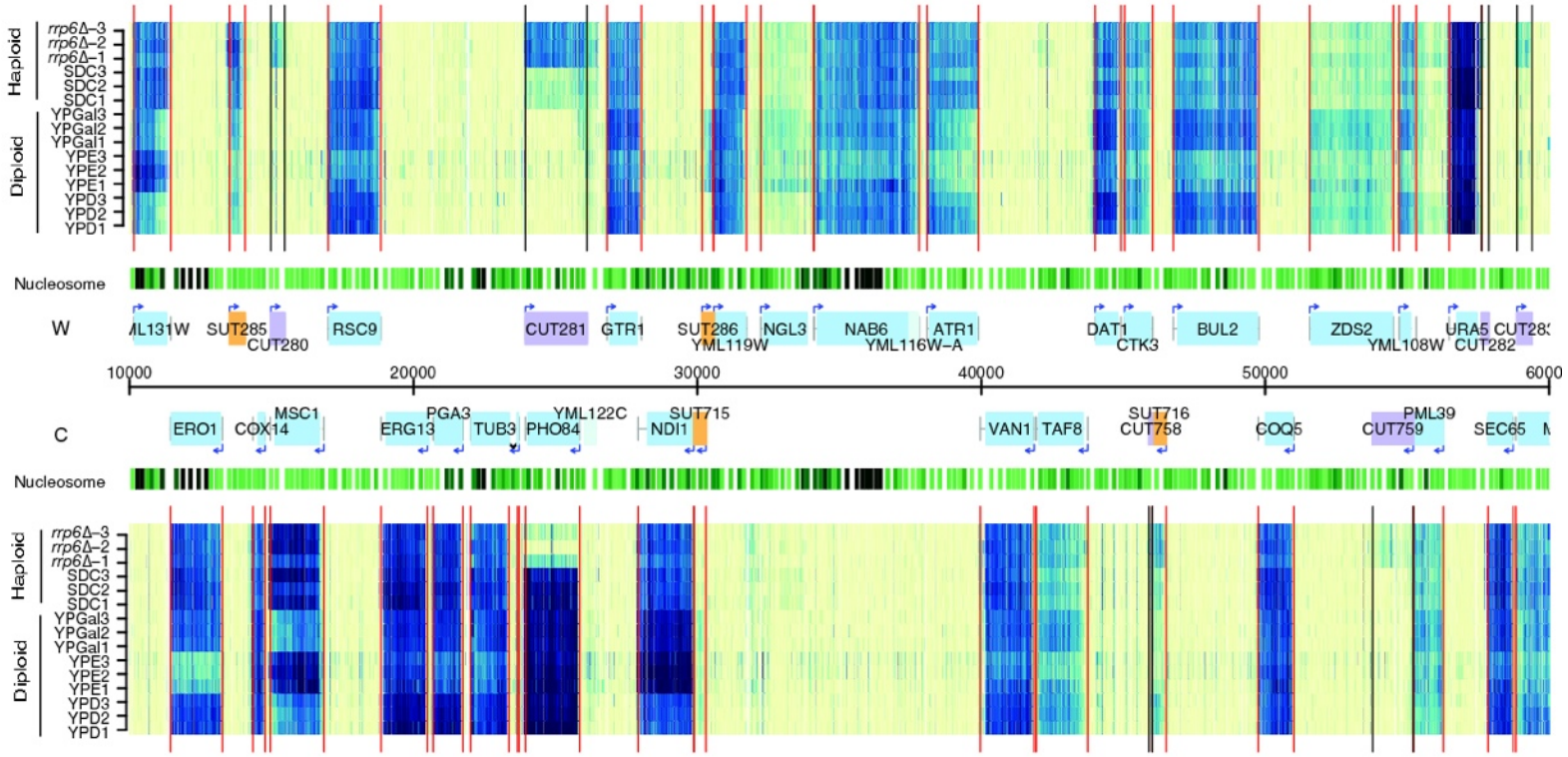
What you still might want to know about microarrays



Brixen, 24 June 2013
Wolfgang Huber
EMBL

Brief history

Late 1980s: Lennon, Lehrach: cDNAs spotted on nylon membranes

1990s: Affymetrix adapts microchip production technology for in situ oligonucleotide synthesis (commercial, patent-fenced)

1990s: Brown lab in Stanford develops two-colour spotted array technology (open and free)

1998: Yeast cell cycle expression profiling on spotted arrays (Spellmann) and Affymetrix (Cho)

1999: Tumor type discrimination based on mRNA profiles (Golub)

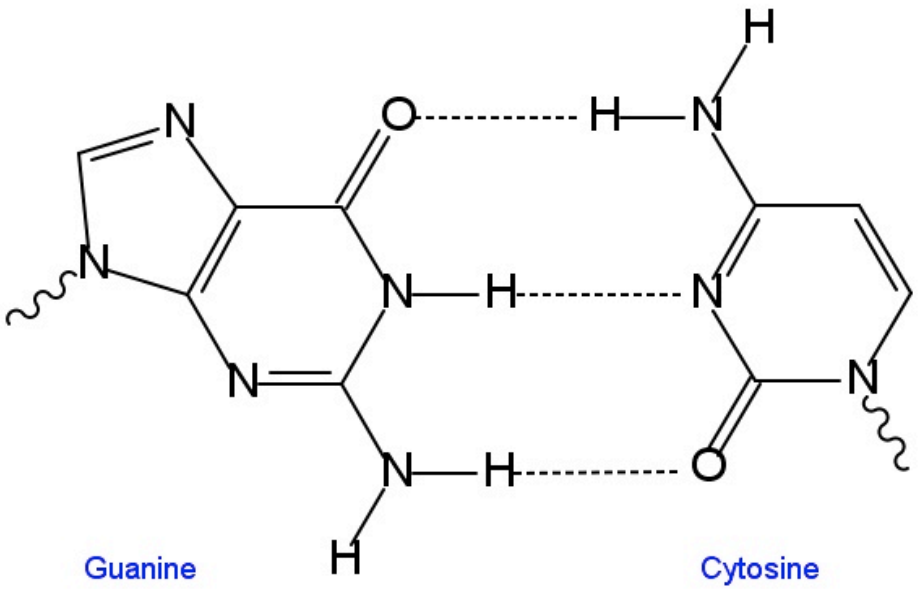
2000-ca. 2004: Affymetrix dominates the microarray market

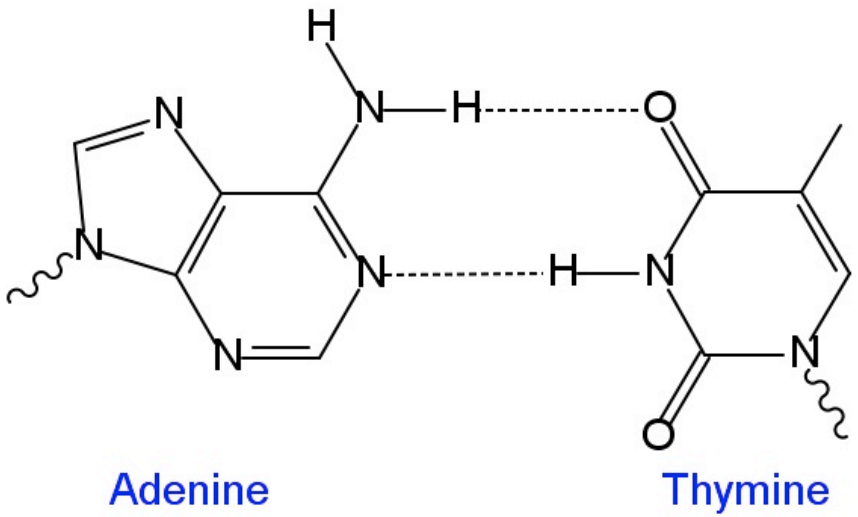
Since ~2003: Nimblegen, Illumina, Agilent (and others)

Throughout 2000‘s: CGH, CNVs, SNPs, ChIP, tiling arrays

Since ~2007: 2nd-generation sequencing (454, Solexa)

Base Pairing

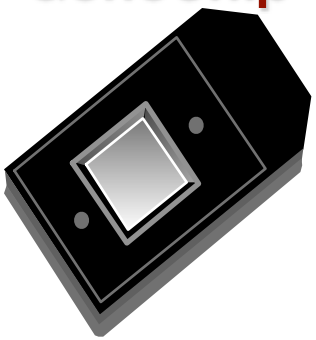




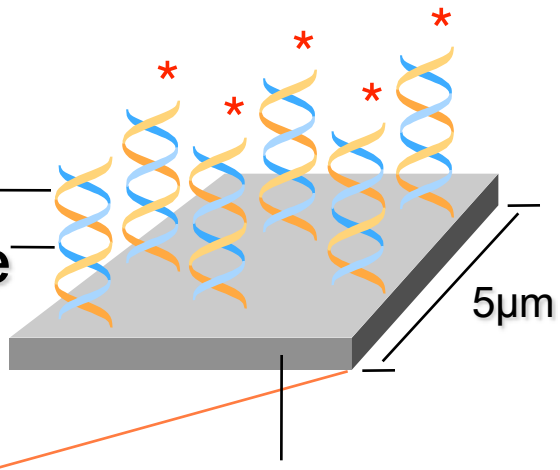
Ability to use hybridisation for constructing specific + sensitive probes at will is unique to DNA (cf. proteins, RNA, metabolites)

Oligonucleotide microarrays

GeneChip



Hybridized Probe Cell



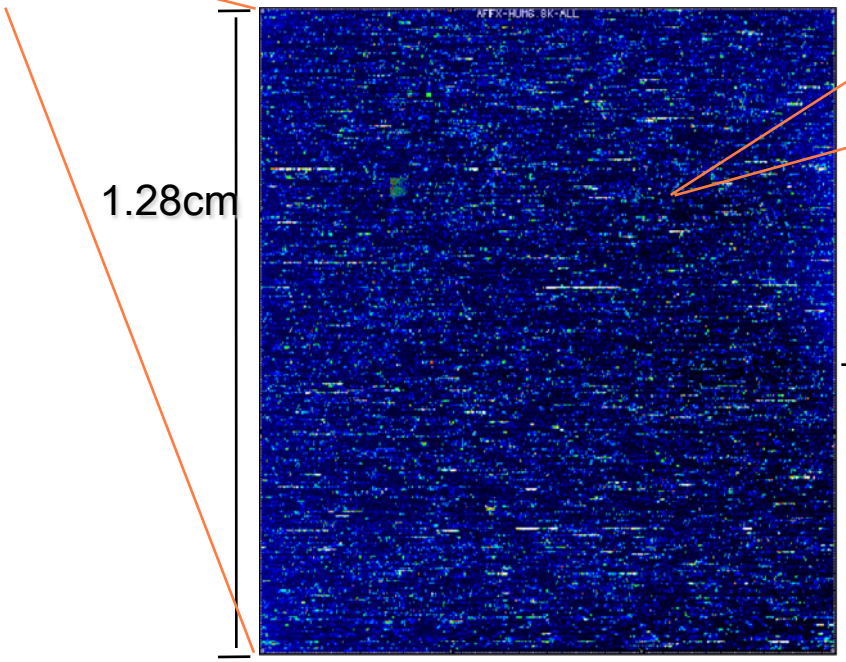
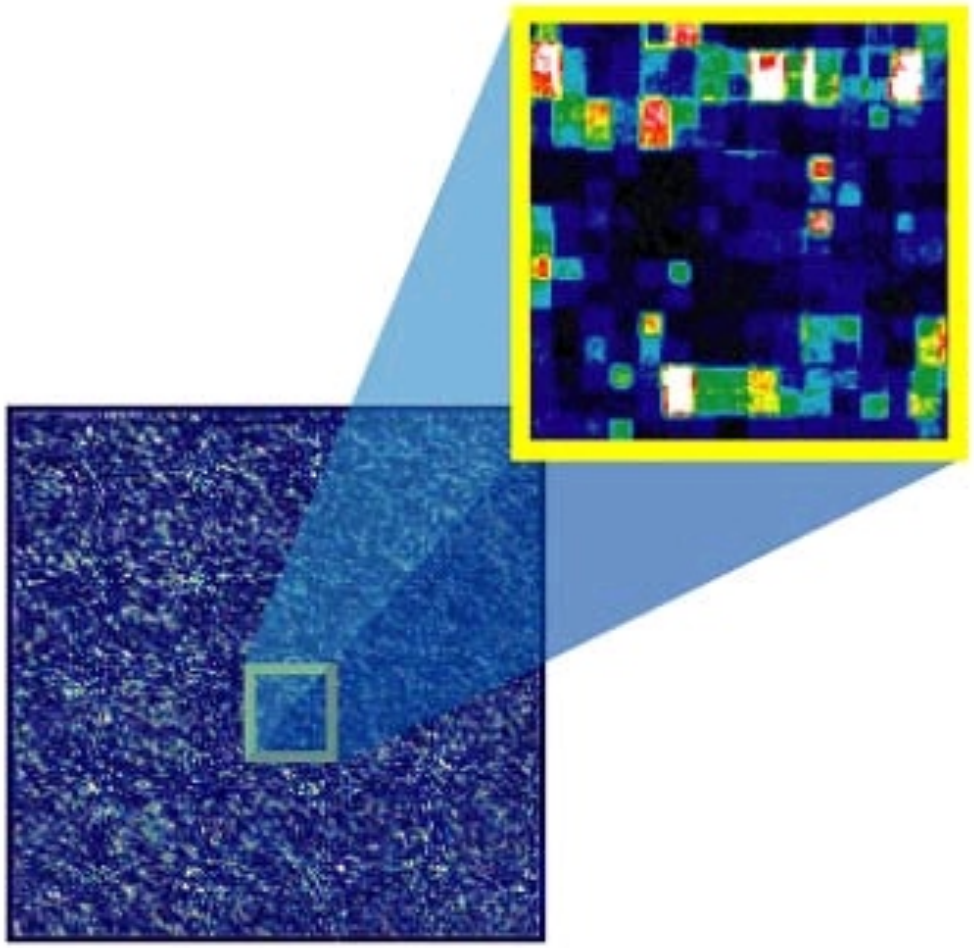


