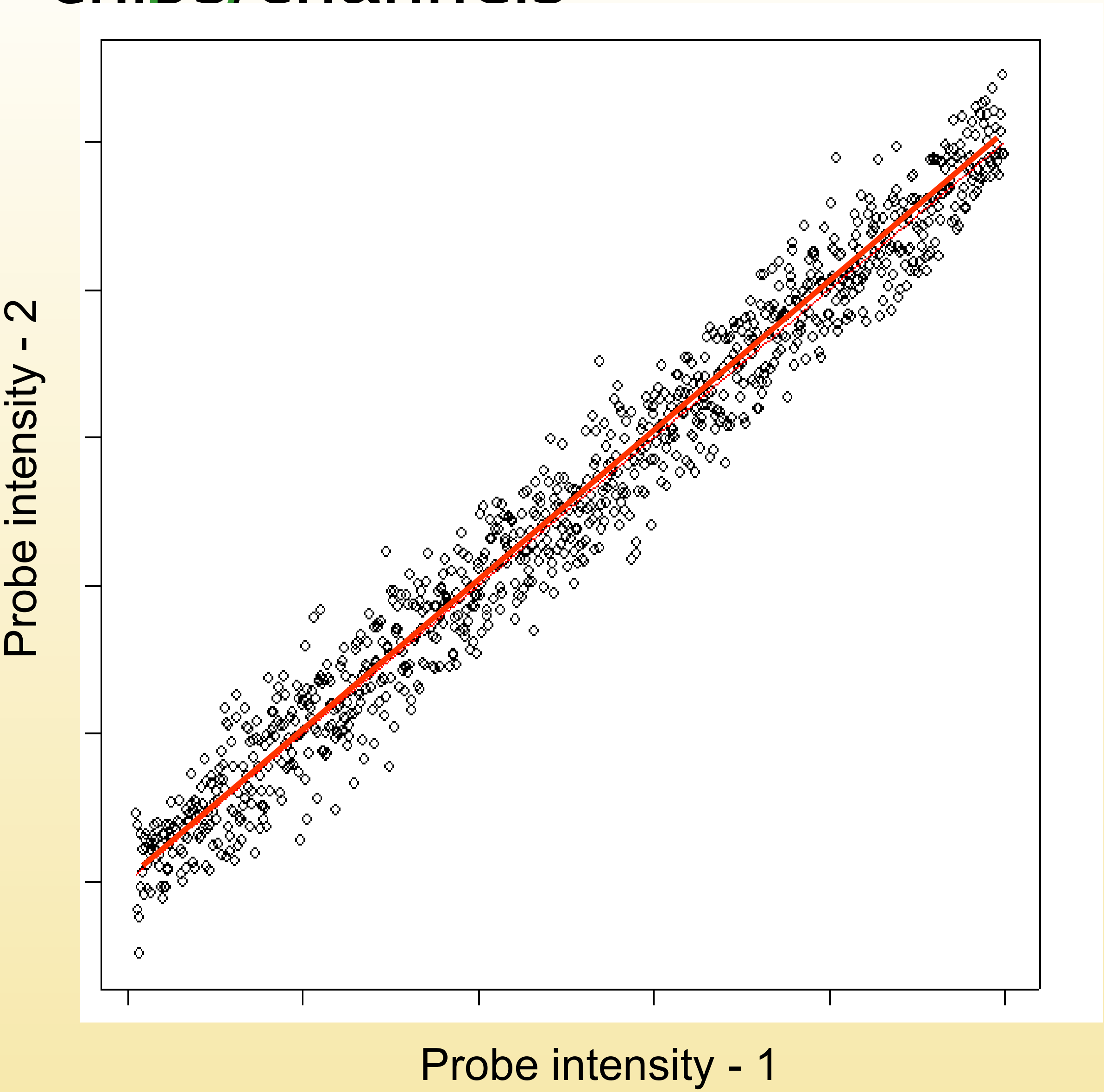
Normalization

Outline

- What is normalization
- Why is normalization needed
- Three quantitative methods for normalization
- Software tools

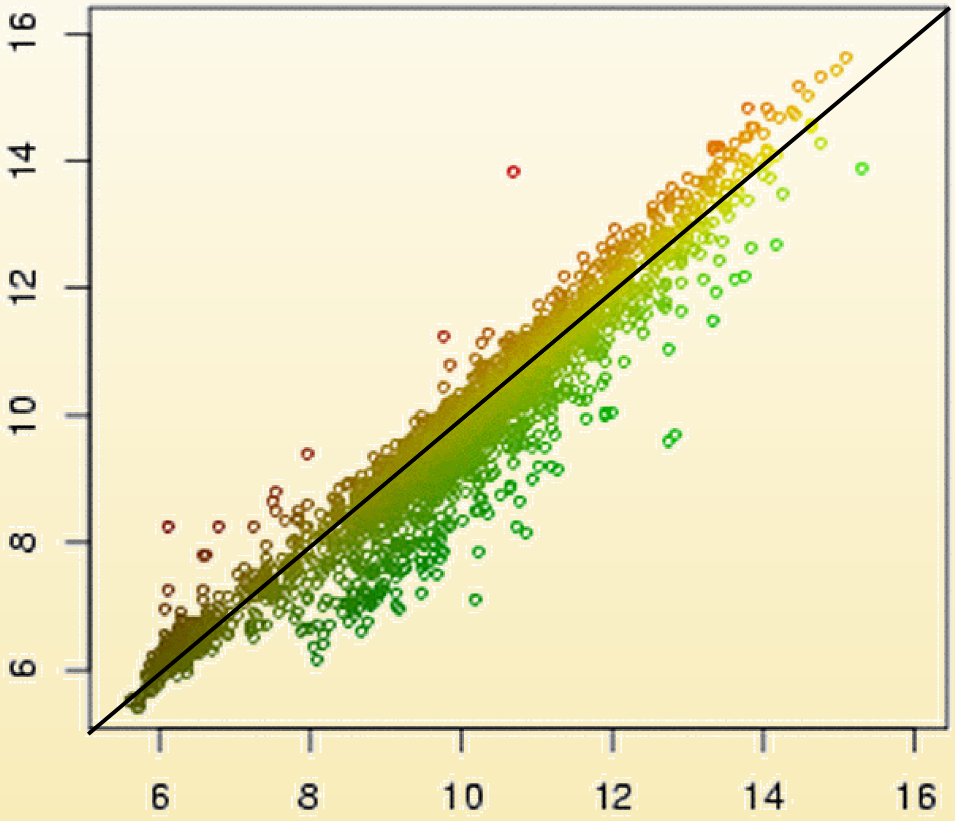
Hybridization of the same sample to 2 chips/channels

- Ideally: scatter plot coincides with the x=y diagonal
- Due to Random errors: we expect to see a ‘cloud’ around the x=y diagonal.

