Microarray Preprocessing

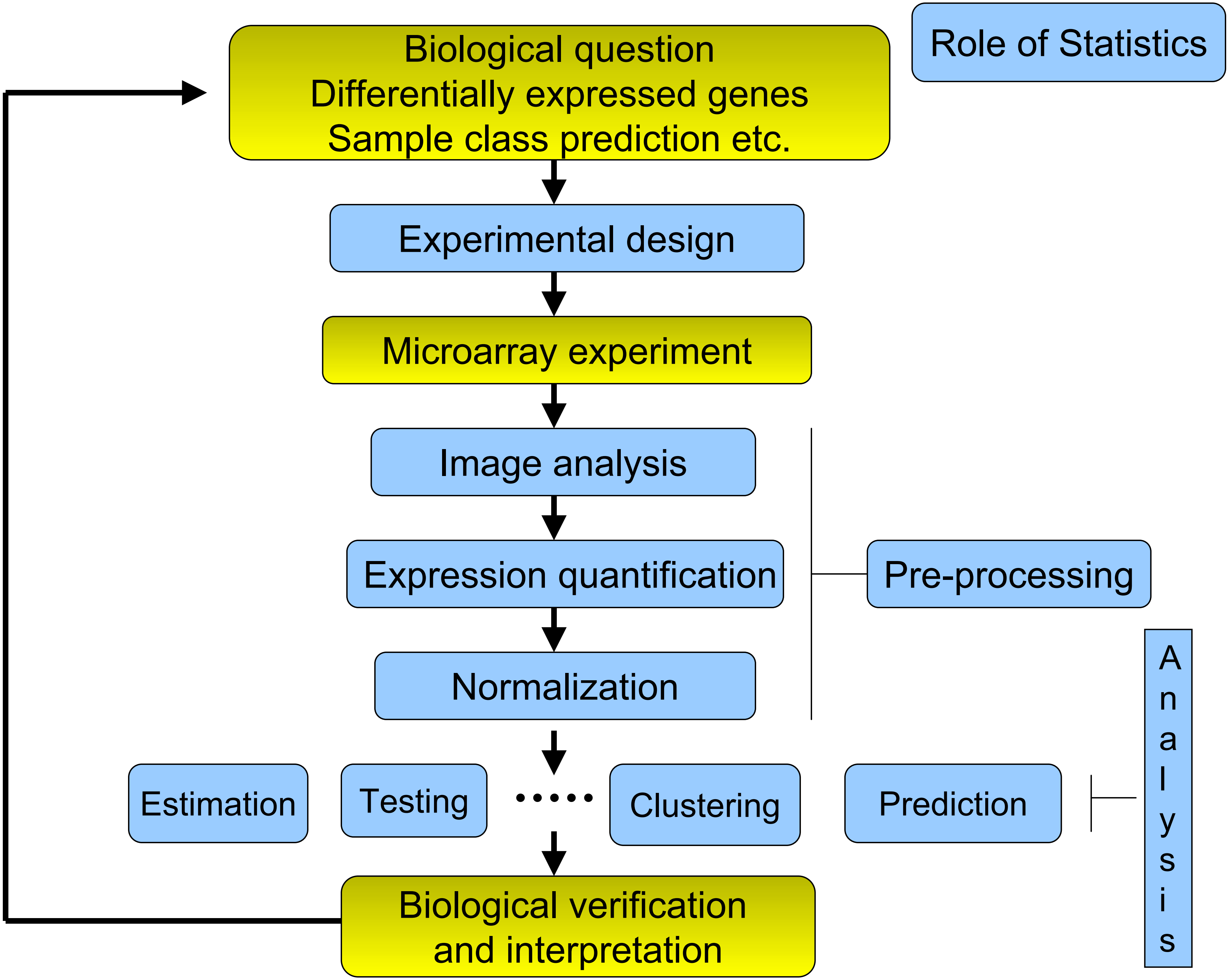
Benno Pütz

Acknowledgment

These slides are based –and have heavily borrowed from– the material available at the BioConductor web site at

www.bioconductor.org

The techniques presented in the following are mainly based on the works of Sandrine Dudoit, Yee Hwa Yang, Anja v. Heydebreck, and Wolfgang Huber



Role of Statistics

Biological question
Differentially expressed genes
Sample class prediction etc.

Experimental design

Microarray experiment

Image analysis

Expression quantification

Normalization

Pre-processing

Estimation

Testing

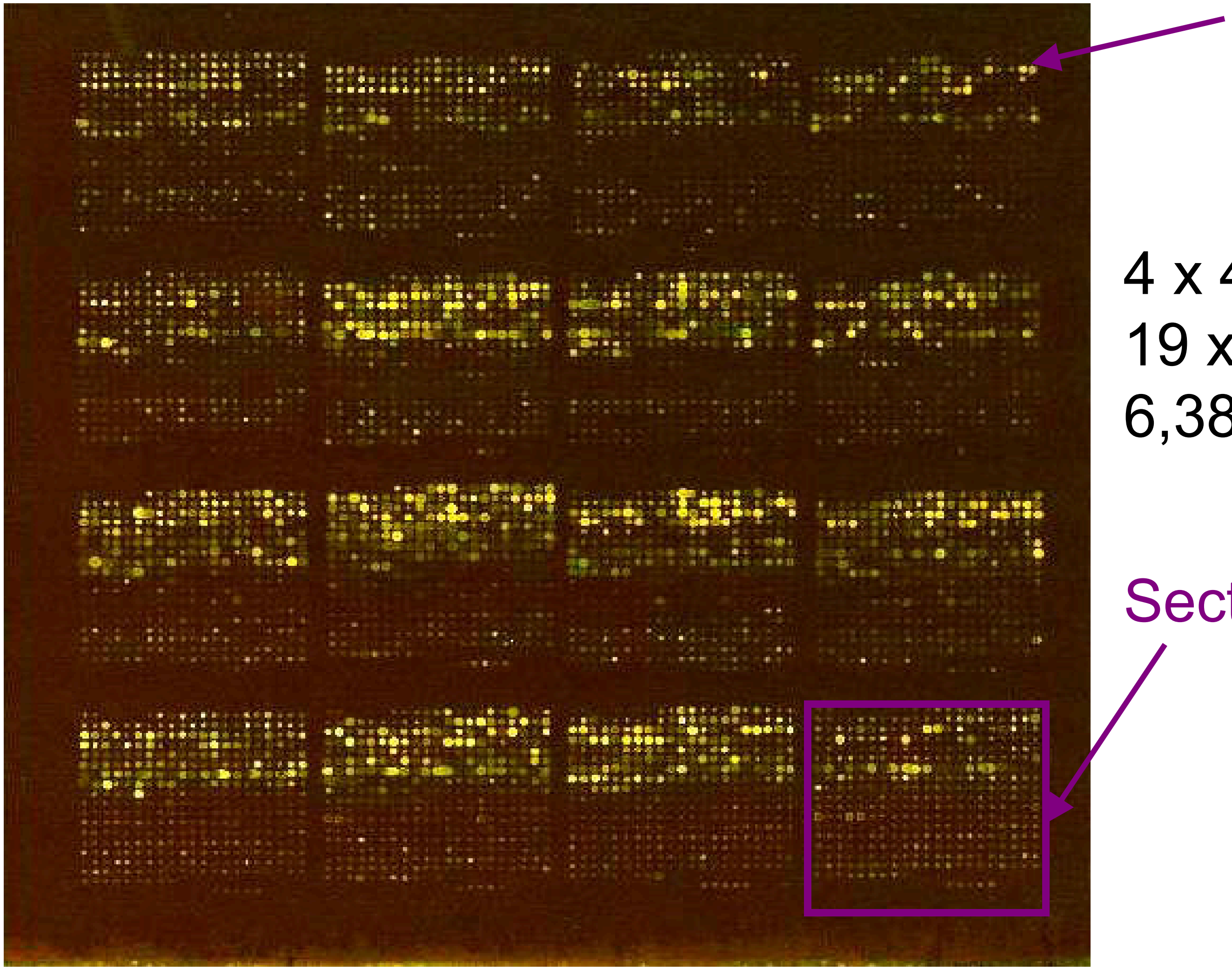
Clustering

Prediction

Analysis

Biological verification
and interpretation

RGB overlay of Cy3 and Cy5 images



Probe

Sector

4 x 4 sectors
19 x 21 probes/sector
6,384 probes/array

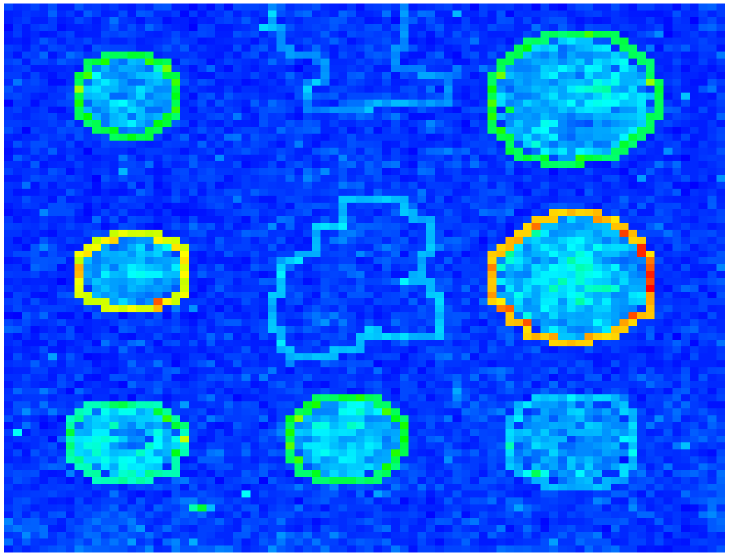
Terminology

- Target: DNA hybridized to the array, mobile substrate.
- Probe: DNA spotted on the array,

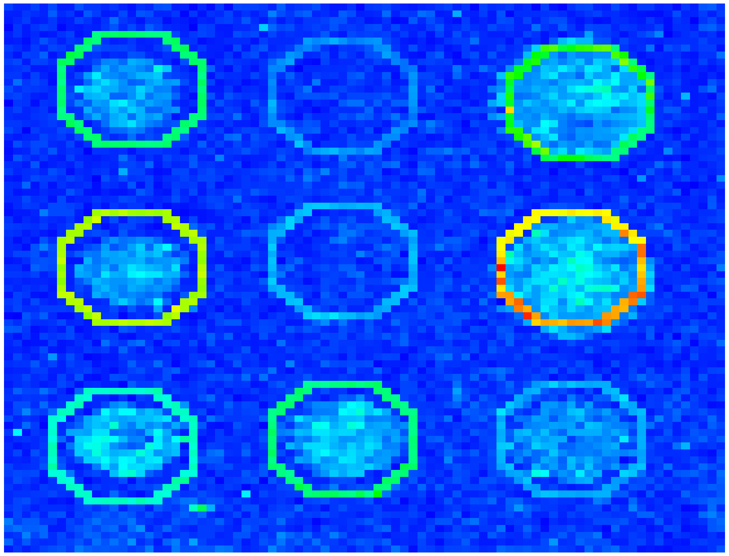
  aka. spot, immobile substrate.
- Sector: collection of spots printed using the same print-tip (or pin),

  aka. print-tip-group, pin-group, spot matrix, grid.
- The terms slide and array are often used to refer to the printed microarray.
- Batch: collection of microarrays with the same probe layout.
- Cy3 = Cyanine 3 = green dye.
- Cy5 = Cyanine 5 = red dye.

Segmentation



Adaptive segmentation, SRG



Fixed circle segmentation

Spots usually vary in size and shape.

Seeded region growing

- **Adaptive** segmentation method.
- Requires the input of **seeds**, either individual pixels or groups of pixels, which control the formation of the regions into which the image will be segmented.

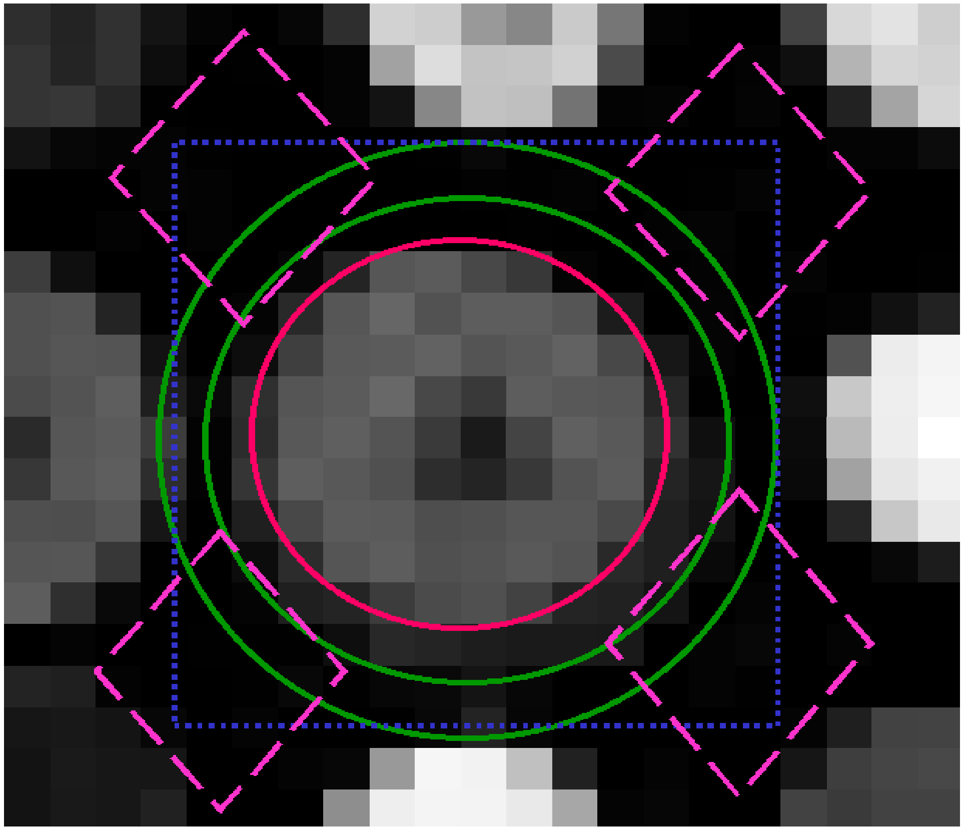
  Here, based on fitted foreground and background **grids** from the addressing step.
- The decision to add a pixel to a region is based on the absolute gray-level difference of that pixel's intensity and the average of the pixel values in the neighboring region.
- Done on combined red and green images.
- Ref. Adams & Bischof (1994)

Local background

---- GenePix

---- QuantArray

---- ScanAnalyze



What is (local) background?

