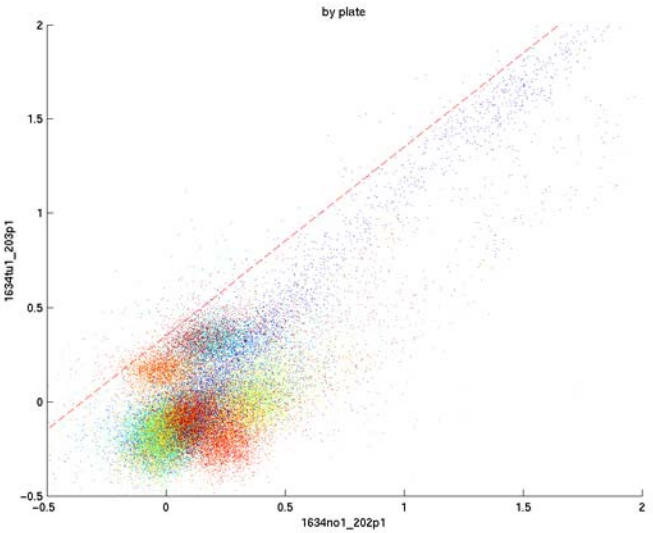
First analysis steps

- quality control and optimization
- calibration and error modeling
- data transformations

Wolfgang Huber

Dep. of Molecular Genome Analysis (A. Poustka)

DKFZ Heidelberg



Acknowledgements

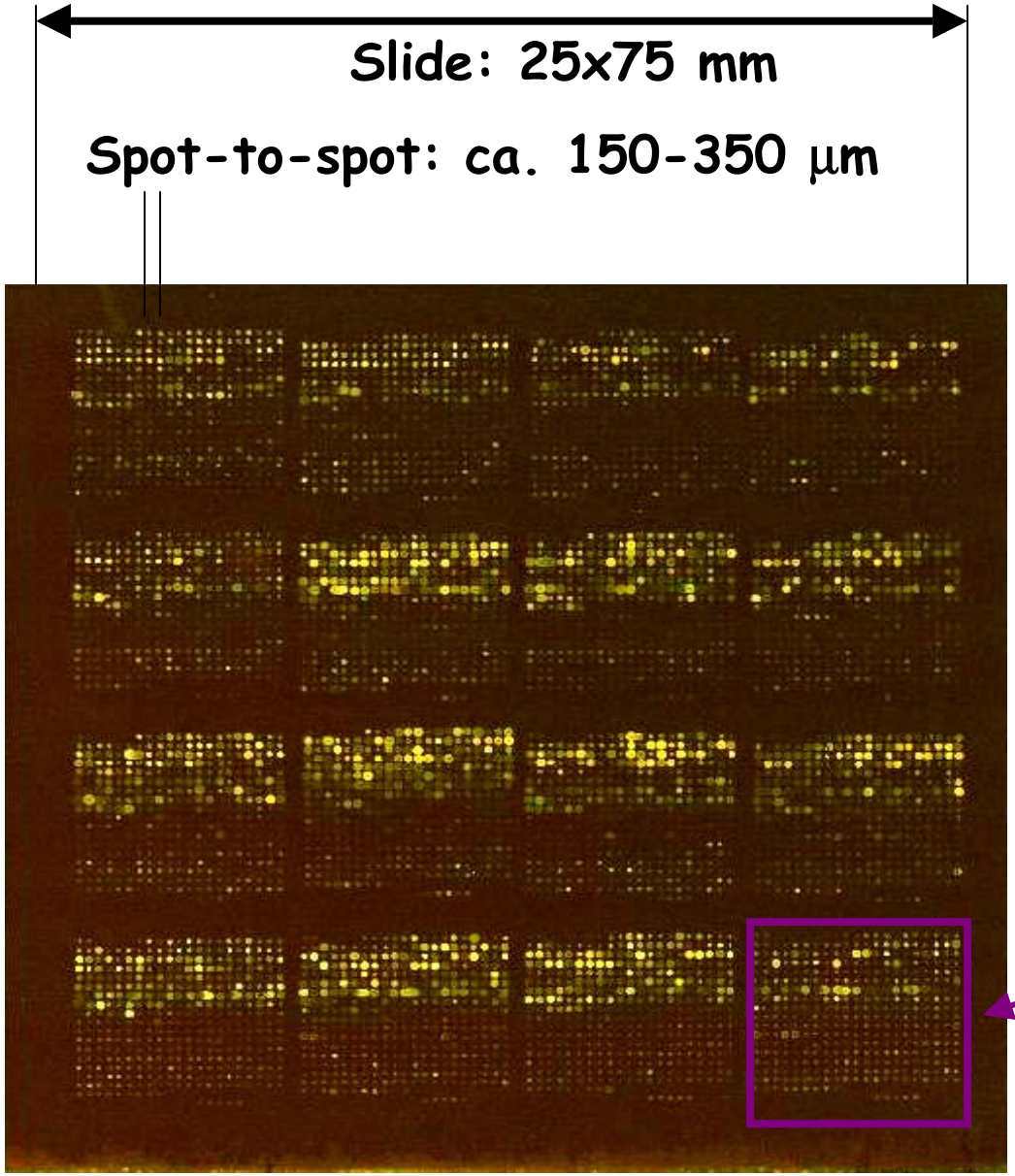
Anja von Heydebreck
Günther Sawitzki

Holger Sültmann, Klaus Steiner, Markus Vogt, Jörg Schneider, Frank Bergmann, Florian Haller, Katharina Finis, Stephanie Süß, Anke Schroth, Friederike Wilmer, Judith Boer, Martin Vingron, Annemarie Poustka

Sandrine Dudoit, Robert Gentleman, Rafael Irizarry and Yee Hwa Yang: Bioconductor short course, summer 2002

and many others

a microarray slide



4 x 4 or 8x4 sectors

17...38 rows and columns per sector

ca. 4600...46000 probes/array

sector: corresponds to one print-tip

Terminology

sample: RNA (cDNA) hybridized to the array, aka target, mobile substrate.

probe: DNA spotted on the array, aka spot, immobile substrate.

sector: rectangular matrix of spots printed using the same print-tip (or pin), aka print-tip-group

plate: set of 384 (768) spots printed with DNA from the same microtitre plate of clones

slide, array

channel: data from one color (Cy3 = cyanine 3 = green, Cy5 = cyanine 5 = red).

batch: collection of microarrays with the same probe layout.

Raw data

scanner signal

resolution:

5 or 10 mm spatial,

16 bit (65536) dynamical per channel

ca. 30-50 pixels per probe (60 μm spot size)

40 MB per array

Raw data

scanner signal

resolution:

5 or 10 mm spatial,

16 bit (65536) dynamical per channel

ca. 30-50 pixels per probe (60 μm spot size)

40 MB per array

Raw data

scanner signal

resolution:

5 or 10 mm spatial,

16 bit (65536) dynamical per channel

ca. 30-50 pixels per probe (60 μm spot size)

40 MB per array

spot intensities

2 numbers per probe (~100-300 kB)

… auxiliaries: background, area, std dev, …

Image analysis

1. Addressing. Estimate location of spot centers.

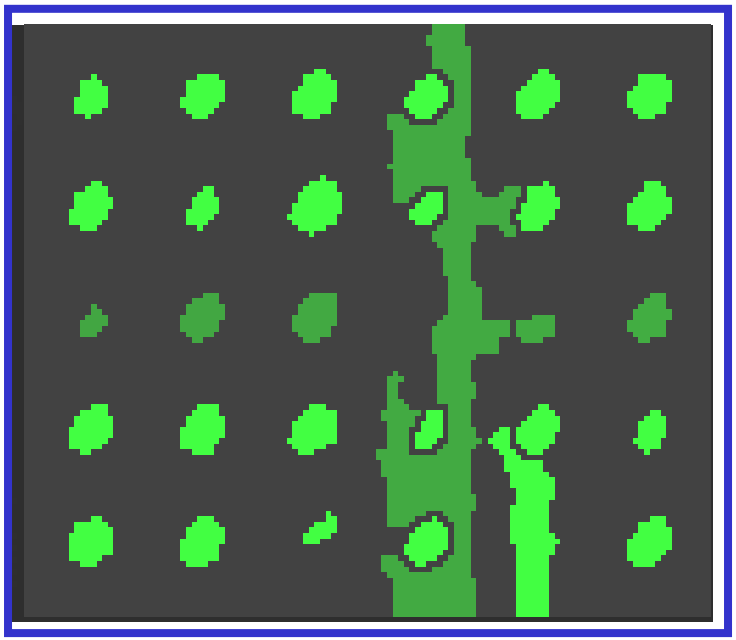


Image analysis

1. Addressing. Estimate location of spot centers.

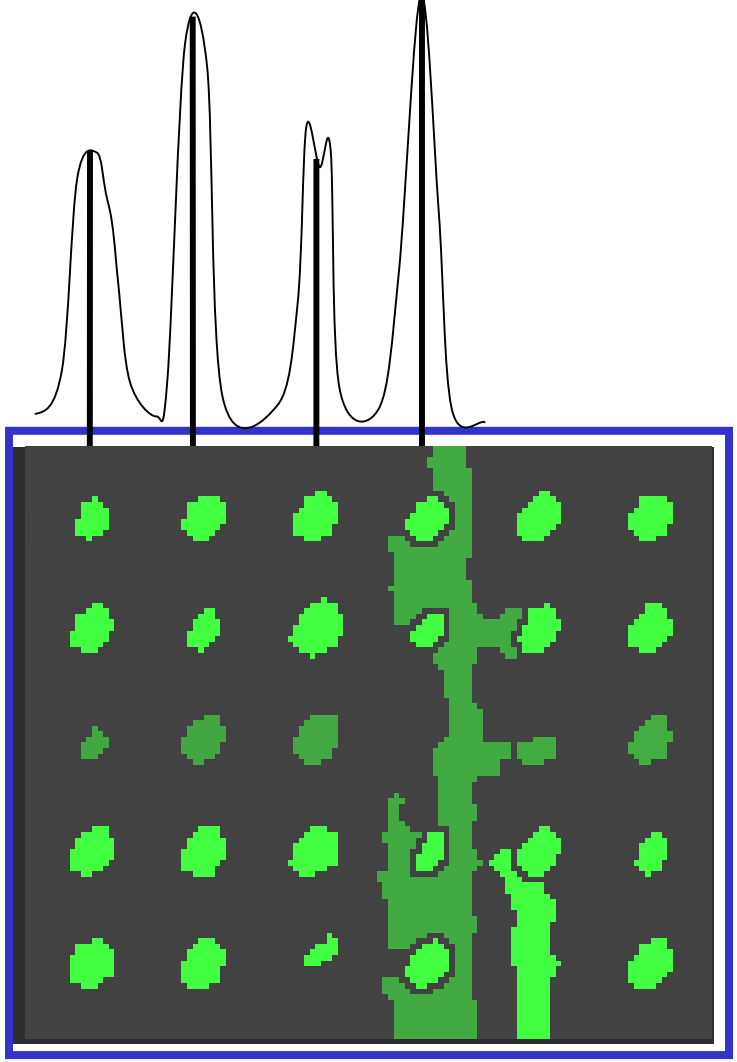


Image analysis

1. Addressing. Estimate location of spot centers.

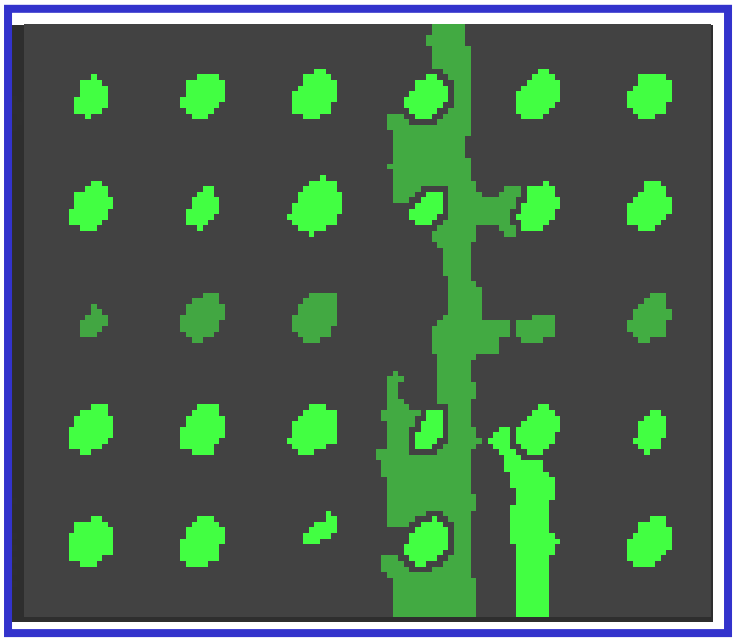


Image analysis

1. Addressing. Estimate location of spot centers.

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

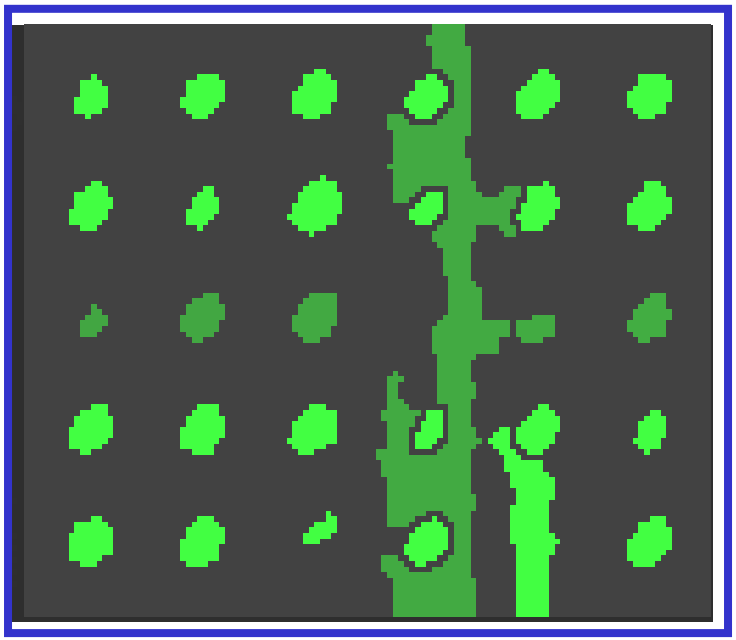


Image analysis

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
   - foreground intensities;
   - background intensities;
   - quality measures.

