Monday
19. September 05

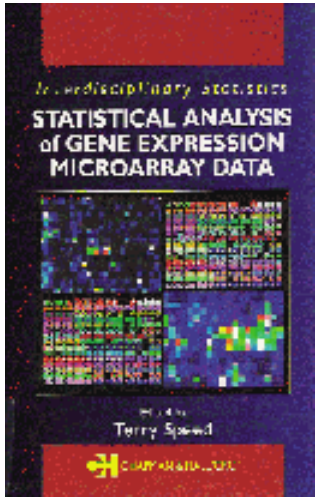
Microarray: Quality Control, Normalization and Design

Tim Beißbarth
Deutsches Krebsforschungszentrum
Molecular Genome Analysis
Bioinformatics

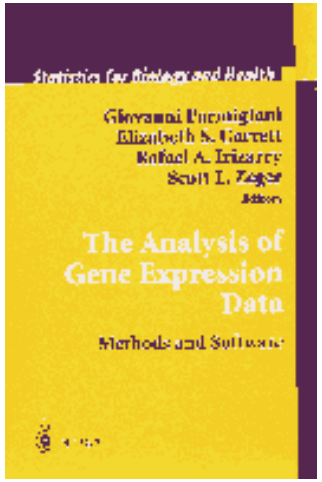
Heidelberg, INF 580
t.beissbarth@dkfz.de

Books

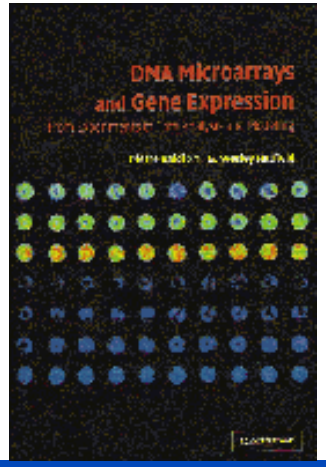
- Terry Speed, „Statistical Analysis of Gene Expression Microarray Data". Chapman & Hall/CRC



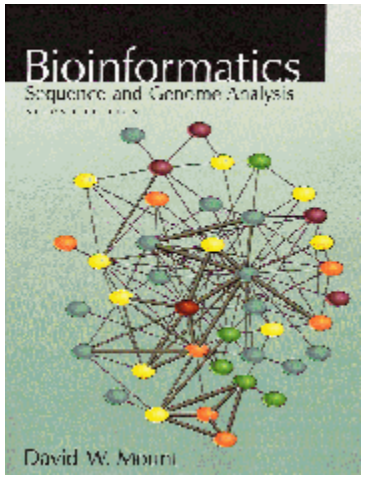
- Giovanni Parmigani et al, „The Analysis of Gene Expression Data", Springer



- Pierre Baldi & G. Wesley Hatfield, „DNA Microarrays and Gene Expression", Cambridge



- David W. Mount, „Bioinformatics", Cold Spring Harbor



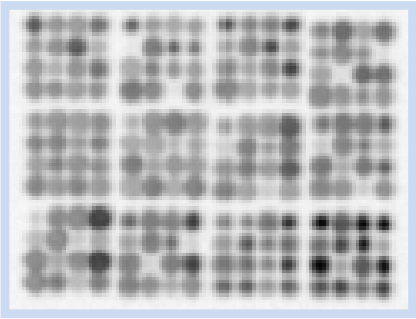
- Reinhard Rauhut, „Bioinformatik", Wiley-VCH

Online

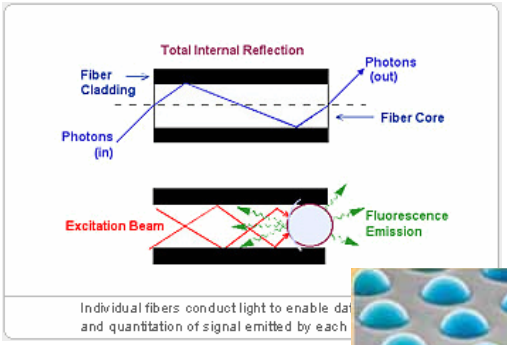
- NGFN Course „Practical DNA Microarray Analysis”: http://compdiag.molgen.mpg.de/lectures.shtml
- Lectures Terry Speed, Berkeley: http://www.stat.berkeley.edu/users/terry/Classes/
- R/Bioconductor Dokumentation (Vignetten): http://www.bioconductor.org
- Google, Pubmed, Wikipedia

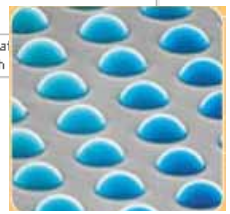
Different Technologies for Measuring Gene Expression

SAGE

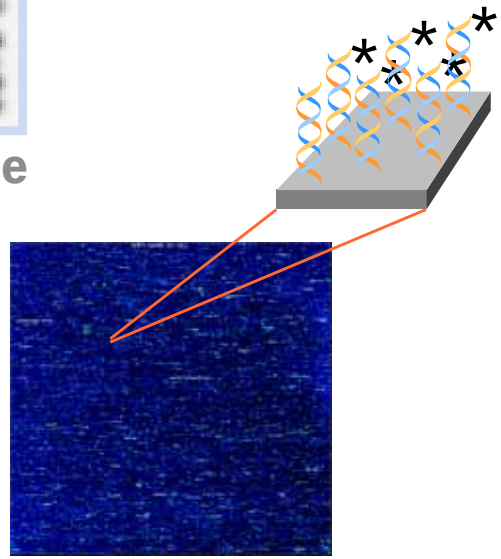


Nylon membrane

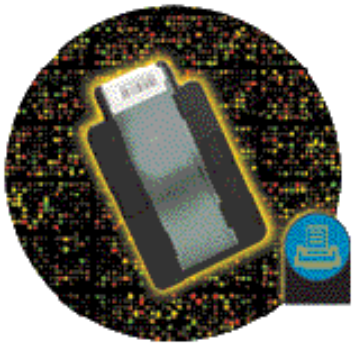




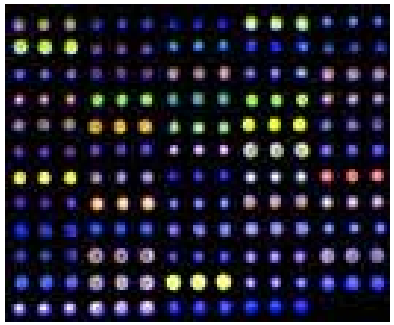
Illumina Bead Array



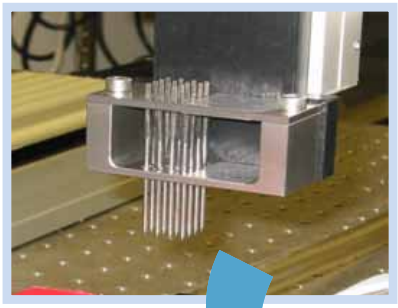
GeneChip Affymetrix



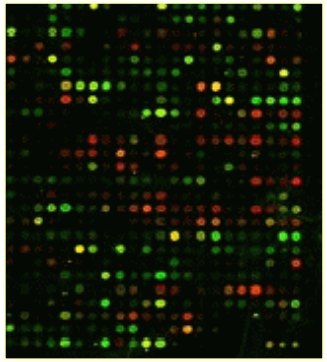
Agilent: Long oligo Ink Jet



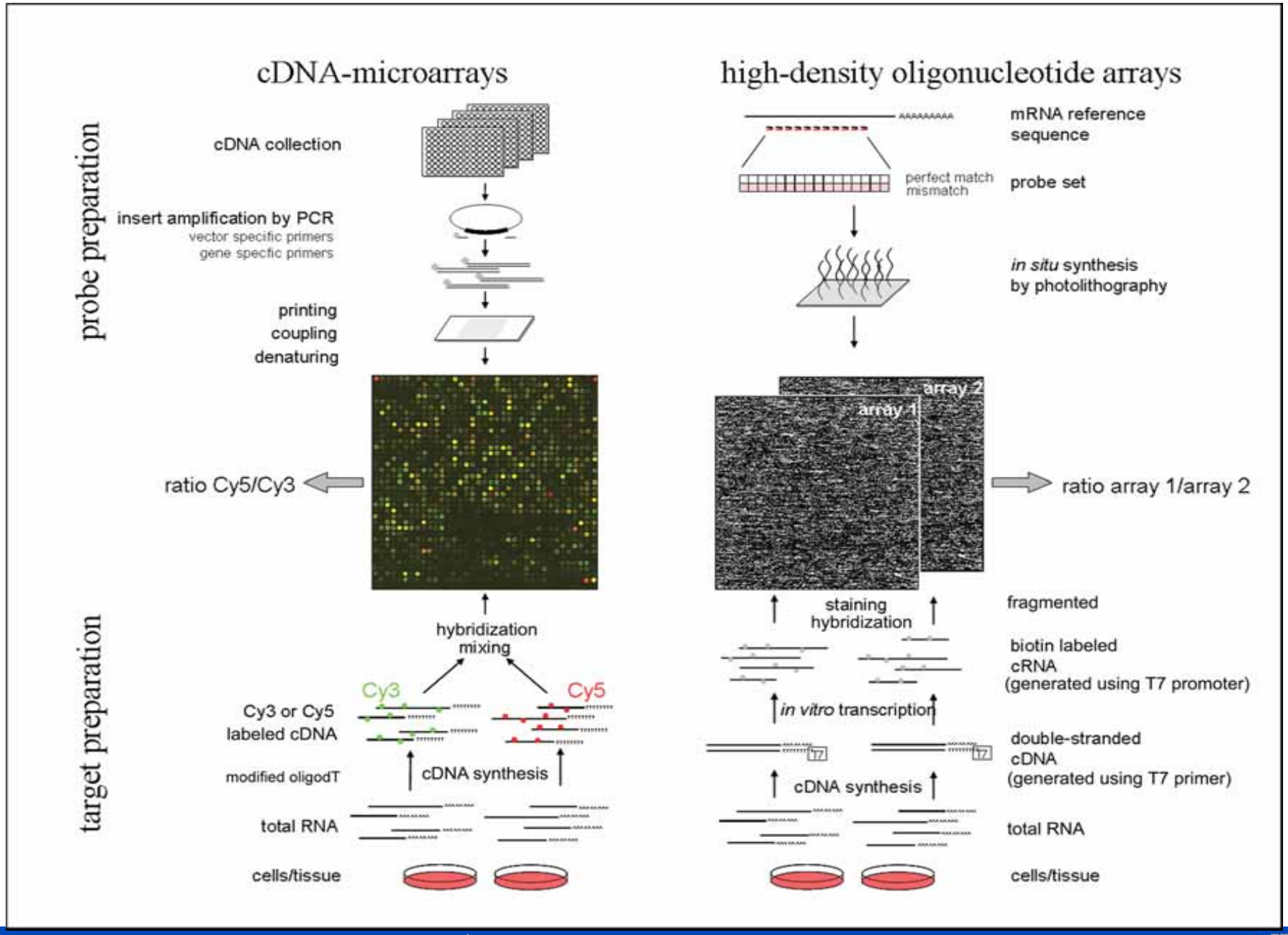
CGH



cDNA microarray

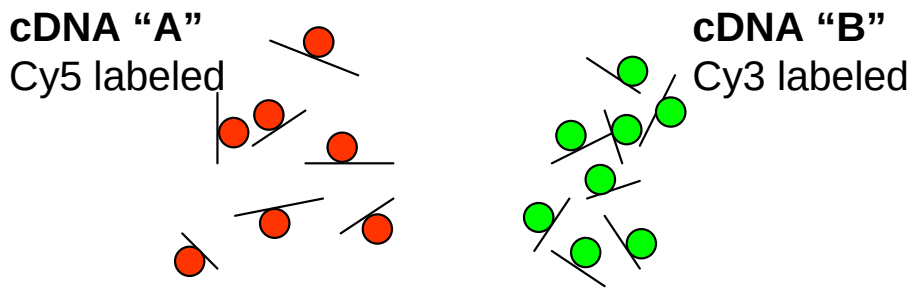


cDNA and Affymetrix (short, 25 bp) Oligo Technologies. Long Oligos (60-75 bp) are used similar to cDNA.

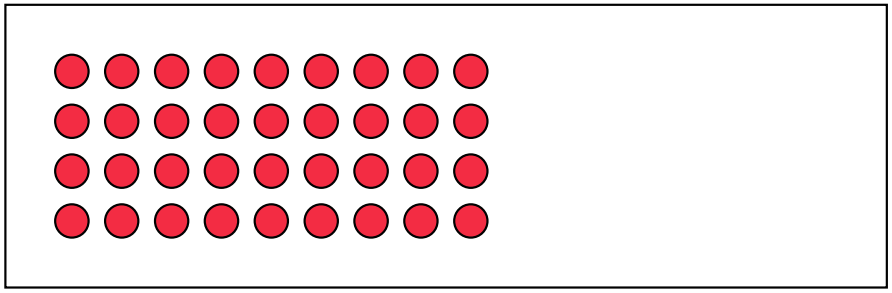


Definition of probe and target

TARGET



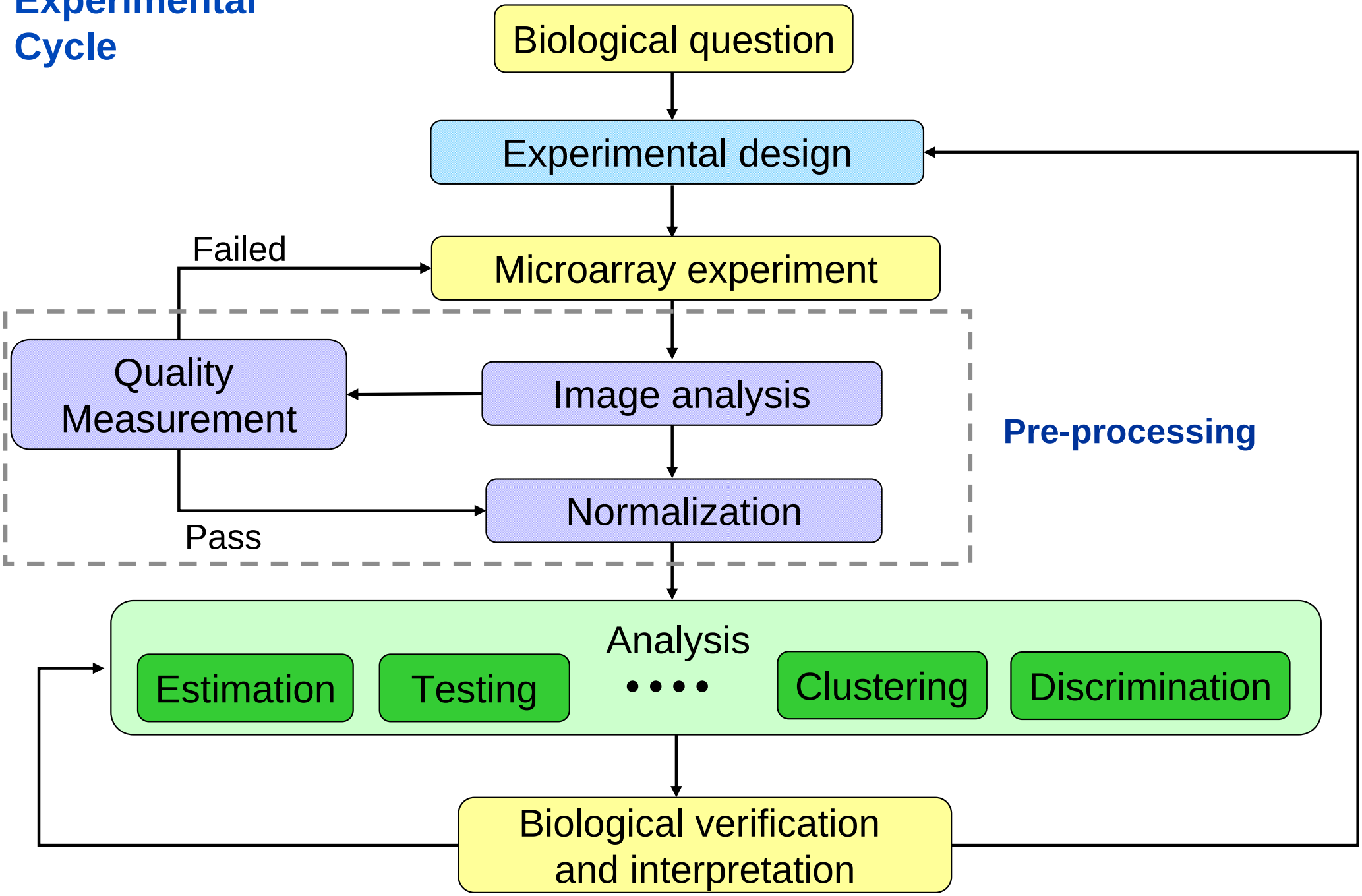
PROBE



Microarrays History

- Based on Southern BIot Technology (Edward Southern, 1975, J. Mol. Biol.)
- 1990: first high-density Nylon filter Arrays (Lennon/Lehrach, 1991, Trends Genet., Review)
- 1995: cDNA-Microarrays described by Schena et al, Science
- 1996: Affymetrix Genechip Technology described by Lockhart et al, Nat. Biotechnol.

Experimental Cycle



Genexpression-Data

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

mRNA Samples

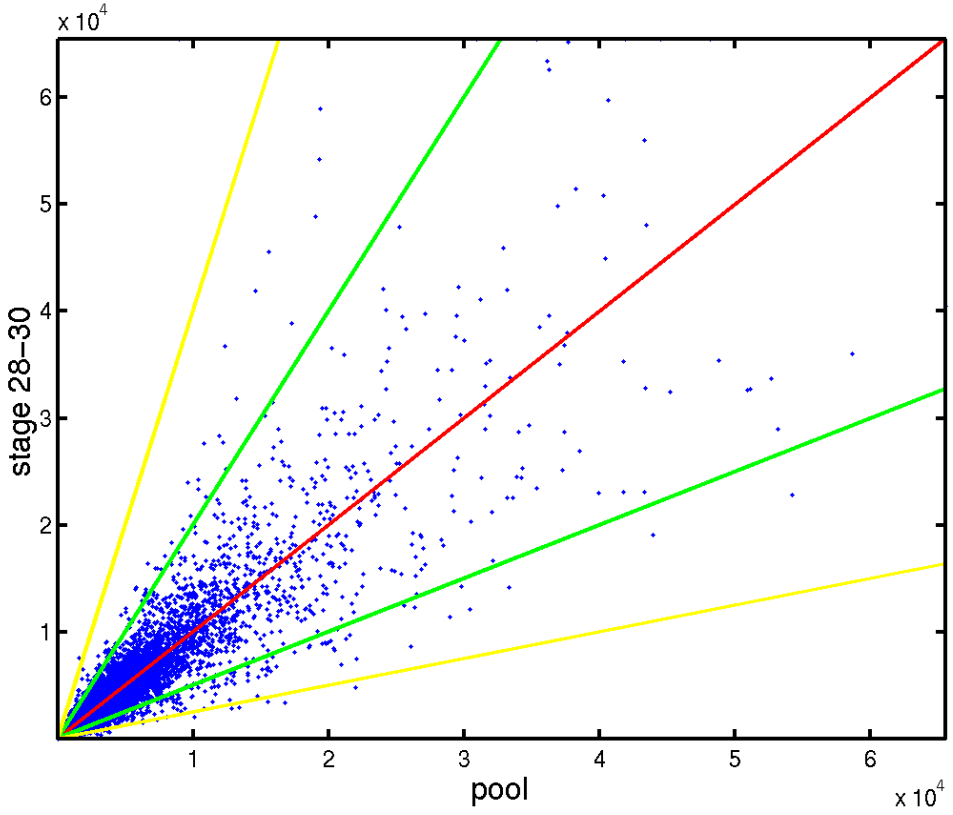
Gene	sample1	sample2	sample3	sample4	sample5	...
1	0.46	0.30	0.80	1.51	0.90	...
2	-0.10	0.49	0.24	0.06	0.46	...
3	0.15	0.74	0.04	0.10	0.20	...
4	-0.45	-1.03	-0.79	-0.56	-0.32	...
5	-0.06	1.06	1.35	1.09	-1.09	...

Genexpression Level for Gen *i* in mRNA sample *j*

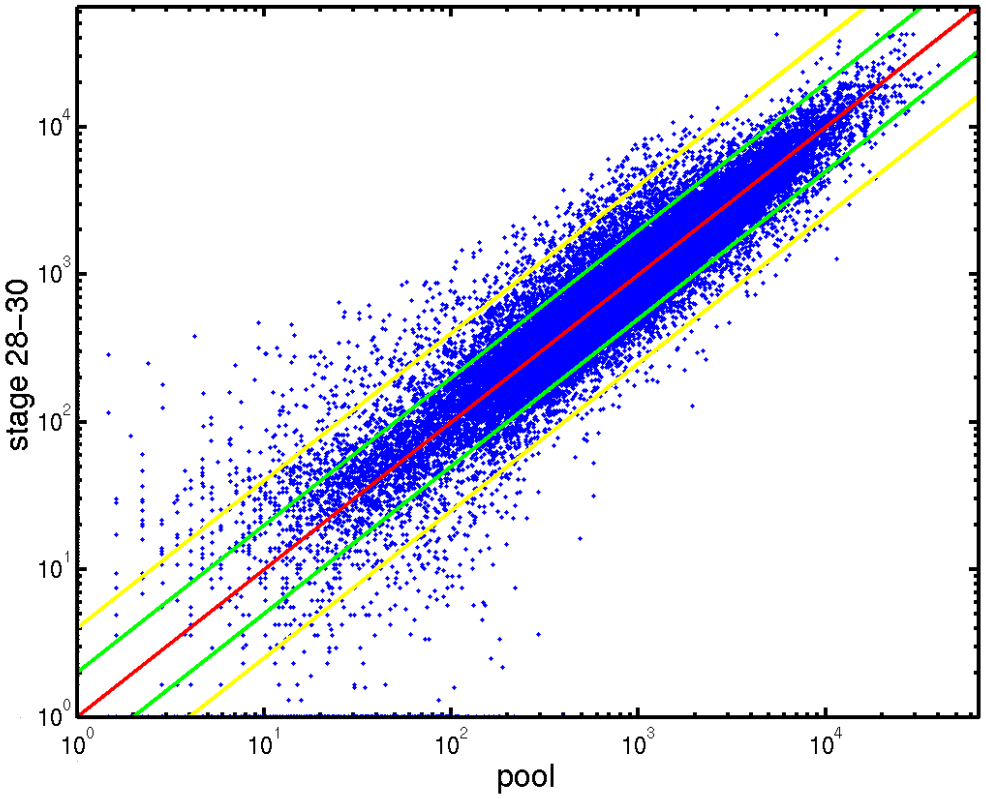
M =
- Log(red intensity / green intensity)
- Function (PM, MM) of MAS, dchip or RMA

Scatterplot

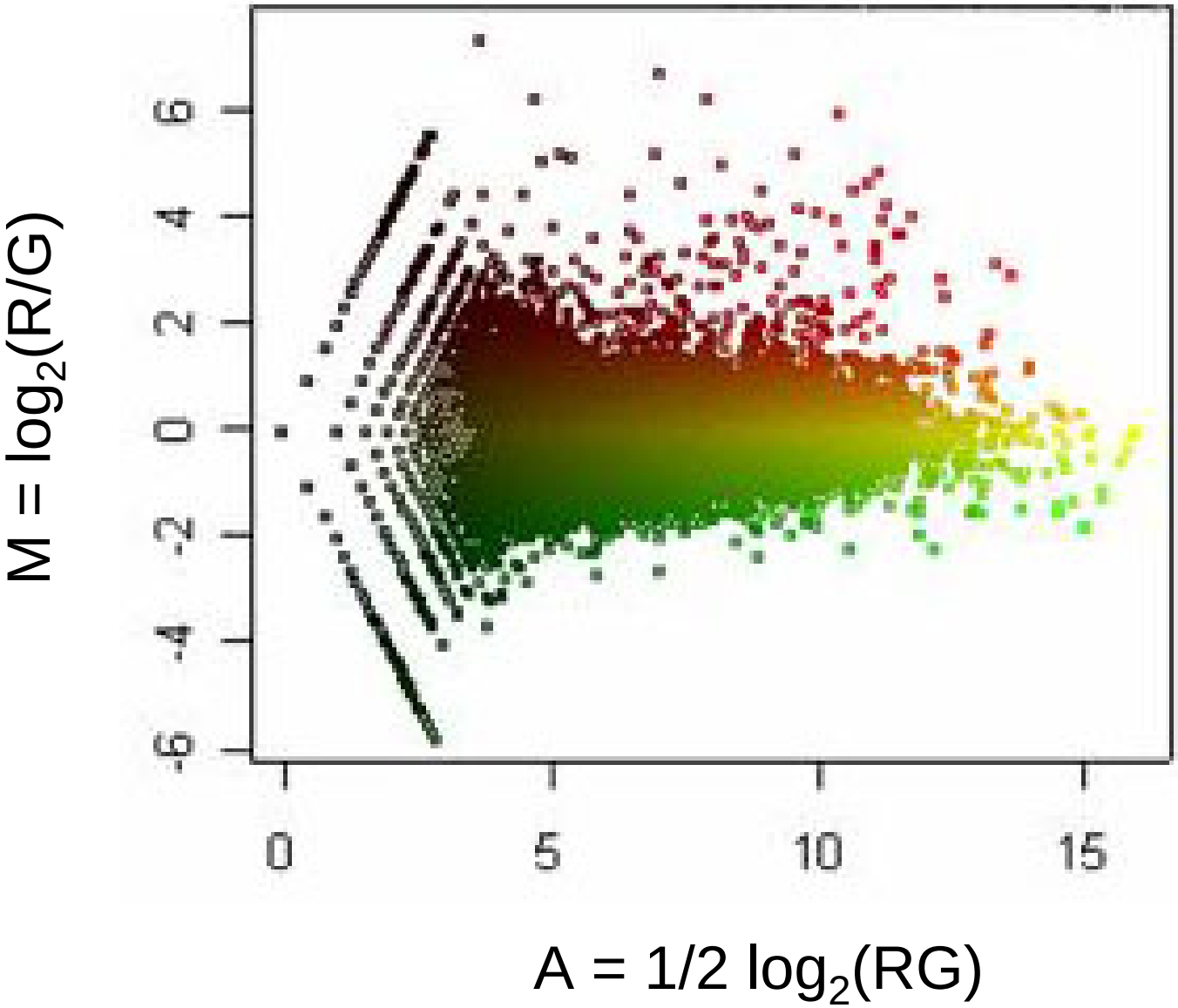
Data



Data (log scale)



MA Plot



Monday
19. September 05

Sources of Variation

- Variance and Bias
- Different Sources of Variation
- Measuring foreground and background signal
- Control quality at different levels

