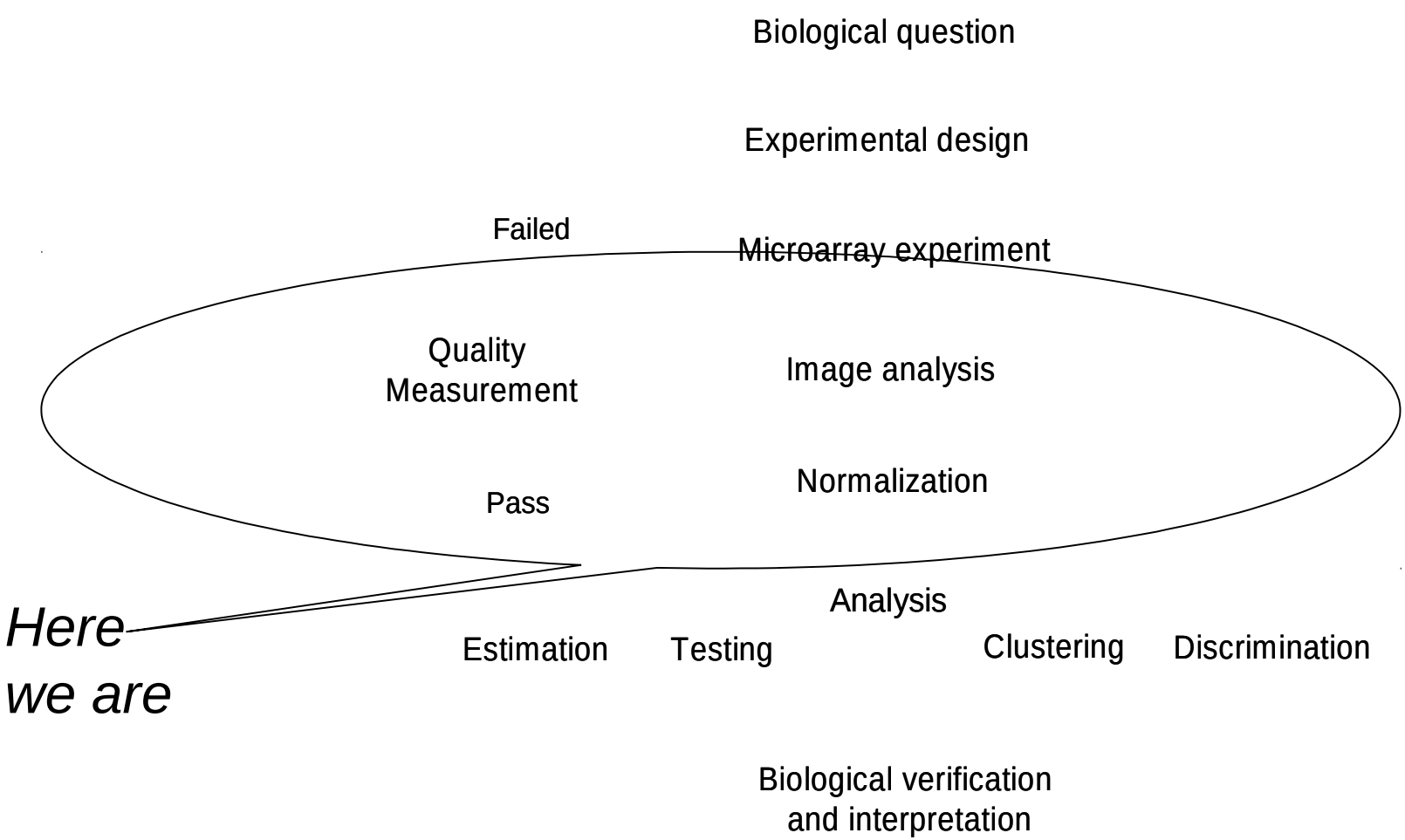
FILTERING AND PREPROCESSING MICROARRAY DATA

- Filtering
- Background correction
- Normalization
- Summarization

Microarray studies life cycle

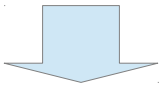


Filtering (1)

There may be errors during hybridization and/or scanning which yield bad spots → These are automatically flagged

Many spots may show very low signals:

- Problems with spotting
- No hybridization in this spot



Bad spots may be removed from the analysis to avoid unnecessary noise (but, some people prefer not to filter to avoid eliminating “good spots” unintentionally)

Filtering (2): Ajust signal

We may filter the data on intensity

- by excluding values where both the red and green channels are less than 100
- by setting the value of an intensity to the minimum in the event only one of the two channel intensities is below the minimum of 100

We may use the flag column imported with the data, and exclude intensities with a flag value not equal to 1

Normalization:Objective

- Achieve a measurement scale such that
  - It has the same origin (zero or other) for all spots
  - It uses the same unit for all spots and microarrays
  - It has a linear relationship with the DNA/RNA biological
  - It has good statistical properties (good for later analyses)

- Deal with the particular characteristics of each platform and experiment
  - Color differences
  - Reference sample
  - Summarize information of each gene
  - Deal with platform characteristics (e.g. “probesets/probepairs”)

Hypotheses

- Most normalization methodologies make two major assumptions about the data.
  - When comparing different samples, only few genes are over-expressed or under-expressed in one array relative to the others.
  - The number of genes over-expressed in a condition is similar to the number of genes under-expressed.
- This assumptions should agree with your experimental context.

