Microarray Data Analysis (1)

Data quality assessment for microarrays

Microarray Raw Data (1): cDNA arrays

One .GPR (for GenePix Results) file per chip containing

- One row per gene but many columns with

– Intensitiy values for each channel (R, G)

– Summary values for intensities

– Quality controls, such as FLAG

Intensity values are converted into a single expression matrix containing

- One column per chip with log(R/G) values
- One row per gene (same rows as .GRP files)
- Gene information stored in a .GAL (for GenePix Array List) file
- Both .GPR and .GAL are ASCII files
- An accurate description of these files is available here

Microarray Raw Data (2): Affymetrix arrays

One .CEL file per chip, containing

- PM and MM values for each probe in the chip
- Presence/Absence calls (one per probeset)

  They can be interpreted as a statistical test of the spot foreground intensity in the experimental sample respect to the background intensity distribution

- Separate PM/MM values are converted into a single expression matrix containing:
  - One column per chip with absolute intensity values
  - One row per probeset
- Gene information stored in the .CDF file
- .CEL file is a binary file
- An accurate description of these files is available here

Looking at microarray data
Diagnostic Plots

Was the experiment a success?

Exploring experimental results

- Microarray experiments generate huge quantities of data.
- It is hard to decide if things “seem to be all right” just by looking at the numbers.
- Standard statistical approach: use plots.
  - Show all data together.
  - Highlight structures,
  - May help detect problems (“unusual patterns”)

Diagnostic plots for microarrays

- Microarray data usually considered at two levels
  - Low-level: Data directly coming from the scanner
  - High-level: Processed from low-level data. Expression values, normalized or not.
- Some diagnostic plots may differ between one and two color arrays, specially for looking at low level values.
- Other may be used for any type of arrays or for any level.

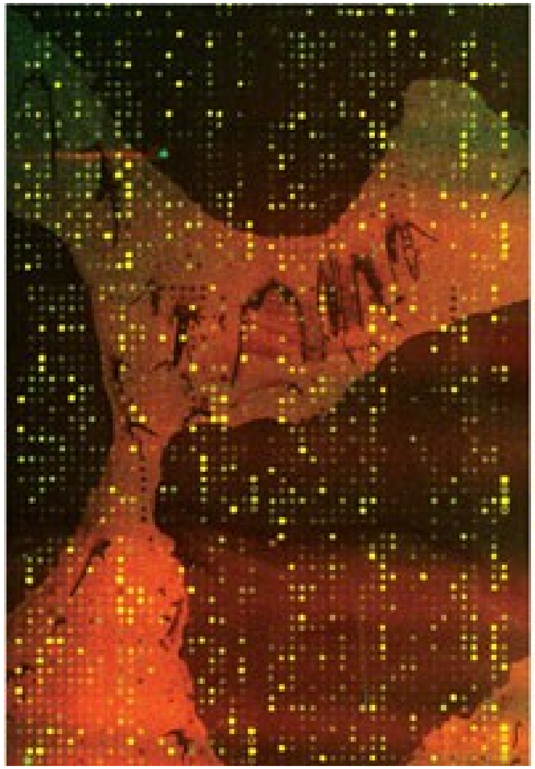
Diagnostic plots

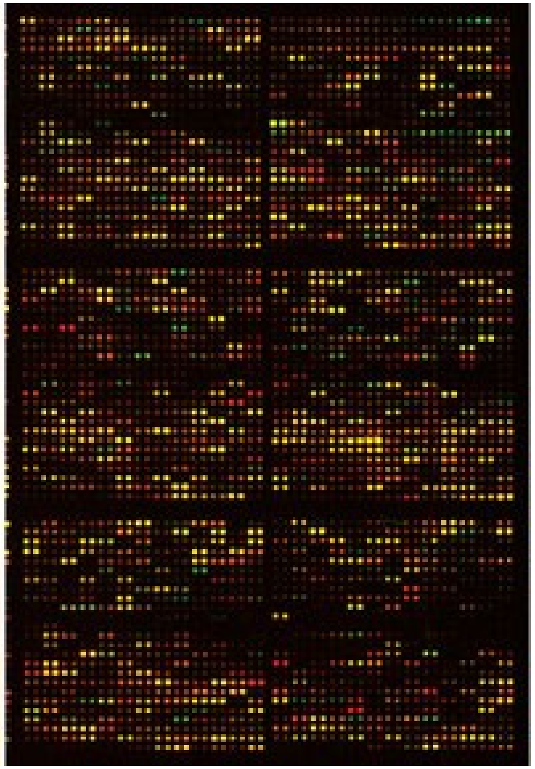
Microarray type *Data type*	**One color**	**Two color**	**General**
Low level **1col: (probe, probe-set)** **2col: (single channel)**	Layout image Degradation plots Density plots *Probeset plots*	Scatterplots R,G MA-plot Signal2Noise plots Layout image (G, R)	PCA Histogram/Density Boxplot
High-level **1 col: Relative expression** **2 col: Absolute expression**	MA-Plots *Model-based plots* *(NUSE, RLE, Residual)*	Layout image (log ratios)	PCA Histogram/Density Boxplot

Diagnostic plots for two color arrays

Red / Green overlay images

- Start by looking at the slides





Provides information on

- colour balance,
- uniformity of hybridization,
- spot uniformity,
- background, and
- artifiacts such as
  - dust or
  - scratches

