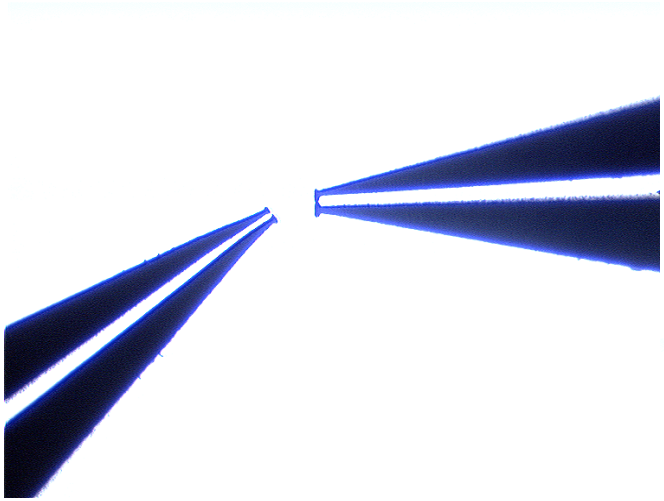
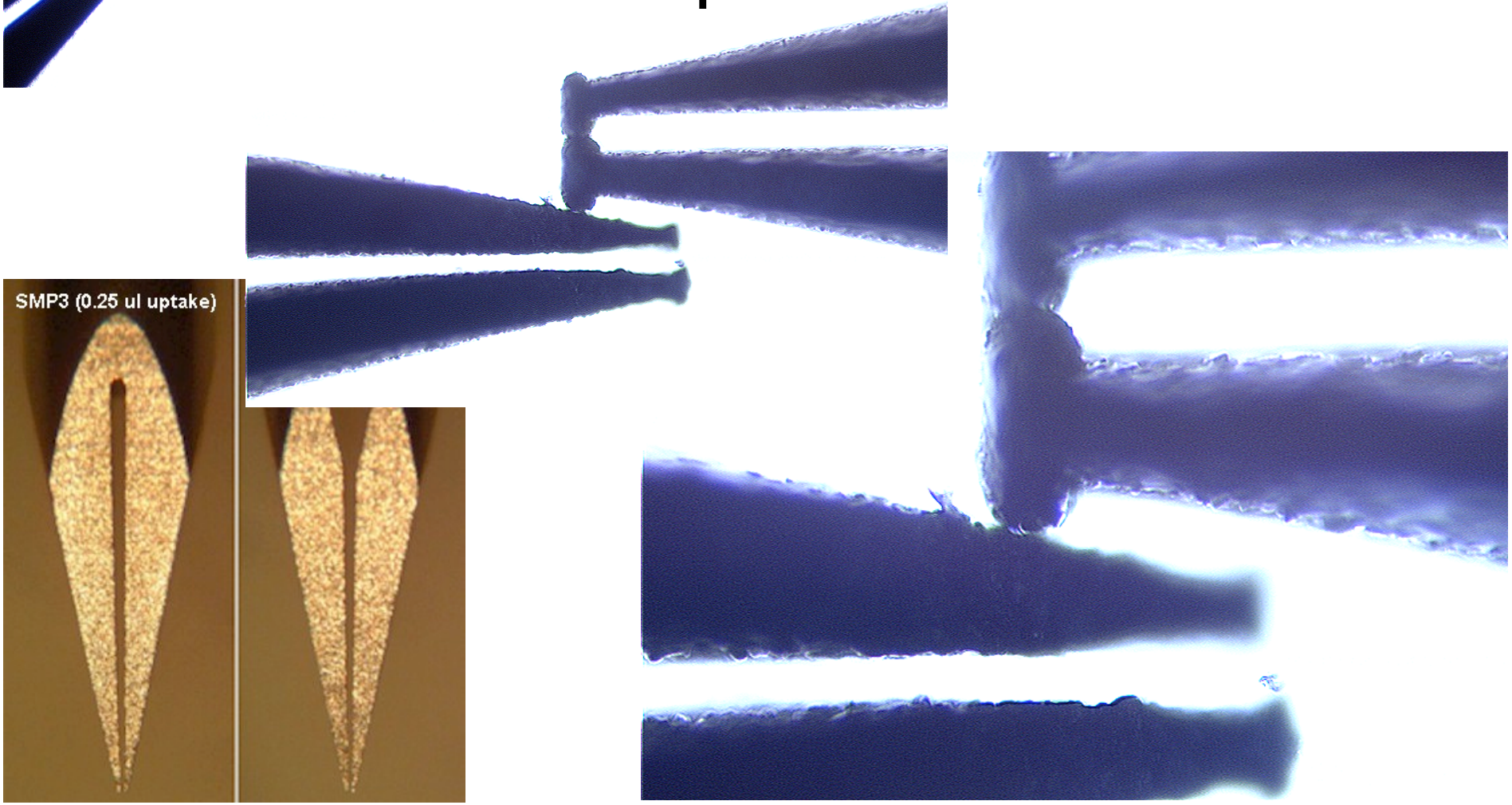


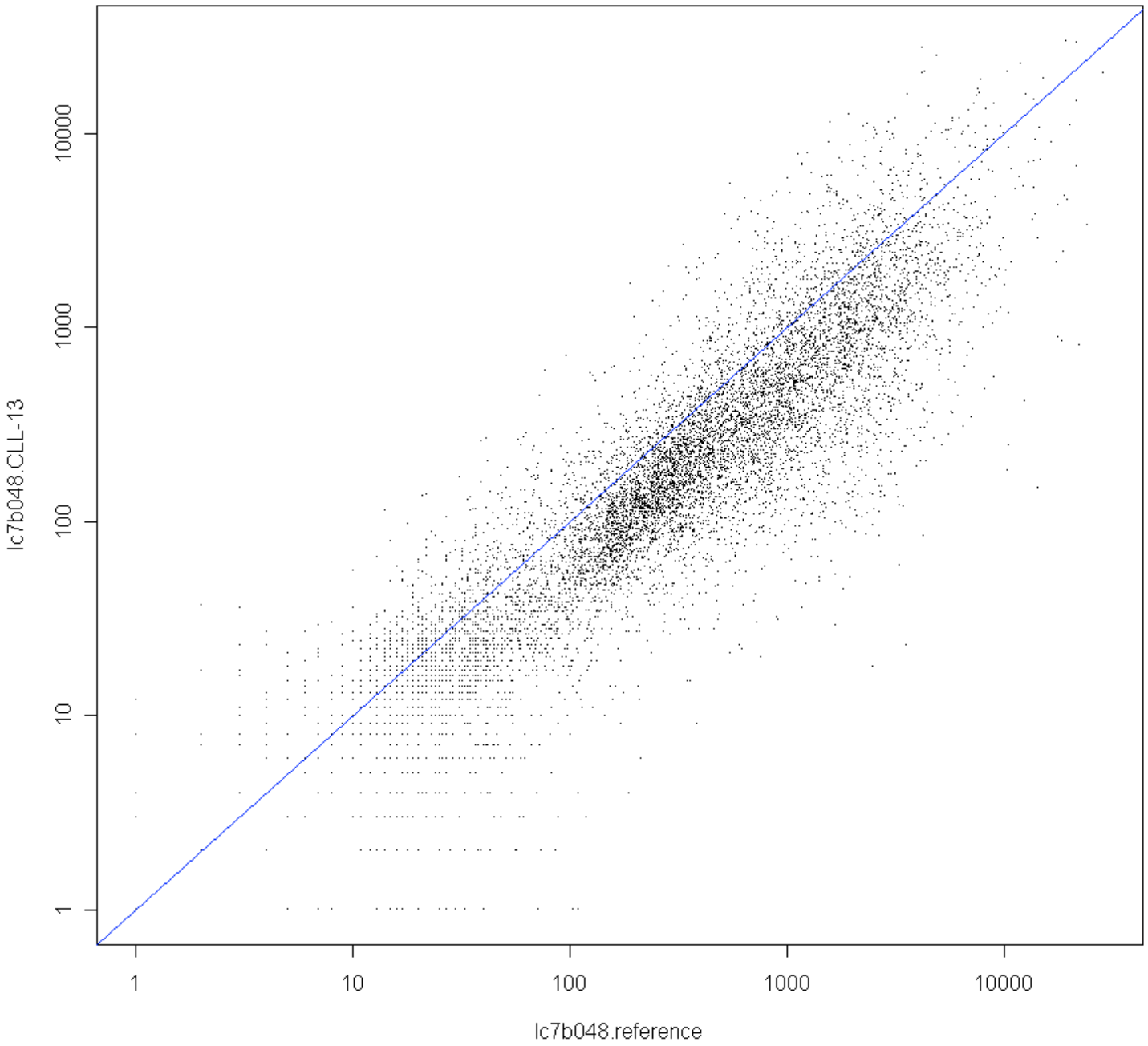
Microarray normalization and error models

Wolfgang Huber
European Bioinformatics Institute



H. Sueltmann DKFZ/MGA

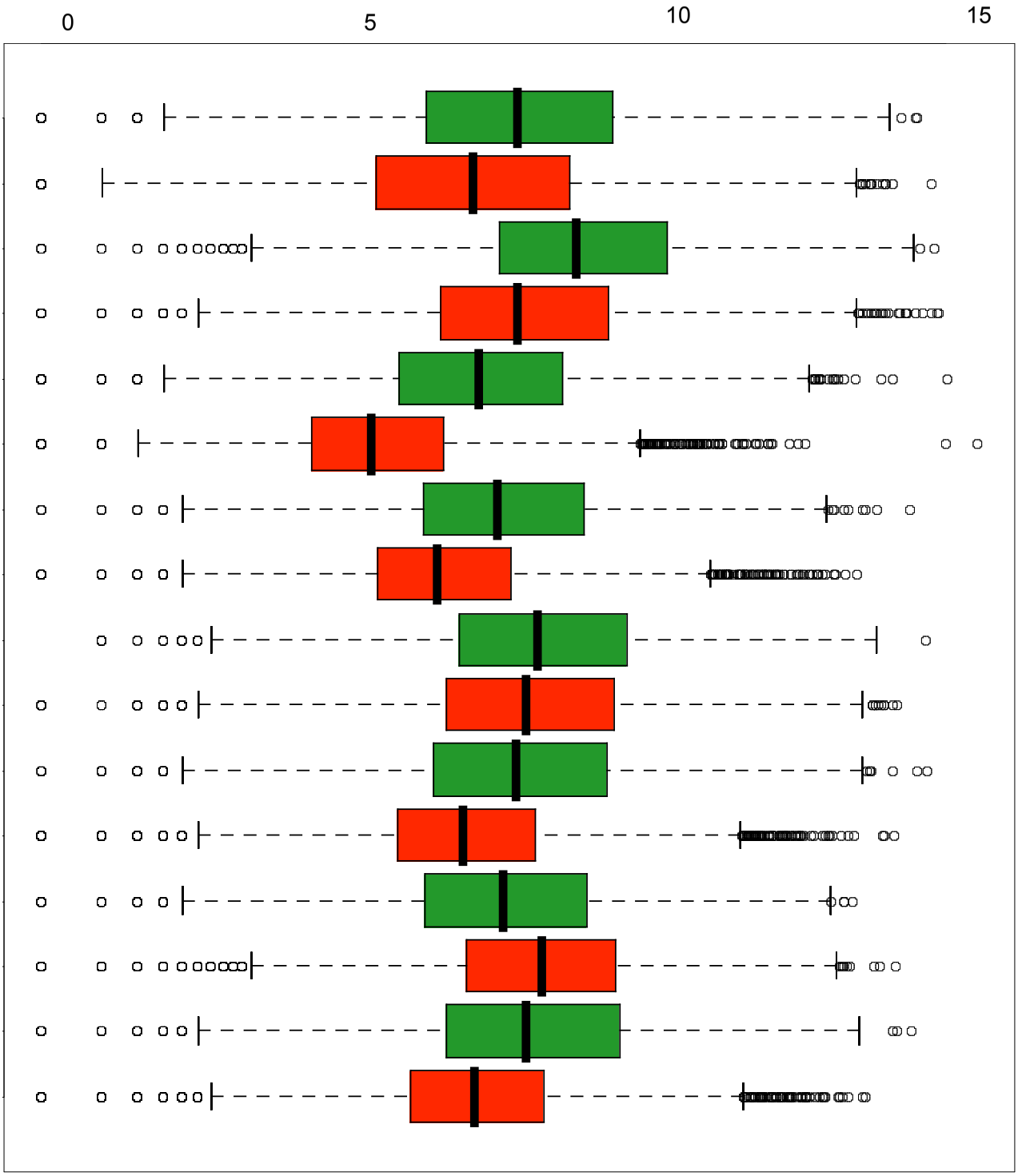
Why do you need normalisation?