General Steps

- Background correction (correcting the scale origin for spots)
- Normalization (standardizing the scale unit - rescaling)
- Adjustments characteristics of each platform or experiment
  - Perfect-Match Mismatch Adjustment (Affymetrix)
  - Correcting for different dye properties (in two color arrays)
  - Adjustments depending on the DNA strands
- Summary of information from several spots into a single measure for each gene
  - Averaging Affymetrix ”probe sets”
  - Averaging duplicated spots
  - Calculating ratios
  - Taking logarithms

Preprocessing two color data

Preprocessing two color data

- Background correction
  - Scanners: separate Signal (Rs, Gs) and Background (Rb, Gb) estimates.
  - Background corrected estimates (Rc, Gc)
    - Rc = Rs – Rb, Gc = Gs – Gb, OR (better)
    - Rc = max(Rs -Rb, 0), Gc = max(Gs -Gb, 0)
- Summarization & Transforms: log-Ratios
  - Estimate relative expression as log(Rc/Gc)

Normalization for two color arrays

- Main source of variation: dye absorbtion
  - Cy3 (G) & Cy5 (Red) absorbed differently.
  - Shown in scatterplots/MAplots
  - WITHIN-ARRAY normalization
    - Estimate transforms from plotted data
- Other sources : different intensities in scanned microarrays.
  - Scale arrays so that they have similar
    - Mean intensity, intensity dispersion
  - BETWEEN ARRAY NORMALIZATION

Global normalization

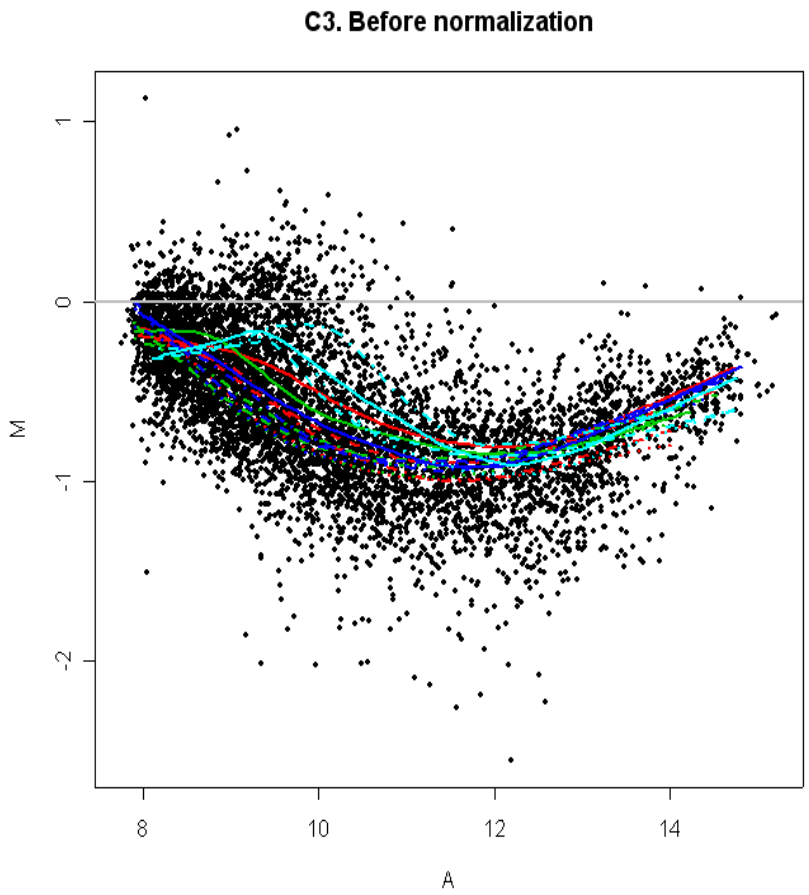
- Based on a *global adjustment*

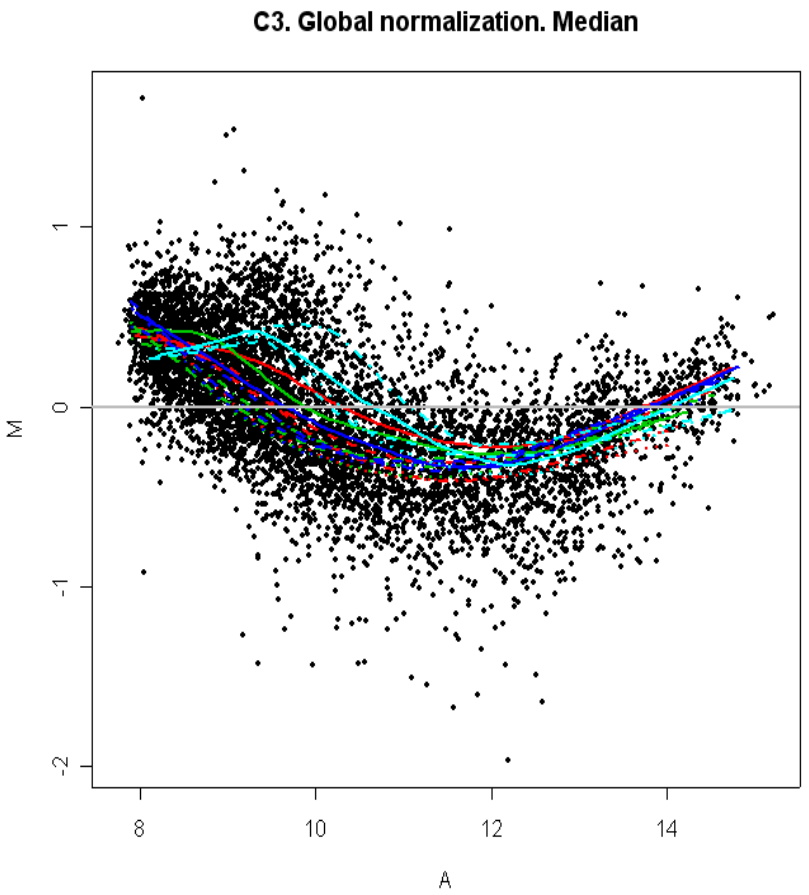
$$log_2 R/G \rightarrow log_2 R/G - c = log_2 R/(kG)$$

- Choices for k or $c = log_2 k$ are
  - c = *median* or *mean* of log ratios for a particular gene set (e.g. control or housekeeping genes)
  - Total intensity normalization, where

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

Example: (Callow et al 2002)
Global median normalization.