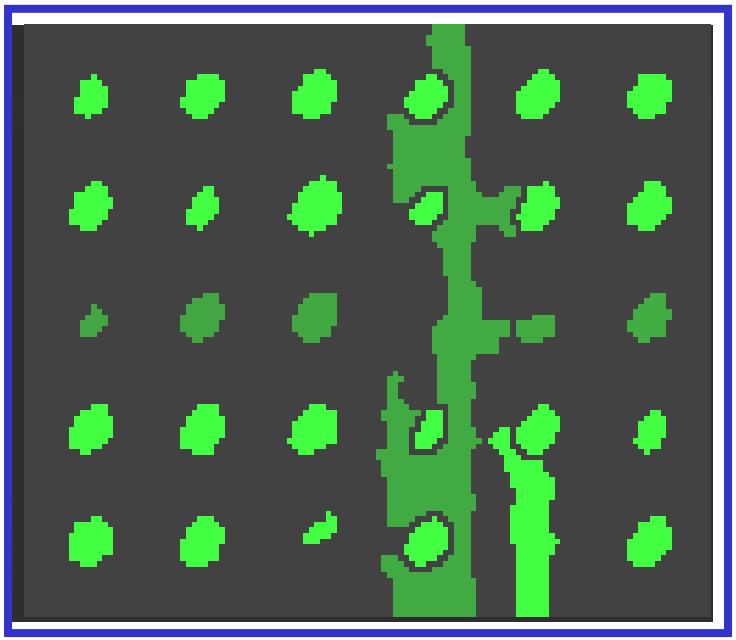
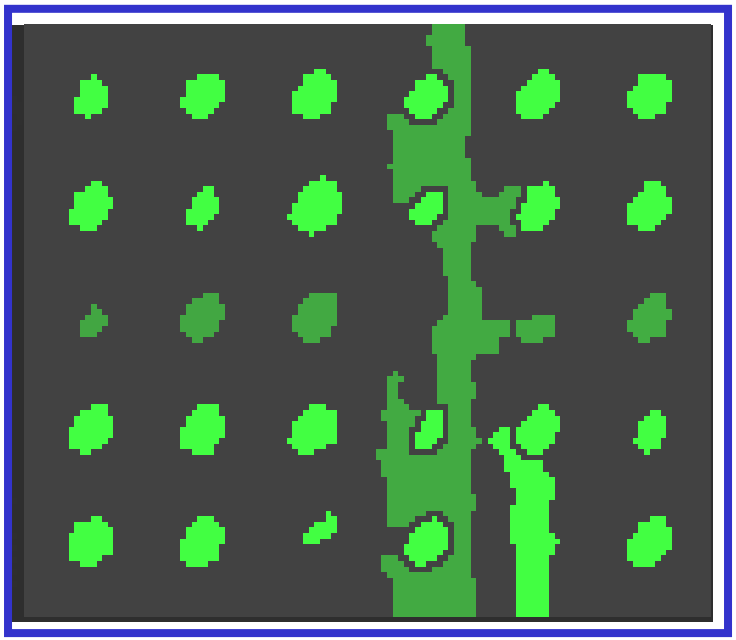


Image analysis

1. Addressing. Estimate location of spot centers.

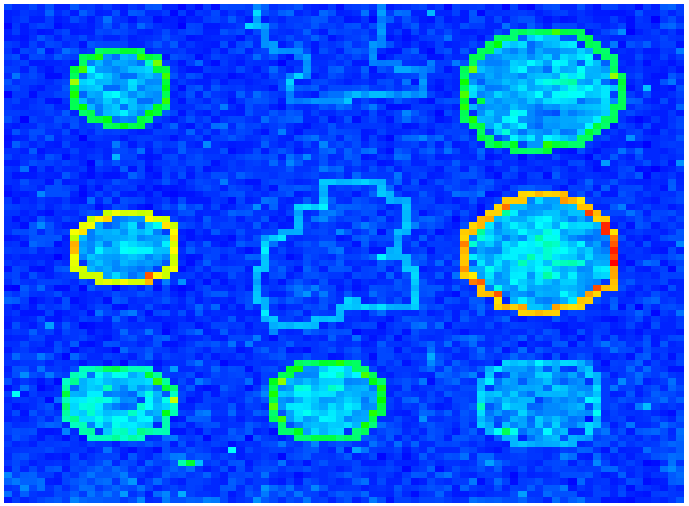
2. Segmentation. Classify pixels as foreground (signal) or background.

3. Information extraction. For each spot on the array and each dye
   - foreground intensities;
   - background intensities;
   - quality measures.

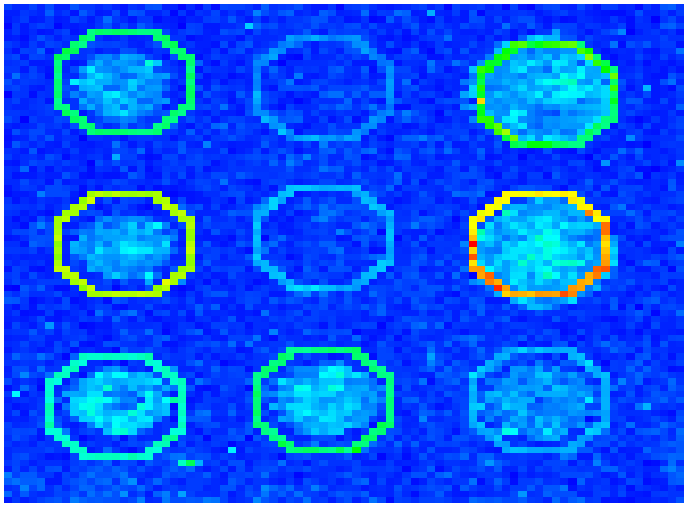


→ R and G for each spot on the array.

Segmentation



adaptive segmentation
seeded region growing



fixed circle segmentation

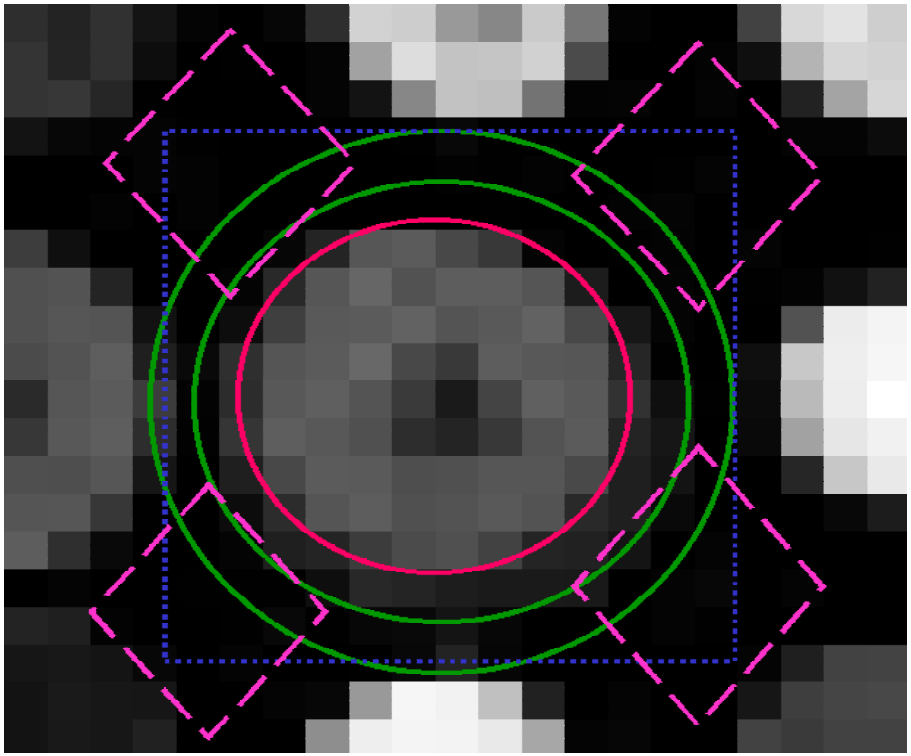
Spots may vary in size and shape.

Local background

---- GenePix

---- QuantArray

---- ScanAlyze



Local background estimation by morphological opening

Image is probed with a window (aka structuring element), eg, a square with side length about twice the spot-to-spot distance.

Local background estimation by morphological opening

Image is probed with a window (aka structuring element), eg, a square with side length about twice the spot-to-spot distance.

Erosion: at each pixel, replace its value by the minimum value in the window around it.

Local background estimation by morphological opening

Image is probed with a window (aka structuring element), eg, a square with side length about twice the spot-to-spot distance.

Erosion: at each pixel, replace its value by the minimum value in the window around it.

followed by

Dilation: same with maximum

Local background estimation by morphological opening

Image is probed with a window (aka structuring element), eg, a square with side length about twice the spot-to-spot distance.

Erosion: at each pixel, replace its value by the minimum value in the window around it.

followed by

Dilation: same with maximum

Do this separately for red and green images. This 'smoothes away' all structures that are smaller than the window

Local background estimation by morphological opening

Image is probed with a window (aka structuring element), eg, a square with side length about twice the spot-to-spot distance.

Erosion: at each pixel, replace its value by the minimum value in the window around it.

followed by

Dilation: same with maximum

Do this separately for red and green images. This 'smoothes away' all structures that are smaller than the window

$\Rightarrow$ Image of the estimated background

What is (local) background?

usual assumption:

total brightness =

background brightness (adjacent to spot)

\+ brightness from labeled sample cDNA

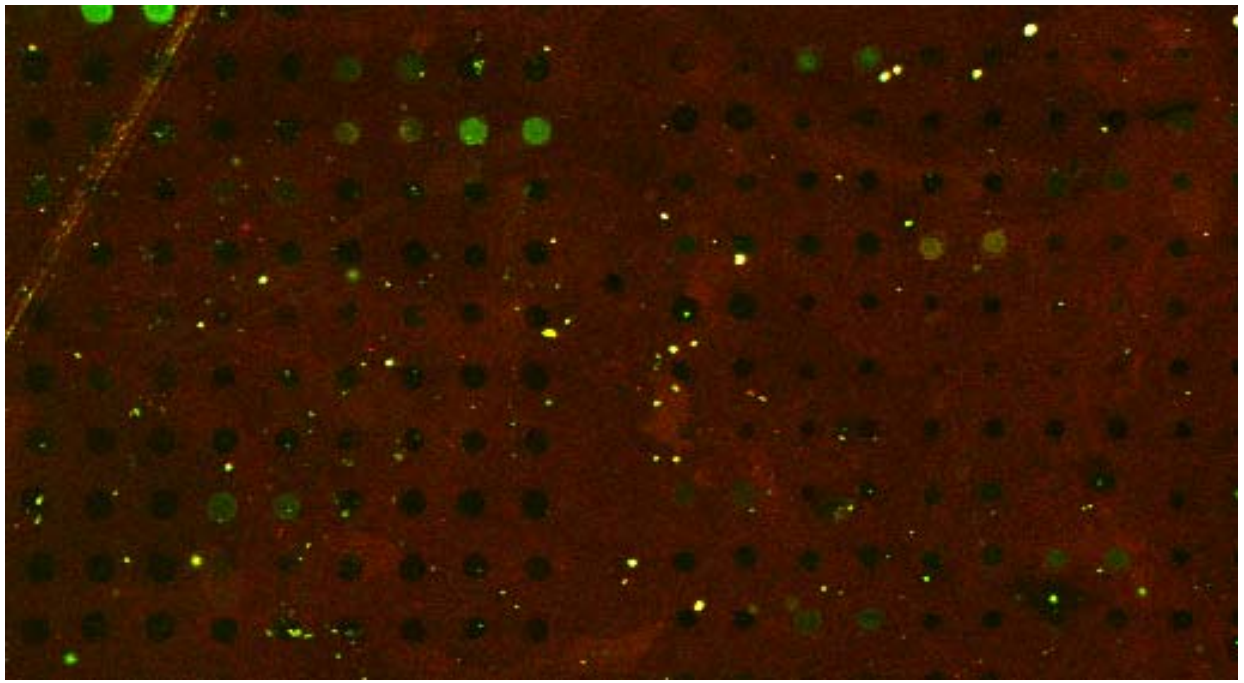
What is (local) background?