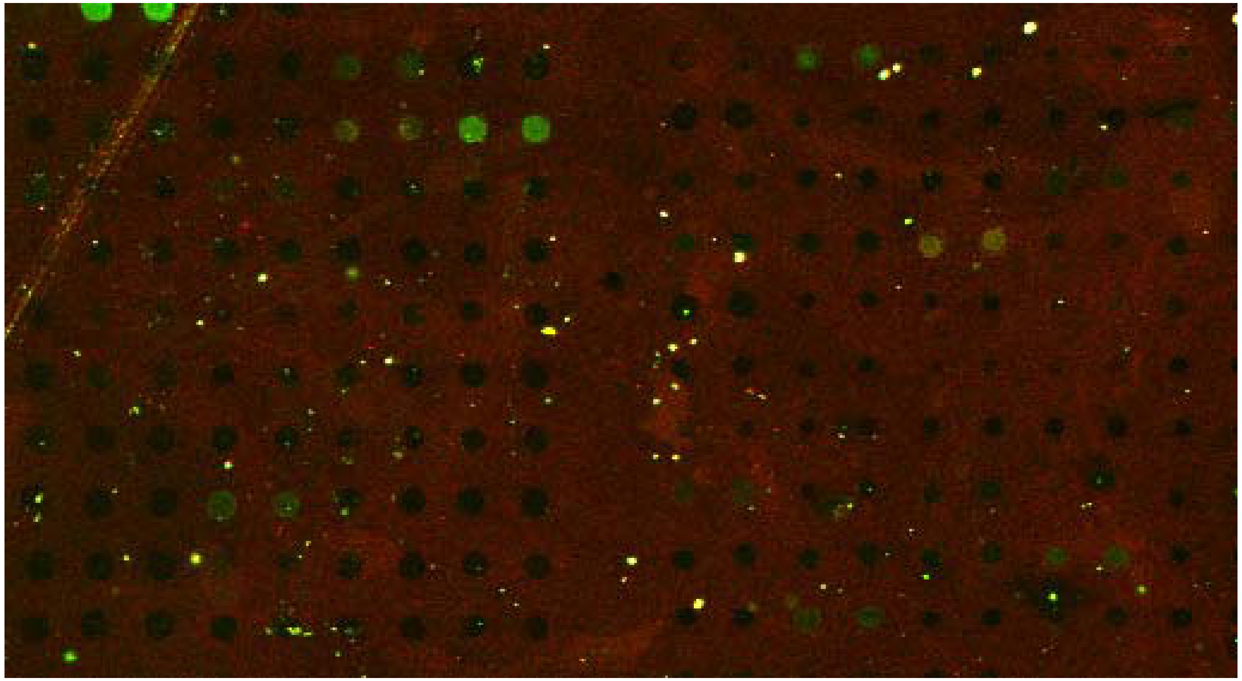
usual assumption:

total brightness =
background brightness (adjacent to spot)
+ brightness from labeled sample cDNA

What is (local) background?

usual assumption:

total brightness =
    background brightness (adjacent to spot)
  + brightness from labeled sample cDNA



?

Quality measures

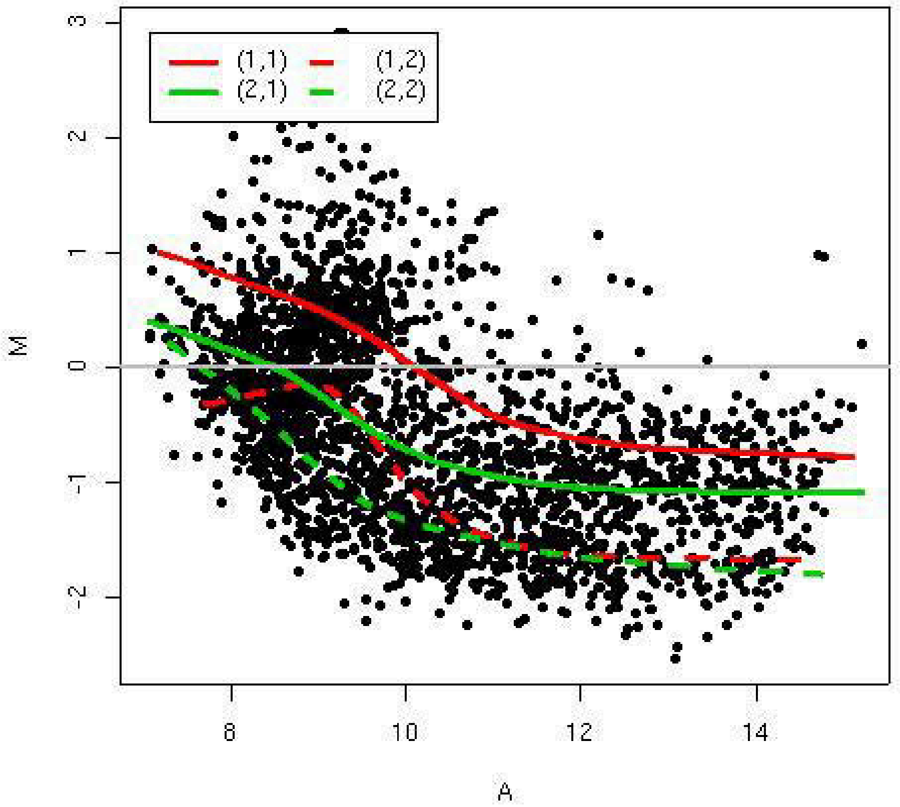
- **Spot quality**
  - **Brightness:** foreground/background ratio;
  - **Uniformity:** variation in pixel intensities and ratios of intensities within a spot;
  - **Morphology:** area, perimeter, circularity.
- **Slide quality**
  - Percentage of spots with no signal;
  - Range of intensities;
  - Distribution of spot signal area, etc.
- How to use quality measures in subsequent analyses?

LOESS-based Normalization

Yang, Dudoit, et al.

Nucl. Acids Res. 30(4):e15, 2002

Normalization



Normalization

- **Purpose.** Identify and remove the effects **of systematic variation** in the measured fluorescence intensities, other than differential expression, for example
  - different labeling efficiencies of the dyes;
  - different amounts of Cy3- and Cy5-labeled mRNA;
  - different scanning parameters;
  - print-tip, spatial, or plate effects, etc.

Normalization

- Normalization is needed to ensure that differences in intensities are indeed due to differential expression, and not some printing, hybridization, or scanning artifact.

- Normalization is necessary before any analysis which involves within or between slides comparisons of intensities, e.g., clustering, testing.

Normalization

- The need for normalization can be seen most clearly in **self-self hybridizations**, where the same mRNA sample is labeled with the Cy3 and Cy5 dyes.
- The imbalance in the red and green intensities is usually **not constant** across the spots within and between arrays, and can vary according to overall spot intensity, location, plate origin, etc.
- These factors should be considered in the normalization.

Single-slide data display

- Usually: R vs. G

  $\log_2 R$ vs. $\log_2 G$.

- Preferred

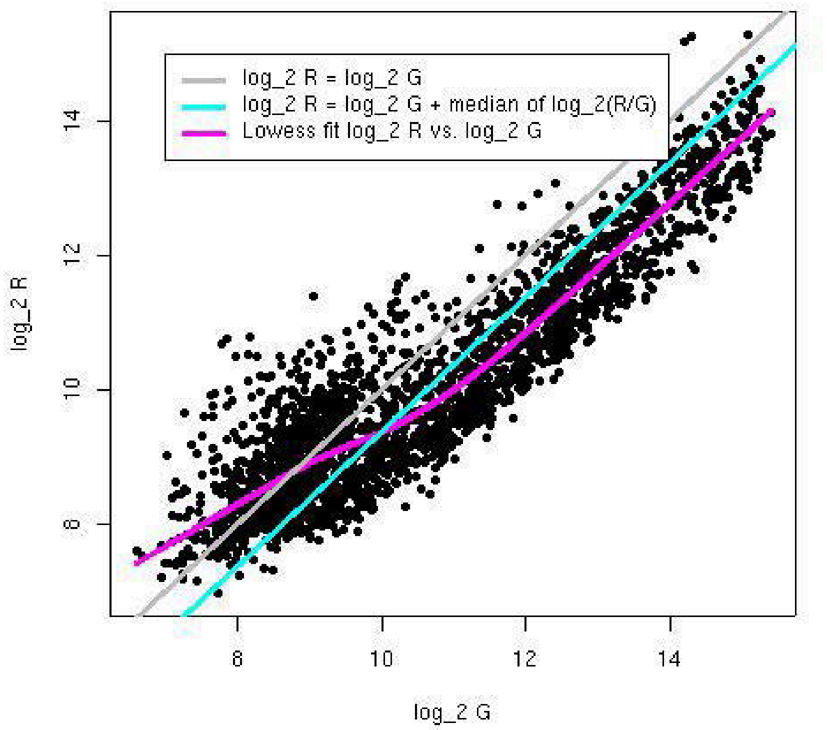
  $$M = \log_2 R - \log_2 G \qquad \text{(ratio)}$$

  vs. $$A = (\log_2 R + \log_2 G)/2. \qquad \text{(geom. mean)}$$

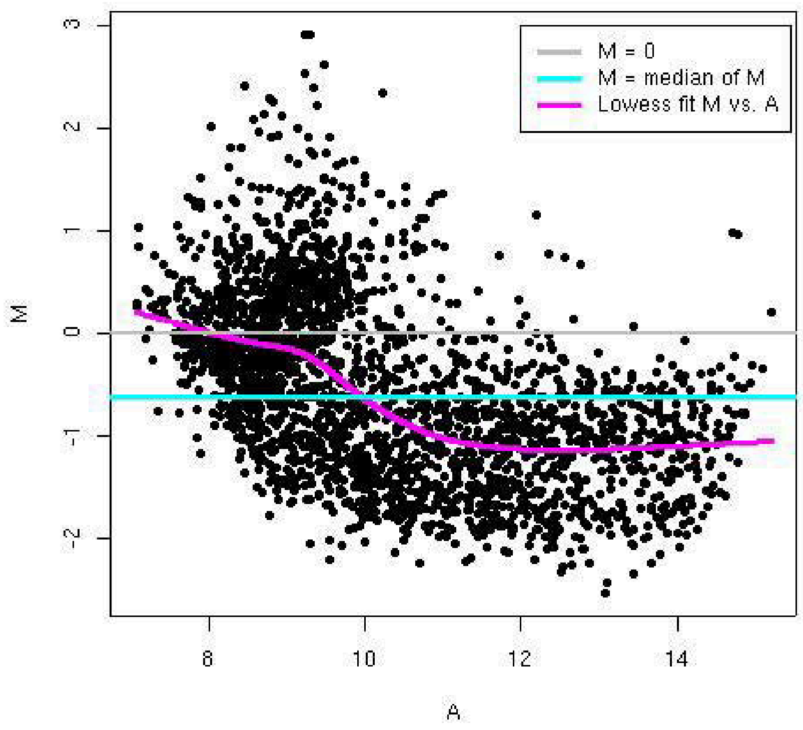
- An MA-plot amounts to a 45^o counterclockwise rotation of a $\log_2 R$ vs. $\log_2 G$ plot followed by scaling.

Self-self hybridization

$\log_2$ R vs. $\log_2$ G



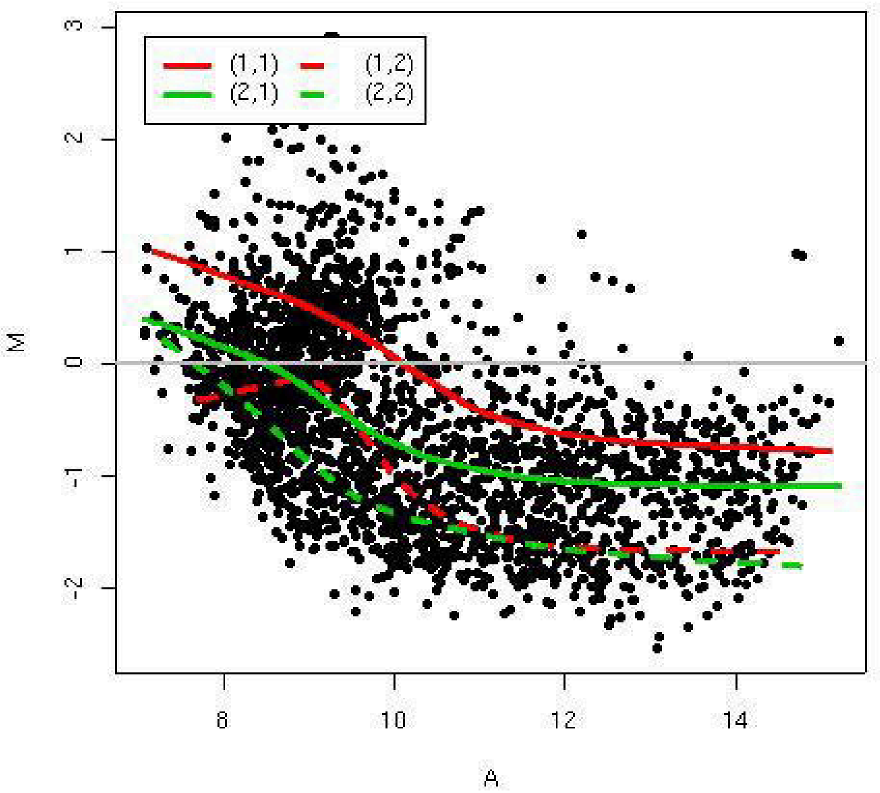
M vs. A



$M = \log_2 R - \log_2 G, \quad A = (\log_2 R + \log_2 G)/2$

Self-self hybridization

M vs. A



Robust local regression within sectors (print-tip-groups) of intensity log-ratio M on average log-intensity A.

$M = \log_2 R - \log_2 G$, $A = (\log_2 R + \log_2 G)/2$

Swirl zebrafish experiment

- **Goal**. Identify genes with altered expression in Swirl mutants compared to wild-type zebrafish.
- 2 sets of dye-swap experiments (n=4).
- Arrays:
  - 8,448 probes (768 controls);
  - 4 x 4 grid matrix;
  - 22 x 24 spot matrices.
- Data available in Bioconductor package **`marrayInput`**.

