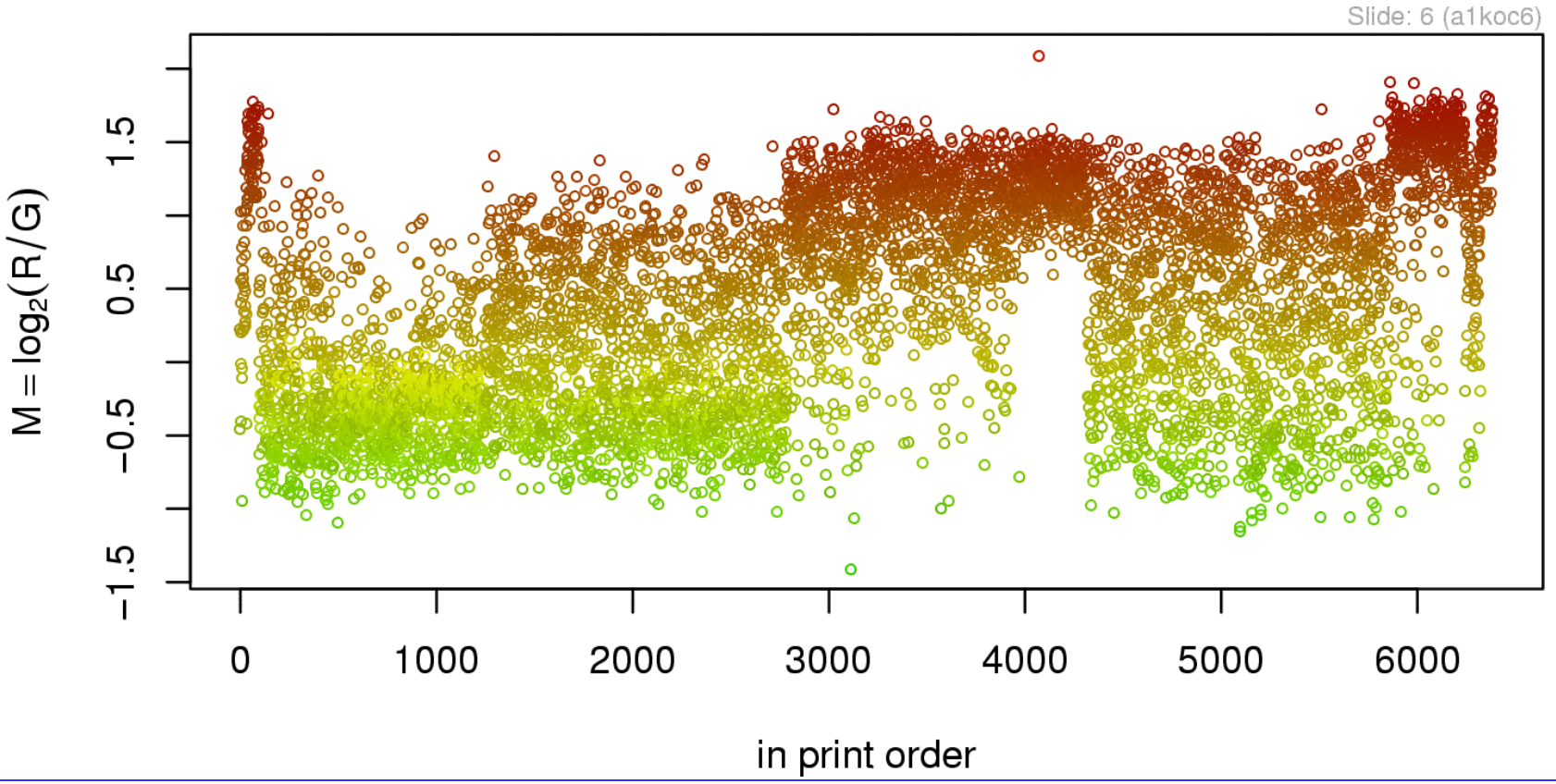
6384 spots printed onto 9 slides in total 399 print turns using 4x4 print-tips...



Print-order plot of log-ratios



The spots are order according to *when* they were spotted/dipped onto the glass slide(s). Note that it takes hours/days to print all spots on *all* arrays.



References

Exploratory data analysis for microarrays:

- **Yee Hwa Yang, Sandrine Dudoit, Percy Luu, and Terence P Speed. *Normalization for cDNA microarray data*. In Michael L. Bittner, Yidong Chen, Andreas N. Dorsel, and Edward R. Dougherty, editors, *Proceedings of SPiE*, volume 4266 of *Microarrays: Optical Technologies and Informatics*, pages 141–152, San Jose, California, June 2001. The International Society for Optical Engineering.**

- **Henrik Bengtsson. *Identification and normalization of plate effects in cDNA microarray data*. Preprints in Mathematical Sciences 2002:28, Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Sweden, 2002.**

- **Gordon Smyth and Terry Speed, *METHODS: Selecting Candidate Genes from DNA Array Screens*, Dec 2003**

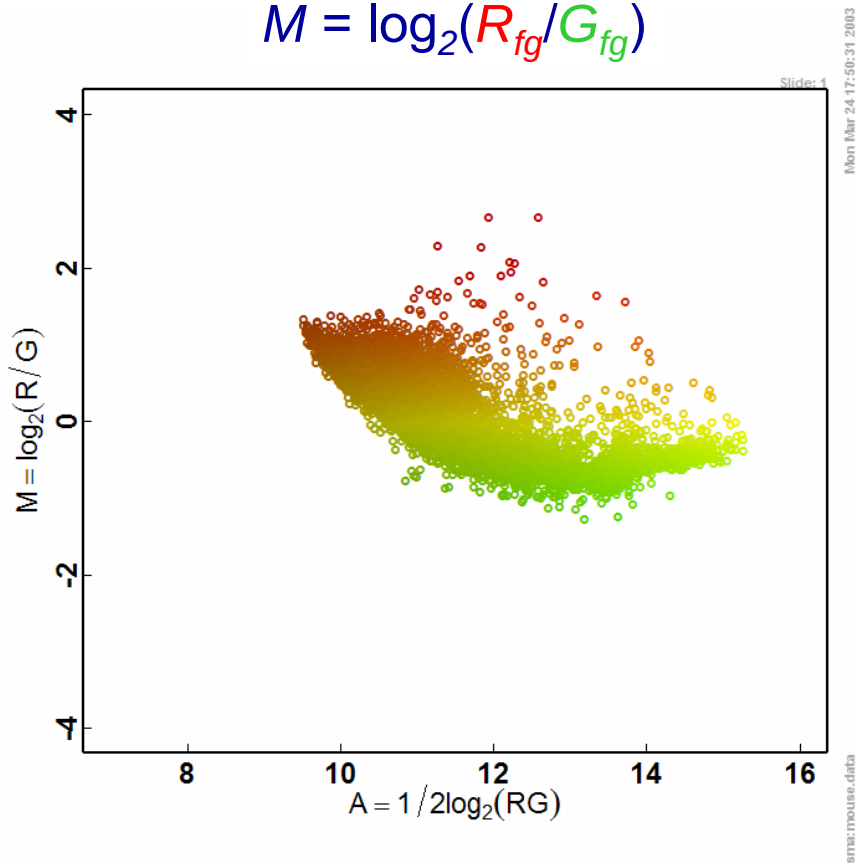
The cDNA microarray technology
-
PART IV:
Processing of data

- Background correction
- Normalization
- Image analysis

Background correction or not?

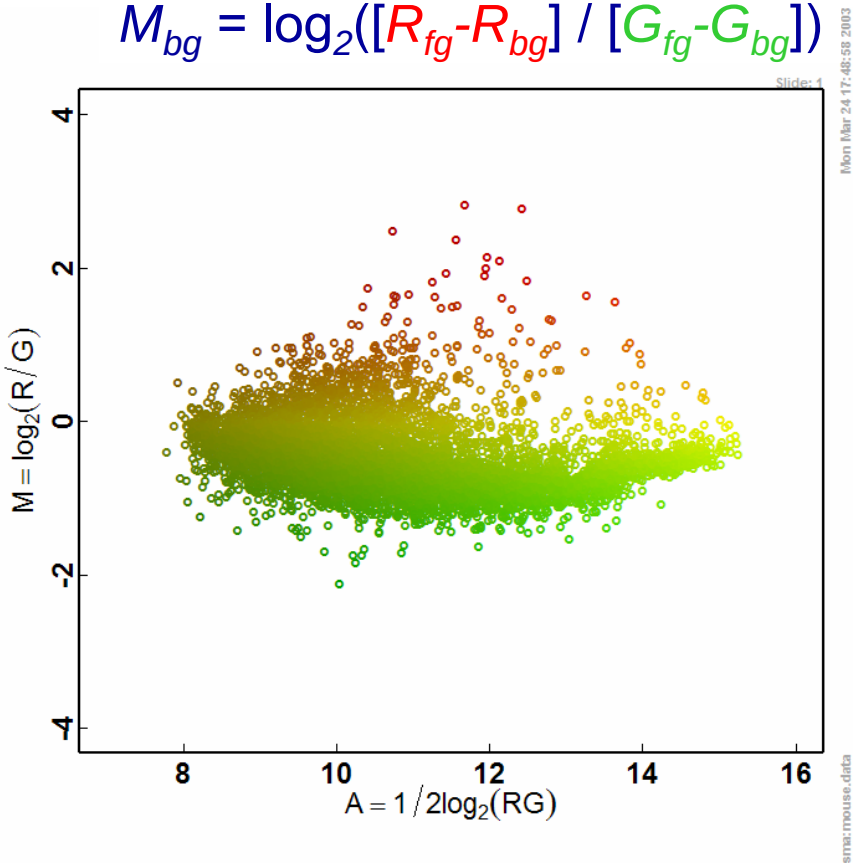
non-background subtracted

$M = \log_2(R_{fg}/G_{fg})$



background subtracted

$M_{bg} = \log_2([R_{fg}-R_{bg}] / [G_{fg}-G_{bg}])$



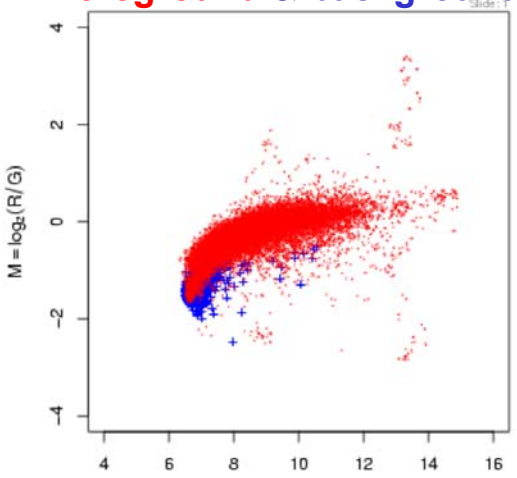
-seems better, but...

