Measuring gene expression with DNA microarrays

02.01.2012 and
04.01.2012

Outline

- Microarrays for the detection of gene expression
  - Technologies for microarrays
  - Normalization
    - Lowess
    - Quantile normalization
    - Variance stabilized normalization
  - Exploratory data analysis
  - Validation

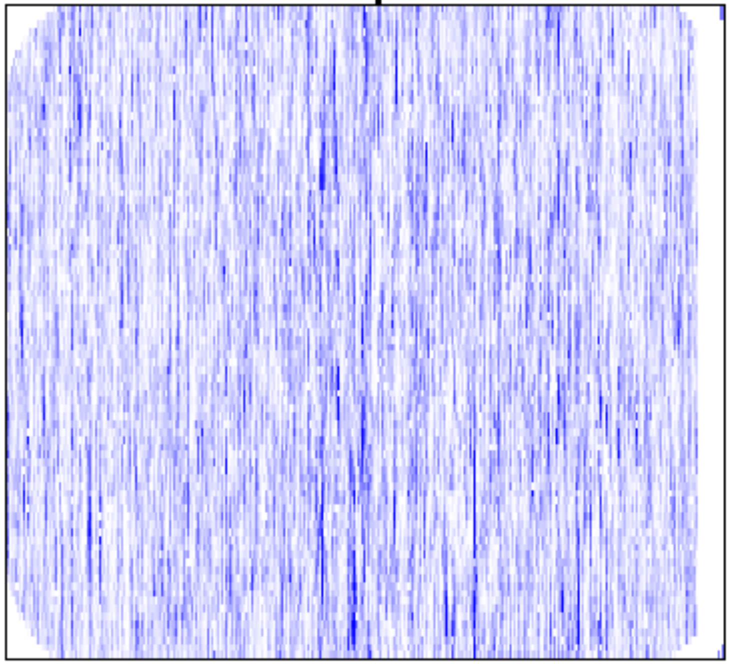
Motivation

- Monitoring gene expression
  - Comparing different samples
    - Tissues
    - Strains of bacteria or yeasts
  - Time series

- Whole genome expression (tiling arrays)
- Pathogen detection
- Resequencing
- Study protein-DNA interaction

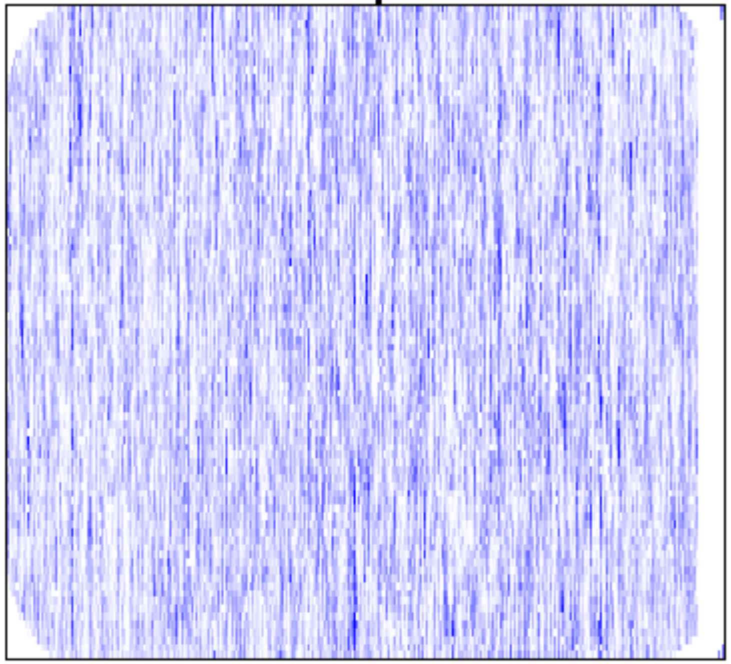
Technologies

Chip1



z-range 84 to 650175 (saturation 84, 487761.1)

Chip2



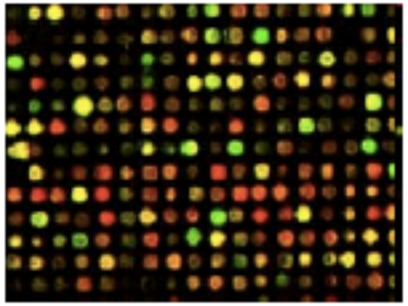
z-range 85.5 to 561152 (saturation 84, 487761.1)

Common technologies

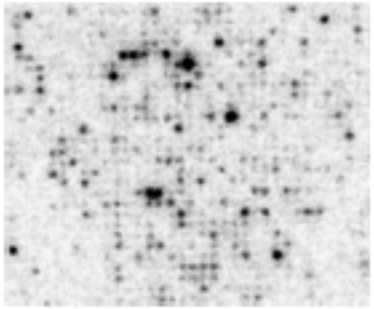
- (spotted) cDNA arrays
  - Custom made
  - Lengths up to 1000 bp
- Oligonucleotide arrays
  - Industrially manufactured (Affymetrix, Agilent, Nimblegen, etc)
  - 25 bp (Affy), ~60 for other technologies

- Single experiments
  - Evaluate intensities
  - Absolute transcript levels
- Two dye experiments
  - Evaluate ratio of intensities
- Different strategies for normalization and analysis

Microarrays

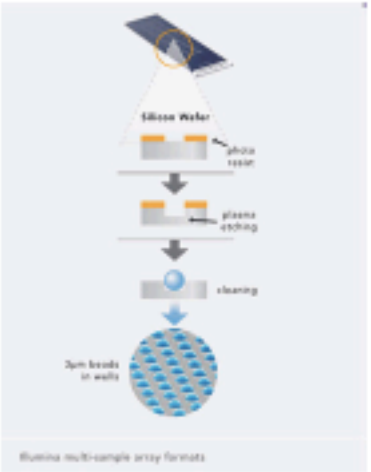


Spotted glas arrays

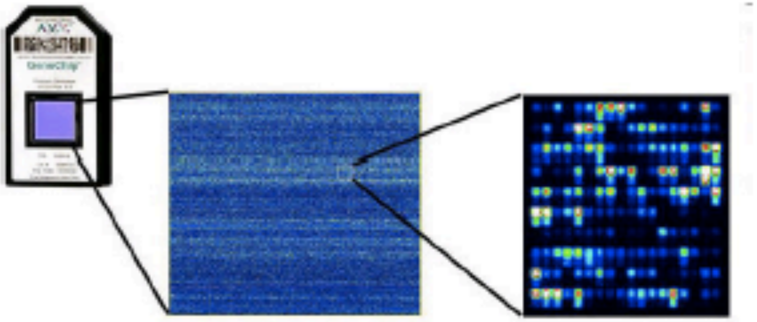


Membrane arrays

cDNA

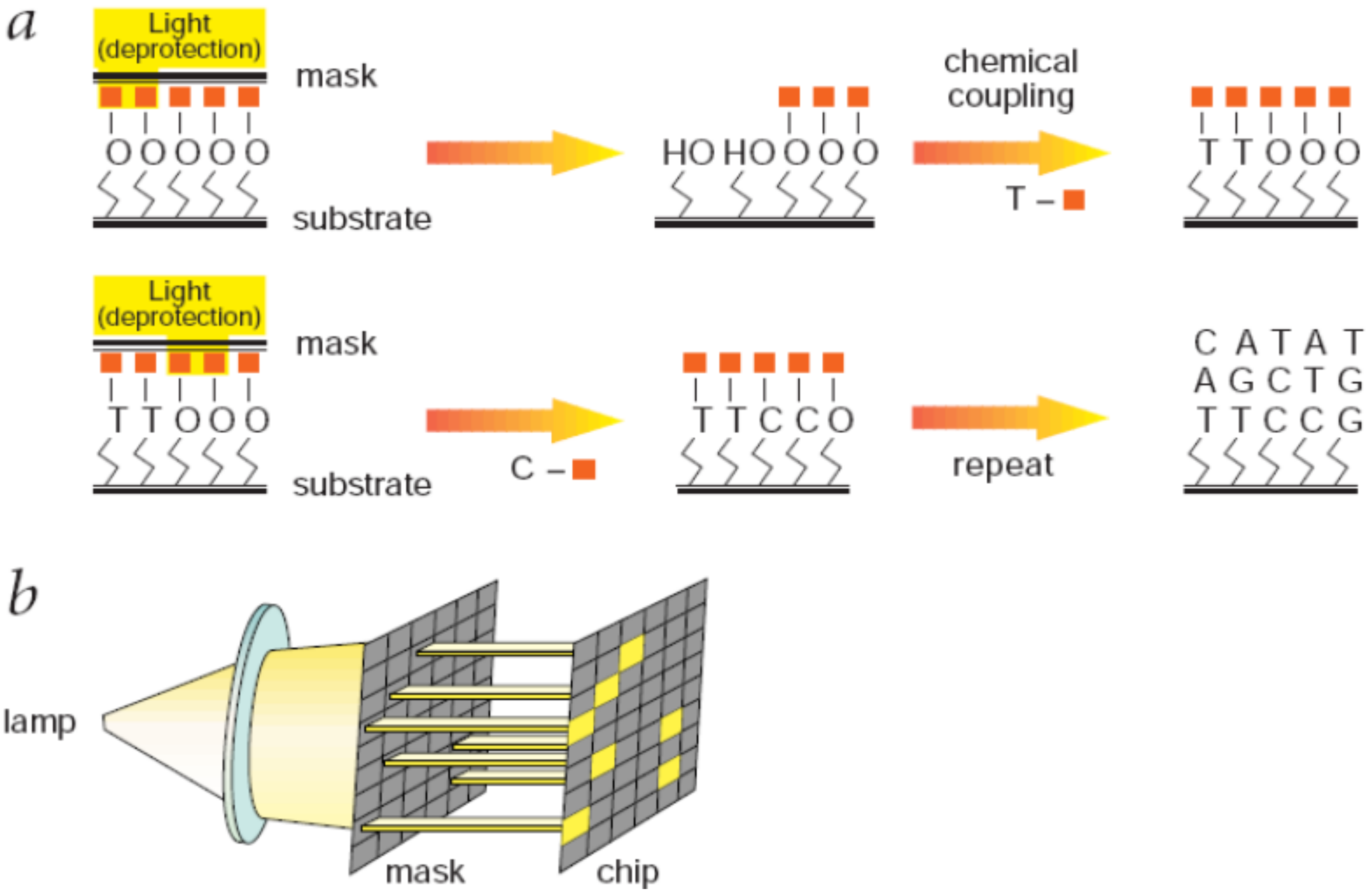


Illumina Bead Arrays



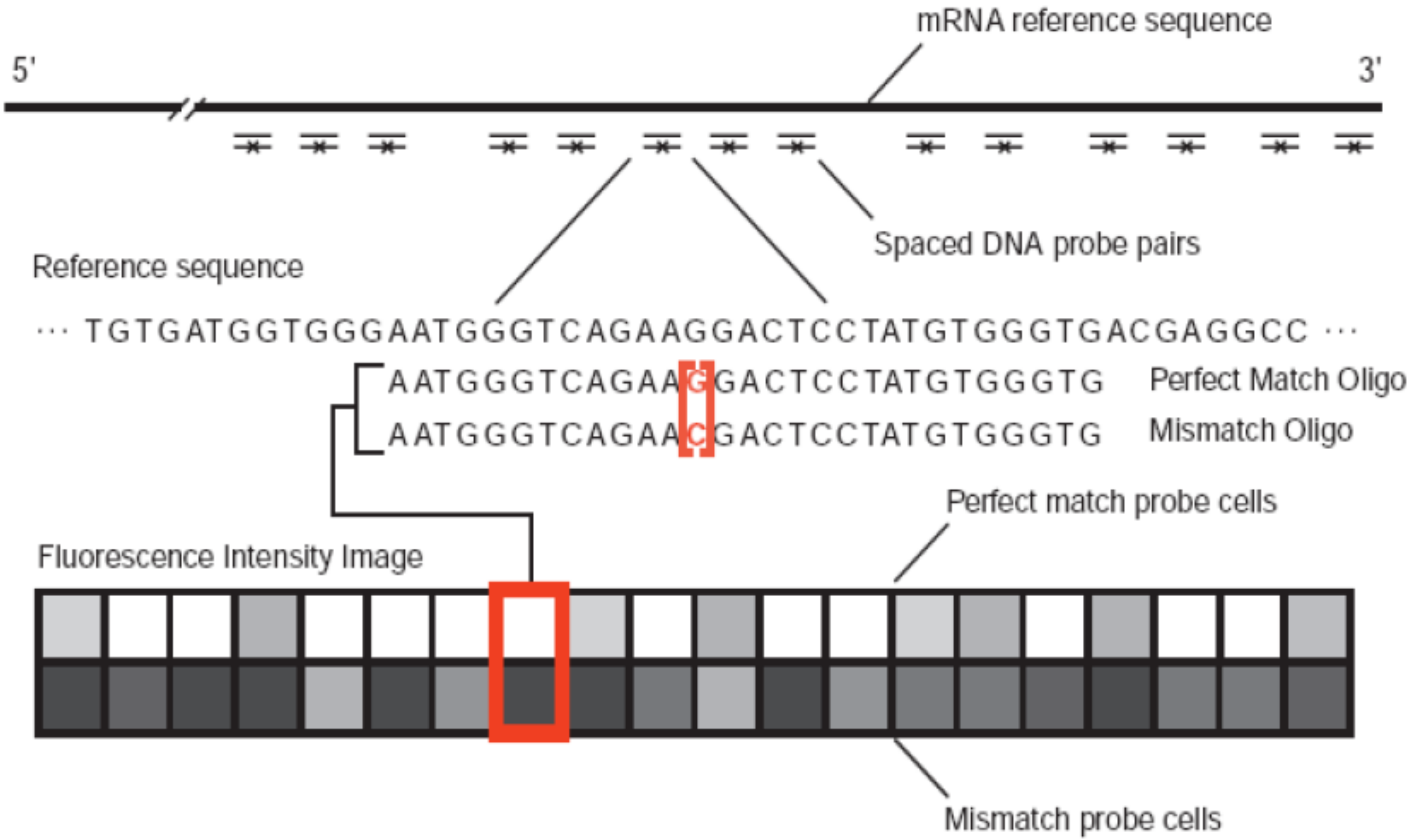
Affymetrix GeneChip

Manufacturing oligonucleotide arrays



AFFYMETRIX
@520019005520560507064010000885092
GeneChip®
H2
Human Genome
U133 Plus 2.0
P/N: 520019
Lot #: 4010008
Exp. Date: 05/07/06
For Research Use Only
AFFYMETRIX
@520029005888411124064015111225338
GeneChip®
S1
Mouse Genome
430 2.0 Array
P/N: 520029
Lot #: 4015112
Exp. Date: 11/24/06
For Research Use Only

Oligonucleotide array design



mRNA reference sequence

5'

3'

Spaced DNA probe pairs

Reference sequence

··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···

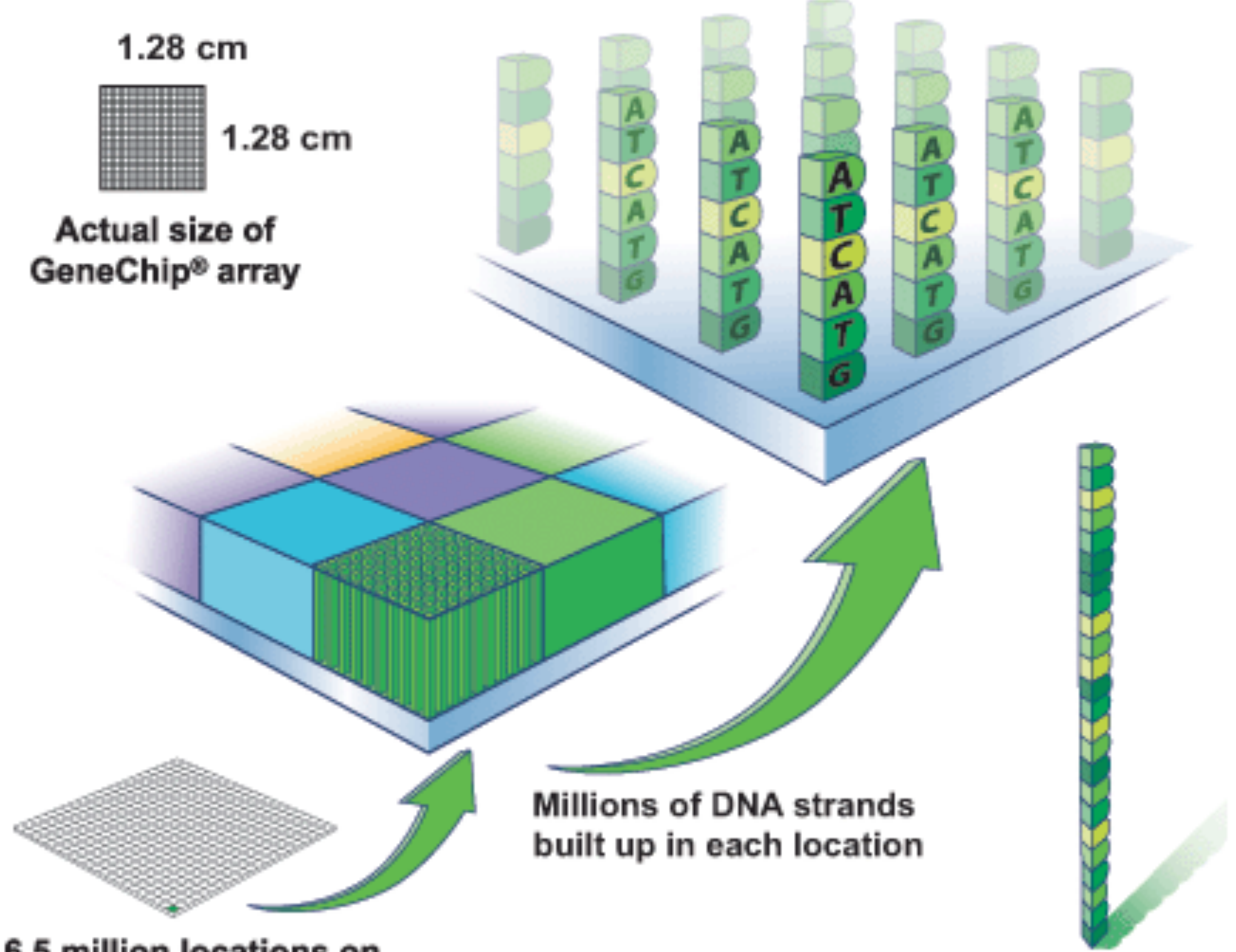
AATGGGTCAGAAGGACTCCTATGTGGGTG Perfect Match Oligo

