Data Normalization and Standardization

the benefits of pre-processing microarray data

Ben Bolstad

Statistics, University of California, Berkeley

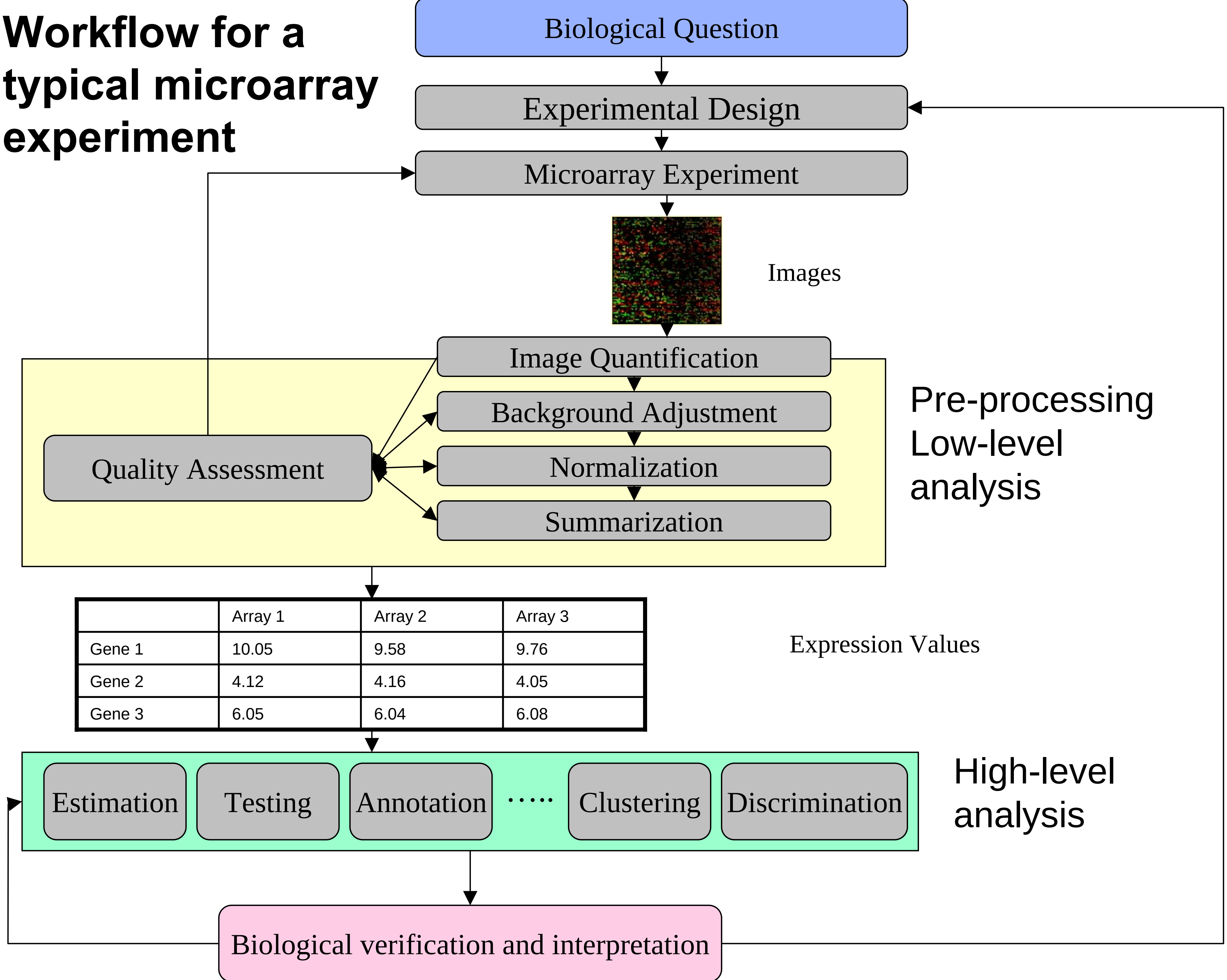
bmb@bmbolstad.com

http://bmbolstad.com

Outline

- Introduction
- Pre-processing methodologies as they relate to
  - Two channel arrays
  - Affymetrix GeneChips (a popular single channel array)

Workflow for a typical microarray experiment



Biological Question

Experimental Design

Microarray Experiment

Images

Image Quantification

Background Adjustment

Normalization

Summarization

Quality Assessment

Pre-processing
Low-level
analysis

	Array 1	Array 2	Array 3
Gene 1	10.05	9.58	9.76
Gene 2	4.12	4.16	4.05
Gene 3	6.05	6.04	6.08

Expression Values

Estimation

Testing

Annotation

·····

Clustering

Discrimination

High-level
analysis

Biological verification and interpretation

Introduction to preprocessing

- Pre-processing typically constitutes the initial (and possibly most important) step in the analysis of data from any microarray experiment
- Often ignored or treated like a black box (but it shouldn’t be)
- Consists of:
  - Data exploration
  - Background correction, normalization, summarization
  - Quality Assessment
- These are interlinked steps

Background Correction/Signal Adjustment

- A method which does some or all of the following:
  - Corrects for background noise, processing effects on the array
  - Adjusts for cross hybridization (non-specific binding)
  - Adjust estimated expression values to fall across an appropriate range

Normalization

“Non-biological factors can contribute to the variability of data ... In order to reliably compare data from multiple probe arrays, differences of non-biological origin must be minimized.“[1]

- Normalization is the process of reducing unwanted variation either within or between arrays. It may use information from multiple chips.
- Typical assumptions of most major normalization methods are (one or both of the following):
  - Only a minority of genes are expected to be differentially expressed between conditions
  - Any differential expression is as likely to be up-regulation as down-regulation (ie about as many genes going up in expression as are going down between conditions)

[1] GeneChip 3.1 Expression Analysis Algorithm Tutorial, Affymetrix technical support

A brief word on the term "Normalization"

- Many use the term "normalization" to refer to everything being discussed in this session. In other words they treat "normalization" and "pre-processing" as being synonymous with each other.
- I view normalization as just one of the steps in the process (although a very important one).

Summarization

- Reducing multiple measurements on the same gene down to a single measurement by combining in some manner.
- Most relevant to Affymetrix Arrays as we will see a little later ….

Quality Assessment

- Need to be able to differentiate between good and bad data.
- Bad data could be caused by poor hybridization, artifacts on the arrays, inconsistent sample handling, …..
- An admirable goal would be to reduce systematic differences with data analysis techniques.
- Sometimes there is no option but to completely discard an array from further analysis. How to decide …..

Two-channel arrays

Image analysis for two color arrays

- The **raw data** from a cDNA microarray experiment consist of pairs of **image files**, 16-bit TIFFs, one for each of the dyes.

- Image analysis is required to extract measures of the red and green fluorescence intensities for each spot on the array.

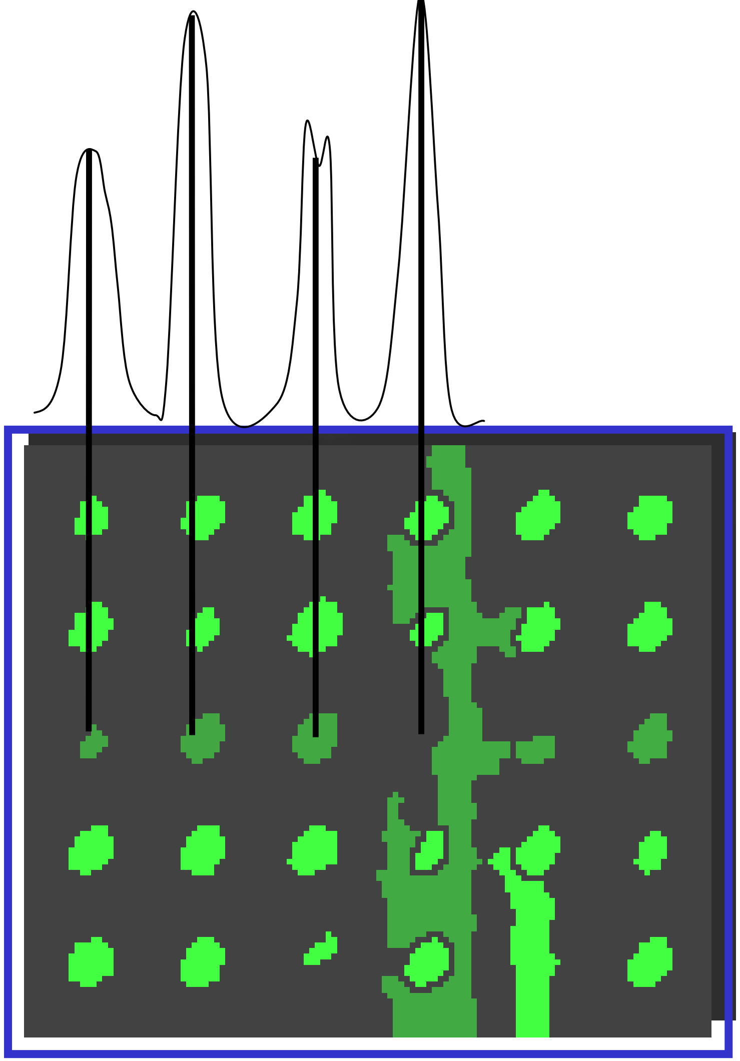
Image analysis

1. Addressing. Estimate location of spot centers.

2. Segmentation. Classify pixels as foreground (signal) or background.

3. Information extraction. For each spot on the array and each dye

- signal intensities;
- background intensities;
- quality measures.

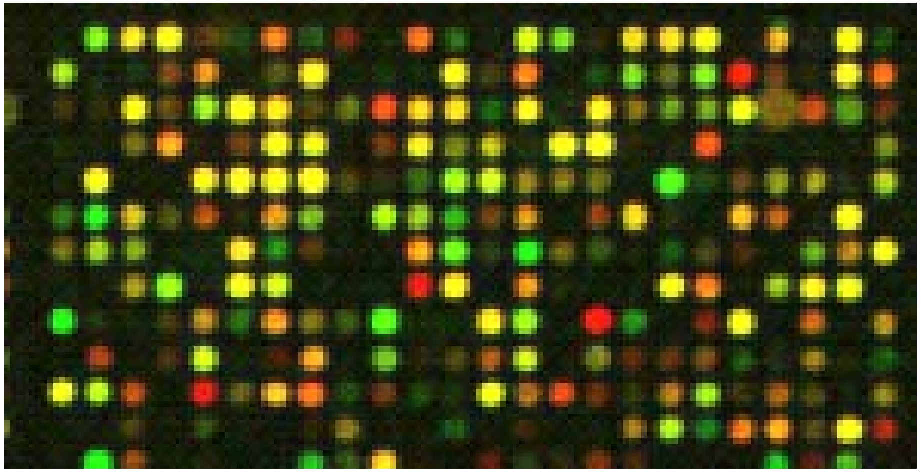


→ **R and G for each spot on the array.**

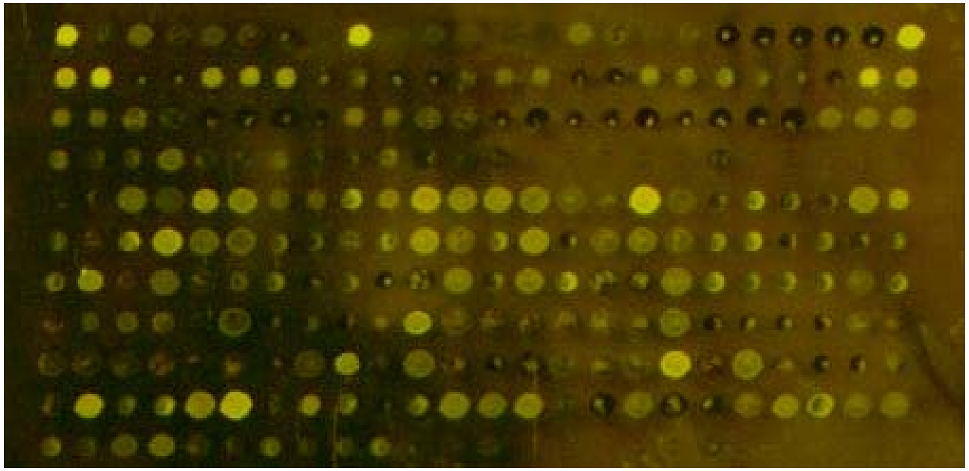
Red/Green overlay images

Co-registration and overlay offers a quick visualization, revealing information on colour balance, uniformity of hybridization, spot uniformity, background, and artifiacts such as dust or scratches

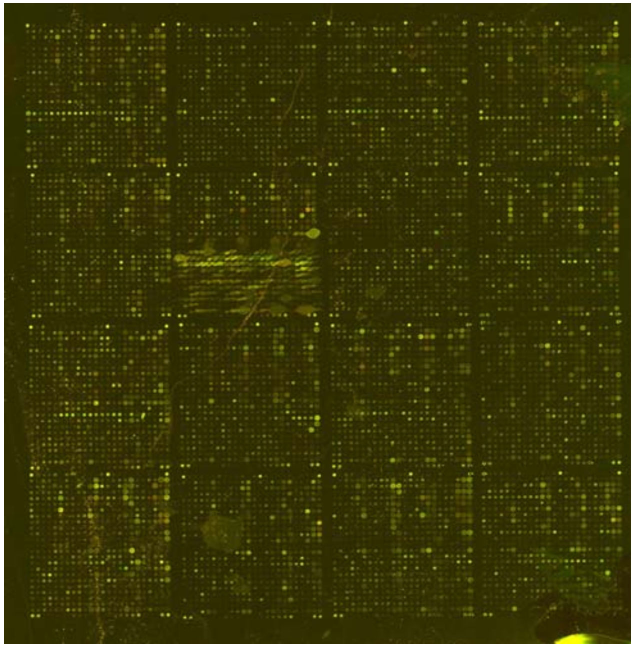
Good: low bg, lots of d.e.

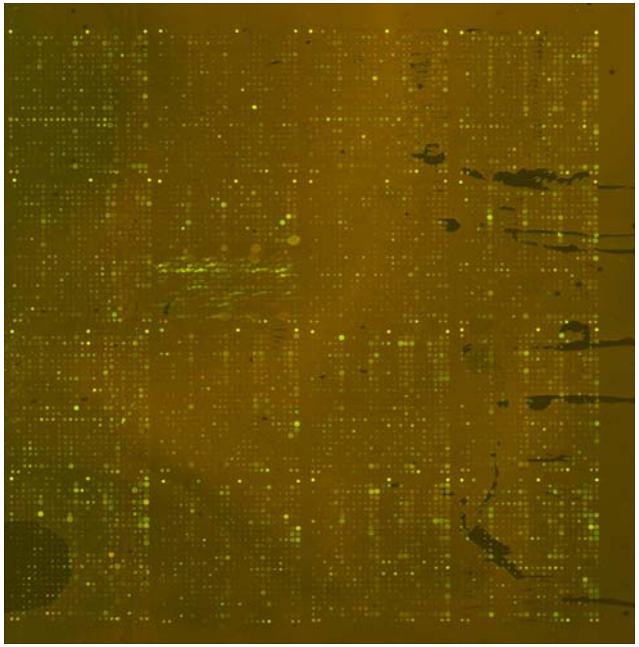


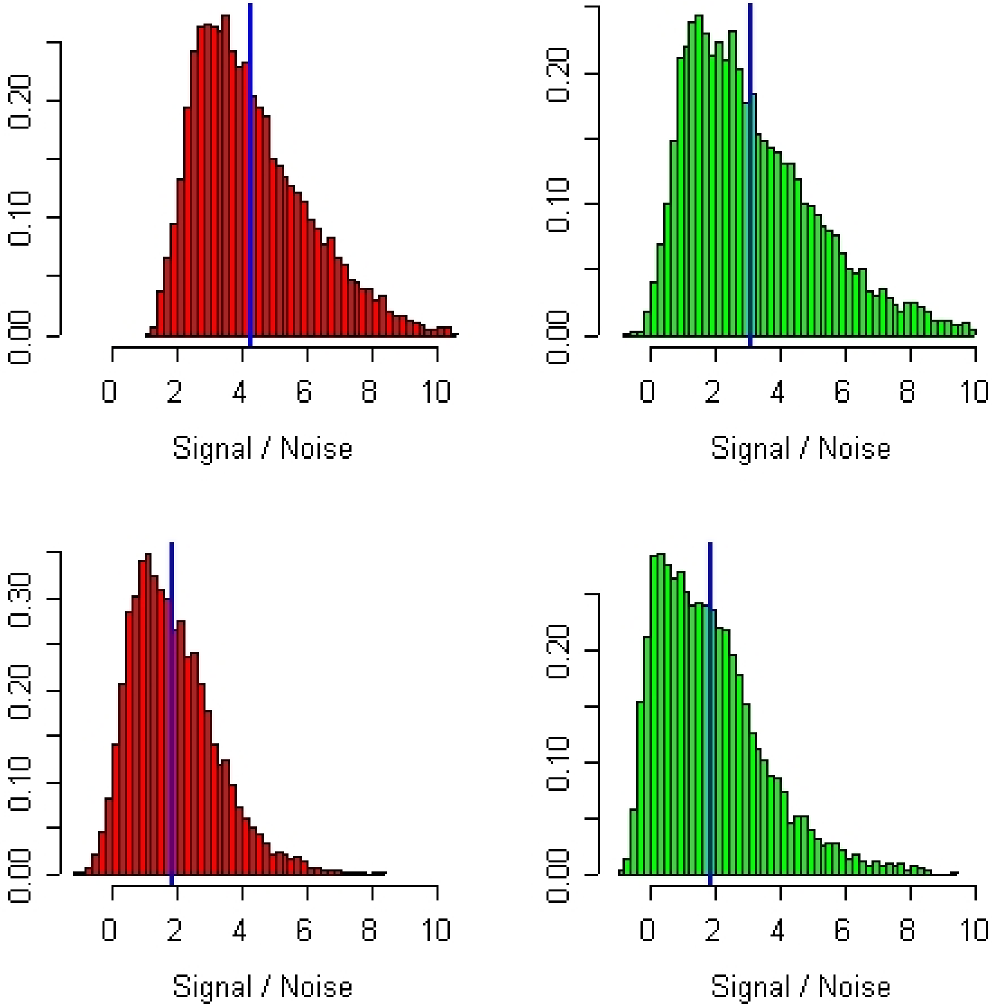
Bad: high bg, ghost spots, little d.e.



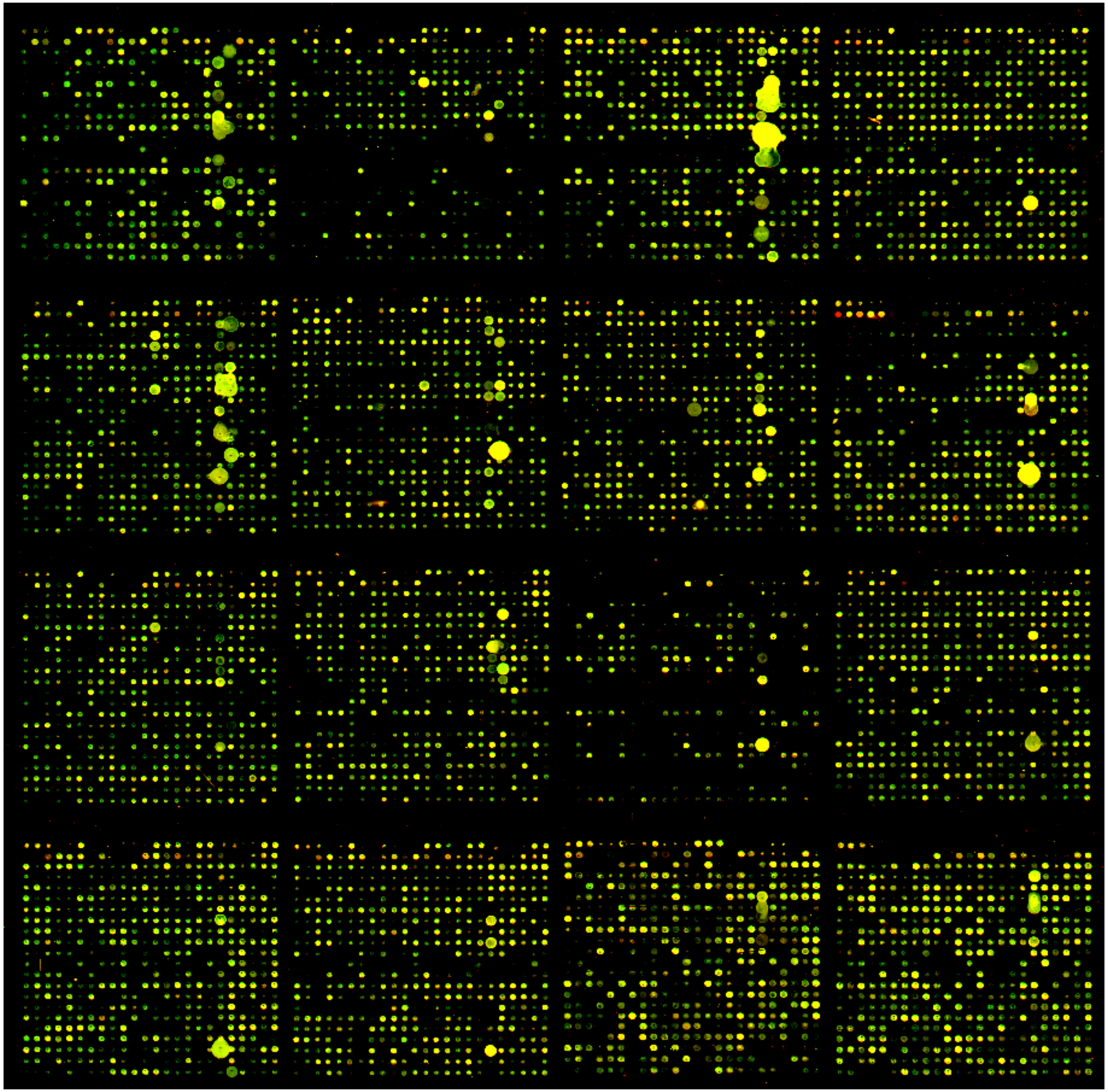
Histograms







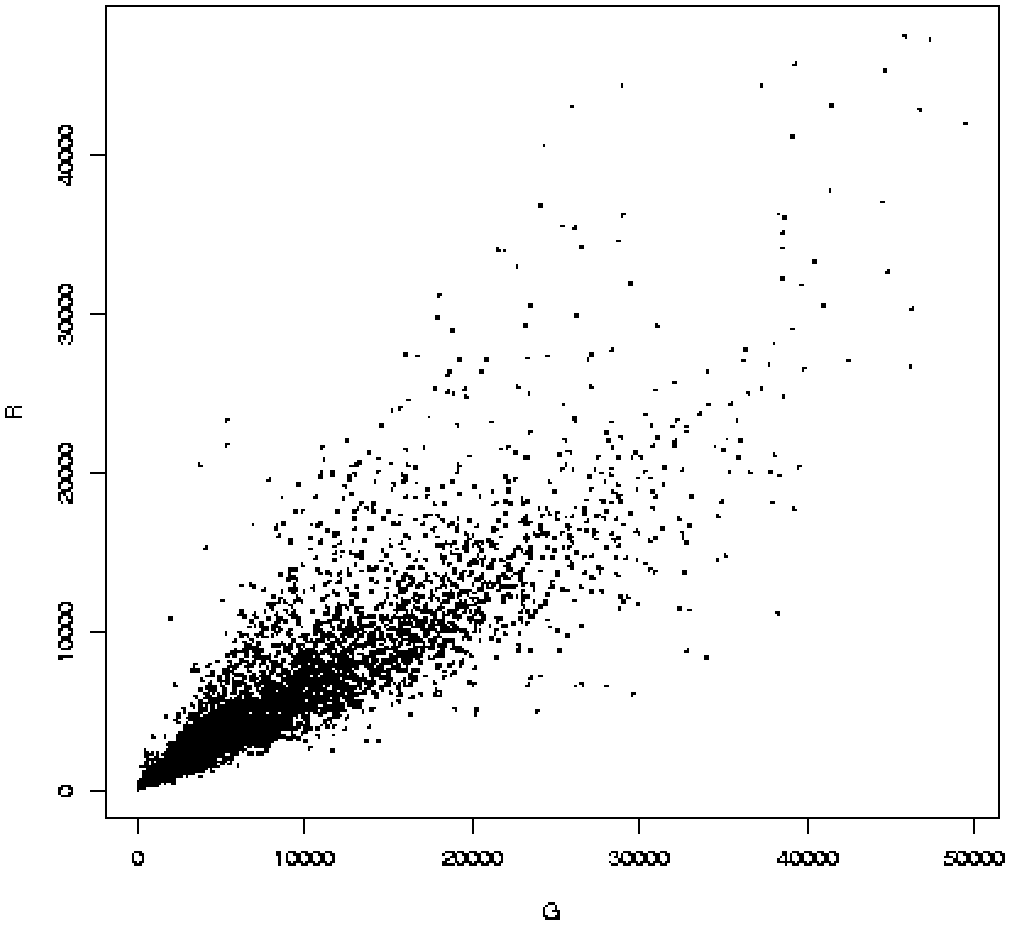
Signal/Noise = log$_2$(spot intensity/background intensity)



Slide 3 of the *swirl* data: used in all that follows.