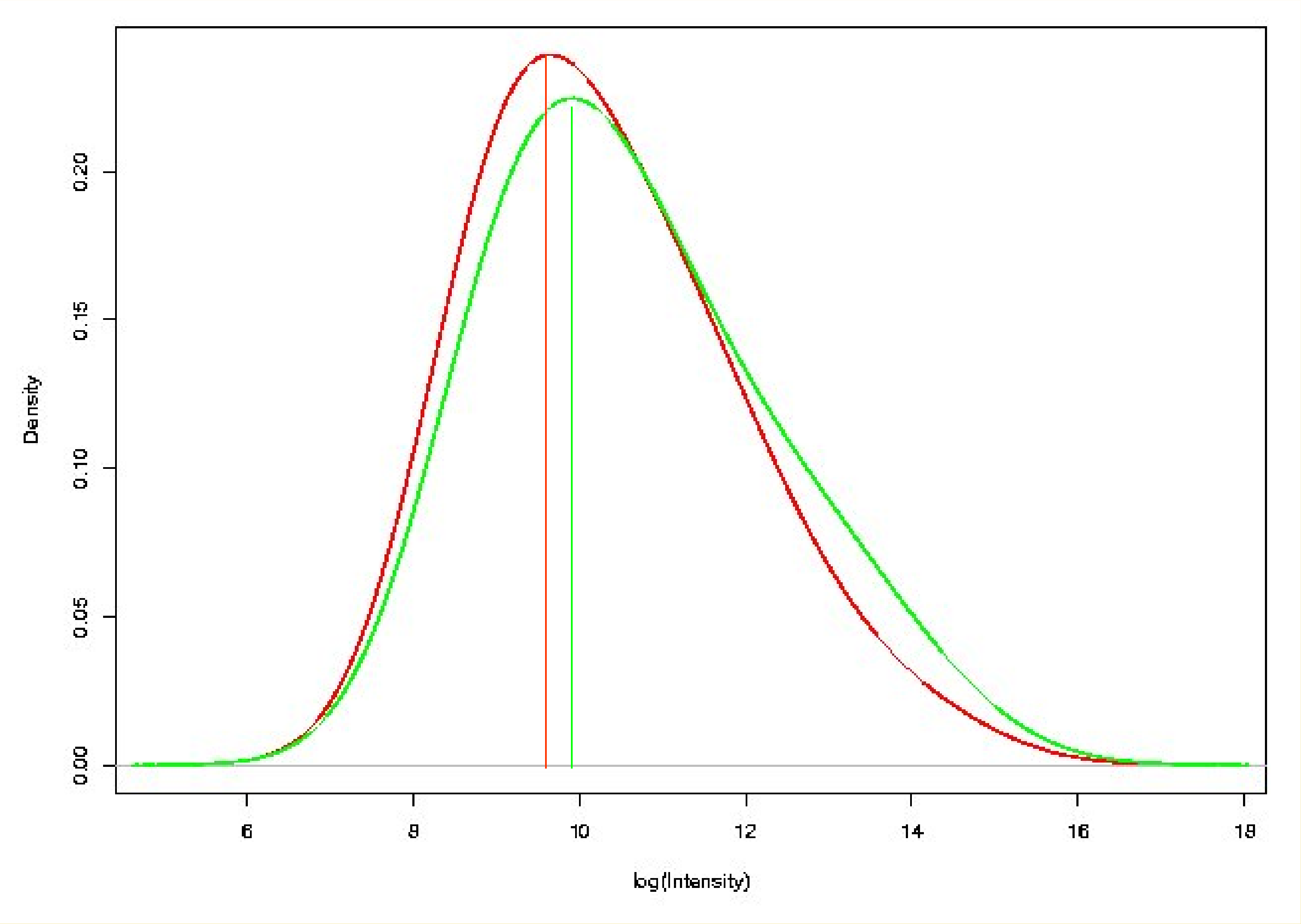


Hybridization of the same sample to 2 chips/channels

- In practice: Both Random and Systematic measurement errors (Bias)
- Due to Biases scatter plots are not centered around the x-y diagonal



Hybridization of the same sample to 2 chips/channels



Normalization – the process of removing systematic errors (biases) from the data

Sources of Systematic Errors

- Different incorporation efficiency of dyes
- Different amounts of mRNA
- Experimenter/protocol issues (comparing chips processed by different labs)
- Different scanning parameters
- Batch bias

Normalization - two problems

I. How to detect biases? Which genes to use for estimating biases among chips/channels?

II. How to remove the biases?

Which Genes to use for bias detection?

1. All genes on the chip
   - Assumption: Most of the genes are equally expressed in the compared samples, the proportion of the differential genes is low (<20%).
   - Limits:
     - Not appropriate when comparing highly heterogeneous samples (different tissues)
     - Not appropriate for analysis of ‘dedicated chips’ (apoptosis chips, inflammation chips etc)

Which Genes to use for bias detection?

1. Housekeeping genes
   - Assumption: based on prior knowledge a set of genes can be regarded as equally expressed in the compared samples
   - Affy novel chips: '*normalization set*' of 100 genes
   - NHGRI's cDNA microarrays: 70 "house-keeping" genes set
   - Limits:
     - The validity of the assumption is questionable
     - Housekeeping genes are usually expressed at high levels, not informative for the low intensities range

Which Genes to use for bias detection?

1. Spiked-in controls from other organism, over a range of concentrations
   - Limits:
     - low number of controls- less robust
     - Can’t detect biases due to differences in RNA extraction protocols
2. “Invariant set”
   - Trying to identify genes that are expressed at similar levels in the compared samples without relying on any prior knowledge:
     - Rank the genes in each chip according to their expression level
     - Find genes with small change in ranks