AATGGGTCAGAACGACTCCTATGTGGGTG Mismatch Oligo

Perfect match probe cells

Fluorescence Intensity Image

Mismatch probe cells



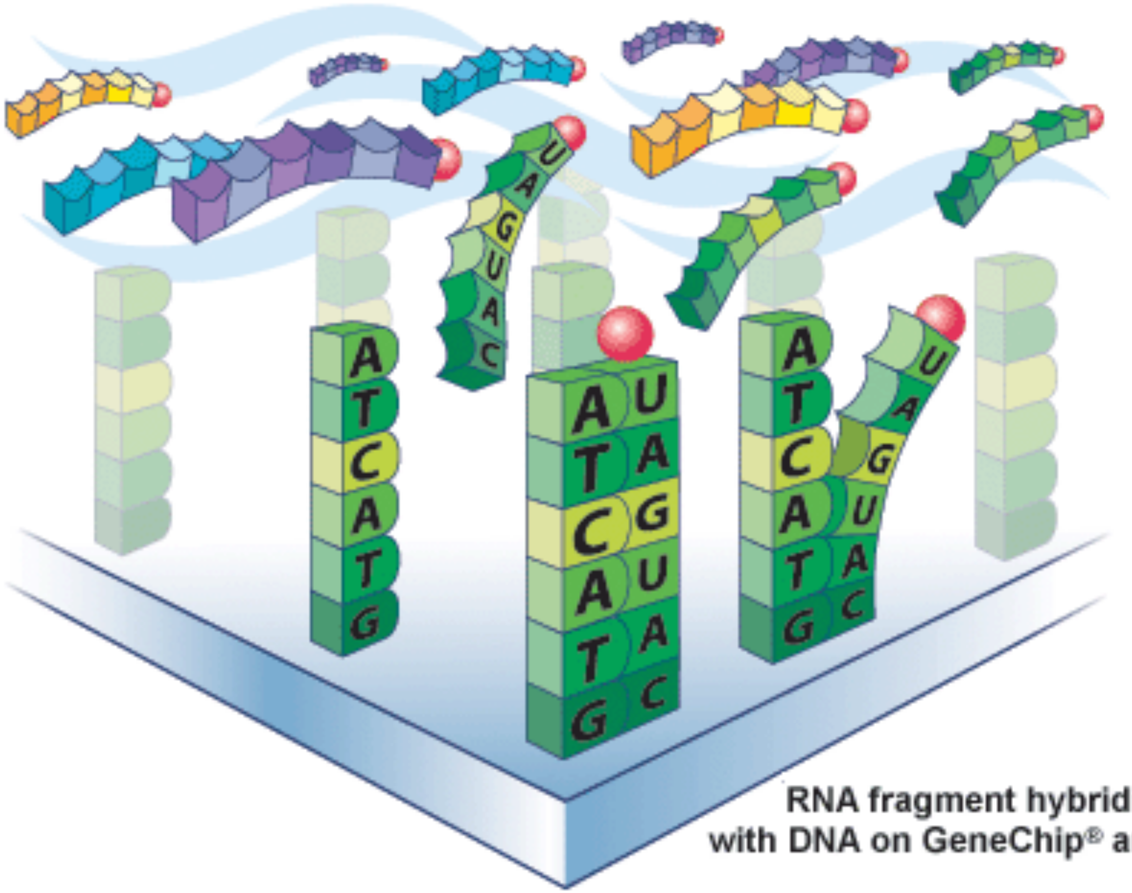
Actual size of GeneChip® array

Millions of DNA strands built up in each location

6.5 million locations on each GeneChip® array

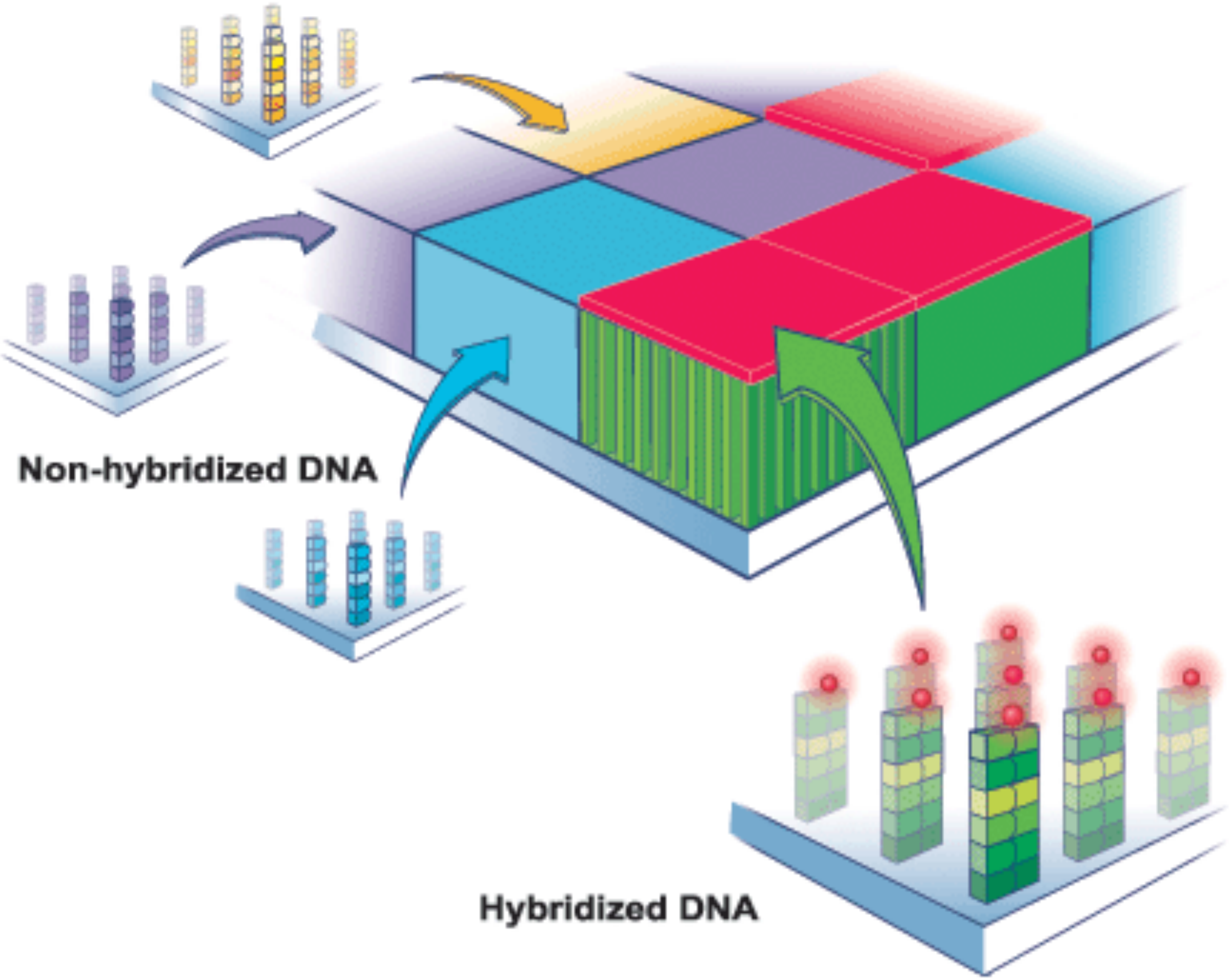
Actual strand = 25 base pairs

RNA fragments with fluorescent tags from sample to be tested



RNA fragment hybridizes
with DNA on GeneChip® array

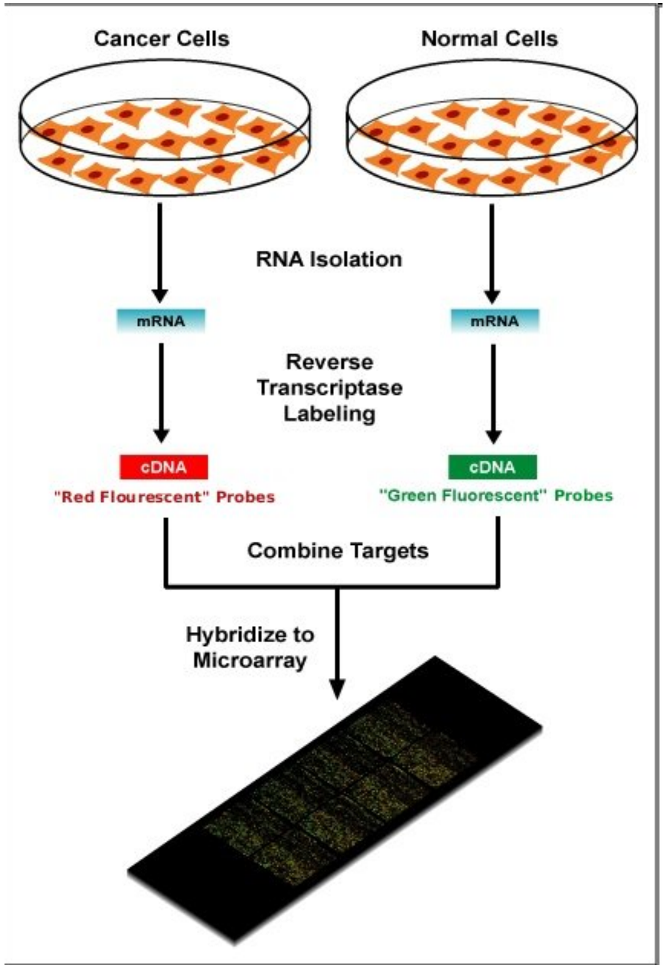
Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow



