From the microarray to the data frame

C. Ambroise

Laboratoire Statistique et Génome
UMR CNRS 8071

Automne 2010

Plan

1. Image analysis

2. Normalisation
   - Background noise
   - Normalisation by histogram specification

Introduction

Two steps before testing for genes which are expressed differently

1. image analysis (measuring the intensity fo the spots)
2. normalisation and descriptive statistics

Plan

Principles I

Image analysis goes throuhg a few steps

1. digitalisation : from microarray to pixels (laser)
2. finding the spots (use two type of information)
   - spatial information
   - intensity
3. Quantification

Principles II

1. Quantification for a cDNA microarray
   - sum over all the pixel intensities of the spot
   - mean of the pixel intensities
   - median
   - mode
2. Quantification for Affymetrix :
   - Principle : for each gene 20 probes of 25 nucleotide
   - 10 PM (Perfect Match)
   - 10 MM (MisMatch, perfect match with one mutated nucleotide),
   - The mean of the difference between PM and MM is the basic output

Quantification I

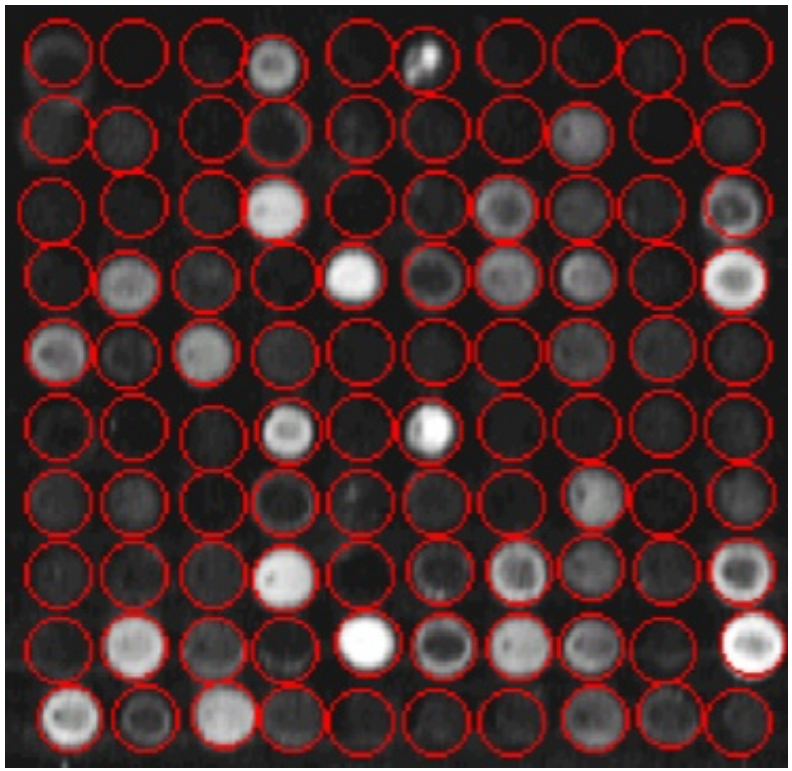