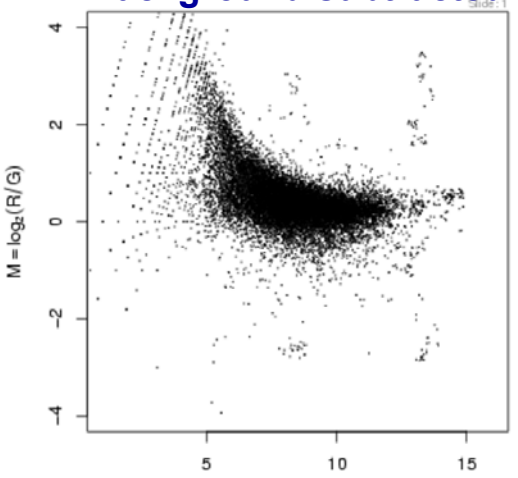
Background problem still not solved!

Foreground & background

Background subtracted

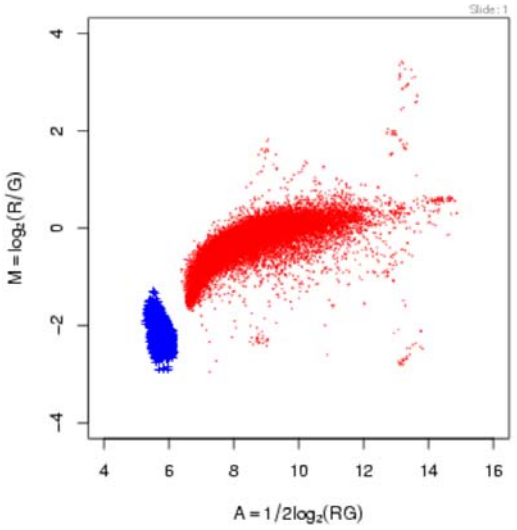
GenePix background:

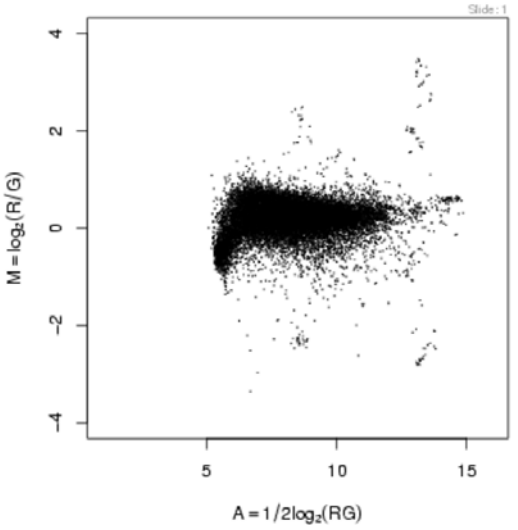




-Curvature in the other direction now! *Too much* background correction?!

Spot background: (morphological opening)



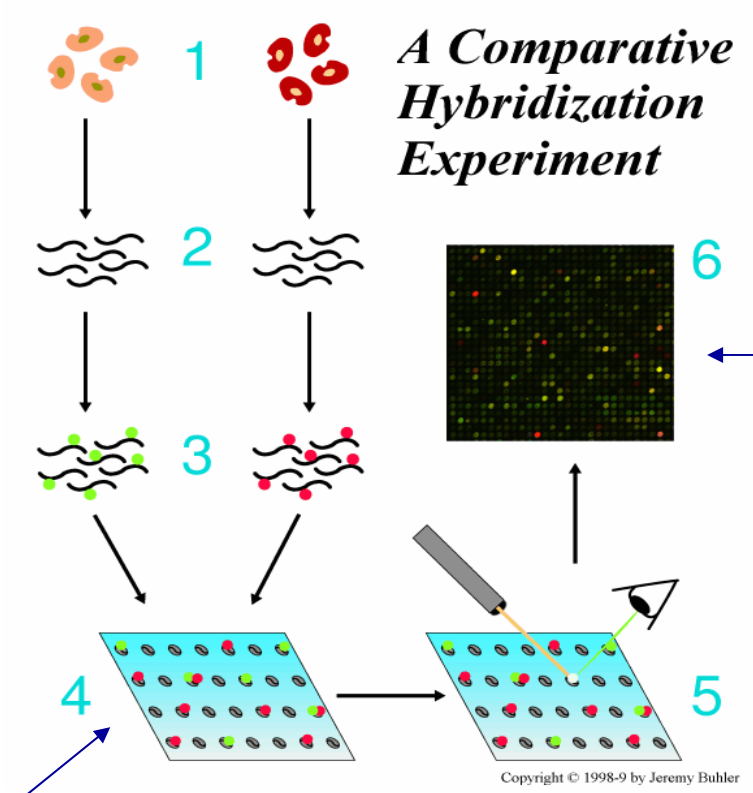


-Still curvature left. *Too little* background correction?!

Microarray *image* analysis

- Quantitation of fluorescence signals
- Data visualisation
- Meta-analysis (clustering)
- More visualisation

Technical



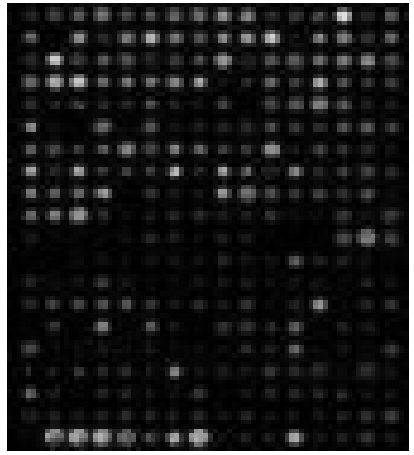
sample
(labelled)

pseudo-colour
image

probe
(on chip)

[image from Jeremy Buhler]

Images from scanner



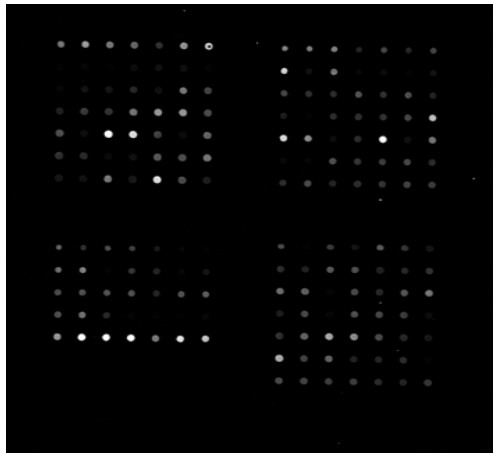
- Resolution
  - standard 10μm [currently, max 5μm]
  - 100μm spot on chip = 10 pixels in diameter
- Image format
  - TIFF (tagged image file format) 16 bit (65’536 levels of grey)
  - 1cm x 1cm image at 16 bit = 2Mb (uncompressed)
  - other formats exist e.g.. SCN (used at Stanford University)
- Separate image for each fluorescent sample
  - channel 1, channel 2, etc.

Images in analysis software

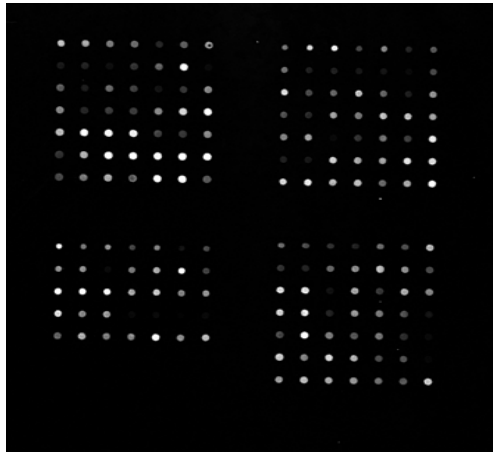
- The two 16-bit images (cy3, cy5) are compressed into 8-bit images
- Goal : display fluorescence intensities for both wavelengths using a 24-bit RGB overlay image
- RGB image :
  - Blue values (B) are set to 0
  - Red values (R) are used for cy5 intensities
  - Green values (G) are used for cy3 intensities
- Qualitative representation of results

Images : examples

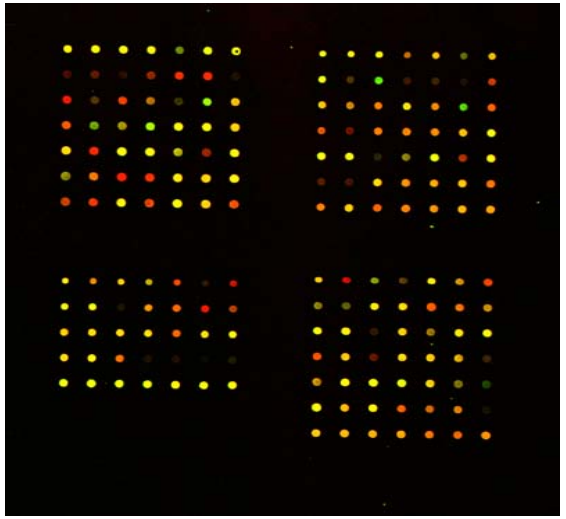
cy3



cy5



Pseudo-color overlay



Spot color	*Signal strength*	*Gene expression*
yellow	Control = perturbed	unchanged
red	Control < perturbed	induced
green	Control > perturbed	repressed

Processing of images

- Addressing or gridding
  - Assigning coordinates to each of the spots
- Segmentation
  - Classification of pixels either as foreground or as background
- Intensity extraction (for each spot)
  - Foreground fluorescence intensity pairs (R, G)
  - Background intensities
  - Quality measures

Addressing

- The basic structure of the images is known (determined by the arrayer)
- Parameters to address the spots positions
  - Separation between rows and columns of grids
  - Individual translation of grids
  - Separation between rows and columns of spots within each grid
  - Small individual translation of spots
  - Overall position of the array in the image
- The measurement process depends on the addressing procedure
- Addressing efficiency can be enhanced by allowing user intervention (slow!)
- Most software systems now provide for both manual and automatic gridding procedures