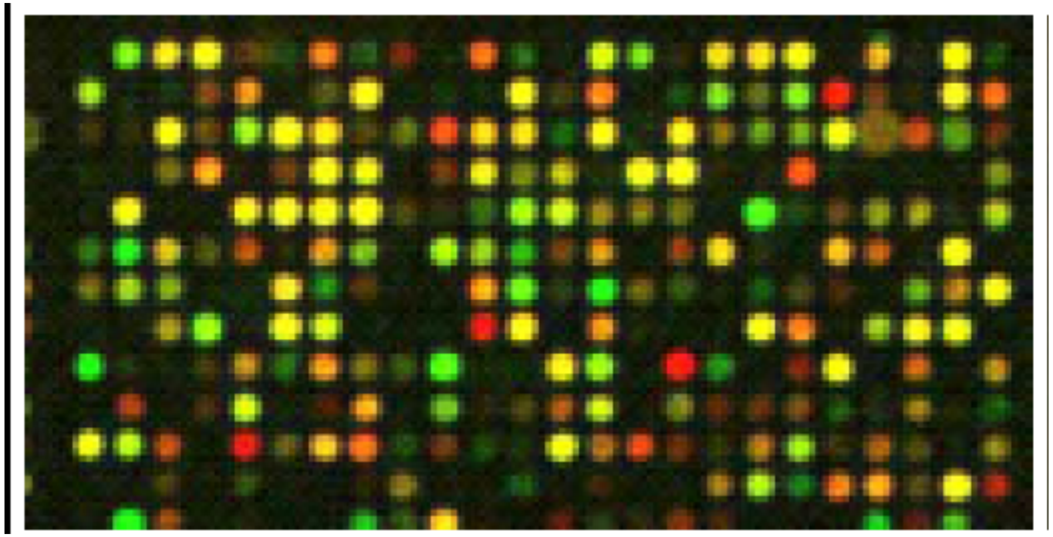
Non-hybridized DNA

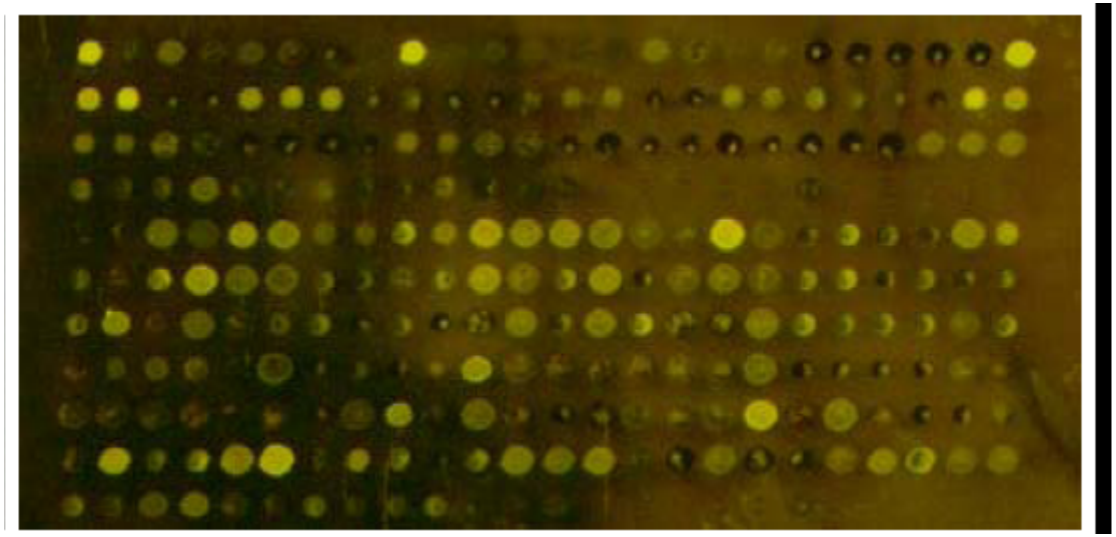
Hybridized DNA

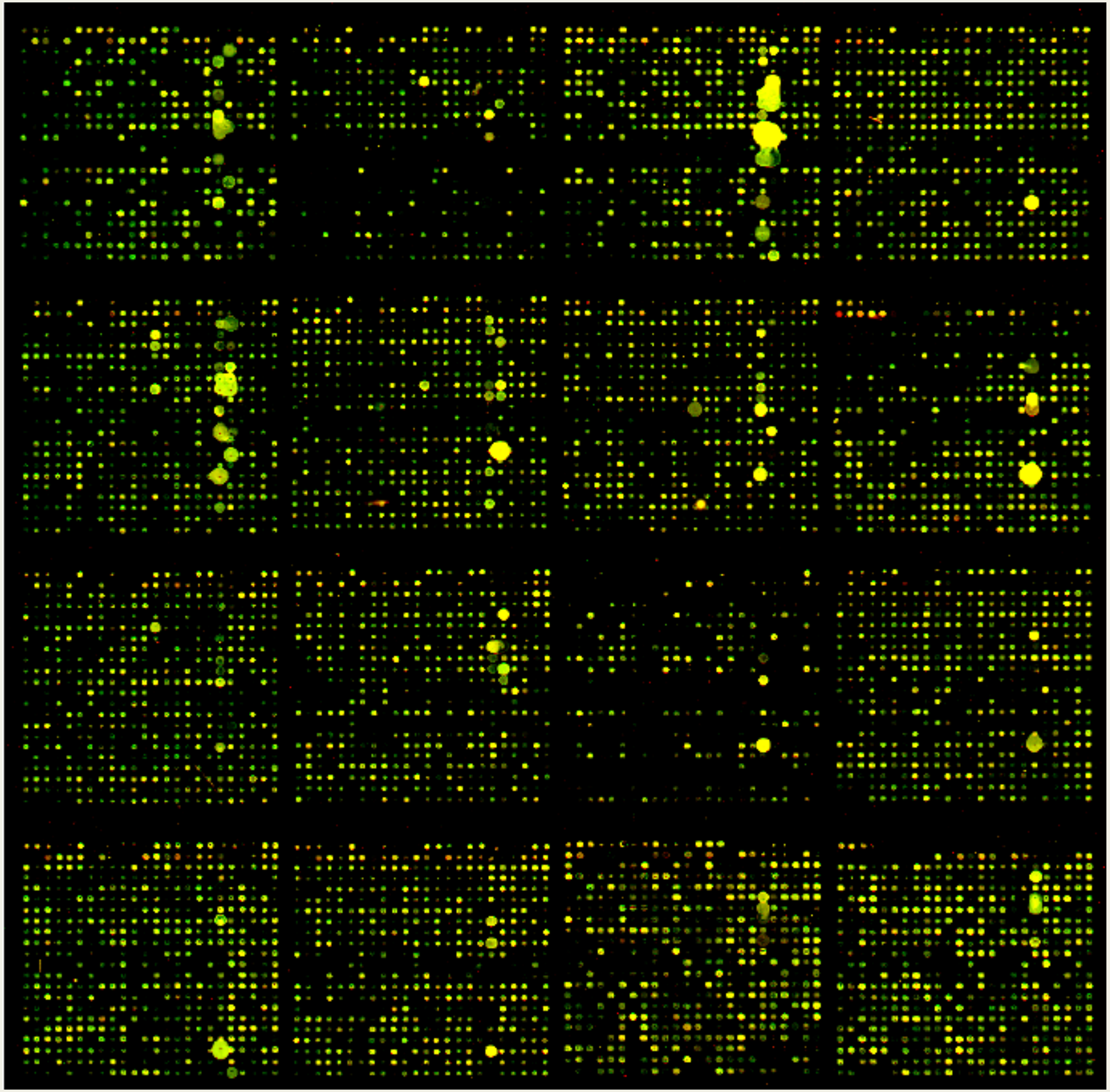
Two colour cDNA array



Red vs green overlay



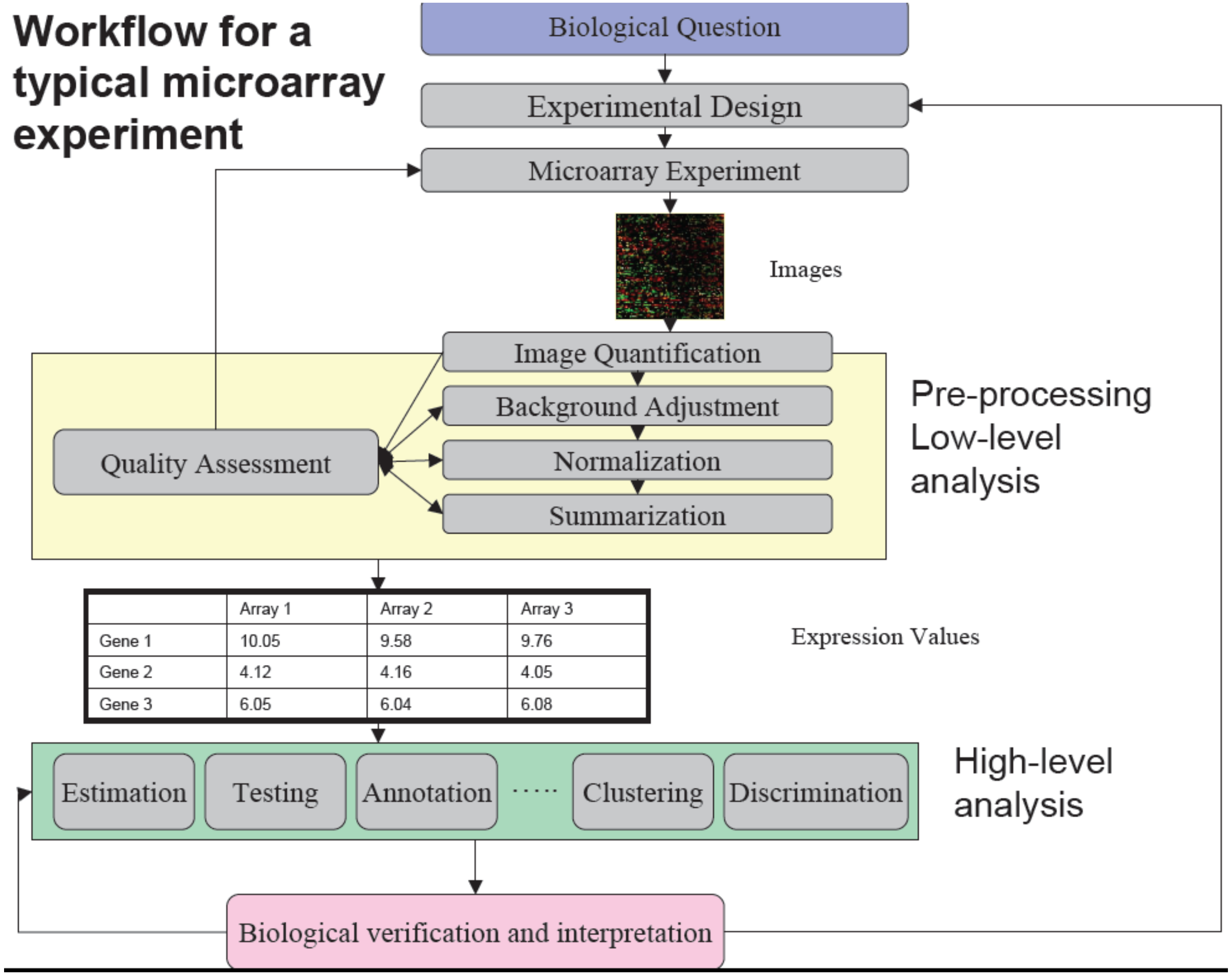




Preliminary data analysis

Plots and strategies

Typical workflow



Workflow for a typical microarray experiment

	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

Pre-processing
Low-level
analysis

High-level
analysis

From Bolstad

Influences

Measuring $Y_{i,k}$ intensity of probe *i* on array *k*

- Total RNA amount
- Total sample amount
- Efficiency of
  - RNA extraction
  - Reverse transcription
  - cDNA amplification
  - cRNA transcription
  - Labeling

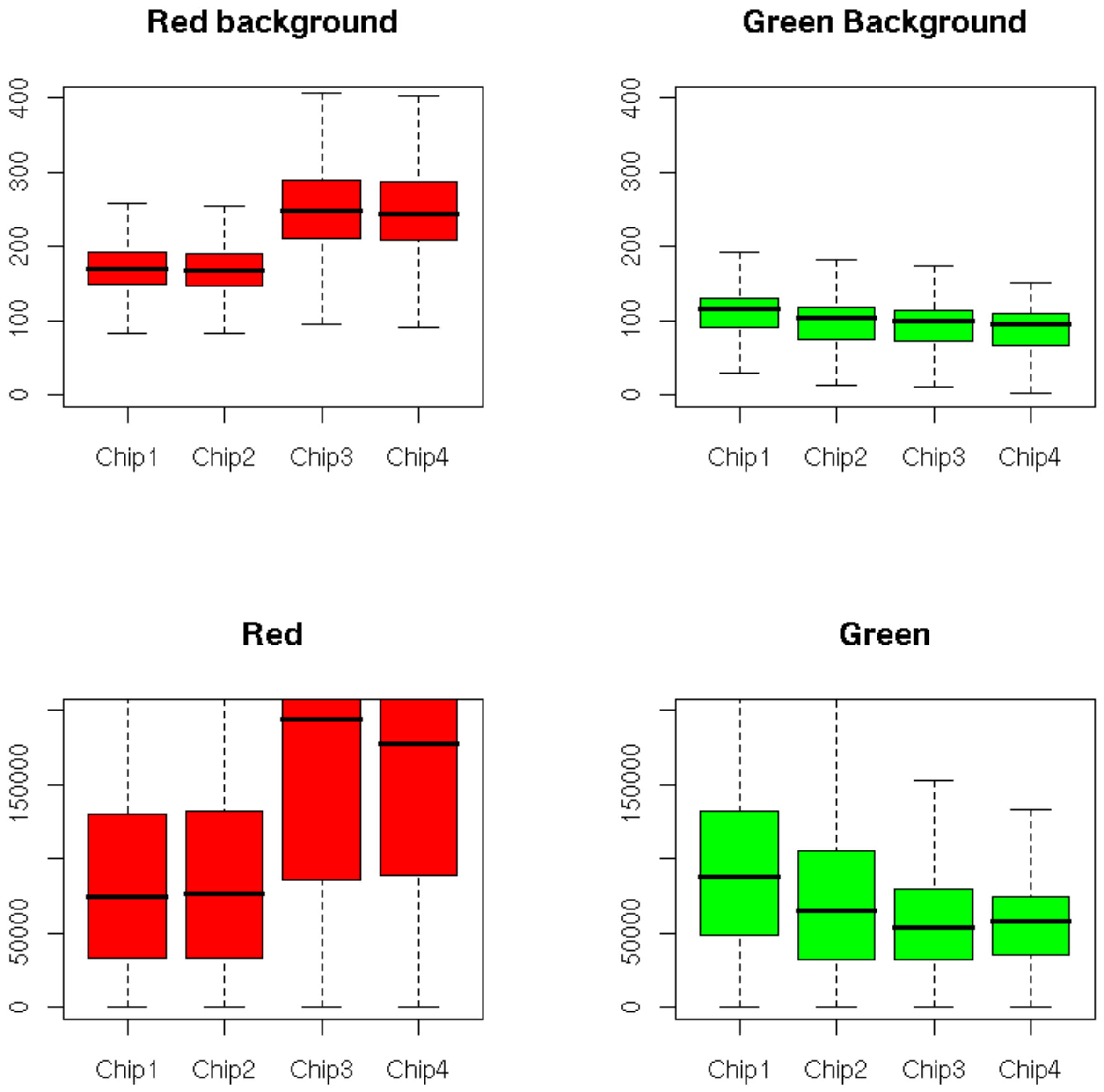
- Hybridization
  - Efficiency
  - Specificity

- Scanner settings

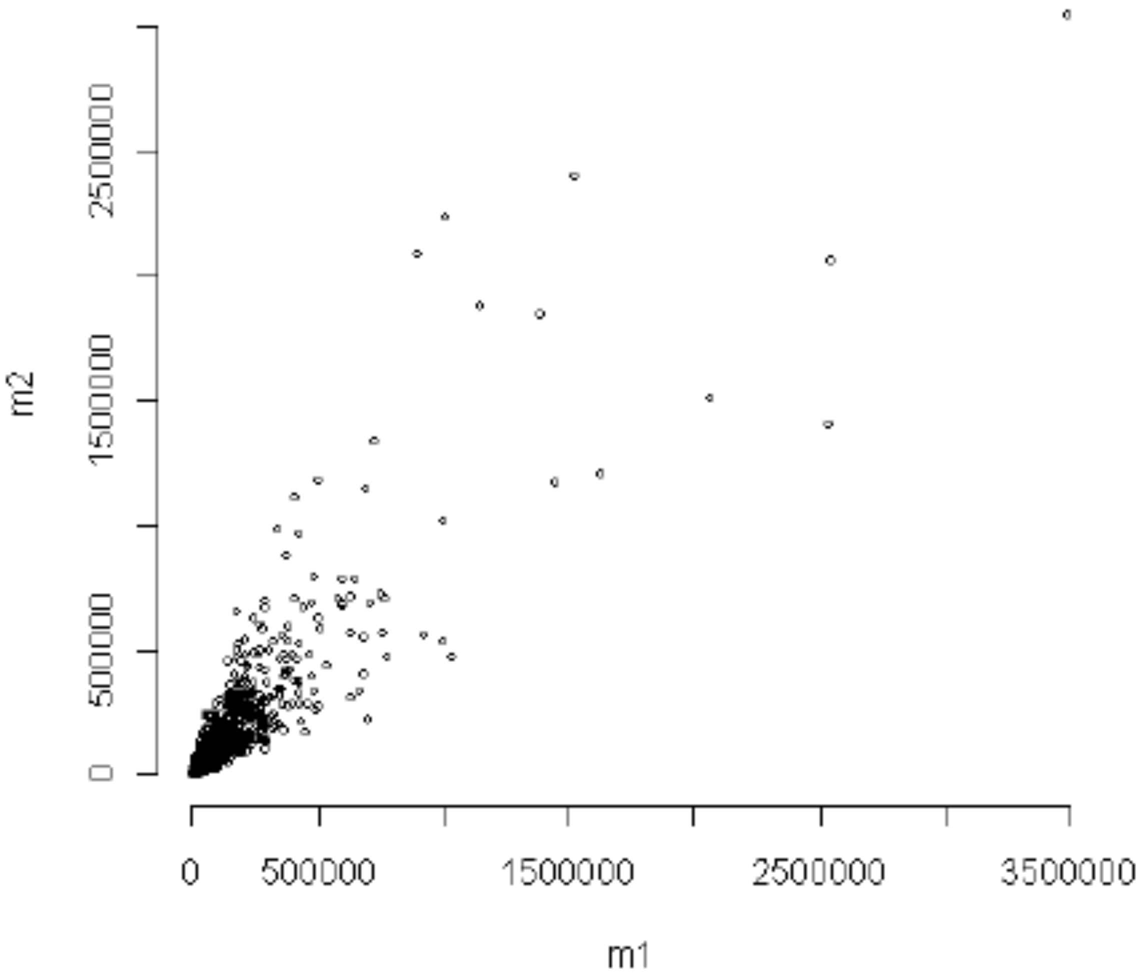
Analysis by inspection

- Box plot
- Scatter plot
- QQ plot
- MvA plot
- sdm plot
- MAD plot

Box plots


Red background
0
100
200
300
400
Chip1
Chip2
Chip3
Chip4
Green Background
0
100
200
300
400
Chip1
Chip2
Chip3
Chip4
Red
0
50000
150000
Chip1
Chip2
Chip3
Chip4
Green
0
50000
150000
Chip1
Chip2
Chip3
Chip4