Image of array after hybridisation and staining

millions of copies of a specific oligonucleotide probe molecule per cell

up to 6.5 Mio different probe cells

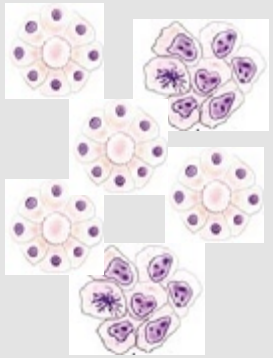
Image analysis



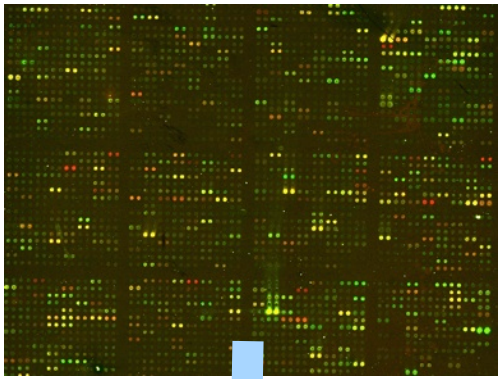
- several dozen pixels per feature
- segmentation
- summarisation into one number representing the intensity level for this feature

→ CEL file

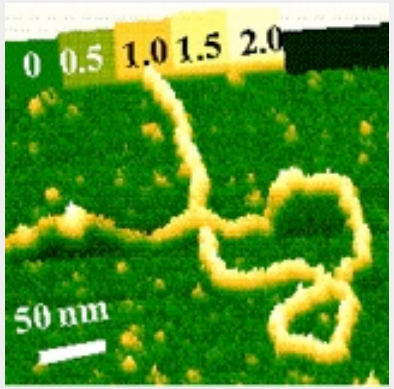
μarray data



samples:
mRNA from tissue biopsies, cell lines



fluorescent detection of the amount of sample-probe binding



arrays:
probes = gene-specific DNA strands

	tissue A	tissue B	tissue C
ErbB2	0.02	1.12	2.12
VIM	1.1	5.8	1.8
ALDH4	2.2	0.6	1.0
CASP4	0.01	0.72	0.12
LAMA4	1.32	1.67	0.67
MCAM	4.2	2.93	3.31

Microarray Analysis Tasks

Data import
reformating and setup/curation of the metadata

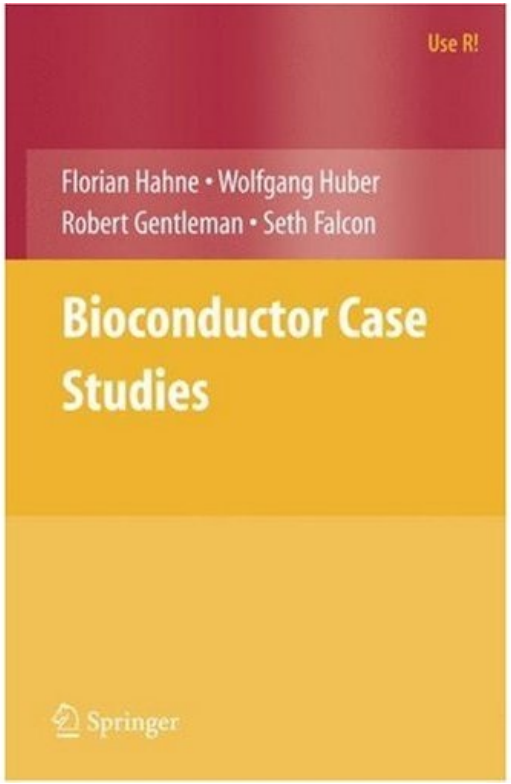
Normalisation
Quality assessment & control

Differential expression

Using gene-level annotation
Gene set enrichment analysis

Clustering & Classification

Integration of other datasets



Platform-specific data import and initial processing

Affymetrix 3' IVT (e.g. Human U133 Plus 2.0, Mouse 430 2.0):

`affy`

Affymetrix Exon (e.g. Human Exon 1.0 ST):

`oligo, exonmap, xps`

Affymetrix SNP arrays:

`oligo`

Illumina bead arrays:

`beadarray, lumi`

http://www.bioconductor.org/docs/workflows/oligoarrays

Flexible data import

Using generic `R` I/O functions and constructors

`Biobase`

`limma`

Chapter *Two Color Arrays* in the useR-book.

`limma` user guide

Normalisation and quality assessment

preprocessCore

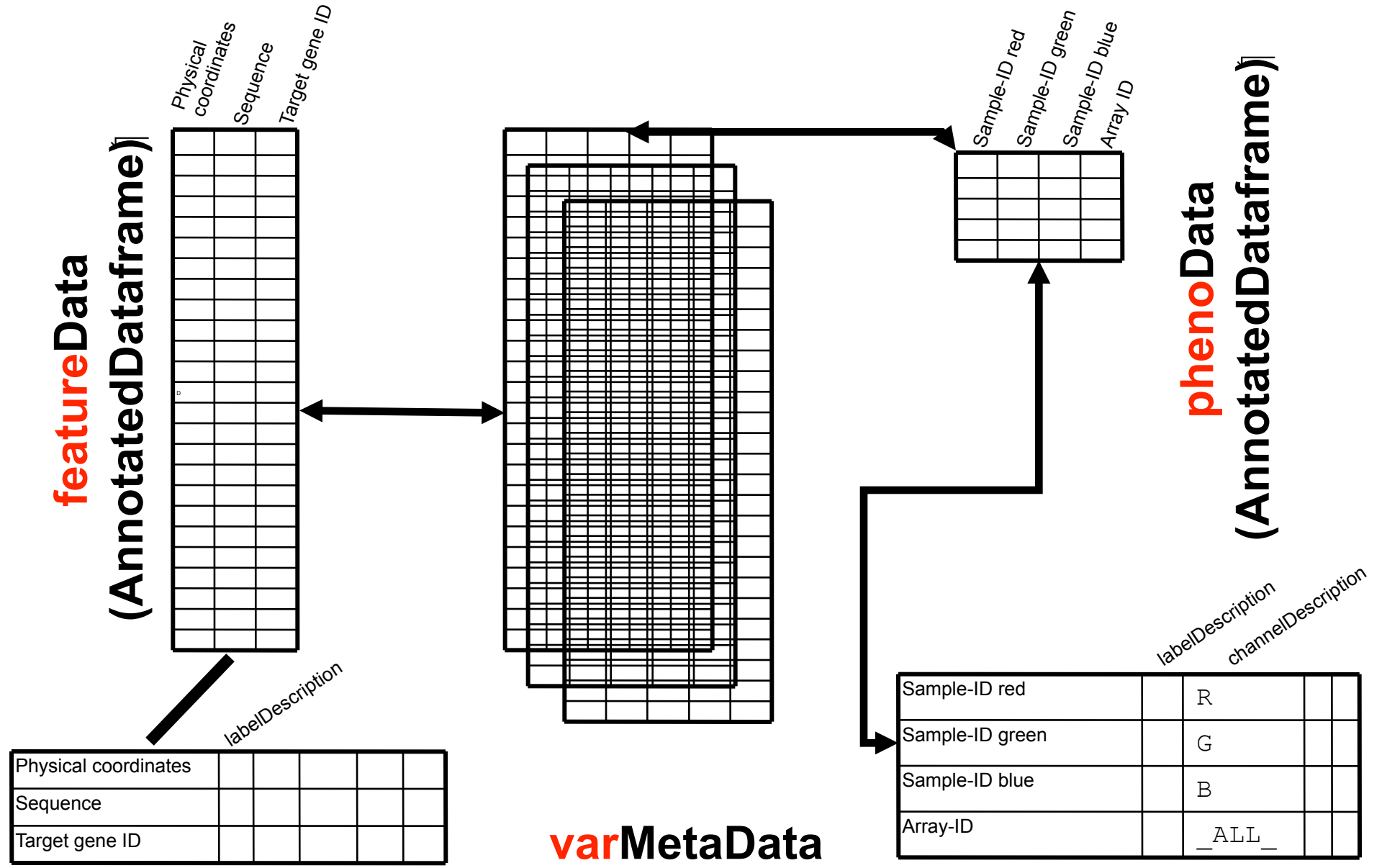
limma

vsn

arrayQualityMetrics

NChannelSet

assayData can contain N=1, 2, ..., matrices of the same size



Annotation / Metadata

Keeping data together with the metadata (about reporters, target genes, samples, experimental conditions, ...) is one of the major principles of Bioconductor

- avoid alignment bugs
- facilitate discovery

→ Matrices with “rich” column and row names.

Annotation infrastructure for Affymetrix

`hgu133plus2probe` nucleotide sequence of the features (for preprocessing e.g. **`gcrma`**; for own annotation)

`hgu133plus2cdf` maps the physical features on the array to probe sets

`hgu133plus2.db` maps probe sets to target genes and provides target gene annotation collected from public databases

What is wrong with microarray data?

Many data are measured in definite units:

- **time in seconds**
- **lengths in meters**
- **energy in Joule, etc.**

Climb Mount Plose (2465 m) from Brixen (559 m) with weight of 76 kg, working against a gravitation field of strength 9.81 m/s^2 :

$$(2465 - 559) \cdot 76 \cdot 9.81 \ \text{m kg m/s}^2$$
$$= 1\,421\,037 \ \text{kg m}^2 \text{s}^{-2}$$
$$= 1\,421.037 \ \text{kJ}$$