Segmentation

- Classification of pixels as foreground or background
  -> *fluorescence intensities* are calculated for each spot as measure of *transcript abundance*
- Production of a *spot mask* : set of foreground pixels for each spot

Segmentation

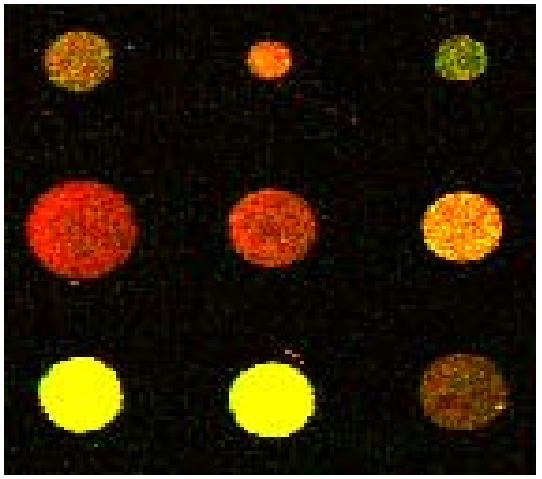
- Segmentation methods :
  - Fixed circle segmentation
  - Adaptive circle segmentation
  - Adaptive shape segmentation
  - Histogram thresholding

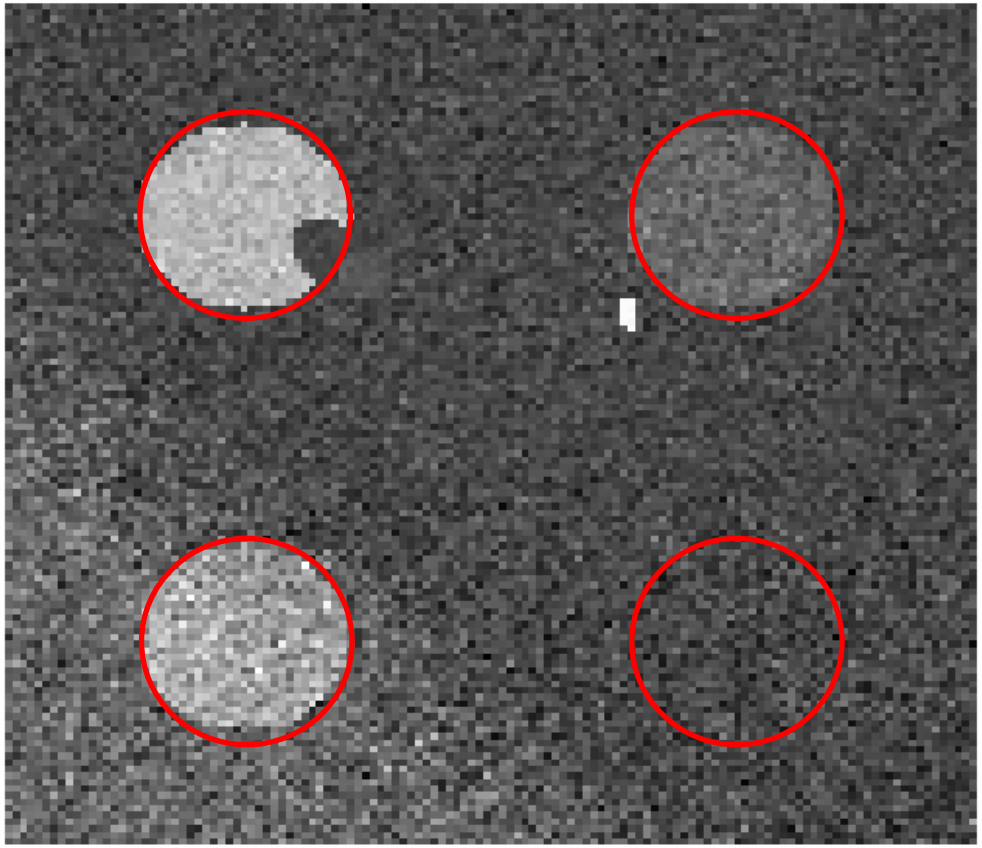
Fixed circle	ScanAlyze, GenePix, QuantArray
Adaptive circle	GenePix, Dapple
Adaptive shape	Spot, region growing and watershed
Histogram method	ImaGene, QuantArraym DeArray and adaptive thresholding

Fixed circle segmentation

- Fits a circle with a constant diameter to all spots in the image
- Easy to implement
- The spots need to be of the same shape and size

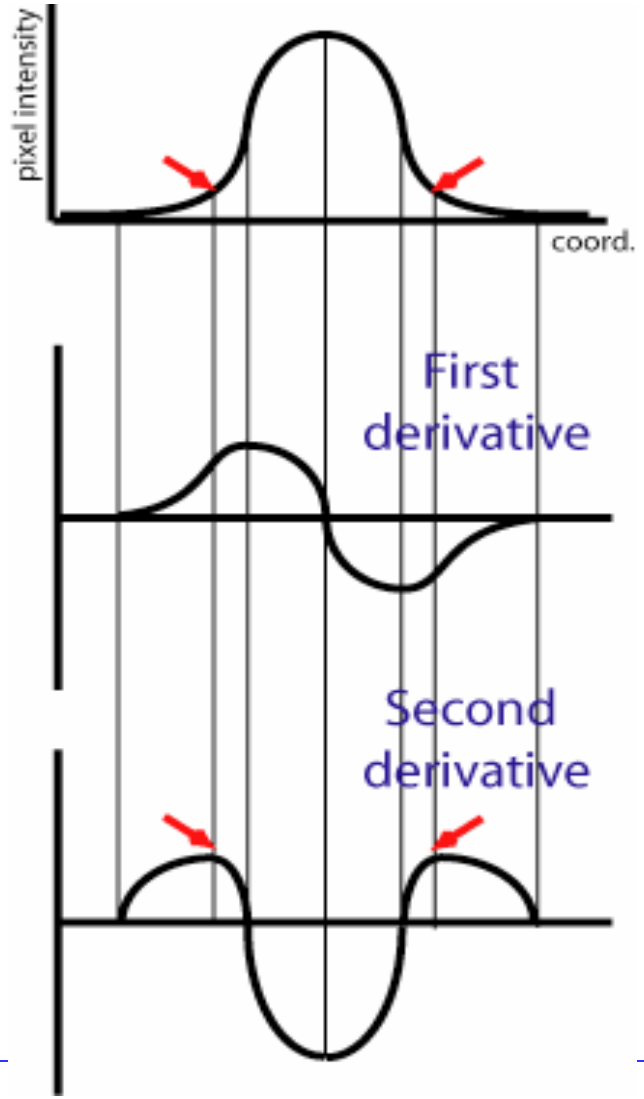
Bad example !





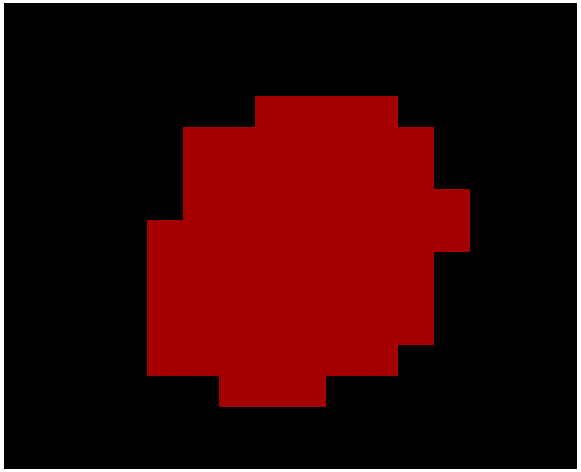
Adaptive circle segmentation

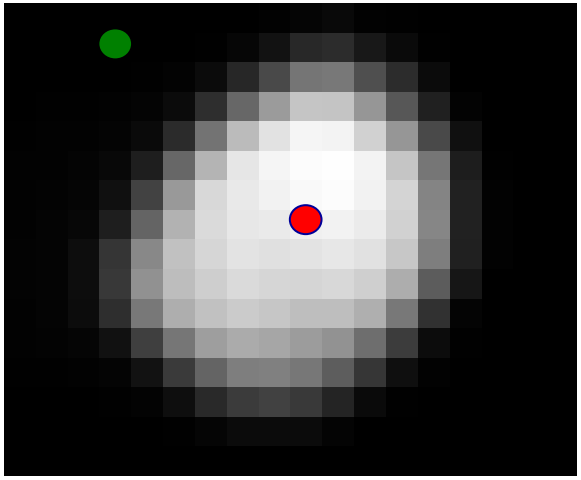
- The circle diameter is estimated separately for each spot
- *Dapple* finds spots by detecting edges of spots (second derivative)
- Problematic if spot exhibits oval shapes



Adaptive shape segmentation

- Specification of starting points or seeds
    - Regions grow outwards from the seed points preferentially according to the difference between a pixel's value and the running mean of values in an adjoining region.

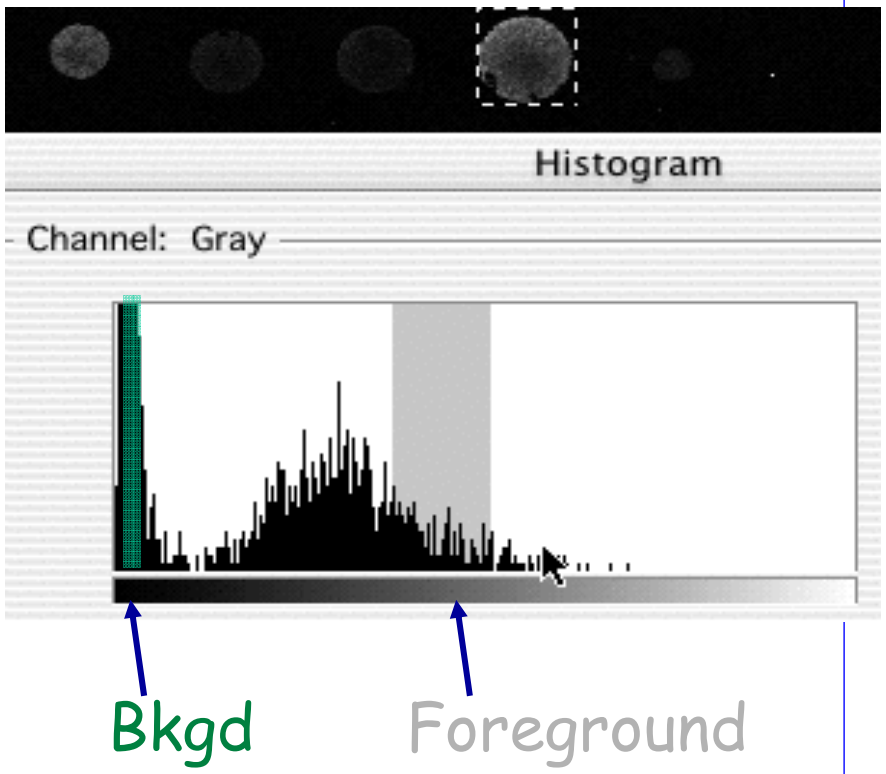




Histogram thresholding

- Uses a target mask chosen to be larger than any other spot
- Foreground and background intensity are determined from the histogram of pixel values for pixels within the masked area
- Example : QuantArray
  - Background : mean between 5th and 20th percentile
  - Foreground : mean between 80th and 95th percentile
- Unstable when a large target mask is set to compensate for variation in spot size

