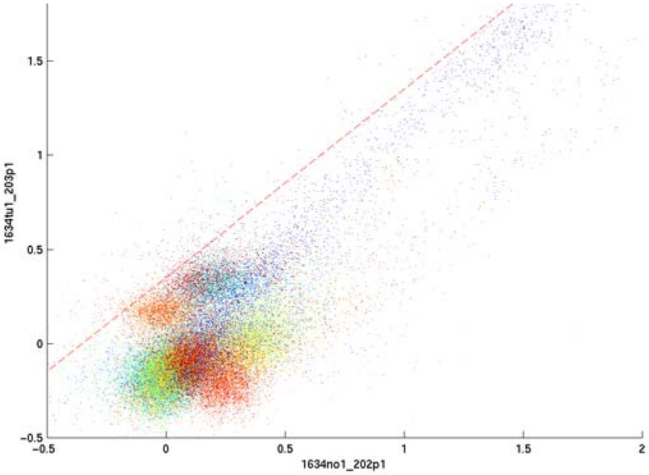
Quality control: artifacts, visualization, QC as residual analysis

Further topics on preprocessing: probe set summaries, physics

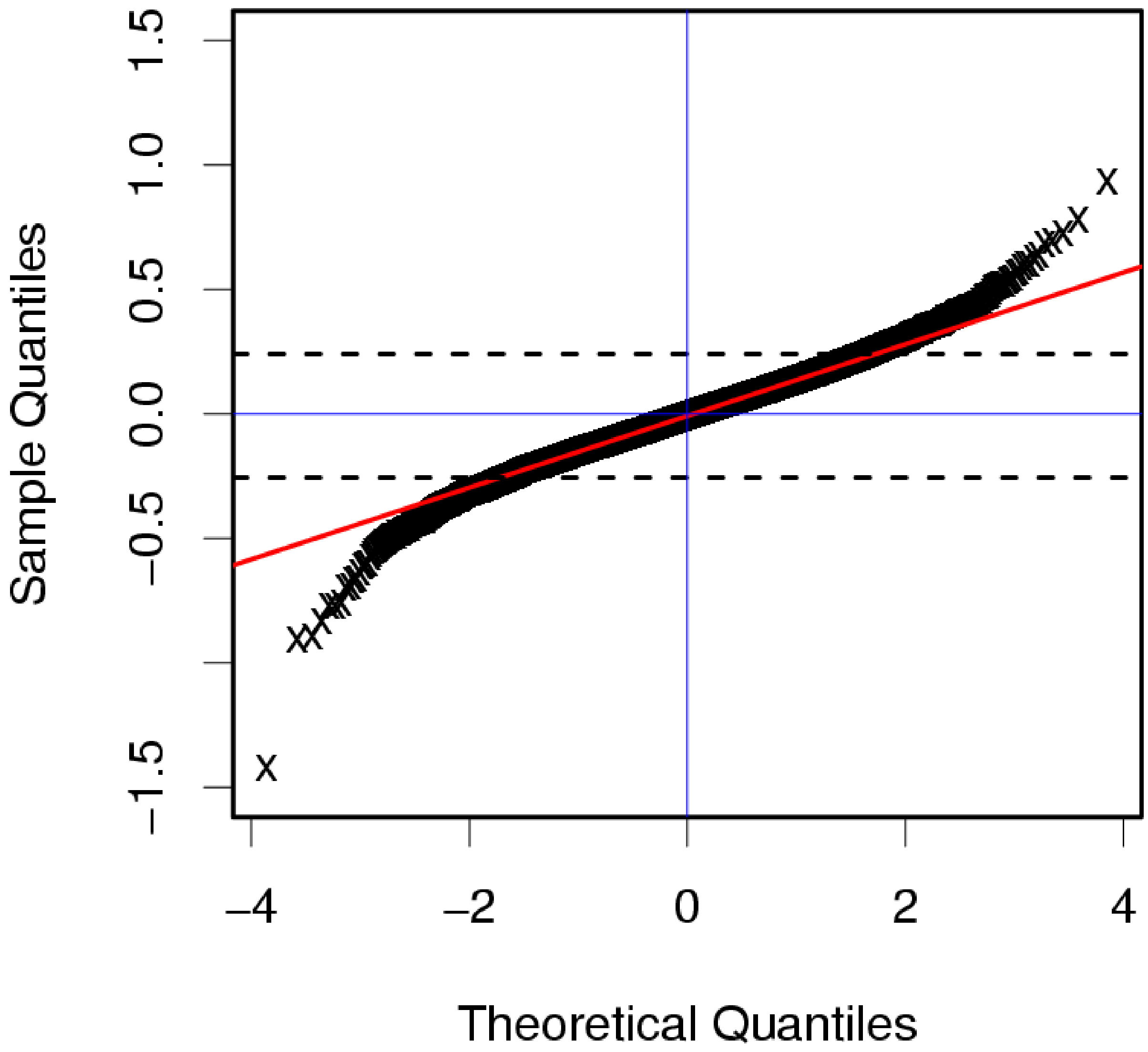
Wolfgang Huber

DKFZ Heidelberg



▶ Normal QQ-plot

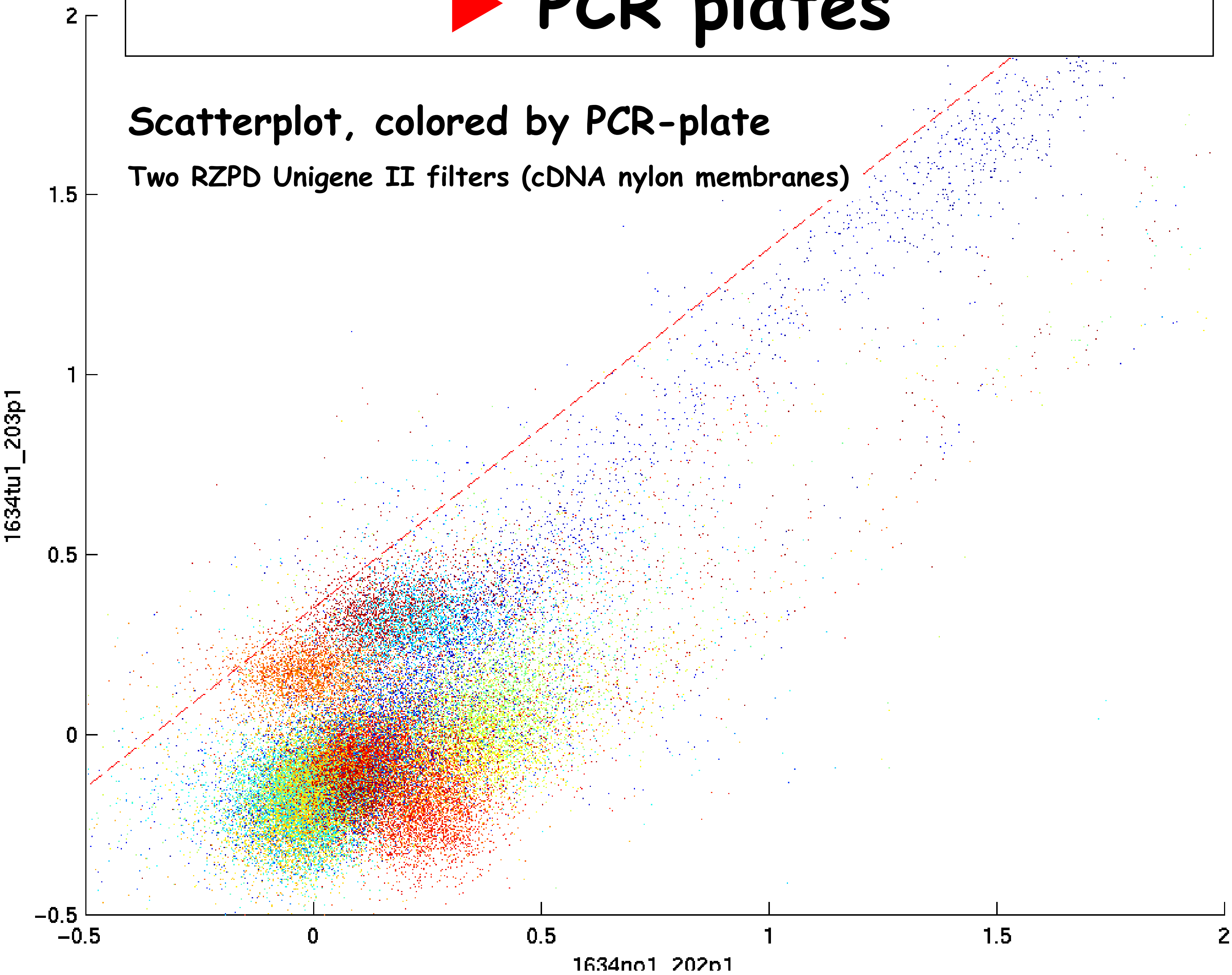
vsn-transformed data



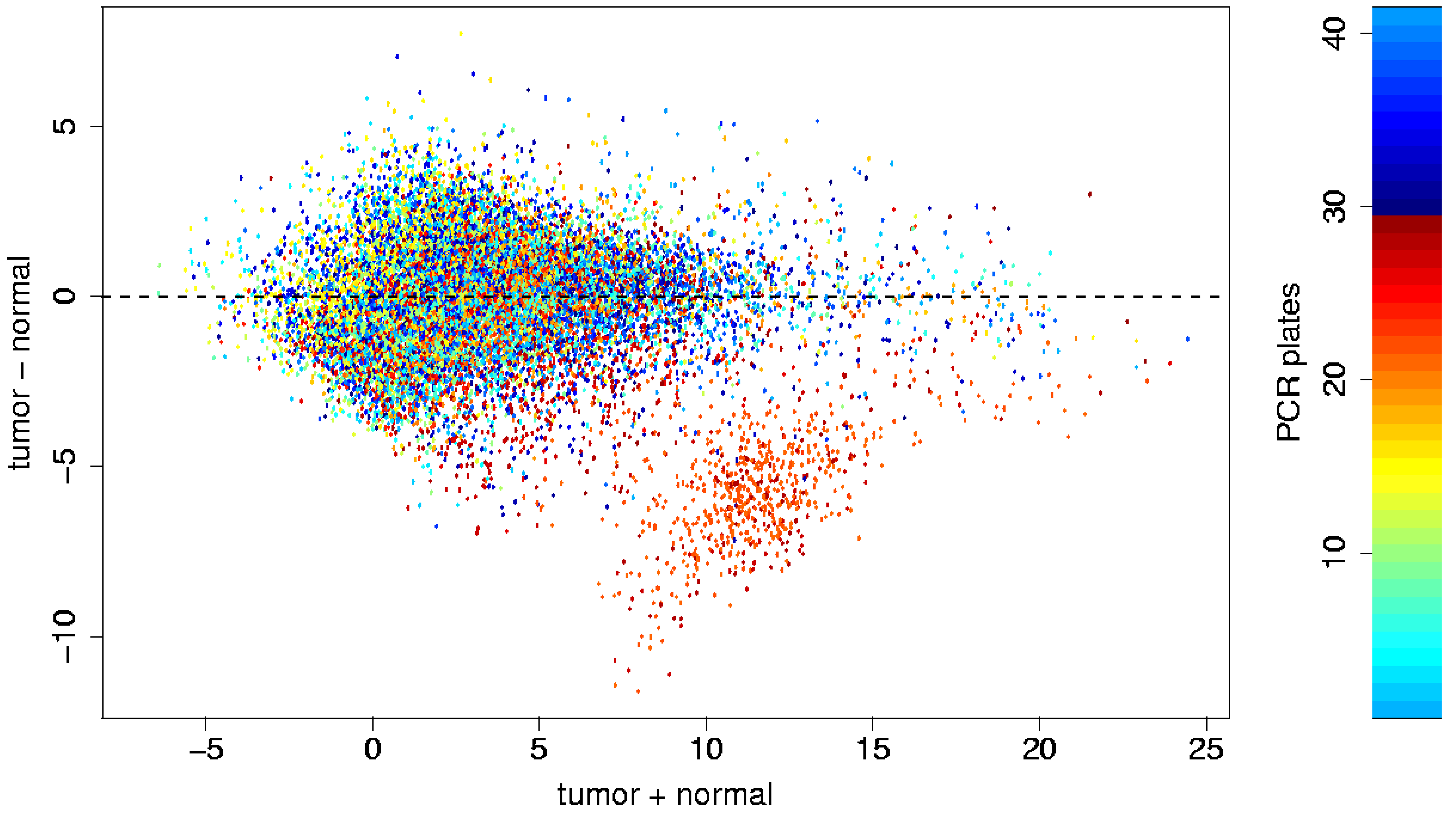
PCR plates

Scatterplot, colored by PCR-plate

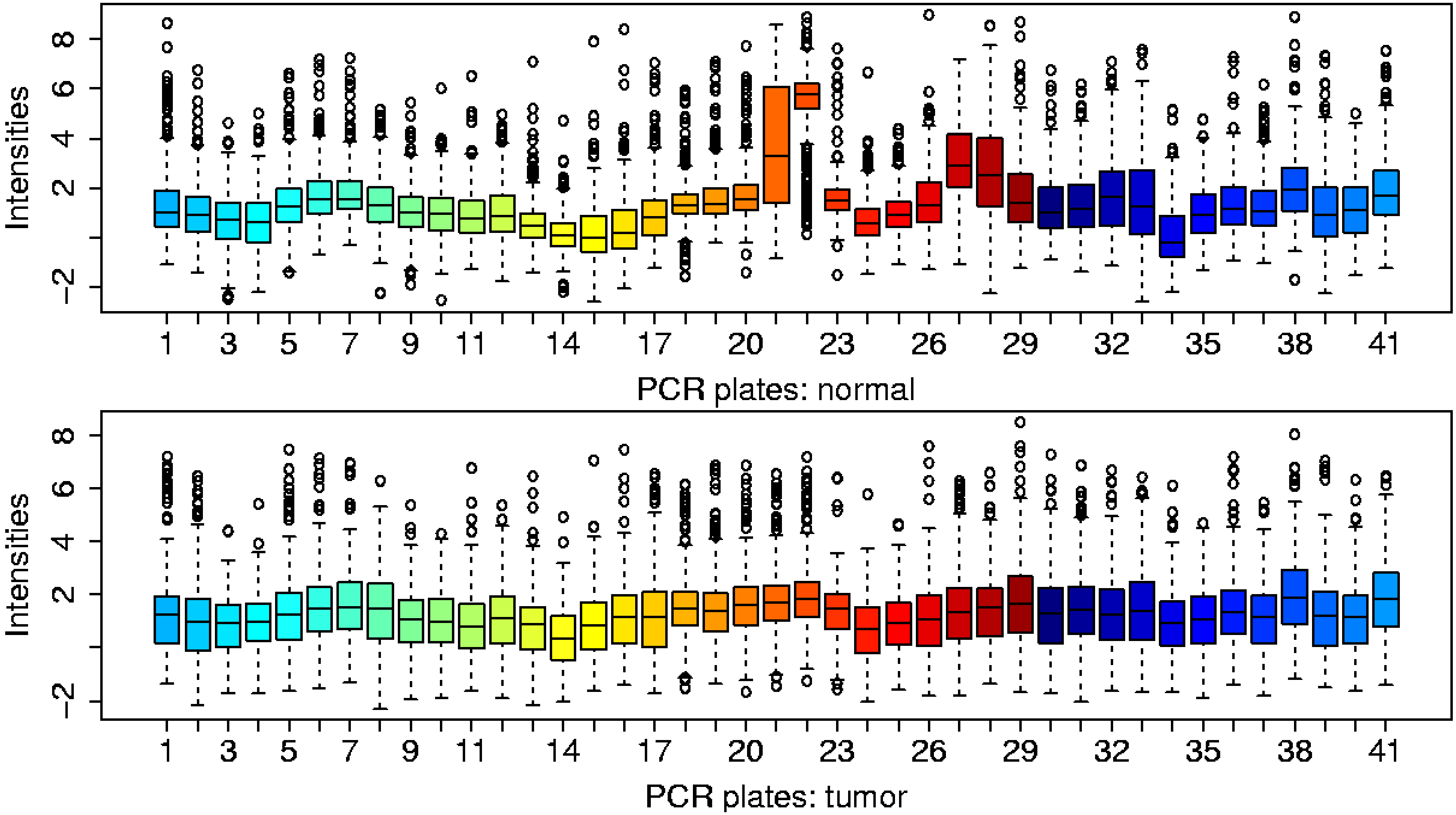
Two RZPD Unigene II filters (cDNA nylon membranes)