Diagnostic plots

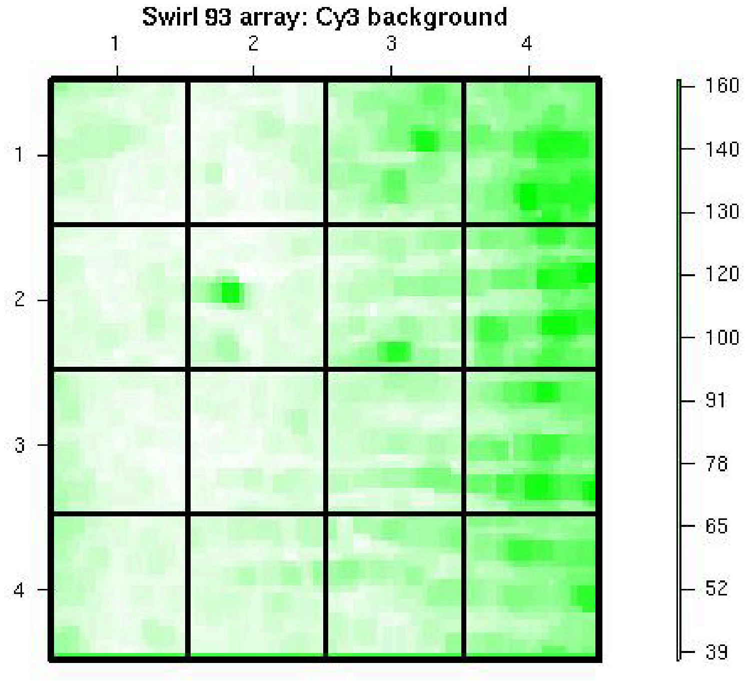
- **Diagnostics plots** of spot statistics

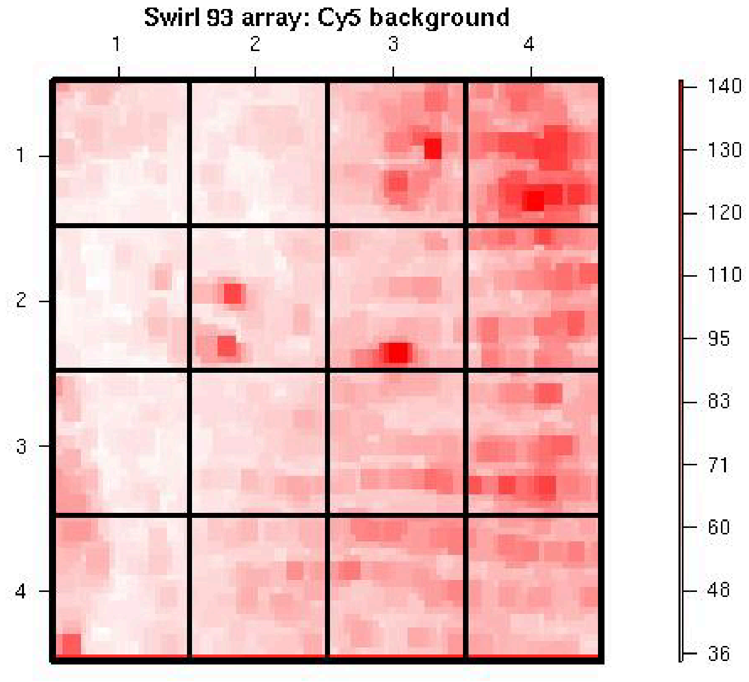
  E.g. red and green log-intensities, intensity log-ratios M, average log-intensities A, spot area.

  - Boxplots;
  - 2D spatial images;
  - Scatter-plots, e.g. MA-plots;
  - Density plots.

- **Stratify** plots according to layout parameters, e.g. print-tip-group, plate.

2D spatial images



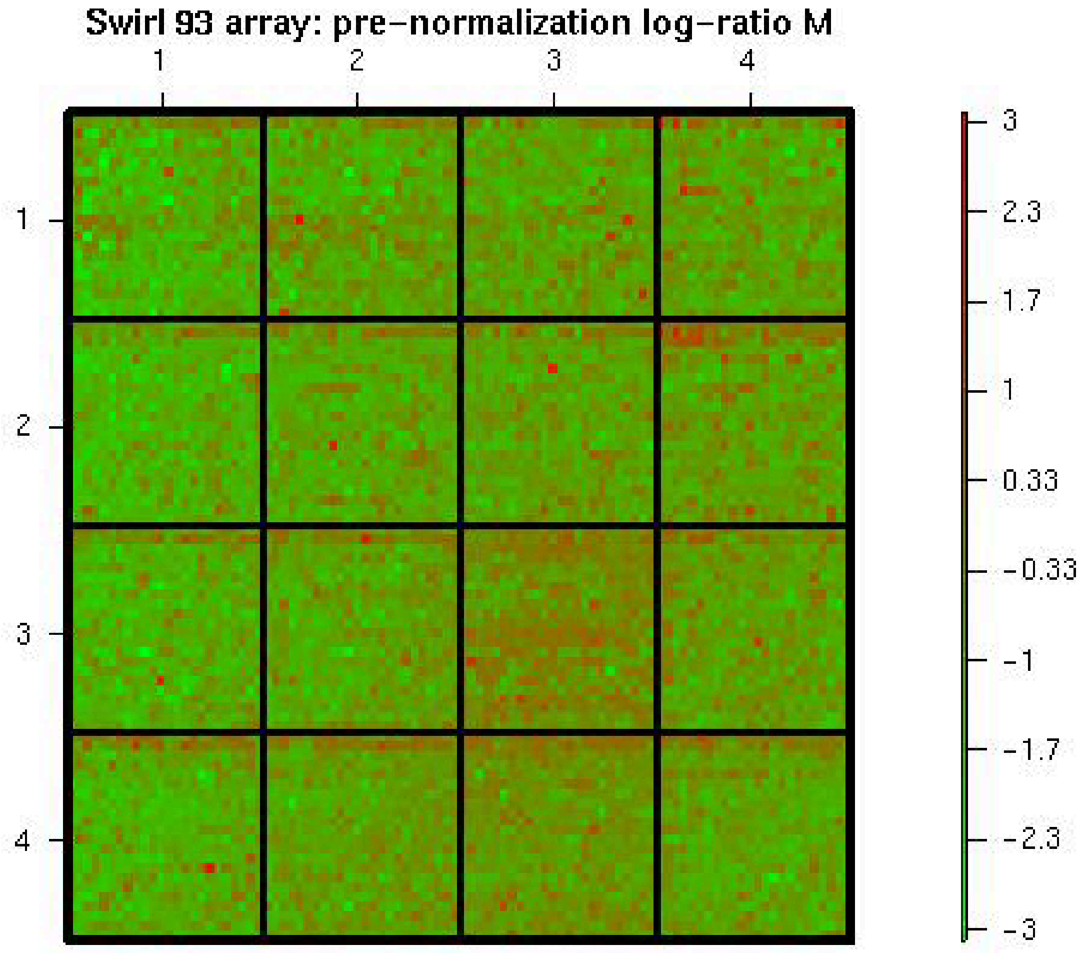


Cy3 background intensity

Cy5 background intensity

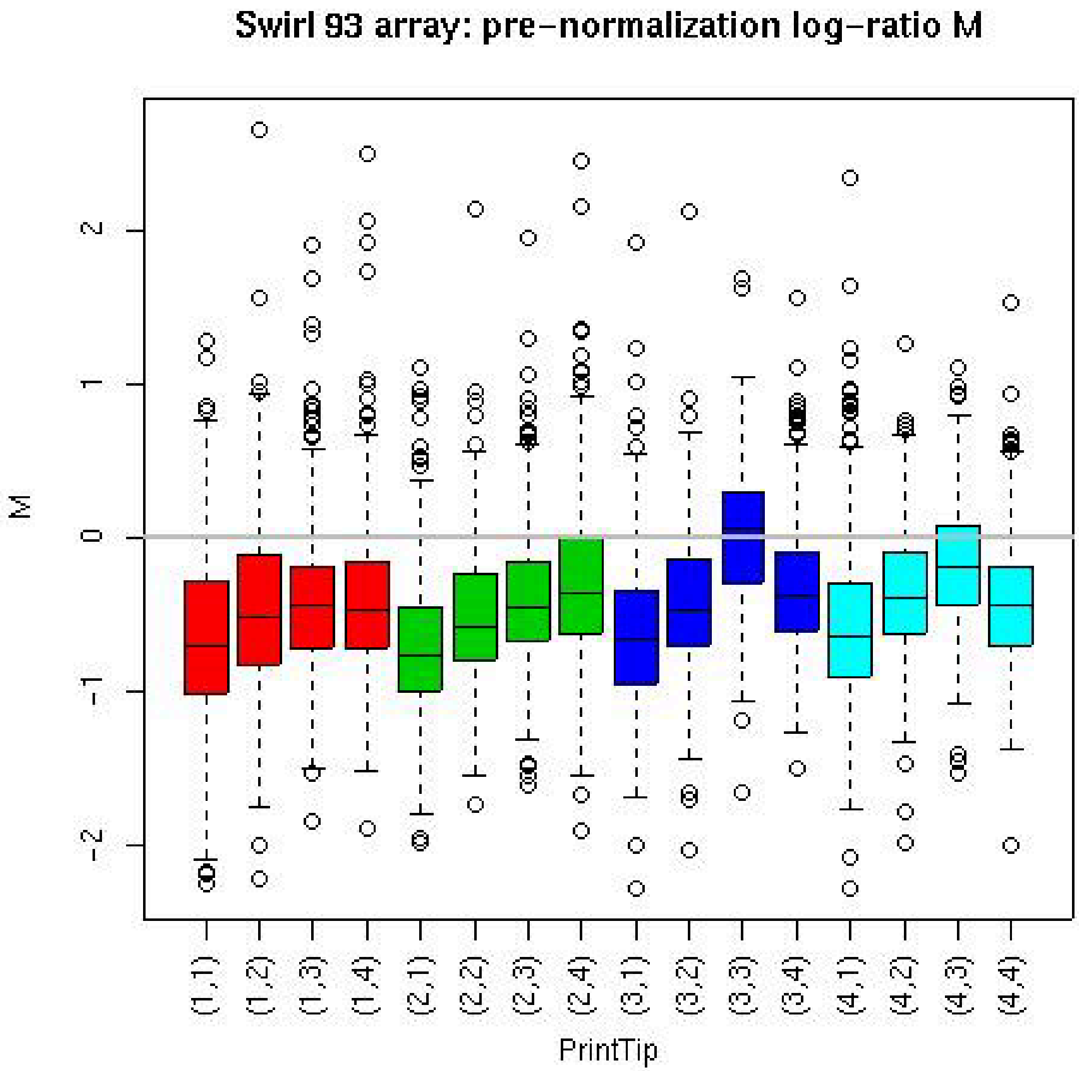
2D spatial images

Intensity
log-ratio, M



Boxplots by print-tip-group

Intensity
log-ratio, M

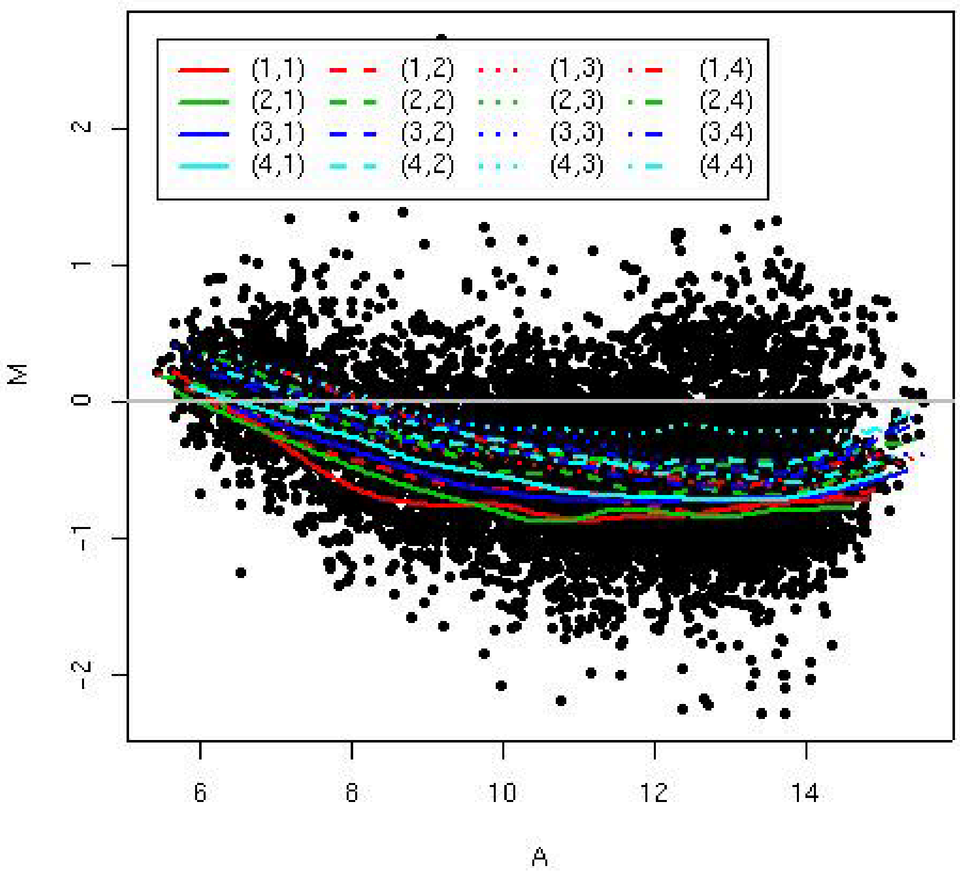


MA-plot by print-tip-group

$M = \log_2 R - \log_2 G, \quad A = (\log_2 R + \log_2 G)/2$

Swirl 93 array: pre-normalization log-ratio M

Intensity log-ratio, M



Average log-intensity, A

Location normalization

$$\log_2 R/G \leftarrow \log_2 R/G - L(\text{intensity, sector, \ldots})$$

- **Constant normalization.** Normalization function L is **constant** across the spots, e.g. mean or median of the log-ratios M.

- **Adaptive normalization.** Normalization function L depends on a number of **predictor variables**, such as spot intensity A, sector, plate origin.

Location normalization

- The normalization function can be obtained by **robust locally weighted regression** of the log-ratios M on predictor variables.

  E.g. regression of M on A within sector.

