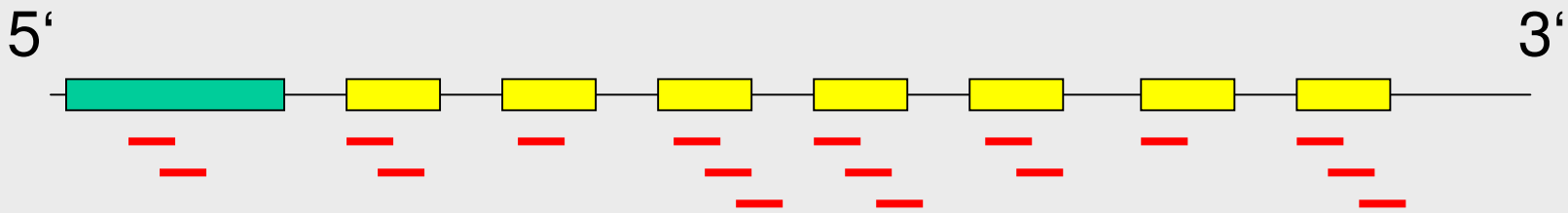
Affy Chips: Cel-file versus summary information

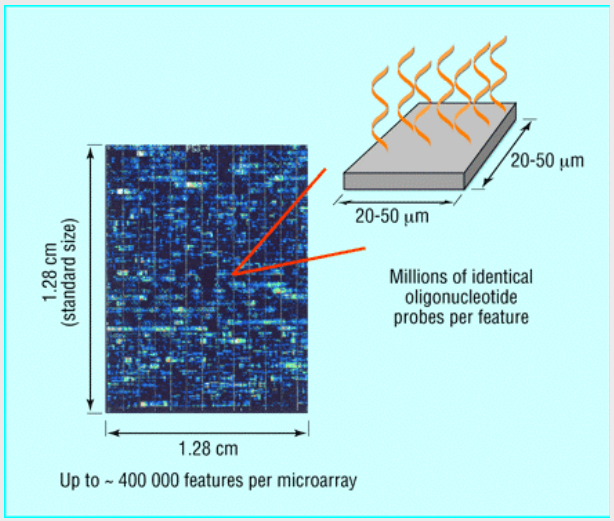
Ulrich Mansmann
Department of Medical Biometrics and Bioinformatics
Medical School, University of Munich

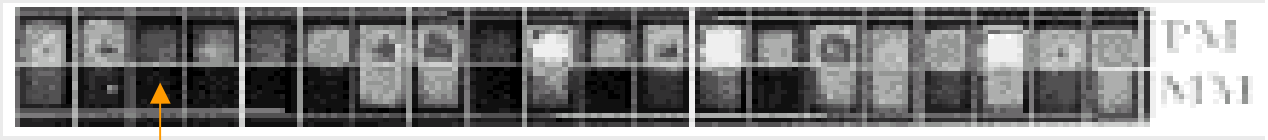
Affymetrix technology



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

Low – level -Analysis

- Preprocessing signals: background correction, normalization, PM-adjustment, summarization.
- Witt E, McClure J (2004) Statistics for Microarray: design, analysis, and inference, Chichester, John Wiley & Sons
- Noise and Bias
- Differences in sample preparation, variation during mRNA extraction and isolation
- Manuifacturing of the array: variation in hybridization efficiency, abundance
- Normalization on probe or probe set level?
- Which probes / probe sets used for normalization
- How to treat PM and MM levels?
- Linear or non-linear normalization?

Expression measures based on d = PM-MM

- AvDiff

$$\text{AvDiff} = \frac{1}{\#A}\sum_{j\in A}(PM_j - MM_j)$$

A contains only probes with d not an outlyer

- Li & Wong (dChip)

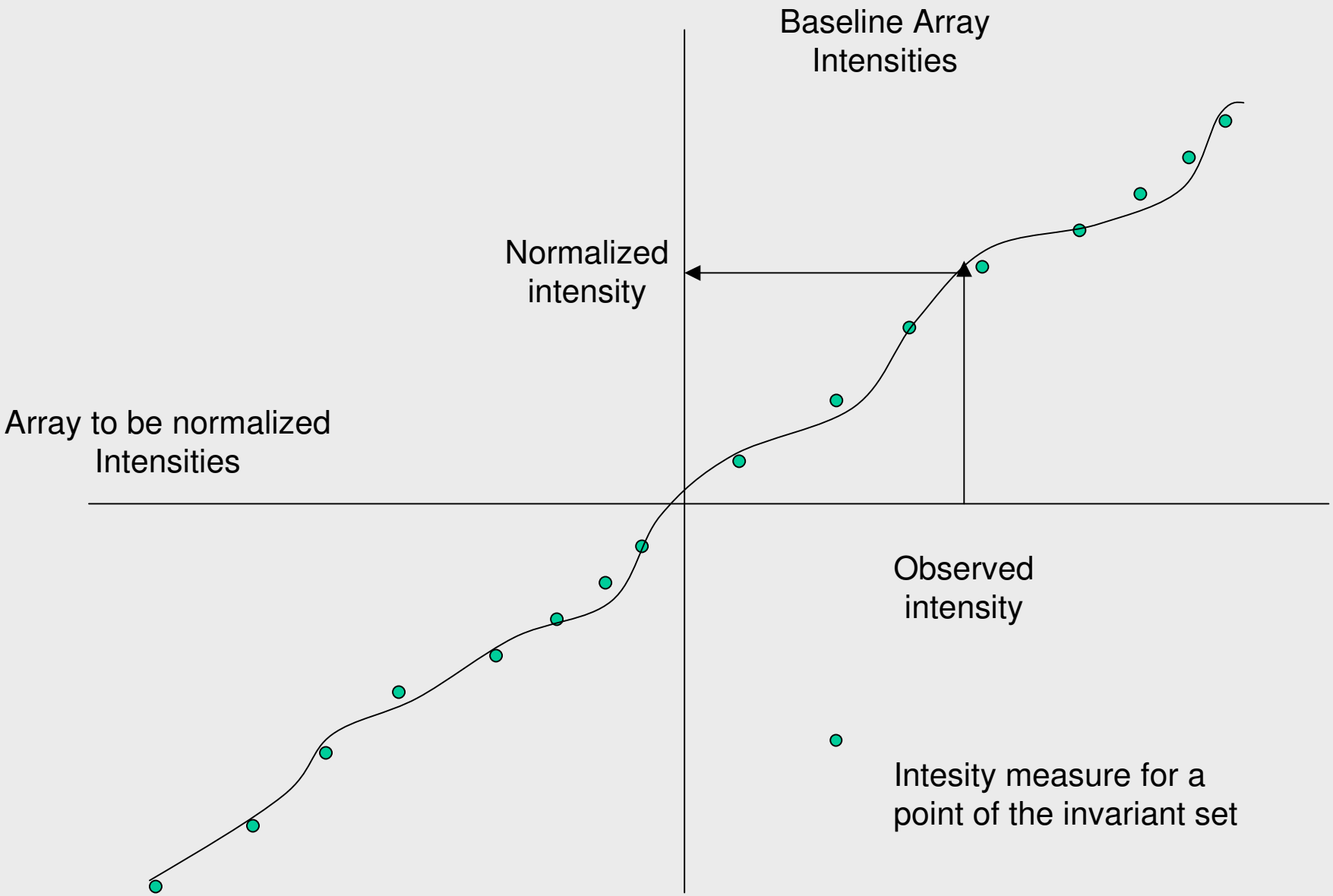
$Pm_{ij} - Mm_{ij} = \theta_i\varphi_j+\varepsilon_{ij}$ $\quad \varepsilon_{ij} \sim N(0,\sigma^2)$
MLE for θ_i gives expression measure