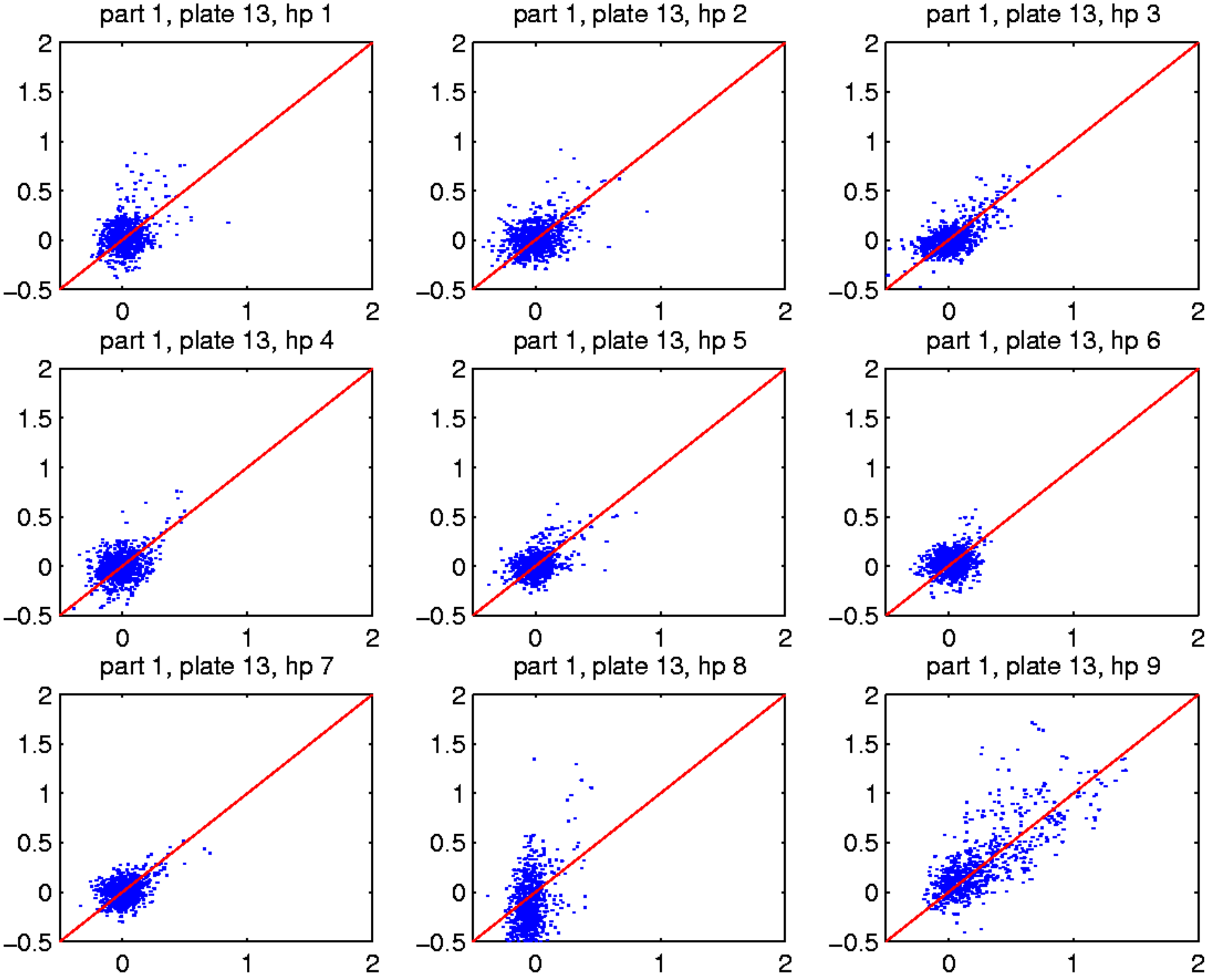
▶ PCR plates


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

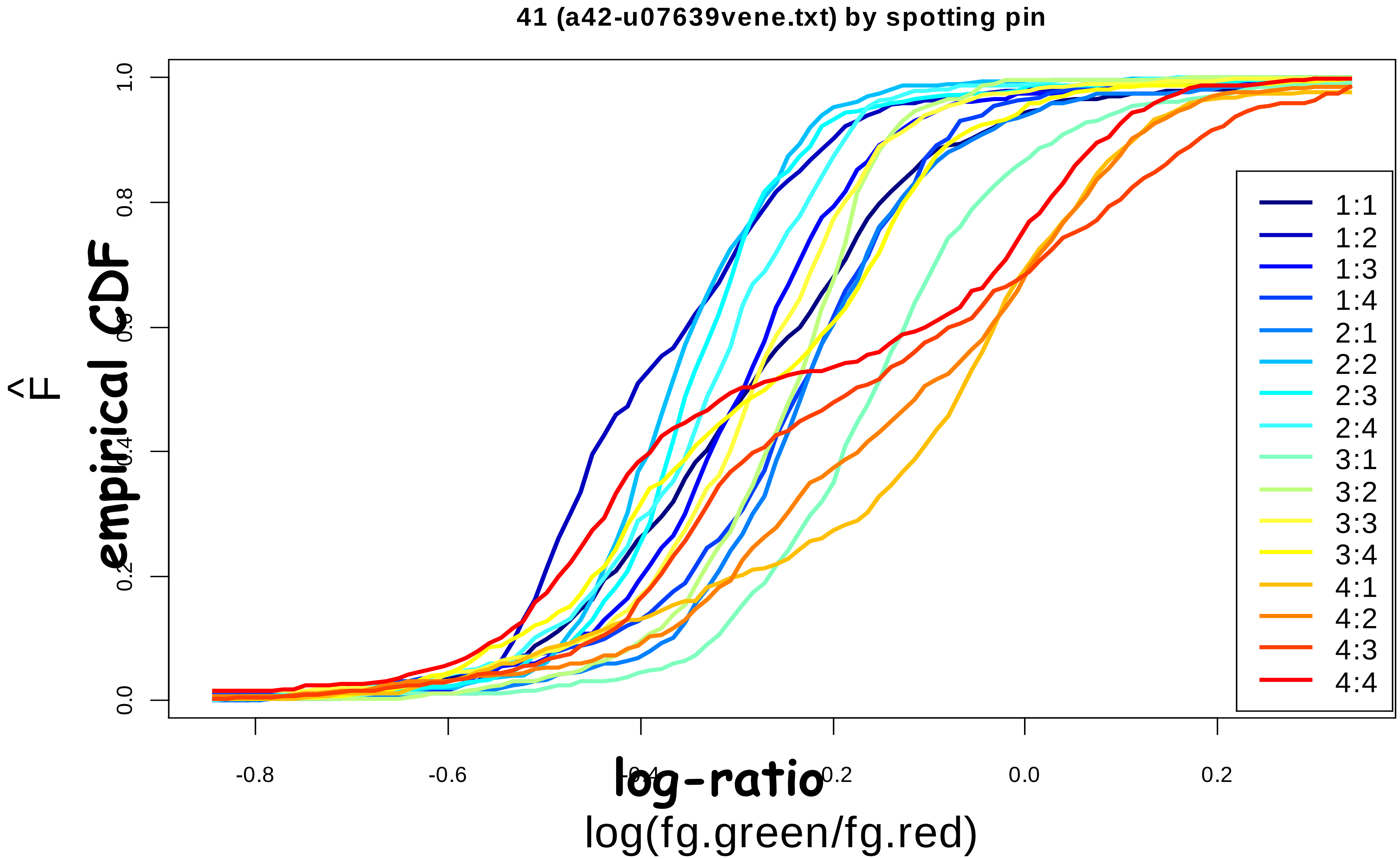
▶PCR plates: boxplots



▶ array batches



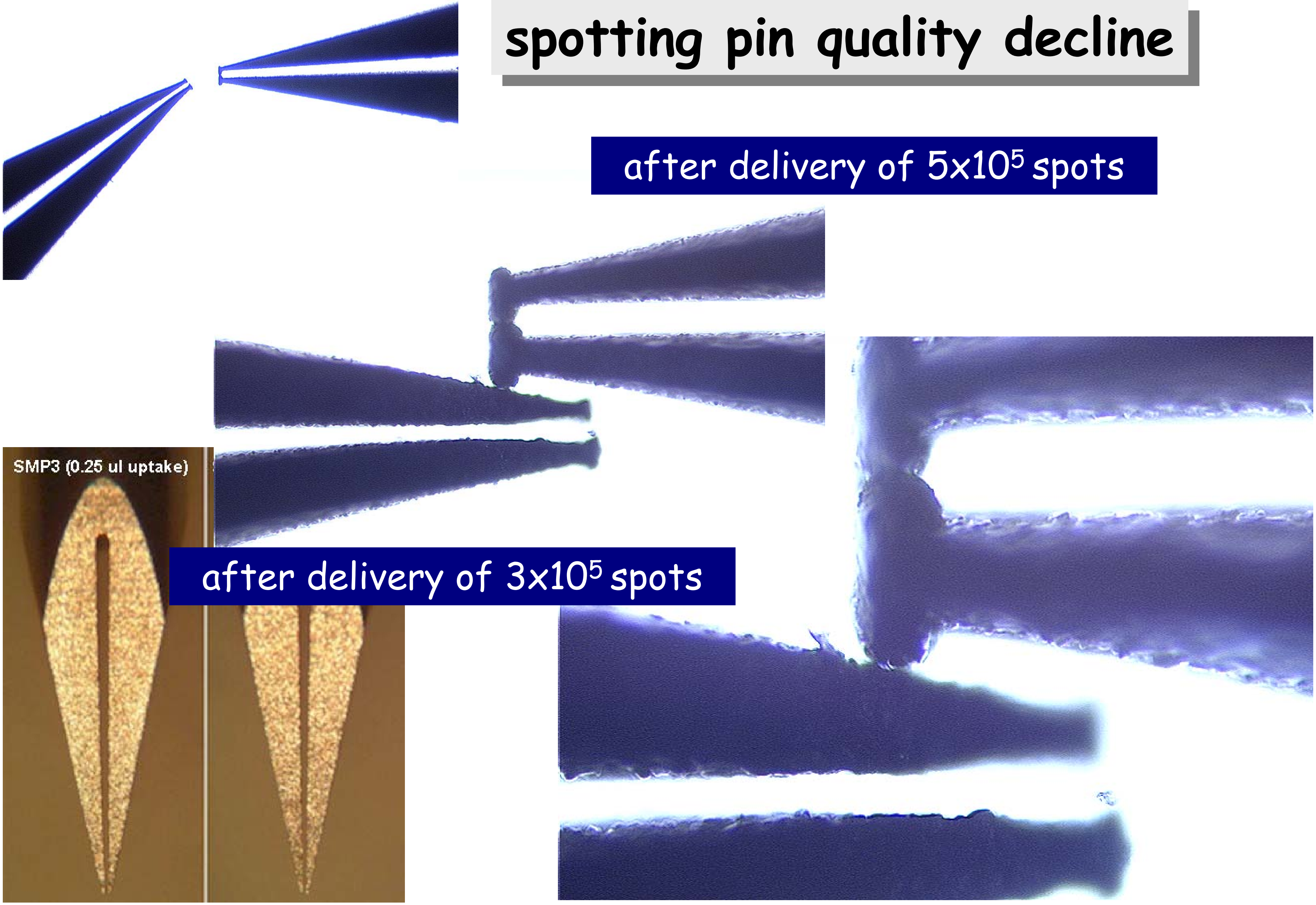
print-tip effects



empirical CDF

log-ratio

spotting pin quality decline