- *A **percentile** (or centile) is the value of a variable below which a certain percent of observations fall. So the 20th percentile is the value (or score) below which 20 percent of the observations may be found.*



Intensity extraction

Spot intensity

- The total amount of hybridization for a spot is proportional to the total fluorescence at the spot
- Spot intensity = sum of pixel intensities within the spot mask
- Since later calculations are based on ratios between cy5 and cy3, we compute the average* pixel value over the spot mask
  - *alternative : use ratios of medians instead of means

Background subtraction

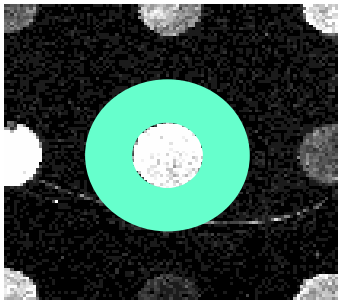
- Spot signal or simply signal is fluorescence intensity due to target molecules hybridized to probe sequences contained in a spot (what we would like to measure) plus background fluorescence (what we would rather not measure)
- Background is fluorescence that may contribute to spot pixel intensities but is not due to fluorescence from target molecules hybridized to spot probe sequences
  - Background may be due to dust particles, stray fluorescent molecules, fluorescence in the slide itself, etc.
- Background will *vary across the slide* so most software packages attempt to measure local background by quantifying pixel intensities around each spot.

Background subtraction

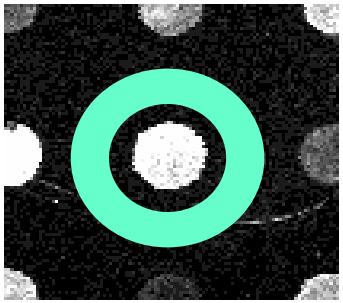
- Thus, spot measured intensity includes a contribution of non-specific hybridization and other chemicals on the glass
- Fluorescence from regions not occupied by DNA should by different from regions occupied by DNA
  - could be interesting to use local negative controls (spotted DNA that should not hybridize)
- Different background methods
  - Local background
  - Morphological opening
  - Constant background
  - No adjustment

Local background

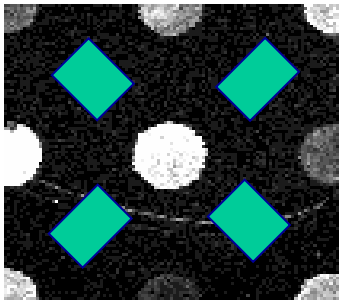
- Focusing on small regions surrounding the spot mask
- Median of pixel values in this region
- Most software package implement such an approach



ScanAlyze



ImaGene



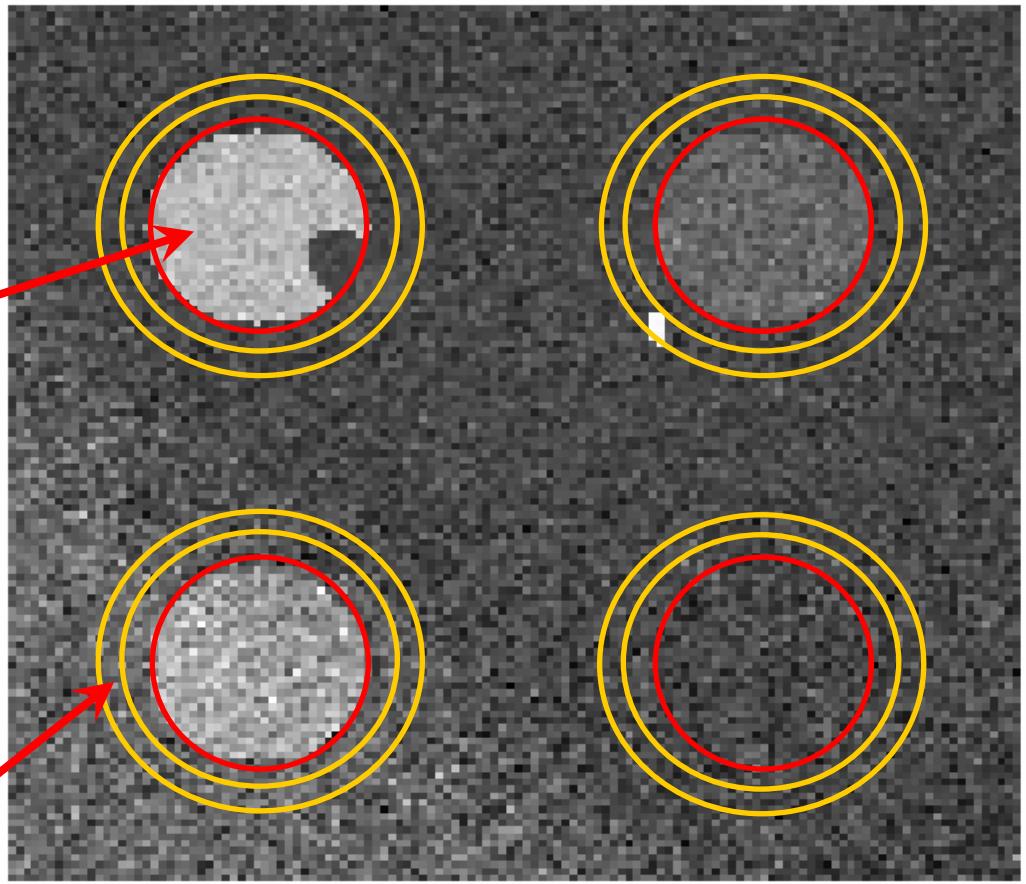
Spot, GenePix

- By not considering the pixels immediately surrounding the spots, the background estimate is less sensitive to the performance of the segmentation procedure

Software locates spots using info about grid.

Pixels in red circle may be segmented as signal.

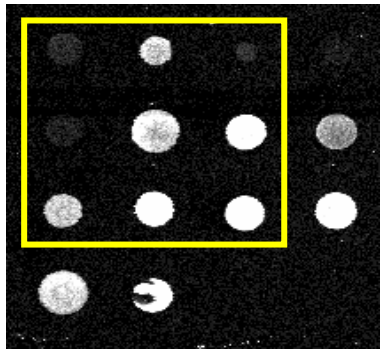
Pixels between gold lines may be segmented as background.

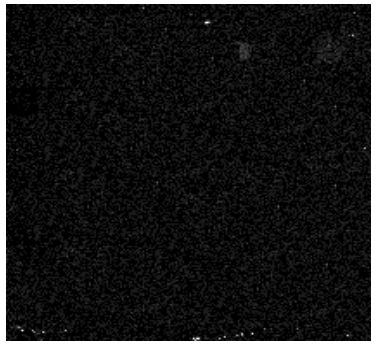


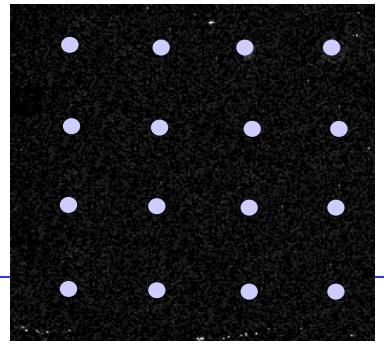
Segmentation algorithms vary in complexity and effectiveness.

Morphological opening (*spot*)

- Applied to the original images R and G
- Use a square structuring element with side length at least twice as large as the spot separation distance
- Remove all the spots and generate an image that is an estimate of the background for the entire slide
- For individual spots, the background is estimated by *sampling* this background image at the nominal center of the spot
- Lower background estimate and less variable







Constant background

- *Global* method which subtracts a constant background for all spots
- Some findings suggests that the binding of fluorescent dyes to ‘negative control spots’ is lower than the binding to the glass slide
- -> More meaningful to estimate background based on a set of negative control spots
  - If no negative control spots : approximation of the average background = third percentile of all the spot foreground values

No adjustment

- Do not consider the background

Quality measures

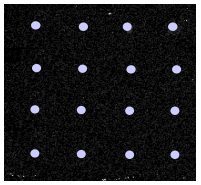
- How good are foreground and background measurements ?
  - Variability measures in pixel values within each spot mask
  - Spot size
  - Circularity measure
  - Relative signal to background intensity
  - *b-value* : fraction of background intensities less than the median foreground intensity
  - *p-score* : extent to which the position of a spot deviates from a rigid rectangular grid
- Based on these measurements, one can flag a spot, namely define a quality index and consider such a spot as “good” or “bad” with respect to such metric

Summary

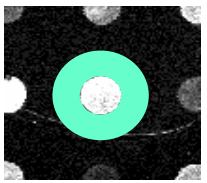
- The choice of background correction method has a larger impact on the log-intensity ratios than the segmentation method used
- The morphological opening method provides a better estimate of background than other methods
  - Low within- and between-slide variability of the log2 R/G
- Background adjustment has a larger impact on low intensity spots