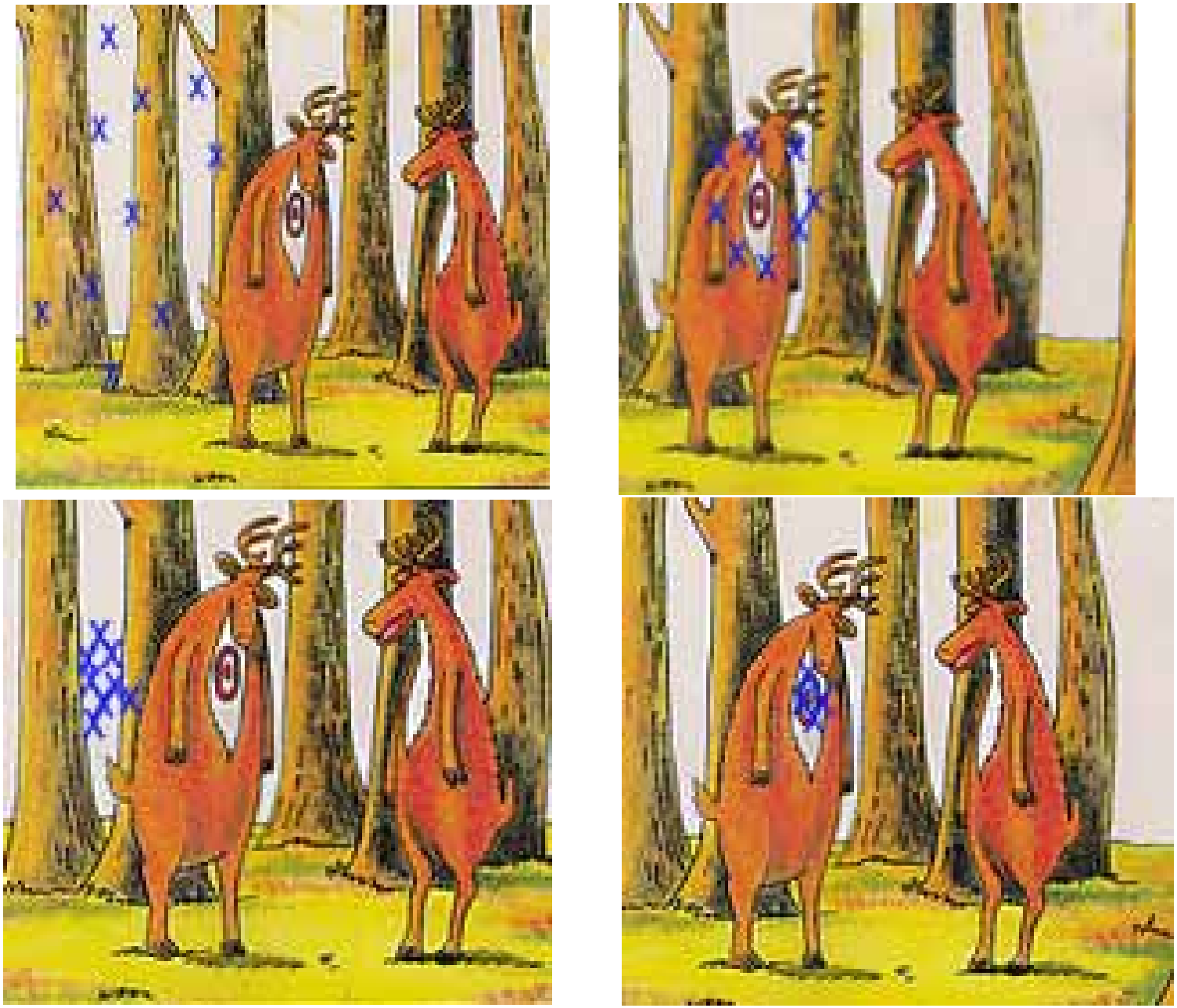
Raw data are not mRNA concentrations

- tissue contamination
- RNA degradation
- amplification efficiency
- reverse transcription efficiency
- Hybridization efficiency and specificity
- clone identification and mapping
- PCR yield, contamination
- spotting efficiency
- DNA support binding
- other array manufacturing related issues
- image segmentation
- signal quantification
- “background” correction

Measurements should be unbiased and precise

high noise

low noise



biased

unbiased

Sources of Variation for Microarray-Data

- amount of RNA in biopsy
- efficiency of:
  - RNA extraction
  - reverse transcription
  - labeling
  - photodetection

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

Systematic

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

Stochastic

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

Normalization

Error model

Quality control: Noise and reliable signal

- Is the signal dominated by noise? Acceptable amount of noise?
- Quantifying noise? (biol. / technol. variability), SNR
- Quantifying quality of a signal;
- Guidelines for reasonable thresholds on the quality of a signal;
- Defining strategies for exclusion of probes:

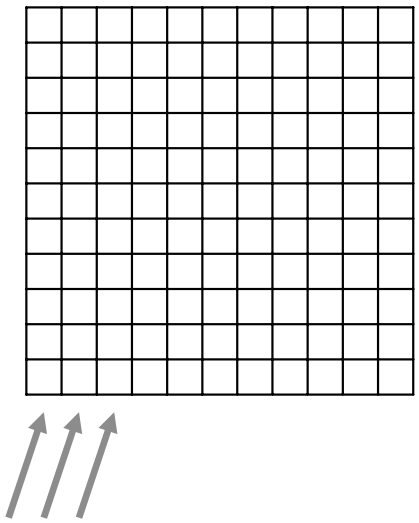
  - Probe level: quality of the expression measurement of one spot on one particular array
  - Array level: quality of the expression measurement on one particular glass slide
  - Gene level: quality of the expression measurement of one probe across all arrays

Probe-level quality control

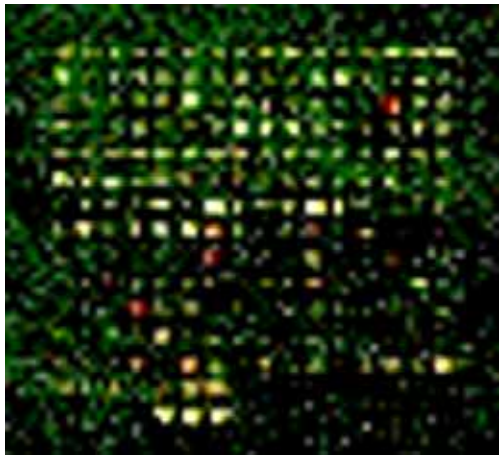
- Individual spots printed on the slide
- Sources:
  - faulty printing, uneven distribution, contamination with debris, magnitude of signal relative to noise, poorly measured spots;
- Visual inspection:
  - hairs, dust, scratches, air bubbles, dark regions, regions with haze
- Spot quality:
  - *Brightness:* foreground/background ratio
  - *Uniformity:* variation in pixel intensities and ratios of intensities within a spot
  - *Morphology:* area, perimeter, circularity.
  - *Spot Size:* number of foreground pixels
- Action:
  - set measurements to NA (missing values)
  - local normalization procedures which account for regional idiosyncrasies.
  - use weights for measurements to indicate reliability in later analysis.

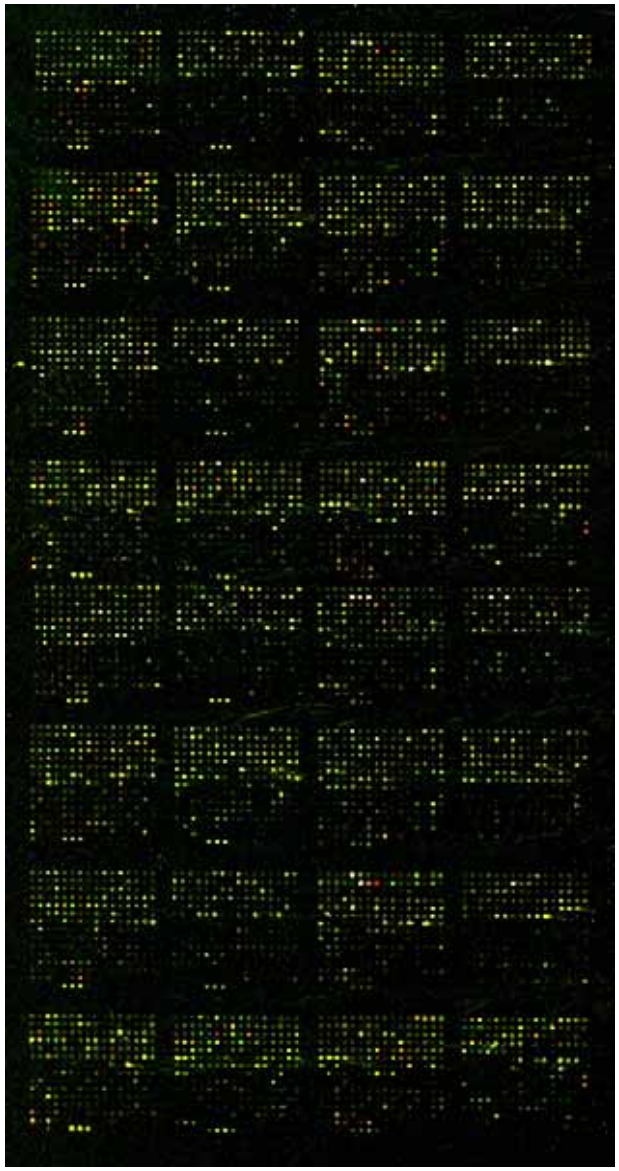
Image Analysis

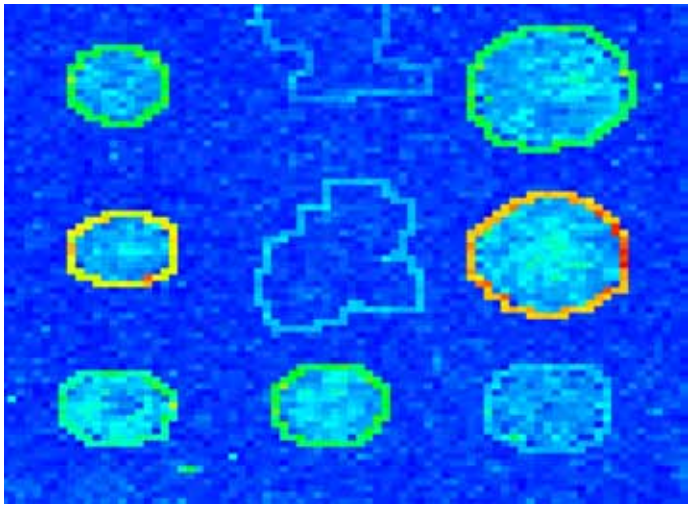
- A grid is overlaid by hand or automatically onto the image.
- Individual spots are recognized, size might be adjusted per spot.
- The signal of the spots is quantified.
- A local background measure might be used.

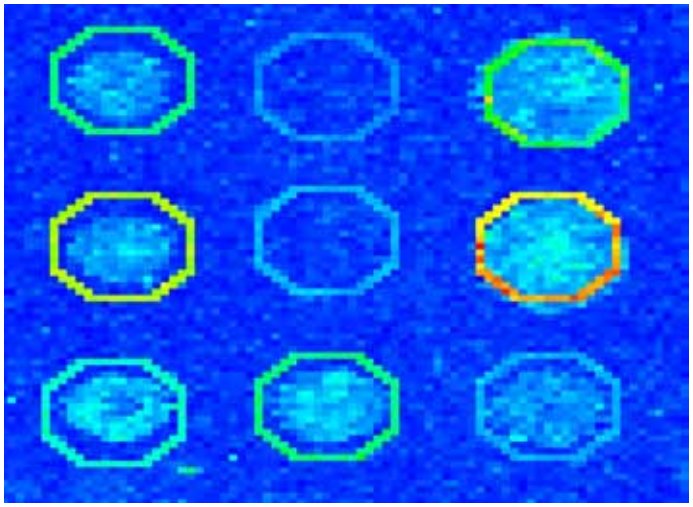


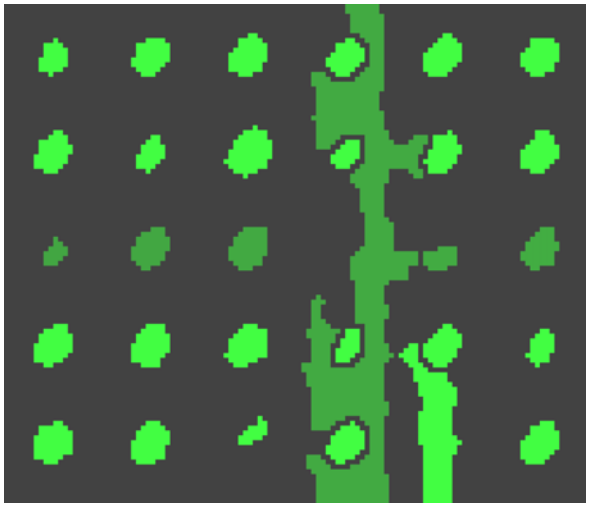
contains Clone-IDs

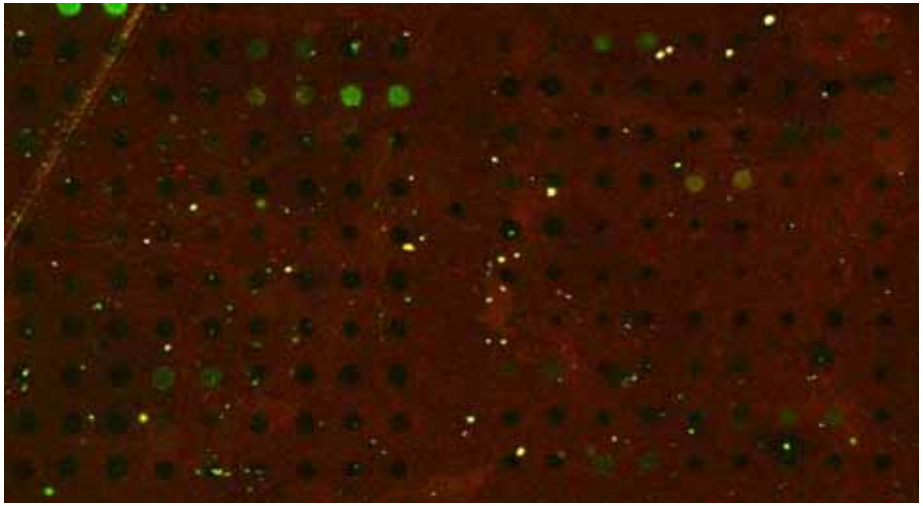










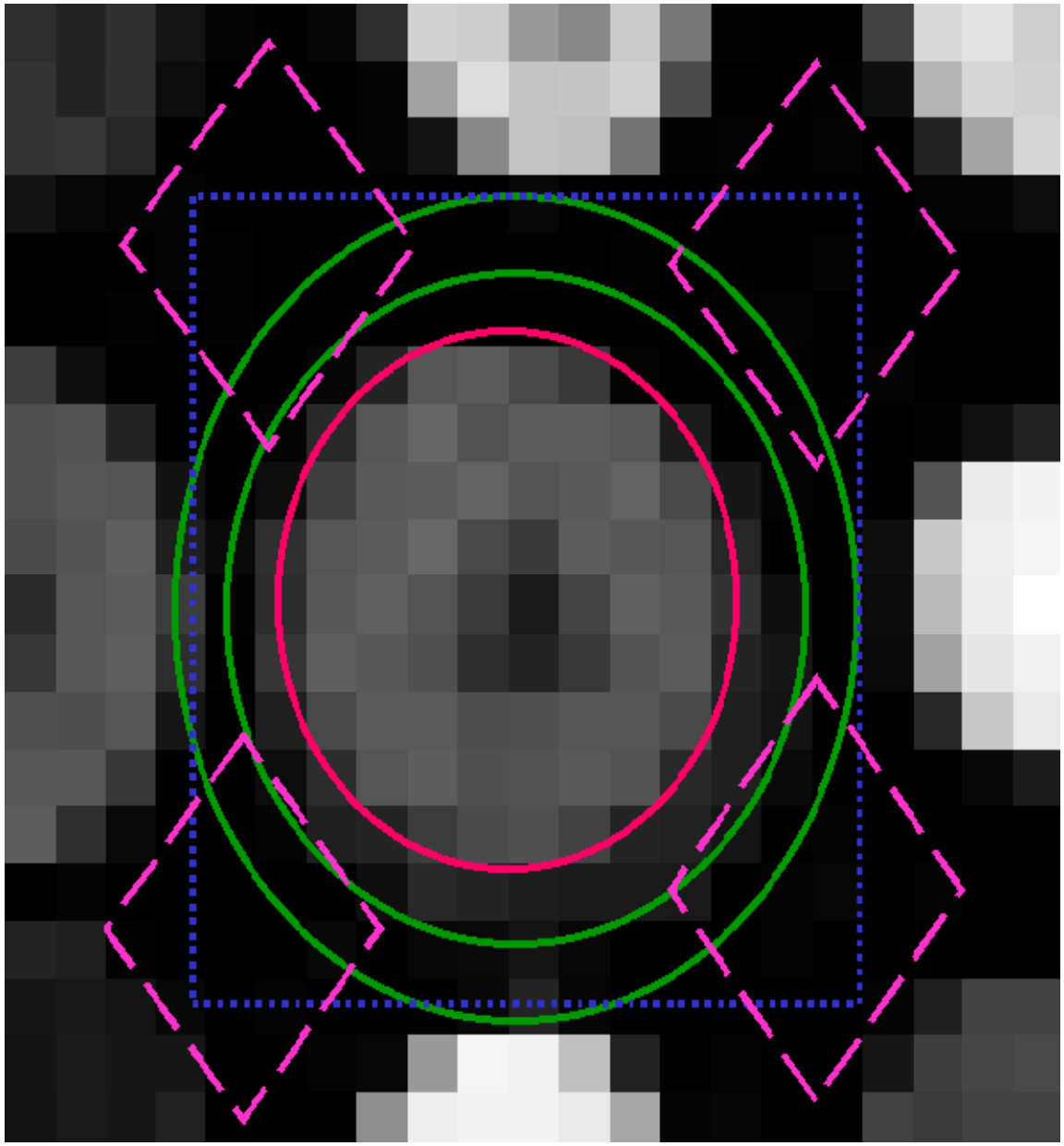


Different Regions around the spot are quantified to measure local background.

GenePix

QuantArray

ScanAlyse

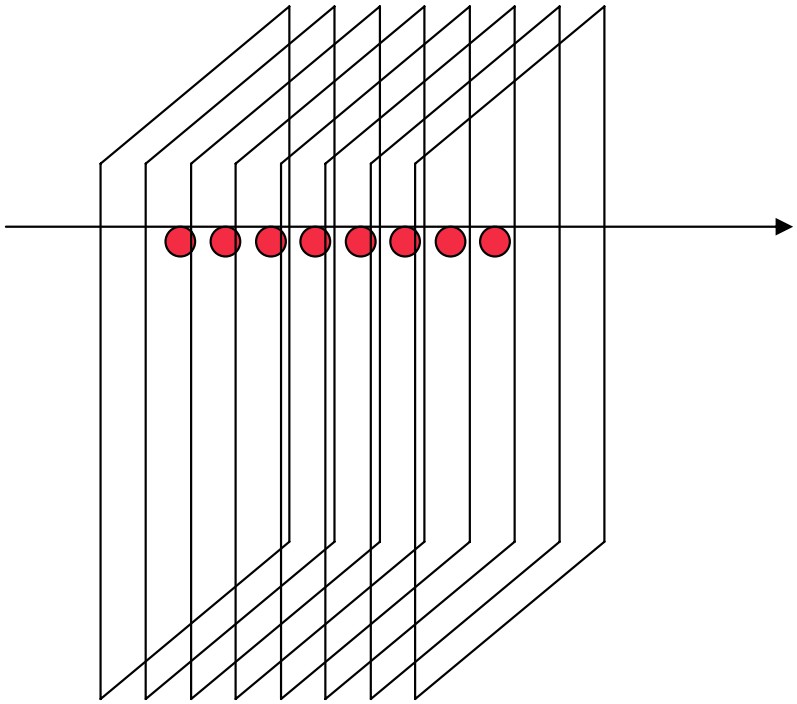


Array-level quality control

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

- Quality measures:
  - Percentage of spots with no signal (~30% exlcuded spots)
  - Range of intensities
  - (Av. Foreground)/(Av. Background) > 3 in both channels
  - Distribution of spot signal area
  - Amount of adjustment needed: signals have to substantially changed to make slides comparable.

Gene-level quality control



Gene g

- Poor hybridization in the reference channel may introduce bias on the fold-change

Gene-level quality control:
Poor Hybridization and Printing

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

**or other appropriate choice*

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

Swirl Data

- Experiment to study early development in zebrafish.
- Swirl mutant vs. wild-type zebrafish.
- 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

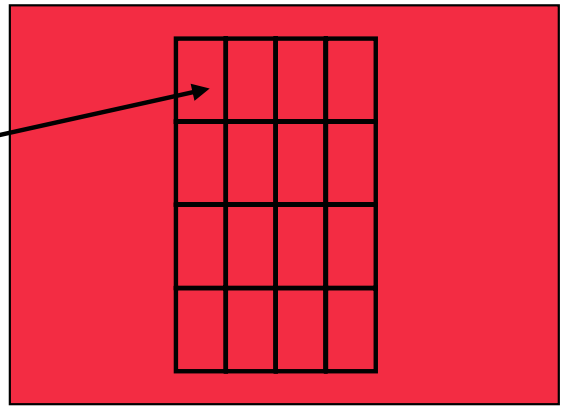
```
> library(marray)
> data(swirl)
> ll()

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

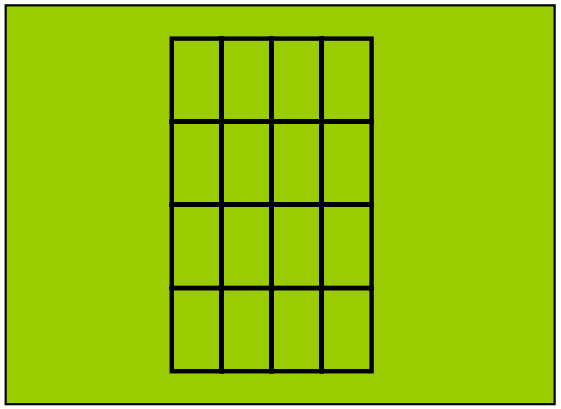
Swirl Data

MT – R
WT - G

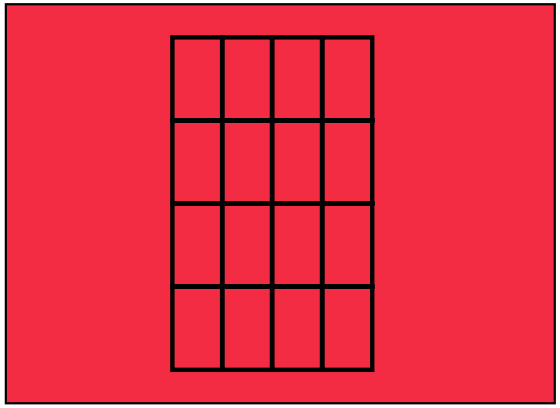
MT – G
WT - R

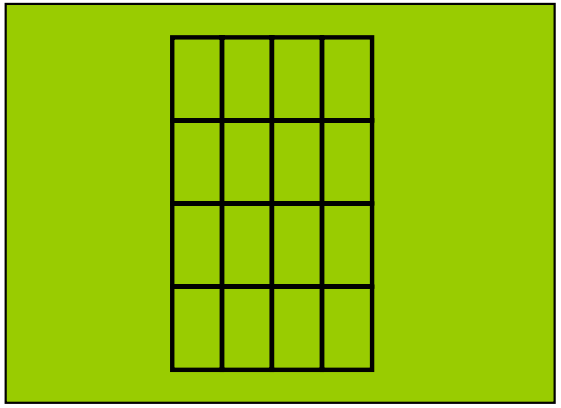
24 x 22
spots per
print-tip





Hybr. I





Hybr. II

technical,
biological
varibility

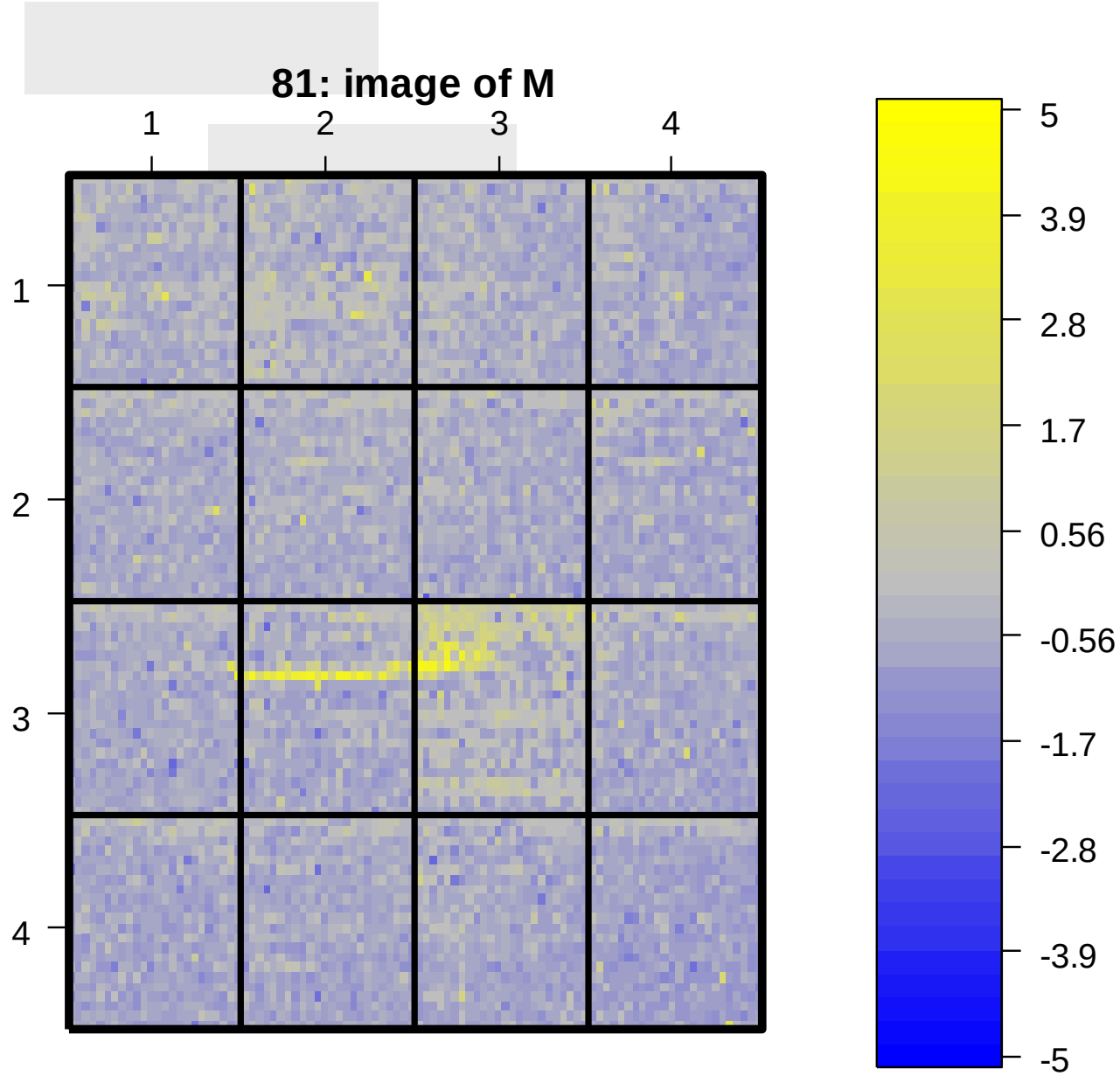
Visual inspection

4 x 4 sectors

Sector:
24 rows
22 columns

8448 spots

Mean signal intensity



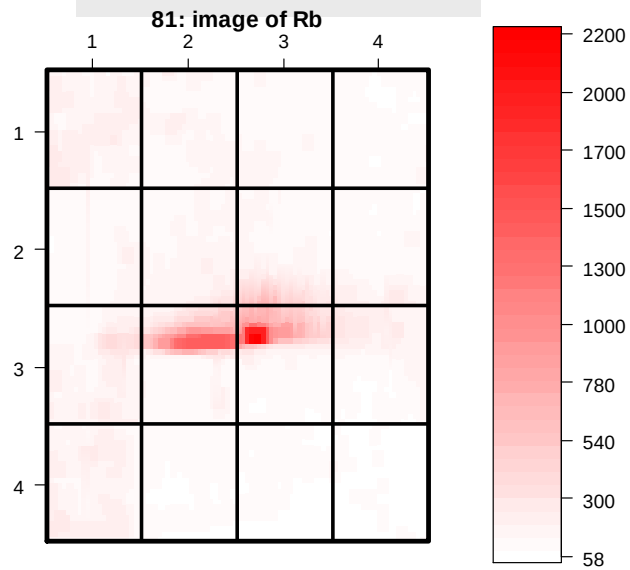
R Console

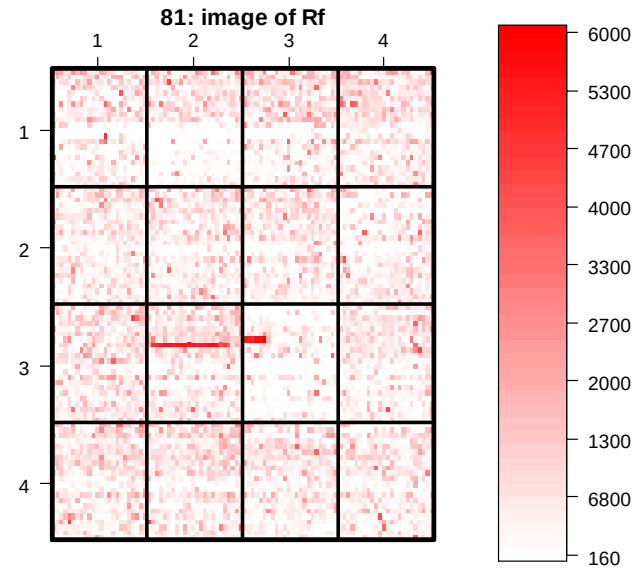
```
> image(swirl[,1])
```

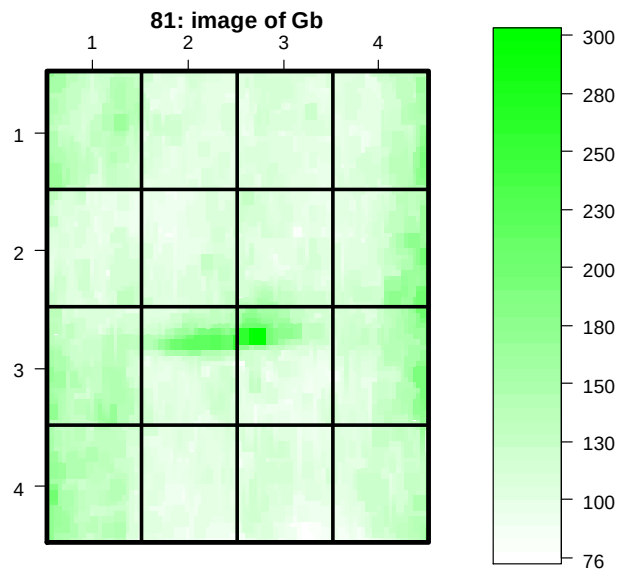
Visual inspection – Foreground and Background intensities

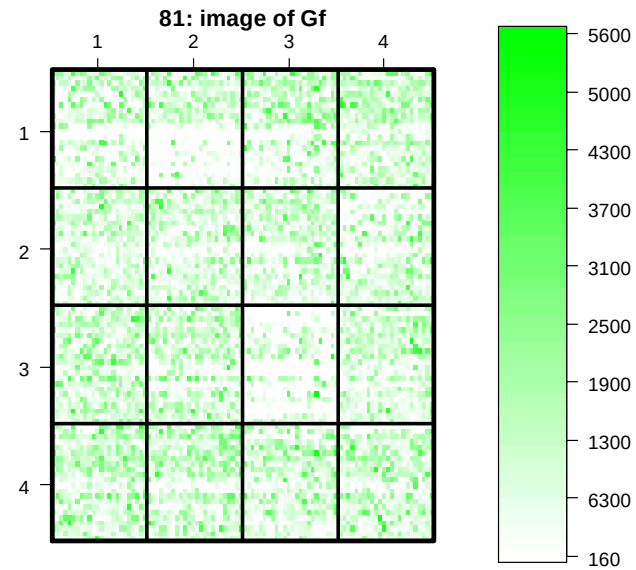
R Console