Bad: high bg

Good: low bg

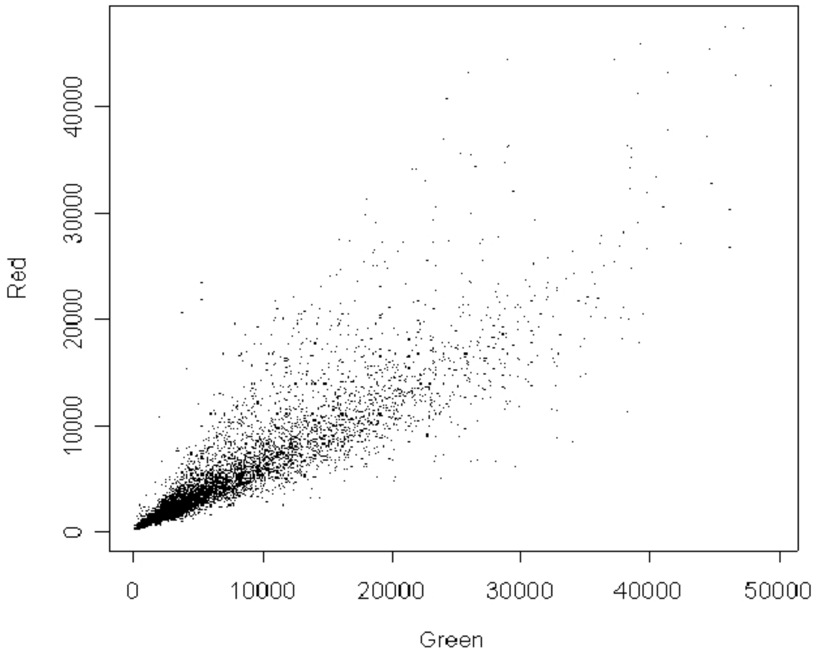
MA-plot (1)

To determine whether correction (normalization) is needed, one can plot R vs G intensities and see whether the slope of the line is around 1

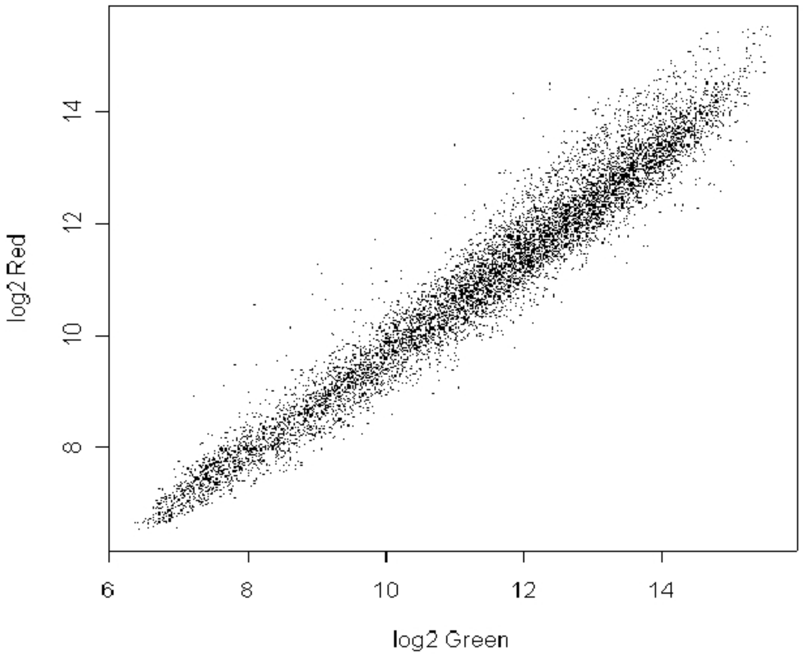
But a better representation of genes with “medium” expression is to take logs...

MA-plot (2)

Biologically, a unit change in log2 represents a 2-fold change



Linear



Log scale

MA-plot (3)

- An improved method of the R vs G plot is the *MA-plot*, which is basically a scaled 45 degree rotation
- It is a plot of the distribution of
- M-value which is the $\log_2$ of the R/G intensity ratio

