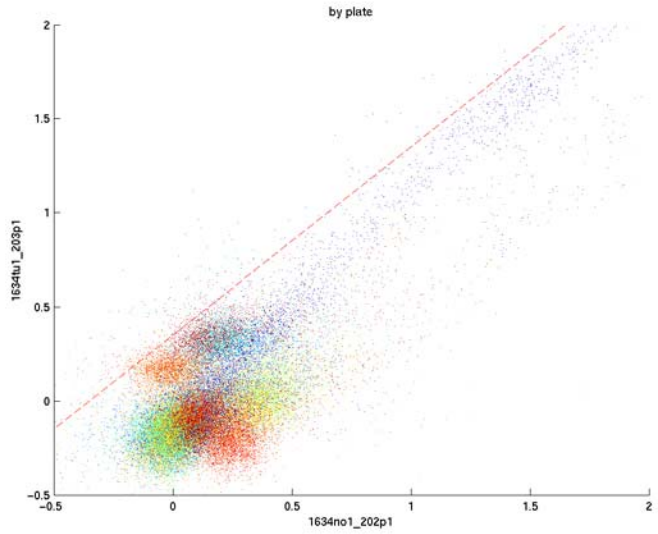
Error models and normalization

Wolfgang Huber

DKFZ Heidelberg



Acknowledgements

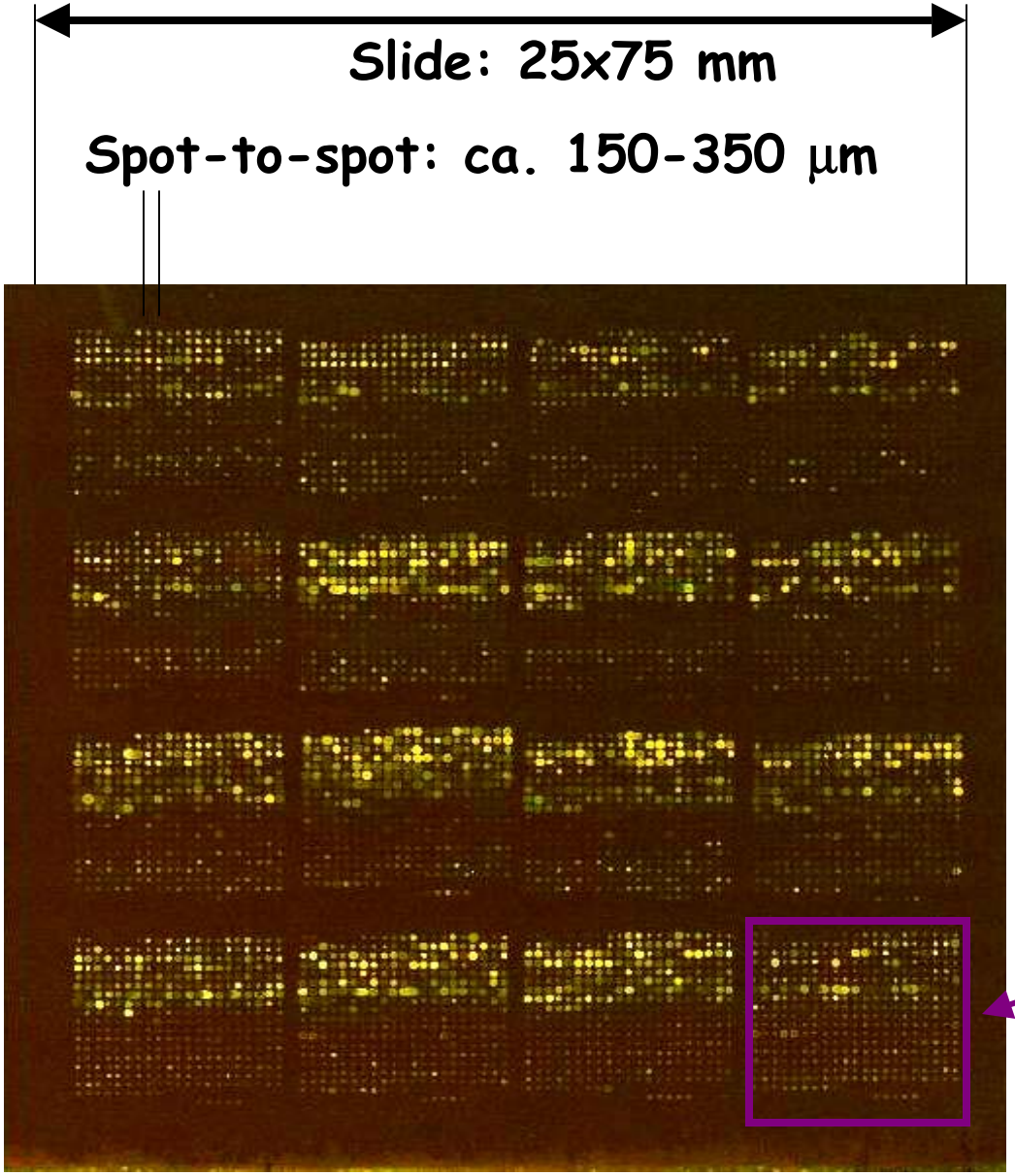
Anja von Heydebreck, Martin Vingron

Andreas Buness, Markus Ruschhaupt, Klaus Steiner, Jörg Schneider, Katharina Finis, Anke Schroth, Friederike Wilmer, Judith Boer, Holger Sültmann, Annemarie Poustka

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

and many others

A microarray slide (spotted)



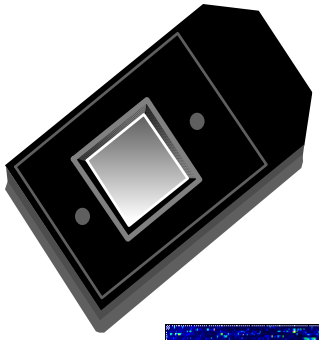
4x4, 8x4, or 12x 4 sectors

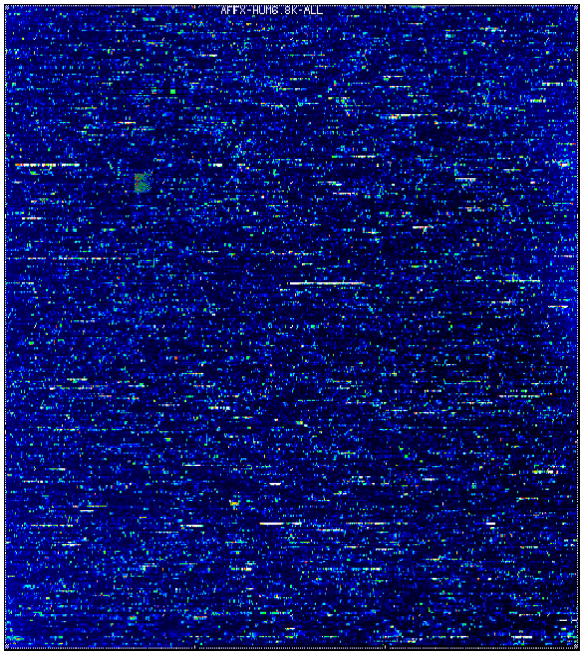
17...38 rows and columns per sector

ca. 4000...46000 probes/array

sector: corresponds to one print-tip

Affymetrix oligonucleotide chips

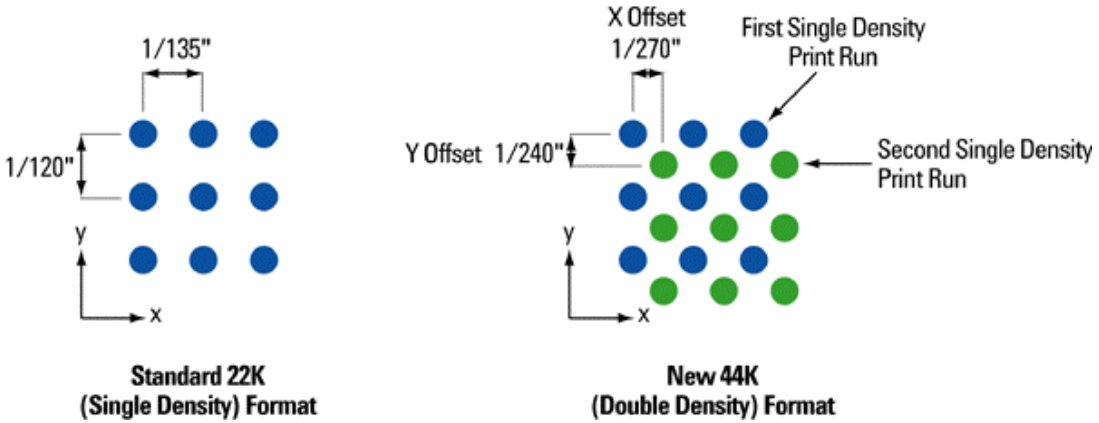




hgU133plus2.0	
Feature size	11μm
No. probes	600,000
No. probe pairs per target sequence	11
Oligonucleotide length	25

Agilent oligonucleotide chips

whole human genome kit (5/2004)	
Feature size	≈100μm
No. probes	44,000
Oligonucleotide length	60



Standard 22K (Single Density) Format

New 44K (Double Density) Format

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.

