Microarrays: Quality Control, Normalization and Experimental Design

Achim Tresch
Institut für medizinische Biometrie, Epidemiologie und Informatik (IMBEI)
Johannes Gutenberg Universität Mainz

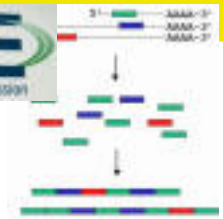
Overview

- Introduction to microarray technologies
- Image Processing:
  Spot Identification, Spot/Background quantification, Quality Measures
- Normalization:
  Scaling, Quantile, Lowess, vsn
- Experimental Design:
  Comparison of typical Designs
- Affy Issues

Different Technologies for Measuring Gene Expression

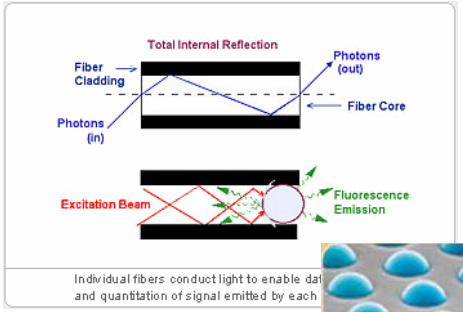
1975: Southern Blotting Technology (Edward Southern)

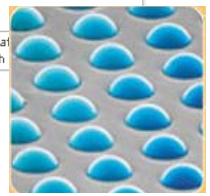
SAGE



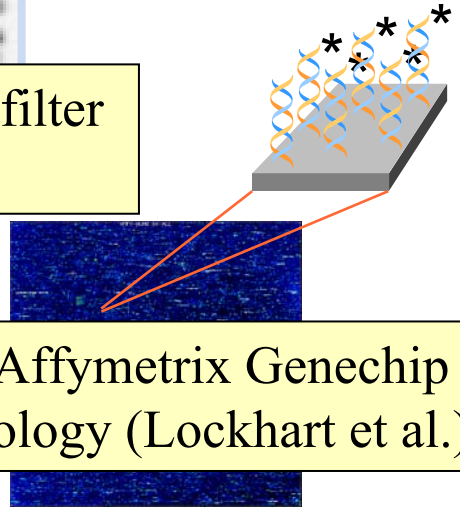
2003: Illumina Bead Arrays

1991: First high-density Nylon filter Arrays (Lennon, Lehrach)



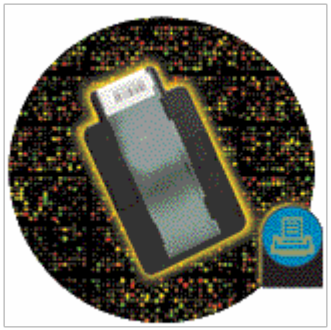


Illumina Bead Array

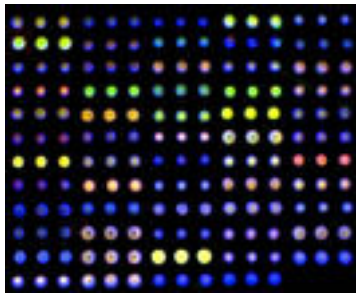


1996: Affymetrix Genechip Technology (Lockhart et al.)

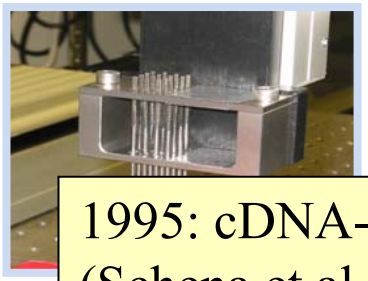
GeneChip Affymetrix



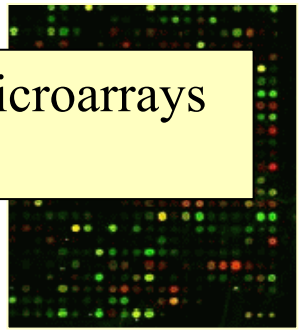
Agilent: Long oligo Ink Jet



CGH



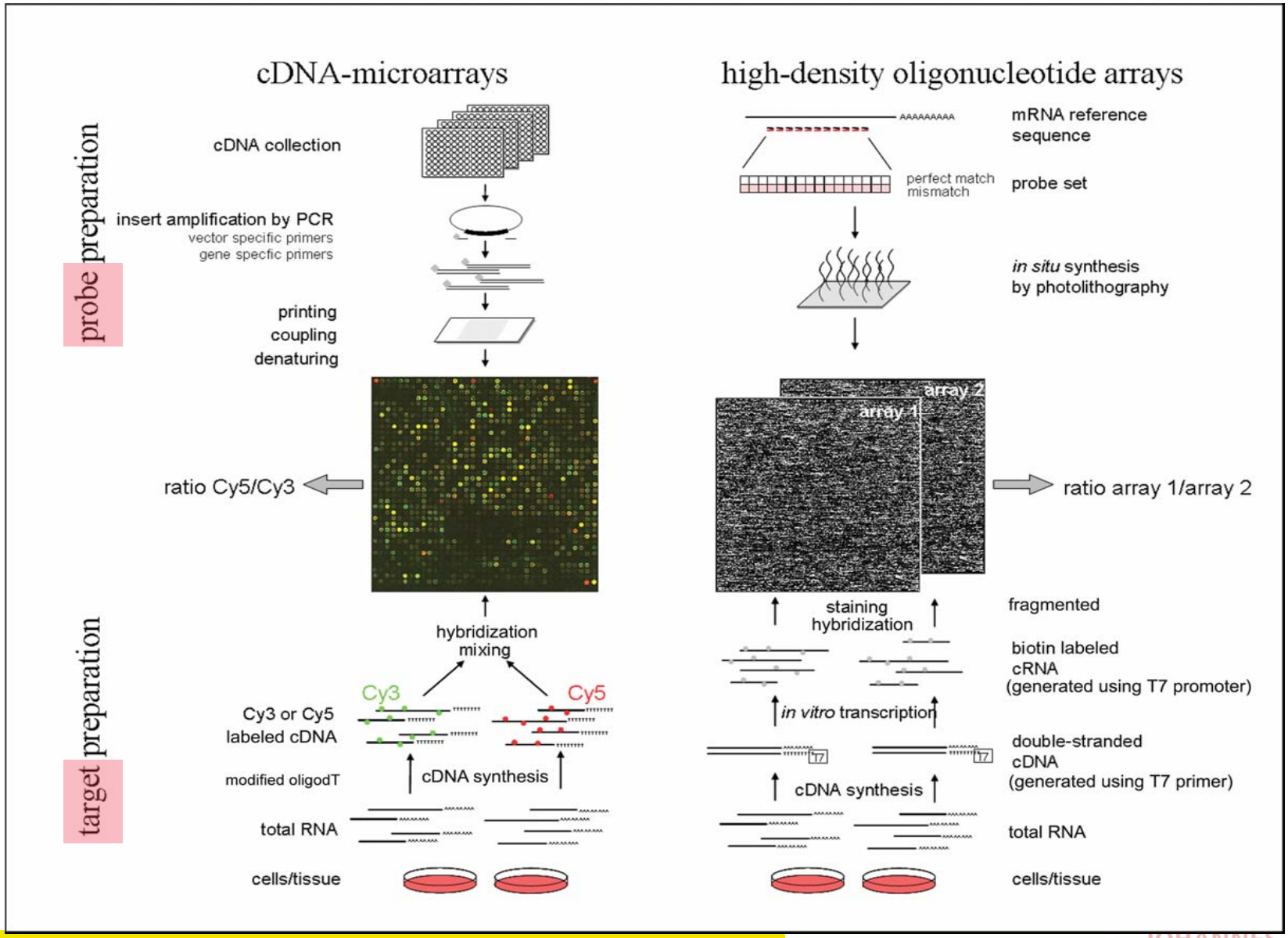
cDNA microarray



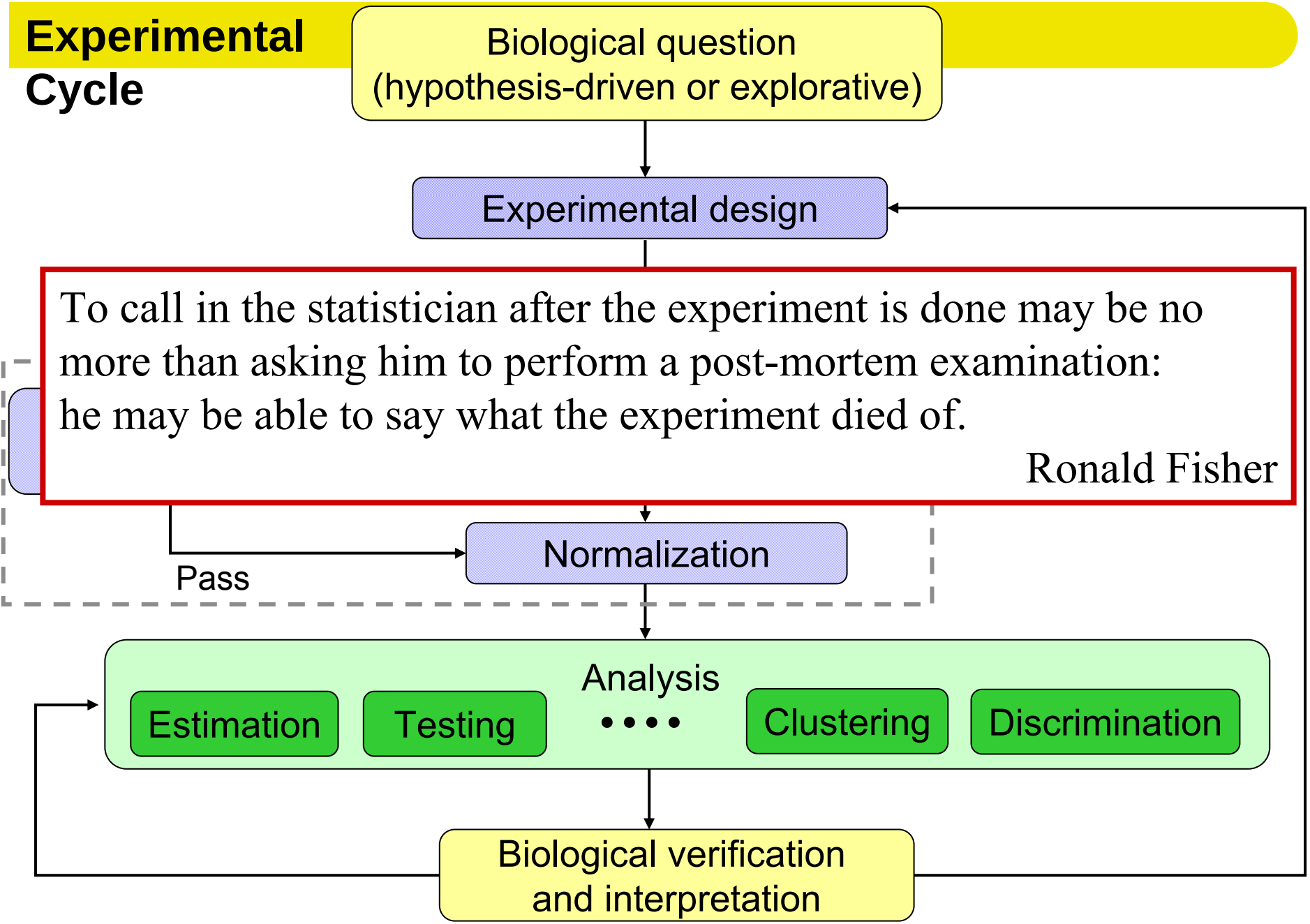
1995: cDNA-Microarrays (Schena et al.)

cDNA and Affymetrix (short, 25 bases) Oligo Technologies.

Long Oligos (60-75 bases) are used similar to cDNA.



Experimental Cycle



Preprocessing result: Gene expression Matrix

Gene expression-Data for **G** Genes and **n** Hybridiyations. Genes times Arrays Data-Matrix:

mRNA Samples

Gene	sample1	sample2	sample3	sample4	sample5	...
1	0.46	0.30	0.80	1.51	0.90	...
2	-0.10	0.49	0.24	0.06	0.46	...
3	0.15	0.74	0.04	0.10	0.20	...
4	-0.45	-1.03	-0.79	-0.56	-0.32	...
5	-0.06	1.06	1.35	1.09	-1.09	...

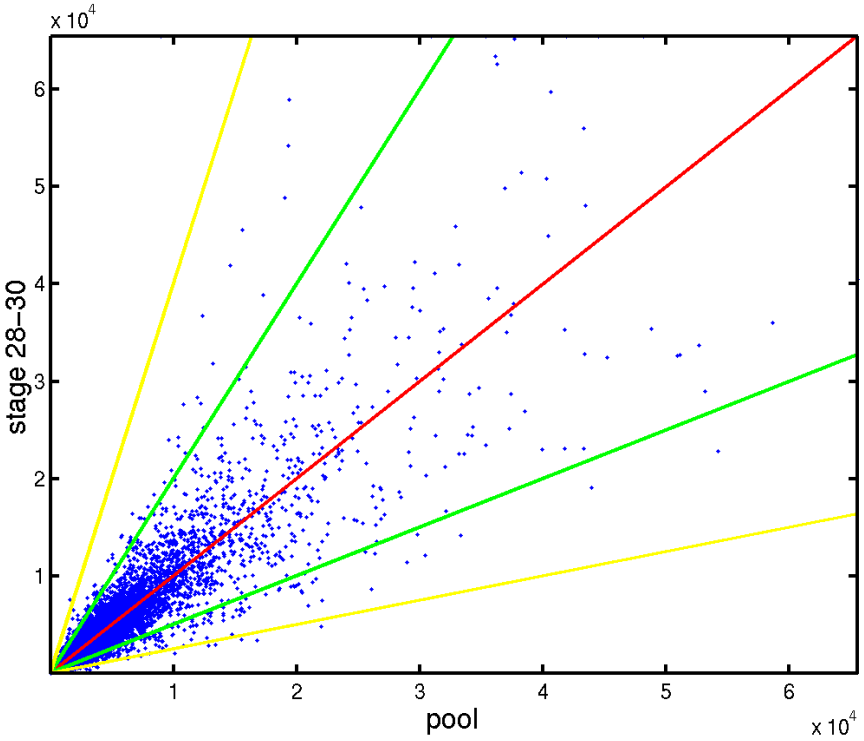
G_{ij}: Gene expression Level for Gen *i* in mRNA sample *j*

Log(red intensity / green intensity)

Function (PM, MM) of MAS, dchip or RMA

Preprocessing result visualization: Scatterplot(s)

Data



Data (log scale)

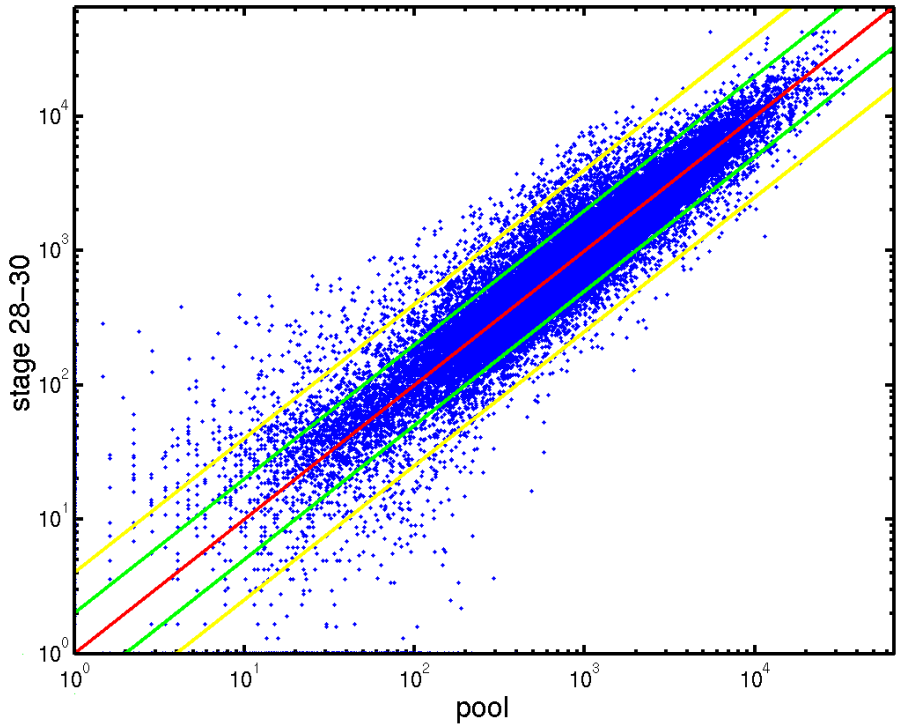


Image Analysis

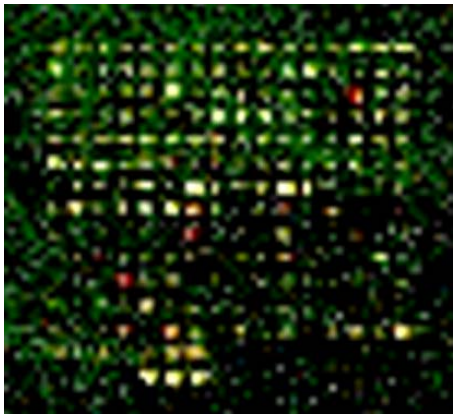
- Spot identification
- Spot quantification
- Probe level quality control
- Gene level quality control
- Array level quality control
- Example

Spot Identification

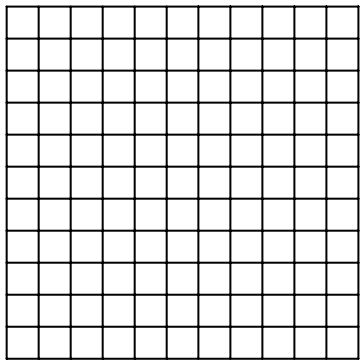
- The grid structure is provided by the manufacturer or generated individually for custom-made microarrays (e.g. GAL-files)
- The grid is overlaid by hand or automatically onto the image (beware of column/row displacement errors!)

Columns

Rows

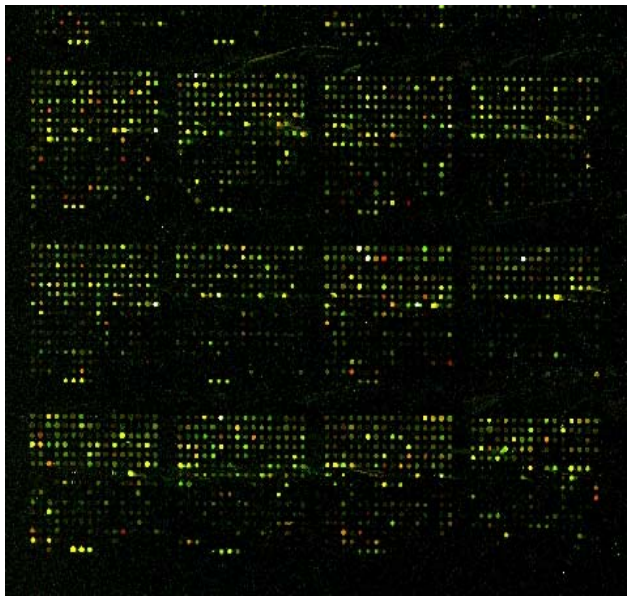


	A	B	C	D	E
1	ATF	1.0			
2	22	5			
3	Type=GenePix ArrayList V1.0				
4	Supplier=Company X				
5	ArrayName=MouseApoptosisProteins 4000				
6	ArrayRevision=2.7				
7	URL=http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Protein&lis				
8	BlockCount=16				
9	Block1= 100, 100, 150, 24, 180, 17, 180				
10	Block2= 4600, 100, 150, 24, 180, 17, 180				
11	Block3= 9100, 100, 150, 24, 180, 17, 180				
12	Block4= 13600, 100, 150, 24, 180, 17, 180				
13	Block5= 100, 4600, 150, 24, 180, 17, 180				
14	Block6= 4600, 4600, 150, 24, 180, 17, 180				
15	Block7= 9100, 4600, 150, 24, 180, 17, 180				
16	Block8= 13600, 4600, 150, 24, 180, 17, 180				
17	Block9= 100, 9100, 150, 24, 180, 17, 180				
18	Block10= 4600, 9100, 150, 24, 180, 17, 180				
19	Block11= 9100, 9100, 150, 24, 180, 17, 180				
20	Block12= 13600, 9100, 150, 24, 180, 17, 180				
21	Block13= 100, 13600, 150, 24, 180, 17, 180				
22	Block14= 4600, 13600, 150, 24, 180, 17, 180				
23	Block15= 9100, 13600, 150, 24, 180, 17, 180				
24	Block16= 13600, 13600, 150, 24, 180, 17, 180				
25	Block	Column	Row	Name	ID
26	1	1	1	MAP-1	11139671
27	1	2	1	bad protein	1083224
28	1	3	1	bcl2-like	6753170
29	1	4	1	interleukin-1	2137456
30	1	5	1	caspase 6	6753286



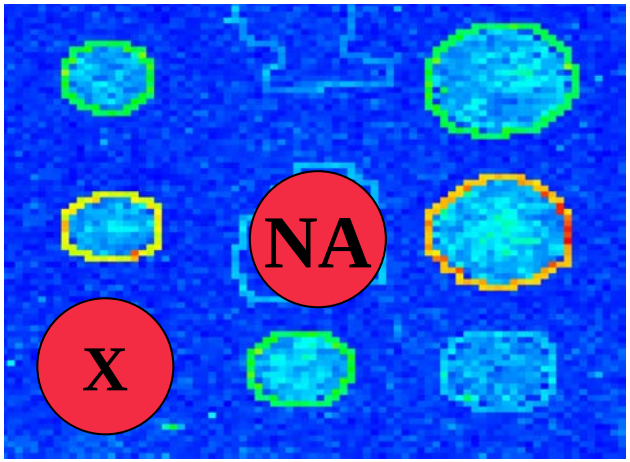
GAL-file contains Clone-IDs and defines their position on the grid

Blocks

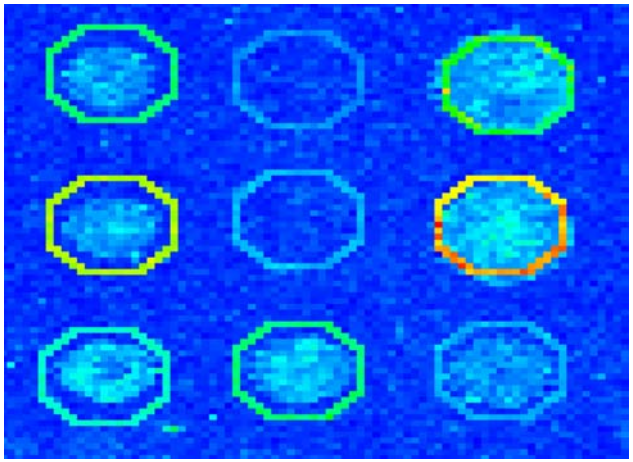


Spot Identification

- Individual spots are recognized, size and shape might be adjusted per spot (automatically fine adjustments by hand).
- Additional manual flagging of bad (X) or non-present (NA) spots



poor spot quality



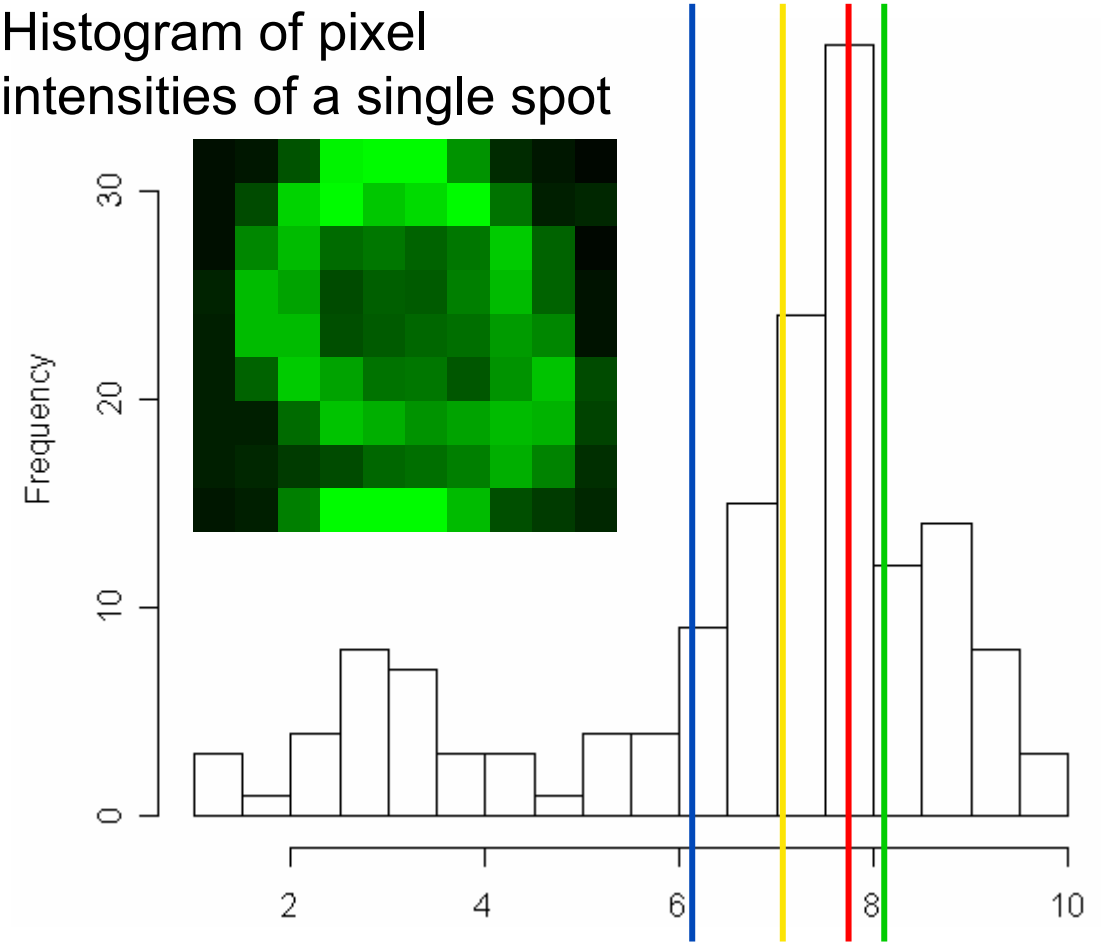
good spot quality

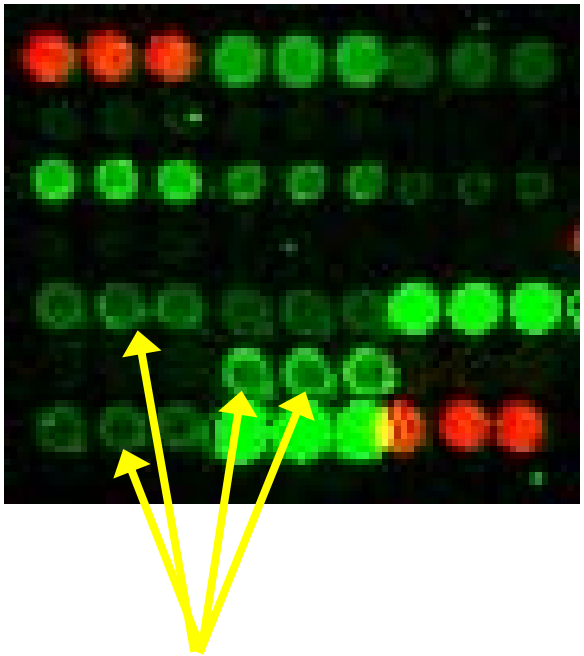
Different Spot identification methods: Fixed circles, circles with variable size, arbitrary spot shape (morphological opening)

Spot identification

- The signal of the spots is quantified.