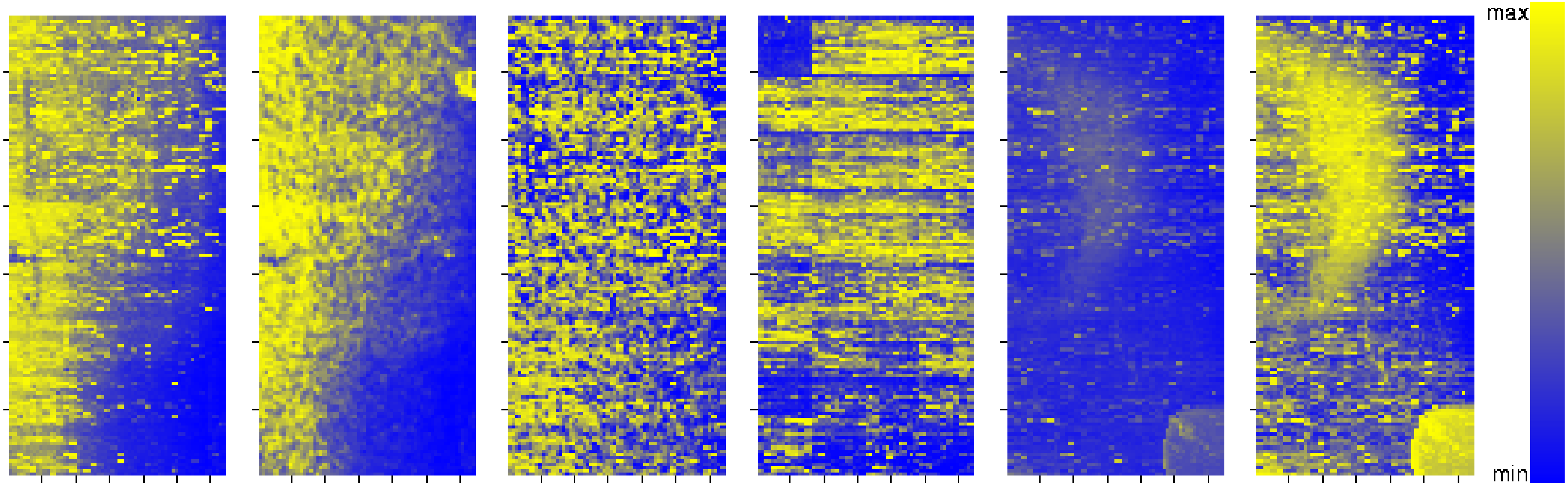


after delivery of 5x10^5 spots

after delivery of 3x10^5 spots

H. Sueltmann DKFZ/MGA

spatial effects



R Rb R-Rb
color scale by rank

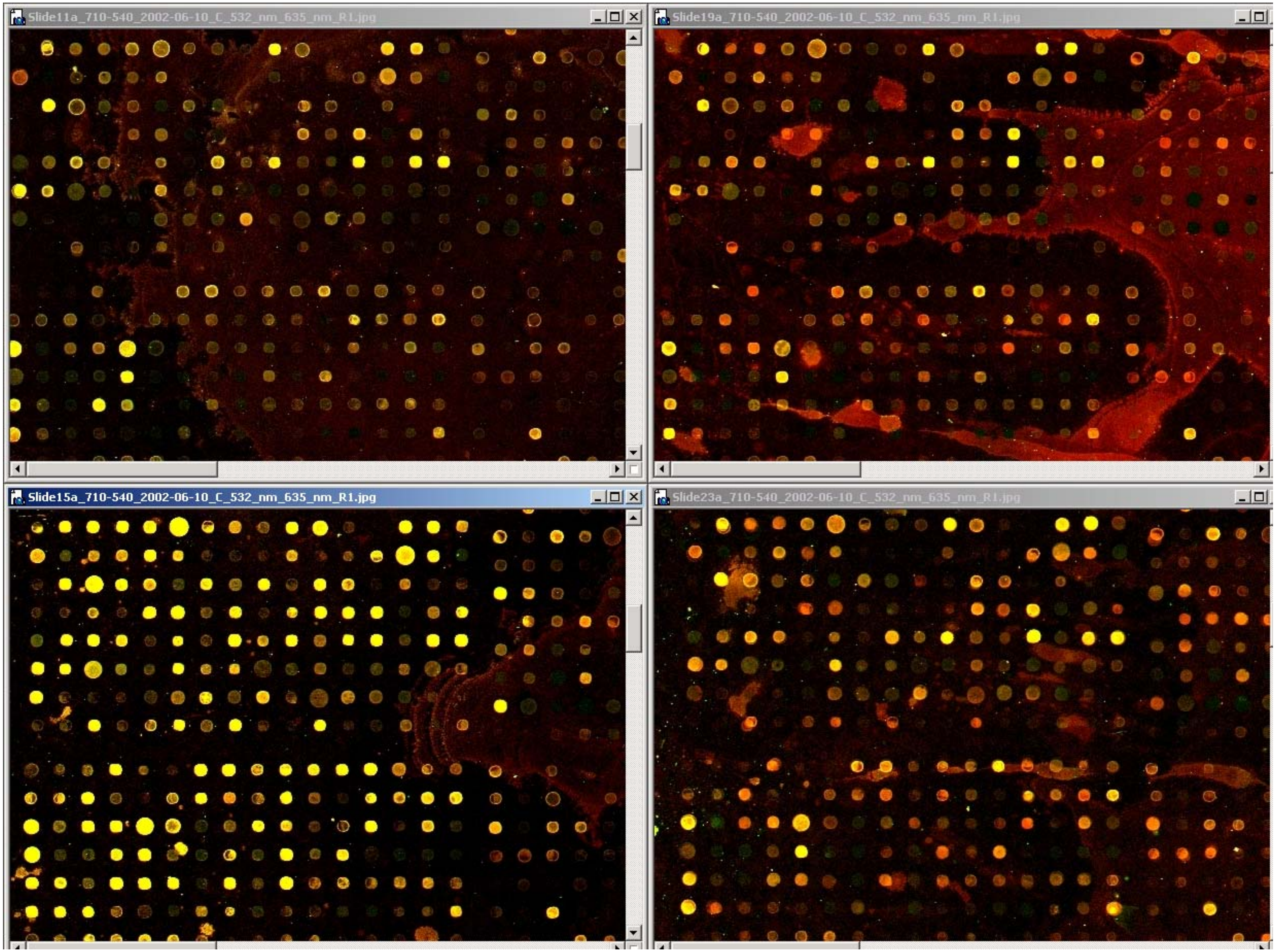
another array:
print-tip

color scale ~ log(G)

color scale ~ rank(G)