Image Analysis

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)

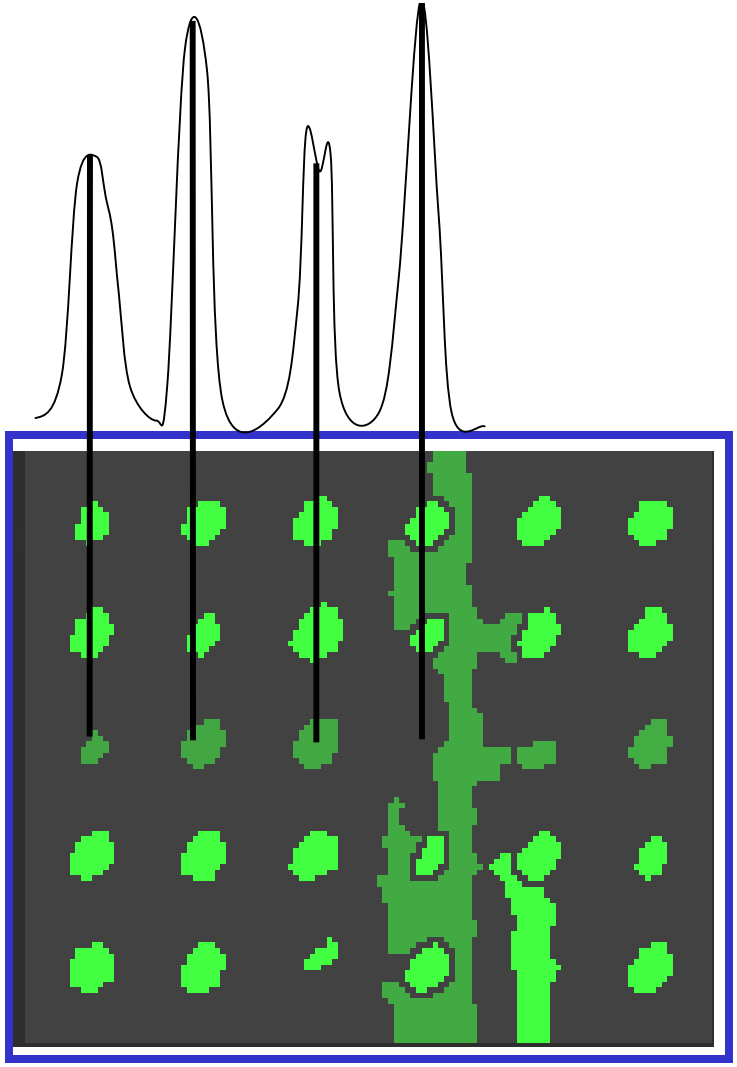
… auxiliaries: background, area, std dev, …

Image analysis

1. Addressing. Estimate location of spot centers.

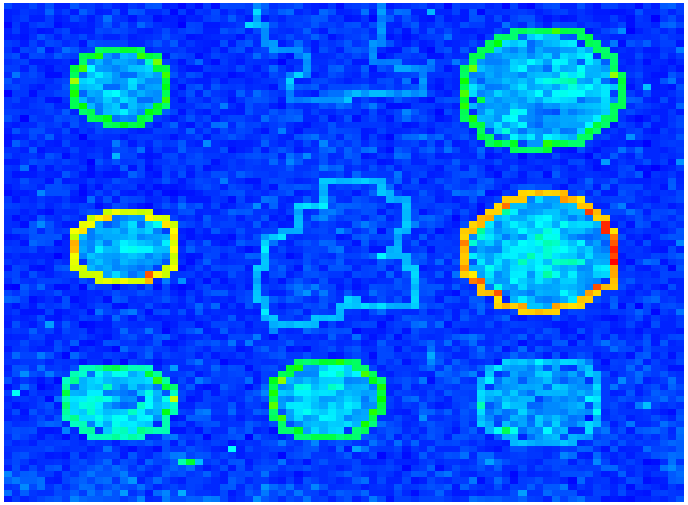
2. Segmentation. Classify pixels as foreground (signal) or background.

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

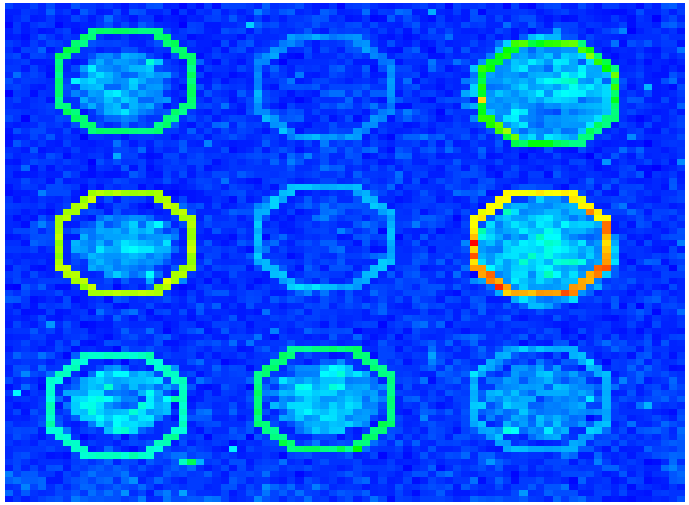


→ R and G for each spot on the array.

Segmentation



adaptive segmentation
seeded region growing



fixed circle segmentation

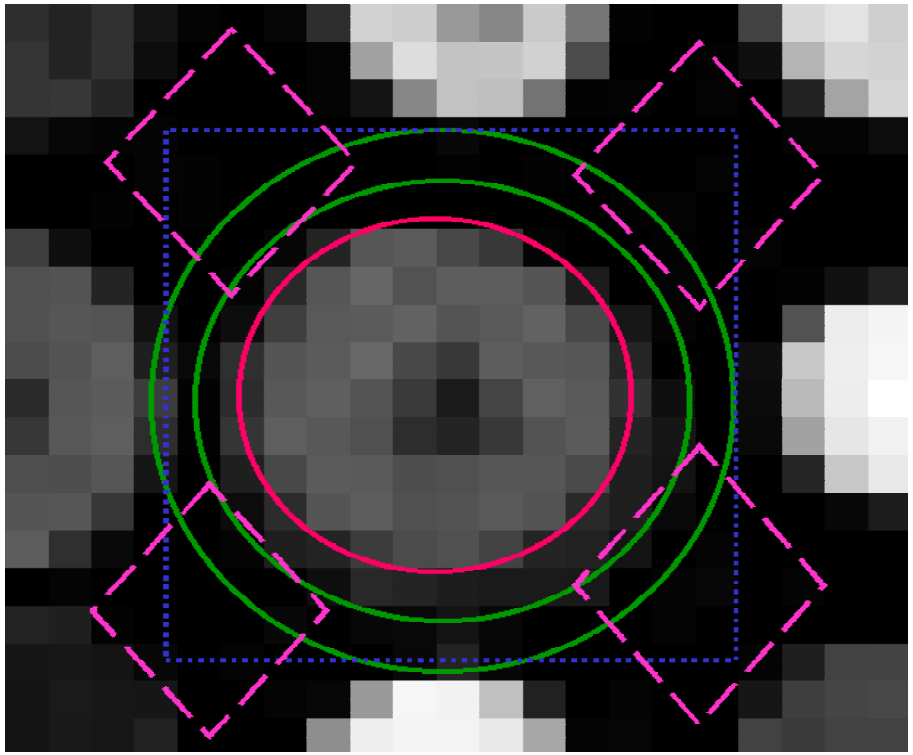
Spots may vary in size and shape.

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.

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

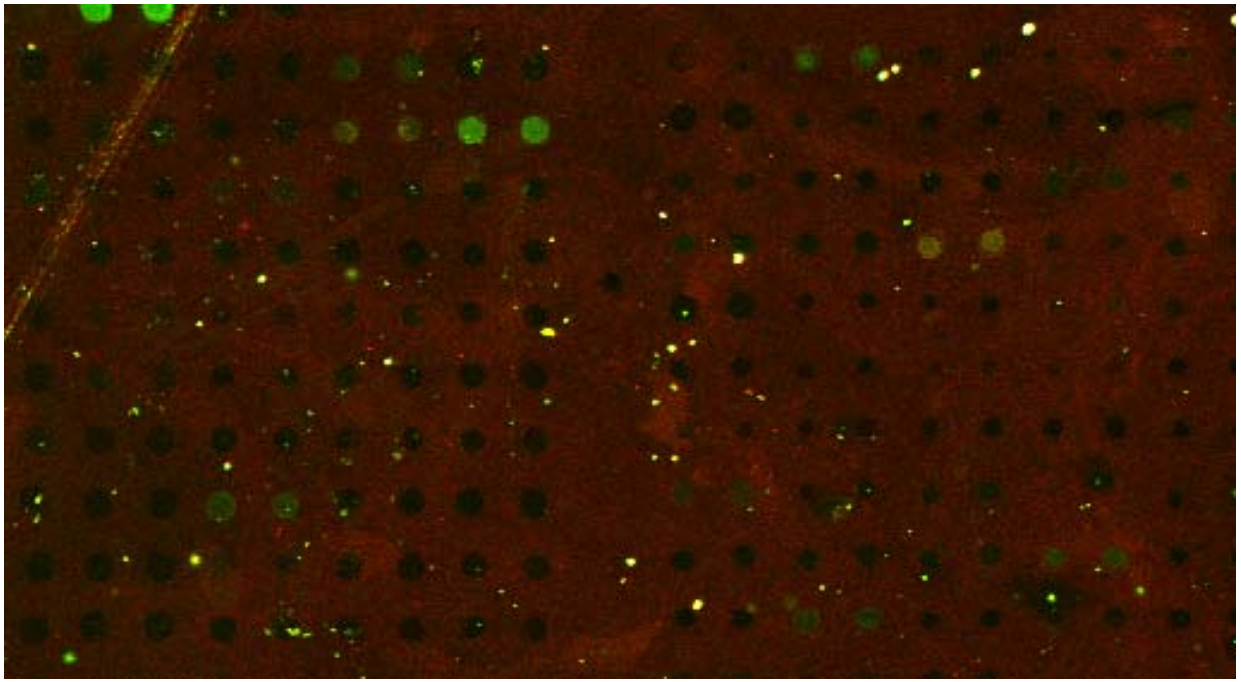
What is (local) background?

usual assumption:

total brightness =

background brightness (adjacent to spot)

\+ brightness from labeled sample cDNA



?