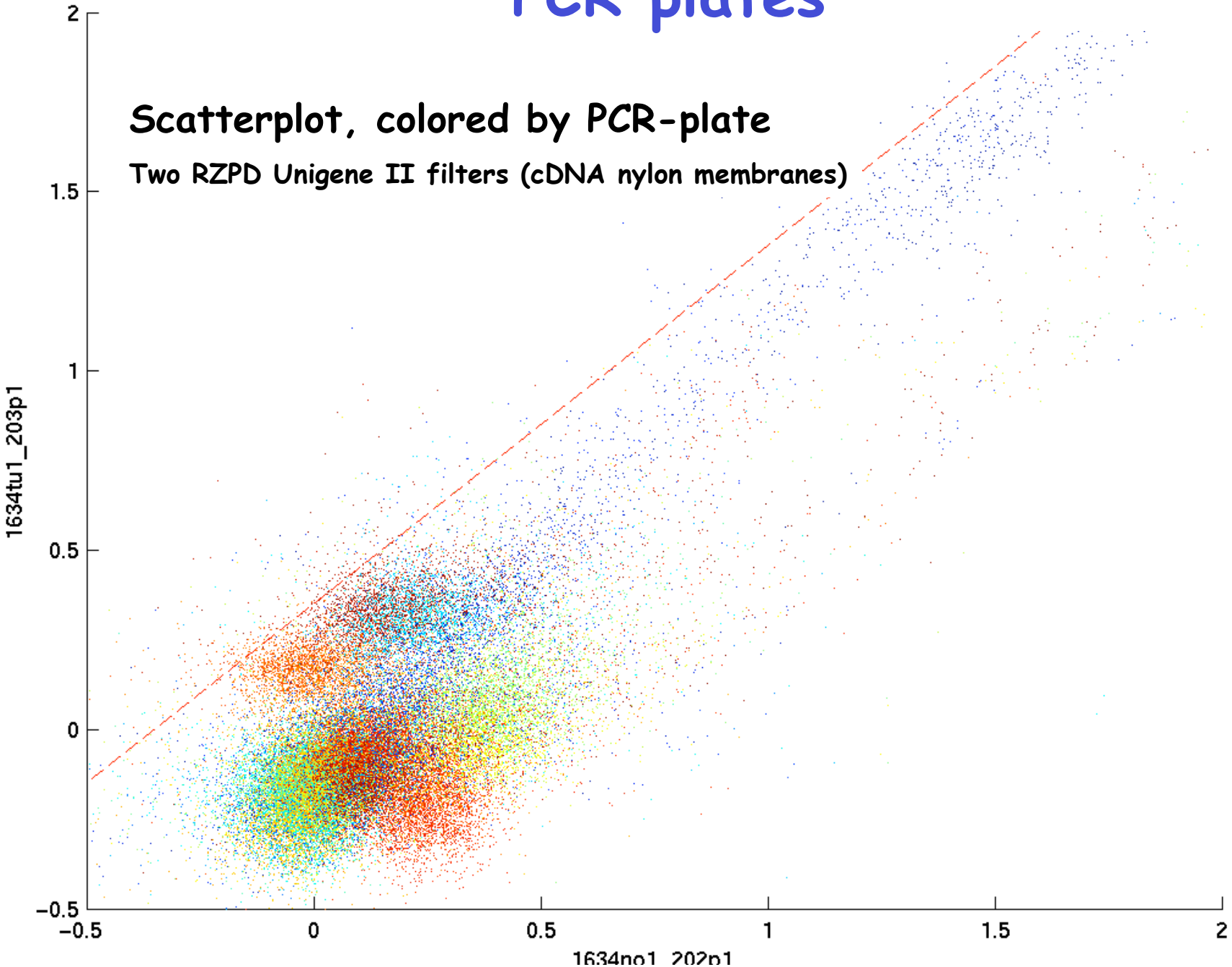
From: lymphoma dataset

vsn package

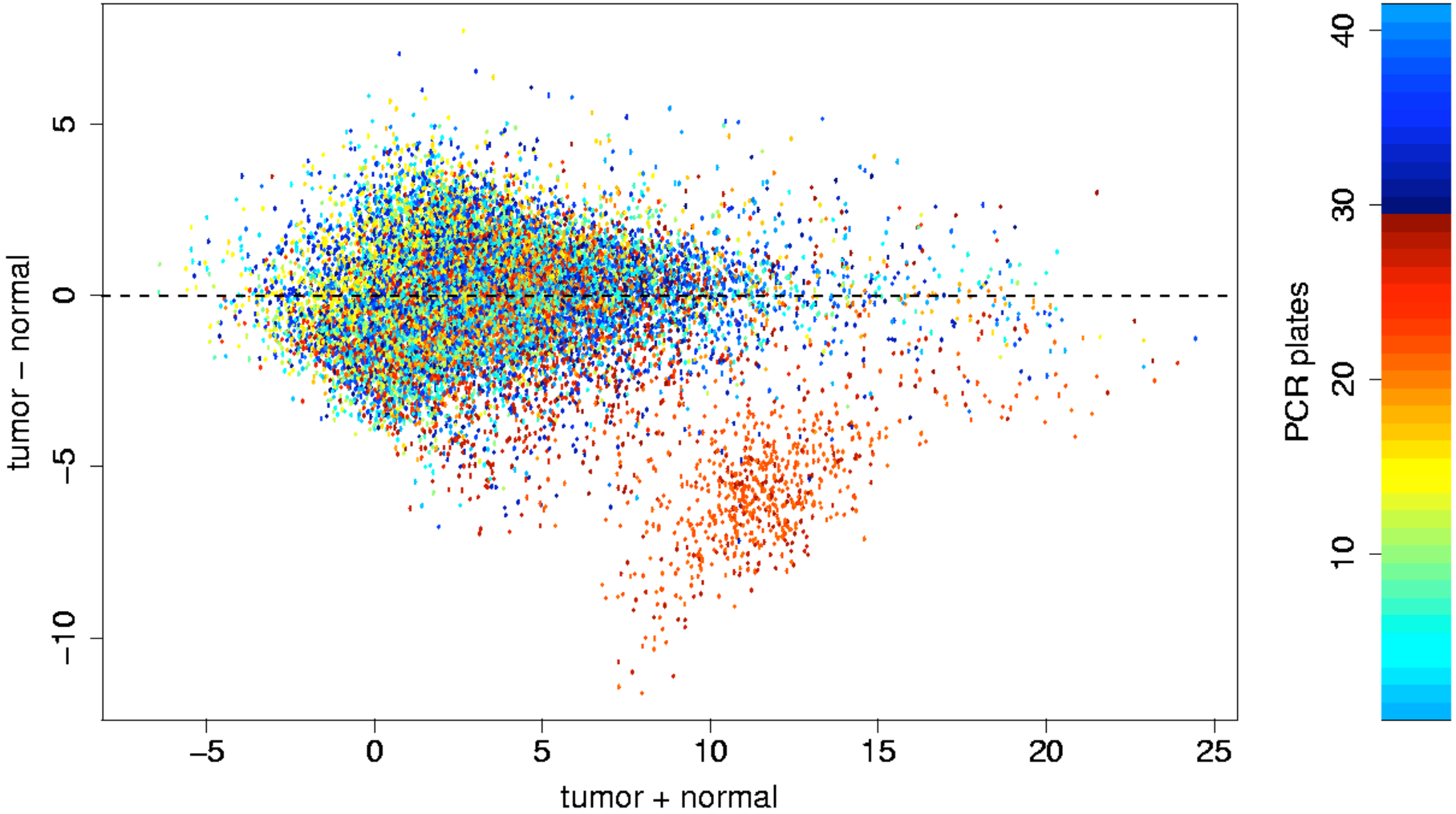
Alizadeh et al., Nature 2000


0
5
10
15

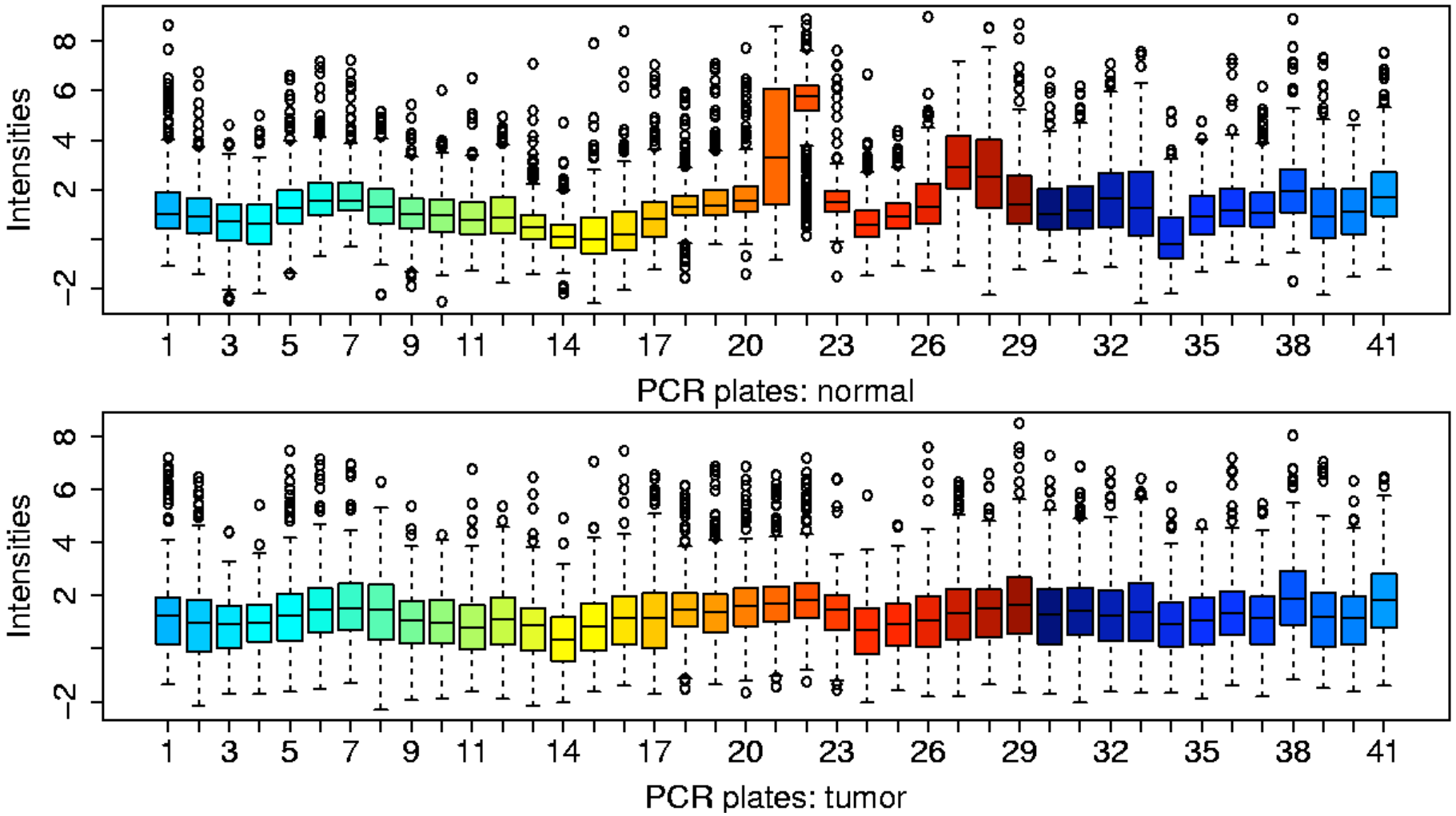
PCR plates


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

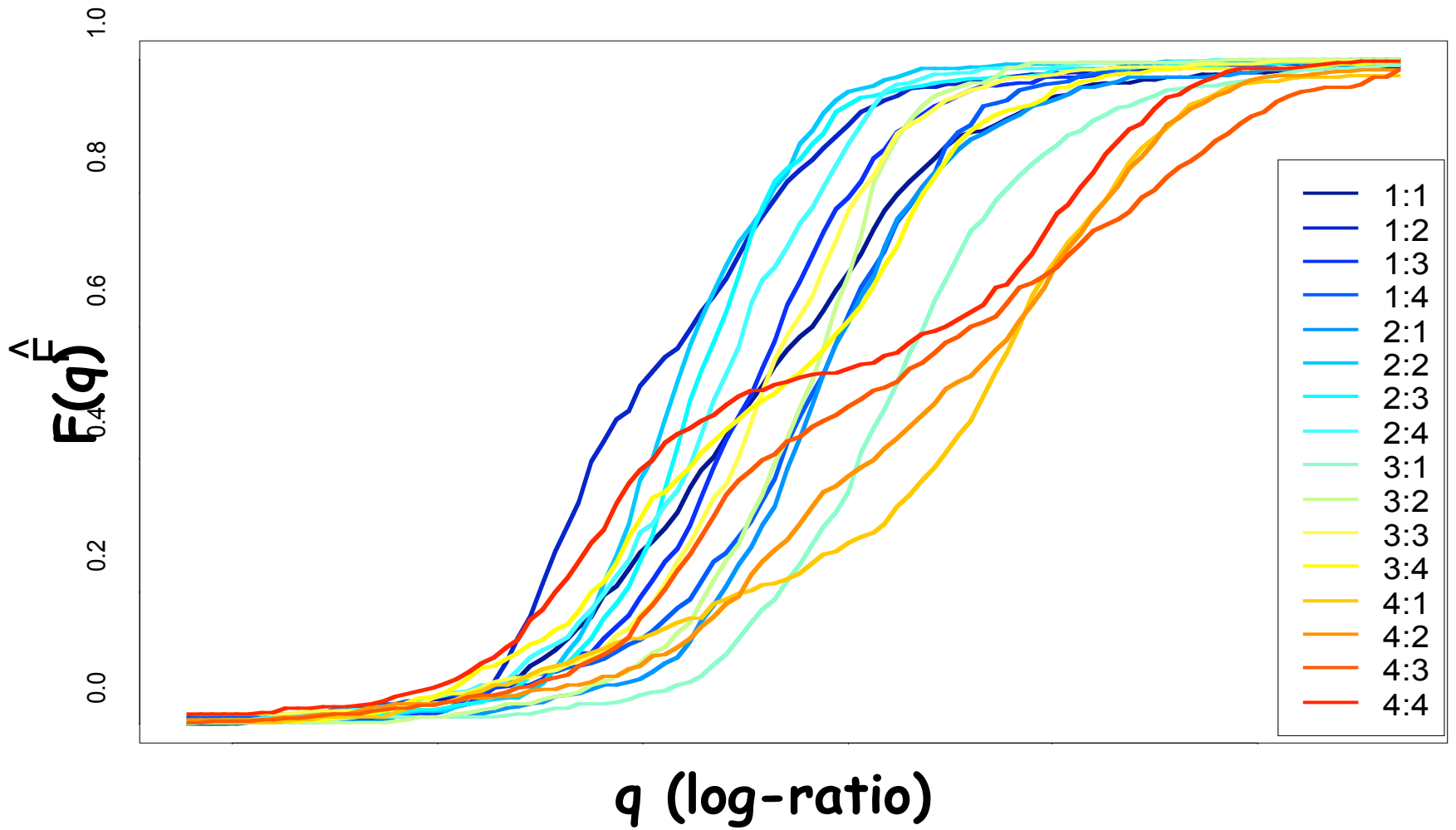
PCR plates


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

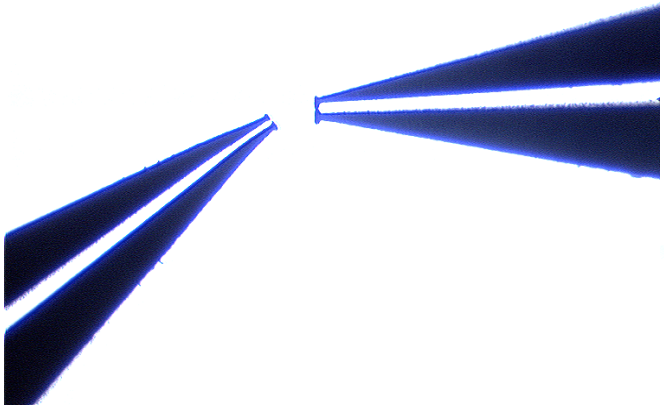
PCR plates: boxplots