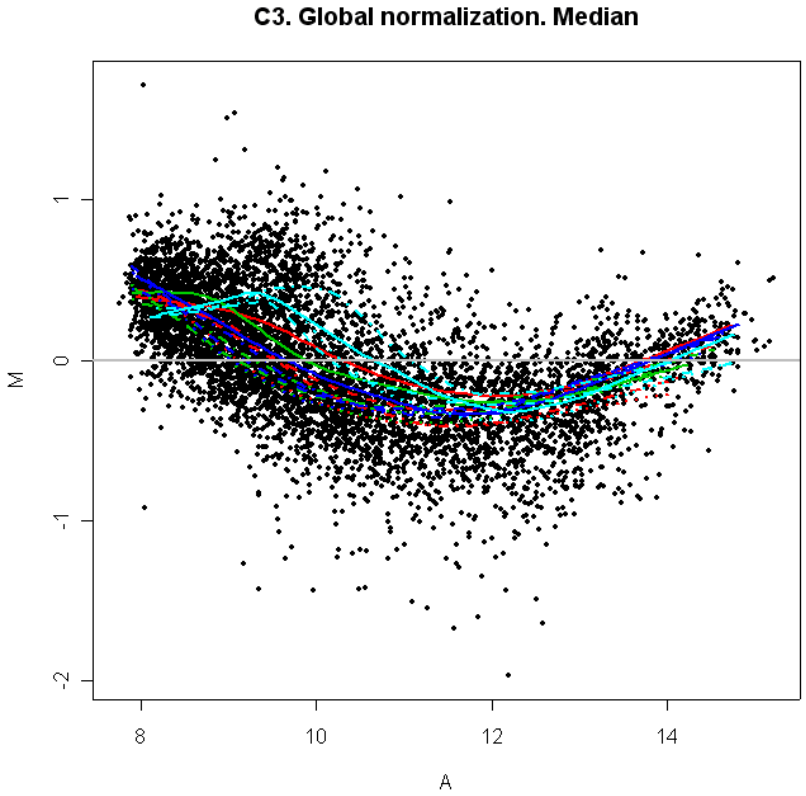
Intensity-dependent normalization

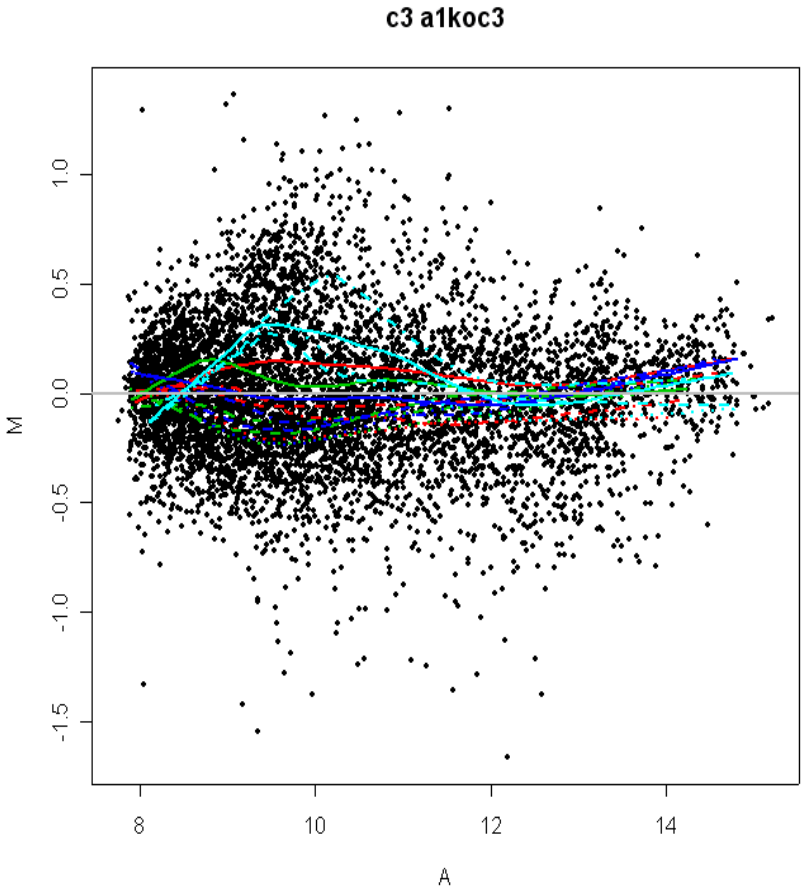
- Run a line through the middle of the MA plot, shifting the M value of the pair (A,M) by c=c(A), i.e.