$M = \log_2 R/G$

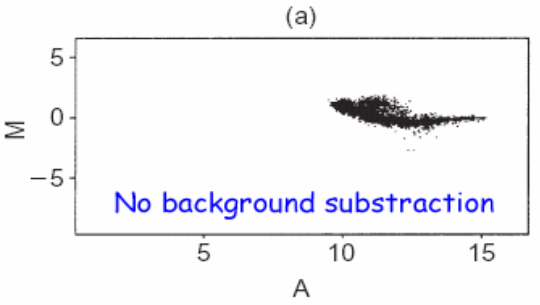
$A = \log_2 \sqrt{(R \cdot G)}$

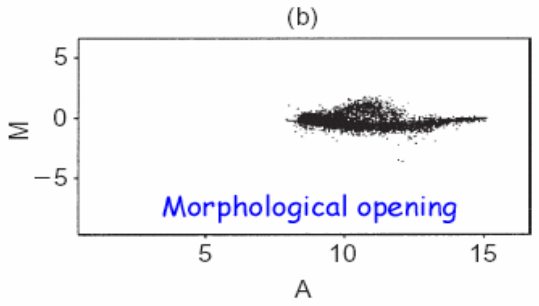


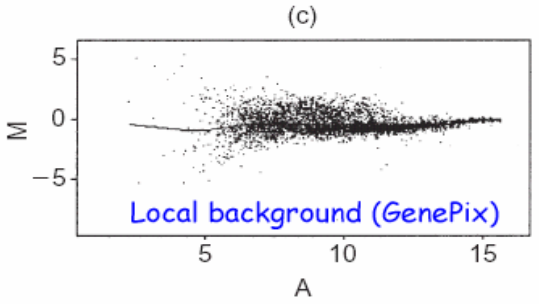
Spot, GenePix

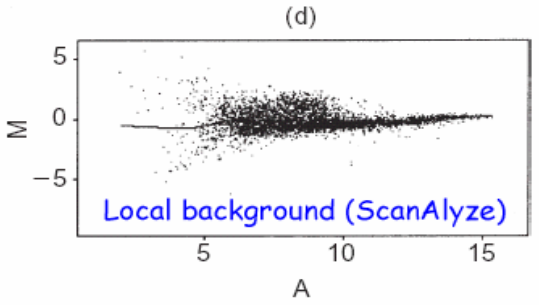


ScanAlyze







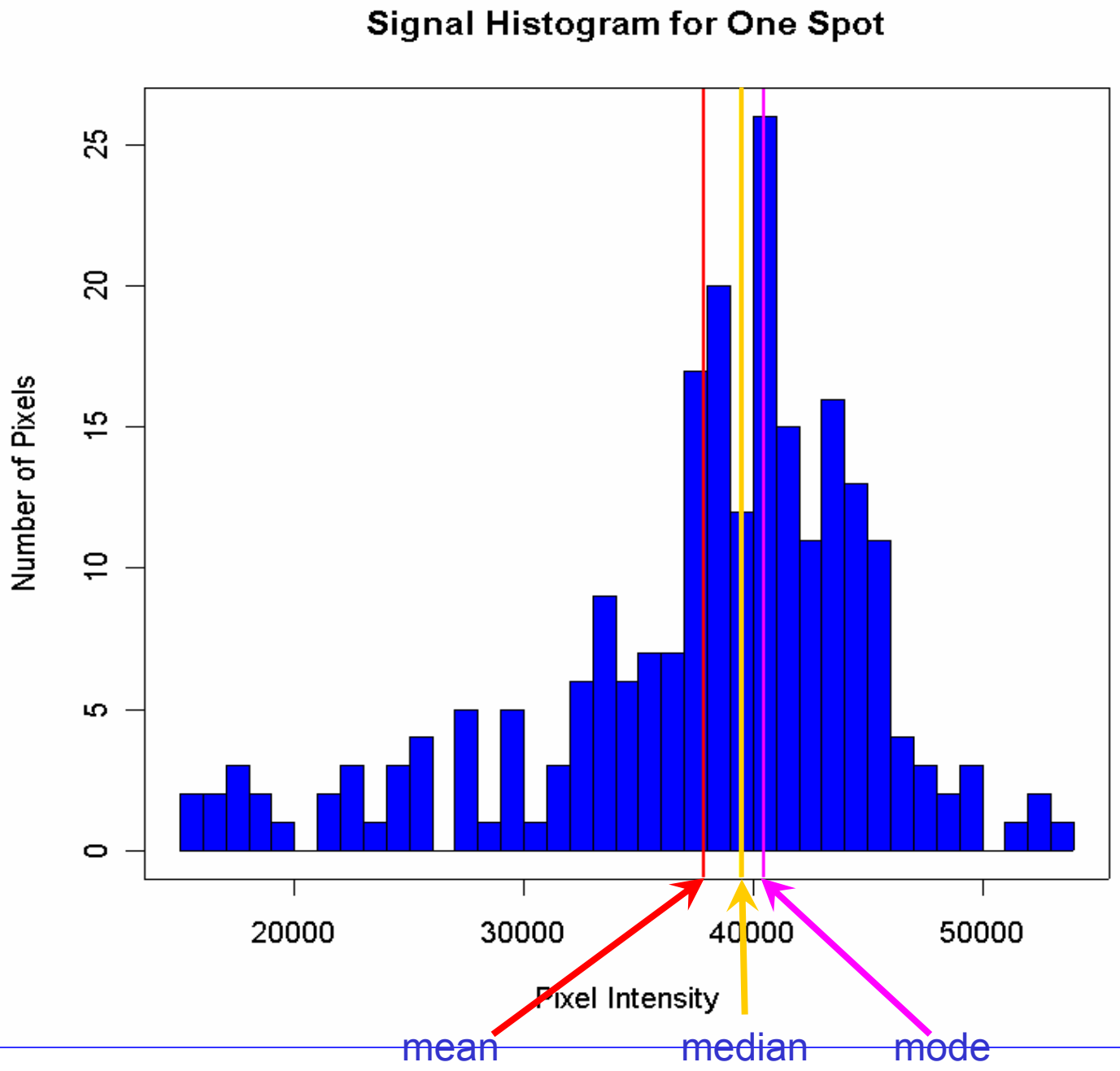


Spot Quality Assessment

- Common quality indexes
  - standard deviation: standard deviation of pixel intensities computed for both signal and background
  - shape regularity: First signal area of a spot is inscribed into a circle. Then the number of non-signal pixels that fall within this circle is computed and divided by the circle area. This ratio subtracted from 1 is defined as “shape regularity”.

Spot Quality Measures

- area to perimeter = (spot area)*4π/perimeter2
  - Ranges from 0 (highly non-circular shape) to 1 (a perfect circle).
  - diameter: diameter of spot's grid circle in pixels
  - saturation: indicates whether some pixels were censored at 2^{16}-1
- signal contamination – indicates whether signal pixels were "contaminated" (contained outliers)
- background contamination – indicates whether background pixels were "contaminated"
- other measures involving spot location


Signal Histogram for One Spot
Number of Pixels
0
5
10
15
20
25
20000
30000
40000
50000
Pixel Intensity
mean
median
mode

Example: Affymetrix GeneChips

- Image processing for Affymetrix GeneChips is typically done using proprietary Affymetrix software.
- The entire surface of a GeneChip is covered with square-shaped cells containing probes.
- Probes are synthesized on the chip in precise locations.
- Thus spot finding and image segmentation are not major issues.

References

Background estimation and correction:

- **Yee Hwa Yang, Michael Buckley, Sandrine Dudoit, and Terry Speed. *Comparison of methods for image analysis on cDNA microarray data*. Technical Report 584, Department of Statistics, University of California at Berkeley, Nov 2000.**

- **Anders Bengtsson. *Microarray image analysis: Background estimation using region and filtering techniques*. Master's Theses in Mathematical Sciences, Mathematical Statistics, Centre for Mathematical Sciences, Lund Institute of Technology, Sweden, December 2003. 2003:E40.**

- **Henrik Bengtsson, Göran Jönsson, and Johan Vallon-Christersson. *Calibration and assessment of channel-specific biases in microarray data with extended dynamical range*. Preprints in Mathematical Sciences 2003:37, Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Sweden, 2003.**

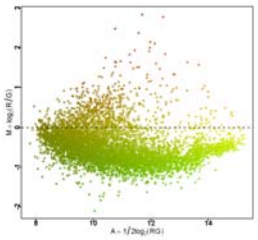
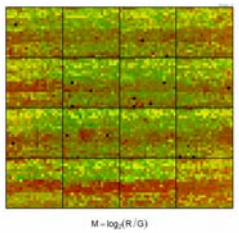
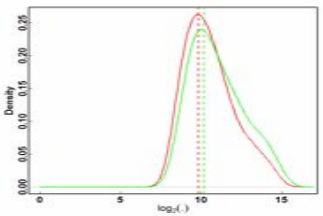
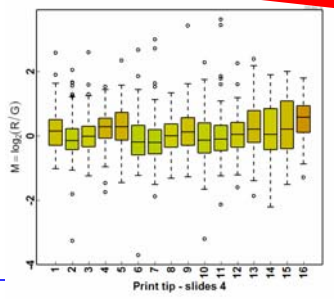
- **Charles Kooperberg, Thomas G. Fazzio, Jeffrey J. Delrow, and Thoshio Tsukiyama. *Improved background correction for spotted DNA microarrays*. *Journal of Computational Biology*, 9:55–66, 2002.**

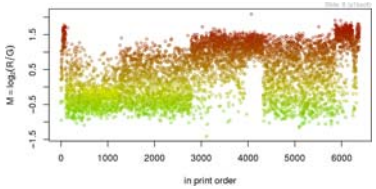
Normalization

- **Expectation: Most genes are non-differentially expressed,**
  - i.e. most of the data points should be around M=0.
- Idea: Do various exploratory plots to see if this assumption is met
  - For example, M vs A, spatial plots, density & boxplots plots, print-order plots etc.
- Result: We commonly observe something else:

  Measured value= real value + systematic errors + noise
- Correction: If so, normalize the data such that the expectations are met:

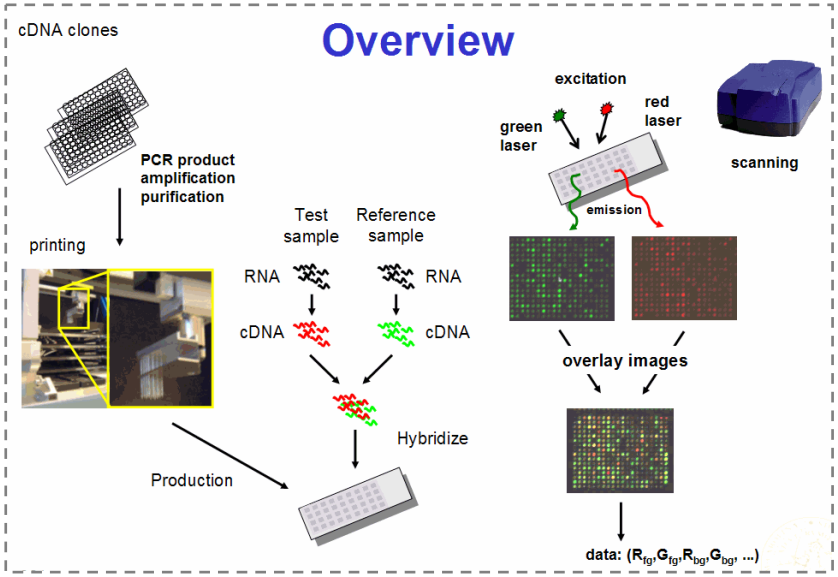
  Corrected value = real value ~~+ systematic errors~~ + noise



