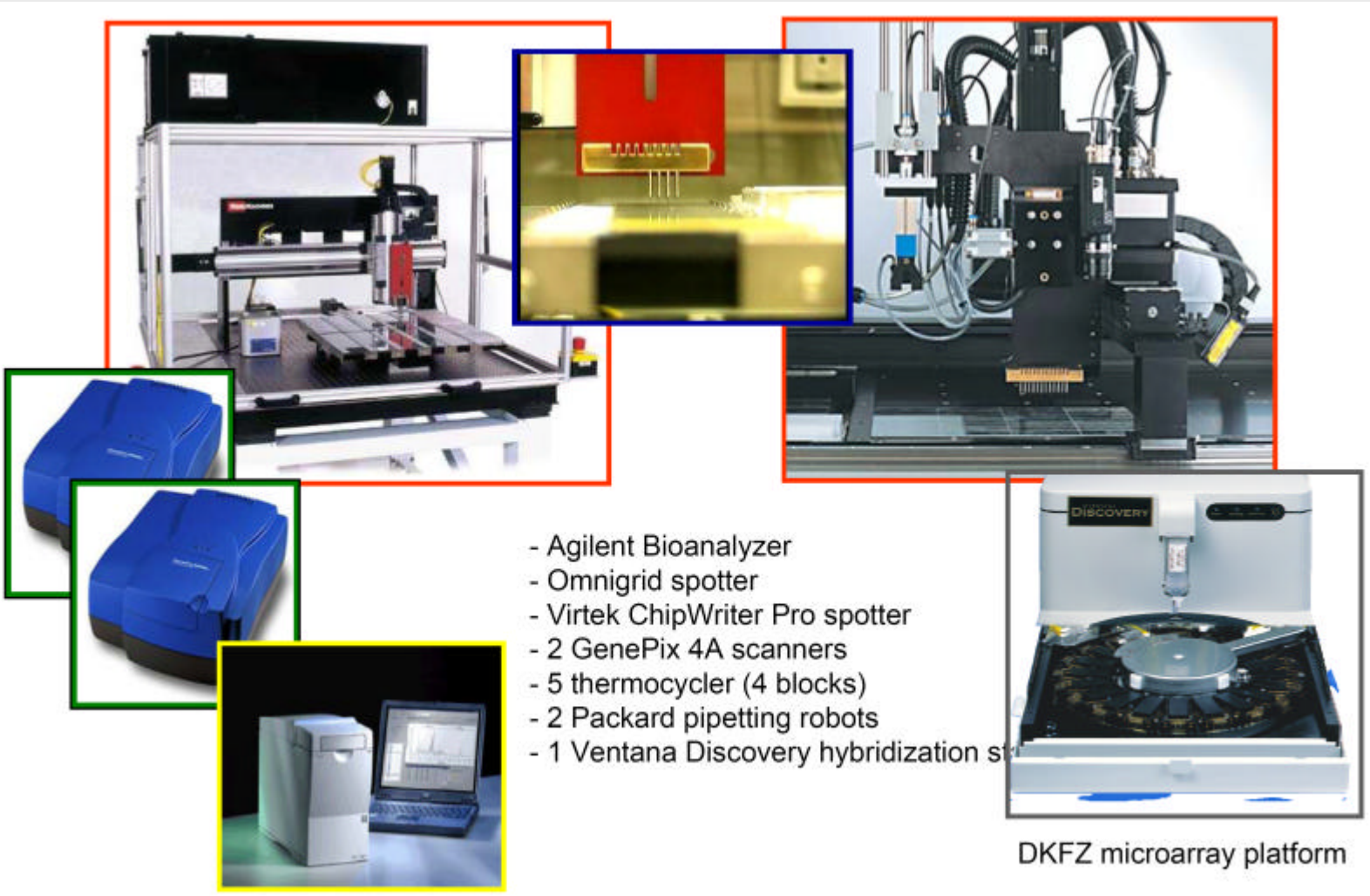
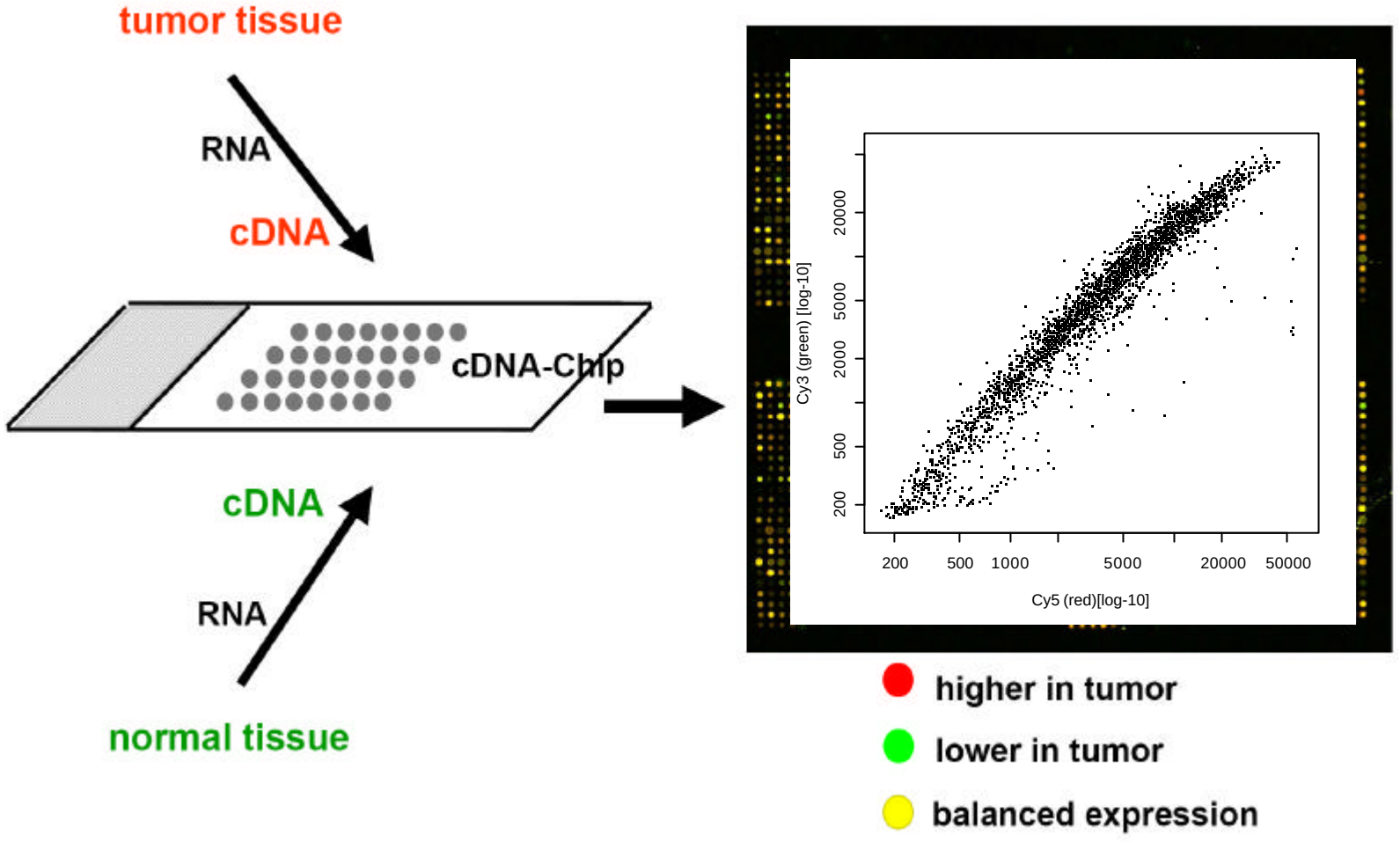
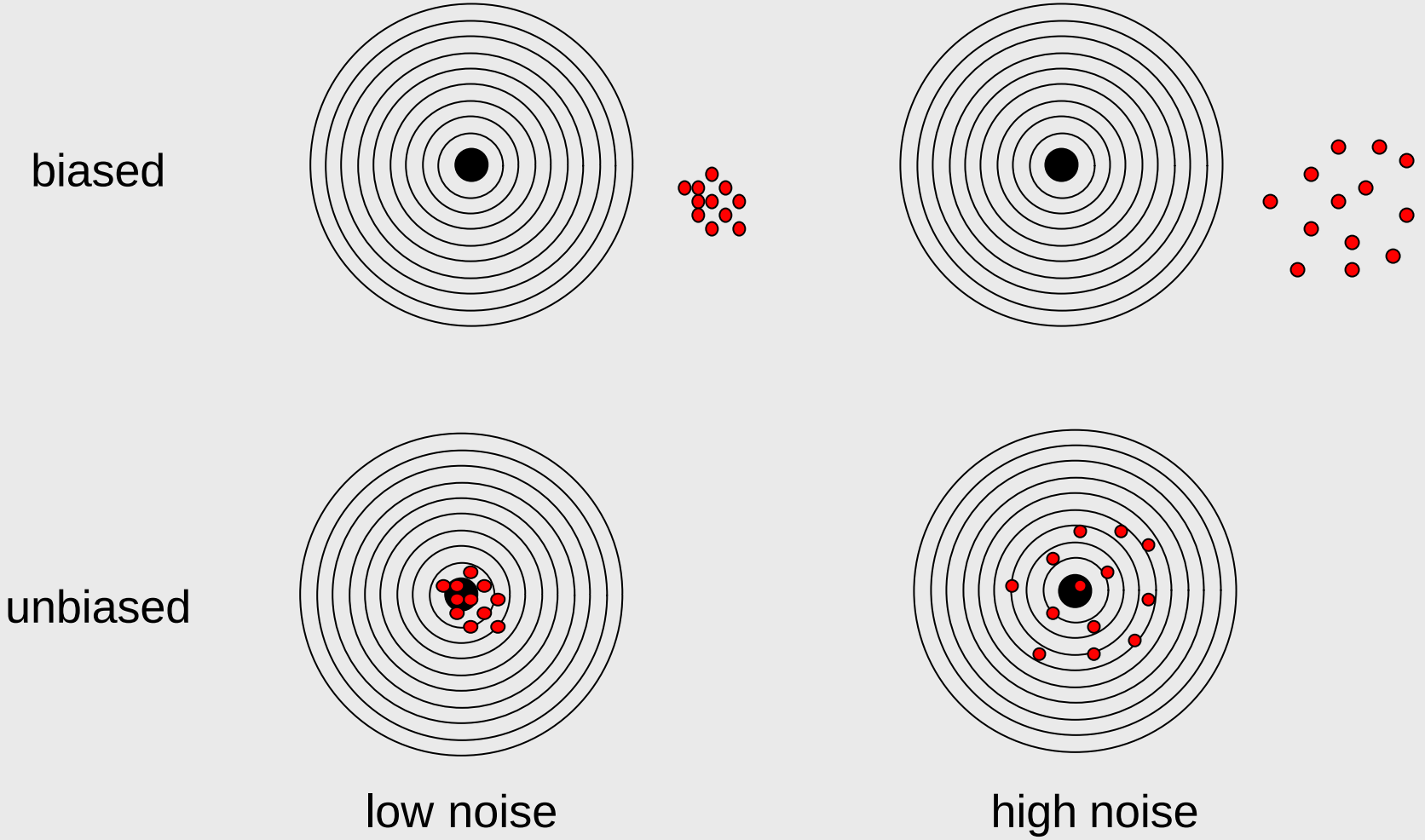
cDNA chips
Quality control and pre-processing