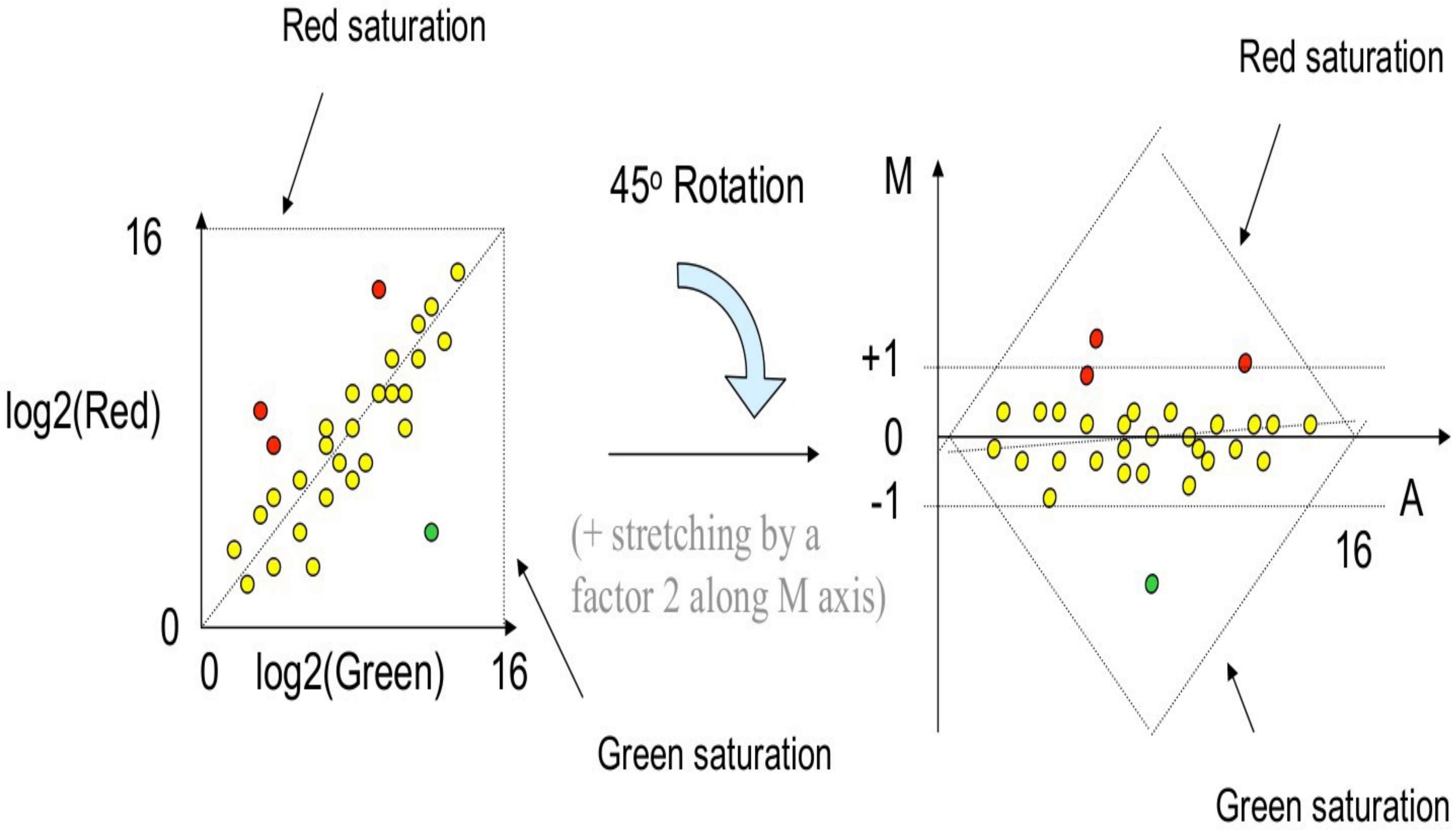
$$M=\log_2(R/G)=\log_2(R)-\log_2(G)$$

A-value which is the $\log_2$ of average intensity

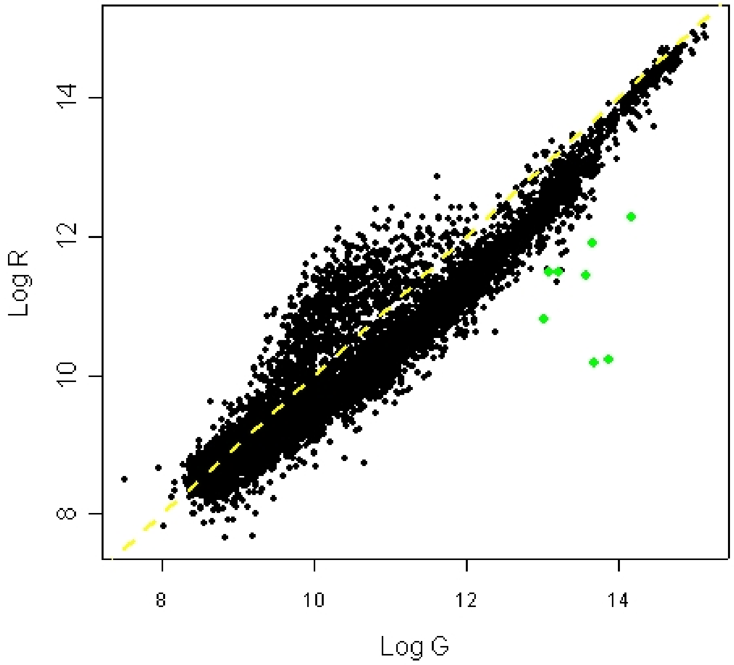
$$A= (\log_2(R)+\log_2(G))/2 = \log_2 \sqrt{RG}$$

- The general assumption is that most of the genes would not see any change in their expression -> the majority of the points on the M would be located at 0, since $\log_2(1) = 0$

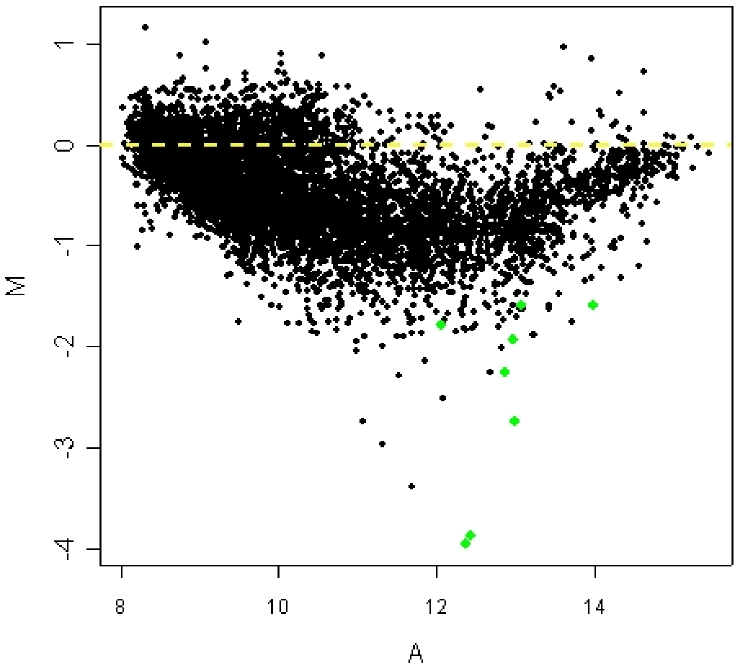
MA-plot (4): M vs A



MA-plot (5)

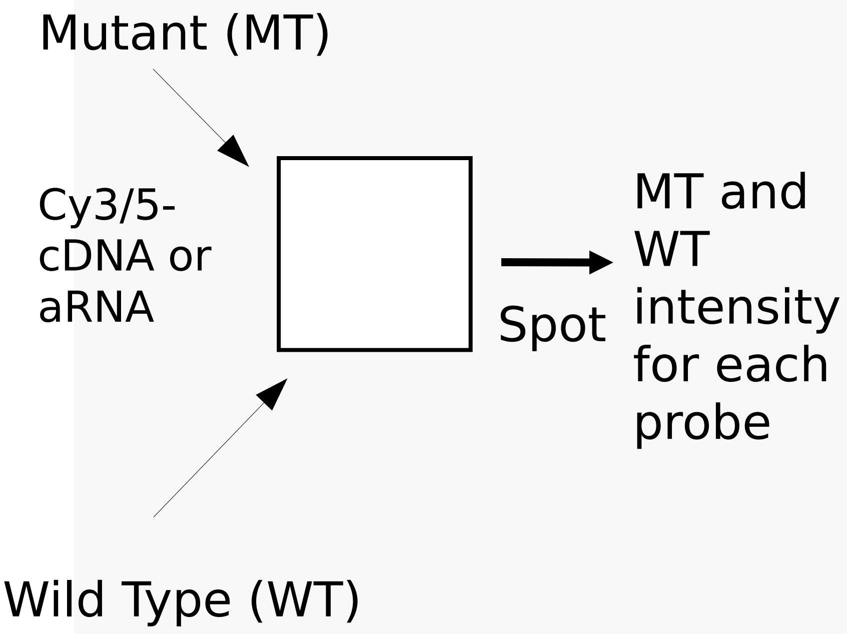


$\log_2 R$ vs $\log_2 G$



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

MA-plot for spotted arrays (2 colors)