Tools for exploring the data

Important: Always log, always rotate

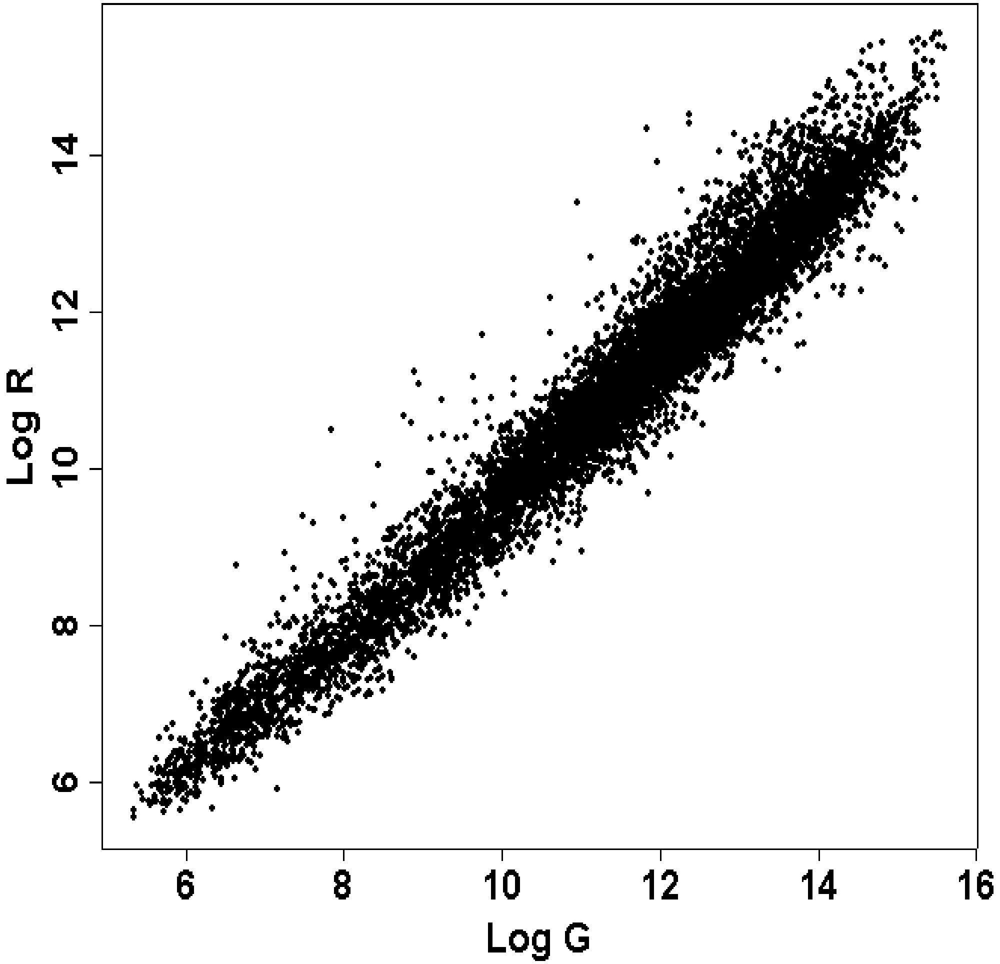


R vs G

Bad

Tools for exploring the data

Important: Always log, always rotate

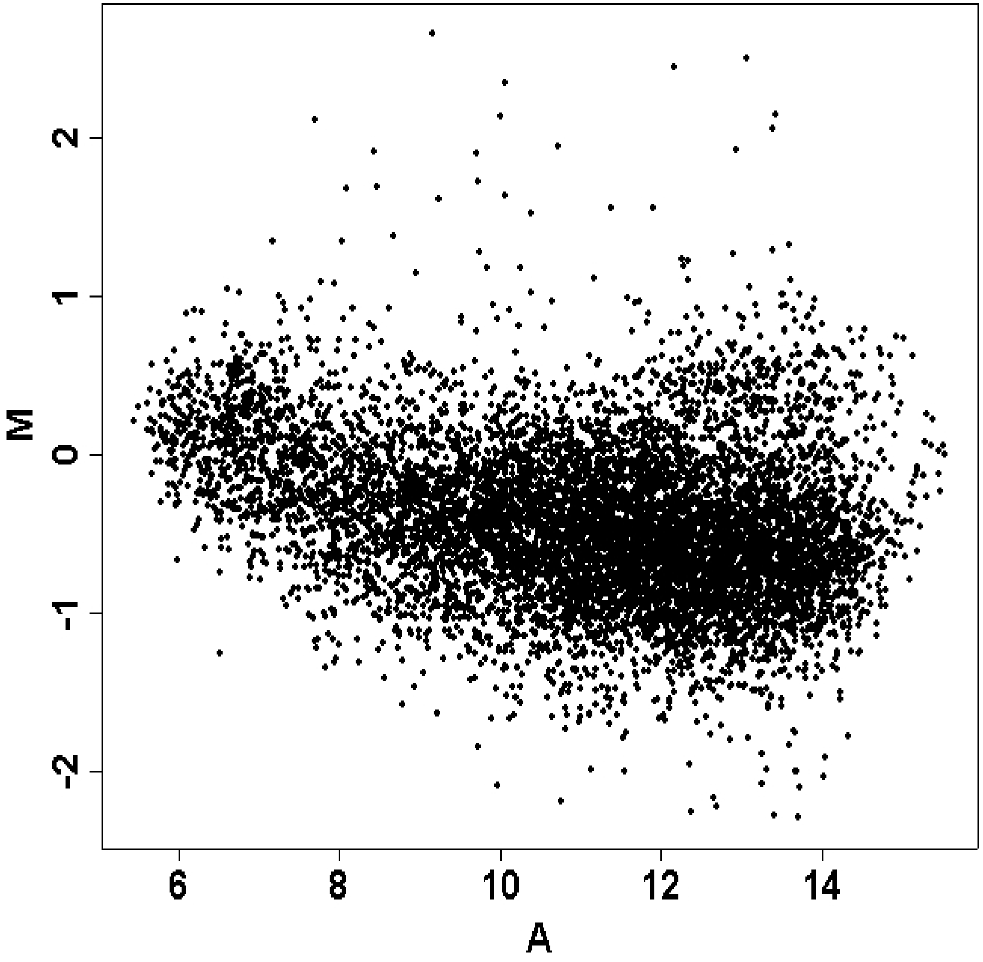


$\log_2 R$ vs $\log_2 G$

Better

Tools for exploring the data

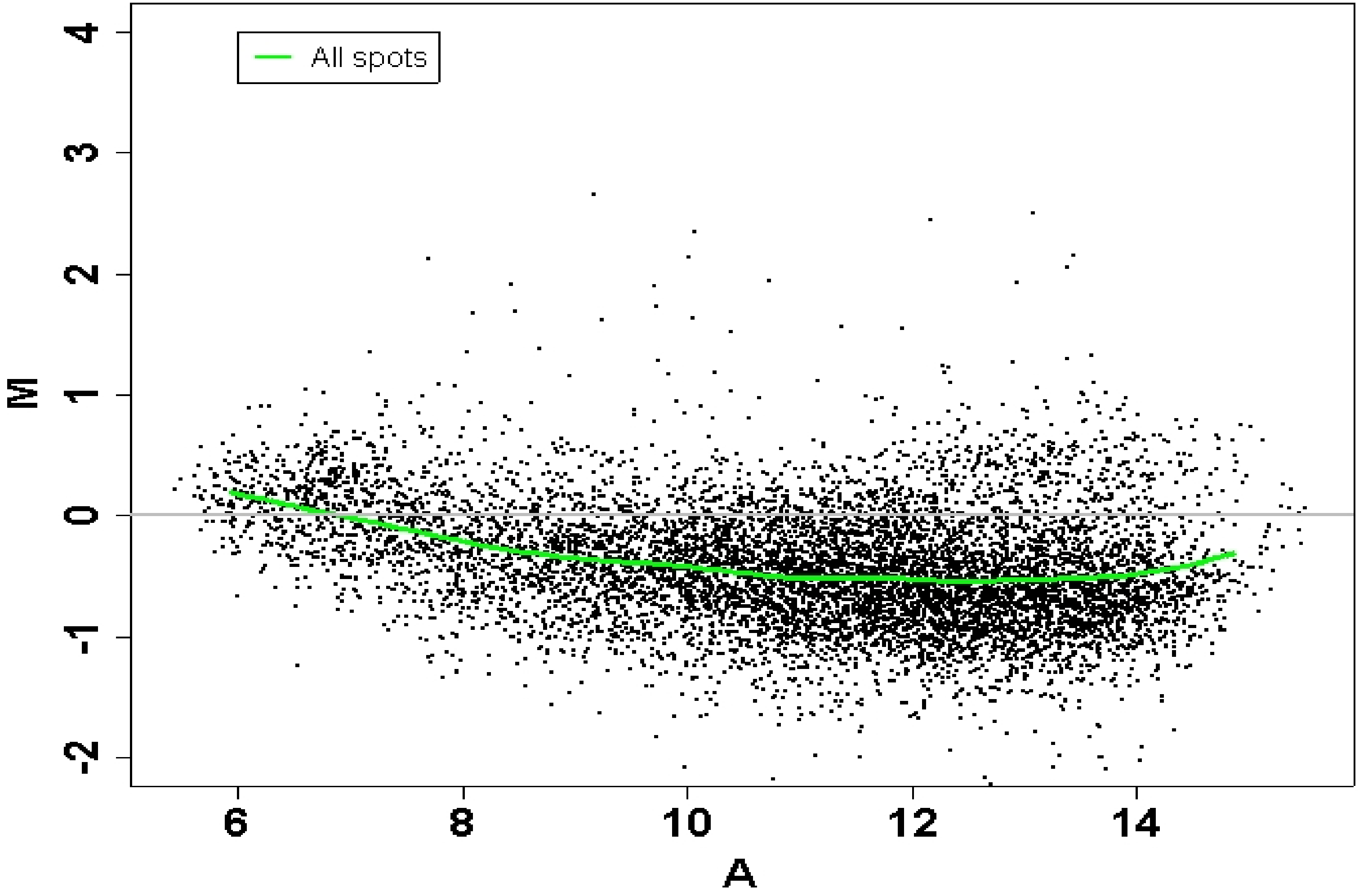
Important: Always log, always rotate



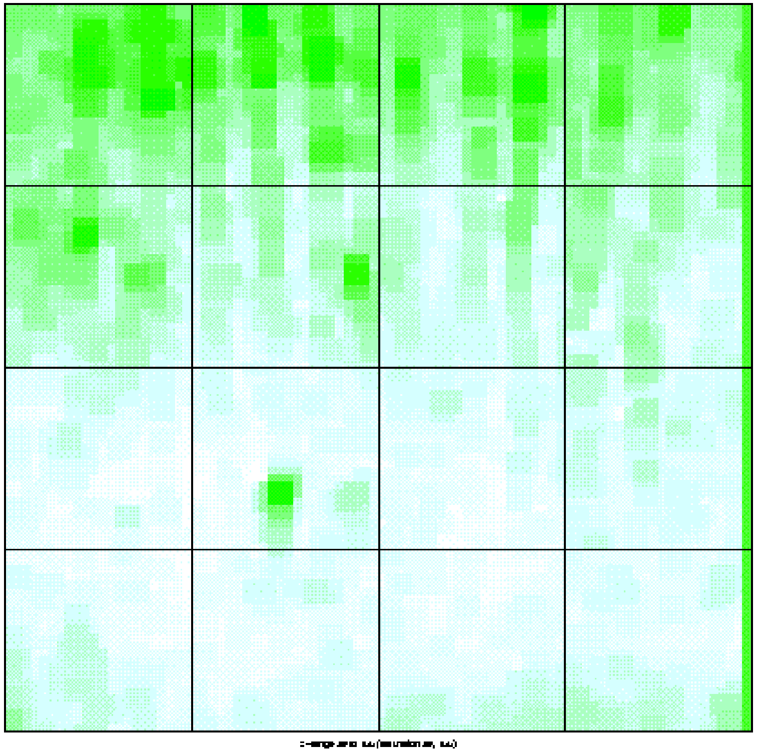
$M=\log_2 R/G$ vs $A=\log_2 \sqrt{RG}$

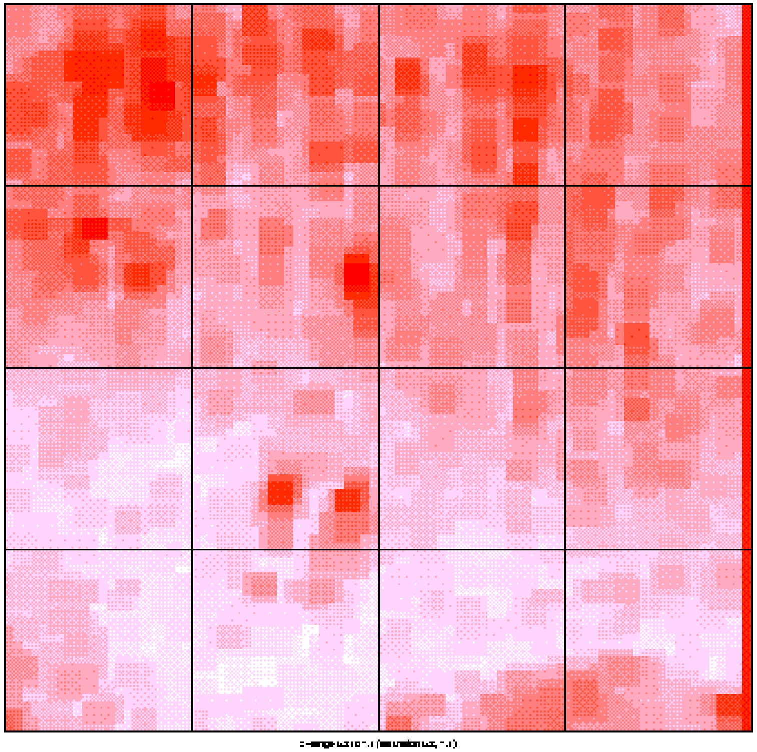
Best

MA-plot

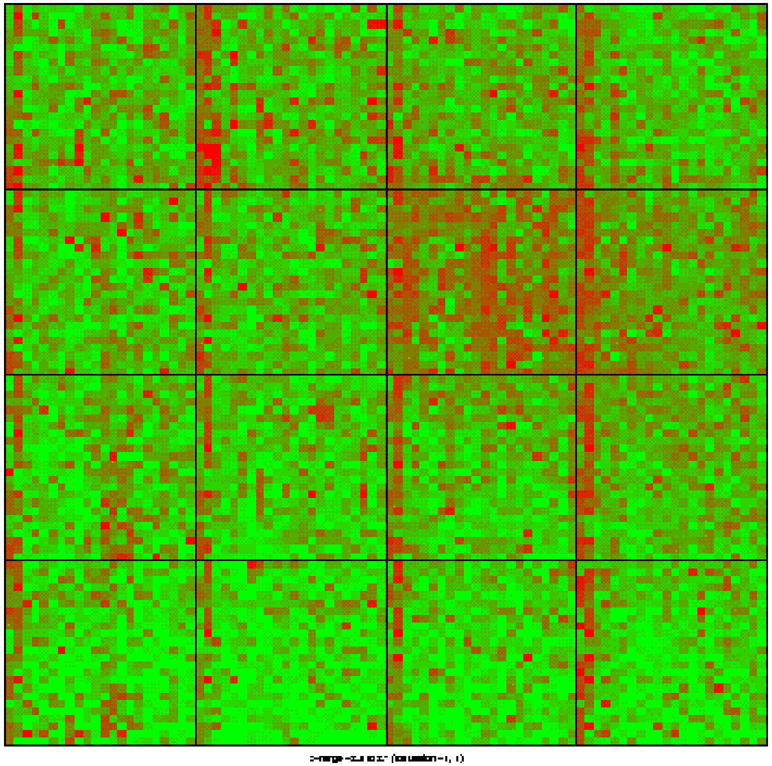


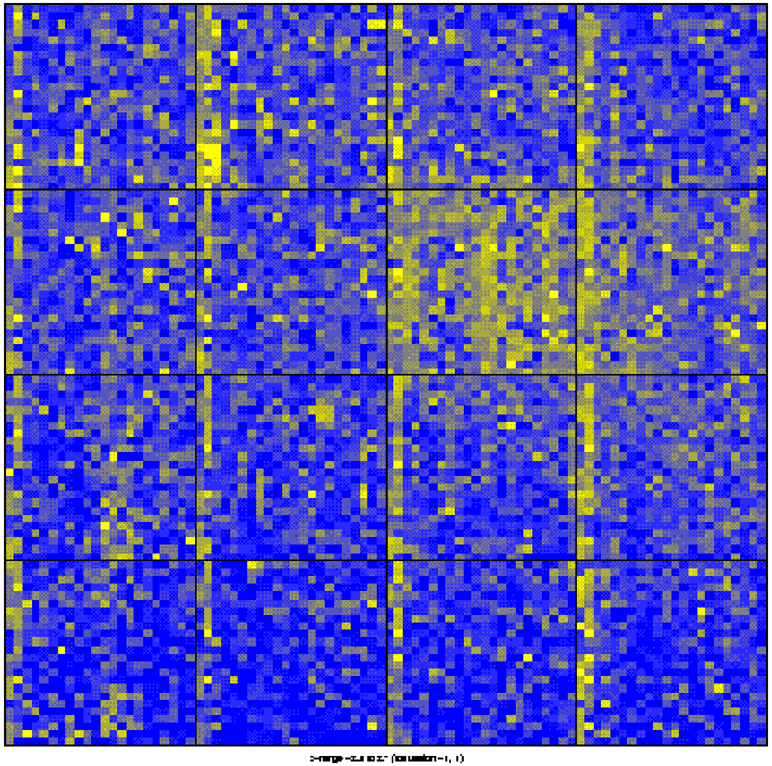
Spatial plots: background





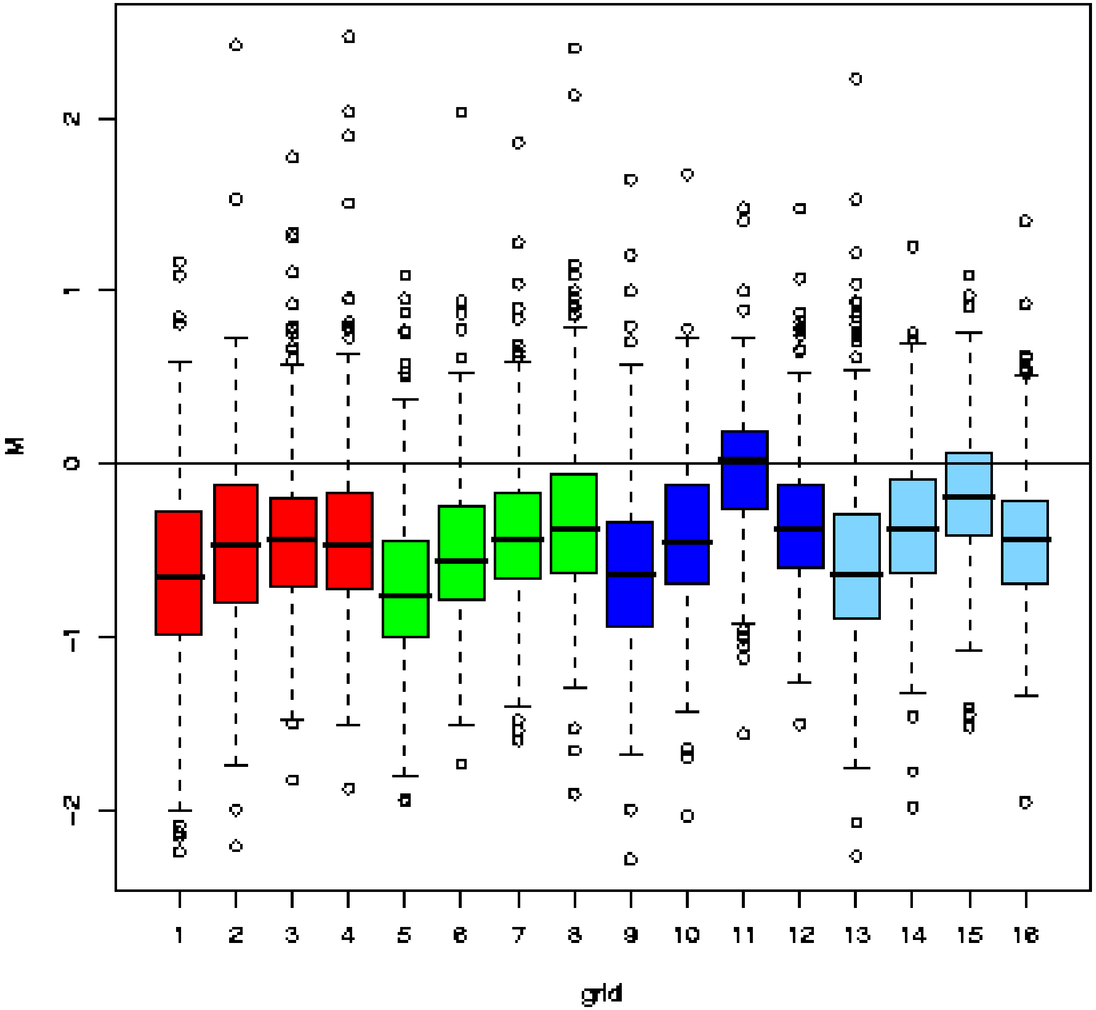
Spatial plots: log ratios (M)





No reason to constrain yourself to red/green when visualizing

Boxplots



Background correction

- Normally this is just a matter of subtracting the background value in the Red channel of the foreground Red intensity and the same for the Green channel intensities for each spot.

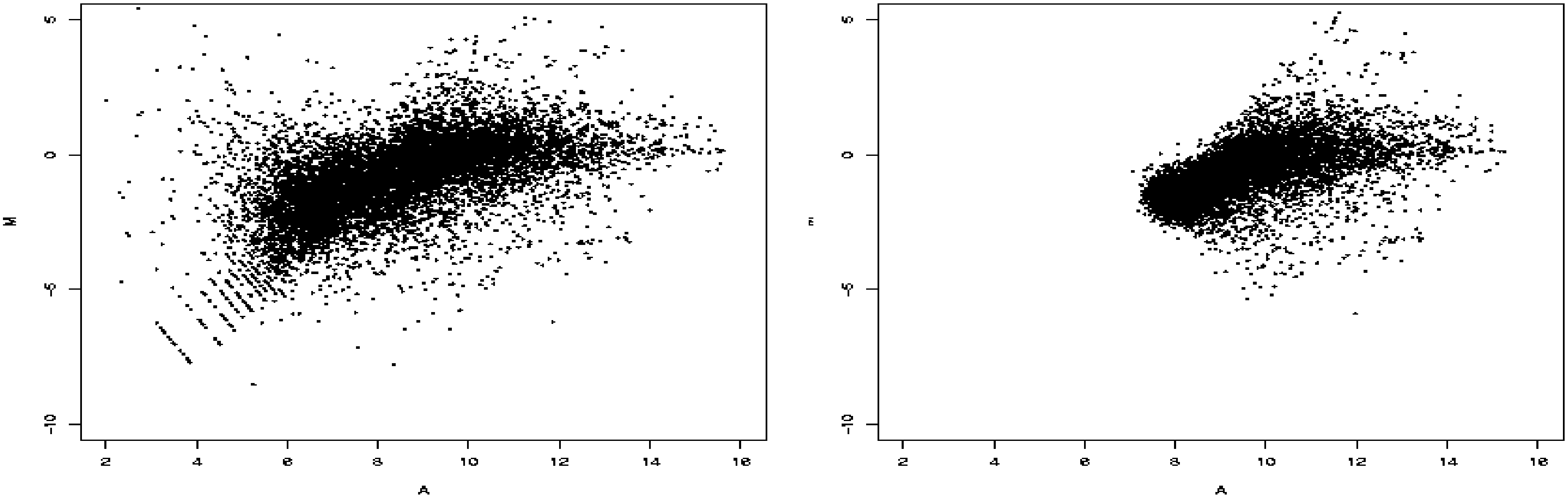
  i.e. $R' = R - Rb$, $G' = G - Gb$

  where R, Rb, G, Gb are all from the output of the image analysis stage (there are some who use models based on these to derive corrections)
- From here on in we will assume that background correction has taken place.

Background Correction

- Note that the image analysis program you use can have quite an impact at this stage by drastically increasing variability, particularly in low intensities.

Same array, different image analysis and background correction



GenePix

Spot

Note this not swirl.3

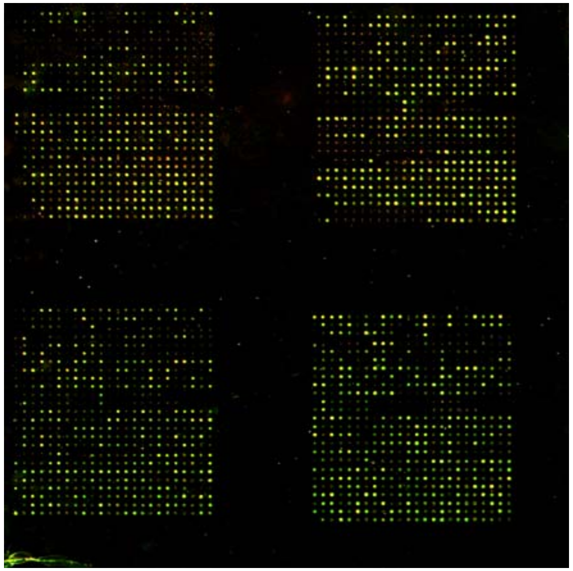
Normalization for two color arrays

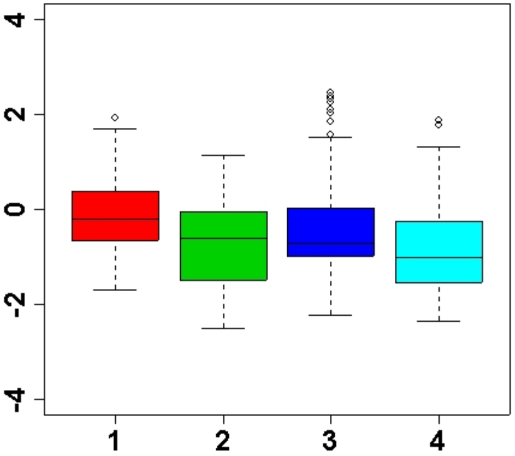
- Why?
  - To correct for systematic differences between samples on the same slide, or between slides, which do not represent true biological variation between samples.
- How do we know it is necessary?
  - By examining self-self hybridizations, where no true differential expression is occurring.
  - We find dye biases which vary with overall spot intensity, location on the array, plate origin, pins, scanning parameters,….