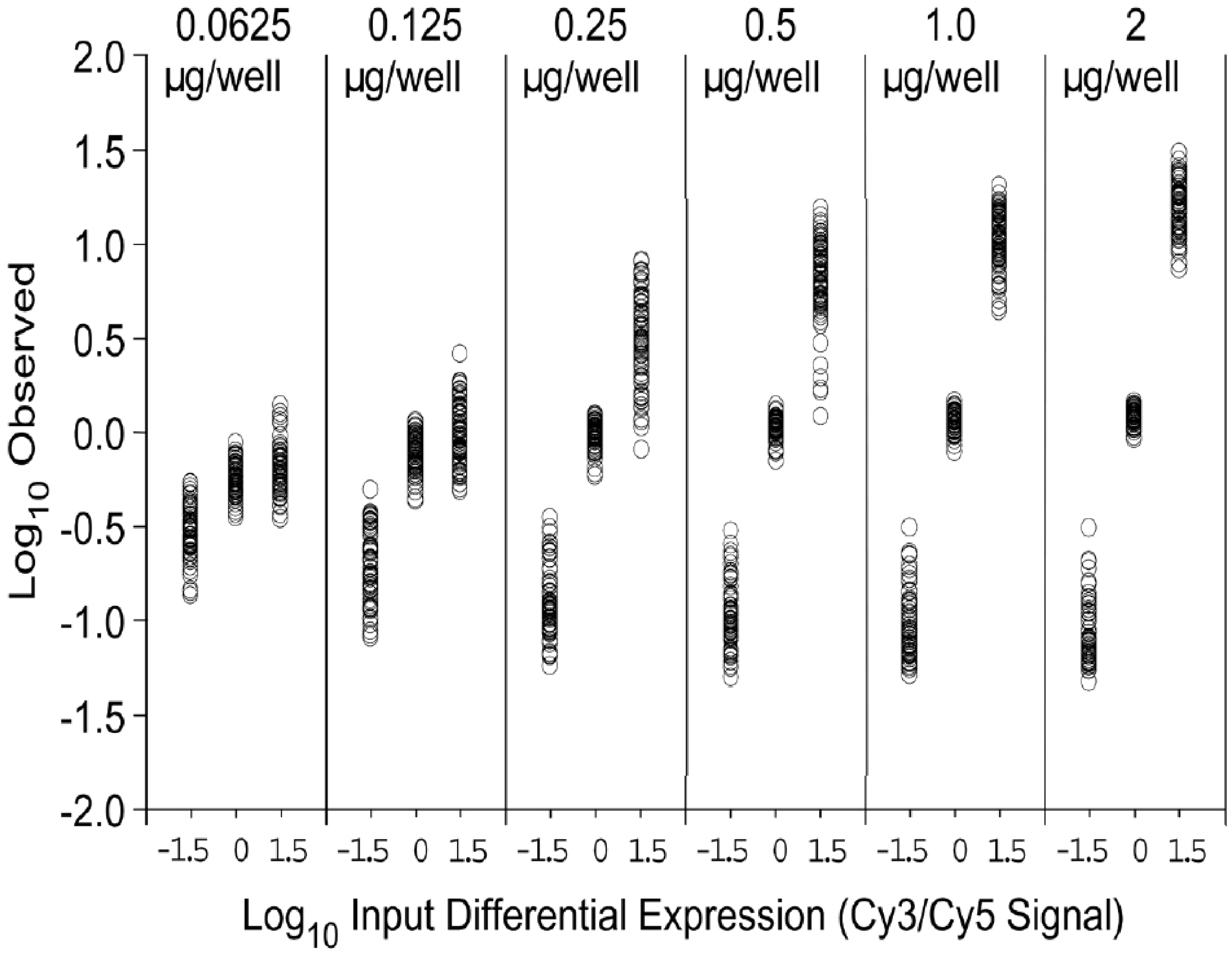
spotted cDNA arrays, Stanford-type

One RNA, four slides


Slide11a_710-540_2002-06-10_C_532_nm_635_nm_R1.jpg
Slide19a_710-540_2002-06-10_C_532_nm_635_nm_R1.jpg
Slide15a_710-540_2002-06-10_C_532_nm_635_nm_R1.jpg
Slide23a_710-540_2002-06-10_C_532_nm_635_nm_R1.jpg

Jörg
Schneider,
DKFZ

▶ Spot DNA concentration: ratio compression



Yue et al., (Incyte Genomics) NAR (2001) 29 e41

Amount of sample mRNA

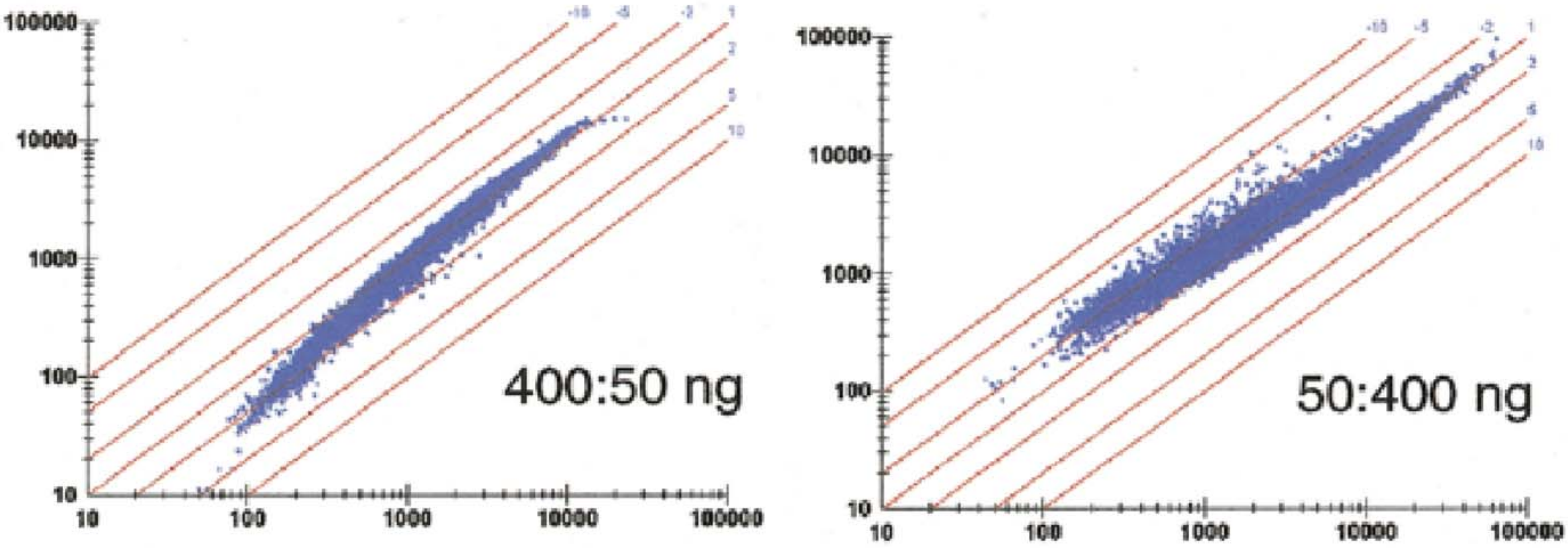


Figure from: H Yue et al. (Incyte), NAR 29: e41 (2001)