Histogram of pixel intensities of a single spot





„Donuts“

Mean / Median / Mode / 75% quantile

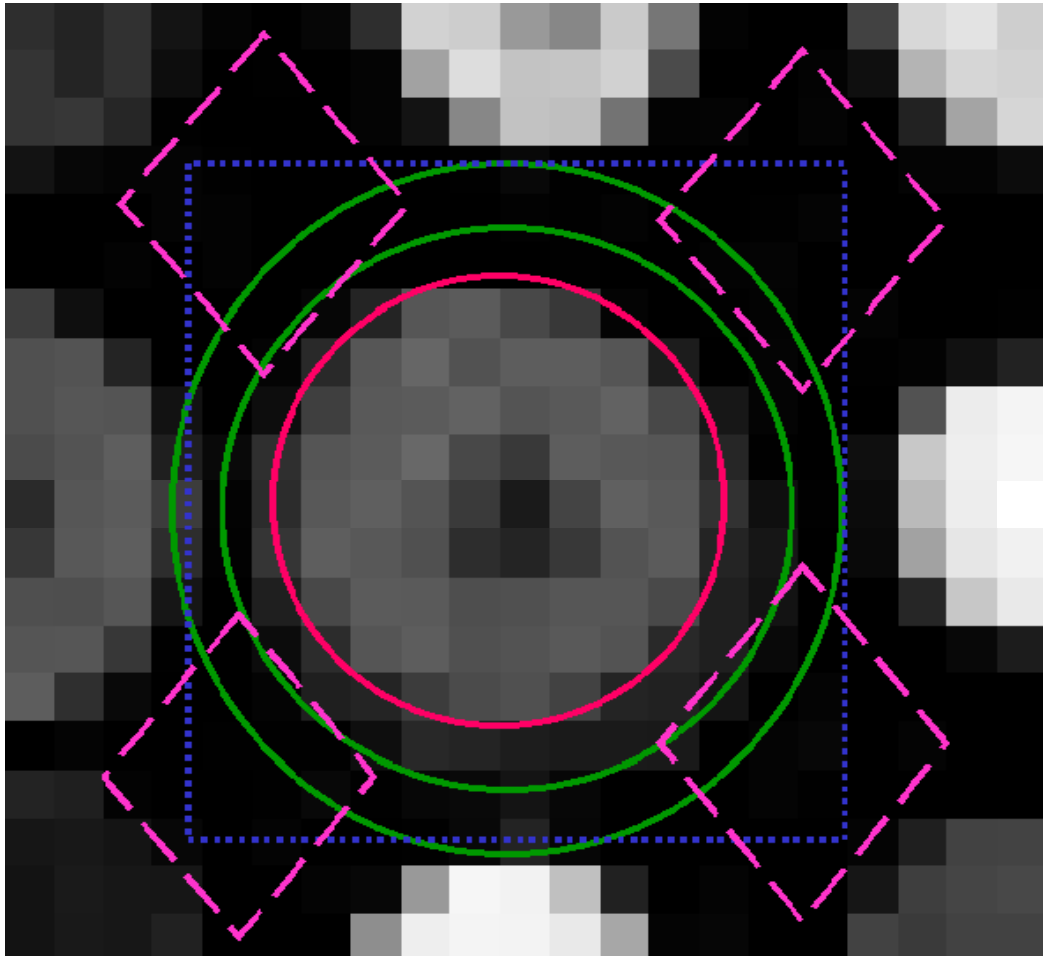
Background correction

- Local background is calculated and subtracted from the spot intensities

GenePix

QuantArray

ScanAlyse



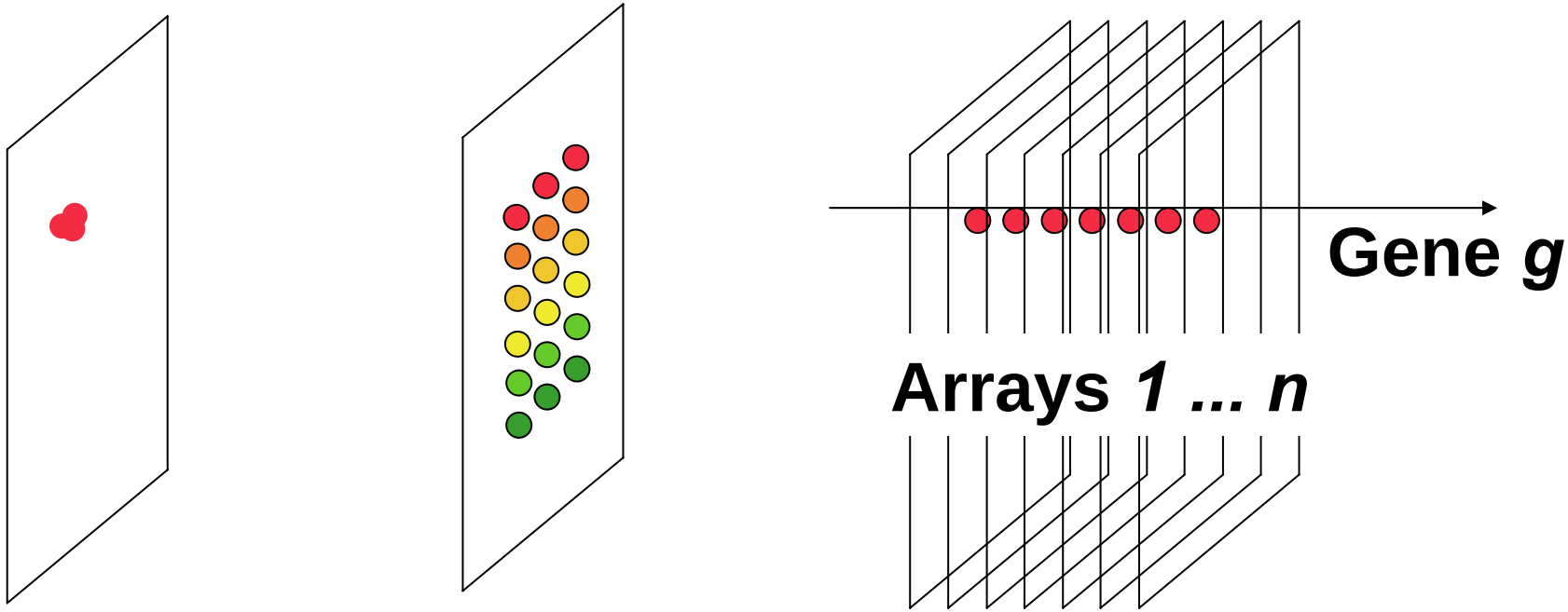
Quality control: Noise and reliable signal

- Is the signal dominated by noise? Acceptable amount of noise?
- Quantify noise (biol./technical variability)
- Quantify quality of a signal
- Guidelines for reasonable thresholds on the quality of a signal
- Defining strategies for exclusion of probes

IMBEI Institut für Medizinische Biometrie, Epidemiologie und Informatik
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Quality control: Noise and reliable signal

Probe level **Array level** **Gene level**



Probe level: quality of the expression measurement of one spot on one particular array

Array level: quality of the expression measurement on one particular glass slide

Gene level: quality of the expression measurement of one probe across all arrays

Probe-level (Individual spots) quality control

- Sources of Variability:
  - faulty printing, uneven distribution of probe material across the spot, contamination with debris
- Visual inspection:
  - hairs, dust, scratches, air bubbles, dark regions, regions with haze
- Spot quality measures:
  - *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
- Action:
  - set measurements to NA (missing values)
  - use weights for measurements to indicate reliability in later analysis.

Gene-level quality control: Poor Hybridization and Printing

- Sources of Variability:
  - Some probes will not hybridize well to the target RNA
  - Printing problems such that all spots of a given print tip will have poor quality.
  - A well may be of bad quality (contamination, wrong RNA)
- Quality measure: Genes with a consistently low signal in the reference channel are suspicious: Median of the background adjusted signal < 200*

  **or other appropriate choice*
- Action: Exclude gene from further analysis

IMBEI Institut für Medizinische Biometrie, Epidemiologie und Informatik
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

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:

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

recommended threshold to assess disagreement: MSDLR > 1

- Disagreement between copies: printing problems, contamination, mislabeling. Not easy if there are only 2 or 3 slides.

Jenssen et al (2002) Nucleic Acid Res, 30: 3235-3244. Theoretical background

Array-level quality control

- Problems:
  - array fabrication defect
  - problem with RNA extraction
  - failed labeling reaction
  - poor hybridization conditions
  - faulty scanner (wrong calibration)

- 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

Swirl Data

- Experiment to study early development in zebrafish.
- Swirl mutant vs. wild-type zebrafish affecting development of dorsal-ventral structures
- Two sets of dye-swap experiments.
- Microarray containing 8448 cDNA probes
- 768 control spots (negative, positive, normalization)
- printed using 4x4 print-tips, each grid contains a 22x24 Spot matrix

R Console

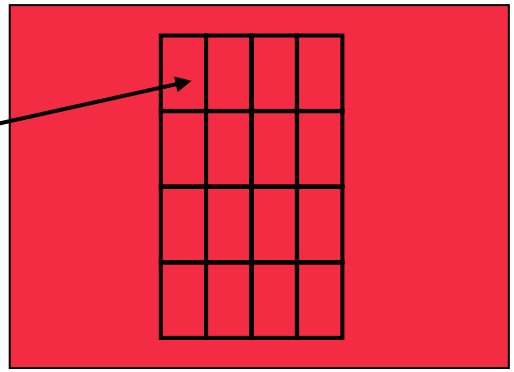
```
> library(marray)
> data(swirl)
> ll()

   member class     mode  dimension
1  swirl  marrayRaw list  c(8448,4)
```

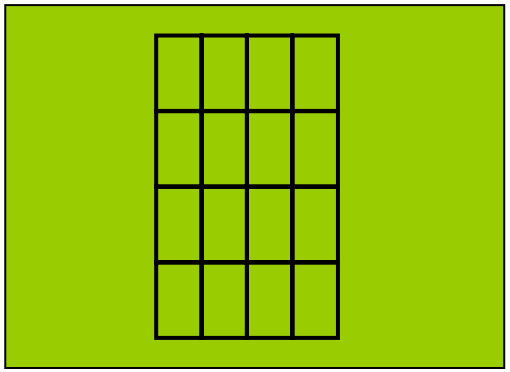
Swirl Data

Mutant Type – R
Wild Type - G

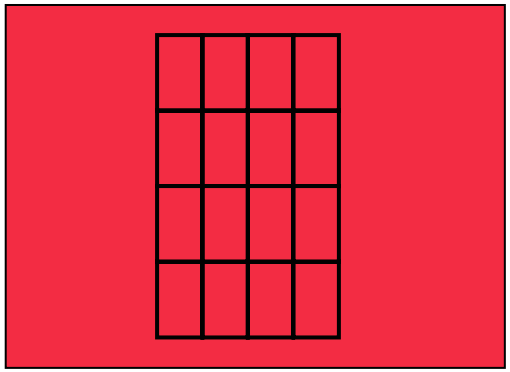
Mutant Type – G
Wild Type - R

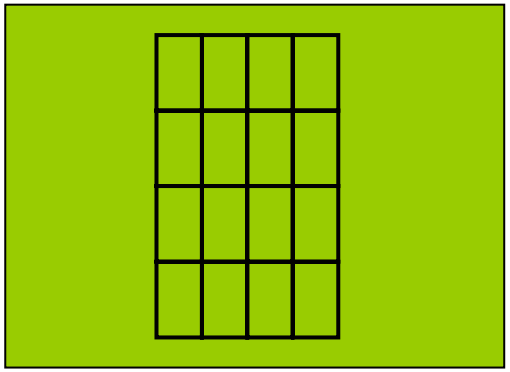
24 x 22 spots per print-tip





Hybr. I





Hybr. II

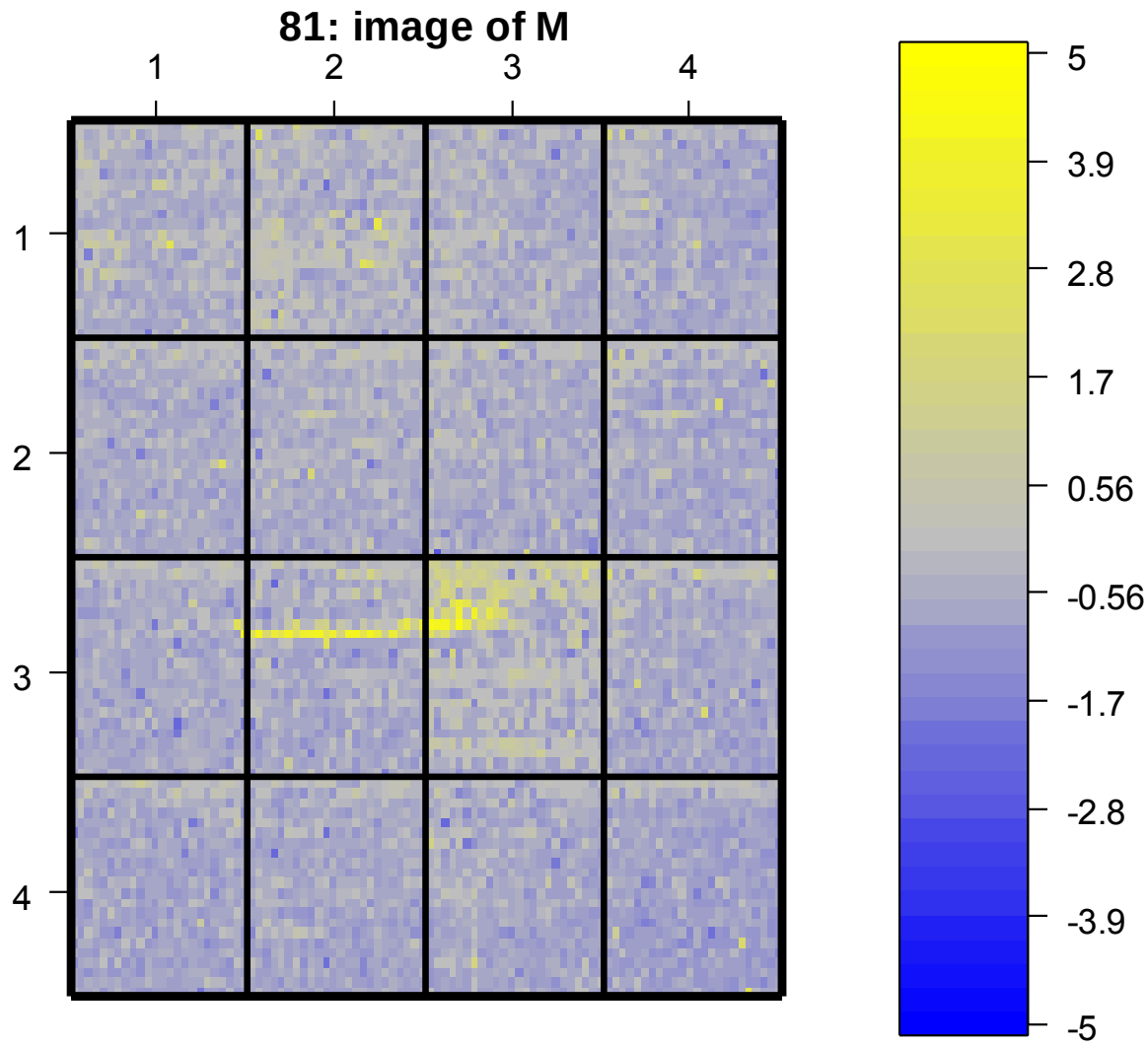
Visual inspection

4 x 4 sectors

Sector:
24 rows
22 columns

8448 spots

Mean signal intensity



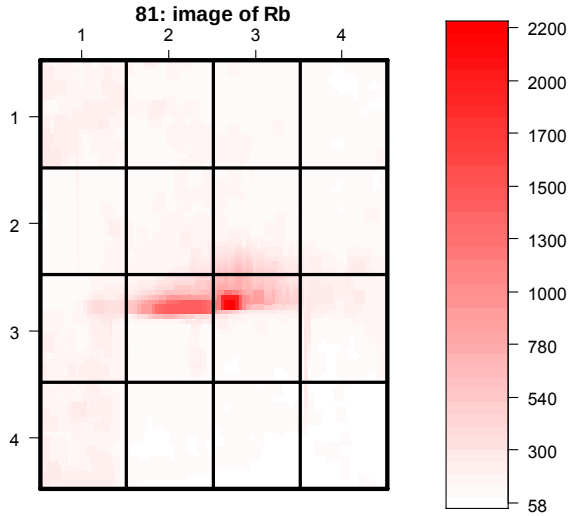
R Console

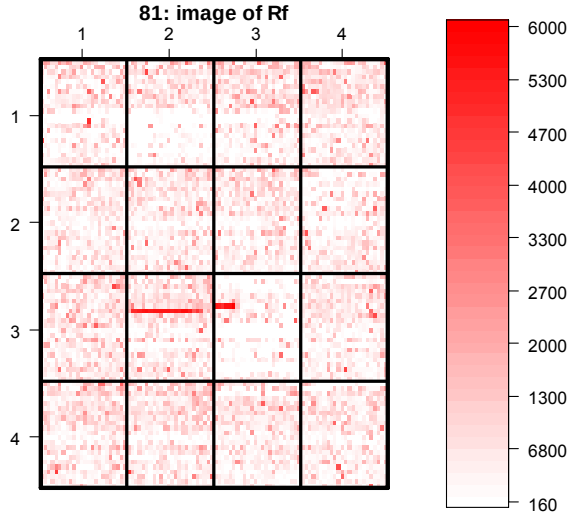
```
> image(swirl[,1])
```

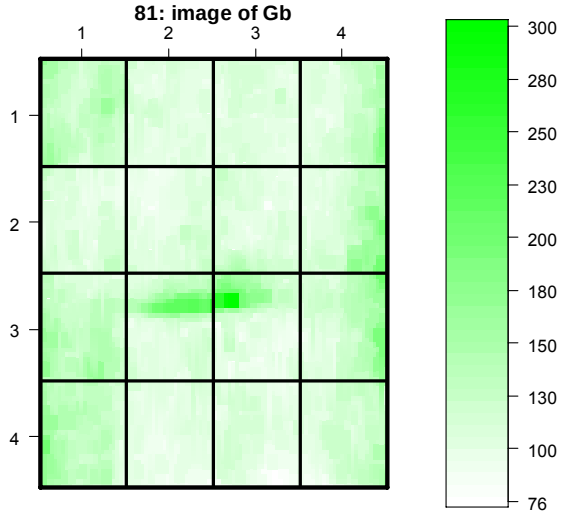
Visual inspection – Foreground and Background intensities

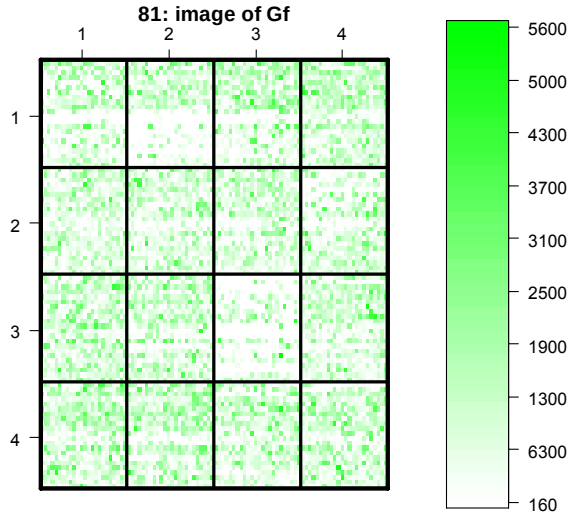
R Console

```
> Gcol <- maPalette(
  low = "white",
  high = "green",
  k = 50)

> Rcol <- maPalette(
  low = "white",
  high = "red",
  k = 50)

> image(swirl[,1]
  xvar="maRb",
  col=Rcol)

> image(swirl[,1]
  xvar="maRf",
  col=Rcol)

> image(swirl[,1]
  xvar="maGb",
  col=Gcol)

> image(swirl[,1]
  xvar="maRf",
  col=Gcol)
```

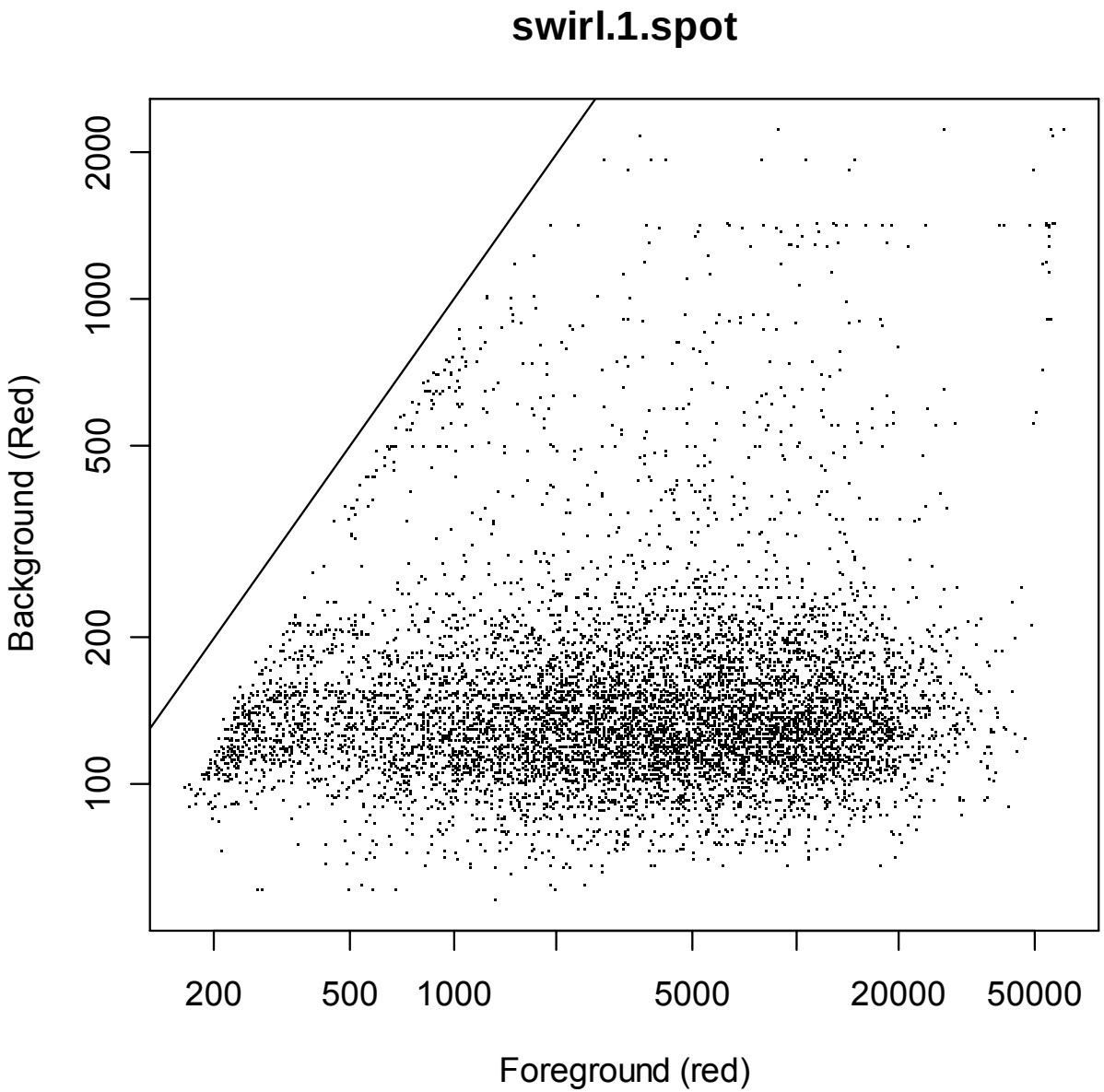








Foreground versus Background intensities



R Console

```
> plot(
  maRf(swirl[,1]),
  maRb(swirl[,1]),
  log="xy")
> abline(0,1)
```

Normalization Methods

- Sources of Variation
- Scaling Methods
- Quantile Normalization
- Lo(w)ess Normalization
- Variance Stabilization

Sources of Variation: Bias and Variance

high noise

low noise

biased

unbiased

Sources of Variation for Microarray-Data

Systematic

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

Stochastic

- Effects on single spots
- random **effects that cannot be estimated**, „noise“

Normalization

Remove bias

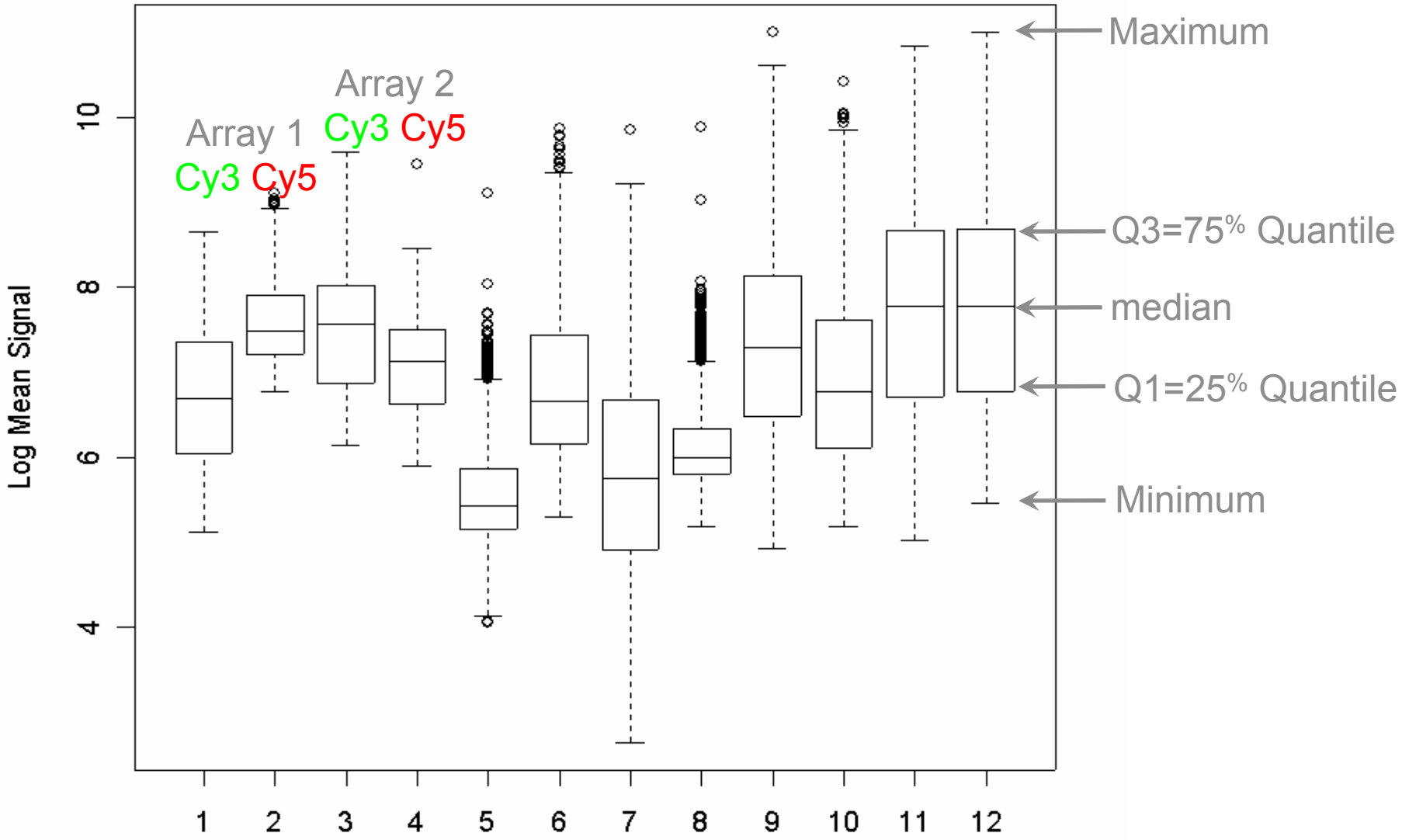
Error model

Quantify variance

Sources of Variation for Microarray-Data

Systematic	Stochastic
amount of RNA in biopsy	
DNA quality	tissue contamination
efficiency of: RNA extraction, reverse transcription, labeling, photodetection	
stray-/background signal	RNA degradation
amplification efficiency	spotting efficiency
hybridization efficiency	and specificity
spot size	
reverse transcription efficiency	DNA support binding

Displaying Variability of Microarray-Data



Aims of normalization:

- Identify and remove sources of systematic variation, other than differential expression, in the measured fluorescence intensities.