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

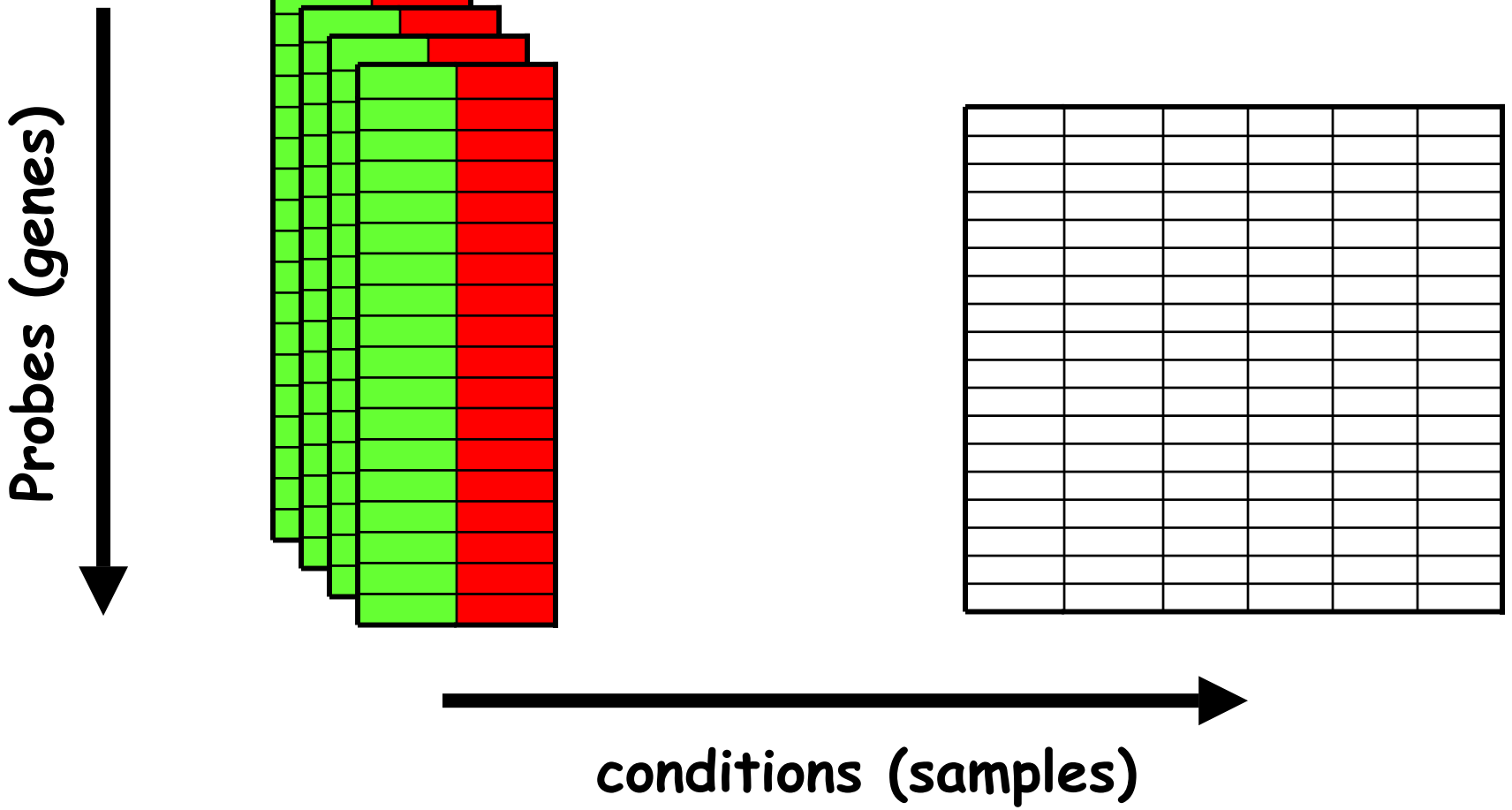
- Percentage of spots with no signal
- Range of intensities
- Distribution of spot signal area, etc.

How to use quality measures in subsequent analyses?

spot intensity data

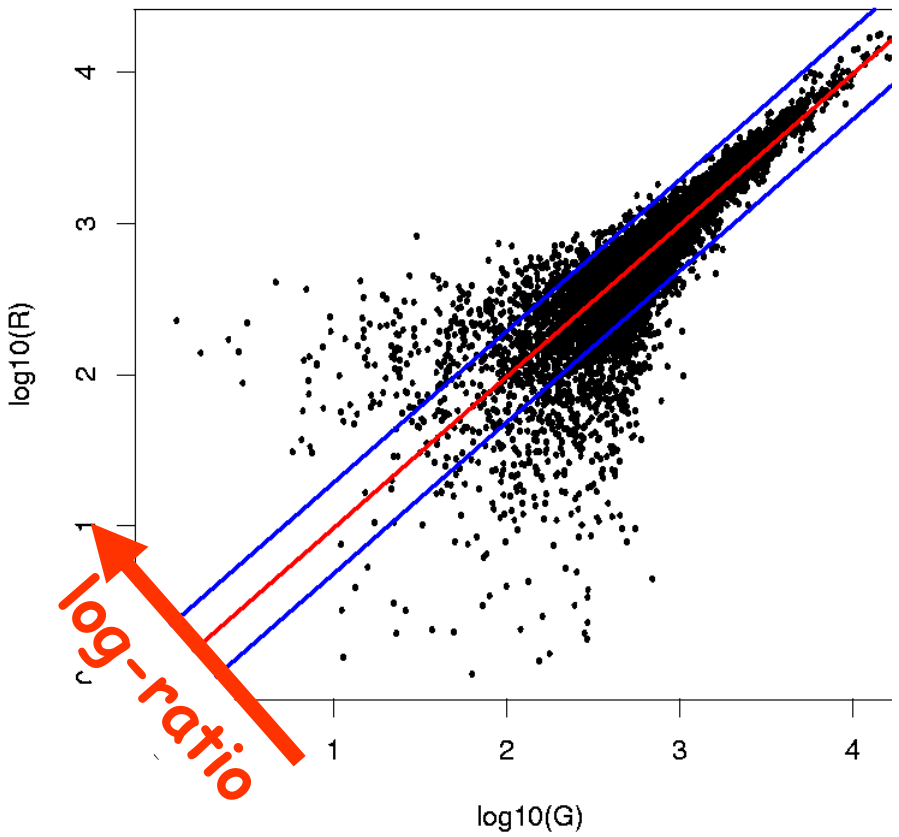
two-color spotted arrays

n one-color arrays
(Affymetrix, nylon)

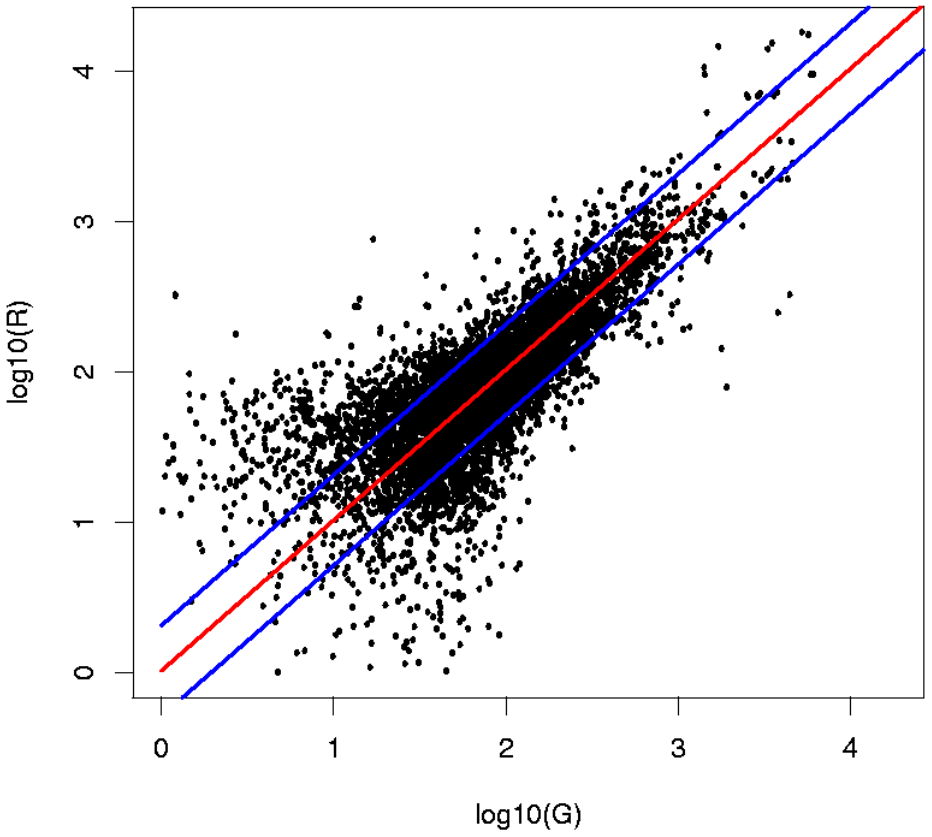


Which genes are differentially transcribed?

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
- clone
- image

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

specificity

related issues

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

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

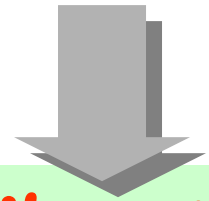
Sources of variation

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

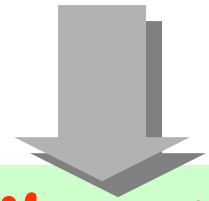
PCR yield
DNA quality
spotting efficiency,
spot size
cross-/unspecific hybridization
stray signal

Systematic

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

Stochastic

- too random to be explicitely accounted for
- "noise"



Calibration

Sources of variation

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

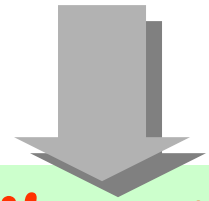
PCR yield
DNA quality
spotting efficiency,
spot size
cross-/unspecific hybridization
stray signal

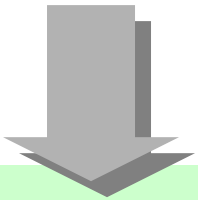
Systematic

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

Stochastic

- too random to be explicitely accounted for
- "noise"





Calibration

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)

$\Rightarrow$ 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)

$\Rightarrow$ 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)

$\Rightarrow$ weighted regression or variance stabilizing transformation

Outliers:

$\Rightarrow$ 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:

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

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

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

the variance-mean dependence

model:

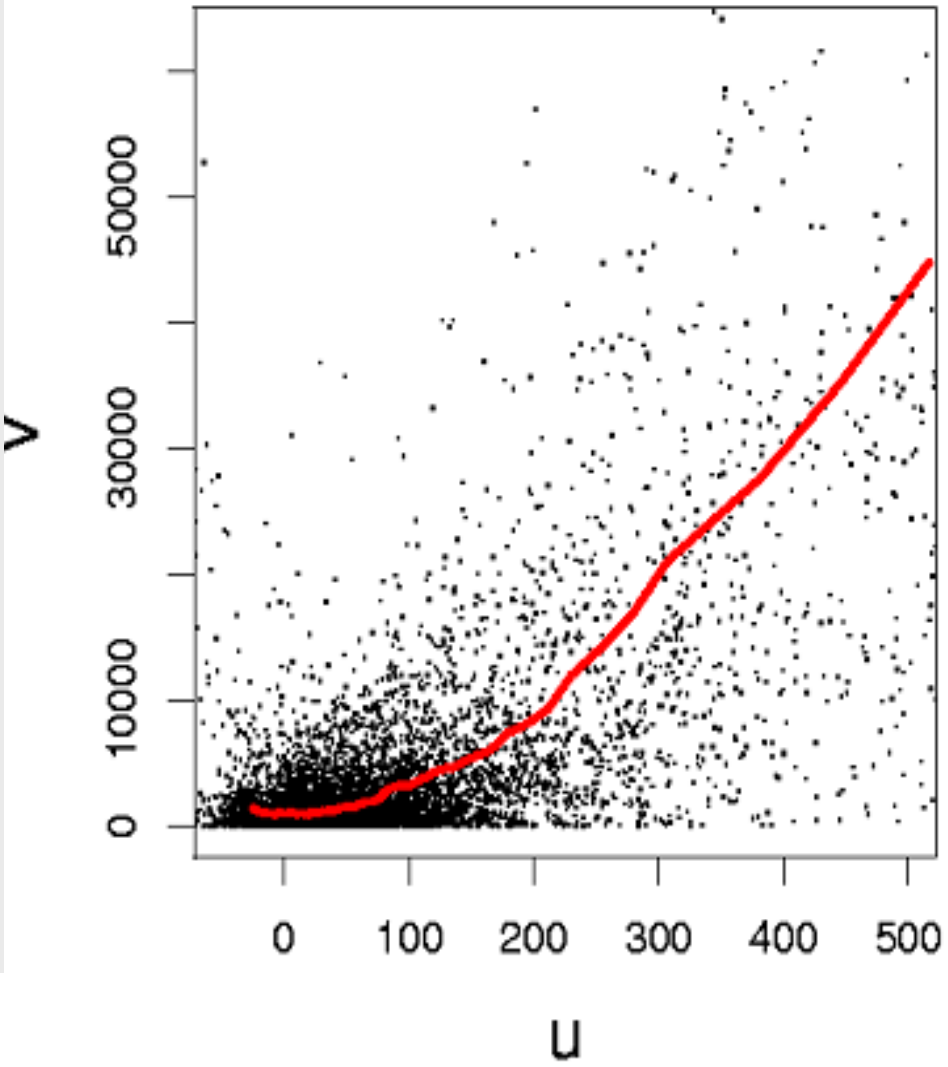
$\Rightarrow$ relation between

$u \equiv E(Y_{ik})$

$v \equiv Var(Y_{ik})$

$$v(u) = c^2(u + u_0)^2 + s^2$$

data (cDNA slide):



variance stabilization

X_u a family of random variables with

$EX_u = u$, $\mathrm{Var} X_u = v(u)$.