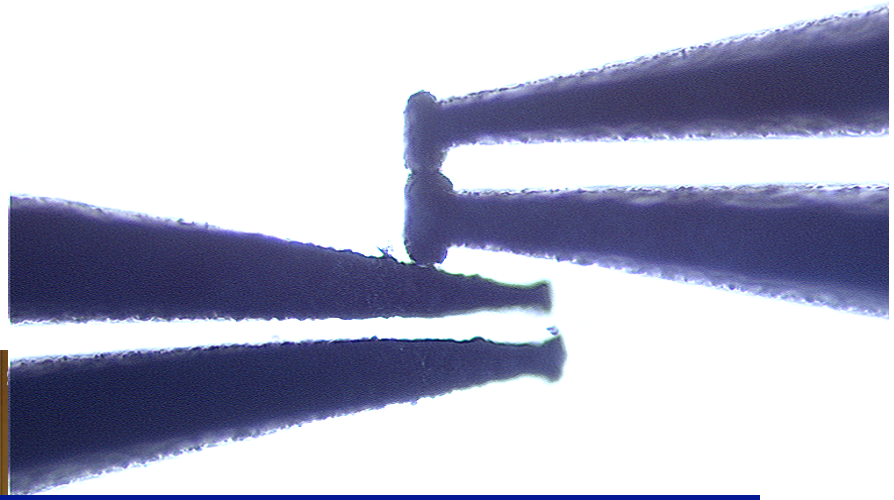
print-tip effects

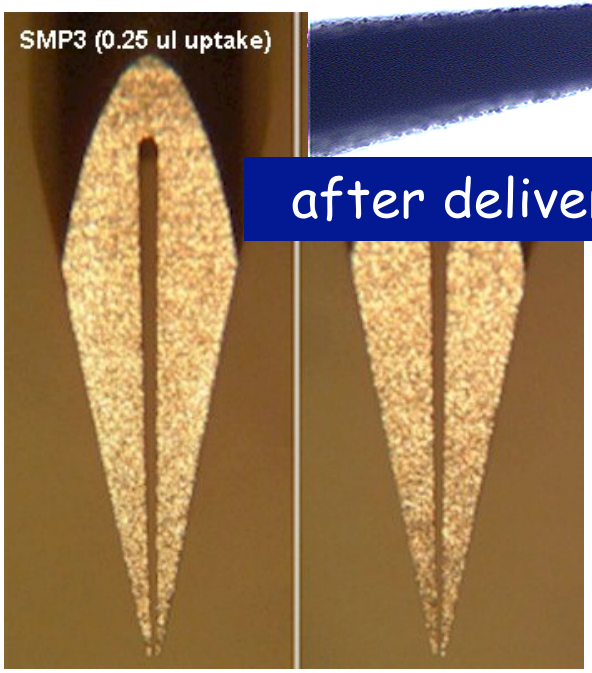


spotting pin quality decline

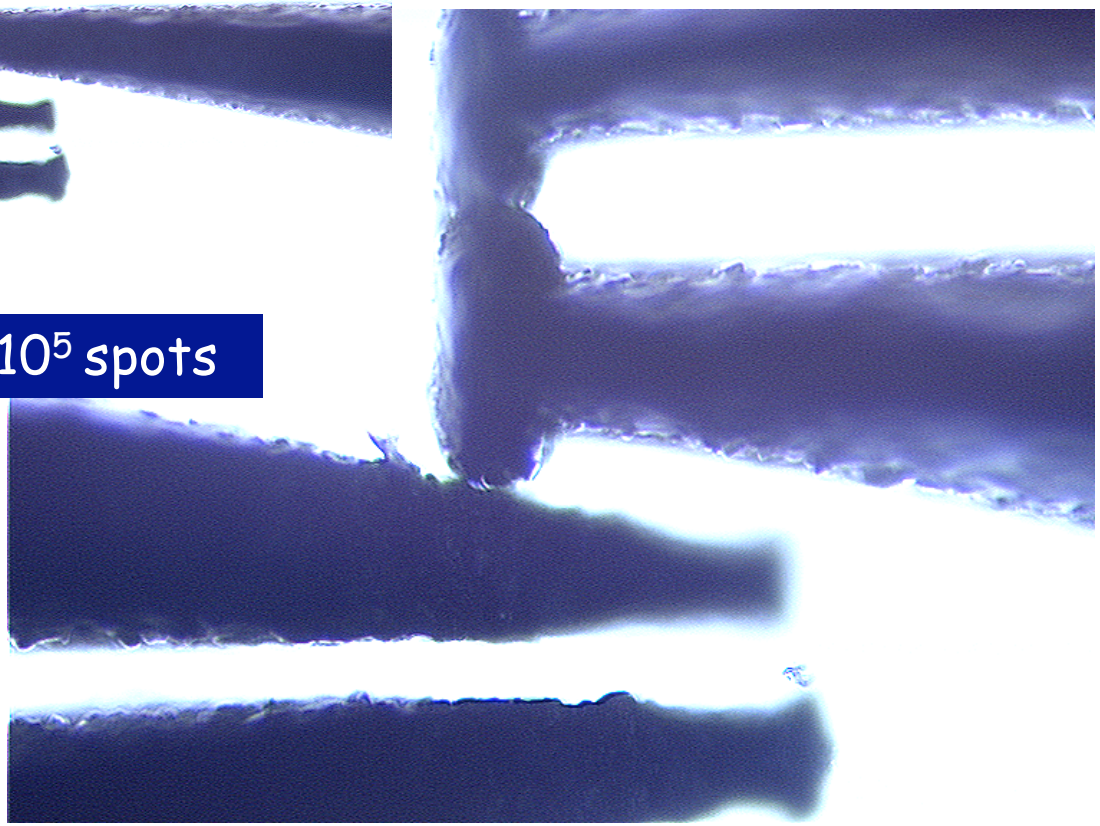


after delivery of $5x10^5$ spots



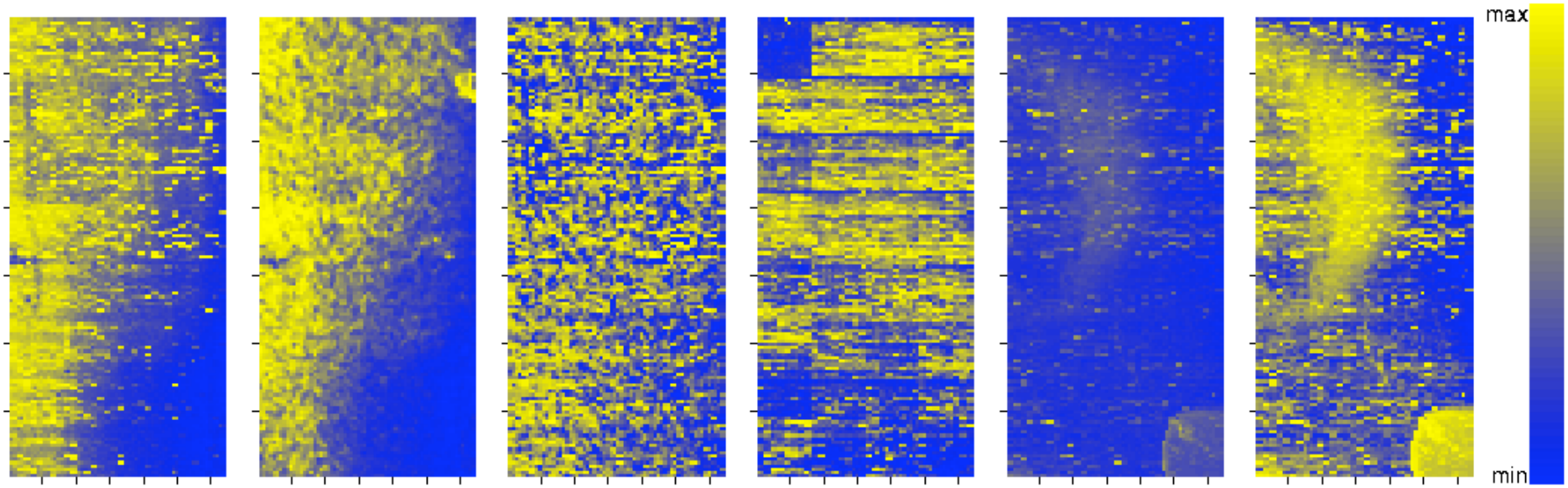


after delivery of $3x10^5$ spots



H. Sueltmann DKFZ/MGA

spatial effects



R Rb R-Rb
color scale by rank

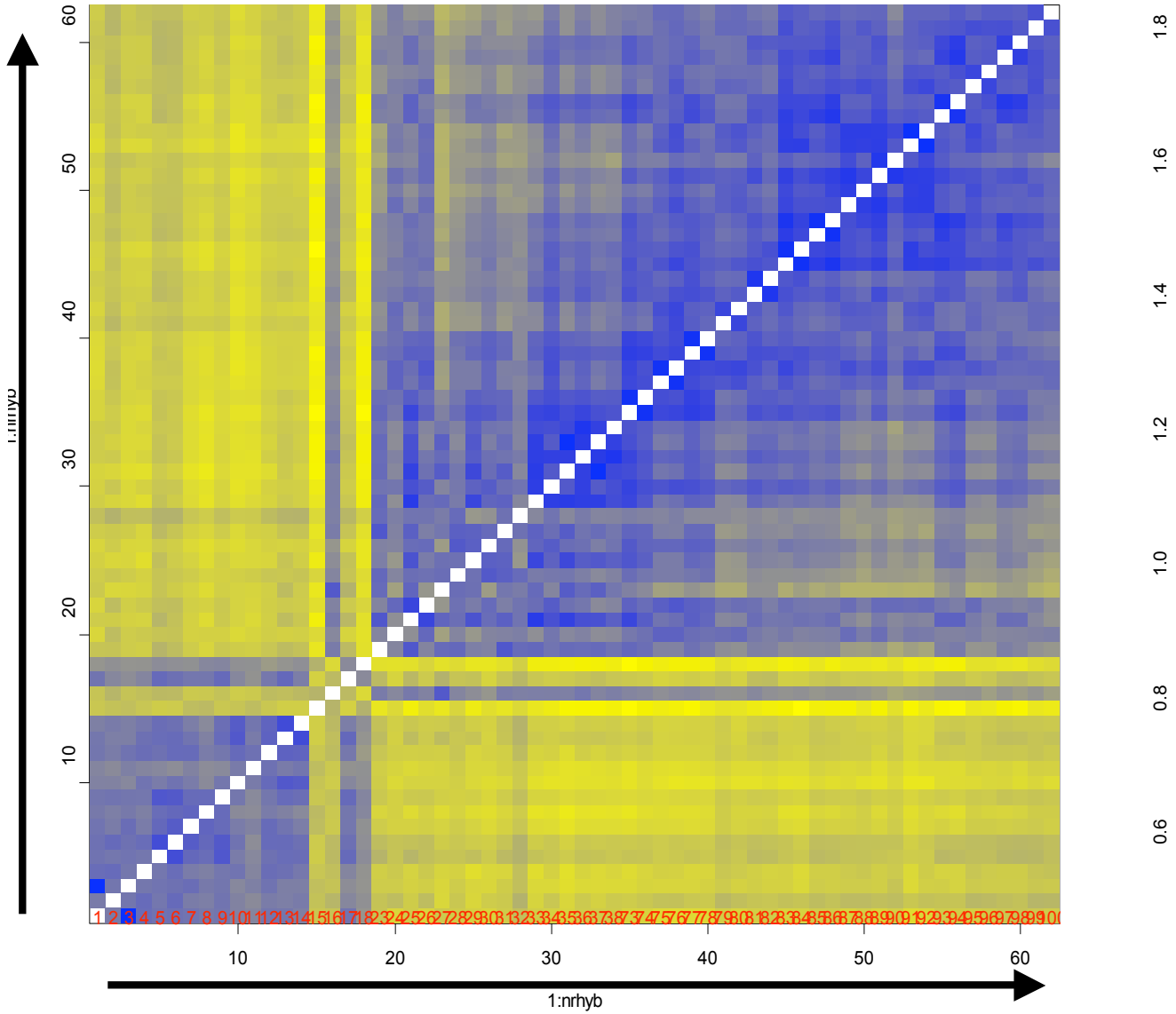
another array: print-tip