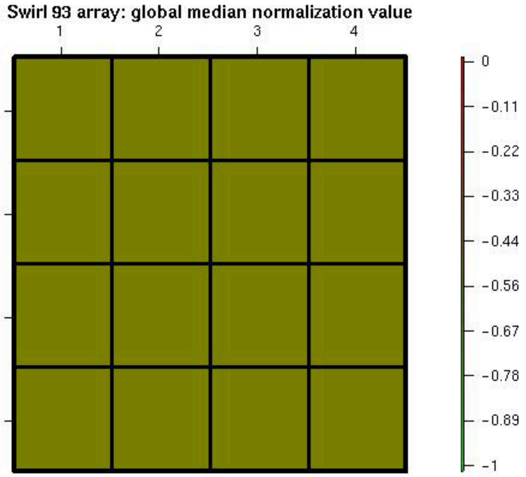
- Regression method: e.g. lowess or loess (Cleveland, 1979; Cleveland & Devlin, 1988).

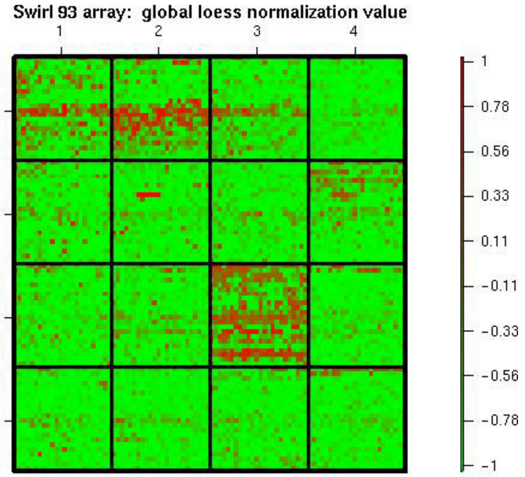
Location normalization

- **Intensity-dependent normalization.**
  Regression of M on A (*global loess*).
- **Intensity and sector-dependent normalization***.*
  Same as above, for each sector separately
  (*within-print-tip-group loess*).
- **2D spatial normalization**.
  Regression of M on 2D-coordinates.
- Other variables: time of printing, plate, etc.
- **Composite normalization**. Weighted average of several normalization functions.

2D images of L values

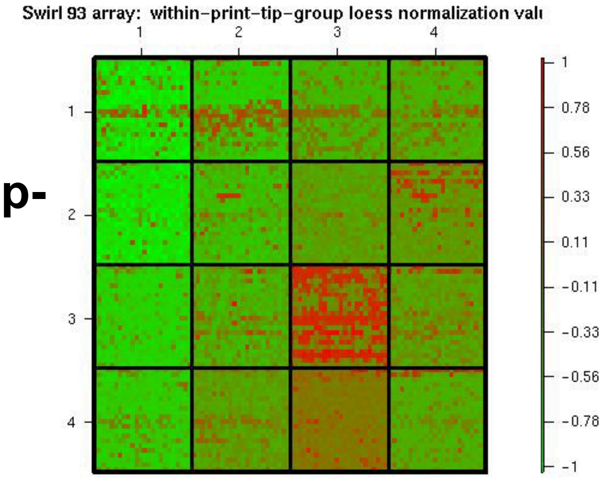
Global median normalization

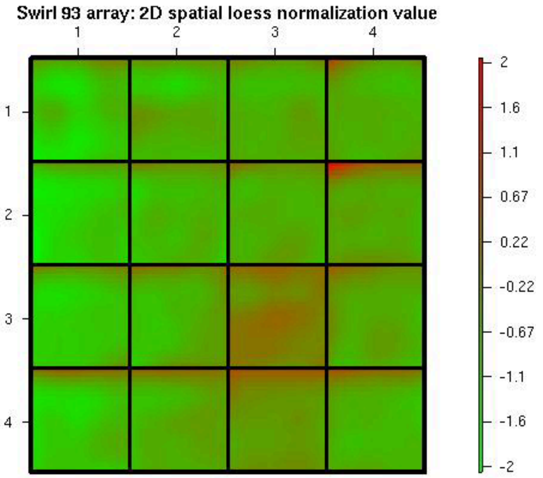




Global loess normalization

Within-print-tip-group loess normalization

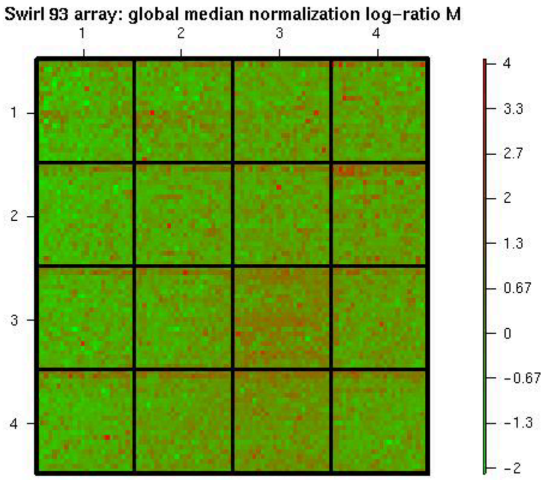


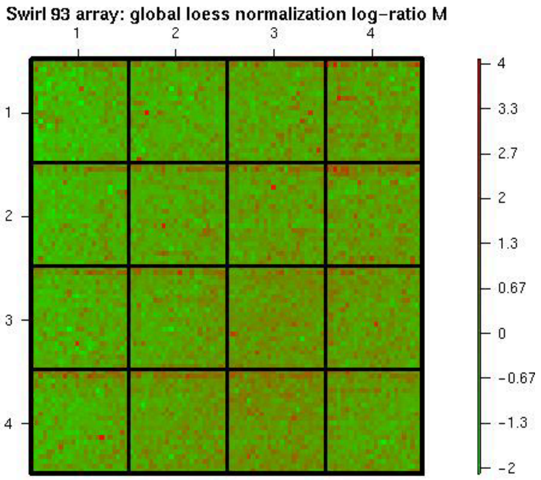


2D spatial normalization

2D images of normalized M-L

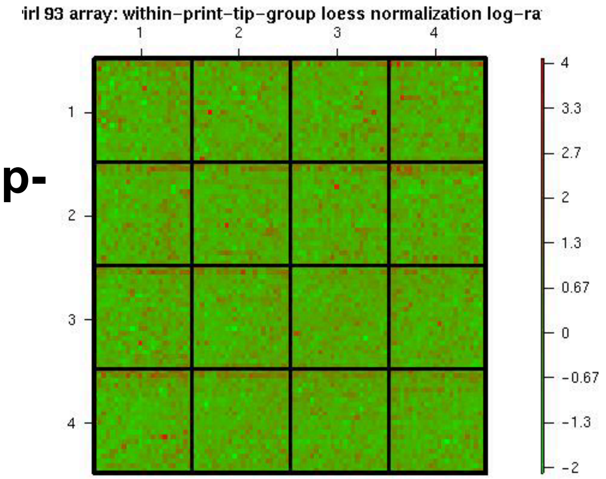
Global median normalization

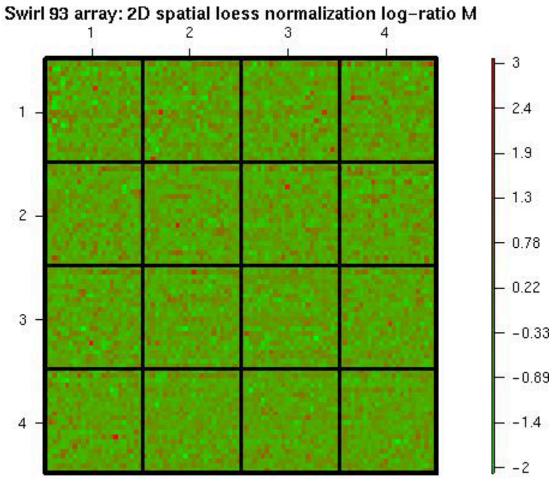




Global loess normalization

Within-print-tip-group loess normalization

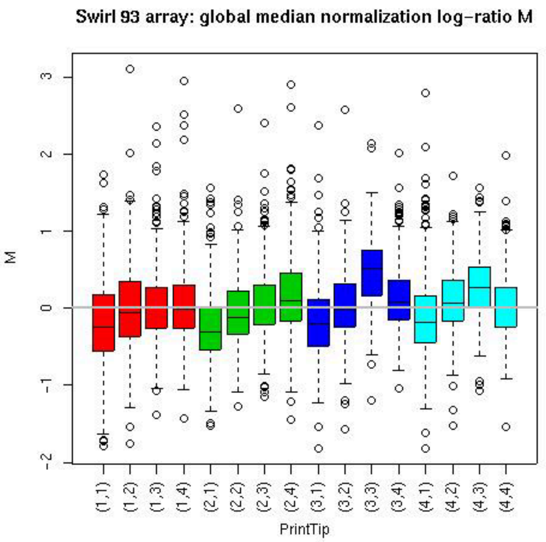




2D spatial normalization

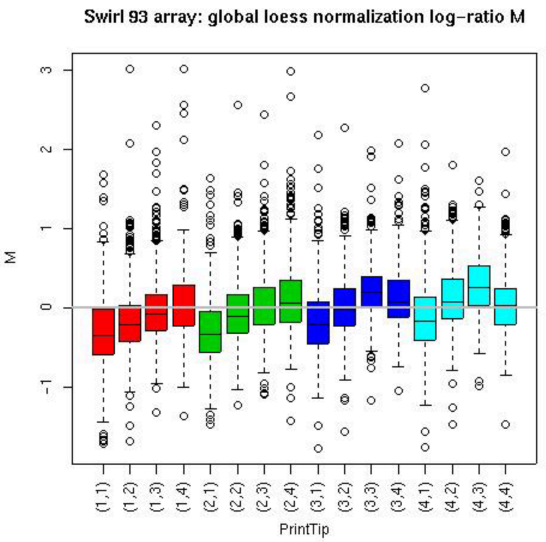
Boxplots of normalized M-L

Global median normalization



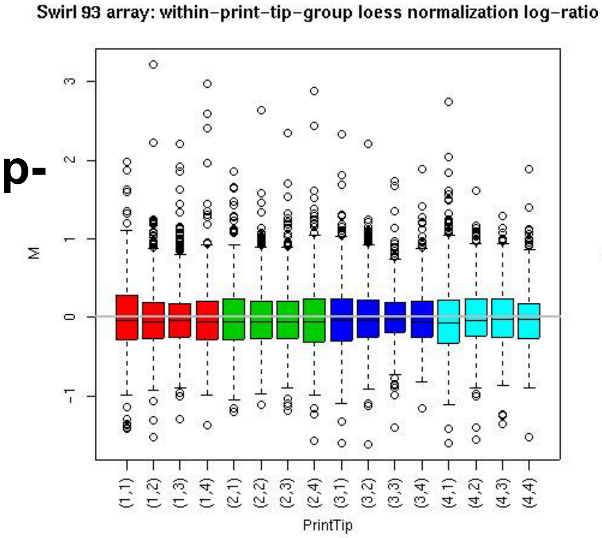
Swirl 93 array: global median normalization log–ratio M

Global loess normalization



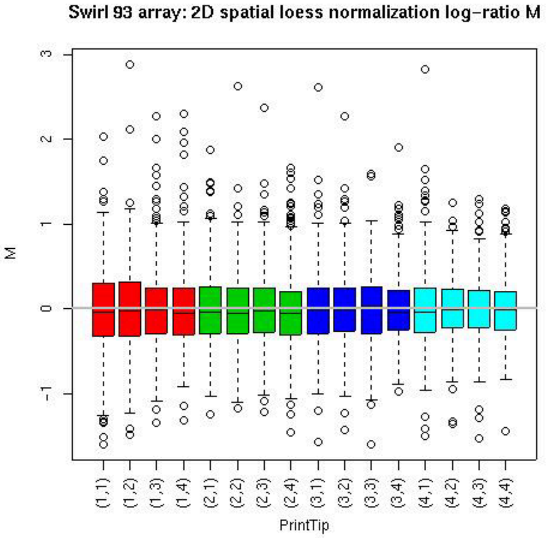
Swirl 93 array: global loess normalization log–ratio M

Within-print-tip-group loess normalization



Swirl 93 array: within–print–tip–group loess normalization log–ratio

2D spatial normalization

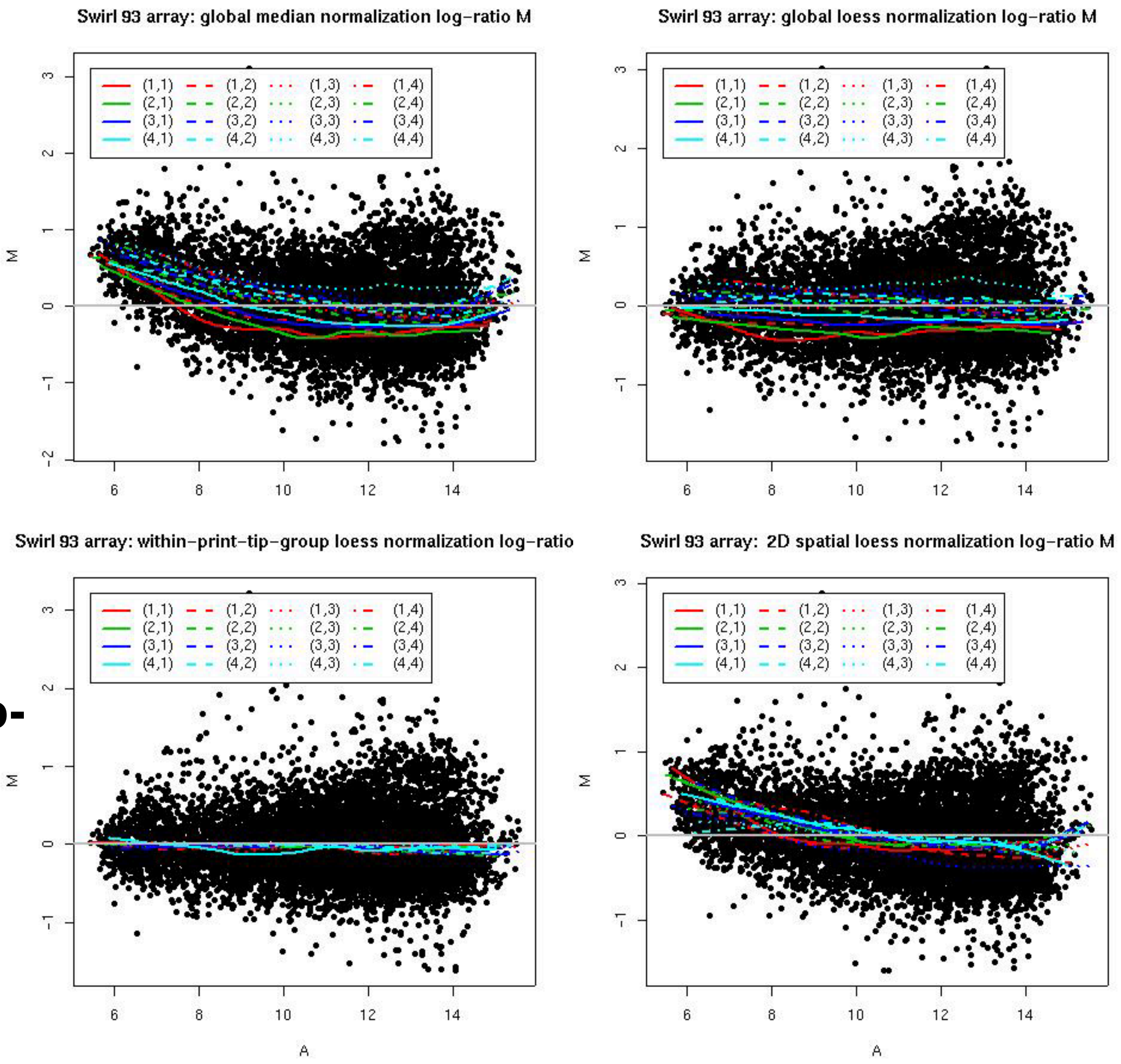


Swirl 93 array: 2D spatial loess normalization log–ratio M

MA-plots of normalized M-L

Global median normalization

Global loess normalization



Swirl 93 array: global median normalization log–ratio M

Swirl 93 array: global loess normalization log–ratio M

Swirl 93 array: within–print–tip–group loess normalization log–ratio

Swirl 93 array: 2D spatial loess normalization log–ratio M

Within-print-tip-group loess normalization

2D spatial normalization

Normalization

- Within-slide
  - **Location** normalization - additive on log-scale.
  - **Scale** normalization - multiplicative on log-scale.
  - **Which spots** to use?
- Paired-slides (dye-swap experiments)
  - Self-normalization.
- Between-slides.