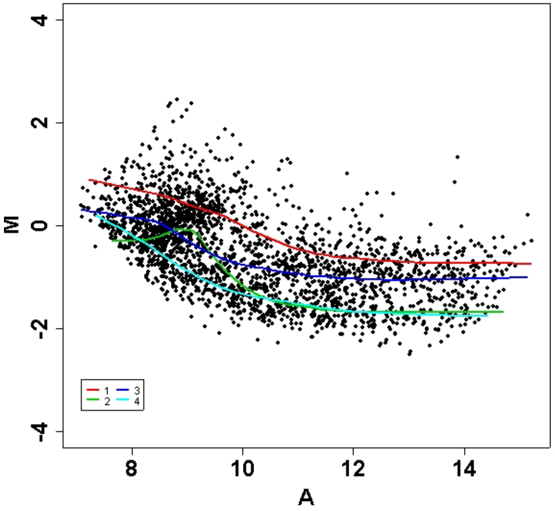
Levels of Normalization for two color arrays

- Within-slides
  - Which genes to use?
  - Location normalization
  - Scale normalization
- Paired-slides (dye-swap)
  - Self-normalization
- Between-slides

Self-self hybridizations







False color overlay

Boxplots within Grid plots

MA-plots

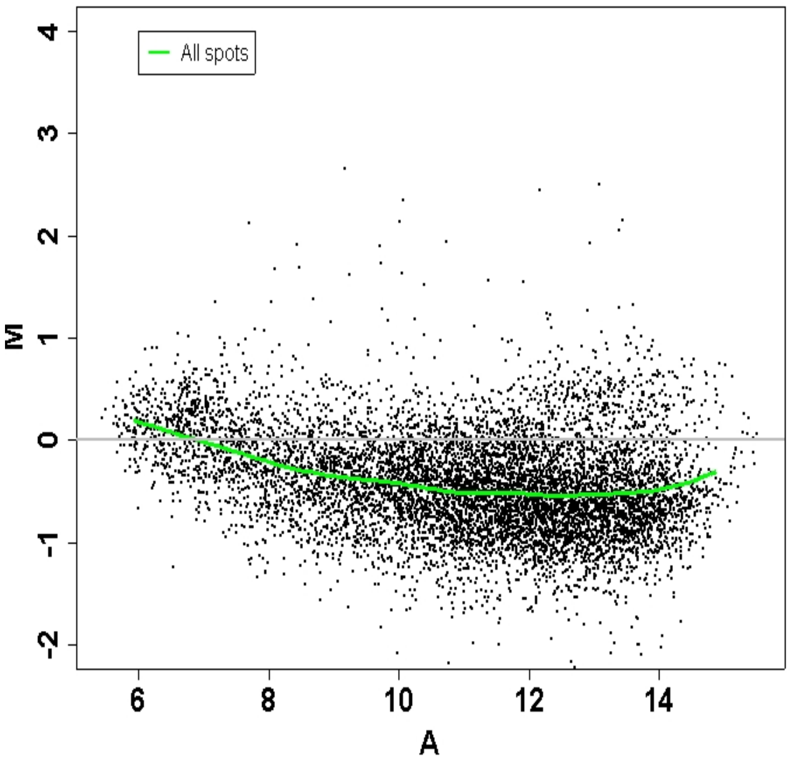
Scaling Normalization

$$log_2R/G \rightarrow log_2R/G - c = log_2R/\ (kG)$$

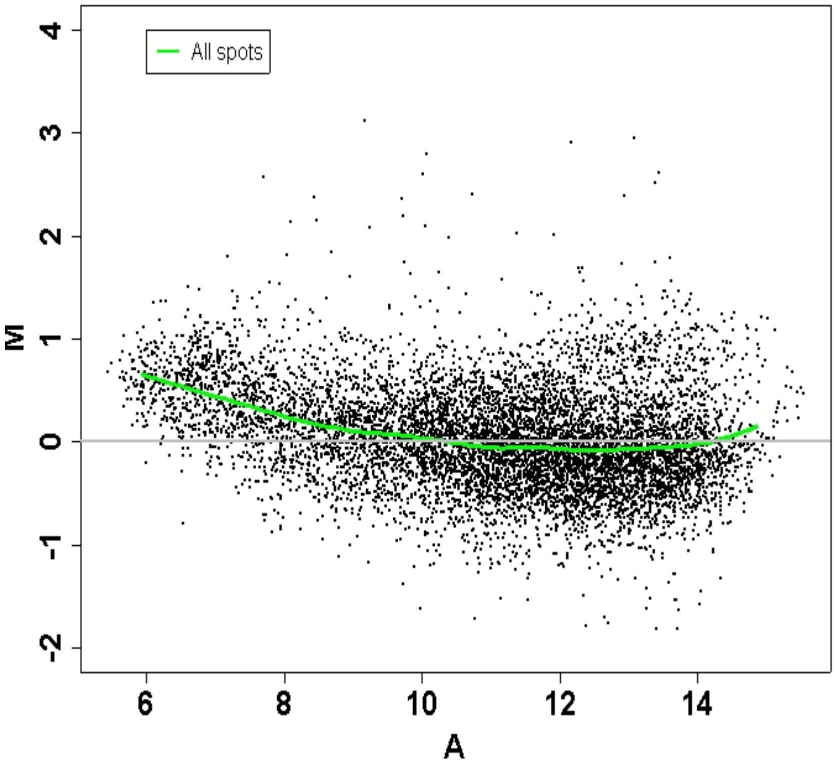
Standard practice (in most software)

c is a constant such as the mean or median log ratio.

MA-plot after scaling



Before Scaling



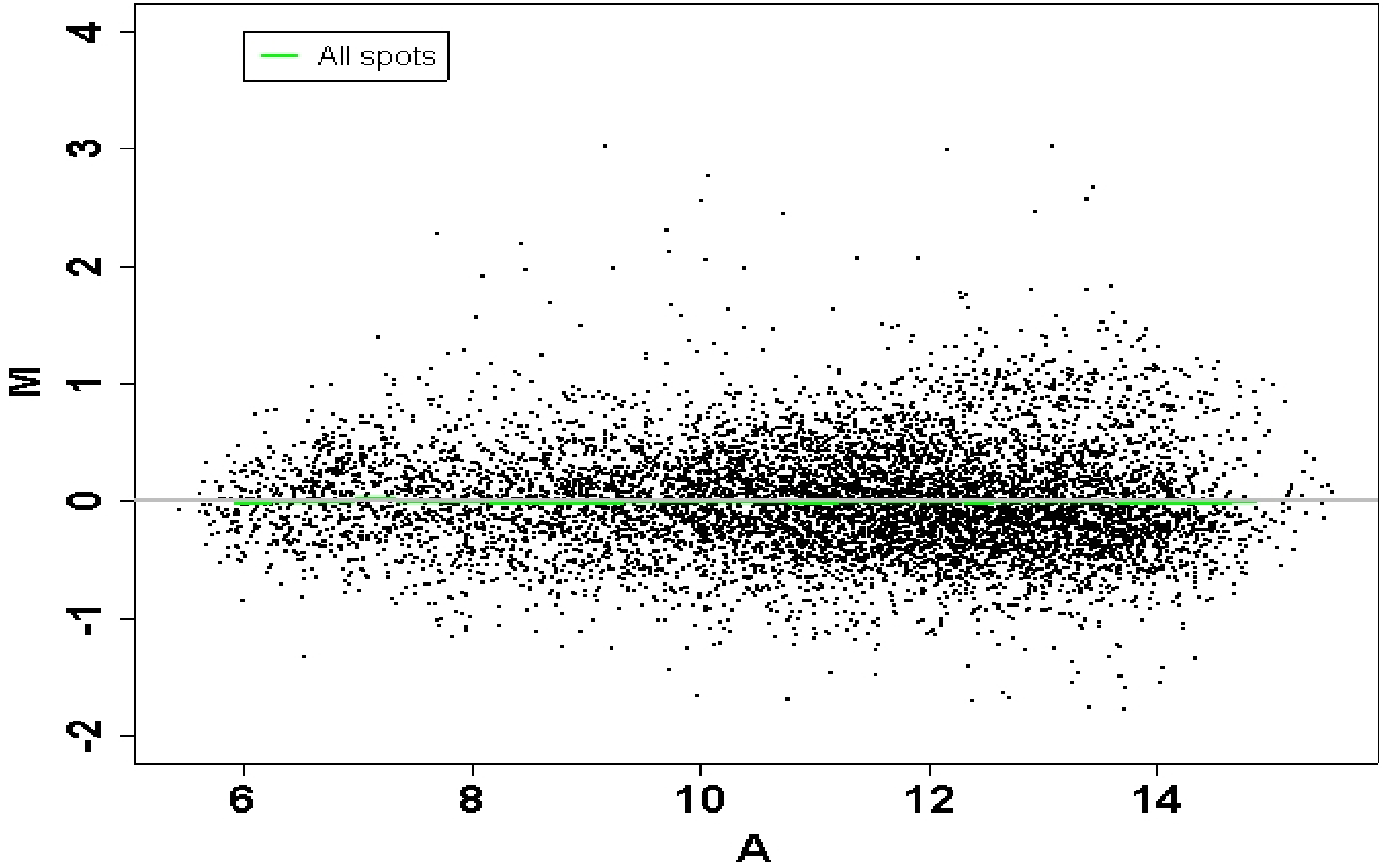
After Scaling

Intensity dependent adjustment

$log_2\ R/G\ \ \text{->}\ \ log_2\ R/G - c(A) = log_2\ R/(k(A)G)$

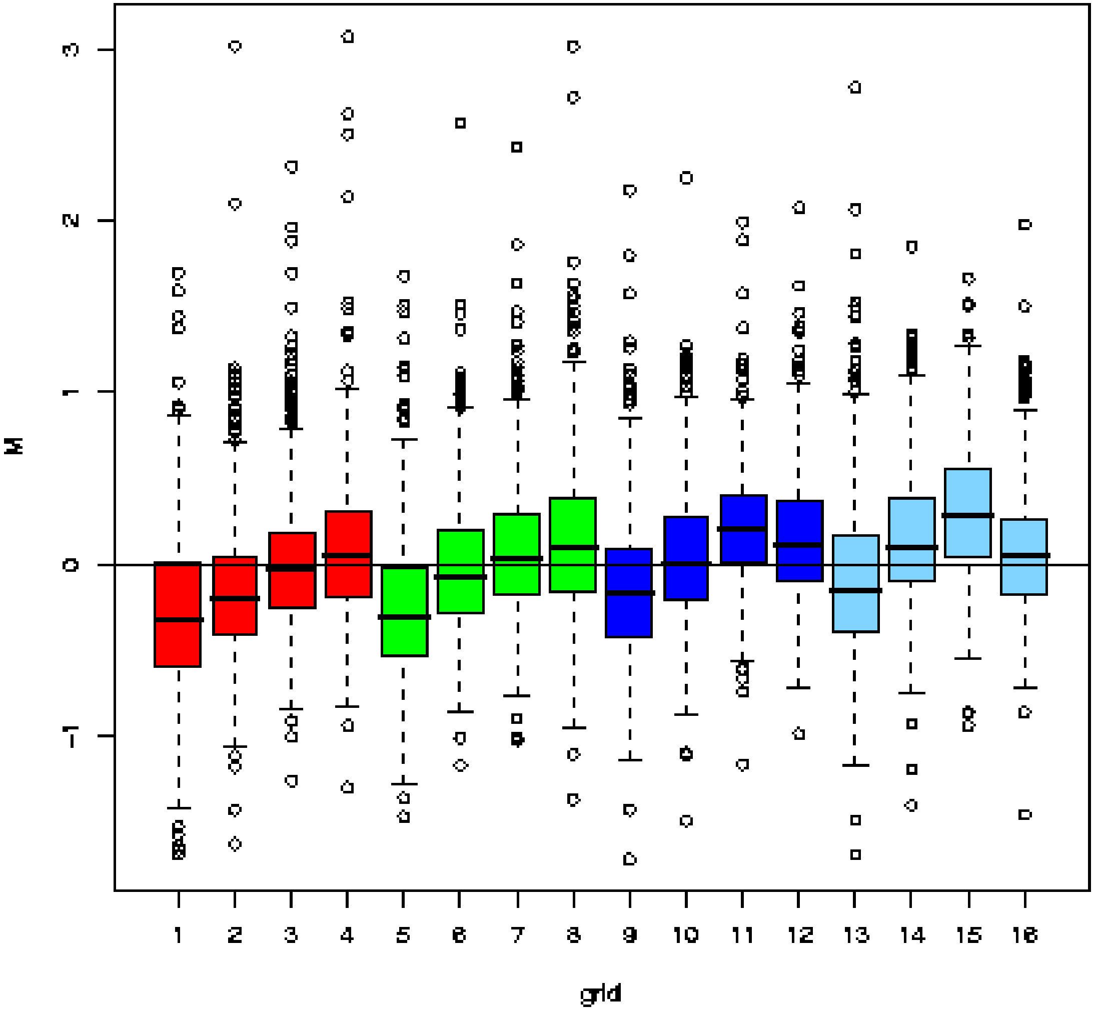
- Compute c by robust locally weighted regression of M on A.
- We typically use a loess curve for this purpose.

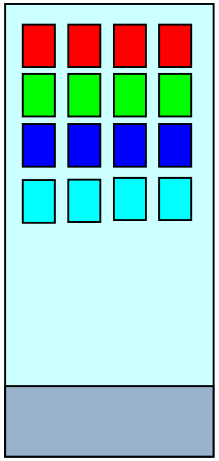
MA-plot after loess normalization



After global loess normalization

Boxplot: print-tip effects remain after global loess normalization





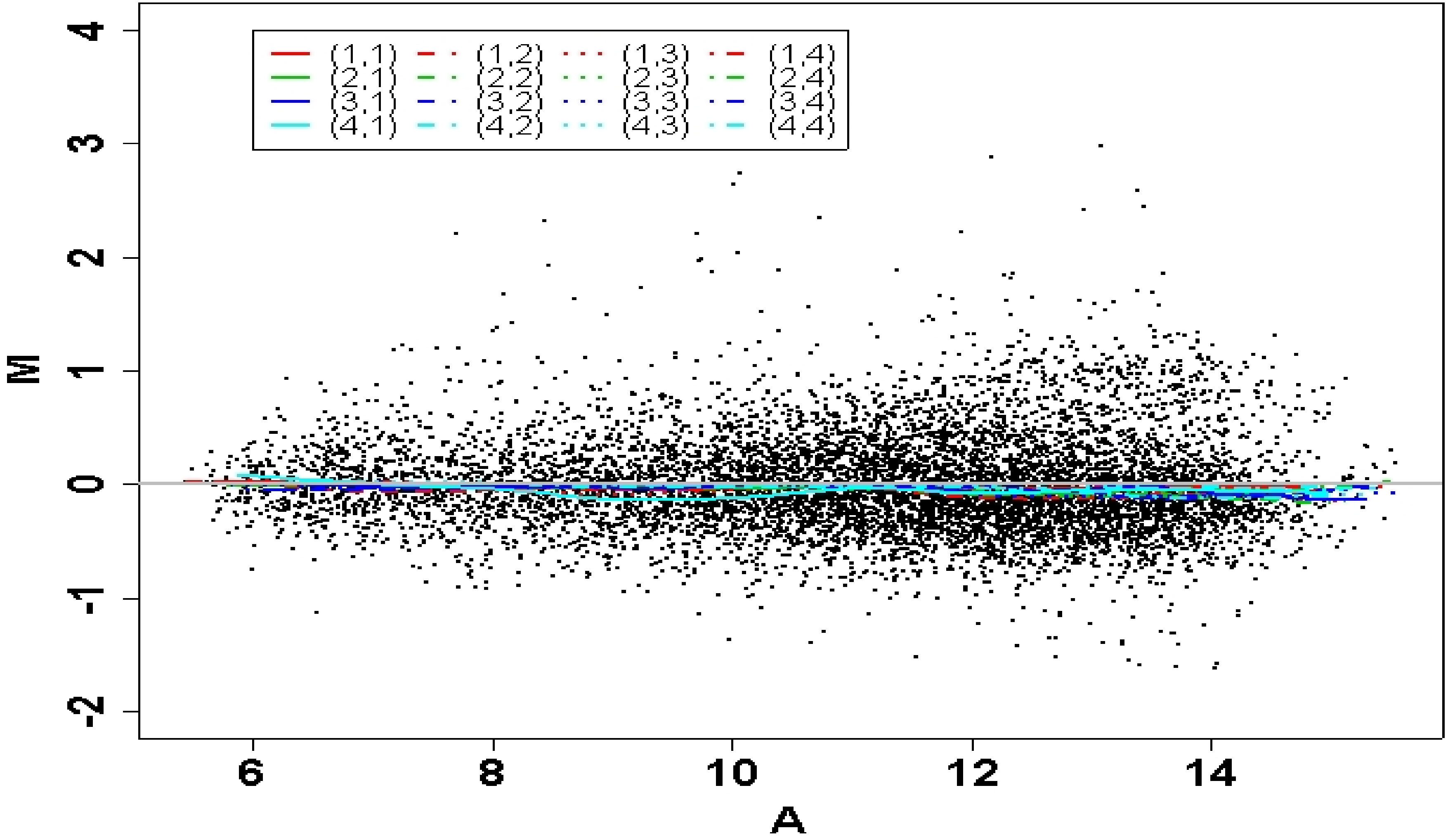
Within print-tip group normalization

- In addition to intensity-dependent variation in log ratios, spatial bias can also be a significant source of systematic error. Most normalization methods do not correct for spatial effects produced by hybridization artifacts or print-tip or plate effects during the construction of the microarrays.
- It is possible to correct for both print-tip and intensity-dependent bias by performing LOWESS fits to the data within print-tip groups, i.e.

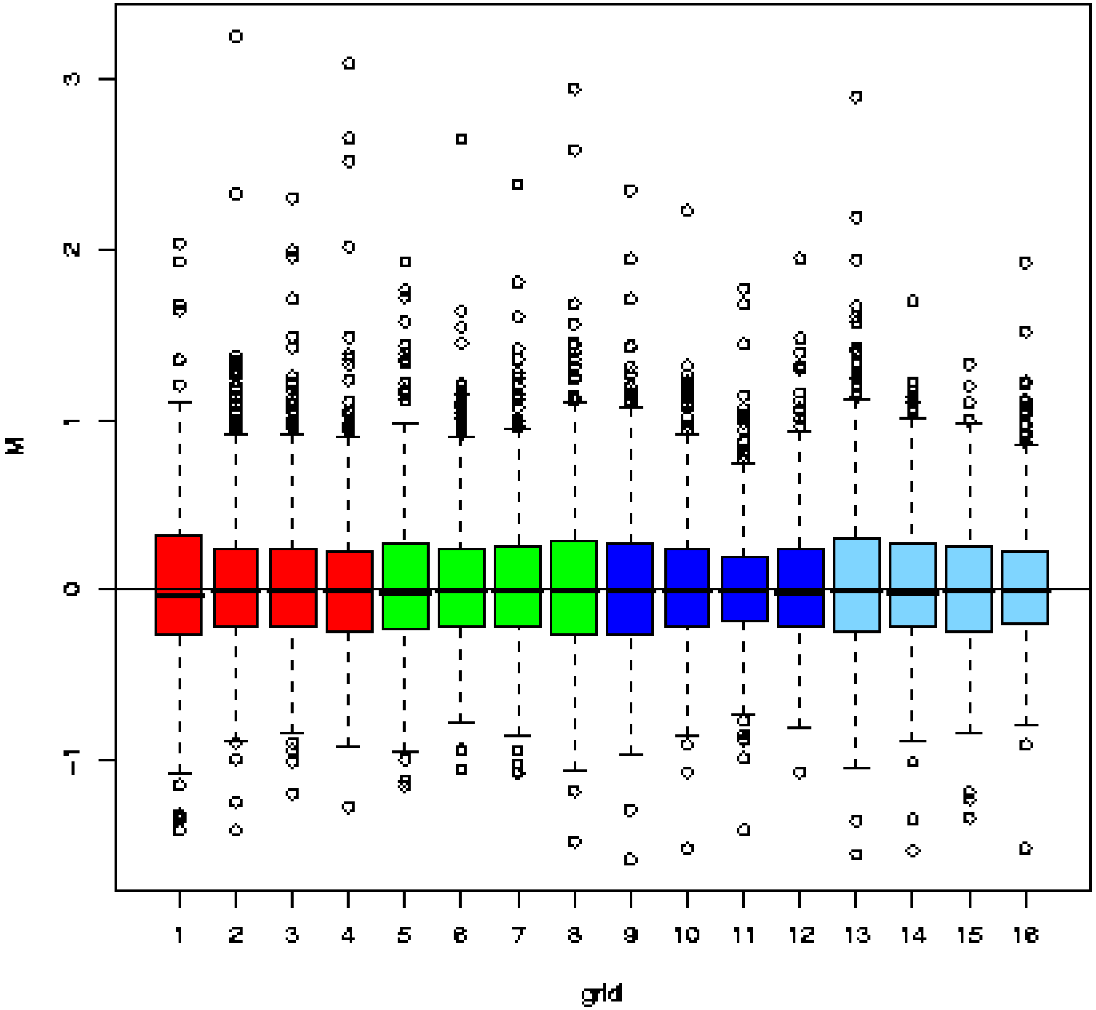
$$\log_2 R/G \;\; -> \;\; \log_2 R/G - c_i(A) = \log_2 R/(k_i(A)G),$$

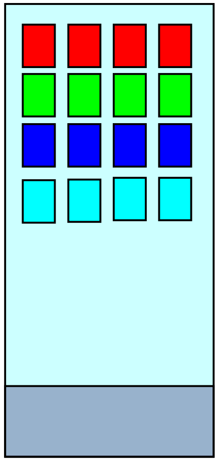
- where $c_i(A)$ is the LOWESS fit to the *MA*-plot for the ith grid only.