```
> 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)
```

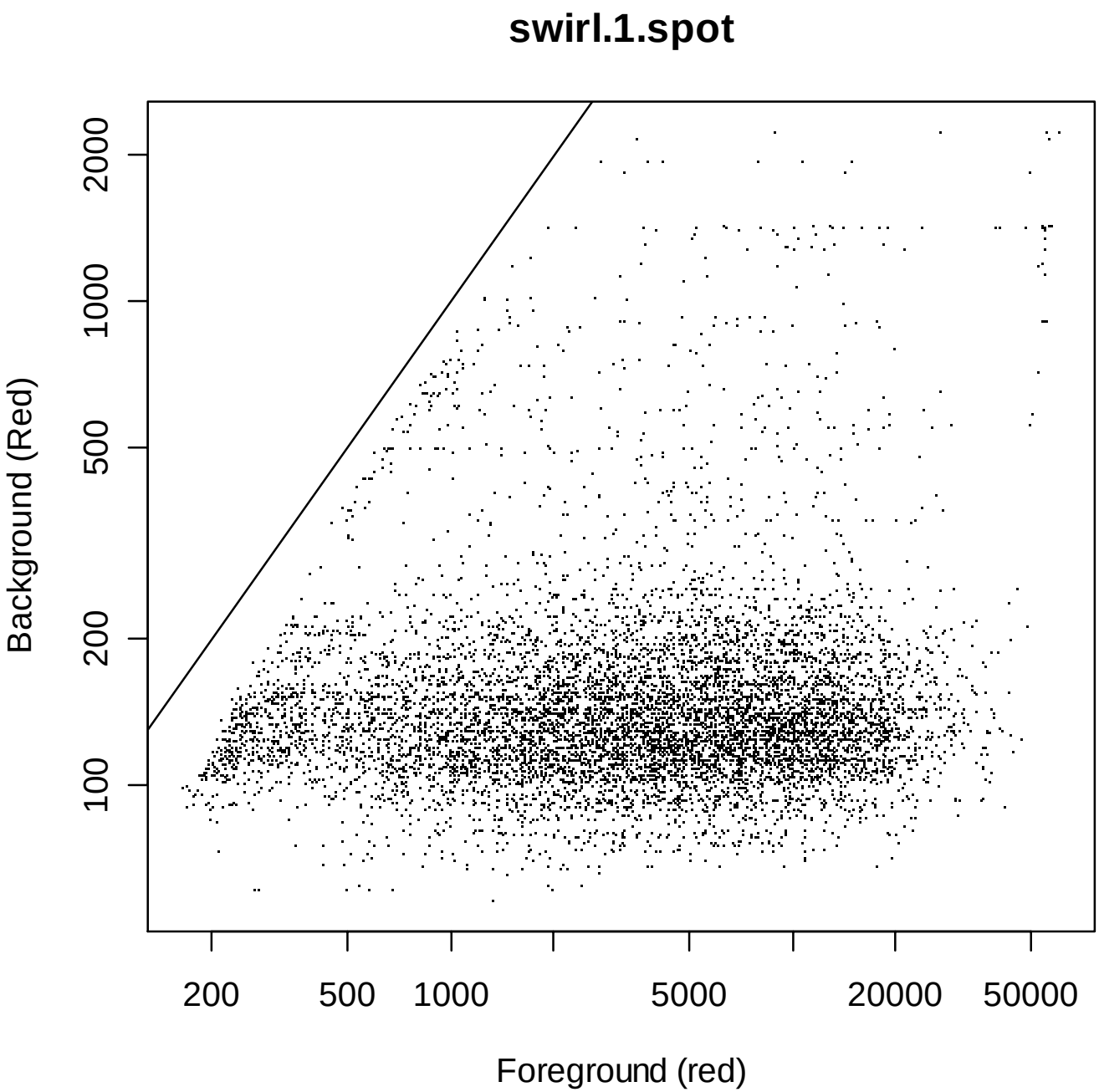








Foreground versus Background intensities



R Console

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

Normalization Methods:

- Scale normalization
- Quantile normalization
- Lowess normalization
- Variance stabilization

Normalization

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

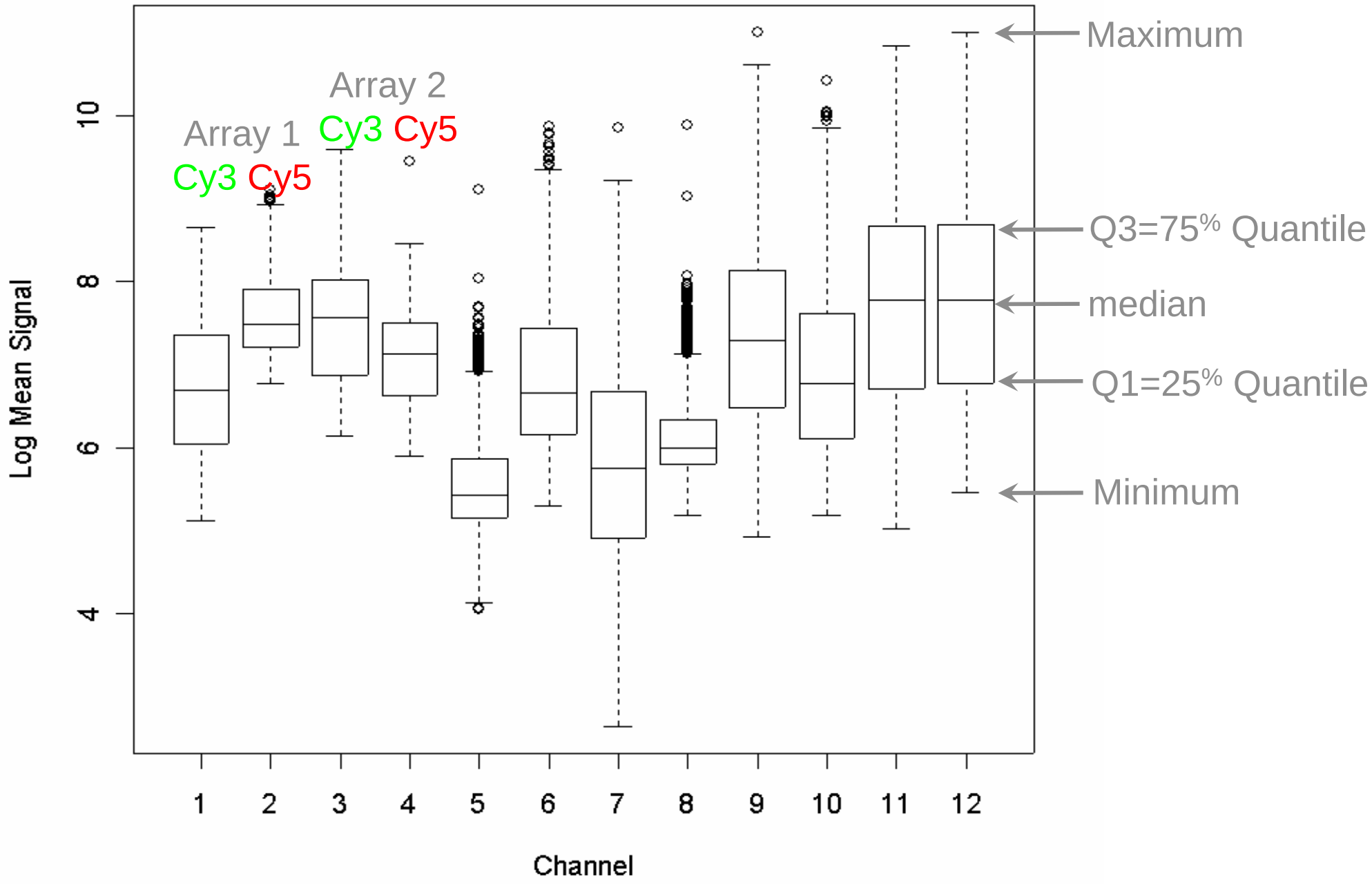
  $$MAD = \text{median}\{ |x_1-m|, \ldots, |x_n-m| \}$$
- Location and scale are basic statistical concepts for data description.
- Normalized log-intensity ratios are given by

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

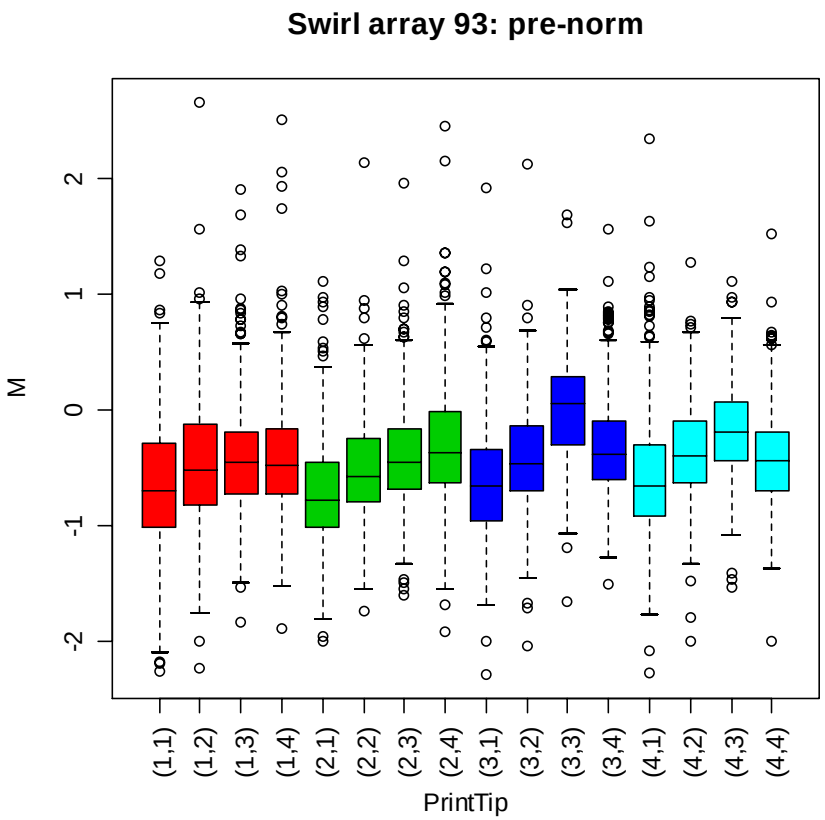
Normalizing the Hybridization-Intensities

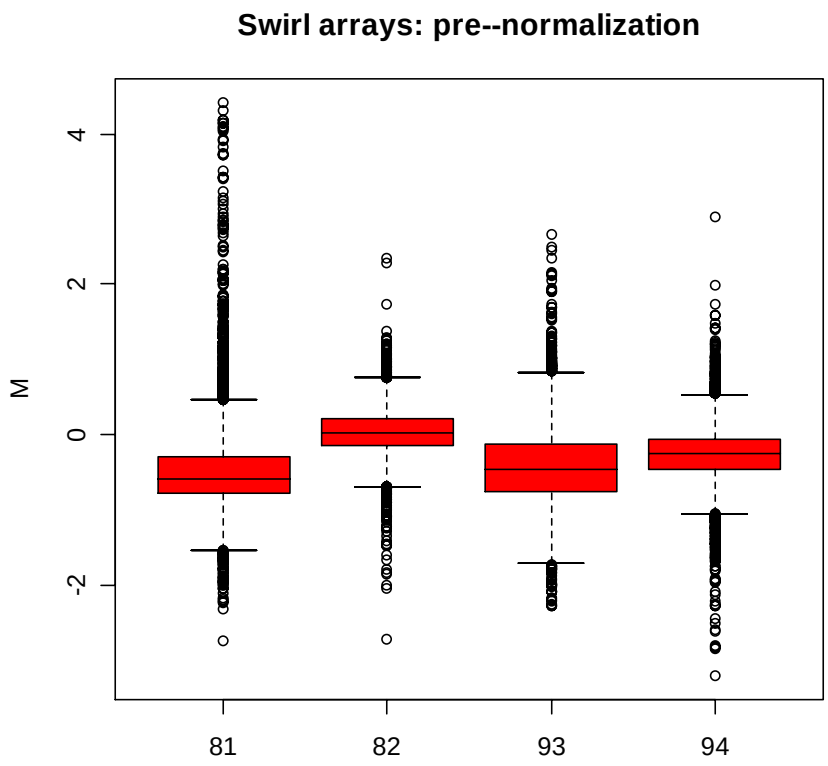
- Background Correction
  - Local Background → Image Analysis
  - Global Background → e.g. 5% Quantile
- Robust estimation of a “rescaling” Factor, e.g. *Median of Differences* based on
  - the majority of genes
  - Housekeeping genes
  - *Spiked* in control genes
- There are many other normalization methods!
  Other methods:
  - Lowess (aka loess)
  - Quantile
  - VSN

Displaying Variability of Microarray-Data



marray – Swirl Data: Pre Normalization

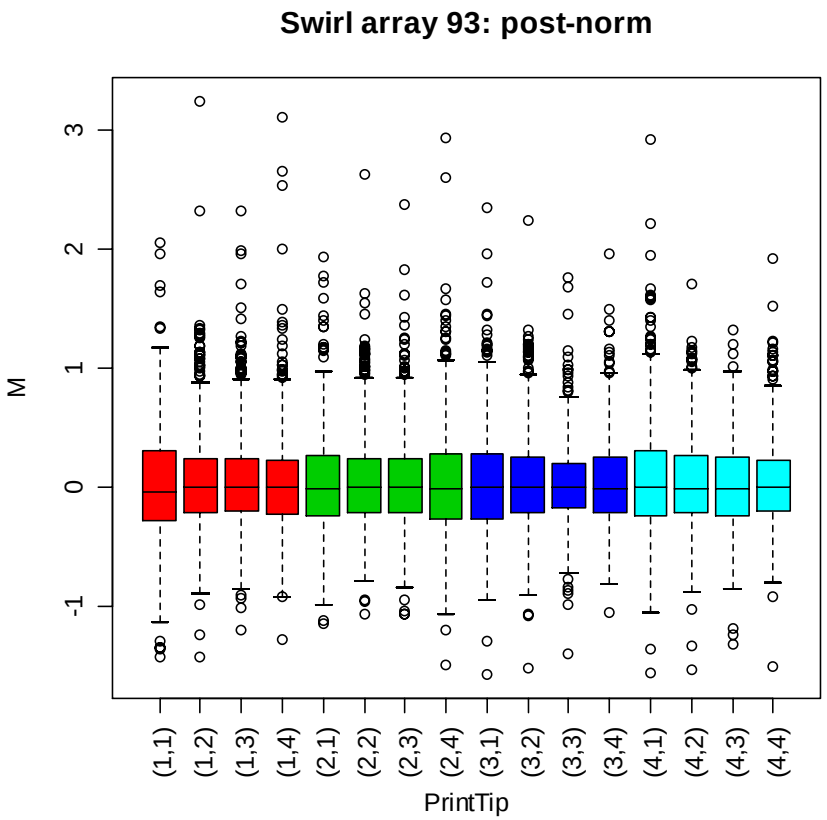


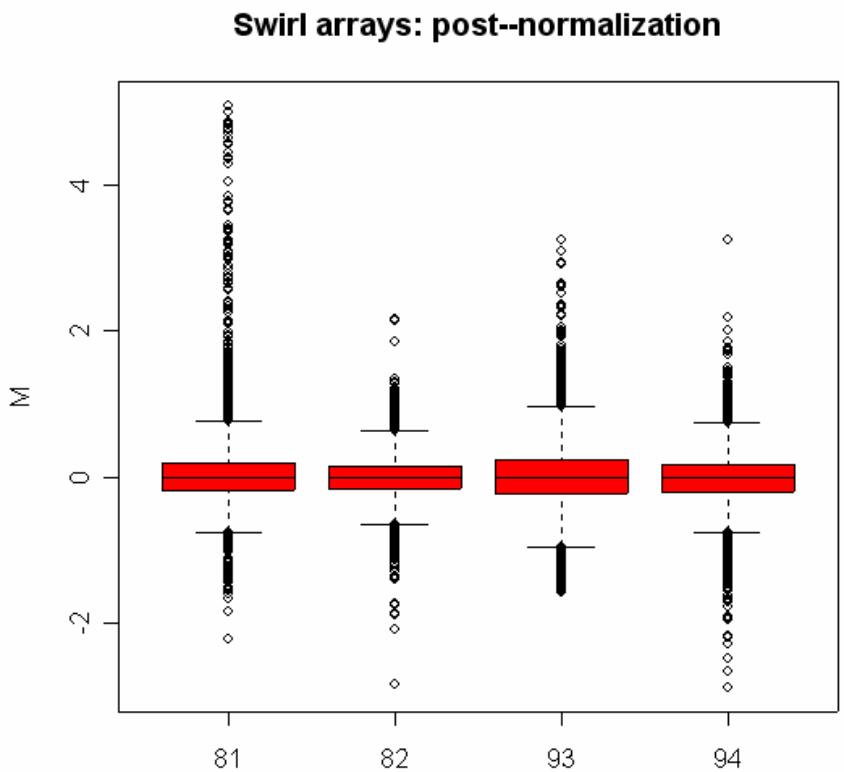


R Console

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

marray – Swirl Data: Post Normalization

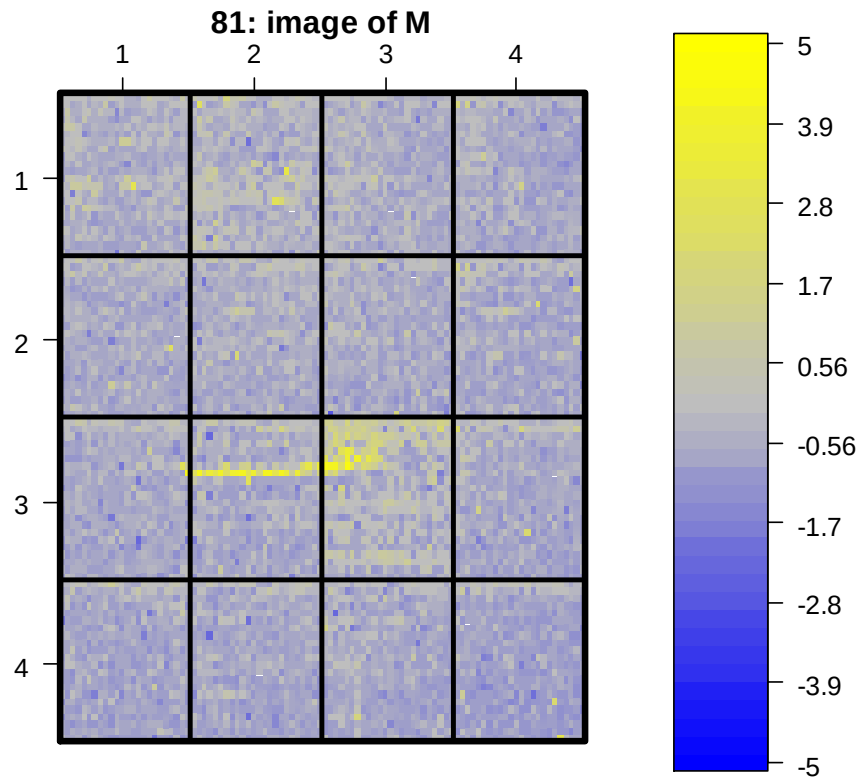


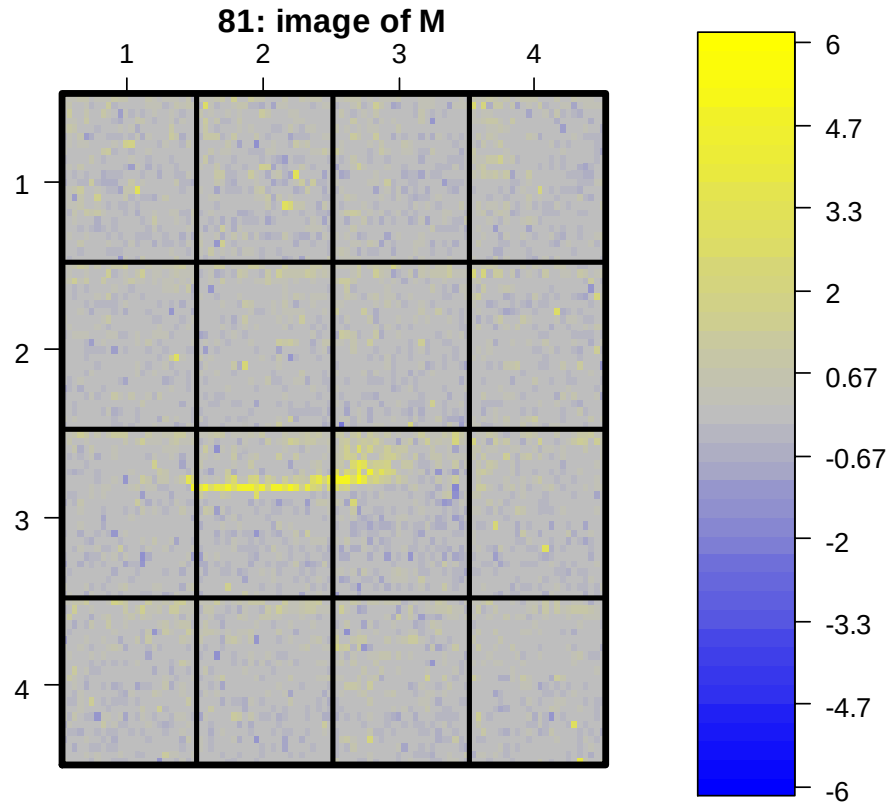


R Console

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

Swirl Data – M values, raw versus normalized





Normalization procedure was not able to remove scratch

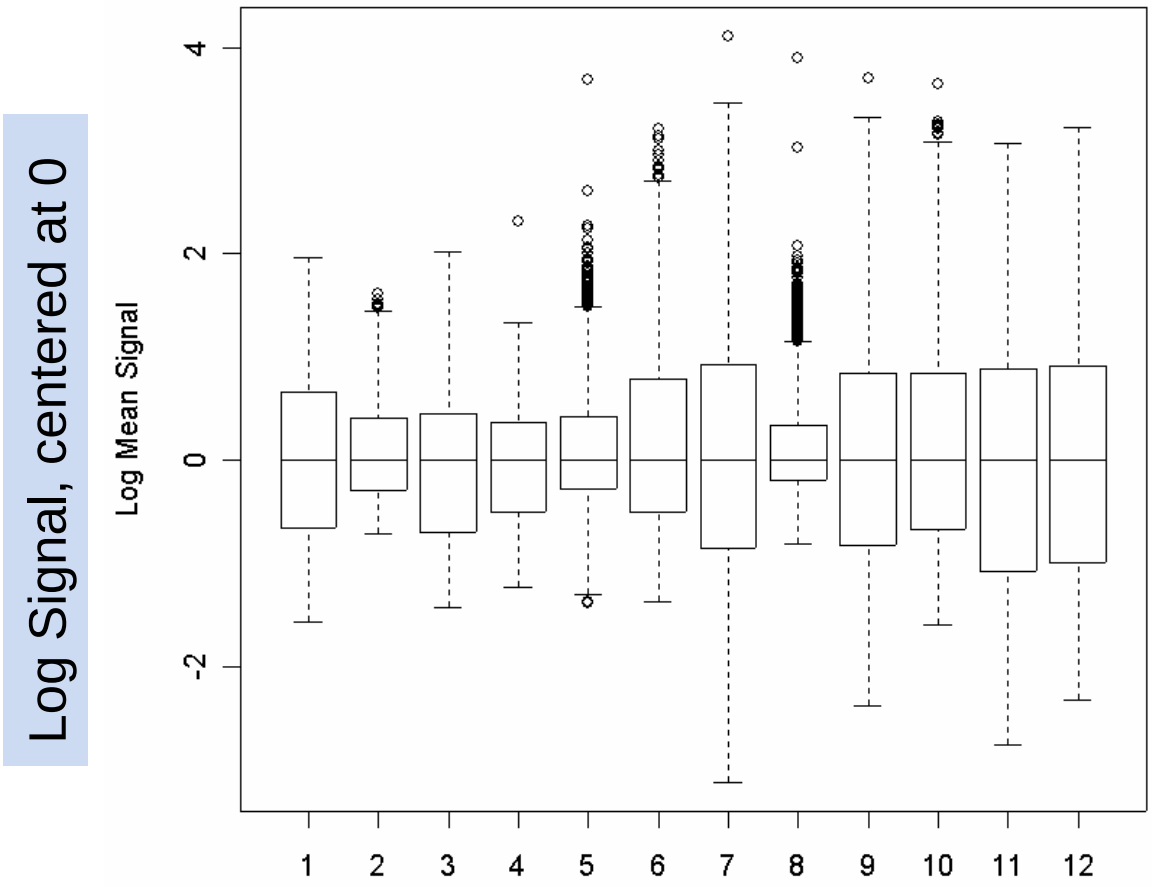
R Console

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

Median-centering

- One of the simplest strategies is to bring all „Centers“ of the array data to the same level.
- Assumption, the majority of genes or the center should not change between conditions.
- the Median is used as a robust measure.

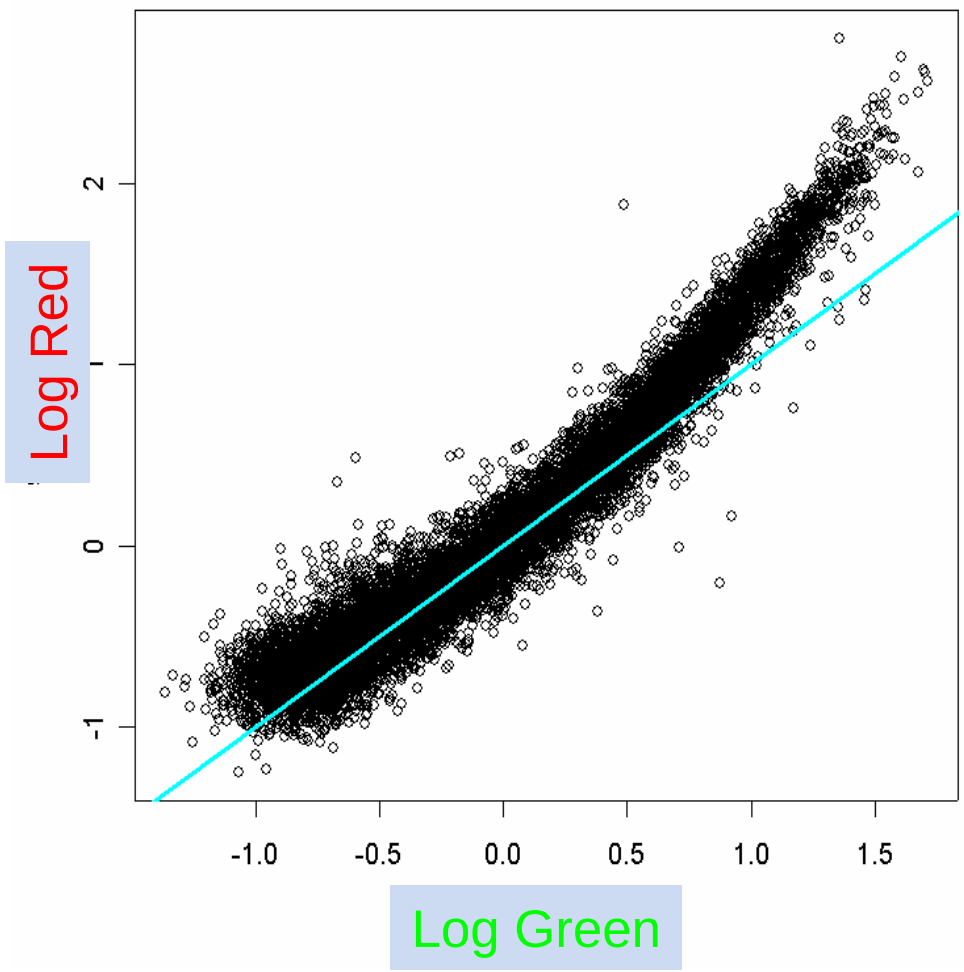
divide all expression measurements of each array by the Median.



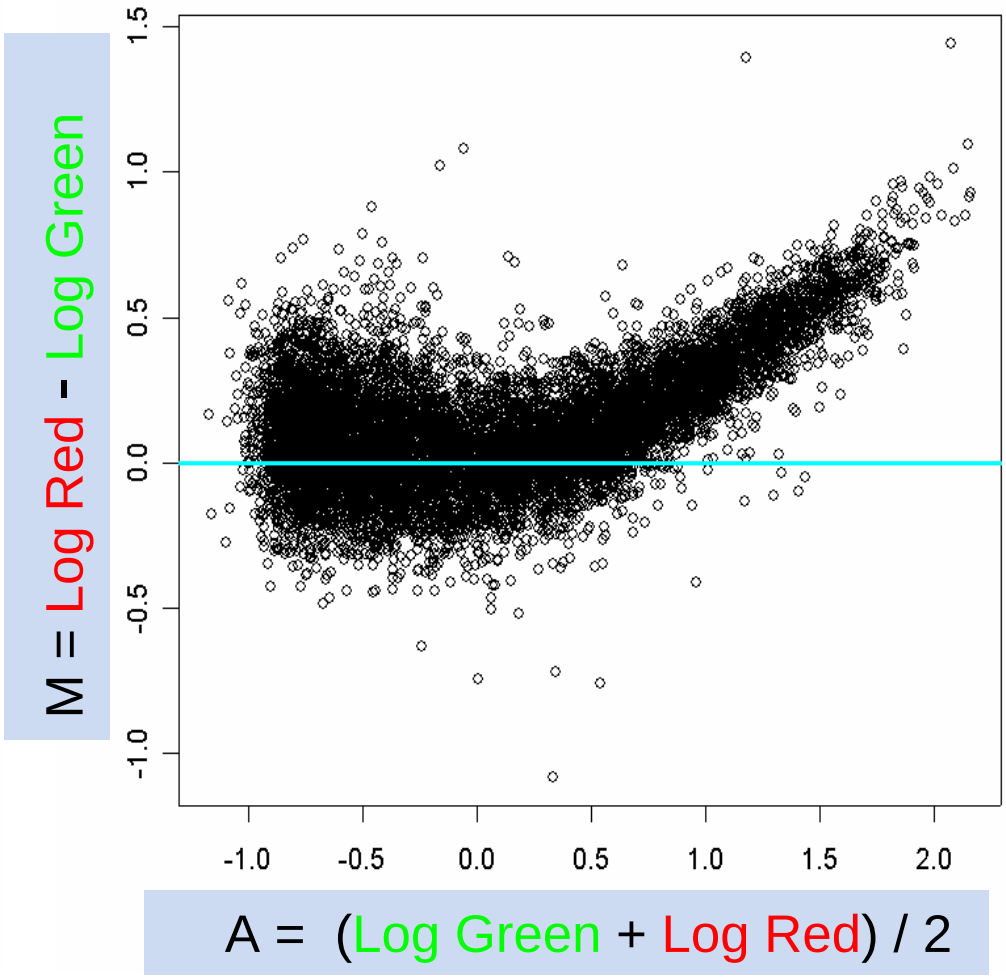
Problems with Median-Centering

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

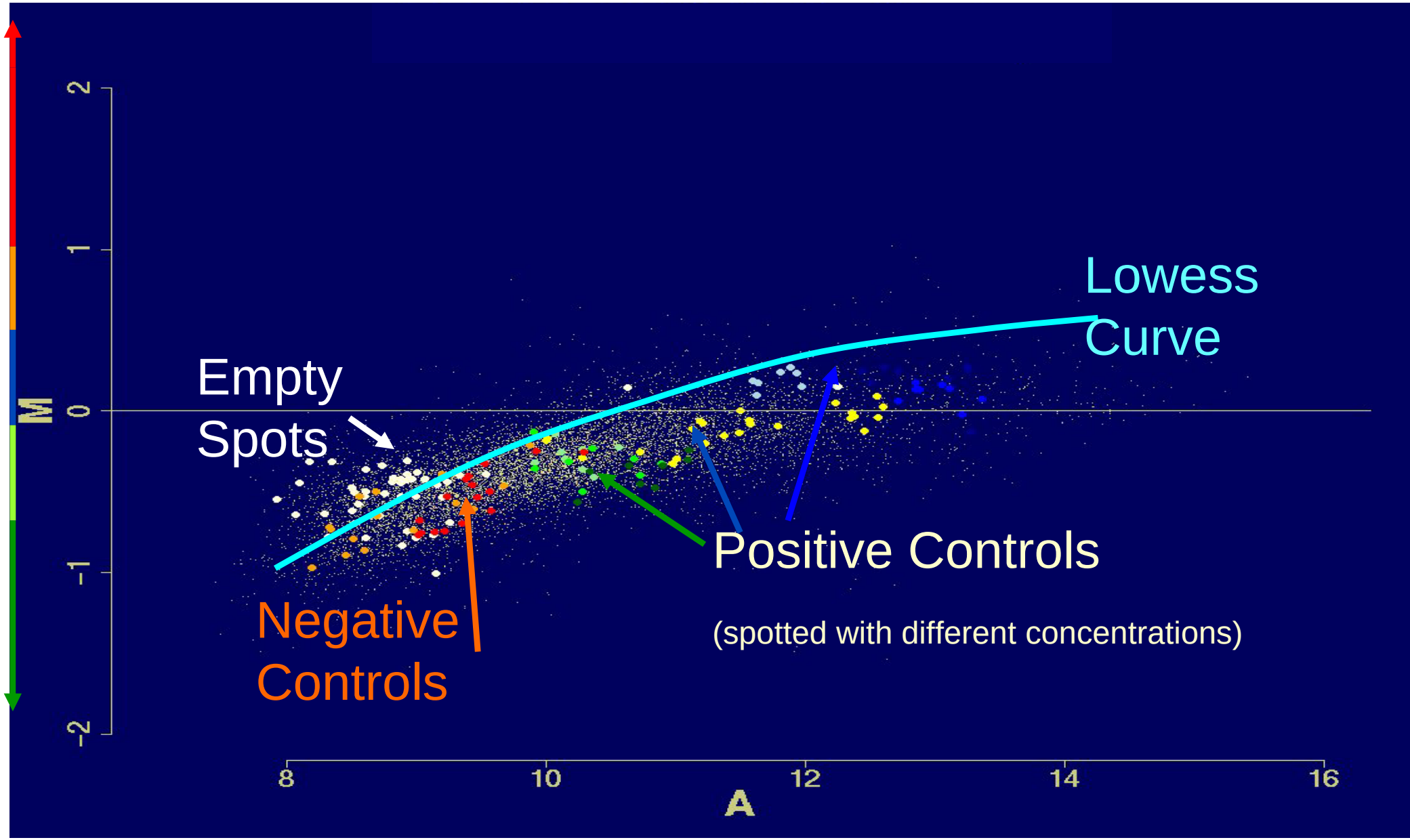
Scatterplot of log-Signals after Median-centering



M-A Plot of the same data

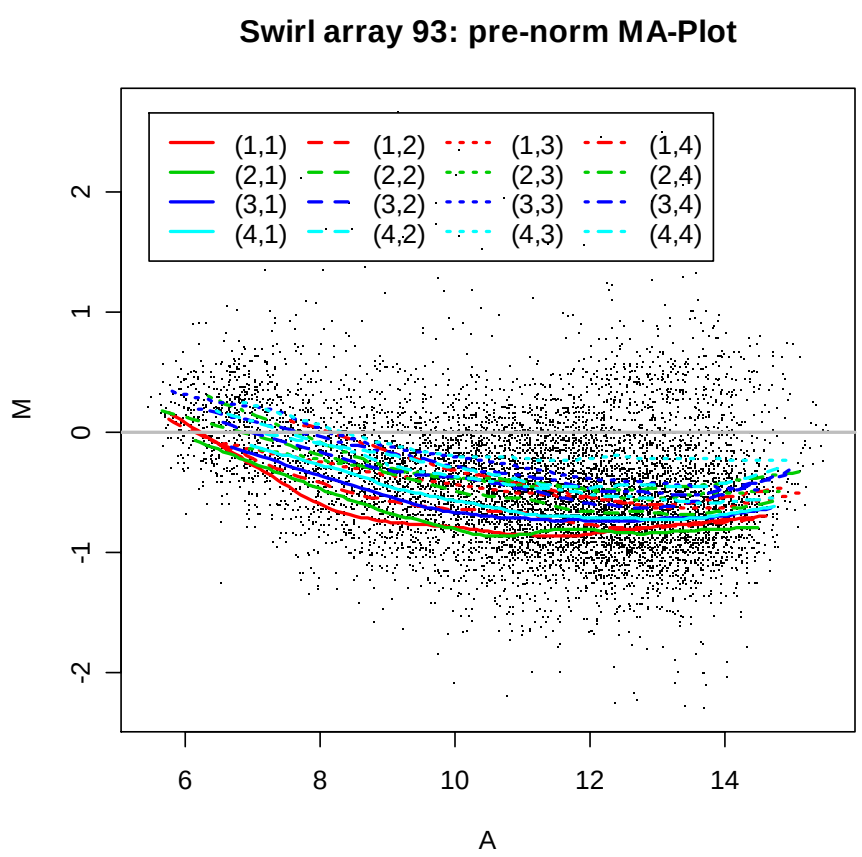


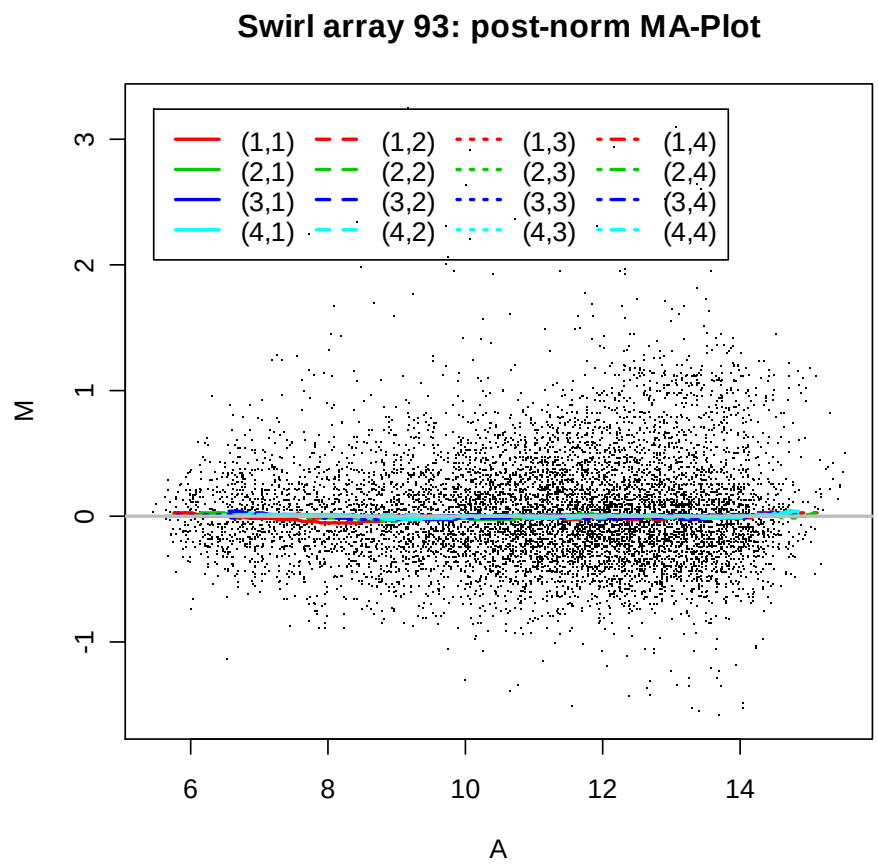
Lowess Normalization



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

marray – Swirl Data: Print-tip lowess Normalization



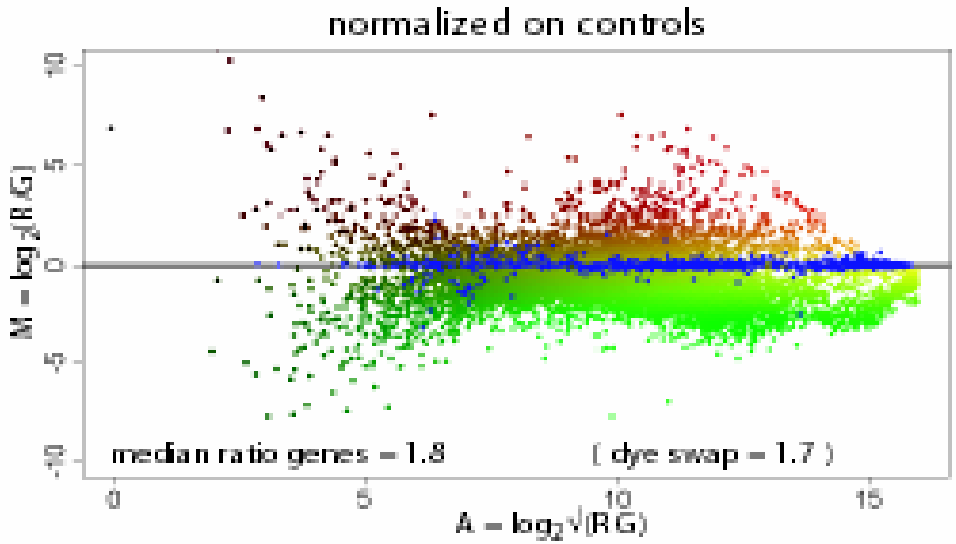


R Console

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

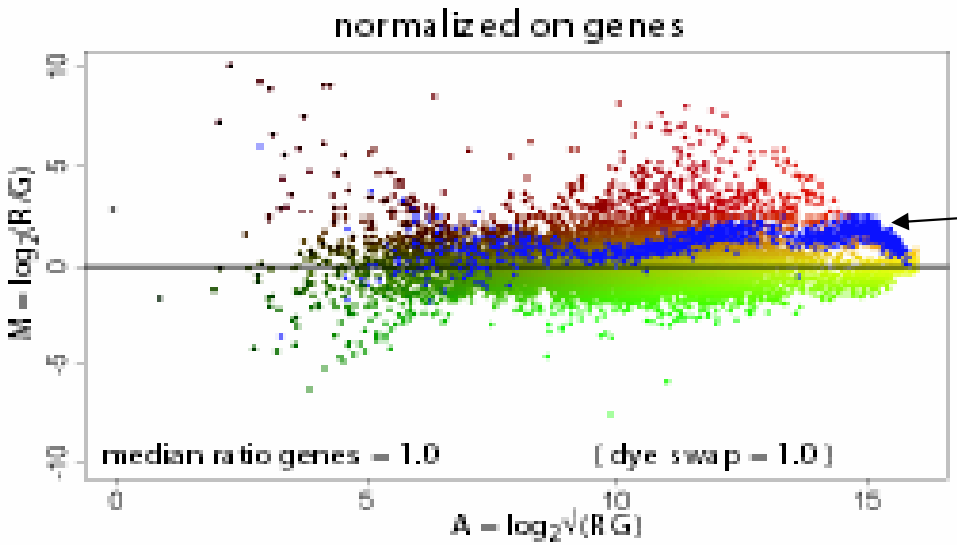