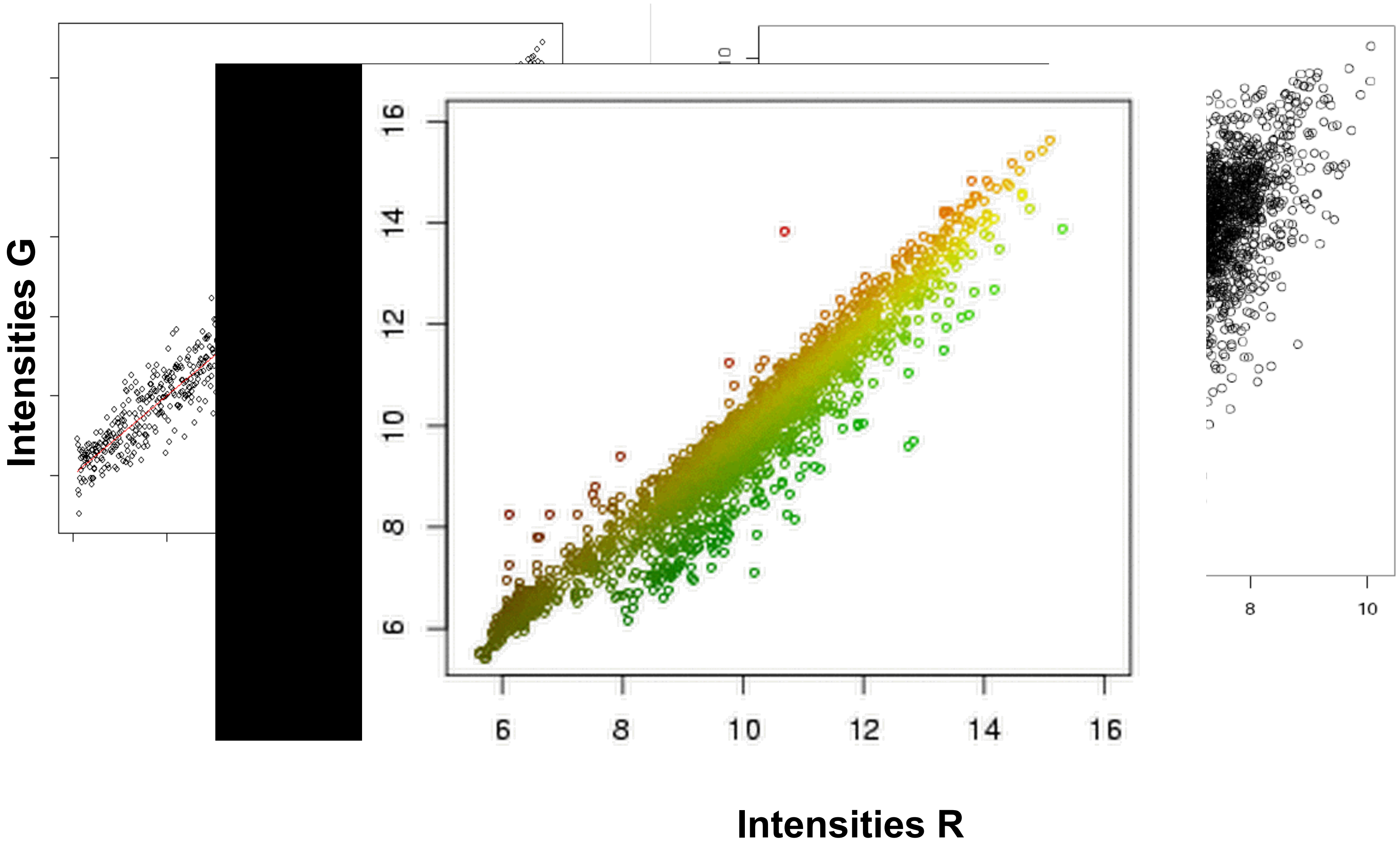
Scatter plot



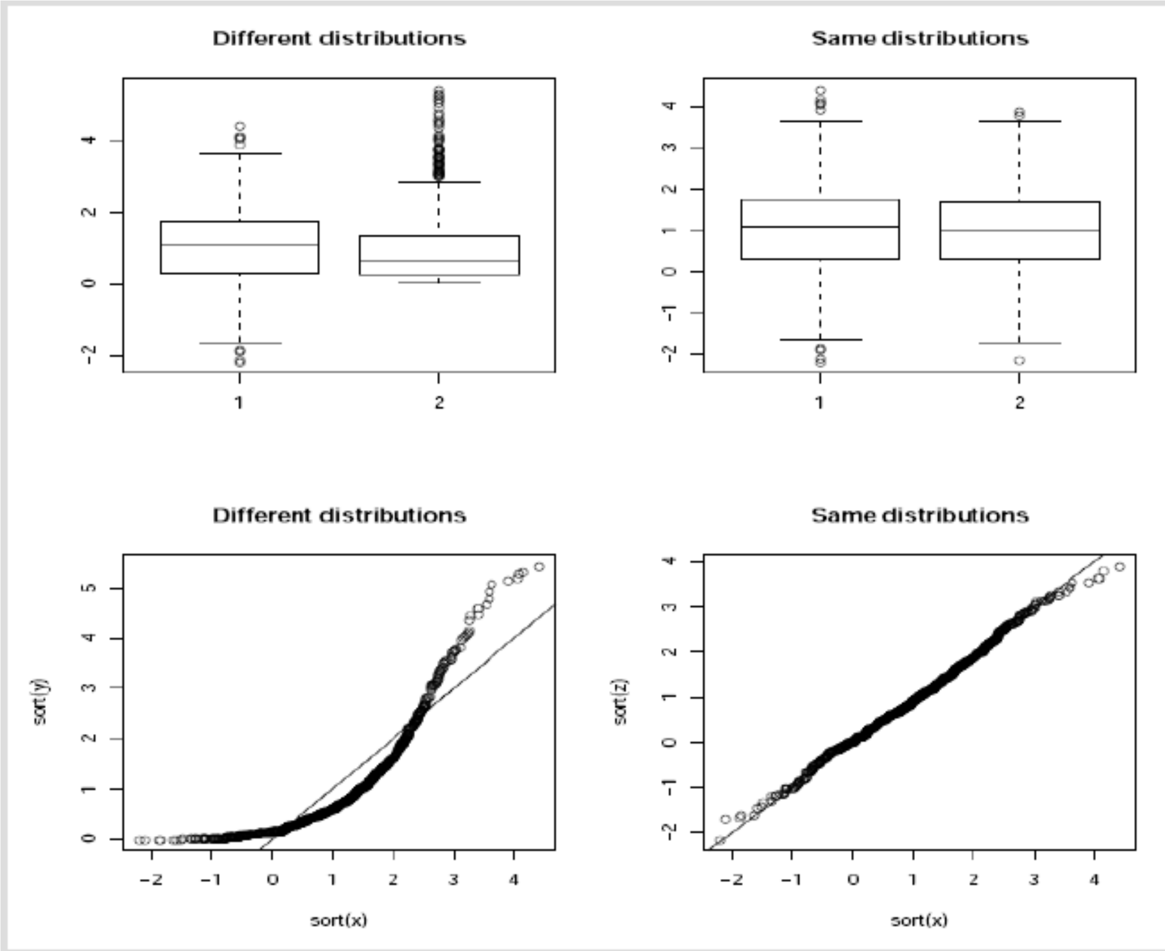
Intensities G

Intensities R

Scatter plot



QQ-plot



Different distributions

Same distributions

Different distributions

Same distributions

MvA plot

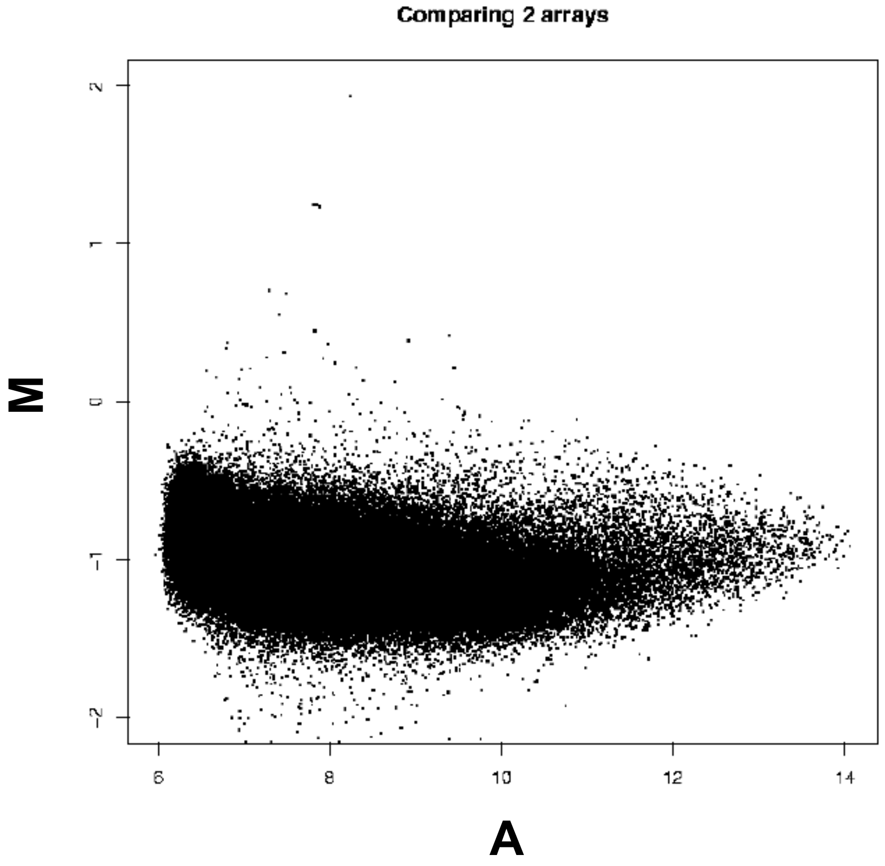
- Comparison of two arrays (Affymetrix) or two samples (e.g. Cy3 and Cy5 labeled)

- X axis: A – average intensity

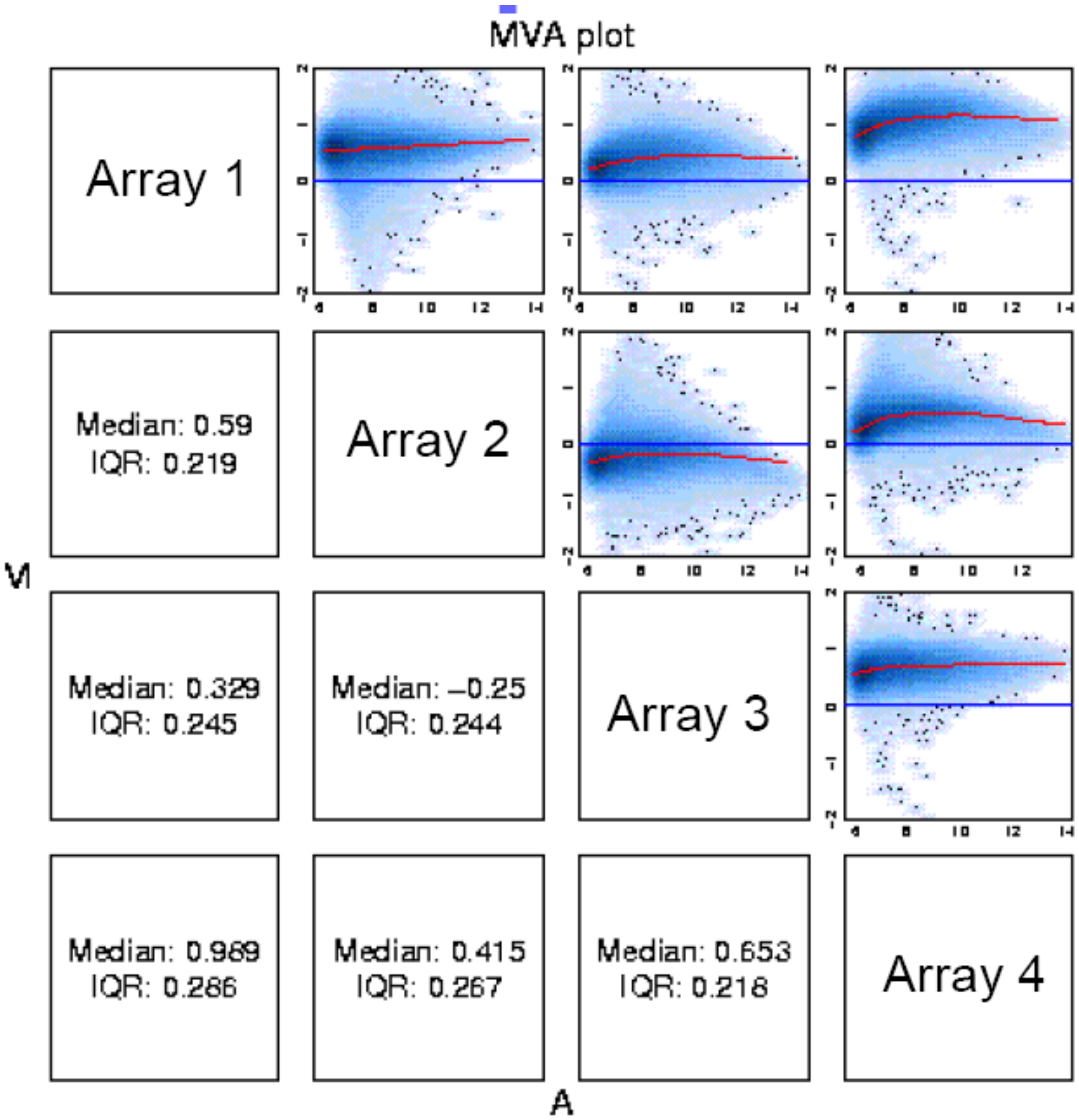
  A = 0.5*(log R + log G)

- Y axis: M – log ratio

  M = log R – log G

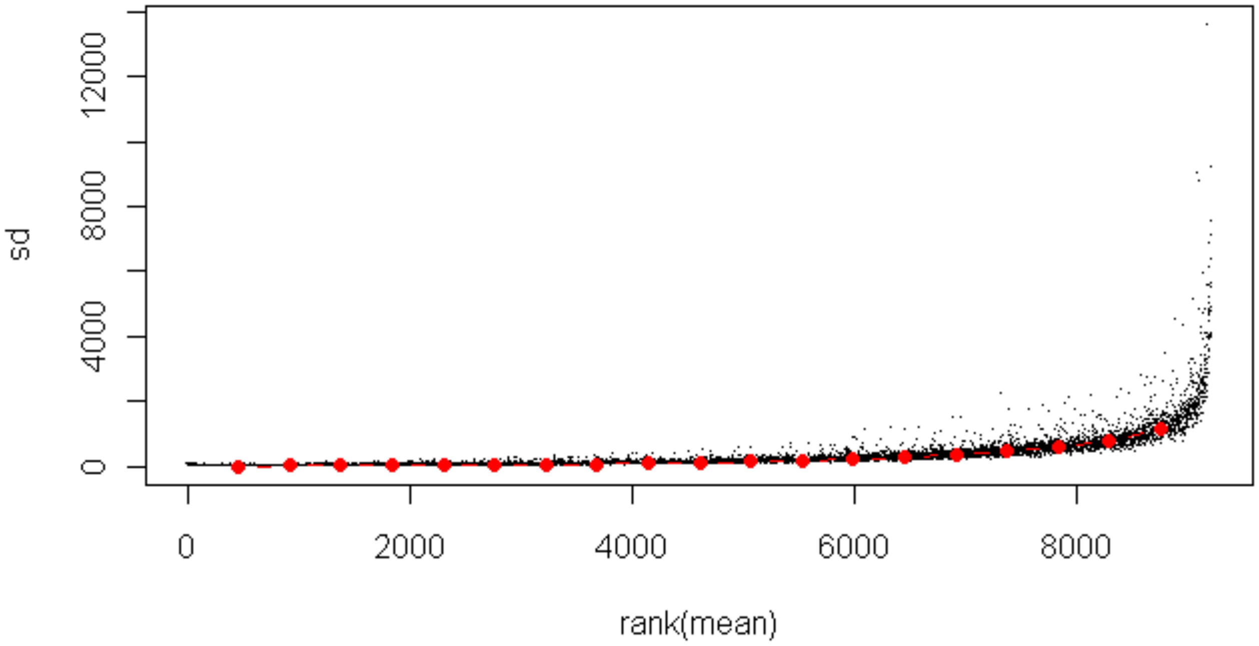


MvA plots


MVA plot
Array 1
Median: 0.59
IQR: 0.219
Array 2
Median: 0.329
IQR: 0.245
Median: –0.25
IQR: 0.244
Array 3
Median: 0.989
IQR: 0.266
Median: 0.415
IQR: 0.267
Median: 0.653
IQR: 0.218
Array 4
M
A
6
8
10
12
14
-2
-1
0
1
2

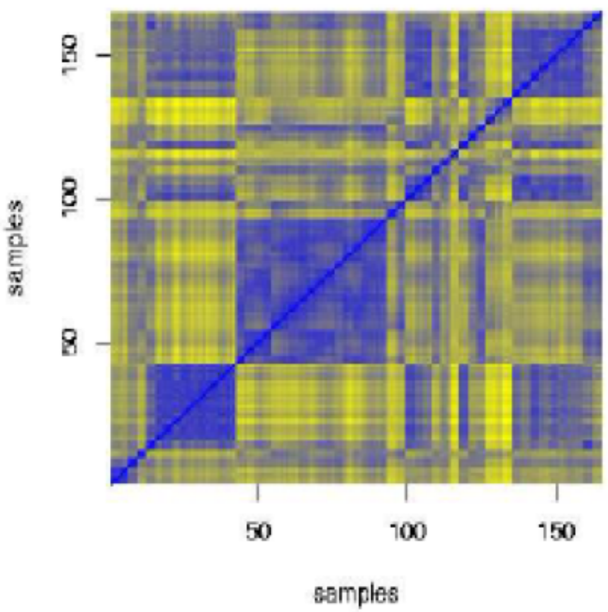
SDM plots

- Standard deviation vs. mean

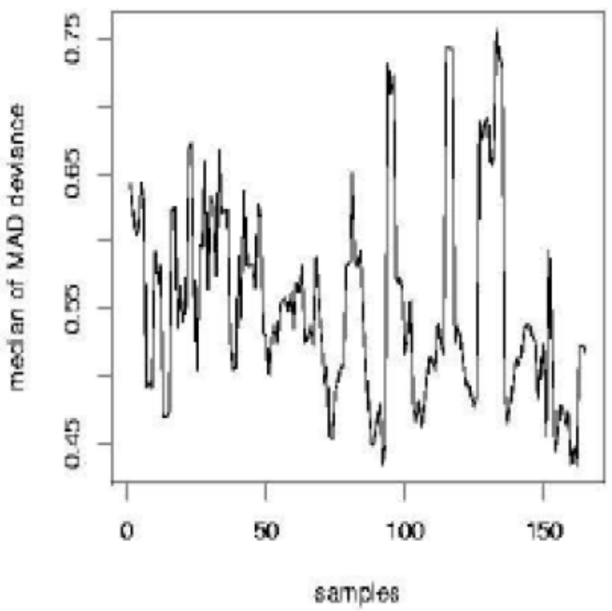


Median absolute deviation

- Comparison between arrays
- $\mathrm{MAD}_{i,j} = \mathrm{median}_j\{|x_{i1} - x_{j1}|,|x_{i2} - x_{j2}|, \ldots)\}$



sorted by tissue type



sorted by tissue type

Normalization

Sources of Artifacts

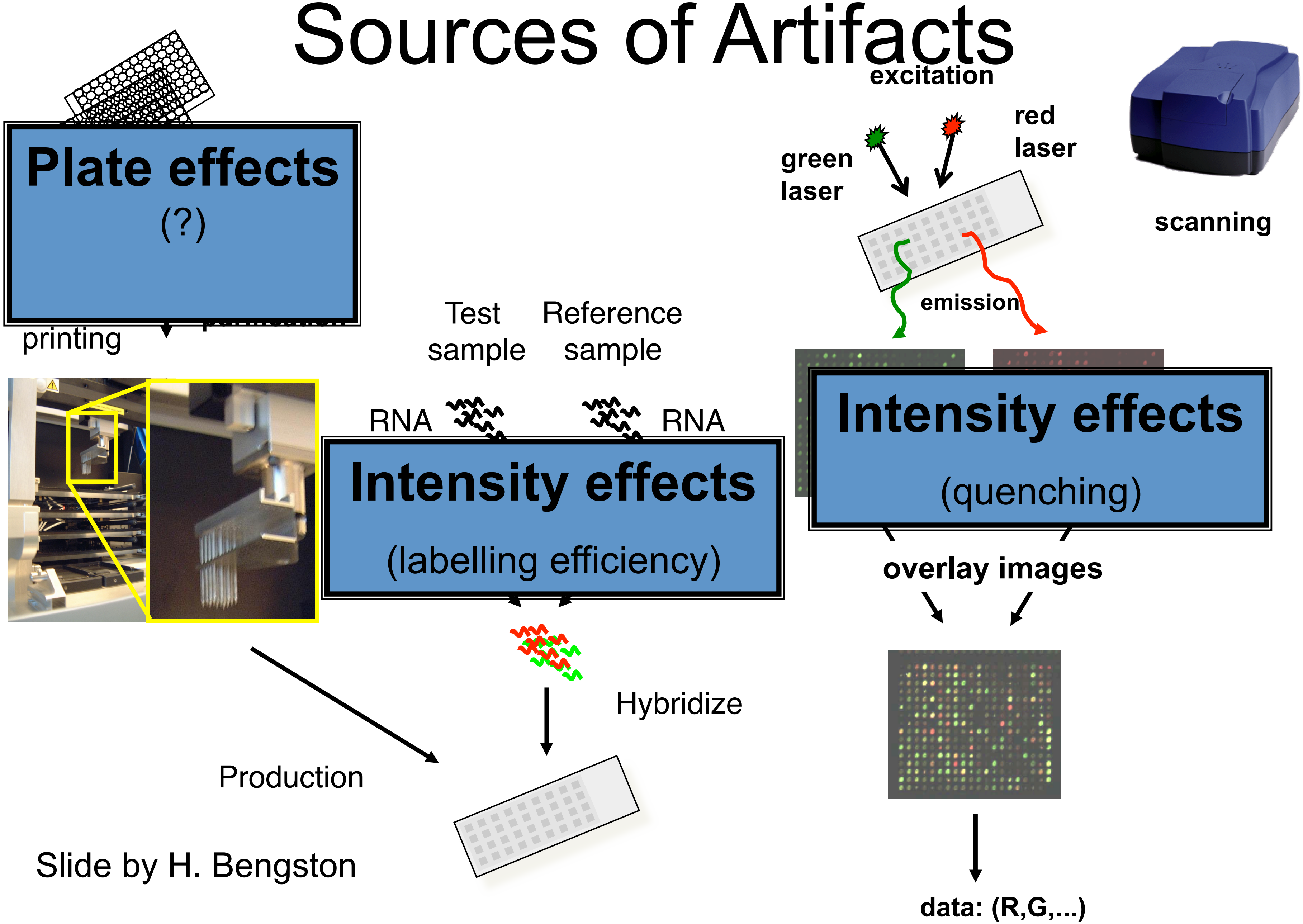


Plate effects

(?)