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

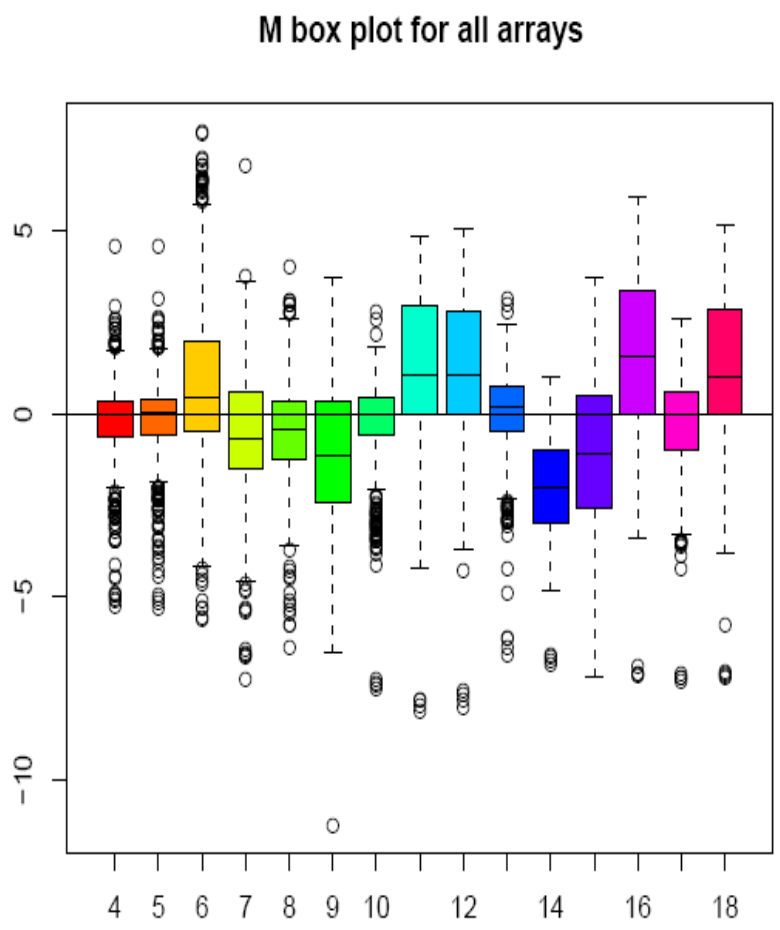
- One estimate of c(A) is made using the LOWESS function of Cleveland (1979): LOcally WEighted Scatterplot Smoothing.

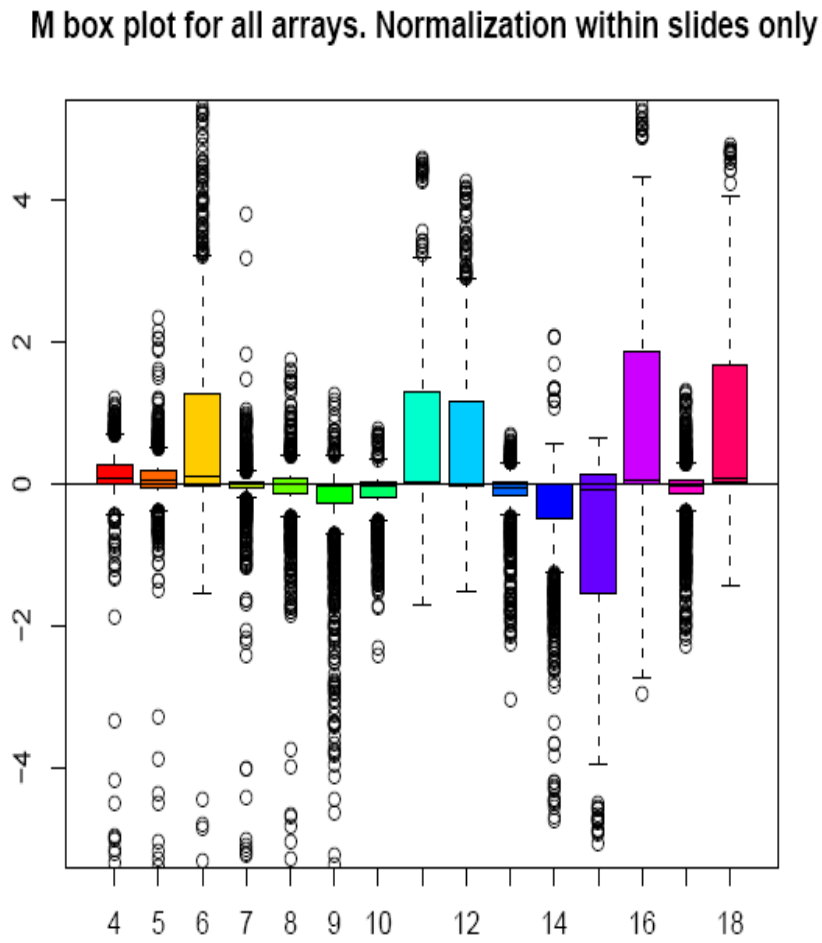
Example: (Callow et al 2002) loess vs median normalization.