Non-parametric smoother: loess, lowess, local regression line, generalizes the concept of moving average.

A.



External Controls

B.



External controls

From Van de Peppel et al, 2003

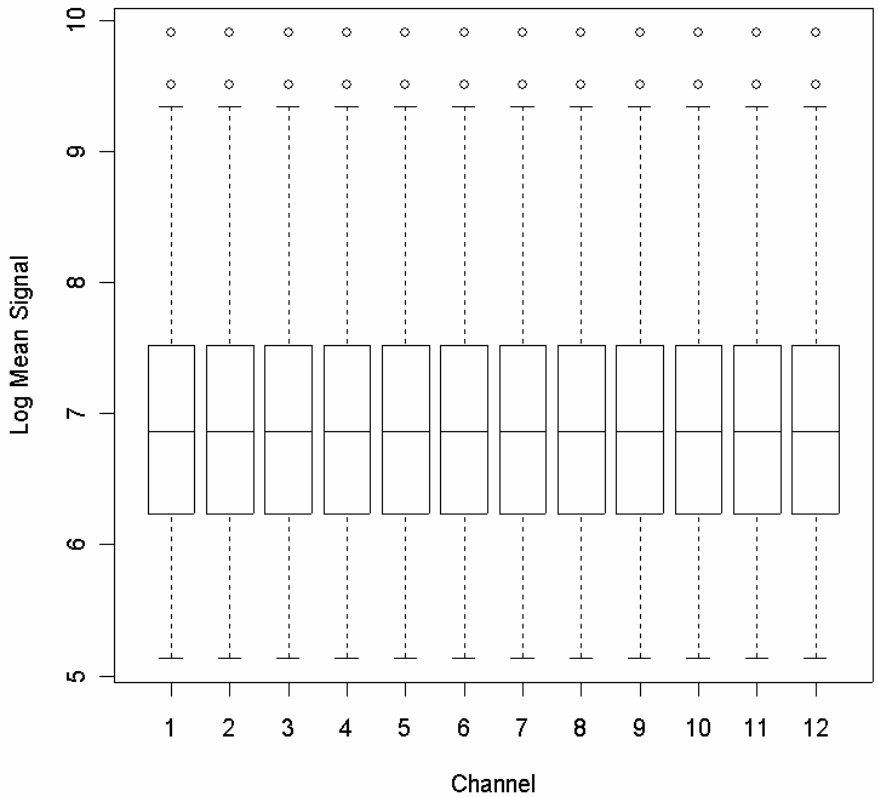
Quantile-Normalization

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.

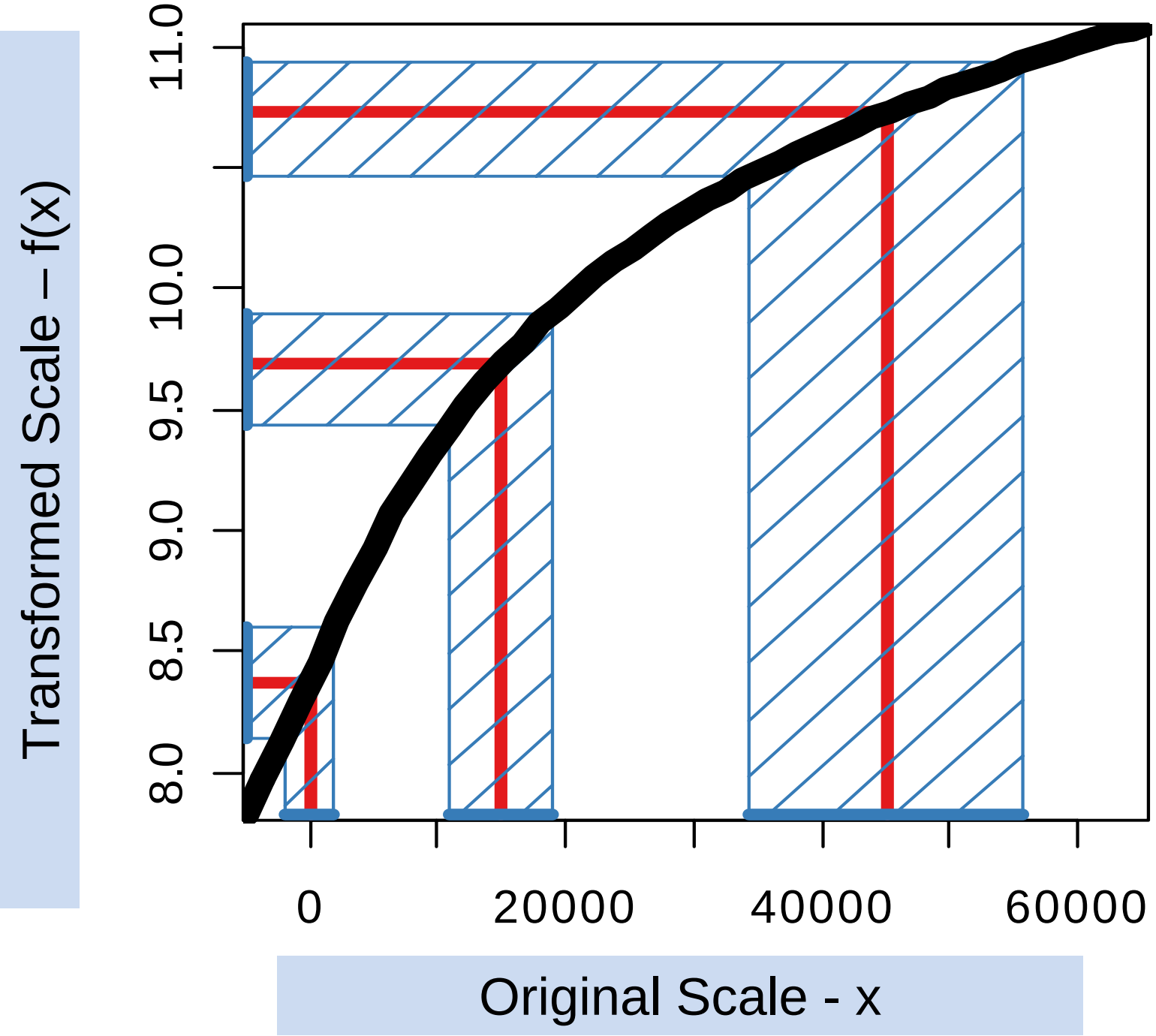
Boxplot after Quantile-normalization



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



variance stabilizing transformations

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

1.) constant variance ('additive') $v(u) = s^2 \quad \Rightarrow \quad f \propto u$

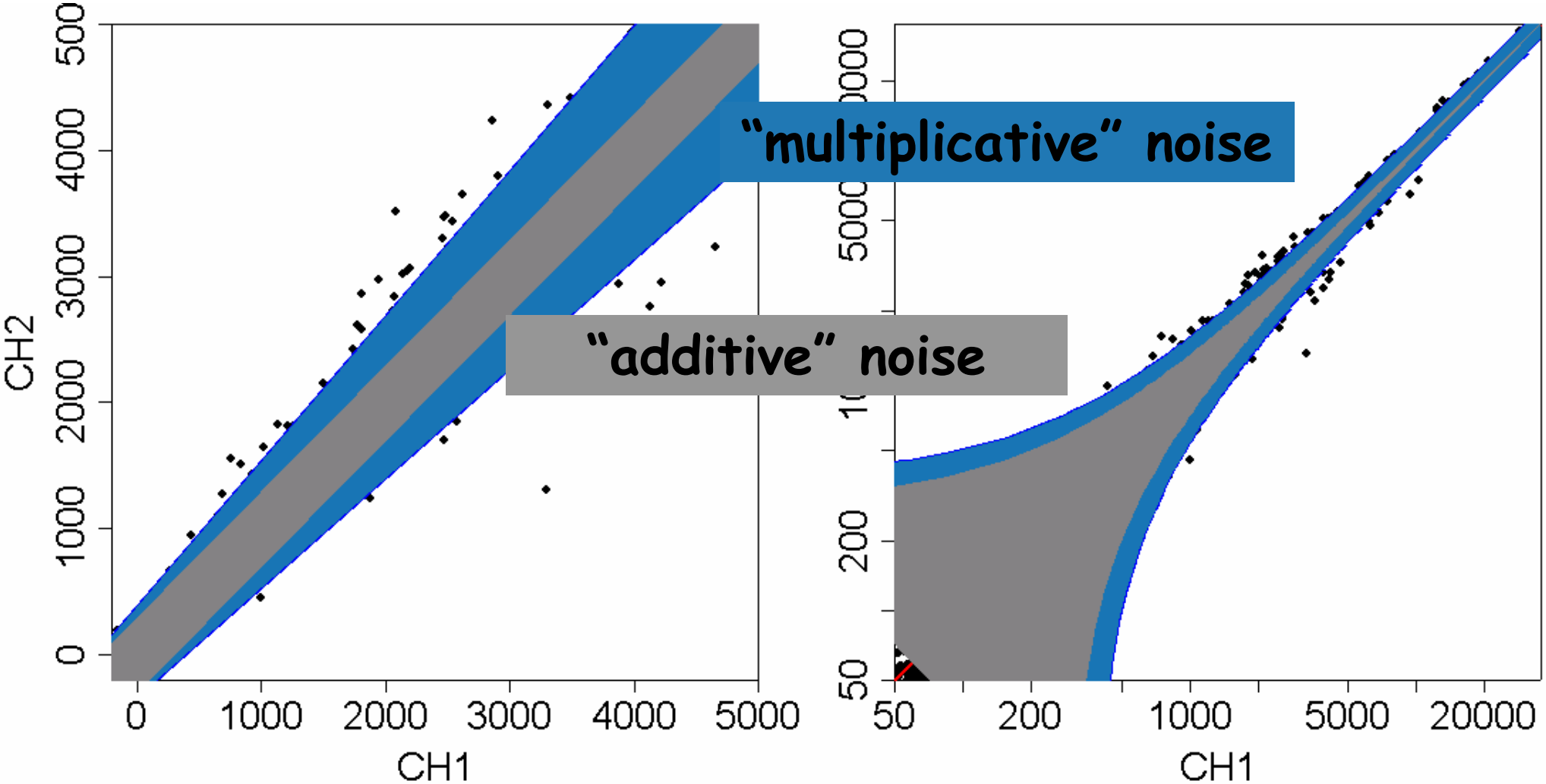
2.) constant CV ('multiplicative') $v(u) \propto u^2 \Rightarrow f \propto \log u$

3.) offset $v(u) \propto (u + u_0)^2 \quad \Rightarrow \quad f \propto \log(u + u_0)$

4.) additive and multiplicative

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

The two-component model



raw scale

log scale

B. Durbin, D. Rocke, JCB 2001

Fitting of an error model

measured intensity = offset + gain × true abundance

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

$$a_{ik} = a_i + \varepsilon_{ik}$$

a_i per-sample offset

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

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

b_i per-sample normalization factor

b_k sequence-wise probe efficiency

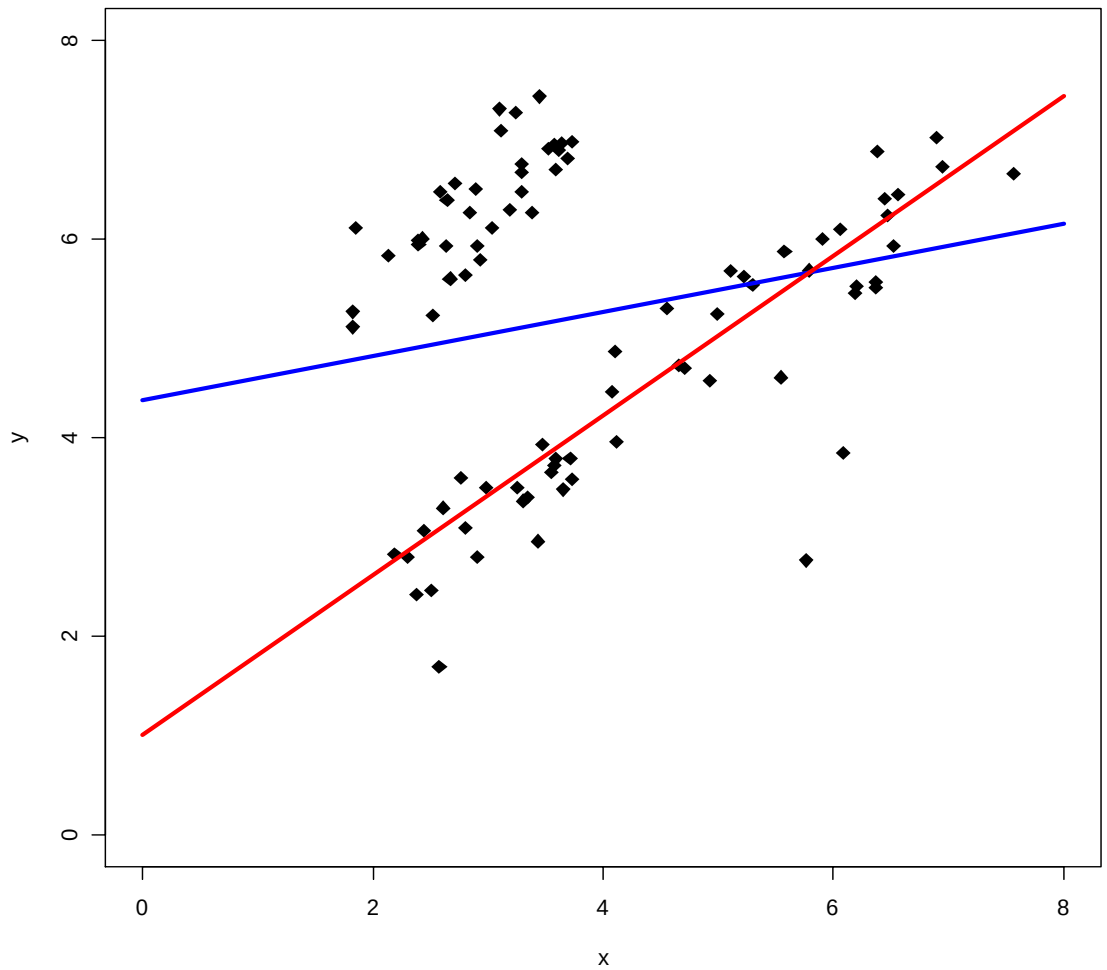
$\eta_{ik} \sim N(0, s_2^2)$
"multiplicative noise"

Parameter-Estimation

$$\operatorname{arsinh}\frac{Y_{ki} - a_i}{b_i} = \mu_k + \varepsilon_{ki}, \qquad \varepsilon_{ki} \sim N(0, c^2)$$

- maximum likelihood estimator: straightforward – but sensitive to deviations from normality
- model holds for genes that are unchanged; differentially transcribed genes act as outliers.
- robust variant of ML estimator, à la Least Trimmed Sum of Squares regression.
- works as long as <50% of genes are differentially transcribed

Least trimmed sum of squares regression



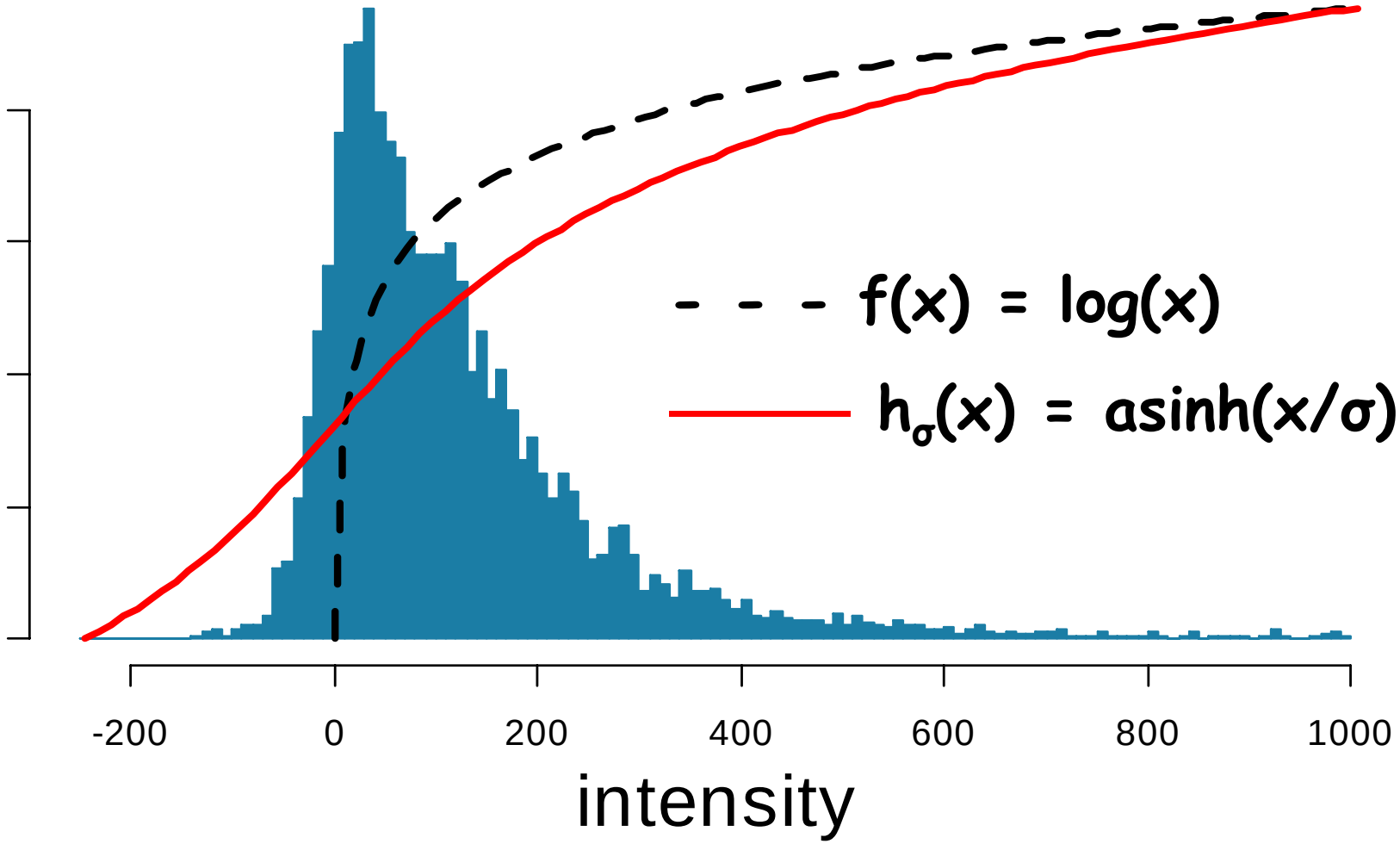
- least sum of squares
- least trimmed sum of squares

minimize

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

P. Rousseeuw, 1980s

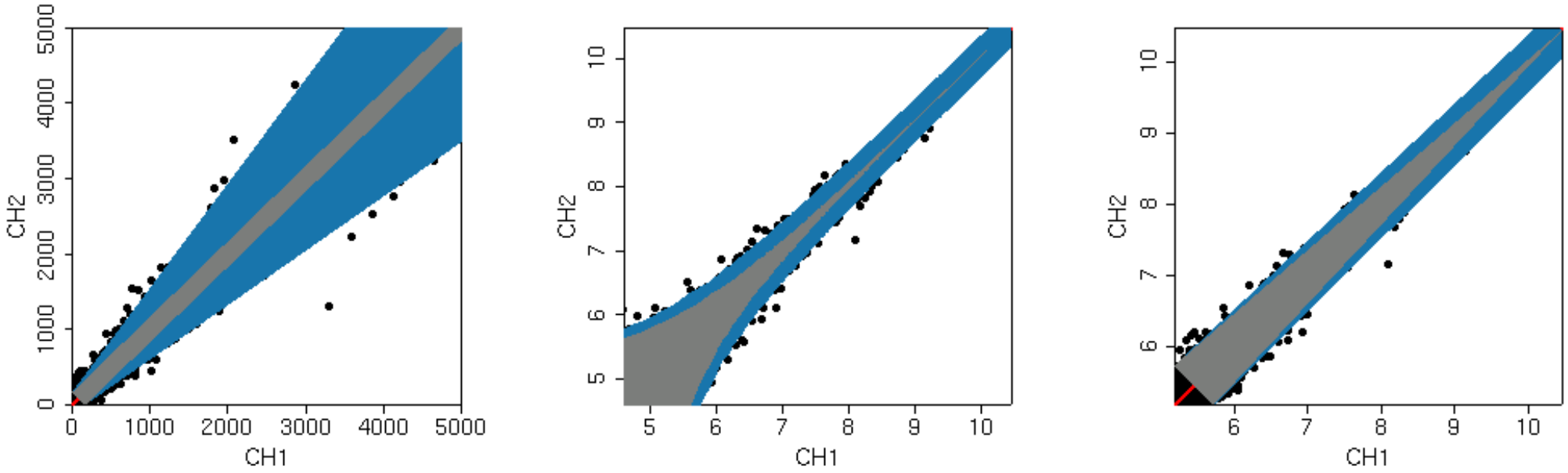
The „glog“-Transformation



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

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

The „glog“-Transformation



Variance:

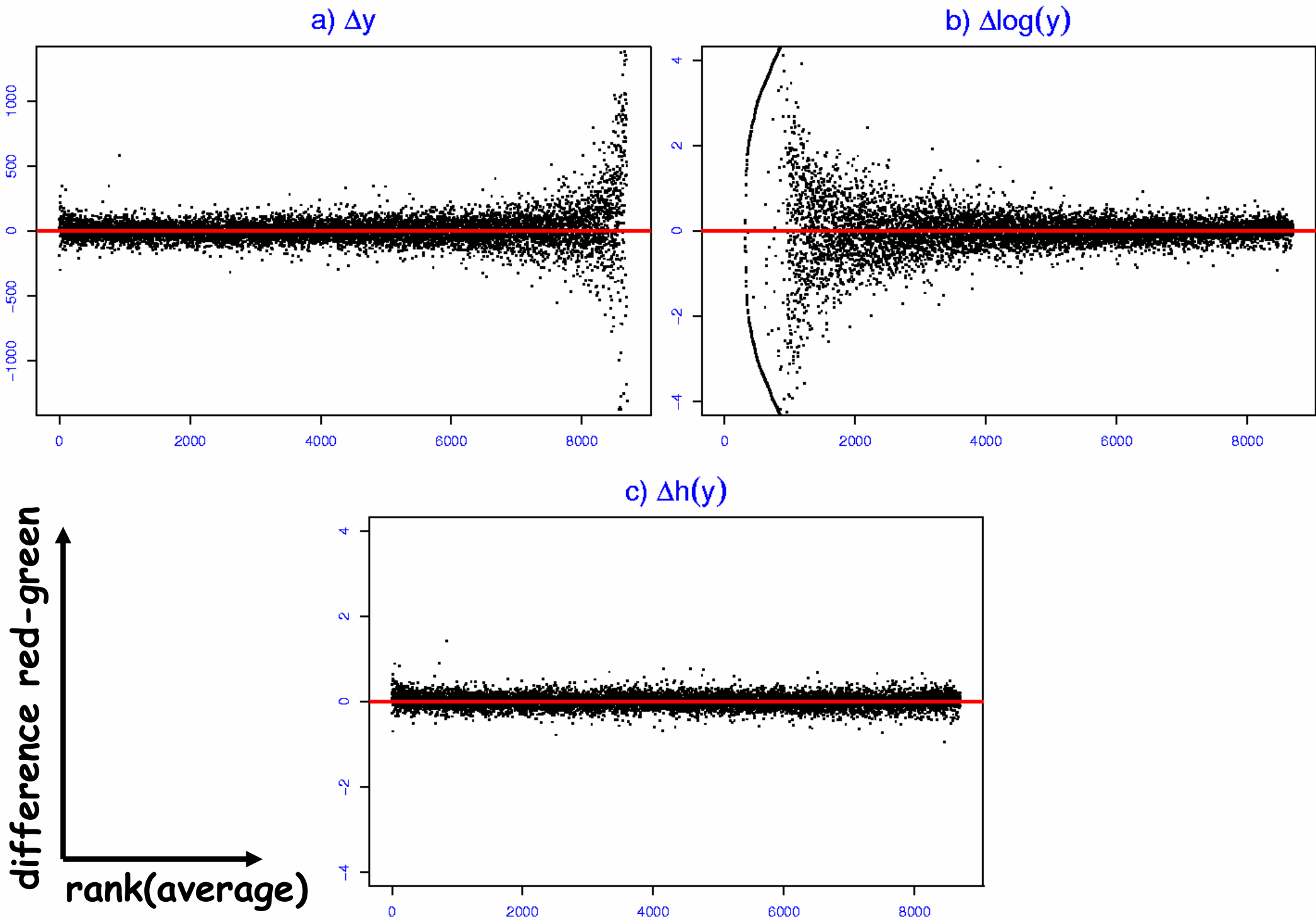
Additive component

multiplicative component

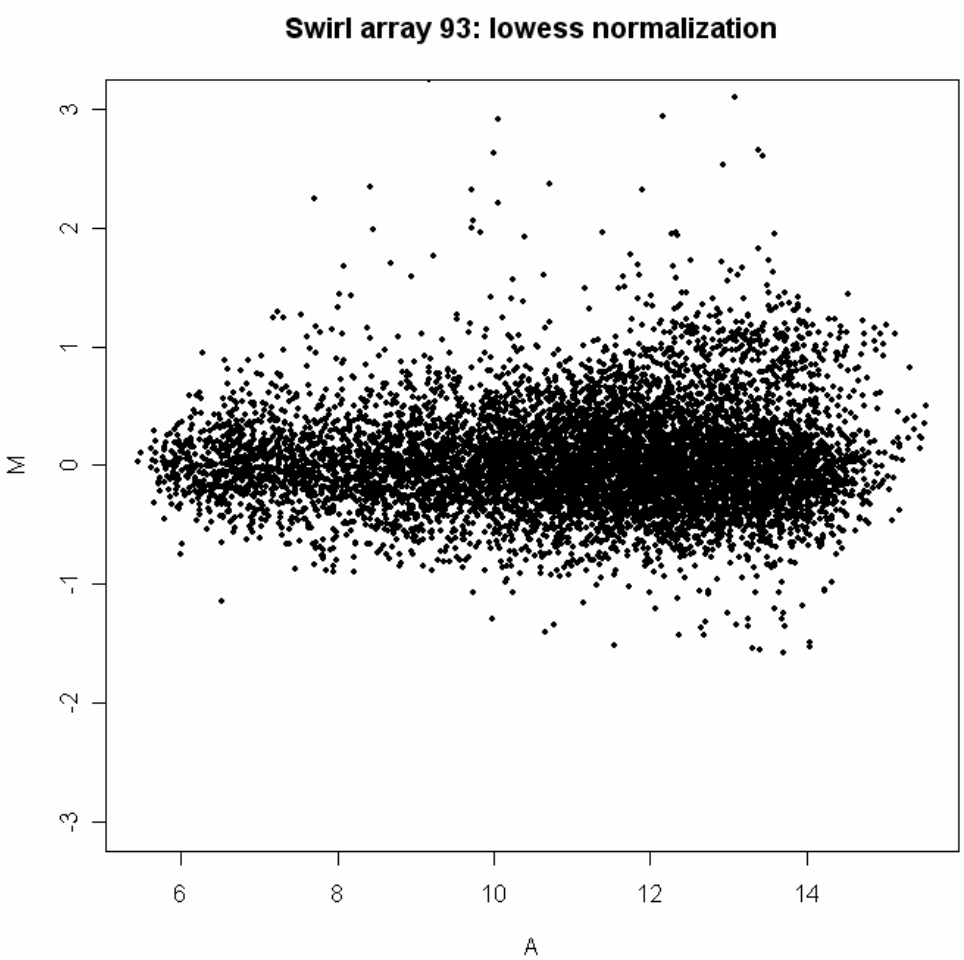
P. Munson, 2001

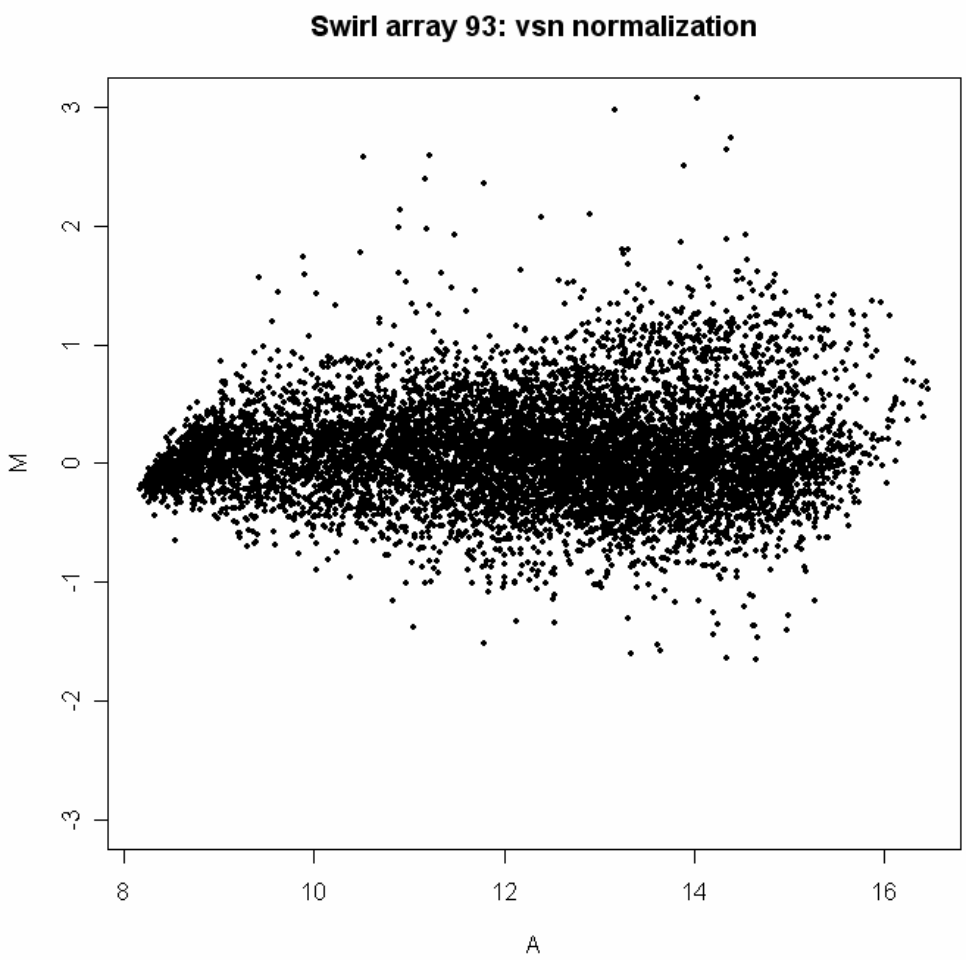
D. Rocke & B. Durbin, ISMB 2002

W. Huber et al., ISMB 2002


a) Δy
b) Δlog(y)
c) Δh(y)
difference red-green
rank(average)
1000
500
0
−500
−1000
4
2
0
−2
−4
0
2000
4000
6000
8000

Swirl Data: Lowess versus VSN





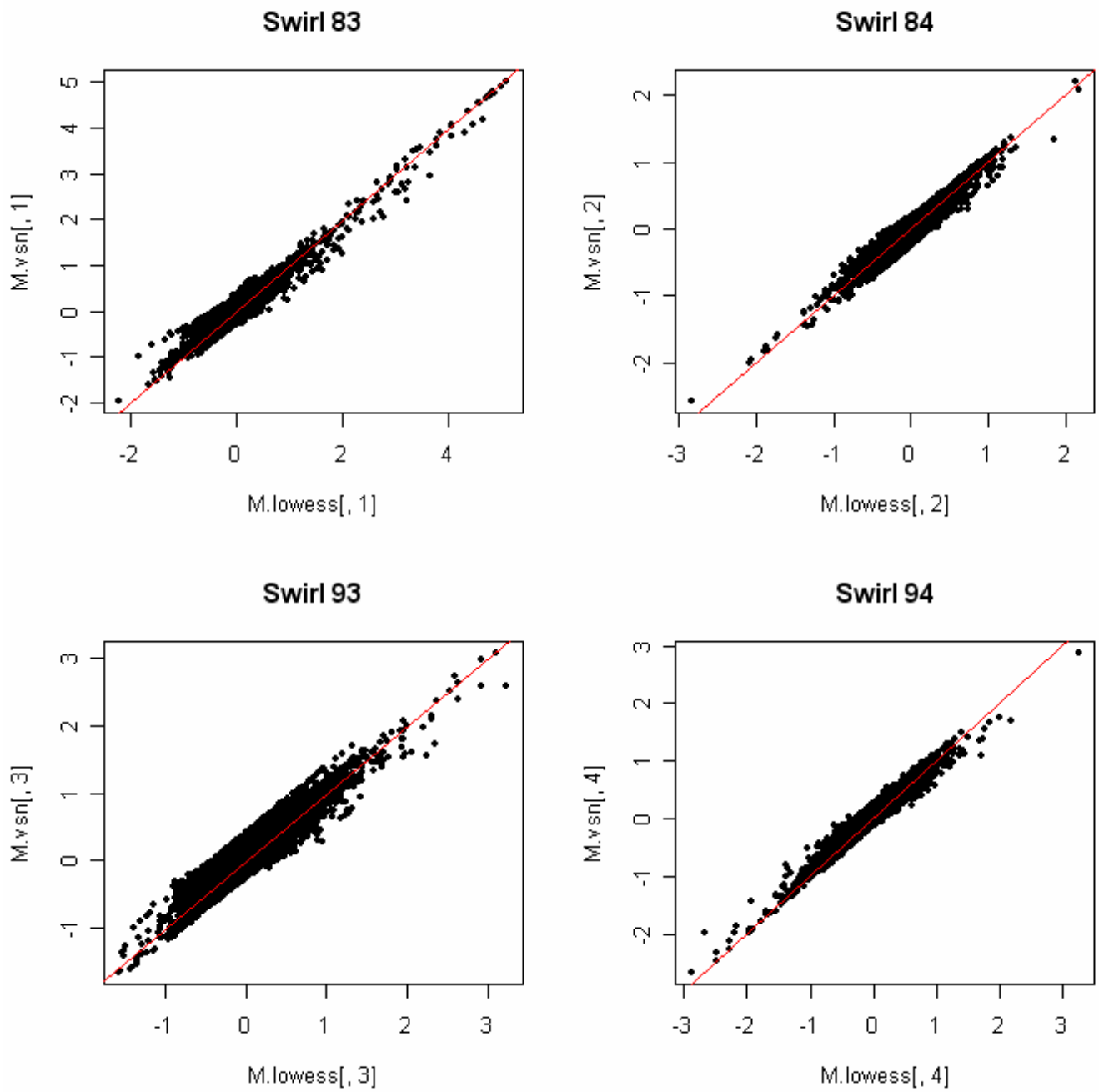
R Console