Print-tip normalized data: MA-plot

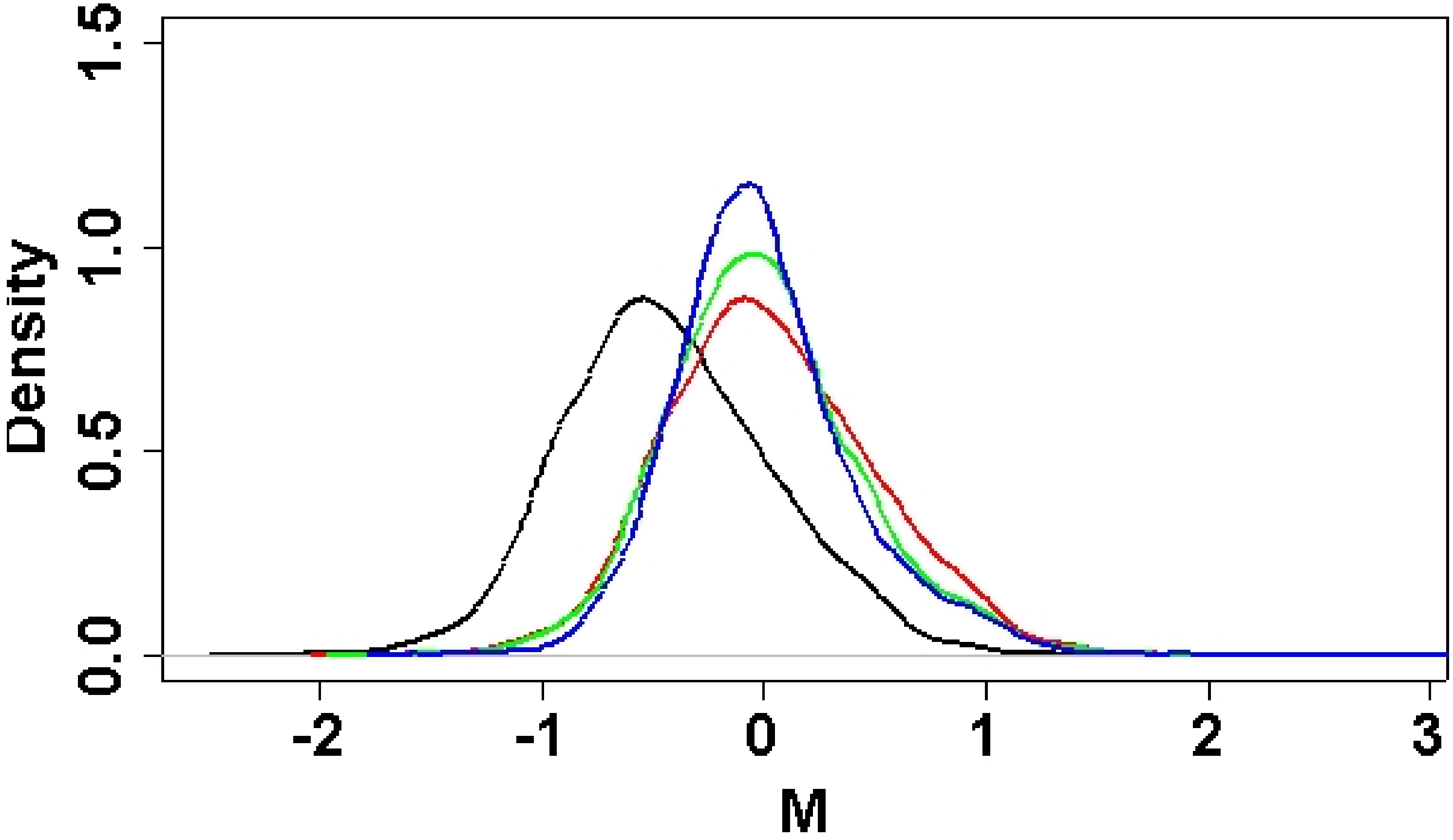


Print-tip normalized data: boxplot



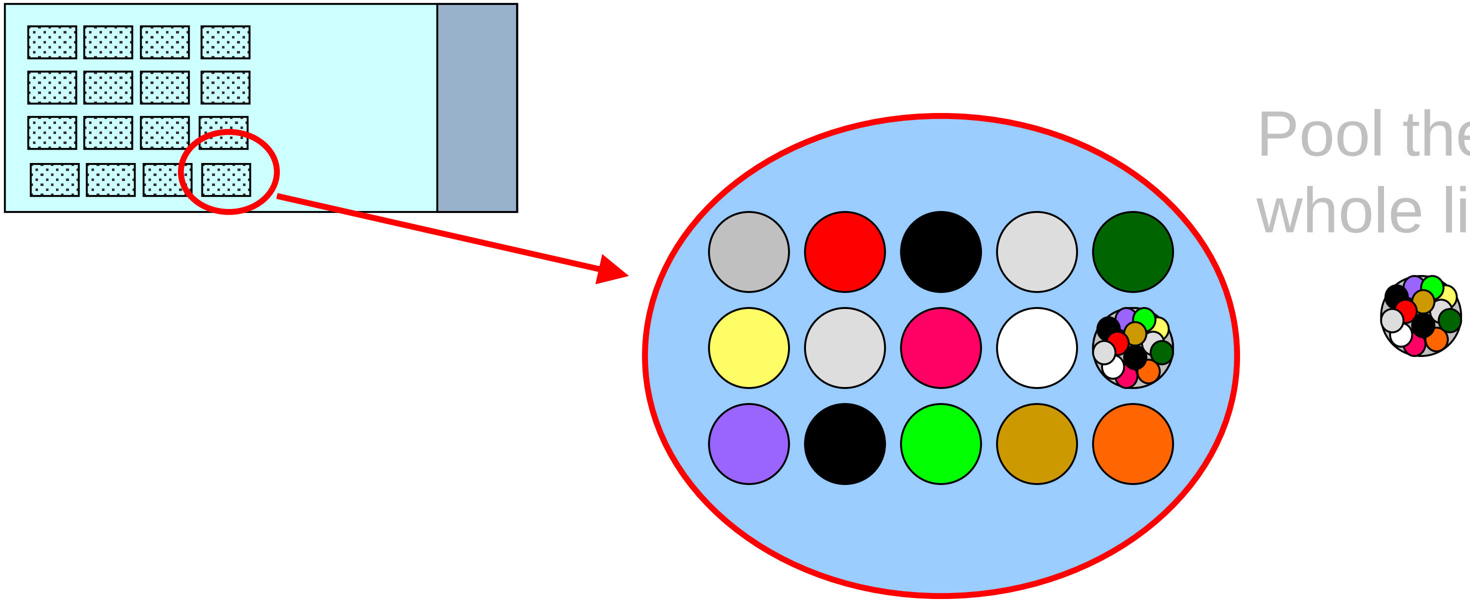


Smoothed histograms of M values



Black: unnormalized; red: global median; green: global lowess; blue: print-tip lowess

MSP titration series (Microarray Sample Pool)

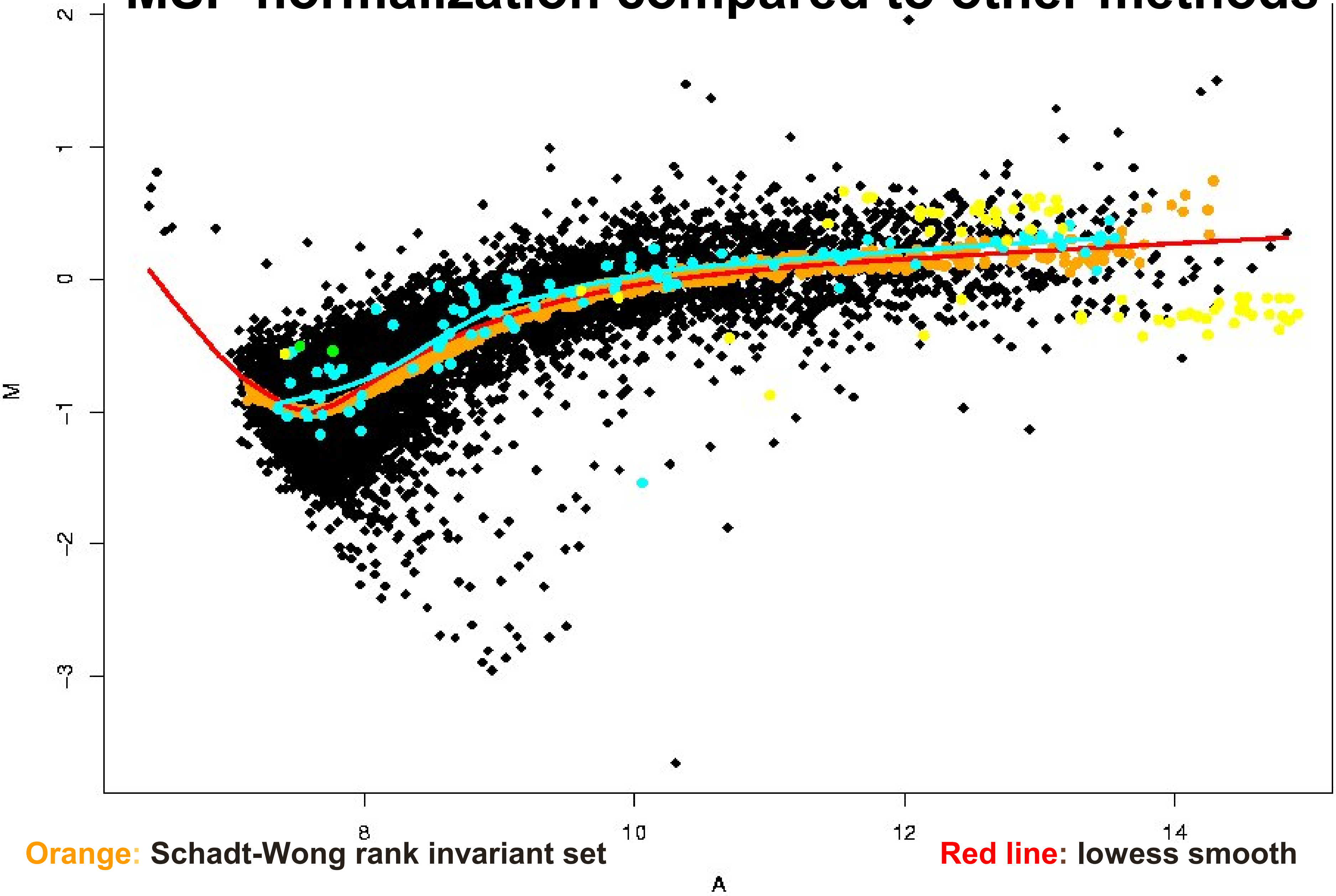


Control set to aid intensity- dependent normalization

Different concentrations

Spotted evenly spread across the slide

MSP normalization compared to other methods



Orange: Schadt-Wong rank invariant set

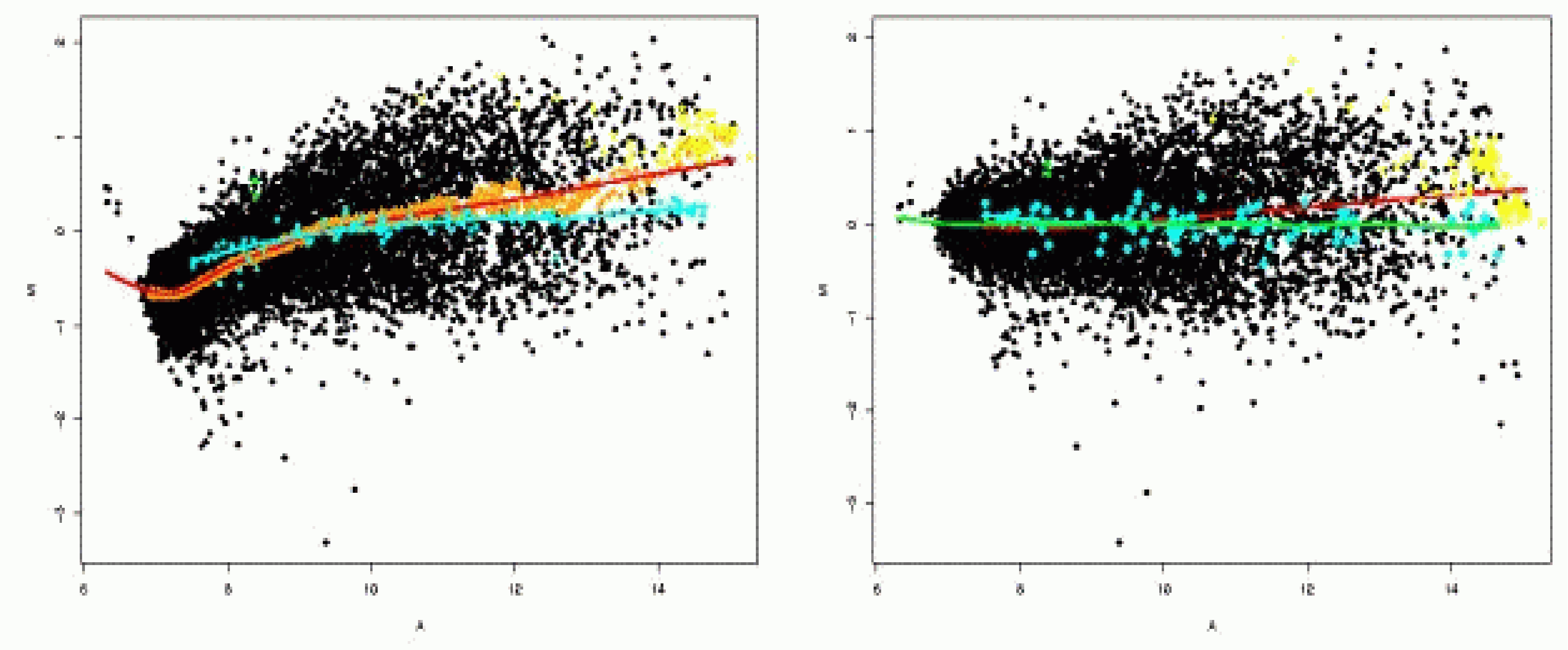
Red line: lowess smooth

Yellow: GAPDH, tubulin

Light blue: MSP pool / titration

Composite normalization

$$c_i(A)=\alpha_A g(A)+(1-\alpha_A)f_i(A)$$



Before and after composite normalization

-MSP lowess curve

-Global lowess curve

-Composite lowess curve

(Other colours control spots)

Paired-slides: dye-swap

- Slide 1, $M = \log_2 (R/G) - c$
- Slide 2, $M' = \log_2 (R'/G') - c'$

Combine by subtracting the normalized log-ratios:

$$[\ (\log_2 (R/G) - c) - (\log_2 (R'/G') - c')\ ] / 2$$

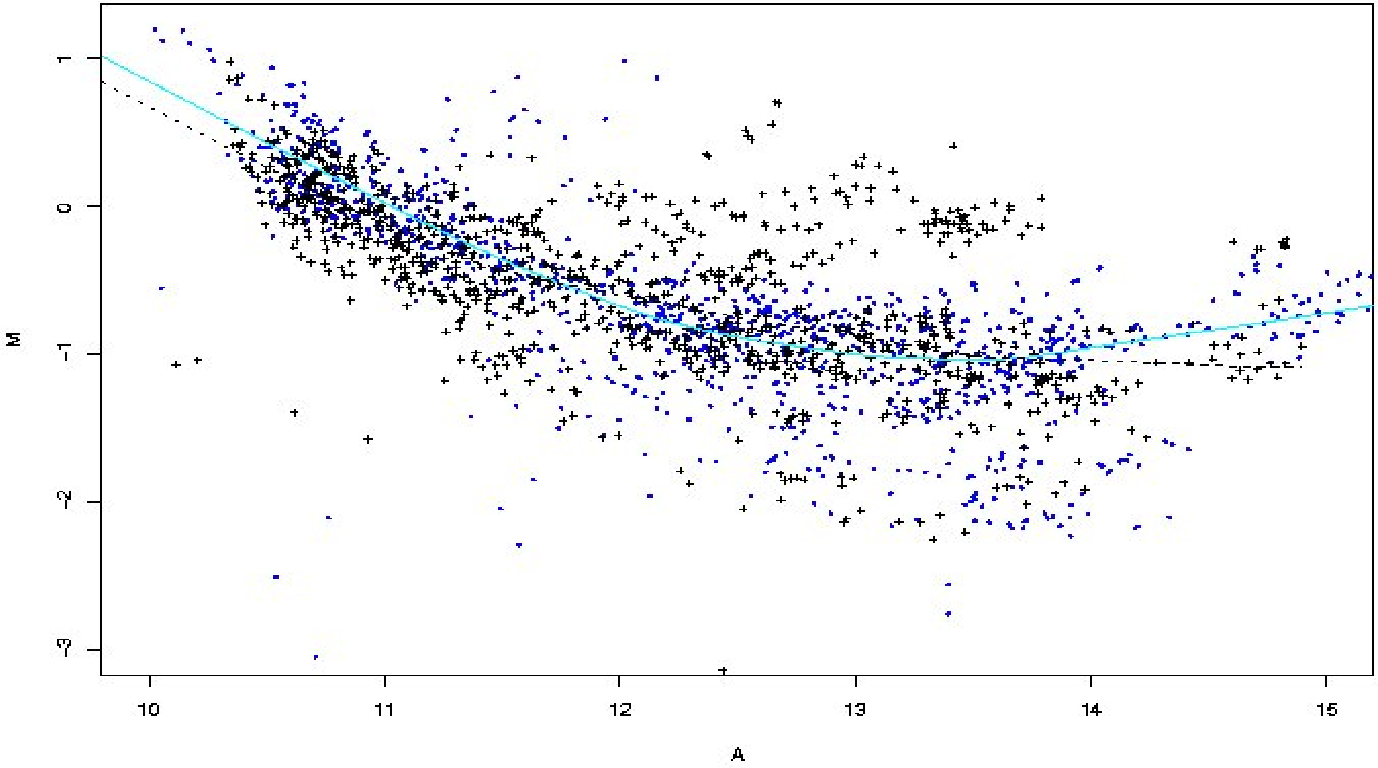
$$\approx [\ \log_2 (R/G) + \log_2 (G'/R')\ ] / 2$$

$$\approx [\ \log_2 (RG'/GR')\ ] / 2$$

provided $c = c'$.

Assumption: the normalization functions are the same for the two slides.

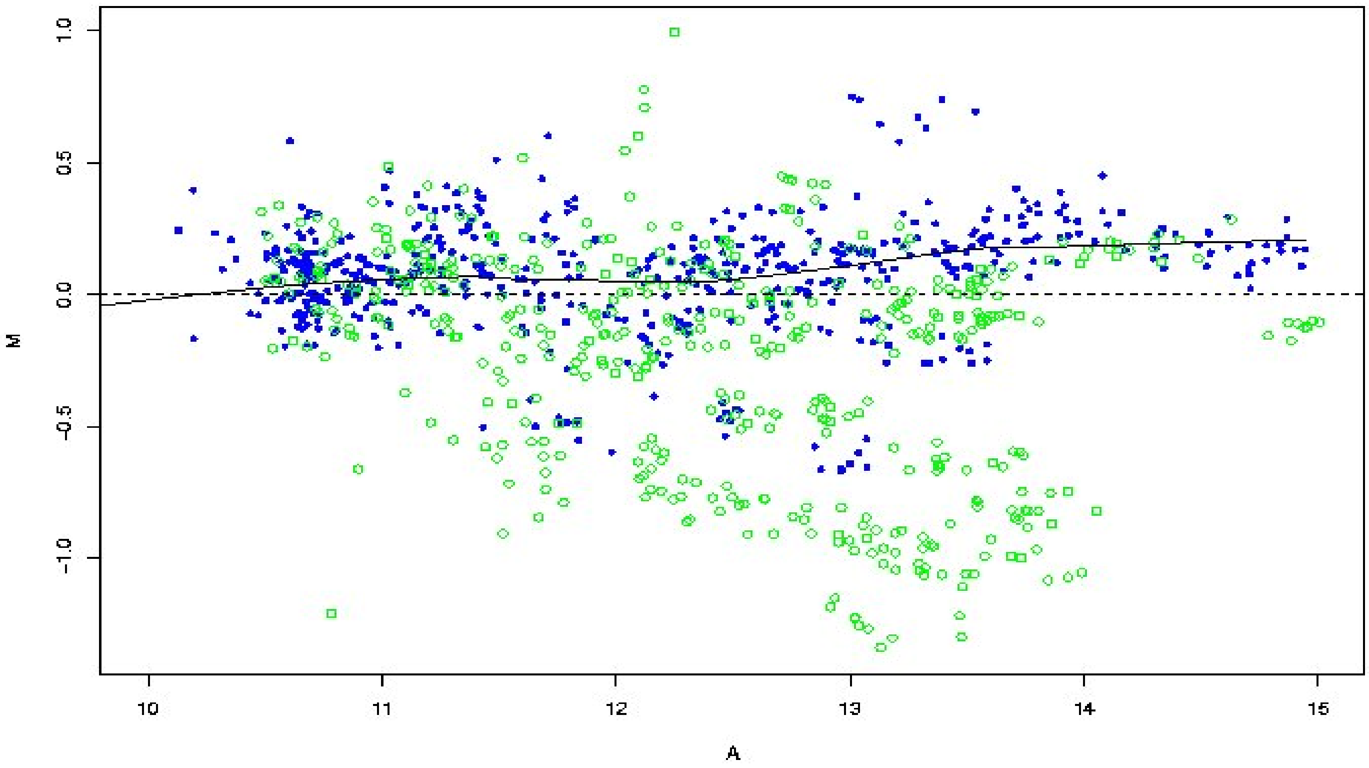
Checking the assumption



MA plot for slides 1 and 2: it isn’t always like this.

Result of self-normalization

(M - M')/2 vs. (A + A')/2



One way of taking scale into account

Assumption: All slides have the same spread in M

True log ratio is m_{ij} where i represents different slides and j represents different spots.

Observed is M_{ij}, where

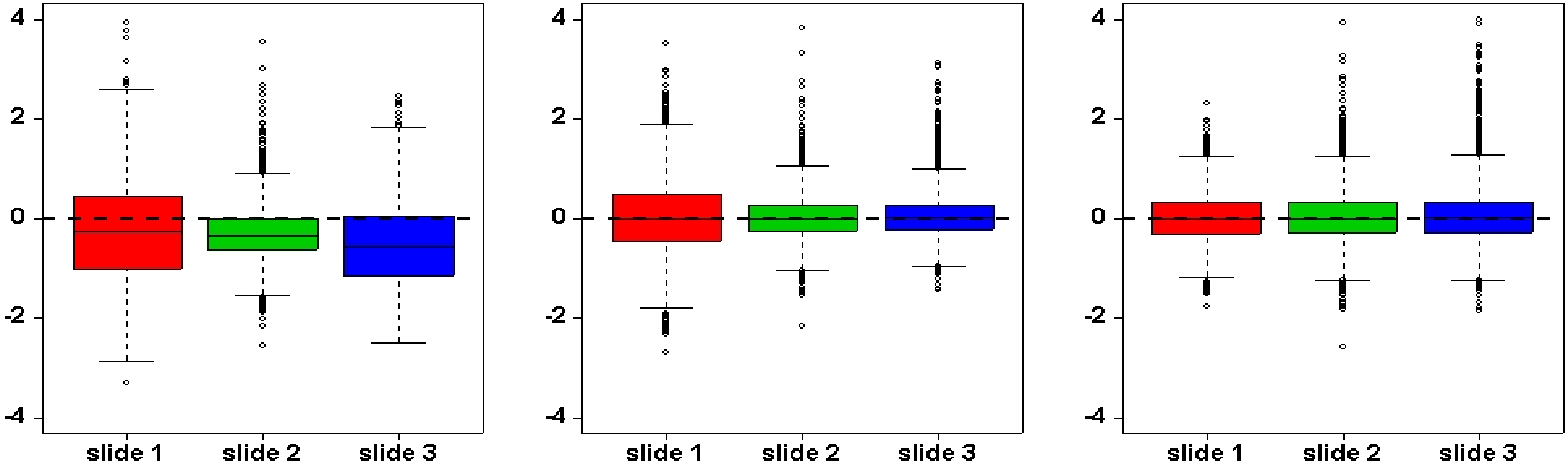
$$M_{ij} = a_i \, m_{ij}$$

Robust estimate of a_i is

$$\frac{MAD_i}{\sqrt[I]{\prod_{i=1}^{I} MAD_i}}$$

$$MADi = \text{median}_j \{ \, |y_{ij} - \text{median}(y_{ij}) \, | \, \}$$

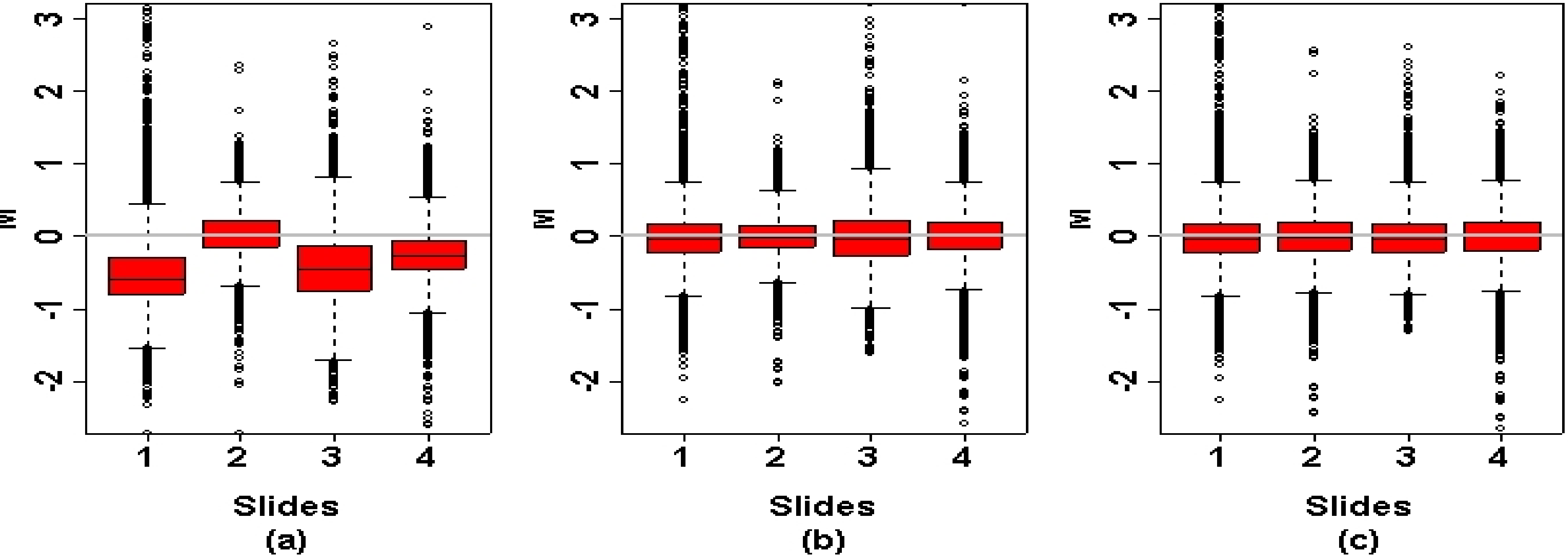
Scale normalization: between slides



Before normalization After location normalization After scale normalization

Boxplots of log ratios from 3 replicate self-self hybridizations.