Ulrich Mansmann
Department of Medical Biometrics and Informatics
University of Heidelberg


- Agilent Bioanalyzer
- Omnigrid spotter
- Virtek ChipWriter Pro spotter
- 2 GenePix 4A scanners
- 5 thermocycler (4 blocks)
- 2 Packard pipetting robots
- 1 Ventana Discovery hybridization st
DISCOVERY
DKFZ microarray platform


tumor tissue
RNA
cDNA
cDNA-Chip
cDNA
RNA
normal tissue
Cy3 (green) [log-10]
200
500
2000
5000
20000
200
500
1000
5000
20000
50000
Cy5 (red)[log-10]
higher in tumor
lower in tumor
balanced expression

Measurements should be unbiased and precise



Quality control: Noise and reliable signal

- Is the signal dominated by noise? Acceptable amount of noise? Quantifying noise?
- Quantifying quality of a signal;
- Guidelines for reasonable thresholds on the quality of a signal;
- Defining strategies for exclusion of probes;
- Probe level: quality of the expression measurement on one particular array
- Gene level: quality of the expression measurement accross all arrays
- Array level: quality of the expression measurement on one particular glass slide

Swirl Data

This experiment was carried out using zebrafish as a model organism to study early development in vertebrates. Swirl is a point mutant in the BMP2 gene that affects the dorsal/ventral body axis. Ventral fates such as blood are reduced, whereas dorsal structures such as somites and notochord are expanded.

A goal of the Swirl experiment is to identify genes with altered expression in the swirl mutant compared to wild-type zebrafish. Two sets of dye-swap experiments were performed, for a total of four replicate hybridizations. For each of these hybridizations, target cDNA from the swirl mutant was labeled using one of the Cy3 or Cy5 dyes and the target cDNA wild-type mutant was labeled using the other dye.

Target cDNA was hybridized to microarrays containing 8,448 cDNA probes, including 768 controls spots (e.g. negative, positive, and normalization controls spots). Microarrays were printed using 4 times 4 print-tips and are thus partitioned into a 4 times 4 grid matrix. Each grid consists of a 22 times 24 spot matrix that was printed with a single print-tip. Here, spot row and plate coordinates should coincide, as each row of spots corresponds to probe sequences from the same 384 well-plate.

Probe-level quality control

- Individual spots printed on the slide
- Sources: faulty printing, uneven distribution, contamination with debris, size of signal relative to noise, poorly measured spots;
- Visual inspection: hairs, dust, scratches, air bubbles, dark regins, regions with haze
  $\rightarrow$ set points to NA
  $\rightarrow$ local normalization procedures which account for regional idiosyncrasies.