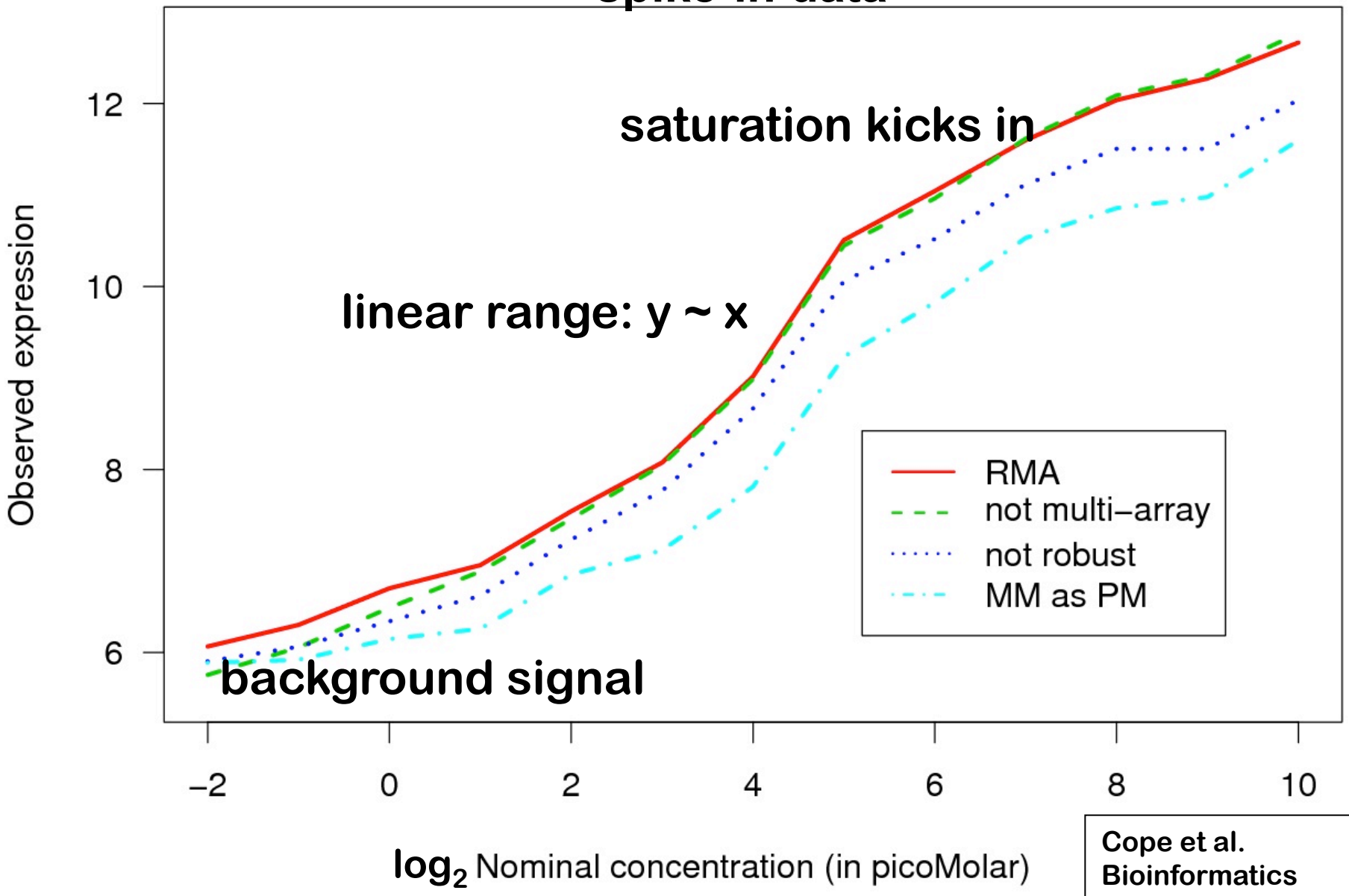
A complex measurement process lies between mRNA concentrations and intensities

- RNA degradation
- quality of actual probe sequences
- image segmentation
- a
- eff
- r
- tra
- eff
- h
- eff
- spe
- labeling efficiency
- optical noise

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

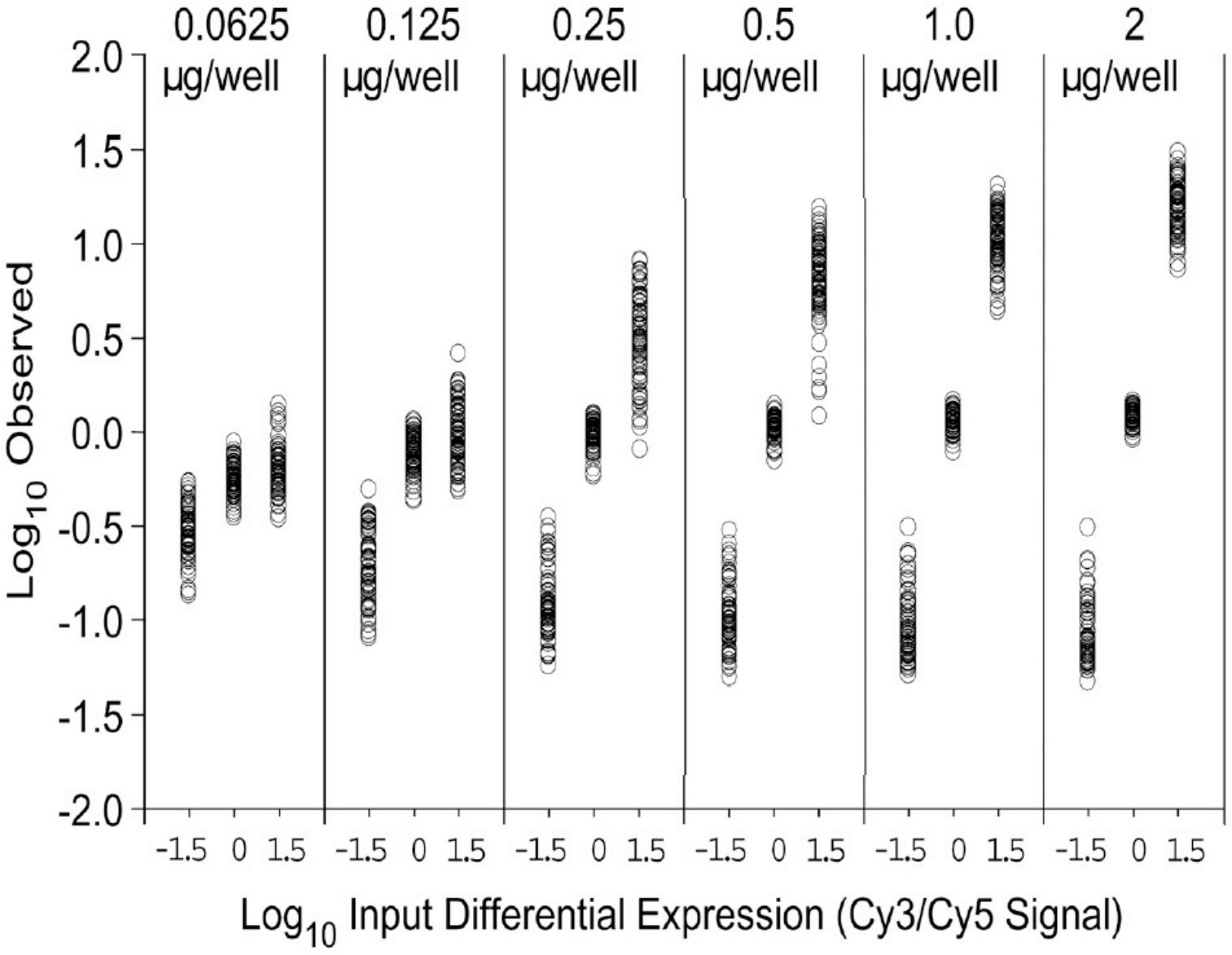
Background signal and non-linearities

"mild" non-linearity



Cope et al.
Bioinformatics
2003

ratio compression



nominal 3:1

nominal 1:1

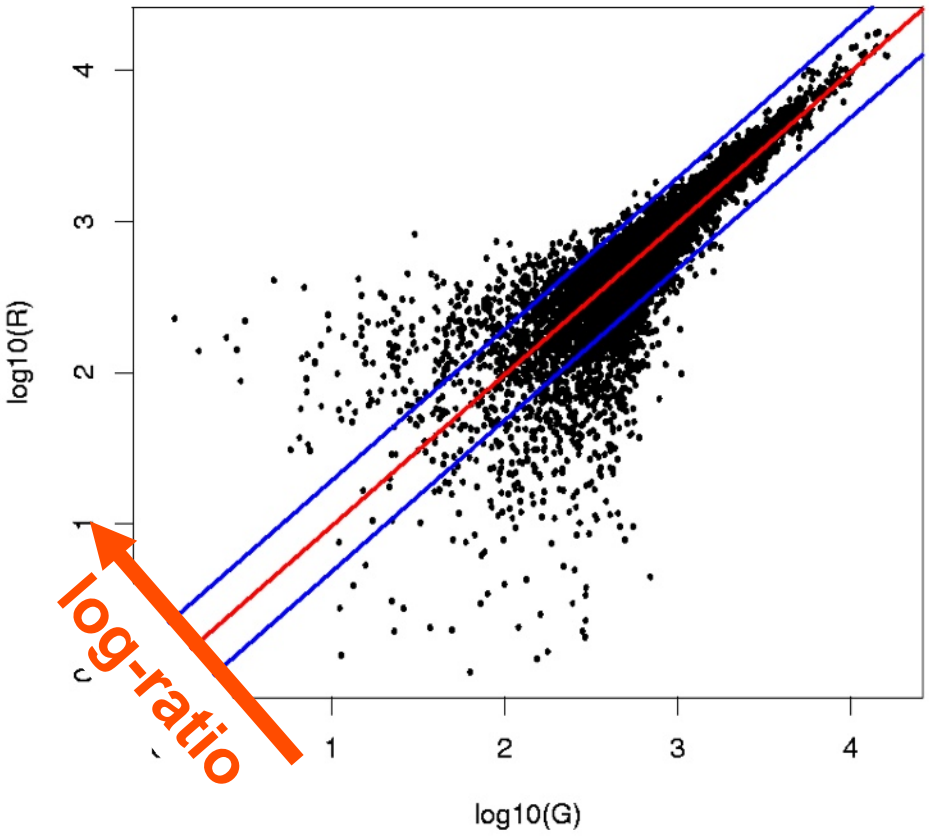
nominal 1:3

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

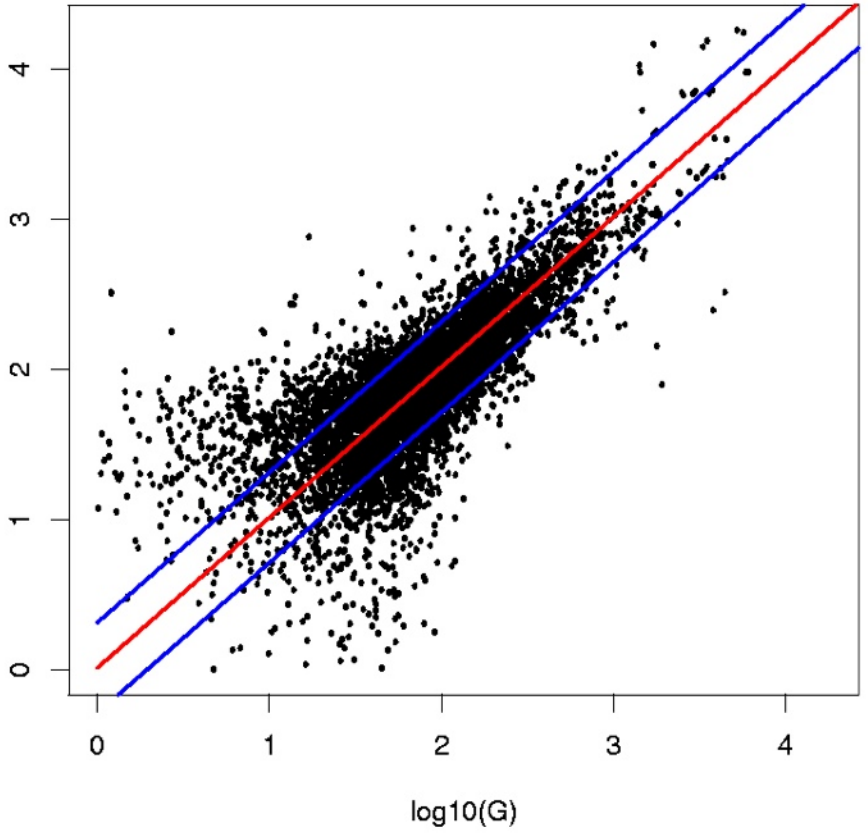
Statistical issues

Which genes are differentially transcribed?