Scale normalization: swirl dataset



Before normalization

After location normalization

After scale normalization

Other between slide normalizations

- Quantile normalization applied separately to R and G channels (after within chip normalization)

Two Channel Summary

- Background Correction
  - Taking too much off can greatly increase variability
- Normalization
  - Reduces systematic (not random) effects
  - Makes it possible to compare several arrays
  - Use logratios (M vs A-plots)
  - Lowess normalization (dye bias)
  - MSP titration series – composite normalization
  - Pin-group location normalization
  - Pin-group scale normalization
  - Between slide scale normalization

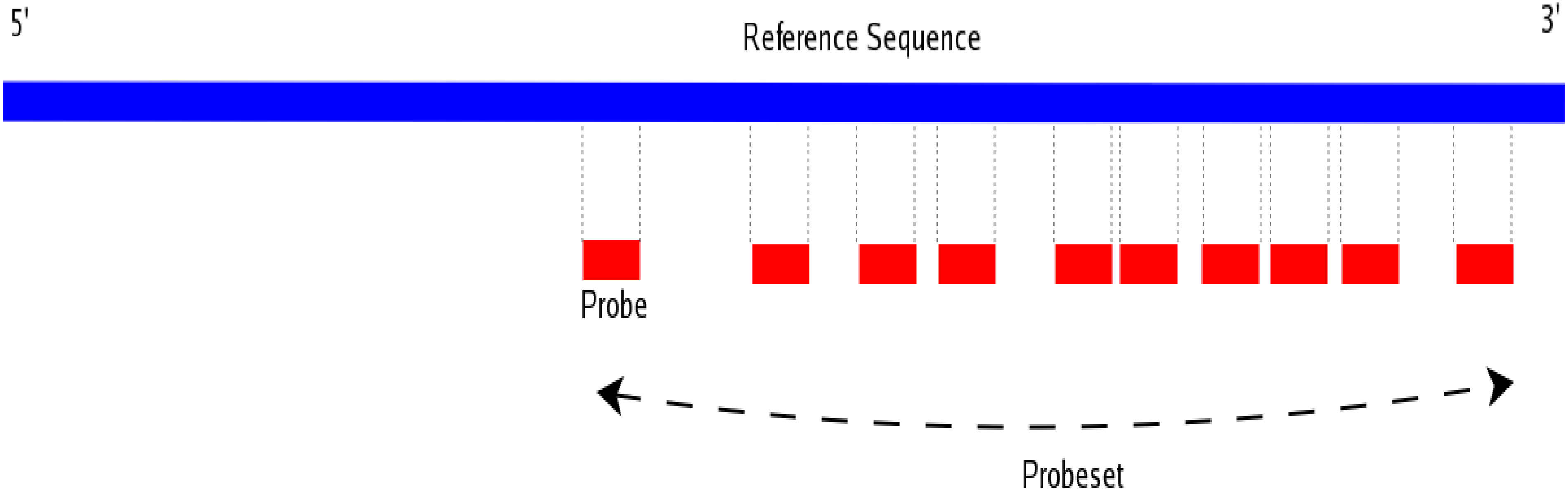
Single-channel arrays

Affymetrix GeneChip

- Commericial mass produced high density oligonucleotide array technology developed by Affymetrix http://www.affymetrix.com
- Single channel microarray

Image courtesy of Affymetrix.

Probes and Probesets



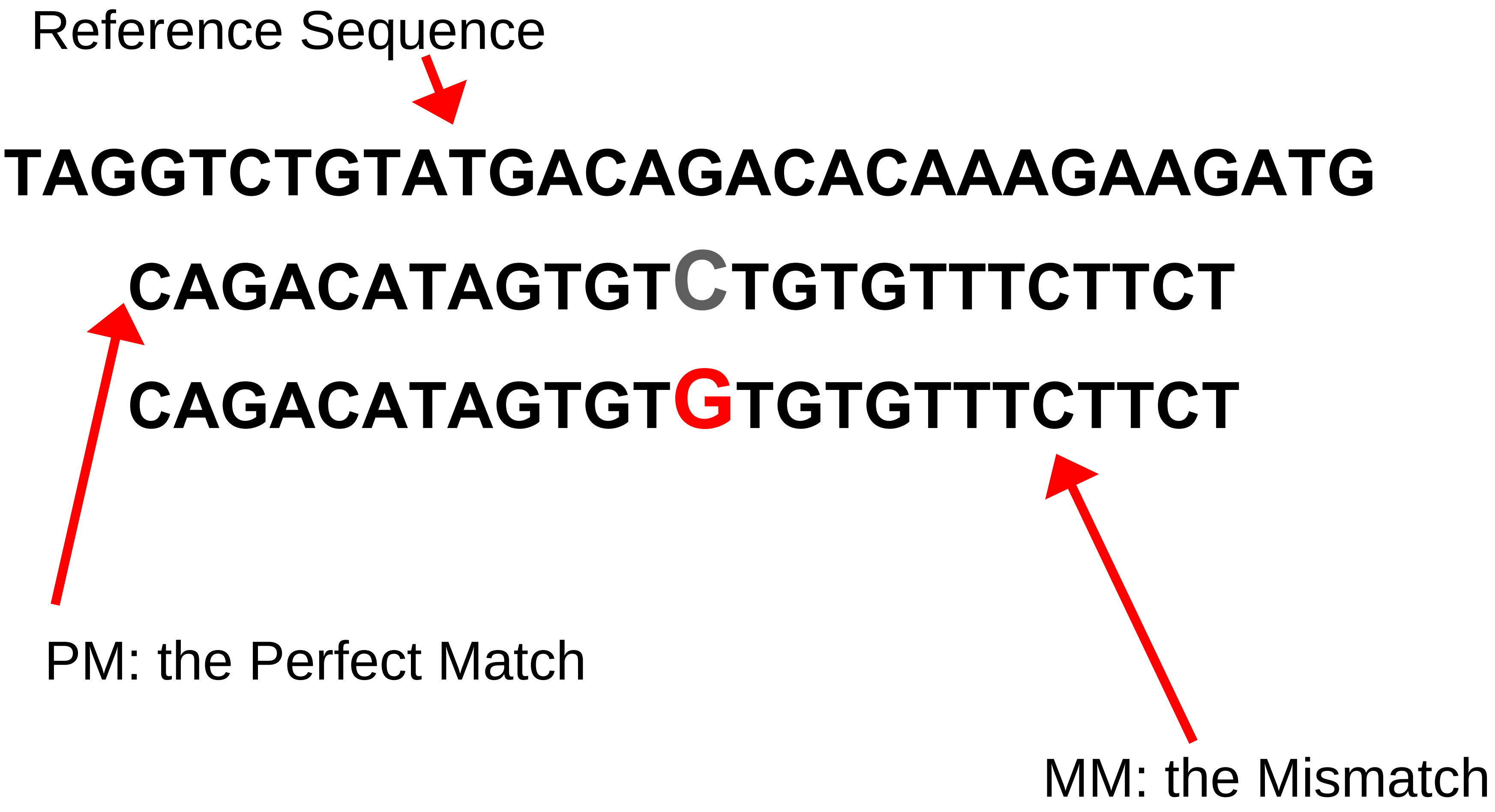
Typically 11 probe(pairs) in a probeset

Latest GeneChips have as many as:

54,000 probesets

1.3 Million probes

Two Probe Types



Reference Sequence

TAGGTCTGTATGACAGACACAAAGAAGATG

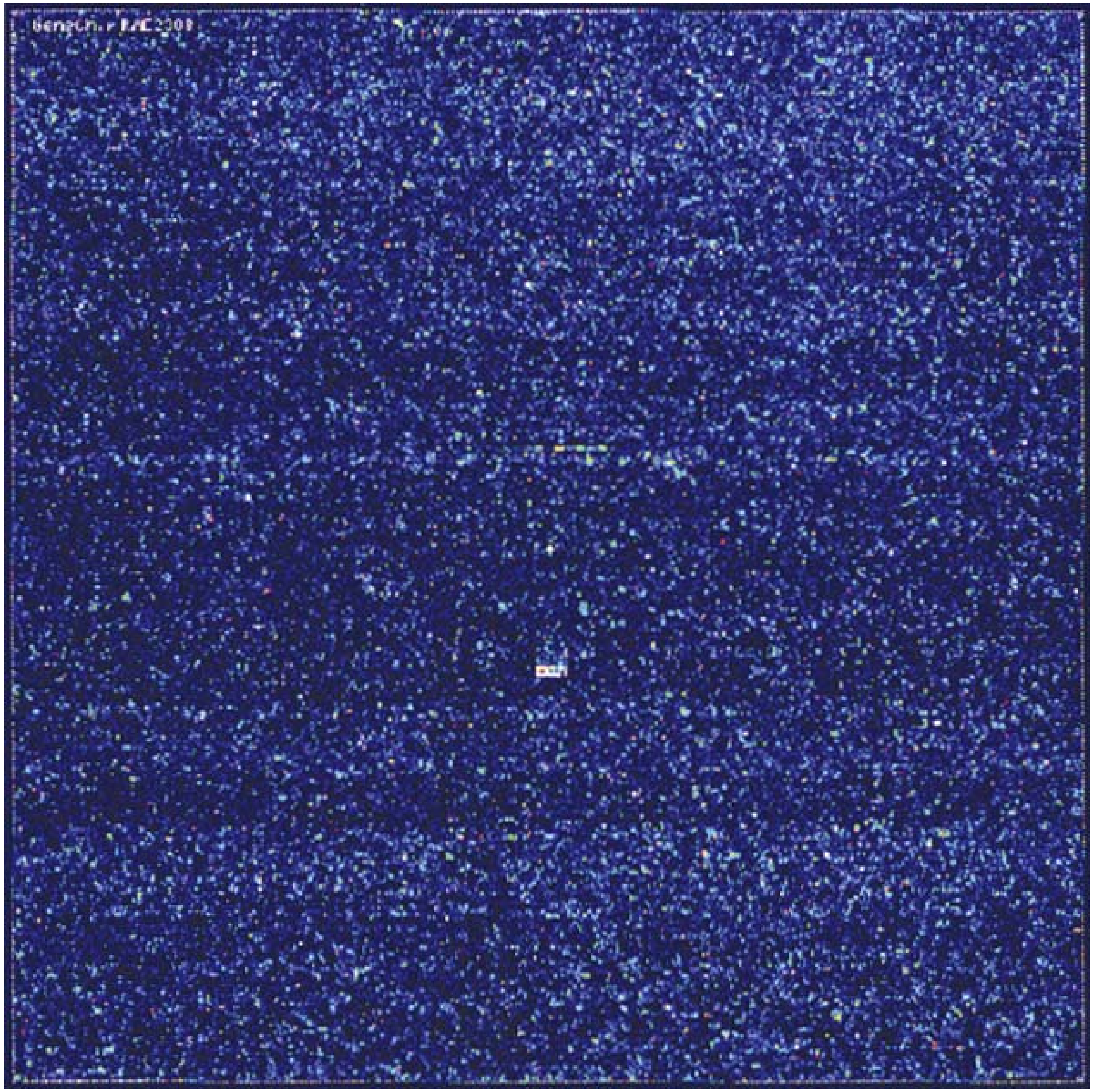
CAGACATAGTGTCTGTGTTTCTTCT

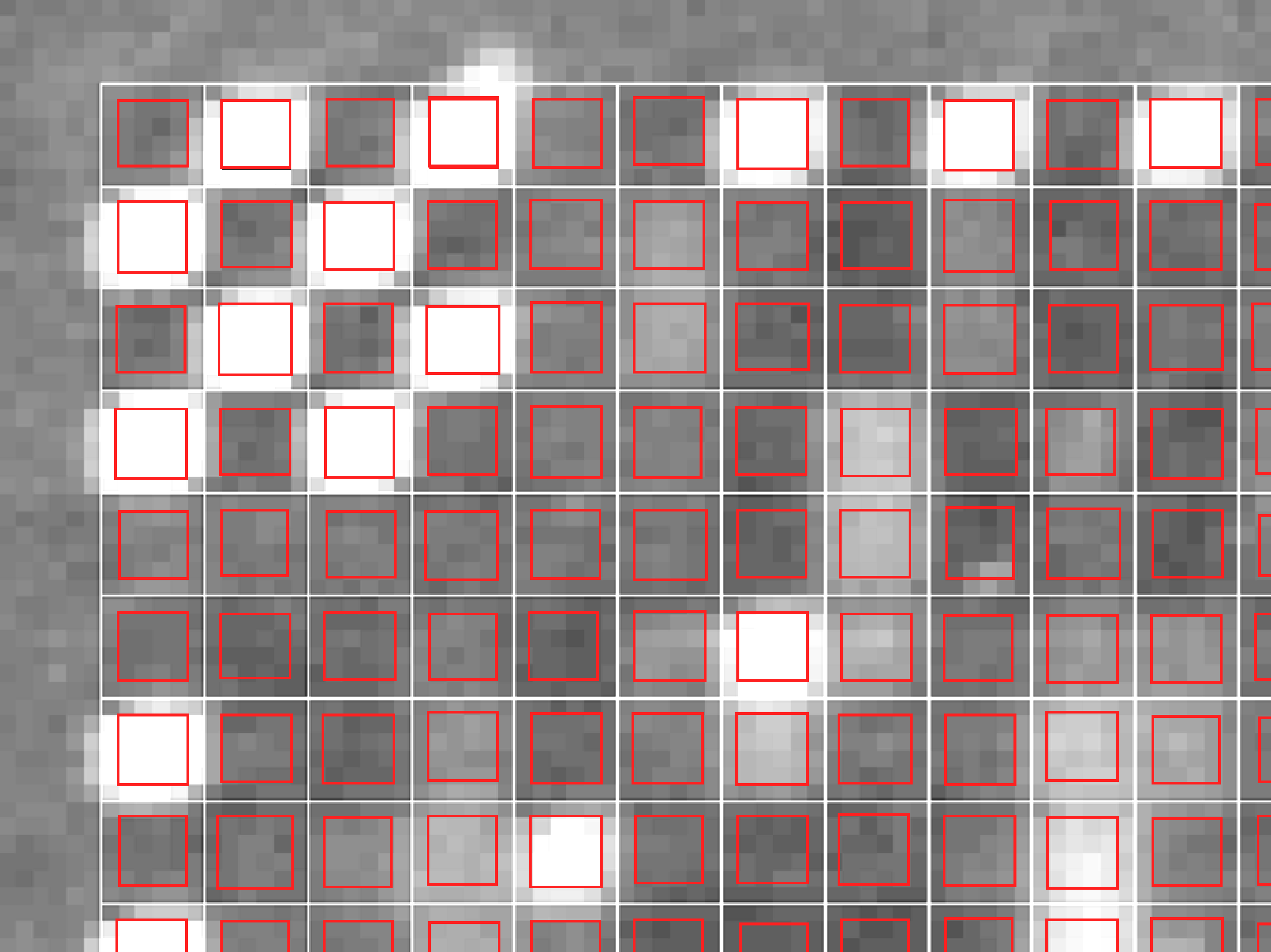
CAGACATAGTGTGTGTGTTTCTTCT

PM: the Perfect Match

MM: the Mismatch

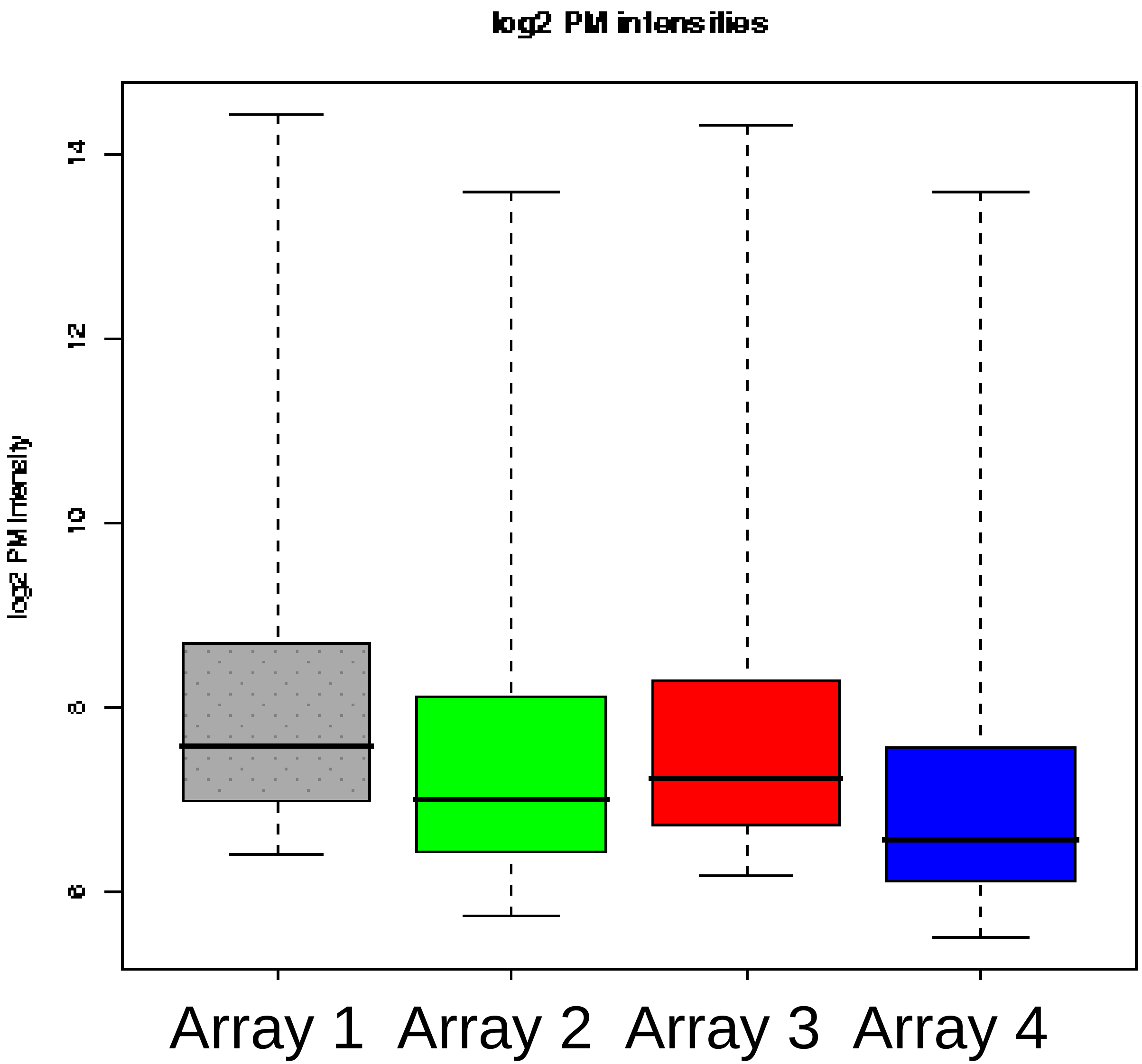
Image Analysis



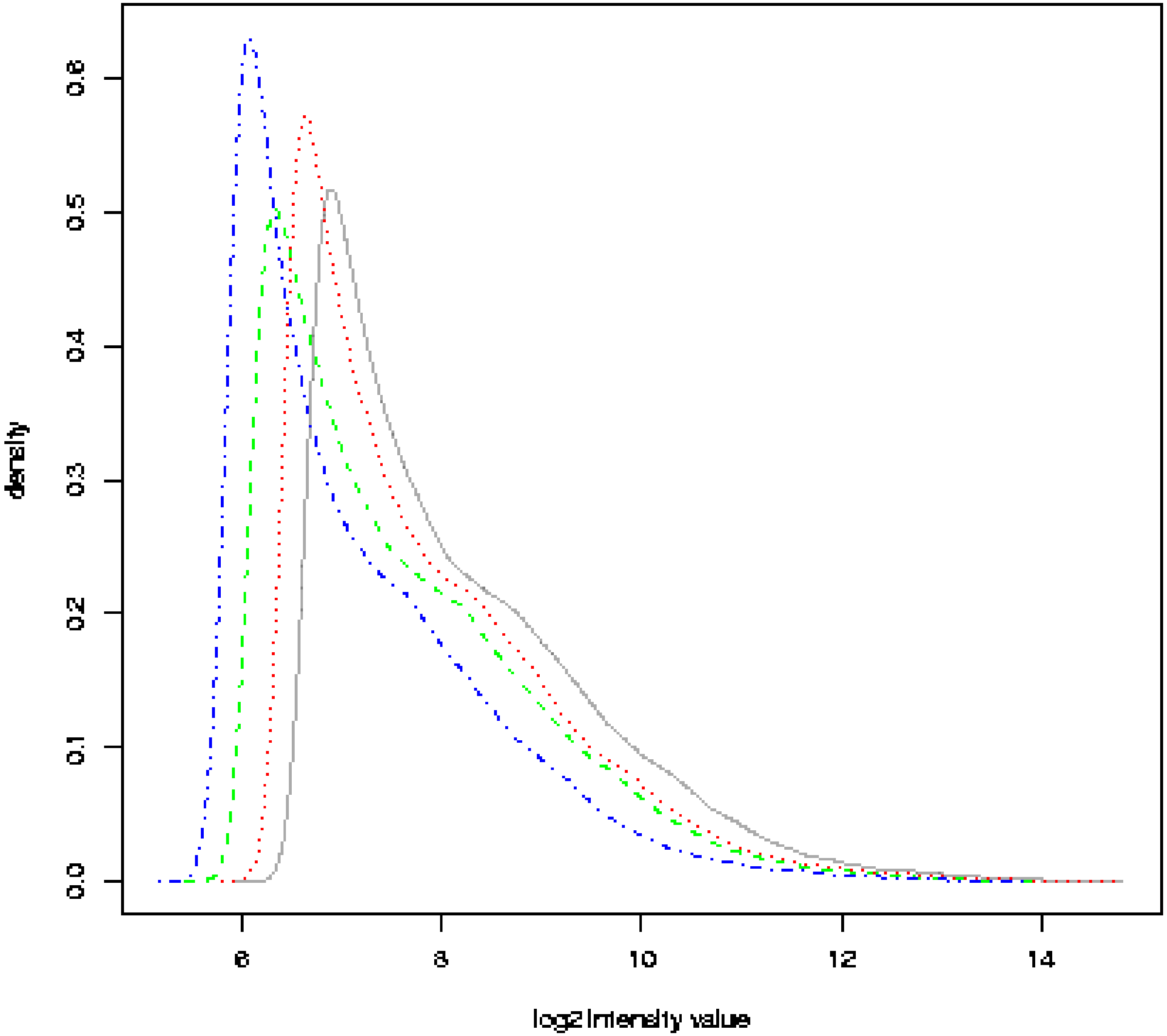


Courtesy: F. Colin

Boxplot raw intensities

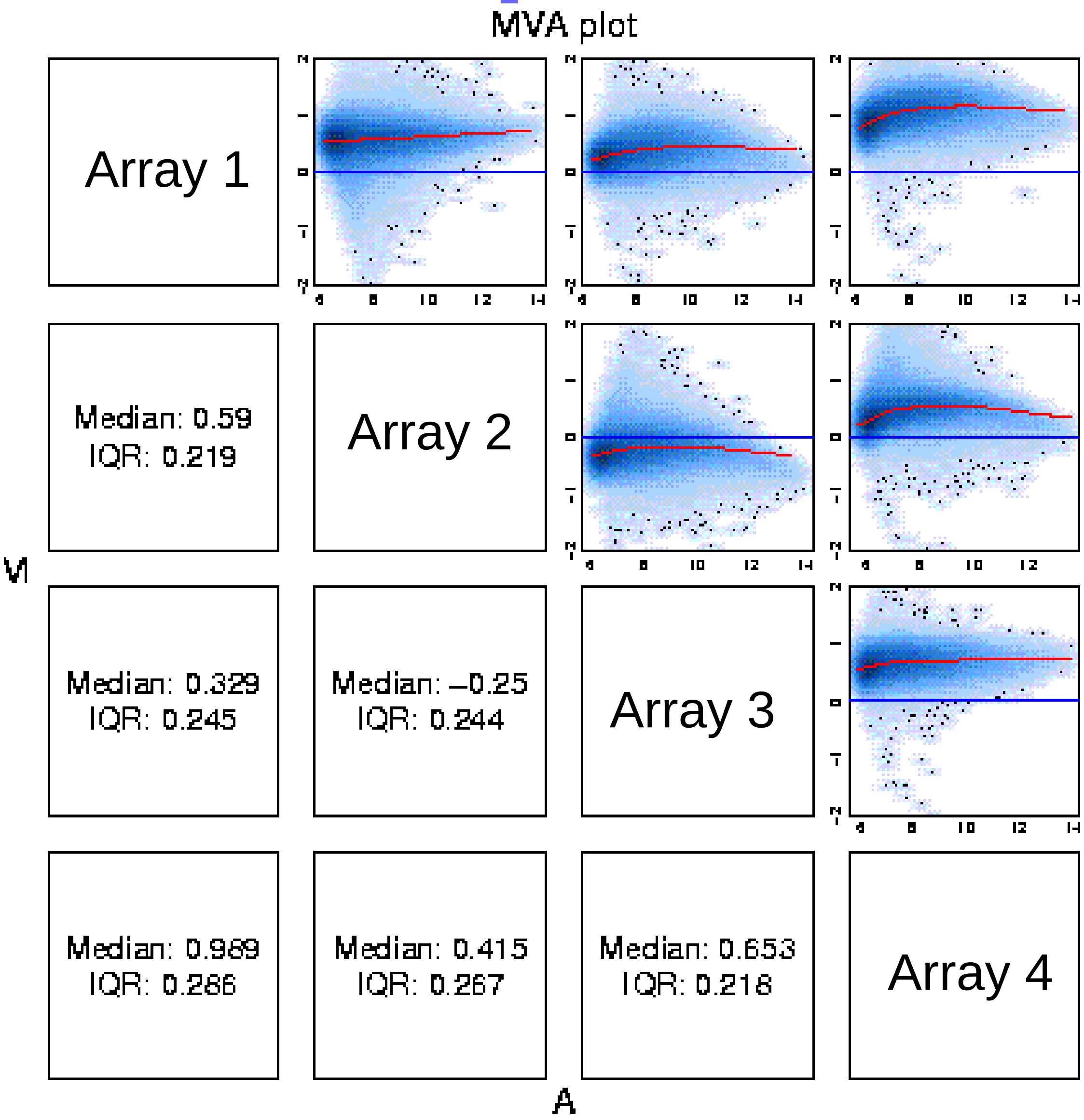


Density plots

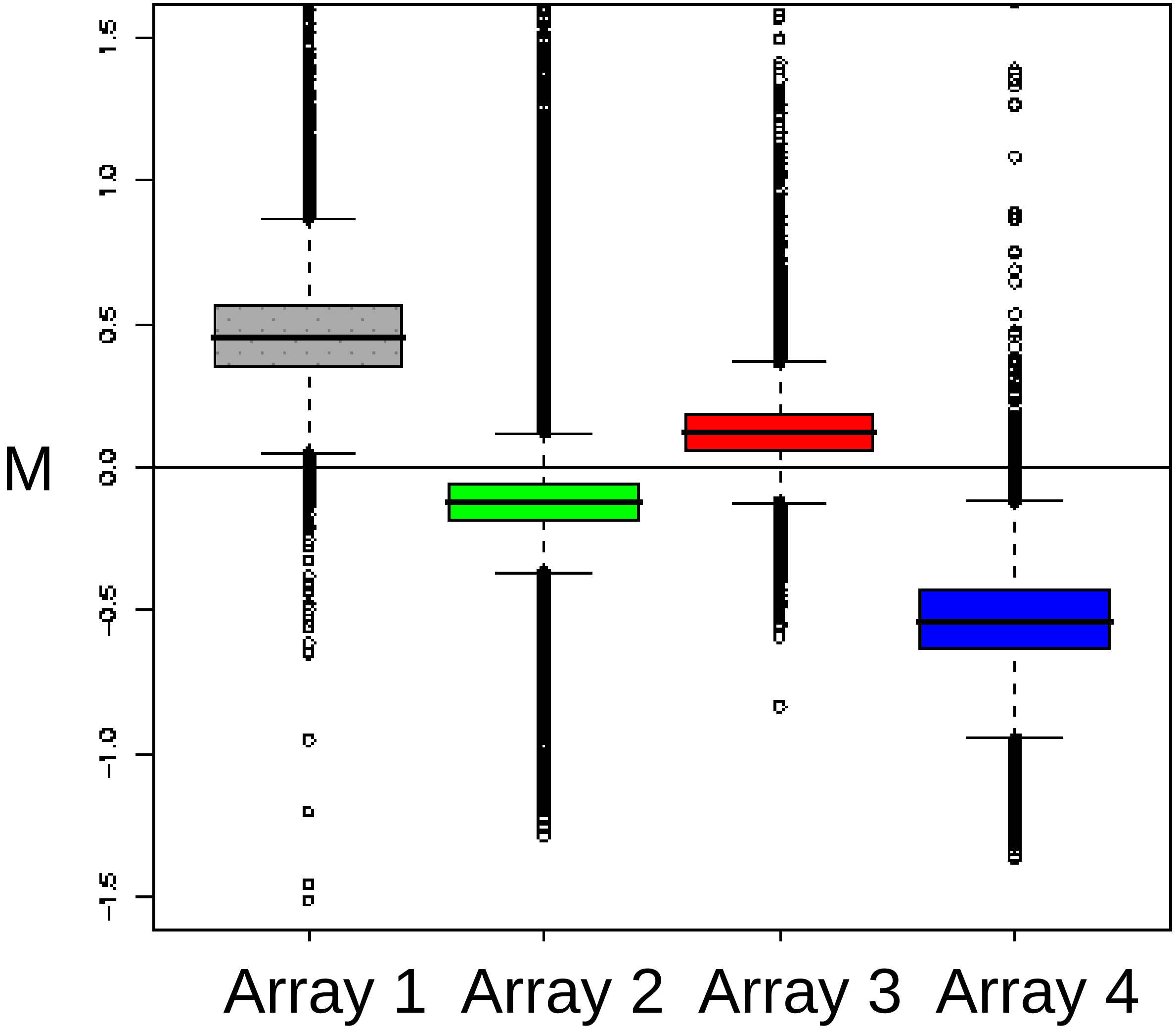


Pairwise MA plots

$M=\log_2 \text{array}_i/\text{array}_j$
$A=1/2*\log_2(\text{array}_i*\text{array}_j)$

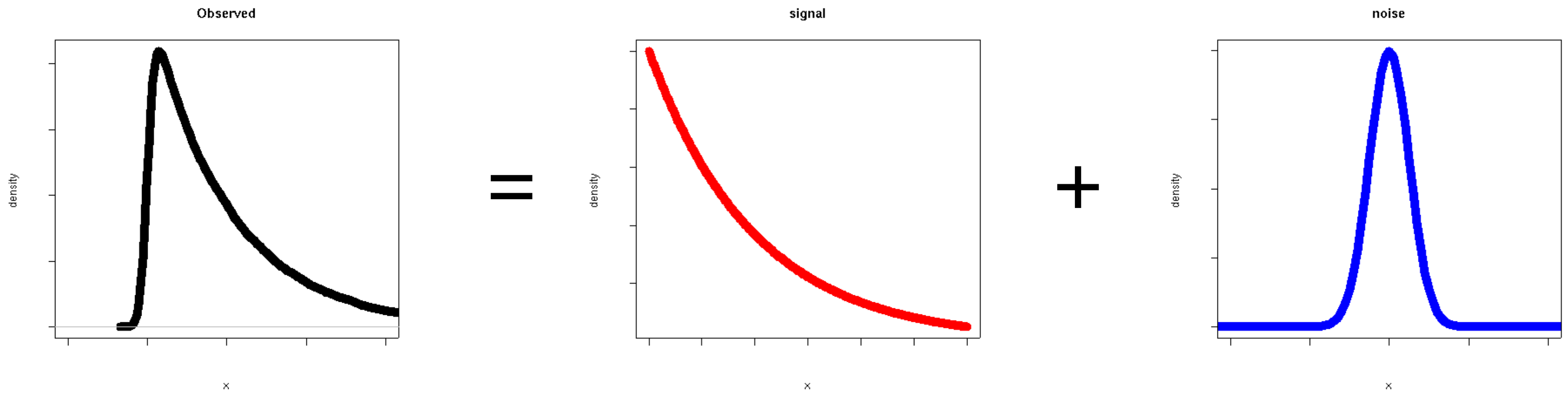


Boxplots comparing M



RMA Background Approach

- Convolution Model



Observed PM

Signal S

Noise N

$$\mathrm{E}\left(S \middle| PM = pm\right) = a + b\,\frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{pm-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) + \Phi\left(\frac{pm-a}{b}\right) - 1}$$

GCRMA Background Approach

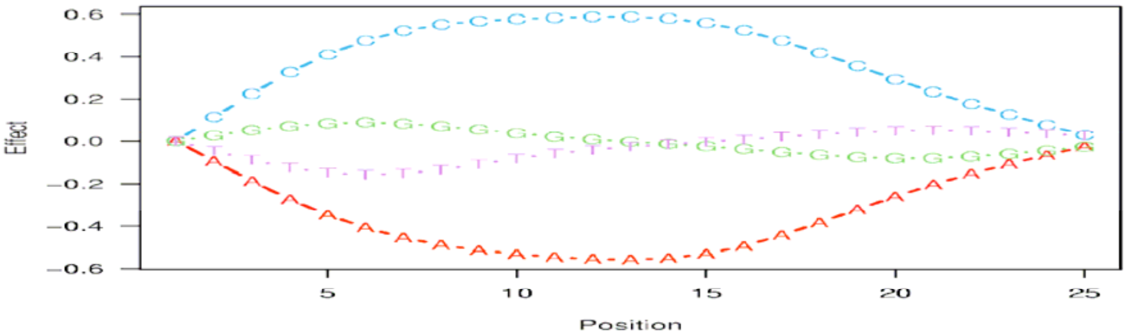
- $PM=O_{pm}+N_{pm}+S$
- $MM=O_{mm}+N_{mm}$

- O – Optical noise
- N – non-specific binding
- S – Signal

- Assume O is distributed Normal
- $\log(N_{pm})$and $\log(N_{mm})$ are assumed bi-variate normal with correlation 0.7
- log(S) assumed exponential(1)

GCRMA continued

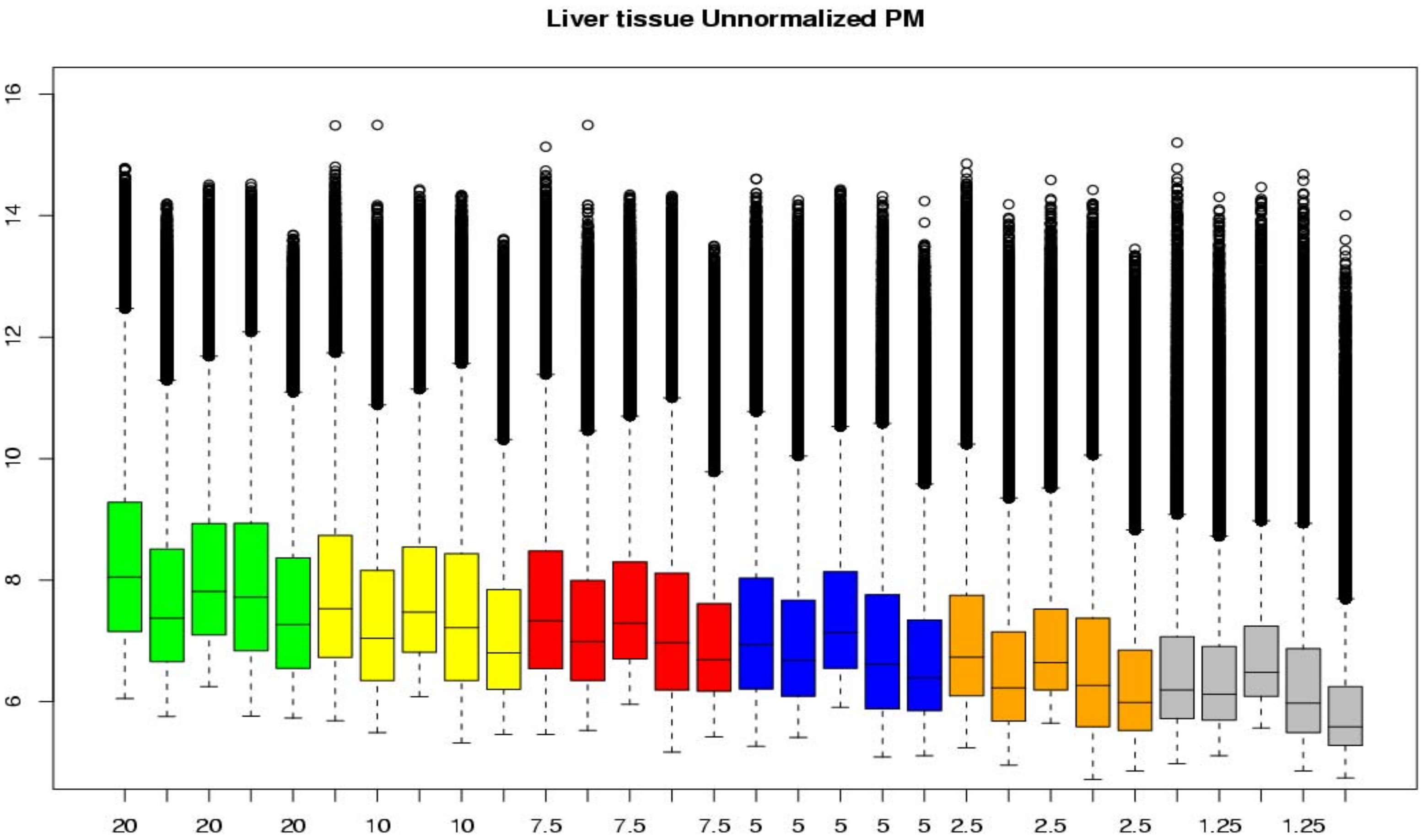
- An experiment was carried out where yeast RNA was hybridized to human chips, so all binding expected to be non specific.
- Fitted a model to predict log intensity from sequence composition gives base and position effects



- Uses these effects to predict an affinity for any given sequence call this A. The means of the distributions for the $N_{pm,}$ N_{mm} terms are functions of the affinities.

Non-Biological variability is a problem for single channel arrays

Log2 PM intensity



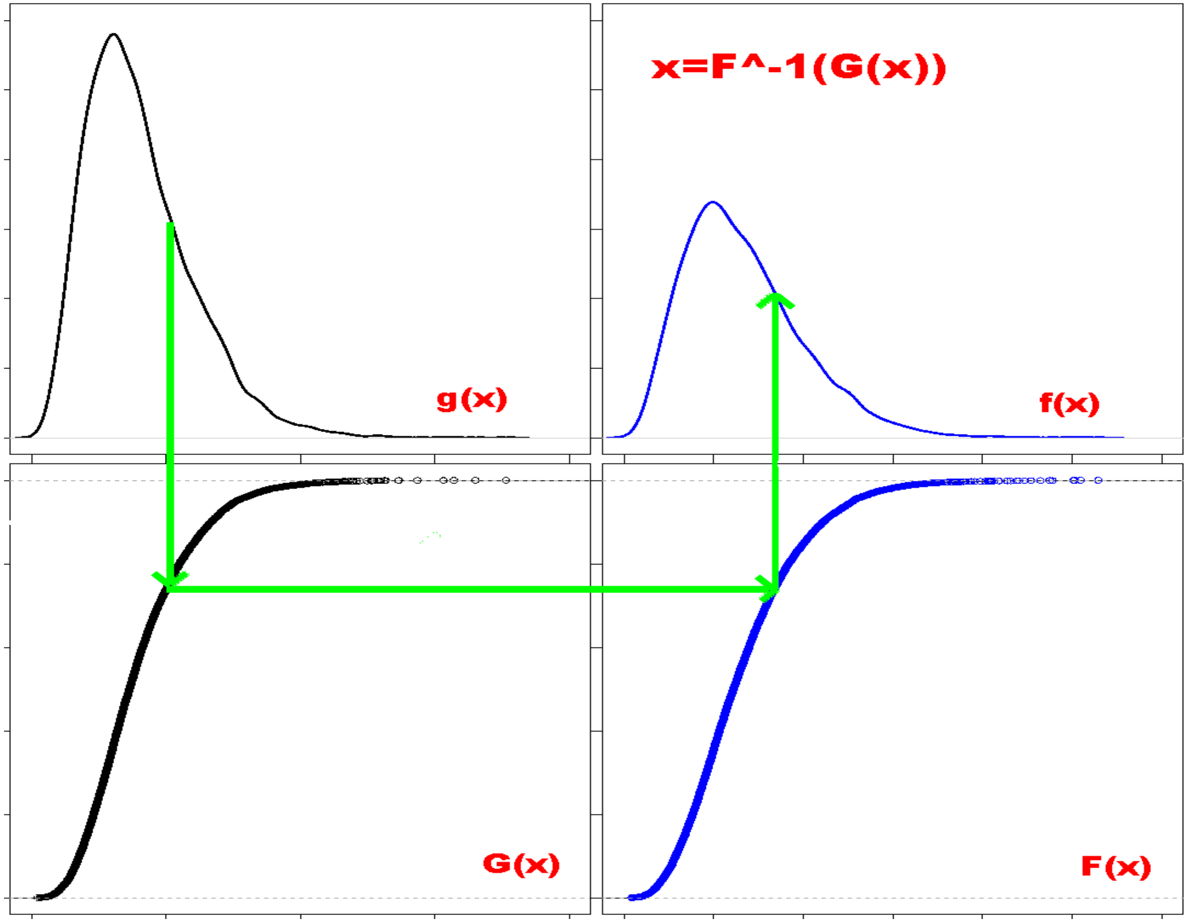
5 scanners for 6 dilution groups

Normalization

- In case of single channel microarray data this is carried out only across arrays.
- Could generalize methods we applied to two color arrays, but several problems:
  - Typically several orders of magnitude more probes on an Affymetrix array then spots on a two channel array
  - With single channel arrays we are dealing with absolute intensities rather than relative intensities.
- Need something fast

Quantile Normalization

- Normalize so that the quantiles of each chip are equal. Simple and fast algorithm. Goal is to give same distribution to each chip.

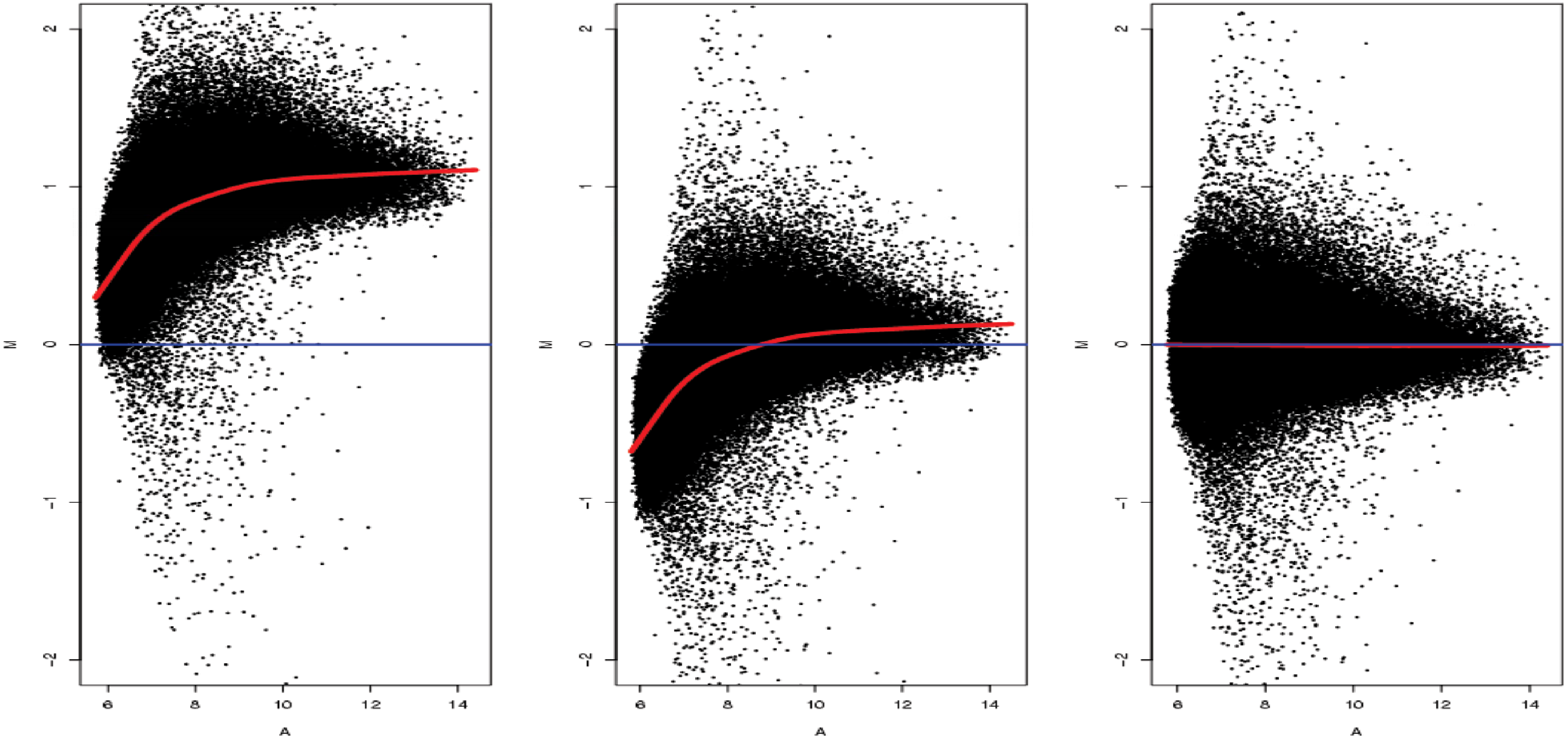


It works!!

Unnormalized

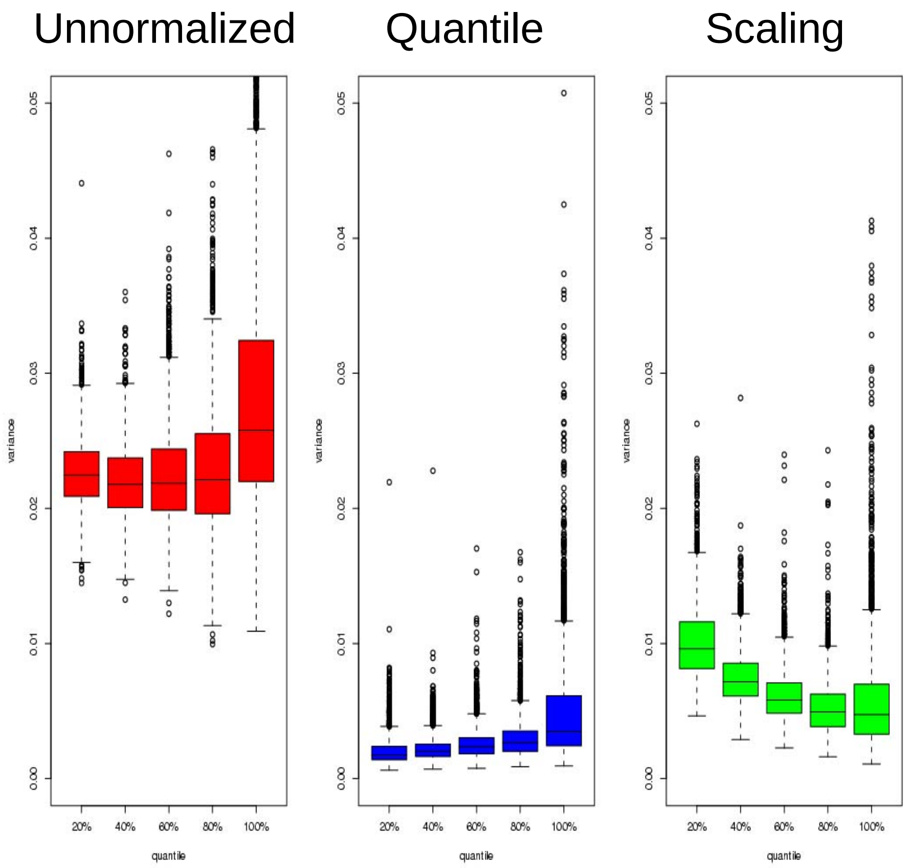
Scaling

Quantile Normalization

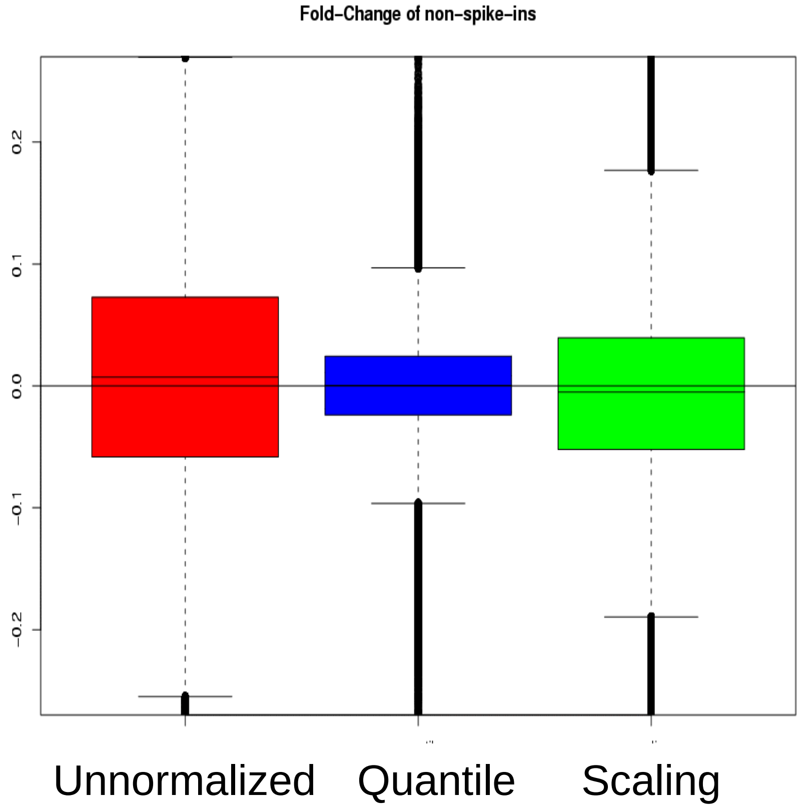


It Reduces Variability

Expression Values



Fold change



Also no serious bias effects. For more see Bolstad et al (2003)

Summarization

- Problem: Calculating gene expression values.
- How do we reduce the 11-20 probe intensities for each probeset on to a gene expression value?
- Our Approach
  - RMA – a robust multi-chip linear model fit on the log scale
- Some Other Approaches
  - Single chip
    - AvDiff (Affymetrix) – no longer recommended for use due to many flaws
    - Mas 5.0 (Affymetrix) – use a 1 step Tukey-biweight to combine the probe intensities in log scale
  - Multiple Chip
    - MBEI (Li-Wong dChip) – a multiplicative model on natural scale