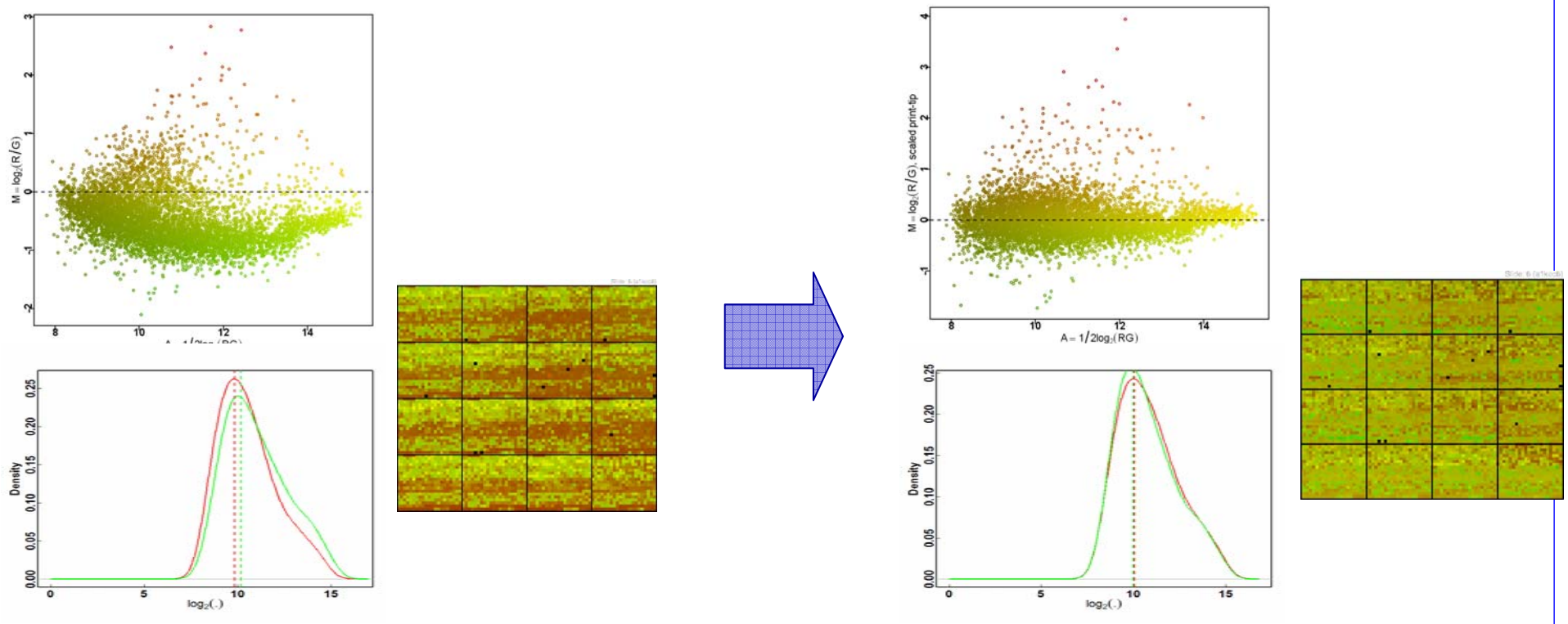
Normalization

- Sources of systematic effects, but also noise and natural variability
  - Biological variability
  - RNA extraction
  - Probe labeling
    - Ex: dye differences
  - Printing
    - Ex: print-order, plate-order, clone variation
  - Hybridization
    - Ex: temperature, time, mixing
  - Human
    - Ex: variation between lab researchers
  - Scanning
    - Ex: laser & detector, chemistry of the fluorescent label
  - Image analysis
    - Ex: identification, quantification, background methods



Normalization

- Assume that all systematic effects have been removed



“loess” normalization

- Print-tip loess normalization provides a well-tested general purpose normalization method which has given good results on a wide range of arrays
  - The method may be refined by using quality weights for individual spots
  - The method is best combined with diagnostic plots of the data which display the spatial and intensity trends.
- When diagnostic plots show that biases still remain in the data after normalization, further normalization steps such as plate-order normalization or scale normalization between the arrays may be undertaken
- Composite normalization may be used when control spots are available which are known to be not differentially expressed
- Variations on loess normalization include global loess normalization and 2D normalization

loess

- Each M-value is normalized by subtracting from it the corresponding value of the tip group loess curve. The normalized log-ratios N are the residuals from the tip group loess regressions, i.e.,

$$N = M - loess_i(A)$$

  - where $loess_i(A)$ is the loess curve as a function of A for the th tip group
  - Each loess curve is constructed by performing a series of *local regressions*, one local regression for each point in the scatterplot
  - This allows to account for *both spatial and intensity variations*

loess: example

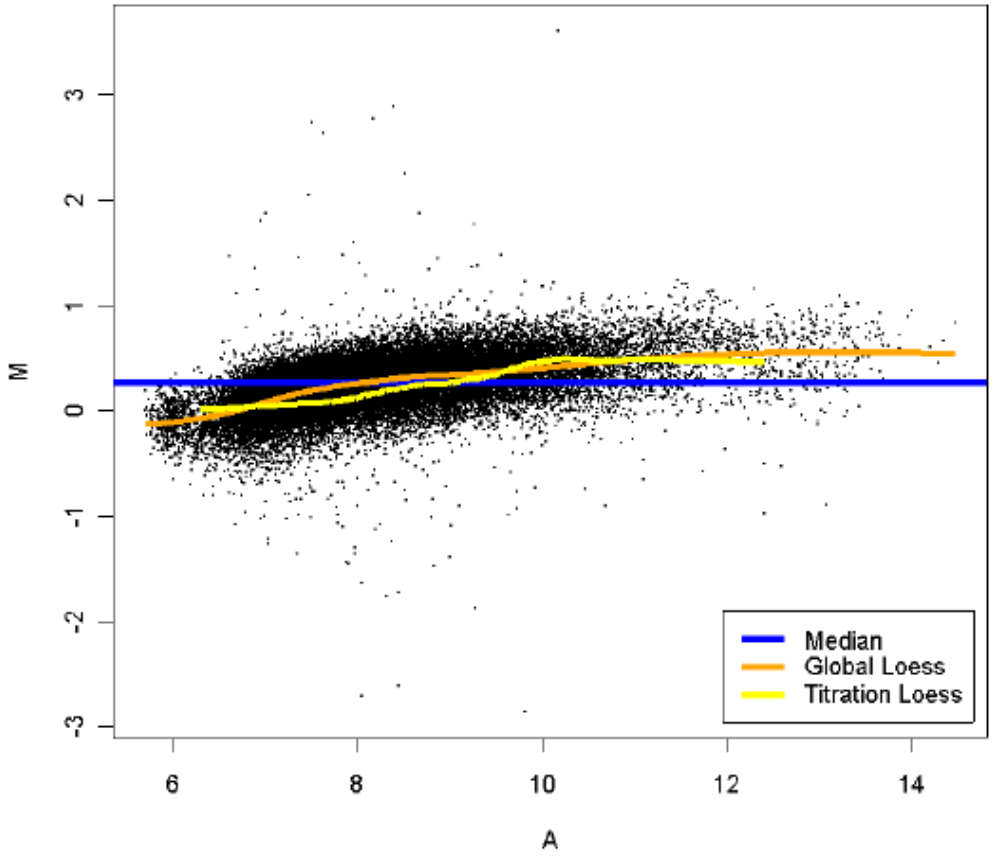
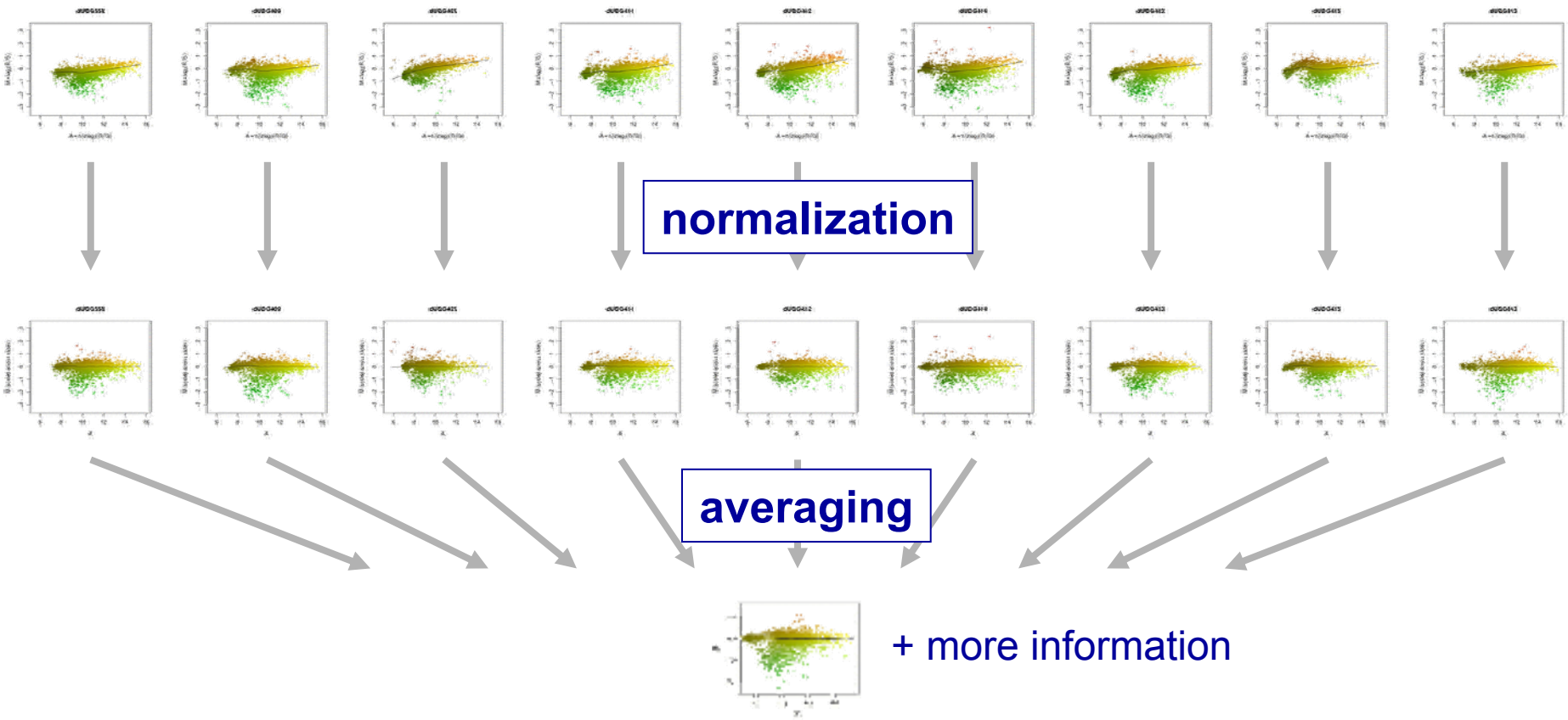


Figure 1. MA-plot showing three different trend lines. The horizontal blue line shows the median of the M-values. The continuous orange curve shows the overall trend line as estimated by loess regression. The yellow curve shows the loess curve through a set of control spots known to be not differentially expressed.