Enable the estimation of

- **True fold changes**
- **Significance of differential expression**

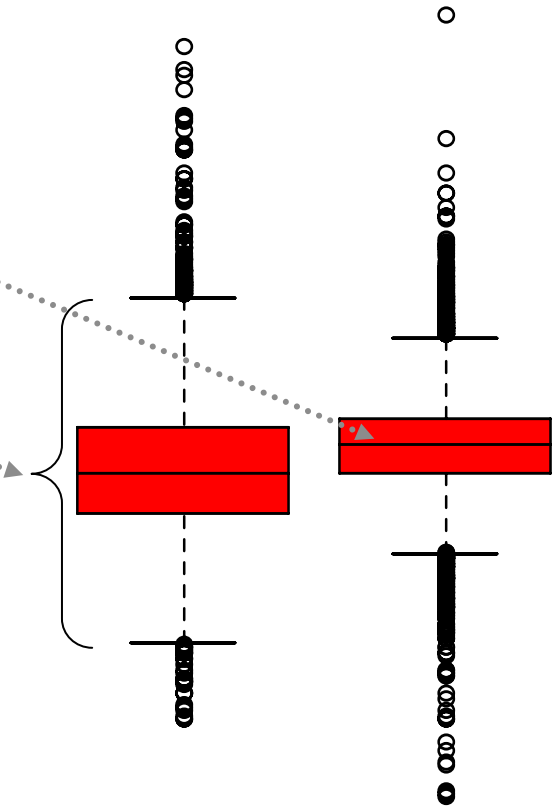
These aims can be adverse! Depending on the further analysis steps, different normalization strategies may be appropriate!

Normalization via rescaling

- Location and scale are basic statistical concepts for data description:

Location
normalization: corrects for spatial or dye bias

Scale
normalization: homogenizes the variability across arrays



Normalized log-intensity ratios are given by

$$M_{norm} = (M\text{-location}) / \text{scale}$$

"Location and scale of different MA measurements should be (approximately) the same."

boxplots

Normalization via rescaling

- **Location:** Robust estimation of a “rescaling” Factor, e.g. based on the median of gene expression values on the chip. The underlying assumption is that the majority of genes and hence the center of the expression values should not change between different measurements. The median is used as a robust measure for the center of a dataset.
- **Scale:** Use some measure for the variability of the data, e.g.

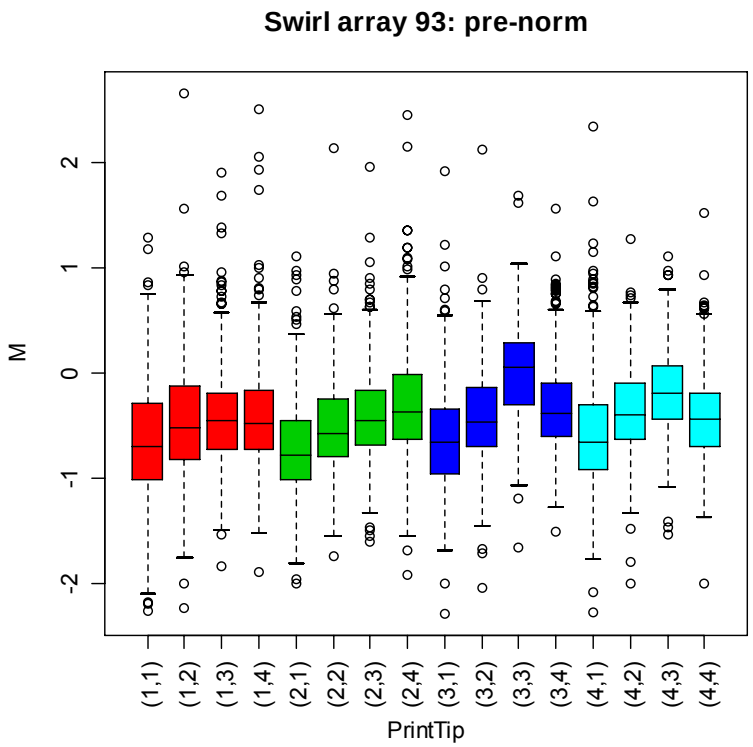
$$\mathbf{MAD} = \mathbf{MedianAbsoluteDifference}$$
$$= \mathit{median}\{\, |x_1 - \mathit{median}|, \ldots, |x_n - \mathit{median}| \,\}$$

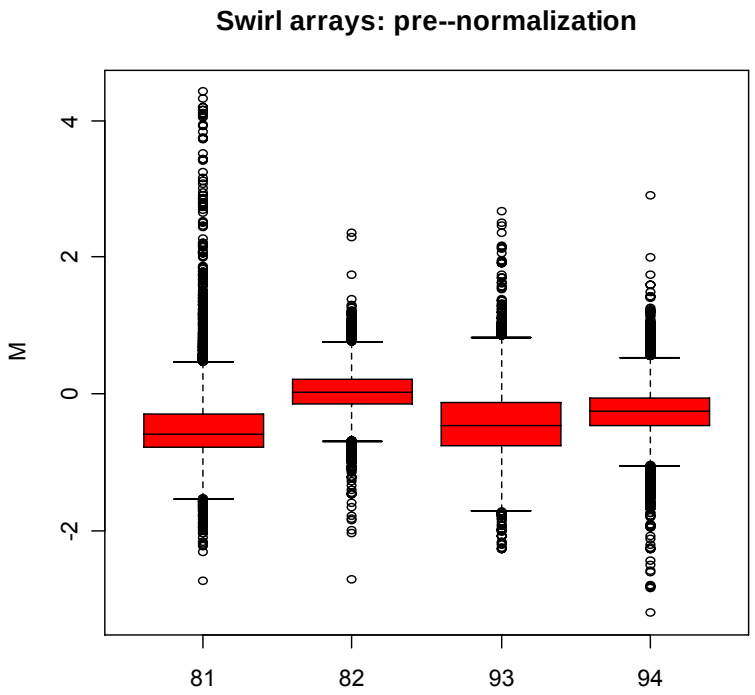
(the MAD is a more robust measure of scale than the variance)

→ **Median centering**: Subtract the median of all expression values of one chip and divide by the MAD.

- ~~Housekeeping genes~~
- ~~*Spiked* in control genes~~

marray – Swirl Data: Raw data

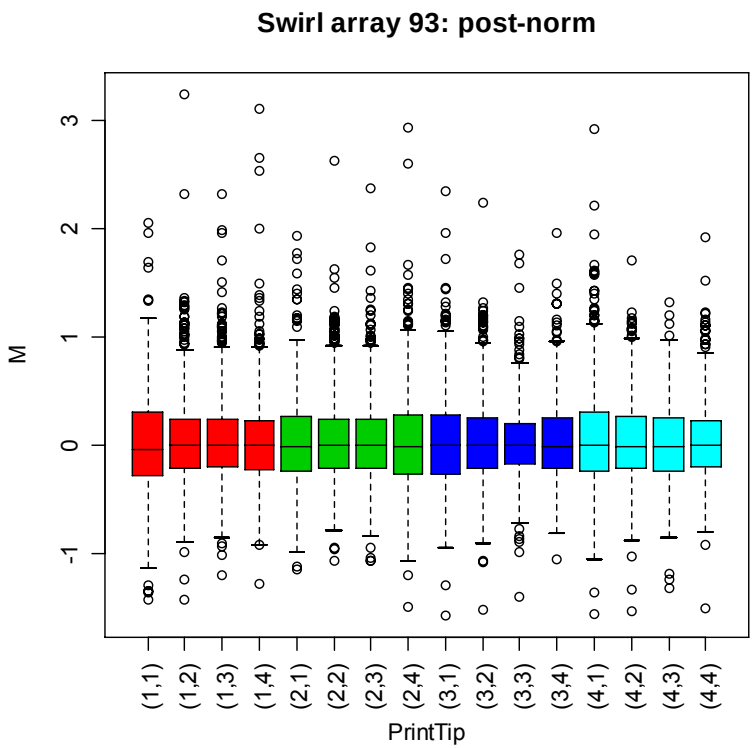


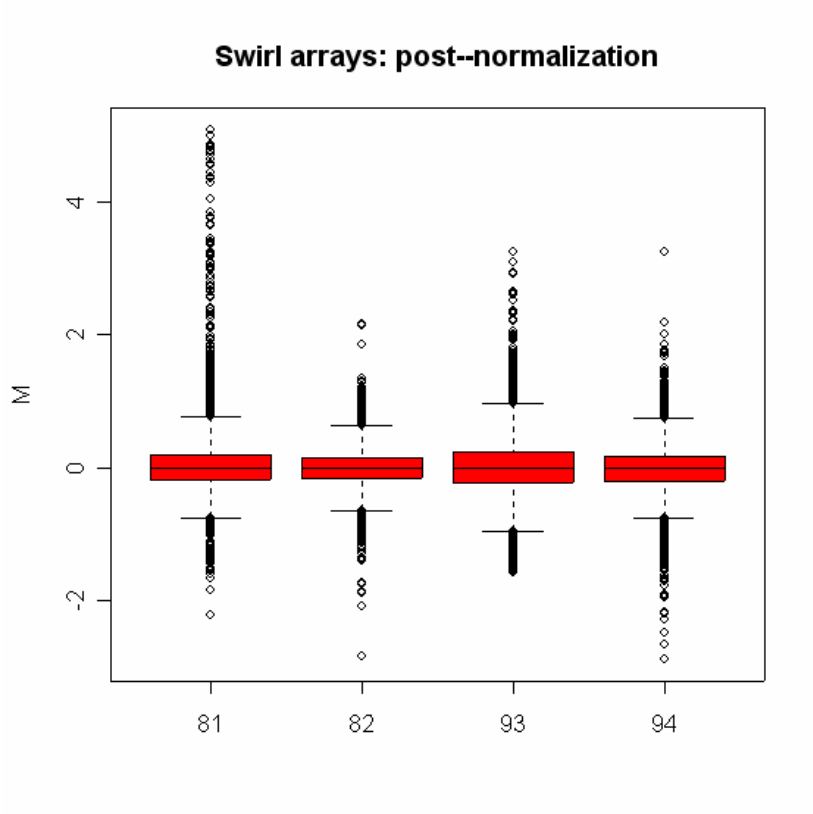


R Console

```
> boxplot(swirl[, 3], xvar = "maPrintTip", yvar = "maM")
> boxplot(swirl, yvar = "maM")
```

marray – Swirl Data: Post Normalization

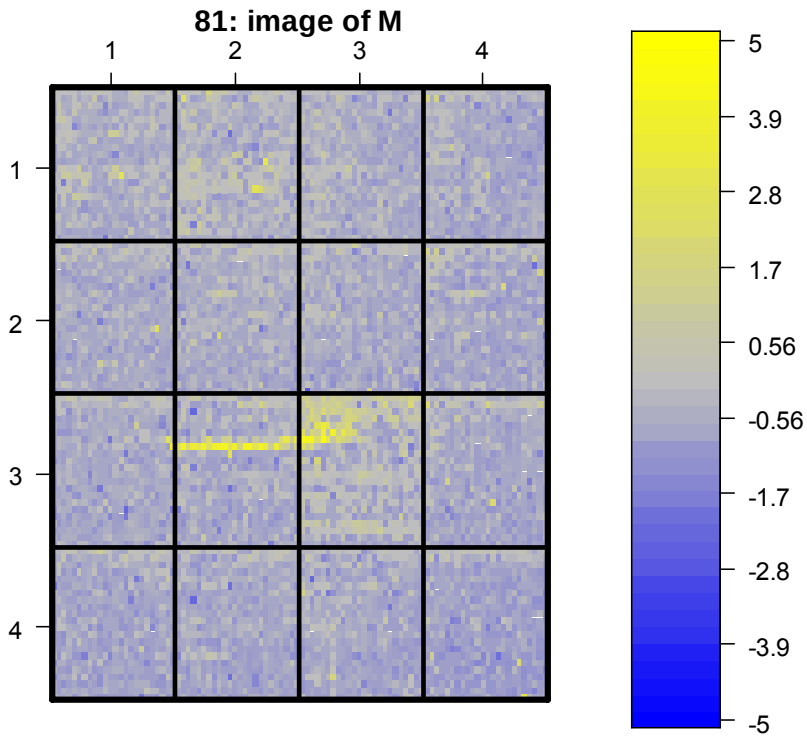


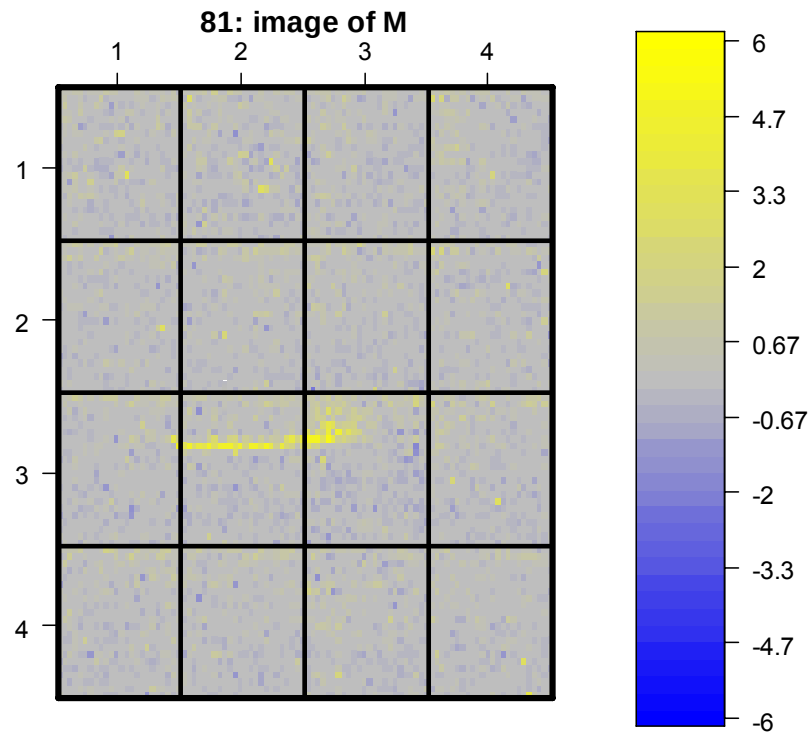


R Console

```
> swirl.norm <- maNorm(swirl, norm = "p")
> boxplot(swirl.norm[, 3], xvar = "maPrintTip", yvar = "maM")
> boxplot(swirl.norm, yvar = "maM")
```

Swirl Data – M values, raw vs preprocessed and rescaled





Normalization procedure was not able to remove scratch

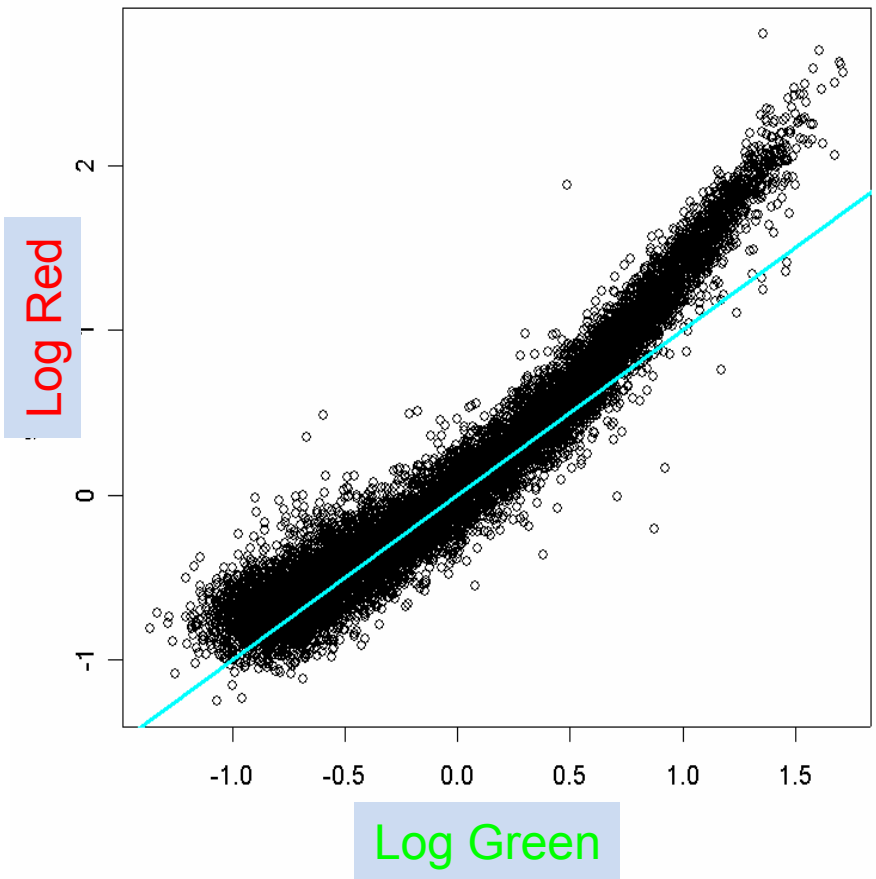
R Console

```
> image(swirl[,1])
> image(swirl.norm[,1])
```

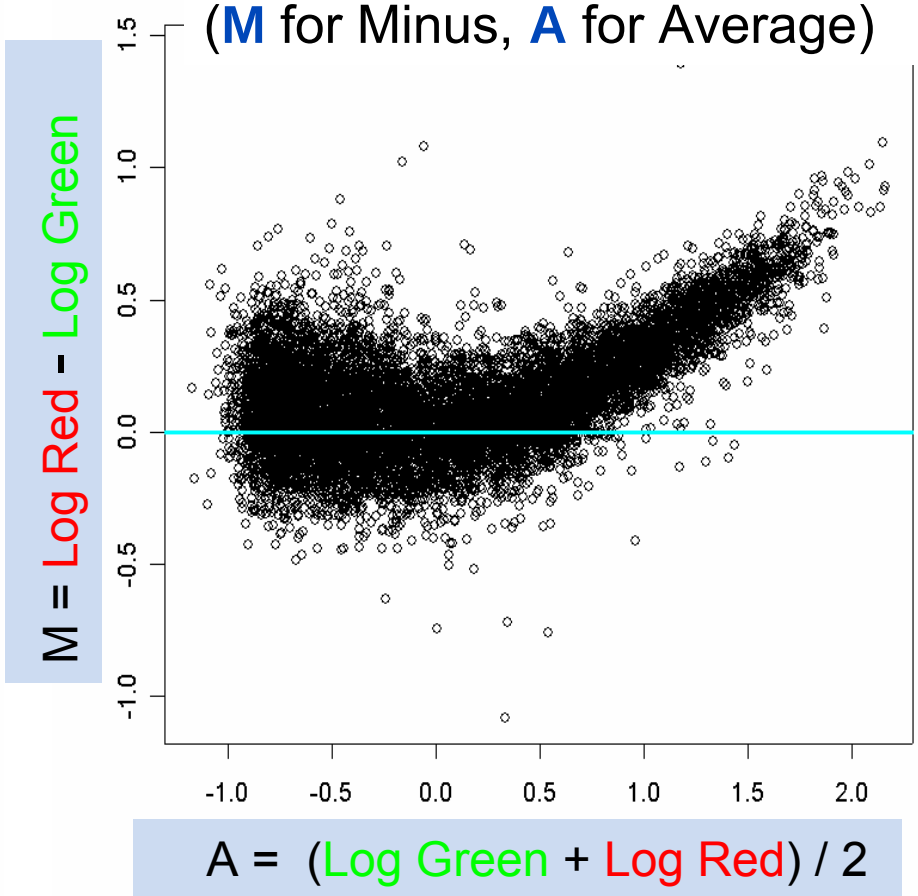
Problems with Median-Centering

Median-Centering is a **global Method**. It does not adjust for local effects, intensity dependent effects, print-tip effects, etc.

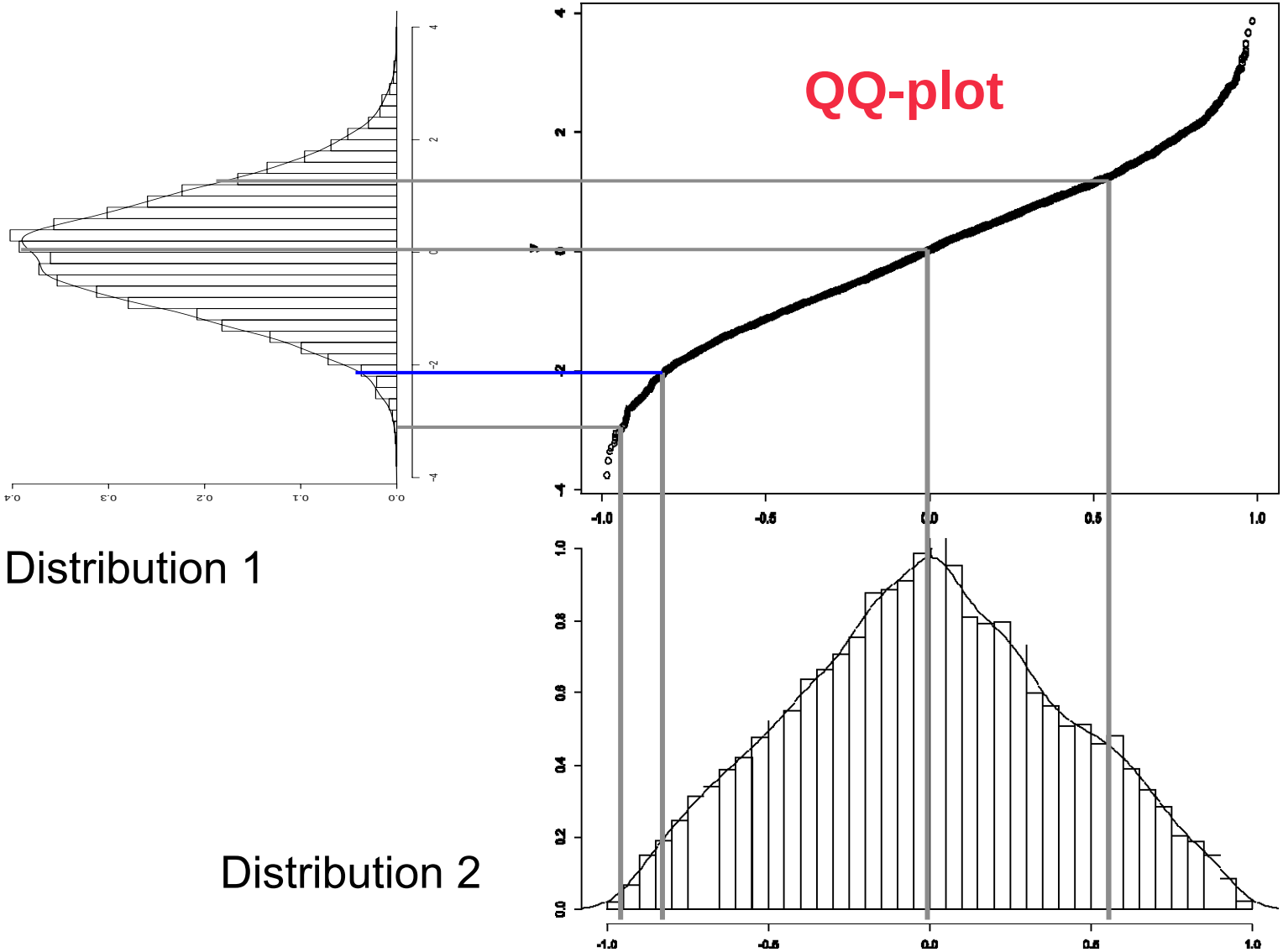
Scatterplot of log-Signals after Median-centering



M-A Plot of the same data (**M** for Minus, **A** for Average)



Quantile Normalization



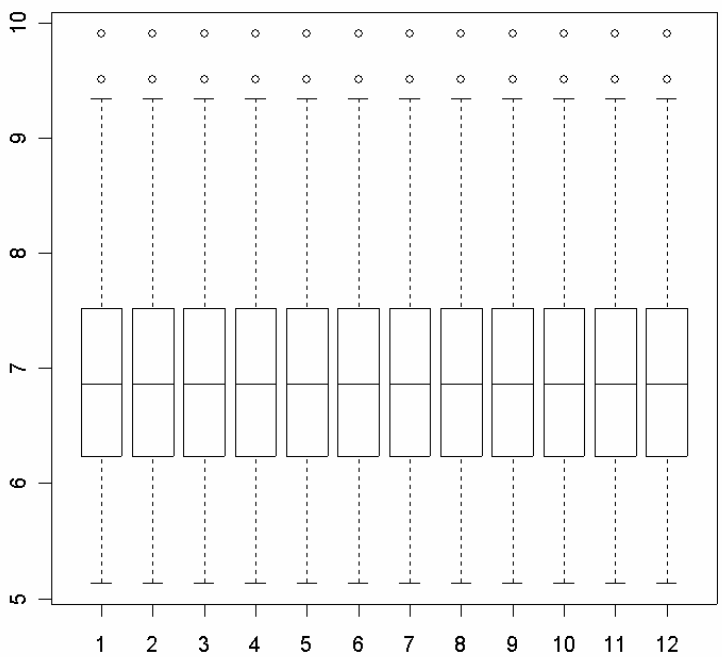
Quantile Normalization

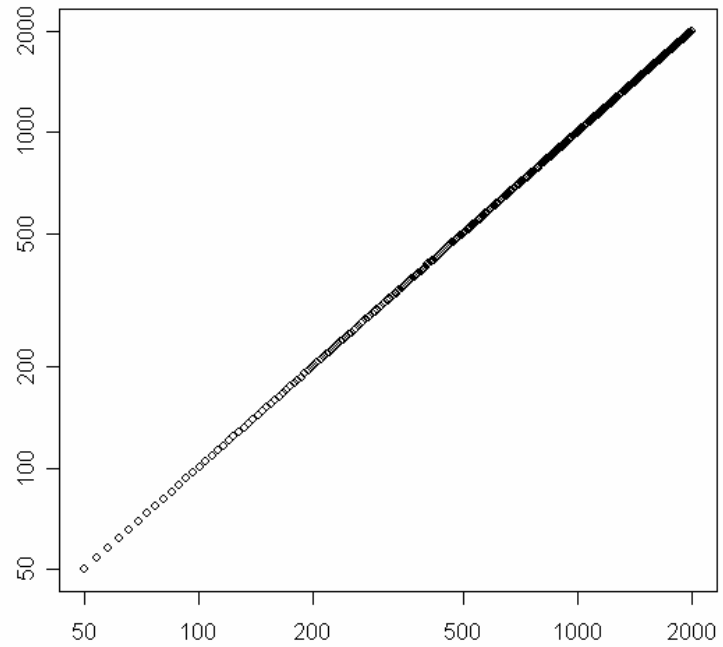
The basic idea of Quantile-Normalization is very simple:

„The Histograms of all Slides are made identical“

Tightens the idea of Median-Centering. Not only the 50%-Quantile is adjusted, but *all* Quantiles.