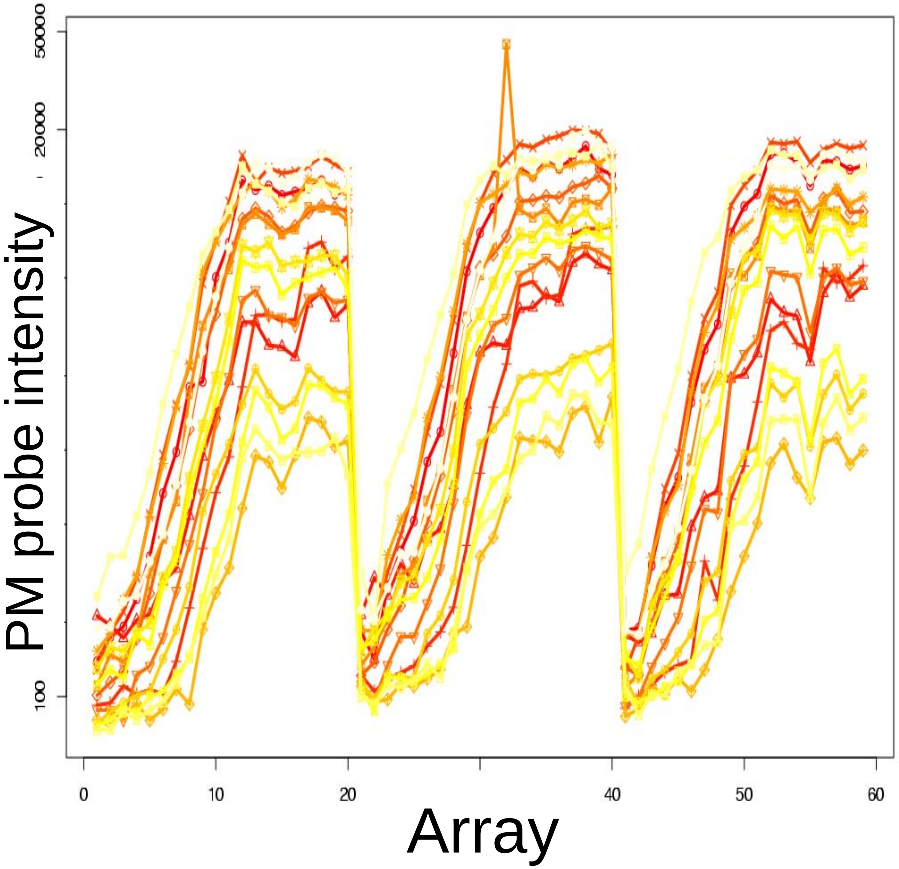
General Probe Level Model

$$y_{kij} = \mathrm{f}(\mathbf{X}) + \varepsilon_{kij}$$

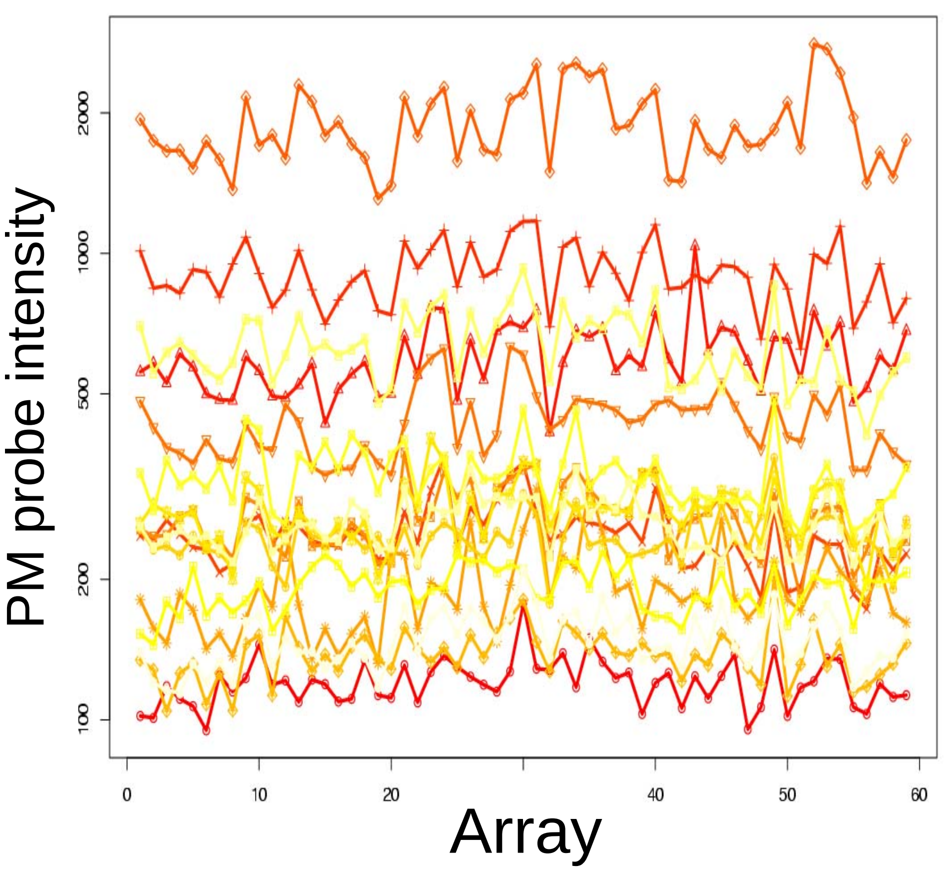
- Where f(X) is function of factor (and possibly covariate) variables (our interest will be in linear functions)
- y_{kij} is a pre-processed probe intensity (usually log scale)
- Assume that $\mathrm{Var}\left[\varepsilon_{kij}\right] = \sigma_k{}^2$

Parallel Behavior Suggests Multi-chip Model

Differentially expressing

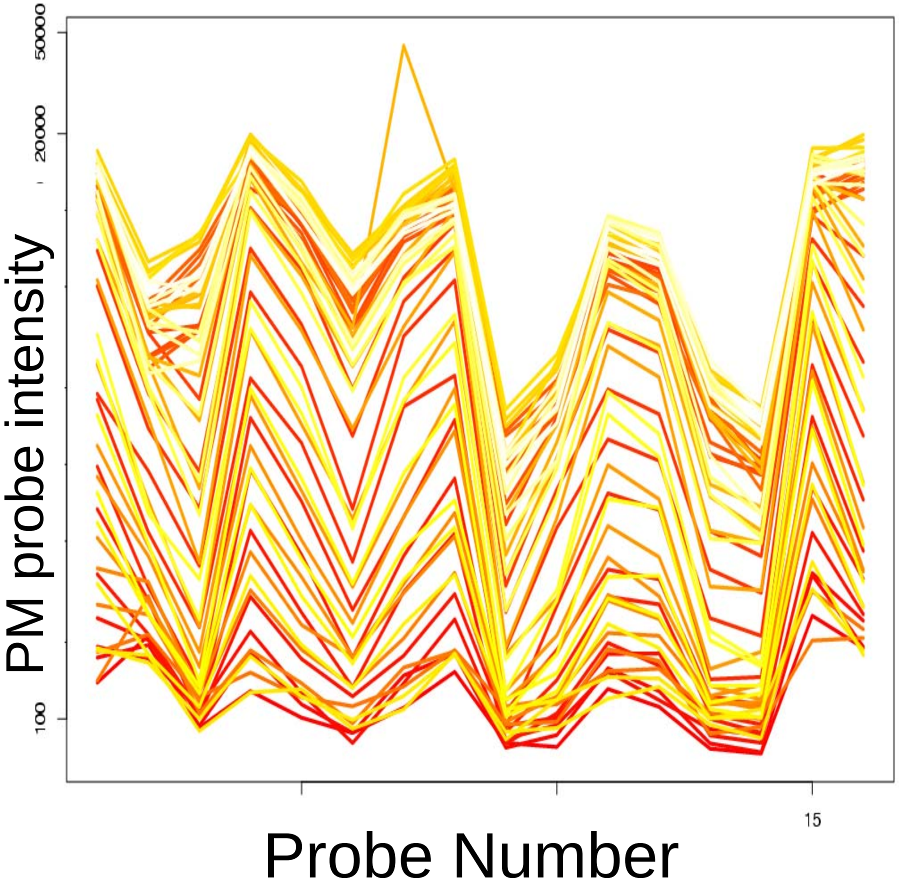


Non Differential

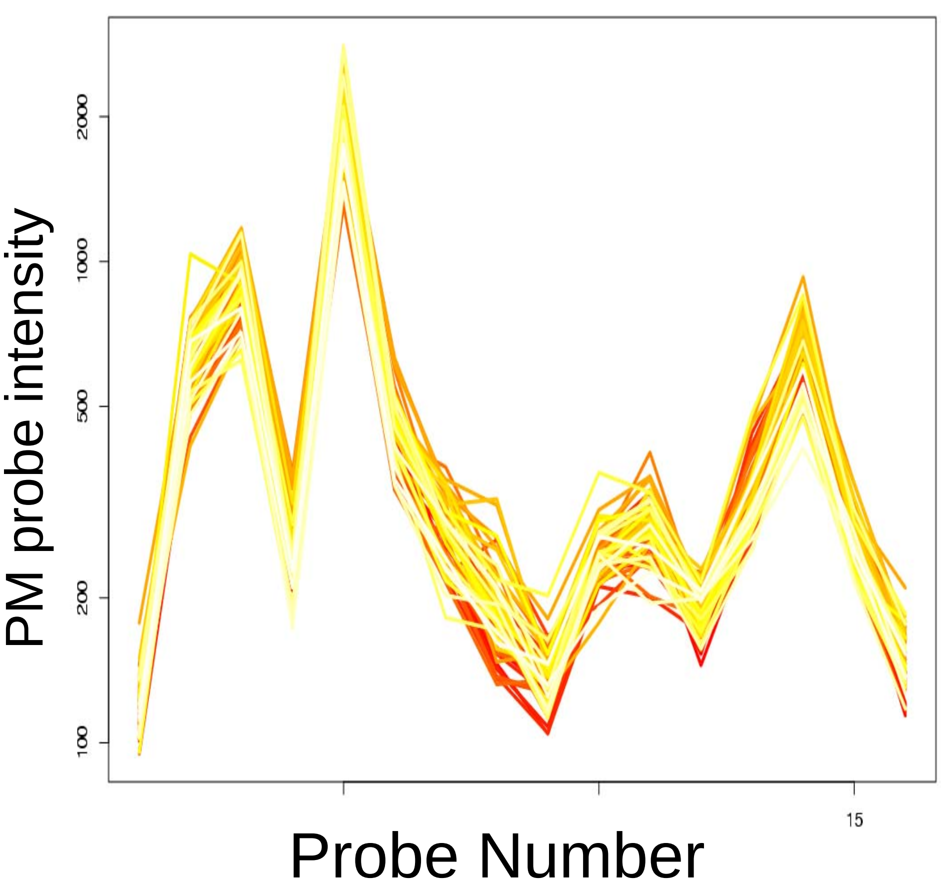


Probe Pattern Suggests Including Probe-Effect

Differentially expressing

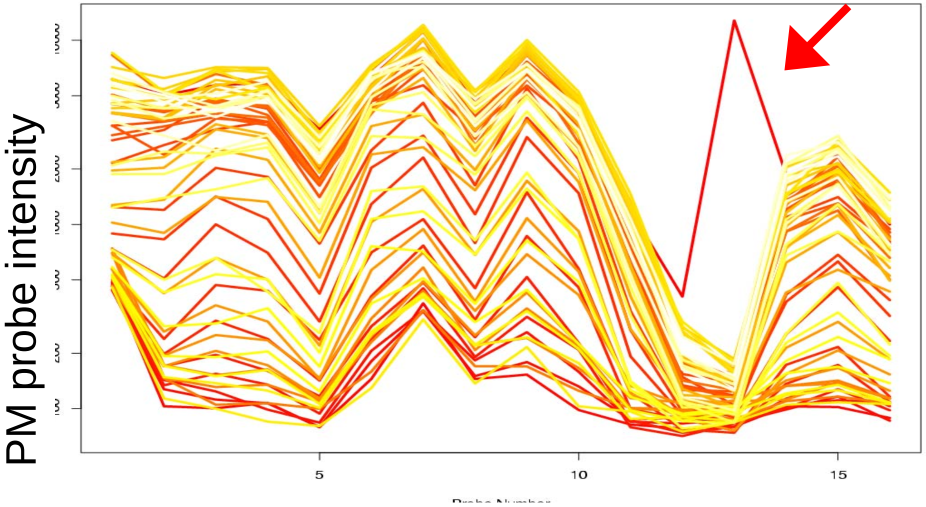


Non Differential

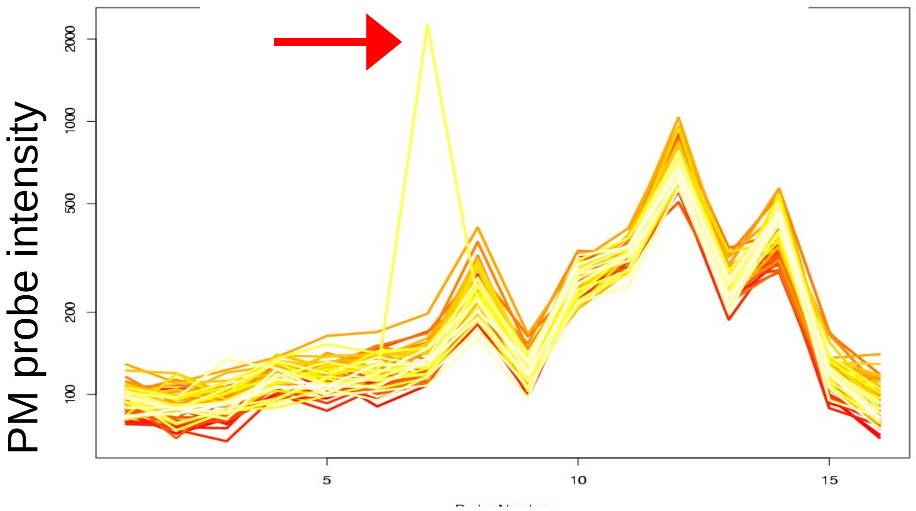


Also Want Robustness

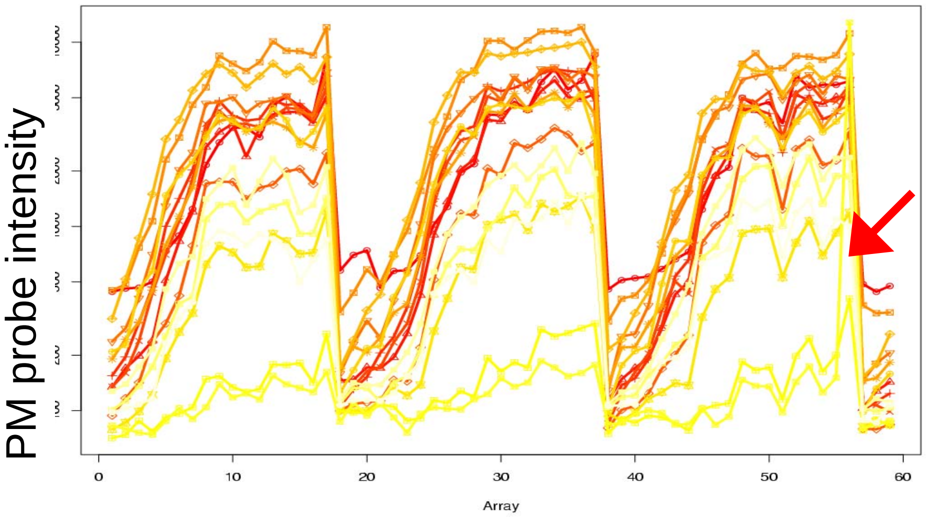
Differentially expressing



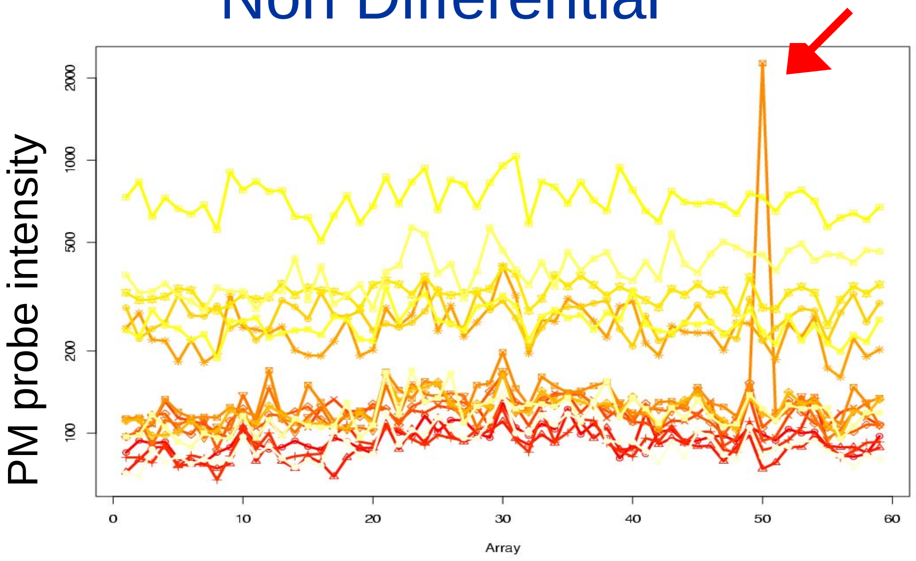
Non Differential



Differentially expressing



Non Differential



The RMA model

$$y_{kij} = m_k + \alpha_{ki} + \beta_{kj} + \varepsilon_{kij}$$

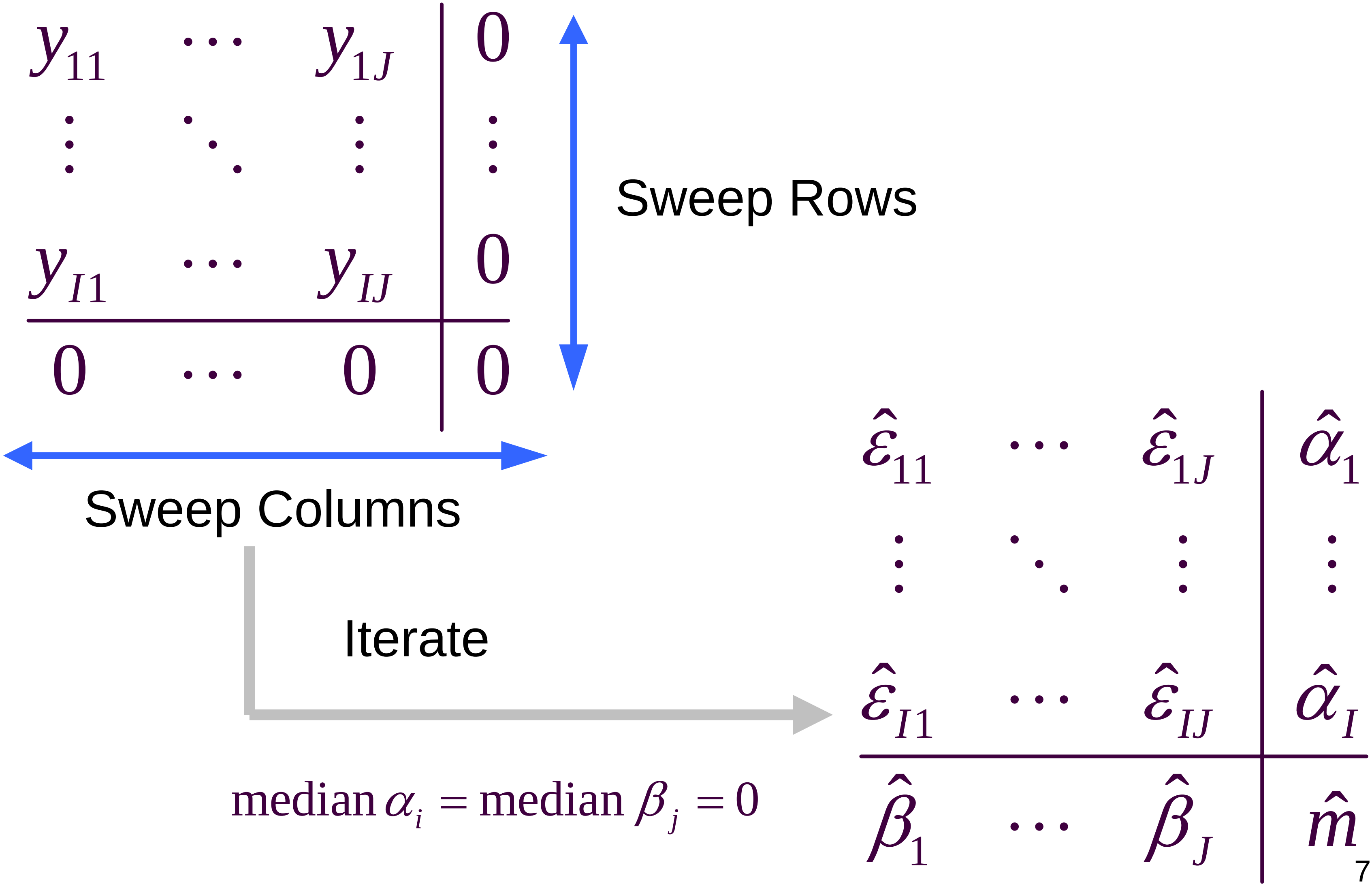
where $y_{kij} = \log_2 \mathrm{N}\left(\mathrm{B}\left(PM_{kij}\right)\right)$

α_{ki} is a probe-effect i= 1,…,I

β_{kj} is chip-effect ($m_k + \beta_{kj}$ is log2 gene expression on array *j)* j=1,…,J

k=1,…,K is the number of probesets

Median Polish Algorithm



$$\begin{array}{ccc|c} y_{11} & \cdots & y_{1J} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ y_{I1} & \cdots & y_{IJ} & 0 \\ \hline 0 & \cdots & 0 & 0 \end{array}$$

Sweep Rows

Sweep Columns

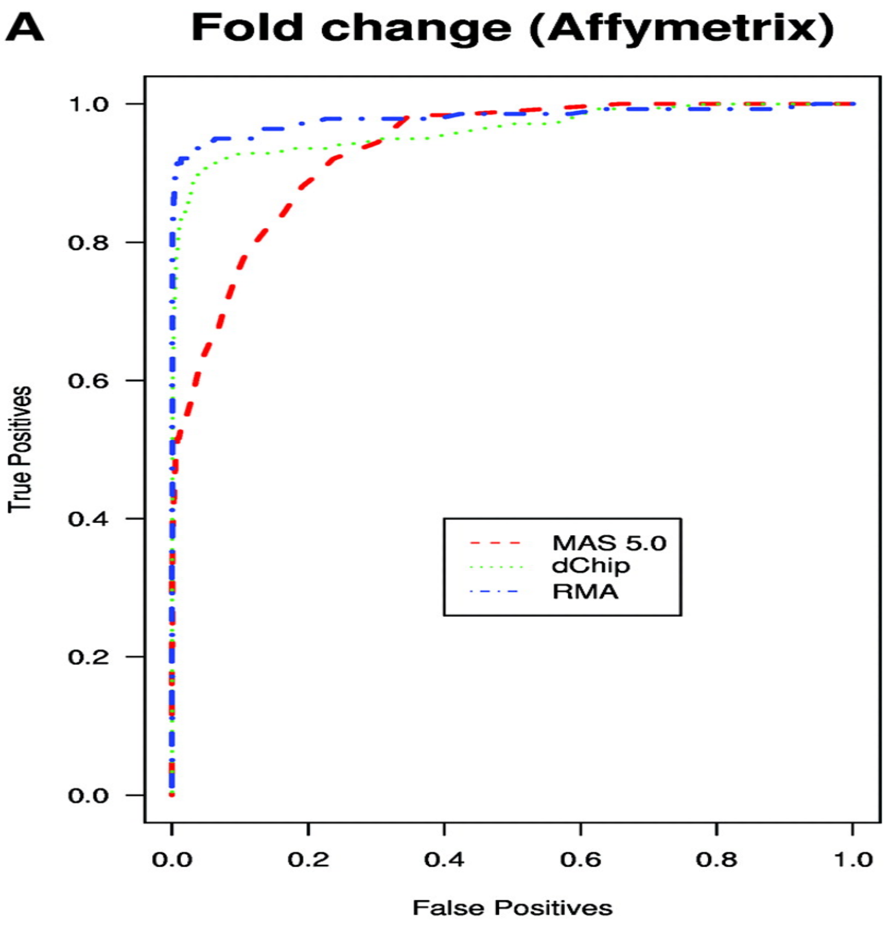
Iterate

$$\begin{array}{ccc|c} \hat{\varepsilon}_{11} & \cdots & \hat{\varepsilon}_{1J} & \hat{\alpha}_1 \\ \vdots & \ddots & \vdots & \vdots \\ \hat{\varepsilon}_{I1} & \cdots & \hat{\varepsilon}_{IJ} & \hat{\alpha}_I \\ \hline \hat{\beta}_1 & \cdots & \hat{\beta}_J & \hat{m} \end{array}$$