printing

purification

Test sample

Reference sample

RNA

RNA

Intensity effects

(labelling efficiency)

Hybridize

Production

excitation

green laser

red laser

emission

scanning

Intensity effects

(quenching)

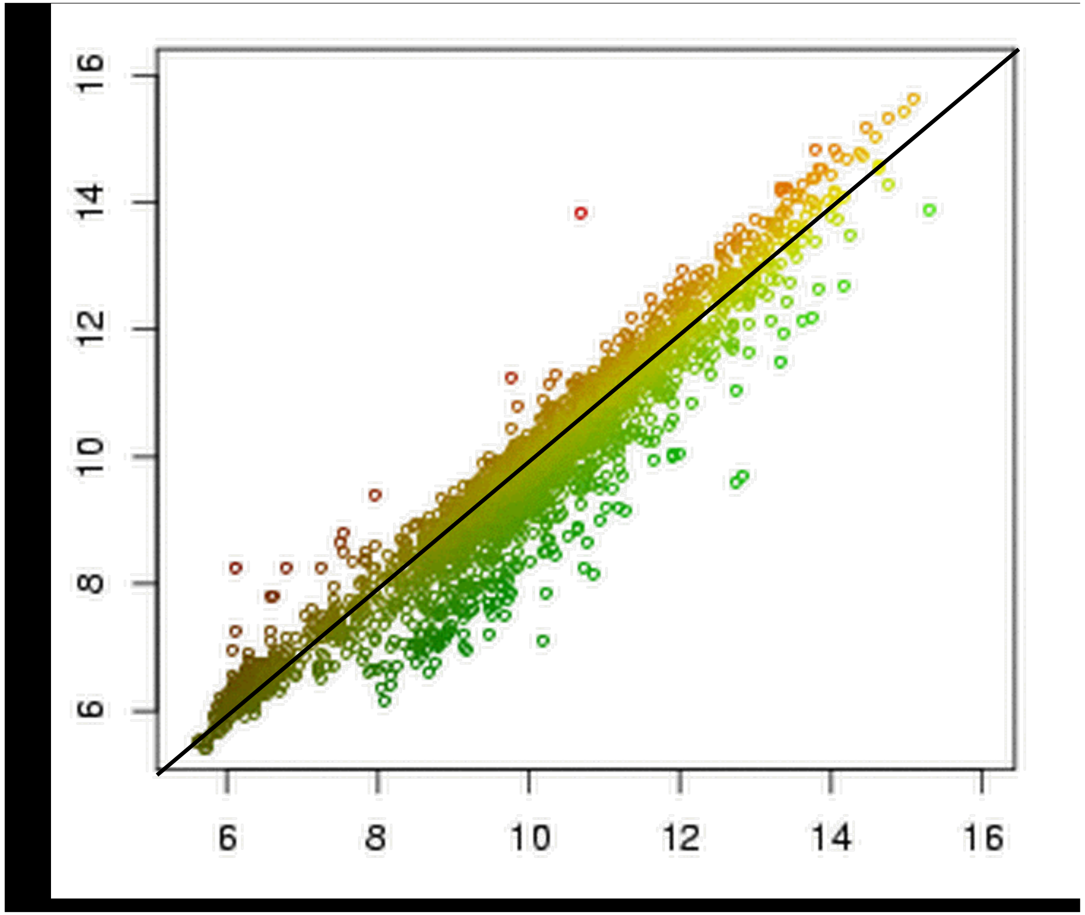
overlay images

data: (R,G,...)

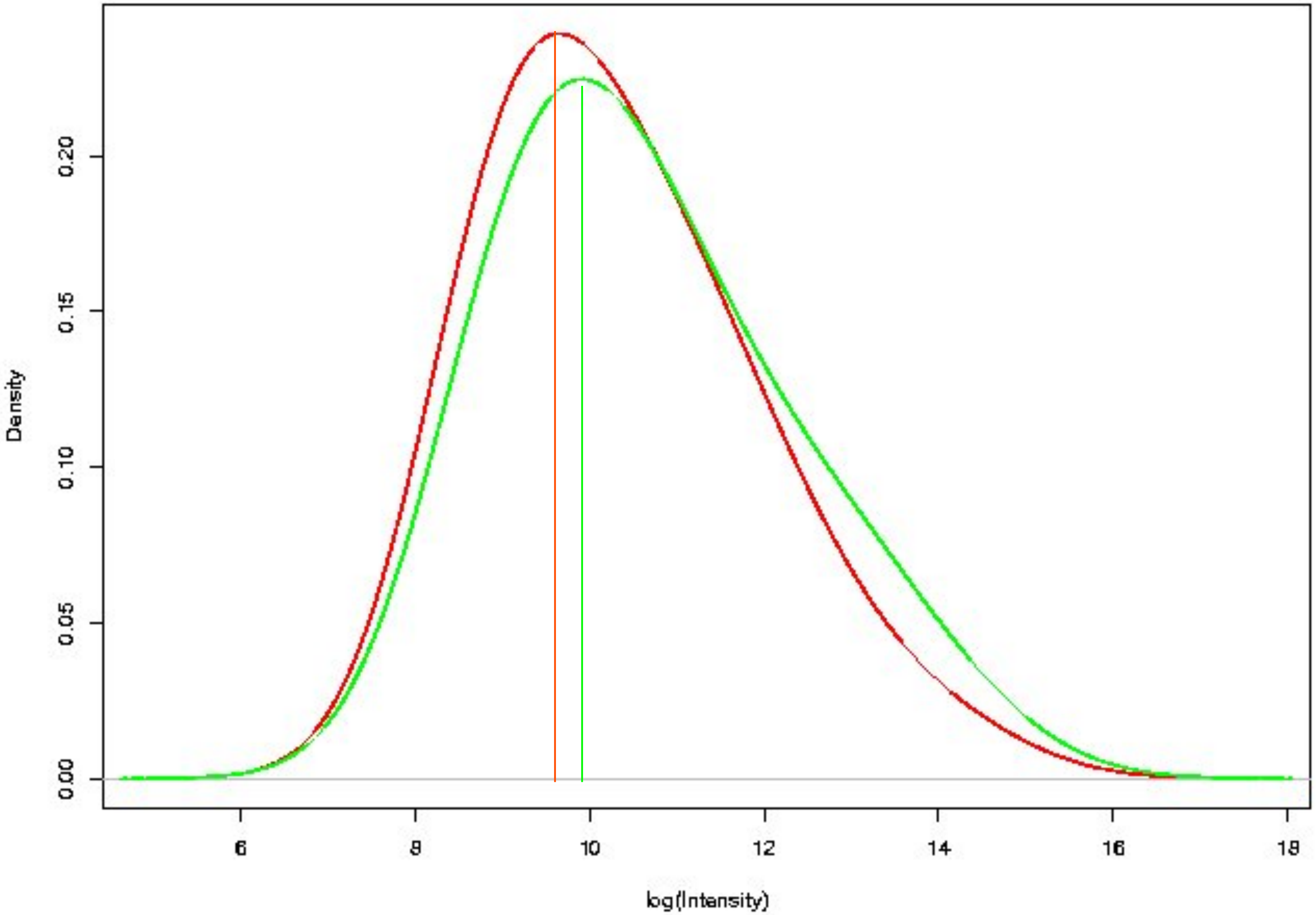
Slide by H. Bengston

Hybridization of the same sample to 2 chips/ channels

- Random and systematic measurement errors
- Biases result in scatter plots not centered around the x-y diagonal



Hybridization of the same sample to 2 chips/ channels



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?

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)

House keeping genes

- 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

Bias detection

- 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
- “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

Influence parameters

Commonly used approaches

- Global intensity scaling
- LOESS
- Quantil normalization
- Variance stabilized normalization (vsn)

Global normalization (Scaling)

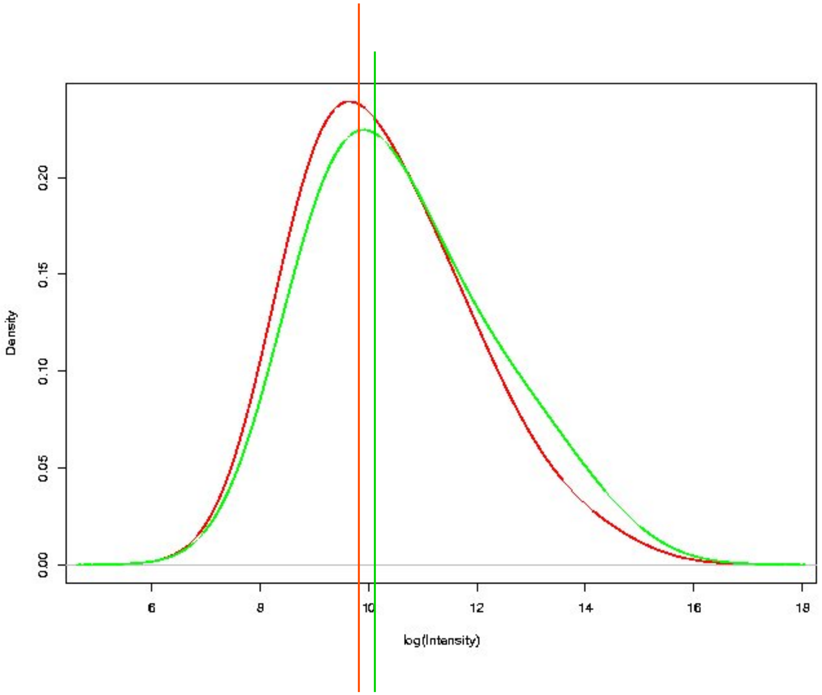
- A single normalization factor (k) is computed for balancing chips\channels:

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

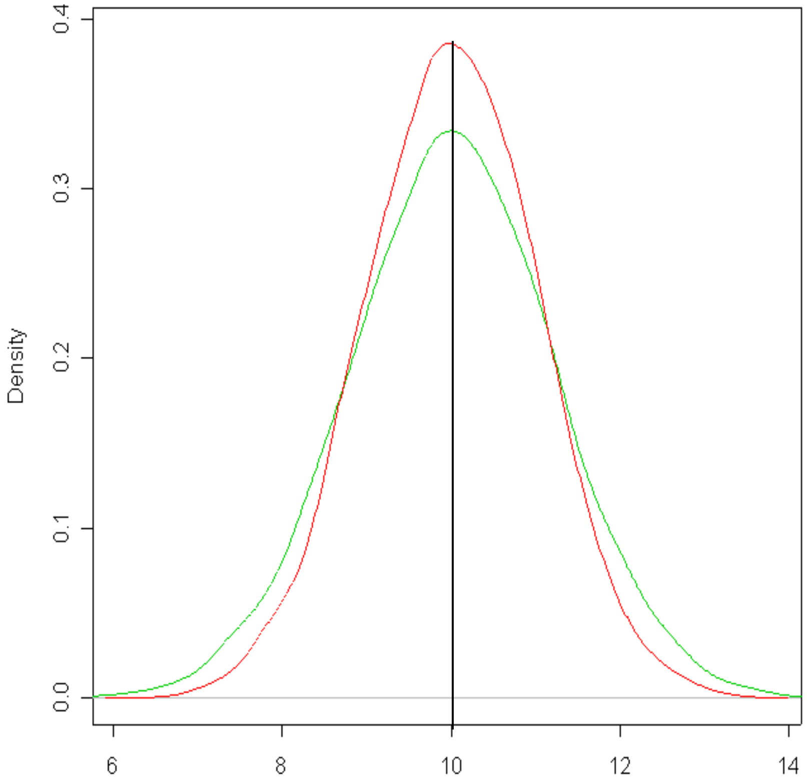
- Multiplying intensities by this factor equalizes the mean (median) intensity among compared chips
- Found in many papers, not recommended

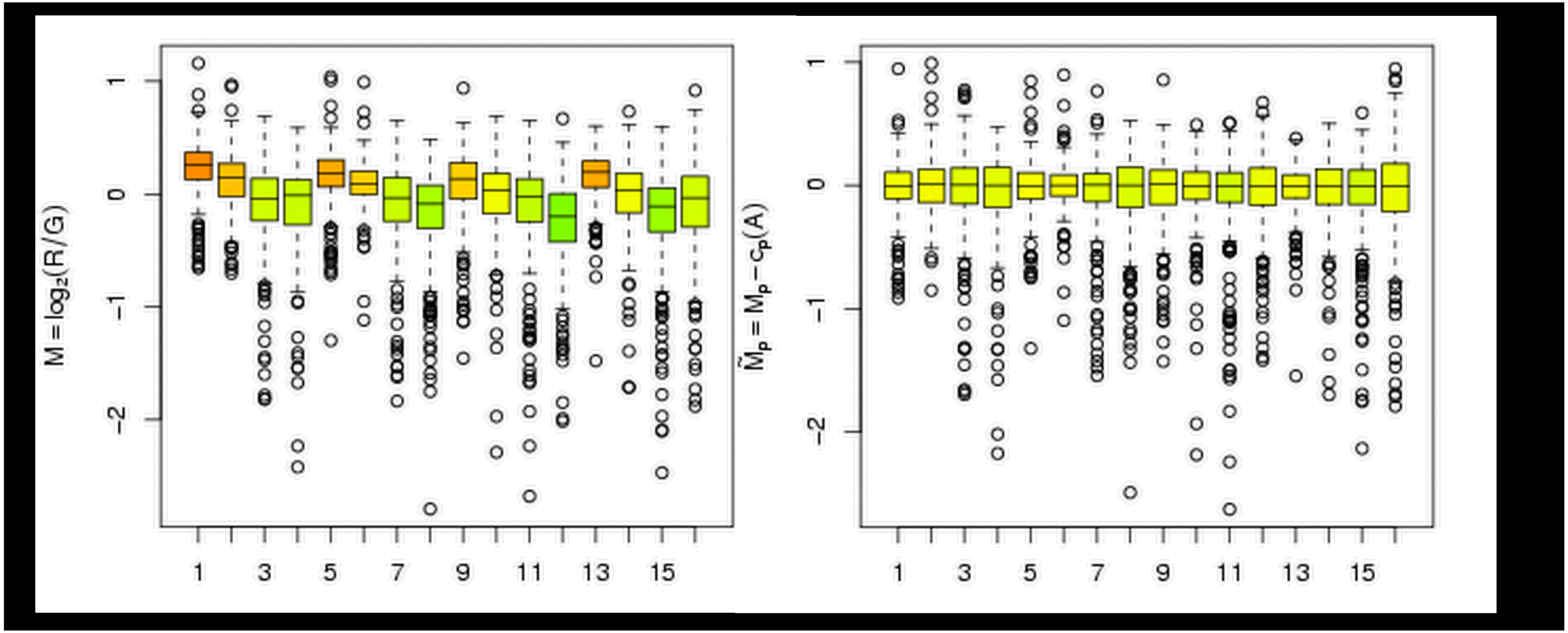
Global Normalization

Before



After





Before Normalization

After Scaling

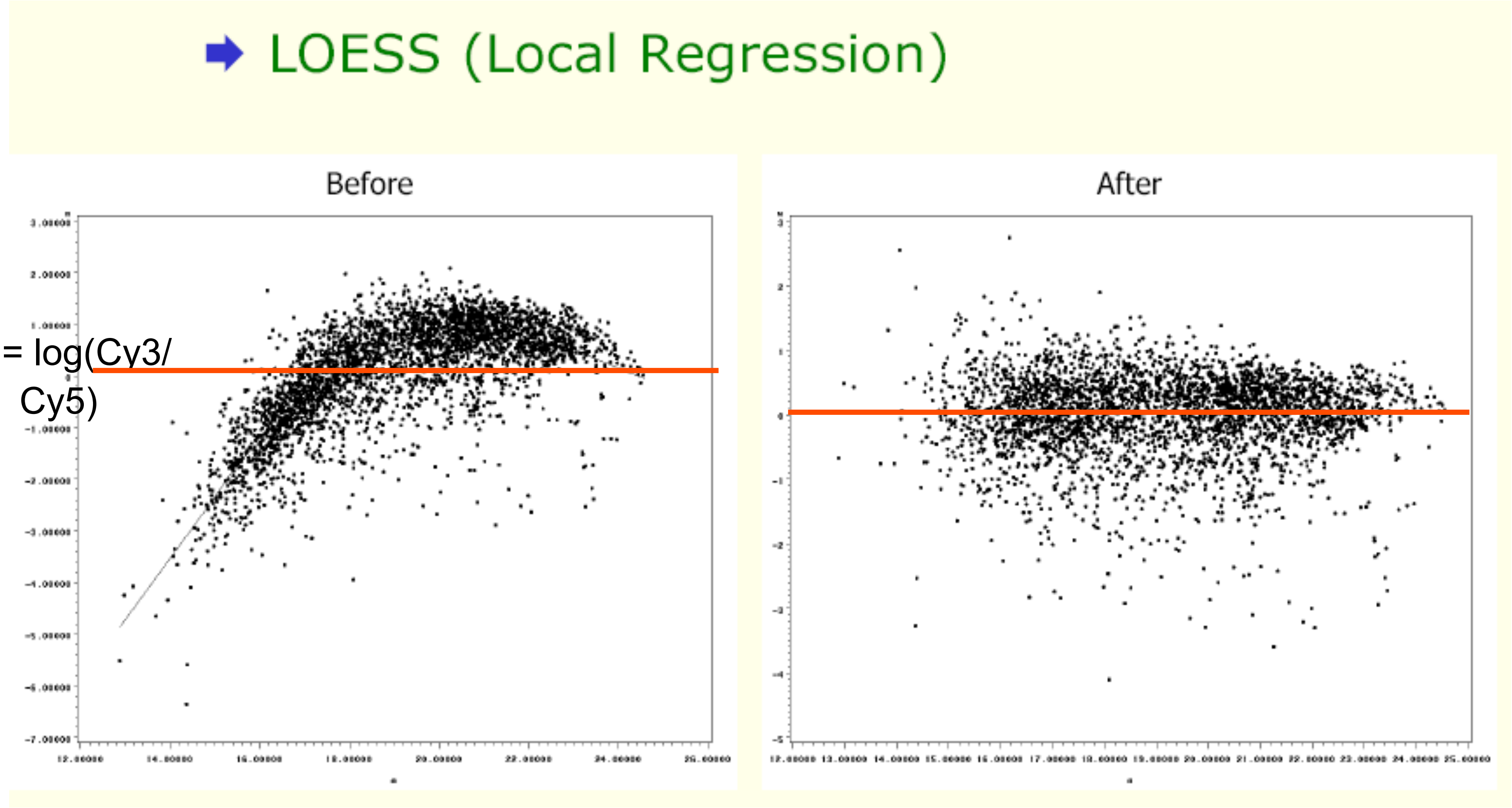
LOESS

- Locally weighted scatter plot smoothing
- Synonymous with *lowess*

- Compensate for intensity-dependent biases

- Separate the data into windows of a given size
- Apply a regression function to the segmented data