Figure: Segmentation of a microarray from Kevin R. Coombes and Keith A. Baggerly

Quantification II

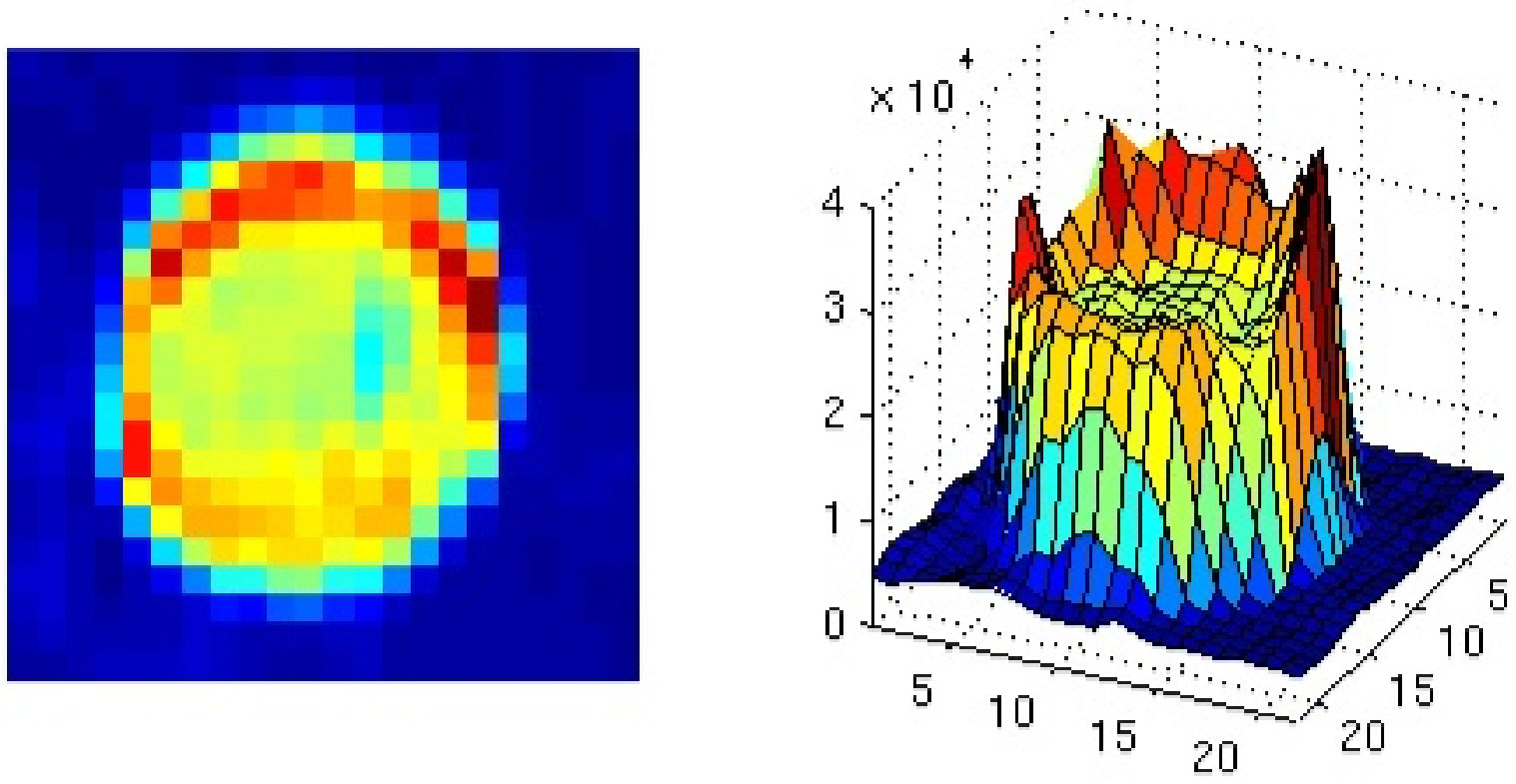

Figure: Un spot localized from Kevin R. Coombes and Keith A. Baggerly

Outcome

- The background noise is also measured. It can be estimated using all the pixel not classified as belonging to the spot
- The image analysis produce an output file with a huge amount of information summarized by 2 data frames :
  - $G = (G_{ij})_{i=1\ldots p;j=1\ldots n}$ measured intensities in Cy3 (conventionally the control sample) for p spots and n microarray
  - $R = (R_{ij})_{i=1\ldots p;j=1\ldots n}$ measured intensities in Cy5

Plan

Introduction

Objective

Transform the initial tables to get a working data set

- Detetect outliers, problems using descriptive statititicstechniques descriptives
- Correction technical artefacts

Hypothesis

- most gene do behave in the same way across conditions. The mean of the difference should be zero.
- Systematic artefacts are not related to an interesting biologically meaningful information (otherwise normalizing is equivalent to eliminating useful information !).