▶ Factors that affect measurements

Arrays

PCR yield: plate bias
ratio compression

Spotting / wear of pins: pin bias

Batch effects: density and steric accessibility of probes

Hybridization chamber asymmetries: spatial gradients

Samples

Ascertainment: RNA degradation
contamination

Amplification

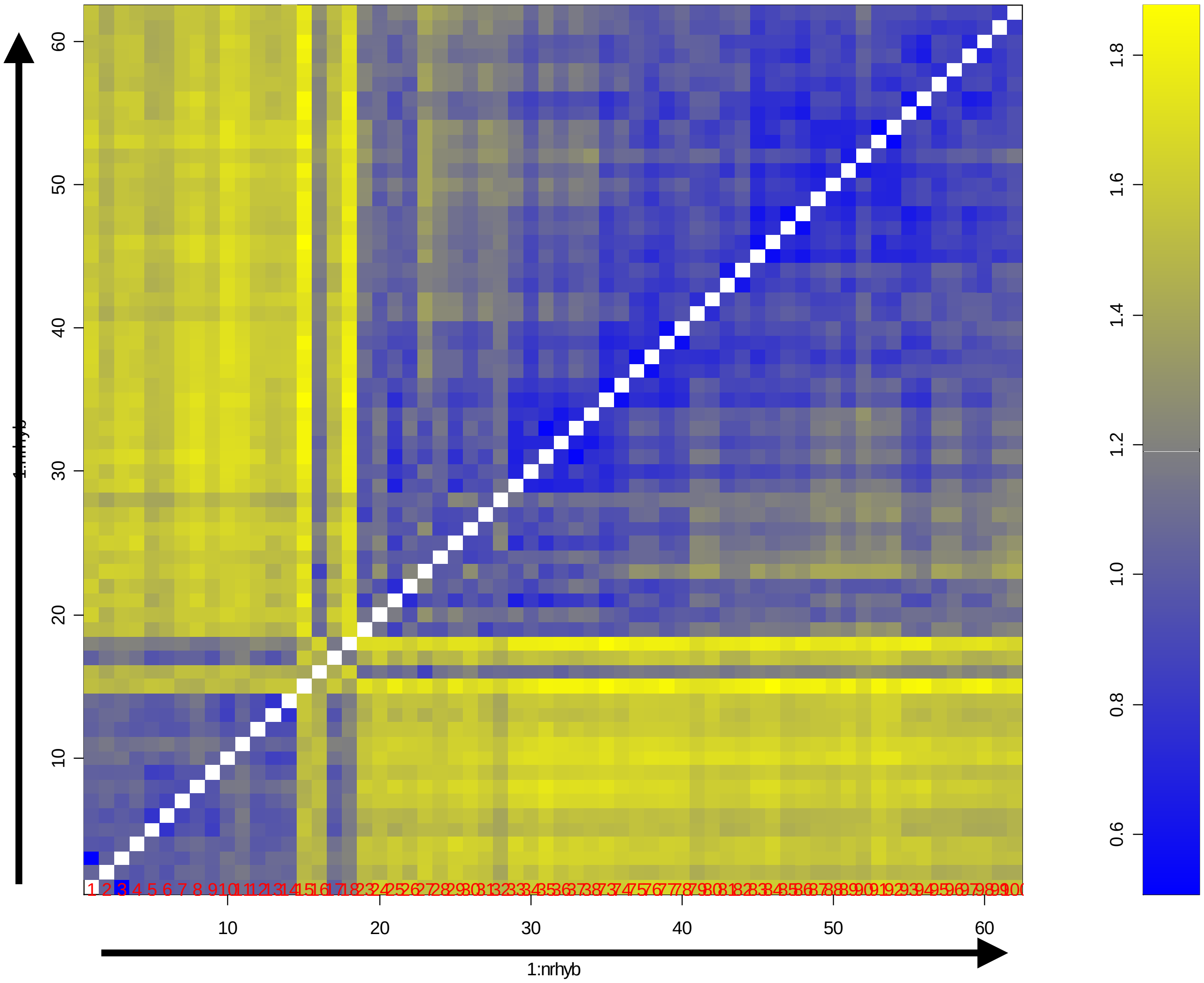
RNA purification

Labeling

Washing

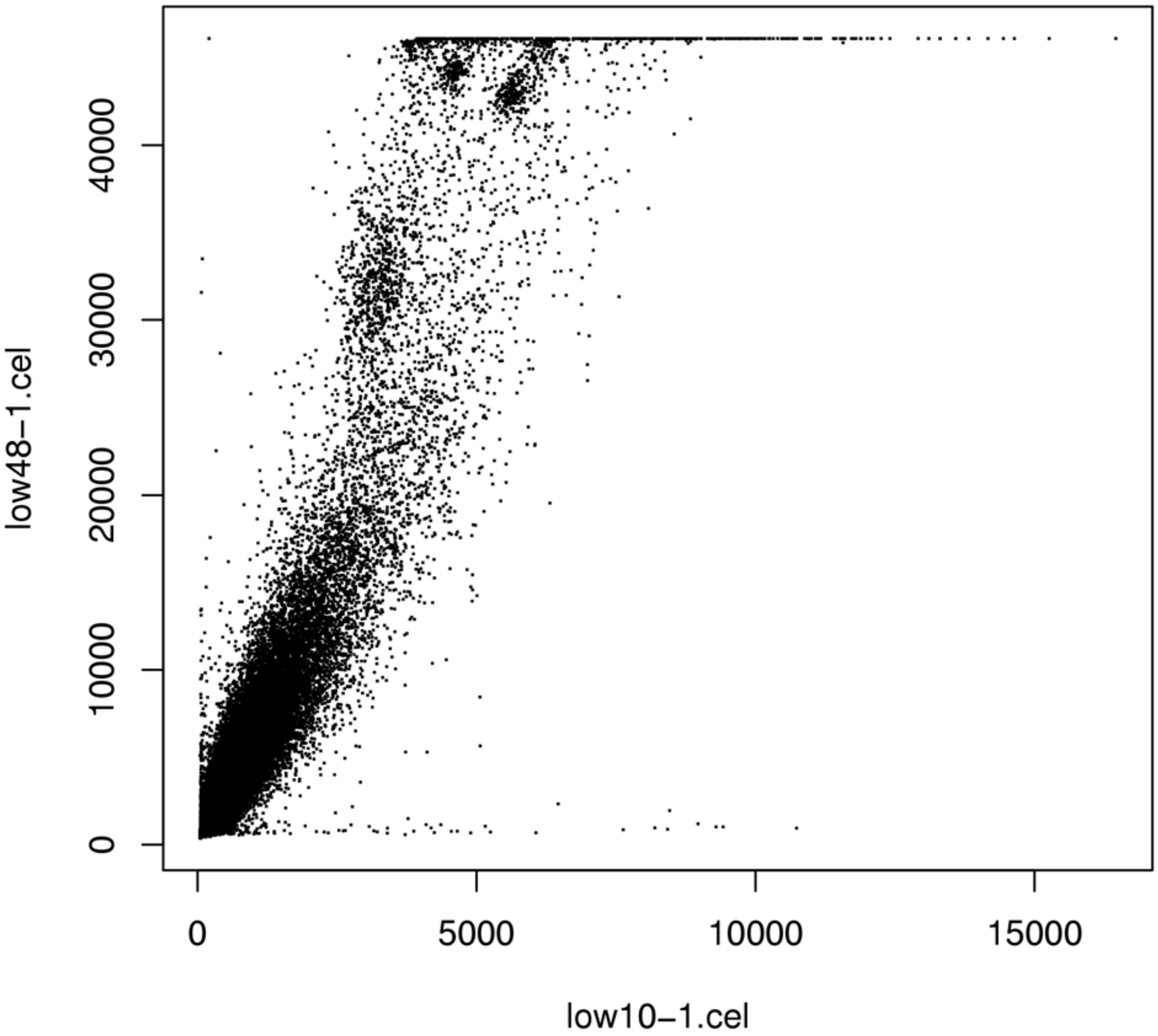
Scanner

Batches: array to array differences $d_{ij} = \text{mad}_k(h_{ik} - h_{jk})$

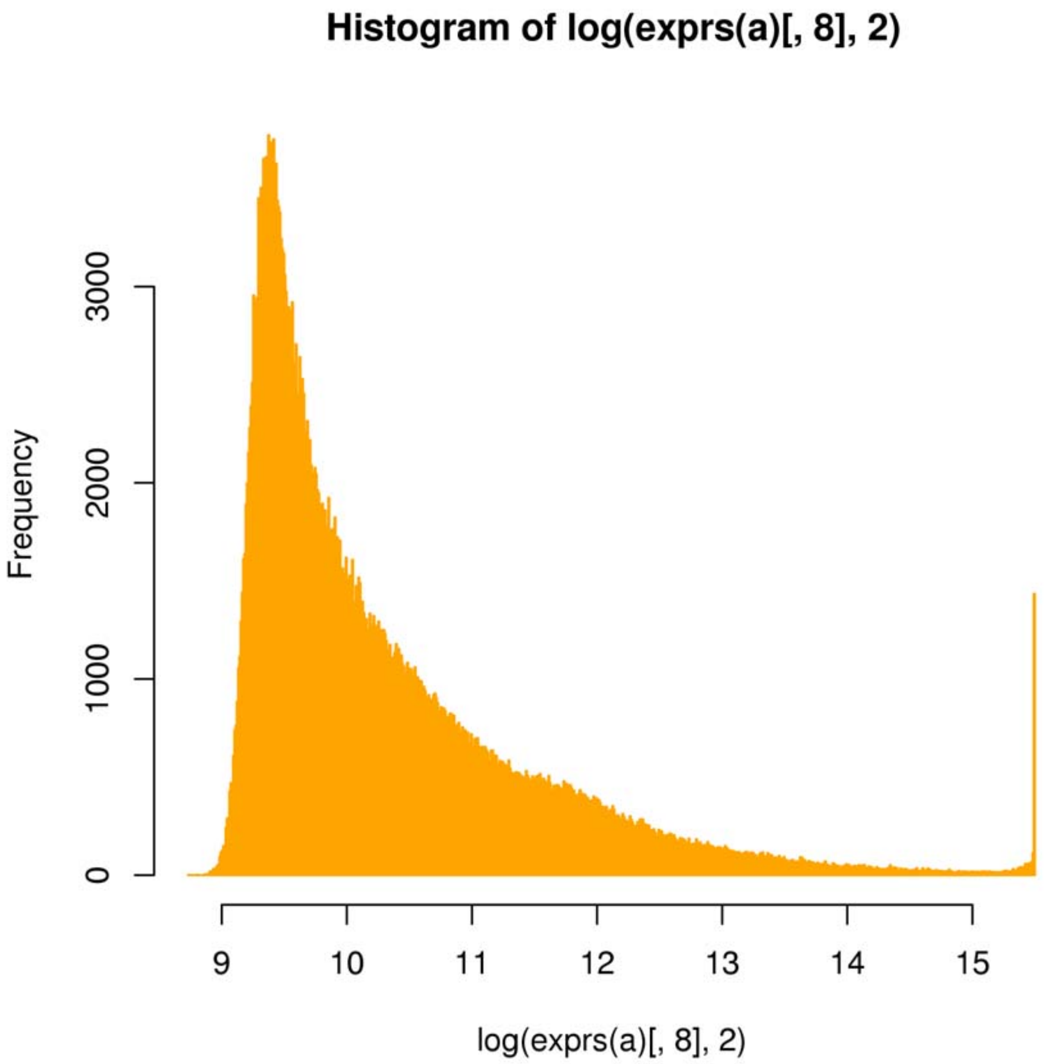


arrays i=1…63; roughly sorted by time

▶ Scatterplots

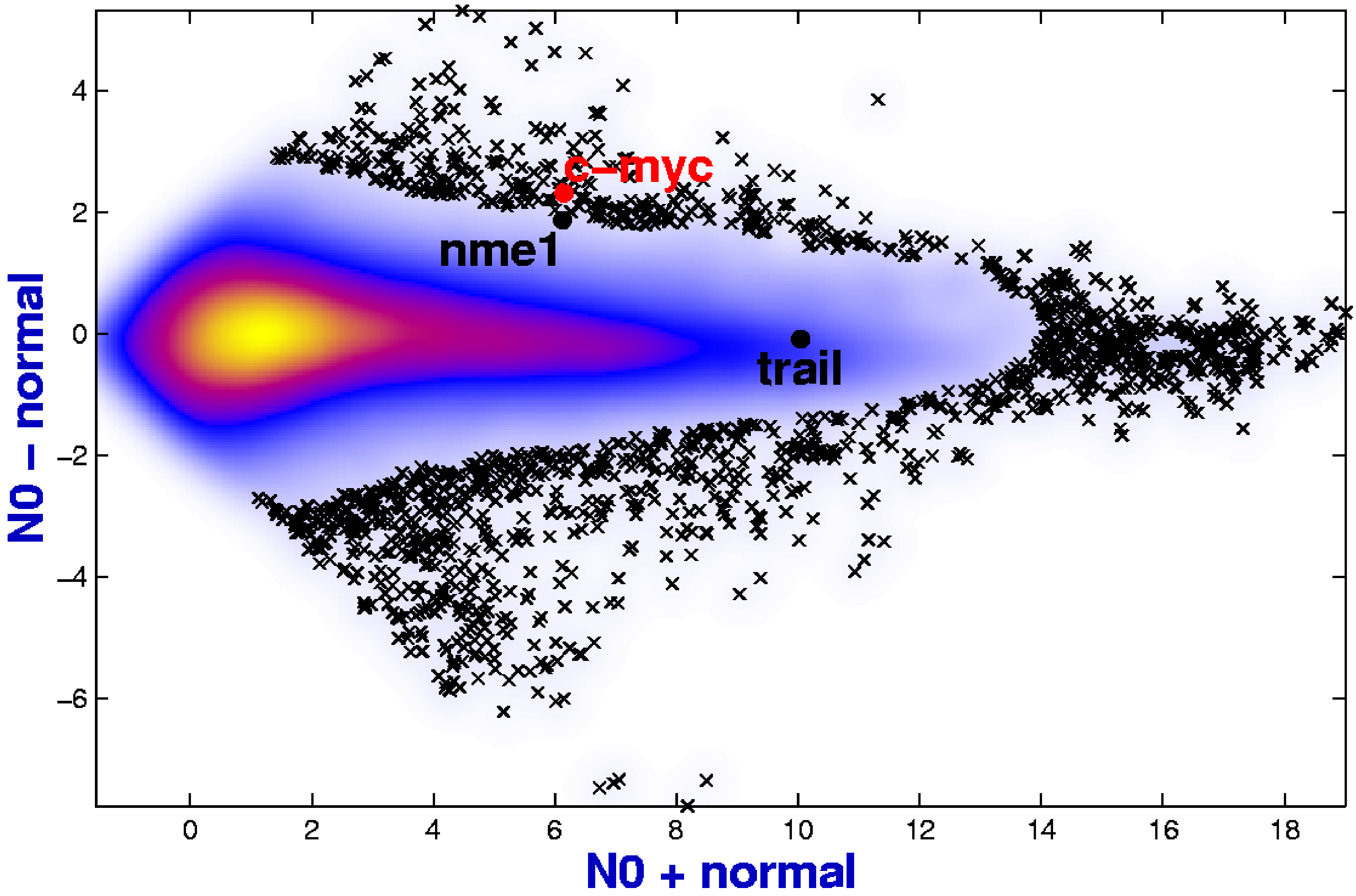


▶ Histogram


Histogram of log(exprs(a)[, 8], 2)
Frequency
0
1000
2000
3000
9
10
11
12
13
14
15
log(exprs(a)[, 8], 2)