We expect the M vs A plot to look like:



LOESS (Local Regression)

M = log(Cy3/ Cy5)

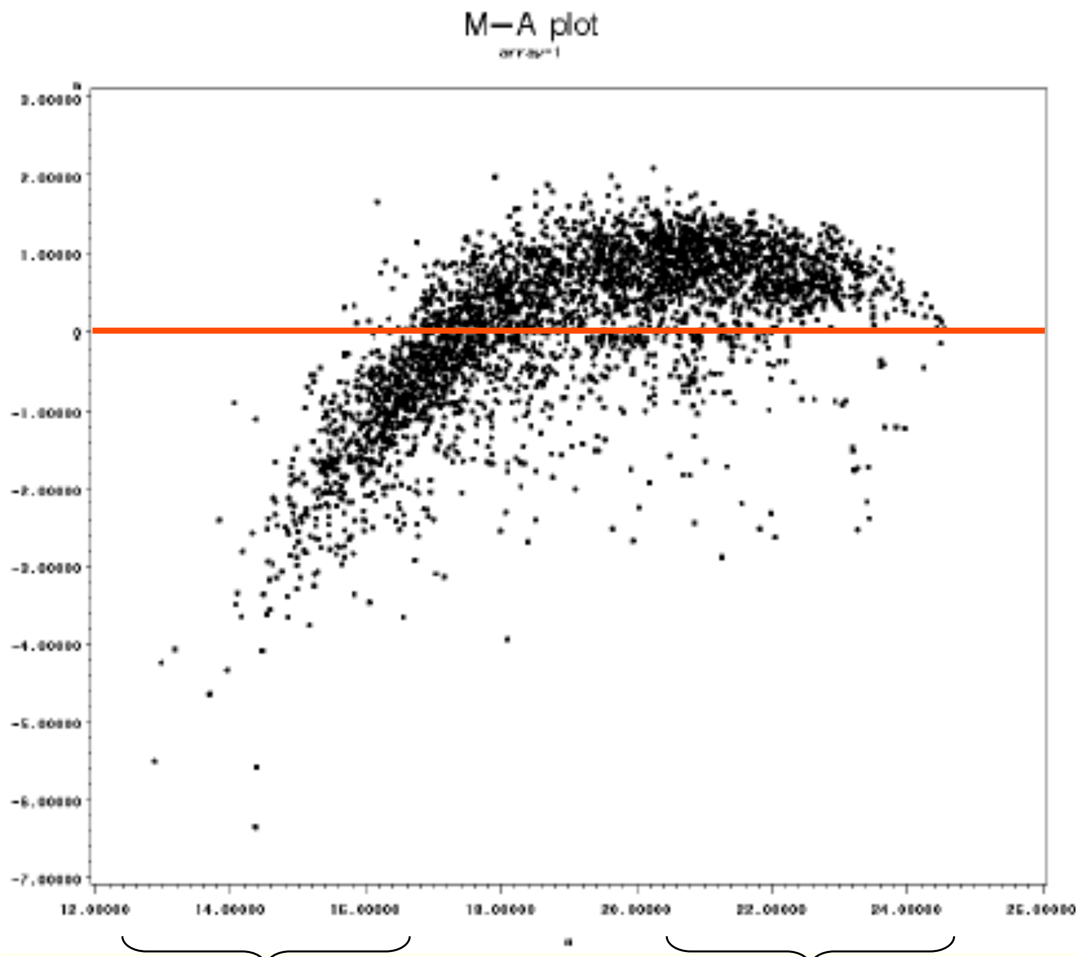
A

Intensity-dependent bias

M>0:
Cy3>Cy5

M = log(Cy3/Cy5)

M<0:
Cy3<Cy5

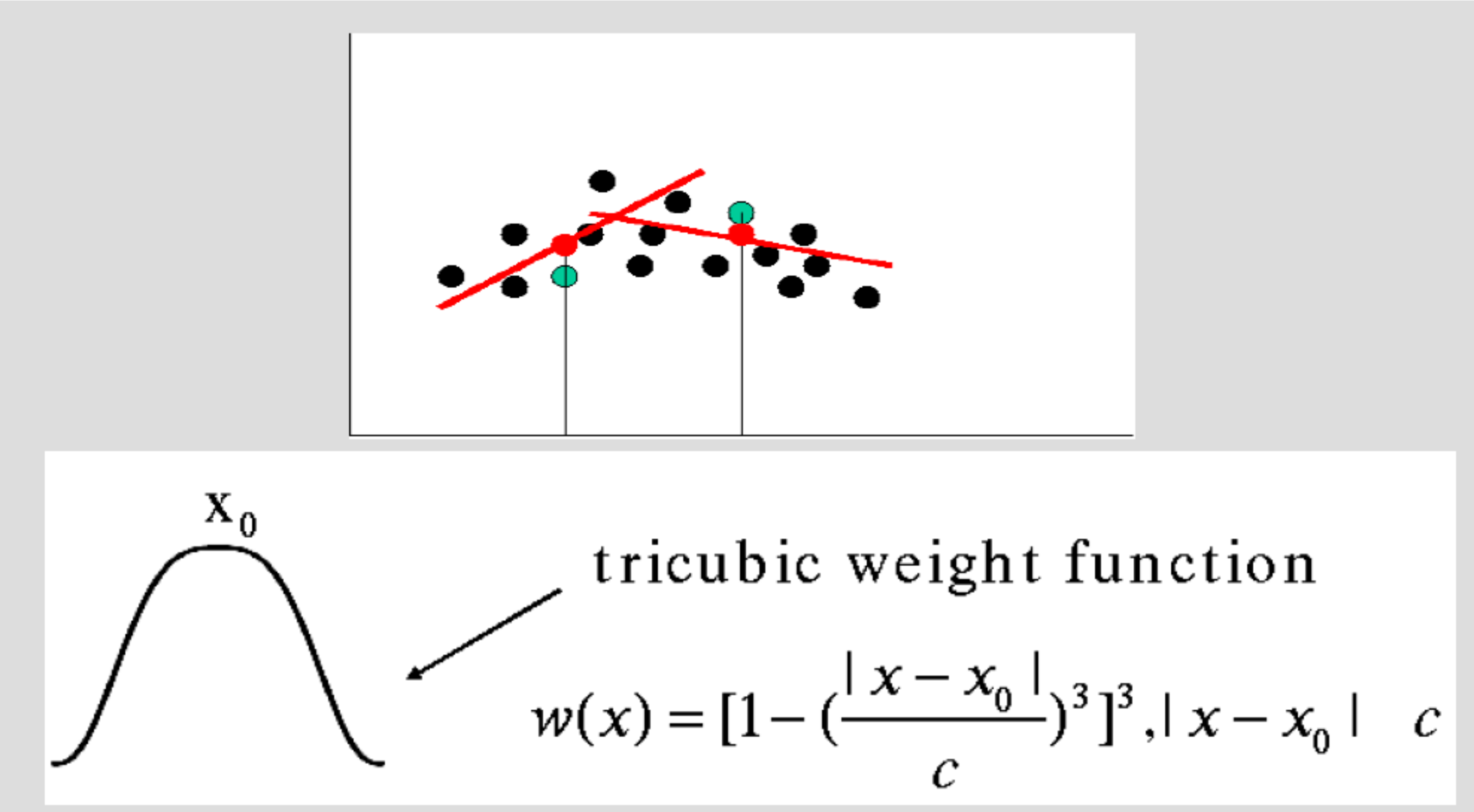


Low intensities

A

High intensities

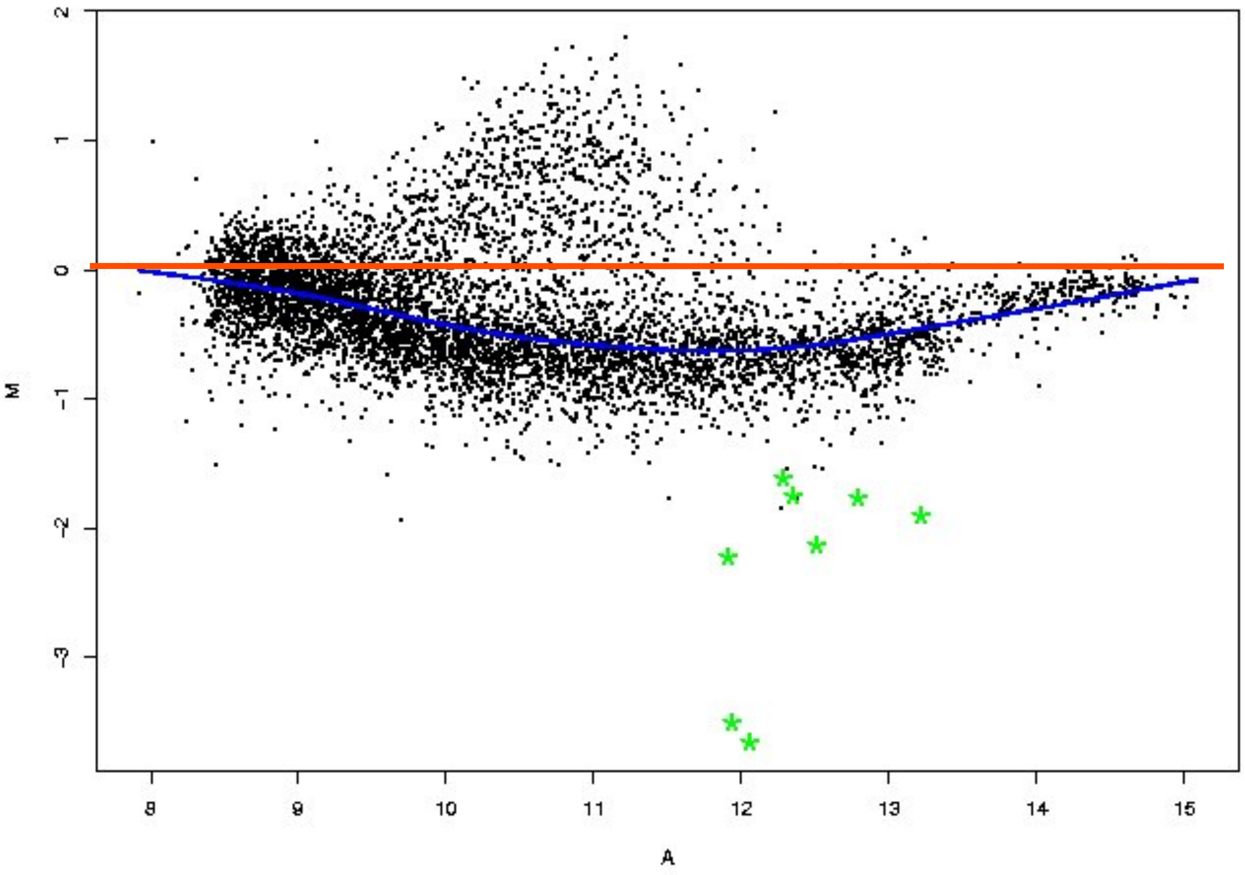
Separate data



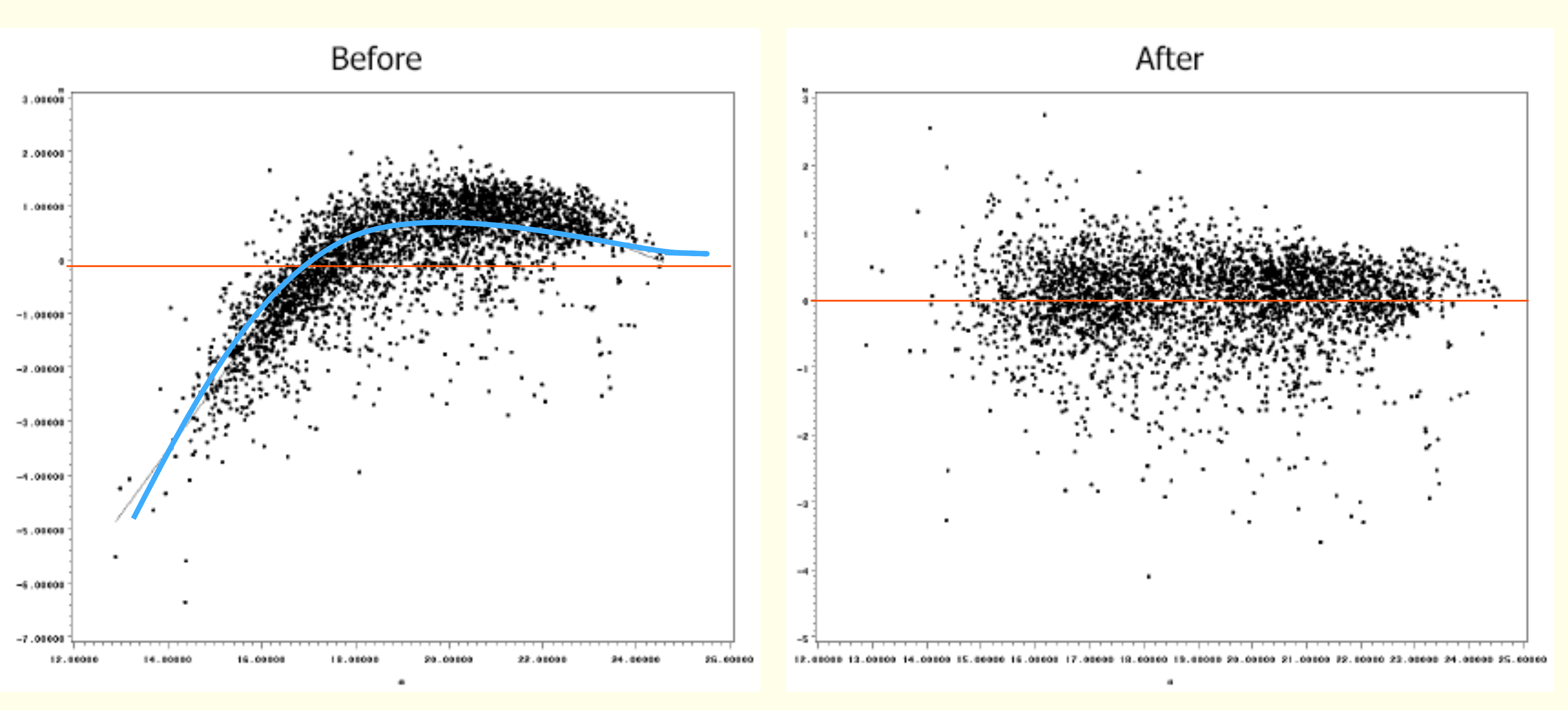
Intensity-Dependent Normalization

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

Lowess – fitting local regression curve – c(A)