Background noise correction

- pixel intensity in the background zone is a random variate with mean μ_b
- intensity of each pixel in the signal zone is a random variate with mean $\mu_f = \mu_t + \mu_b \ \mu_t \geq 0$.
- Simple estimation $\hat{\mu}_t = X_f - X_b$

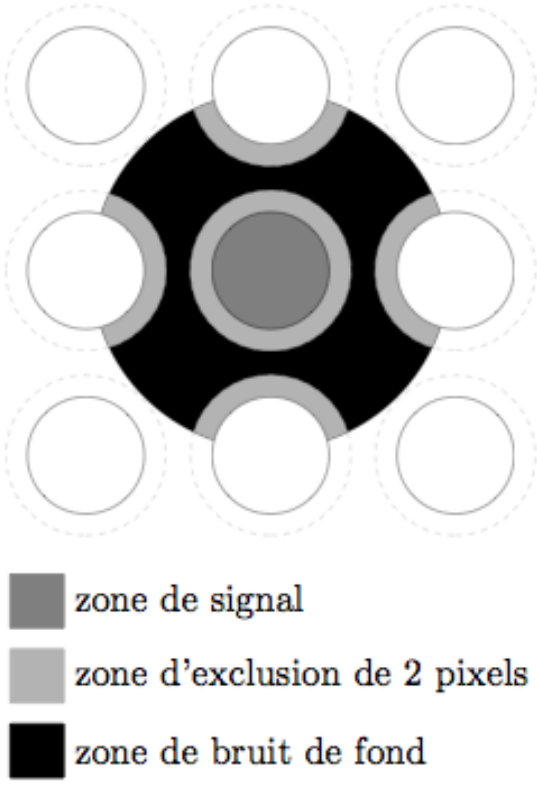

Figure: Zone d'estimation du bruit de fond (d'après la thèse de Julie Peyre)

Taking into account the noise

- Signal can become negative !
- Increase of the variability of the signal
- A simple explanation ($var(X_f - \bar{X_b}) = var(X_f) + var(\bar{X_b})$ if signal and noise are independent .
- When raw data is available, be careful with the noise.

Bias detection using Analysis of Variance

- Objective : Quantify the different bias and choose the one which are to be taken into account
- Problem : need of a minimum of **repetition**

Example

Main effect gene, slide, fluorochrome et condition and limiting to interaction of order 3 :

$$
\begin{aligned}
X_{glft} = \mu \quad & +\alpha_g + \beta_l + \gamma_f + \delta_t && \text{principal effects} \\
& +(\alpha\beta)_{gl} + (\alpha\gamma)_{gf} + (\alpha\delta)_{gt} && \text{interactions of order two with t} \\
& +(\beta\gamma)_{lf} + (\beta\delta)_{lt} + (\gamma\delta)_{ft} && \text{other order two interaction with} \\
& +(\alpha\beta\gamma)_{glf} + \ldots && \text{order 3 interactions} \\
& +E_{glft}
\end{aligned}
$$

logarithmic transform

Absolute or relative

- Which genes are differently expressed between different conditions ?)
- ratio of the expression level is more informative than the expression level themselves