```
> 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)
```

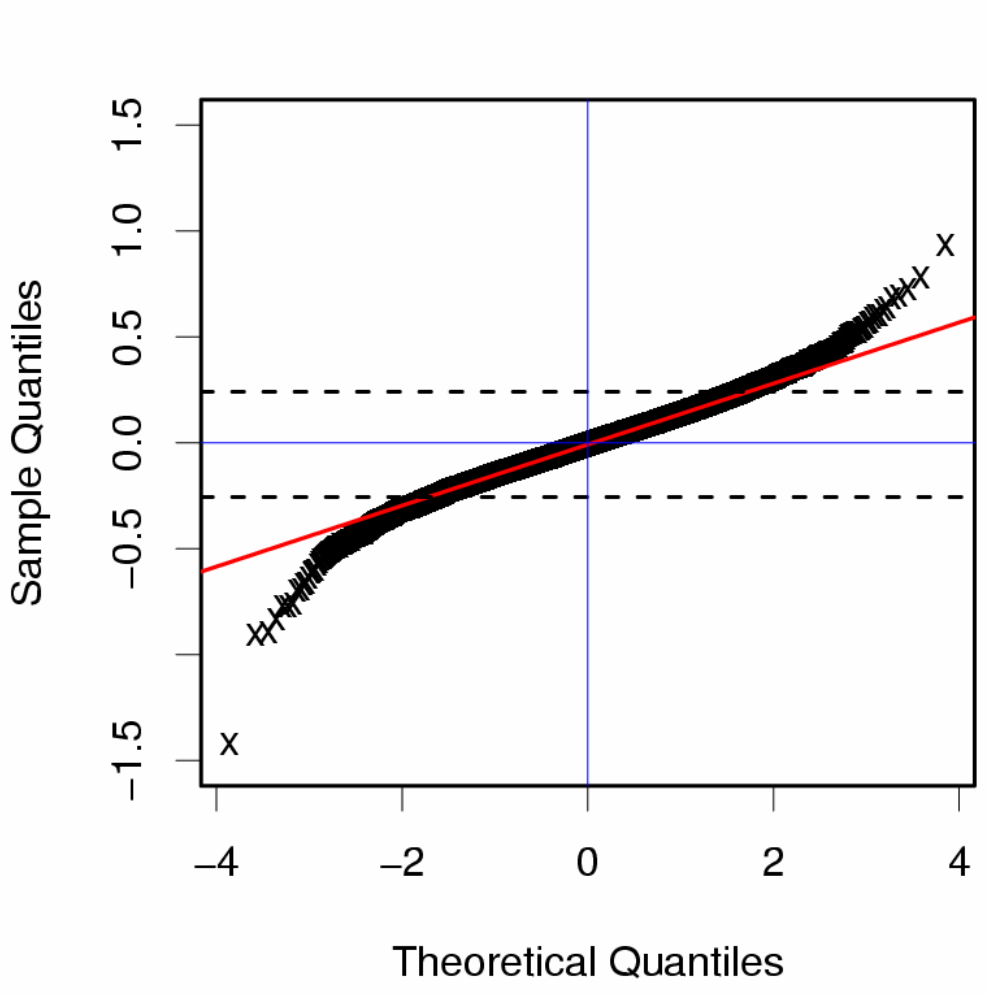
Swirl: LOWESS versus VSN

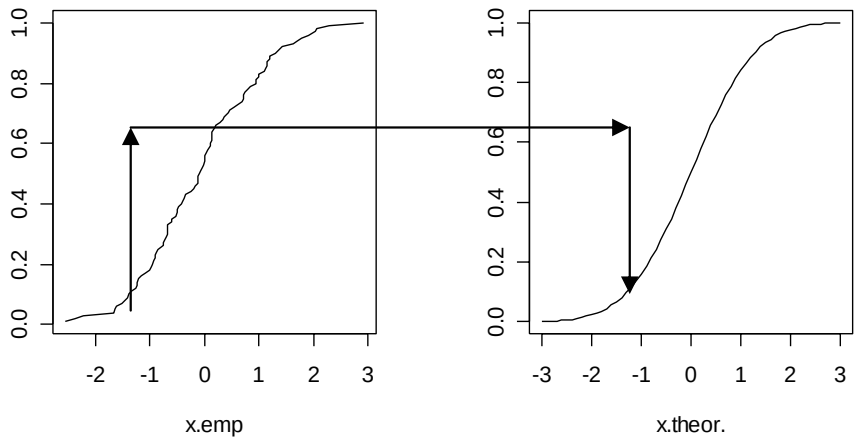
R Console