Mutant (MT)

Cy3/5-cDNA or aRNA

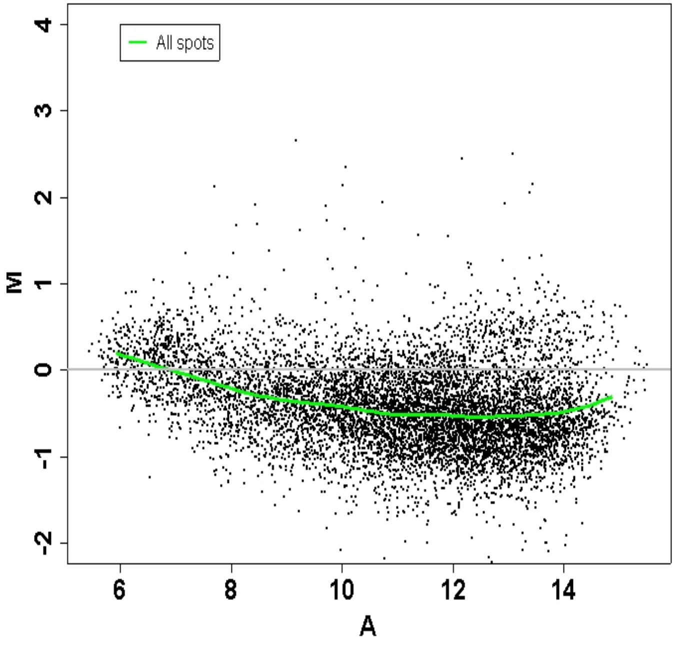
Spot

MT and WT intensity for each probe

Wild Type (WT)

M

$\log_2$ (MT/WT)

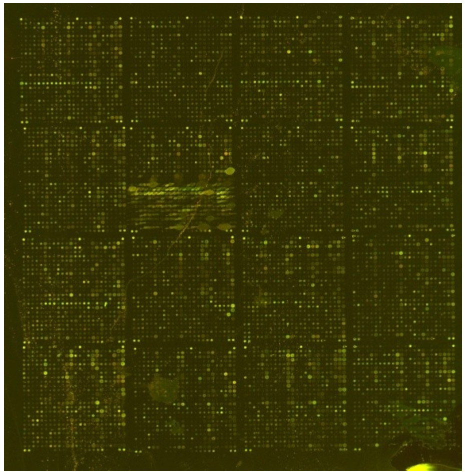


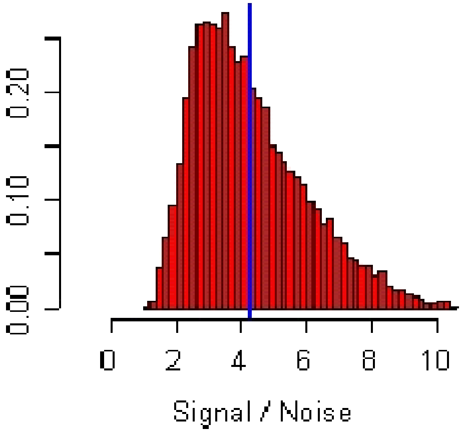
A

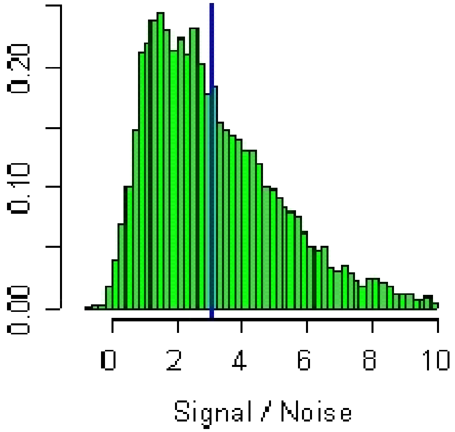
$\log_2$(MT*WT) / 2

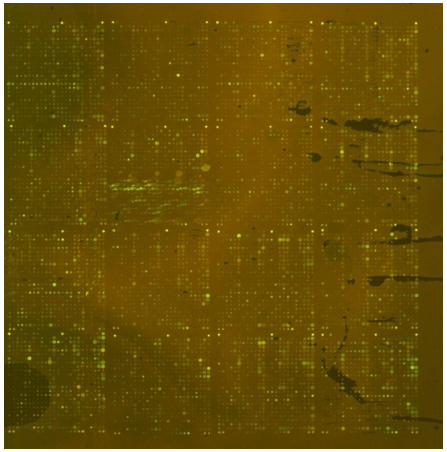
(signal strength)