- MAS5

signal = Tuckey Biweight[$PM_j - CT_j$]
$CT_j = \min(MM_j, PM_j)$

Normalization – Baseline Array

- **Scaling:**
  First array is baseline: m_{base} mean intensity of baseline array, m_i (i=1,…,n) mean intensity of array i, scale factor for array i: $\beta_i = m_{base} / m_i$.
  Normalized intensity an array i: $x_{i,norm} = x_i \cdot \beta_i$.
  Two options: apply normalisation to probes or after summarization to probe set measures.
- **Invariant set:**
  PM probe values are used only.
  Probes which are not differentially expressed (unknown). It is assumed that PM probe signals which are not differentially expressed in two arrays have similar intensity ranks (r).
  Point's proportion rank difference (PRD): $|(r_{k,i}-r_{k,base})|(\#probes)$
  Small $PRD_{k,i}$ (<0.003, 0.007), include probe into invariant set, cycle through all arrays, use invariant set to create array specific calibration curve by running median.
  Estimate the non-differential expressed genes.


Baseline Array
Intensities
Normalized
intensity
Array to be normalized
Intensities
Observed
intensity
Intesity measure for a
point of the invariant set

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)
- Contrast
- RMA
- VSN

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

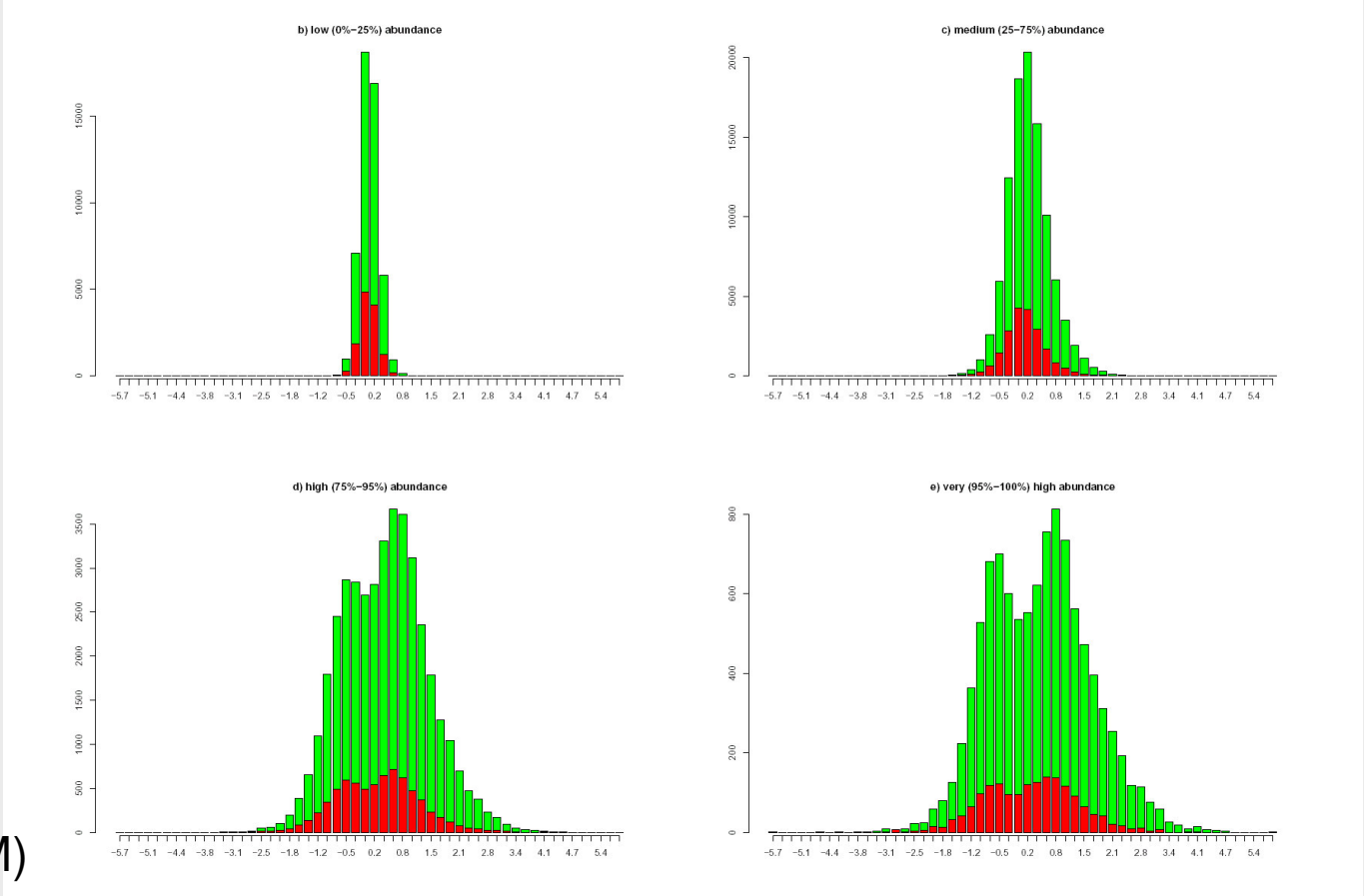
Cope LM, Irizarry RM, Jaffee H, Wu Z, Speed TP, ***A Benchmark for Affymetrix GeneChip Expression Measures,***
Bioinformatics, 2004, 20:323-31

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.