```
> 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")
```



Normality: QQ-plot





Summary

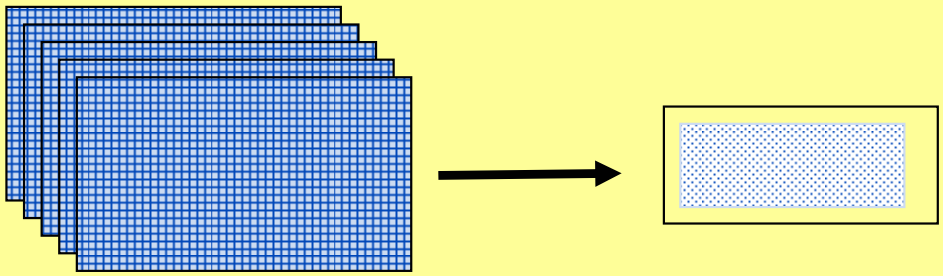
- What makes a good measurement: Precision and Unbiasednes
- Need to normalize.
- Normalization is not something trivial, has many practical and theoretical implications which need to be considered.
- What is the best way to normalize?
- How dependent is the result of your analysis from the normalization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

Experimental Design:

- Different levels of replication
- Pooling vs. non pooling
- different strategies to pair hybridization targets on cDNA arrays
- direct vs. indirect comparisons

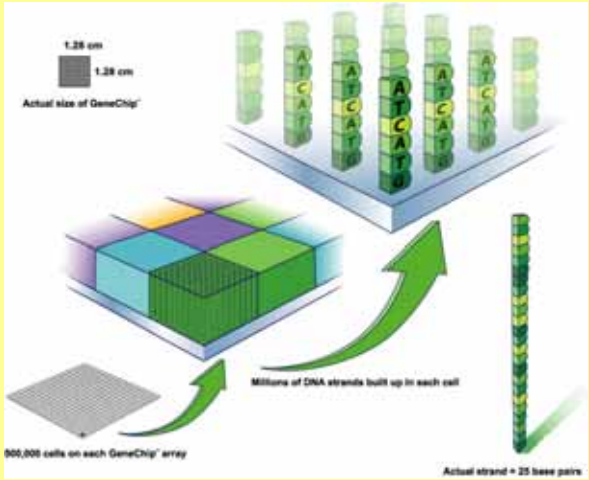
Two main aspects of array design

Design of the array



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

Spot as microarray on glass slides

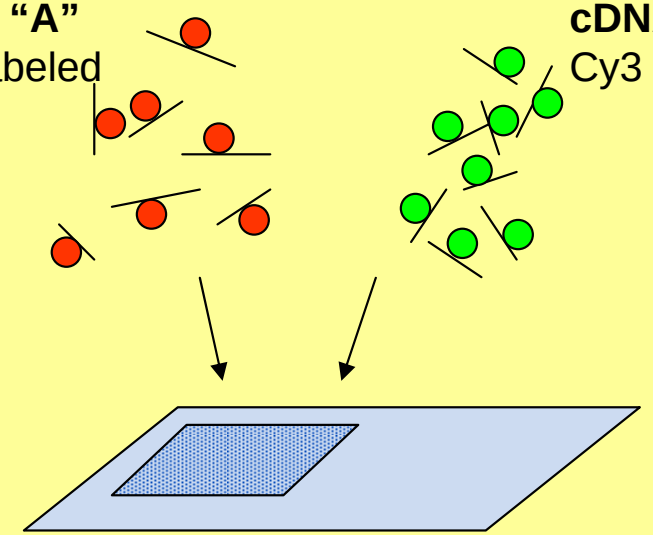


Allocation of mRNA samples to the slides

cDNA

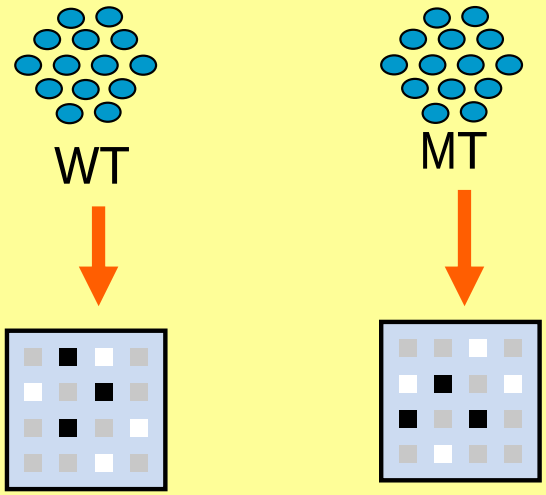
cDNA "A"
Cy5 labeled

cDNA "B"
Cy3 labeled



Hybridization

affy



Some aspects of design

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 array.

B Different design layout

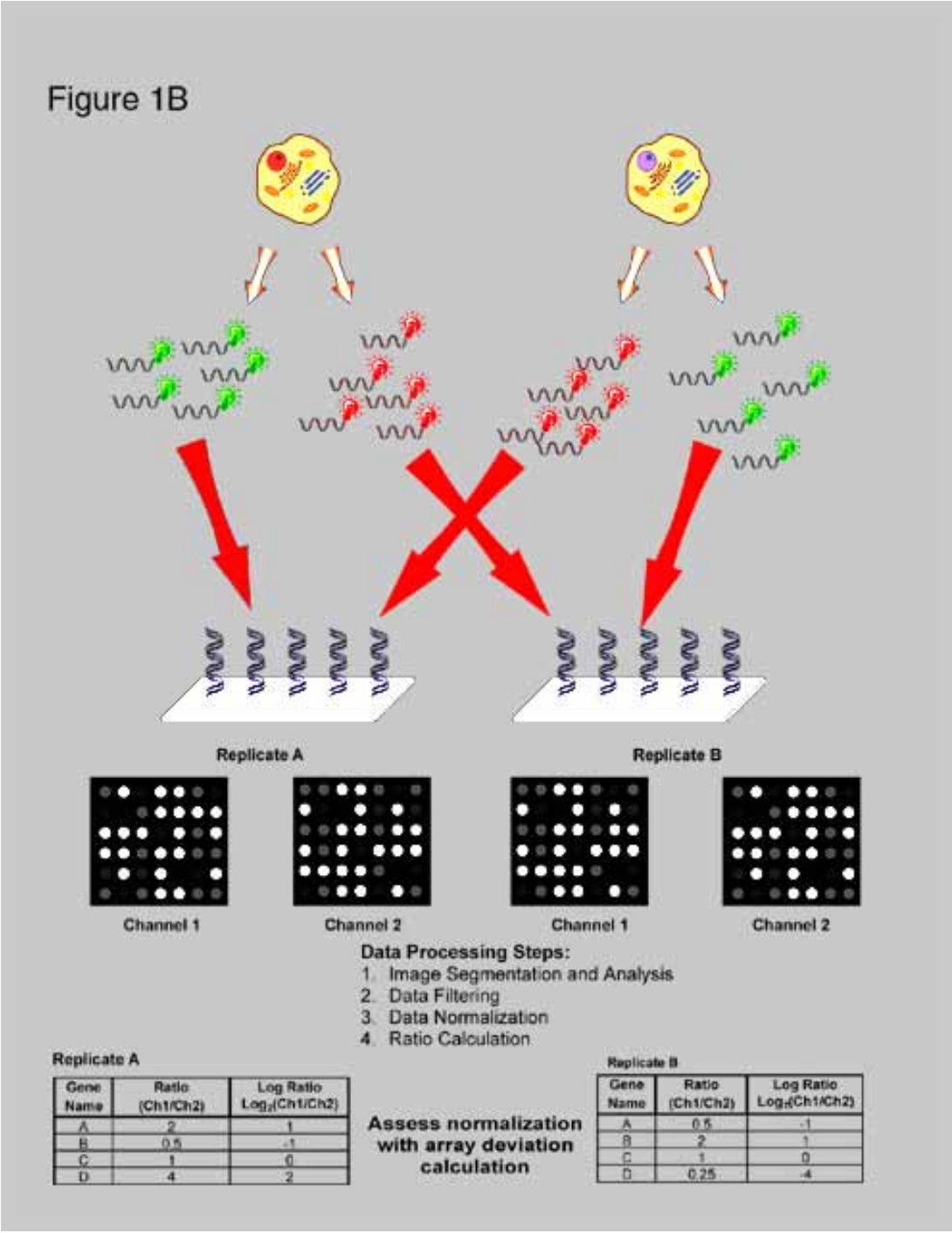
- Scientific aim of the experiment.
- Robustness.
- Extensibility.
- Efficiency.

Taking physical limitations or cost into consideration:

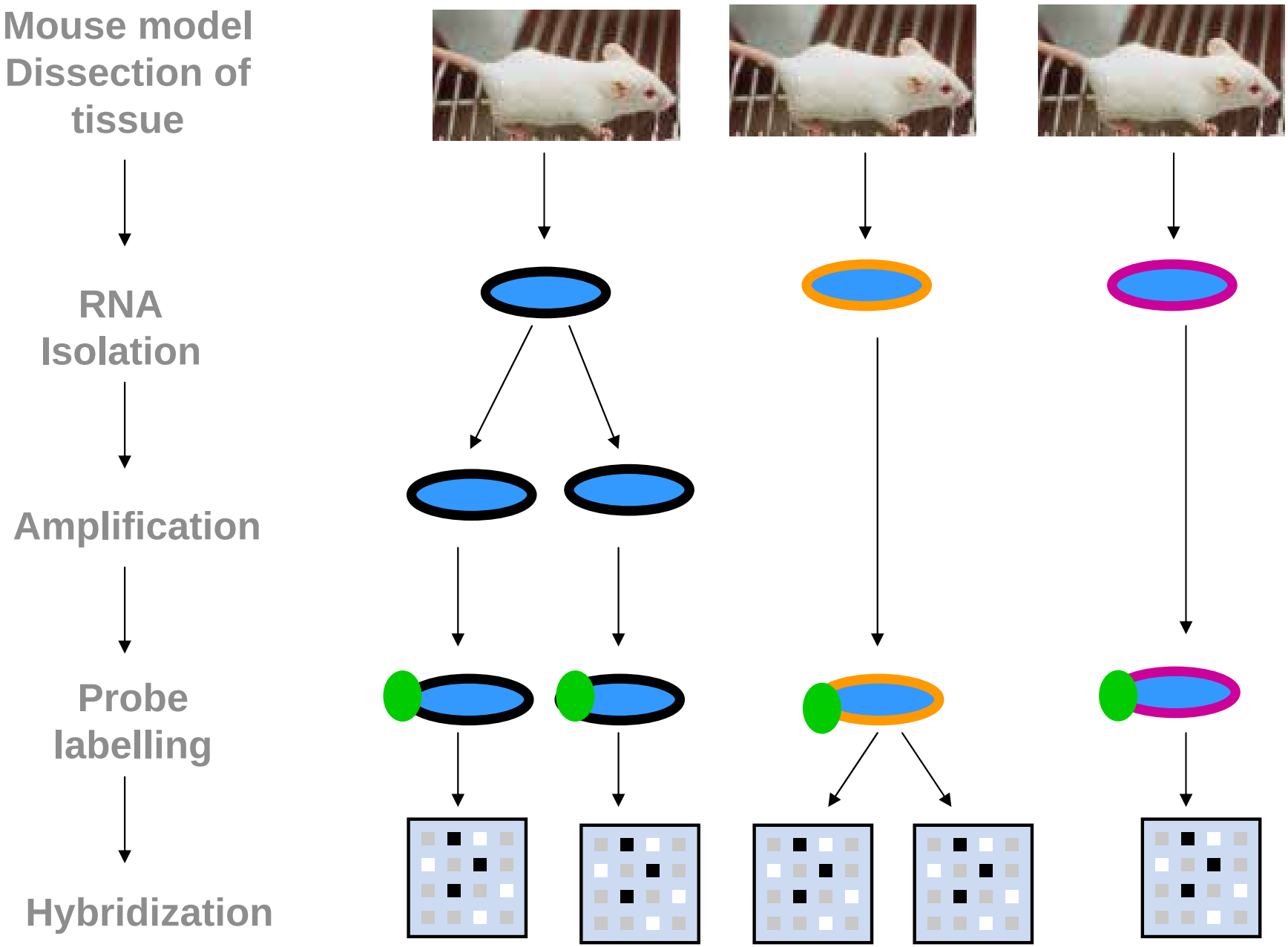
- the number of slides.
- the amount of material.

Design of a Dye-Swap Experiment

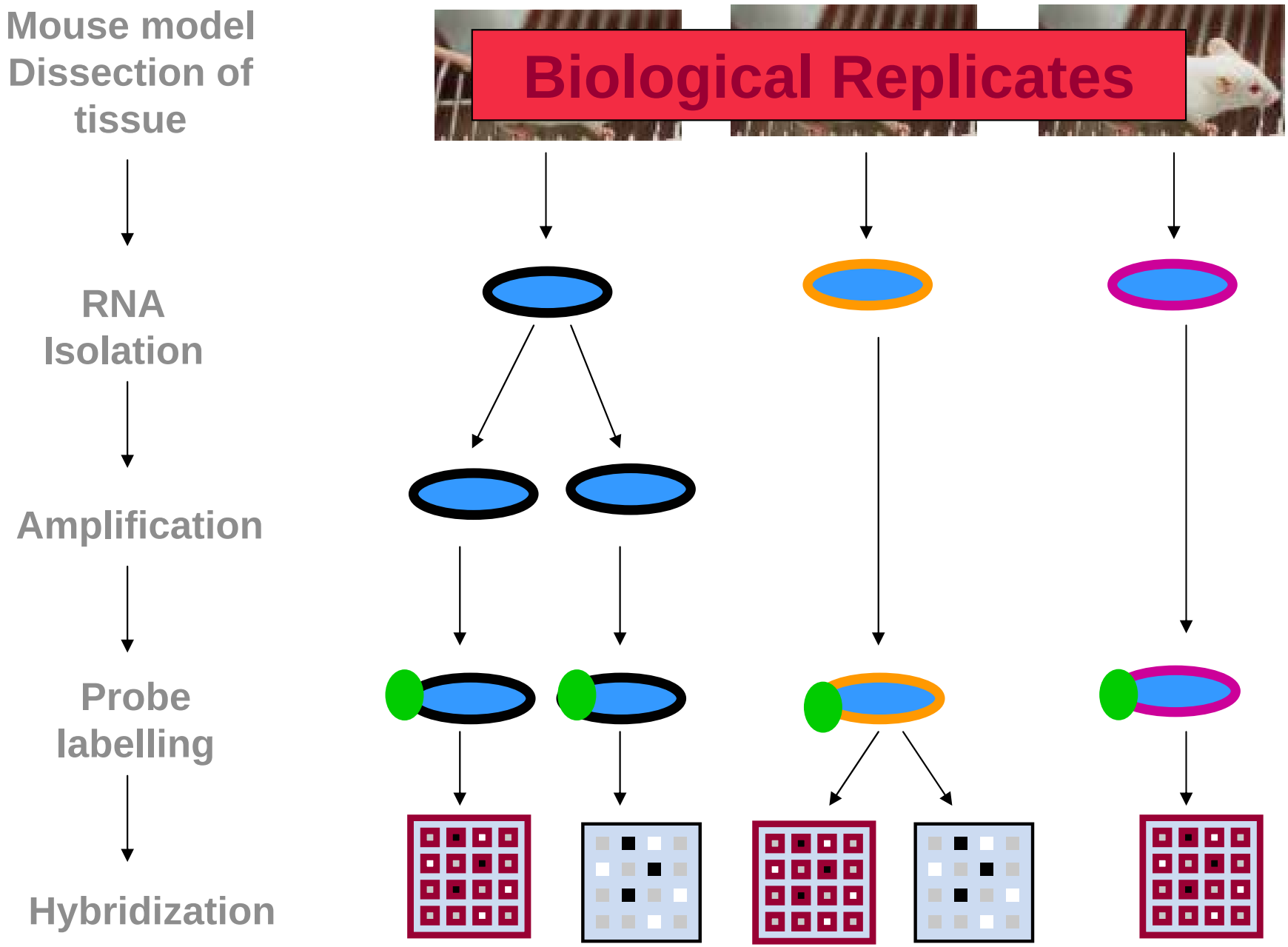
- Repeats are essential to control the quality of an experiment.
- One example for Replicates is the Dye-Swap, i.e. Replicates with the same mRNA Pool but with swapped labels.
- Dye-Swap shows whether there is a dye-bias in the Experiment.



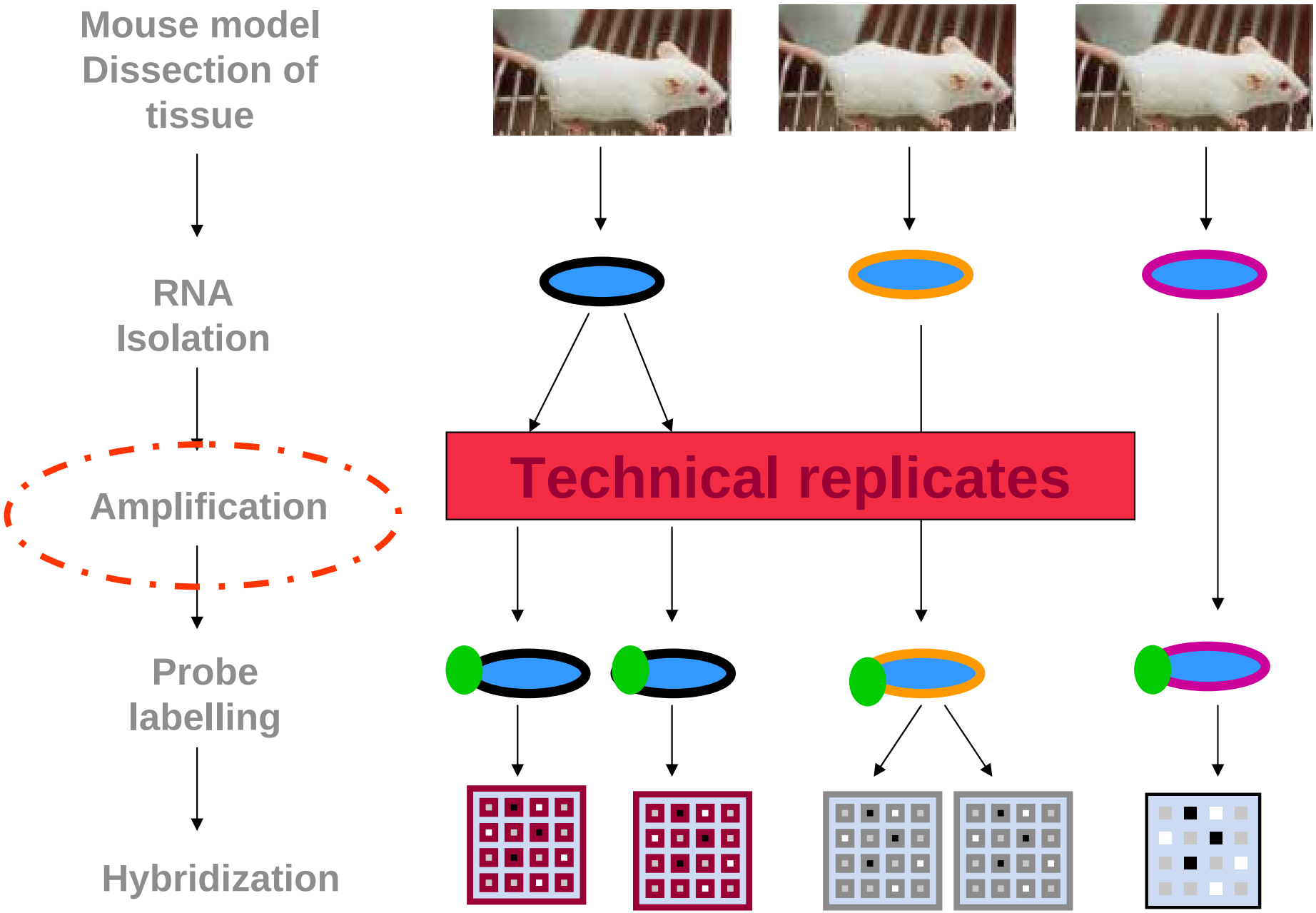
Preparing mRNA samples:



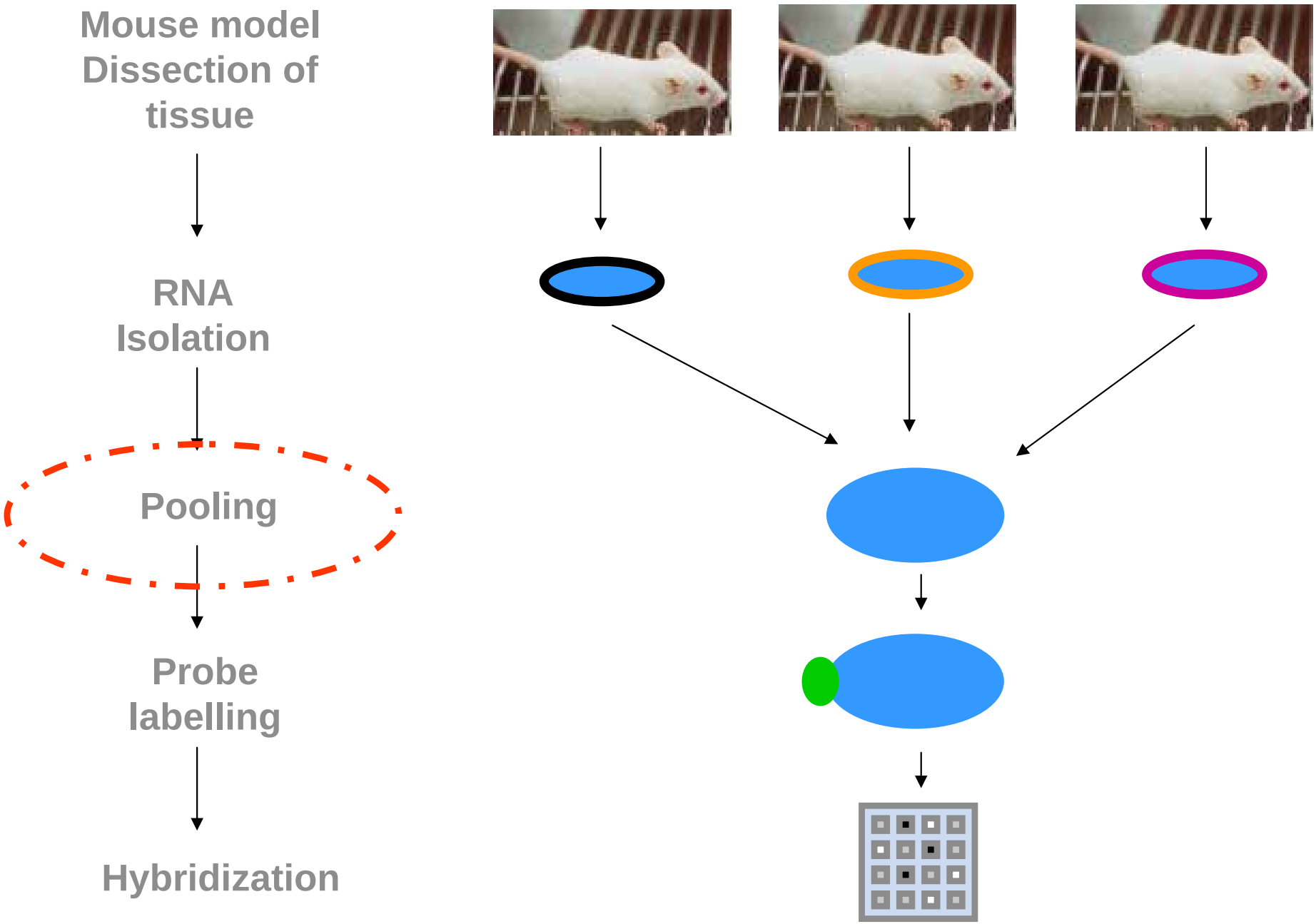
Preparing mRNA samples:



Preparing mRNA samples:

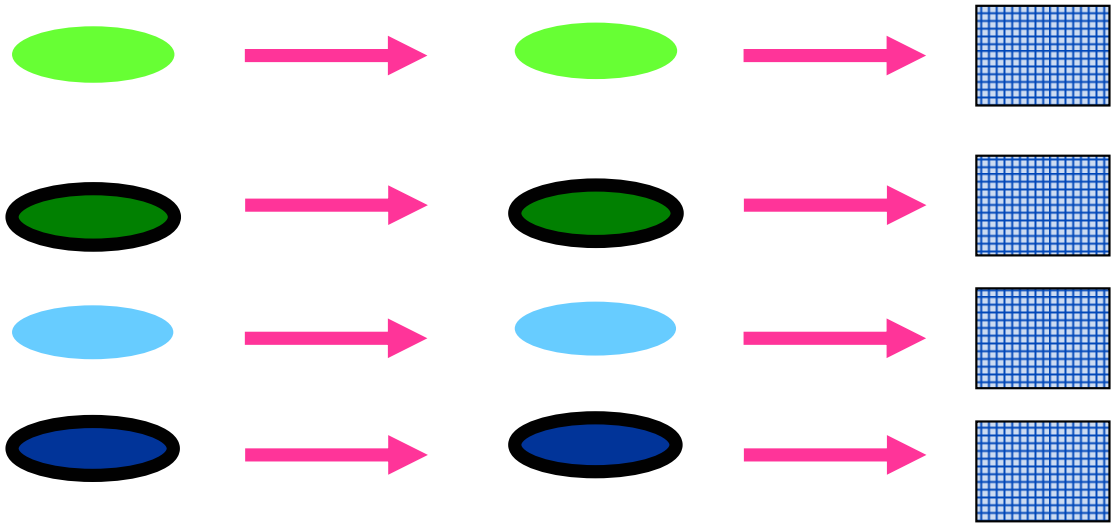


Pooling: looking at very small amount of tissues

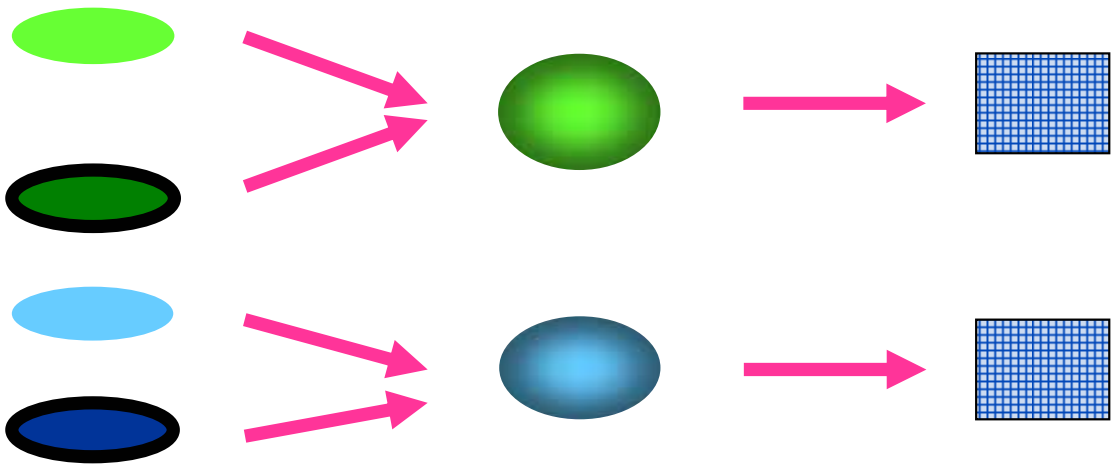


Pooled vs Individual samples

Design 1



Design 2



Taken from
Kendziorski etl al (2003)

Pooled versus Individual samples

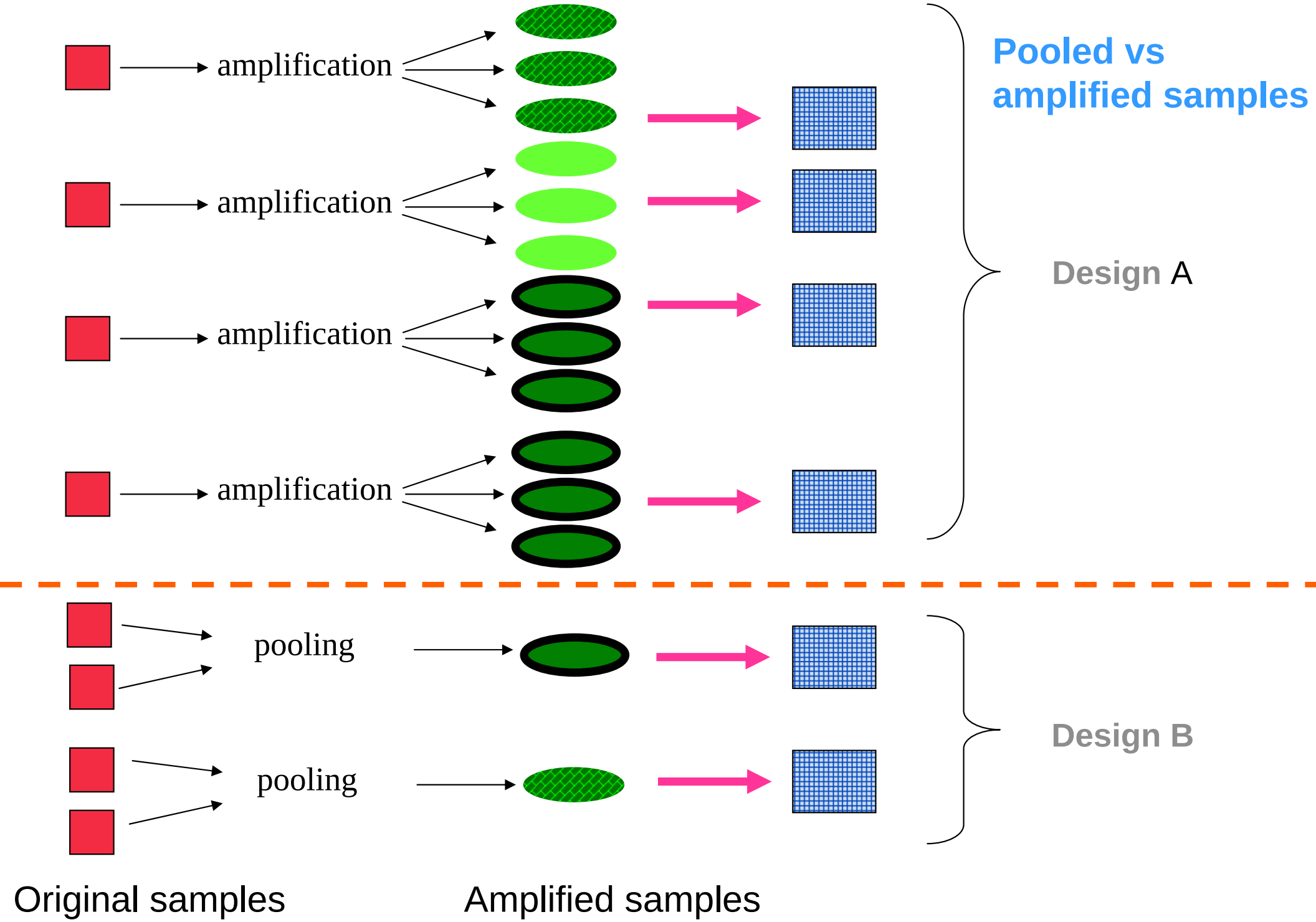
- Pooling is seen as “biological averaging”.

- Trade off between
  - Cost of performing a hybridization.
  - Cost of the mRNA samples.

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

- **References:**

  - Han, E.-S., Wu, Y., Bolstad, B., and Speed, T. P. (2003). A study of the effects of pooling on gene expression estimates using high density oligonucleotide array data. Department of Biological Science, University of Tulsa, February 2003.
  - Kendziorski, C.M., Y. Zhang, H. Lan, and A.D. Attie. (2003). The efficiency of mRNA pooling in microarray experiments. *Biostatistics* **4**, 465-477. 7/2003
  - Xuejun Peng, Constance L Wood, Eric M Blalock, Kuey Chu Chen, Philip W Landfield, Arnold J Stromberg (2003). Statistical implications of pooling RNA samples for microarray experiments. *BMC Bioinformatics* 4:26. 6/2003