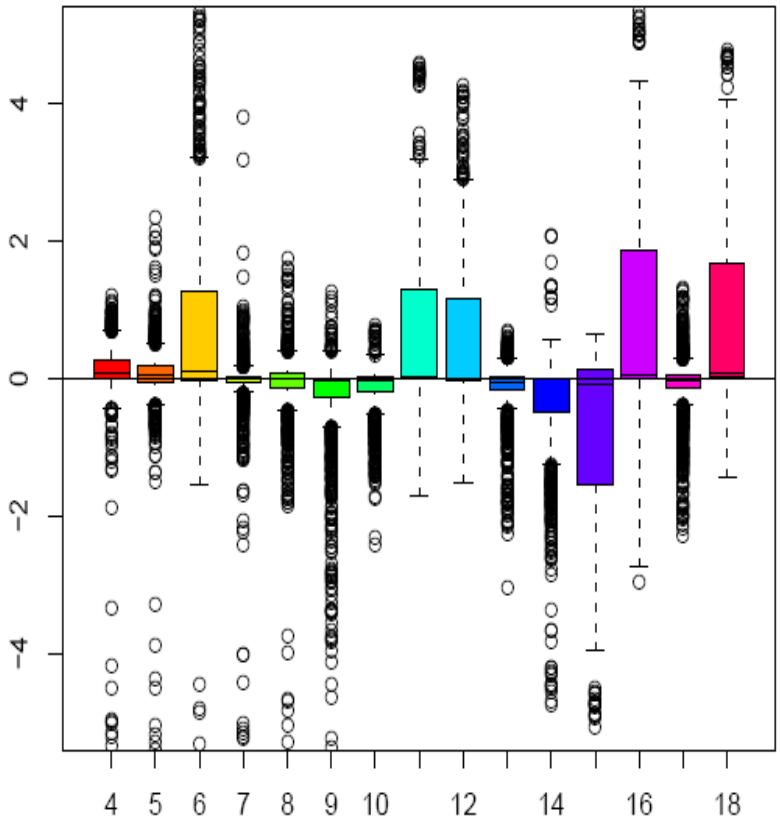
Effect of within-slide normalization



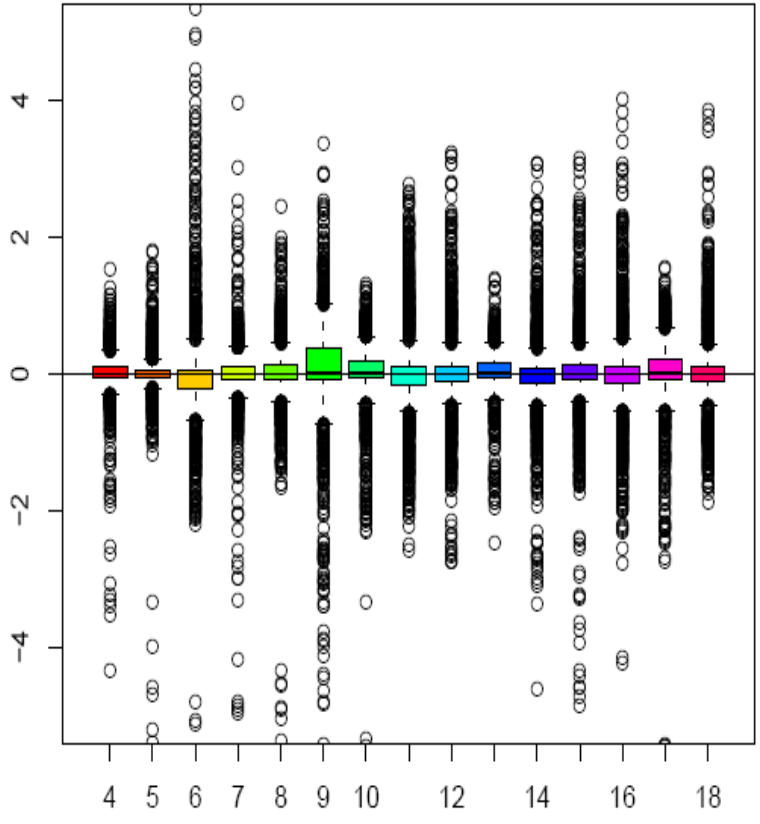


Effect of between-slide normalization

M box plot for all arrays. Normalization within slides only



M box plot for all arrays. Normalization within& between slide



Preprocessing one color data

Preprocessing one color data

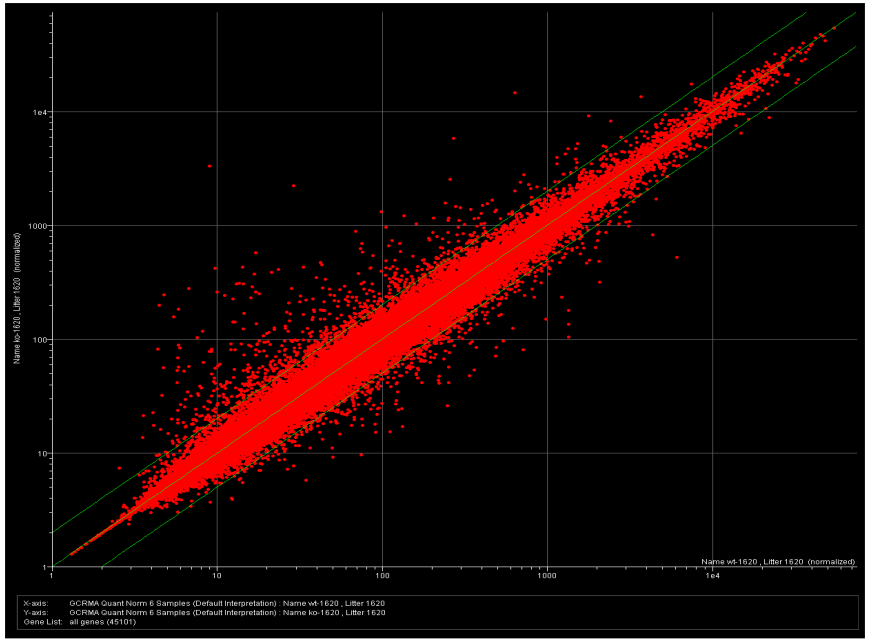
- The Big Four algorithms for correcting, normalizing,
- and summarizing probe level data
- ● MAS 5: Microarray Analysis Suite 5.0, Affymetrix [1, 2]
- ● RMA: Robust Multichip Analysis, Irizarry et al. [3, 4]
- ● dChip: Model Based Expression Index, Li and Wong [5, 6]
- ● SAM: Significance Analysis of Microarrays, Tusher et al.