Feature of probe level data

- MM grows with PM
- Many MM >> PM
- log scale stabilises variance



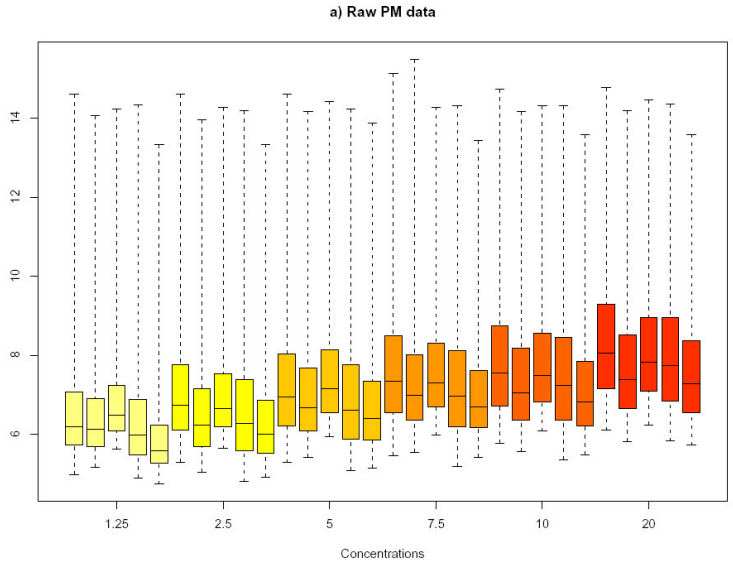
$M = \log_2(PM/MM)$

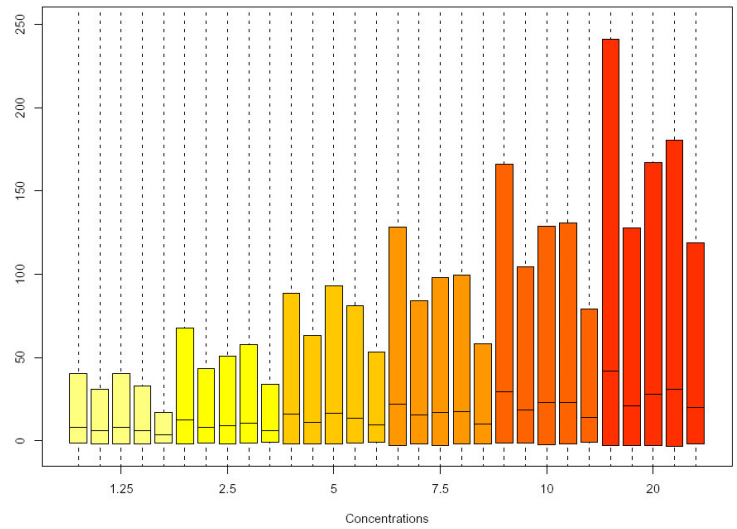
$A = 0.5 \cdot \log_2(PM \cdot MM)$

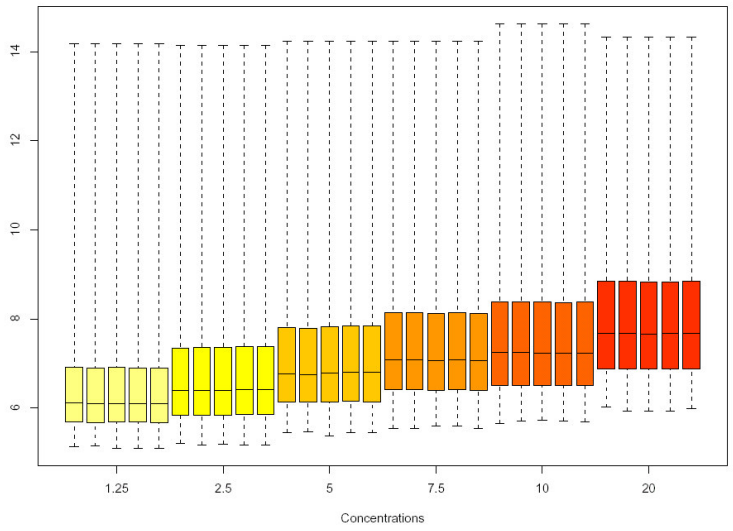
abundance

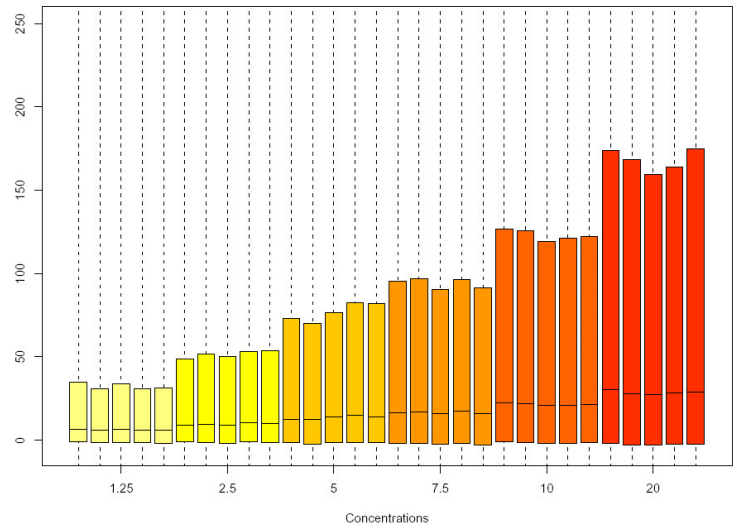
red: incorrect sequenced probes; green: correct probes, x-axis: M