Affymetrix files

Main software from Affymetrix:

MAS - MicroArray Suite.

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

CEL file: probe intensities, ~500,000 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.

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

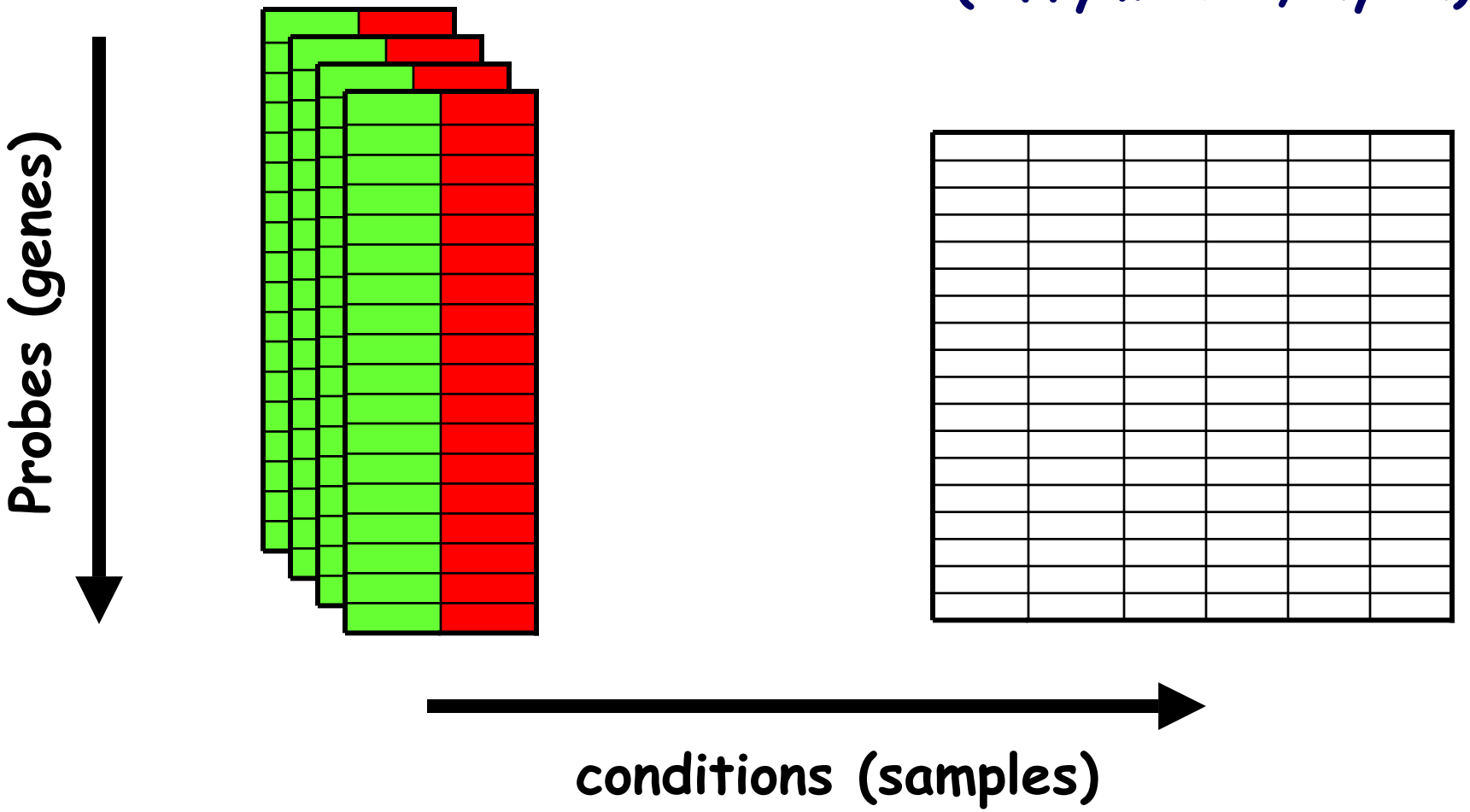
- Percentage of spots with no signal
- Range of intensities
- Distribution of spot signal area, etc.

How to use quality measures in subsequent analyses?

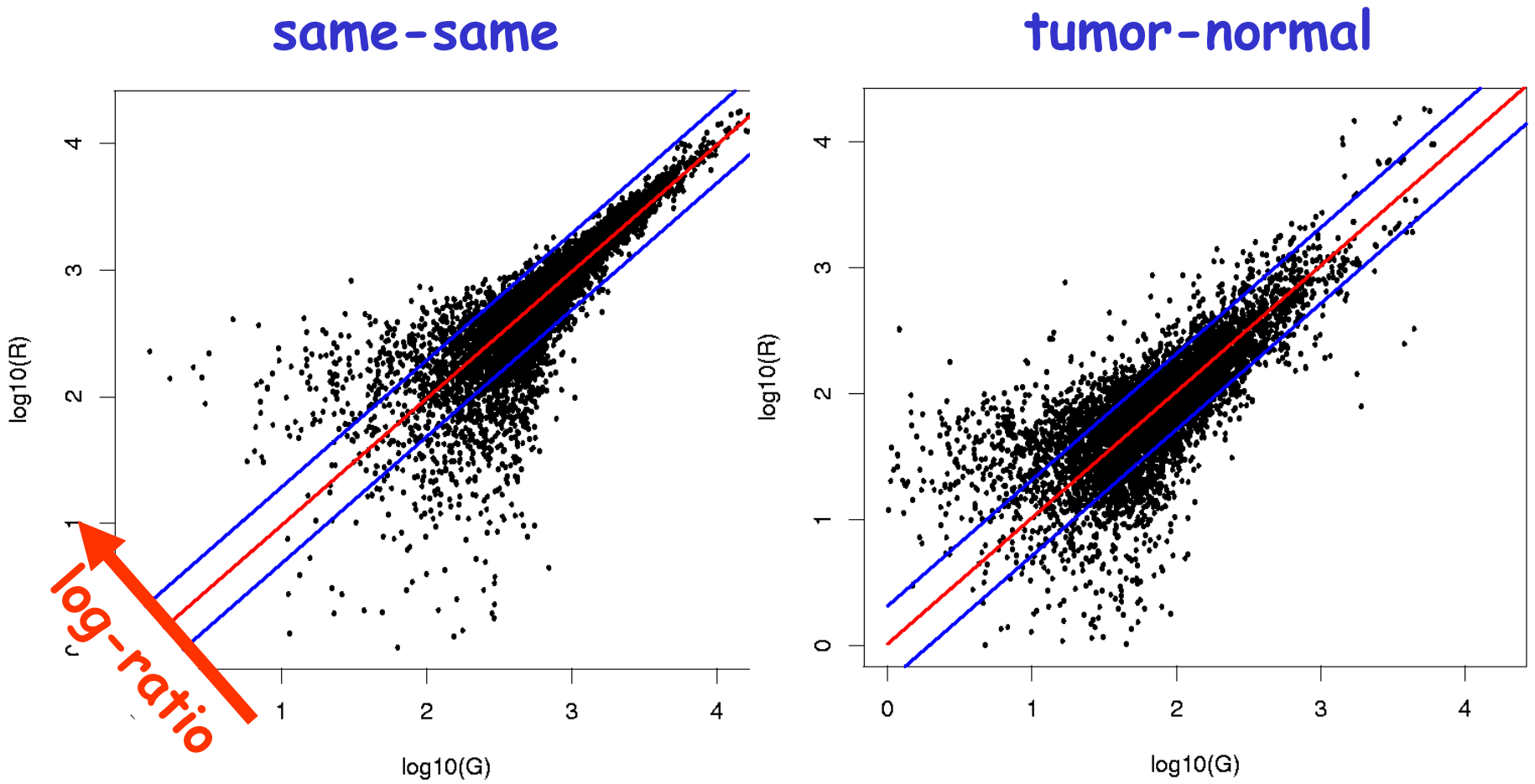
spot intensity data

two-color spotted arrays

n one-color arrays
(Affymetrix, nylon)

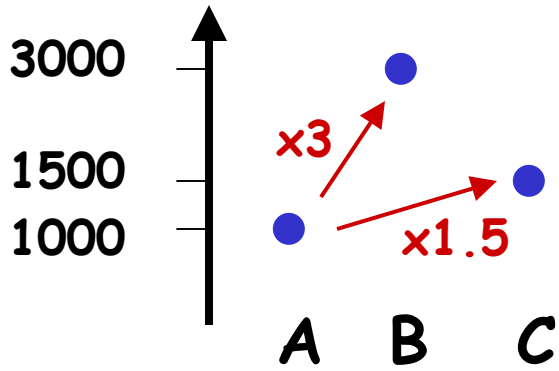


Which genes are differentially transcribed?