Robust Multiarray Average (RMA)

Subtraction of MM data corrects for NSB, but introduces noise.

Want a method that gives positive intensity values.

Normalising at probe level avoids the loss of information.



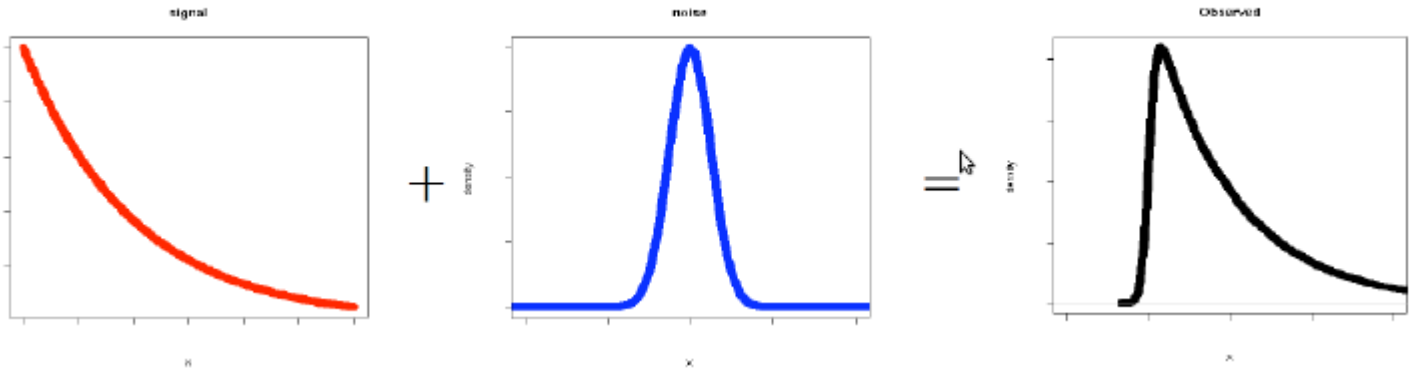
Robust Multiarray Average (RMA)

1) **Background correction.**
2) Normalization (across arrays).
3) Probe set summarization.

RMA: Backgroud correction

Assumes PM data is combination of background and signal

- *PM = Signal + Background*, where
  - Signal: $S \sim exp(\lambda)$ and
  - Background: $B \sim N(\mu,\sigma^2)$



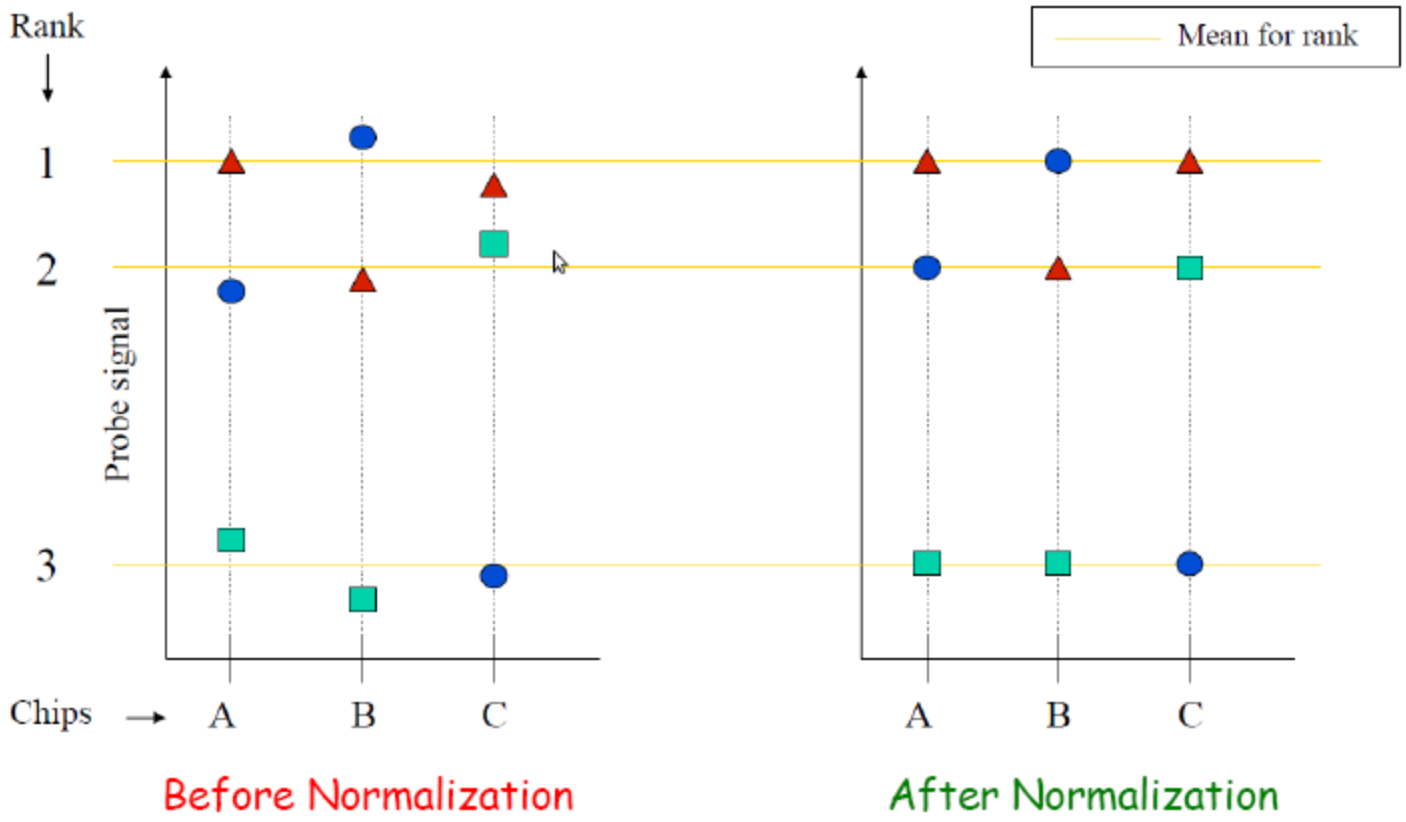
RMA

1) Background correction.
2) **Normalization (across arrays).**
3) Probe set summarization.