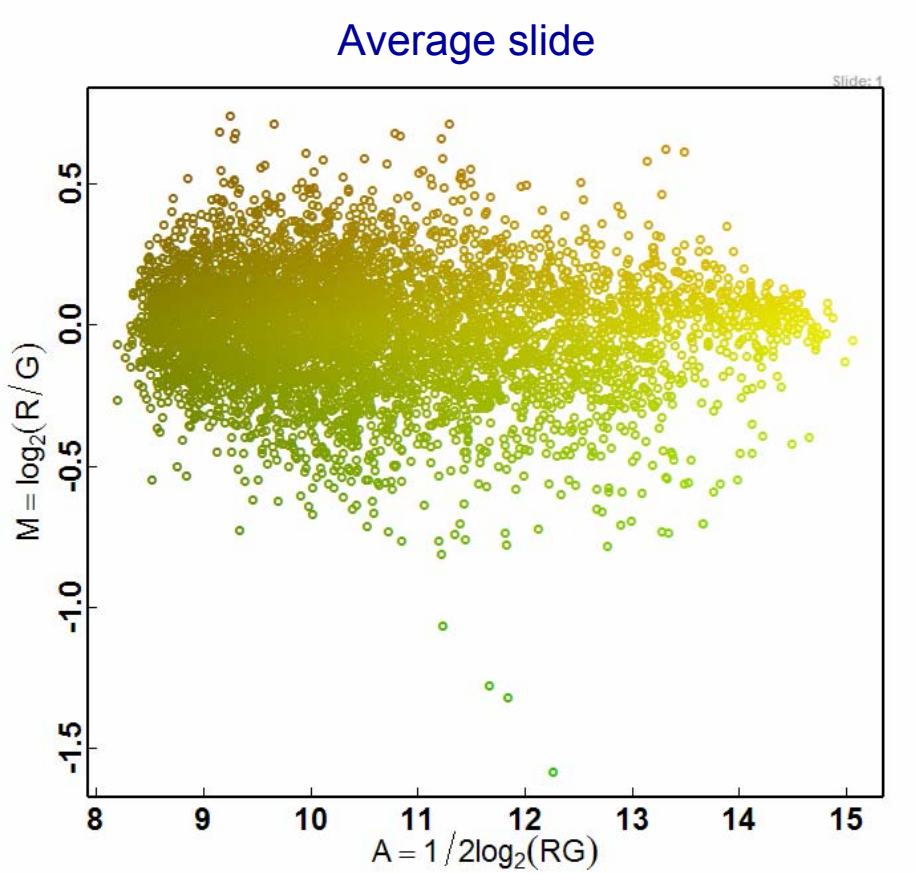
The cDNA microarray technology
-
PART V:
Identifying differentially expressed genes

1. Cut-off by M values
2. The t-statistics and cut-off by T values
3. Multiple testing and adjusted the p-values
4. Validation

Combining data from several slides

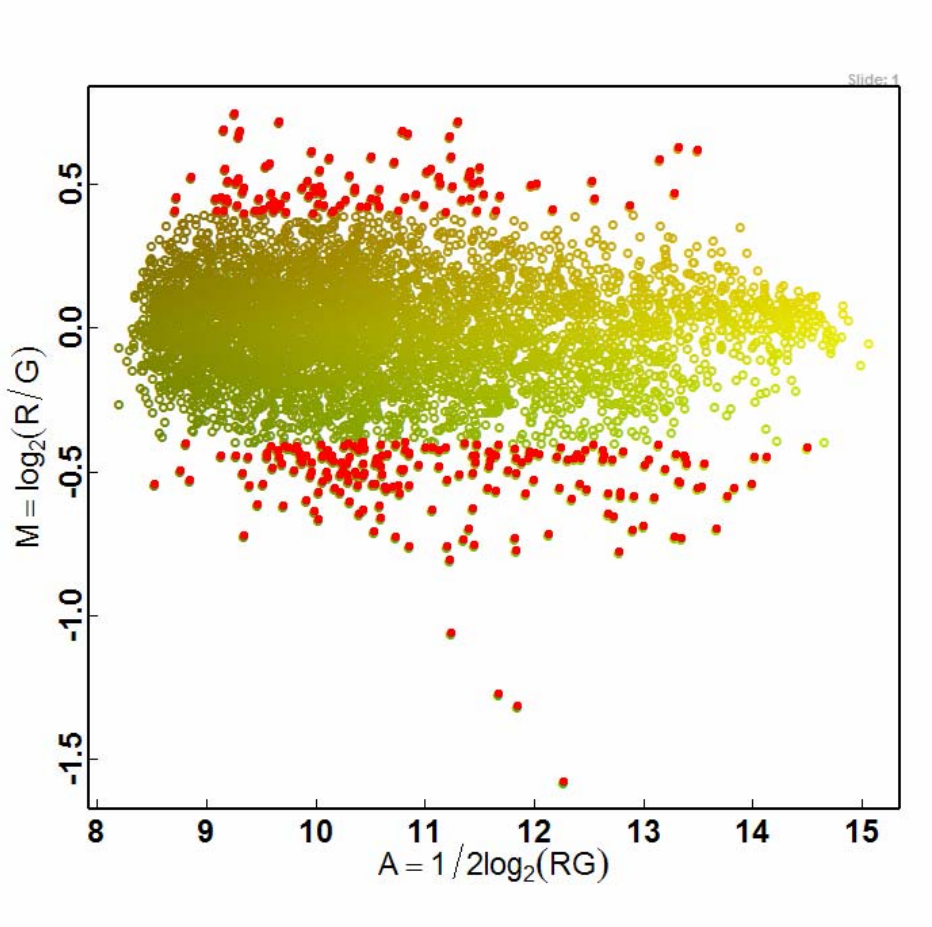


Average of all normalized slides



Cut-off by log-ratios (naive)

Top 5% of the absolute *M* values:



Finding differentially expressed genes

For *each* gene *i* we have the hypothesis test:

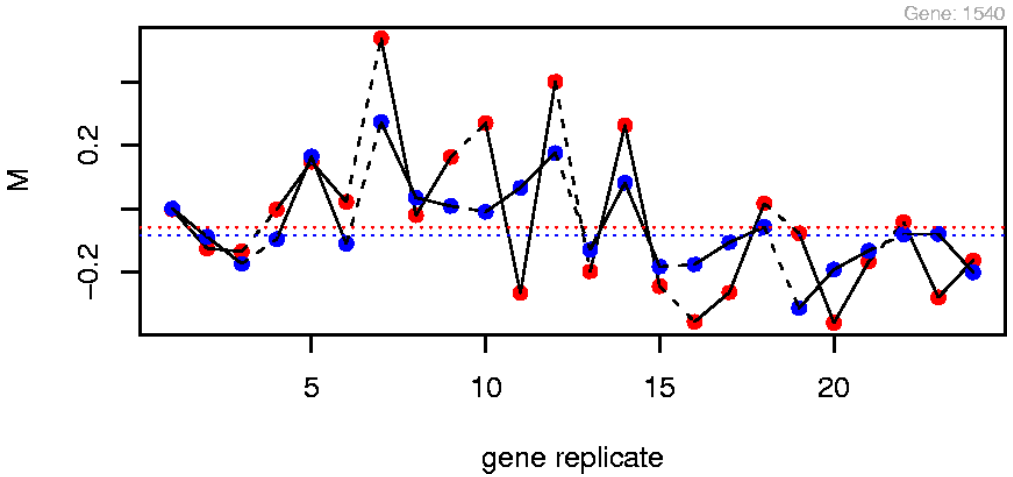
Null (neutral) hypothesis $H_{0,i}: M_i = 0$
Alternativ hypothesis $H_{1,i}: M_i \neq 0$

Risk level: Allow α=5% test to reject H_0 even if it is true.

If we are far enough away of M = 0, then we can reject H_0, otherwise we assume it is true.

The t-statistics

- Idea: For replicated data, i.e. multiple measurements of the same thing, we trust the estimate of the average (mean or median) more if the deviation (std.dev. or MAD) is small. If the deviation is large, we do not trust it that much.



Example: The blue and the red genes have almost the same average log-ratio, but we are more *confident* with the measure of the blue gene since its variability across replicates is smaller.