Scale normalization

- The log-ratios M from different sectors, plates, or arrays may exhibit different spreads and some **scale** adjustment may be necessary.

$$\log_2 R/G \leftarrow (\log_2 R/G - L)/S$$

- Can use a robust estimate of scale such as the **median absolute deviation (MAD)**

$\text{MAD} = \text{median} \mid M - \text{median}(M) \mid.$

Scale normalization

- For print-tip-group scale normalization, assume all print-tip-groups have the same spread in M.
- Denote ***true*** and ***observed*** log-ratio by μ_{ij} and M_{ij}, resp., where $M_{ij} = a_i \mu_{ij}$, and i indexes print-tip-groups and j spots. Robust estimate of a_i is

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

where MAD_i is MAD of M_{ij} in print-tip-group i.

- Similarly for between-slides scale normalization.

Which genes to use?

- **All spots on the array**:
  - Problem when many genes are differentially expressed.

- **Housekeeping genes**: Genes that are thought to be constantly expressed across a wide range of biological samples (e.g. tubulin, GAPDH). Problems:
  - sample specific biases (genes are actually regulated),
  - do not cover intensity range.

Which genes to use?

- **Genomic DNA titration series**:
  - fine in yeast,
  - but weak signal for higher organisms with high intron/exon ratio (e.g. mouse, human).

- **Rank invariant set** (Schadt et al., 1999; Tseng et al., 2001): genes with same rank in both channels. Problems: set can be small.

Microarray sample pool

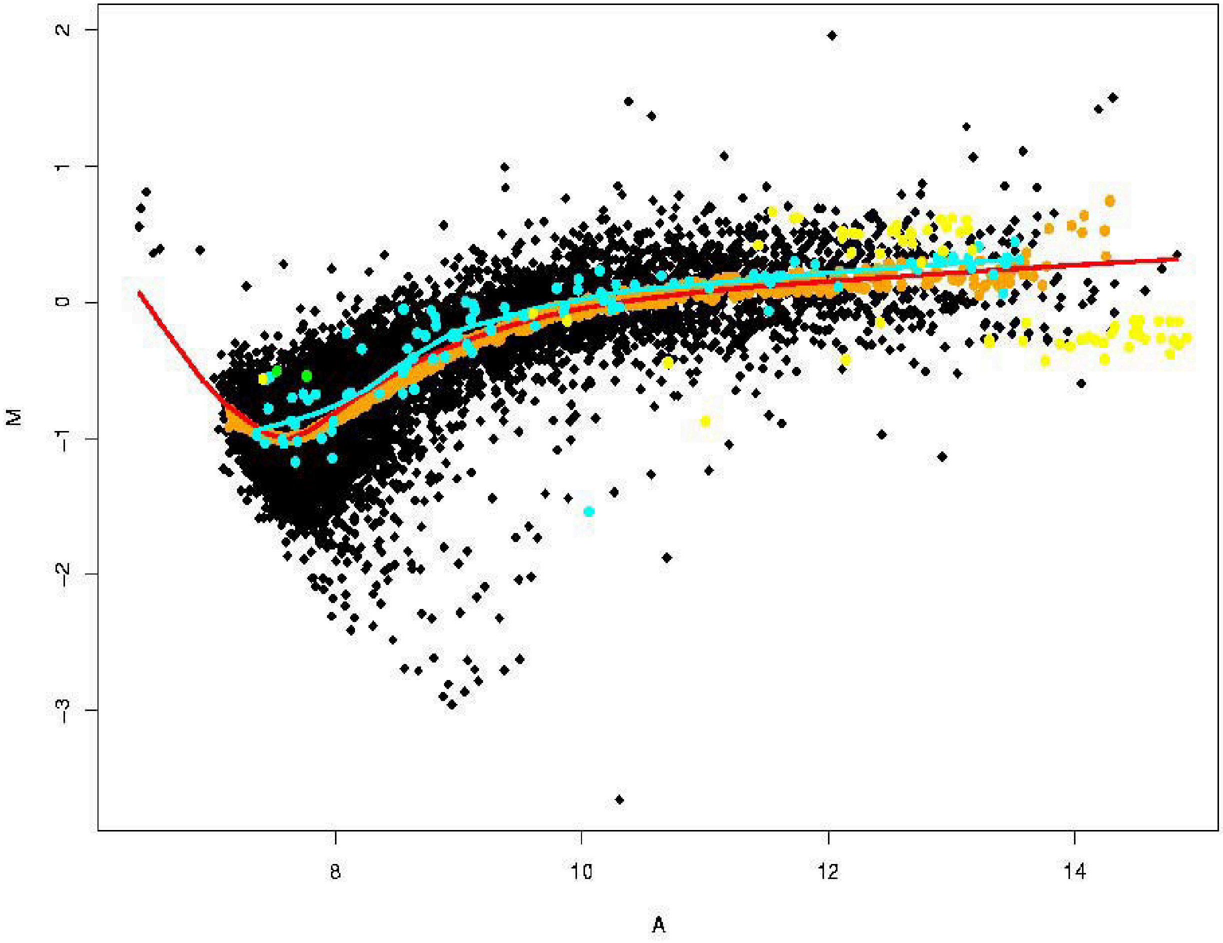
- **Microarray Sample Pool**, **MSP**: Control sample for normalization, in particular, when it is not safe to assume most genes are equally expressed in both channels.
- MSP: **pooled** all 18,816 ESTs from RIKEN release 1 cDNA mouse library.
- Six-step **dilution series** of the MSP.
- MSP samples were spotted in middle of first and last row of each sector.
- Ref. Yang et al. (2002).

Microarray sample pool

MSP control spots

- provide potential probes for every target sequence;
- are constantly expressed across a wide range of biological samples;
- cover the intensity range;
- are similar to genomic DNA, but without intron sequences → better signal than genomic DNA in organisms with high intron/exon ratio;
- can be used in composite normalization.

Microarray sample pool



MSP
Rank invariant
Housekeeping
Tubulin, GAPDH

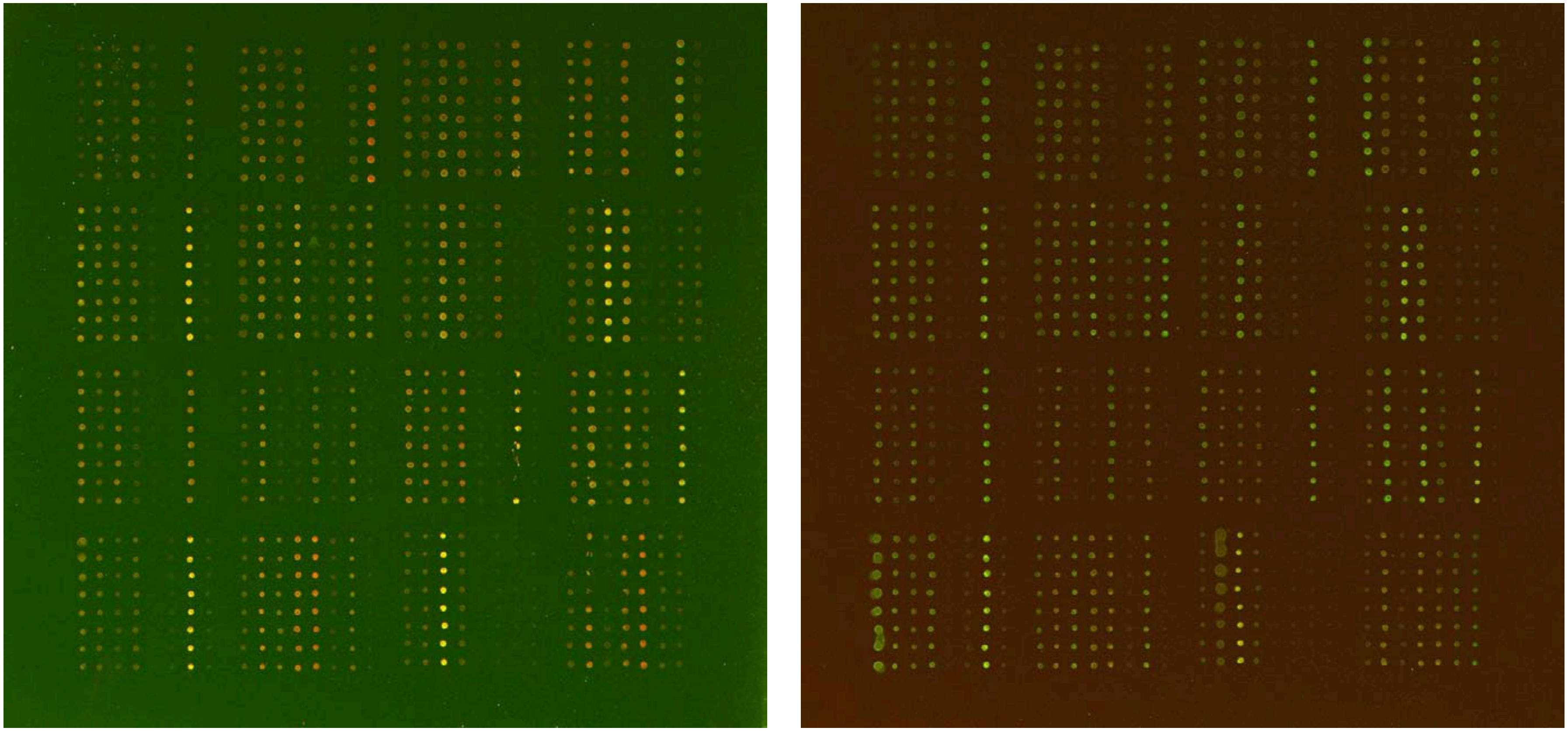
Dye-swap experiment

- Probes
  - 50 distinct clones thought to be differentially expressed in apo AI knock-out mice compared to inbred C57Bl/6 control mice (largest absolute t-statistics in a previous experiment).
  - 72 other clones.
- Spot each clone 8 times .
- Two hybridizations with dye-swap:

  Slide 1: trt → red, ctl → green.

  Slide 2: trt → green, ctl → red.

Dye-swap experiment



Self-normalization

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

Combine by **subtracting** the normalized log-ratios:

$$M - M'$$

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

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

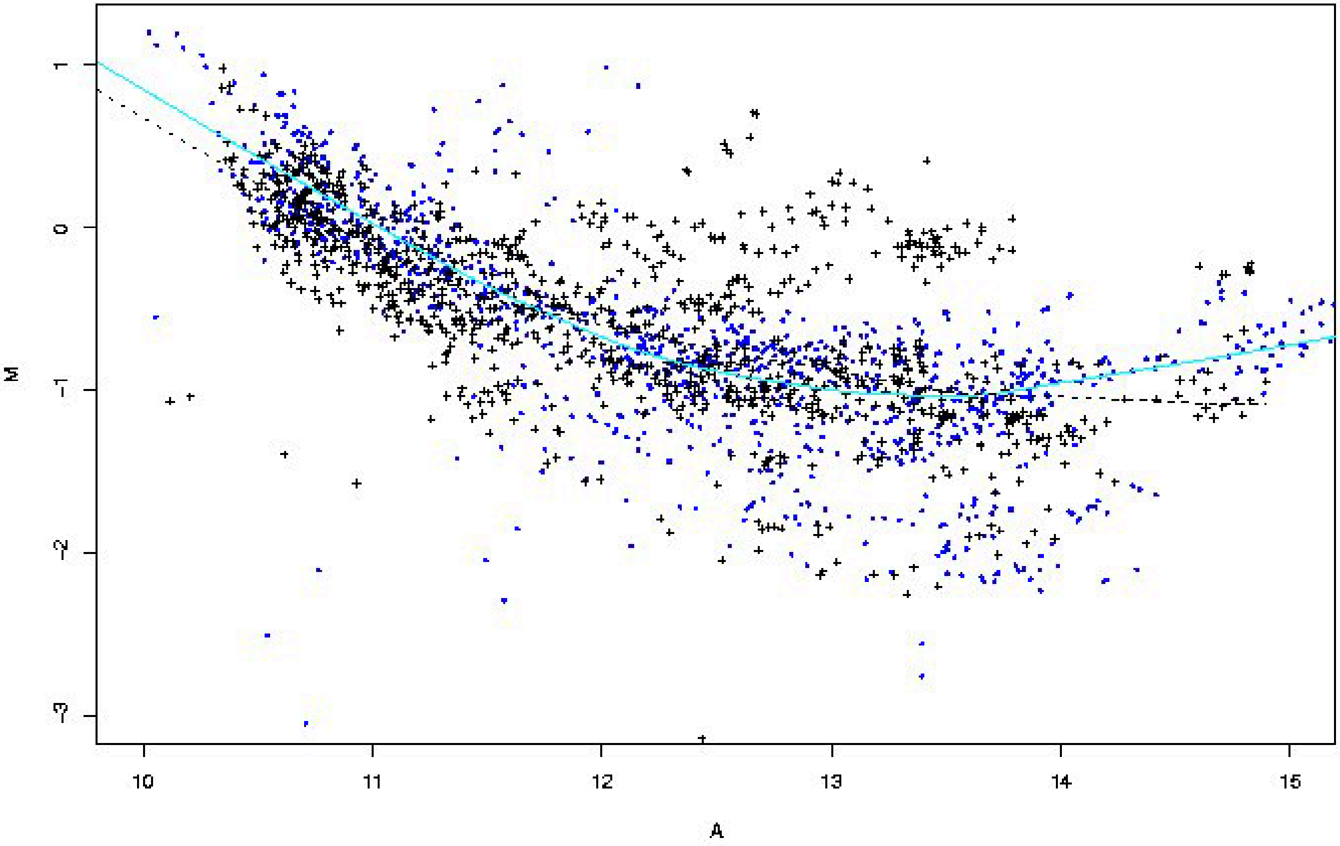
$$\approx [\log_2 (RG'/GR')] / 2$$

provided $L = L'$.