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

A complex measurement process lies between mRNA concentrations and intensities

o ti
cor

o R
deg

o a
effi

o r
tra
efficiency

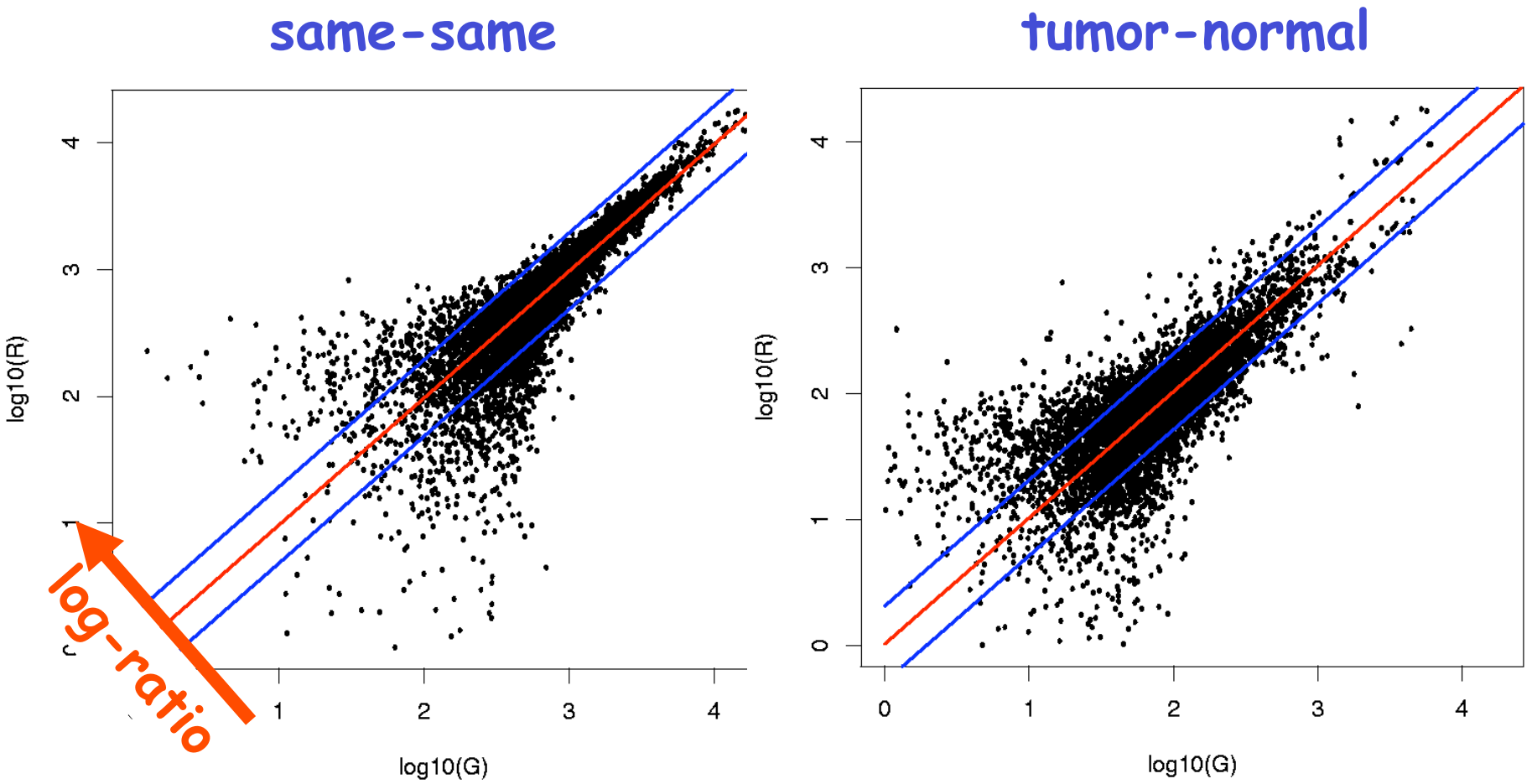
o hybridization efficiency and specificity

o other array manufacturing-related issues

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

Why do you need statistics?

Which genes are differentially transcribed?