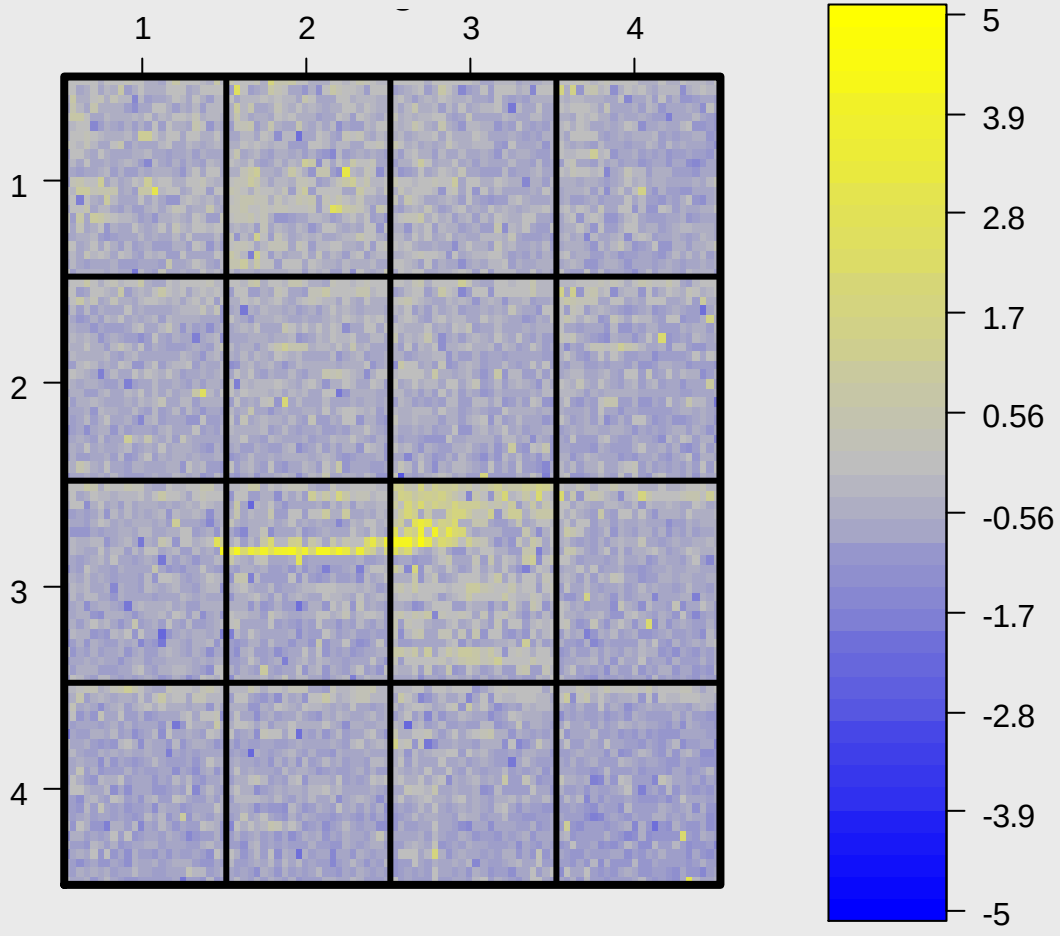
Visual inspection

4 x 4 sectors

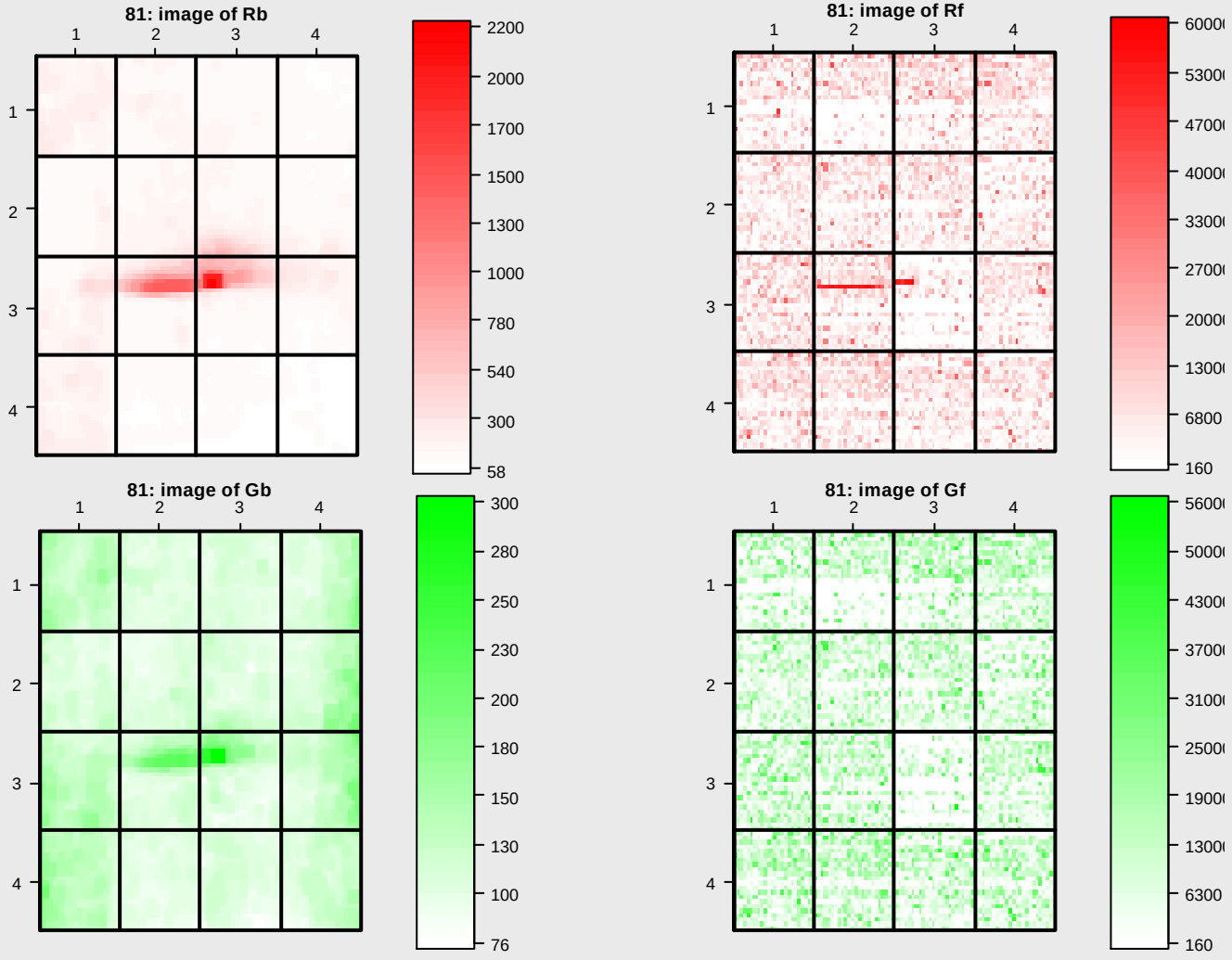
Sector:
24 rows
22 columns

8448 spots

Mean signal intensity



Visual inspection



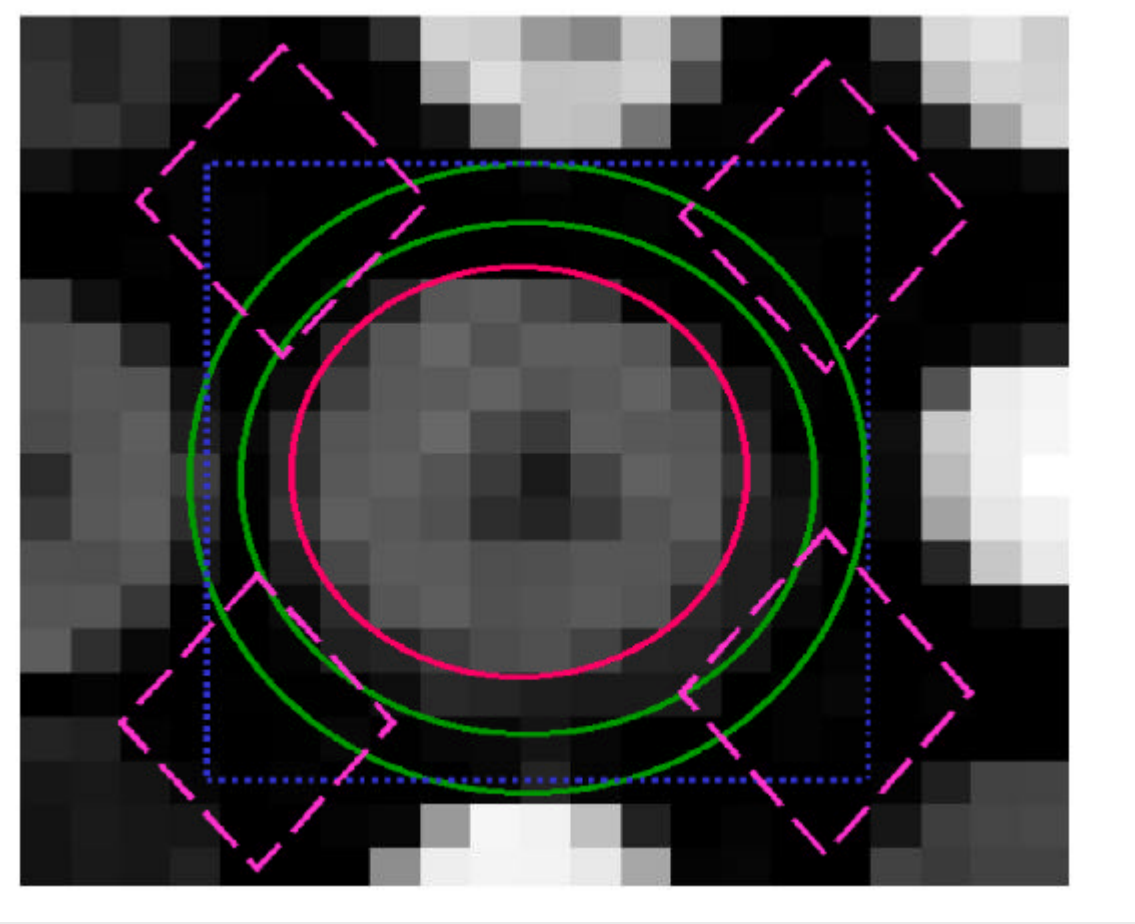
Probe-level quality control

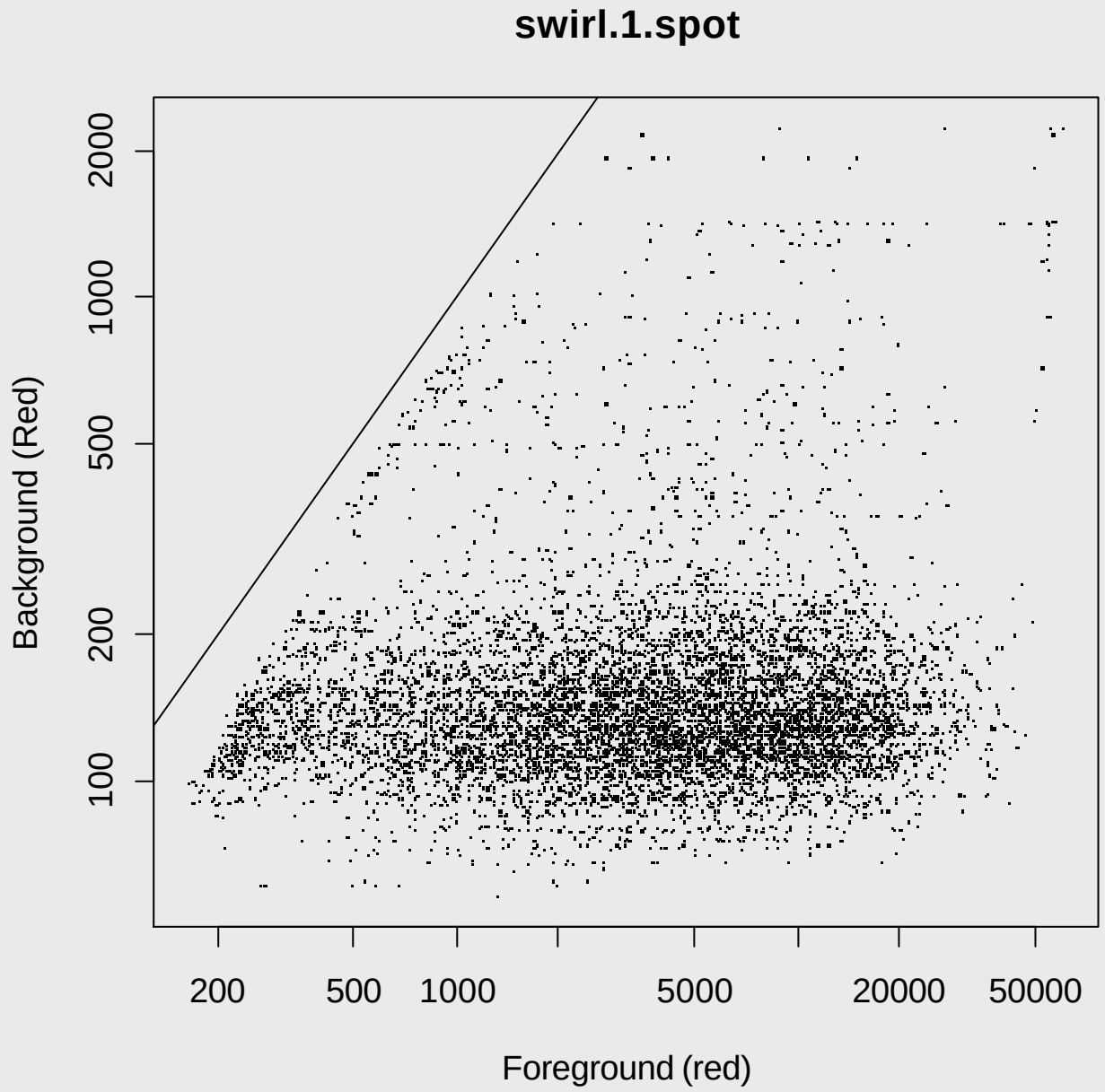
- Individual spots printed on the slide
- Sources: faulty printing, uneven distribution, contamination with debris, size of signal relative to noise, poorly measured spots;
- Visual inspection
- Spot quality
  Brightness: foreground/background ratio
  Uniformity: variation in pixel intensities and ratios of intensities within a spot
  Morphology: area, perimeter, circularity.
  Spot Size: number of foreground pixels

GenePix

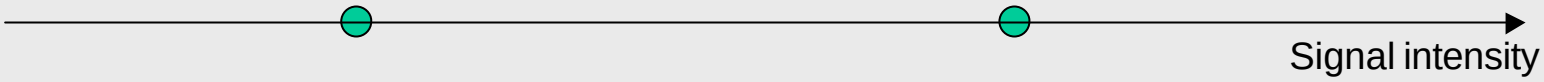
QuantArray

ScanAlyse




swirl.1.spot
Background (Red)
100
200
500
1000
2000
200
500
1000
5000
20000
50000
Foreground (red)