$$\text{median}\,\alpha_i = \text{median}\,\beta_j = 0$$

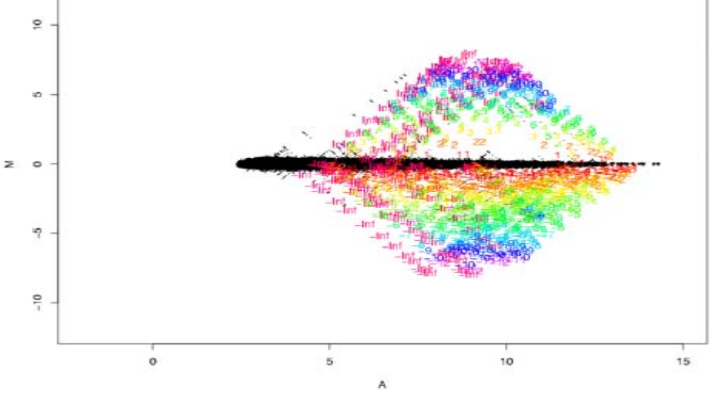
RMA mostly does well in practice

Detecting Differential Expression

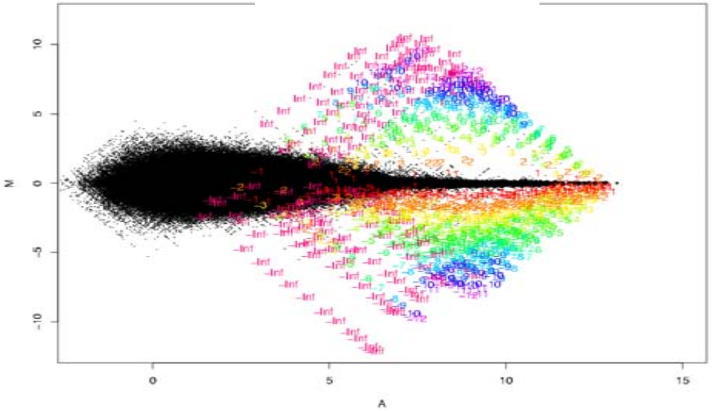


Not noisy in low intensities

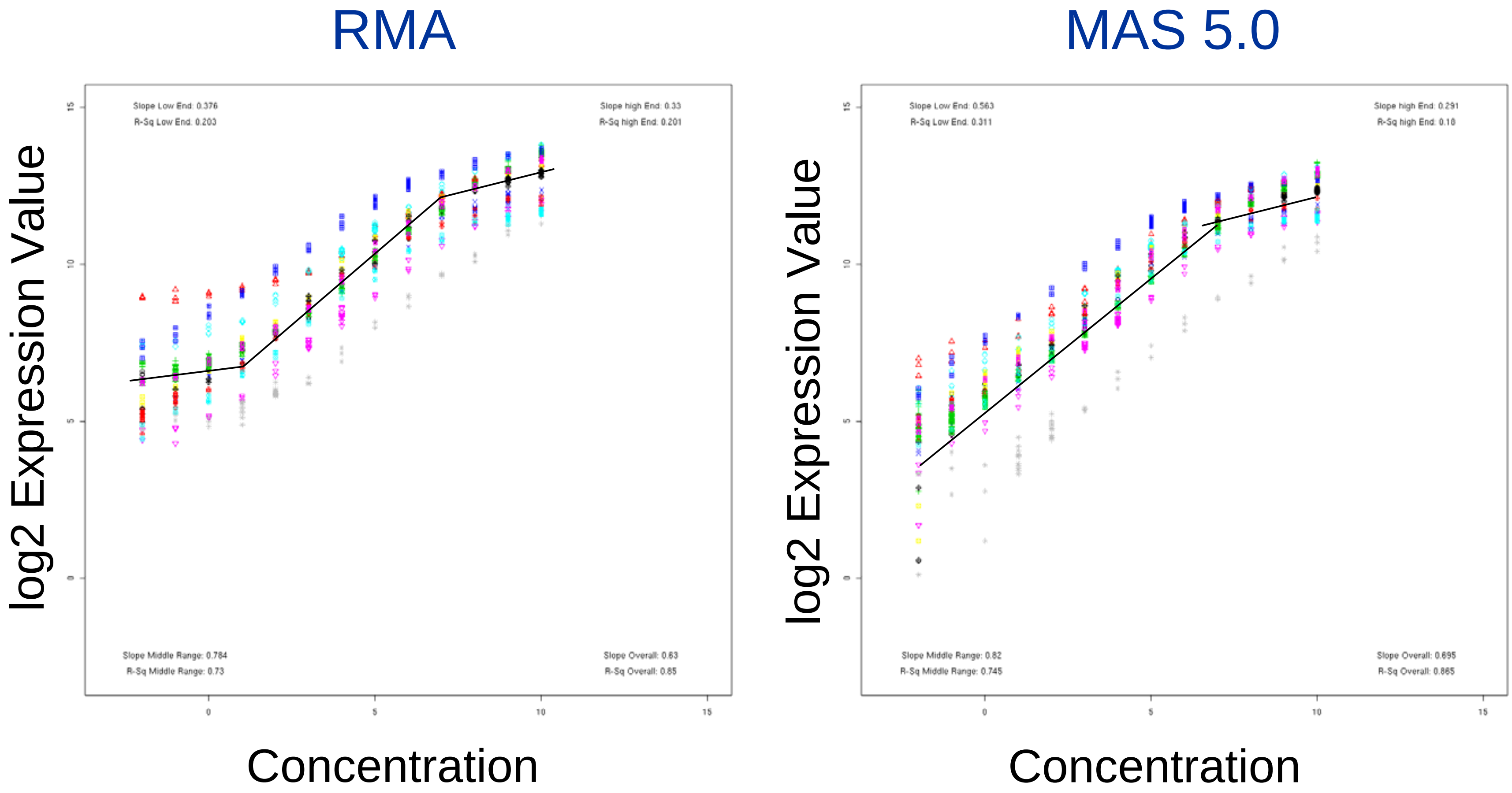
RMA



MAS 5.0

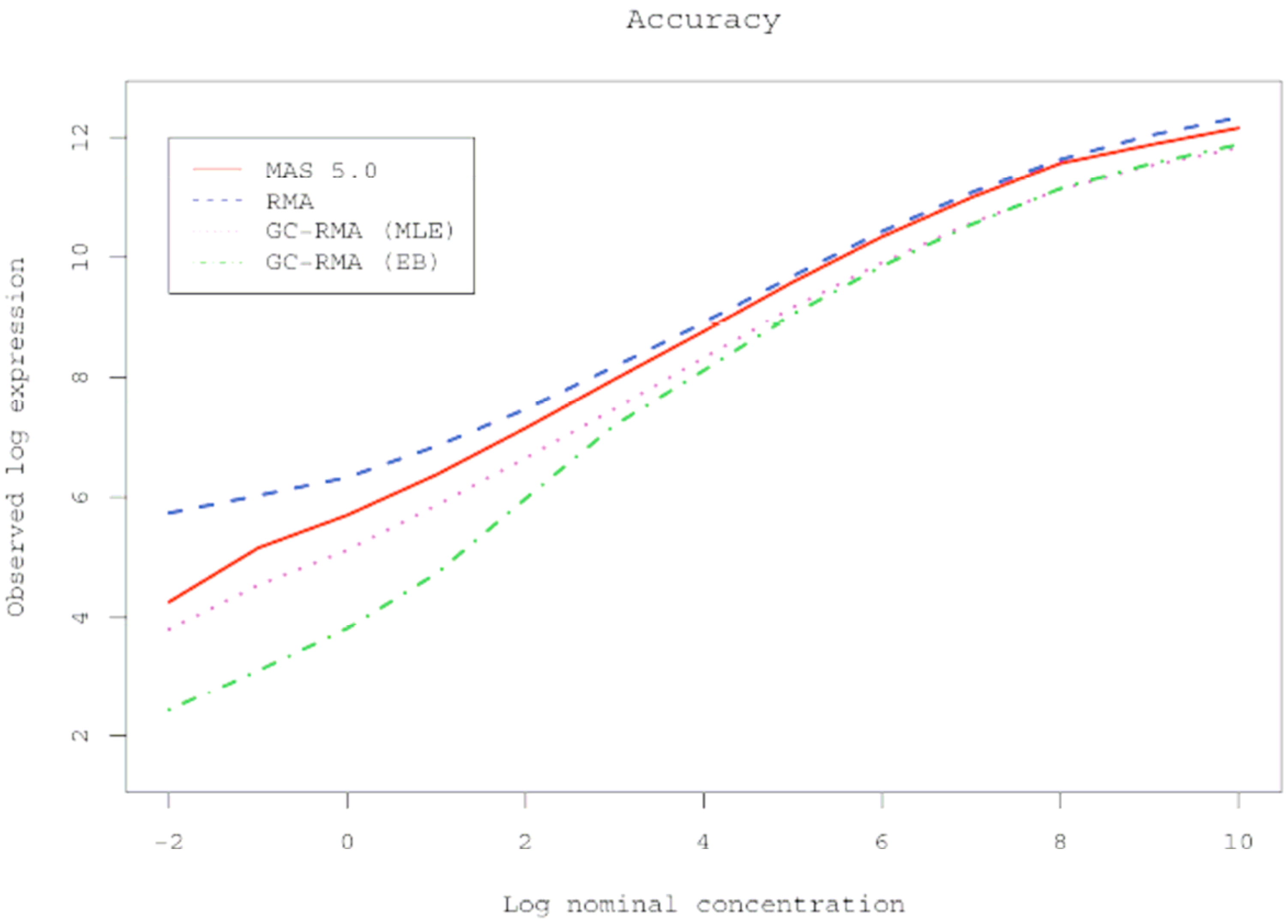


One Drawback



Linearity across concentration. GCRMA fixes this problem

GCRMA improve linearity



An Alternative Method for Fitting a PLM

- Robust regression using M-estimation
- In this talk, we will use Huber's influence function. The software handles many more.
- Fitting algorithm is IRLS with weights dependent on current residuals $\frac{\psi\left(r_{kij}\right)}{r_{kij}}$

Variance Covariance Estimates

- Suppose model is $Y = X\beta + \varepsilon$
- Huber (1981) gives three forms for estimating variance covariance matrix

$$\kappa^2 \frac{1/\left(n-p\right)\sum_i \psi\left(r_i\right)^2}{\left[1/n\sum_i \psi'\left(r_i\right)\right]^2}\left(X^T X\right)^{-1}$$

$$\kappa \frac{1/(n-p)\sum_i \psi\left(r_i\right)^2}{1/n\sum_i \psi'\left(r_i\right)} W^{-1}$$

We will use this form

$$\frac{1}{\kappa} 1/\left(n-p\right)\sum_i \psi\left(r_i\right)^2 W^{-1}\left(X^T X\right) W^{-1}$$

We Will Focus on the Summarization PLM

- Array effect model

Array Effect

$$y_{kij} = \alpha_{ki} + \beta_{kj} + \varepsilon_{kij}$$

Pre-processed
Log PM intensity

Probe Effect

With constraint

$$\sum_{i=1}^{I} \alpha_{ki} = 0$$

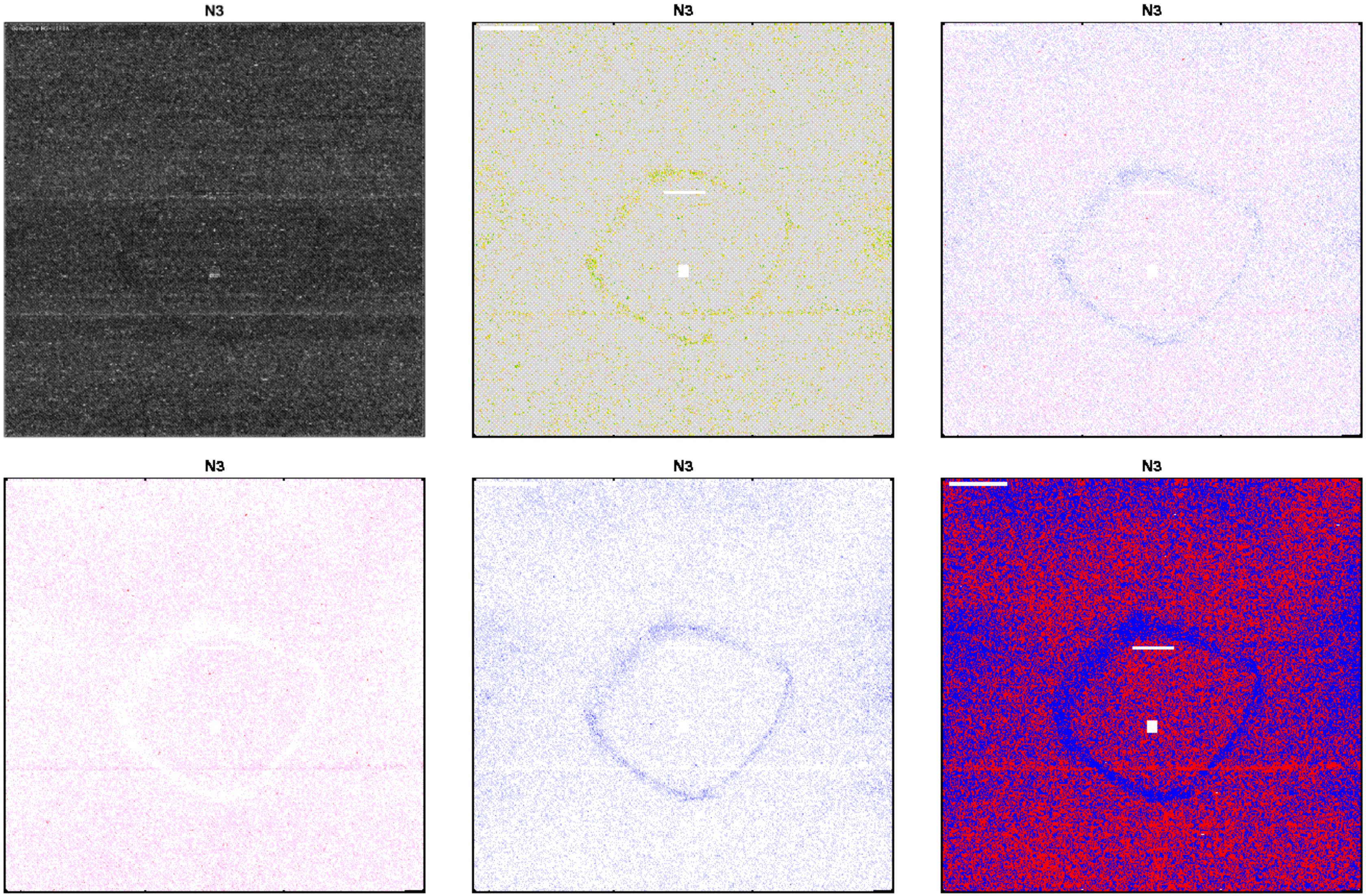
Quality Assessment

- Problem: Judge quality of chip data

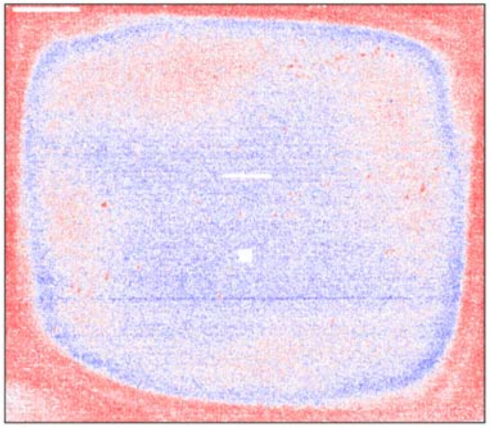
- Question: Can we do this with the output of the Probe Level Modeling procedures?

- Answer: Yes. Use weights, residuals, standard errors and expression values.

Chip pseudo-images

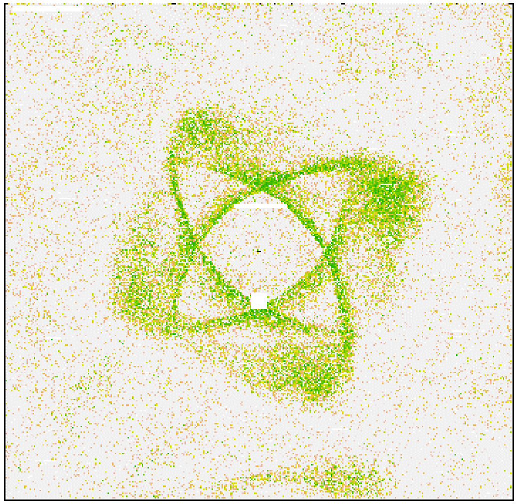


An Image Gallery

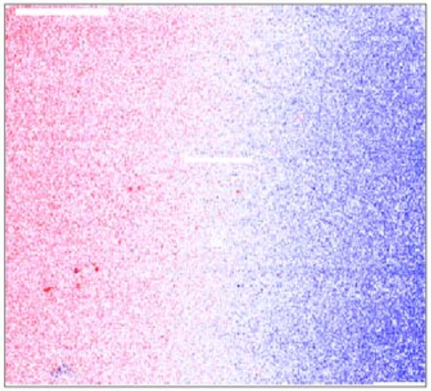


“Ring of Fire”

“Crop Circles”



“Tricolor”



http://PLMImageGallery.bmbolstad.com

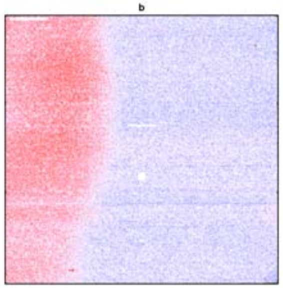
NUSE Plots

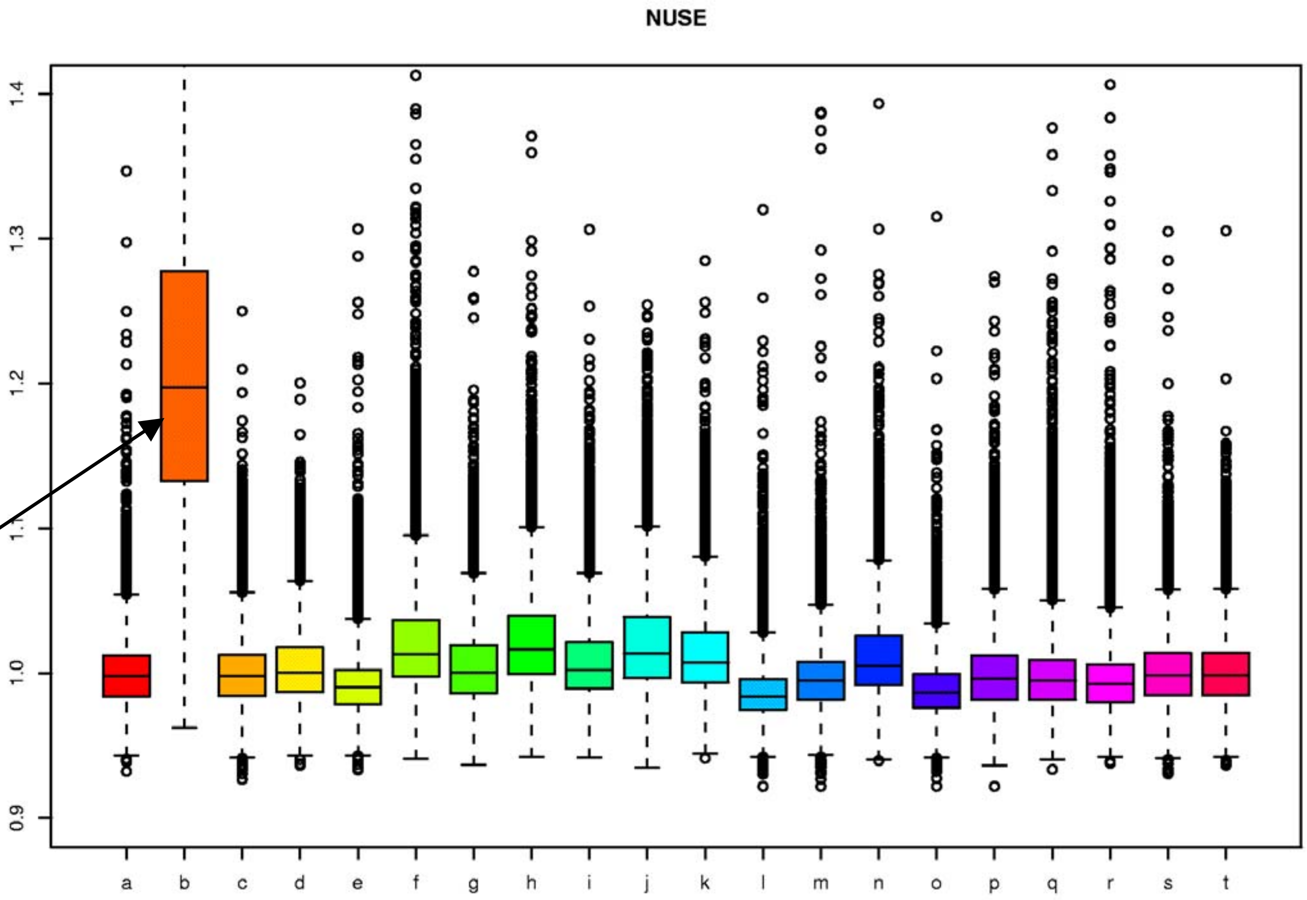
Normalized

Unscaled

Standard

Errors



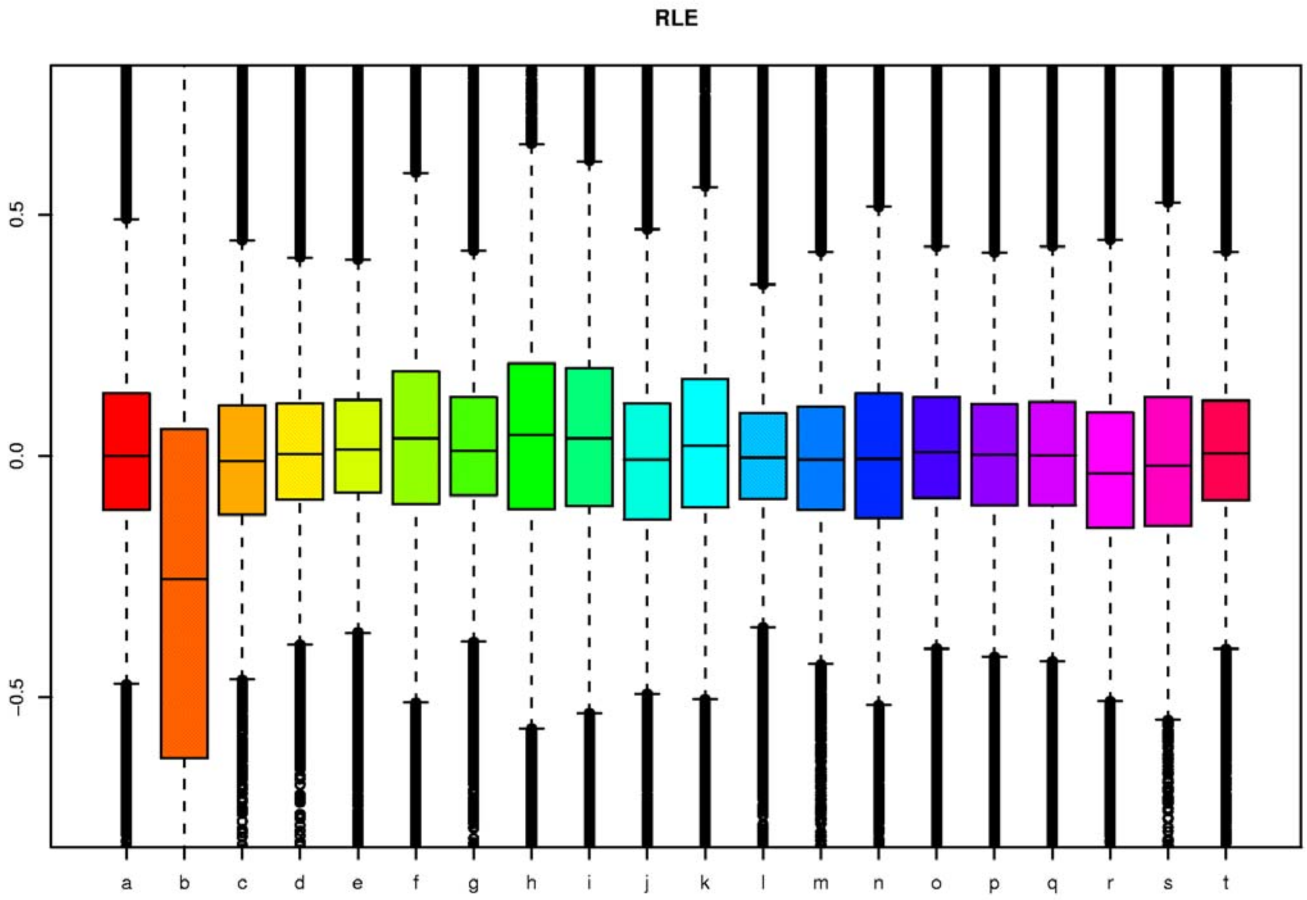


RLE Plots

Relative

Log

Expression



Summary of One Channel Arrays

- Background correction
  - RMA model
  - GCRMA model
- Normalization
  - Quantile normalization
- Summarization
  - Robust multi-chip probe level modeling
- Quality Assessment

Acknowledgements

- Terry Speed
- Rafael Irizarry
- Julia Brettschneider
- Francois Colin
- Jean Yang
- Zhijin (Jean) Wu
- Gordon Smyth
- James Westenhall
- Any one else …

References

- 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.
- Yang, Y. H., Buckley, M. J., Dudoit, S., and Speed, T. P. (2002). Comparison of methods for image analysis on cDNA microarray data. Journal of Computational and Graphical Statistics, 11 (1), 108-136.
- Smyth, G. K., Thorne, N. P. and Wettenhall J. (2004) limma: Linear Models for Microarray Data User's Guide. The Walter and Eliza Hall Institute of Medical Research.
- Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics, 19, 185 (2003).
- Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, and Speed TP. Summaries of Affymetrix GeneChip Probe Level Data. Nucleic Acids Research, 31(4):e15, 2003.
- Bolstad BM, Collin F, Brettschneider J, Simpson K, Cope L, Irizarry RA, and Speed TP. (2005) Quality Assessment of Affymetrix GeneChip Data in Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Gentleman R, Carey V, Huber W, Irizarry R, and Dudoit S. (Eds.), Springer
- Wu, Z., Irizarry, R., Gentleman, R., Martinez Murillo, F. Spencer, F. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of American Statistical Association 99, 909-917 (2004)