same-same



tumor-normal



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

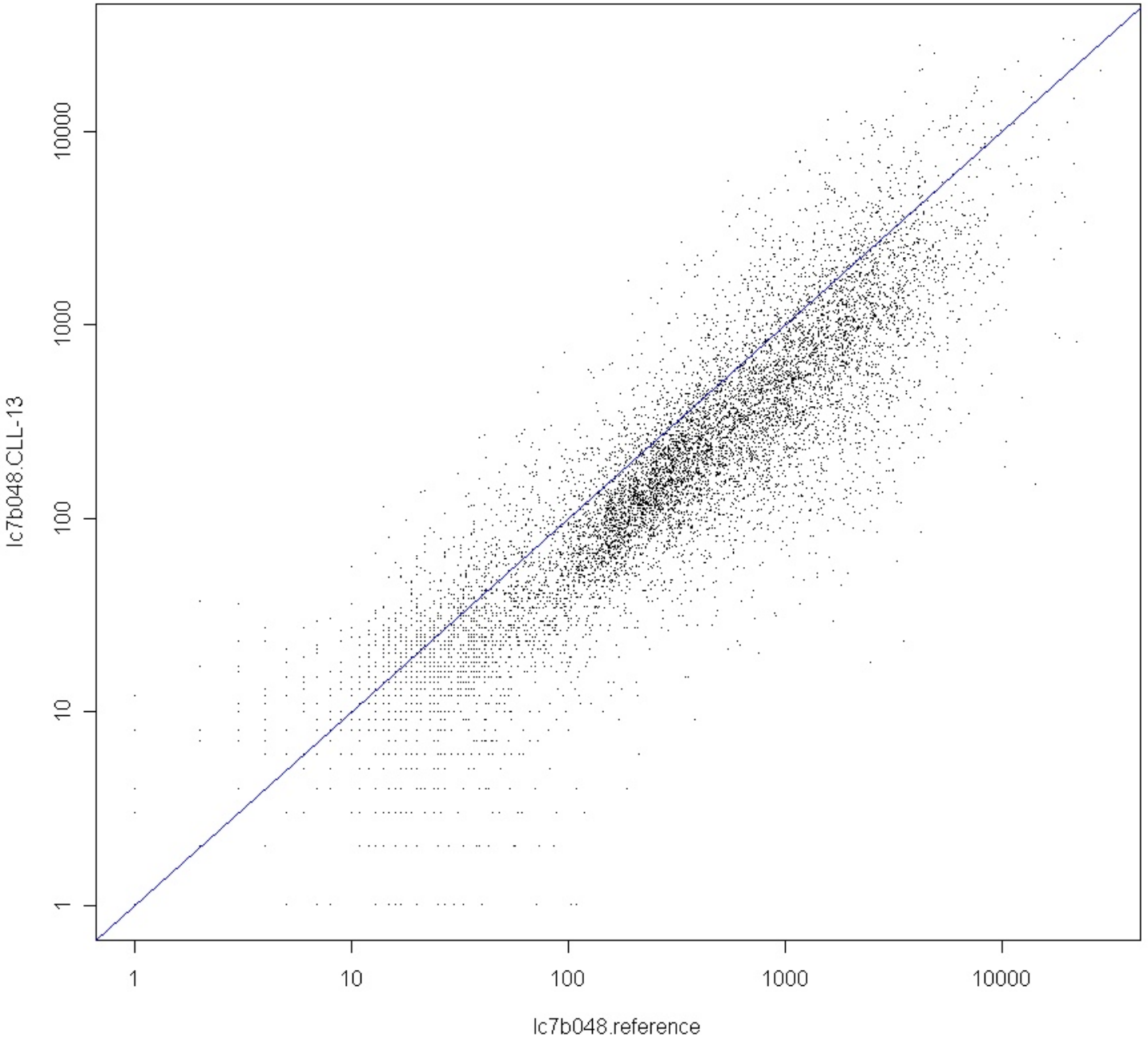
- too random to be explicitely accounted for
- remain as “noise”

Calibration

Error model

Why do you need ‘normalisation’ (a.k.a. calibration)?

Systematic effects



From: lymphoma dataset

vsn package

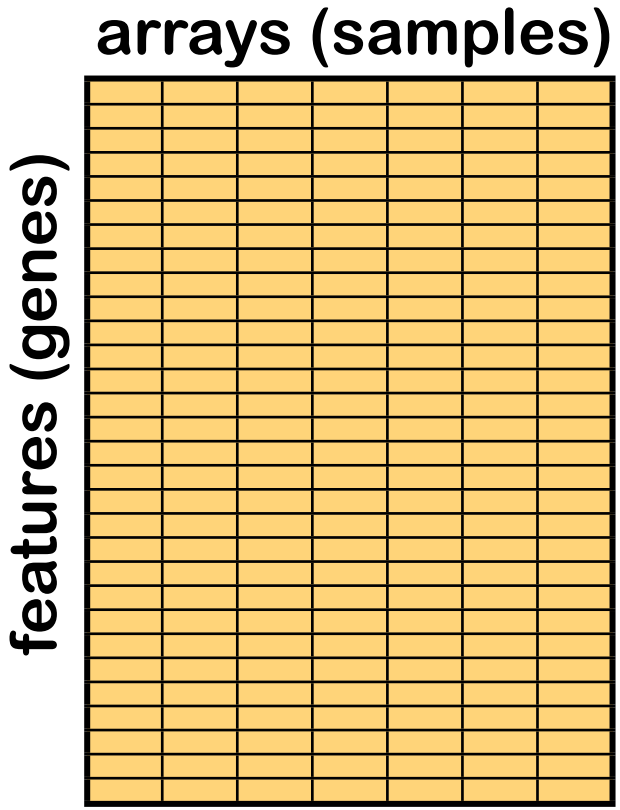
Alizadeh et al., Nature 2000

Quantile normalisation

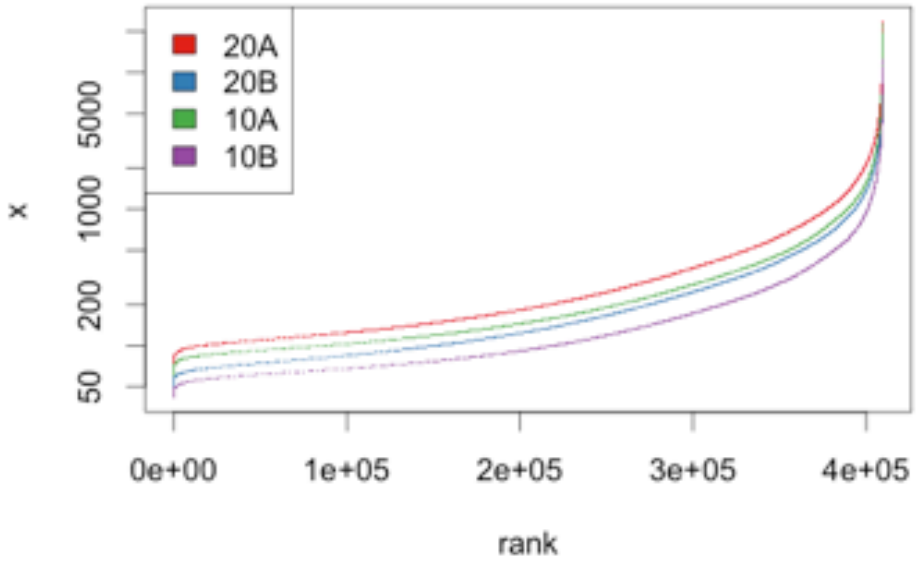
Within each column (array), replace the intensity values by their rank

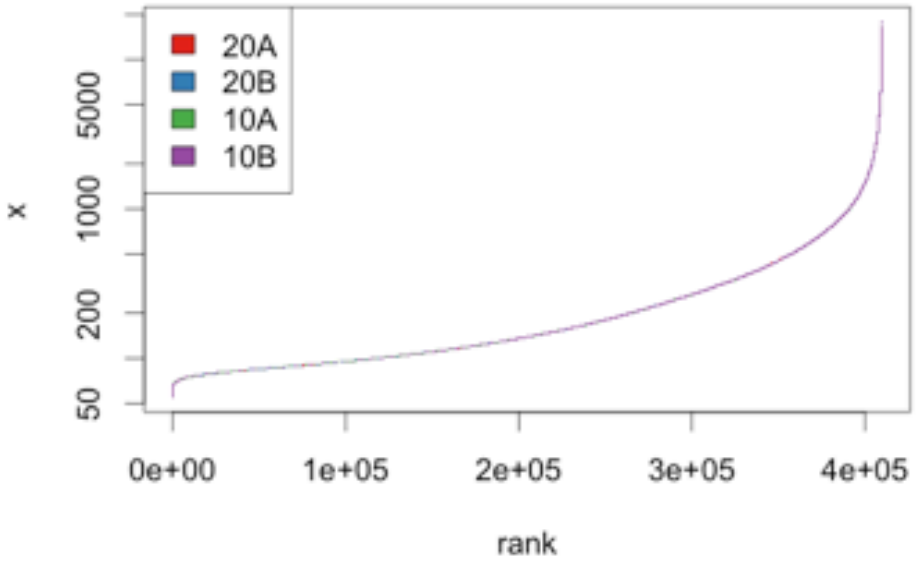
For each rank, compute the average of the intensities with that rank, across columns (arrays)

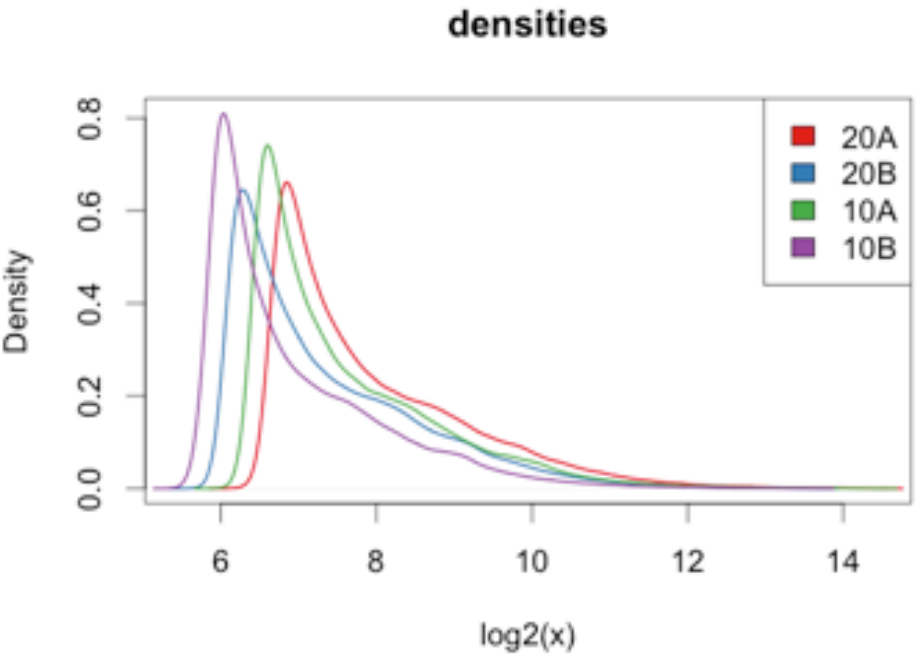
Replace the ranks by those averages

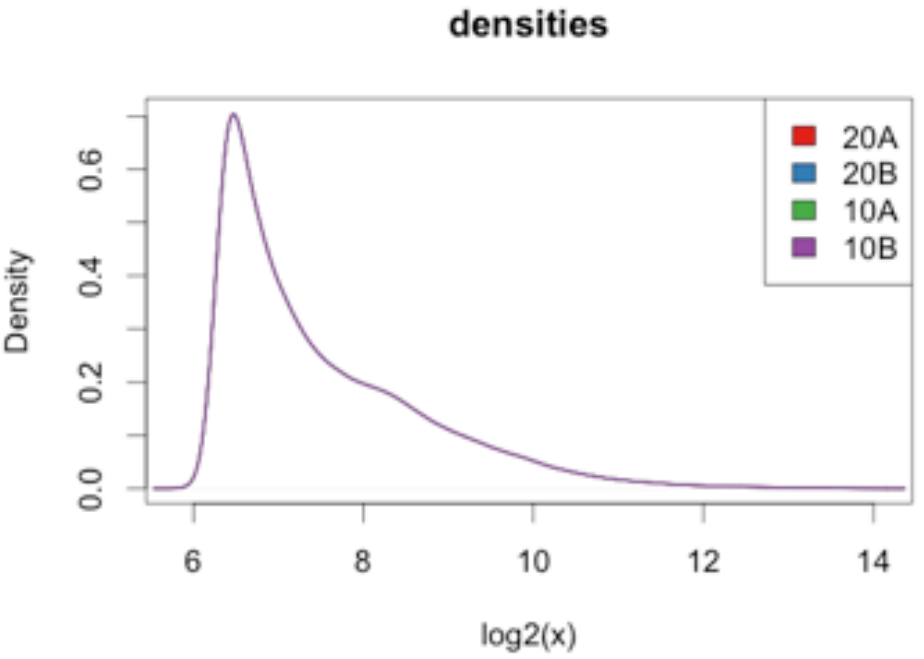


Ben Bolstad 2001


20A
20B
10A
10B
x
50
200
1000
5000
0e+00
1e+05
2e+05
3e+05
4e+05
rank


20A
20B
10A
10B
x
50
200
1000
5000
0e+00
1e+05
2e+05
3e+05
4e+05
rank


densities
20A
20B
10A
10B
Density
0.0
0.2
0.4
0.6
0.8
6
8
10
12
14
log2(x)


densities
20A
20B
10A
10B
Density
0.0
0.2
0.4
0.6
6
8
10
12
14
log2(x)

Quantile normalisation

+ Simple, fast, easy to implement
+ Always works, needs no user interaction / tuning
+ Non-parametric: can correct for quite nasty non-linearities (saturation, background) in the data

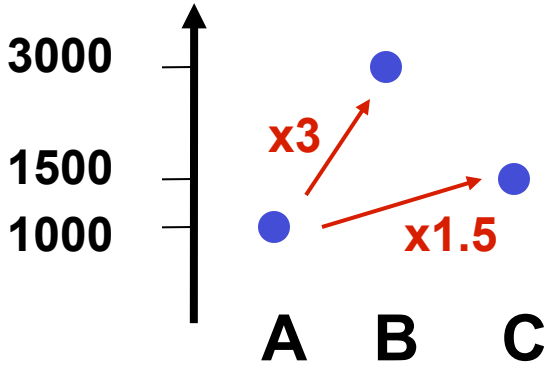
- Always "works", even if data are bad / inappropriate
- May be conservative: rank transformation looses information - may yield less power to detect differentially expressed genes
- Aggressive: if there is an excess of up- (or down) regulated genes, it removes not just technical, but also biological variation

Less aggressive methods exist, e.g. loess, vsn

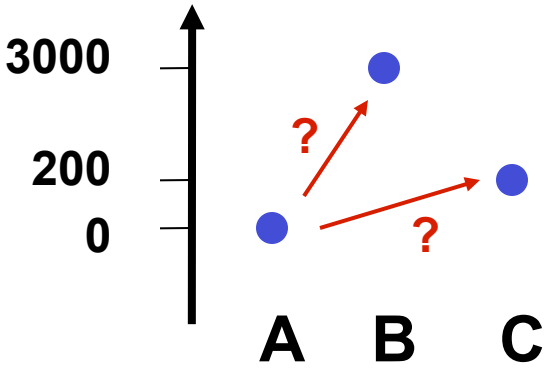
Estimating relative expression (fold-changes)

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 $2^{-2} = ¼$**

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.