Pooled vs amplified samples
amplification
amplification
amplification
amplification
Design A
pooling
pooling
Design B
Original samples
Amplified samples

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.
  - Able to perform more hybridizations.

- Pooling
  - Less replicates hybridizations.

Some aspects of design

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

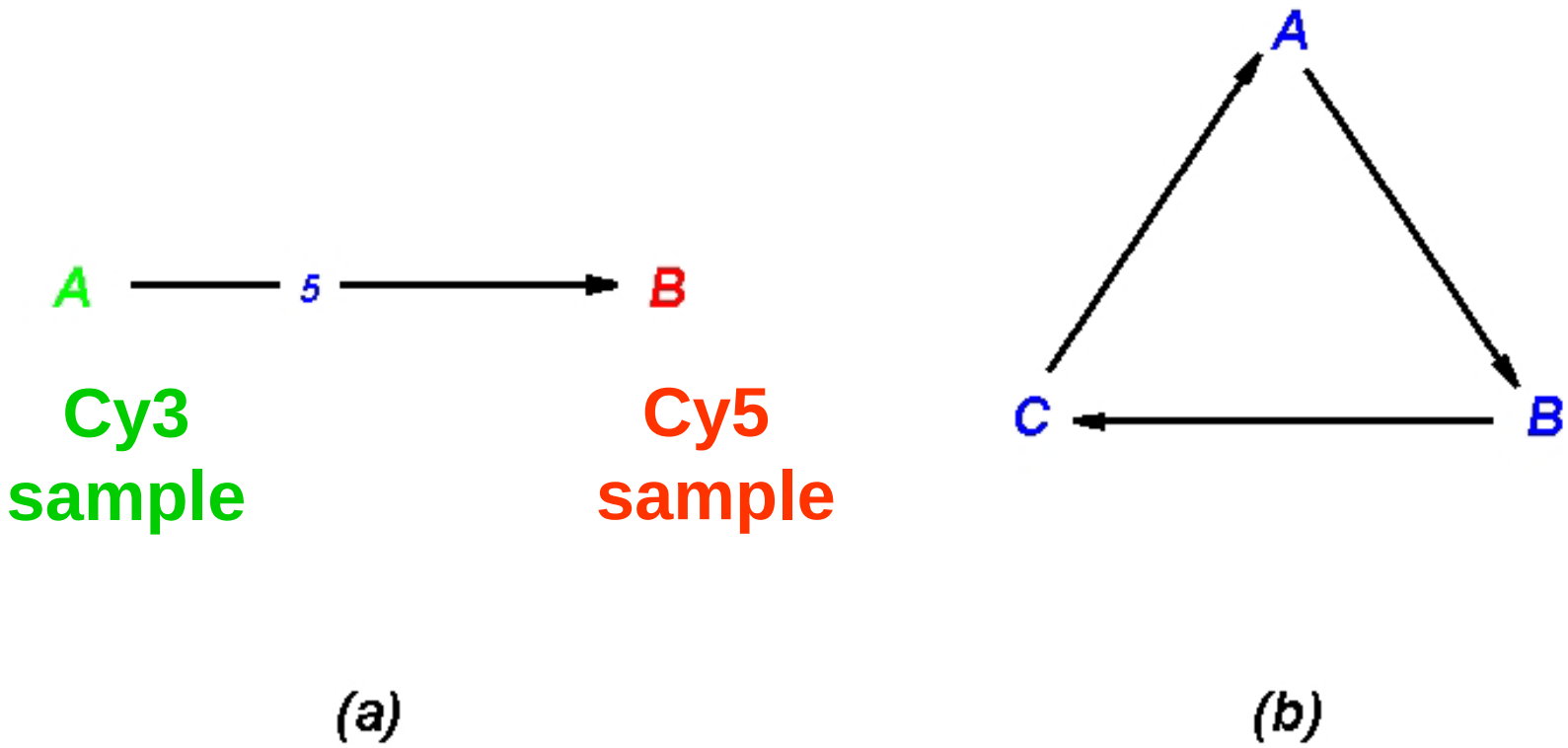
- Scientific aim of the experiment.
- Robustness.
- Extensibility.
- Efficiency.

Taking physical limitation or cost into consideration:

- the number of slides.
- the amount of material.

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.

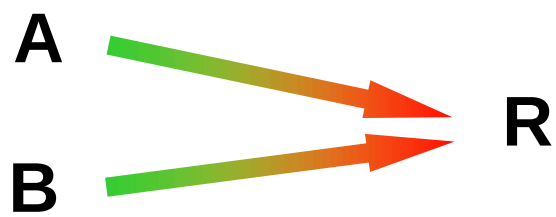
The simplest design question:
Direct versus indirect comparisons

Two samples (A *vs* B)

e.g. KO *vs.* WT or mutant *vs.* WT

Direct

Indirect



$$\text{average}\,(\log\,(A/B))$$

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

$$\sigma^2\,/2$$

$$2\sigma^2$$

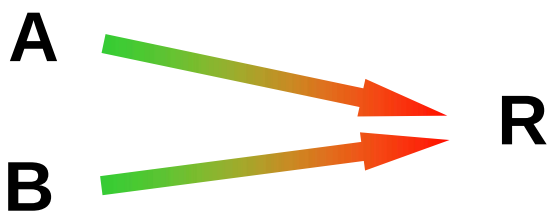
These calculations assume independence of replicates: the reality is not so simple.

Direct vs Indirect - revisited

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

Direct

Indirect



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

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

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

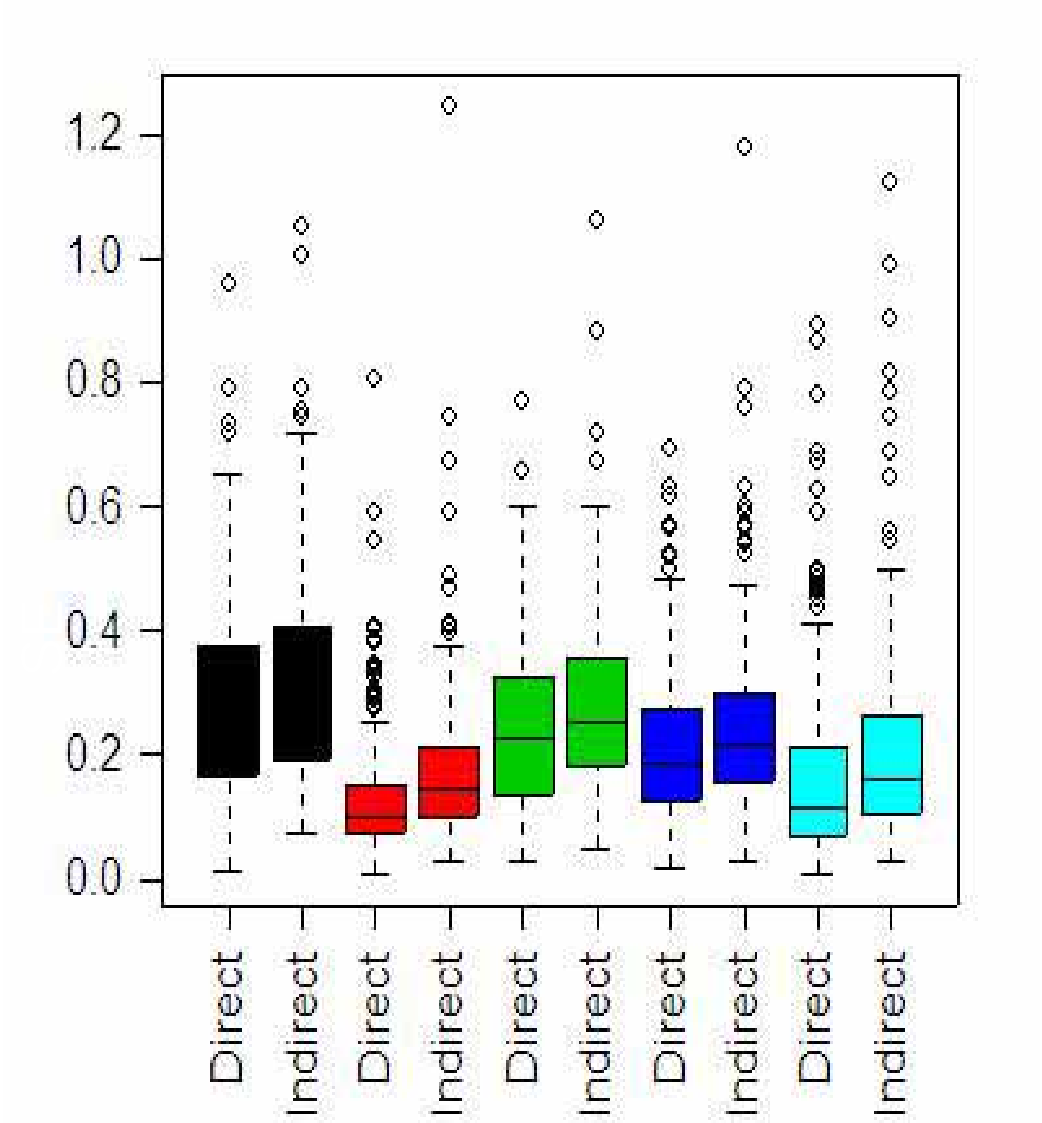
$Var(y) = 2\sigma^2 - 2\chi_1$

$\sigma^2 = 2\chi_1$ 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)
  A) SE for $aveM_{mt}$
  B) SE 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.

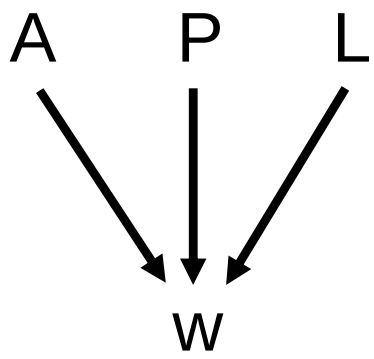
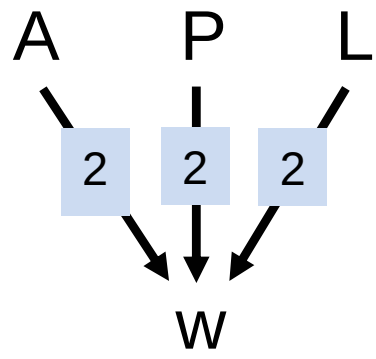
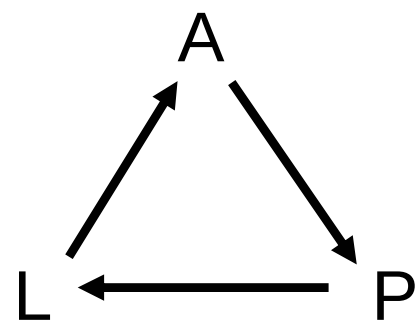
- Not advocating the use of technical replicates in place of biological replicates for samples of interest.

- 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.

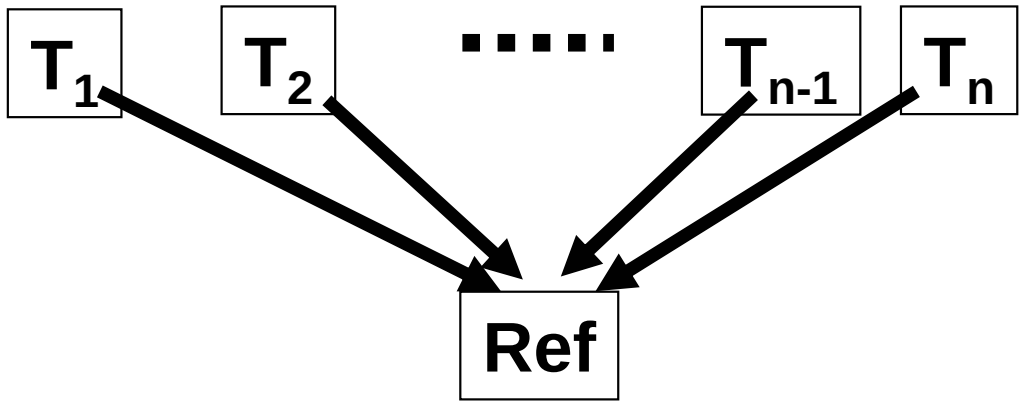
  E.g. which main effects, which interactions.

- 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.

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

For k = 3, 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.

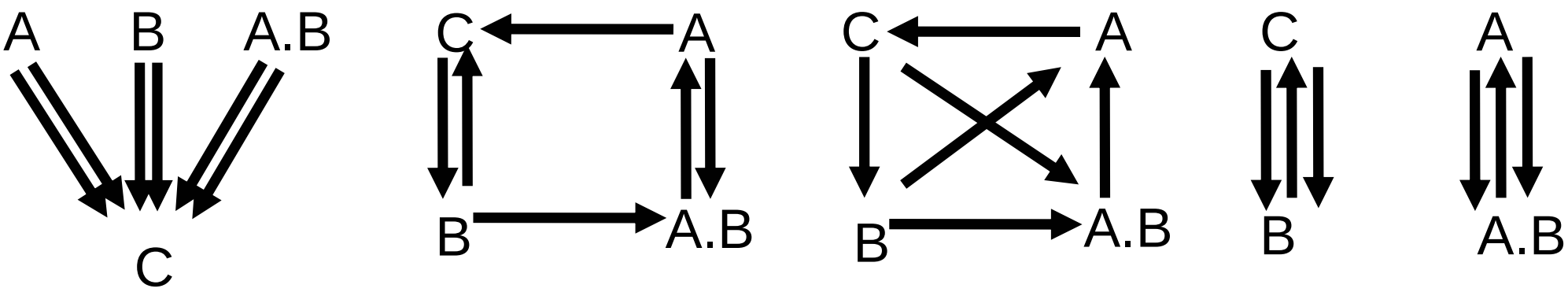
- Advantages:

  Ease of interpretation.

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

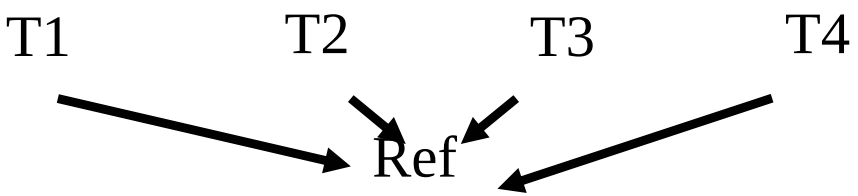
Experiment for which a number of designs are suitable for use

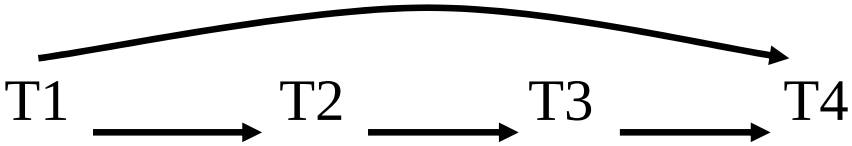
4 samples

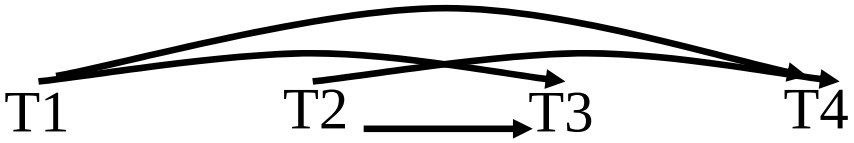


Experiment for which a number of designs are suitable for use

Time Series







2 x 2 factorial

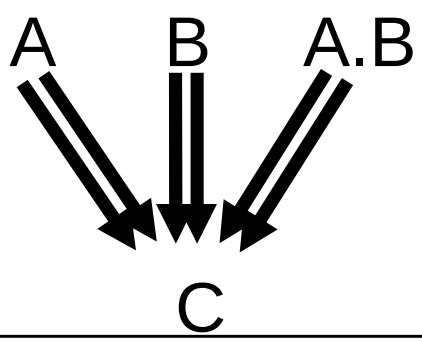
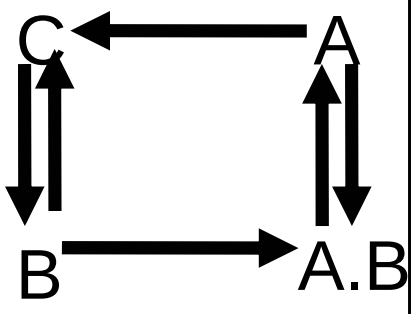
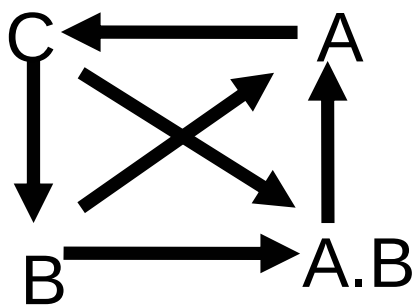
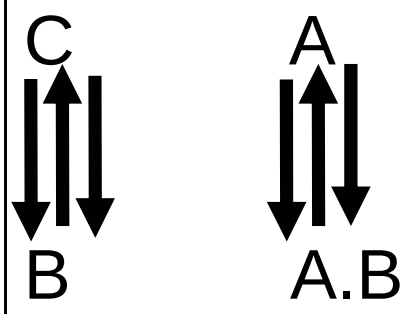
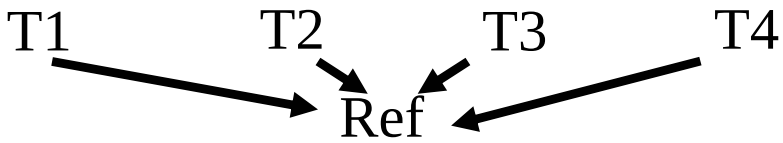
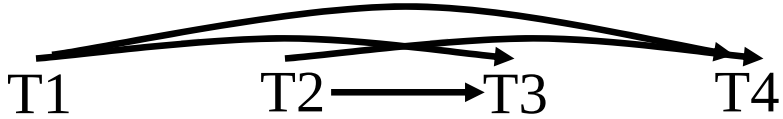
	Indirect	A balance of direct and indirect		
	I) 	II) 	III) 	IV) 
# Slides	N = 6			
Main effect A	0.5	0.67	0.5	NA