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

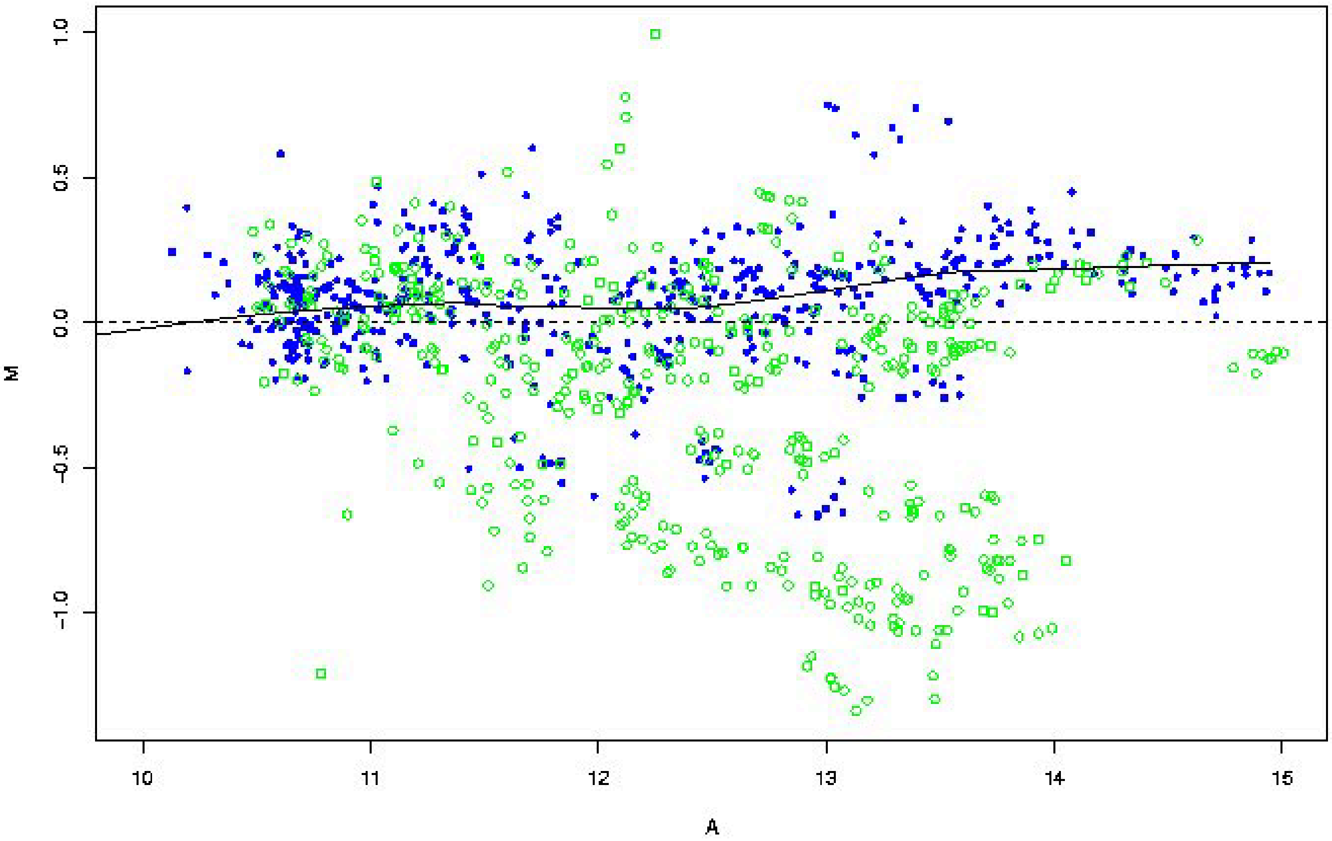
Checking the assumption

MA-plot for slides 1 and 2



Result of self-normalization

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



Summary

Case 1. Only a few genes are expected to change.

Within-slide

- Location: intensity + sector-dependent normalization.
- Scale: for each sector, scale by MAD.

Between-slides

- An extension of within-slide scale normalization.

Case 2. Many genes are expected to change.

- Paired-slides: Self-normalization.
- Use of controls or known information, e.g. MSP.
- Composite normalization.

Pre-processing cDNA microarray data

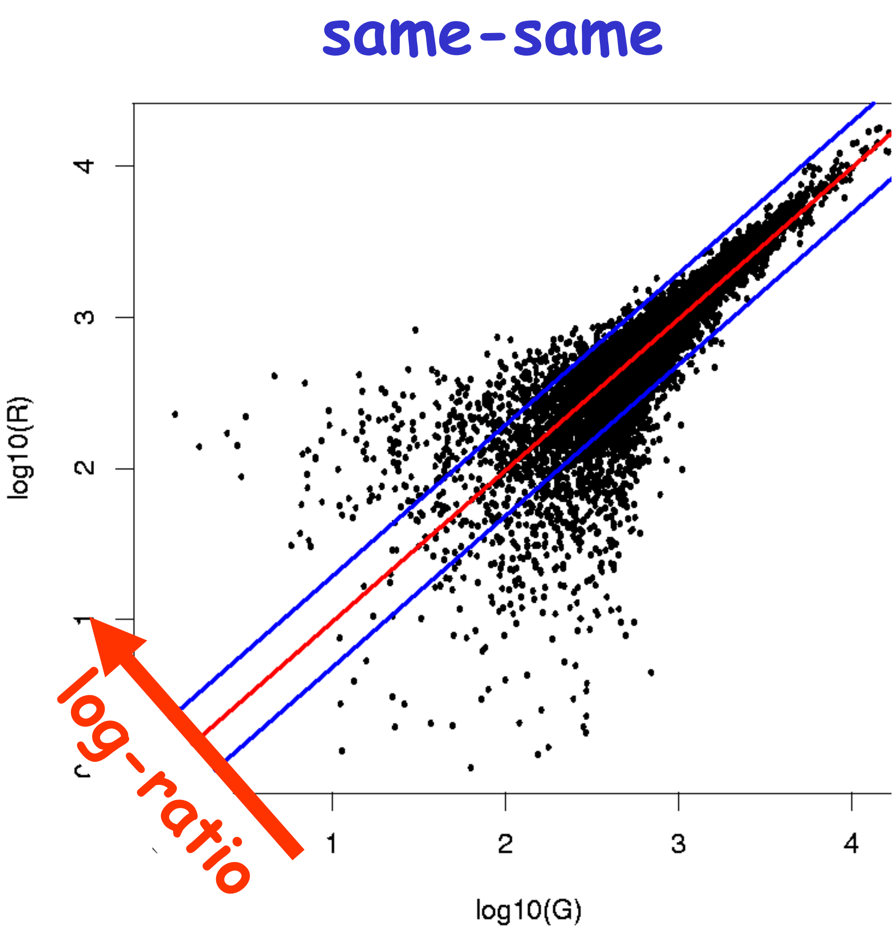
- **`marrayClasses`**:
  - class definitions for cDNA microarray data;
  - basic methods for manipulating microarray objects: printing, plotting, subsetting, class conversions, etc.
- **`marrayInput`**:
  - reading in intensity data and textual data describing probes and targets;
  - automatic generation of microarray data objects;
  - widgets for point & click interface.
- **`marrayPlots`**: diagnostic plots.
- **`marrayNorm`**: robust adaptive location and scale normalization procedures.

Variance Stabilization

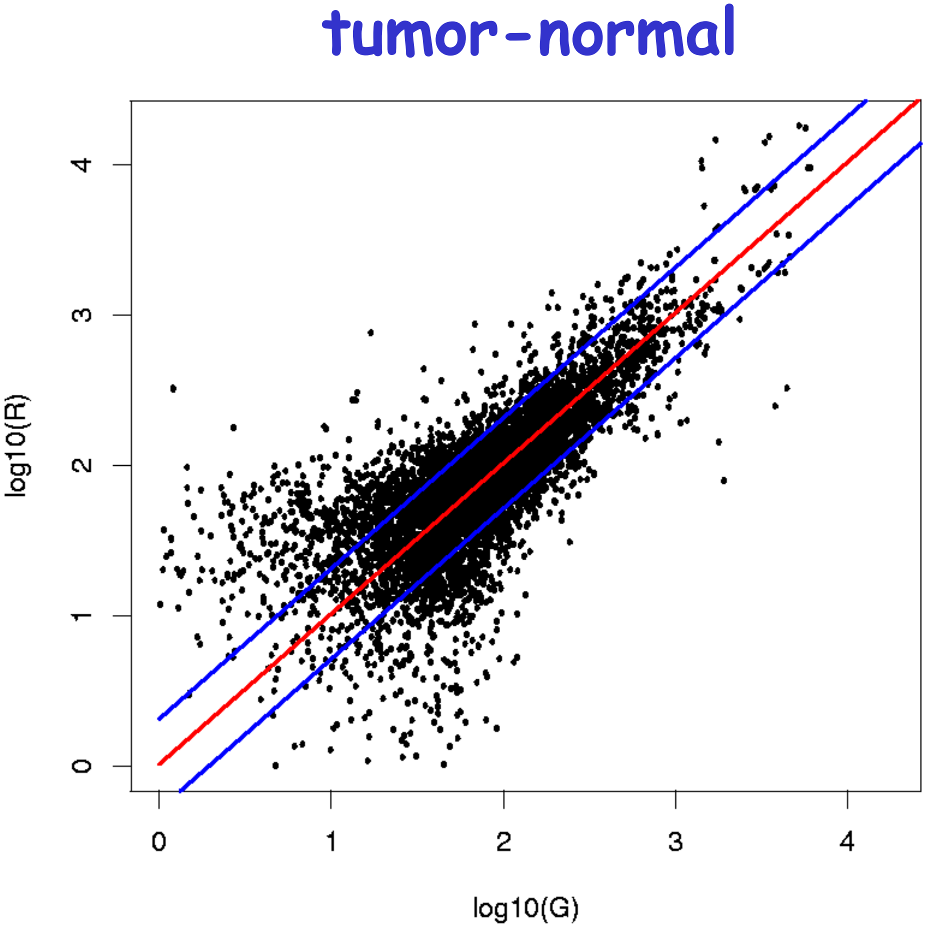
Huber, v. Heydebreck, et al.

Bioinformatics 18 suppl. 1 (2002), S96-S104

Which genes are differentially transcribed?



same-same



tumor-normal

Raw data are not mRNA concentrations

- tissue contamination
- RNA degradation
- amplification efficiency
- reverse transcription efficiency
- hybridization efficiency and specificity

- clone identification and mapping
- PCR yield, contamination
- spotting efficiency
- DNA-support binding
- other array manufacturing-related issues

- image segmentation
- signal quantification
- ‘background’ correction

Raw data are not mRNA concentrations

- tissue con
- R deg
- a eff
- r tra eff
- h eff specificity

- clone

related issues

- image

The problem is less that these steps are 'not perfect'; it is that they may vary from array to array, experiment to experiment.

Sources of variation

amount of RNA in the biopsy
efficiencies of
-RNA extraction
-reverse transcription
-labeling
-photodetection

PCR yield
DNA quality
spotting efficiency,
spot size
cross-/unspecific hybridization
stray signal

Sources of variation

amount of RNA in the biopsy
efficiencies of
-RNA extraction
-reverse transcription
-labeling
-photodetection

PCR yield
DNA quality
spotting efficiency,
spot size
cross-/unspecific hybridization
stray signal

Systematic

- o similar effect on many measurements
- o corrections can be estimated from data

Sources of variation

amount of RNA in the biopsy
efficiencies of
-RNA extraction
-reverse transcription
-labeling
-photodetection

PCR yield
DNA quality
spotting efficiency,
spot size
cross-/unspecific hybridization
stray signal

Systematic

- similar effect on many measurements
- corrections can be estimated from data

↓

Calibration

Sources of variation

amount of RNA in the biopsy
efficiencies of
-RNA extraction
-reverse transcription
-labeling
-photodetection

PCR yield
DNA quality
spotting efficiency,
spot size
cross-/unspecific hybridization
stray signal