Boxplot and QQ-plot after Quantile normalization





Quantile Normalization

The Algorithm:

- For each array, sort the genes by expression
- Let M_n be the mean value of the nth genes of each array. Replace the values for the nth gene by M_n in each array.
- Do this for all positions n.

Disadvantage: For genes at the extreme ends of the distribution, the expression values of the nth genes have a high variance, so the mean may vary strongly. In general, quantile normalization tends to underestimate expression values at the high end and vice versa at the low end.

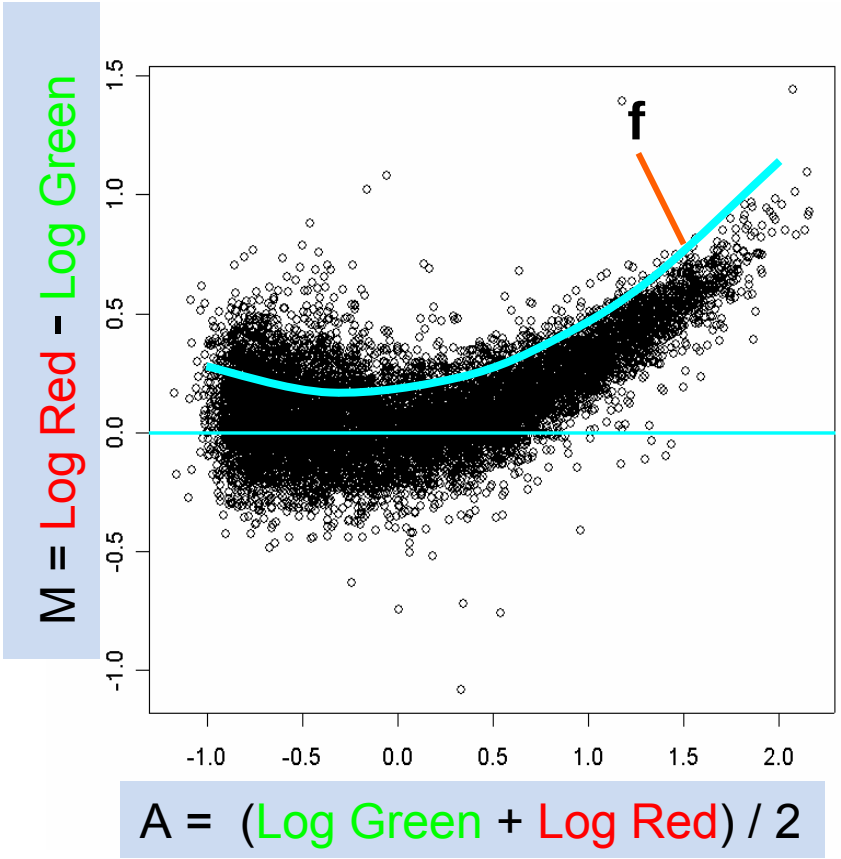
Before using quantile normalization, measurement data for each chip should be on the same scale!

Lo(w)ess Normalization

Assumption: There is an intensity-dependent bias of the fold change,

$$M = f(A)$$

and hence $y_j = f(x_j) + \delta_j$
where δ_j is the "true" log fold change for gene j. The true fold change distribution is approximately a zero-symmetric normal distribution.



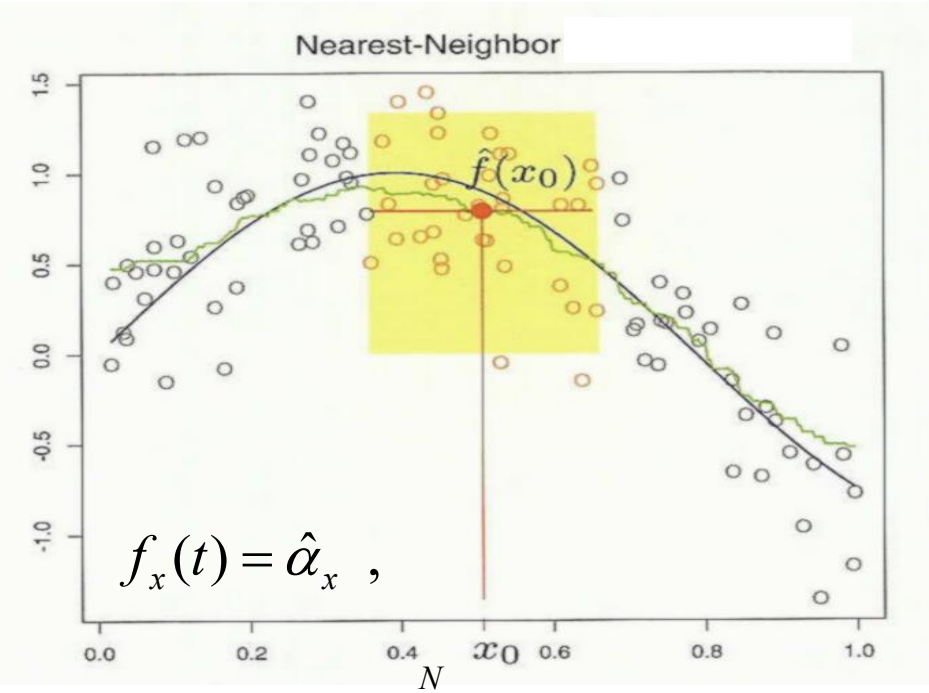
Task: Find f, replace y_j by $y_j - f(x_j)$.

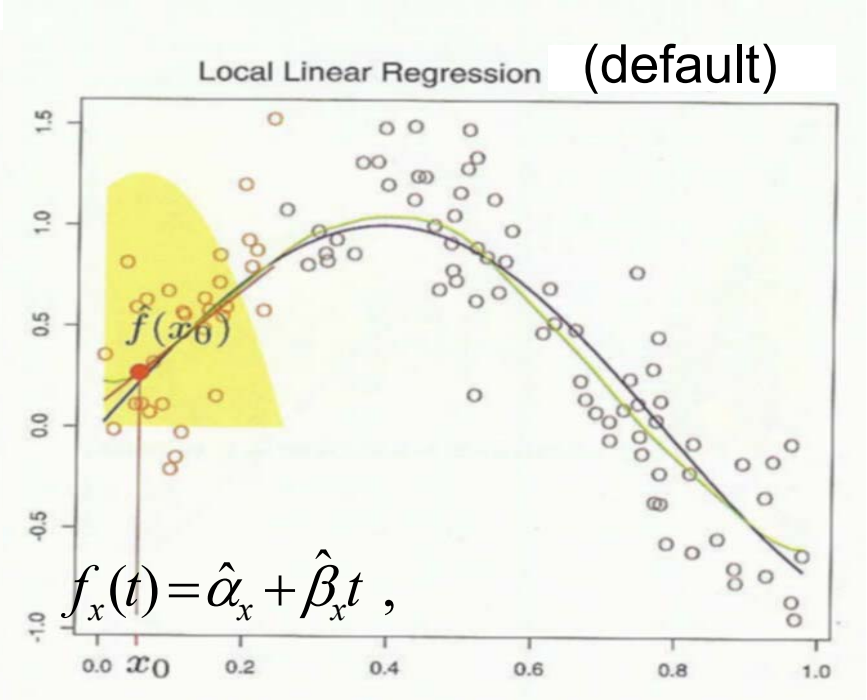
The **idea of local regression** is that f can be estimated locally at a point x by a simple (and easy-to-fit) function f_x. For each point x, we then estimate f by

$$\hat{f}(x) = f_x(x)$$

Lo(w)ess Normalization

In practice, f_x is a polynomial of low order (≤ 2). Which points (and with which weights) are used to estimate f_x is determined by a kernel weight function K.





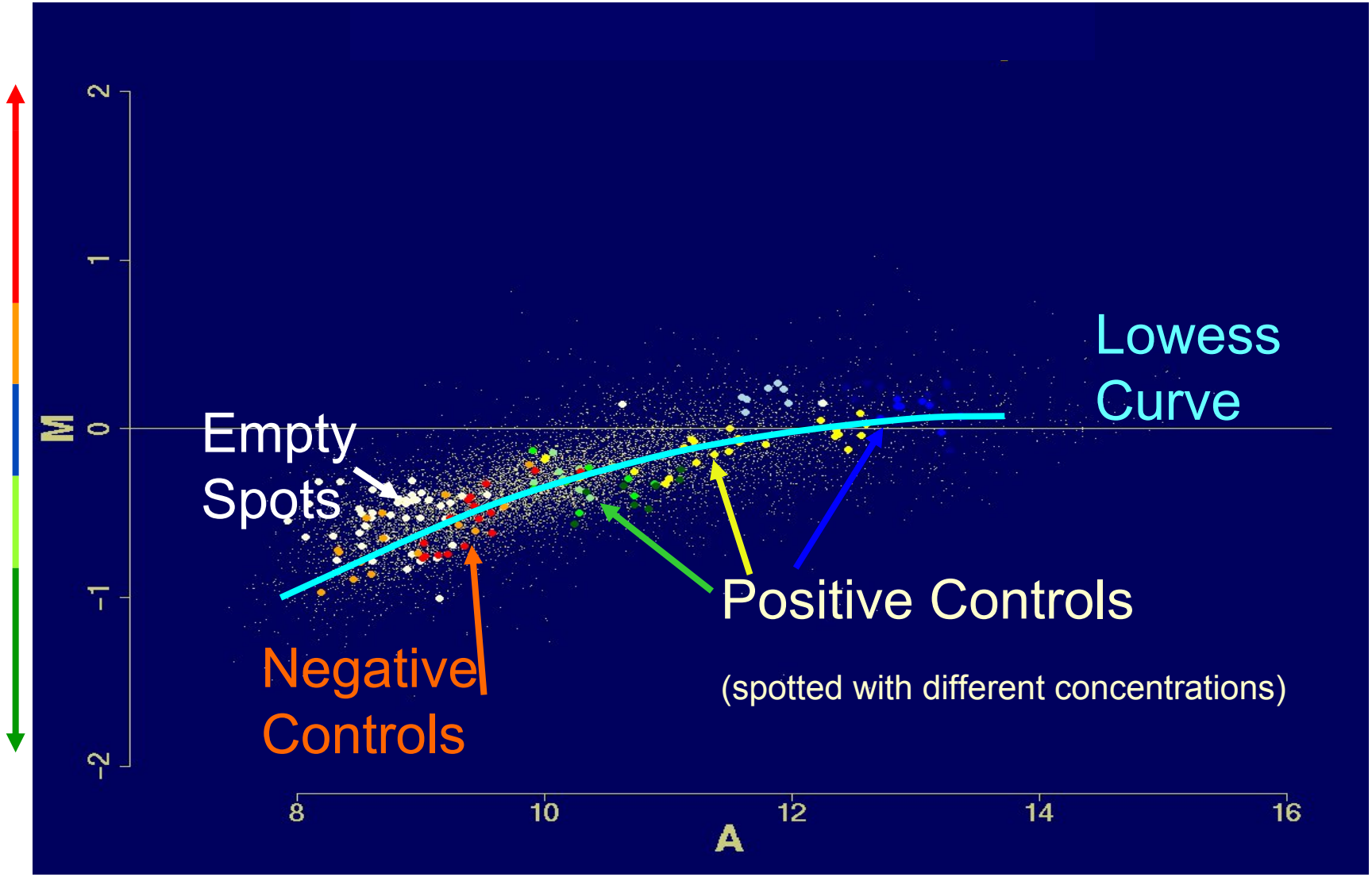
$$\hat{\alpha}_x = \underset{\alpha_x}{\operatorname{argmin}} \sum_{j=1}^{N} K(x, x_j)[y_j - \alpha_x]^2$$

$$(\hat{\alpha}_x, \hat{\beta}_x) = \underset{(\alpha_x, \beta_x)}{\operatorname{argmin}} \sum_{j=1}^{N} K(x, x_j)[y_j - (\alpha_x + \beta_x x_j)]^2$$

lowess = **LO**cally **W**eighted regr**ESS**ion

Taken from Tibshirani et al., „Elements of Statistical Learning“

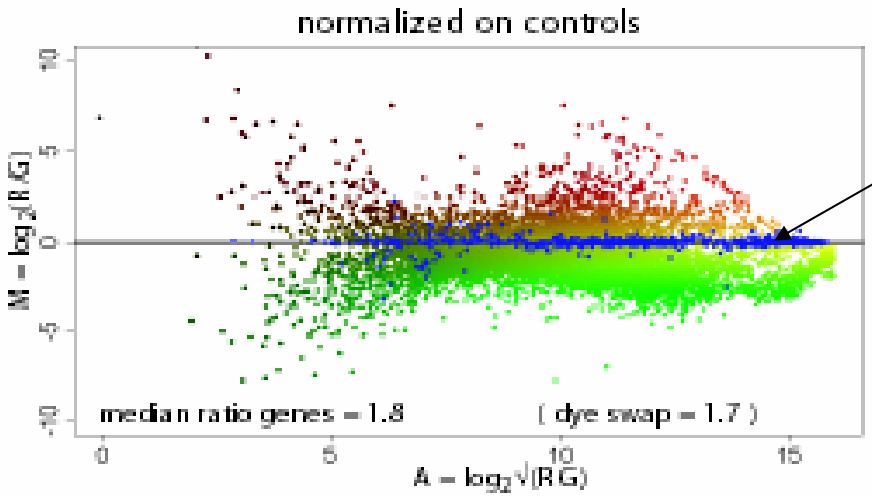
Lo(w)ess Normalization on all Genes vs. Spike-ins



$M = \log R/G = \log R - \log G$ $\quad$ $A = (\log R + \log G)/2$

Lo(w)ess Normalization on all Genes vs. Spike-ins

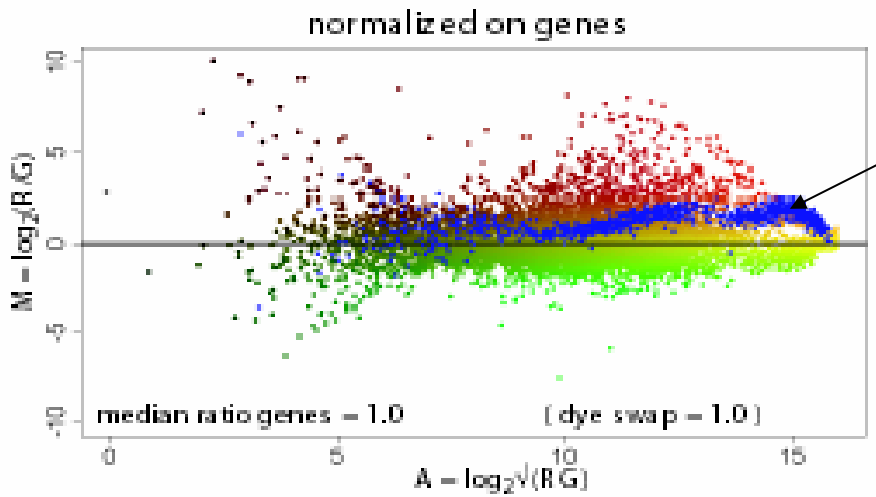
A.



External Controls

Lowess Regression fitted to spike-ins

B.

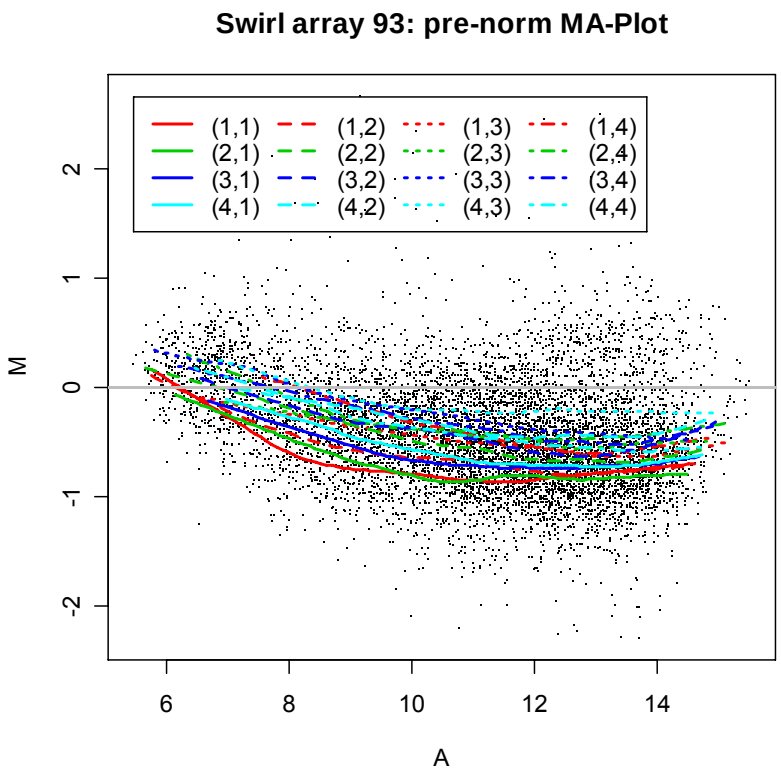


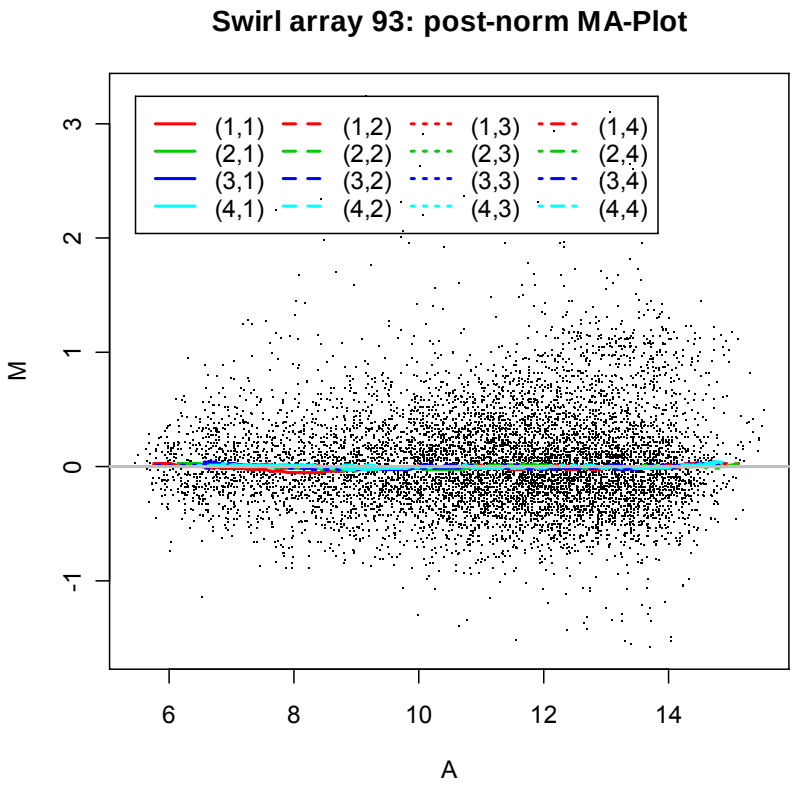
External Controls

Lowess Regression fitted to all genes

From Van de Peppel et al, 2003

marray – Swirl Data: Print-tip lowess Normalization





R Console

```
> plot(swirl[, 3], xvar = "maA", yvar = "maM",
  zvar = "maPrintTip")
> plot(swirl.norm[, 3], xvar = "maA", yvar = "maM",
  zvar = "maPrintTip")
```

Non-parametric smoother: loess, lowess, local regression line, generalizes the concept of moving average.

Variance Stabilizing Normalization (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 least trimmed sum of squares regression.
- vsn–normalized data behaves close to the normal distribution

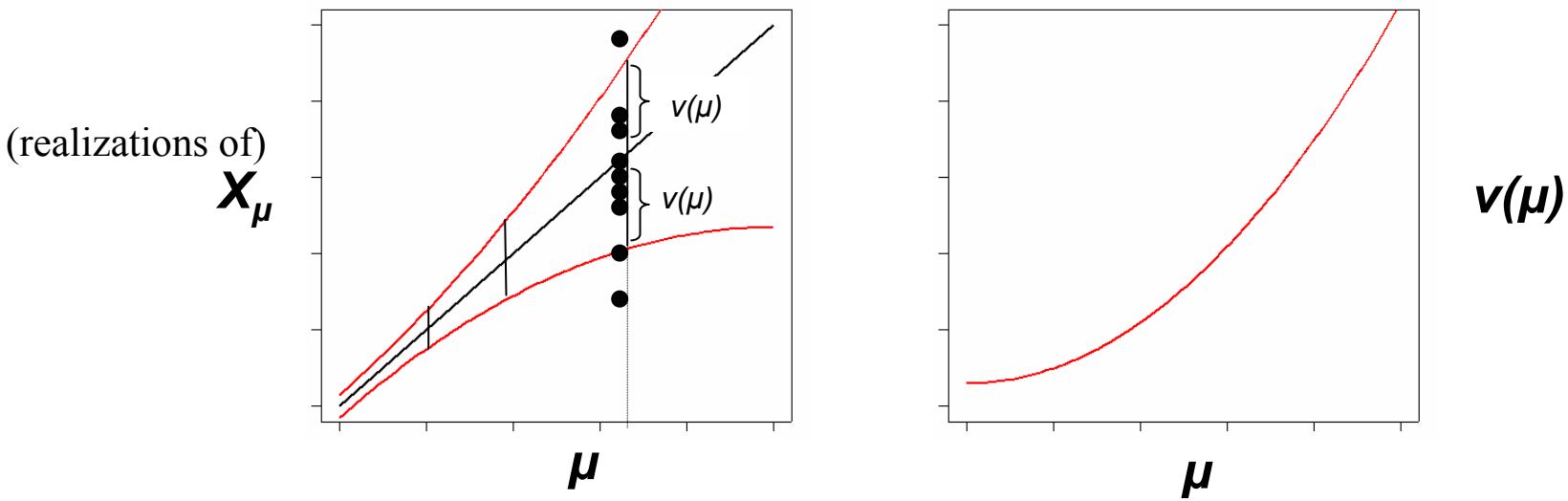
Variance stabilizing transformations

- Let X_μ, $\mu \in [a,b]$, be a family of random variables X_μ with expectation value

$$E(X_\mu) = \mu$$

and variance

$$Var(X_\mu) = v(\mu).$$



We seek a transformation $T: IR \rightarrow IR$ such that

$$Var(T(X_\mu)) \approx const.$$

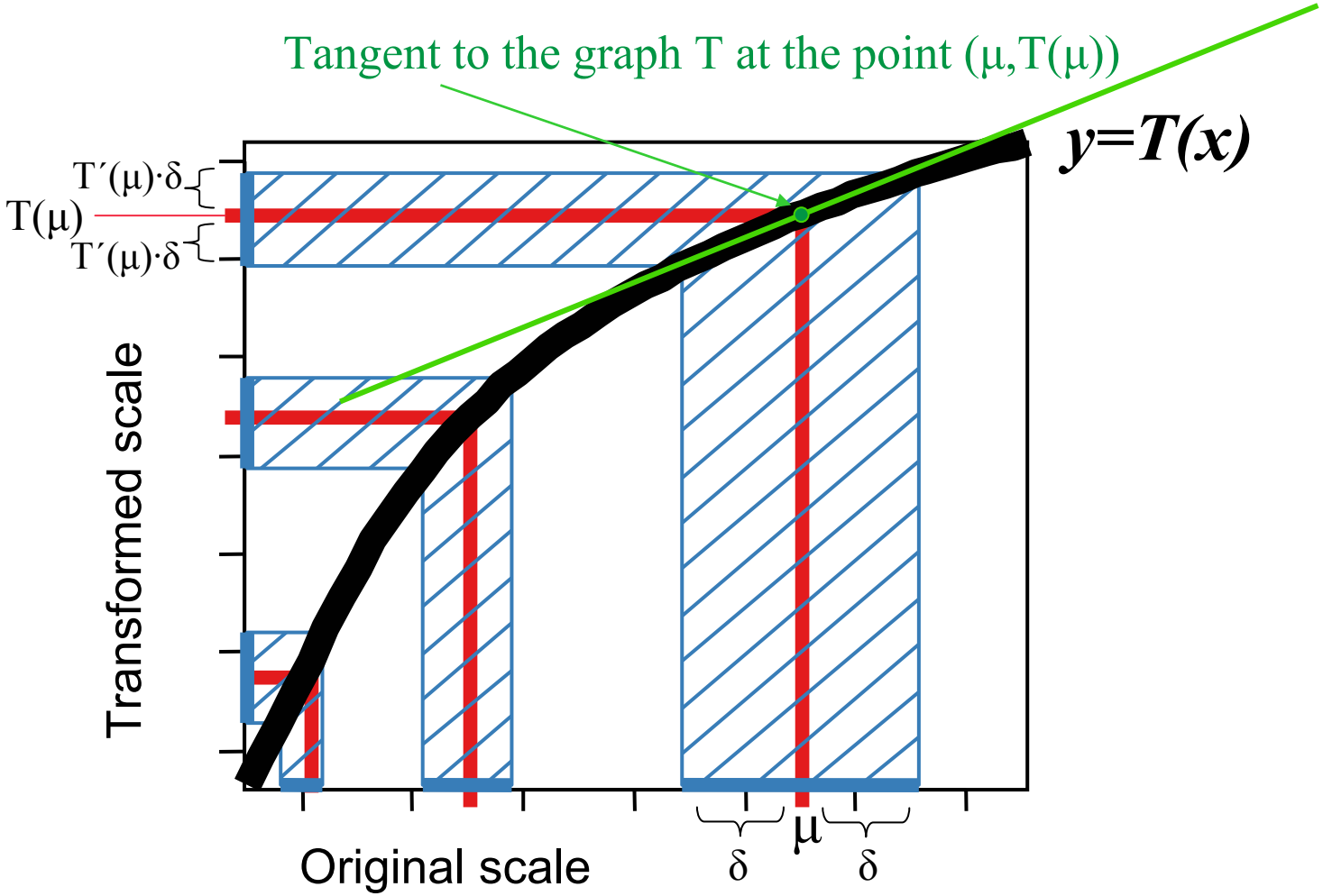
Variance Stabilizing Transformations

What are variance stabilizing transformations good for?