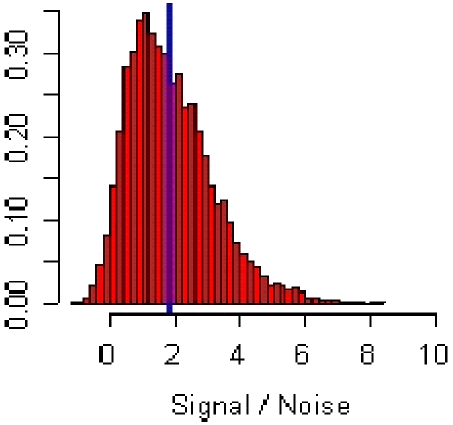
Signal/Noise histograms

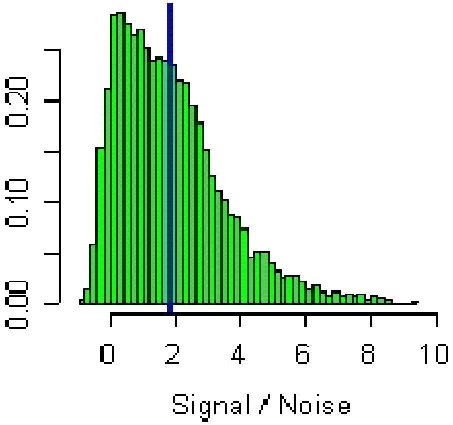






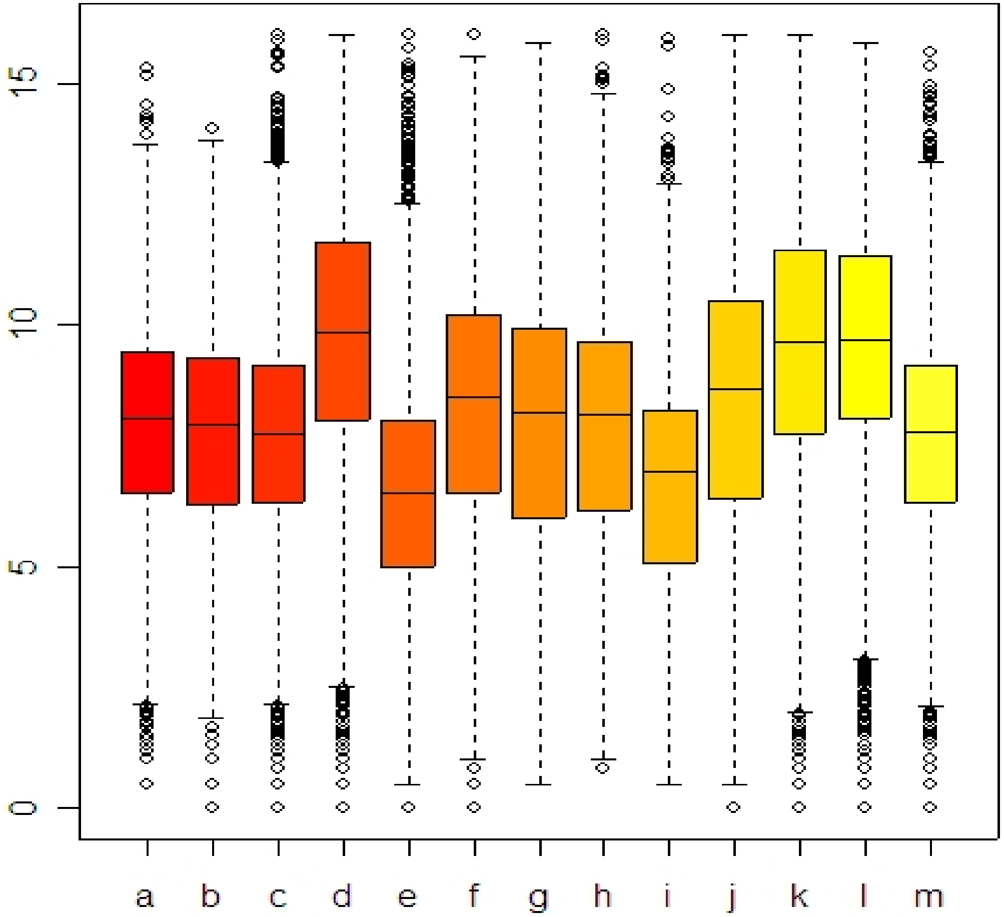






Images with high background tend to have lower $\log_2$(signal/noise) ratios

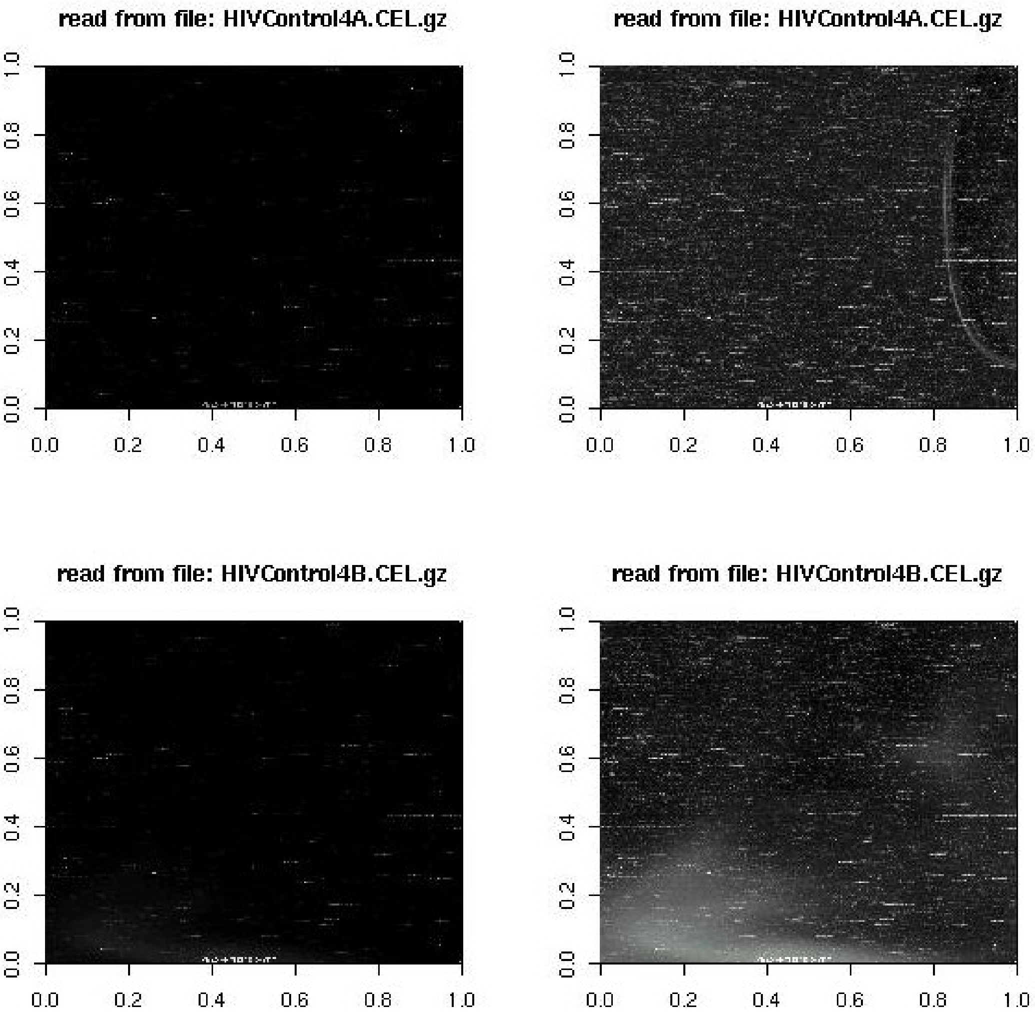
Quality between slides