Systematic

- similar effect on many measurements
- corrections can be estimated from data

Calibration

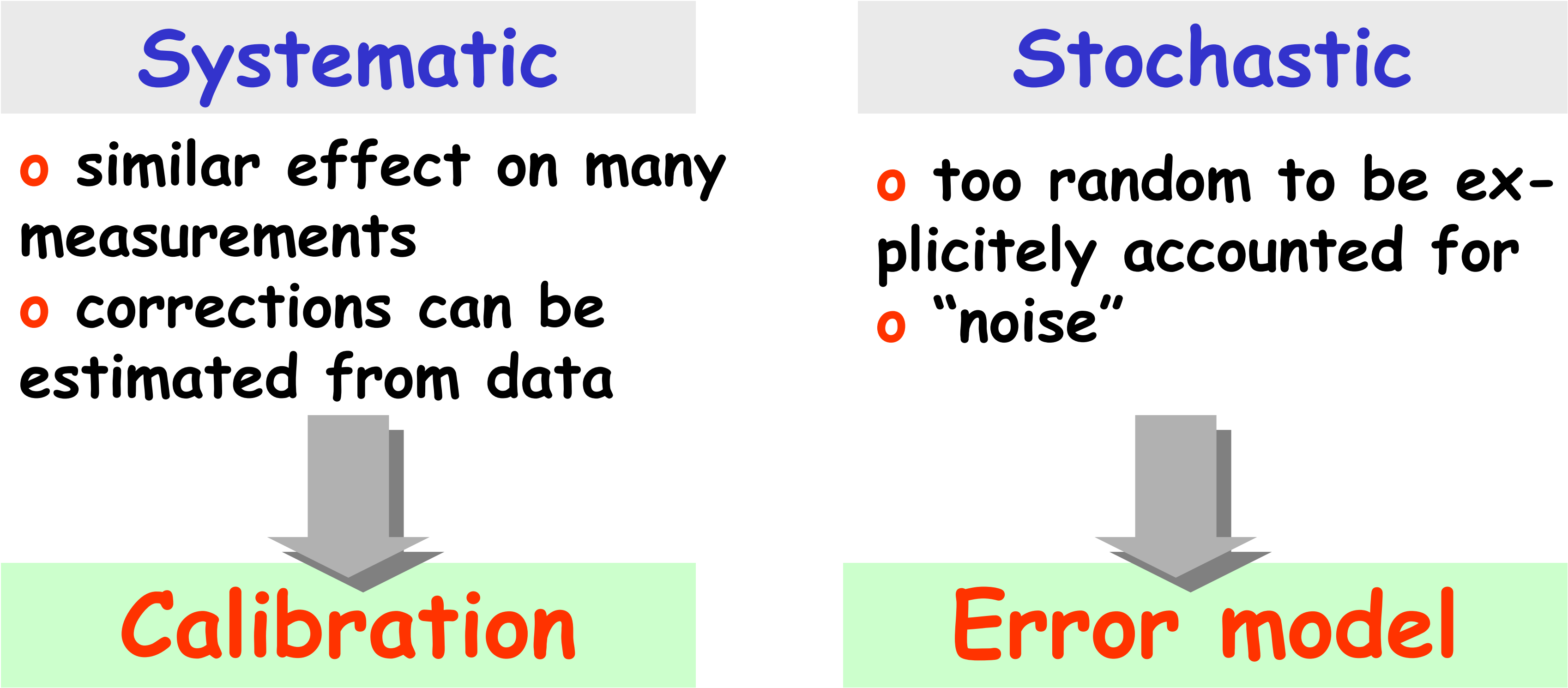
Stochastic

- too random to be explicitely accounted for
- “noise”

Sources of variation

amount of RNA in the biopsy
efficiencies of
-RNA extraction
-reverse transcription
-labeling
-photodetection

PCR yield
DNA quality
spotting efficiency,
spot size
cross-/unspecific hybridization
stray signal



Systematic

- similar effect on many measurements
- corrections can be estimated from data

Calibration

Stochastic

- too random to be explicitely accounted for
- “noise”

Error model

measured intensity = offset + gain * true abundance

$$y_{ik} = a_{ik} + b_{ik} x_{ik}$$

measured intensity = offset + gain * true abundance

$$y_{ik} = a_{ik} + b_{ik} x_{ik}$$

$$b_{ik} = b_i \, b_k \, \exp(\eta_{ik})$$

b_i per-sample normalization factor

b_k sequence-wise labeling efficiency

$\eta_{ik} \sim N(0, s_2^2)$ “multiplicative noise”

measured intensity = offset + gain * true abundance

$$y_{ik} = a_{ik} + b_{ik} x_{ik}$$

$$a_{ik} = a_i + L_{ik} + \varepsilon_{ik}$$

a_i per-sample offset

L_{ik} local background provided by image analysis

$\varepsilon_{ik} \sim N(0, b_i^2 s_1^2)$

"additive noise"

$$b_{ik} = b_i \, b_k \, \exp(\eta_{ik})$$

b_i per-sample normalization factor

b_k sequence-wise labeling efficiency

$\eta_{ik} \sim N(0, s_2^2)$

"multiplicative noise"

Calibration
("normalization")

Correct for systematic variations.

To do: fit appropriate "correction parameters" a_i, b_i, and apply to the data.

Calibration ("normalization")

Correct for systematic variations.
To do: fit appropriate "correction parameters" a_i, b_i, and apply to the data.

"Heteroskedasticity" (unequal variances)

Calibration
("normalization")

Correct for systematic variations.
To do: fit appropriate "correction parameters" a_i, b_i, and apply to the data.

"Heteroskedasticity" (unequal variances)
⇒ weighted regression or variance stabilizing transformation

Calibration ("normalization")

Correct for systematic variations.
To do: fit appropriate "correction parameters" a_i, b_i, and apply to the data.

"Heteroskedasticity" (unequal variances)
⇒ weighted regression or variance stabilizing transformation

Outliers:

Calibration ("normalization")

Correct for systematic variations.
To do: fit appropriate "correction parameters" a_i, b_i, and apply to the data.

"Heteroskedasticity" (unequal variances)
⇒ weighted regression or variance stabilizing transformation

Outliers:
⇒ use a robust method

Ordinary regression

Minimize the sum of squares

$$SoS = \sum_{\text{all } i} (\text{residual } i)^2$$

residual:= "fit" - "data"

Ordinary regression

Minimize the sum of squares

$$SoS = \sum_{\text{all } i} (\text{residual } i)^2$$

residual:= "fit" - "data"

Problem: all data points get the same weight, even if they come with different variance ('precision') - this may greatly distort the fit!

Ordinary regression

Minimize the sum of squares

$$SoS = \sum_{\text{all } i} (\text{residual } i)^2$$

residual:= "fit" - "data"

Problem: all data points get the same weight, even if they come with different variance ('precision') - this may greatly distort the fit!

Solution: weight them accordingly (some weights may be zero)

Weighted regression

$$SoS = \sum_{\text{all i}} w_i \times (\text{residual i})^2$$

If w_i = 1/variance(i), then minimizing SoS produces the maximum-likelihood estimate for a model with normal errors.

$$w(i) = \begin{cases} 1/\text{variance(i)} & \text{if residual(i)} \le \text{median(residuals)} \\ 0 & \text{otherwise} \end{cases}$$

Weighted regression

$$SoS = \sum_{\text{all } i} w_i \times (\text{residual } i)^2$$

If w_i = 1/variance(i), then minimizing SoS produces the maximum-likelihood estimate for a model with normal errors.

Least Median Sum of Squares Regression:

$$w(i) = \begin{cases} 1/\text{variance(i)} & \text{if residual(i)} \le \text{median(residuals)} \\ 0 & \text{otherwise} \end{cases}$$

But what is the variance of a measured spot intensity?

To estimate the variance of an individual probe, need many replicates from biologically identical samples. Often unrealistic.

Idea:

- o use pooled estimate from several probes who we expect to have about the same true (unknown) variance

$$\mathrm{var}_{\mathrm{pooled}} = \mathrm{mean}\ (\mathrm{var}_{\mathrm{individual\ probes}})$$

- o there is an obvious dependence of the variance on the mean intensity, hence stratify (group) probes by that.

the variance-mean dependence

model:

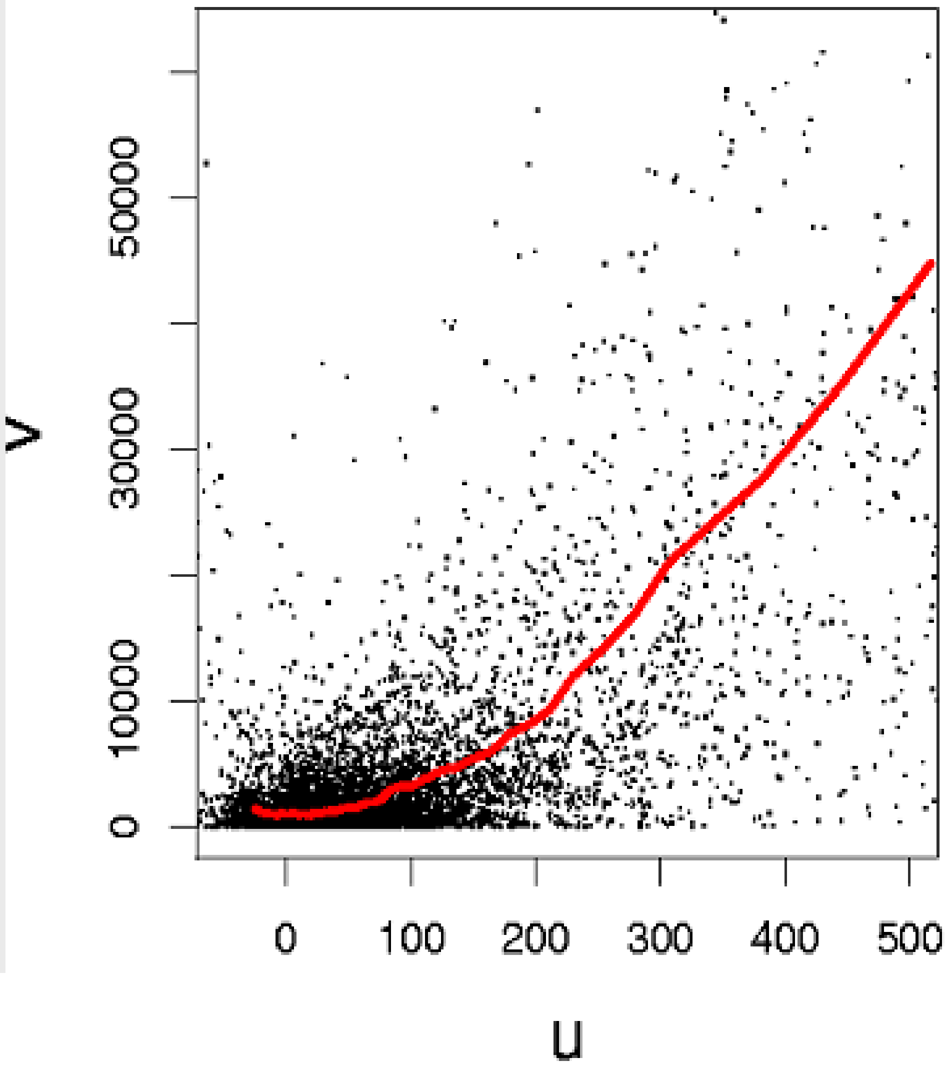
⇒ relation between

u≡E(Y_{ik})

v≡Var(Y_{ik})

$$\boxed{v(u) = c^2(u+u_0)^2 + s^2}$$

data (cDNA slide):



variance stabilization

X_u a family of random variables with EX_u=u, VarX_u= $v(u)$.

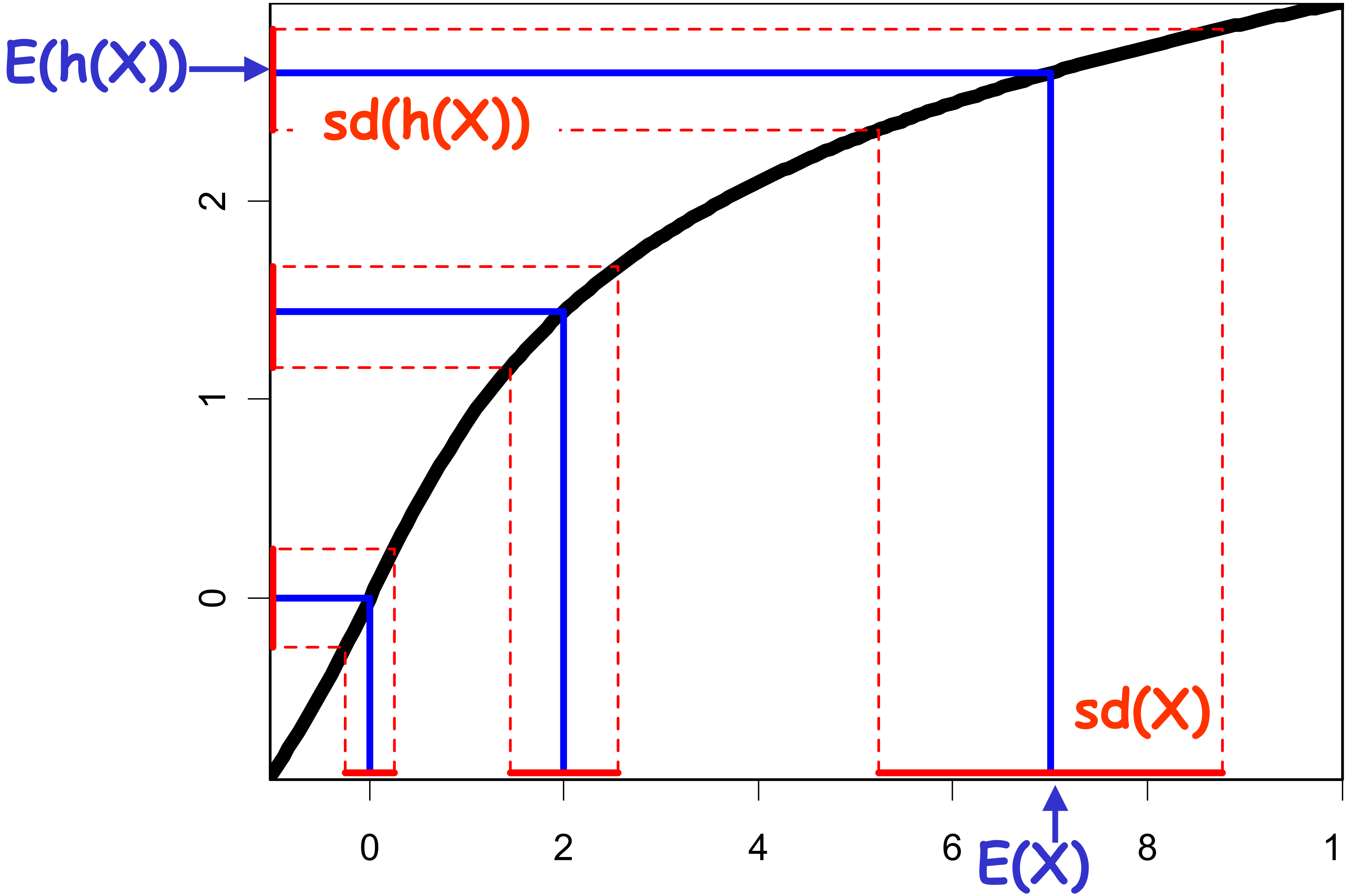
Define

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

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

derivation: linear approximation

variance stabilizing transformation



variance stabilizing transformations

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

variance stabilizing transformations

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

1.) constant variance $\qquad v(u) = \text{const} \quad \Rightarrow \quad f \propto u$

variance stabilizing transformations

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

1.) constant variance $v(u) = \text{const} \Rightarrow f \propto u$

2.) const. coeff. of variation $v(u) \propto u^2 \Rightarrow f \propto \log u$

variance stabilizing transformations

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

1.) constant variance $v(u) = \text{const} \quad \Rightarrow \quad f \propto u$

2.) const. coeff. of variation $v(u) \propto u^2 \quad \Rightarrow \quad f \propto \log u$