After variance stabilization with $\boldsymbol{T}$ the data are **homoskedastic**, i.e. the variance of the transformed random variables $\boldsymbol{T(X_\mu), \mu \in [a,b],}$ is (approximately) constant (the antonym of homosketasticity is **heteroskedasticity**. Regarding the replicate measurements of the expression of a gene with mean expression $\boldsymbol{\mu}$ as realizations of a random variable $\boldsymbol{X_\mu}$, the $\boldsymbol{X_\mu, \mu \in [a,b],}$ are heteroskedastic).

Homoskedastic data enable the application of more powerful statistical tests. E.g. the requirements for the application of the t-test as a test for differential expression are better fulfilled with transformed, homoskedastic data.

Deduction of the Variance Stabilizing Transformation



A differentiable function ***T:IR→IR*** can be approximated linearly in the neighourhood of ***μ*** by

$$T(x) \approx T(\mu) + T'(\mu)\cdot(x-\mu)$$

Deduction of the Variance Stabilizing Transformation

Hence for given Transformation ***T*** we have:

$$T(X_\mu) \approx T(\mu) + T'(\mu)\cdot(X_\mu - \mu)$$

And we can calculate the variance of $T(X_\mu)$ as

$$\begin{aligned} Var(T(X_\mu)) &\approx Var(\, T(\mu) + T'(\mu)\cdot(X_\mu - \mu)\,) \\ &= (T'(\mu))^2\, Var(X_\mu - \mu) \\ &= (T'(\mu))^2\, Var(X_\mu) \\ &= (T'(\mu))^2\, v(\mu) \end{aligned}$$

All that rests is to "whish" $Var(T(X_\mu))$ to be constant, 1 say, and solve the resulting equation for ***T***.

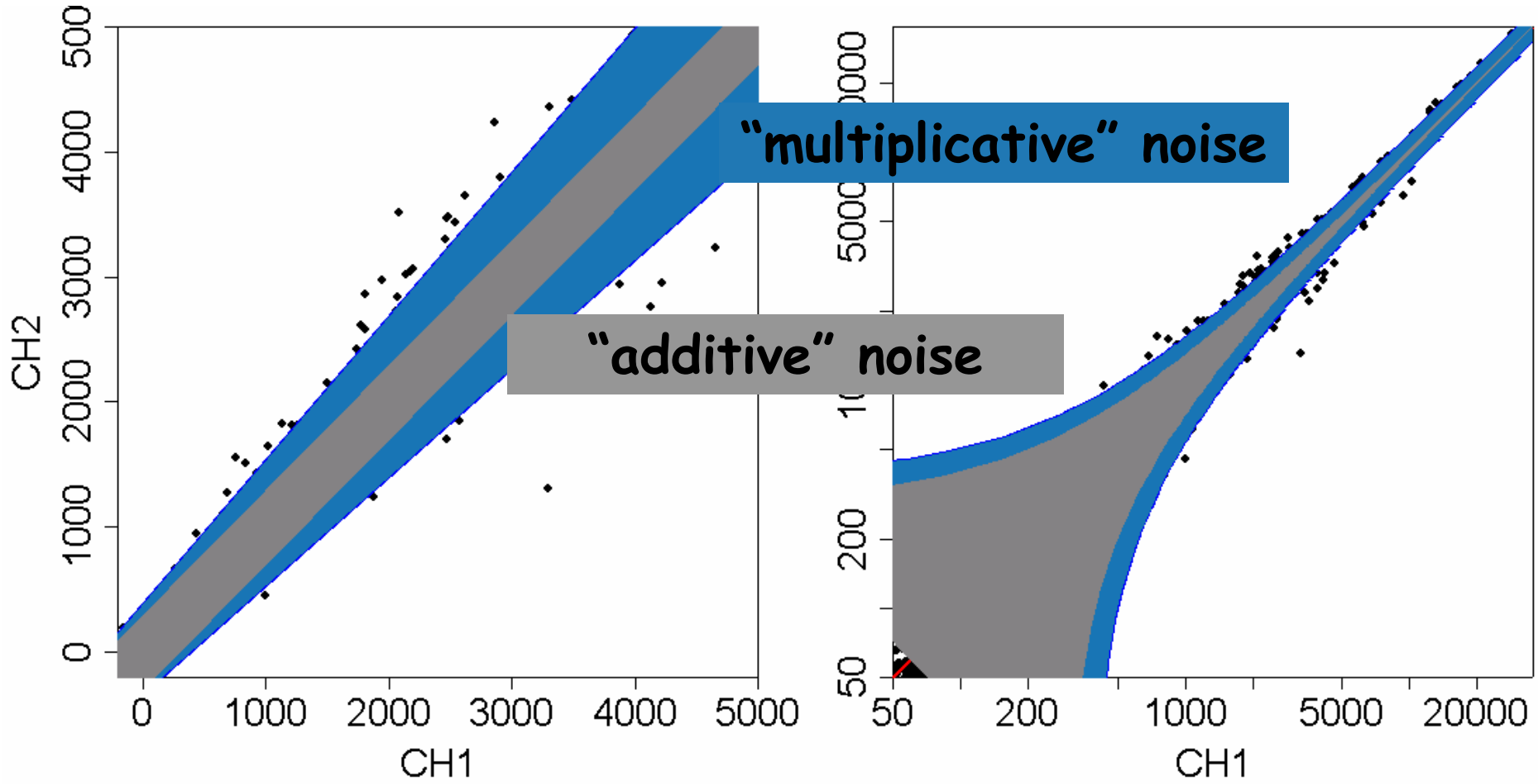
$$1 = Var(T(X_\mu)) \approx (T'(\mu))^2 v(\mu)$$

$$\rightarrow \quad T'(\mu) = 1\big/\sqrt{v(\mu)}$$

$$\rightarrow \quad T(\mu) = \int_0^\mu 1\big/\sqrt{v(t)}\,\mathrm{d}t$$

(modulo an additive constant)

Determination of *v(μ)*: The Two-Component Error Model



raw scale

log scale

B. Durbin, D. Rocke, JCB 2001

The Two-Component Error Model (for one gene)

μ : "true" gene expression

X_μ : measured gene expression

$$X_\mu = a + \varepsilon + b \cdot \mu \cdot (1 + \eta)$$

$$X_\mu = a + \varepsilon + b \cdot \mu \cdot \exp^{\eta}$$

For small η, both variants are practically equivalent

a constant background	Constant for all probes of one array and one colour, varies with array and colour (Cy5/Cy3)
ε background noise	iid for each spot
b constant amplification factor	Constant for all probes of one array and one colour, varies with array and colour (Cy5/Cy3)
η random amplification fluctuations	iid for each spot

Calculation of the variance stabilizing transformation for different model specifications

$$\left.\begin{aligned} X_\mu &= a + \varepsilon + b \cdot \mu \cdot (1+\eta) \\ \varepsilon &\sim N(0,\sigma^2)\ ,\ \eta \sim N(0,\tau^2) \end{aligned}\right\}$$

Specified error model

a) No multiplicative noise ($\tau = 0$) :

$$v(\mu) = Var(X_\mu) = Var(a + \varepsilon + b \cdot \mu)$$

$$= Var(\varepsilon) = \sigma^2$$

$$\Rightarrow\ T(\mu) = \int_0^\mu 1/\sqrt{v(t)}\,\mathrm{d}t = \int_0^\mu 1/\sqrt{\sigma^2}\,\mathrm{d}t = \frac{\mu}{\sigma}$$

T is merely a proportional rescaling

Calculation of the variance stabilizing transformation for different model specifications

b) No additive noise ($\sigma = 0$) :

$$\begin{pmatrix} X_\mu = a + \varepsilon + b \cdot \mu \cdot (1+\eta) \\ \varepsilon \sim N(0,\sigma^2)\ ,\ \eta \sim N(0,\tau^2) \end{pmatrix}$$

$$v(\mu) = Var(X_\mu) = Var(a + b \cdot \mu \cdot (1+\eta))$$

$$= b^2 \mu^2 Var(\eta) = b^2 \mu^2 \tau^2$$

$$\Rightarrow \quad T(\mu) = \int_1^\mu 1/\sqrt{v(t)}\, \mathrm{d}t = \int_1^\mu 1/(bt\tau)\, \mathrm{d}t$$

$$= \frac{\log(b\tau\mu)}{b\tau} + \text{const.} = \frac{\log(\mu)}{b\tau} + \text{const}'.$$

T is (up to rescaling) the logarithmic transformation

Calculation of the variance stabilizing transformation for different model specifications

c) Unrestricted model :

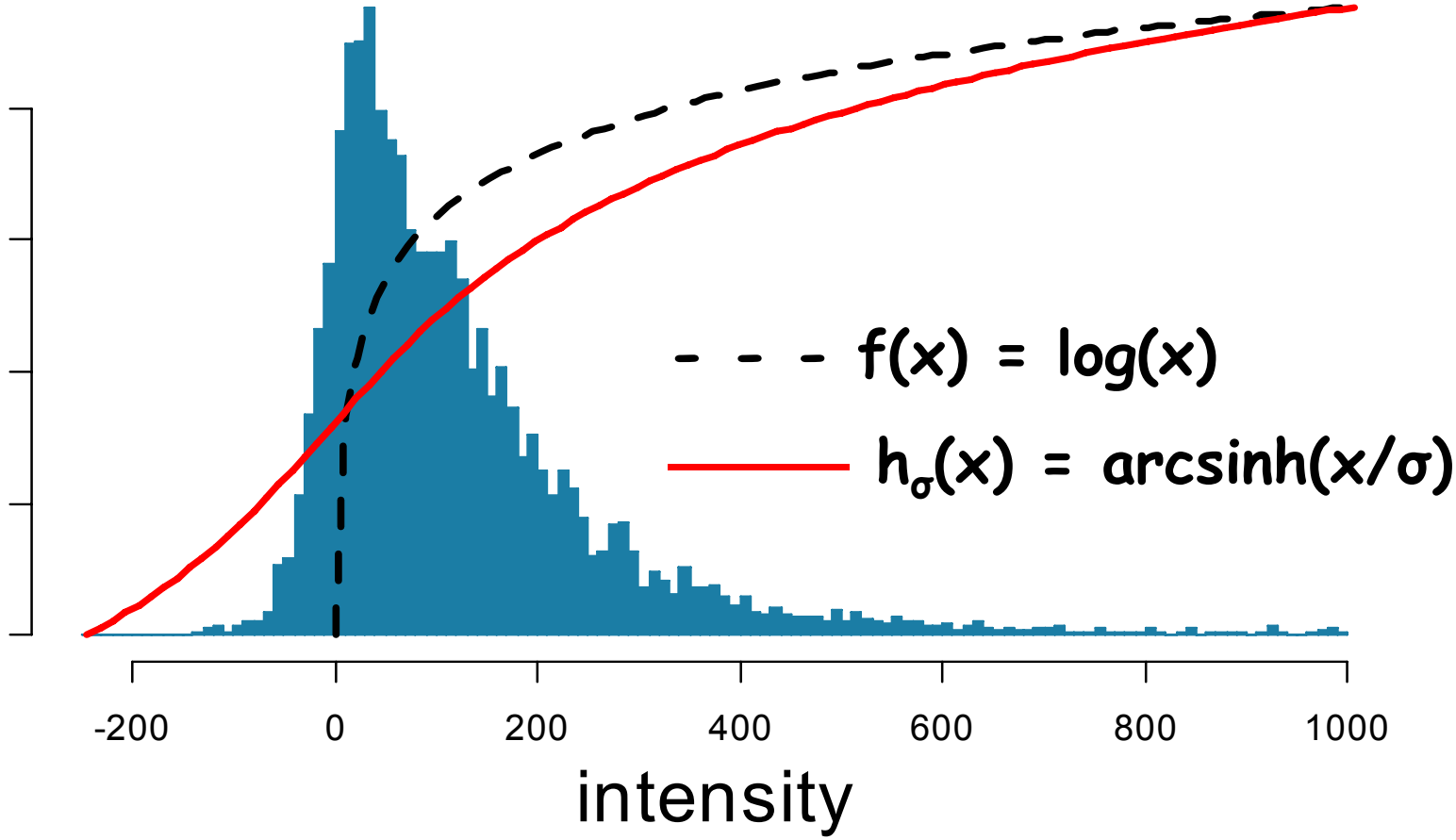
$$v(\mu) = Var(a + \varepsilon + b \cdot \mu \cdot (1 + \eta)) = \sigma^2 + b^2 \mu^2 \tau^2$$

$$\Rightarrow \quad T(\mu) = \int_1^{\mu} 1/\sqrt{v(t)}\, \mathrm{d}t = \int_1^{\mu} 1/\sqrt{\sigma^2 + b^2 t^2 \tau^2}\, \mathrm{d}t$$

$$\overset{\text{up to rescaling}}{=} \quad \operatorname{arcsinh}\left(\frac{\mu}{\sigma}\right)$$

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

The „glog“-Transformation



$$\operatorname{arcsinh}(x) = \log(x + \sqrt{x^2 + 1})$$

$$\lim\,(\operatorname{arcsinh}(x) - \log(x)) \xrightarrow{x \to \infty} \log(2)$$

The Two-Component Model for the whole Array

measured intensity = offset + gain × true abundance

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

a_{ik}

a_i per

$\varepsilon_{ik} \sim N$

"add

(η_{ik})

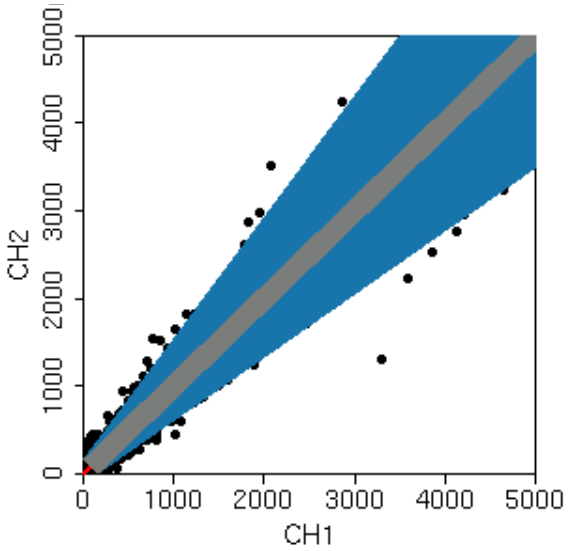
tor

"multiplicative noise"

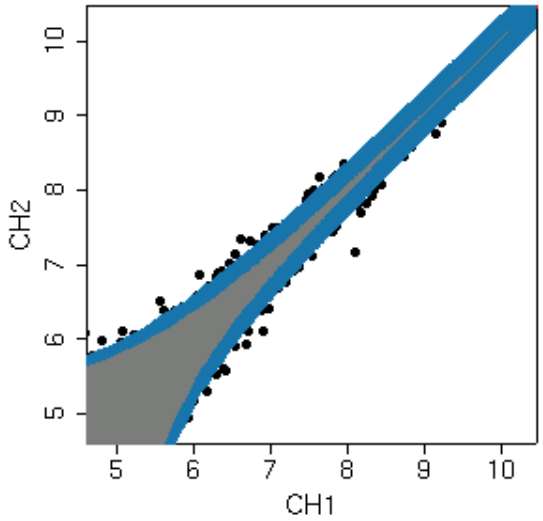
Cave: This model applies only to the unaltered genes, which are supposed to account for at least 50% of all genes.
A robust fitting method for the estimation of the parameters a_i, b_i, s_1, s_2 has been developed by W.Huber and A.v.Heydebreck.

The resulting transformation method has been implemented in the ***R*** package ***vsn***.

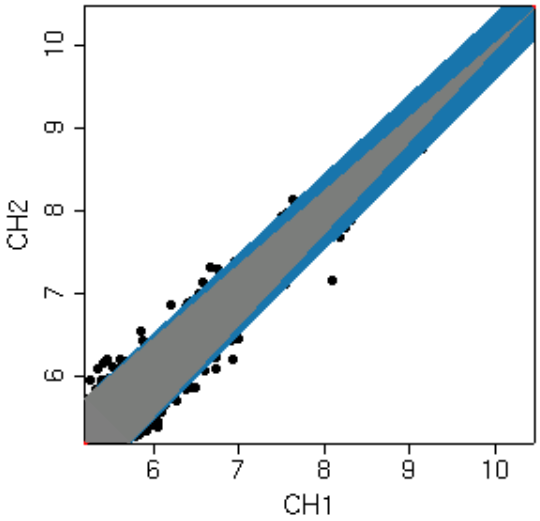
The „glog“-Transformation



no transformation



log transformed



glog transformed

Variance:

Additive component

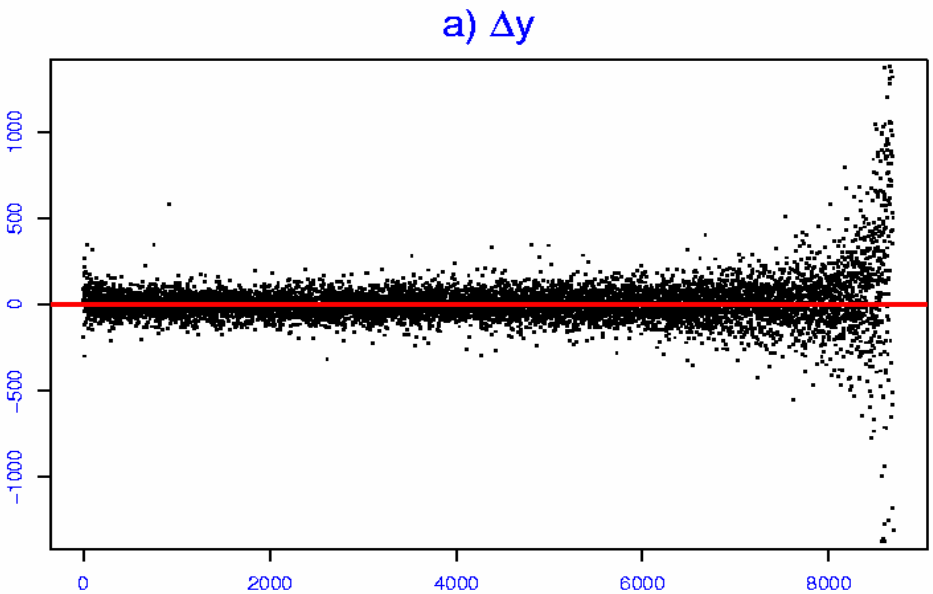
multiplicative component

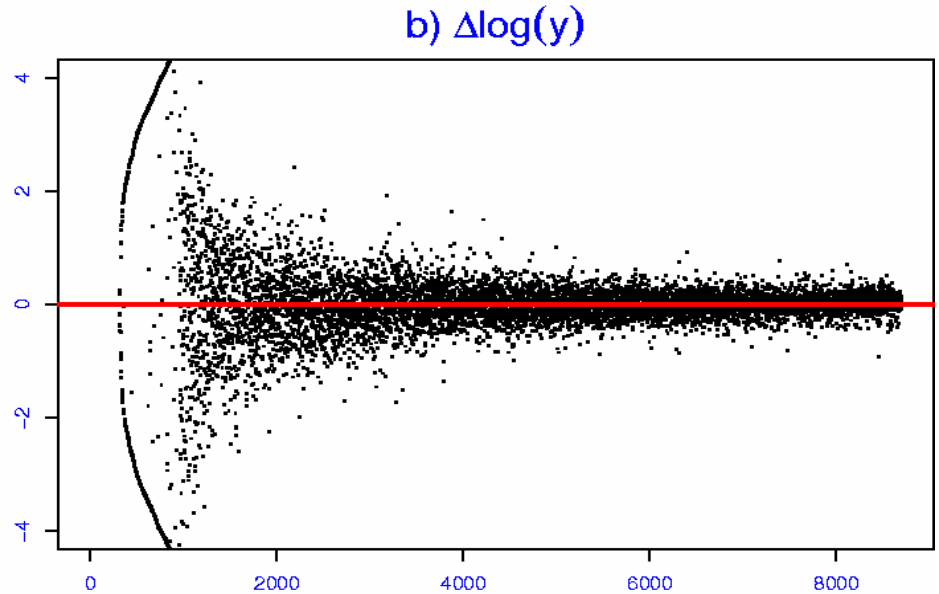
P. Munson, 2001

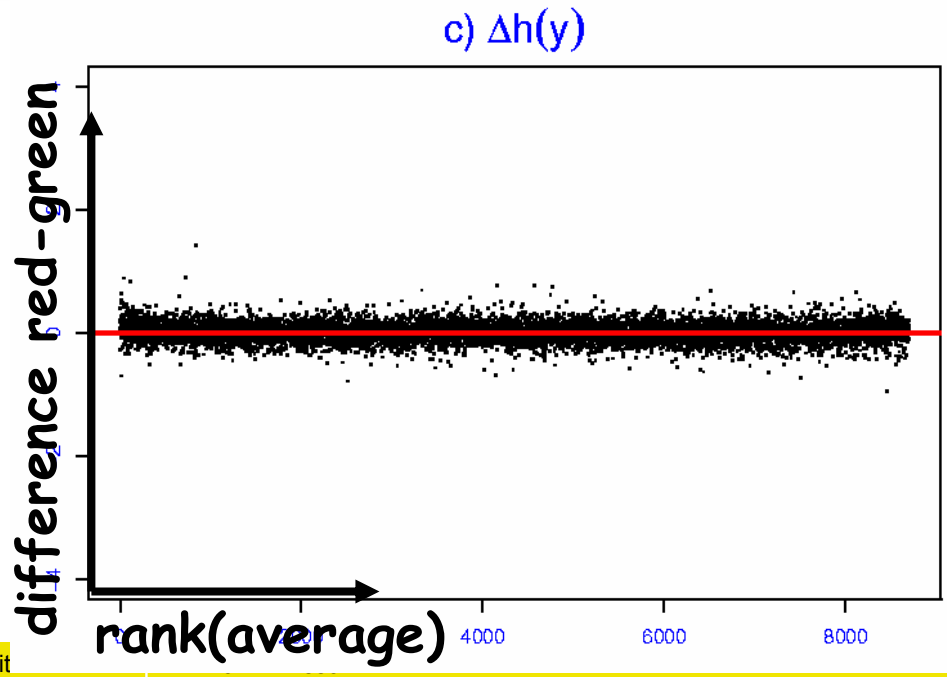
D. Rocke & B. Durbin, ISMB 2002

W. Huber et al., ISMB 2002

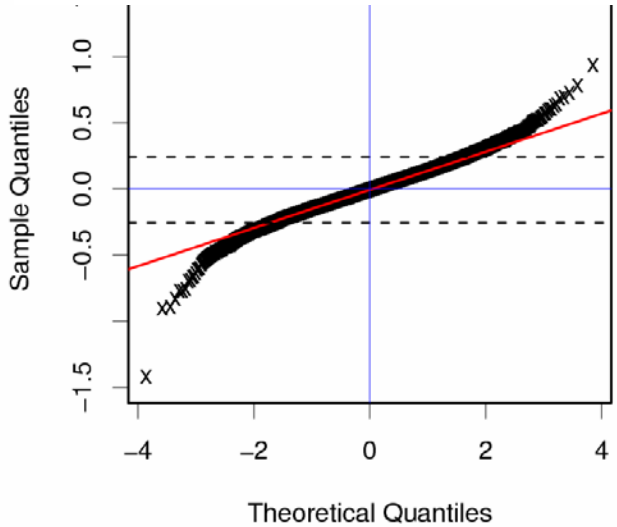
Evalutation: Effects of different Data Transformations