Diagnostic plots for affy chips

Image plots for affymetrix chips

GOOD


read from file: HIVControl4A.CEL.gz

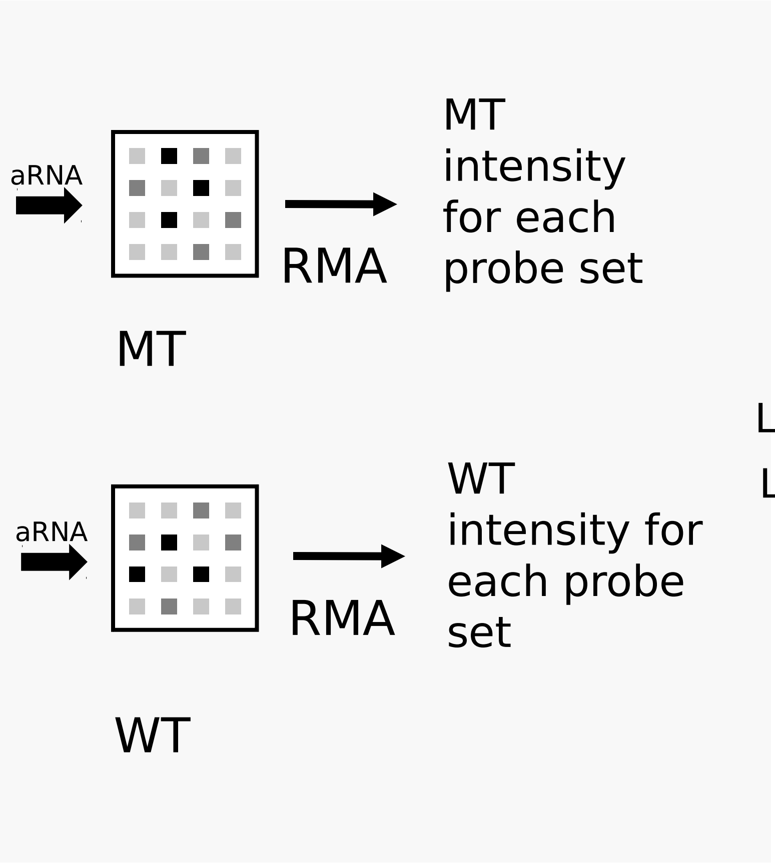
read from file: HIVControl4A.CEL.gz

read from file: HIVControl4B.CEL.gz

read from file: HIVControl4B.CEL.gz

BAD

MA-plot for GeneChip arrays (1 color)