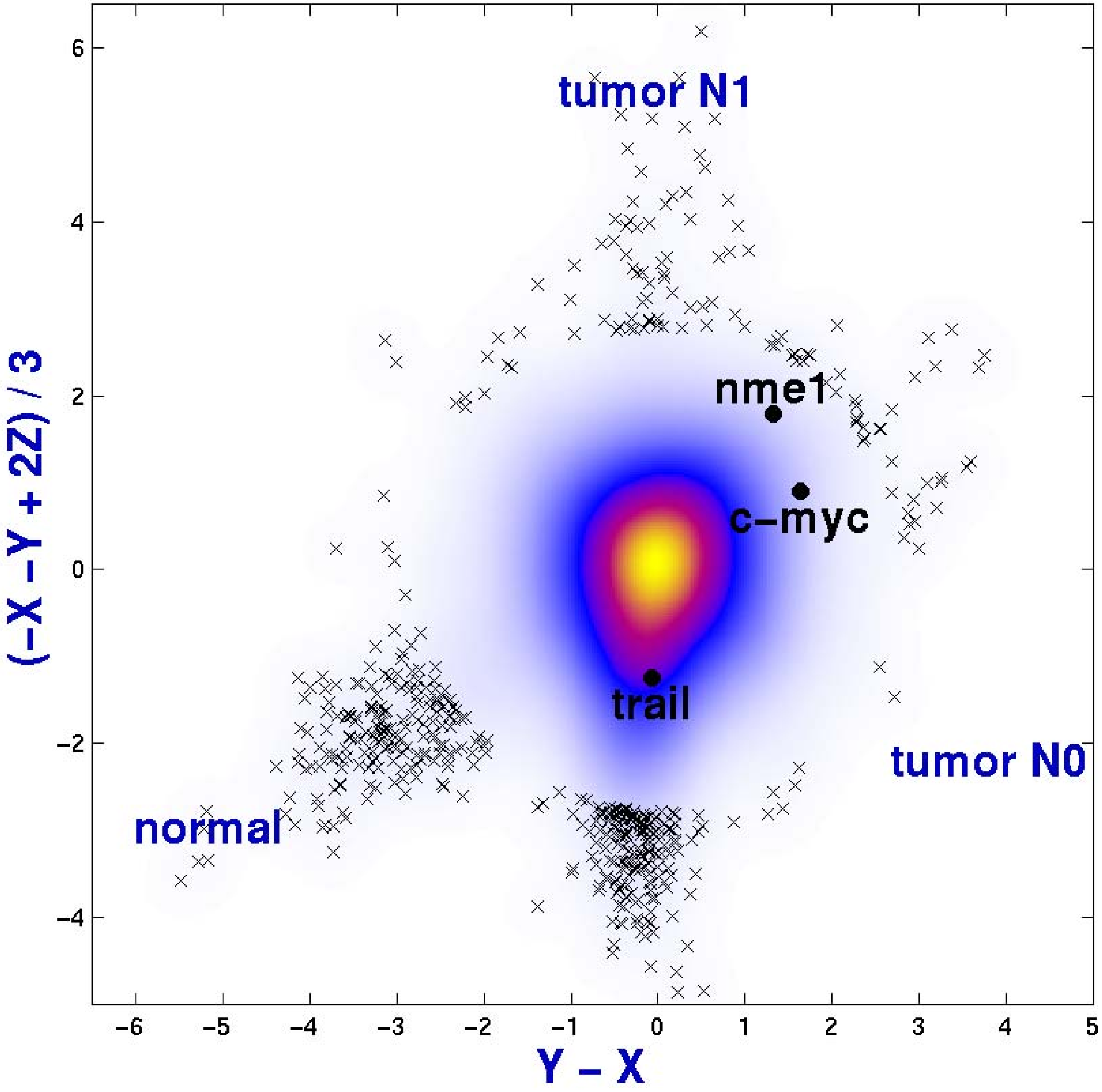
▶ Density representation of the scatterplot

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



▶ Density representation of the scatterplot

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



▶ Quantities that can be used for QC

Control data:

Positive controls (e.g. metallothioneins in kidney)
Negative controls (e.g. nonhomologous probes)
(Spikein cDNA)

Hot data:

reproducibility / similarity:

replicate probes per array
replicate arrays per sample
multiple probes per transcript
multiple samples per biological condition
Absence of correlation with technical factors (enzyme-bacth, spatial location on array, …)

signal:

amplitude / quantity of differences between samples known to be biological different

▶ Quantities that can be used for QC

Essential:

Experimental design that
minimizes role of technical effects
biological groups are balanced/randomized

▶ A model-based approach to QC

Make theoretically and/or empirically founded modelling assumptions on the data, then see if a given set of data fits. If no, the data is bad.

Examples:

- additive-multiplicative error model with affine chip effects
- additive-multiplicative error model with affine chip- und pin-effects
- Li-Wing model with probe- and sample effects
- affyPLM (later … first we need some background on Affymetrix)

▶ Affymetrix expression measures

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.

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

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

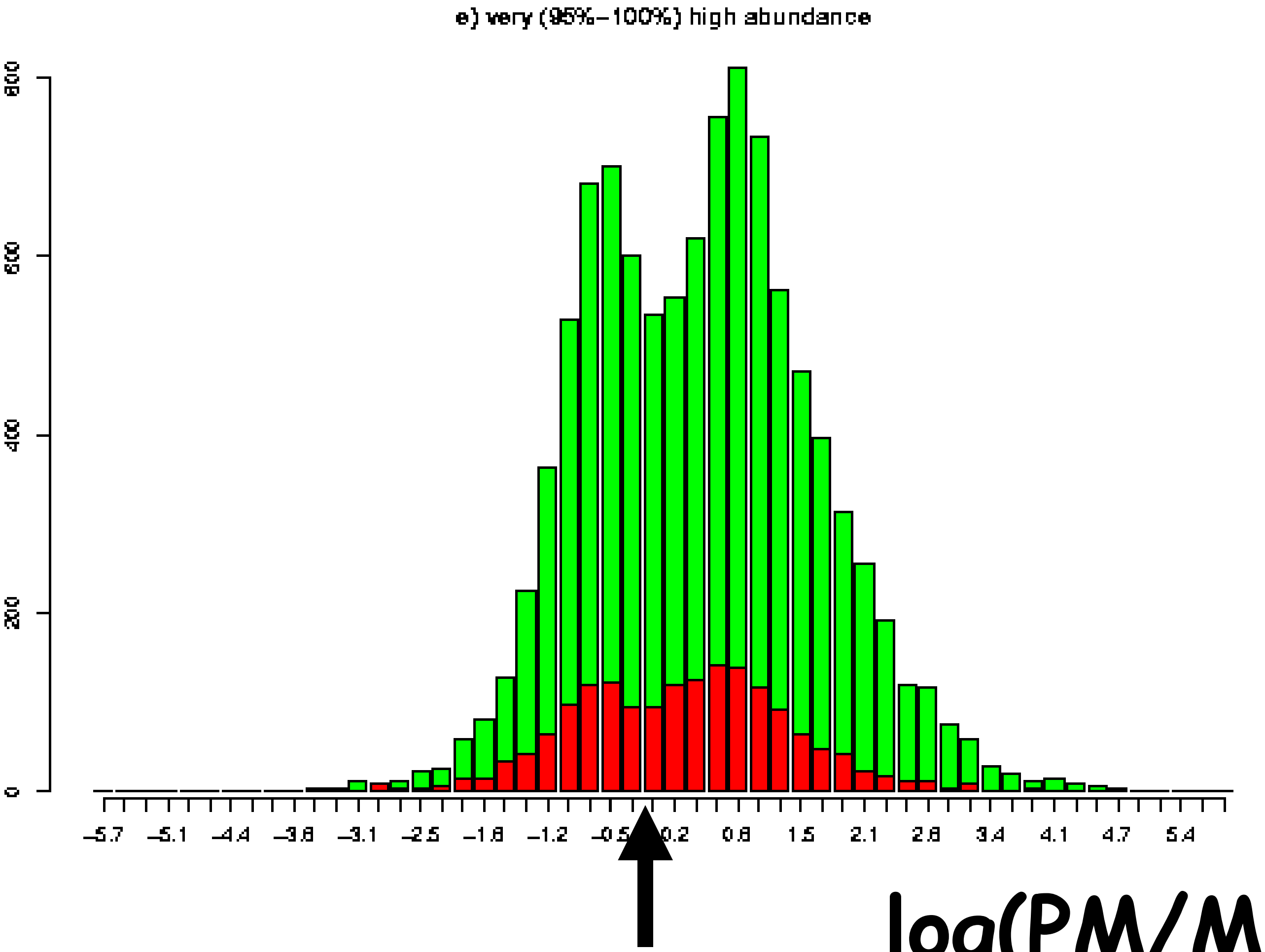
with A a set of "suitable" pairs.

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

▶ $I_{PM} = I_{MM} + I_{specific}$?



0

log(PM/MM)

From: R. Irizarry et al.,
Biostatistics 2002

▶ Physico-chemical modeling of the probe intensities: the riddle of the bright mismatches

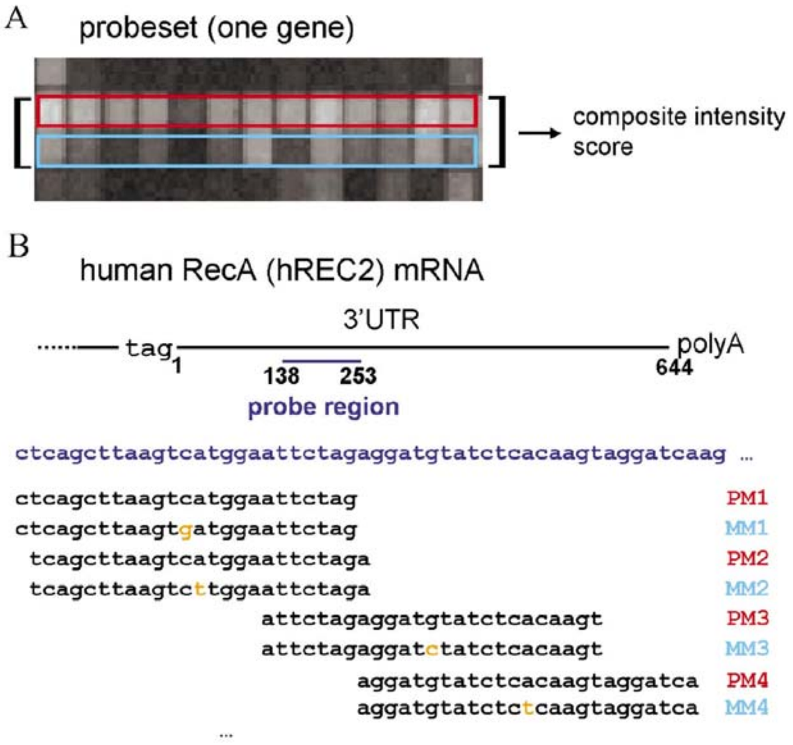

A
probeset (one gene)
composite intensity score
B
human RecA (hREC2) mRNA
3'UTR
tag
1
polyA
644
138
253
probe region
ctcagcttaagtcatggaattctagaggatgtatctcacaagtaggatcaag …
ctcagcttaagtcatggaattctag PM1
ctcagcttaagtgatggaattctag MM1
tcagcttaagtcatggaattctaga PM2
tcagcttaagtcttggaattctaga MM2
attctagaggatgtatctcacaagt PM3
attctagaggatctatctcacaagt MM3
aggatgtatctcacaagtaggatca PM4
aggatgtatctctcaagtaggatca MM4
…

FIG. 1. (Color) Probeset design. (A) The raw scanned image of a typical probeset, with the PM (MM) on the top (bottom) row; higher brightness (white) corresponds to higher abundance of bound RNA molecules. The large variability in probe brightness is clearly visible. (B) Arrangement of probe sequences along the target transcript for the human recA gene in the HG-U95A array. Here the probe region (blue) is 116 bases long; it is typical that probes lie in the 3' UnTRanslated region, namely, between the stop triplet (codon) "tag" and the polyadenylation signal. The first four probes are shown explicitly; notice the overlap in their sequences.