Statistics 101:

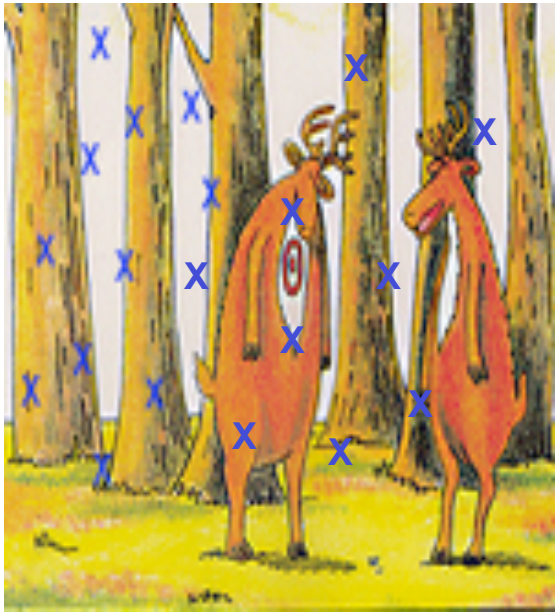
←bias accuracy→

variance→

← precision

Basic dogma of data analysis

Can always increase sensitivity on the cost of specificity,

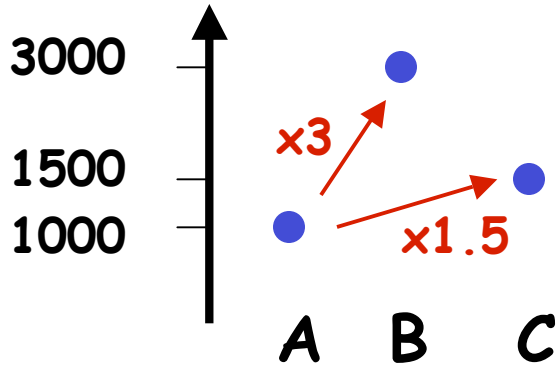


or vice versa,

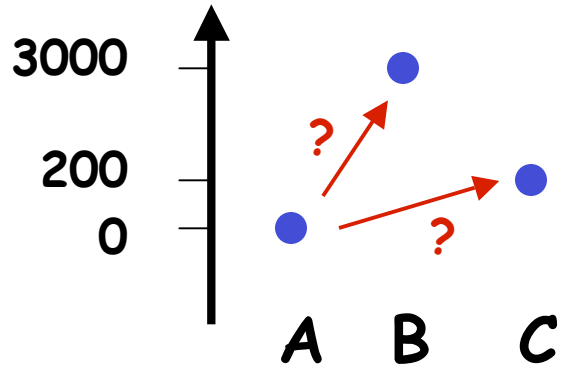
the art is to
- optimize both
- then find the best trade-off.

ratios and fold changes

Fold changes are useful to describe continuous changes in expression



But what if the gene is "off" (below detection limit) in one condition?



ratios and fold changes

The idea of the log-ratio (base 2)

0: no change

+1: up by factor of $2^1 = 2$

+2: up by factor of $2^2 = 4$

-1: down by factor of $2^{-1} = 1/2$

-2: down by factor of $2^{-2} = \frac{1}{4}$

A unit for measuring changes in expression: assumes that a change from 1000 to 2000 units has a similar biological meaning to one from 5000 to 10000.

What about a change from 0 to 500?

- conceptually
- noise, measurement precision

Questions

- How to compare microarray intensities with each other?
- How to address measurement uncertainty ("variance")?
- How to calibrate ("normalize") for biases between samples?

Sources of variation

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

probe purity and length distribution
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
- remain as "noise"

Calibration

Error model

Error models

describe the possible outcomes of a set of measurements

Outcomes depend on:

-true value of the measured quantity
(abundances of specific molecules in biological sample)

-measurement apparatus
(cascade of biochemical reactions, optical detection system with laser scanner or CCD camera)

Error models

Purpose:

1. **Data compression: summary statistic instead of full empirical distribution**
2. **Quality control**
3. **Statistical inference: appropriate parametric methods have better power than non-parametric (this has practical, financial, and ethical aspects)**

The two component model

measured intensity = offset + gain × true abundance

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

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

a_i per-sample offset

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

"additive noise"

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

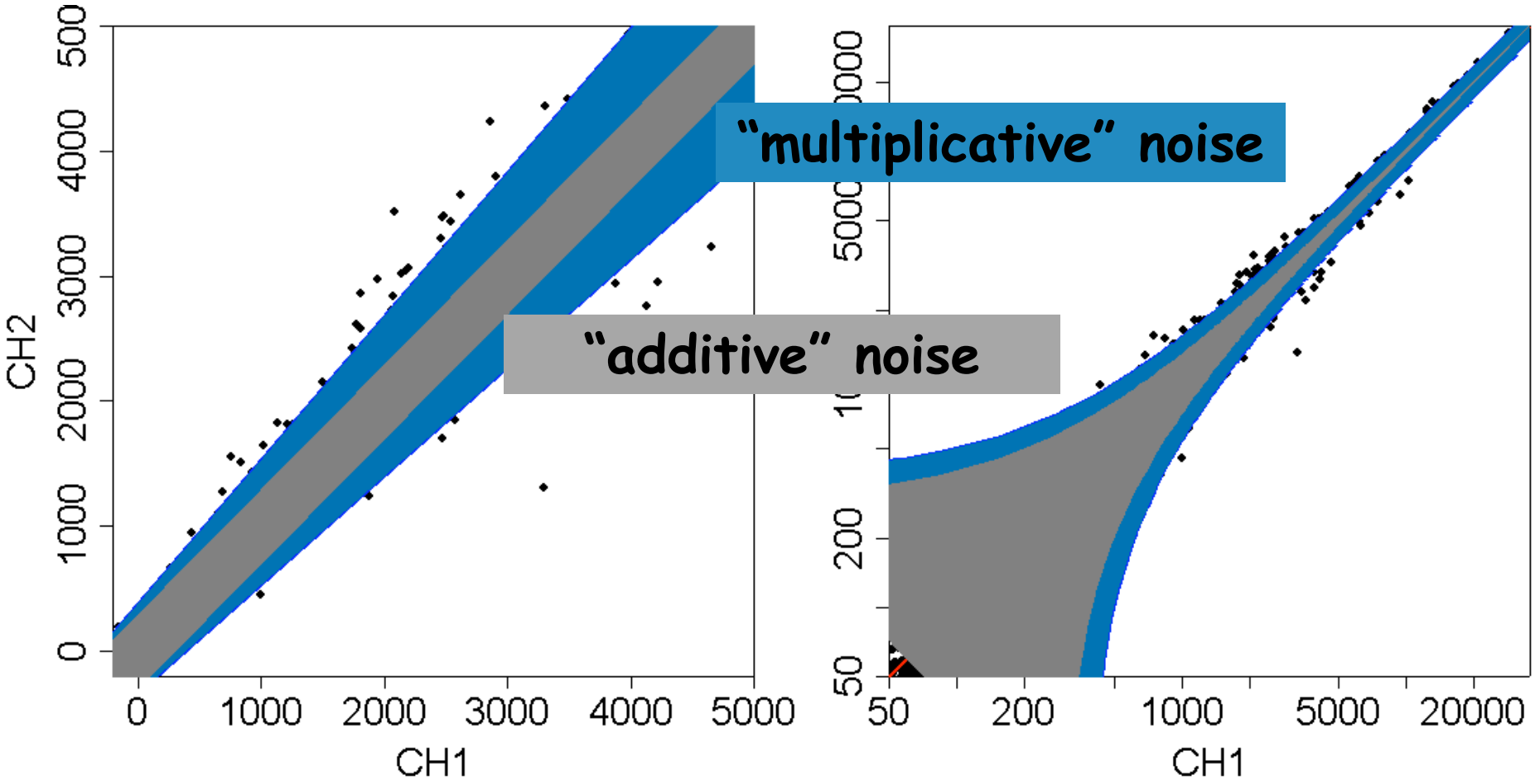
b_i per-sample normalization factor

b_k sequence-wise probe efficiency

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

"multiplicative noise"

The two-component model



raw scale

log scale

B. Durbin, D. Rocke, JCB 2001

Parameterization

$$y = a + \varepsilon + b \cdot x \cdot (1 + \eta)$$

$$y = a + \varepsilon + b \cdot x \cdot e^{\eta}$$

two practically equivalent forms ($\eta << 1$)

a systematic background	same for all probes (per array x color)	per array x color x print-tip group
ε random background	iid in whole experiment	iid per array
b systematic gain factor	per array x color	per array x color x print-tip group
η random gain fluctuations	iid in whole experiment	iid per array

Important issues for model fitting

Parameterization

variance vs bias

"Heteroskedasticity" (unequal variances)

⇒ weighted regression or variance stabilizing transformation

Outliers

⇒ use a robust method

Algorithm

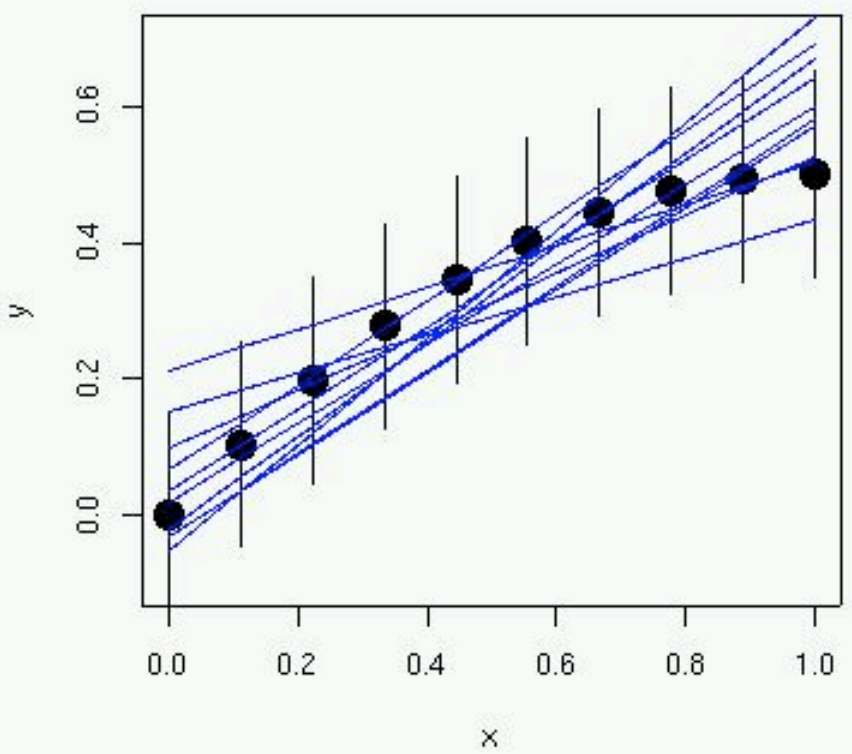
If likelihood is not quadratic, need non-linear optimization. Local minima / concavity of likelihood?

Models are never correct, but some are useful

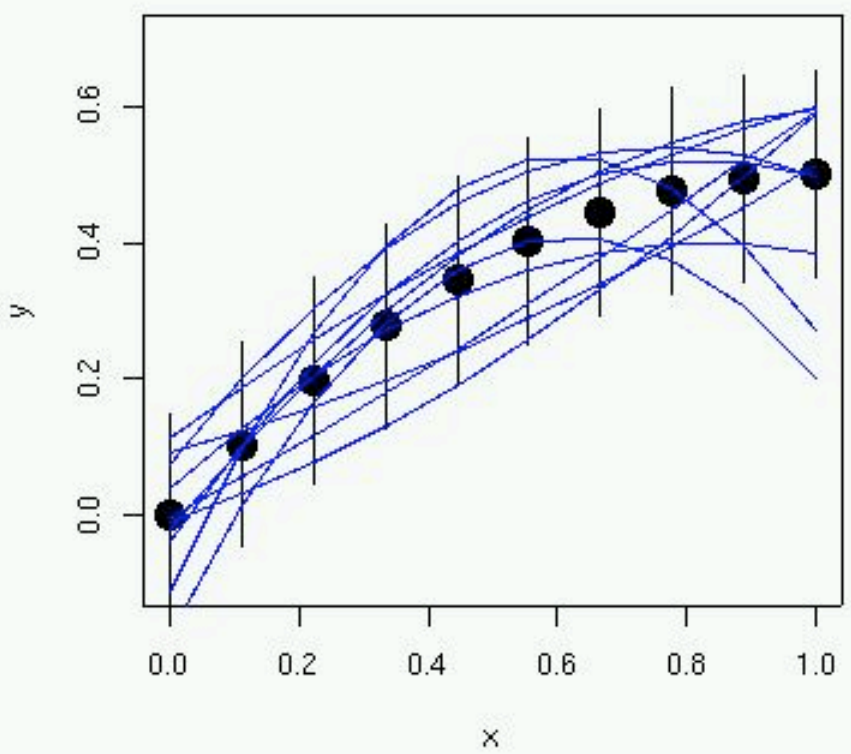
True relationship:

$$y = x - \frac{1}{2}x^2 + \varepsilon \qquad \varepsilon \sim N(0,\ 0.15^2)$$

Model: linear dependence



Model: quadratic dependence



variance stabilizing transformations

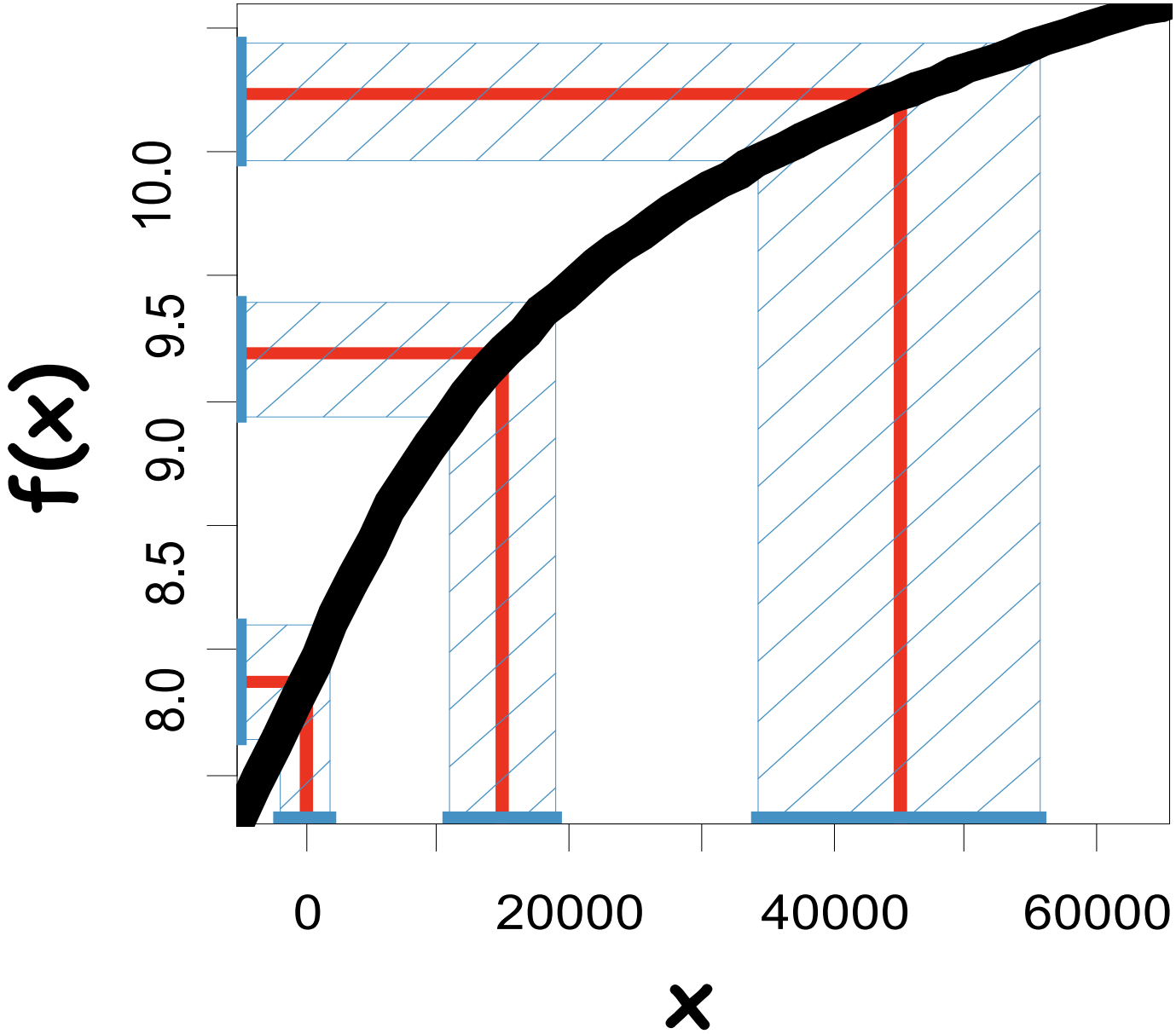
X_u a family of random variables with $EX_u = u$, $\mathrm{Var} X_u = v(u)$. Define

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

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

derivation: linear approximation

variance stabilizing transformations



variance stabilizing transformations

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

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

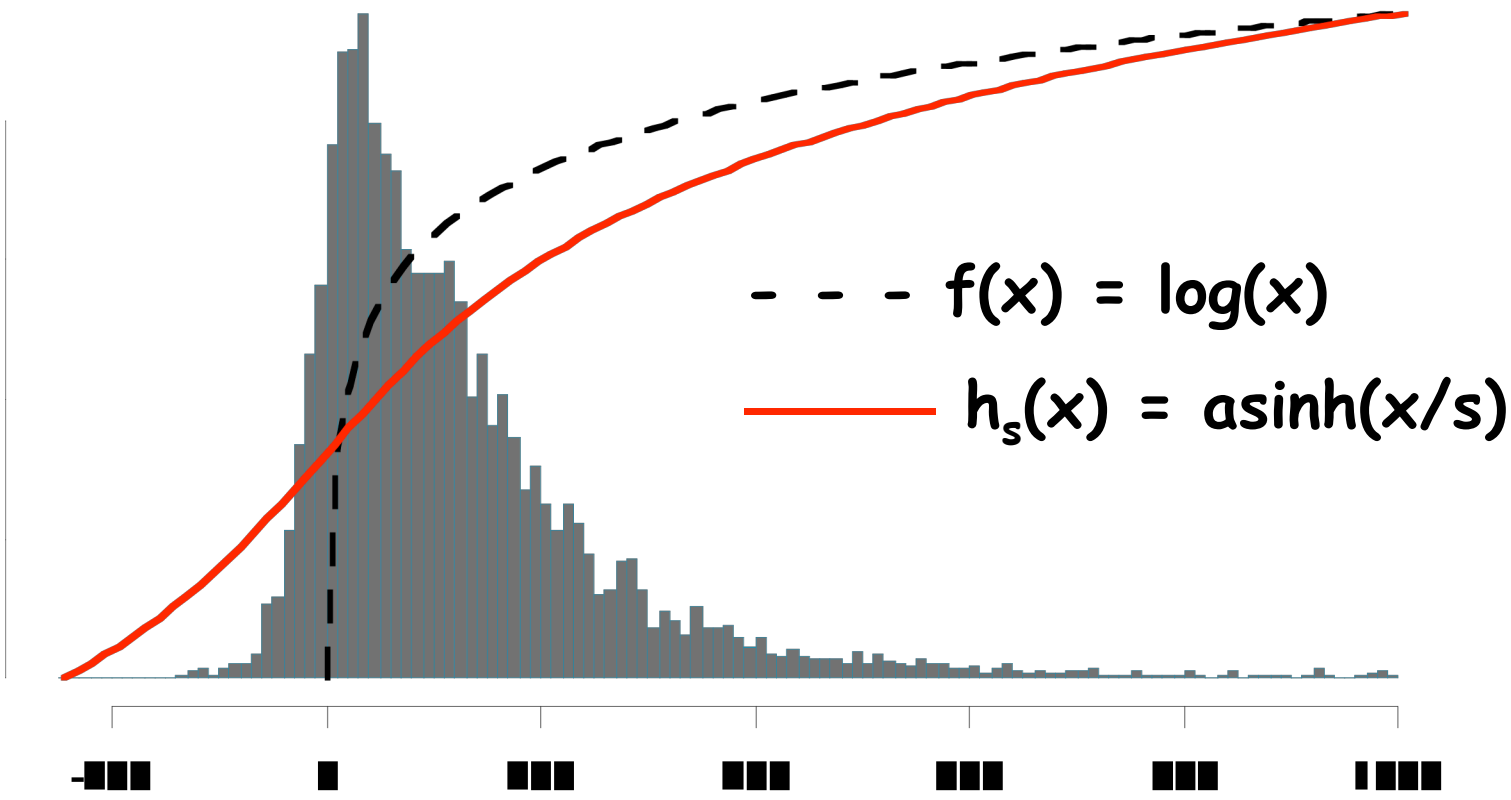
2.) constant CV ('multiplicative') $v(u) \propto u^2 \Rightarrow f \propto \log u$

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

4.) additive and multiplicative

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

the "glog" transformation



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

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

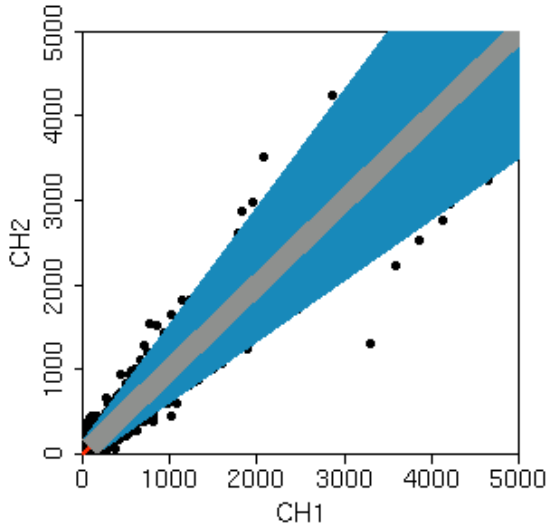
P. Munson, 2001

D. Rocke & B. Durbin, ISMB 2002

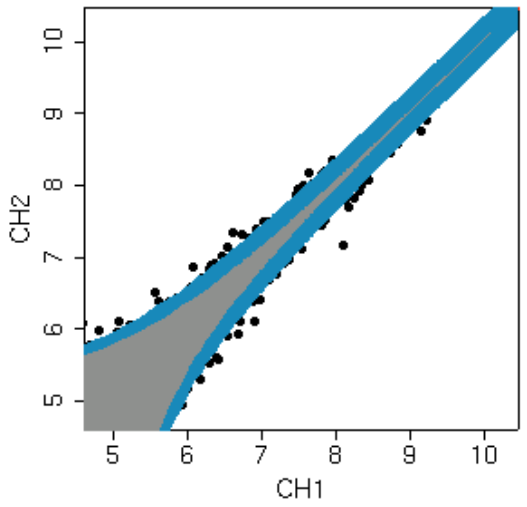
W. Huber et al., ISMB 2002

glog

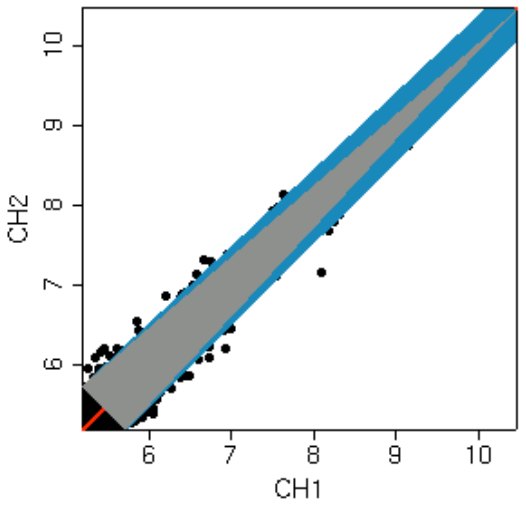
raw scale



log



glog



variance:

constant part

proportional part

Parameter estimation

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

- maximum lik… d –
  but sensitive
- model holds
  differentially
- robust varic
  Trimmed Sum
- works well
  differentially

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"