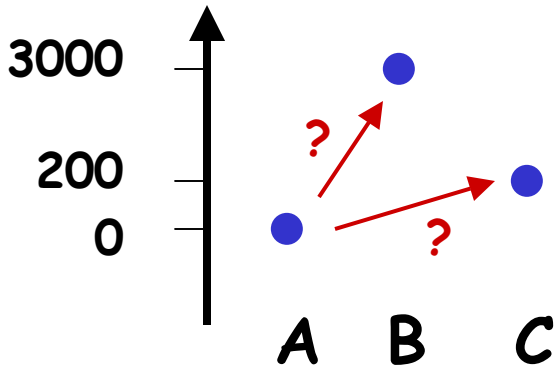


ratios and fold changes

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

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

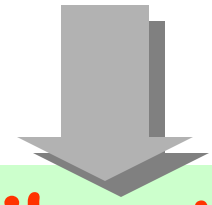
PCR yield
DNA quality
spotting efficiency,
spot size
cross-/unspecific hybridization
stray signal

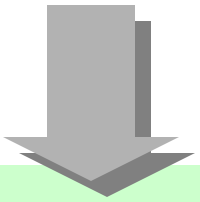
Systematic

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

Stochastic

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





Calibration

Error model

Error models

Definition:

description of the possible outcomes of a measurement

Depends 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. statistical inference (appropriate parametric methods have better power)
2. summarization (summary statistic instead of full empirical distribution)
3. quality control

Derivation of additive-multiplicative error model

$$y = f(x,u)$$

y measurement

f measurement apparatus

x true underlying quantity

u further factors that can influence the measurement ("environment")

Derivation of additive-multiplicative error model

$$y = f(x,u)$$

generic observation eqn.
(x=true value, u=environment)

$$y = f(0,u) + f'(0,u)\cdot x + O(x^2)$$

first order approximation of x-dependence of f

$$f(0,u) \approx f(0,\overline{u}) + \sum_i \frac{\partial f\ (0,u)}{\partial u_i}(u_i - \overline{u}_i)$$

first order approximation of u-dependence of f

$$f'(0,u) \approx f'(0,\overline{u}) + \sum_i \frac{\partial f'(0,u)}{\partial u_i}(u_i - \overline{u}_i)$$

first order approximation of u-dependence of f'

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

model environment fluctuations as noise

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	array **x** color **x** print-tip group
ε random background	iid in whole experiment	iid per array
b systematic gain factor	array **x** color	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