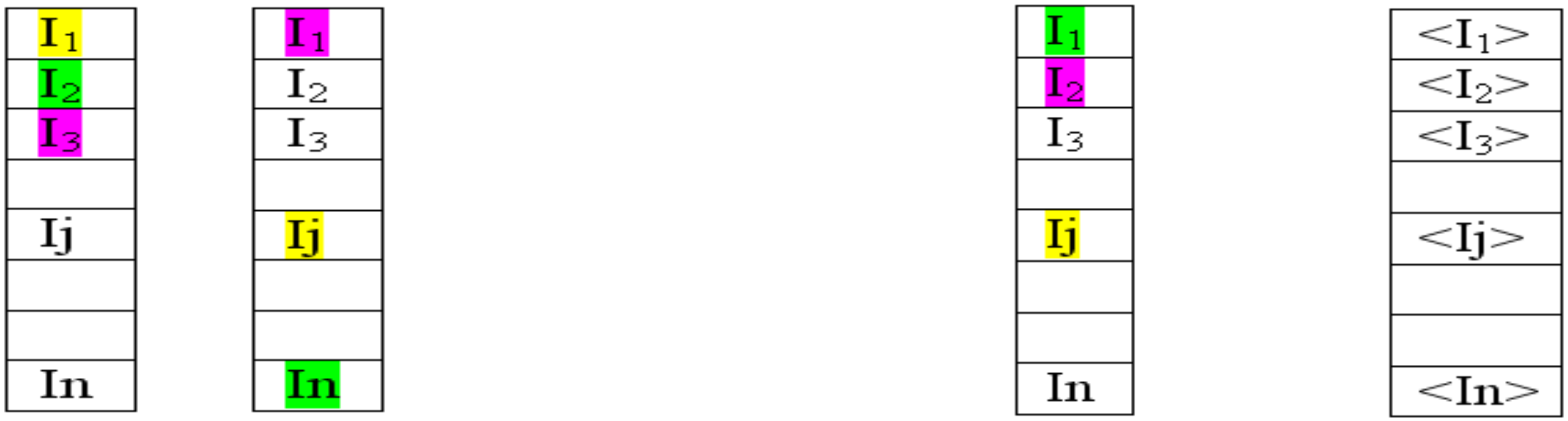
LOWESS normalization


Before

After

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

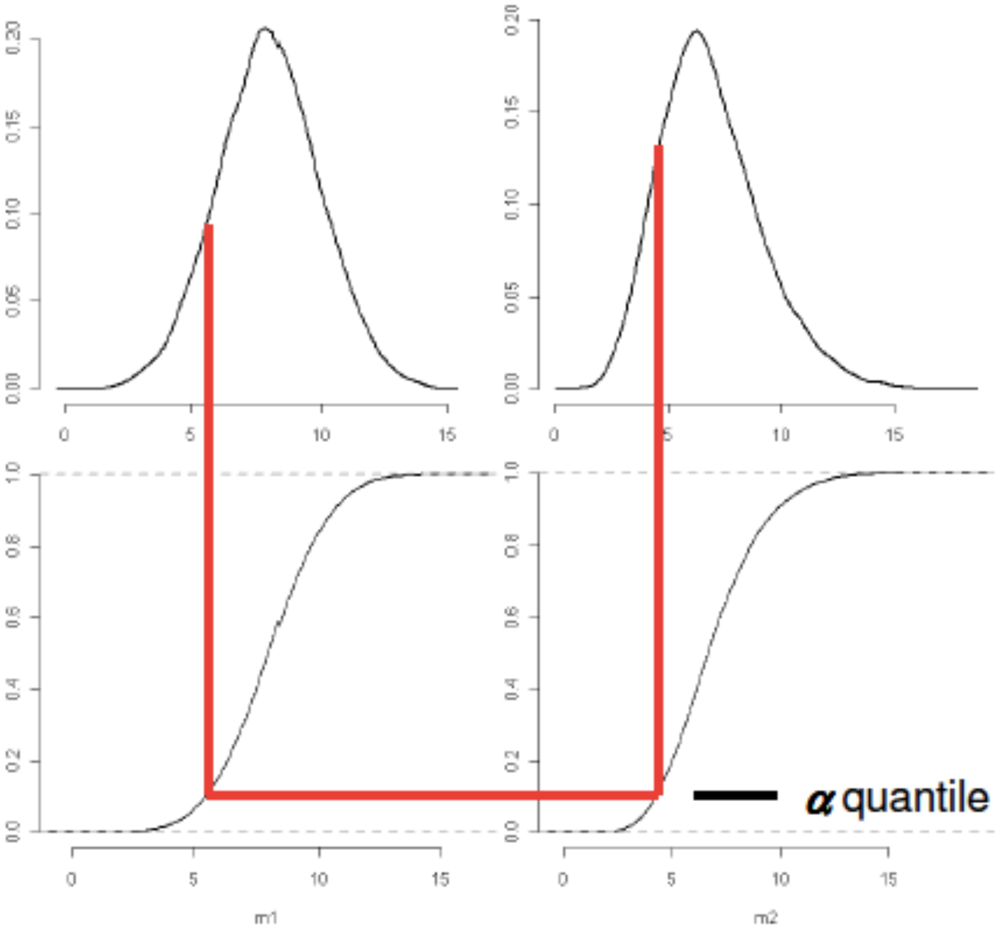


Chip #1 Chip #2 Chip #3 Average chip

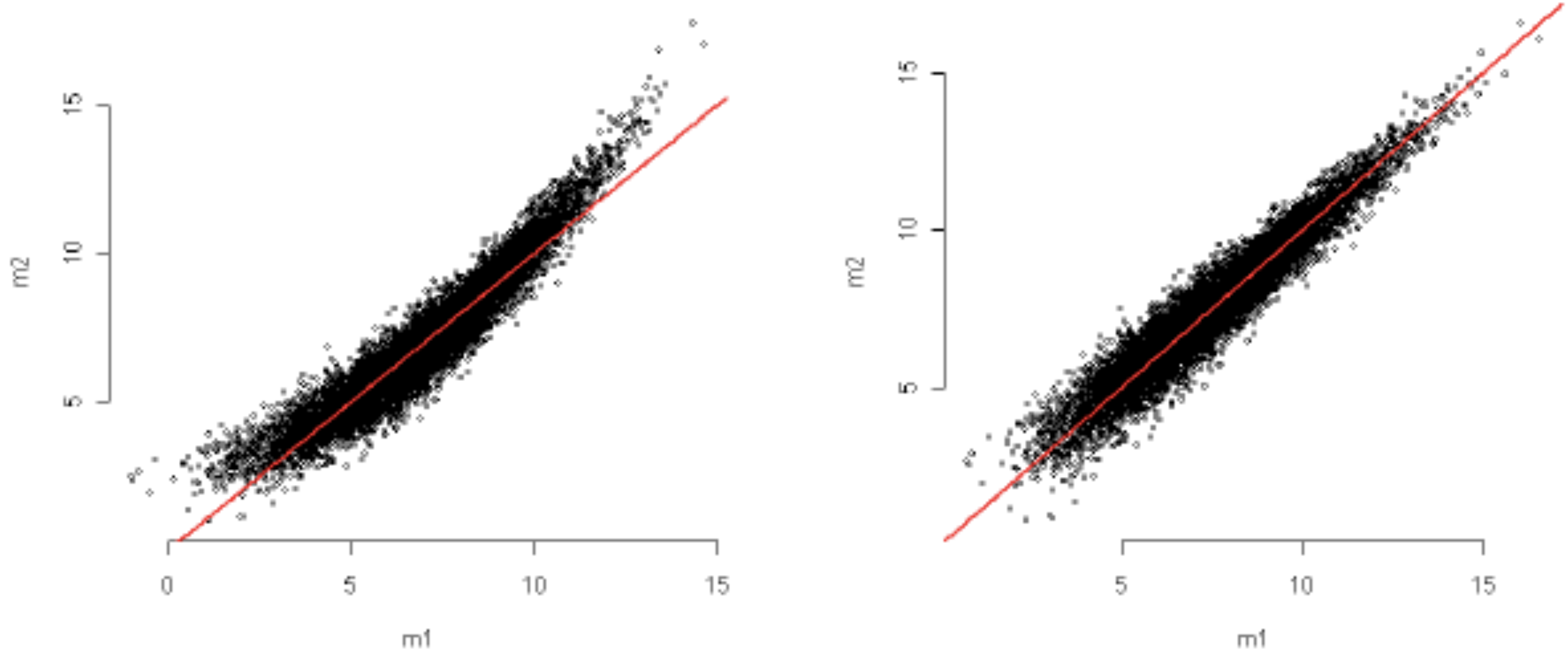
Quantile normalization

density

cummulative distribution

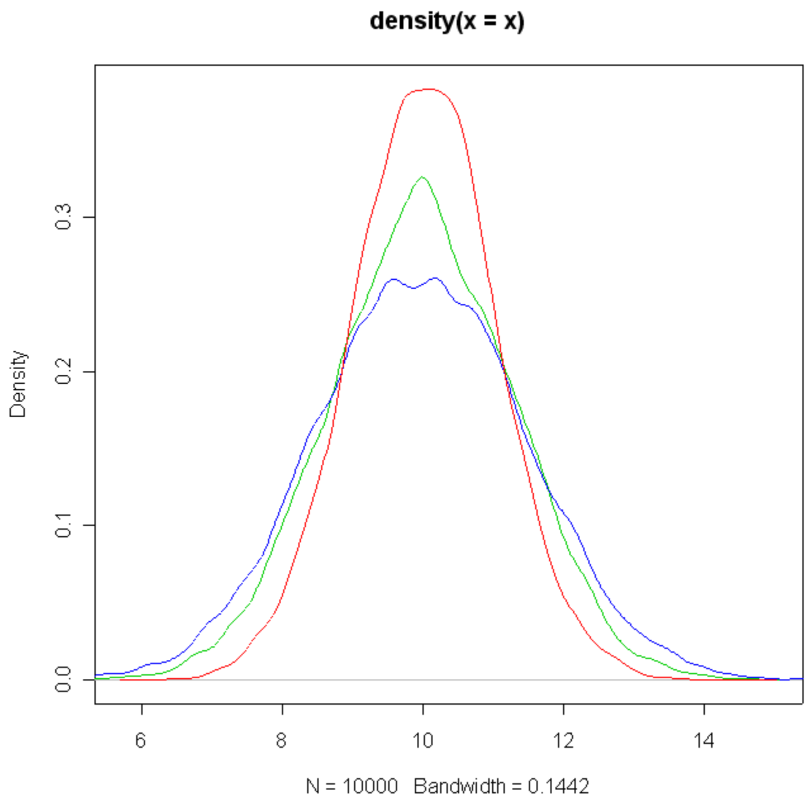


Quantile normalization

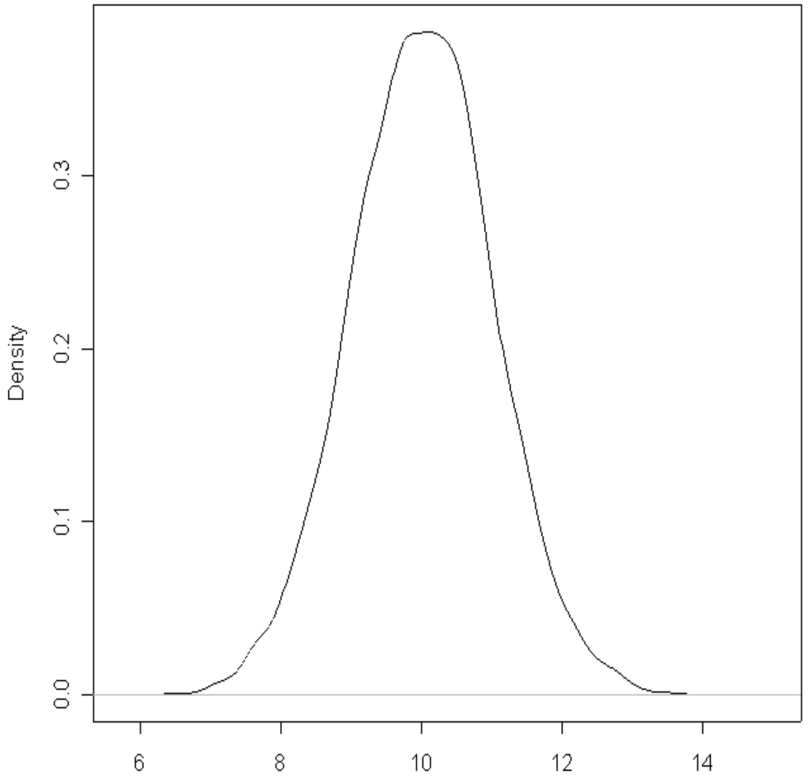


Comparison

After lowess normalization



After quantile normalization



Variance stabilized normalization

Measured intensity = offset + gain x true abundance

$$Y_{ik} = \alpha_{ik} + \beta_{ik}\, x_k$$

$\alpha_{ik} = \alpha_i + e_{ik}$

a_i : per sample offset

e_{ik} : additive noise ~ N $(0, b_i s_1^2)$

$\beta_{ik} = \beta_i \beta_k \exp(n_{ik})$

β_i : per sample normalization factor

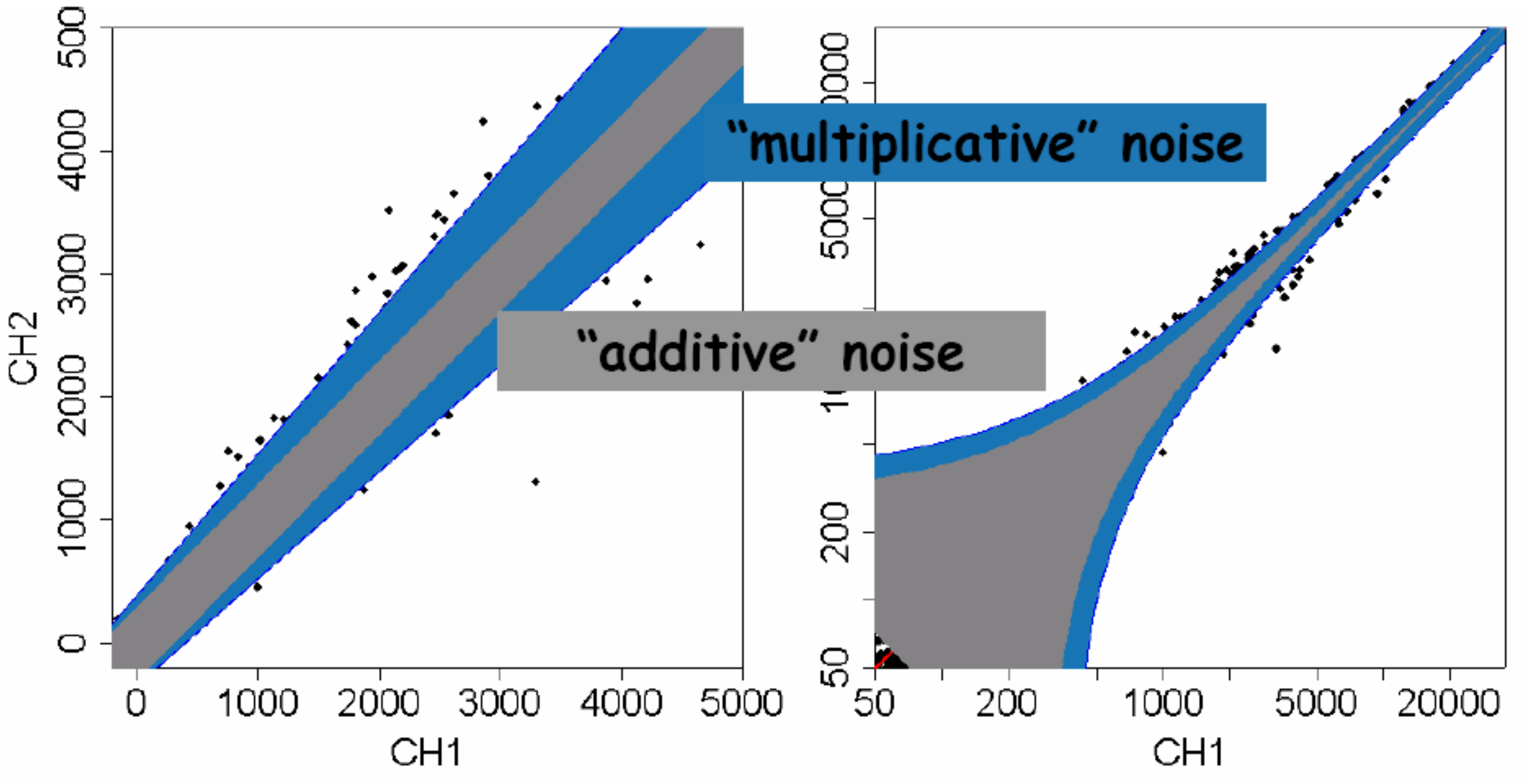
β_k : sequence-wise labeling efficiency

n_{ik} ~ N(0, $s_2 2$) : multiplicative noise

Variance stabilizing normalization

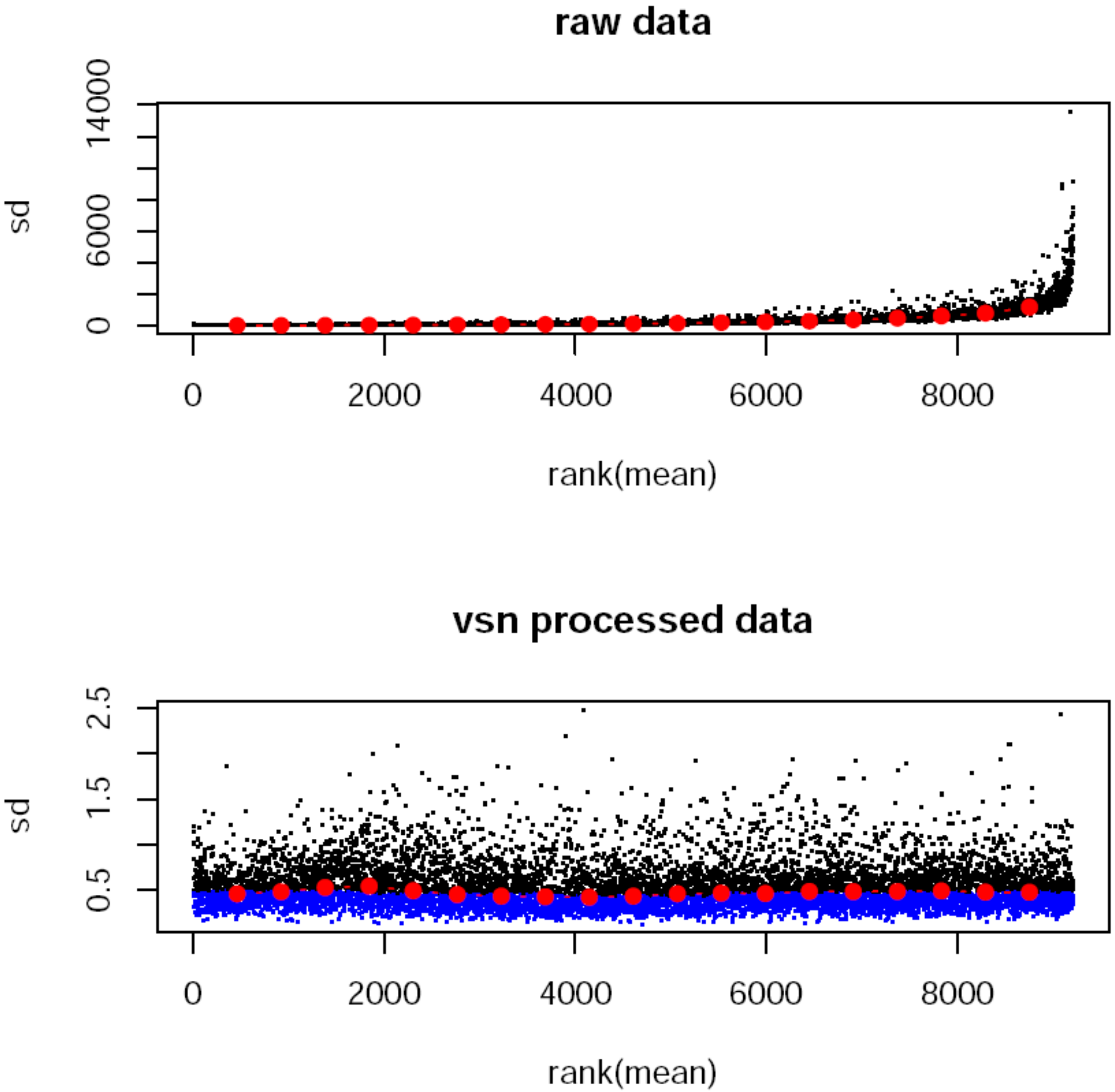
- Powerful method incorporating
  - Background substraction
  - Error model
  - Analysis of significantly expressed genes
- Typically employed in the analysis of ratios
  - Many genes are lowly expressed

Additive vs. multiplicative noise

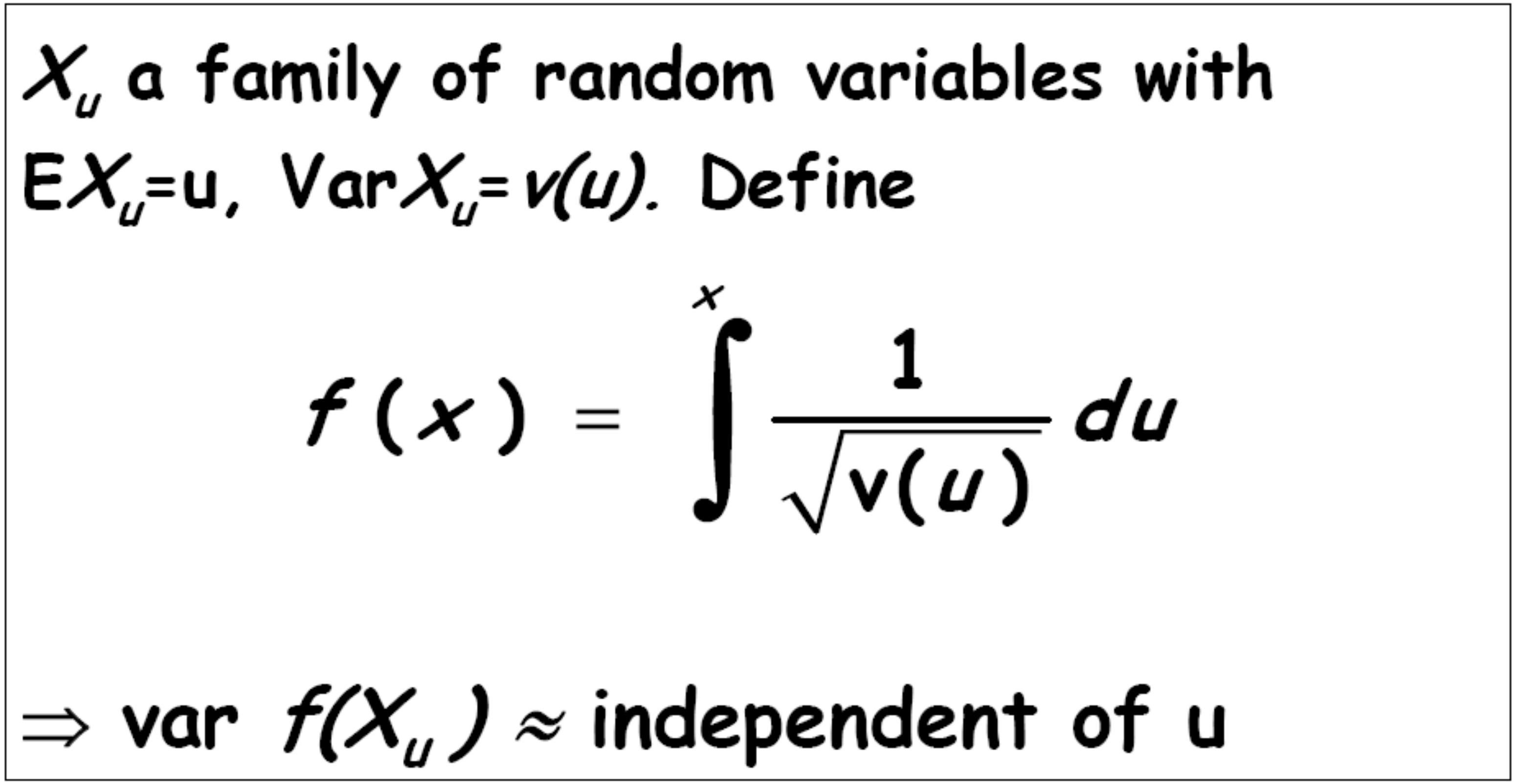


raw scale **log scale**

From Huber


raw data
sd
0
6000
14000
0
2000
4000
6000
8000
rank(mean)
vsn processed data
sd
0.5
1.5
2.5
0
2000
4000
6000
8000
rank(mean)

Variance stabilizing transformation