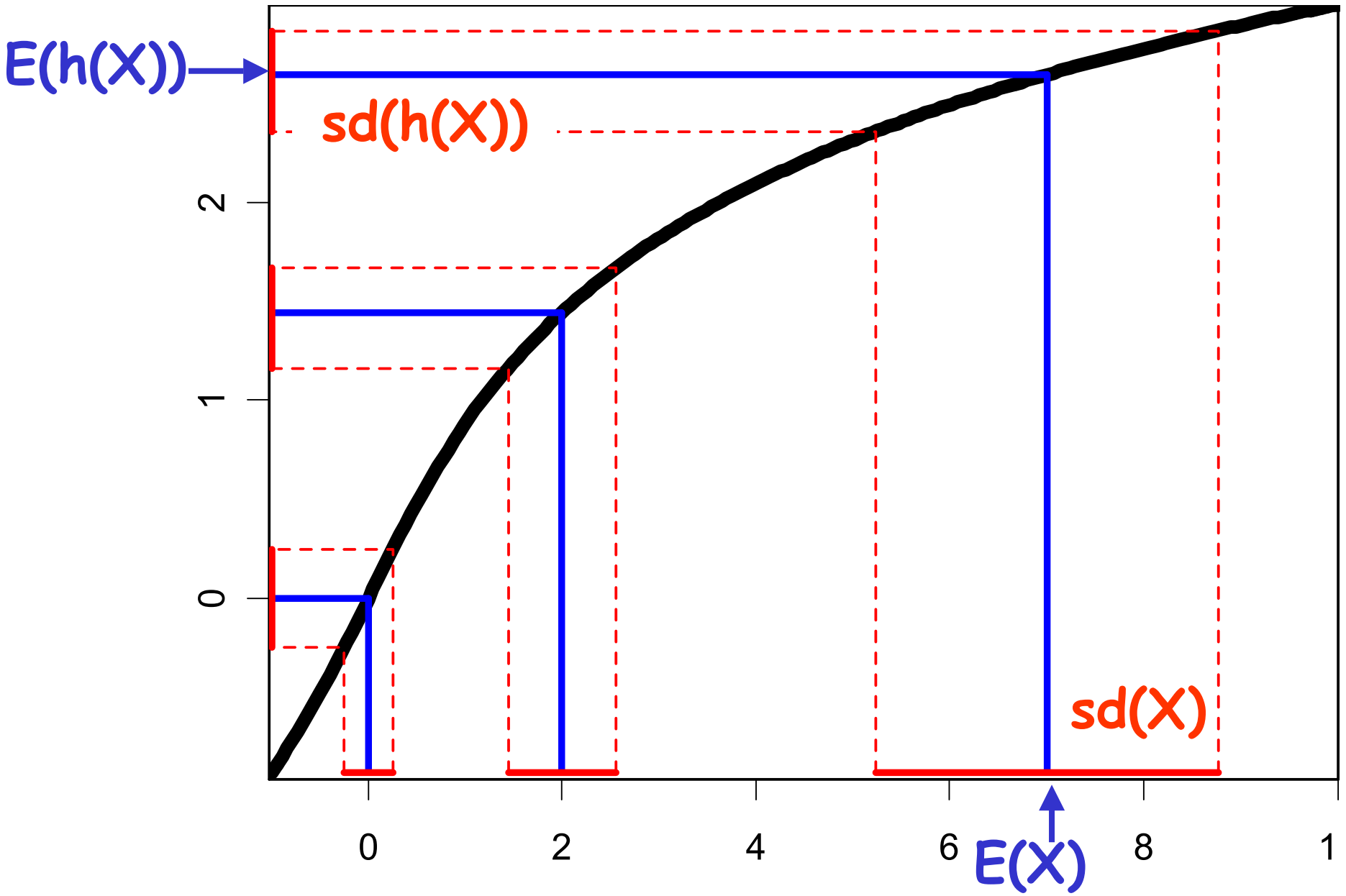
Define

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

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

derivation: linear approximation

variance stabilizing transformation



variance stabilizing transformations

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

variance stabilizing transformations

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

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

variance stabilizing transformations

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

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

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

variance stabilizing transformations

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

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

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

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

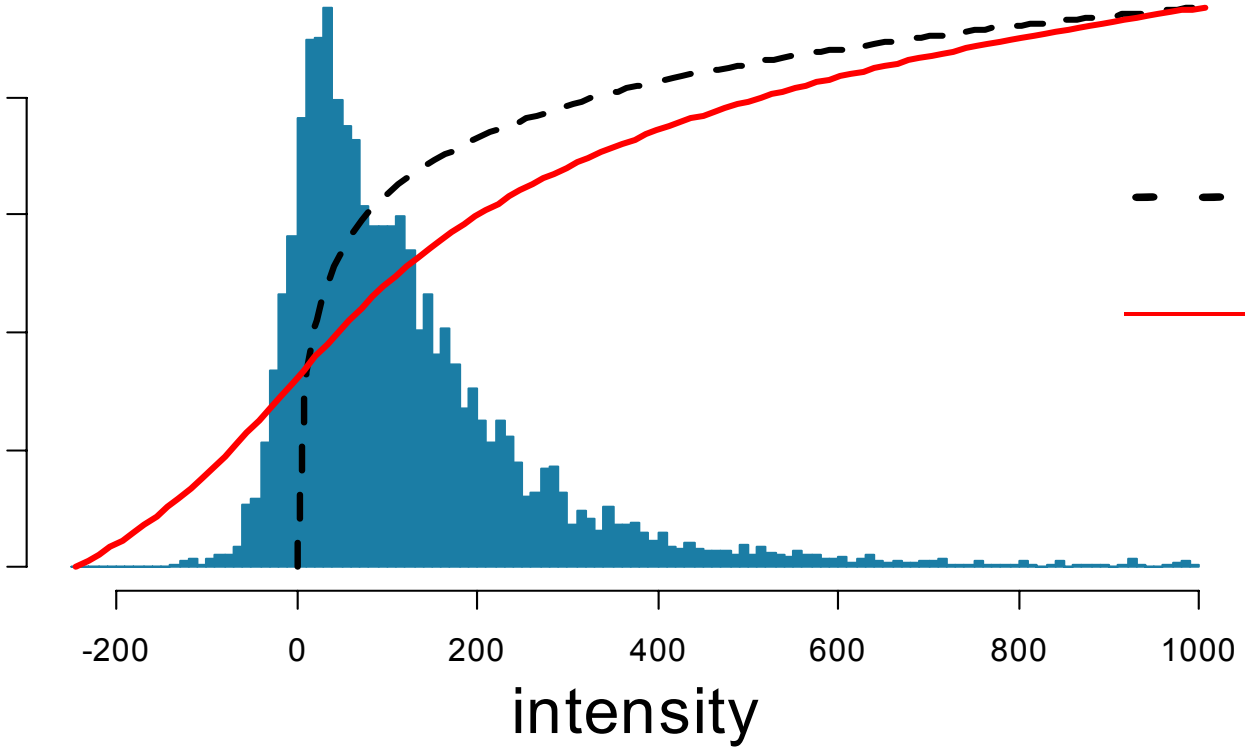
2.) const. coeff. of variation $v(u) \propto u^2 \quad \Rightarrow \quad f \propto \log u$

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

4.) microarray

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

the arsinh transformation



- - - log u

——— arsinh((u+u_0)/c)

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

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

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

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

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

- 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

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

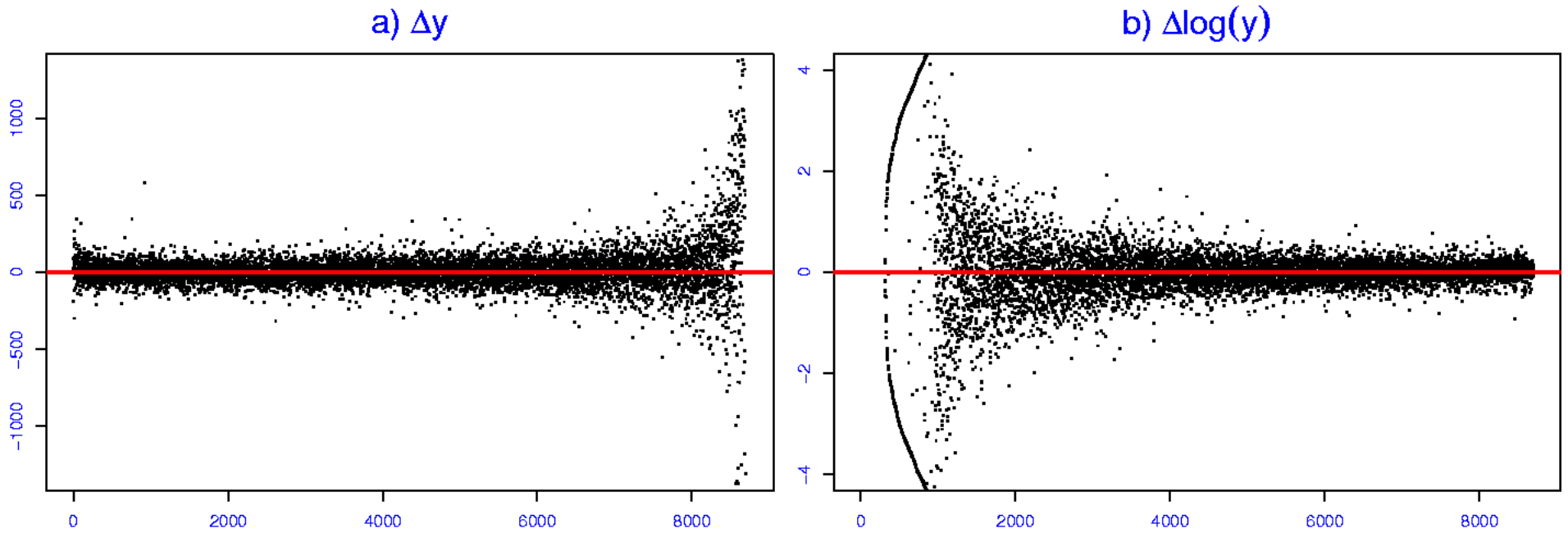
- maximum likelihood estimator: straightforward – but sensitive to deviations from normality