Normalization Methods

1. Global normalization (Scaling)

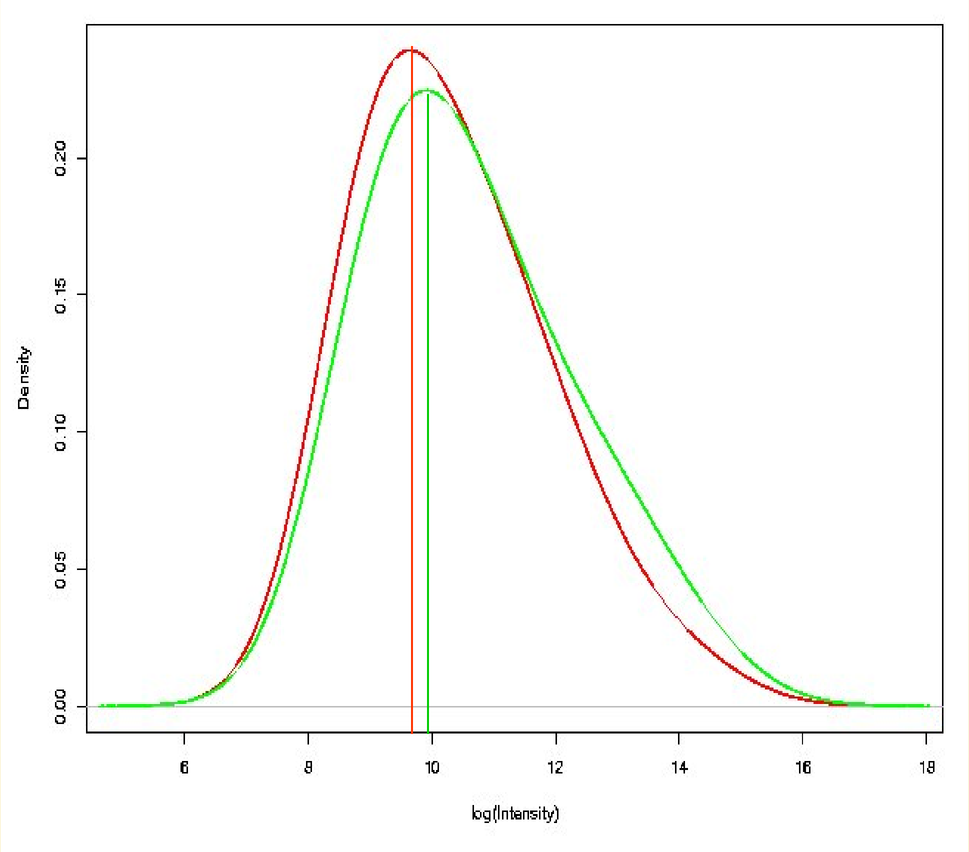
- A single normalization factor (k) is computed for balancing chips\channels:

$$X_i^{norm} = k*X_i$$

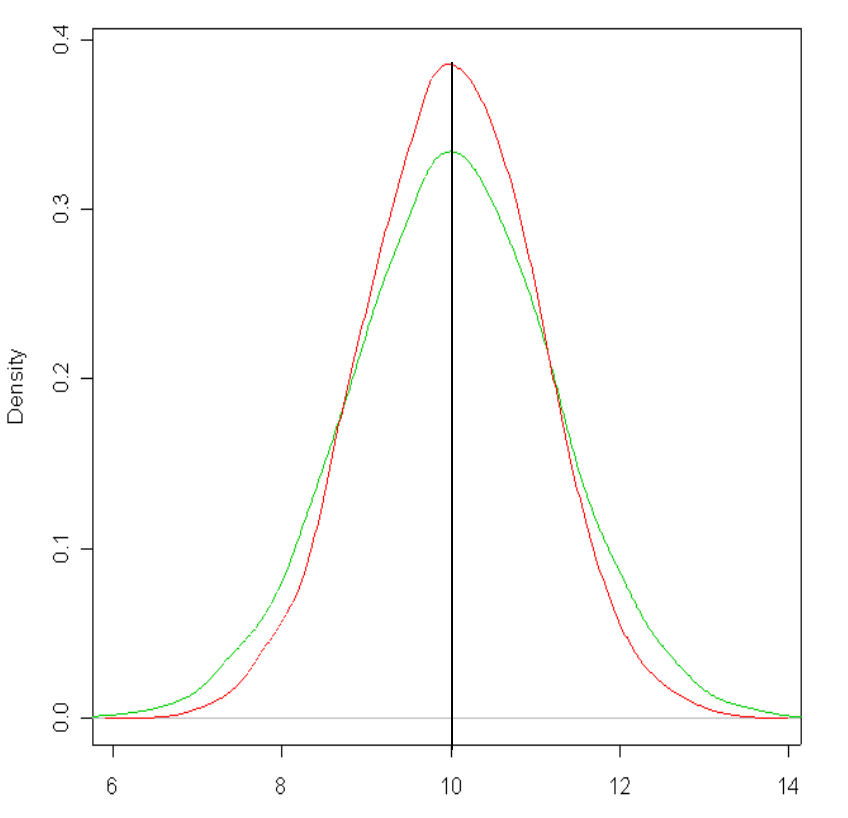
- Multiplying intensities by this factor equalizes the mean (median) intensity among compared chips