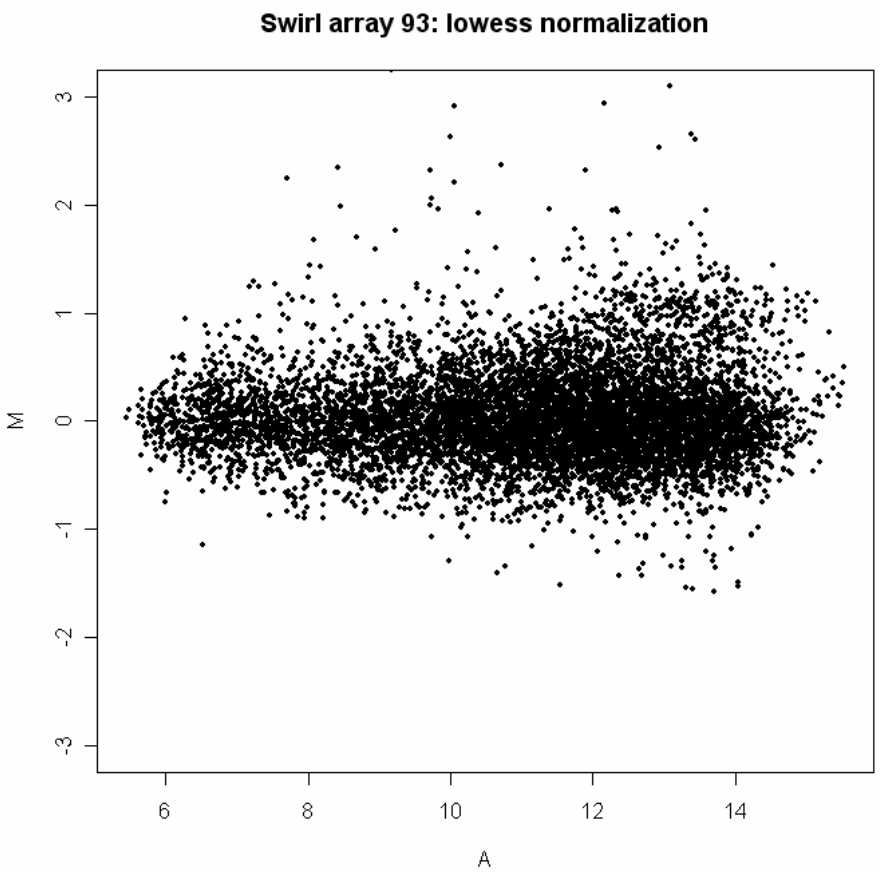


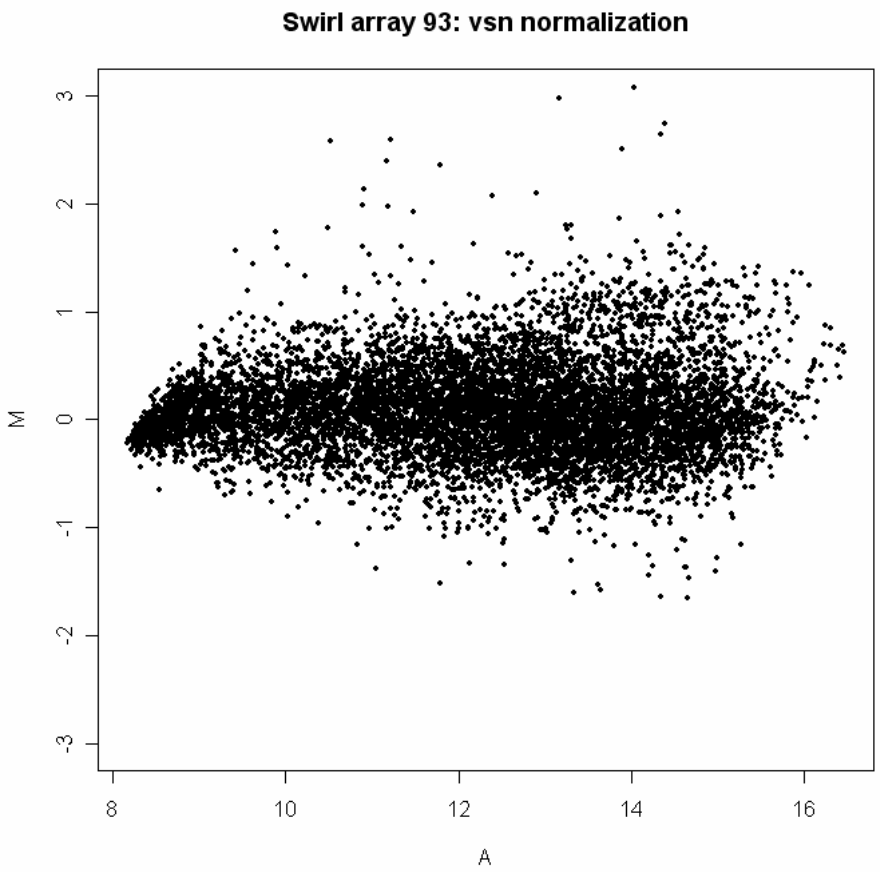


Normality of residuals: QQ-plot



Swirl Data: Lowess versus VSN





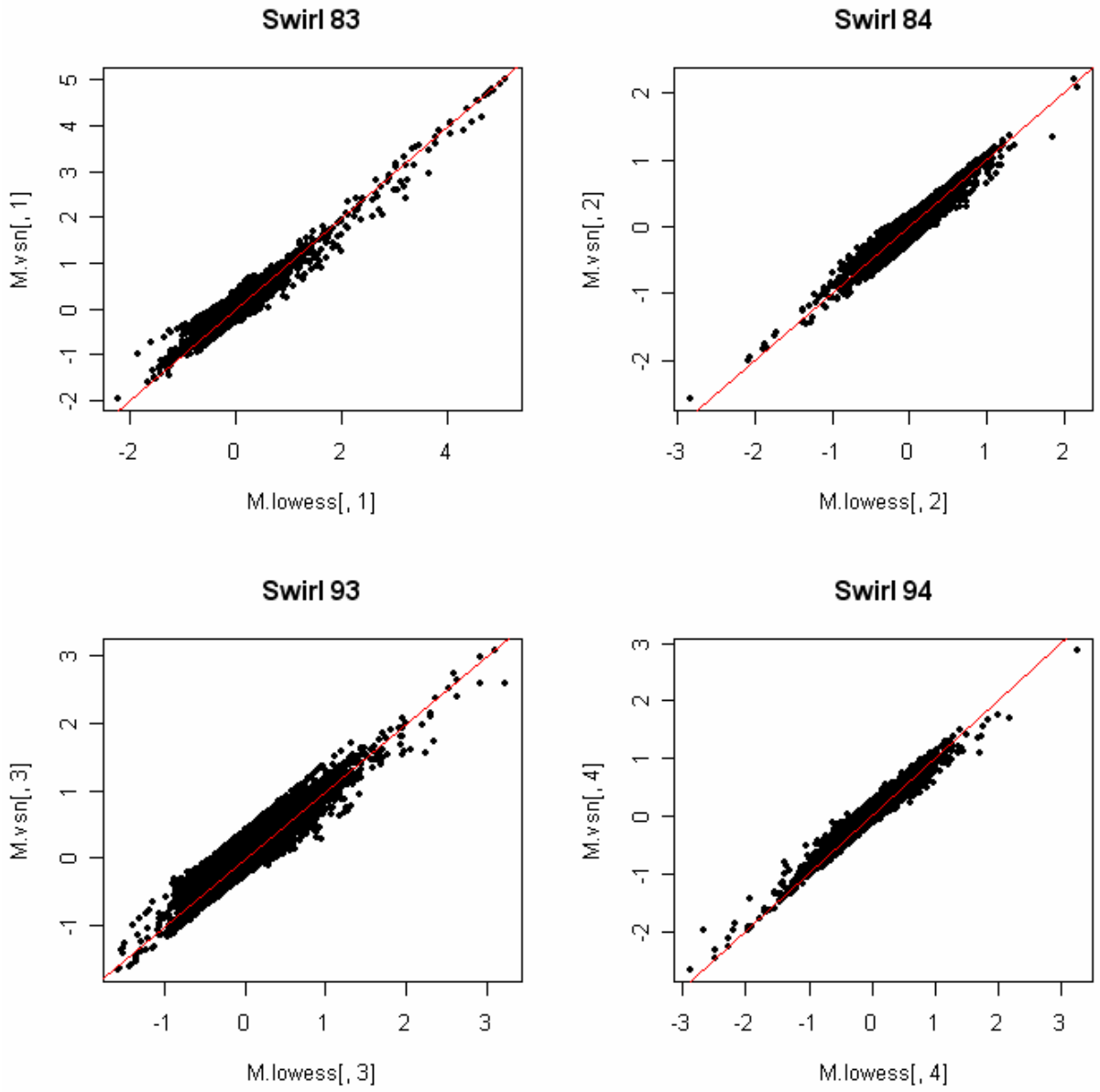
R Console

```
> plot(maA(swirl.norm[,3]), maM(swirl.norm[,3]), ylim=c(-3,3))
> library(vsn); library(limma);
> A.vsn<-log2(exp(exprs(swirl.vsn[,6])+exprs(swirl.vsn[,5])))/2
> M.vsn<-log2(exp(exprs(swirl.vsn[,6])-exprs(swirl.vsn[,5])))
> plot(A.vsn, M.vsn, ylim=c(-3,3)
```

Swirl: LOWESS versus VSN

R Console

```
> M.lowess<-maM(swirl.norm)
> M.vsn<-log2(exp(exprs(
  swirl.vsn[,c(2,4,6,8)])-
  exprs(swirl.vsn[,c(1,3,5,7)]
  )))
> par(mfrow=c(2,2))
> plot(M.lowess[,1},
       M.vsn[,1], pch=20)
> abline(0,1, col="red")
> plot(M.lowess[,1},
       M.vsn[,1], pch=20)
> abline(0,1, col="red")
> plot(M.lowess[,1},
       M.vsn[,1], pch=20)
> abline(0,1, col="red")
> plot(M.lowess[,1},
       M.vsn[,1], pch=20)
> abline(0,1, col="red")
```



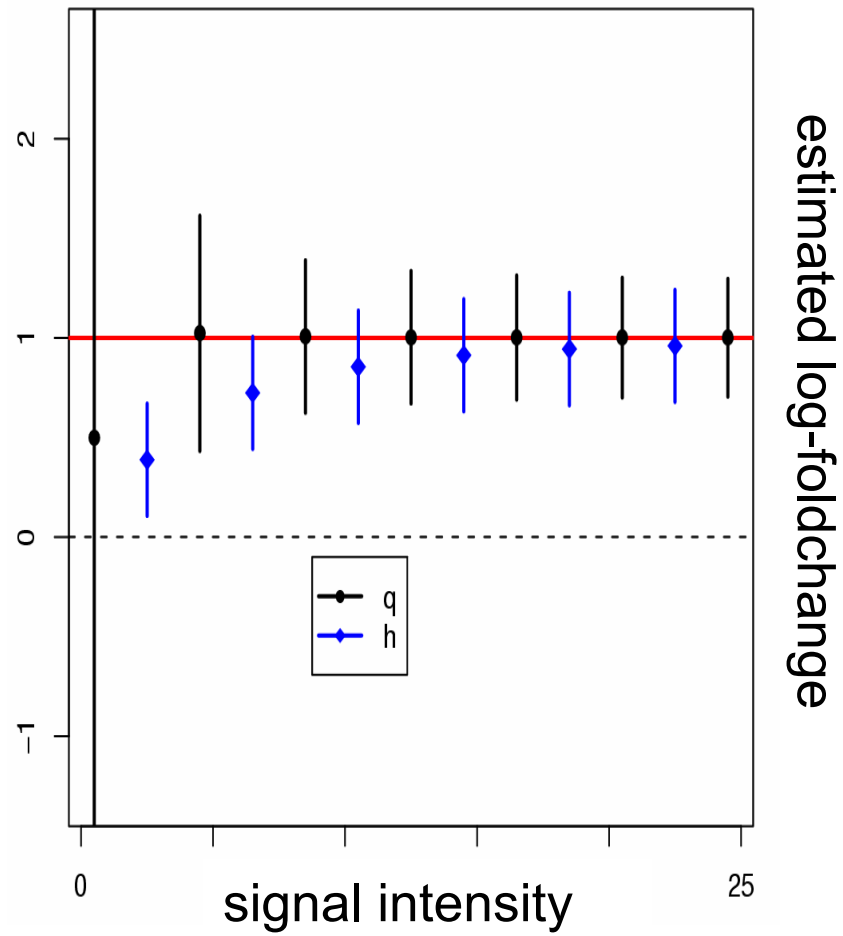
Fold change Estimation: Bias-Variance tradeoff

The traditional log-ratio $q = \log \frac{x_1}{x_2}$ is replaced by the „glog“-ratio

$$h = \log \frac{x_1 + \sqrt{x_1^2 + c_1^2}}{x_2 + \sqrt{x_2^2 + c_2^2}}$$

(c_1, c_2 parameters estimated by *vsn*)

The glog-ratio is a so-called shrinkage estimator: In exchange of an increased bias towards zero (relative to the log ratio), the variance of the glog ratio is smaller than that of the log ratio. Such an estimator is particularly useful in the case of low replicate numbers and thus large expected variances.



Summary

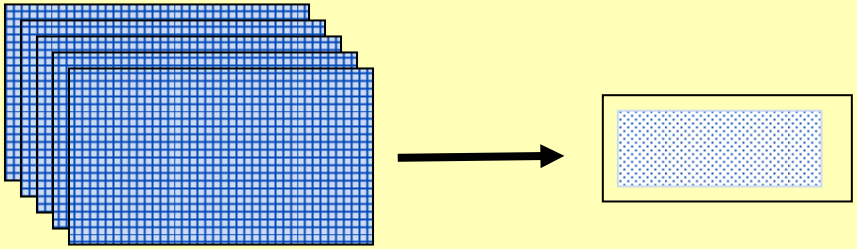
- What makes a good measurement: Precision and Unbiasednes
- Need to normalize.
- Normalization is not something trivial, has many practical and theoretical implications which need to be considered.
- What is the best way to normalize?
- How dependent is the result of your analysis from the normalization procedure?

Experimental Design

- Different levels of Replication
- Pooling vs. non Pooling
- Different Strategies to pair hybridization Targets on cDNA Arrays
- Direct vs. indirect Comparisons

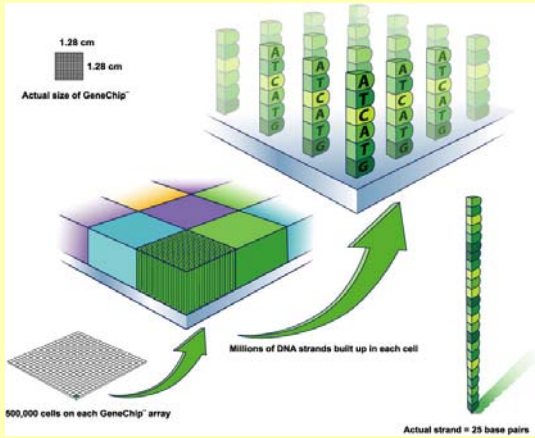
Two main aspects of array design

Design of the array



Arrayed Library
(96 or 384-well plates of bacterial glycerol stocks)

Spot as microarray on glass slides

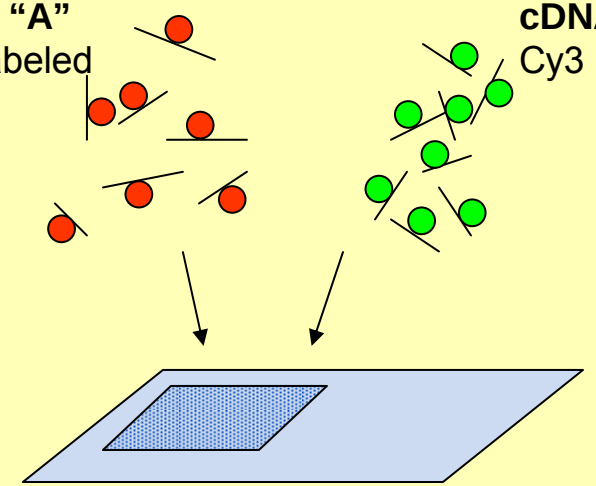


Allocation of mRNA samples to the slides

cDNA

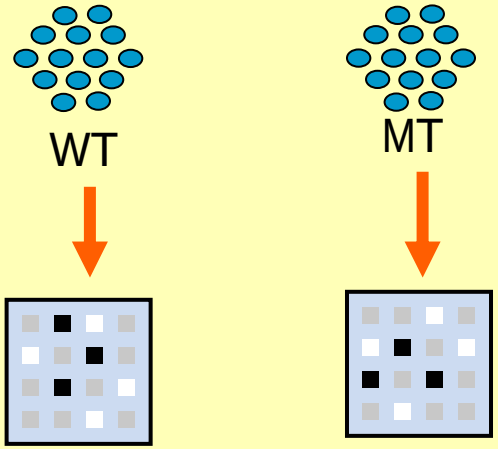
cDNA "A"
Cy5 labeled

cDNA "B"
Cy3 labeled



Hybridization

affy



2. Allocation of samples to the slides

A Types of Samples

- Replication – technical, biological
- Pooled vs individual samples
- Pooled vs amplification samples

This relates to both Affymetrix and two color spotted arrays

B Different design layout

- Scientific aim of the experiment
- Robustness
- Extensibility
- Efficiency

Applies to two color spotted arrays only

Preparing mRNA samples:

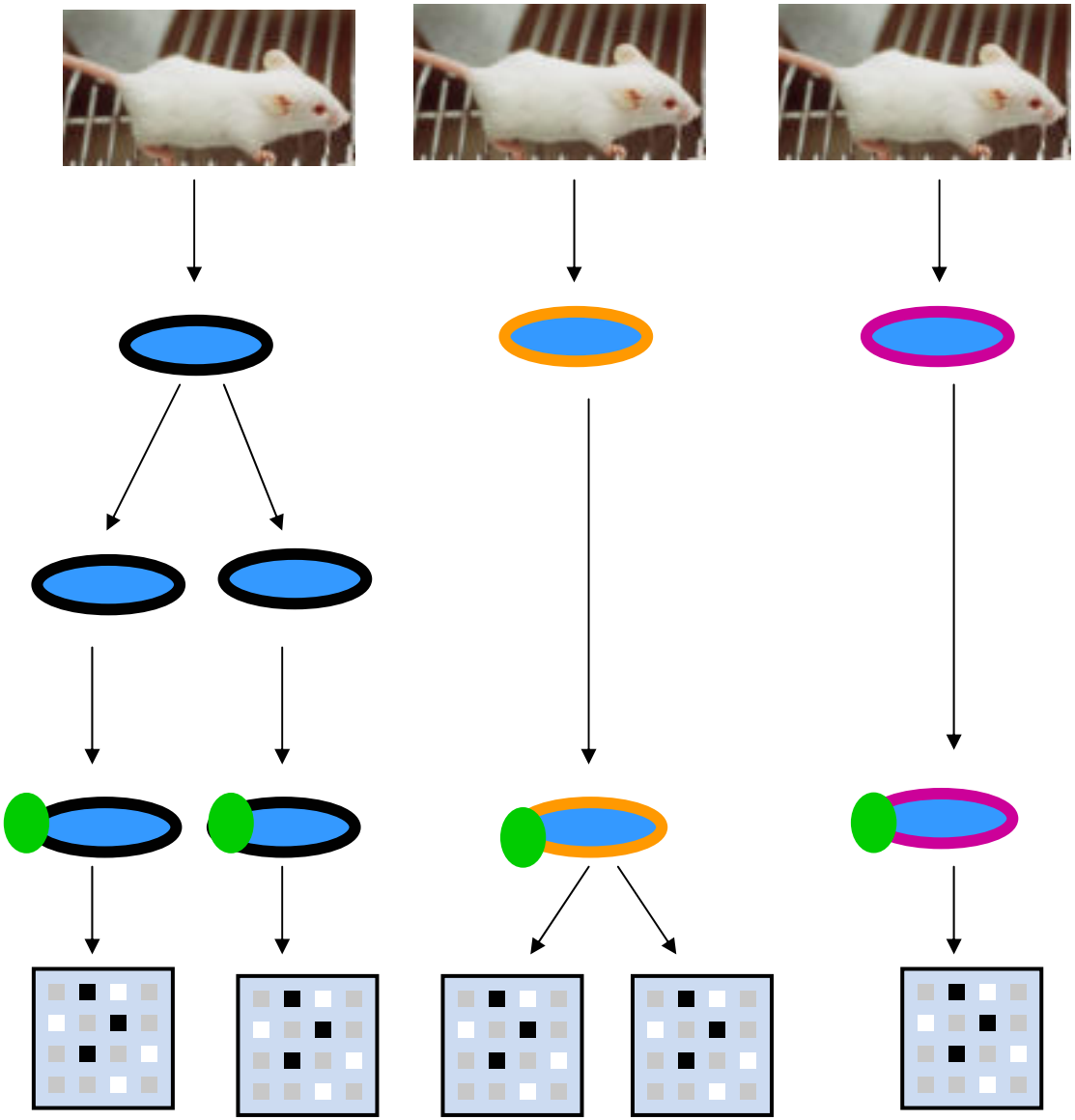
Mouse model
Dissection of tissue

RNA Isolation

Amplification

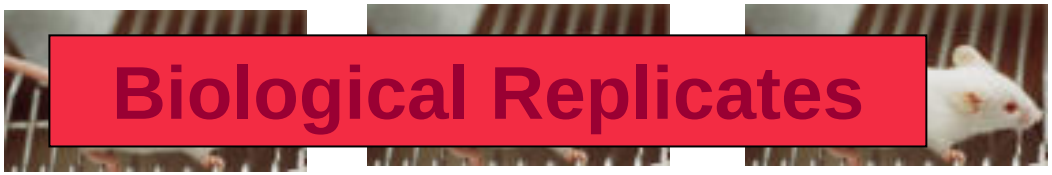
Probe labelling

Hybridization



Preparing mRNA samples:

Mouse model
Dissection of tissue

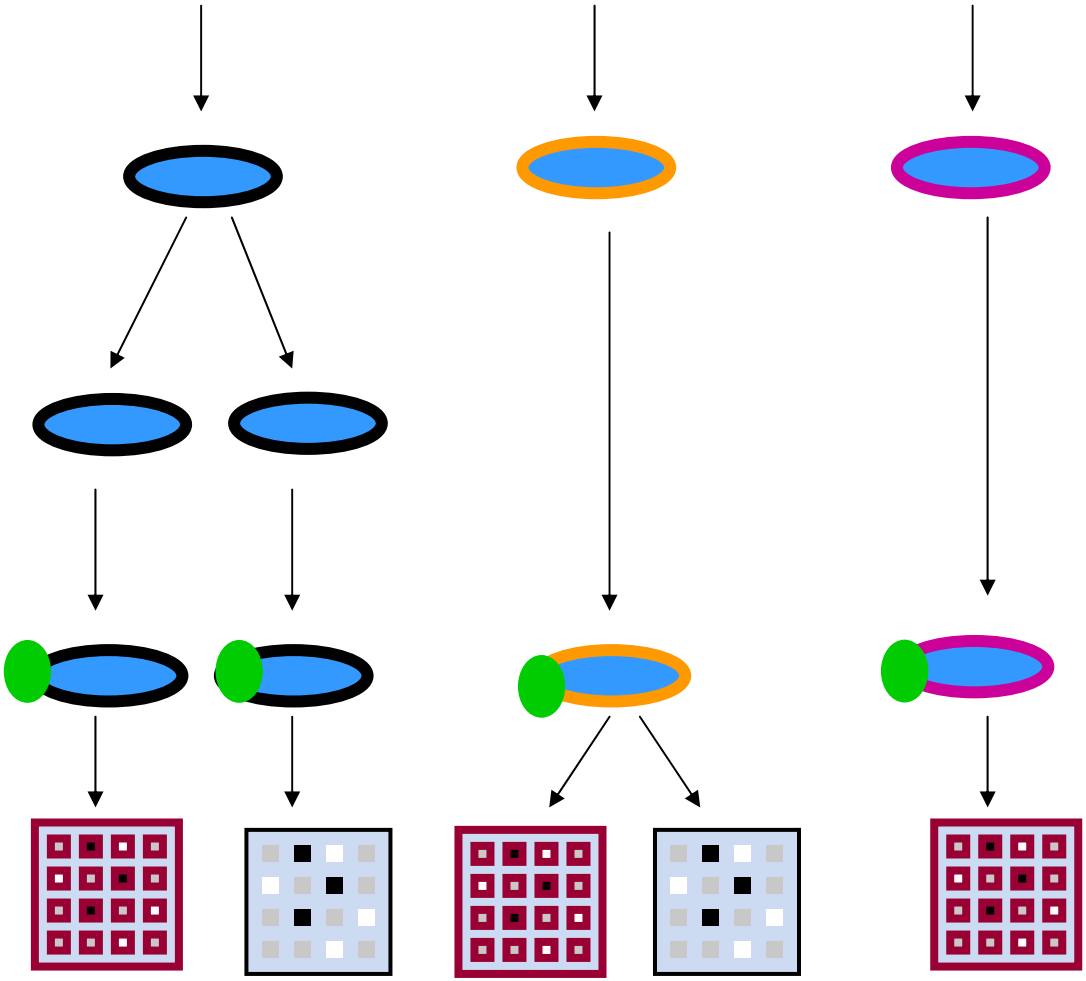


RNA Isolation

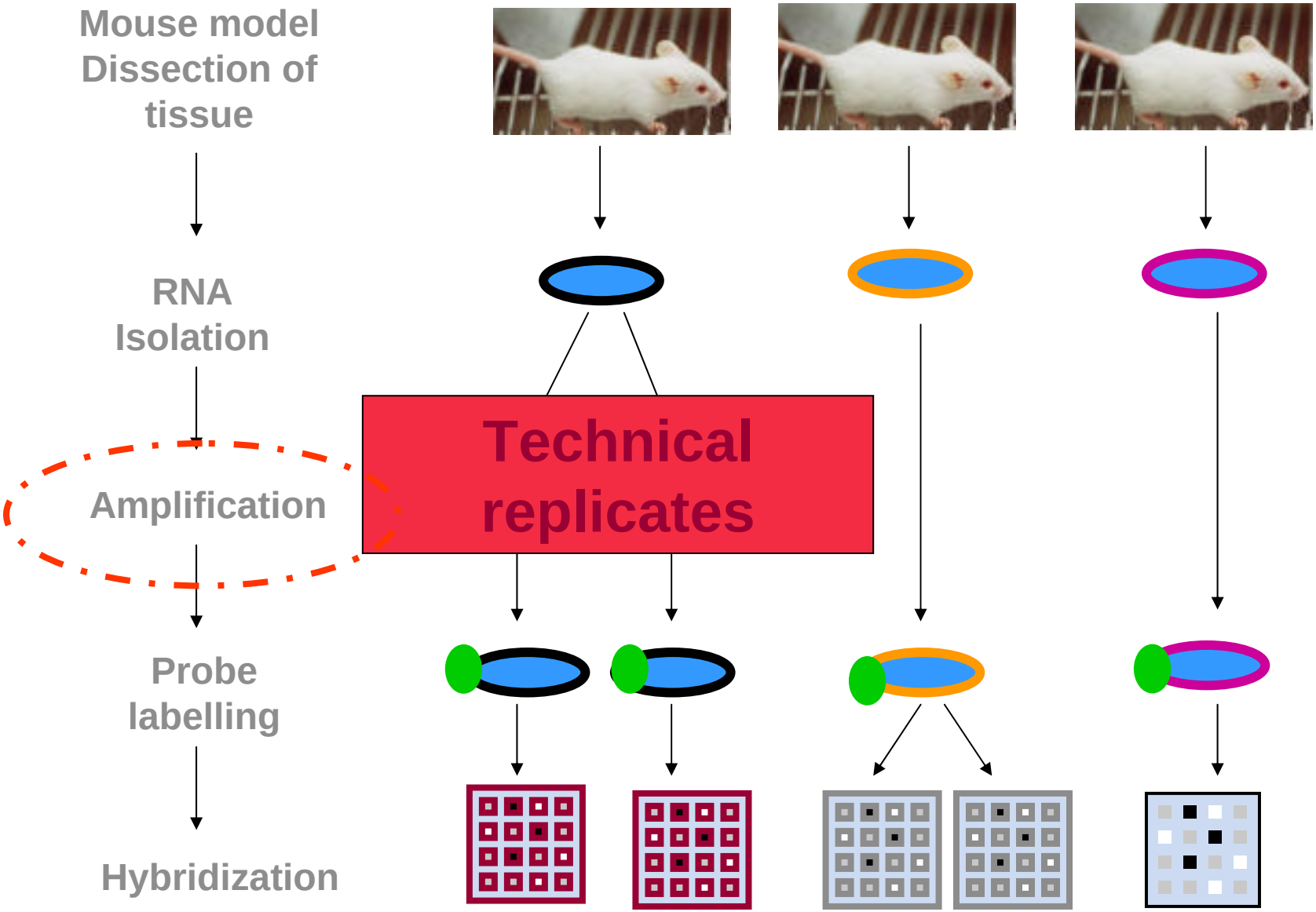
Amplification

Probe labelling

Hybridization



Preparing mRNA samples:



Pooling: looking at very small amount of tissues

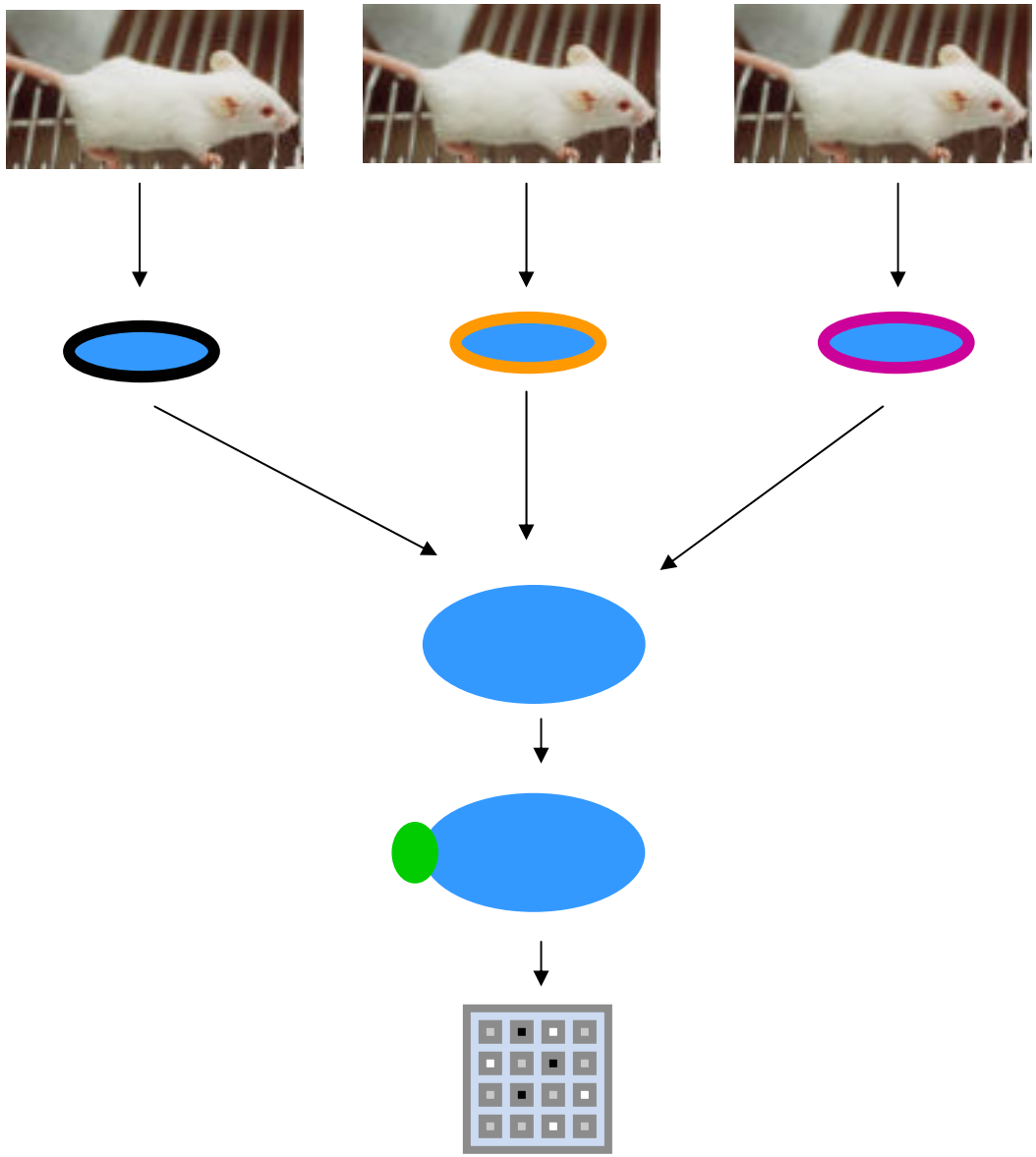
Mouse model
Dissection of tissue

RNA Isolation

Pooling

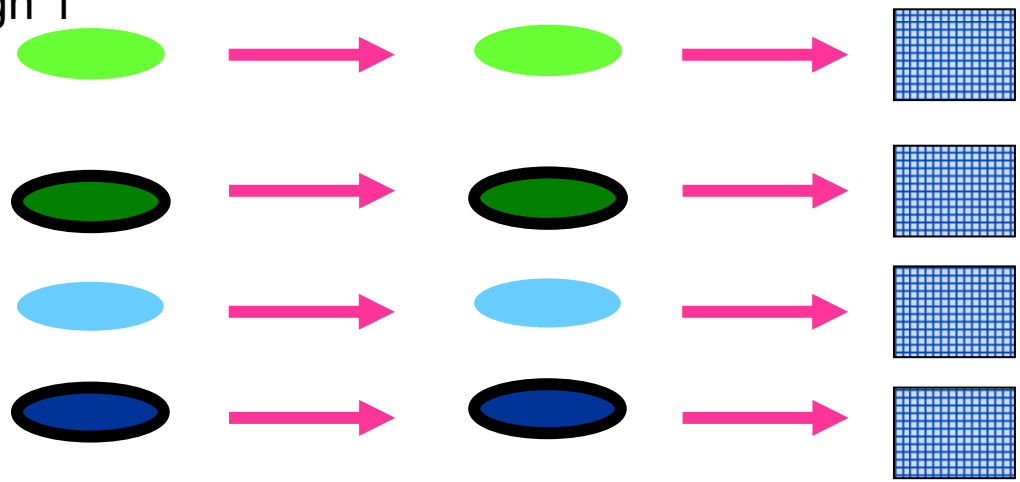
Probe labelling

Hybridization



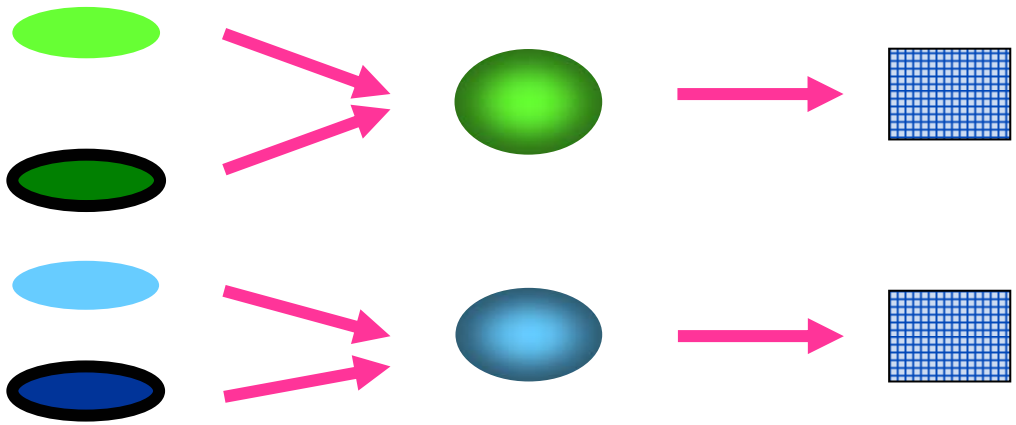
Pooled vs. Individual samples

Design 1



Bottleneck: Not enough chips available

Design 2



Taken from
Kendziorski etl al (2003)

Pooled versus Individual samples

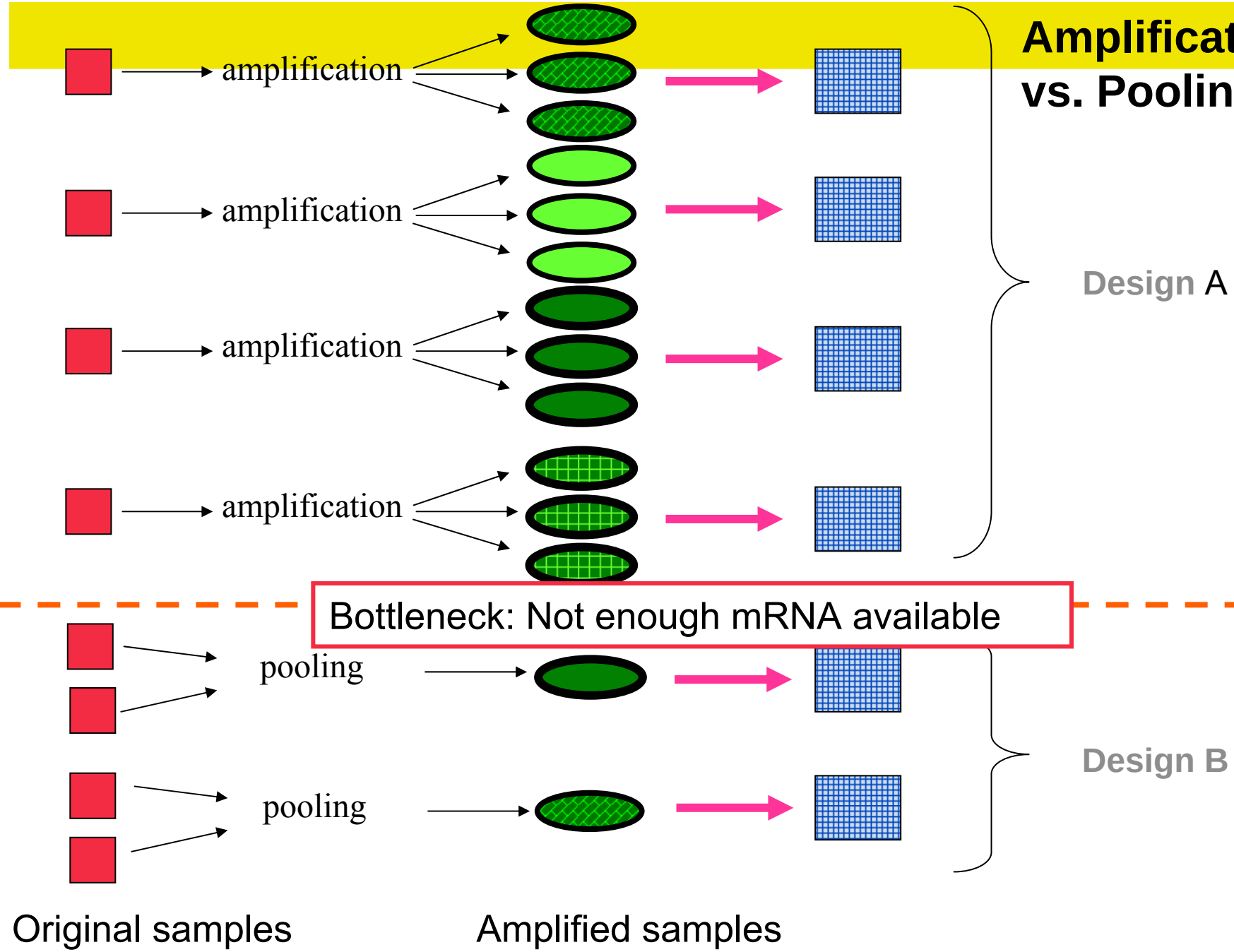
Pooling is seen as “biological averaging”.

Trade off between

- Cost of performing a hybridization.
- Cost of the mRNA samples.

- Case 1: Cost or mRNA samples << Cost per hybridization
  Pooling can assist reducing the number of hybridizations.
- Case 2: Cost or mRNA samples >> Cost per hybridization
  Hybridize every sample on an individual array to get the maximum amount of information.

Amplification vs. Pooling



Pooled vs Amplified samples

- In the cases where we **do not** have enough material from one biological sample to perform one array (chip) hybridizations, pooling or amplification are necessary.

- Amplification
  - Introduces more noise.
  - Non-linear amplification (??), different genes amplified at different rate.
  - Enables to perform more hybridizations.

- Pooling
  - Increased effort to obtain sufficiently large number of samples

2. Allocation of samples to the slides

A Types of Samples

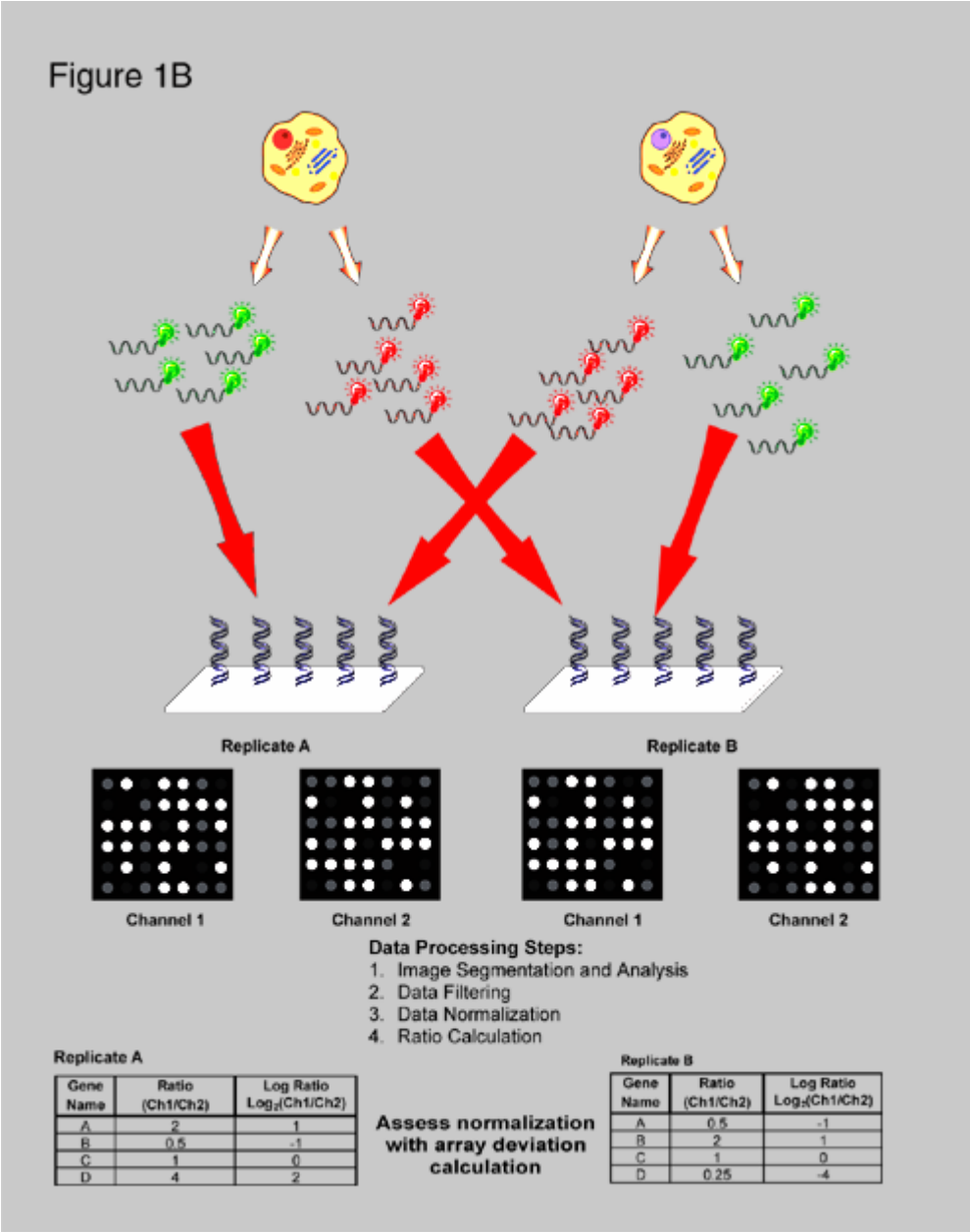
- Replication – technical, biological
- Pooled vs individual samples
- Pooled vs amplification samples

B Different design layout

- Scientific aim of the experiment
- Robustness
- Extensibility
- Efficiency

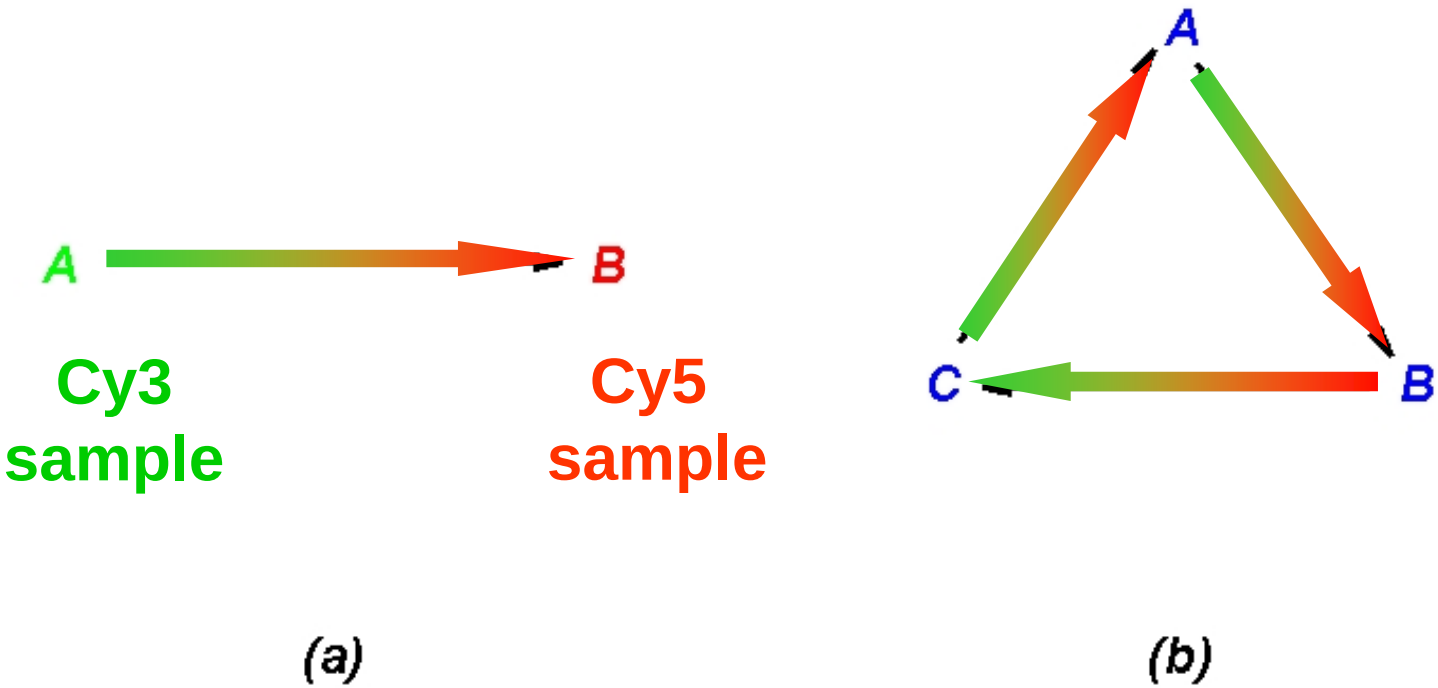
Design of a Dye-Swap Experiment

- Repeats are essential to control the quality of an experiment.
- One example for Replicates is the Dye-Swap, i.e. Replicates with the same mRNA Pool but with swapped labels.
- Dye-Swap shows whether there is a dye-bias in the Experiment.



Graphical representation

Vertices: mRNA samples;
Edges: hybridization;
Direction: dye assignment.



Graphical representation

- The structure of the graph determines which effects can be estimated and the precision of the estimates.
  - Two mRNA samples can be compared only if there is a path joining the corresponding two vertices.
  - The precision of the estimated contrast then depends on the number of paths joining the two vertices and is inversely related to the length of the paths.

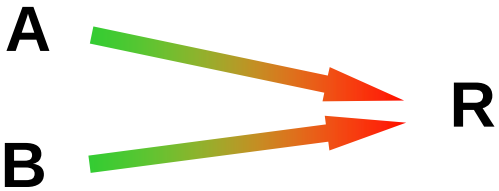
- Direct comparisons **within slides** yield more precise estimates than indirect ones between slides.

The first design question:
Direct versus indirect comparisons

Two samples (A *vs* B)
e.g. KO *vs.* WT or mutant *vs.* WT

Direct

Indirect



$$(\log(A/B) + \log(B/A)) / 2$$

$$\sigma^2 / 2$$

$$\log(A / R) - \log(B / R)$$

$$2\sigma^2$$

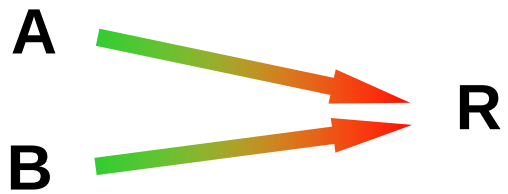
These calculations assume independence of replicates: the reality is not so simple.

Direct vs. Indirect - revisited

Two samples (A vs B)
e.g. KO *vs.* WT or mutant *vs.* WT

Direct

Indirect



$y = (a - b) + (a' - b')$

$Var(y/2) = \sigma^2/2 + \chi_1$

$y = (a - r) - (b - r')$

$Var(y) = 2\sigma^2 - 2\chi_1$

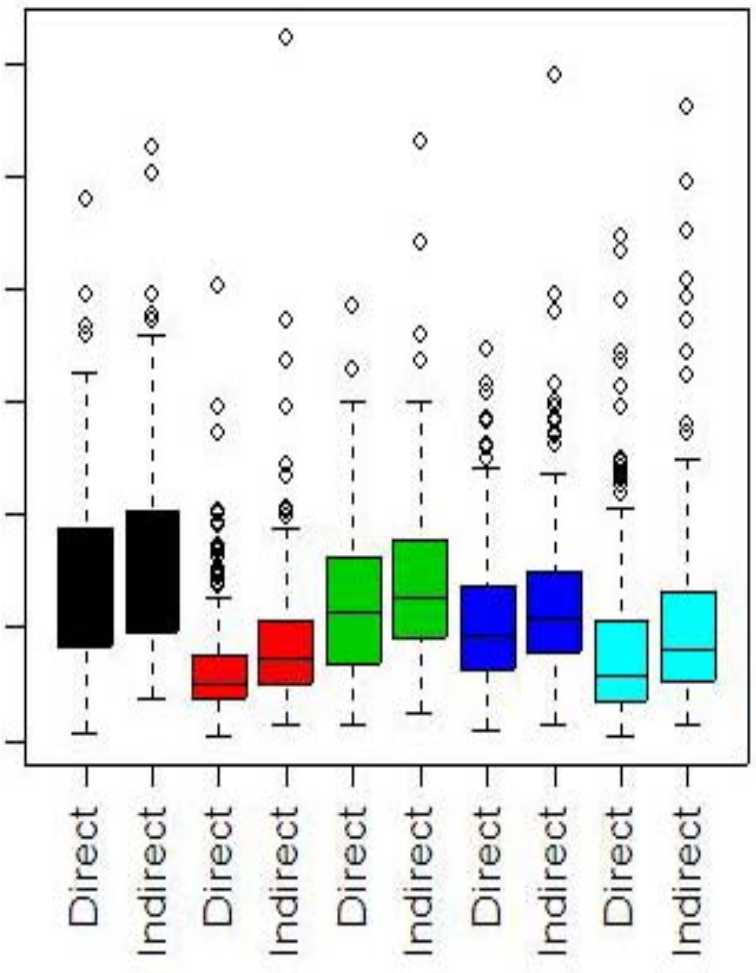
χ_1 = Correlation of replicates

$\sigma^2/2 = \chi_1$ efficiency ratio (Indirect / Direct) = 1

$\chi_1 = 0$ efficiency ratio (Indirect / Direct) = 4

Experimental results

- 5 sets of experiments with similar structure.
- Compare (Y axis)
  Direct) StdErr for $aveM_{mt}$
  Indirect) StdErr for $aveM_{mt} - aveM_{wt}$
- Theoretical ratio of (A / B) is 1.6
- Experimental observation is 1.1 to 1.4.



Experimental design

- Create highly correlated reference samples to overcome inefficiency in common reference design.

- Not advocating the use of technical replicates in place of biological replicates for samples of interest.

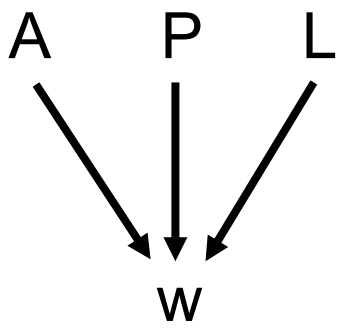
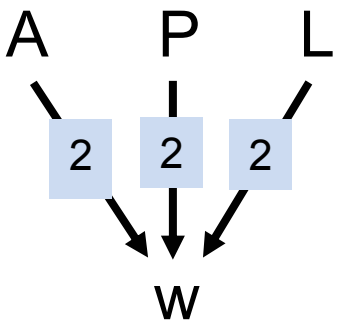
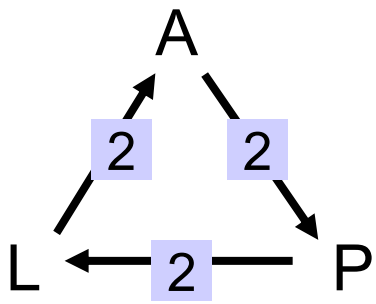
- Efficiency can be measured in terms of different quantities
  - number of slides or hybridizations;
  - units of biological material, e.g. amount of mRNA for one channel.

- In addition to experimental constraints, design decisions should be guided by the knowledge of which effects are of greater interest to the investigator.

  E.g. which main effects, which interactions.

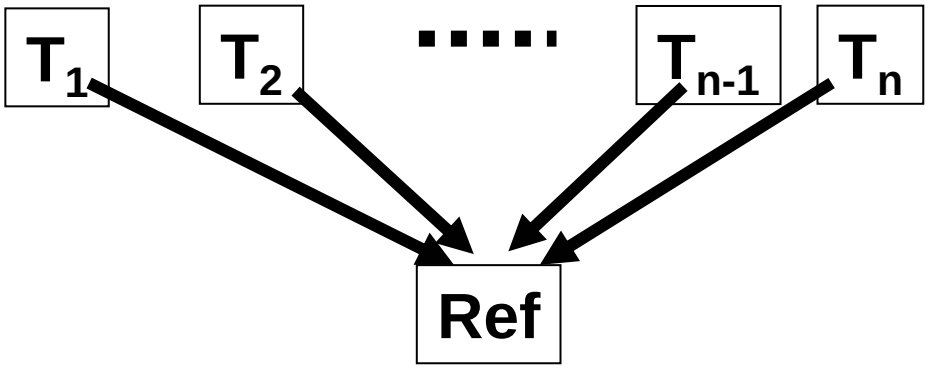
- The experimenter should thus decide on the comparisons for which he wants the most precision and these should be made **within slides** to the extent possible.

Experimental design

	I (a) Common reference	I (b) Common reference	II Direct comparison
			
Number of Slides	N = 3	N=6	N=6
mean Variance	2	1	0.67
used Material	A = P = L = 1	A = P = L = 2	A = P = L = 2

Efficiency rate (Design I(b) / Design II) = 1.5

Common reference design



- Experiment for which the common reference design is appropriate

  Meaningful biological control (C) Identify genes that responded differently / similarly across two or more treatments relative to control.

  Large scale comparison. To discover tumor subtypes when you have many different tumor samples.

- Advantages:

  Ease of interpretation.

  Robustness against failure of microarrays

  Extensibility - extend current study or to compare the results from current study to other array projects.