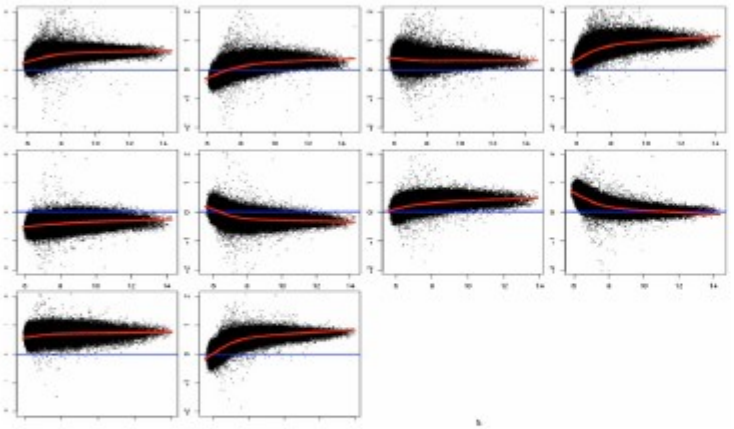
RMA: Normalization

- The purpose of normalization is to remove artifact differences between arrays (e.g. differences in total intensity)
- Quantile normalization makes the empirical distribution of probe intensities the same for every chip
- The common distribution is obtained by averaging each quantile across chips
- RMA normalization reduces the variability without loosing the ability to detect differential expression

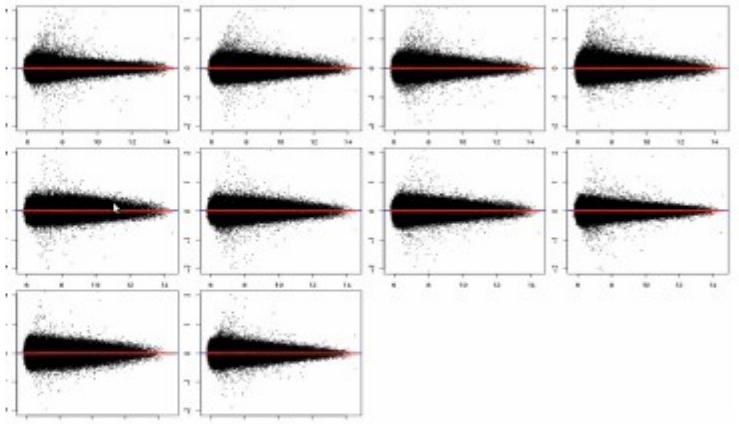
RMA:Quantile normalization outlined



MA-plots before and after RMA



Before Normalization



After Normalization

RMA

1) Background correction.
2) Normalization (across arrays).
3) **Probe set summarization.**

RMA: summarization (1)

After quantile normalization across chips at probe level, assess the probeset value (i.e. gene expression)

- Given a probeset, PM values are based on the linear model

$$\log_2 Normalized\,(PM_{ij} - BG) = \alpha_i + \beta_j + \epsilon_{ij}$$

α_i chip-specific contribution (chip-specific probeset value we are interested in)

β_j probe-specific contribution (probe intrinsic, "explainable" variability)

ϵ_{ij} unexplainable noise contribution

Estimate chip effects (log expression) α_i and probe effects β_j by using a robust method

- Median polish, which is quick
- Robust linear model, which yields quality diagnostics

RMA: summarization(2)

- Combine intensity values from the probes in the probe set to get a single intensity value for each gene (*probeset*).
- Uses '*Median Polishing*'.
  - Each chip normalised to its median.
  - Each gene normalised to its median.
  - Repeated until medians converge.
  - Maximum of 5 iterations to prevent infinite loops.

An Example

Suppose the following are background-adjusted, $\log_2$-transformed, quantile-normalized PM intensities for a single probe set. Determine the final RMA expression measures for this probe set.

GeneChip	Probe 1	2	3	4	5
1	4	3	6	4	7
2	8	1	10	5	11
3	6	2	7	8	8
4	9	4	12	9	12
5	7	5	9	6	10

An Example (continued)

4	3	6	4	7	4
8	1	10	5	11	8
6	2	7	8	8	7
9	4	12	9	12	9
7	5	9	6	10	7

row medians

0	-1	2	0	3
0	-7	2	-3	3
-1	-5	0	1	1
0	-5	3	0	3
0	-2	2	-1	3

matrix after removing row medians

An Example (continued)

0	-1	2	0	3
0	-7	2	-3	3
-1	-5	0	1	1
0	-5	3	0	3
0	-2	2	-1	3

0 -5 2 0 3

column medians

0	4	0	0	0
0	-2	0	-3	0
-1	0	-2	1	-2
0	0	1	0	0
0	3	0	-1	0

matrix after subtracting column medians

An Example (continued)

0	4	0	0	0	0
0	-2	0	-3	0	0
-1	0	-2	1	-2	-1
0	0	1	0	0	0
0	3	0	-1	0	0

row medians

0	4	0	0	0
0	-2	0	-3	0
0	1	-1	2	-1
0	0	1	0	0
0	3	0	-1	0

matrix after removing row medians

An Example (continued)

0	4	0	0	0
0	-2	0	-3	0
0	1	-1	2	-1
0	0	1	0	0
0	3	0	-1	0

0	1	0	0	0

column medians

0	3	0	0	0
0	-3	0	-3	0
0	0	-1	2	-1
0	-1	1	0	0
0	2	0	-1	0

matrix after subtracting column medians

An Example (continued)

0	3	0	0	0
0	-3	0	-3	0
0	0	-1	2	-1
0	-1	1	0	0
0	2	0	-1	0

All row medians and column medians are 0. Thus the median polish procedure has converged. This above is the residual matrix that we will subtract from the original matrix to obtain the fitted values.