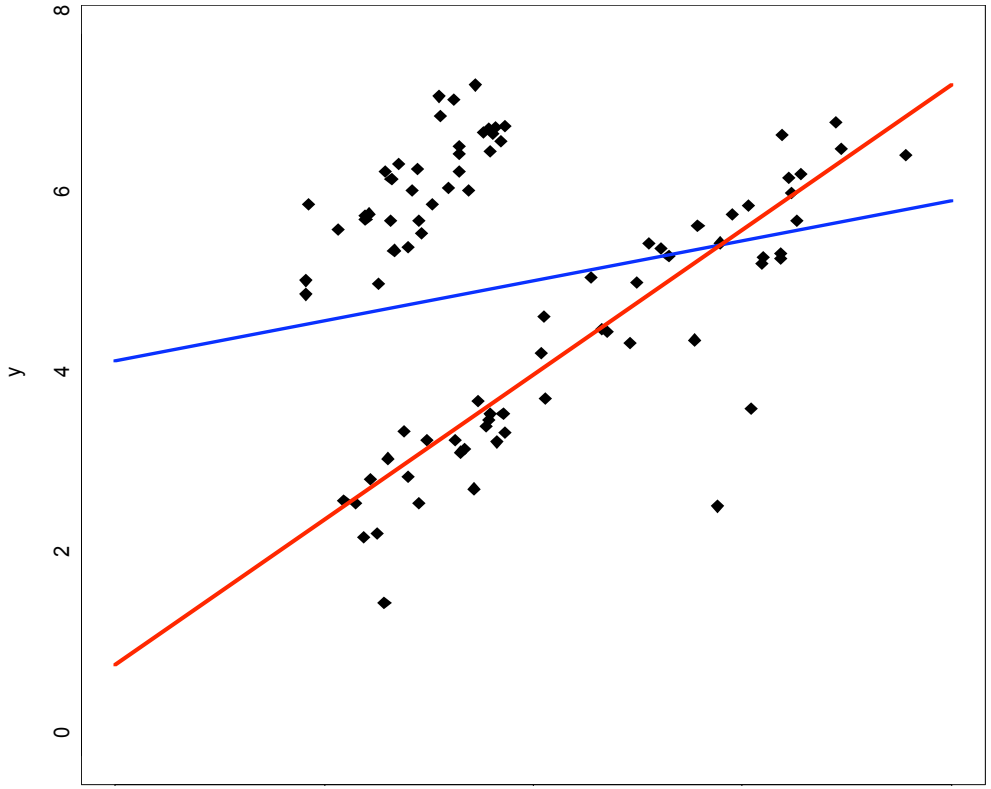
$b_{ik} = b_i\, b_k \exp(\eta_{ik})$

b_i per-sample normalization factor

b_k sequence-wise labeling efficiency

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

Least trimmed sum of squares regression



minimize

$$\sum_{i=1}^{n/2} \left(y_{(i)} - f(x_{(i)}) \right)^2$$

P. Rousseeuw, 1980s

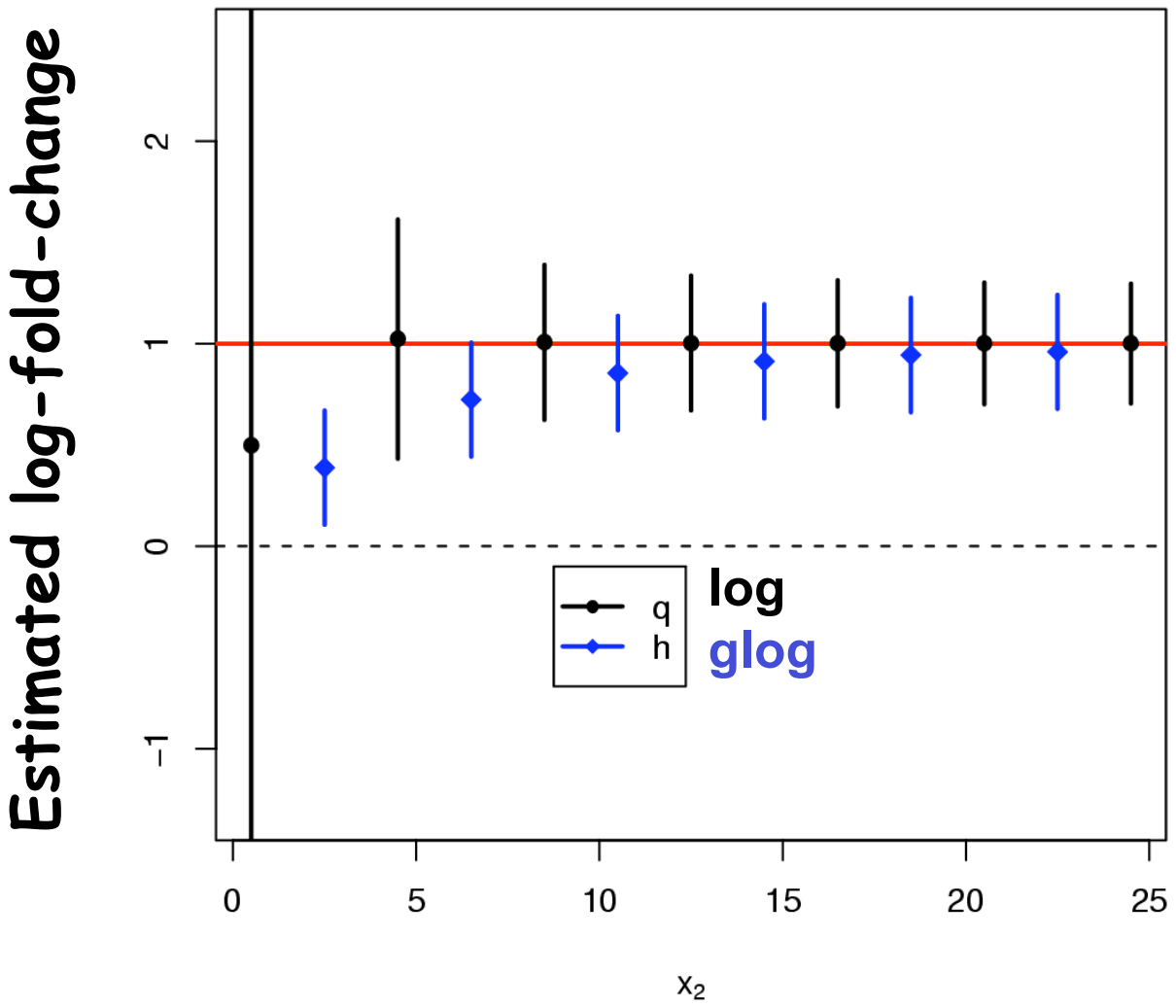
- least sum of squares
- least trimmed sum of squares

"usual" log-ratio $\log \frac{x_1}{x_2}$

'glog' (generalized log-ratio) $\log \frac{x_1 + \sqrt{x_1^2 + c_1^2}}{x_2 + \sqrt{x_2^2 + c_2^2}}$

c_1, c_2 are experiment specific parameters (~level of background noise)

Variance Bias Trade-Off



Signal intensity

Variance-bias trade-off and shrinkage estimators

Shrinkage estimators:

pay a small price in bias for a large decrease of variance, so overall the mean-squared-error (MSE) is reduced.

Particularly useful if you have few replicates.

Generalized log-ratio:

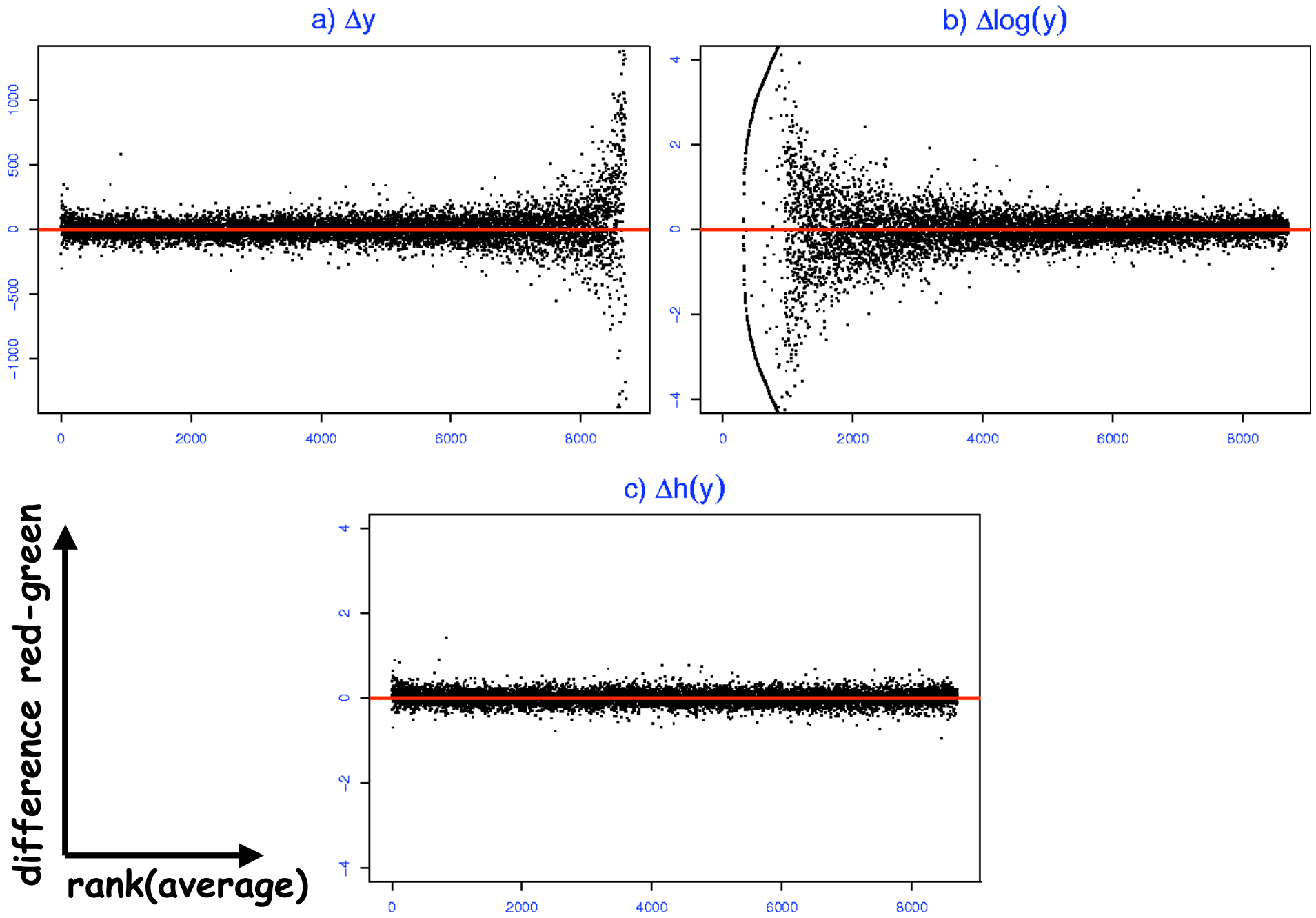
= a shrinkage estimator for fold change

There are many possible choices, we chose "variance-stabilization":