Global Normalization

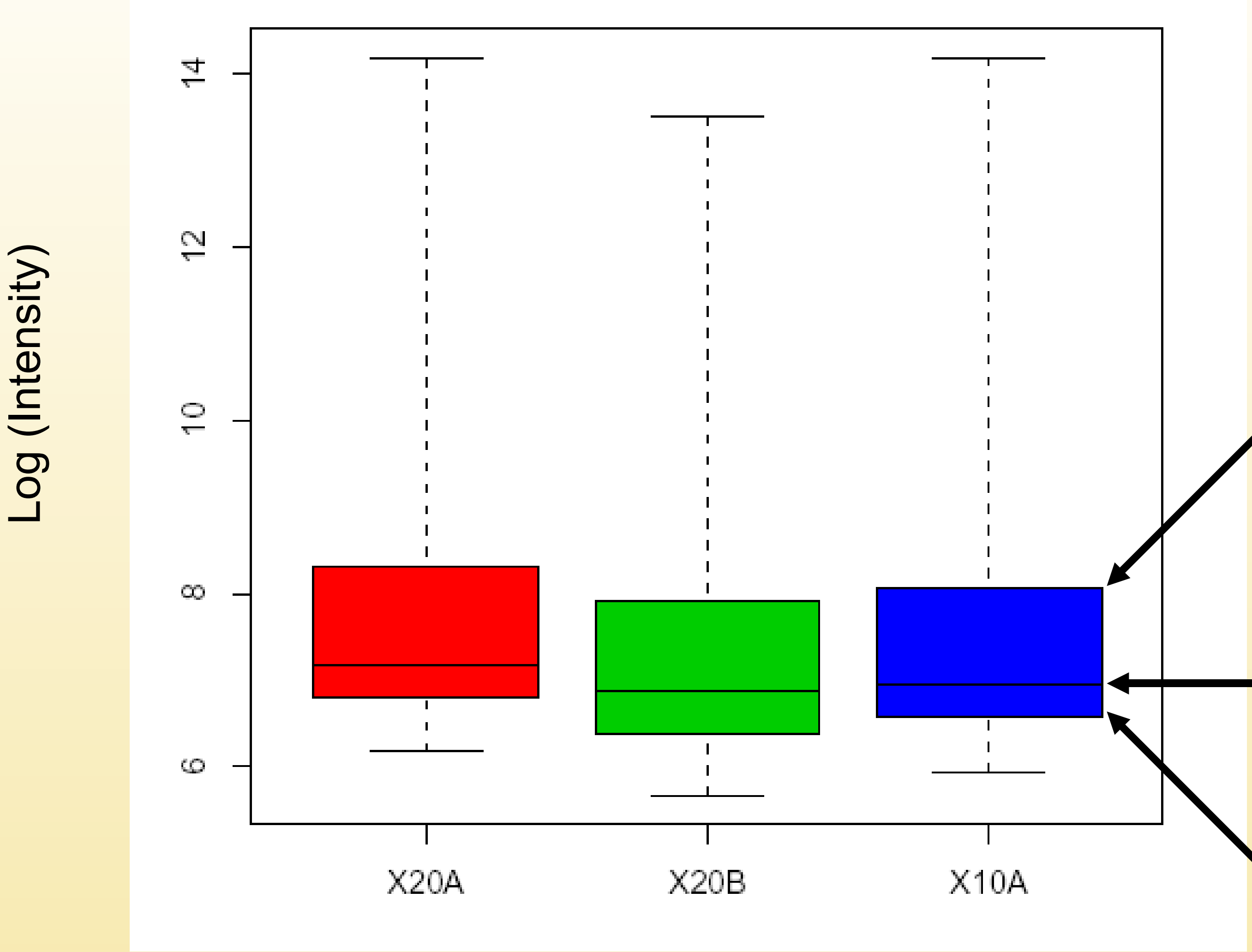
Before



After



Boxplots

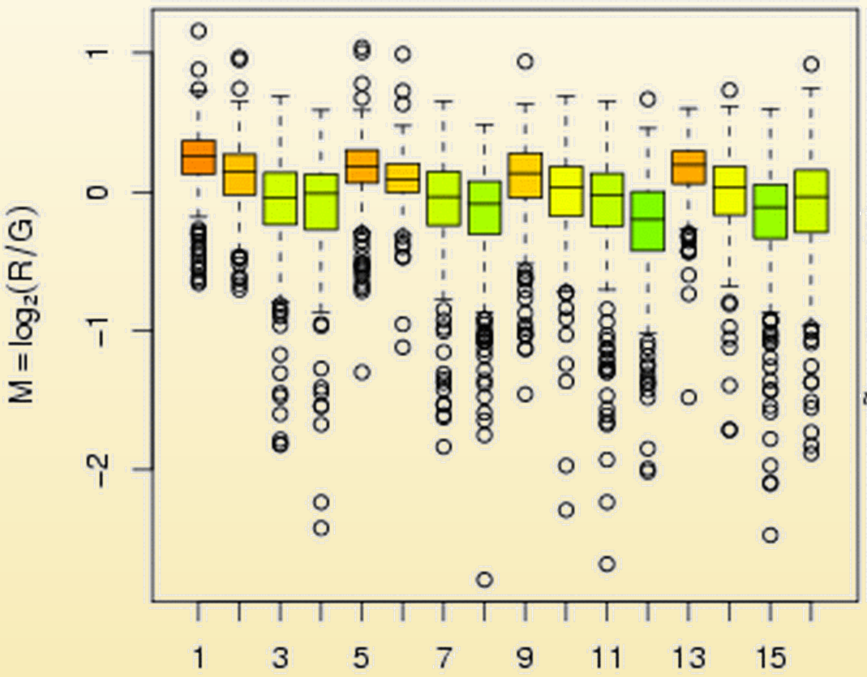


Upper quartile

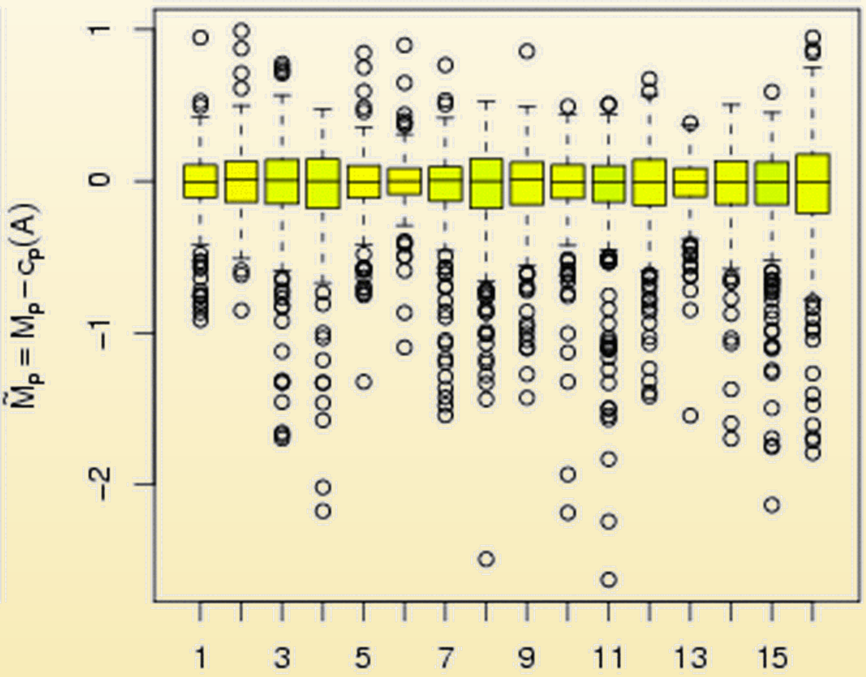
Median intensity

Lower quartile

Before Normalization



After Scaling



2. Intensity-dependent normalization (Yang, Speed)

(Lowess – local linear fit)