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

variance stabilizing transformations

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

1.) constant variance $v(u) = \text{const} \Rightarrow f \propto u$

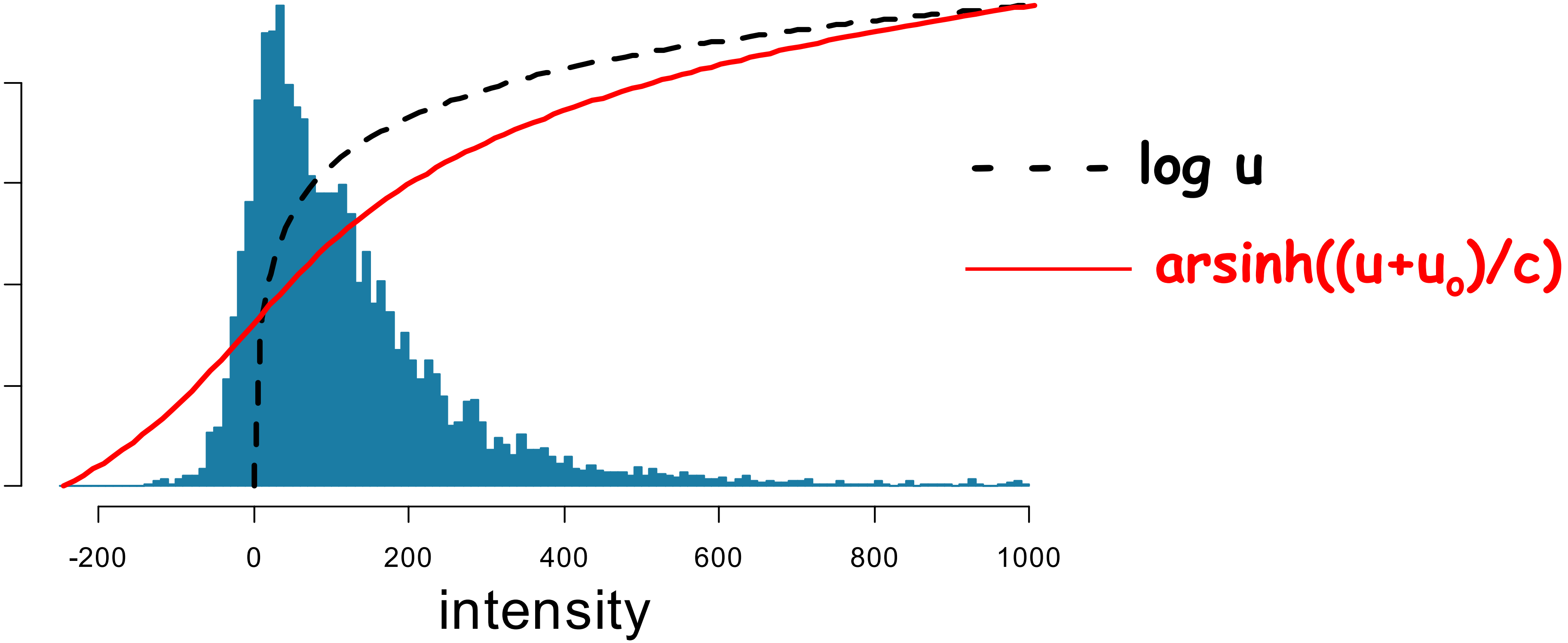
2.) const. coeff. of variation $v(u) \propto u^2 \Rightarrow f \propto \log u$

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

4.) microarray

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

the arsinh transformation



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

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

parameter estimation

$$\operatorname{arsinh}\frac{y_{ki}-a_i}{b_i}=\mu_k+\varepsilon_{ki}, \qquad \varepsilon_{ki}\sim N(0,c^2)$$

parameter estimation

$$\operatorname{arsinh}\frac{y_{ki} - a_i}{b_i} = \mu_k + \varepsilon_{ki}, \qquad \varepsilon_{ki} \sim N(0, c^2)$$

- maximum likelihood estimator: straightforward

– but sensitive to deviations from normality

parameter estimation

$$\operatorname{arsinh}\frac{y_{ki}-a_i}{b_i}=\mu_k+\varepsilon_{ki}, \qquad \varepsilon_{ki}\sim N(0,c^2)$$

o **maximum likelihood estimator**: straightforward – but sensitive to deviations from normality

o models holds for genes that are unchanged; differentially transcribed genes act as **outliers**.

parameter estimation

$$\operatorname{arsinh}\frac{y_{ki}-a_i}{b_i}=\mu_k+\varepsilon_{ki},\qquad \varepsilon_{ki}\sim N(0,c^2)$$

- **maximum likelihood estimator**: straightforward – but sensitive to deviations from normality

- models holds for genes that are unchanged; differentially transcribed genes act as outliers.

- robust variant of ML estimator, à la *Least Trimmed Sum of Squares* regression.

parameter estimation

$$\operatorname{arsinh}\frac{y_{ki}-a_i}{b_i}=\mu_k+\varepsilon_{ki},\qquad \varepsilon_{ki}\sim N(0,c^2)$$

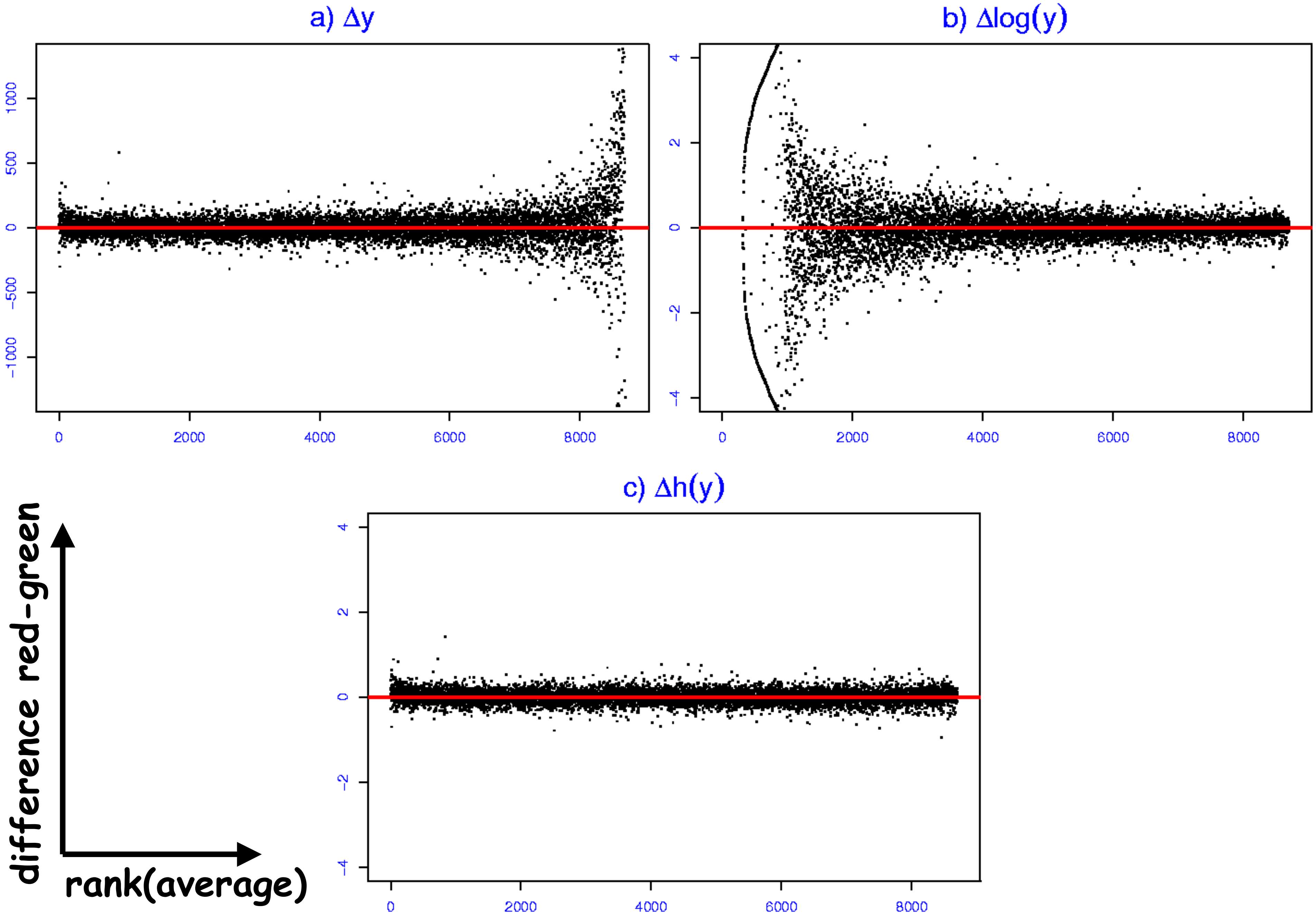
- **maximum likelihood estimator**: straightforward – but sensitive to deviations from normality

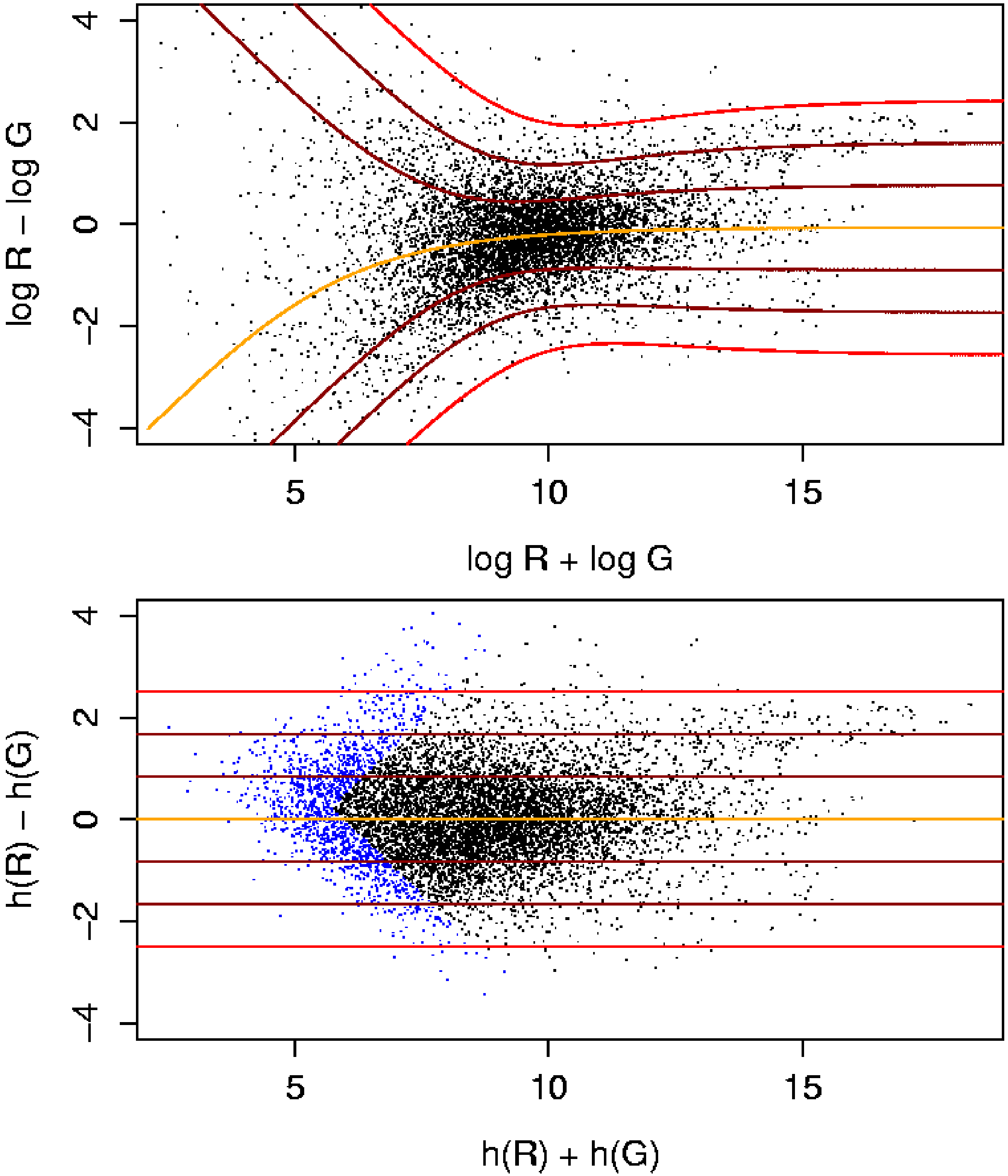
- models holds for genes that are unchanged; differentially transcribed genes act as **outliers**.

- **robust** variant of ML estimator, à la *Least Trimmed Sum of Squares* regression.

- works as long as <50% of genes are differentially transcribed

evaluation: effects of different data transformations





Coefficient of variation

cDNA slide:
H. Sueltmann

Summary

log-ratio

$$\boxed{\log\frac{y_{k1}-a_1}{b_1}-\log\frac{y_{k2}-a_2}{b_2}}$$

'generalized' log-ratio

$$\boxed{\operatorname{arsinh}\frac{y_{k1}-a_1}{b_1}-\operatorname{arsinh}\frac{y_{k2}-a_2}{b_2}}$$

- advantages of variance-stabilizing data-transformation: generally better applicability of statistical methods (hypothesis testing, ANOVA, clustering, classification…)
- R package vsn

References

Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. YH Yang, S Dudoit, P Luu, DM Lin, V Peng, J Ngai and TP Speed. *Nucl. Acids Res.* 30(4):e15, 2002.

Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. W.Huber, A.v.Heydebreck, H.Sültmann, A.Poustka, M.Vingron. *Bioinformatics*, Vol.18, Supplement 1, S96-S104, 2002.

A Variance-Stabilizing Transformation for Gene Expression Microarray Data. : Durbin BP, Hardin JS, Hawkins DM, Rocke DM. *Bioinformatics*, Vol.18, Suppl. 1, S105-110.

Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP (2002). Accepted for publication in *Biostatistics.* http://biosun01.biostat.jhsph.edu/~ririzarr/papers/index.html

A more complete list of references is in:

Elementary analysis of microarray gene expression data. W. Huber, A. von Heydebreck, M. Vingron, manuscript.

http://www.dkfz-heidelberg.de/abt0840/whuber/