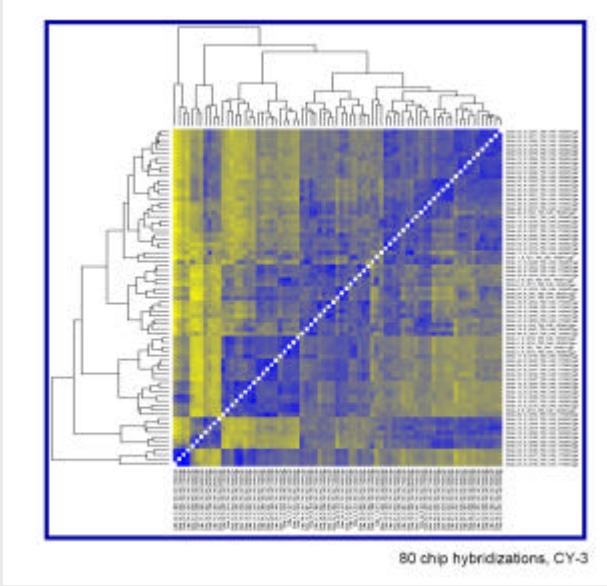
Weak Signal

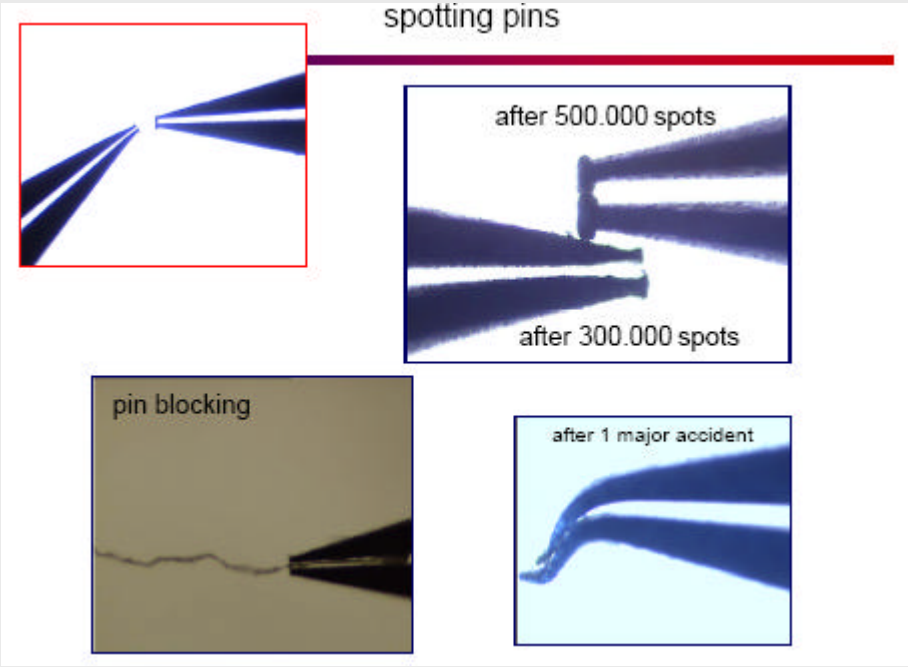


$\lambda_{low} = 200$ $\lambda_{high} = 500$

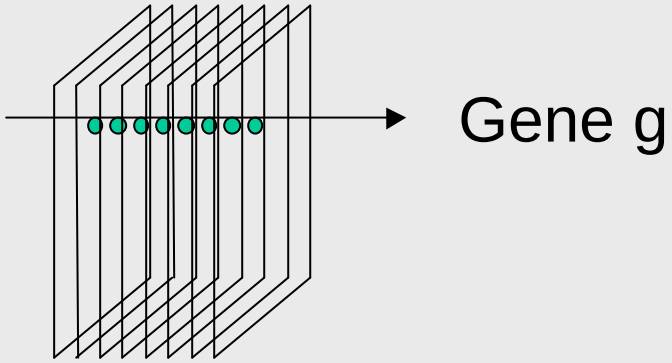
CY5 channel	CY3 channel	Measure
FR-BR < λ_{low}	FG-BG < λ_{high}	exclude
FR-BR < λ_{high}	FG-BG < λ_{low}	exclude
FR-BR < λ_{low}	FG-BG > λ_{high}	$Log_2[\lambda_{low} / (FG-BG)]$
FR-BR > λ_{high}	FG-BG < λ_{low}	$Log_2[(FR-BR) / \lambda_{low}]$
FR-BR > λ_{high}	FG-BG > λ_{high}	$Log_2[(FR-BR) / (FG-BG)]$

$Log_2[510/490] = 0.057$, $Log_2[30/10] = 1.58$, FG=2030, BG=2000


80 chip hybridizations, CY-3


spotting pins
after 500.000 spots
after 300.000 spots
pin blocking
after 1 major accident

Gene-level quality control



- Test of poor hybridization in the reference channel
- Test for exclusion of low variance genes.

Gene-level quality control: Poor Hybridization and Printing

- Some probes will not hybridize well to the target RNA
- Printing problems such that all spots of a given inventory well have poor quality.
- A well may be of bad quality – contamination
- Genes with a consistently low signal in the reference channel are suspicious: Median of the background adjusted signal < 200

Gene-level quality control: Probe quality control based on duplicated spots

- Printing different probes that target the same gene or printing multiple copies of the same probe.
- Mean squared difference of $\log_2$ ratios between spot r and s:

$$MSDLR = \Sigma(x_{jr} - x_{js})^2/J \quad \text{sum over arrays } j = 1, \ldots, J$$

  recommended threshold to assess disagreement: MSDLR > 1
- Disagreement between copies: printing problems, contamination, mislabelling. Not easy if there are only 2 or 3 slides.
- Jenssen et al (2002) Nucleic Acid Res, 30: 3235-3244. Theoretical background

Gene-level quality control: Low variance genes

- Good for normalization, but uninformative for the analysis
- Threshold for fold change:
  $\log_2(\max) - \log_2(\min)$, this difference may depend on sample size
  taking sample size into account: $\log_2(q_{0.95}) - \log_2(q_{0.05})$
- Variance based criterion

Array-level quality control

- Problems: array fabrication defect
  problem with RNA extraction
  failed labelling reaction
  poor hybridization conditions
  faulty scanner
- Quality measures:
  Percentage of spots with no signal (~30% exlcuded spots)
  Range of intensities
  (Av. Foreground)/(Av. Background) > 3 in both channels
  Distribution of spot signal area
  Amount of adjustment needed: signals have to substantially changed to make slides comparable.

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

W. Huber

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

W. Huber

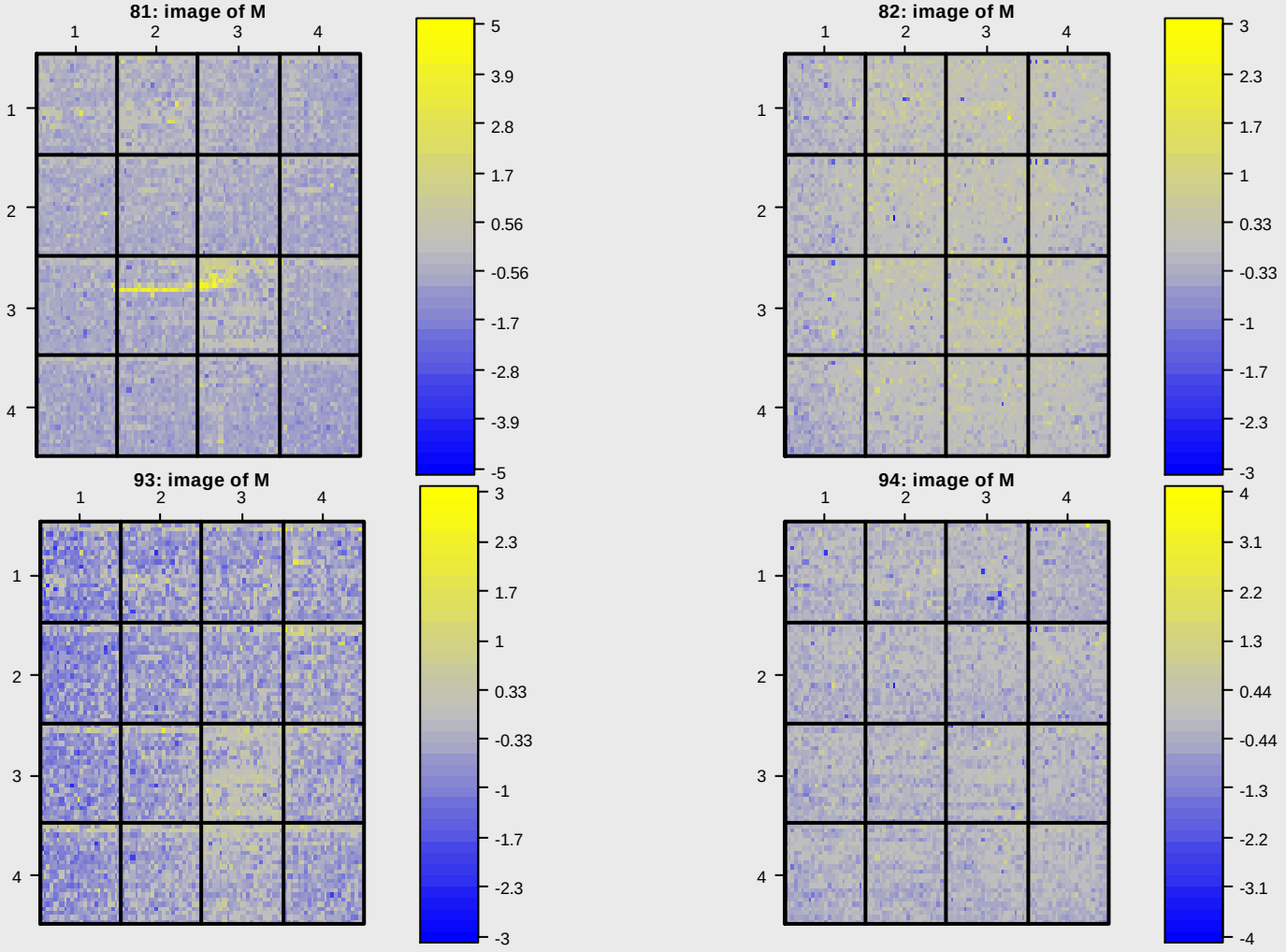
Normalization

- Identify and remove sources of systematic variation, other than differential expression, in the measured fluorescence intensities.
- Normalization is necessary before each analysis, in order to ensure that differences in intensities are indeed due to differential expression and not experimental artefacts.
- Location normalisation: corrects for spatial or dye bias
- Scale normalisation: homogenises the variability across arrays

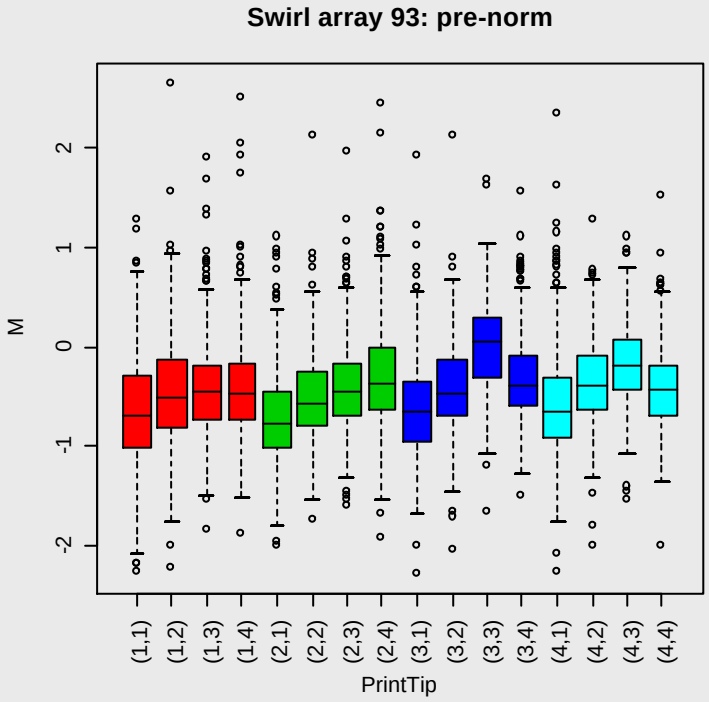
  $MAD = median\{ |x_1-m|, \ldots, |x_n-m| \}$
- Normalised log-intensity ratios are given by