Logarithm

1. Heavy right tail densities
2. $\log_2$ transform is often useful with ratio

$$\log_2 2 = 1$$
$$\log_2 1 = 0$$
$$\log_2 \frac{1}{2} = -1$$

Example of log transform

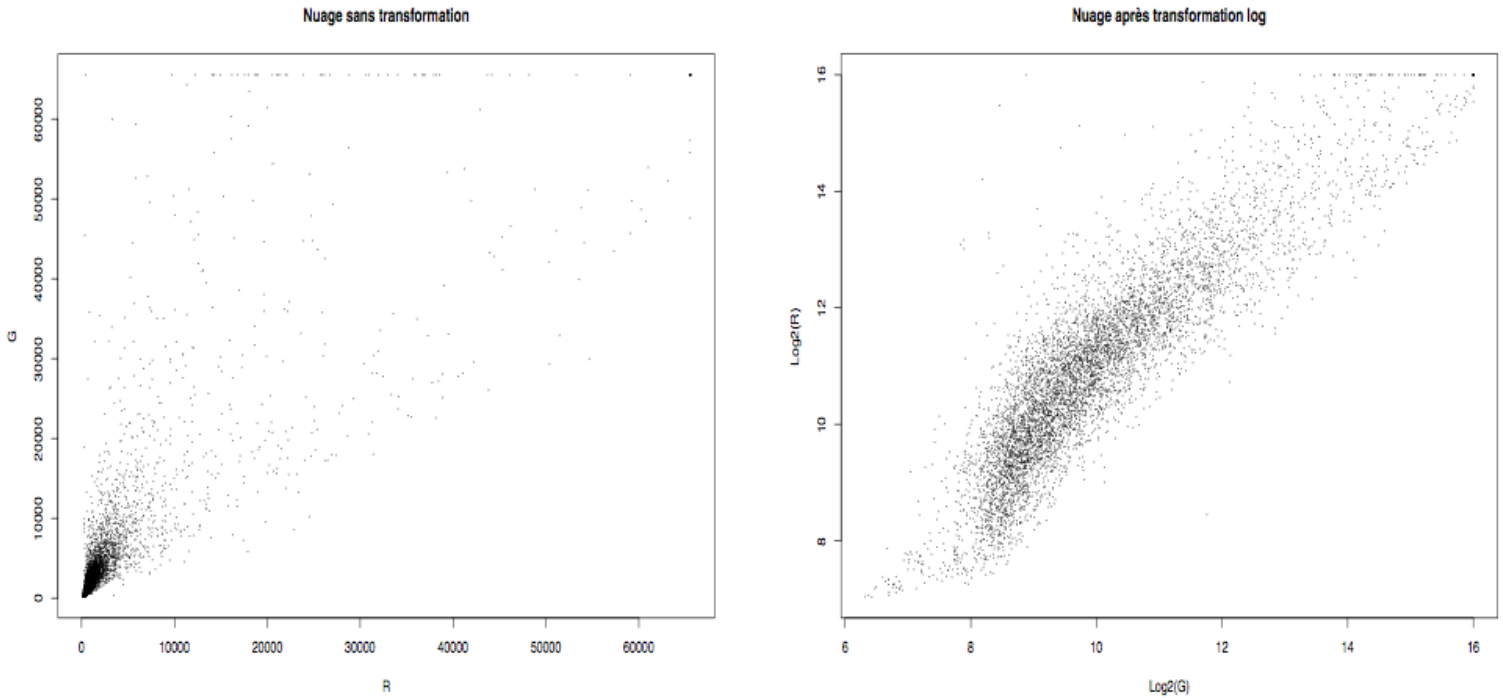

Figure: Cloud of red intensities before and after log transform from Julie Peyre PhD thesis

Global Normalisation

Principle

- Hypothesis green and red intensities are proportional, $R = kG$.
- Centering all the log ratio by determining the factor k :