aRNA

RMA

MT intensity for each probe set

MT

aRNA

RMA

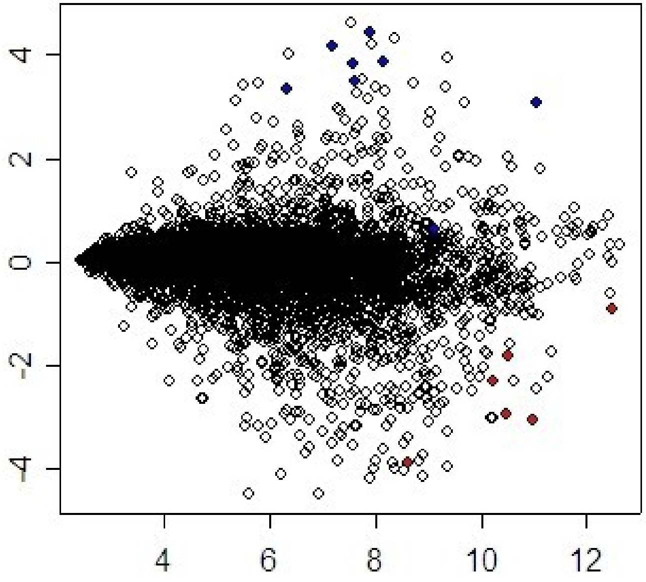
WT intensity for each probe set

WT

M

$\text{Log}_2(\text{MT})\text{-}$

$\text{Log}_2(\text{WT})$



A

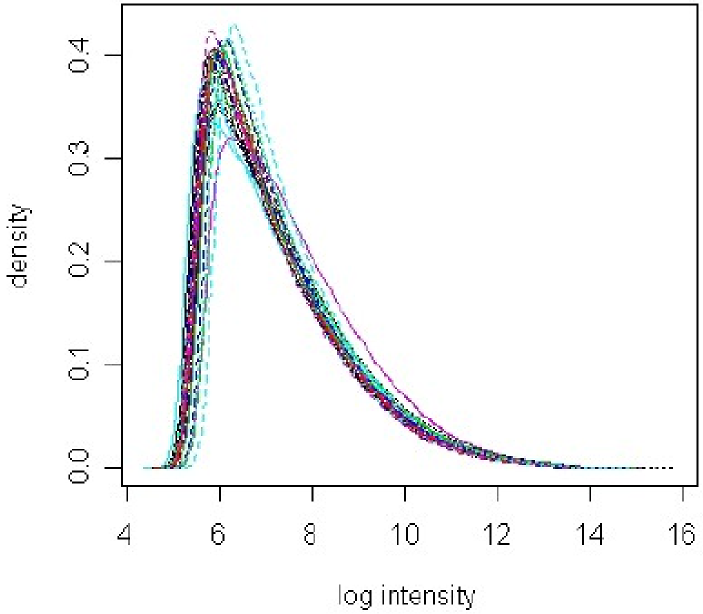
$\text{Log}_2(\text{MT*WT}) / 2$

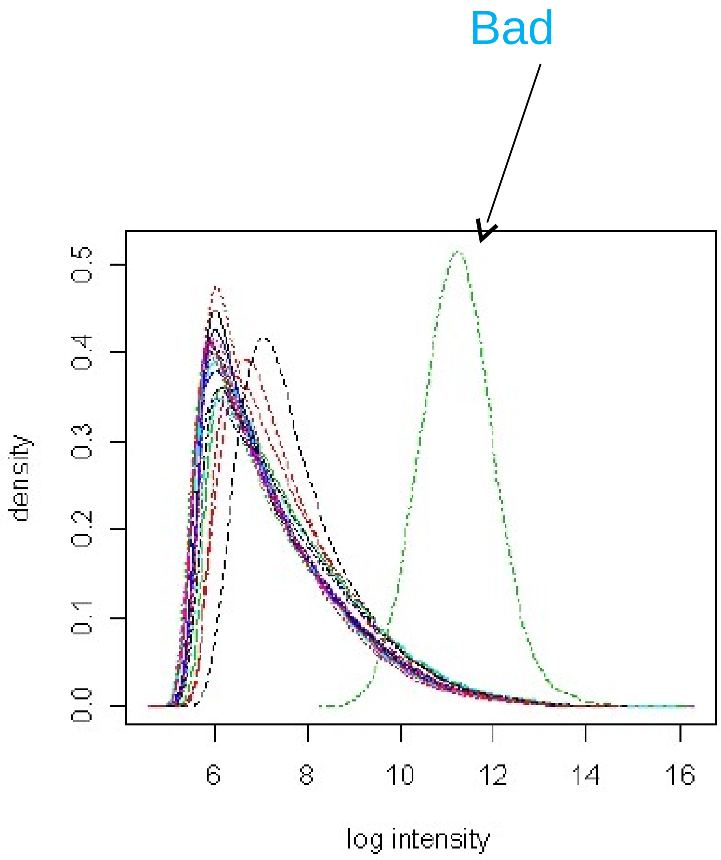
(signal strength)

Density plots (1)

- Density plots of probe intensities are useful to visualize differences in distribution arrays
- Some hints for the interpretation:
  - Density is skewed to right if genes have high expression values
  - If the shape of the bell has a sharp central point and fat tails there are genes with high expression values
  - If the shape of the bell has a flat central point and skinny tails there are few genes with high expression values

Density plots (2)

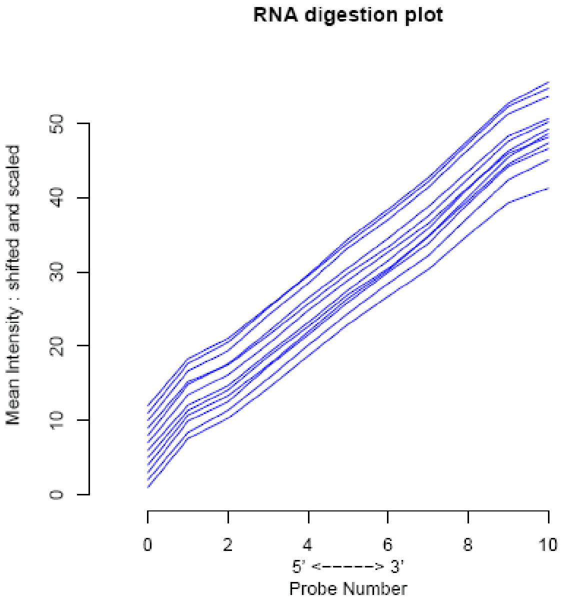


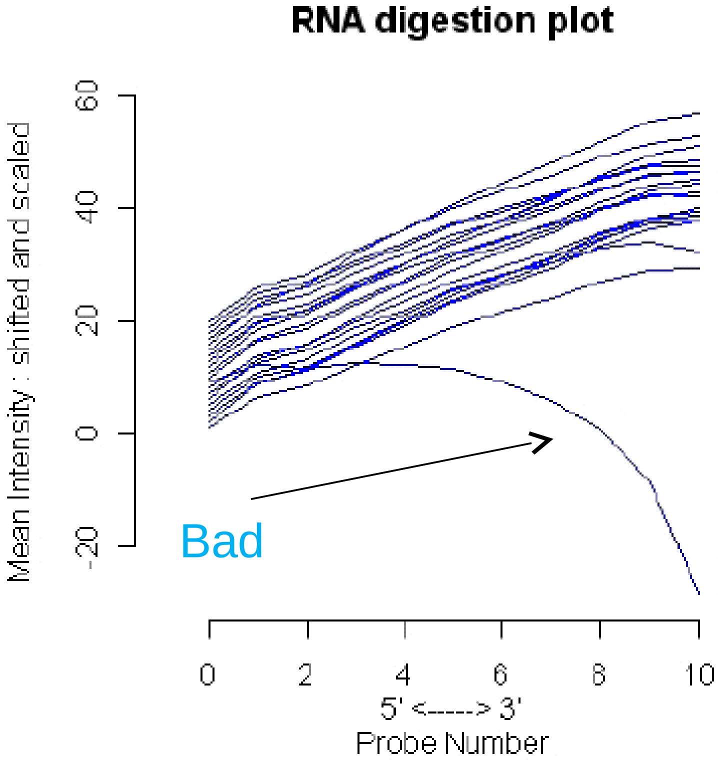


Digestion or (degradation) plots (1)