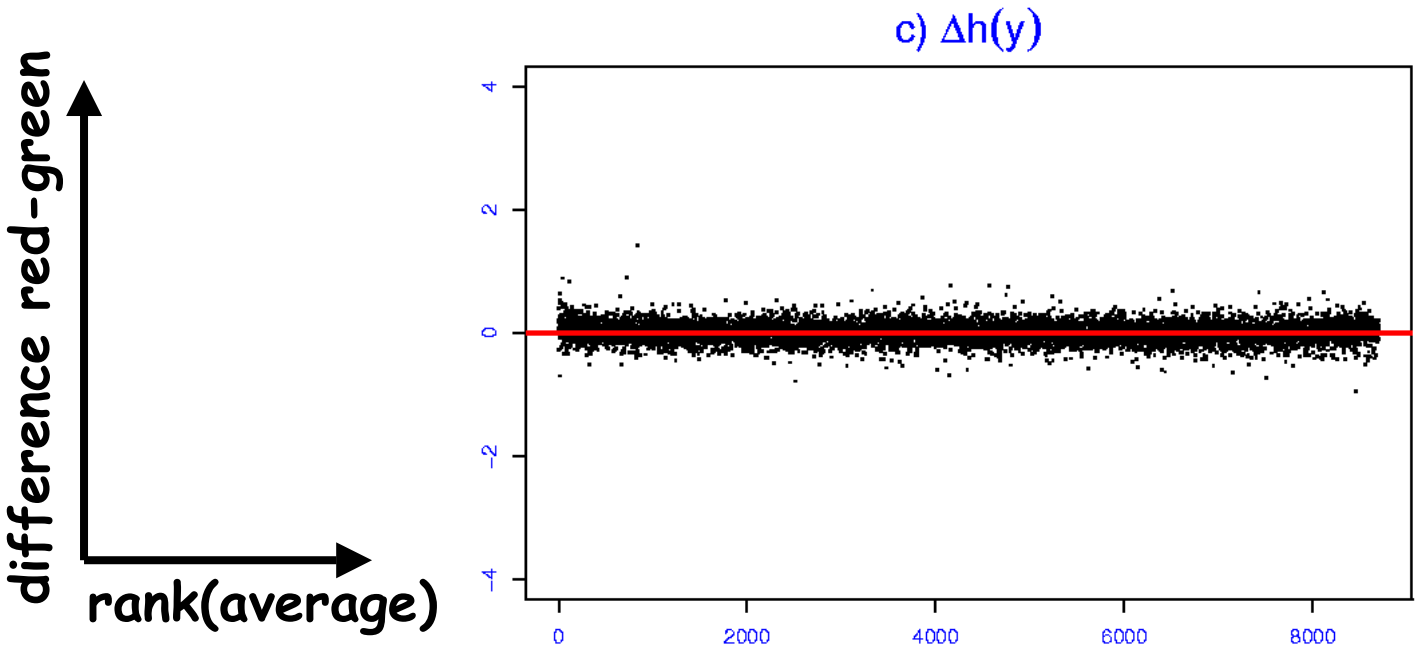
- models holds for genes that are unchanged; differentially transcribed genes act as outliers.

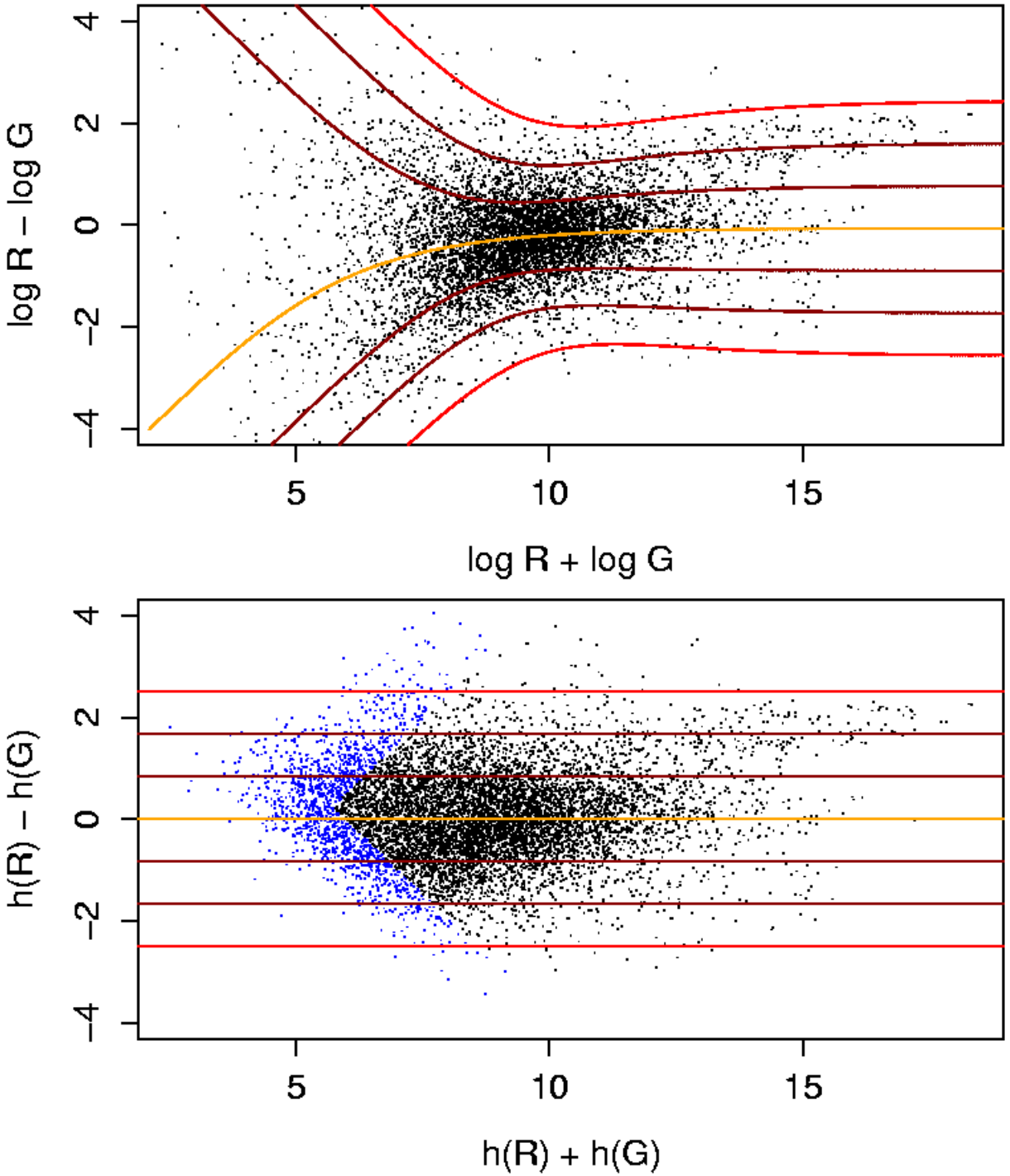
- robust variant of ML estimator, à la *Least Trimmed Sum of Squares* regression.

- 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

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

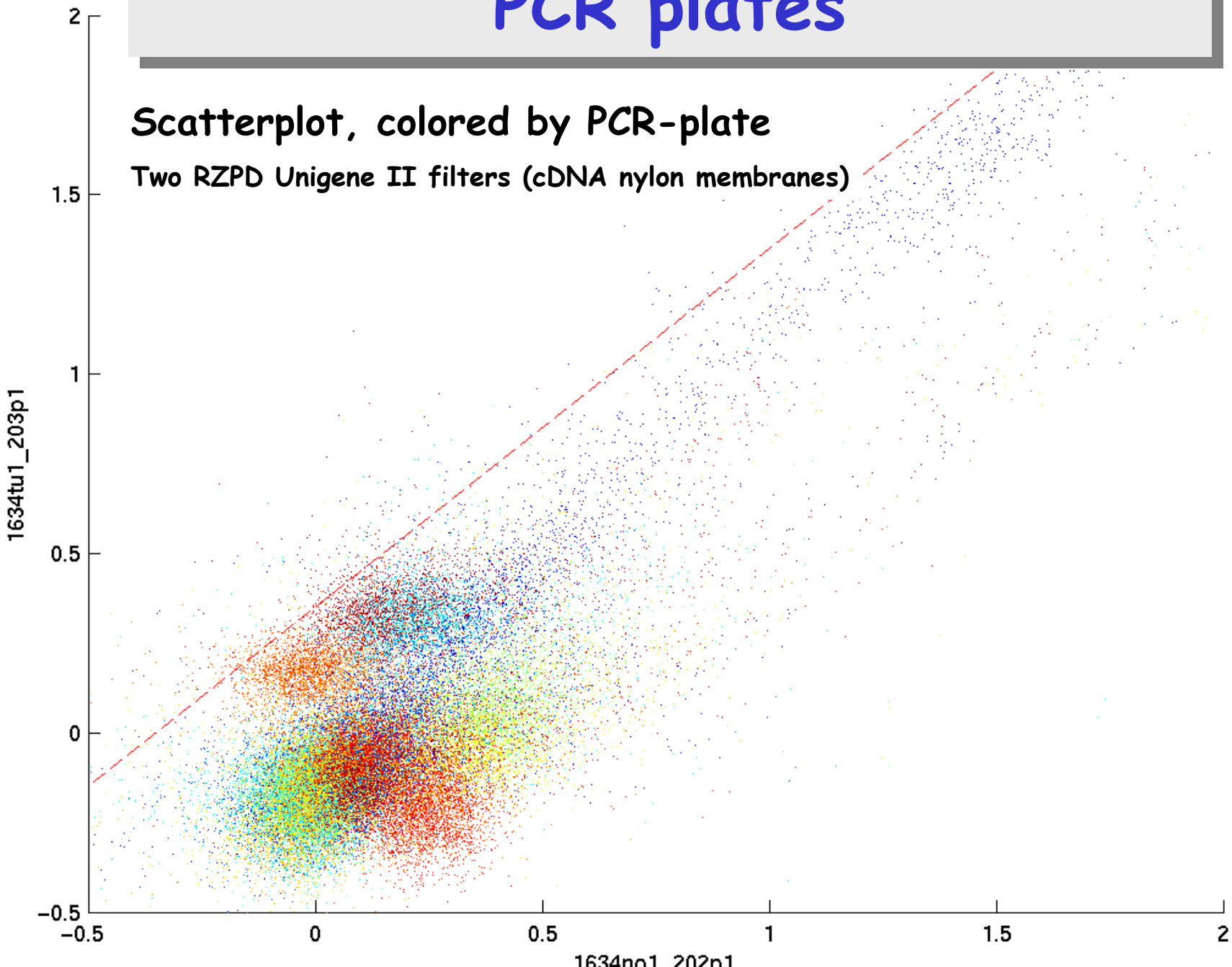
'generalized' log-ratio

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

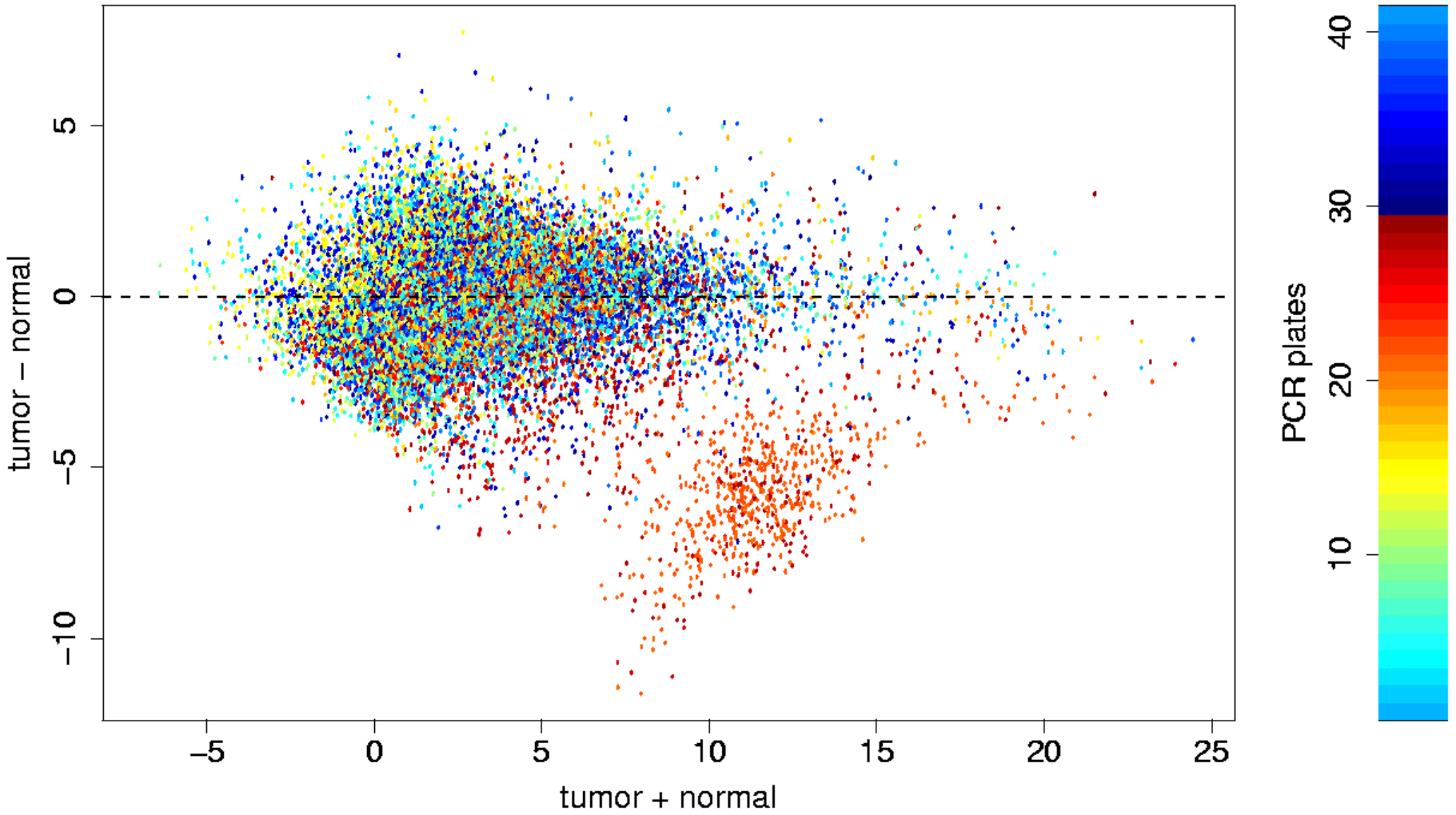
- advantages of variance-stabilizing data-transformation: generally better applicability of statistical methods (hypothesis testing, ANOVA, clustering, classification…)
- R package vsn