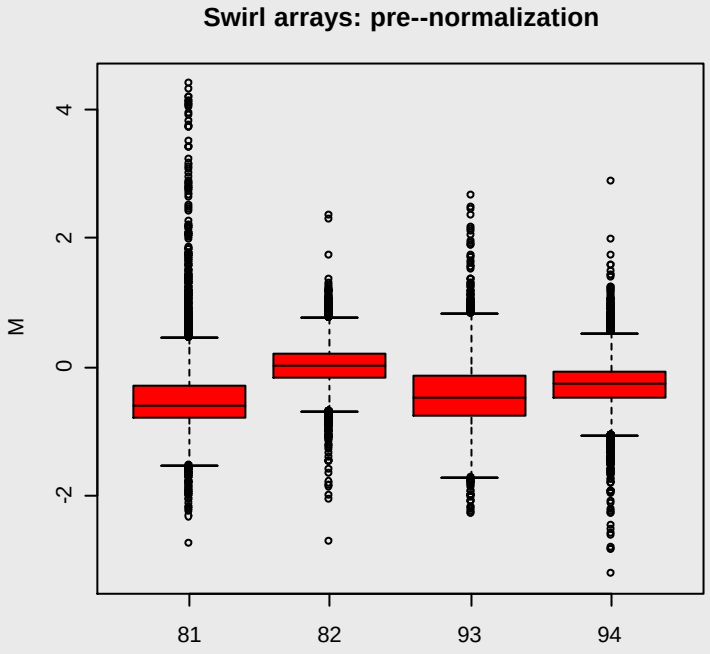
  $M_{norm} = (M-loc)/s$
- Normalisation: within arrays (marray) or between arrays (vsn),
  single channels between arrays,
  log expression ratios, etc

Swirl Data

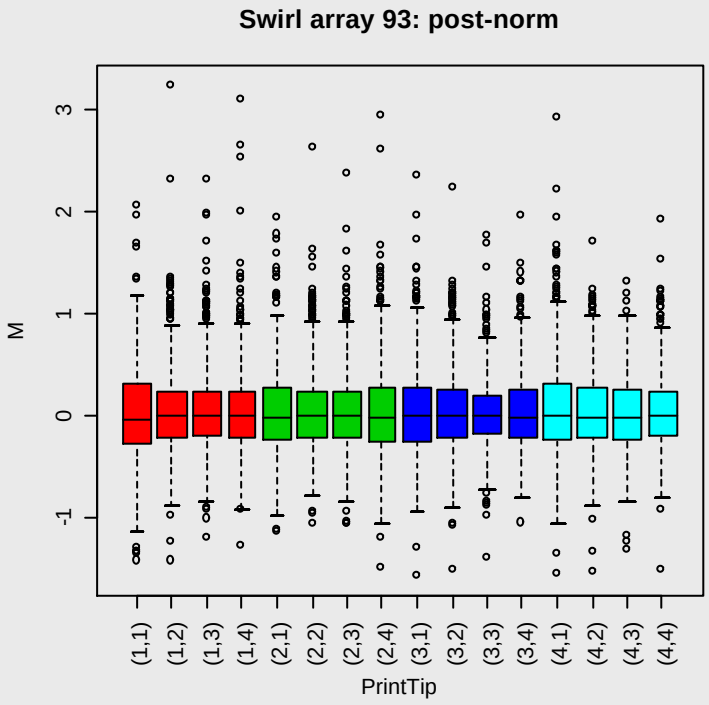


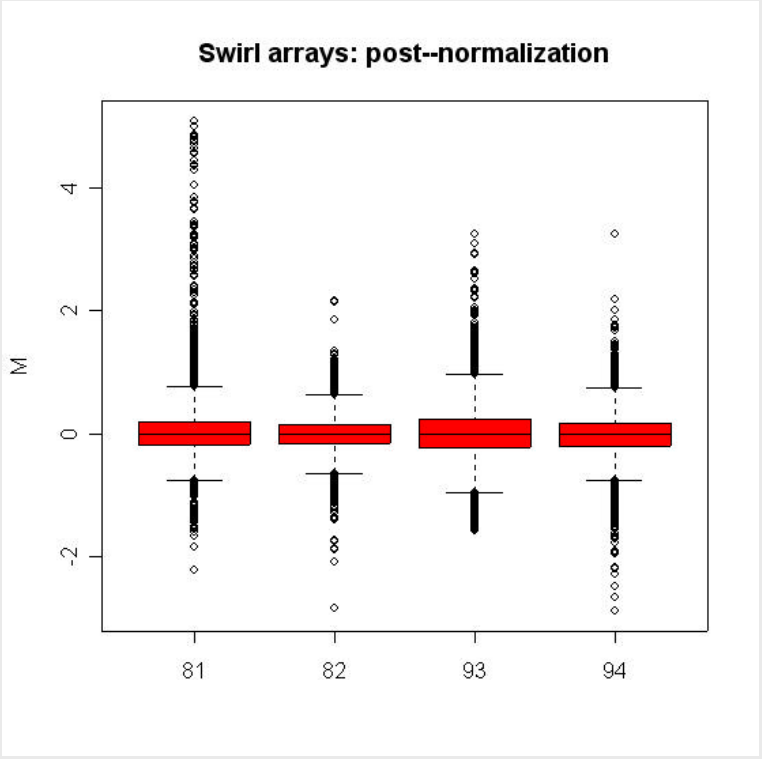
marray: Pre Normalization



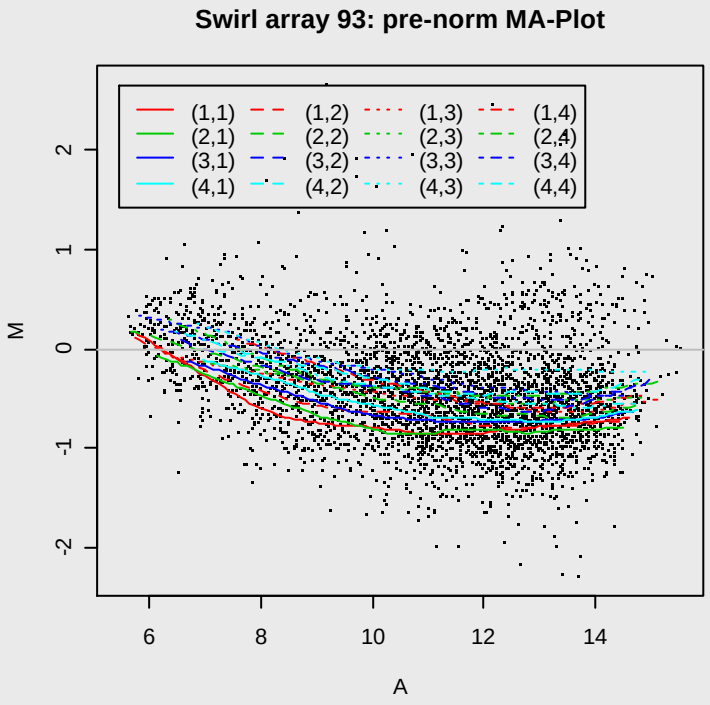


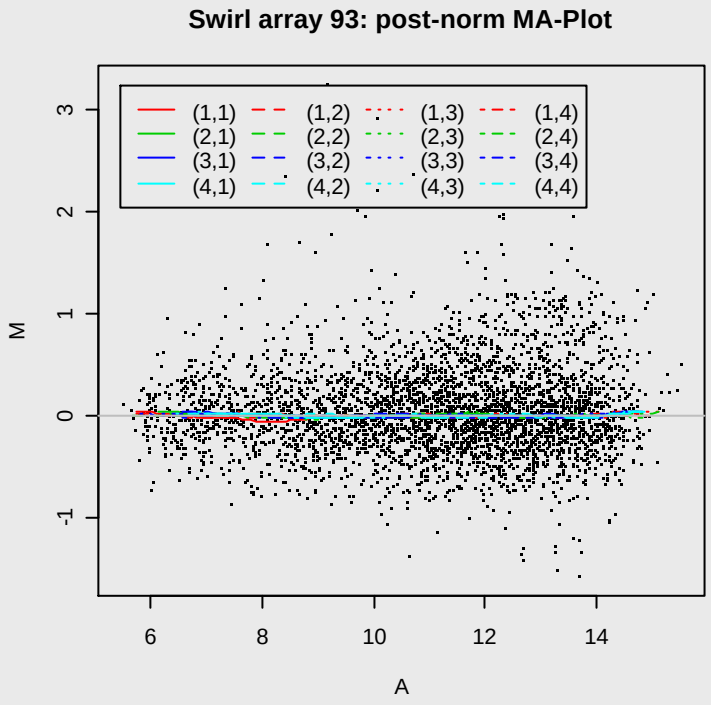
marray: Post Normalization



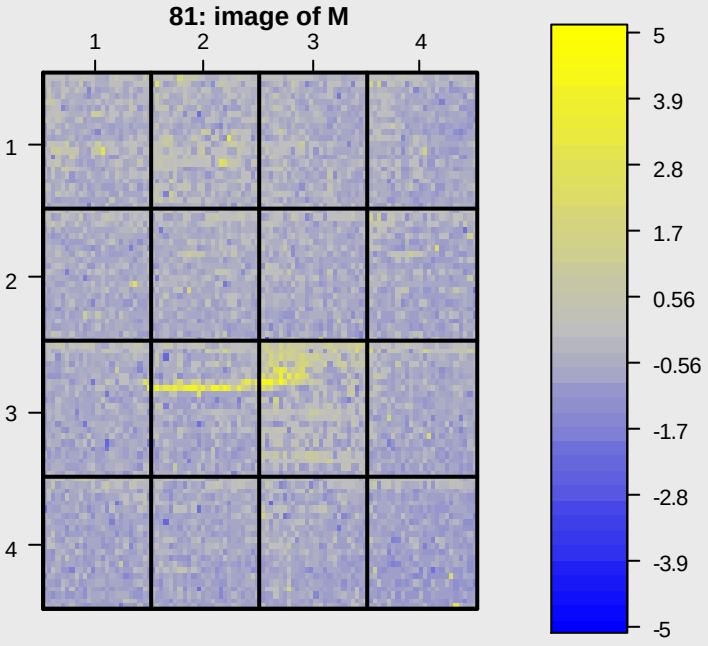


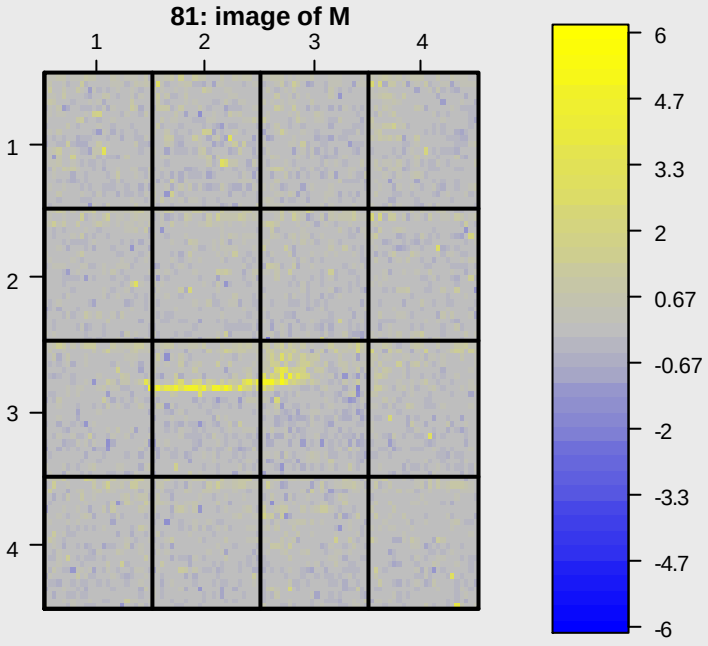
marray: MA Plots





marray:
Signals - raw versus normalised





VSN: model and theory

- Huber et al. (2002) Bioinformatics, 18:S96–S104
- Model for measured probe intensity
  Rocke DM, Durbin B (2001) Journal of Computational Biology, 8:557–569
- log-transformation is replaced by a transformation (arcsinh) based on theoretical grounds.
- Estimation of transformation parameters (location, scale) based on ML paradigm and numerically solved by a leats trimmed sum of squares regression.
- vsn–normalised data behaves close to the normal distribution
- Following slides are borrowed from Wolfgang Huber

u modeling ansatz

measured intensity = offset + gain ´ true abundance

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

$$a_{ik} = a_i + e_{ik}$$

a_i per-sample offset

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

"additive noise"