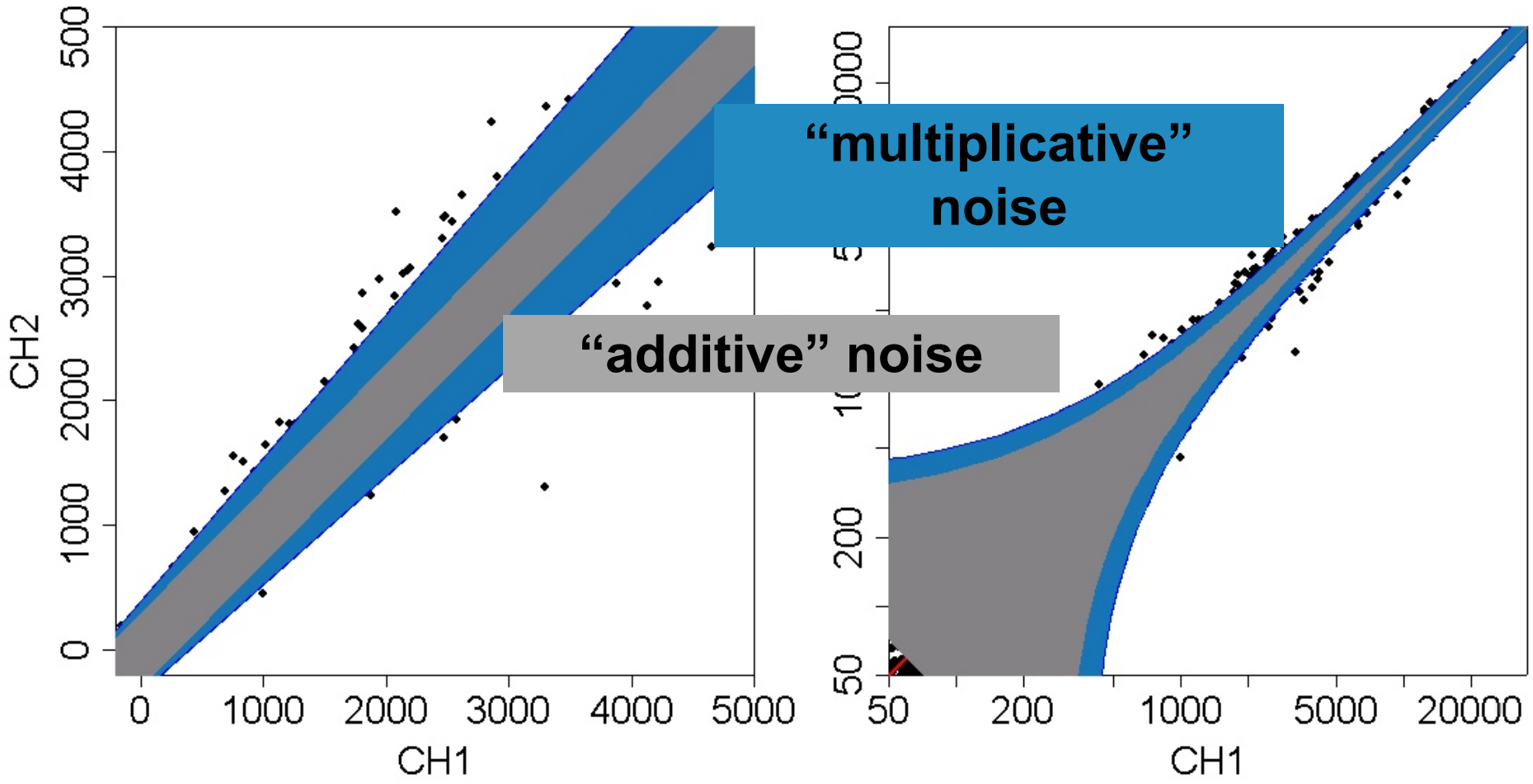
…. data reduction

What about a change from 0 to 500?

- conceptually

- noise, measurement precision

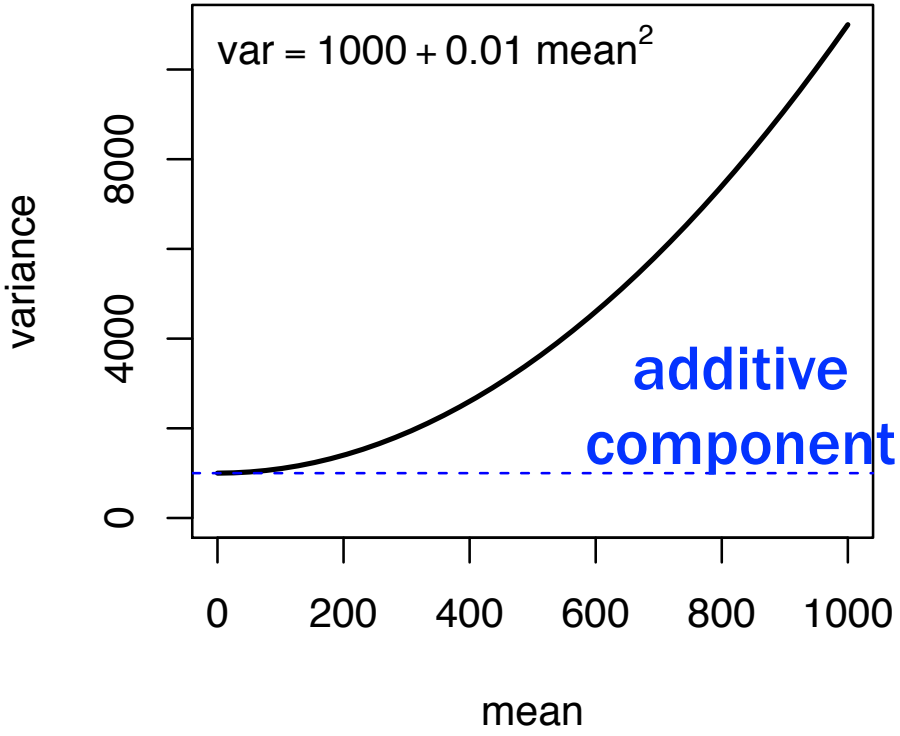
The two-component model for microarray data

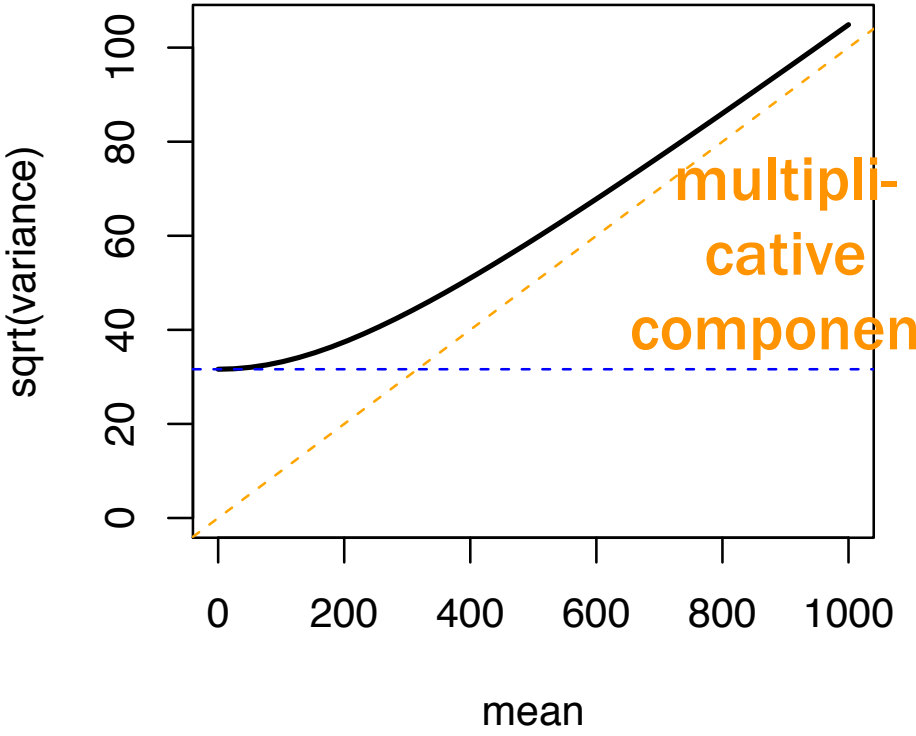


raw scale **log scale**

B. Durbin, D. Rocke, JCB 2001

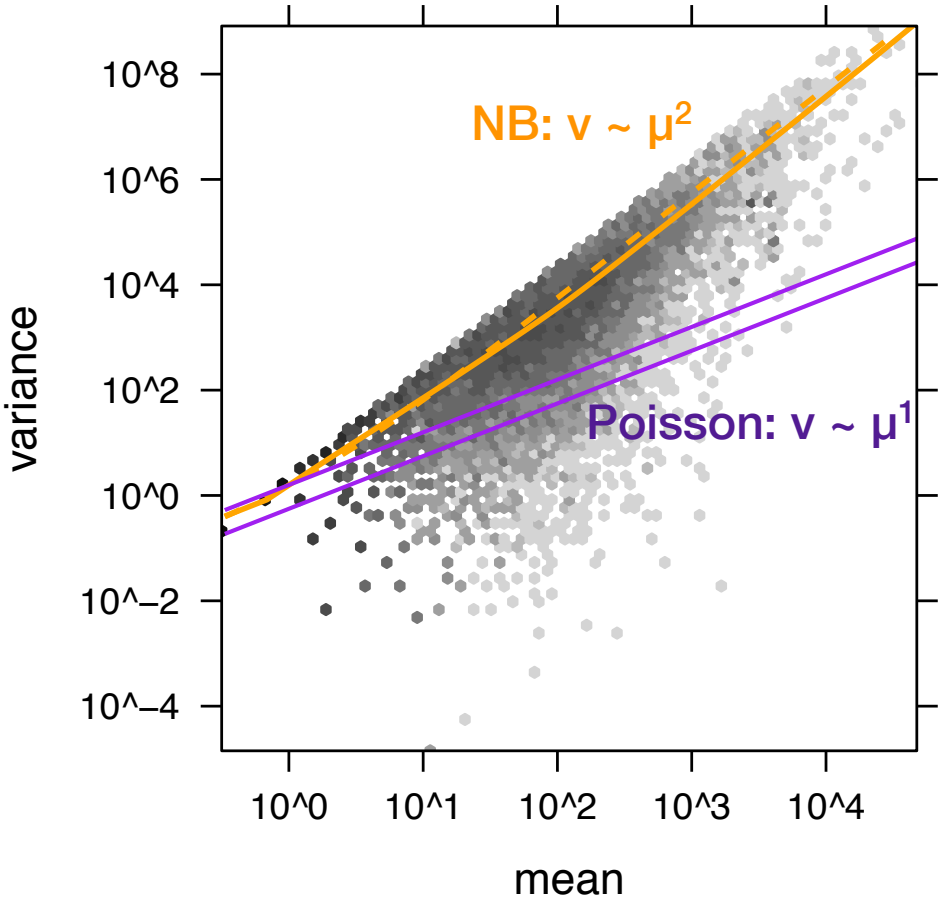
The additive-multiplicative error model





Trey Ideker et al.: JCB (2000)
David Rocke and Blythe Durbin: JCB (2001), Bioinformatics (2002)
For robust affine regression normalisation: W. Huber et al. Bioinformatics (2002)
For background correction in RMA: R. Irizarry et al., Biostatistics (2003)