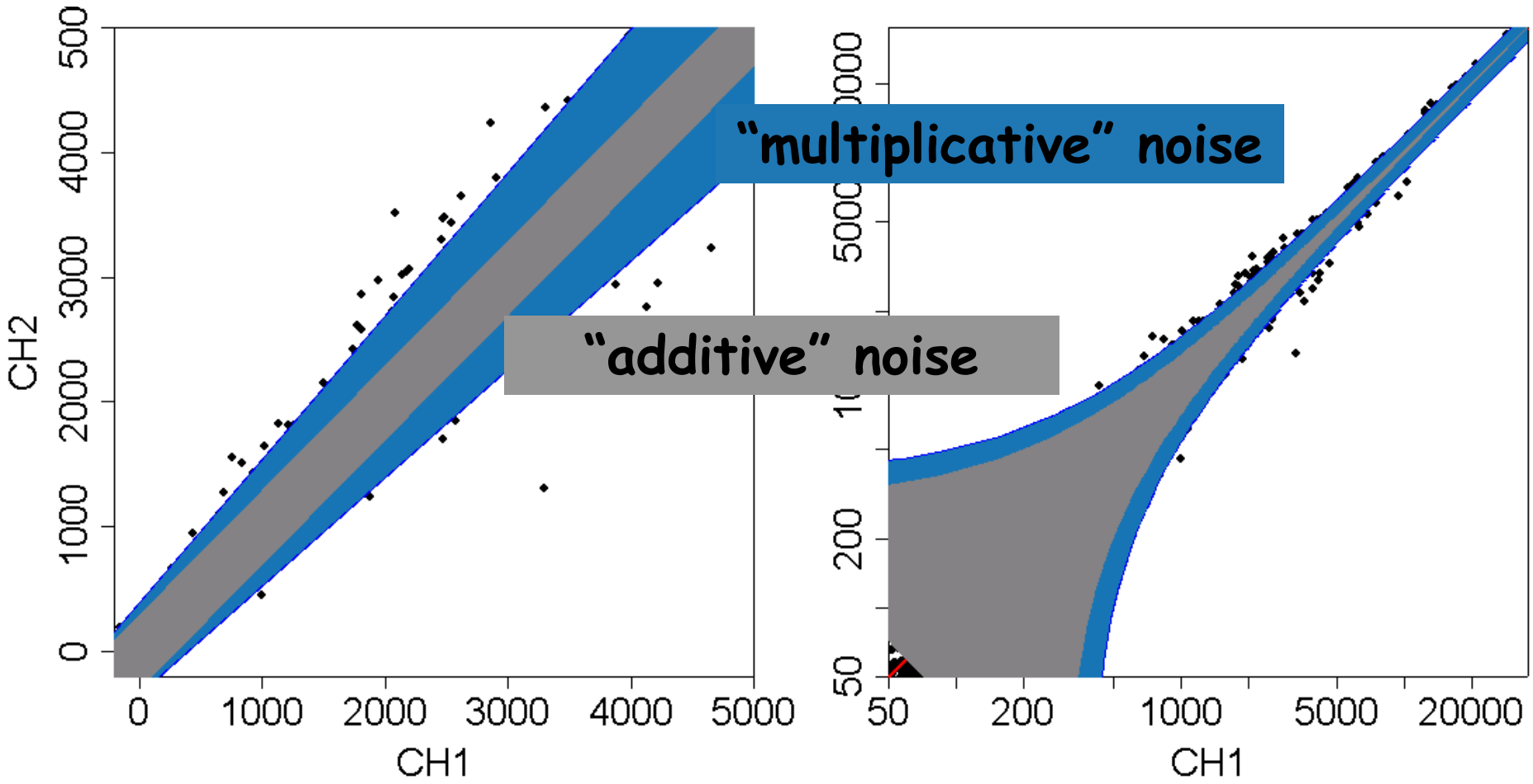
Outliers

⇒ use a robust method

Algorithm

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

The two-component model



raw scale

log scale

B. Durbin, D. Rocke, JCB 2001

Nesting

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

e.g. replicate hybridization

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

e.g. replicate RNA isolation

$$\Downarrow$$

$$y \approx a'' + \varepsilon'' + b'' \cdot z \cdot (1 + \eta'')$$

overall

variance stabilization

X_u a family of random variables with

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

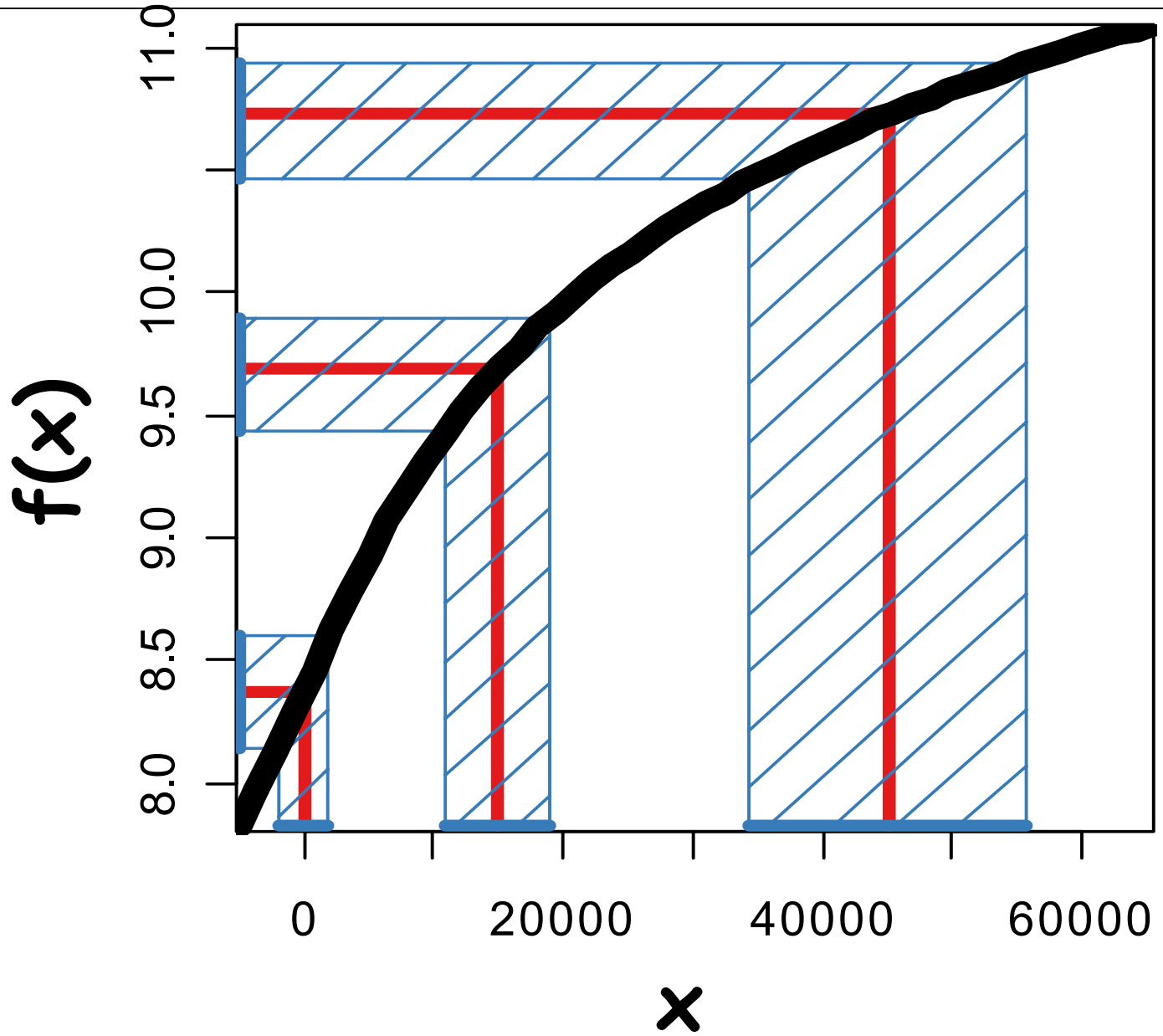
Define

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

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

derivation: linear approximation

variance stabilizing transformation



variance stabilizing transformations

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

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

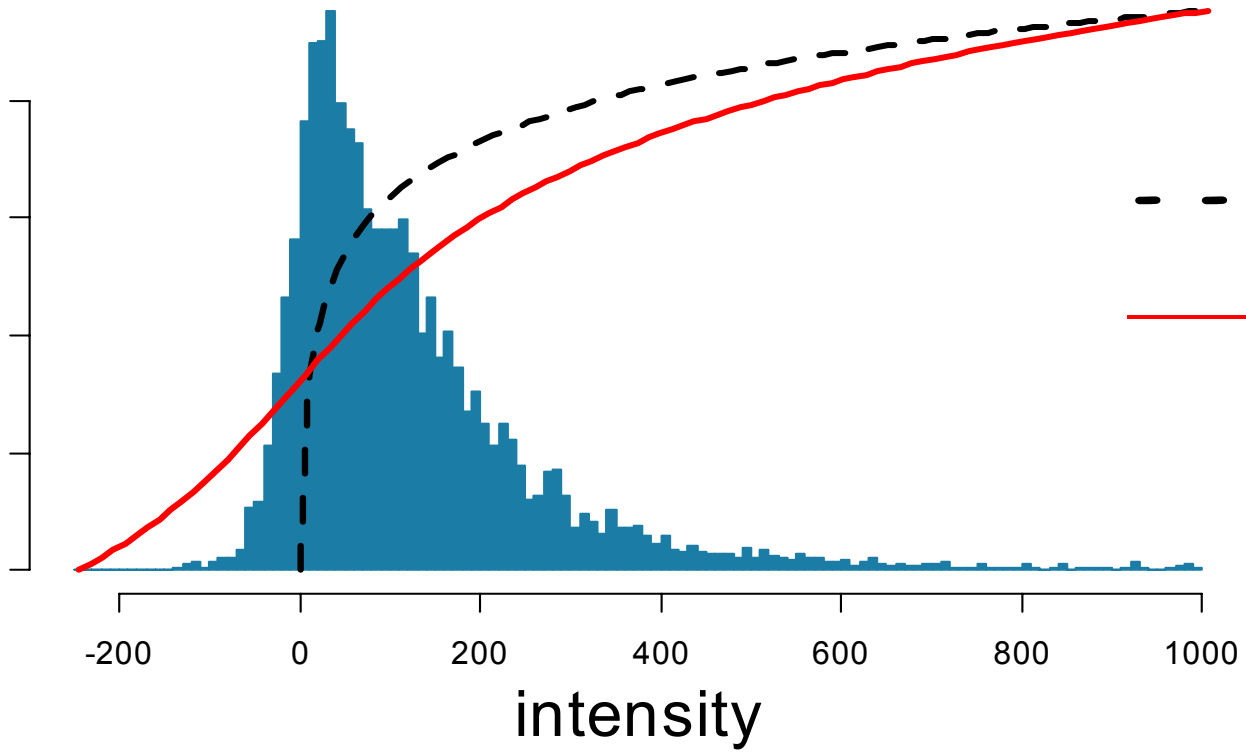
2.) const. coeff. of variation $v(u) \propto u^2 \quad \Rightarrow \quad f \propto \log u$

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



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

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

the transformed model

$$\operatorname{arsinh}\frac{Y_{ki} - a_{si}}{b_{si}} = \mu_k + \varepsilon_{ki}$$

$$\varepsilon_{ki} : \quad N(0, c^2)$$

i: arrays
k: probes
s: probe strata (e.g. print-tip, region)

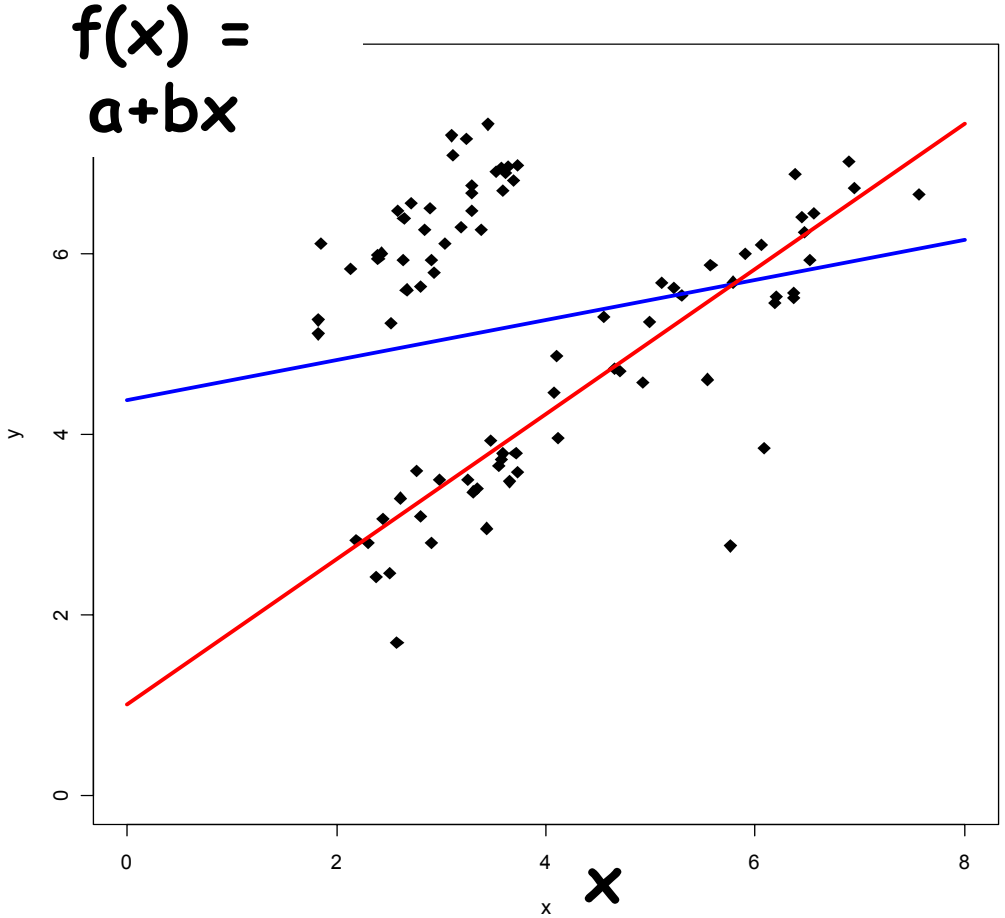
profile log-likelihood

$$pll(a,b) = \sup_{c,\mu} ll(a,b,c,\mu)$$

Here:

$$\begin{aligned}
&\mathrm{pll}(a_1, b_1, \ldots, a_d, b_d) = \\
&= \quad -nd \log \hat{\sigma} + \sum_{k=1}^{n} \sum_{i=1}^{d} \log h_i'(y_{ki}) \\
&= \quad -\frac{nd}{2} \log \left(\sum_{k=1}^{n} \sum_{i=1}^{d} (h_i(y_{ki}) - \hat{\mu}_k)^2 \right) + \sum_{k=1}^{n} \sum_{i=1}^{d} \log h_i'(y_{ki})
\end{aligned}$$