An Example (continued)

original matrix

$$\begin{pmatrix} 4 & 3 & 6 & 4 & 7 \\ 8 & 1 & 10 & 5 & 11 \\ 6 & 2 & 7 & 8 & 8 \\ 9 & 4 & 12 & 9 & 12 \\ 7 & 5 & 9 & 6 & 10 \end{pmatrix}$$

residuals from median polish

$$- \begin{pmatrix} 0 & 3 & 0 & 0 & 0 \\ 0 & -3 & 0 & -3 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 2 & 0 & -1 & 0 \end{pmatrix}$$

matrix of fitted values

$$= \begin{pmatrix} 4 & 0 & 6 & 4 & 7 \\ 8 & 4 & 10 & 8 & 11 \\ 6 & 2 & 8 & 6 & 9 \\ 9 & 5 & 11 & 9 & 12 \\ 7 & 3 & 9 & 7 & 10 \end{pmatrix}$$

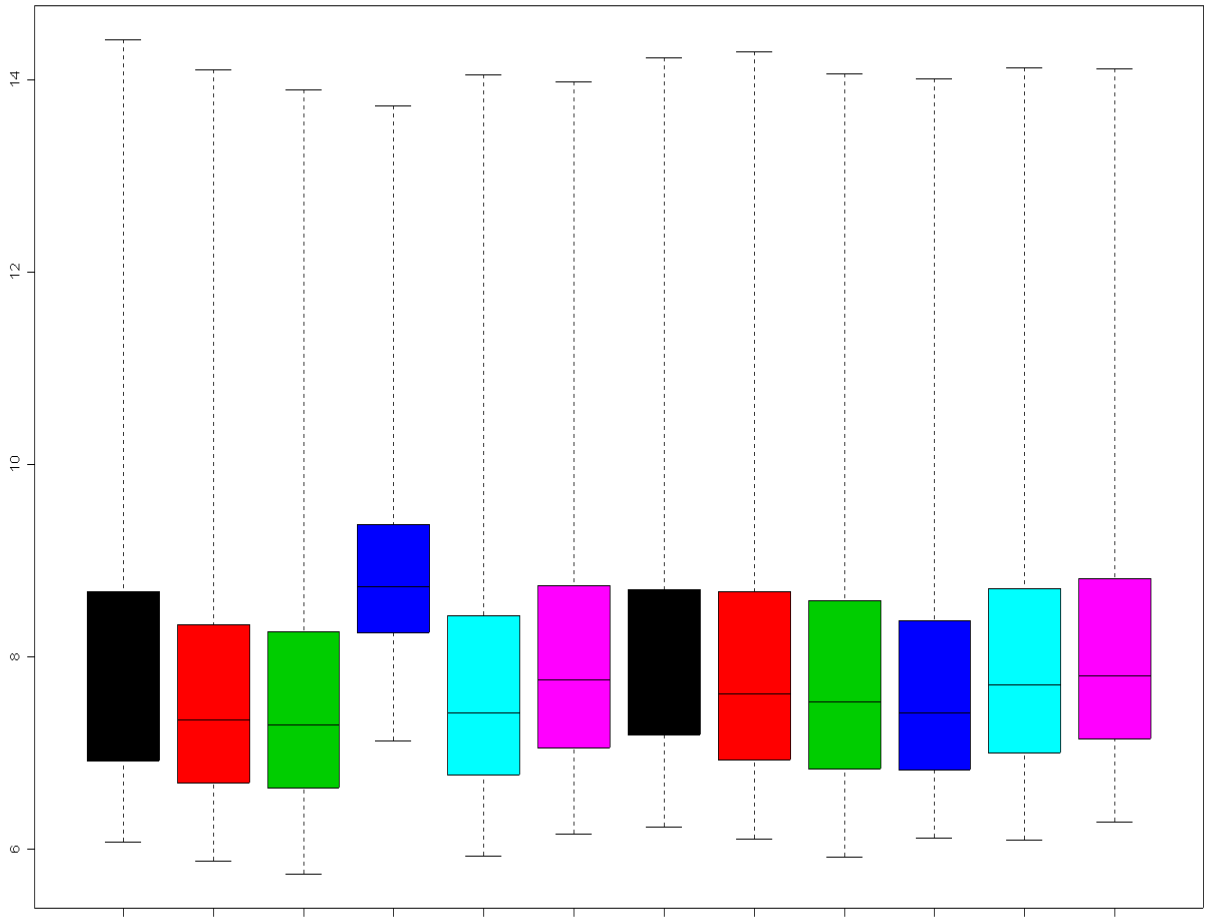
row means

$$\begin{matrix} 4.2 \\ 8.2 \\ 6.2 \\ 9.2 \\ 7.2 \end{matrix} \quad = \quad \begin{matrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \\ \hat{\mu}_4 \\ \hat{\mu}_5 \end{matrix}$$

RMA expression measures for the 5 GeneChips

RMA: results

Pre-Normalisation



RMA: results

Post-Normalisation