Naef et al., Phys Rev E 68 (2003)

▶ Physico-chemical modeling of the probe intensities: the riddle of the bright mismatches

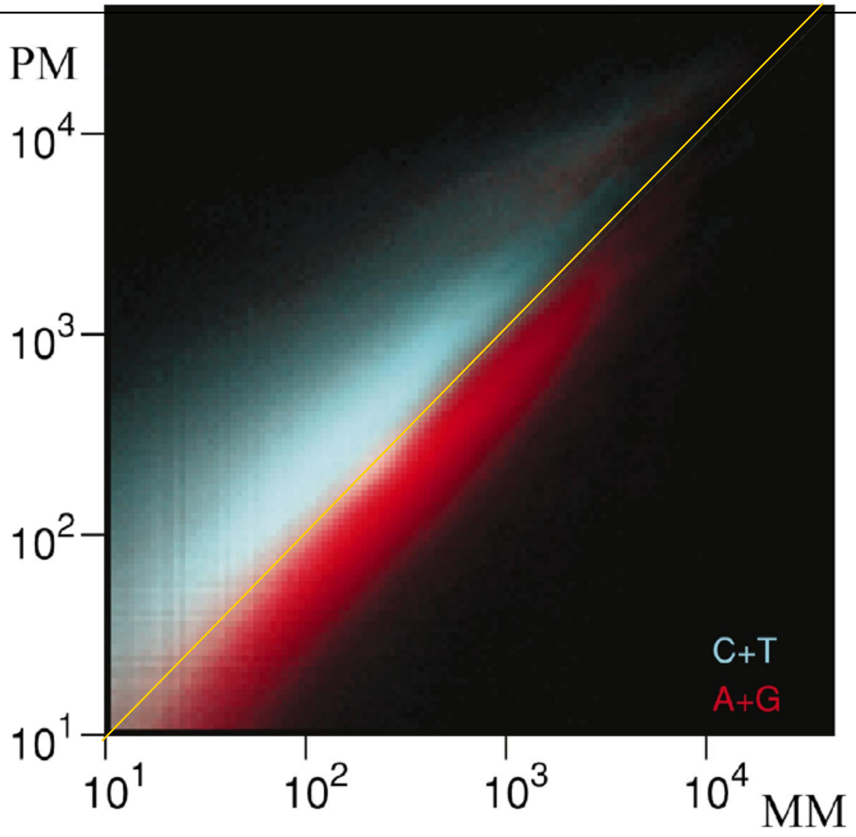


FIG. 2. (Color) PM vs MM histogram from 86 human HG-U95A arrays. The joint probability distribution for PM and MM shows strong sequence specificity. In this diagram, all 17×10^6 (PM,MM) pairs in a dataset were used to construct a two-dimensional histogram. Pairs whose PM middle letter is a pyrimidine (C or T) are shown in cyan, and purines (A or G) in red. 33% of all probe pairs are below the PM=MM diagonal; 95% of these have a purine as their middle letter.

purines

2 rings

MM: 2 large molecules -> steric hindrance

pyrimidines

1 ring

MM: 2 small molecules -> no problem

This explains the existence of two populations, but not their location

Felix Naef et al., Phys Rev E 68 (2003)

▶ Physico-chemical modeling of the probe intensities: the riddle of the bright mismatches

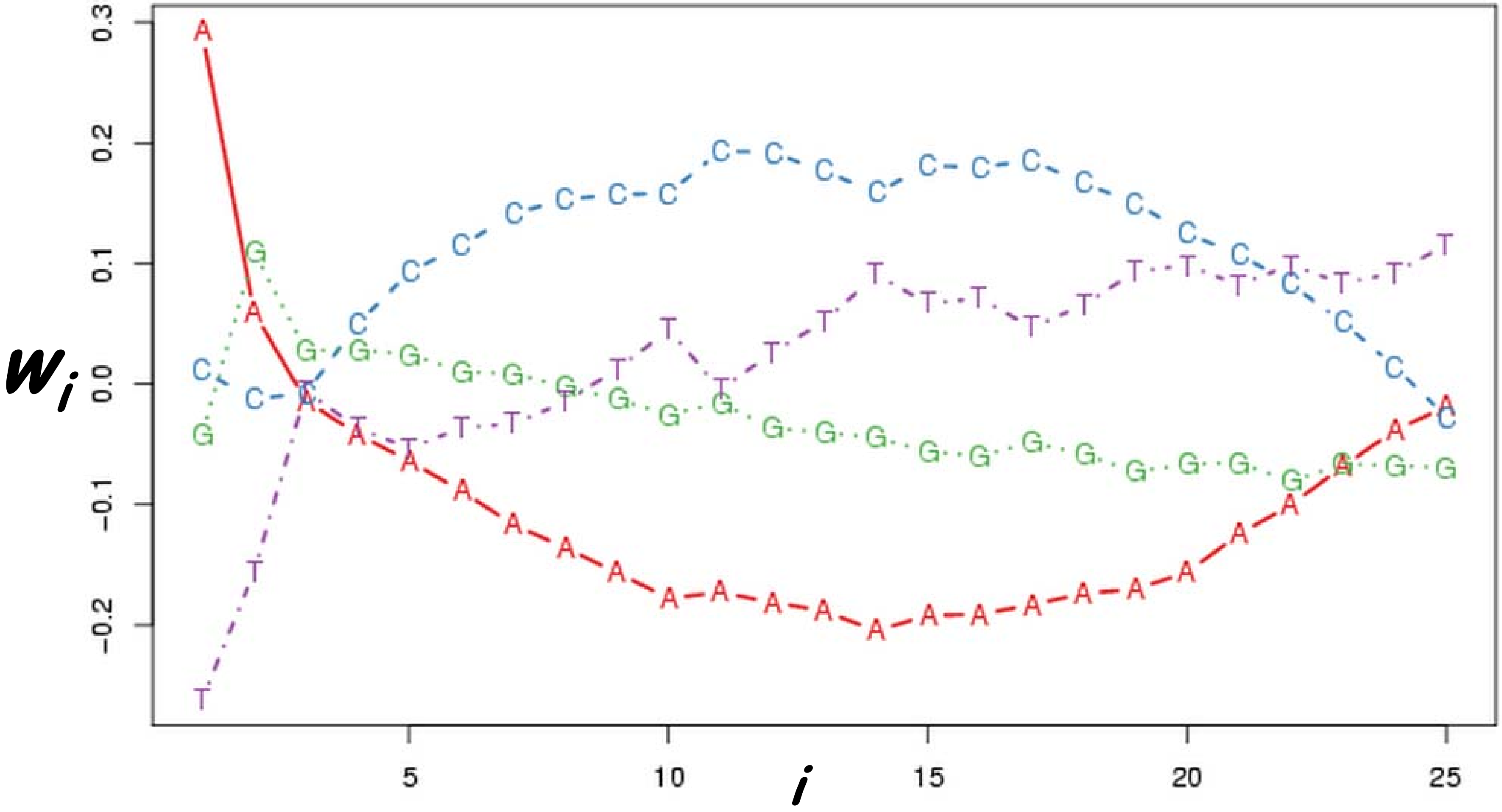
Fit a statistical model for the deviation of a probe's intensity from its probe set's median intensity

$$\log\left(\frac{PM}{\underset{i}{med}(PM_i)}\right) \sim s_1 + s_2 + \ldots + s_{25}$$

s_i: factor representing nucleotide (A, C, G, T) at i-th position

Naef et al., Phys
Rev E 68 (2003)

▶ Physico-chemical modeling of the probe intensities: the riddle of the bright mismatches



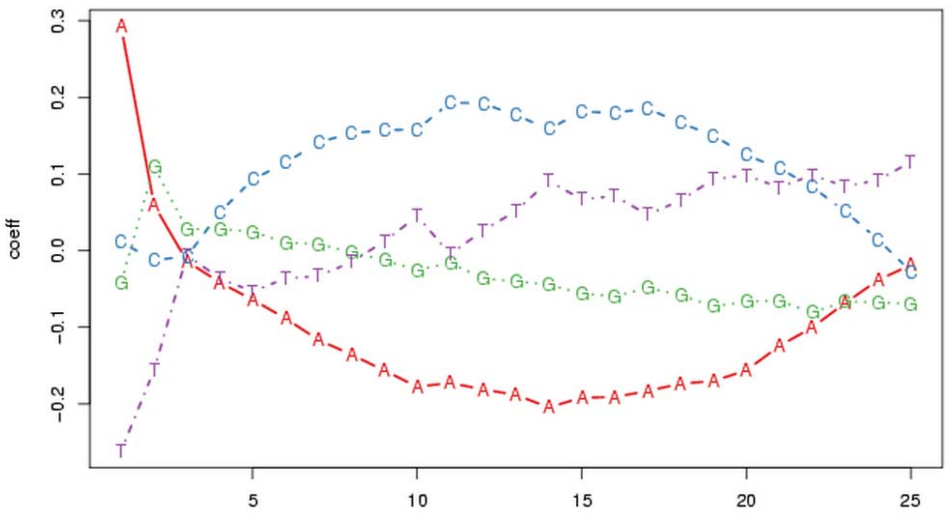
▶ Physico-chemical modeling of the probe intensities: the riddle of the bright mismatches

- *Changing one A into C in the middle of the probe: $e^{0.4}$~1.5*

- *Left/right asymmetry*

- *Asymmetry A vs T: A-T bonds are not equivalent to T-A bonds! (similar for G vs C).*

- *Labels are at U and C*



G-C (PM) dimmer than C-C* (MM)*

▶ affyPLM package

Fitting linear models to probe set intensities across mutliple arrays

$$y_{pi} \sim p + a_i + \ldots$$

Y_{pi}	intensity of probe p (e.g. 1…11) on array i
p	probe ID (factor)
a_i	array effect
…	further biological factors!

▶ affyPLM package

affy::fitPLM

example: robust linear model for Dilution data with effect for liver dilution level and scanner

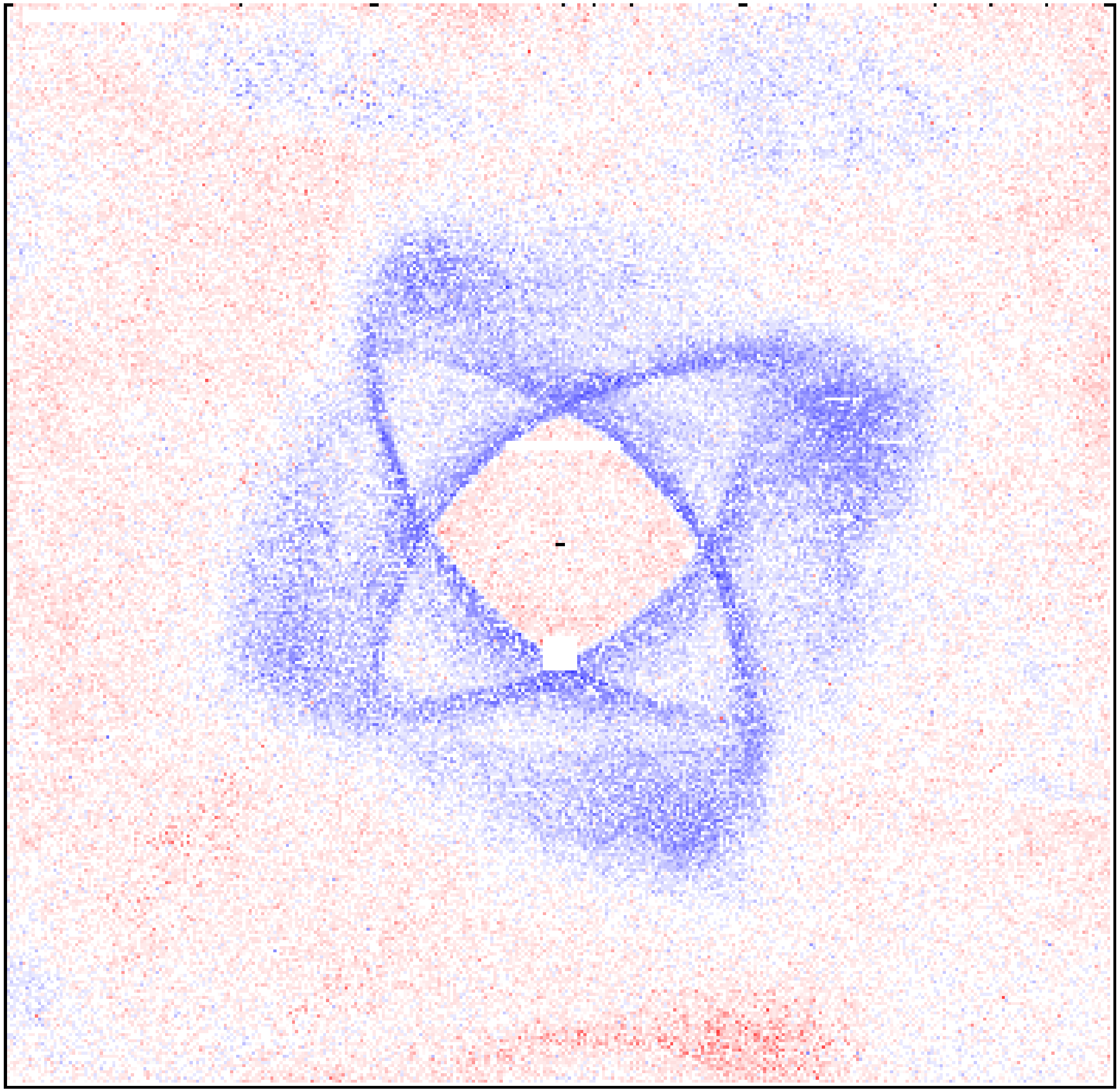
```
Pset <- fitPLM(Dilution,
        model = PM ~ -1 + probes + liver + scanner,
        normalize = FALSE, background = FALSE)
```