Least trimmed sum of squares regression



$$r_i = y_i - f(x_i)$$

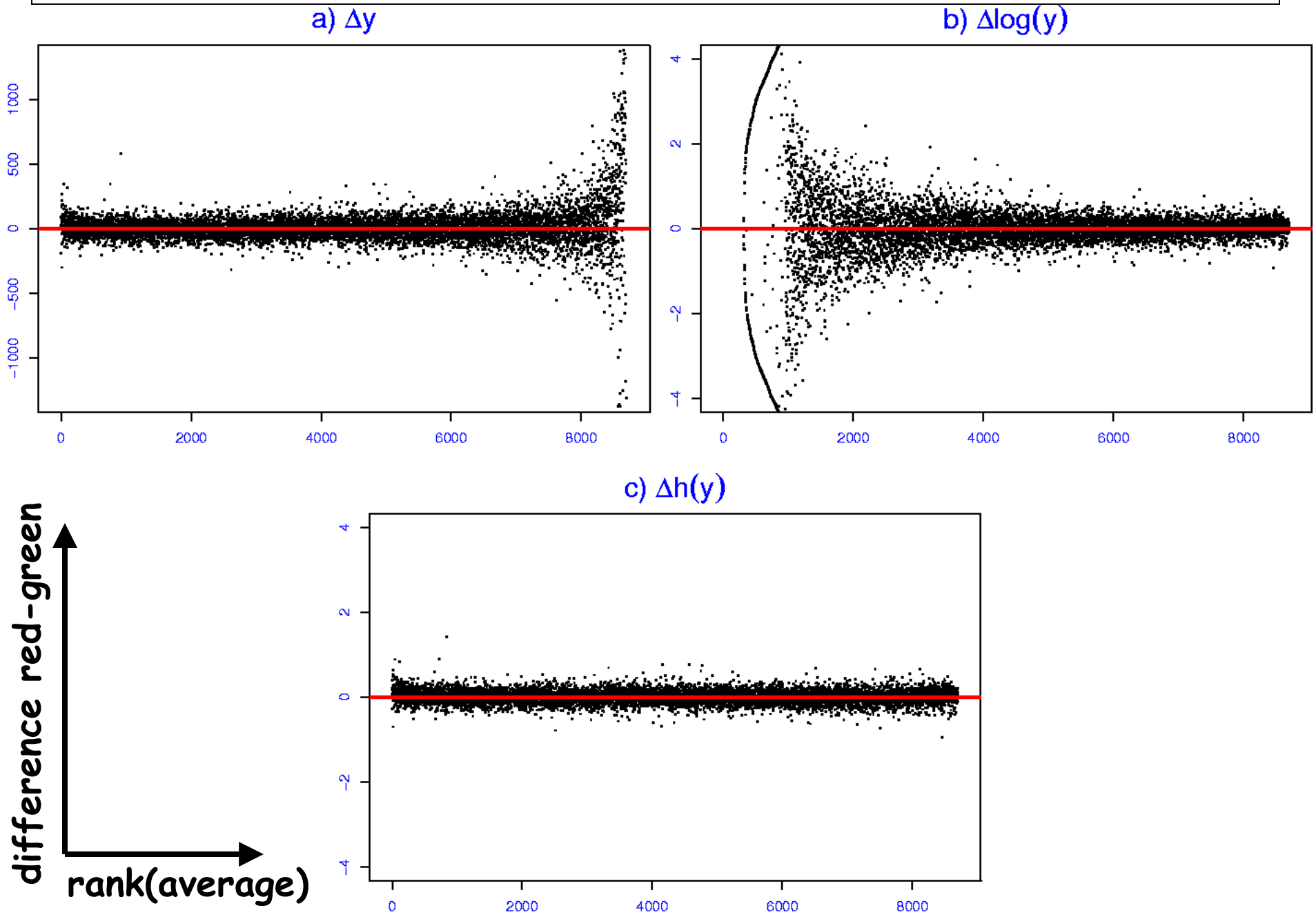
$$LS : \sum_i r_i^2 \rightarrow \min$$

$$LTS : \sum_{i \in I} r_i^2 \rightarrow \min$$

$$I = \{ i \mid r_i^2 < \operatorname{med}_k r_k^2 \}$$

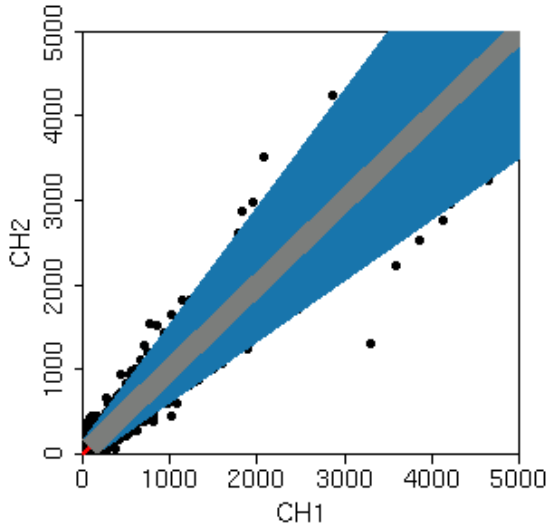
- least sum of squares (LS): Gauss, Legendre ~ 1790
- least trimmed sum of squares (LTS): Rousseeuw 1984

evaluation: effects of different data transformations

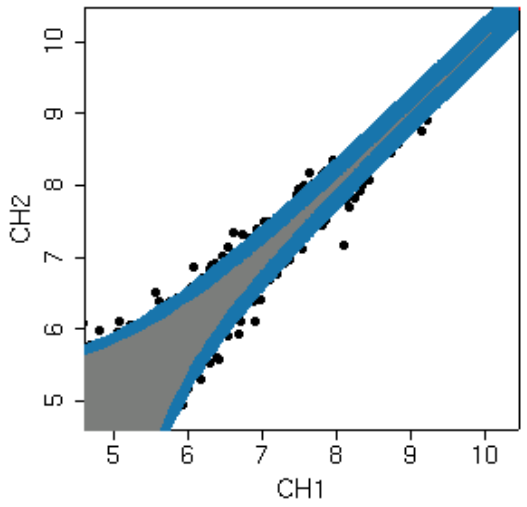


glog

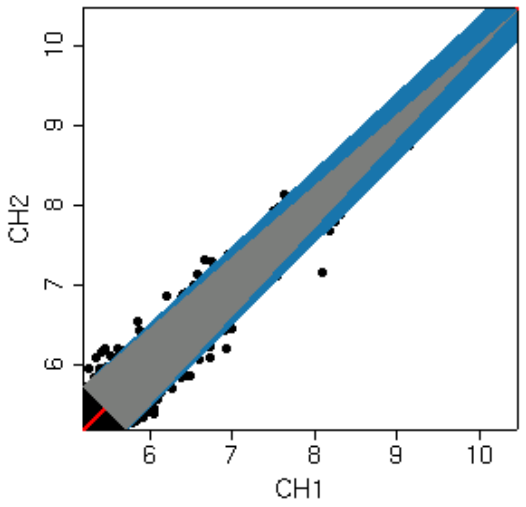
raw scale



log



glog



variance:

constant part

proportional part

Motivation for the generalized log-ratio

z_1, z_2 ~ additive-multiplicative error model
Search function h that fulfills

(*i*) $h(z_1, z_2) = -h(z_2, z_1)$

(*ii*) $\mathrm{Var}(h(z_1, z_2)) \approx \text{const.}$

$$\Rightarrow h(z_1, z_2) = a\sinh\left(\frac{z_1 - a}{b}\right) - a\sinh\left(\frac{z_s - a}{b}\right)$$

Properties of the generalized log-ratio

$$h(z_1, z_2) = \operatorname{asinh}\left(\frac{z_1 - a}{b}\right) - \operatorname{asinh}\left(\frac{z_2 - a}{b}\right)$$