Quality control: diagnostic plots and artifacts

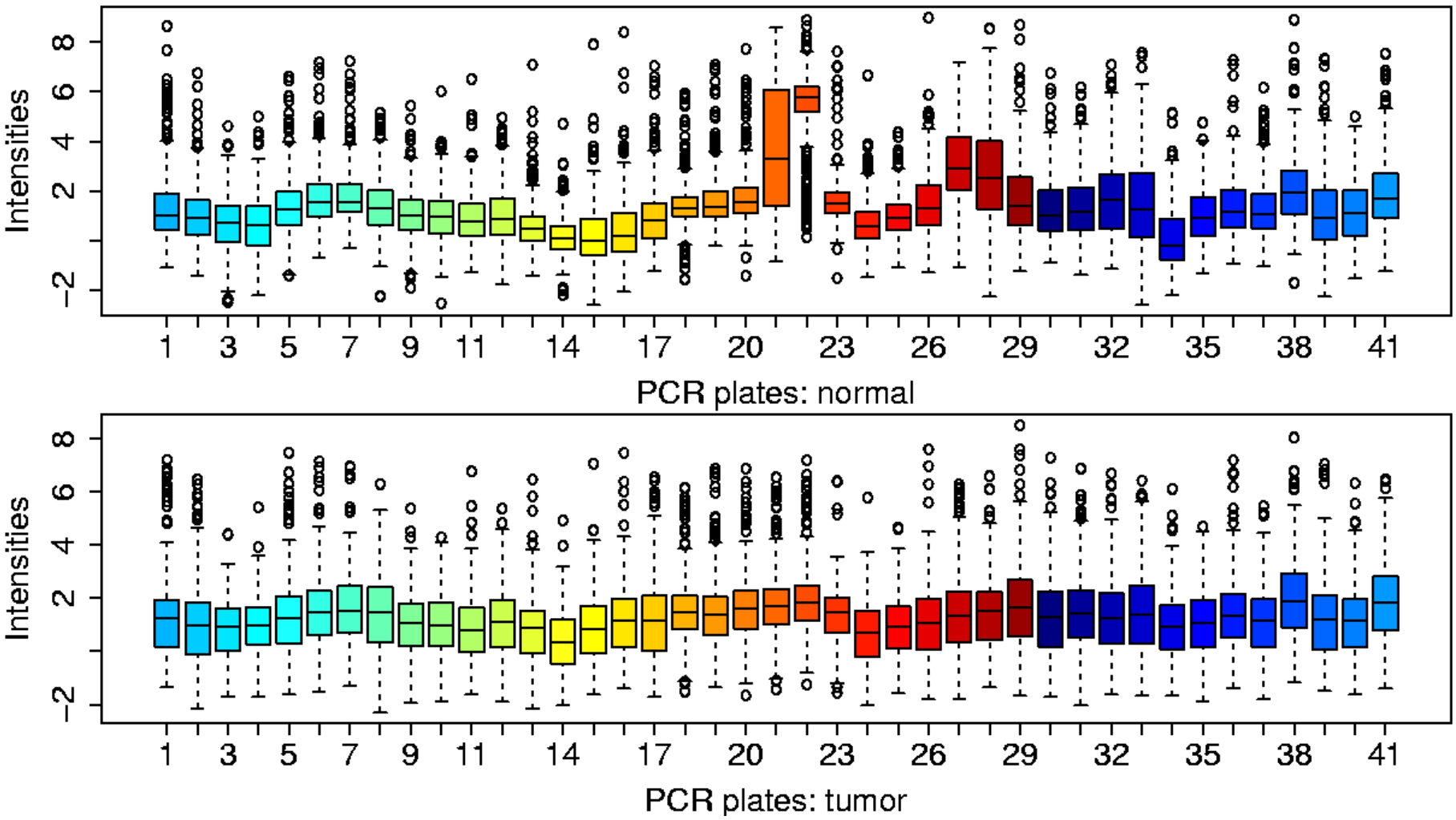
PCR plates


Scatterplot, colored by PCR-plate
Two RZPD Unigene II filters (cDNA nylon membranes)
1634tu1_203p1
1634no1_202p1
2
1.5
1
0.5
0
−0.5
−0.5
0
0.5
1
1.5
2

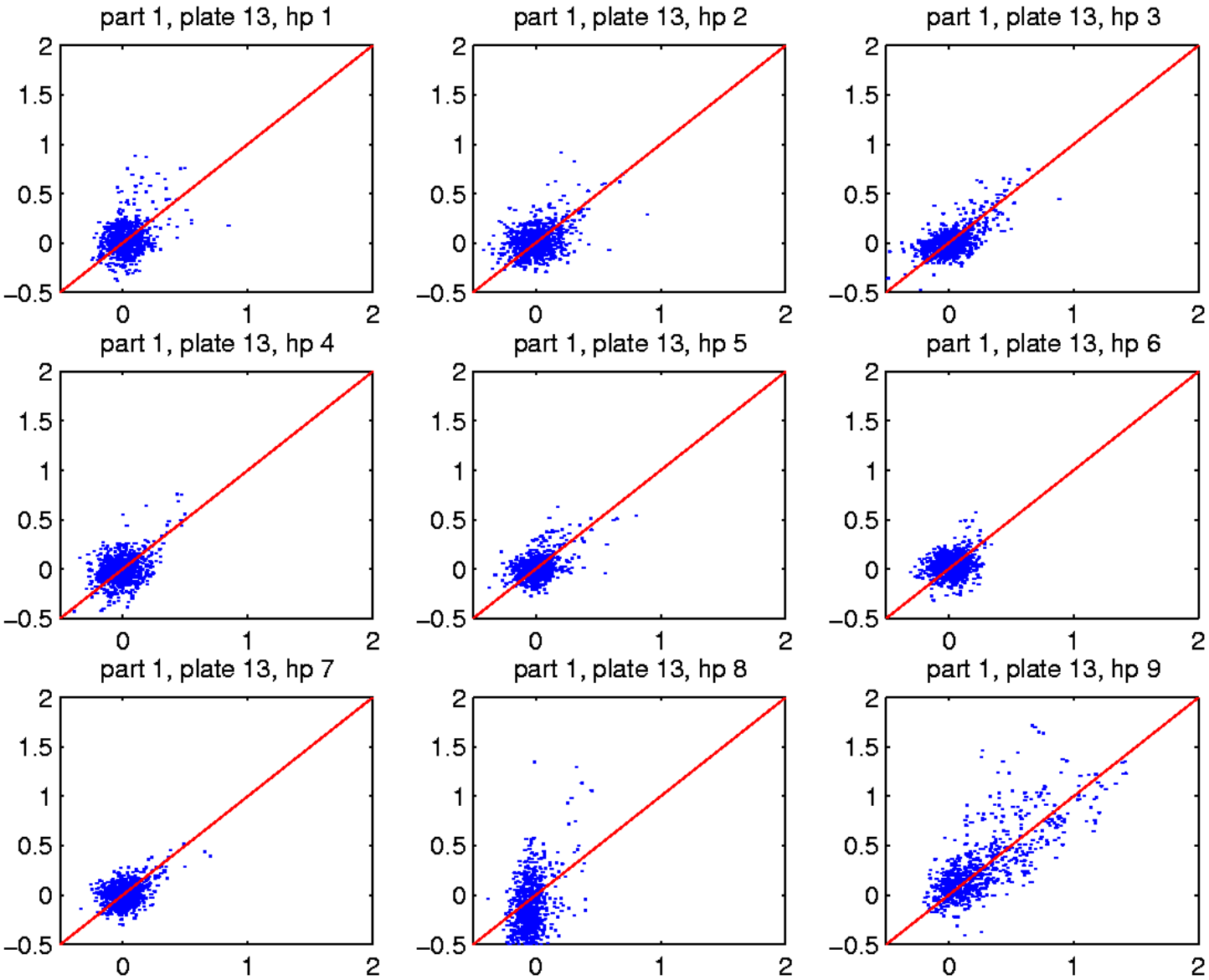
PCR plates


tumor – normal
5
0
−5
−10
−5
0
5
10
15
20
25
tumor + normal
PCR plates
40
30
20
10