- Compensate for intensity-dependent biases

Detect Intensity-dependent Biases: M vs A plots

- X axis: A – average intensity

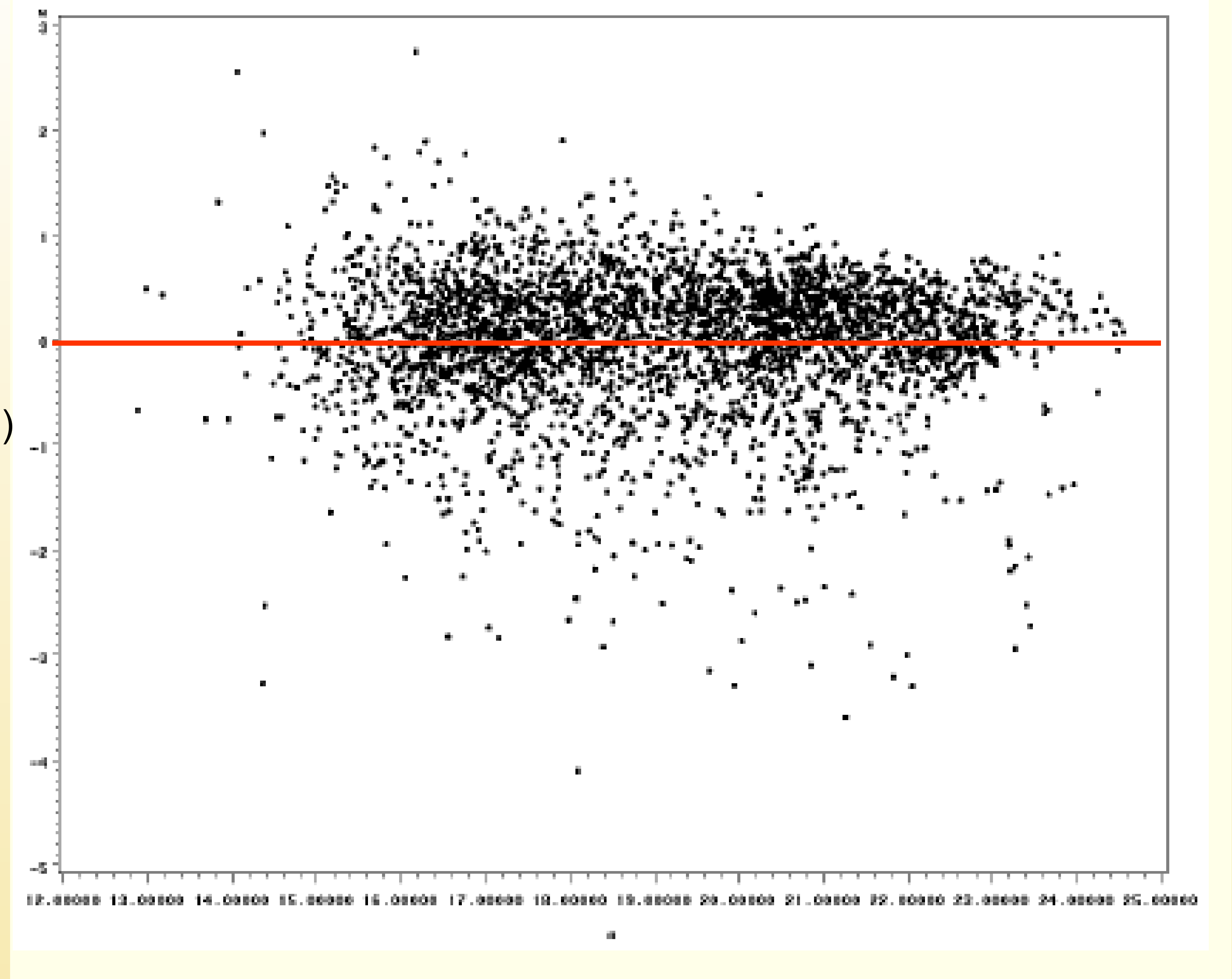
  A = 0.5*log(Cy3*Cy5)

- Y axis: M – log ratio

  M = log(Cy3/Cy5)

We expect the M vs A plot to look like:

M = log(Cy3/Cy5)



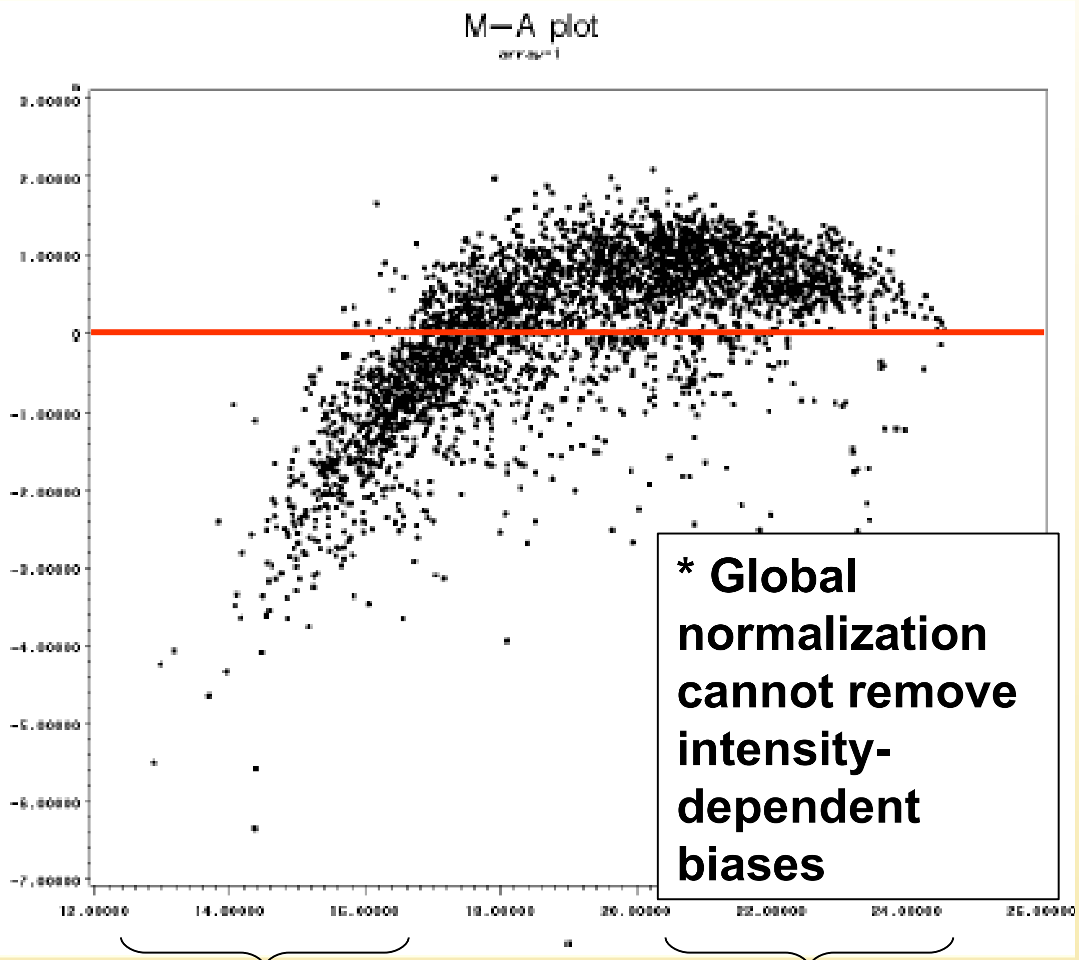
A

Intensity-dependent bias

M>0:
Cy3>Cy5

M =
log(Cy3/Cy5)