Result:

For each probe: weight

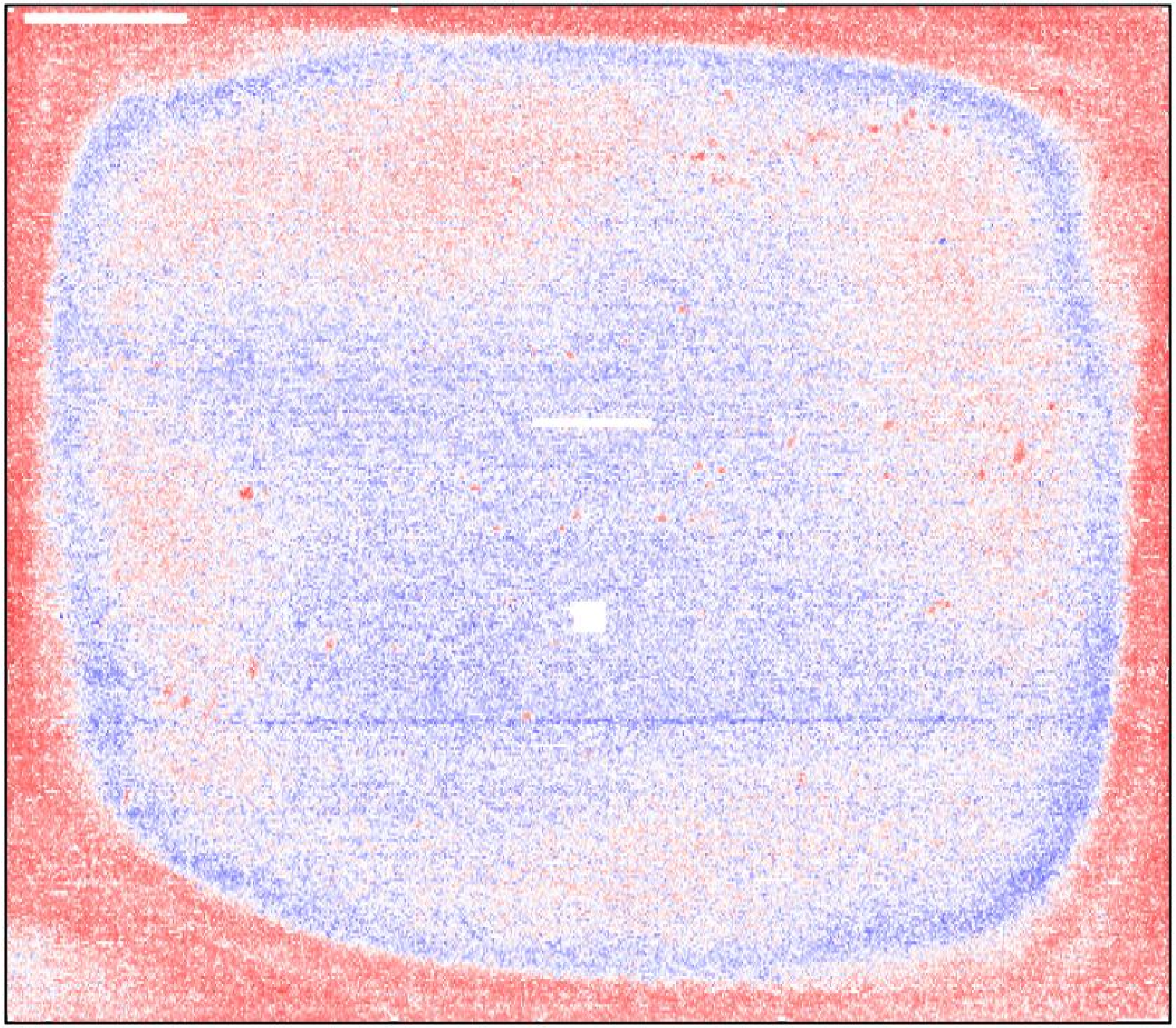
For measurement (probe*chip): residual

Ben Bolstad's PLM Image Hall of Fame



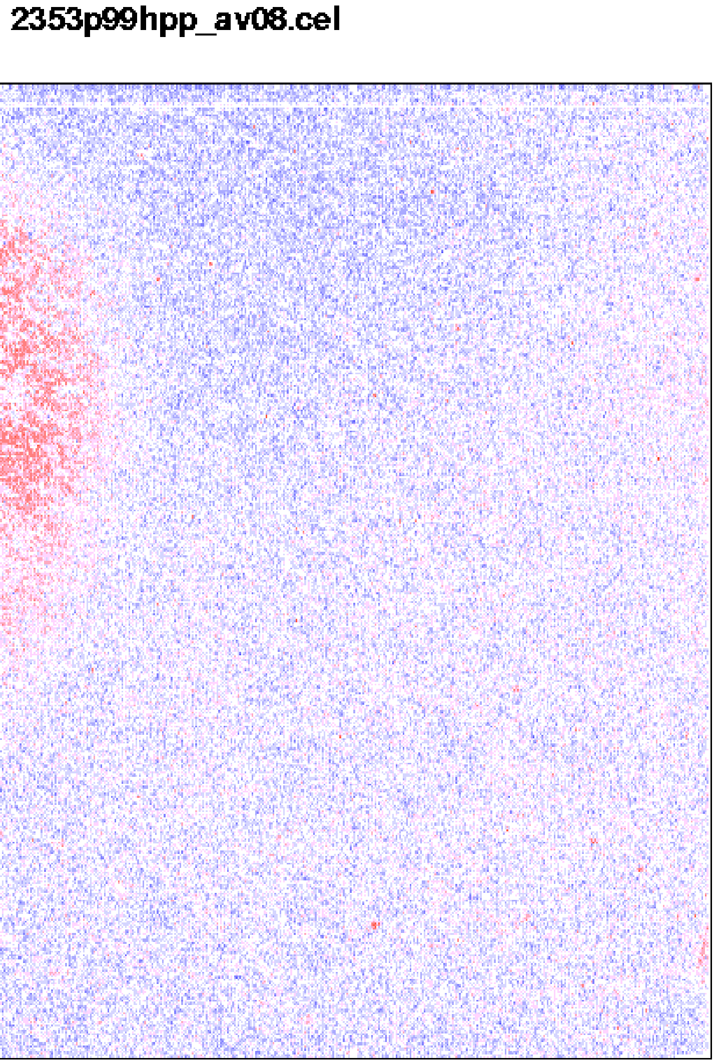
residuals

Ben Bolstad's PLM Image Hall of Fame



residuals

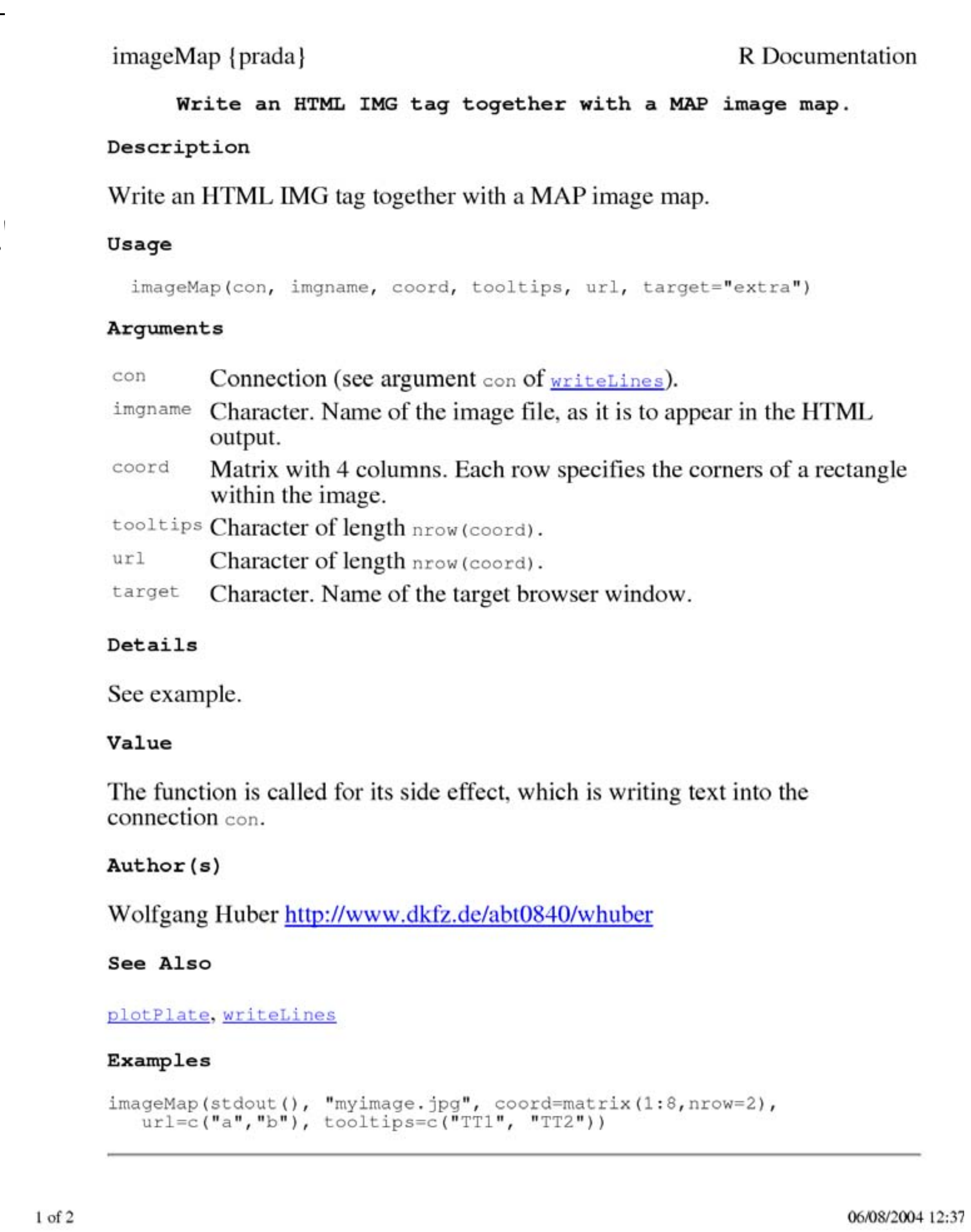
Ben Bolstad's PLM Image Hall of Fame



from Affymetrix' HGU95a latin square spike-in data set

Clickable plots via client side imagemaps

1. Plate plots
2. Domain combination gra
3. prada



imageMap {prada} R Documentation

Write an HTML IMG tag together with a MAP image map.

Description

Write an HTML IMG tag together with a MAP image map.

Usage

```
imageMap(con, imgname, coord, tooltips, url, target="extra")
```

Arguments

`con`	Connection (see argument `con` of `writeLines`).
`imgname`	Character. Name of the image file, as it is to appear in the HTML output.
`coord`	Matrix with 4 columns. Each row specifies the corners of a rectangle within the image.
`tooltips`	Character of length `nrow(coord)`.
`url`	Character of length `nrow(coord)`.
`target`	Character. Name of the target browser window.

Details

See example.

Value

The function is called for its side effect, which is writing text into the connection `con`.

Author(s)

Wolfgang Huber http://www.dkfz.de/abt0840/whuber

See Also

`plotPlate`, `writeLines`

Examples

```
imageMap(stdout(), "myimage.jpg", coord=matrix(1:8,nrow=2),
   url=c("a","b"), tooltips=c("TT1", "TT2"))
```

1 of 2 06/08/2004 12:37

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

W. Huber, A.v. Heydebreck, M. Vingron, Error models for microarray intensities (PDF file on the course CD)