X_u a family of random variables with

EX_u=u, VarX_u=$v(u)$. Define

$$f(x) = \int^{x} \frac{1}{\sqrt{v(u)}}\,du$$

$\Rightarrow$ var $f(X_u) \approx$ independent of u

derivation: linear approximation

vsn transformation

$$\boxed{f(x) = \int^{x} \frac{1}{\sqrt{v(u)}} \, du}$$

1.) constant variance ('additive') $v(u) = s^2 \quad \Rightarrow \quad f \propto u$

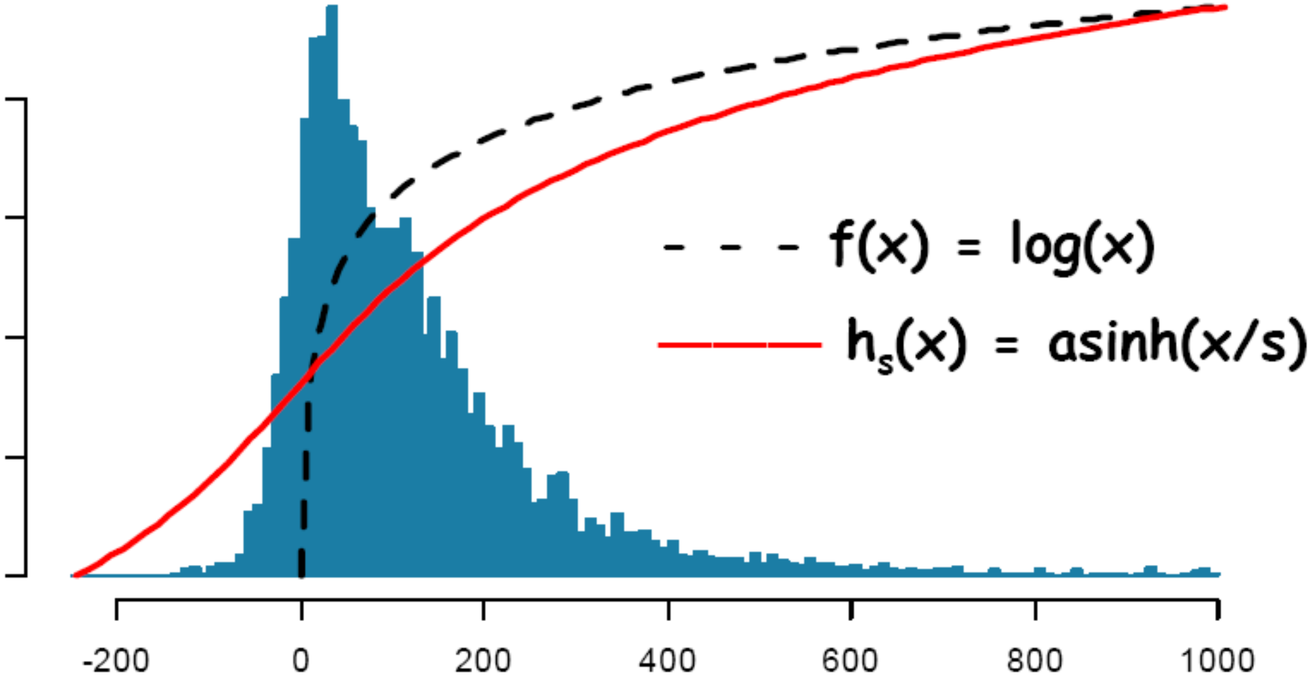
2.) constant CV ('multiplicative') $v(u) \propto u^2 \Rightarrow f \propto \log u$

3.) offset $v(u) \propto (u + u_0)^2 \quad \Rightarrow \quad f \propto \log(u + u_0)$

4.) additive and multiplicative

$$v(u) \propto (u + u_0)^2 + s^2 \Rightarrow f \propto \operatorname{arsinh} \frac{u + u_0}{s}$$

arsinh and log



$$\operatorname{arsinh}(x) = \log\left(x + \sqrt{x^2 + 1}\right)$$

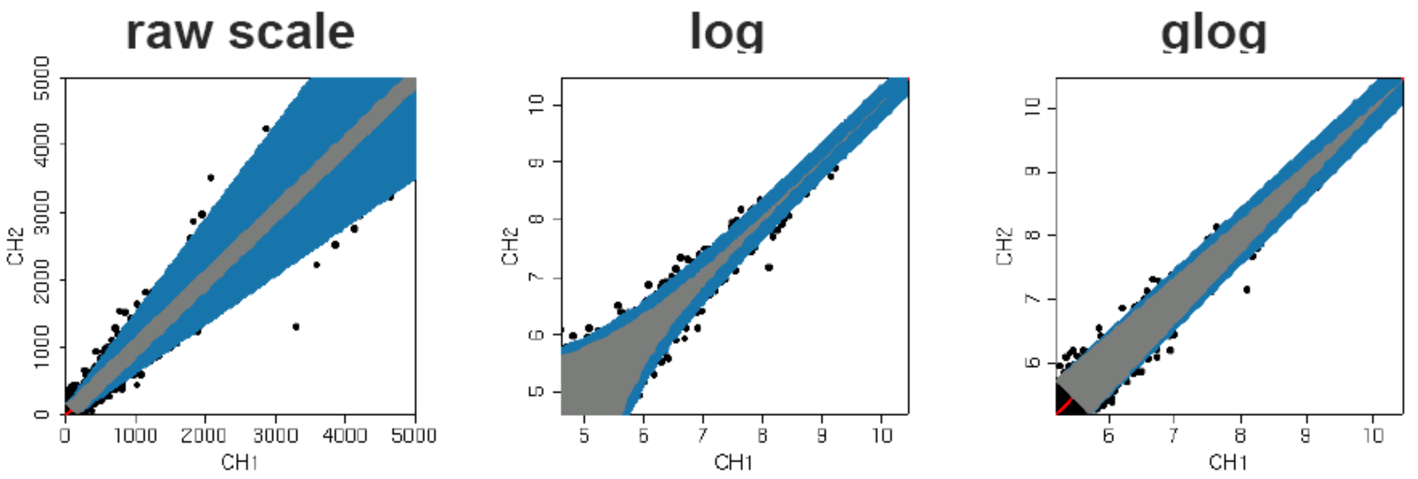
$$\lim_{x\to\infty}\left(\operatorname{arsinh} x - \log x - \log 2\right) = 0$$

P. Munson, 2001

D. Rocke & B. Durbin, ISMB 2002

W. Huber et al., ISMB 2002

Generalized logarithm



variance:

- ▨ **constant part**
- 🟦 **proportional part**

Huber

Exploratory data analysis

Fold change

ANOVA

Median polish

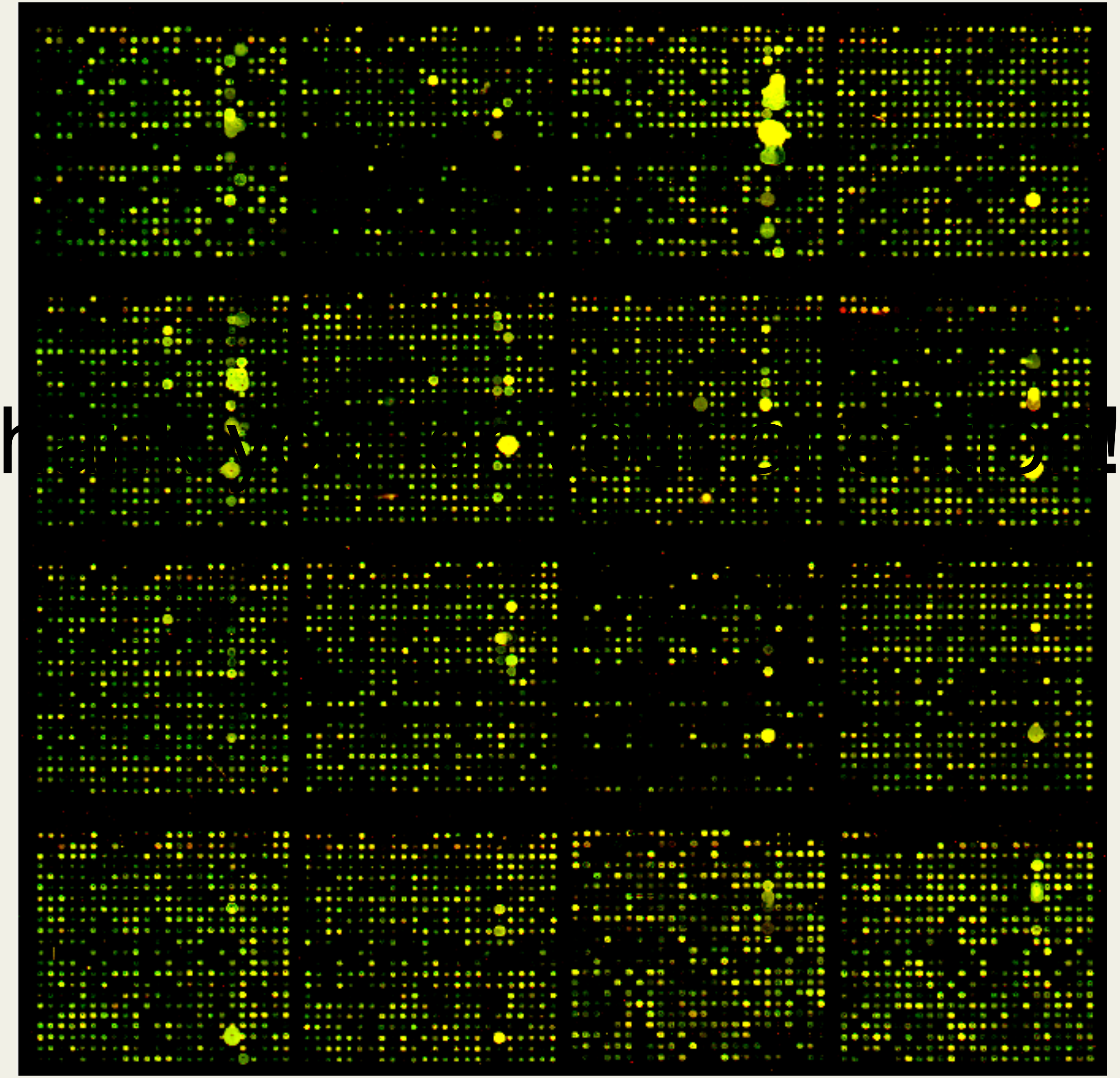
Validation

Sensitivity, Specificity

ROC curves

Receiver operating characteristic

- A framework to compare the performance of binary classifiers
- Plot of *false positive rate (sensitivity) vs true positive rate* (1-specificity)
- TPR = TP/P
- FPR = FP/N



Thank you for your attention!