2x2 Factorial experiments

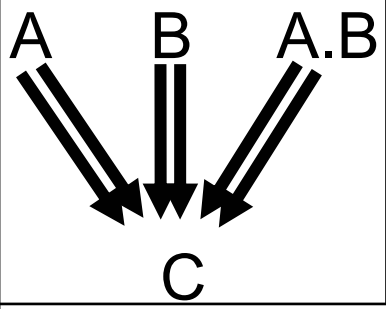
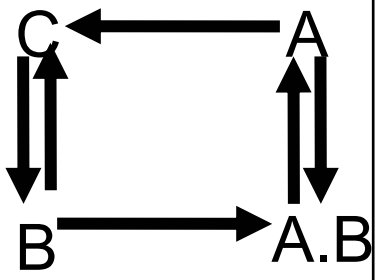
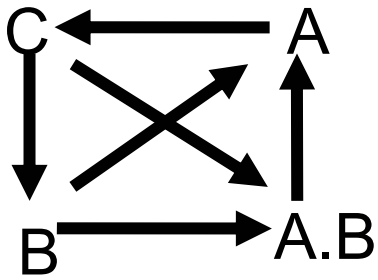
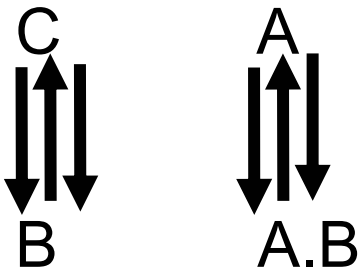
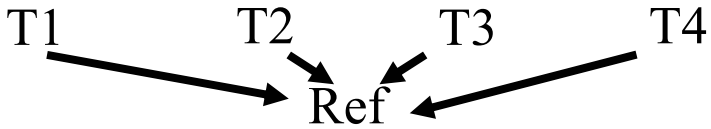
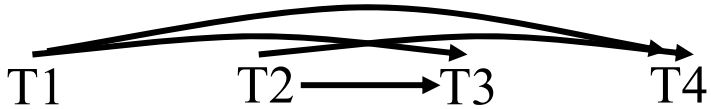
	Indirect	A balance of direct and indirect		
	I) 	II) 	III) 	IV) 
# Slides	N = 6			
Main effect A	0.5	0.67	0.5	NA
Main effect B	0.5	0.43	0.5	0.3
Interaction A.B	1.5	0.67	1	0.67

Table entry: variance

Ref: Glonek & Solomon (2002)

Time series experiments

		t vs t+1			t vs t+2			
		T1T2	T2T3	T3T4	T1T3	T2T4	T1T4	Ave
N=3	A) T1 as common reference 	1	2	2	1	2	1	1.5
	B) Direct Hybridization 	1	1	1	2	2	3	1.67
N=4	C) Common reference 	2	2	2	2	2	2	2
	D) T1 as common ref + more 	.67	.67	1.67	.67	1.67	1	1.06
	E) Direct hybridization choice 1 	.75	.75	.75	1	1	.75	.83
	F) Direct Hybridization choice 2 	1	.75	1	.75	.75	.75	.83

References

- T. P. Speed and Y. H Yang (2002). Direct versus indirect designs for cDNA microarray experiments. *Sankhya : The Indian Journal of Statistics*, Vol. 64, Series A, Pt. 3, pp 706-720

- Y.H. Yang and T. P. Speed (2003). Design and analysis of comparative microarray Experiments In T. P Speed (ed) **Statistical analysis of gene expression microarray data,** *Chapman & Hall.*

- R. Simon, M. D. Radmacher and K. Dobbin (2002). **Design of studies using DNA microarrays**. *Genetic Epidemiology 23:21-36.*

- F. Bretz, J. Landgrebe and E. Brunner (2003). **Efficient design and analysis of two color factorial microarray experiments**. *Biostaistics.*

- G. Churchill (2003). **Fundamentals of experimental design for cDNA microarrays**. *Nature genetics review 32:490-495.*

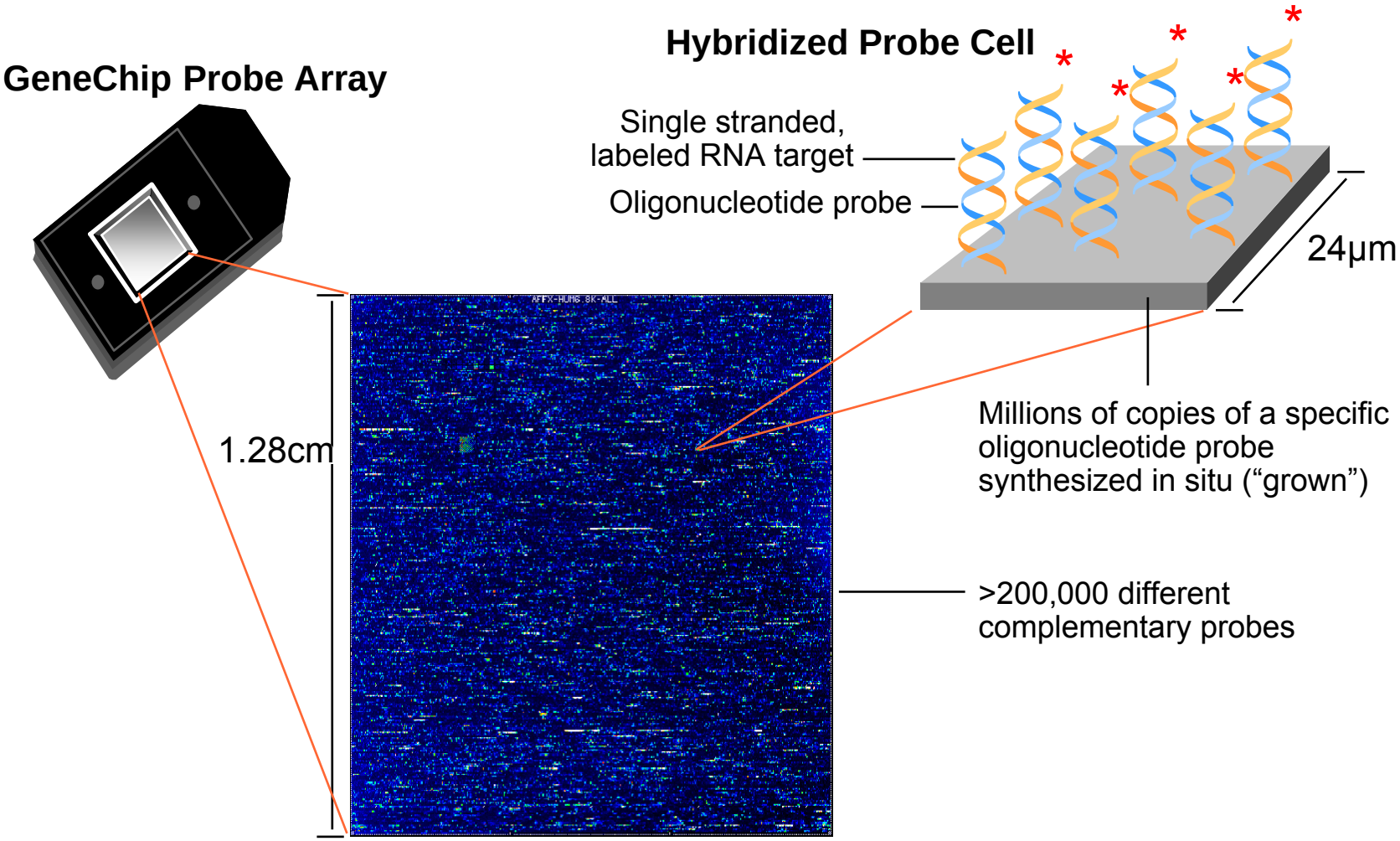
- G. Smyth, J. Michaud and H. Scott (2003) **Use of within-array replicate spots for assessing differential experssion in microarray experiments**. Technical Report In WEHI.

- Glonek, G. F. V., and Solomon, P. J. (2002). Factorial and time course designs for cDNA microarray experiments. Technical Report, Department of Applied Mathematics, University of Adelaide. 10/2002

Affy Chips:

PM versus MM and summary information

Affymetrix GeneChips: Technical details



GeneChip® Expression Array Design

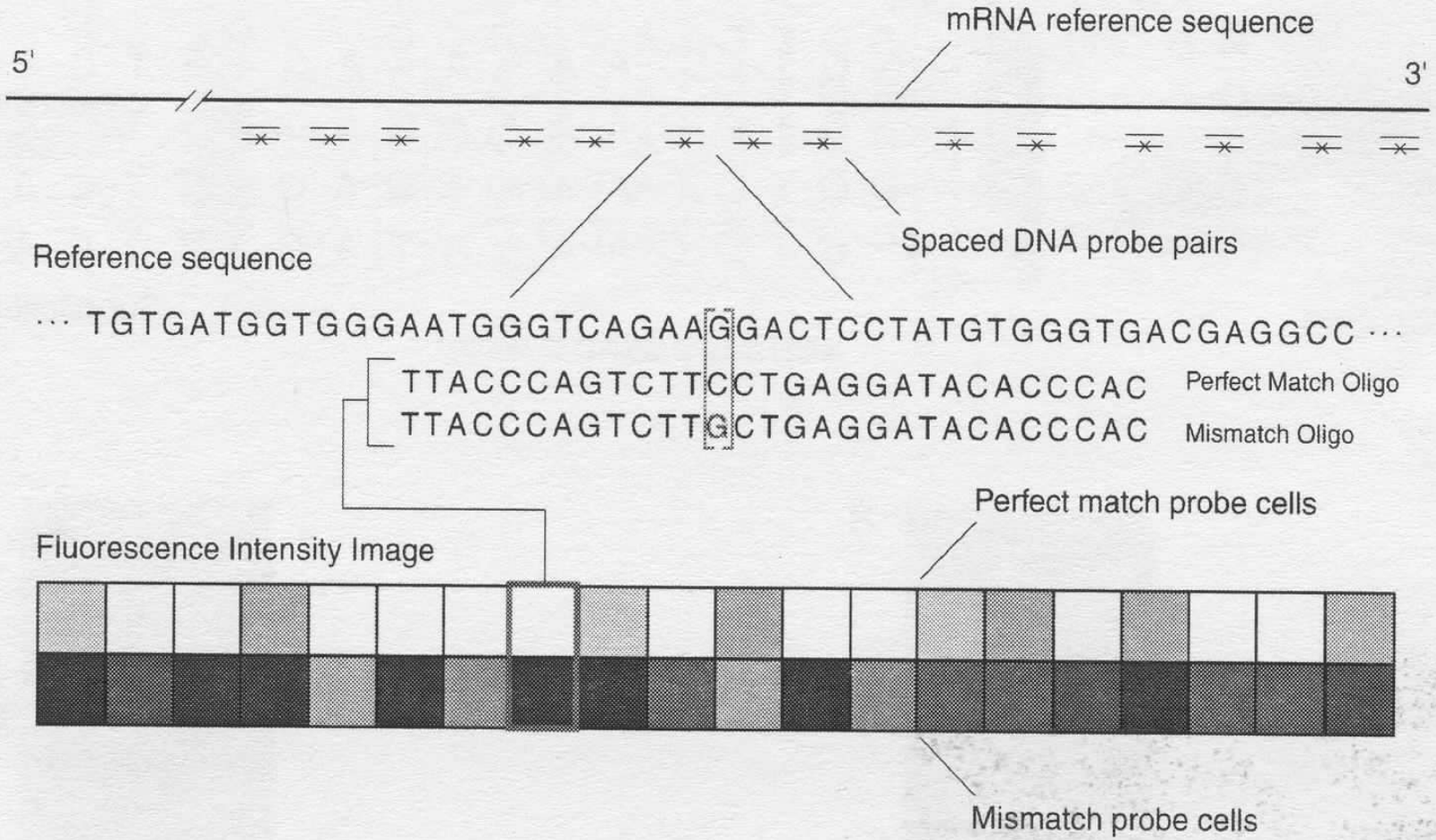
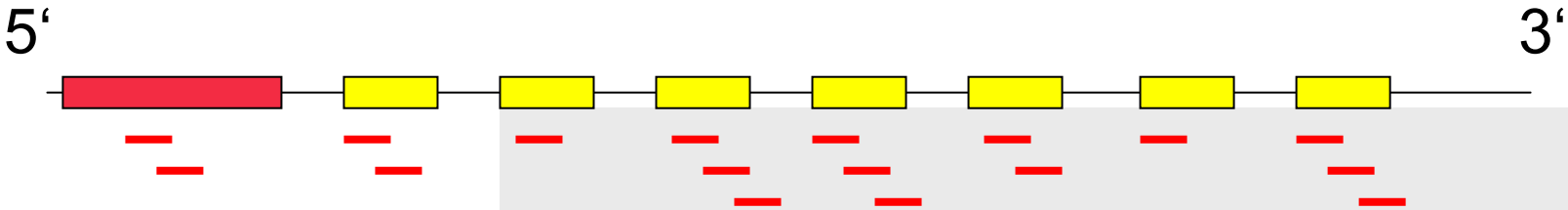


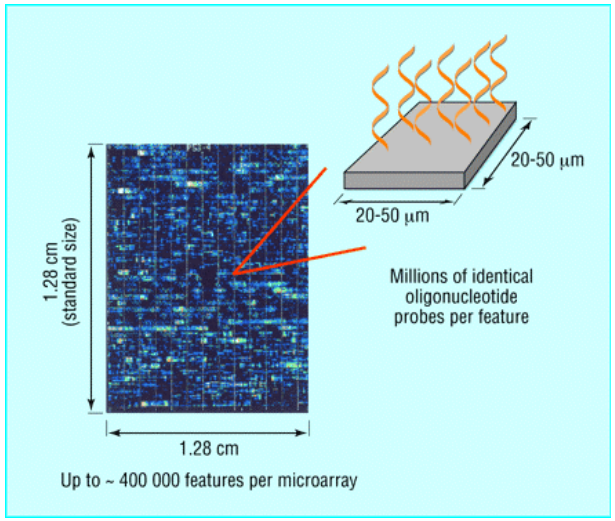
Figure 1-3 Expression tiling strategy

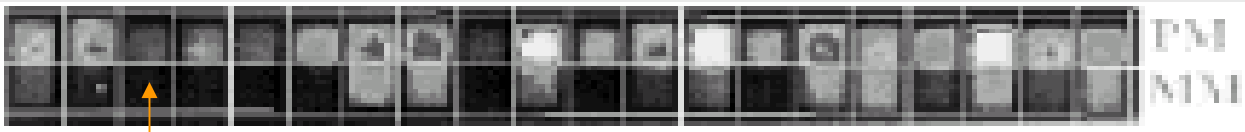
Affymetrix technology



16-20 *probe pairs* per gene

PM: ATGAGCTGTACCAATGCCAACCTGG
MM: ATGAGCTGTACCTATGCCAACCTGG





64 pixels; Signal intensity is upper quartile of the 36 inner pixels

Stored in CEL file

16-20 probe pairs: HG-U95a
11 probe pairs: HG-U133

Affymetrix expression measures

- PM_{ijg}, MM_{ijg} = Intensity for perfect match and mismatch probe j for gene g in chip i.
  - i = 1,…, n one to hundreds of chips
  - j = 1,…, J usually 16 or 20 probe pairs
  - g = 1,…, 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.

Low – level -Analysis

- Preprocessing signals: background correction, normalization, PM-adjustment, summarization.
- Normalization on probe or probe set level?
- Which probes / probe sets used for normalization
- How to treat PM and MM levels?

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)

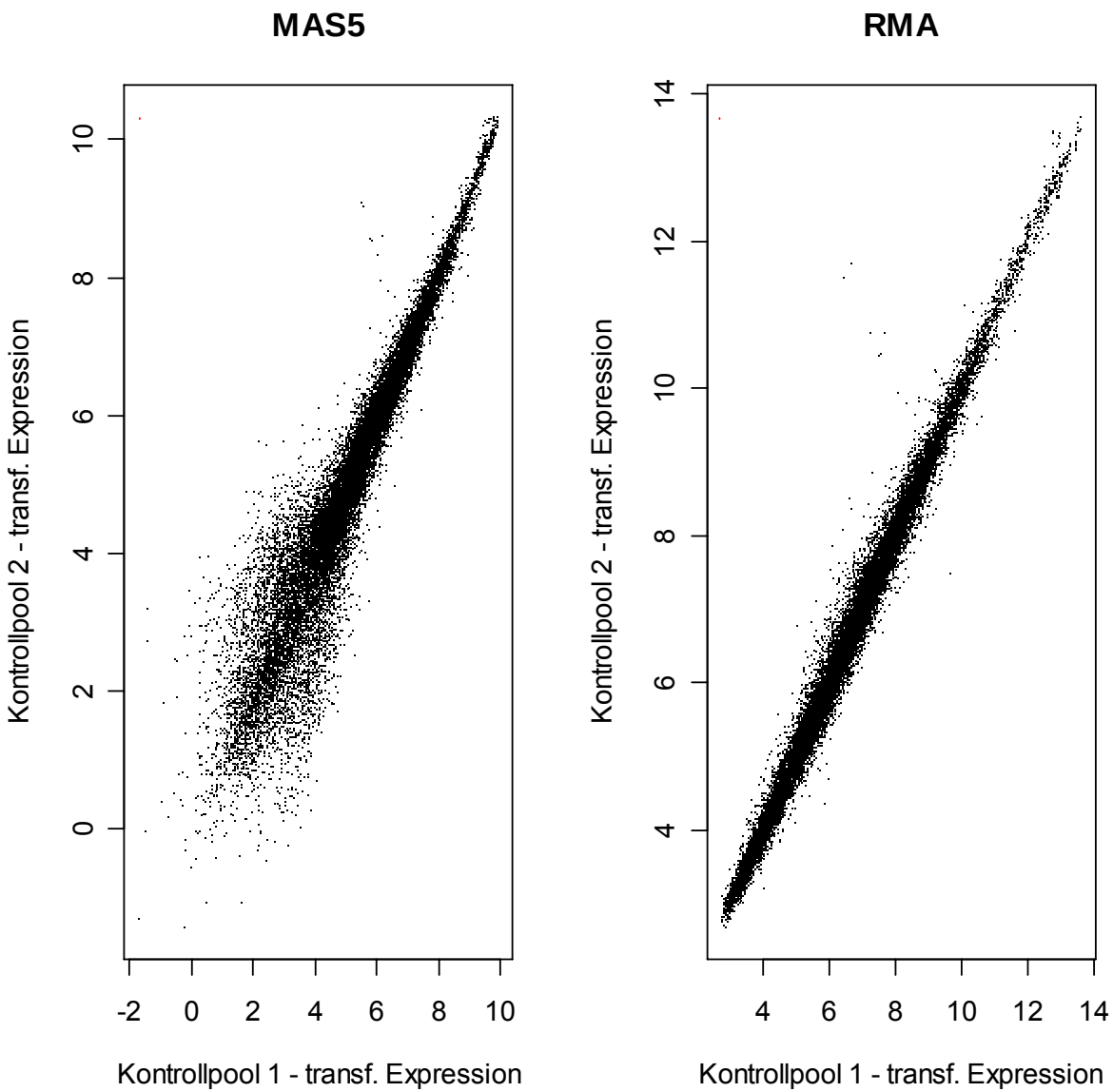
- Estimate one global background value b=mode(MM). No probe-specific background!
- Assume: $PM = s_{true} + b$

  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.

Arguments against the use of d = PM-MM

- Difference is more variable. Is there a gain in bias to compensate for the loss of precision?
- MM detects signal as well as PM
- PM / MM results in a bias.
- Subtraction of MM is not strong enough to remove probe effects, nothing is gained by subtraction

Example LPS: Expression Summaries



How to approach the quantification of gene expression:Three data sets to learn from

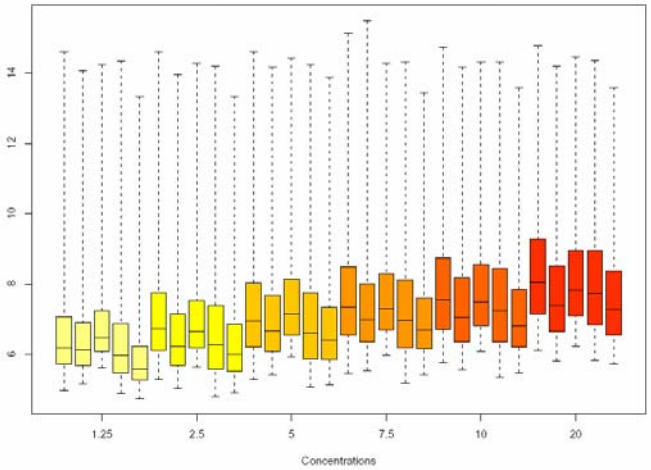
- **Mouse Data Set (A)**
  5 MG-U74A GeneChip® arrays, 20% of the probe pairs were incorrectly sequenced, measurements read for these probes are entirely due to non-specific binding
- **Spike-In Data Set (B)**
  11 control cRNAs were spiked-in at different concentrations
- **Dilution Data Set (C)**
  Human liver tissues were hybridised to HG-U95A in a range of proportions and dilutions.

IMBEI Institut für Medizinische Biometrie, Epidemiologie und Informatik
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

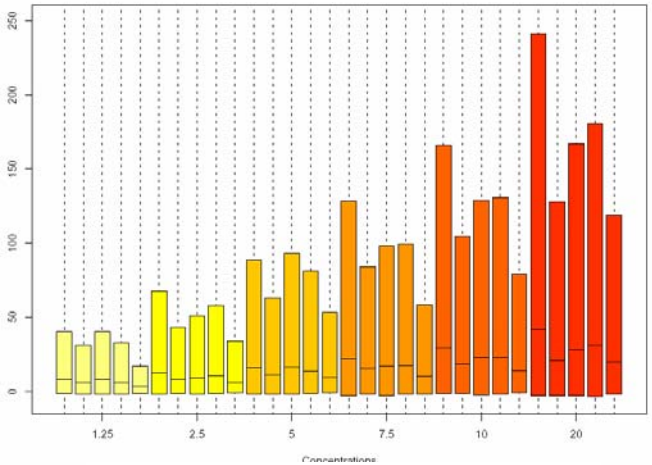
Normalization – Baseline Array

Data C

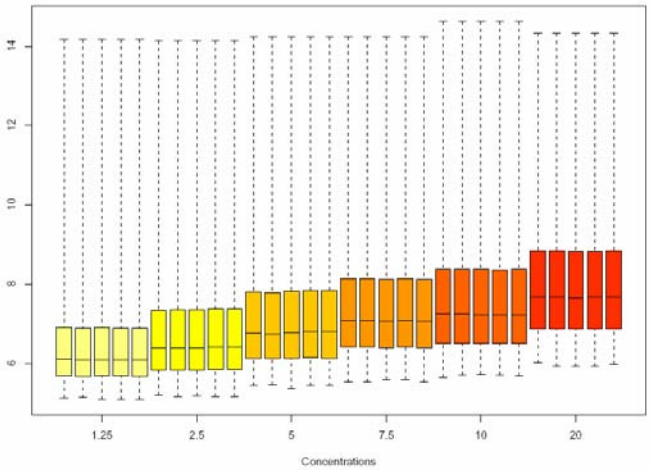
Raw PM



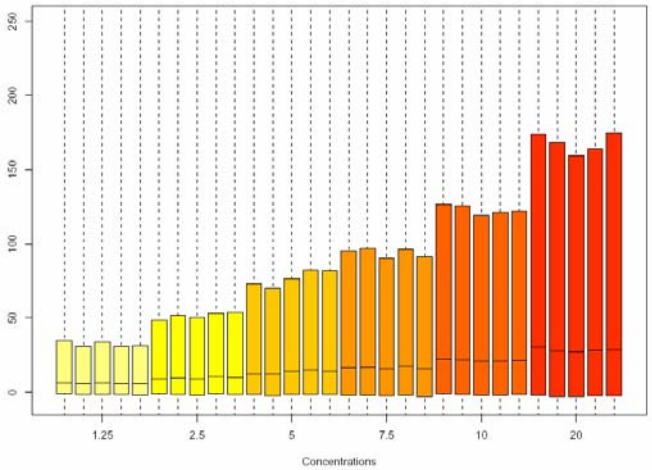
Raw PM-MM



Normalized PM



Normalized PM-MM



AffyComp

- Graphical tool to evaluate summaries of Affymetrix probe level data.
- Plots and summary statistics
- Comparison of competing expression measures
- Selection of methods suitable for a specific investigation
- Use of benchmark data sets

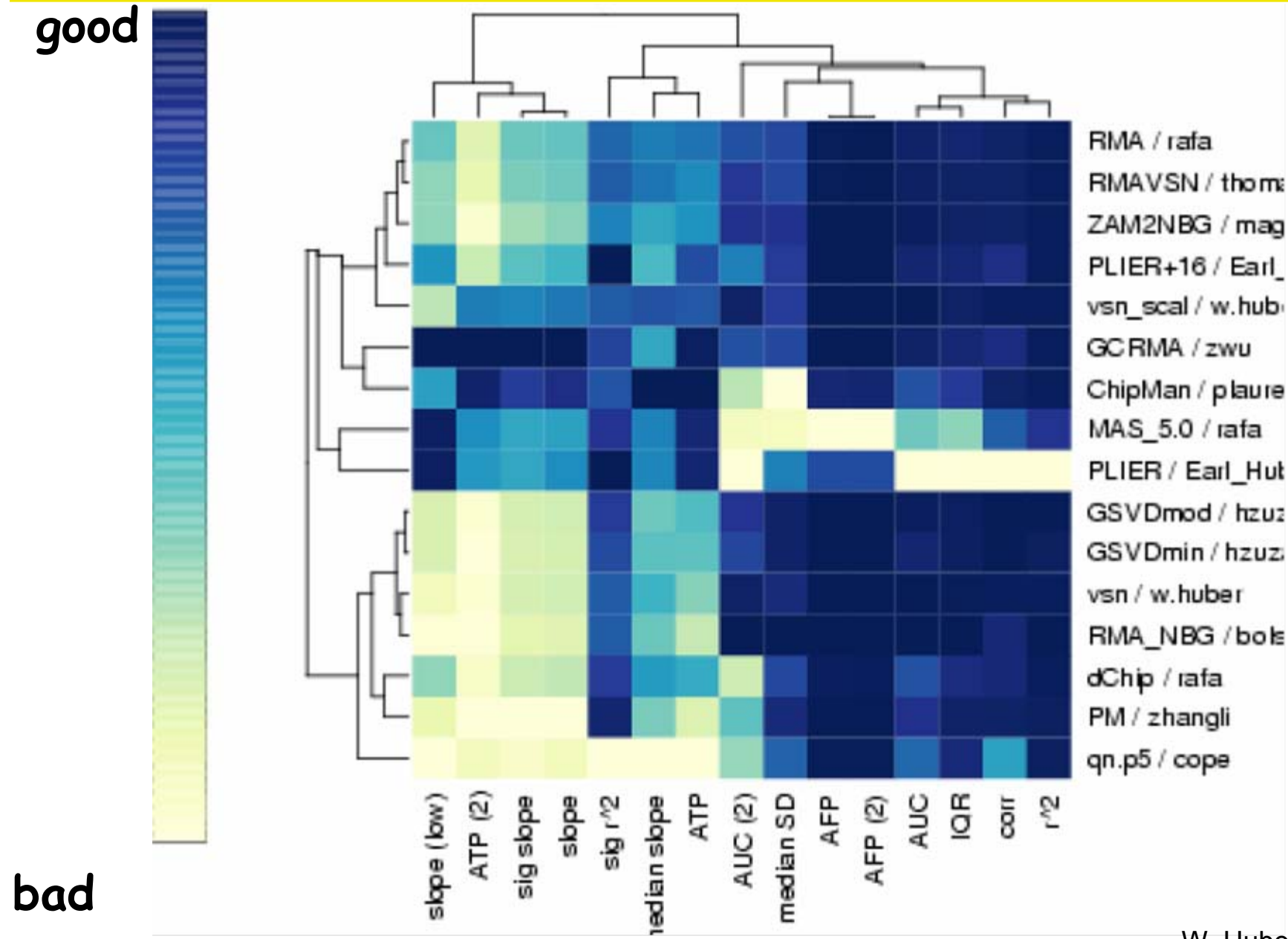
What makes a good expression measure: leads to good and precise answers to a research question.

AffyComp

```
> affycompTable(rma.assessment, mas5.assessment)
```

	RMA	MAS.5.0	whatsgood	Figure
Median SD	0.08811999	2.920239e-01	0	2
R2	0.99420626	8.890008e-01	1	2
1.25v20 corr	0.93645083	7.297434e-01	1	3
2-fold discrepancy	21.00000000	1.226000e+03	0	3
3-fold discrepancy	0.00000000	3.320000e+02	0	3
Signal detect slope	0.62537111	7.058227e-01	1	4a
Signal detect R2	0.80414899	8.565416e-01	1	4a
Median slope	0.86631340	8.474941e-01	1	4b
AUC (FP<100)	0.82066051	3.557341e-01	1	5a
AFP, call if fc>2	15.84156379	3.108992e+03	0	5a
ATP, call if fc>2	11.97942387	1.281893e+01	16	5a
FC=2, AUC (FP<100)	0.54261364	6.508575e-02	1	5b
FC=2, AFP, call if fc>2	1.00000000	3.072179e+03	0	5b
FC=2, ATP, call if fc>2	1.71428571	3.714286e+00	16	5b
IQR	0.30801579	2.655135e+00	0	6
Obs-intended-fc slope	0.61209902	6.932507e-01	1	6a
Obs-(low)int-fc slope	0.35950904	6.471881e-01	1	6b

affycomp results (28 Sep 2003)



W. Huber

Acknowledgements – Slides borrowed from

- Wolfgang Huber
- Ulrich Mansmann
- Terry Speed
- Benedikt Brors
- Anja von Heydebreck
- Tim Beissbarth
- Rainer König