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

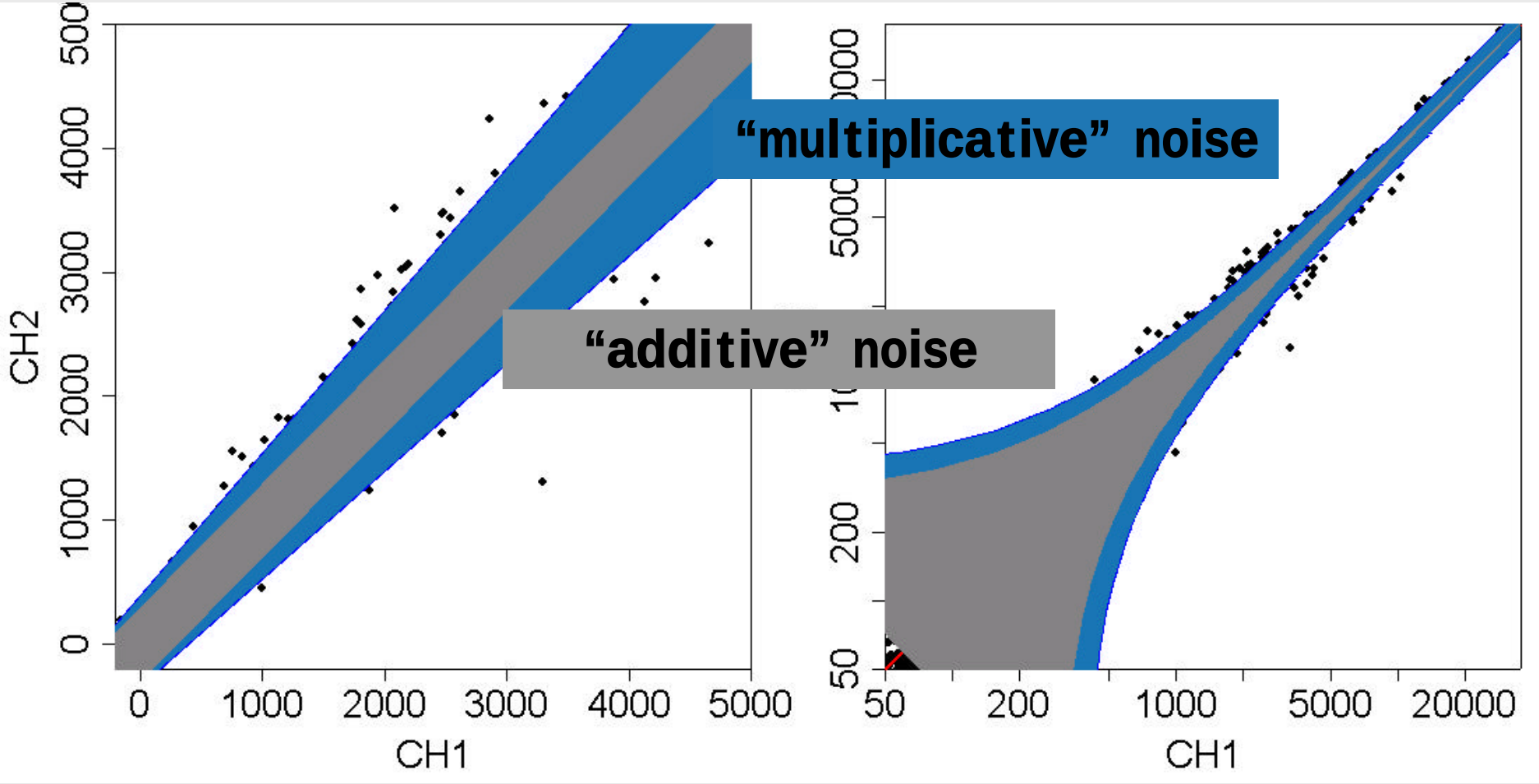
b_i per-sample normalization factor

b_k sequence-wise probe efficiency

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

"multiplicative noise"

u The two-component model



B. Durbin, D. Rocke, JCB 2001

u variance stabilizing transformations

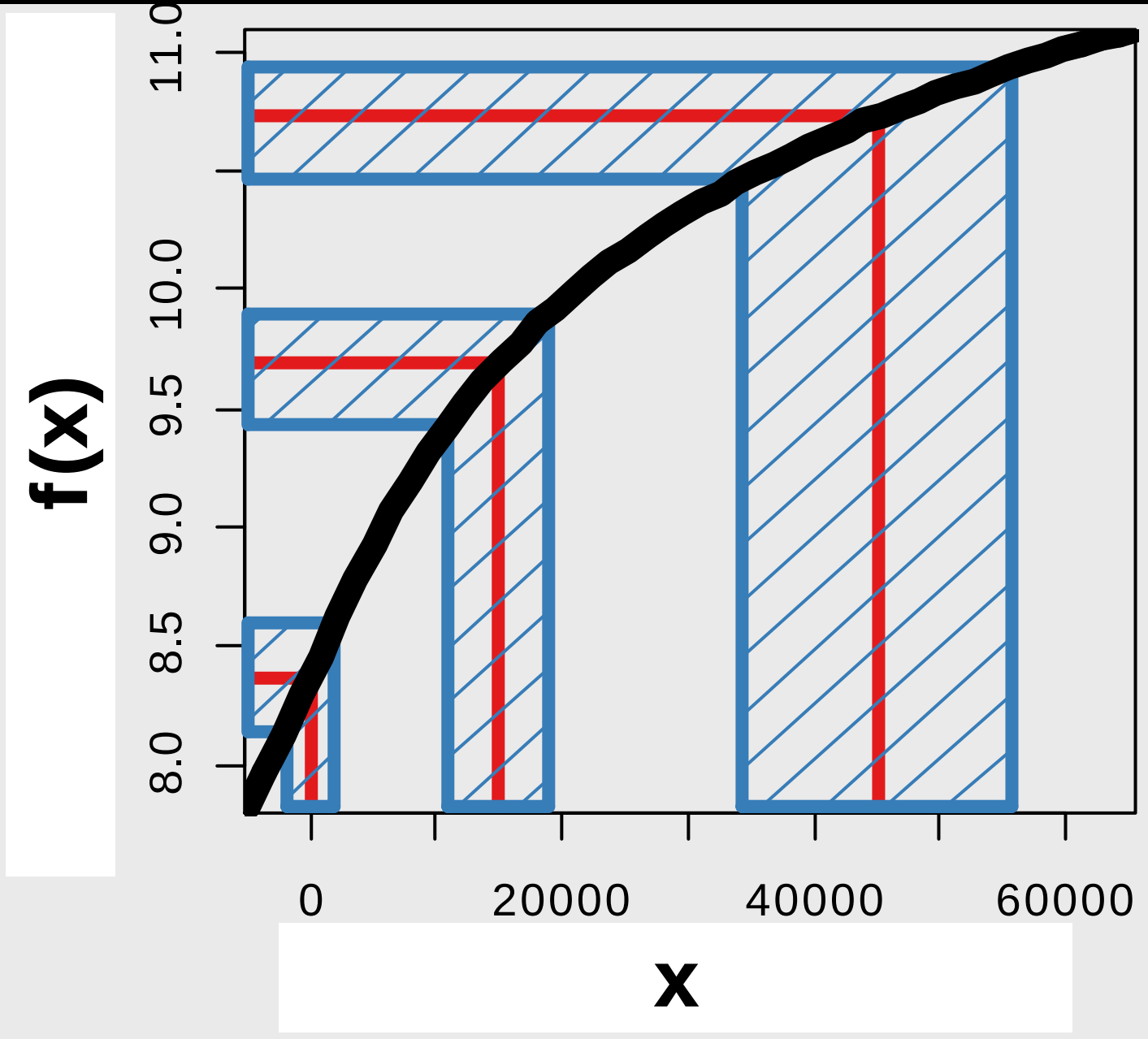
X_u a family of random variables with $EX_u = u$, $\mathrm{Var}X_u = v(u)$. Define