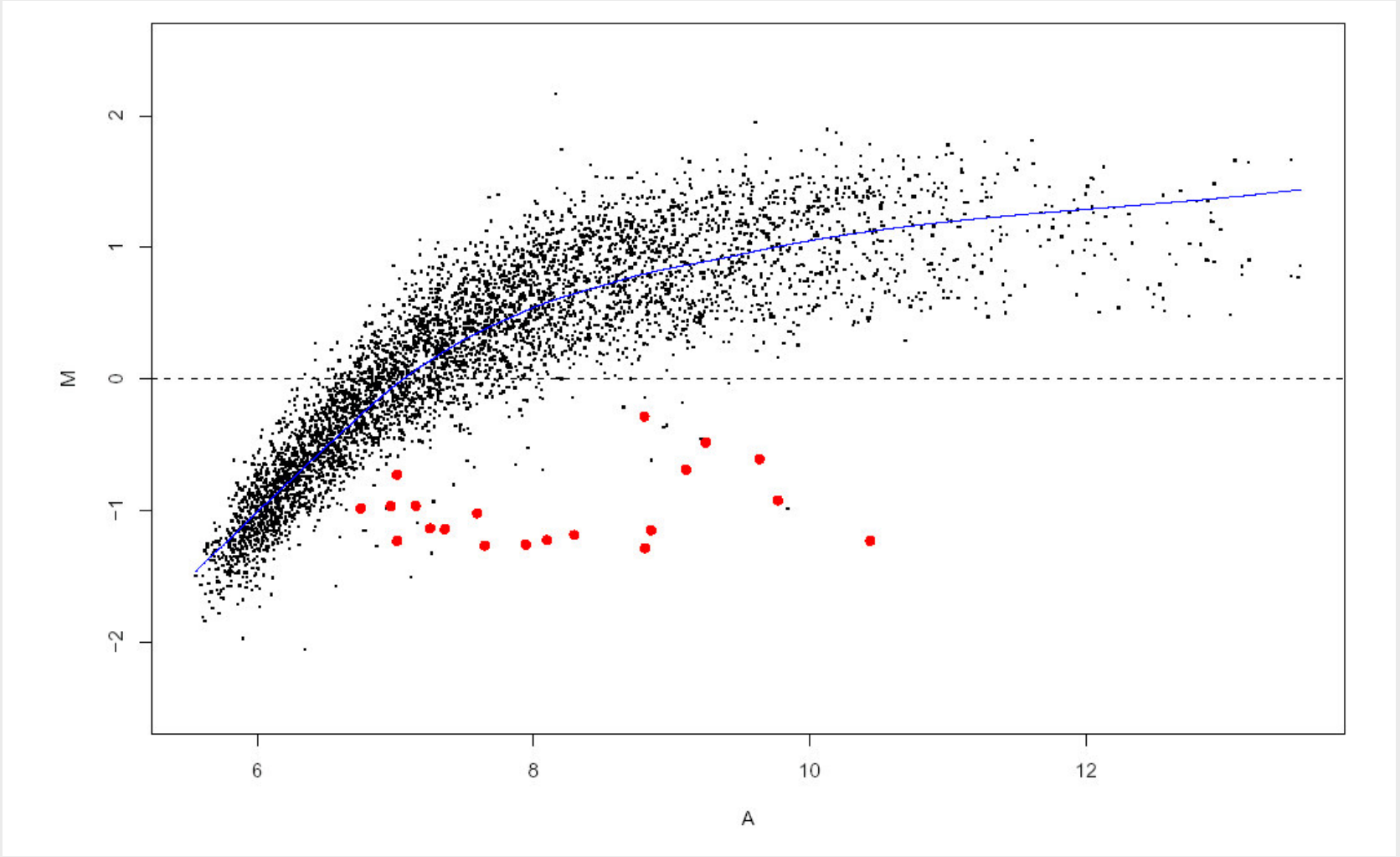
Dilution data









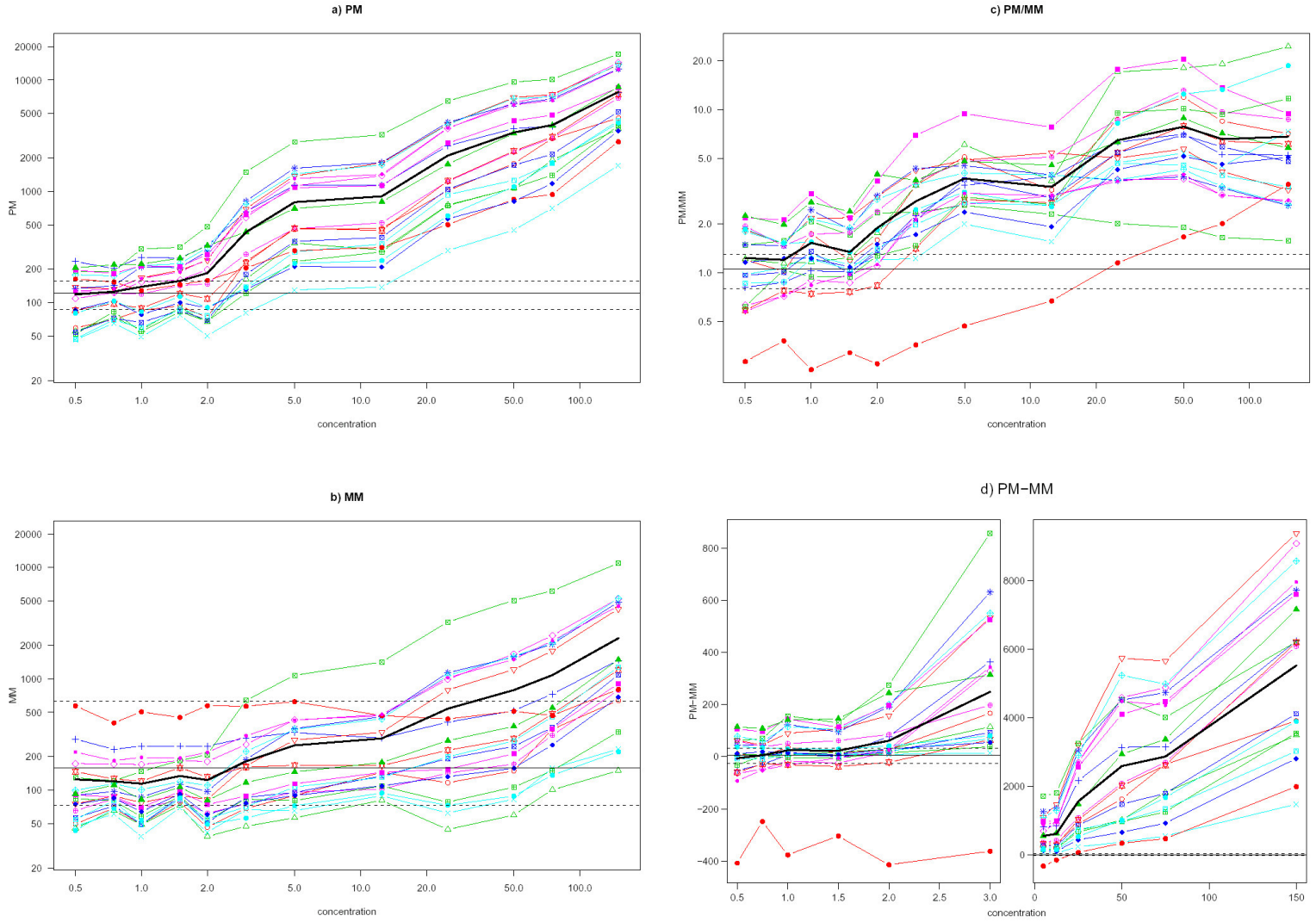


$M = \log_2(PM_1 / PM_2)$ $A = 0.5 \cdot \log_2(PM_1 \cdot PM_2)$

Bland-Altman plots

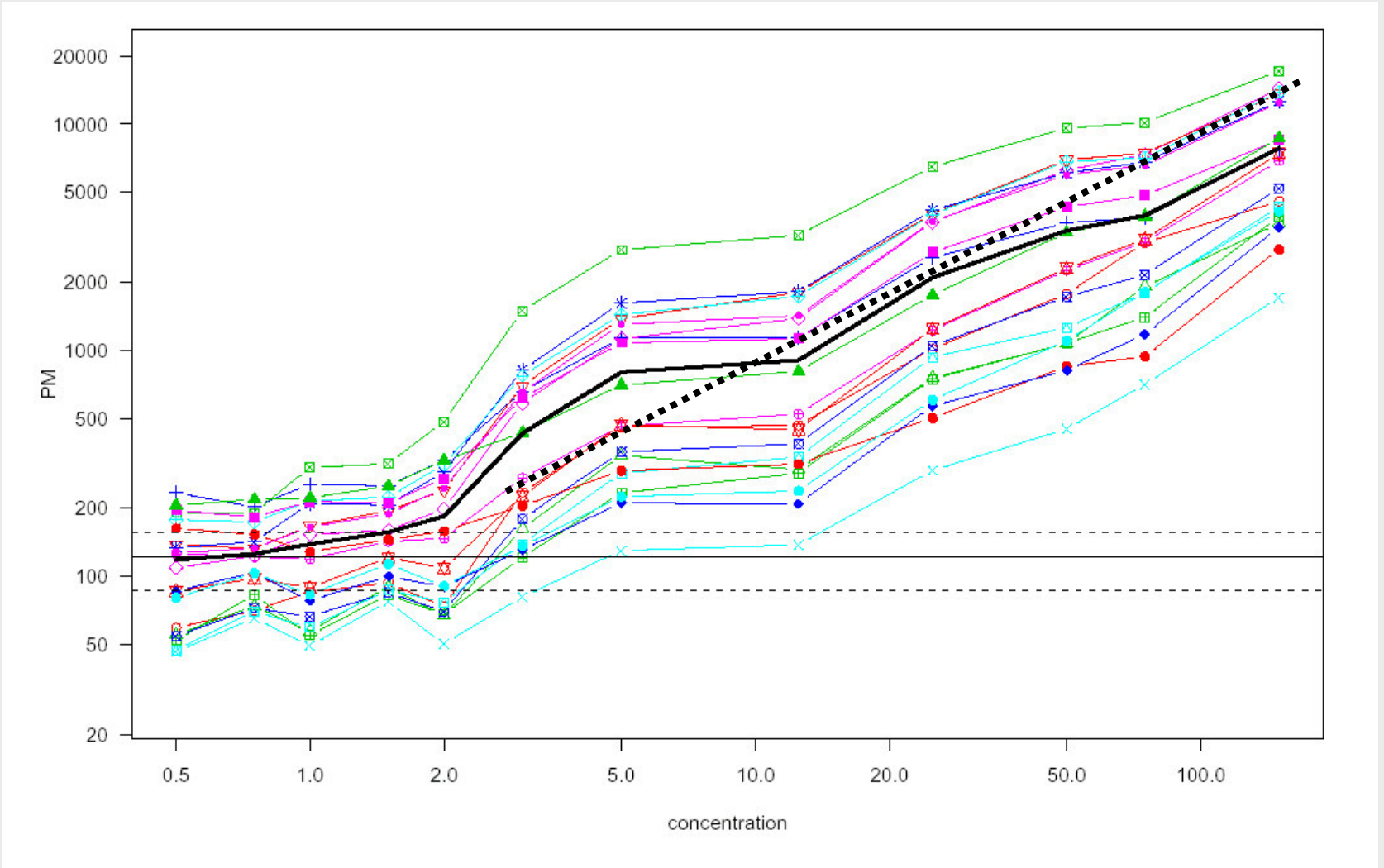
Arguments against the use of d = PM-MM