Two component error models



Microarrays

$\mathrm{var}(\mu) = b + c \cdot \mu^2$

b: background

c: asymptotic coefficient of variation

Sequencing counts

early edgeR:

$\mathrm{var}(\mu) = \mu + \alpha \cdot \mu^2$

μ: from Poisson

α: dispersion

DESeq

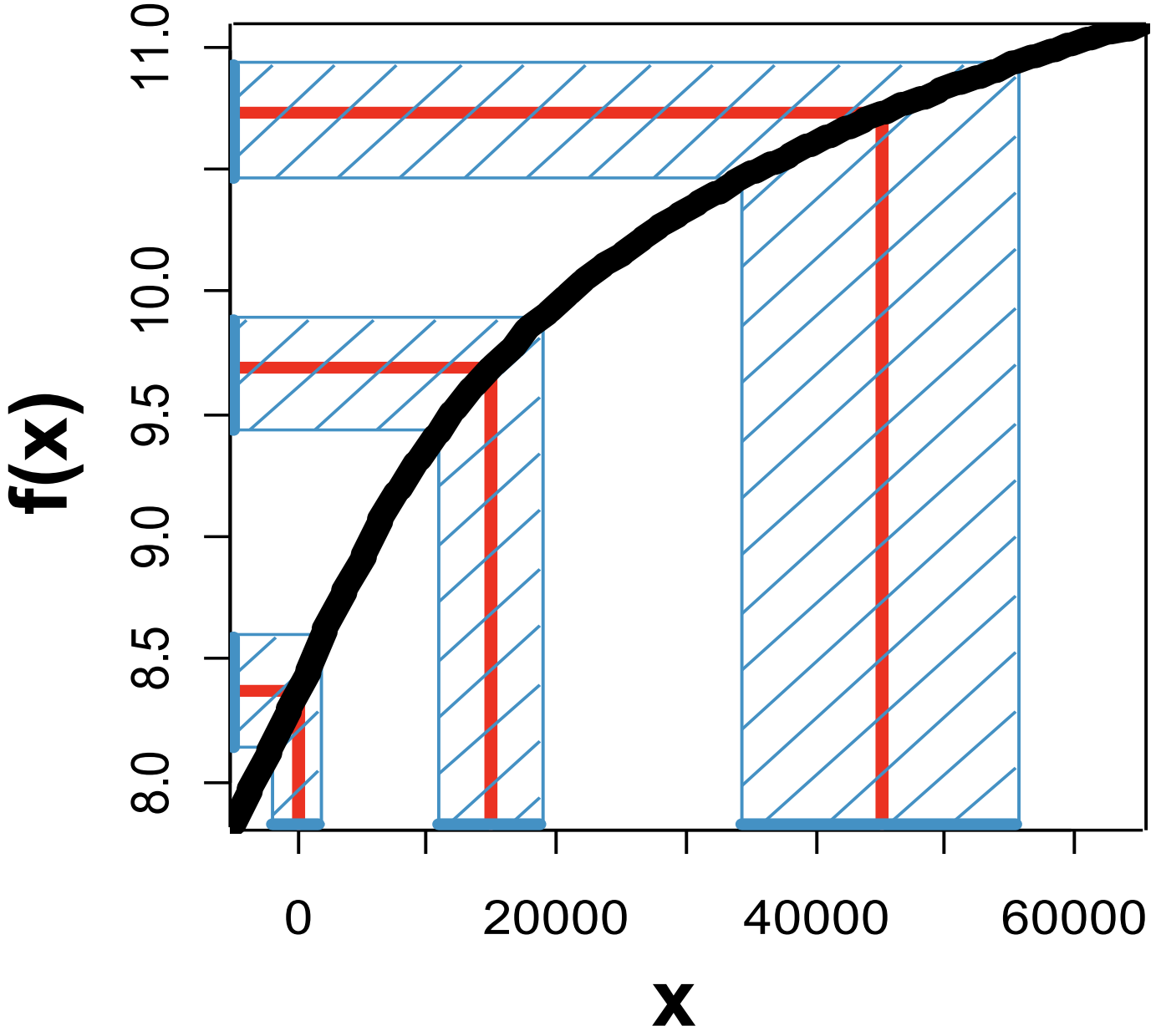
$\mathrm{var}(\mu) = \mu + \alpha(\mu) \cdot \mu^2$

DESeq parametric option

$\alpha(\mu) = a_1/\mu + a_0 \quad \Leftrightarrow$

$\mathrm{var}(\mu) = \mu + a_1 \cdot \mu + a_0 \cdot \mu^2$

variance stabilizing transformation



variance stabilizing transformations

X_u a family of random variables with

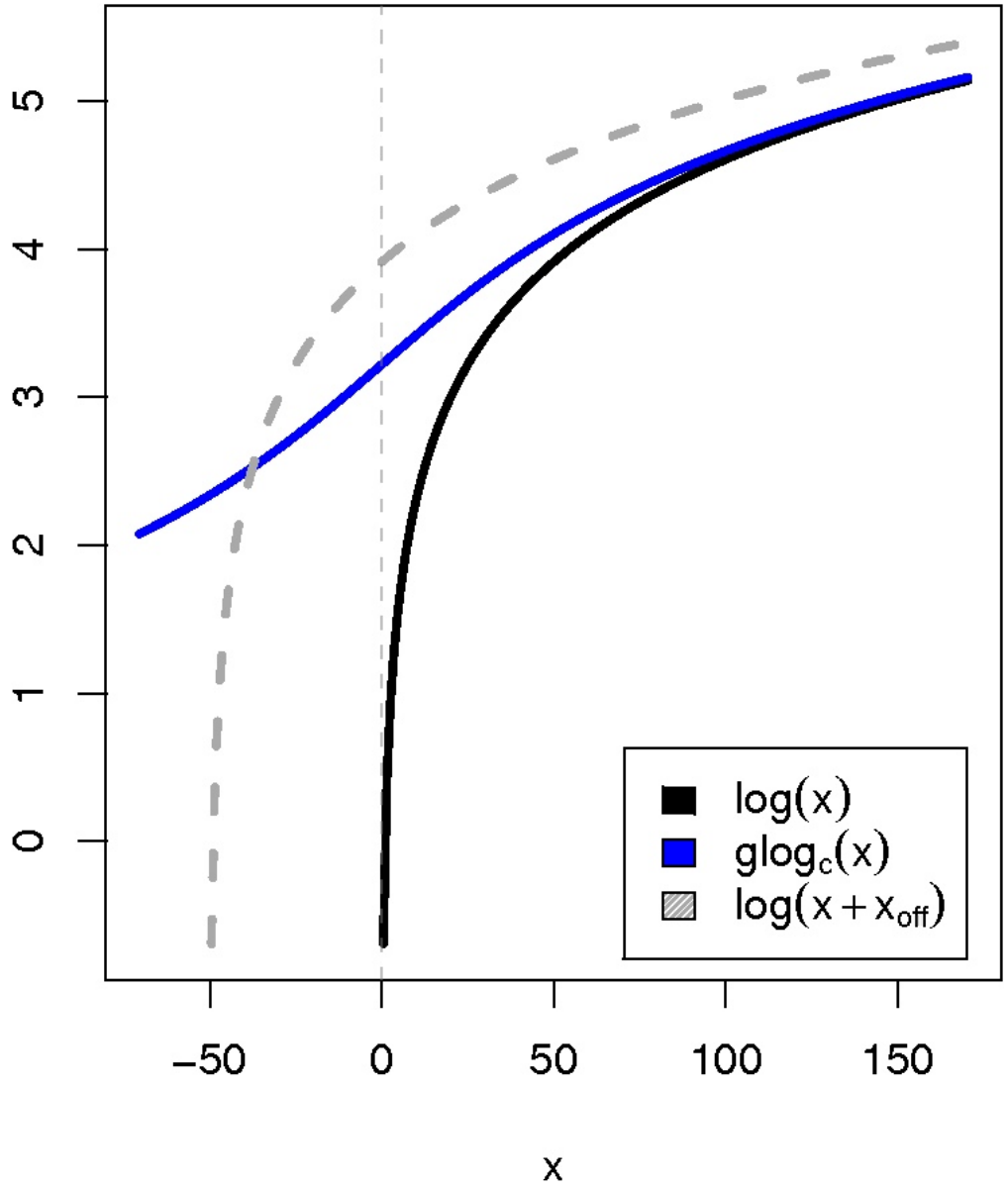
$E(X_u) = u$ and $Var(X_u) = v(u)$. Define

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

Then, var $f(X_u) \approx$ does not depend on u

Derivation: linear approximation,
relies on smoothness of $v(u)$.

the “glog” transformation



$$\mathrm{glog}_2(x,c) = \log_2\left(\frac{x+\sqrt{x^2+c^2}}{2}\right)$$

$$\mathrm{glog}_e(x,1) + \log_e 2 = \mathrm{arsinh}(x)$$

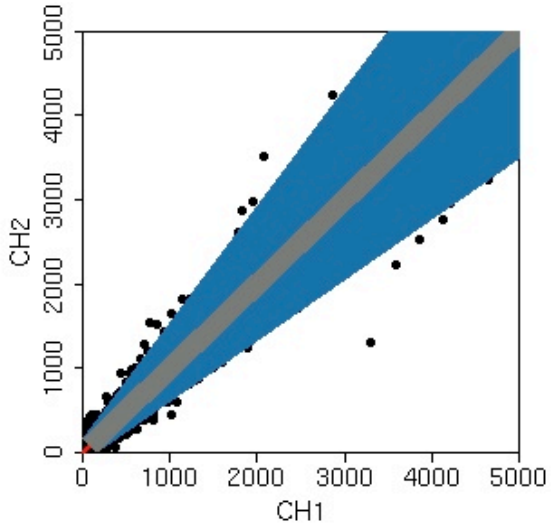
P. Munson, 2001

D. Rocke & B. Durbin, ISMB 2002

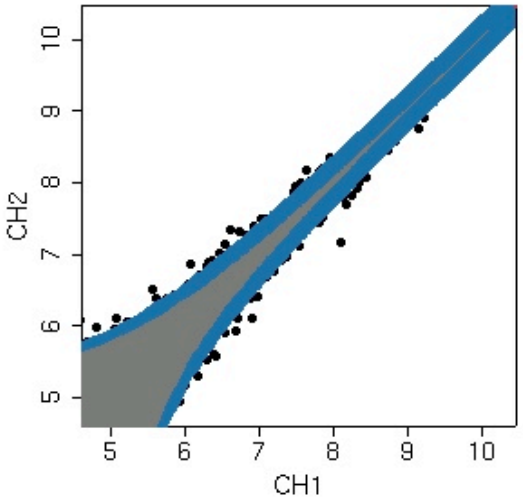
W. Huber et al., ISMB 2002

glog

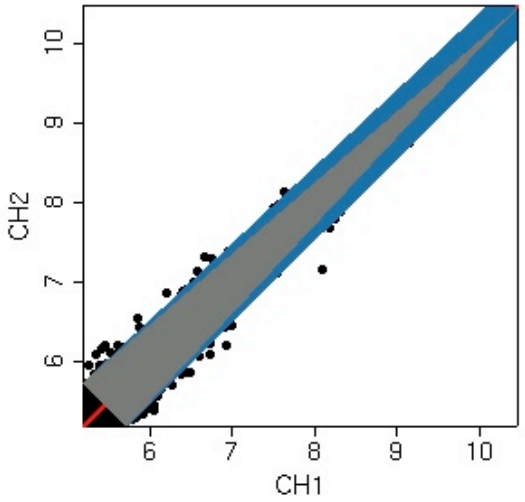
raw scale



log



glog



variance:

constant part

proportional part

“usual” log-ratio

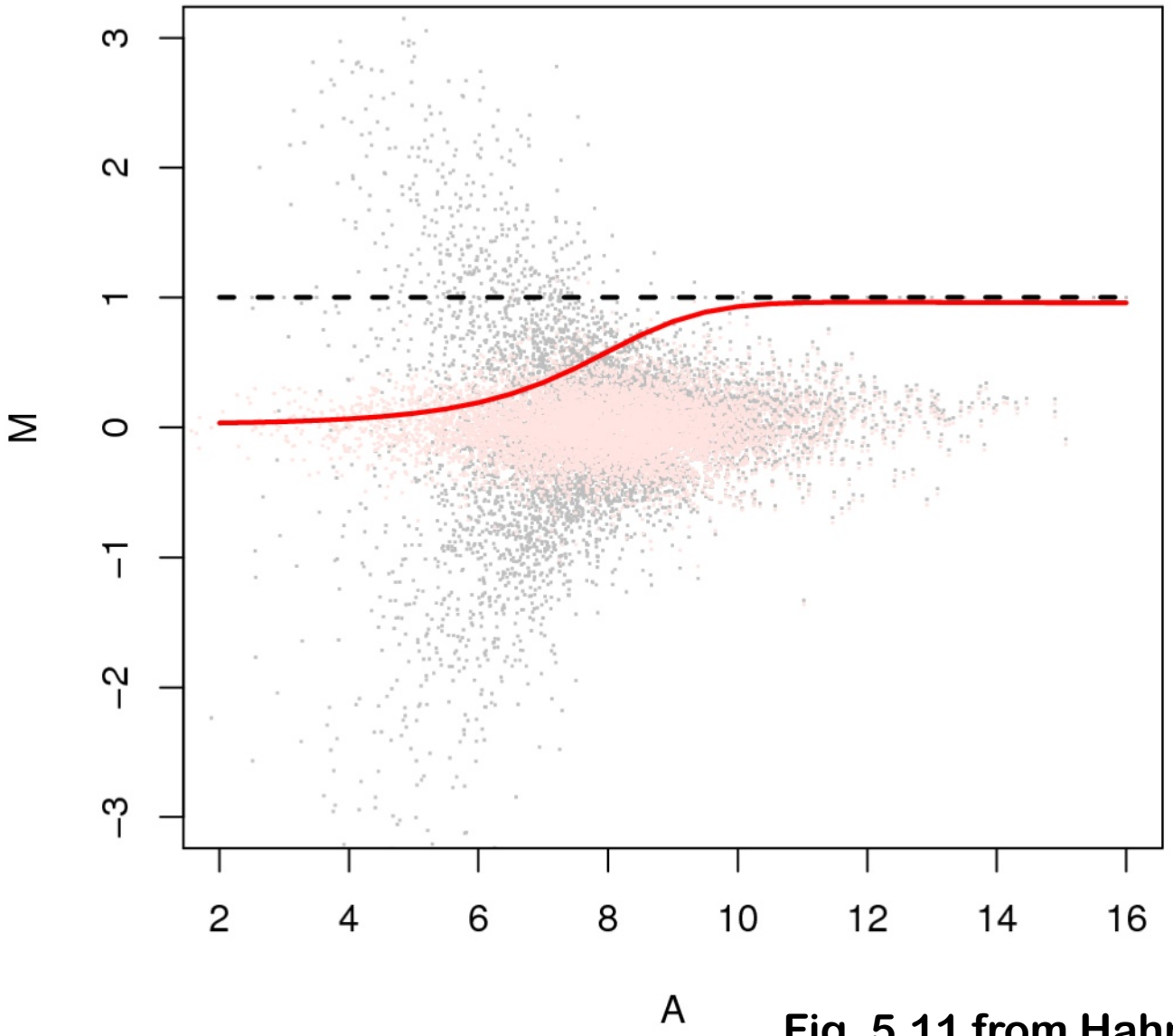
$$\log \frac{x_1}{x_2}$$

'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 and shrinkage estimators



Same-same comparison

log-ratio

glog-ratio

Lines: 29 data points with *observed* ratio of 2

Fig. 5.11 from Hahne et al. (useR book)

Variance-bias trade-off and shrinkage estimators

Shrinkage estimators:

a general technology in statistics:
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 is a shrinkage estimator for log fold change

Quality assessment

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).

Normalization and analysis of DNA microarray data by self-consistency and local regression. T.B. Kepler, L. Crosby, K. Morgan. Genome Biology. 3(7):research0037 (2002)

Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. S. Dudoit, Y.H. Yang, M. J. Callow, T. P. Speed. Technical report # 578, August 2000 (UC Berkeley Dep. Statistics)

A Benchmark for Affymetrix GeneChip Expression Measures. L.M. Cope, R.A. Irizarry, H. A. Jaffee, Z. Wu, T.P. Speed. Bioinformatics (2003).

….many, many more…

Acknowledgements

Anja von Heydebreck (Merck, Darmstadt)
Robert Gentleman (Genentech, San Francisco)
Günther Sawitzki (Uni Heidelberg)
Martin Vingron (MPI, Berlin)
Rafael Irizarry (JHU, Baltimore)
Terry Speed (UC Berkeley)
Lars Steinmetz (EMBL Heidelberg)
Audrey Kauffmann (Novartis, Basel)
David Rocke (UC Davis)

Summaries for Affymetrix genechip probe sets

Data and notation

PM_{ikg} , MM_{ikg} = Intensities for perfect match and mismatch probe *k* for gene *g* on chip *i*

i = 1,…, *n* one to hundreds of chips

k = 1,…, *J* usually 11 probe pairs

g = 1,…, *G* tens of thousands of probe sets.

Tasks:

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

summarize for each probe set the probe level data, i.e., 11 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 used AvDiff, a trimmed mean:

$$AvDiff = \frac{1}{\#K} \sum_{k \in K} (PM_k - MM_k)$$

- **sort $d_k = PM_k - MM_k$**
- **exclude highest and lowest value**
- **K := 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***

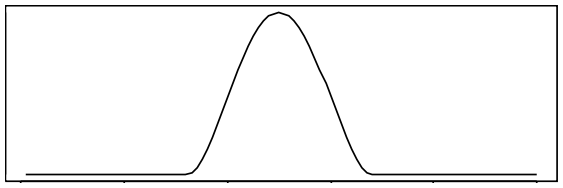
= PM / "typical log-ratio" **if *MM>=PM***

Signal = Weighted mean of the values log(PM-CT)

weights follow Tukey Biweight function

(location = data median,

scale a fixed multiple of MAD)



Expression measures: Li & Wong

dChip **fits a model for each gene**

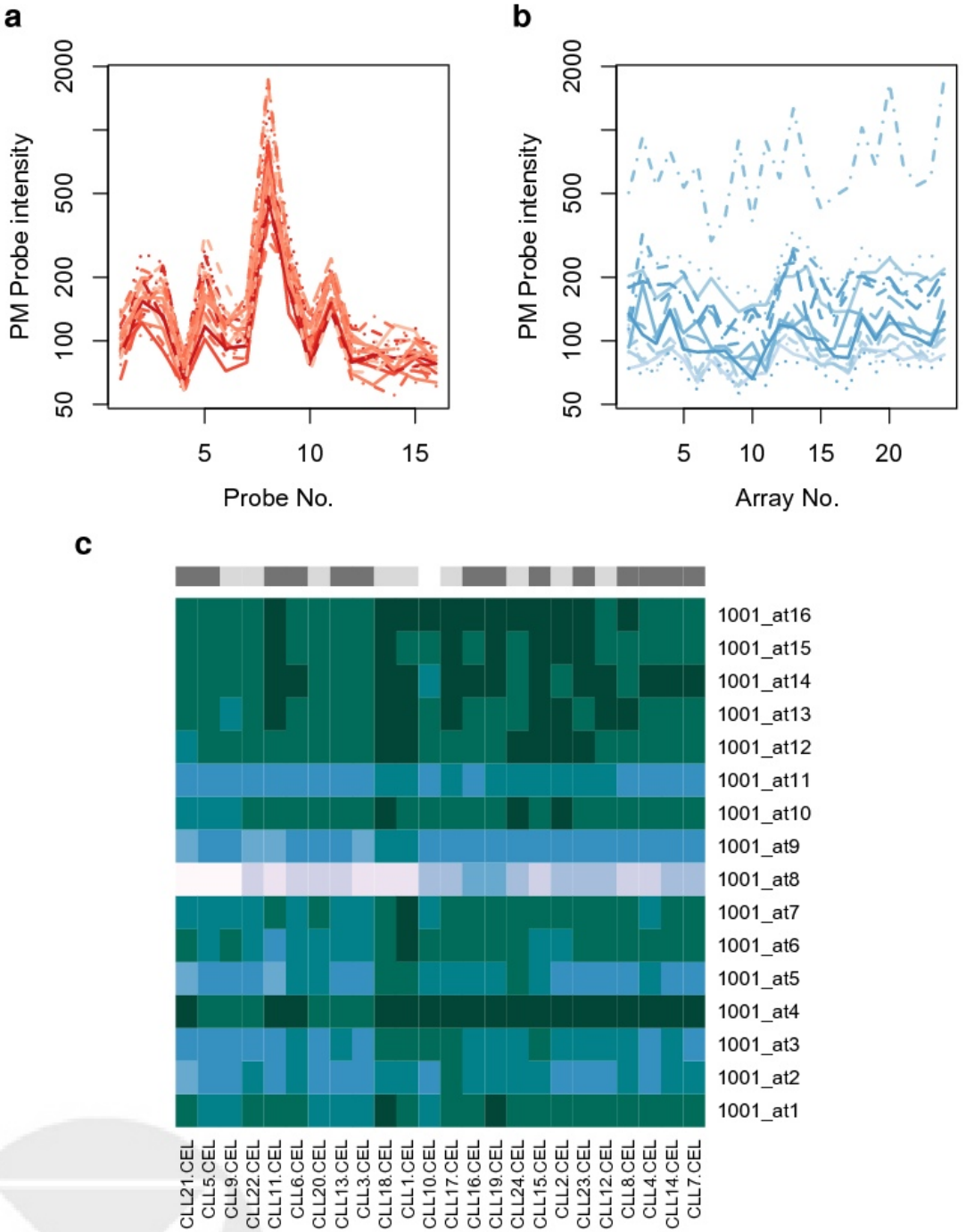
$$PM_{ki} - MM_{ki} = \theta_k \phi_i + \varepsilon_{ki}, \qquad \varepsilon_{ki} \propto N(0, \sigma^2)$$

where

ϕ_i **: expression measure for the gene in sample *i***

θ_k **: probe effect**

ϕ_i **is estimated by maximum likelihood**


a
PM Probe intensity
2000
500
200
100
50
5
10
15
Probe No.
b
PM Probe intensity
2000
500
200
100
50
5
10
15
20
Array No.
c
1001_at16
1001_at15
1001_at14
1001_at13
1001_at12
1001_at11
1001_at10
1001_at9
1001_at8
1001_at7
1001_at6
1001_at5
1001_at4
1001_at3
1001_at2
1001_at1
CLL21.CEL
CLL5.CEL
CLL9.CEL
CLL22.CEL
CLL11.CEL
CLL6.CEL
CLL20.CEL
CLL13.CEL
CLL3.CEL
CLL18.CEL
CLL1.CEL
CLL10.CEL
CLL17.CEL
CLL16.CEL
CLL19.CEL
CLL24.CEL
CLL15.CEL
CLL2.CEL
CLL23.CEL
CLL12.CEL
CLL8.CEL
CLL4.CEL
CLL14.CEL
CLL7.CEL

Expression measures
RMA: Irizarry et al. (2002)

dChip

$$Y_{ki} = \theta_k \phi_i + \varepsilon_{ki}, \qquad \varepsilon_{ki} \propto N(0, \sigma^2)$$