M<0:
Cy3<Cy5



*** Global normalization cannot remove intensity-dependent biases**

Low intensities

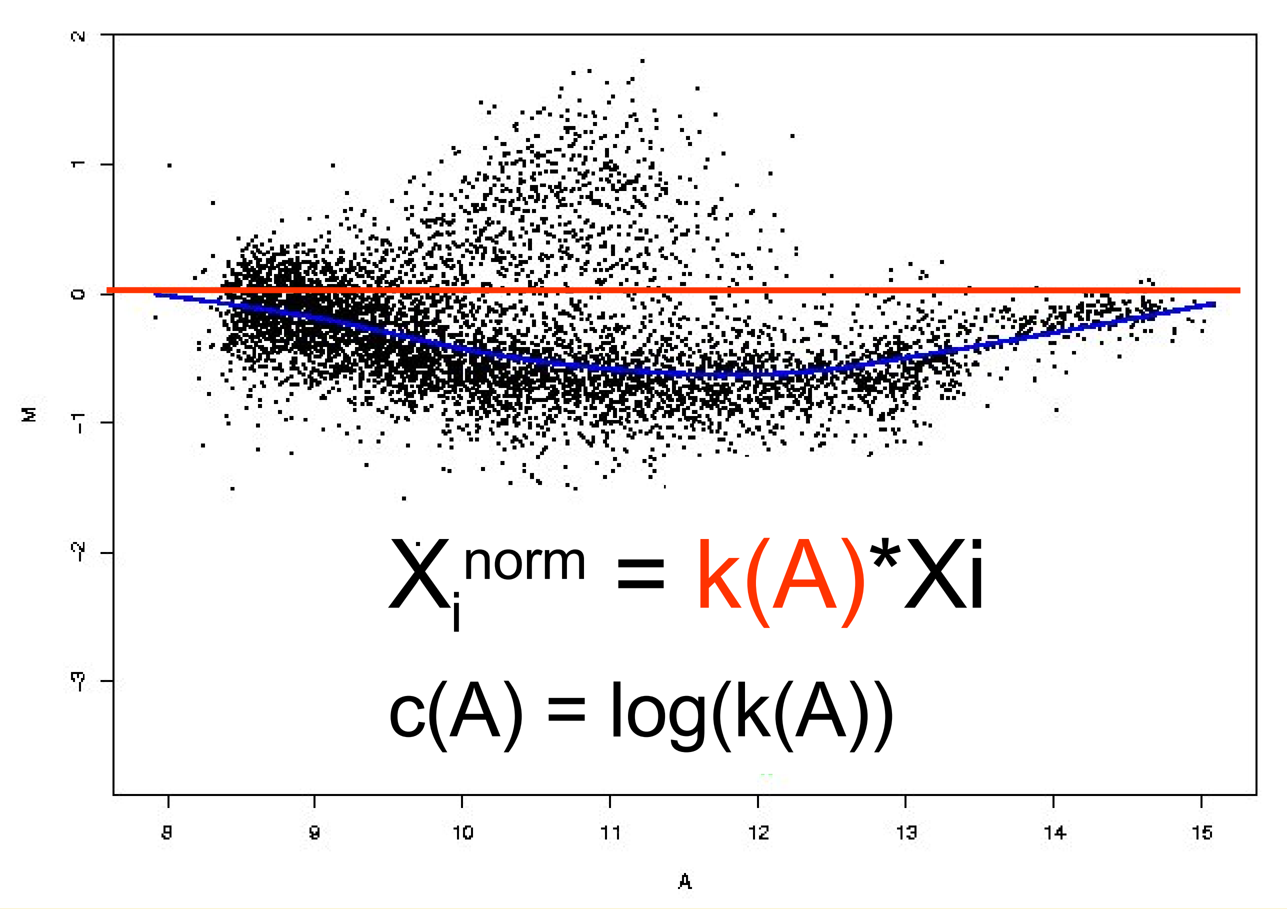
A

High intensities

Intensity-Dependent Normalization

Assumption: Most of the genes are equally expressed at all intensities

Lowess – fitting local regression curve – c(A)