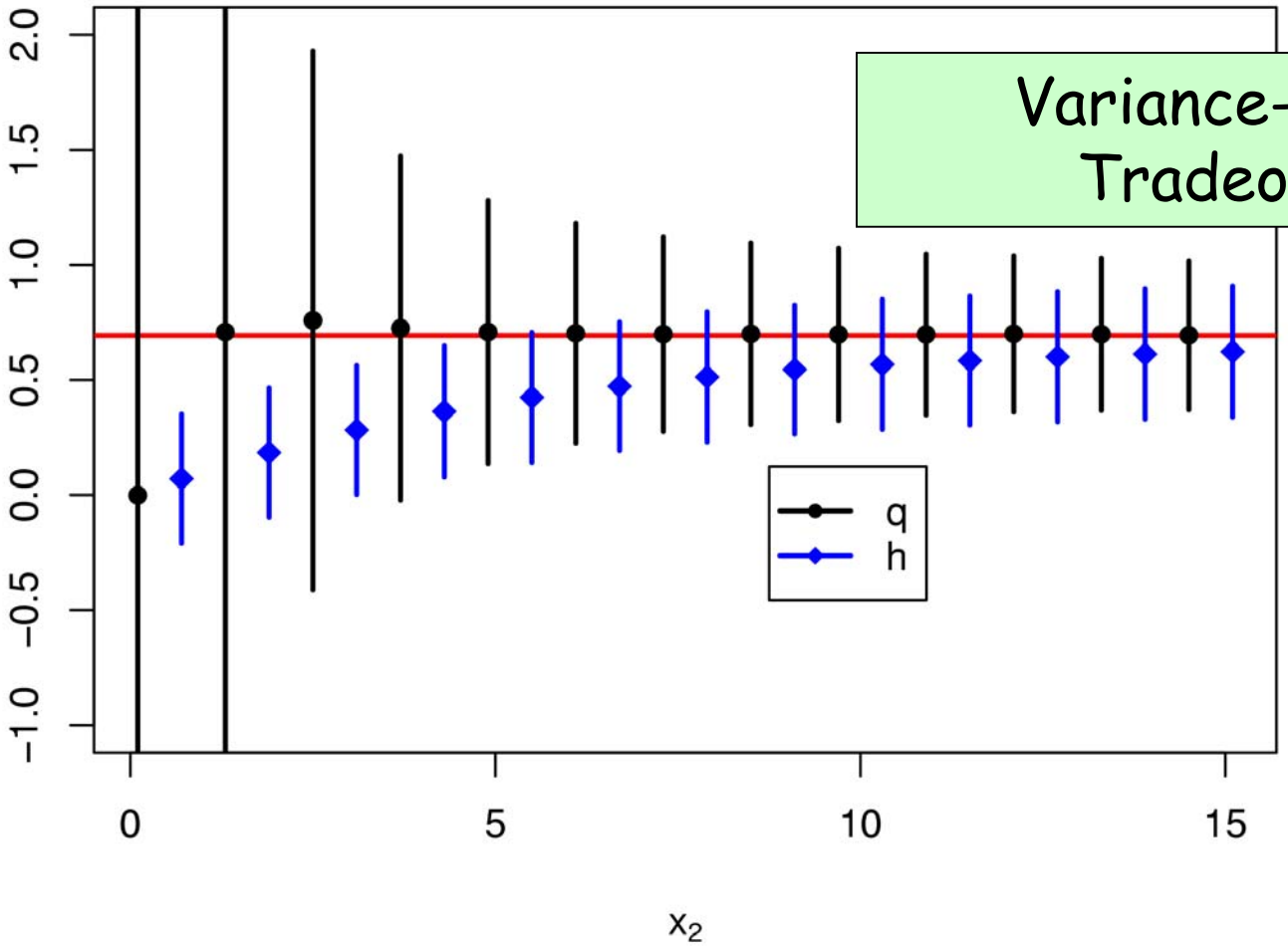
$$q(z_1, z_2) = \log(z_1 - a) - \log(z_2 - a)$$

(i) for z_1, $z_2 >> a$, h and q are the same

(ii) $|h(z_1, z_2)| \leq |q(z_1, z_2)|$

(iii) $\exp(h(z_1, z_2))$ is a shrinkage estimator for fold-change

Properties of the generalized log-ratio



Variance-Bias Tradeoff

$z_i = a + \varepsilon + b x_i \exp(\eta)$

$x_2 = 0.5 \ldots 15$, $x_1 = 2\, x_2$, $a = 0$, $\sigma_a = 1$, $b = 1$, $\sigma_b = 0.1$

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

"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

How to compare and assess different 'preprocessing' methods

Normalization = correction for systematic experimental biases + provision of an expression value that can be used subsequently for testing, clustering, classification, modelling.

Quality trade-off: the better the measurements, the less normalization

Variance-Bias trade-off: how do you weigh measurements that have low signal-noise ratio?

How to compare and assess different 'normalization' methods?

Normalization :=
1. correction for systematic experimental biases
2. provision of expression values that can subsequently be used for testing, clustering, classification, modelling…
3. provision of a measure of measurement uncertainty

Quality trade-off: the better the measurements, the less need for normalization. Need for "too much" normalization relates to a quality problem.

Variance-Bias trade-off: how do you weigh measurements that have low signal-noise ratio?
- just use anyway
- ignore
- shrink

How to compare and assess different 'normalization' methods?

Aesthetic criteria

Logarithm is more beautiful than arsinh

Practical critera

It takes forever to run vsn. Referees will only accept my paper if it uses the original MAS5.

Silly criteria

The best method is that that makes all my scatterplots look like straight, slim cigars

Physical criteria

Normalization calculations should be based on physical/chemical model

Economical/political criteria

Life would be so much easier if everybody were just using the same method, who cares which one

How to compare and assess different 'normalization' methods?

Comparison against a ground truth

But you have millions of numbers – need to choose the metric that measures deviation from truth.

FN/FP: do you find all the differentially expressed genes, and do you not find non-d.e. genes?

qualitative/quantitative: how well do you estimate abundance, fold-change?

Spike-In and Dilution series

… great, but how representative are they of other data?

Implicitely, from resampling the actual experiment of interest

… but isn't that too much like Munchhausen?

evaluation: a benchmark for Affymetrix genechip expression measures

o Data:

Spike-in series: from Affymetrix 59 x HGU95A,
16 genes, 14 concentrations, complex background
Dilution series: from GeneLogic 60 x HGU95Av2,
liver & CNS cRNA in different proportions and amounts

o Benchmark:

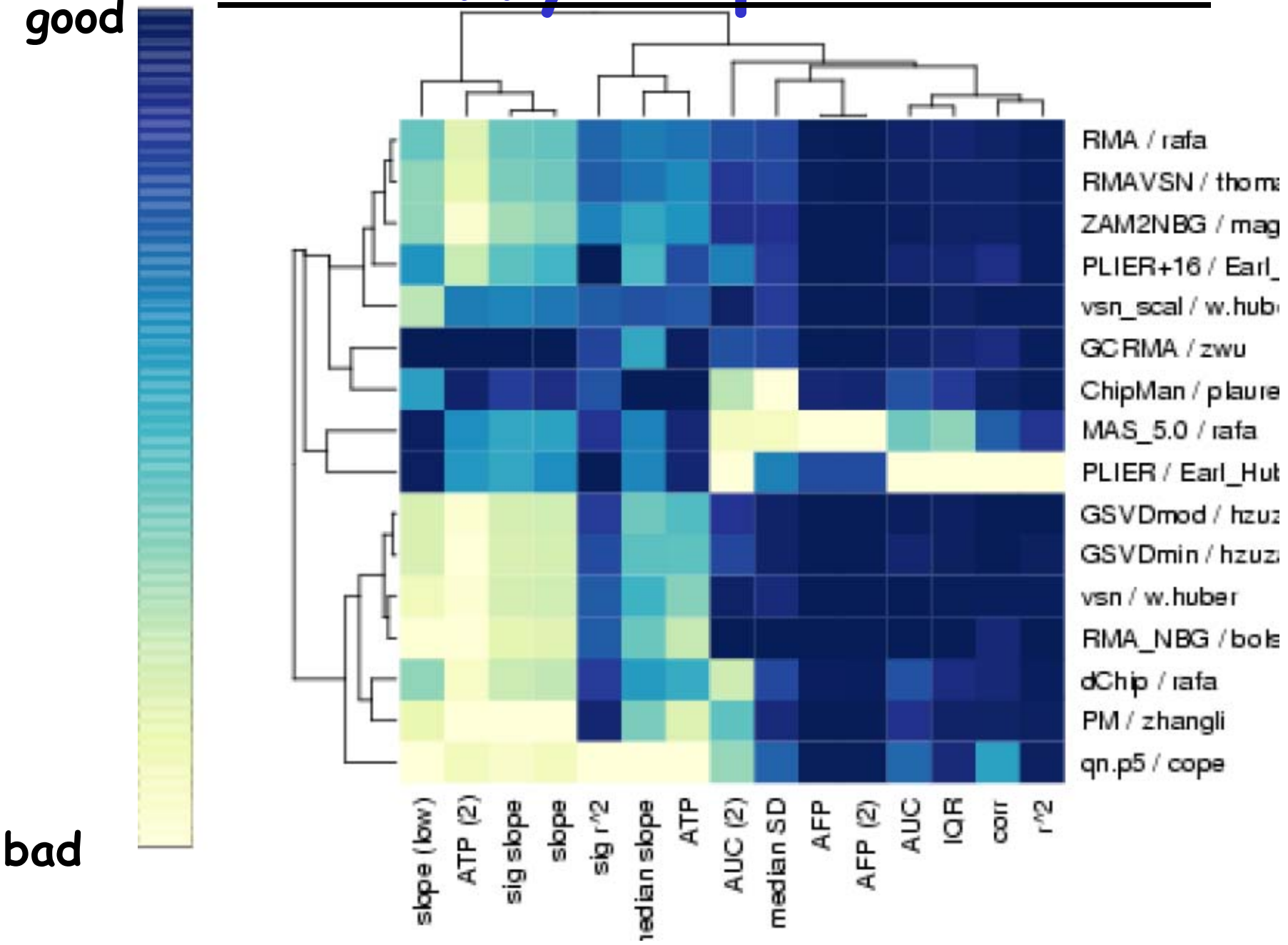
15 quality measures regarding
-reproducibility
-sensitivity
-specificity
Put together by Rafael Irizarry (Johns Hopkins)
http://affycomp.biostat.jhsph.edu

evaluation: a benchmark for Affymetrix genechip expression measures

- Package affycomp (on Bioconductor)