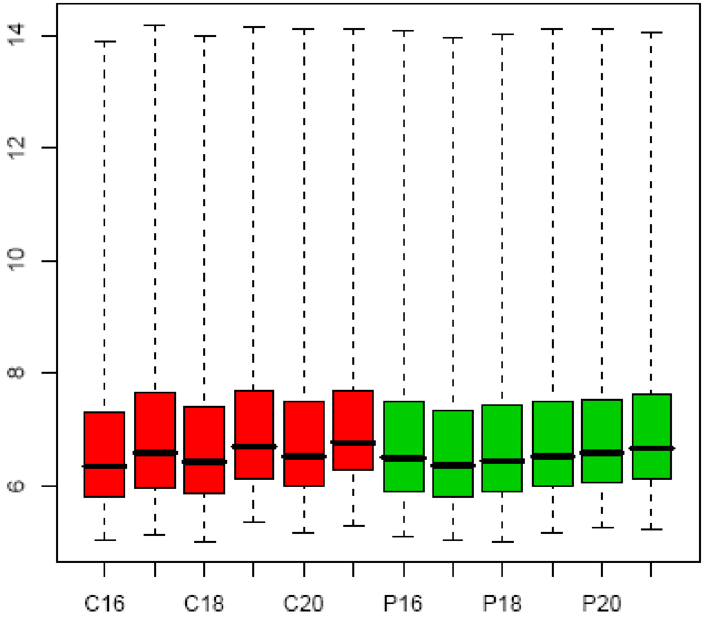
- RNA degradation plots show the mRNA average expression from 5' to 3'
- They allow to asses mRNA quality from biological samples that has been used to perform arrays
- Each curve represents a single array
- Ideally, curves should be flatted as much as possible

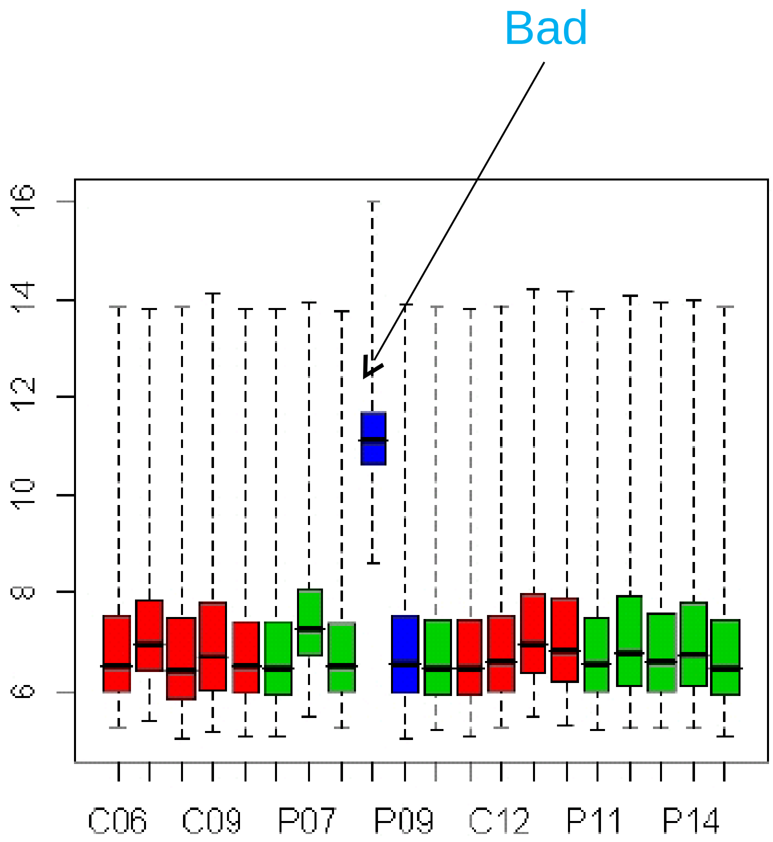
Digestion or (degradation) plots (2)





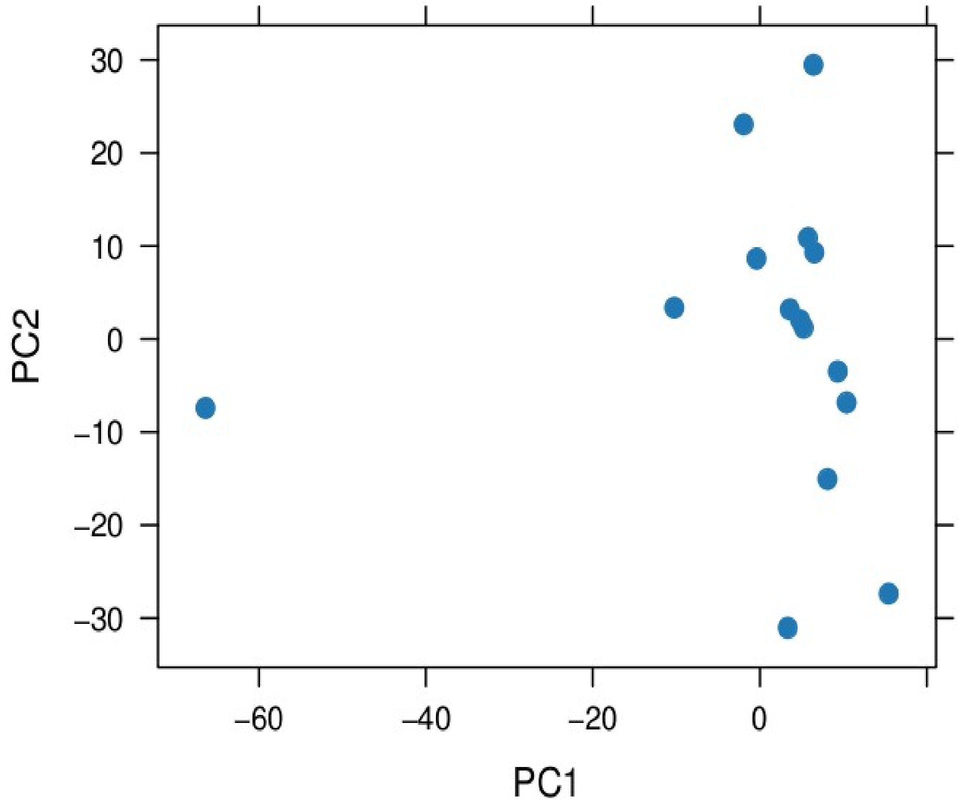
Box plots





This plot can be used in both one and two-color arrays

Principal components (PCA) plot



This plot can be used in both one and two-color arrays