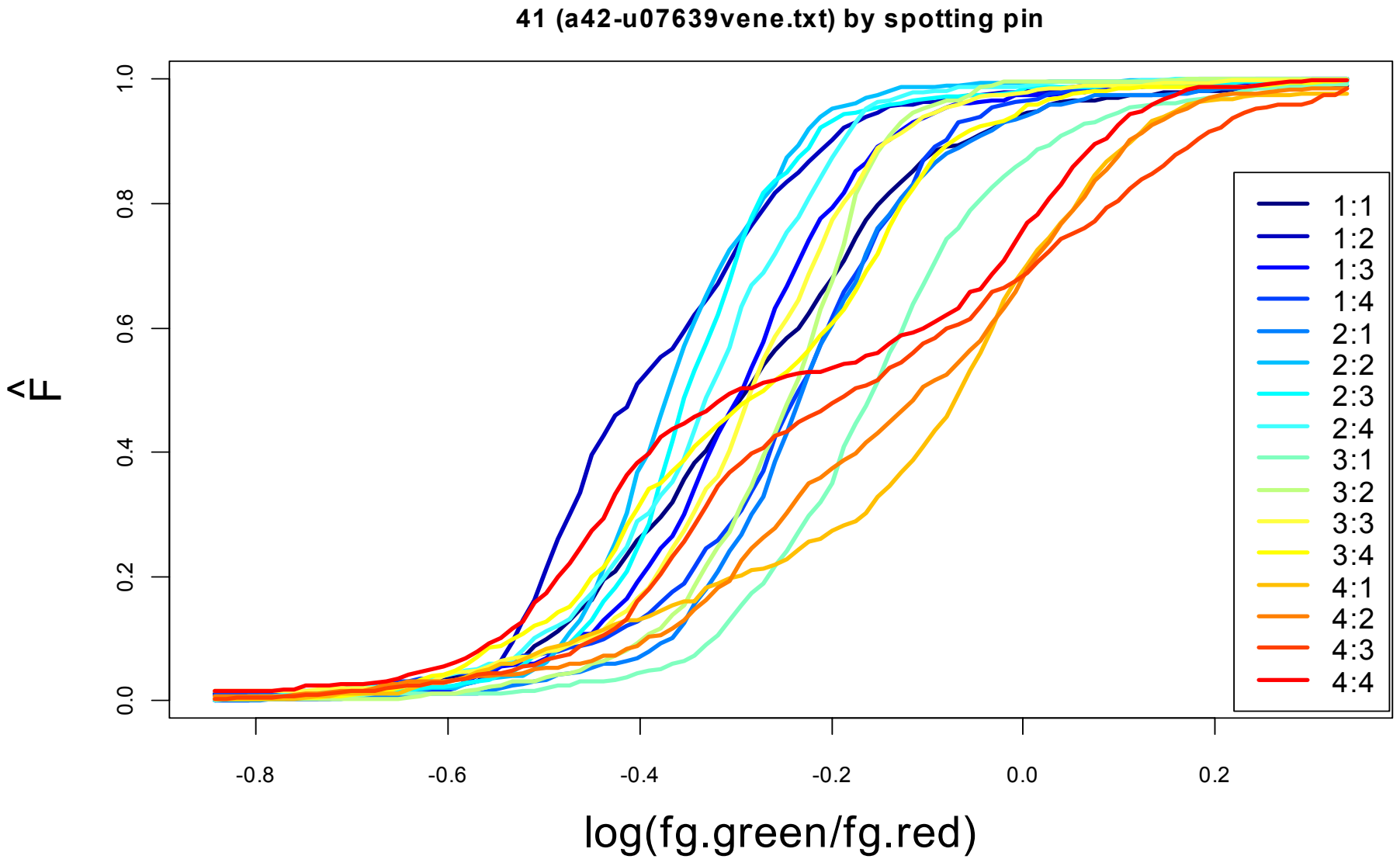
PCR plates: boxplots



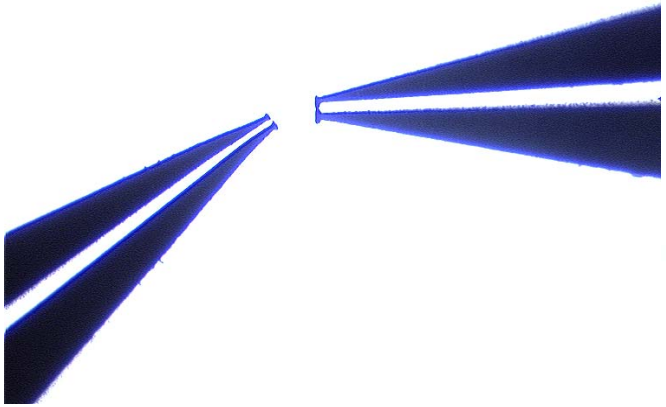
array batches



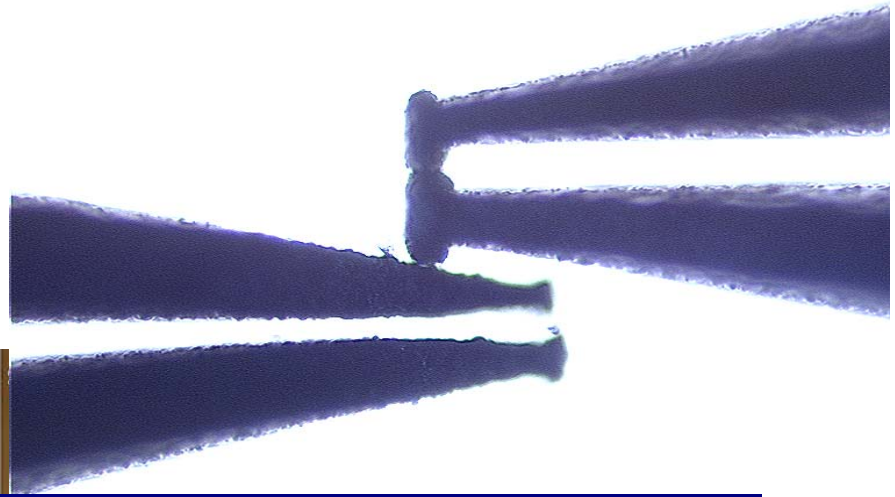
print-tip effects

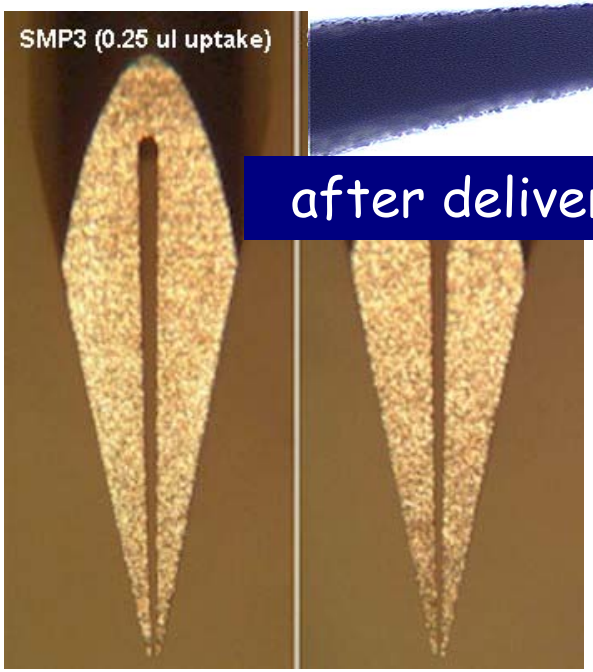


spotting pin quality decline

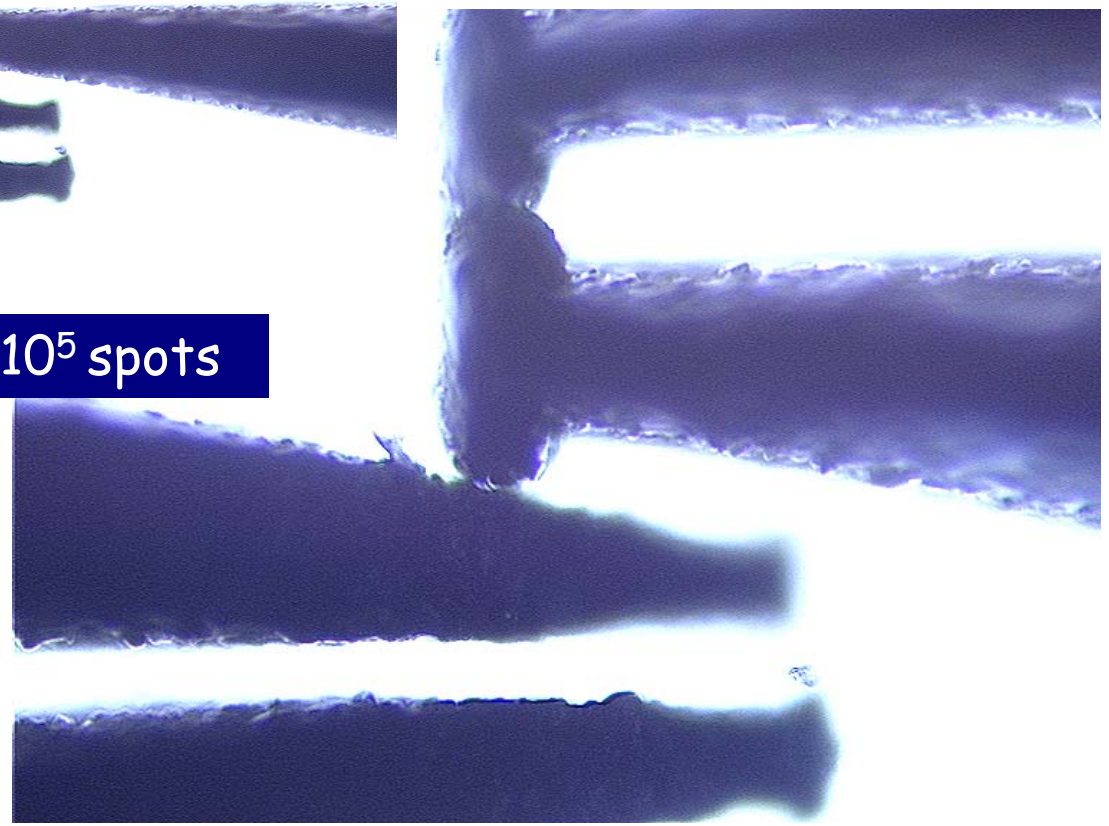


after delivery of $5x10^5$ spots



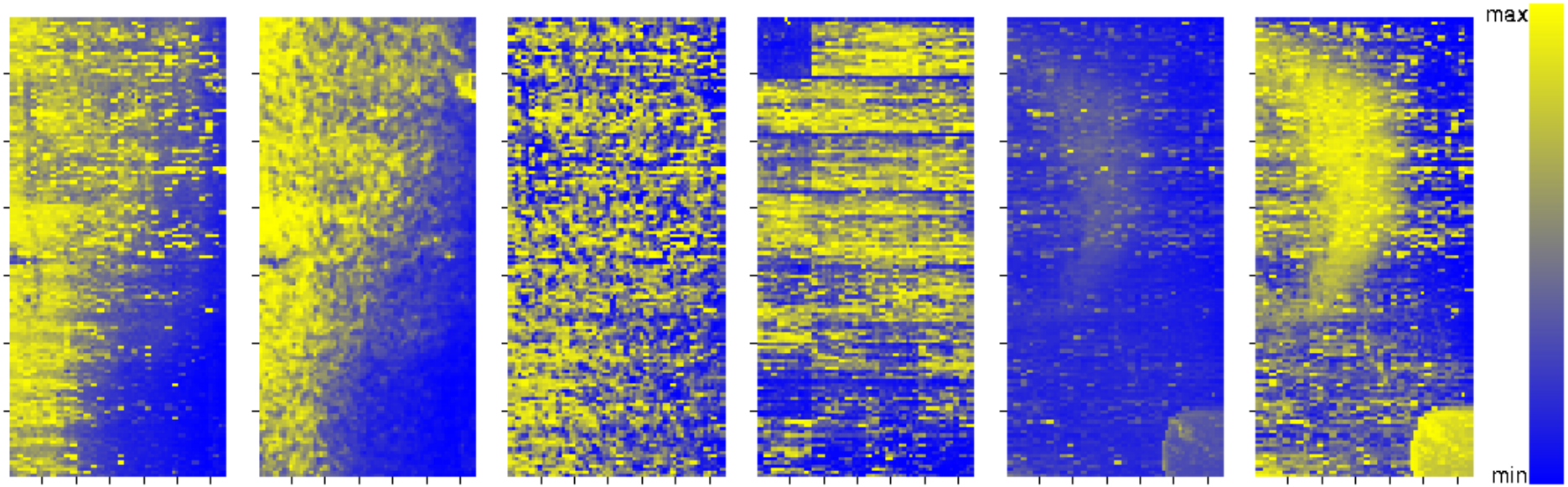


after delivery of $3x10^5$ spots



H. Sueltmann DKFZ/MGA

spatial effects



R Rb R-Rb
color scale by rank

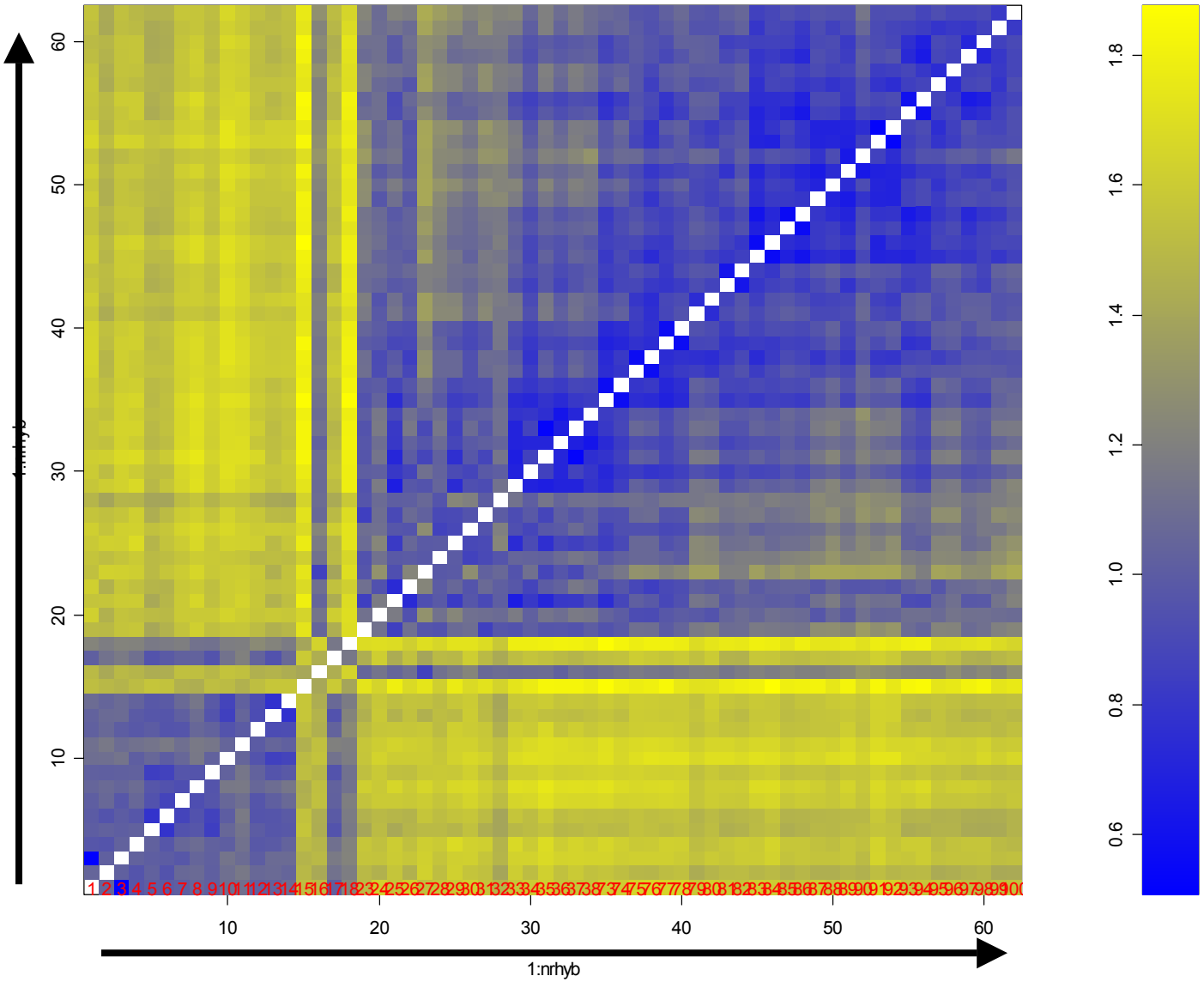
another array: print-tip

color scale ~ log(G)