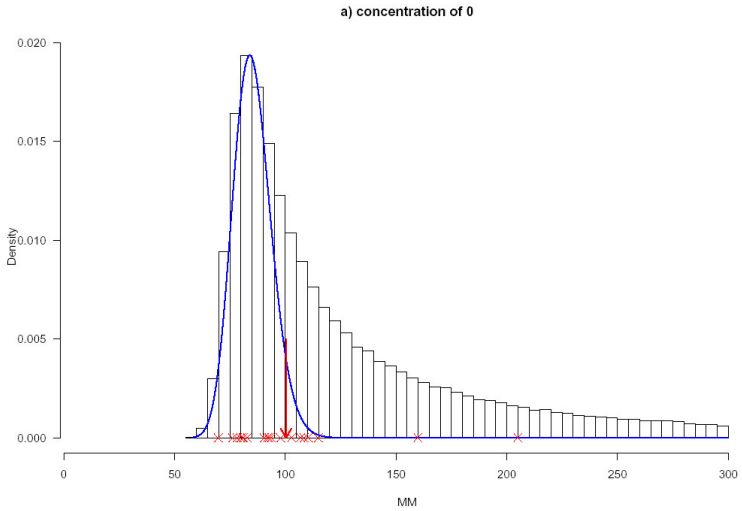
- Difference is more variable. Is there a gain in bias to compensate for the loss of precision?
- MM detects signal as well as PM
- PM / MM results in a bias.
- Subtraction of MM is not strong enough to remove probe effects, nothing is gained by subtraction