Main effect B	0.5	0.43	0.5	0.3
Interaction A.B	1.5	0.67	1	0.67

Table entry: variance

Ref: Glonek & Solomon (2002)

Design choices in time series		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

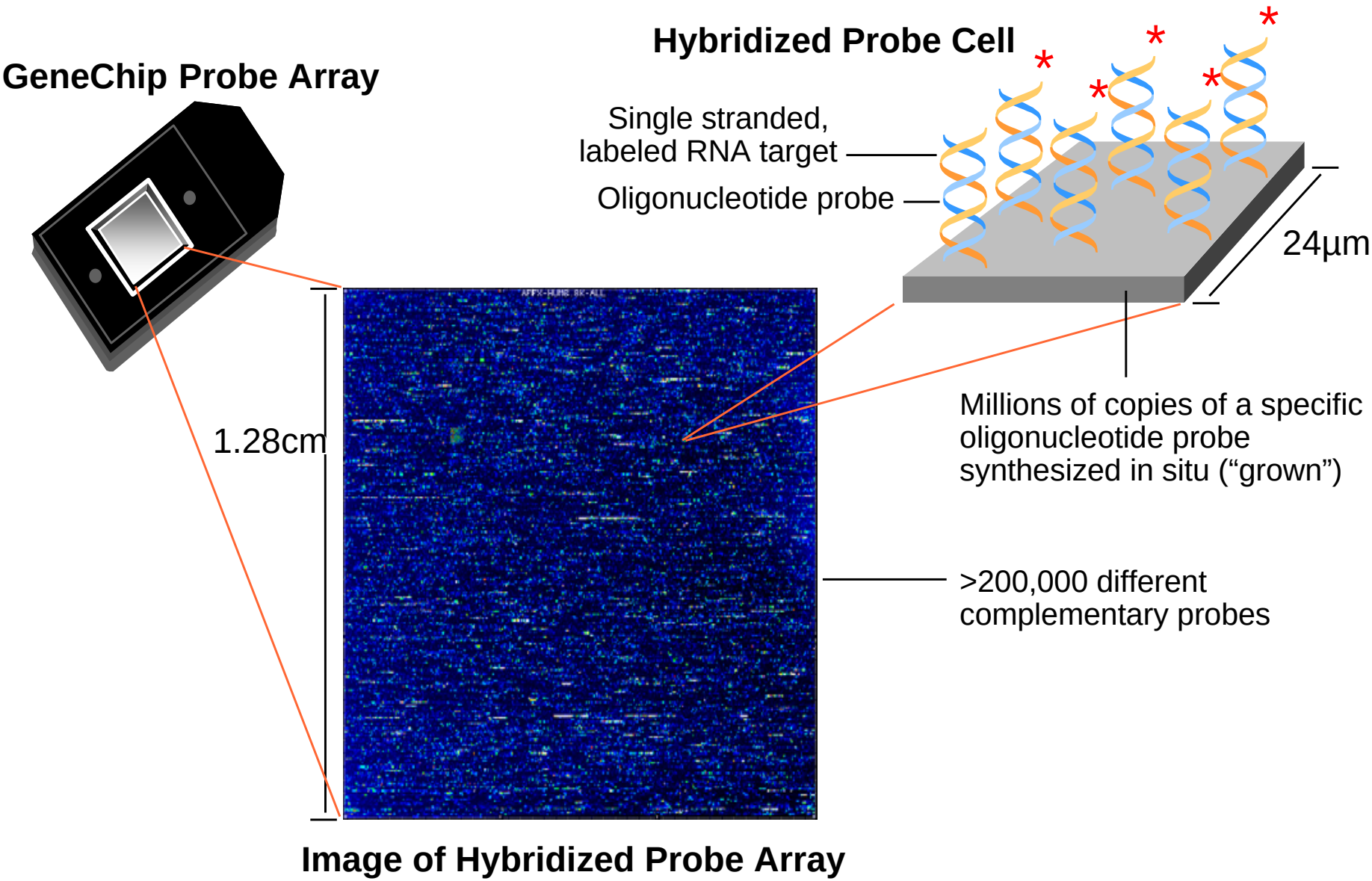
Monday
19. September 05

Affy Chips: PM versus MM and summary information

MGA
Molecular Genome Analysis - Bioinformatics and Quantitative Modeling

dkfz.
DEUTSCHES KREBSFORSCHUNGSZENTRUM
IN DER HELMHOLTZ-GEMEINSCHAFT

Affymetrix GeneChips zusammengefasst



GeneChip® Expression Array Design

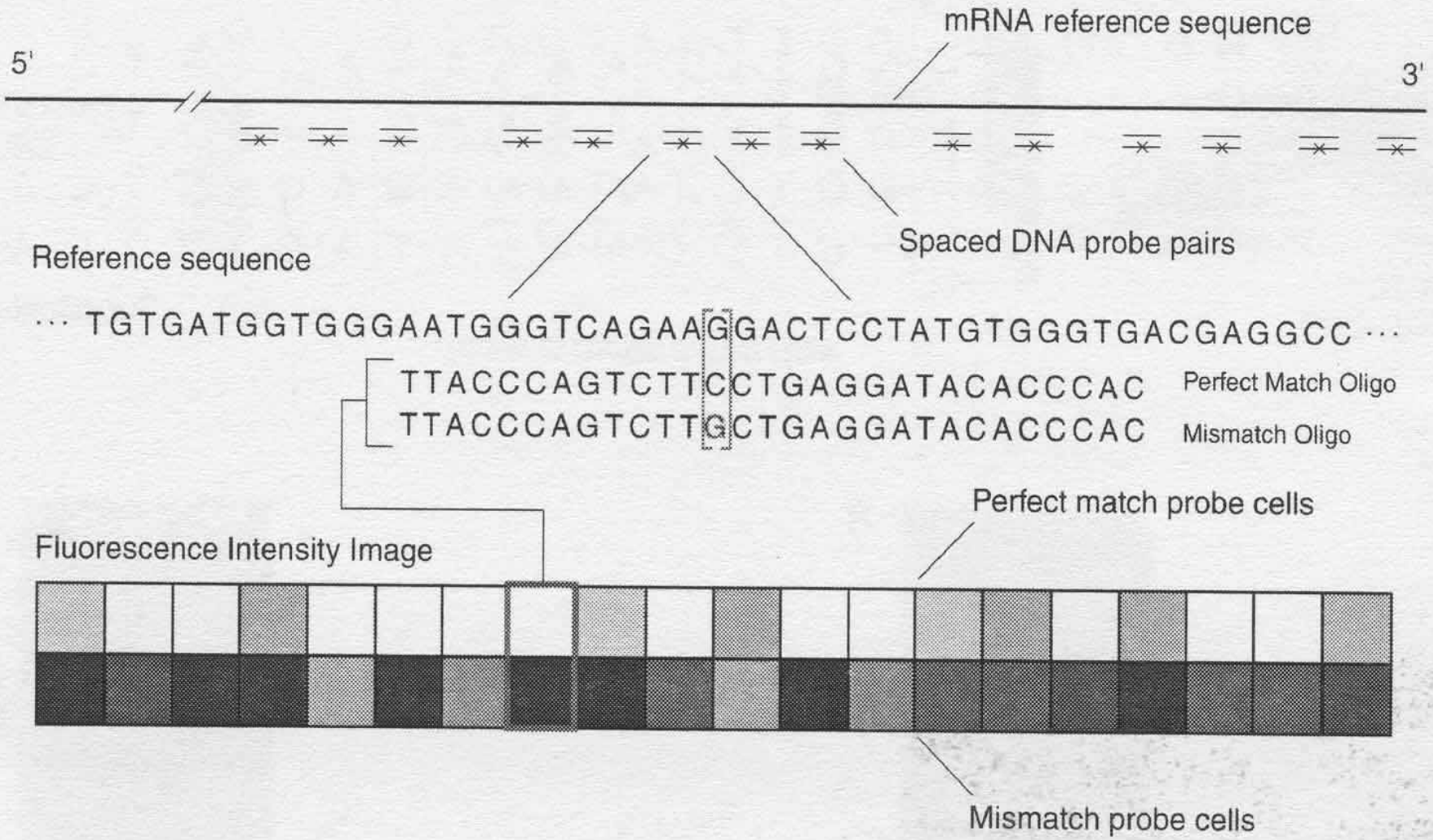
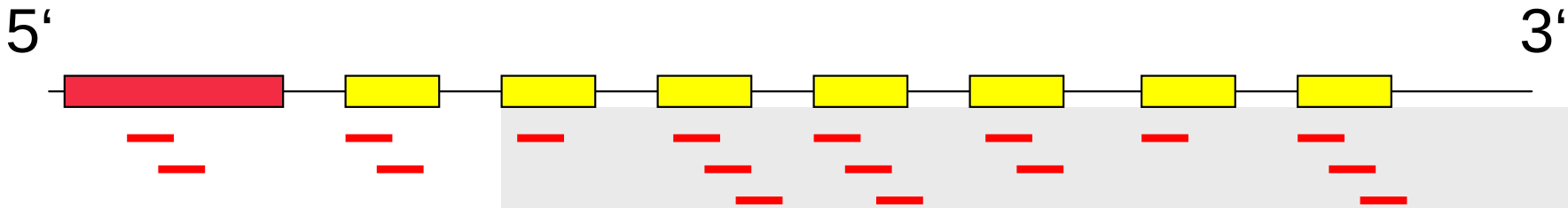


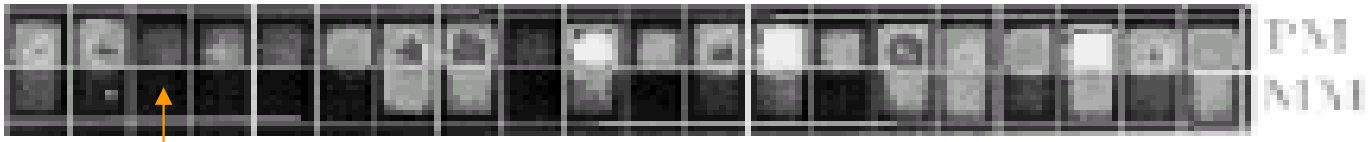
Figure 1-3 Expression tiling strategy

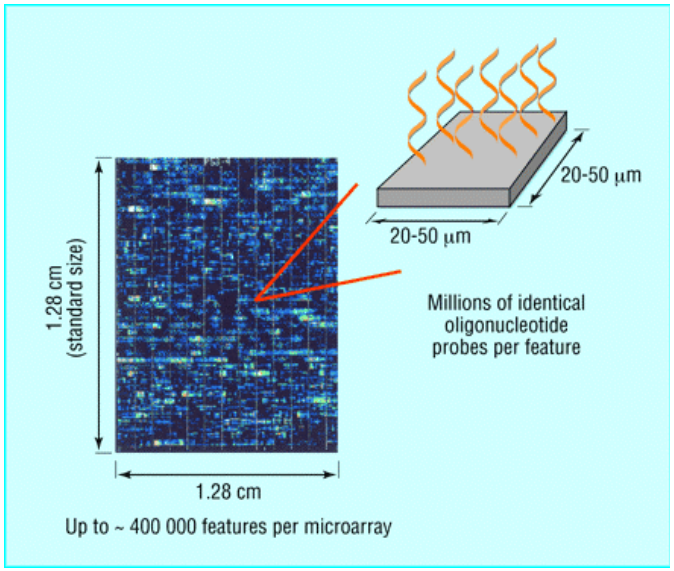
Affymetrix technology



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?

Normalization – complete data methods

- Quantile normalization:
  Make the distribution of probe intensities the same for all arrays.
  $F_{i,normalised}(x) = F^{-1}_{global}(F_i(x))$ (Q-Q-Plot)
- Robust quantile normalization
- Cyclic loess (MA plots of two arrays for log-transformed signals and loess)
- VSN

What is the best approach? Look at criteria provided by the affycomp procedure.

Cope LM, Irizarry RM, Jaffee H, Wu Z, Speed TP, ***A Benchmark for Affymetrix GeneChip Expression Measures,*** *Bioinformatics, 2004, 20:323-31*

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.

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

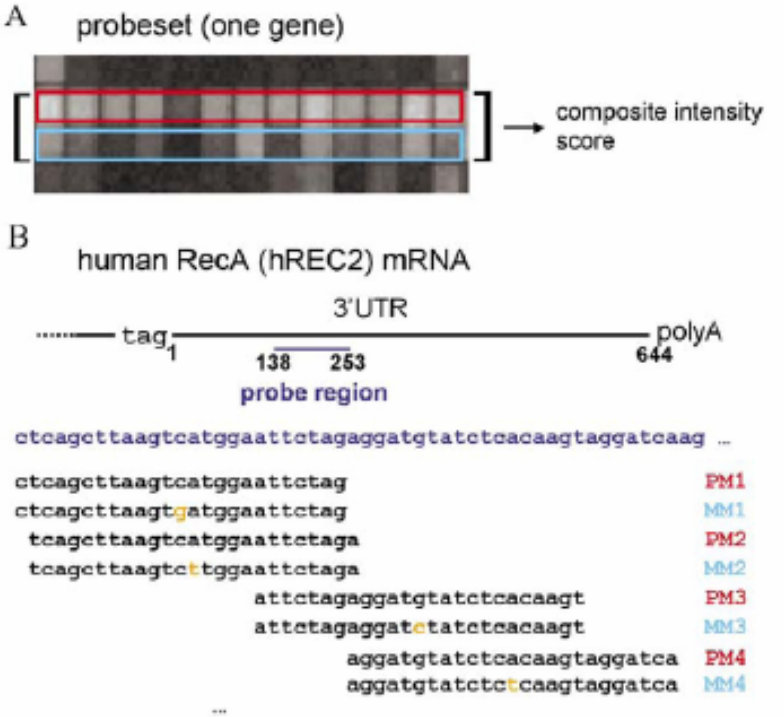


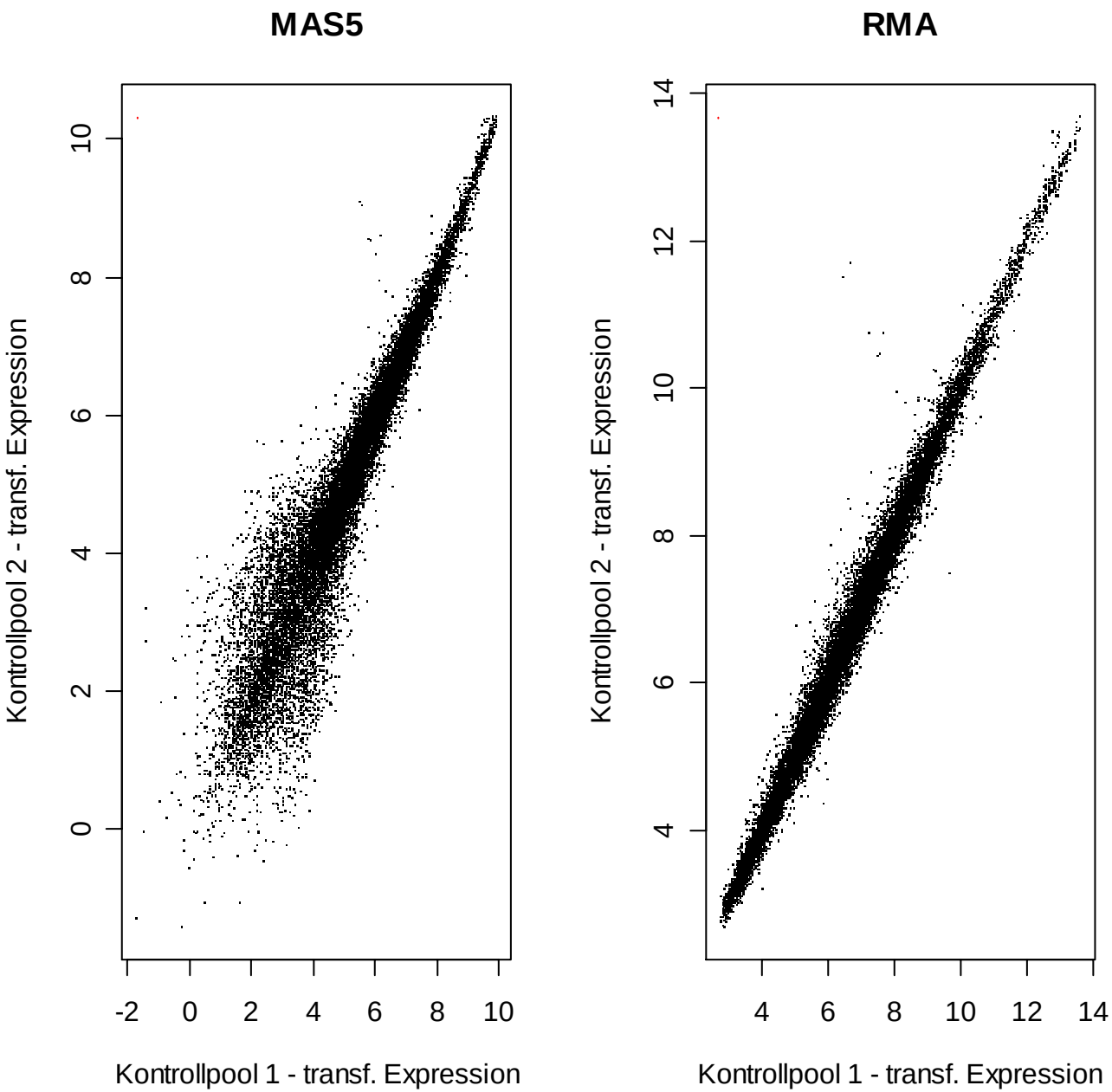
FIG. 1. (Color) Probeset design. (A) The raw scanned image of a typical probeset, with the PM (MM) on the top (bottom) row; higher brightness (white) corresponds to higher abundance of bound RNA molecules. The large variability in probe brightness is clearly visible. (B) Arrangement of probe sequences 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)

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

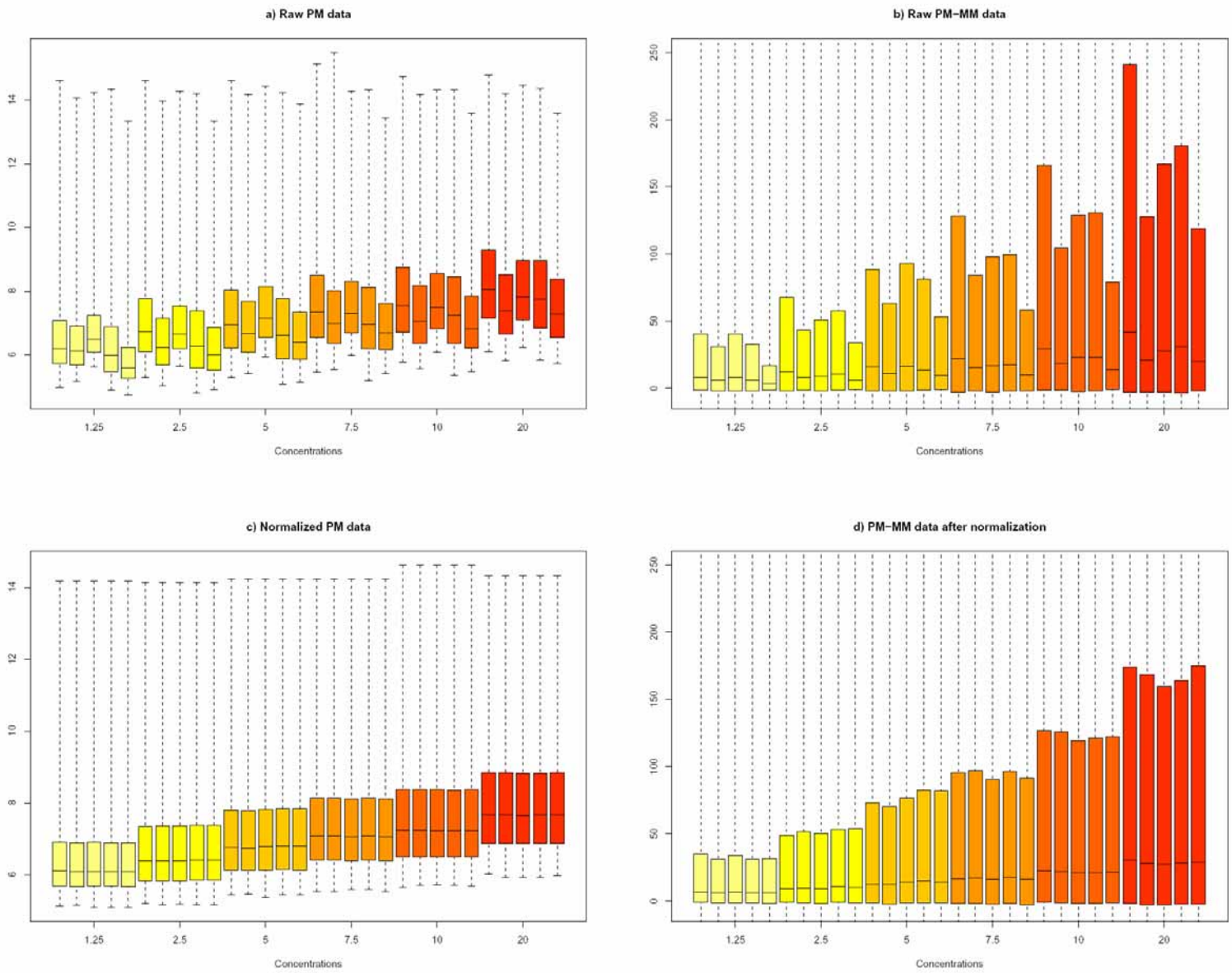


How to approach the quantification of gene expression:Three data sets to learn from

- **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.

Normalization – Baseline Array

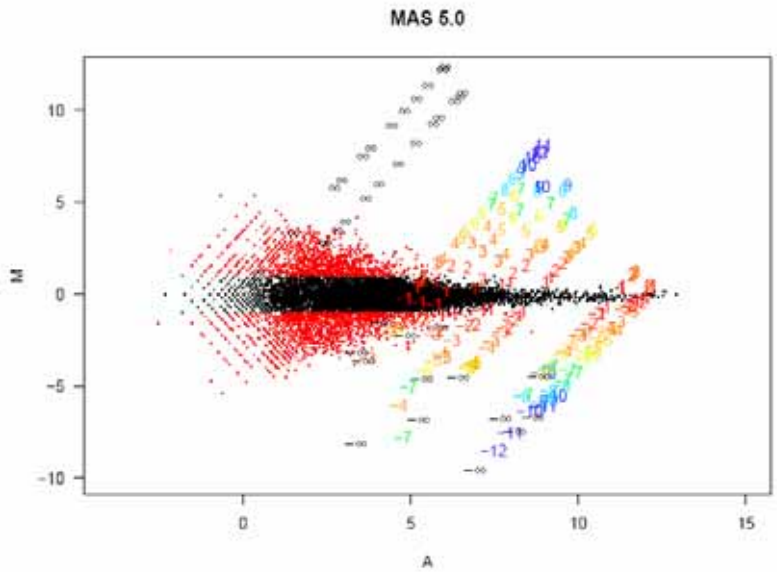
Data C

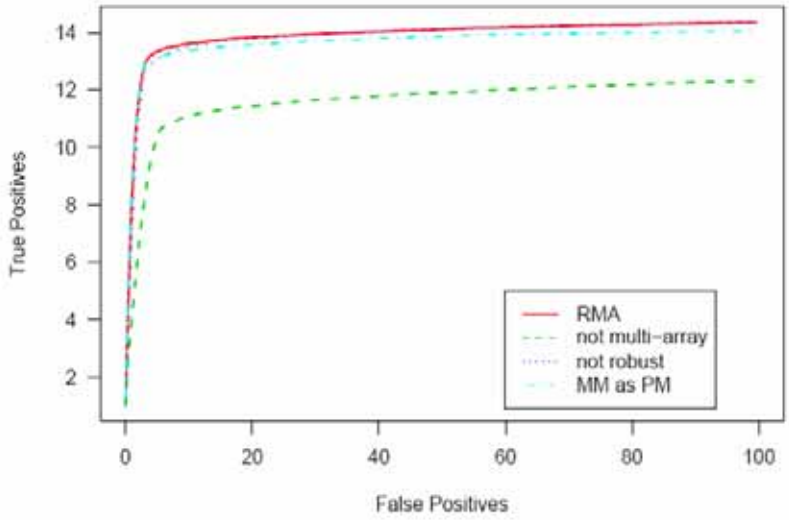


AffyComp

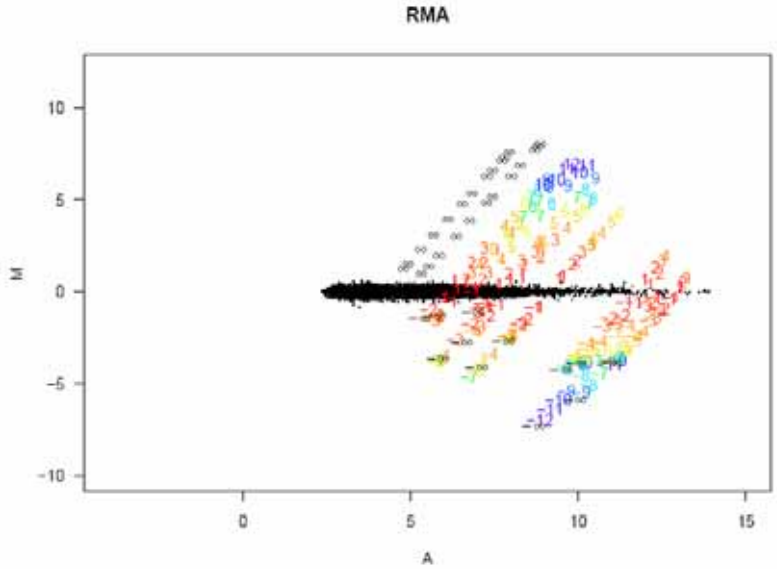
- Graphical tool to evaluate summaries of Affymetrix probe level data.
- Plots and summary statistics
- Comparison of competing expression measures
- Selection of methods suitable for a specific investigation
- Use of benchmark data sets

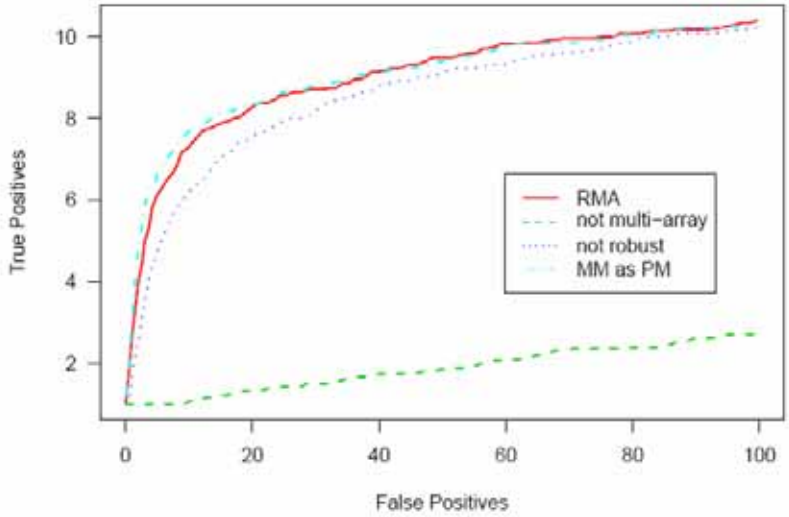
What makes a good expression measure: leads to good and precise answers to a research question.





b) FC=2





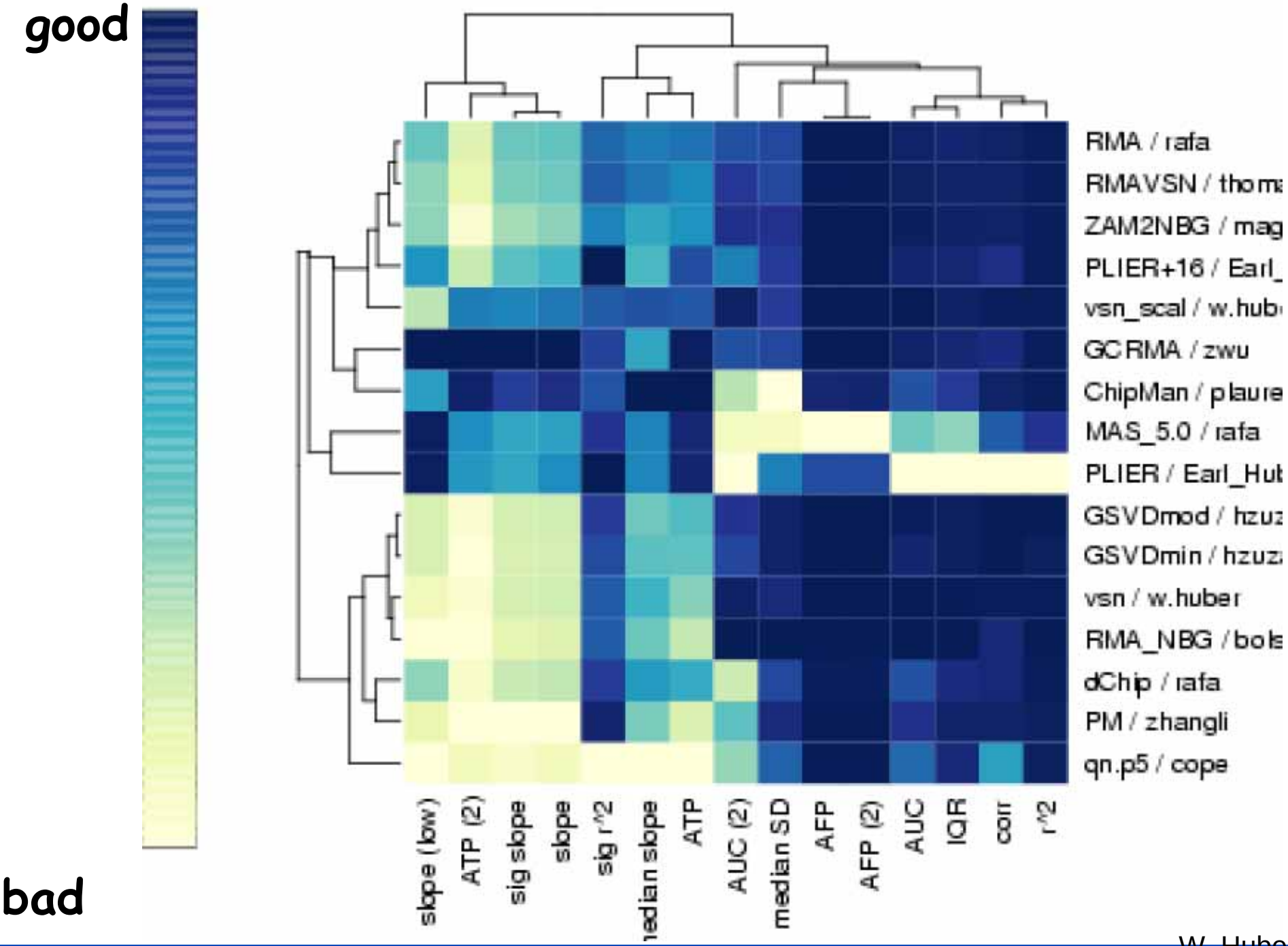
How to create the trapezoid?

ROC changing the cutpoint

```
> 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**
- **Jean Yang**
- **Benedikt Brors**
- **Anja von Heydebreck**
- **Achim Tresch**
- **Rainer König**