- Online competition, accepts contributions via webserver

affycomp results



ROC curves

Figure 5a

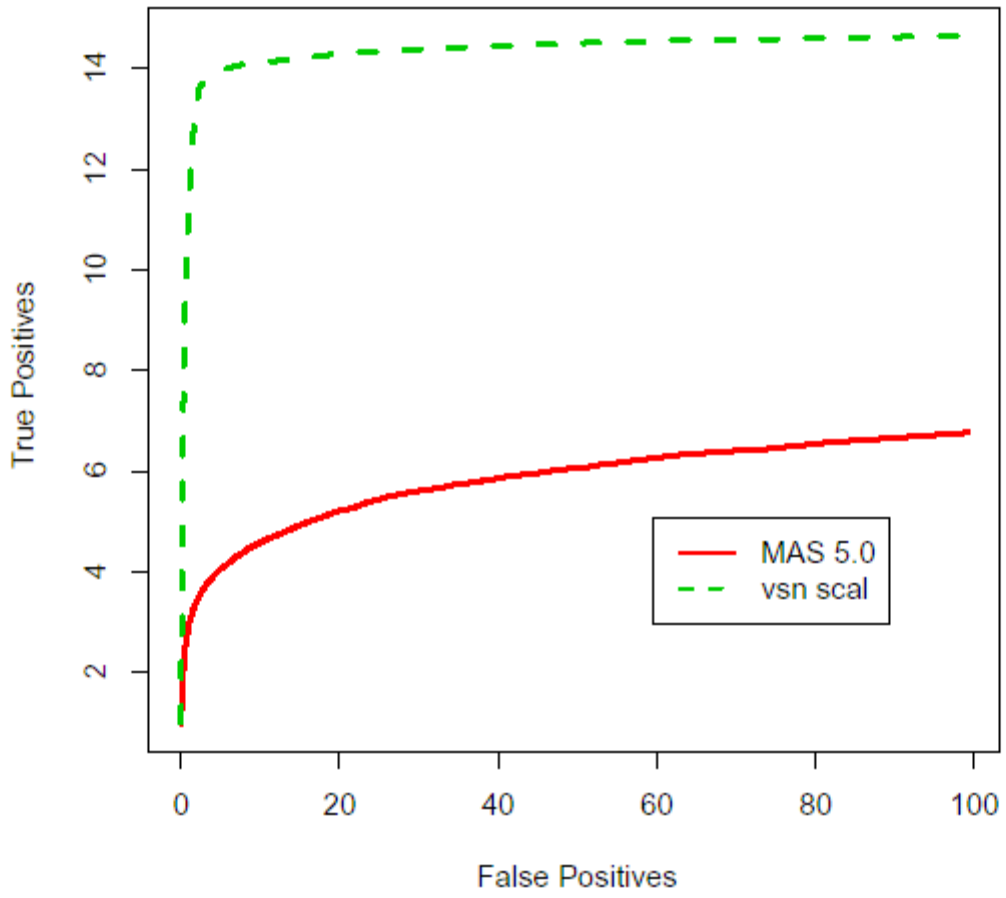


Figure 5a): A typical identification rule for differential expression filters genes with fold change exceeding a given threshold. This figure shows average ROC curves which offer a graphical representation of both specificity and sensitivity for such a detection rule.Average ROC curves based on comparisons with nominal fold changes ranging from 2 to 4096. b) As a) but with nominal fold changes equal to 2.

Limitations

Affymetrix preprocessing involves

(1) PM,MM-synthesis
(2) calibration, transformation
(3) probe set summarization

'vsn-scal' used

(1) ignore MM
(2) vsn
(3) medianpolish (as in RMA, similar to dChip)

This can be improved

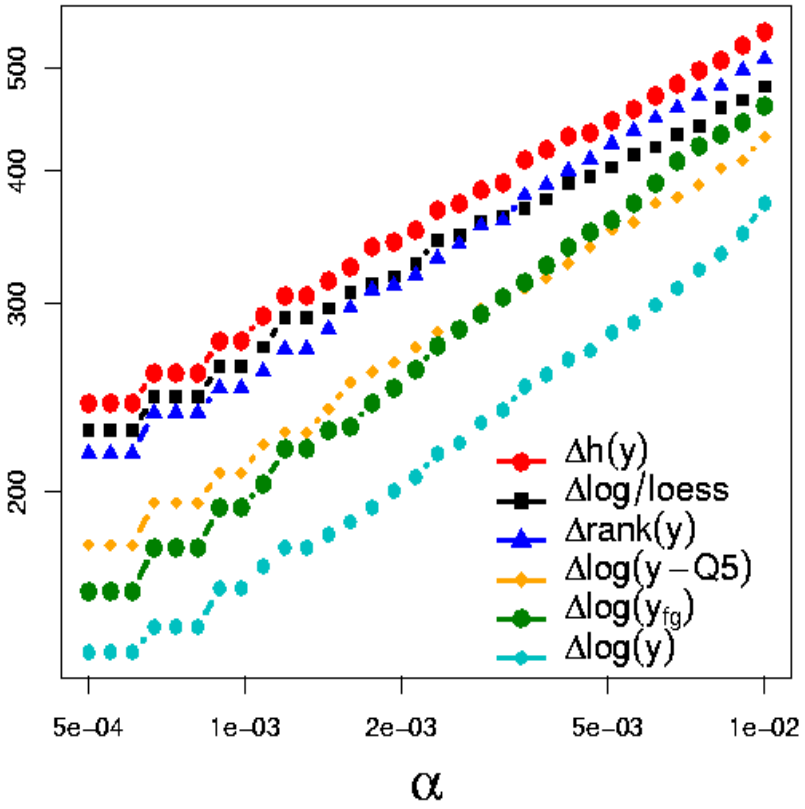
(1) use MM! (but just not simply PM-MM)
(2) stratify by physical probe properties

Resampling method: sensitivity / specificity in detecting differential abundance

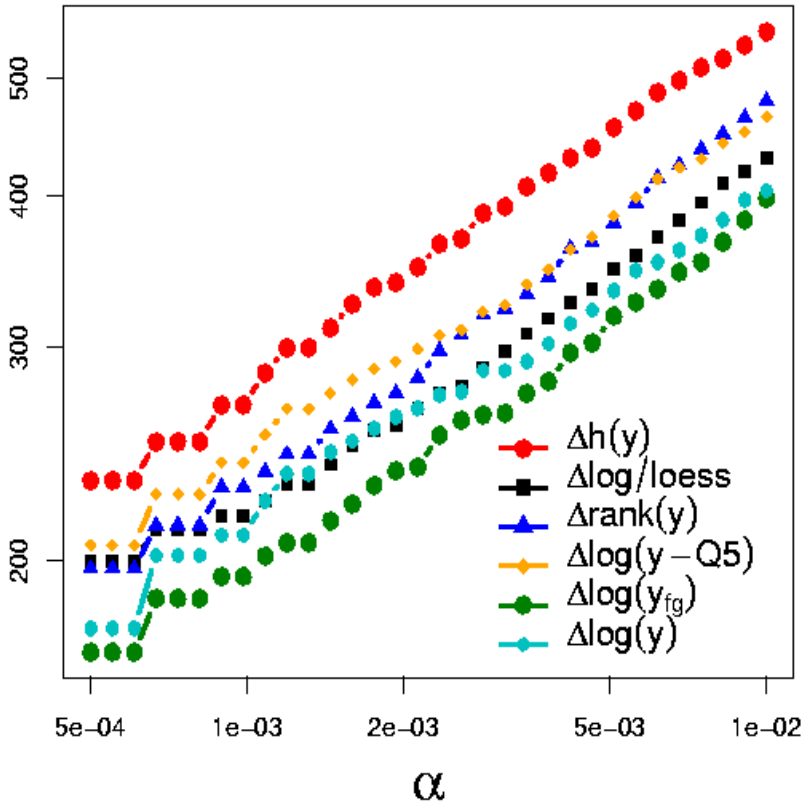
- Data: paired tumor/normal tissue from 19 kidney cancers, in color flip duplicates on 38 cDNA slides à 4000 genes.
- 6 different strategies for normalization and quantification of differential abundance
- Calculate for each gene & each method:
  t-statistics, permutation-*p*
- For threshold α, compare the number of genes the different methods find, $\#\{p_i \mid p_i \leq \alpha\}$

sensitivity vs specificity

one-sided test for up



one-sided test for down



Summary

Measuring microarray data is a complex chain of biochemical reactions and physical measurements.

Systematic and stochastic errors

Calibration and error models

Parameter estimation

Getting preprocessing right is prerequisite for getting reasonable results in the end

Improving preprocessing is just like any other technology improvement

How to choose from the plethora of methods?

What's next

Exercises on data import, diagnostic plots, quality criteria, comparing normalization methods

Lecture on **quality control**, probe set summaries, hybridization physics

Thank you