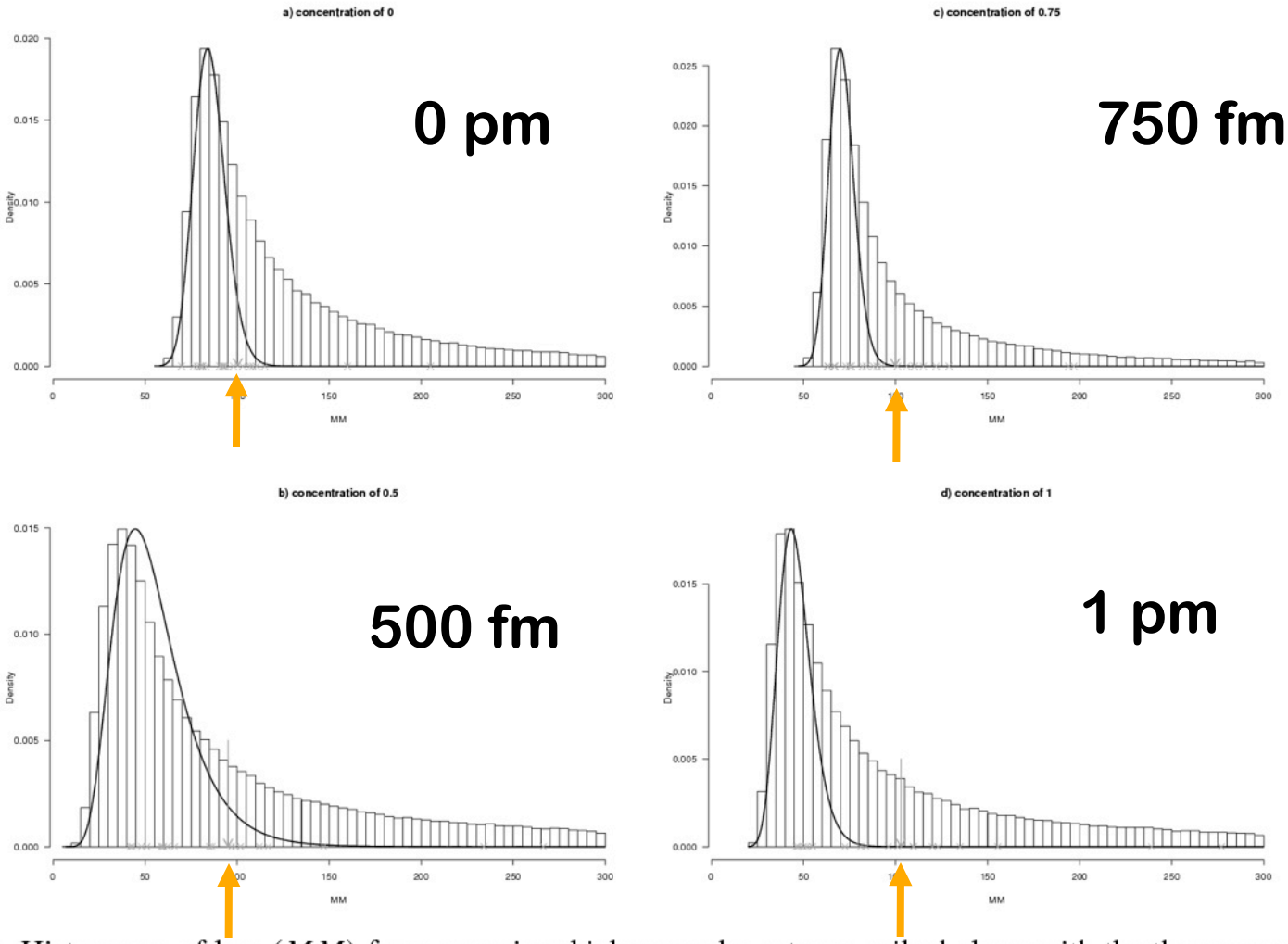
RMA

$$\log_2 Y_{ki} = a_k + b_i + \varepsilon_{ki}$$

b_i is estimated using the robust method median polish (successively remove row and column medians, accumulate terms, until convergence).

further background correction methods

Background correction



Irizarry et al. Biostatistics 2003

Fig. 5. Histograms of $\log_2(MM)$ for a array in which no probe-set was spiked along with the three arrays in which BioB-5 was spiked-in at concentrations of 0.5, 0.75, and 1 pM. The observed PM values for the 20 probes associated with BioB-5 are marked with crosses and the average with an arrow. The black curve represents the log normal distribution obtained from left-of-the-mode data.

RMA Background correction

$PM = B + S$

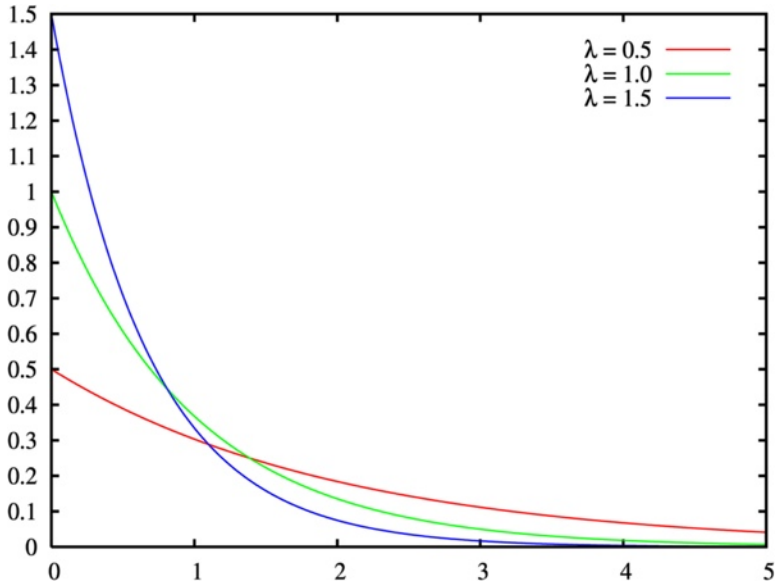
$B \sim$ log-normal with mean and sd read off MM values

$S \sim$ exponential

$\Rightarrow$ closed form expression for $E[S \mid PM]$,

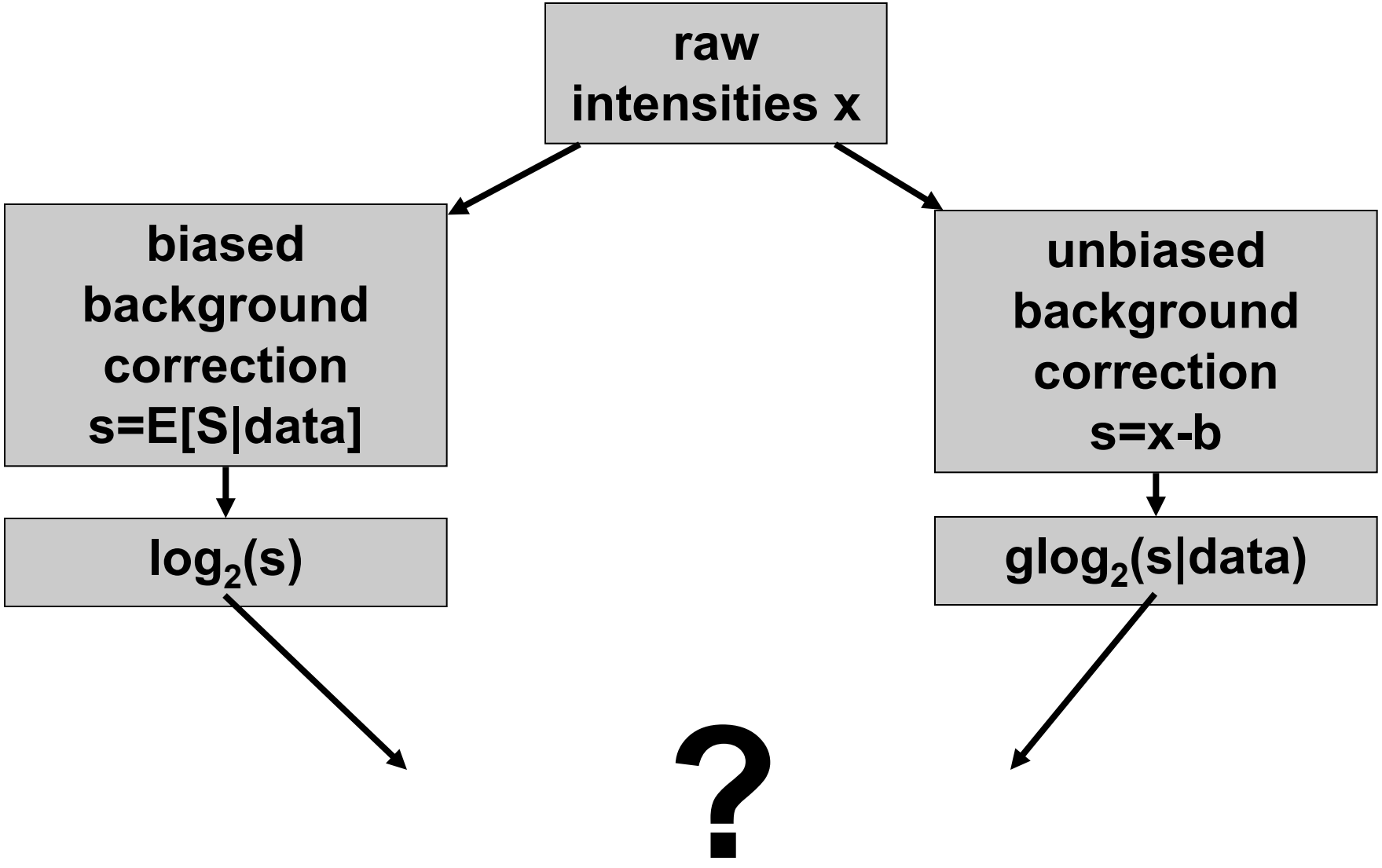
use this as $\hat{s}$ (> 0).

(NB, $P[S > 0] = 1$ is not realistic)

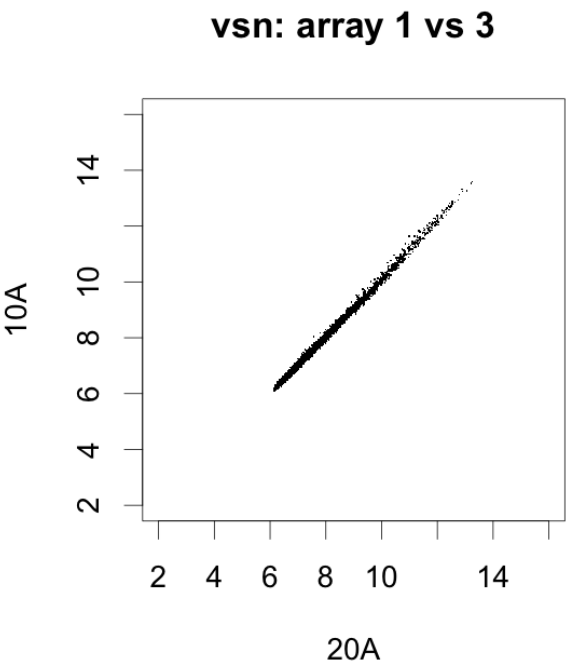


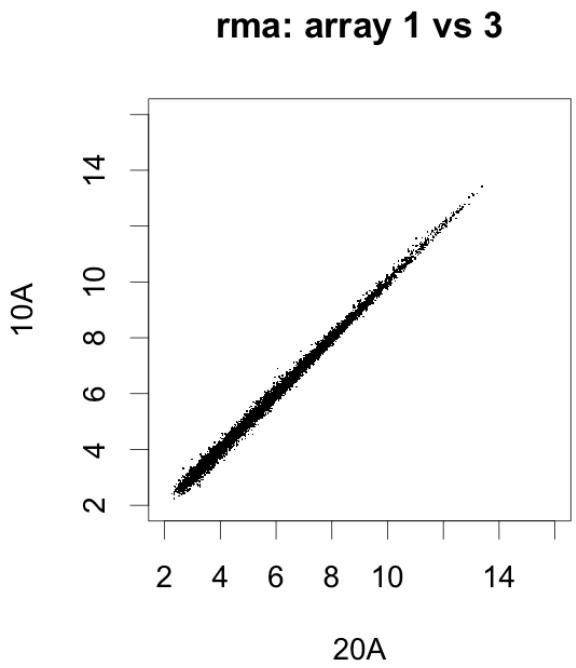
Irizarry et al. (2002)

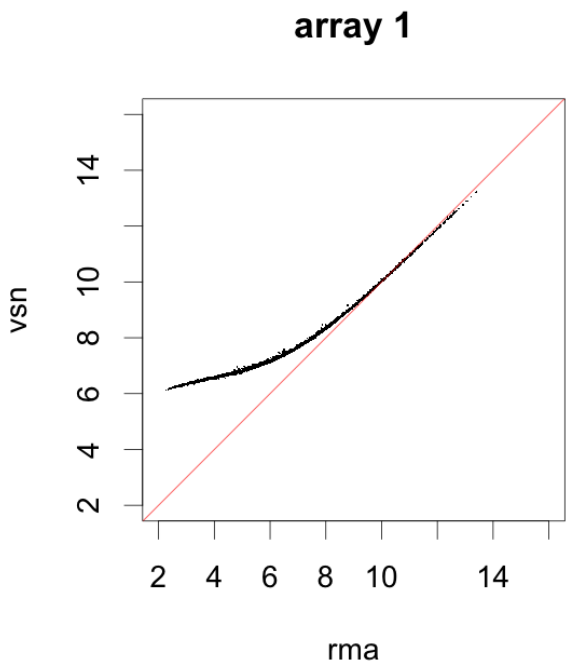
Background correction:



Comparison between RMA and VSN background correction







vsn package vignette