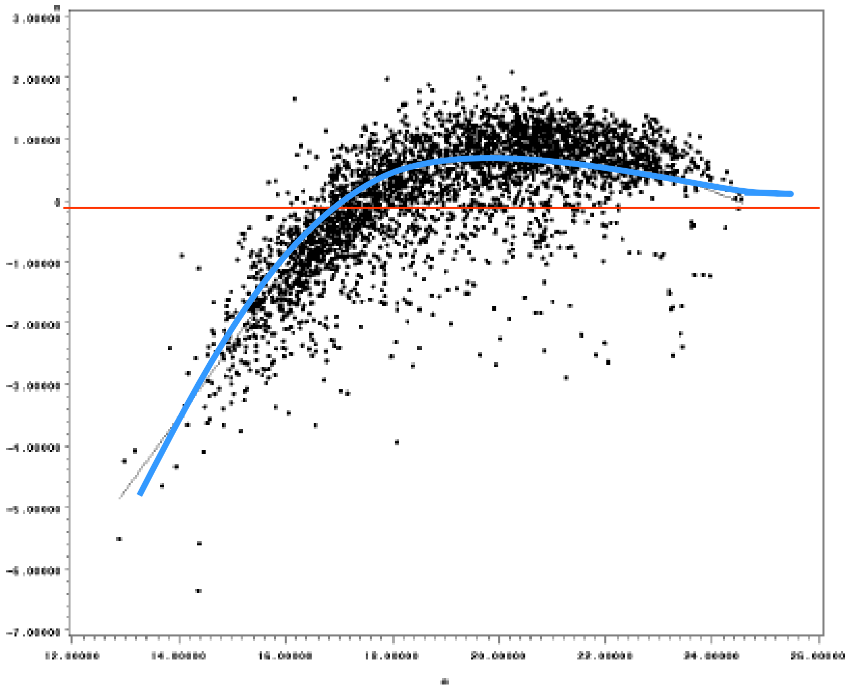


$$X_i^{norm} = k(A)*Xi$$

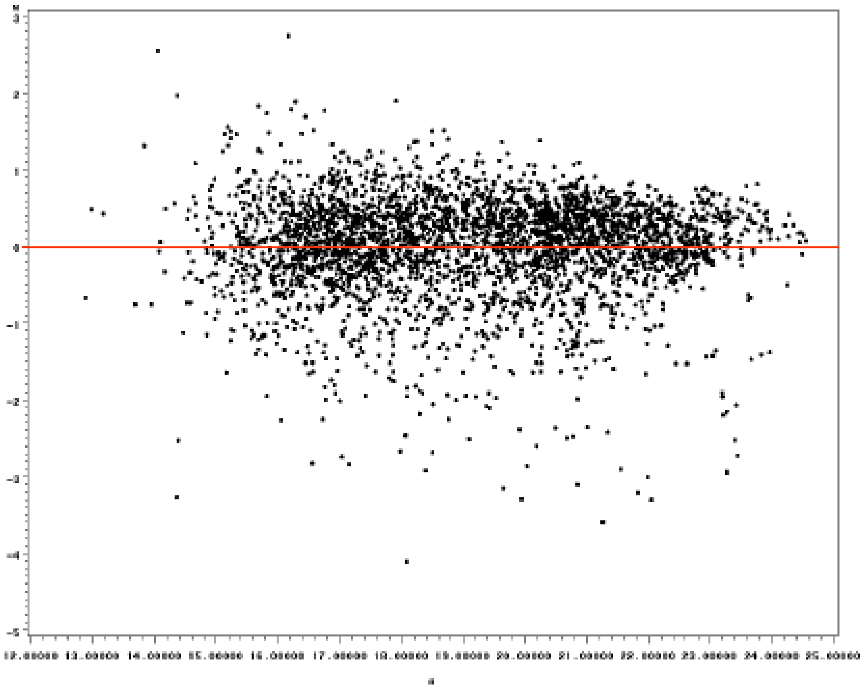
$$c(A) = \log(k(A))$$

LOESS (Local Regression)

Before



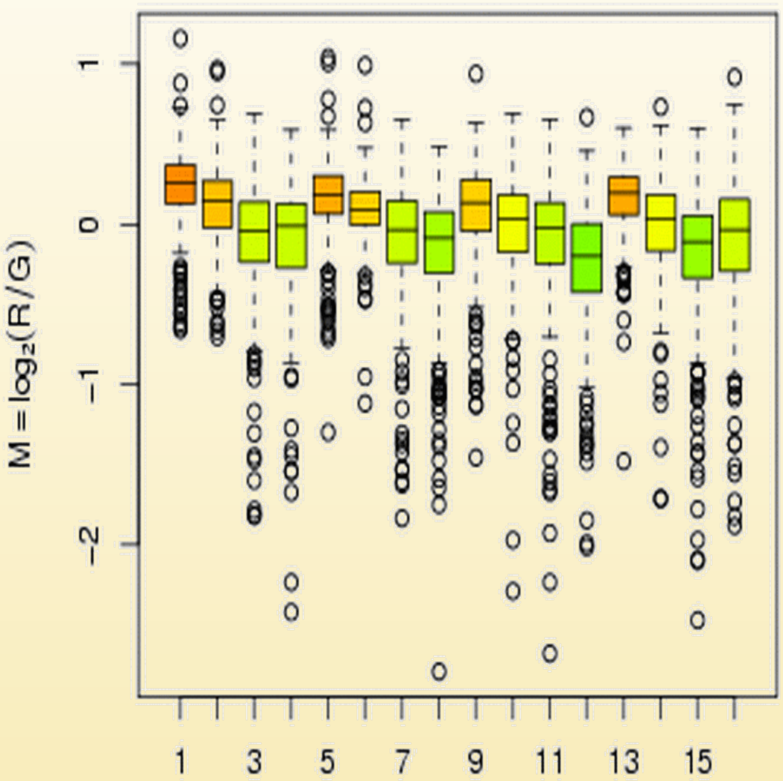
After



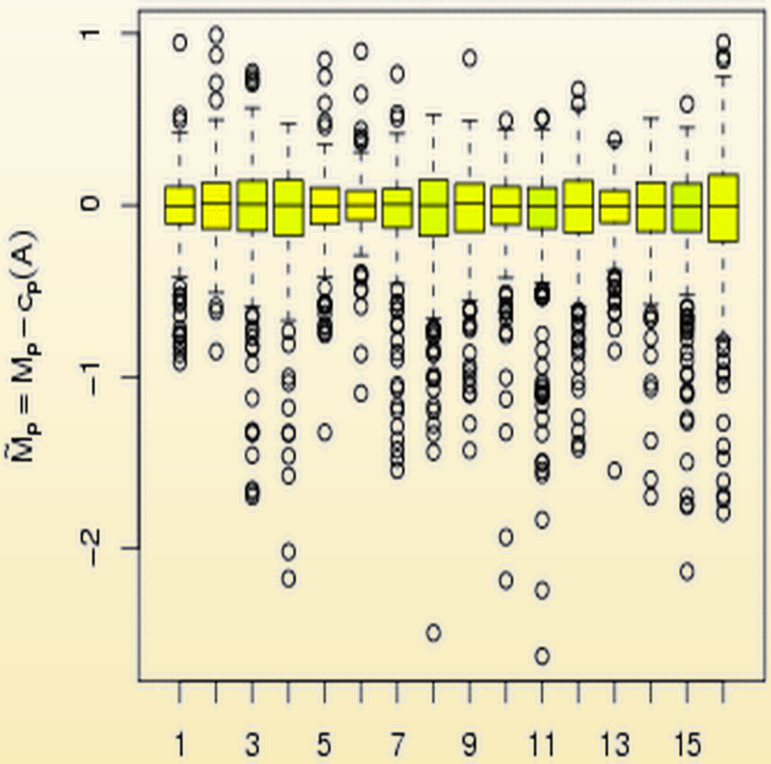
3. Quantile Normalization

- Global normalization - enforces the chips to have equal mean (median) intensity
- Lowess – enforces equal means at all intensities
- Quantile Normalization - enforces the chips to have identical intensity distribution