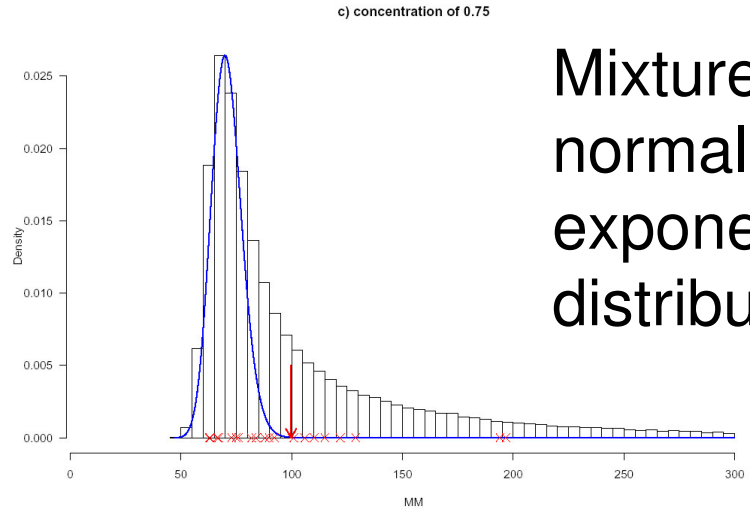

a) PM
PM
concentration
c) PM/MM
PM/MM
concentration
b) MM
MM
concentration
d) PM−MM
PM−MM
concentration

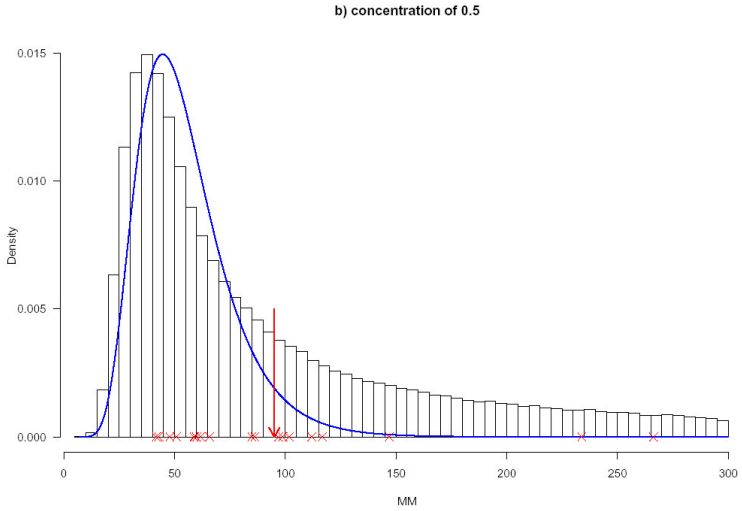
Expression measure based on PM only

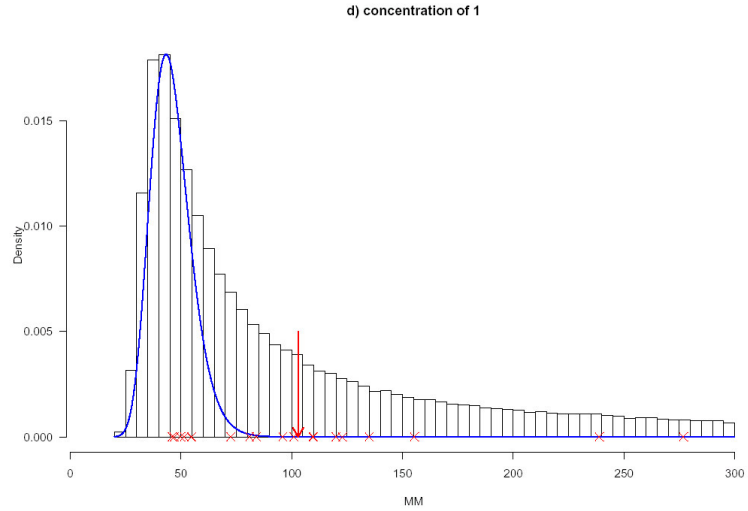
- Use PM values but correct for unspecific binding and background (optical) noise.
  For small signals uncorrected values may give misleading results:
  $\log_2(100+2s)-\log_2(100+s)$ versus $\log_2(2s)-\log_2(s)$
- $PM_{ijn} = bg_{ijn} + s_{ijn}$
- Basic idea:
  Correct by $PM_{ijn} - b_i$ with $\log_2(b_i)$ equal to the mode of $\log_2(MM)$
- Advanced idea:
  $B(PM_{ijn}) = E[s_{ijn}| PM_{ijn}]$ with s_{ijn} exponential and bg_{ijn} normally distributed. This problem has an explicit solution and gives a closed form transformation B.


20000
10000
5000
2000
1000
500
200
100
50
20
PM
0.5
1.0
2.0
5.0
10.0
20.0
50.0
100.0
concentration









Mixture of normal and exponential distribution

The RMA procedure

- Robust multi-array average
- Background corrections for array using transformation B
- Normalise the arrays by using quantile normalisation
- Use the background adjusted, normalised, log-transformed PM intensities (Y) and follow a linear model:

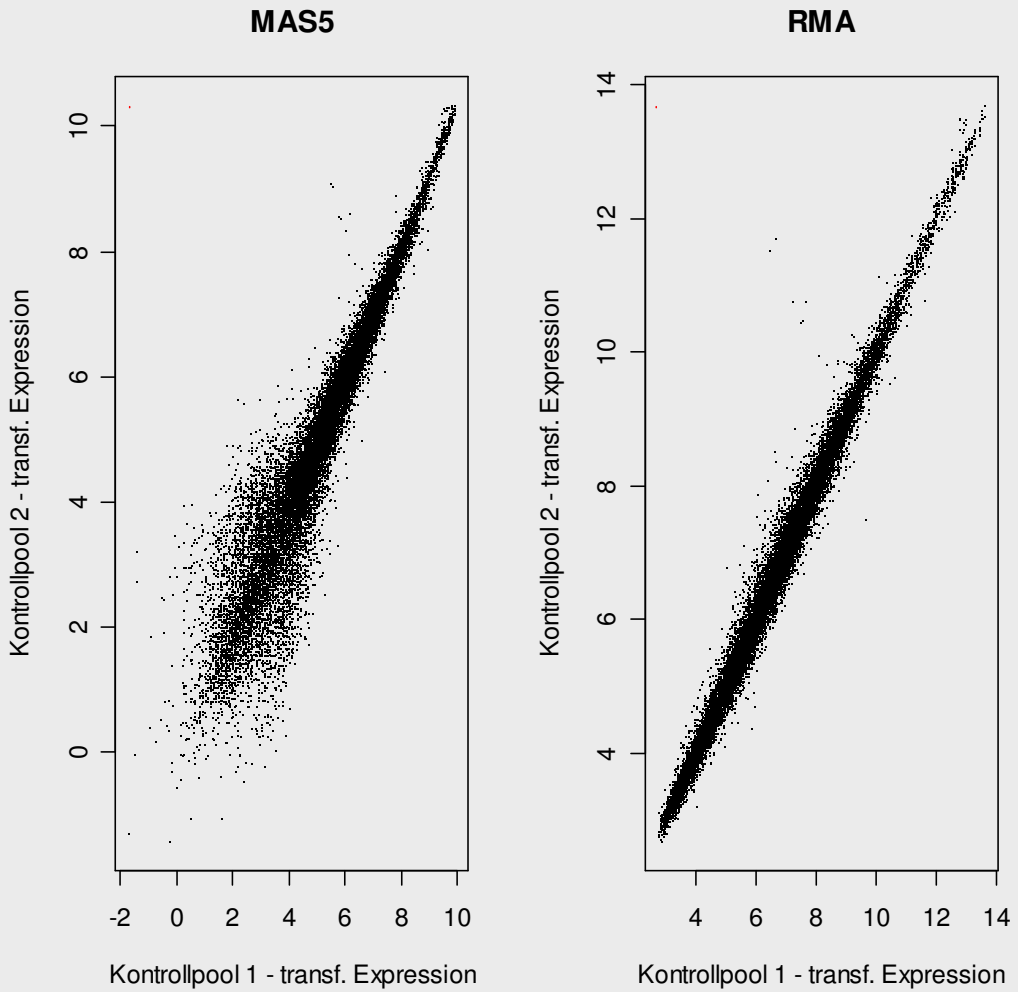
$$Y_{ijn} = \mu_i + \alpha_{ijn} + \varepsilon_{ijn}$$

i – gene
j – probe
n - array

where α_{jn} is the probe affinity, $\Sigma\ \alpha_{ijn} = 0$ for all n
μ_i is the log scale expression level
ε_{ijn} is an error with mean 0