color scale ~ rank(G)

spotted cDNA arrays, Stanford-type

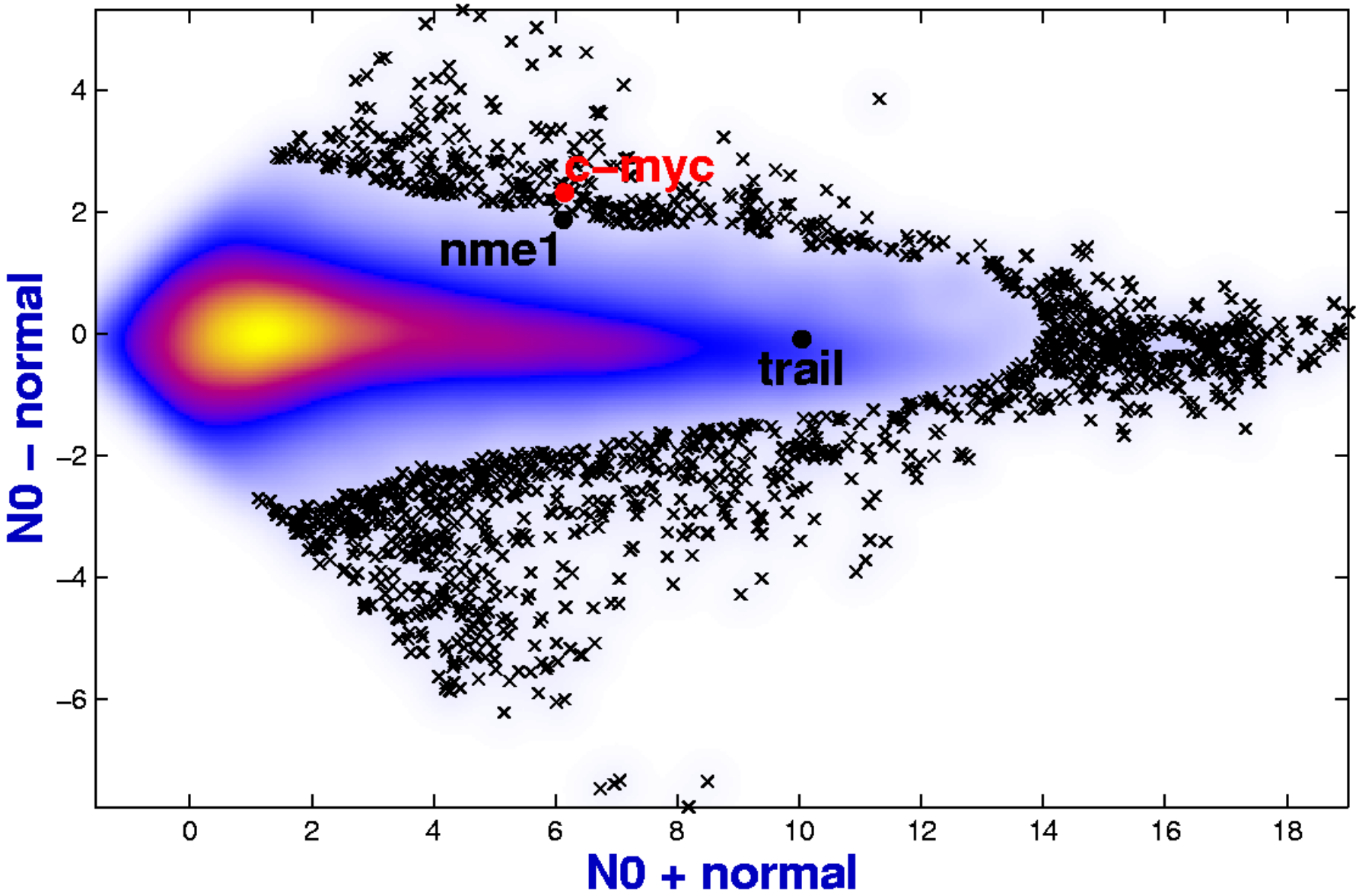
Batches: array to array differences $d_{ij} = mad_k(h_{ik} - h_{jk})$



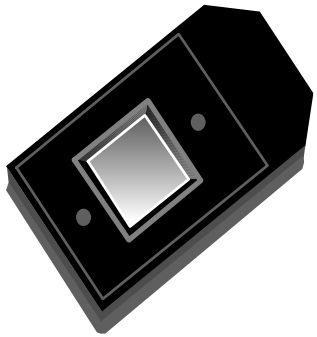
arrays i=1…63; roughly sorted by time

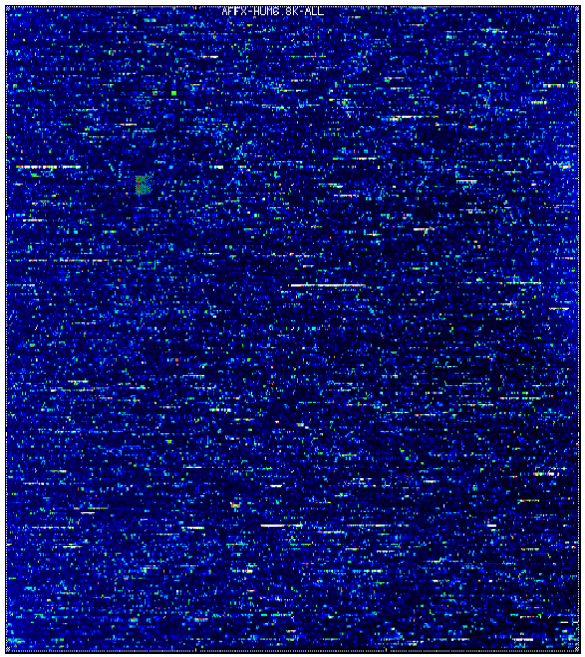
Density representation of the scatterplot

(76,000 clones, RZPD Unigene-II filters)



Oligonucleotide chips





Affymetrix files

Main software from Affymetrix:

MAS - MicroArray Suite.

DAT file: Image file, ~10^7 pixels, ~50 MB.

CEL file: probe intensities, ~400000 numbers

CDF file: Chip Description File. Describes which probes go in which probe sets (genes, gene fragments, ESTs).

Image analysis

DAT image files → CEL files

Each probe cell: 10x10 pixels.

Gridding: estimate location of probe cell centers.

Signal:

- Remove outer 36 pixels → 8x8 pixels.
- The probe cell signal, PM or MM, is the 75th percentile of the 8x8 pixel values.

Background: Average of the lowest 2% probe cells is taken as the background value and subtracted.

Compute also quality values.

Data and notation

PM_{ijg} , MM_{ijg} = Intensity for perfect match and mismatch probe j for gene g in chip i.

$i = 1,\ldots, n$ one to hundreds of chips

$j = 1,\ldots, J$ usually 16 or 20 probe pairs

$g = 1,\ldots, G$ 8…20,000 probe sets.

Tasks:

calibrate (normalize) the measurements from different chips (samples)

summarize for each probe set the probe level data, i.e., 20 PM and MM pairs, into a single expression measure.

compare between chips (samples) for detecting differential expression.

expression measures: MAS 4.0

Affymetrix GeneChip MAS 4.0 software uses AvDiff, a trimmed mean:

$$AvDiff = \frac{1}{\#J} \sum_{j \in J} (PM_j - MM_j)$$

- sort $d_j = PM_j - MM_j$
- exclude highest and lowest value
- J := those pairs within 3 standard deviations of the average

Expression measures
MAS 5.0

Instead of MM, use "repaired" version CT

CT= MM *if MM<PM*

= PM / "typical log-ratio" *if MM>=PM*

"Signal" =

Tukey.Biweight (log(PM-CT))

(... ≈median)

Tukey Biweight: B(x) = (1 – (x/c)^2)^2 if |x|<c, 0 otherwise

Expression measures: Li & Wong

dChip fits a model for each gene

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \propto N(0, \sigma^2)$$

where

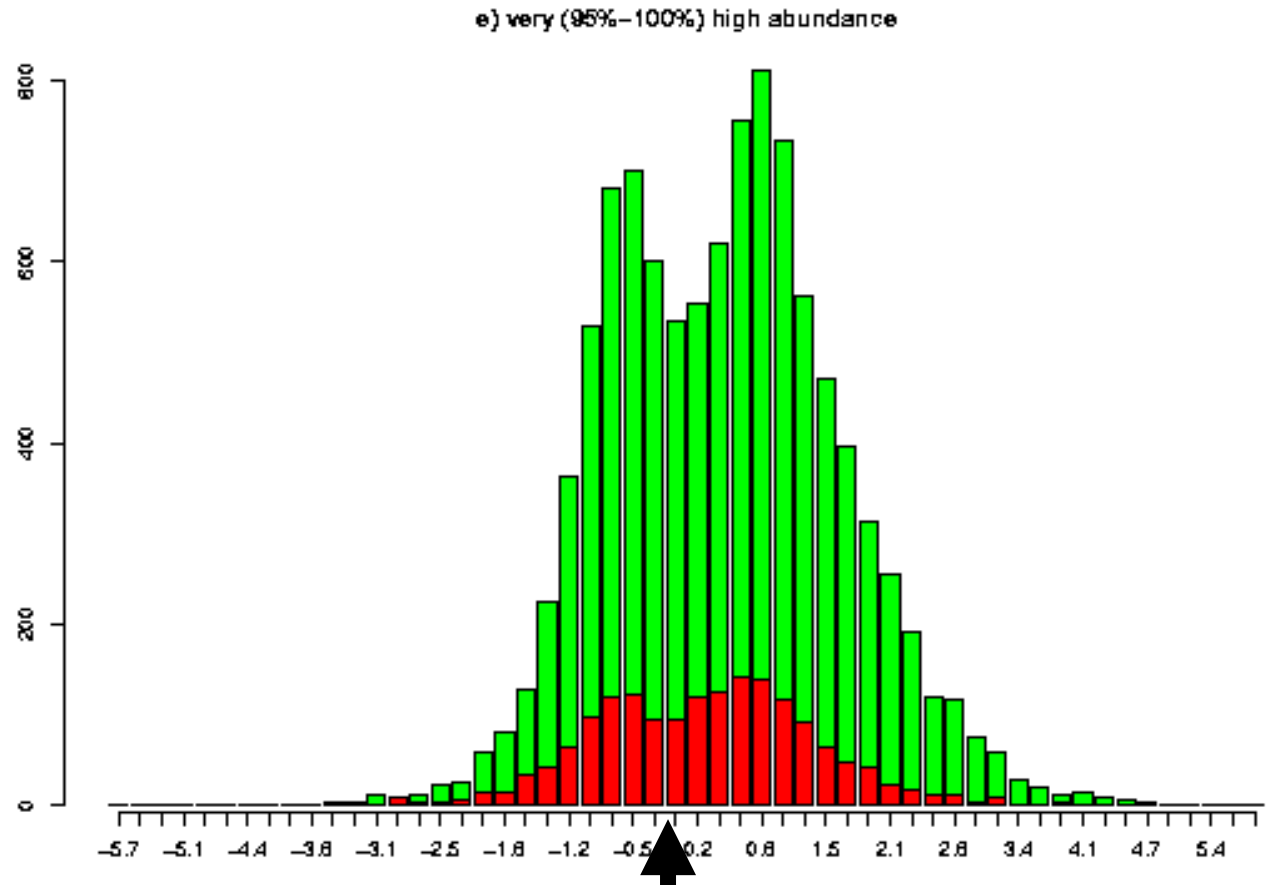
- θ_i: expression index for gene i
- ϕ_j: probe sensitivity

Maximum likelihood estimate of MBEI is used as expression measure of the gene in chip *i*.

Need at least 10 or 20 chips.

Current version works with PMs only.

Affymetrix: $I_{PM} = I_{MM} + I_{specific}$?



0

log(PM/MM)

From: R. Irizarry et al.,
Biostatistics 2002

Expression measures
RMA: Irizarry et al. (2002)

- Estimate one global background value b=mode(MM). No probe-specific background!
- Assume: $PM = s_{true} + b$

  Estimate $s \geq 0$ from PM and b as a conditional expectation $E[s_{true}|PM, b]$.
- Use $\log_2(s)$.
- Nonparametric nonlinear calibration ('quantile normalization') across a set of chips.

Robust expression measures
RMA: Irizarry et al. (2002)

AvDiff-like

$$RMA = \frac{1}{|A|}\sum_{j \in A} \log_2(PM_j - BG_j)$$

with A a set of "suitable" pairs.

Li-Wong-like: additive model

$$\log_2(PM_{ij} - BG) = a_i + b_j + \varepsilon_{ij}$$

Estimate RMA = a_i for chip *i* using robust method median polish (successively remove row and column medians, accumulate terms, until convergence). Works with d>=2

Software for pre-processing of Affymetrix data

- Bioconductor R package affy.
- Background estimation.
- Probe-level normalization.
- Expression measures
- Two main functions: ReadAffy, expresso.
- Can use vsn as a normalization method for expresso.

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/