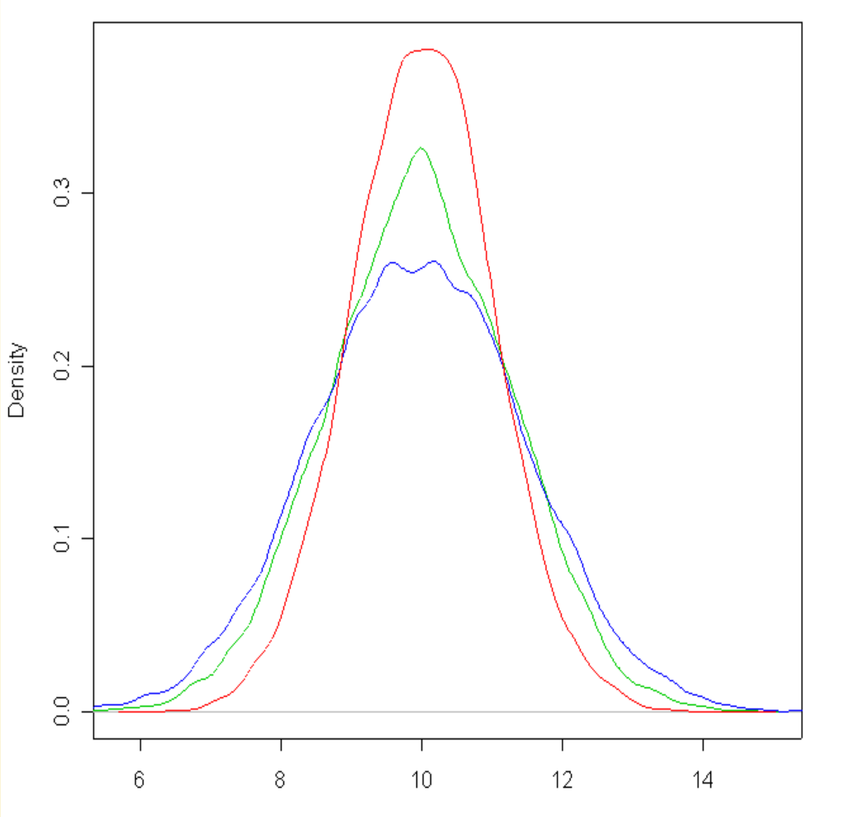
Before Normalization



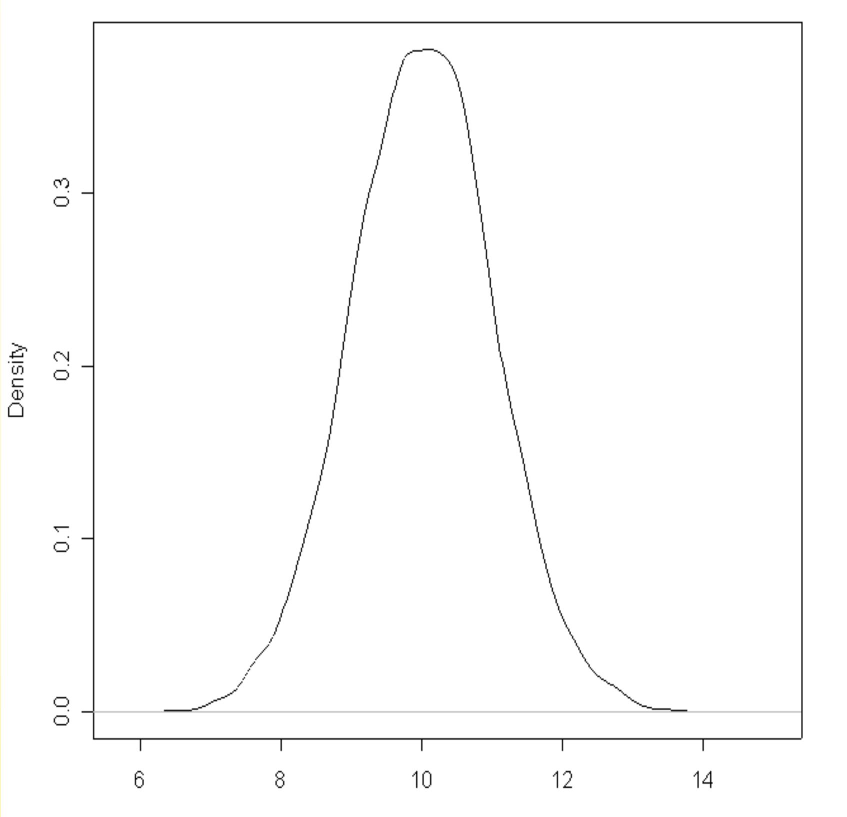
After Scaling



After lowess normalization

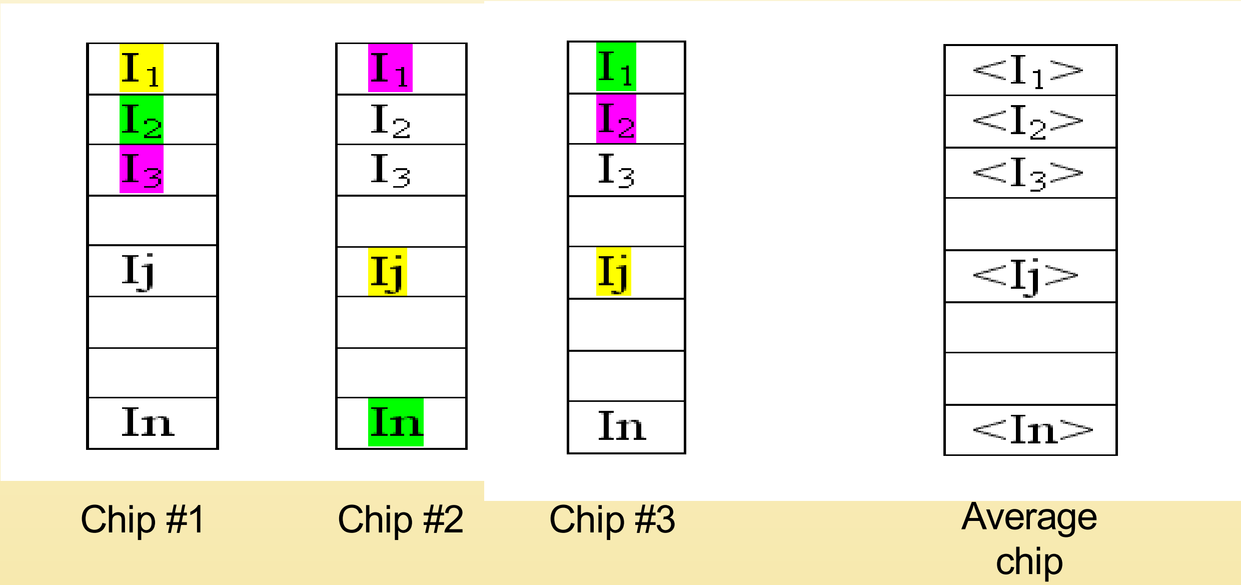


After quantile normalization



Quantile Normalization

- Sort intensities in each chip
- Compute mean intensity in each rank across the chips
- Replace each intensity by the mean intensity at its rank



Recommendation (Bolstad et al, Speed, 2003)

- Quantile normalization performs best
- Lowess is comparable to Quantile
- Scaling is not satisfactory

Normalization - tools

- Bioconductor (both AFFY and cDNA):
  - Packages in R language
- dChip (Affymetrix):
  - Quantile, Invariant set
- Expander (both AFFY and cDNA):
  - Lowess
  - Quantile

Acknowledgements

- Figures in this presentations were taken in part from presentations of:
  - Henrik Bengtsson, Terry Speed
  - Yee Yang, Terry Speed
  - Guilherme J. M. Rosa
  - Laurent Gautier, Rafael Irizarry, Leslie Cope, and Ben Bolstad