$$f(x) = \int_0^x \frac{1}{\sqrt{v(u)}}\,du$$

Þ var $f(X_u)$ » independent of u

derivation: linear approximation

u variance stabilizing transformations



u variance stabilizing transformations

$$f(x) = \int_0^x \frac{1}{\sqrt{v(u)}}\,du$$

1.) constant variance ('additive') $v(u) = s^2$ Þ f µ u

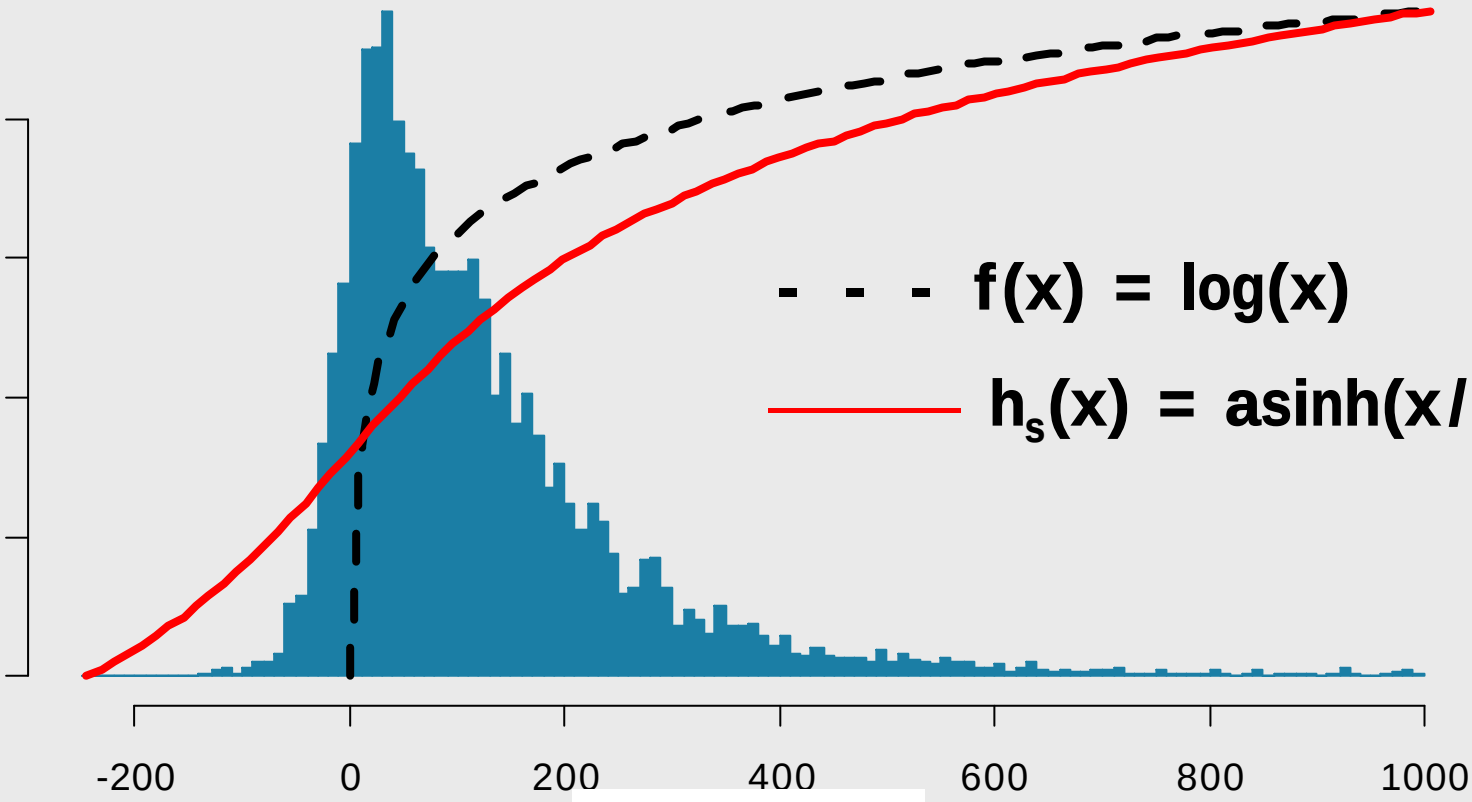
2.) constant CV ('multiplicative') $v(u)$ µ u^2 Þ f µ $\log u$

3.) offset $v(u)$ µ $(u + u_0)^2$ Þ f µ $\log(u + u_0)$

4.) additive and multiplicative

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

u 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

parameter estimation

$$\operatorname{arsinh}\frac{Y_{ki} - a_i}{b_i} = m_k + e_{ki}, \qquad e_{ki} : N(0, c^2)$$

- o maximum lik … d
- – but sensitiv…
- o model holds … differentially …
- o robust varia…
  Trimmed Sum…
- o works as lo… differentially …

measured intensity = offset + gain * true abundance

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

$a_{ik} = a_i + L_{ik} + e_{ik}$

a_i per-sample offset

L_{ik} local background provided by image analysis

$e_{ik} \sim N(0, b_i^2 s_1^2)$ "additive noise"