$$\log_2 R/G \rightarrow \log_2 \frac{R}{G} - c = log_2 \frac{R}{kG}$$

A current choice for c is the median of the log-ratio.
It is different to

1. log transform and center
2. center and then log-transform

Ratios and intensities : MA plot

$$M = \log R - \log G \quad \text{et} \quad A = \frac{\log R + \log G}{2}$$

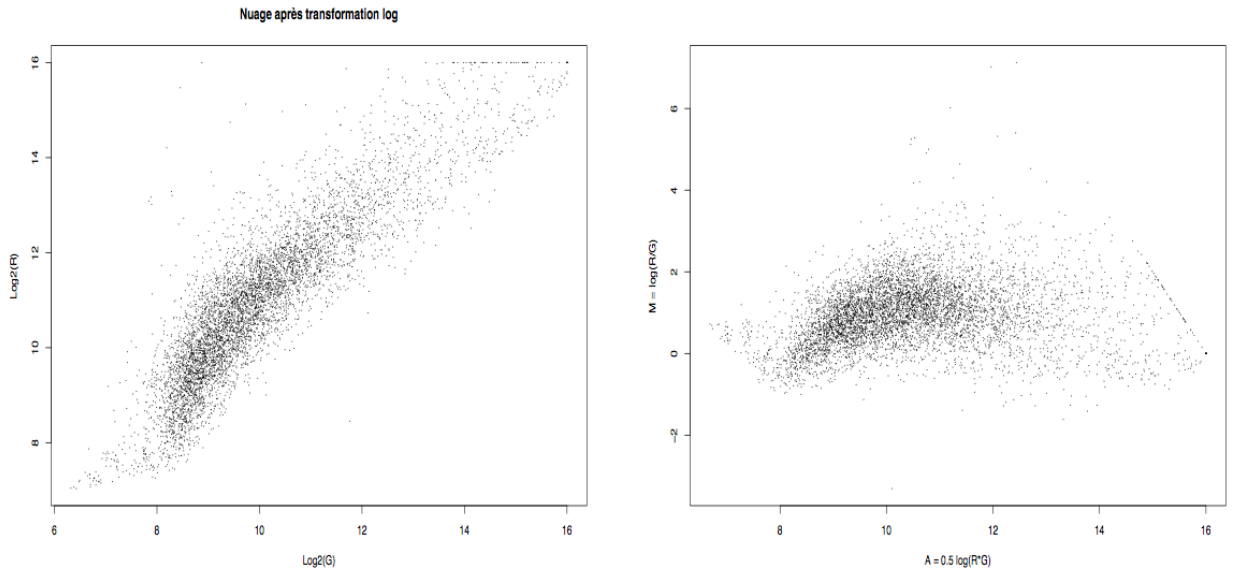

Figure: Cloud after log transform and MA plot from Julie Peyre PhD

MA plot and median normalisation

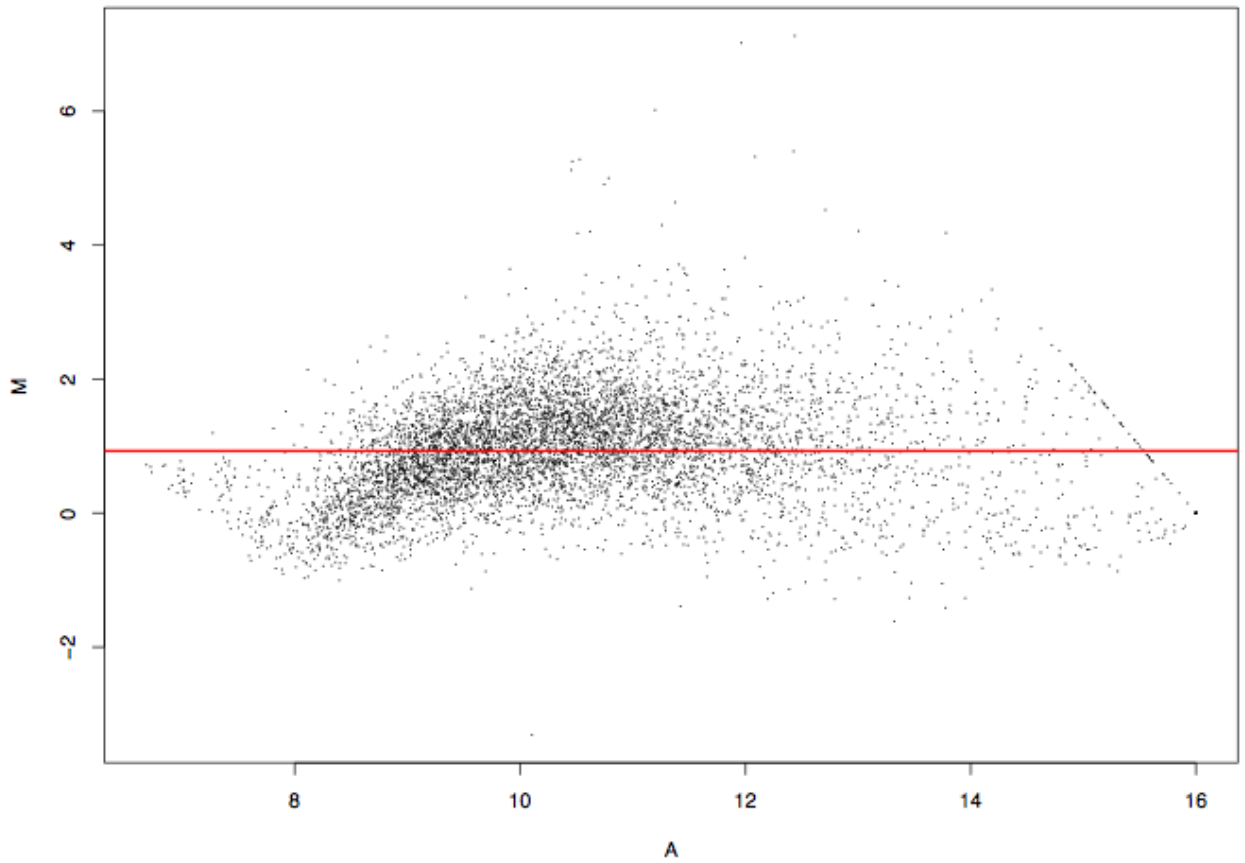

Figure: MA plot and median normalisation)

Normalisation by loess or lowess (Cleveland 1979) I

- The M against A plot for all gene should be centered around the x-axis.
- Observation of a deformation
- Correction :

$$M = c(A) + \epsilon$$

where ϵ is a gaussian noise and c the correction

Normalisation by loess or lowess (Cleveland 1979) II

- The function c is traditionally estimated by loess regression
  - Local approximation of $c()$ by a weigthed regression $M_g = a_g + b_g A$
  - Estimation of coefficients a_g etb_g by least weighted least square from the observation in the neighborhood

$$\sum_{g'} w(g, g')(M_{g'} - a_g + b_g A_{g'})^2$$

  - Weight of an observation is inversly proportional to the distance to the center. $w(g, g')$.
  - For each gene g, $\widehat{c(A_g)}$, the corrected normalised log-ratio $\widehat{M_g}$ de M_g is :