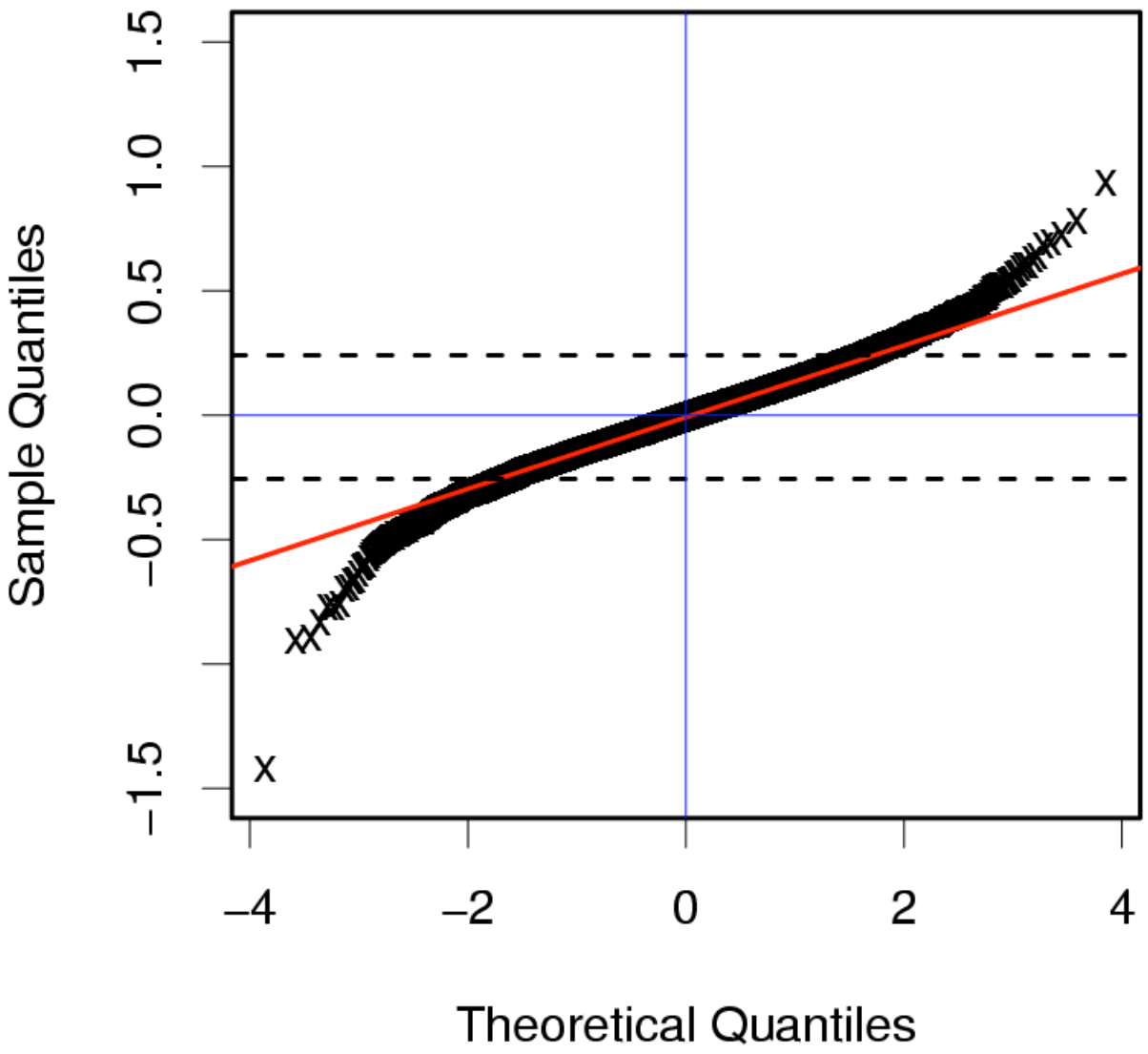
+ interpretable even in cases where genes are off in some conditions

+ can subsequently use standard statistical methods (hypothesis testing, ANOVA, clustering, classification…) with less worries about heteroskedasticity than with many alternative methods

evaluation: effects of different data transformations



Normality: QQ-plot



"Single color normalization"

n red-green arrays (R_1, G_1, R_2, G_2,… R_n, G_n)

within/between slides

for (i=1:n)

calculate $M_i = \log(R_i/G_i)$, $A_i = \frac{1}{2}\log(R_i * G_i)$

normalize M_i vs A_i

normalize $M_1 \ldots M_n$

all at once

normalize the matrix of (R, G)

then calculate log-ratios or any other contrast you like

What about non-linear effects

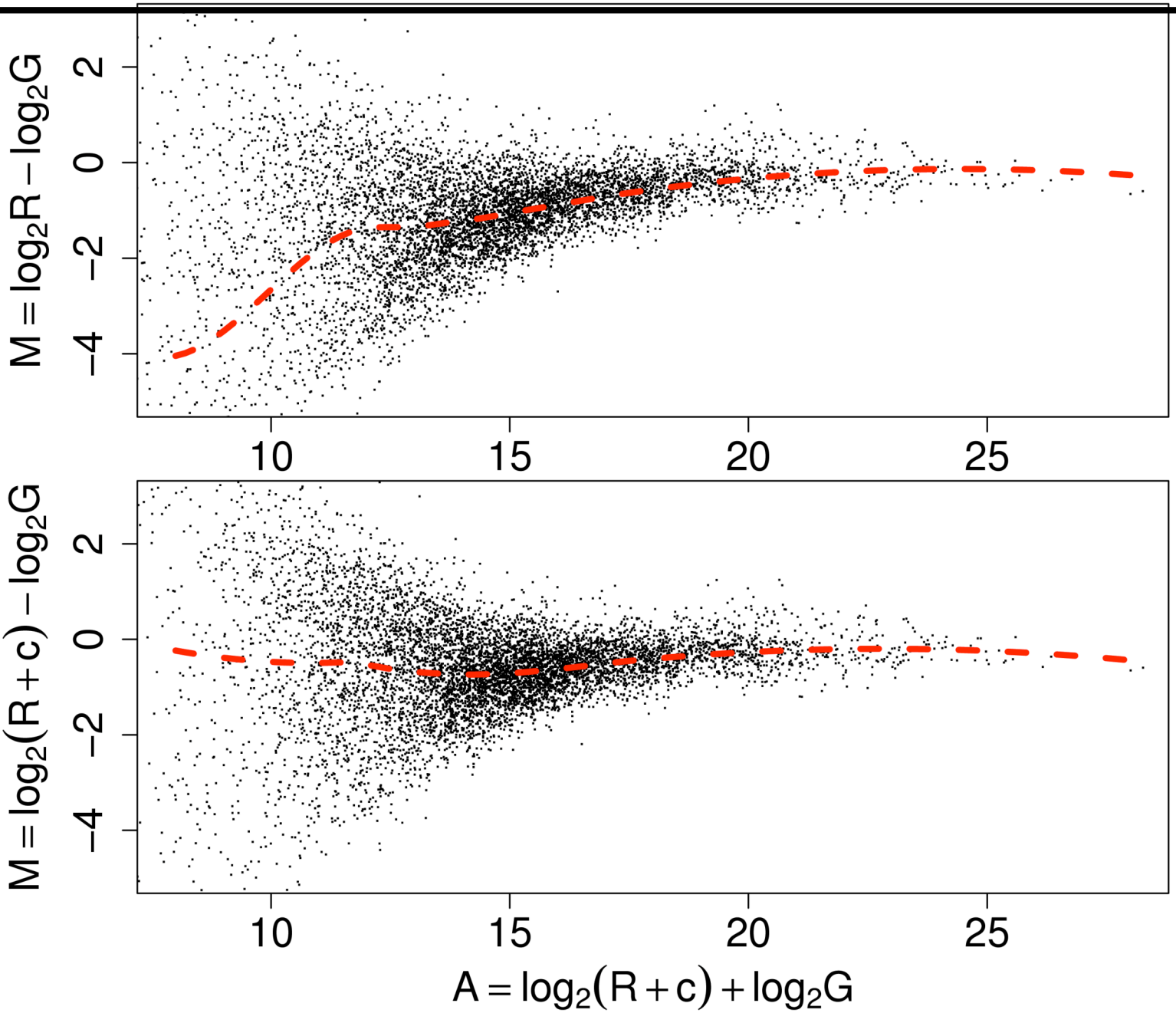
- Microarrays can be operated in a linear regime, where fluorescence intensity increases proportionally to target abundance (see e.g. Affymetrix dilution series)

Two reasons for non-linearity:

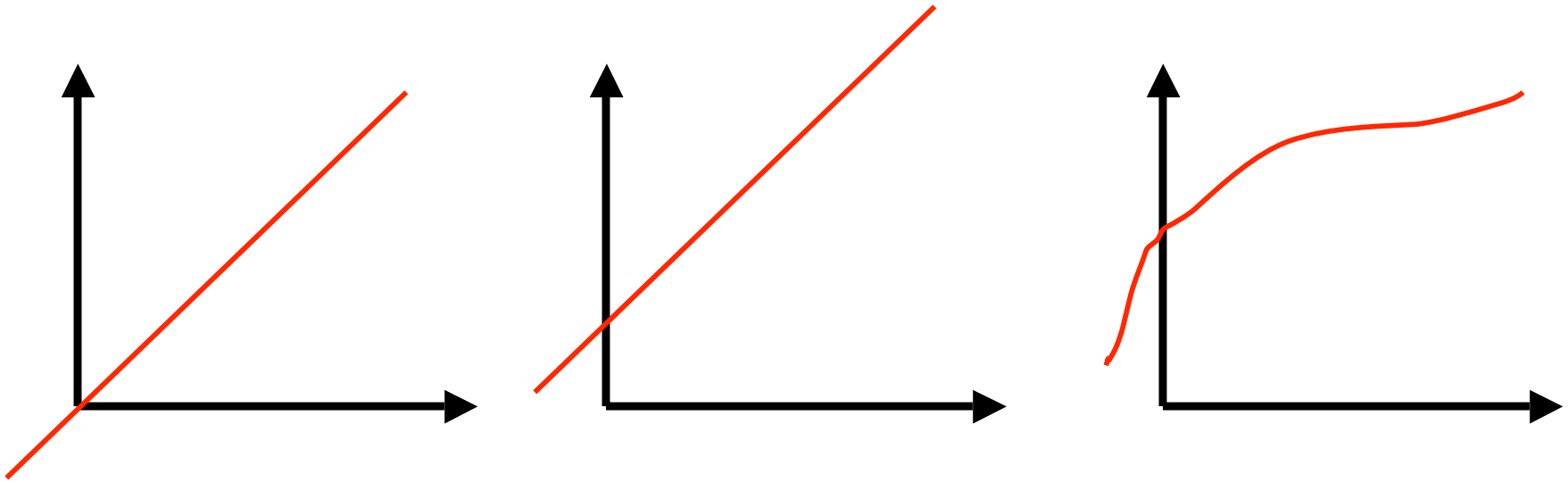
- At the high intensity end: saturation/quenching. This can and should be avoided experimentally - loss of data!

- At the low intensity end: background offsets, instead of $y=k \cdot x$ we have $y=k \cdot x+x_0$, and in the log-log plot this can look curvilinear. But this is an affine-linear effect and can be correct by affine normalization. Non-parametric methods (e.g. loess) risk overfitting and loss of power.

Non-linear or affine linear?



Definitions



linear

affine linear

genuinely non-linear

References

Bioinformatics and computational biology solutions using R and Bioconductor, R. Gentleman, V. Carey, W. Huber, R. Irizarry, S. Dudoit, Springer (2005).

Variance stabilization applied to microarray data calibration and to the quantification of differential expression. W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, M. Vingron. Bioinformatics 18 suppl. 1 (2002), S96-S104.

Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. R. Irizarry, B. Hobbs, F. Collins, …, T. Speed. Biostatistics 4 (2003) 249-264.

Error models for microarray intensities. W. Huber, A. von Heydebreck, and M. Vingron. Encyclopedia of Genomics, Proteomics and Bioinformatics. John Wiley & sons (2005).

Differential Expression with the Bioconductor Project. A. von Heydebreck, W. Huber, and R. Gentleman. Encyclopedia of Genomics, Proteomics and Bioinformatics. John Wiley & sons (2005).

Acknowledgements

Anja von Heydebreck (Darmstadt)
Robert Gentleman (Seattle)
Günther Sawitzki (Heidelberg)
Martin Vingron (Berlin)
Annemarie Poustka, Holger Sültmann, Andreas Buness, Markus Ruschhaupt (Heidelberg)
Rafael Irizarry (Baltimore)
Judith Boer (Leiden)
Anke Schroth (Heidelberg)
Friederike Wilmer (Hilden)