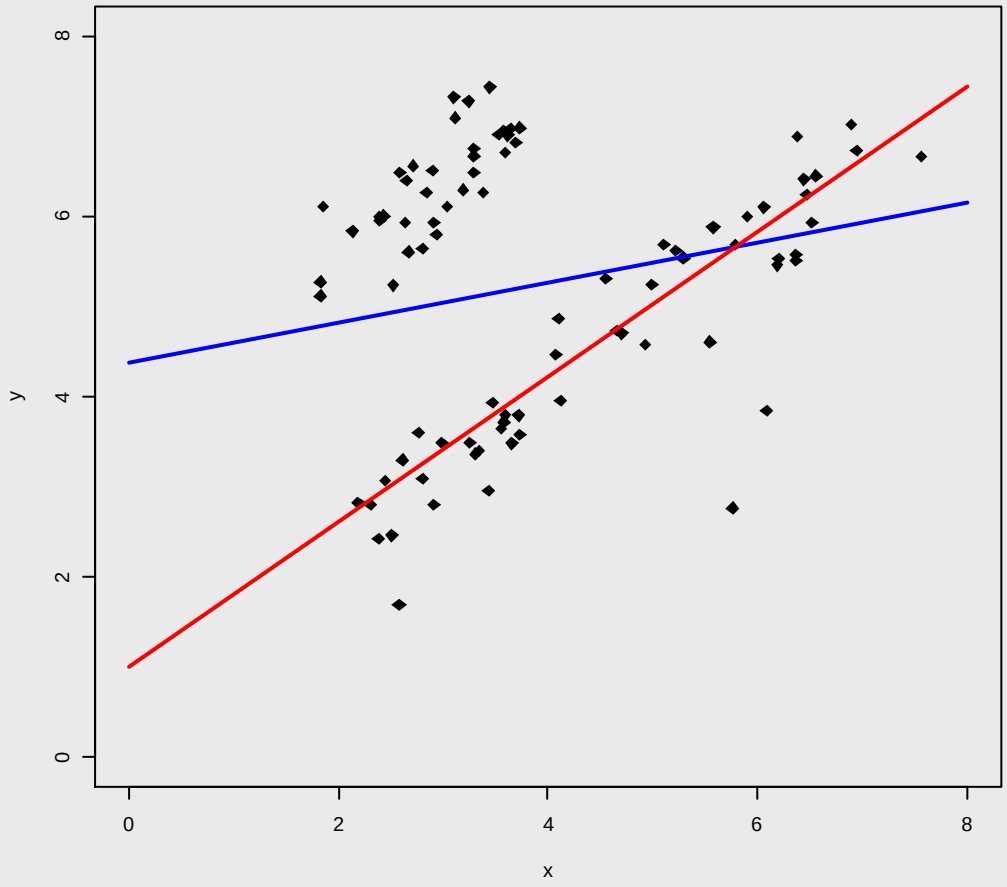
$b_{ik} = b_i \; b_k \exp(h_{ik})$

b_i per-sample normalization factor

b_k sequence-wise labeling efficiency

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

Least trimmed sum of squares regression



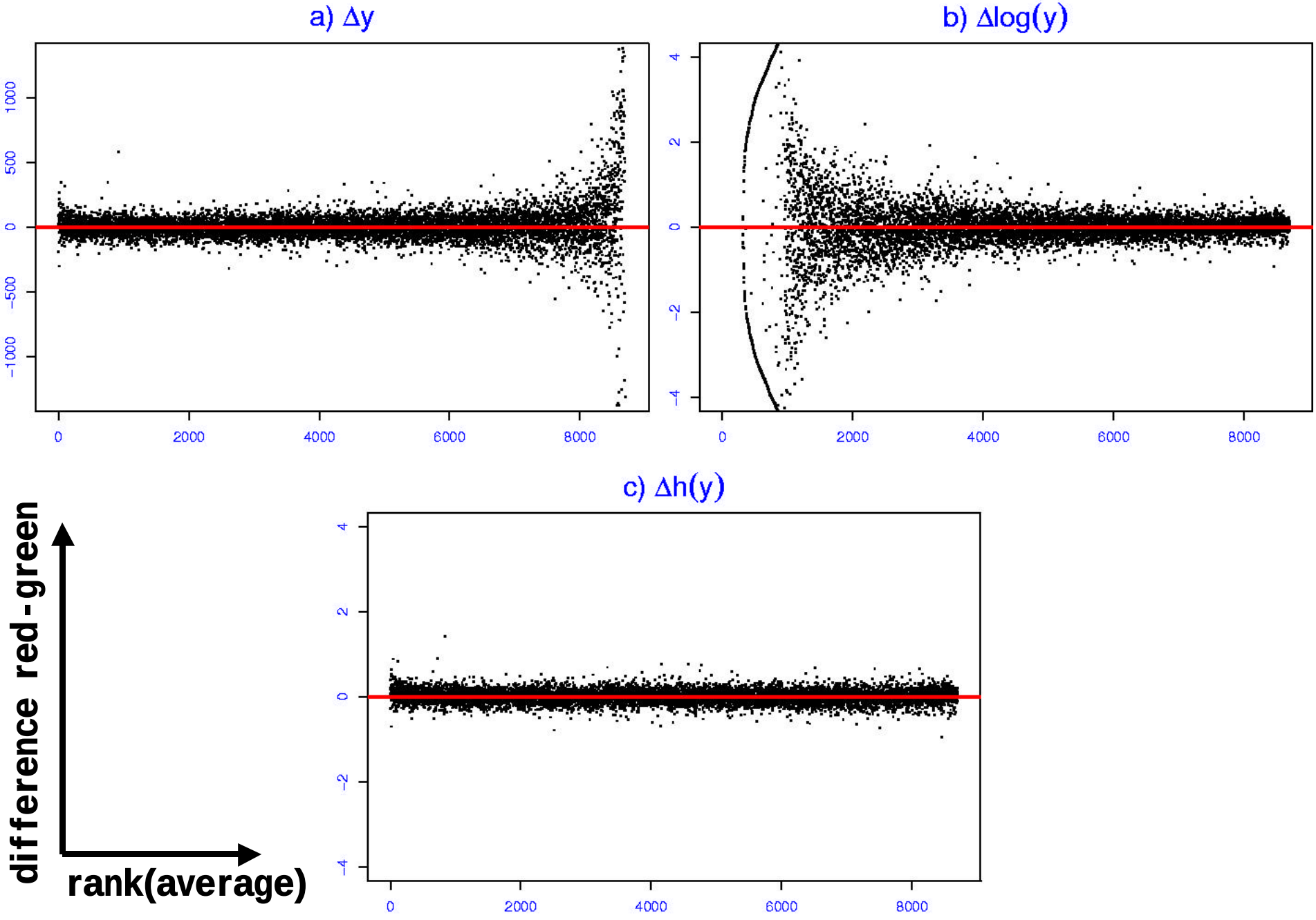
minimize

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

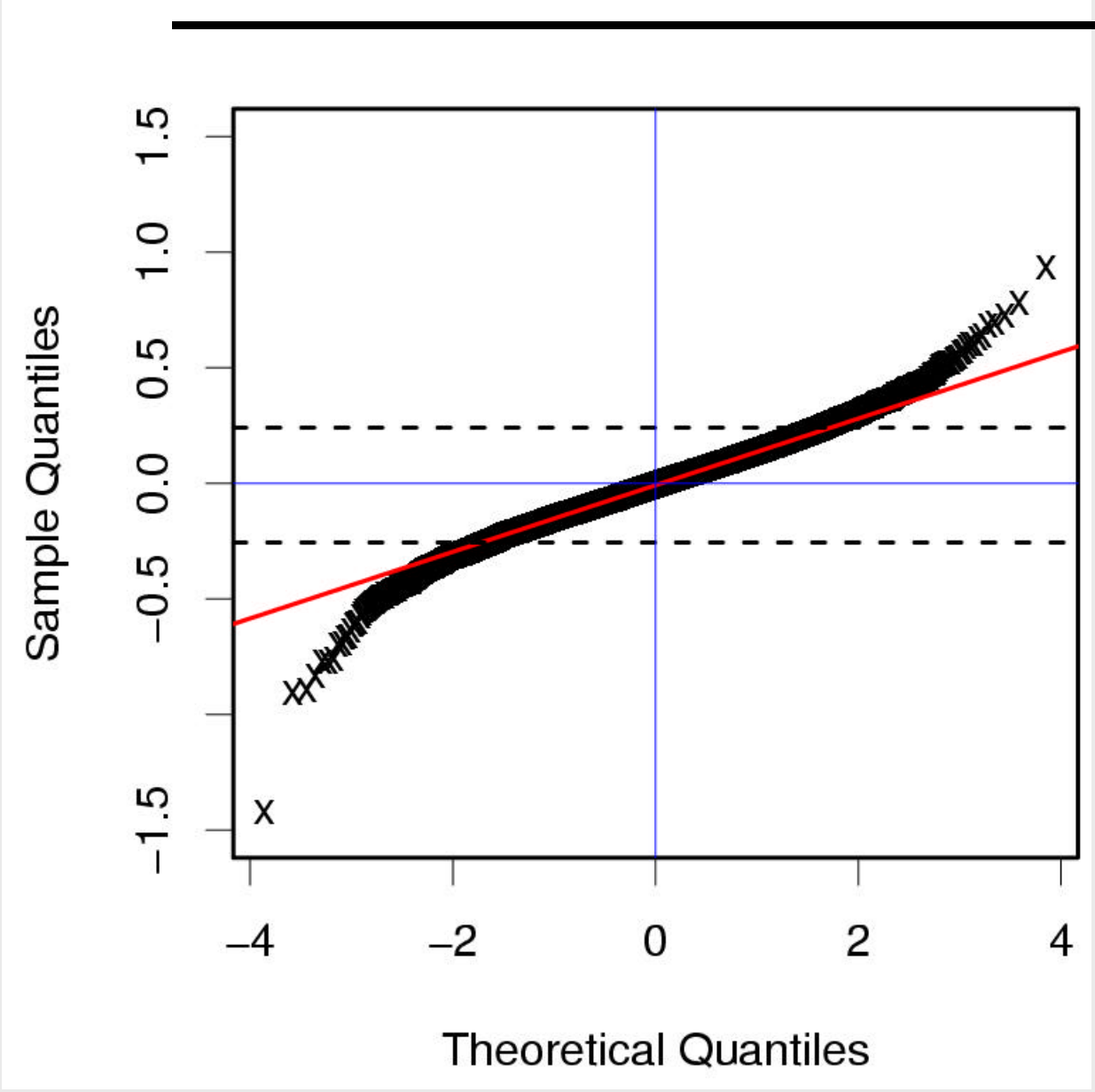
P. Rousseeuw, 1980s

- least sum of squares
- least trimmed sum of squares

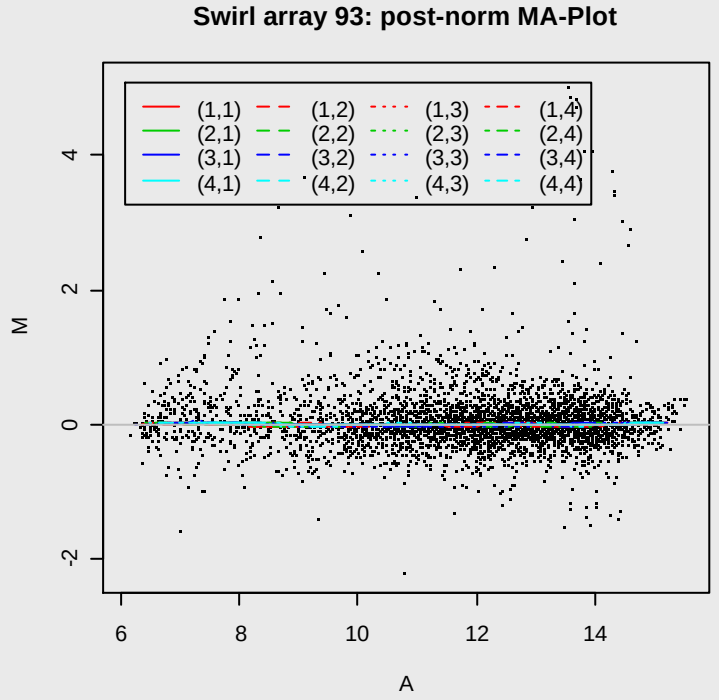
evaluation: effects of different data transformations

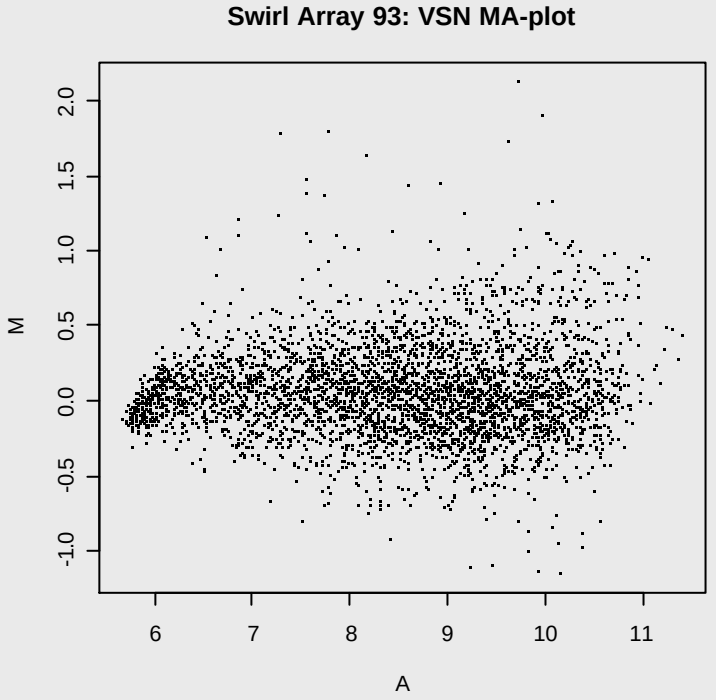


υ Normality: QQ-plot

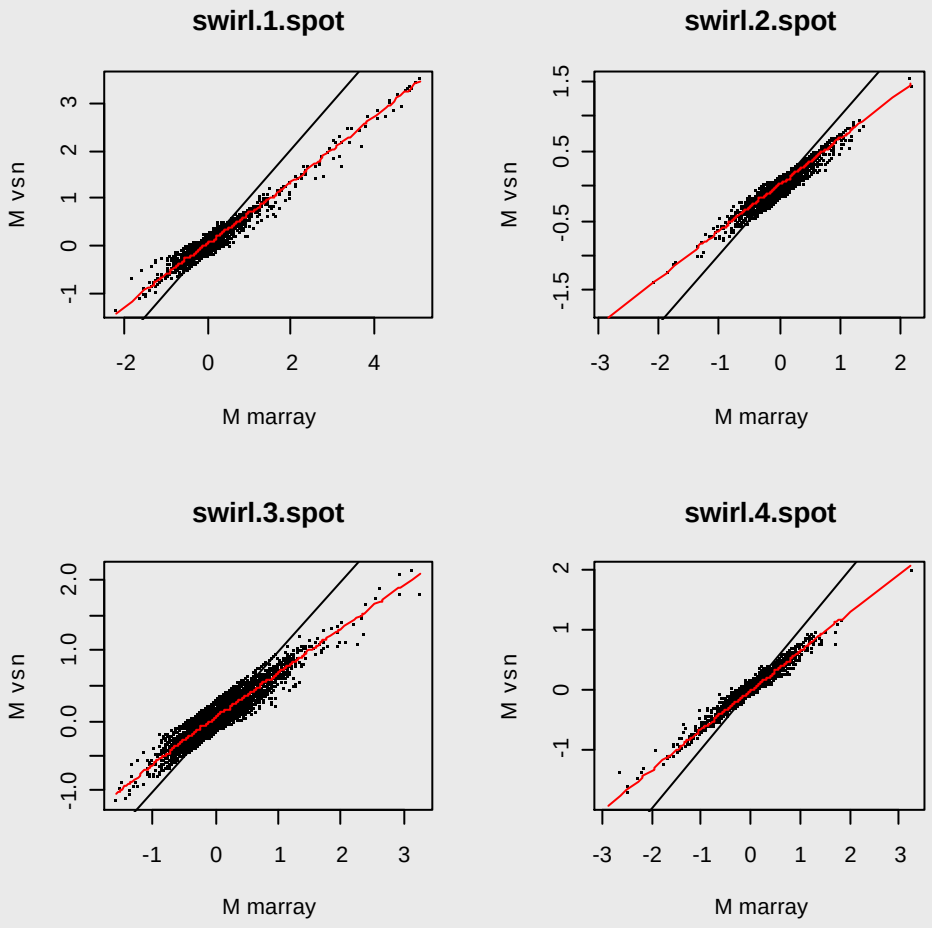


Swirl - VSN


Swirl array 93: post-norm MA-Plot
(1,1) (1,2) (1,3) (1,4)
(2,1) (2,2) (2,3) (2,4)
(3,1) (3,2) (3,3) (3,4)
(4,1) (4,2) (4,3) (4,4)
M
A
-2
0
2
4
6
8
10
12
14


Swirl Array 93: VSN MA-plot
M
A
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
6
7
8
9
10
11

Swirl: marray versus VSN



Acknowledgments

Holger Sültmann
Wolfgang Huber

Simon RM et al. (2003) *Design and Analysis of DANN Microarray Investigations*, Springer New York

Yang YH, Dudoit S. (2004) *Normalization: Bioconductor's marray package*