$$\widehat{M_g} = M_g - \widehat{c(A_g)}$$

Loess

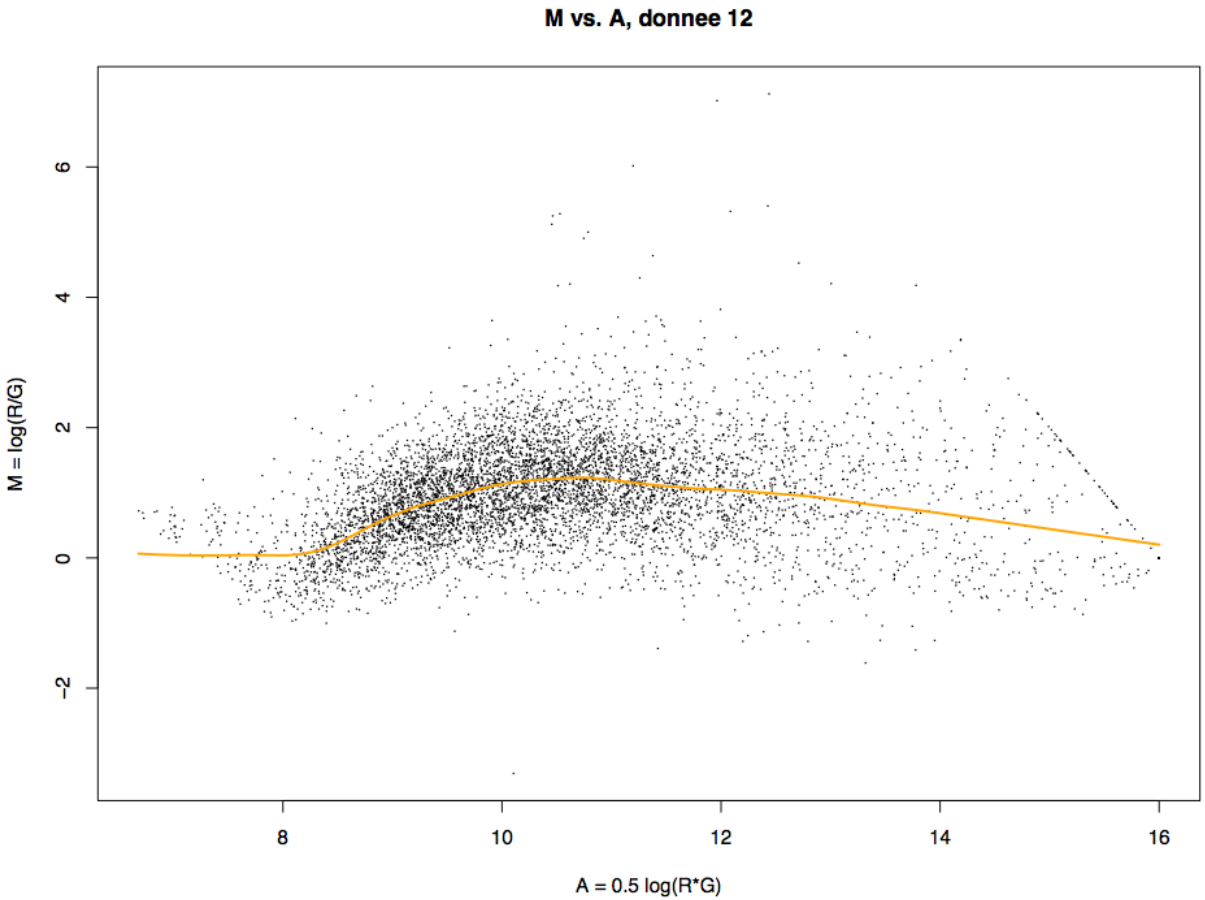

Figure: MA plot and loess normalisation (from Julie Peyre)

Improving over the global Loess

- Problem : Each bloc in a microarry is printed using a different printing head .
- Does it imply a difference between blocs ?
- Solution : Compute a normalisation per bloc.

Scale Normalisation by MAD I

- A scale transformation allows to put all the variables (microarray) on the same scale and thus allows the comparison. It comes after a trend transformation
- Sometimes called " Z transform " by Anglo-Saxons, it produces a zero mean, unit variance variable

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

  - Exemple d'une v.a. normale.
  - Effet sur les lignes, effet sur les colonnes.

- Microarray have some spots with extrem values, which produces poor variance estimation
- *median absolute deviation*, ou MAD :

Scale Normalisation by MAD II

- List of values $x_1, \ldots, x_n$, the MAD is defined as
-

$$m = \text{médiane}(x_1, \ldots, x_n)$$
$$MAD = \text{médiane}(|x_1 - m|, \ldots, |x_n - m|)$$

Normalisation by Normal Score (Lin et al.)

Idea Work directly with fluorescence after normalizing by histogramm specification

Procedure

- $NS(R_i) = \Phi^{-1}(\frac{\tilde{R}_i}{(p+1)})$ with $\tilde{R}_i$ rank of fluorescence R_i
- $$\begin{cases} Asco = NS(R_i) + NS(G_i) \\ Msco = NS(R_i) - NS(G_i) \end{cases}$$

Normal score

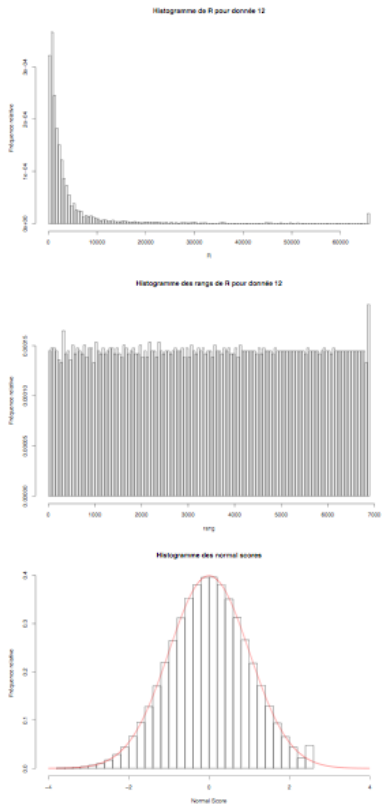

Figure: Normal score transform of the intensities (From Julie Peyre PhD)