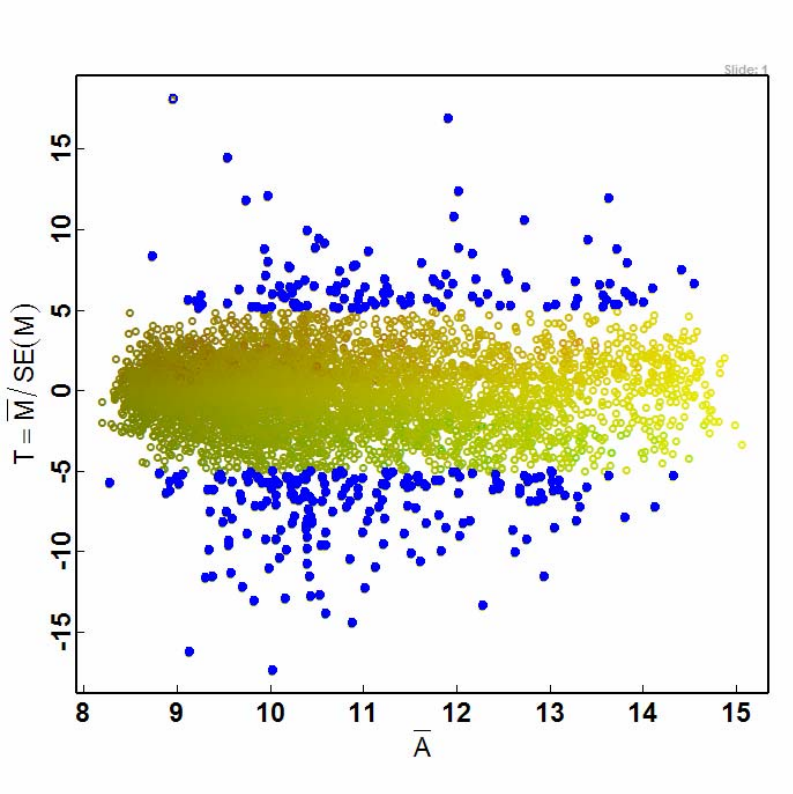
- **The *T statistics* down-weight the importance of the average if the deviation is large and vice versa;**

$$T = \text{mean}(x) / SE(x)$$

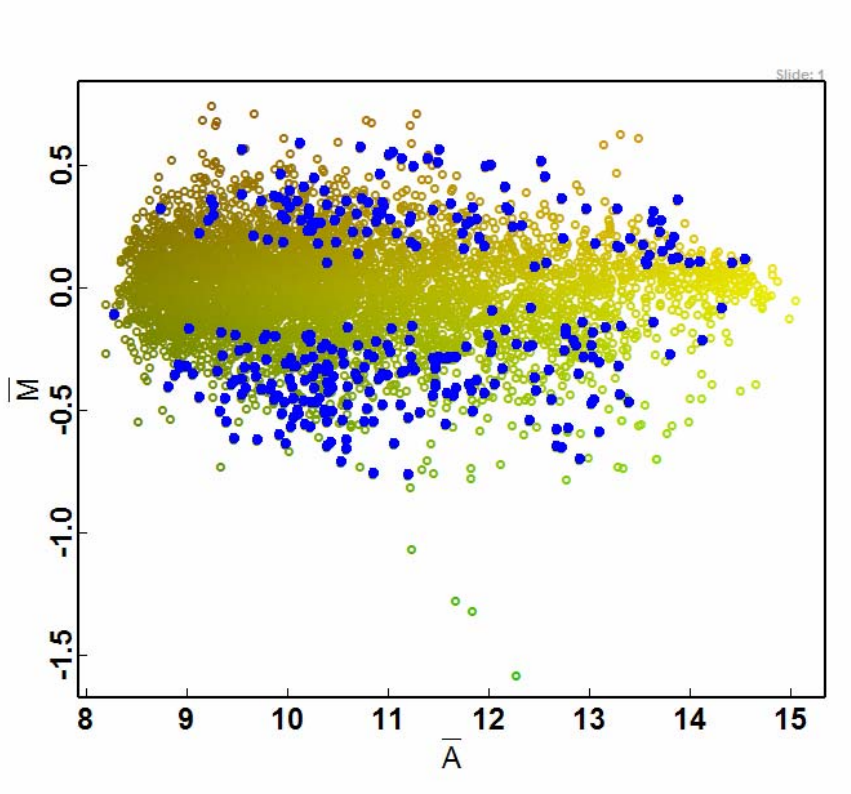
where SE(x)=std.dev(x)/N (standard error of the mean)

Cut-off by T values (better)

Top 5% of the absolute *T* values:



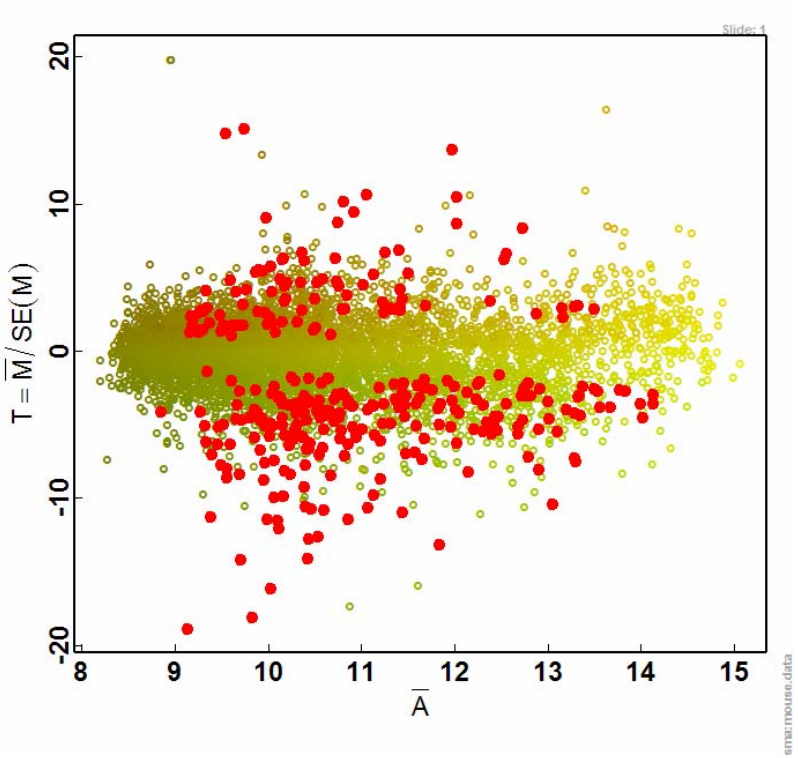
T vs A



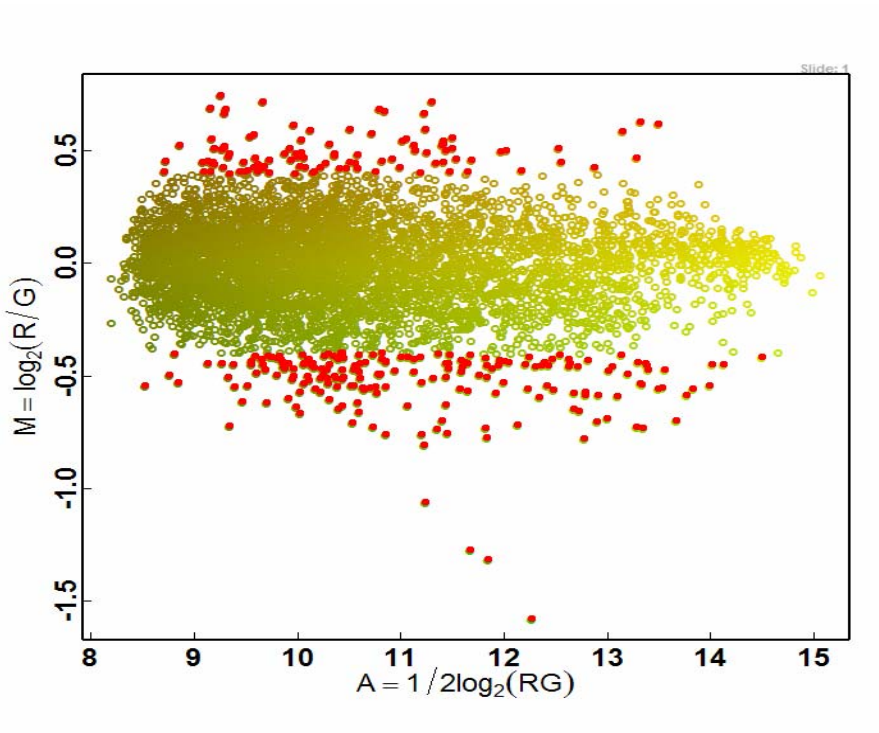
M vs A

Compare

Top 5% of the absolute *M* values in red:



T vs A



M vs A

False positive and false negative

False Positive and False Negative: Statistical test vs. truth

	If truth is **T ≠ 0:**	If truth is **T=0:**
Statistical test decision: Reject H_0: **T=0**.	Correctly reject H_0	False Positive
Statistical test decision: Do not reject H_0: **T=0**.	False Negative	Correctly not reject H_0

In cDNA microarray experiments we commonly test the hypothesis H_0 that T=0 against T≠0 (non-DE or not) ***for every gene separately***. For the genes for which we reject H_0, we say they are differentially expressed.

However, ***by chance*** we will reject H_0 for some genes that are not DE. We call these findings *false positive* (Type I error). Genes that are DE, but for which we do not reject H_0 are called *false negative* (Type II error).

The multiple testing problem

- Pitfalls
  - thousands of tests, i.e. each gene is tested against H0: T=0. By chance some will “fail” (be rejected).
  - false positives problems more serious.
  - need to adjust p-values.
- Different adjustment procedures
  - Bonferroni, Sidak, Duncan, Holm, etc. Not discussed here, but available automatically in the better microarray analysis software.

References

- **Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, and Terence P. Speed. *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments*. Technical Report 578, Department of Statistics, University of California at Berkeley, 2000.**

- **M. Callow, S. Dudoit, E. Gong, T. Speed, and E. Rubin. *Microarray expression profiling identifies genes with altered expression in HDL-deficient mice*. *Genome Research*, 10(12):2022–9, December 2000.**

- **Ingrid Lönnstedt and Terence P. Speed. *Replicated microarray data. Statistical Sinica, 12(1), 2002.***

- **Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, and Terence P. Speed. *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments*. Technical Report 578, Department of Statistics, University of California at Berkeley, 2000.**

- **Y. Ge, S. Dudoit & T. P. Speed, *Resampling-based multiple testing for microarray data hypothesis* (submitted to Test, Spain). Technical Report #633 of UCB Statistics, 2003.**

Summary: Identification of DEs

- You need replication and statistics to find real differences.
- Cutoff by log ratios is not enough/correct.
- Cutoff by t-statistics is much better.
- Multiple testing => must adjust the p values.
- Validate your results by other means!

Take-home messages

- Good image analysis is essential
  - Some software are obsolete and not that good
  - Background correction or not is not solved. Progress has been done, but more research is needed
- Normalization is needed
- Use at least the t-statistics to identify differentially expressed genes
  - Do not rely exclusively on log-ratios.
- Multiple testing must be considered; adjust your p-values.
- Talk to a statistician before doing the experiments!
  - They do think about these kind of problems for a living.