Irizarry et al. (2002) www.biostat.jhsph.edu/~rirzarr/affy

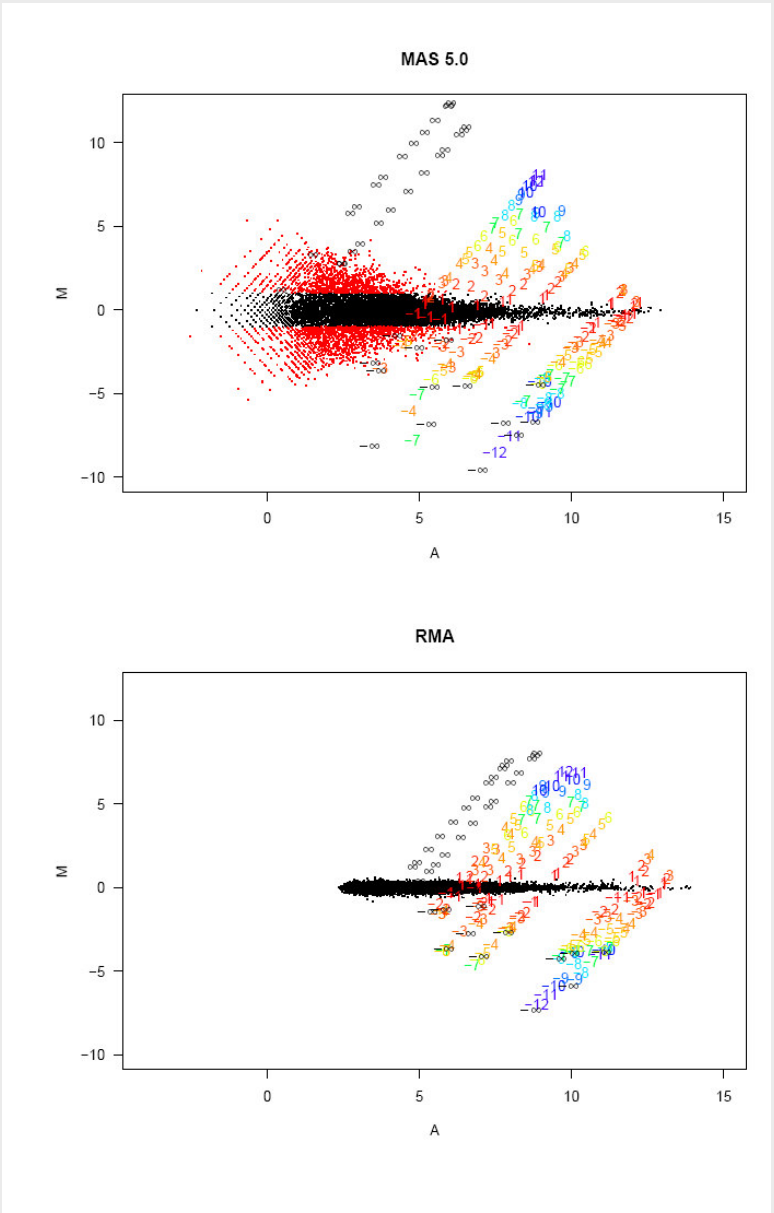
Example LPS: *Expression Summaries*



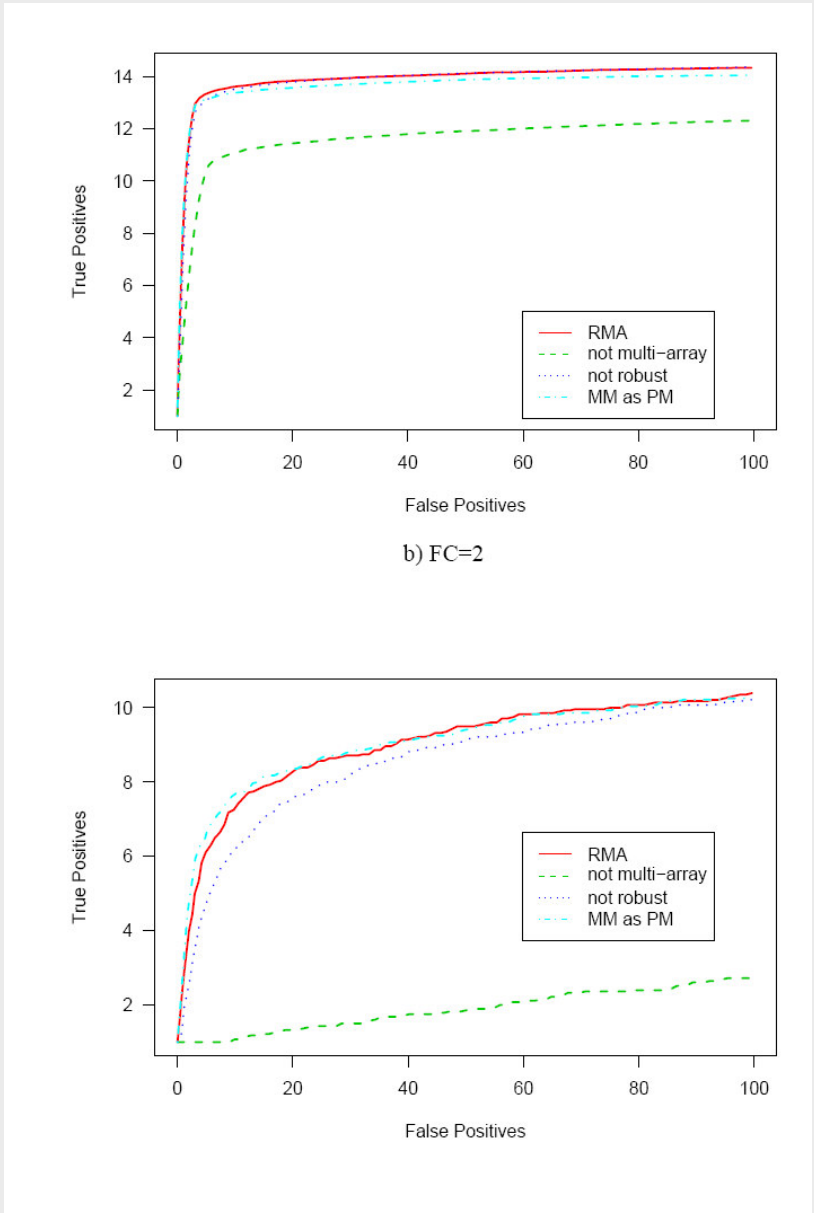
AffyComp

- 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.



How to create the trapezoid?

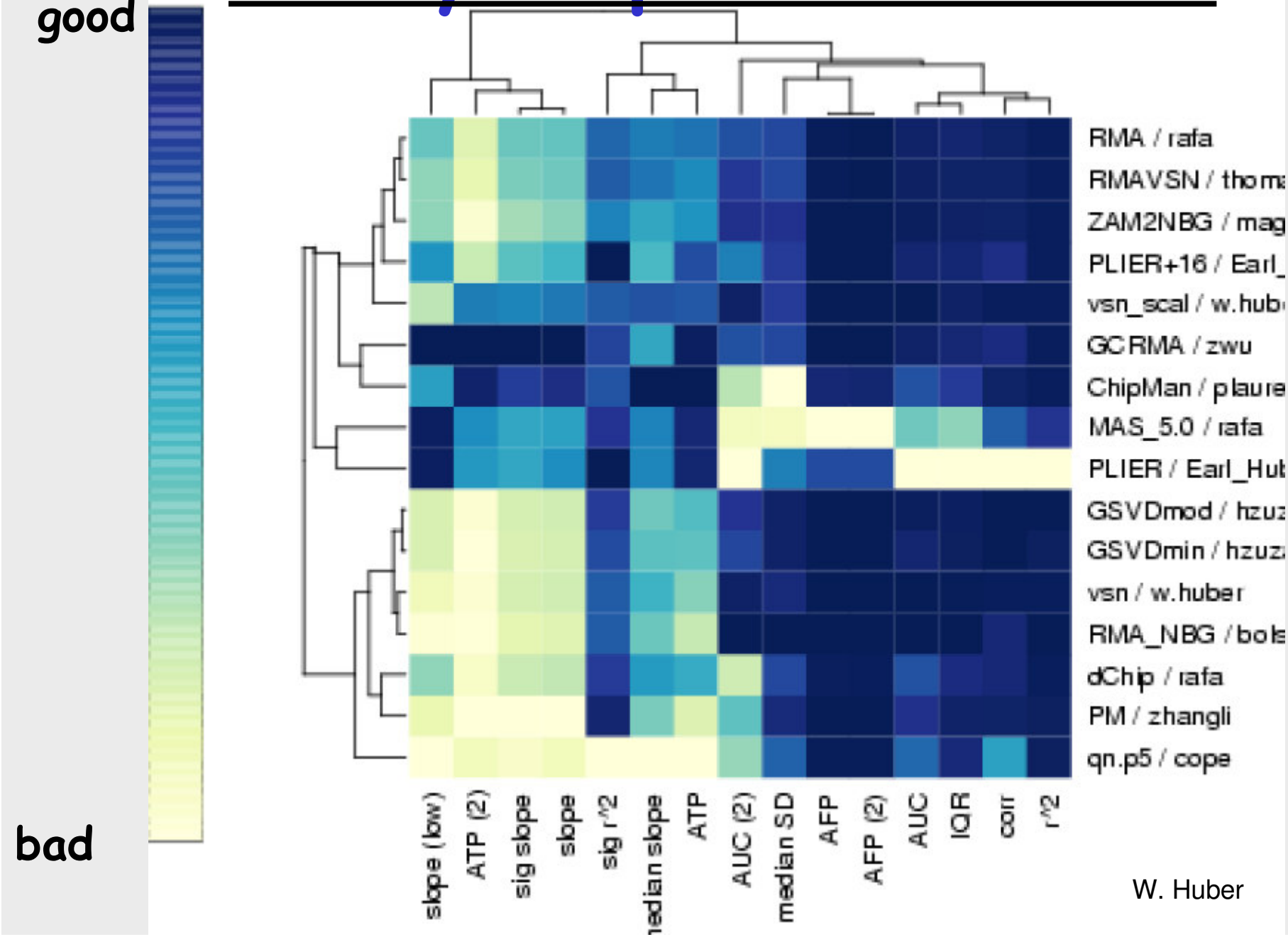


ROC changing the cutpoint

> 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)

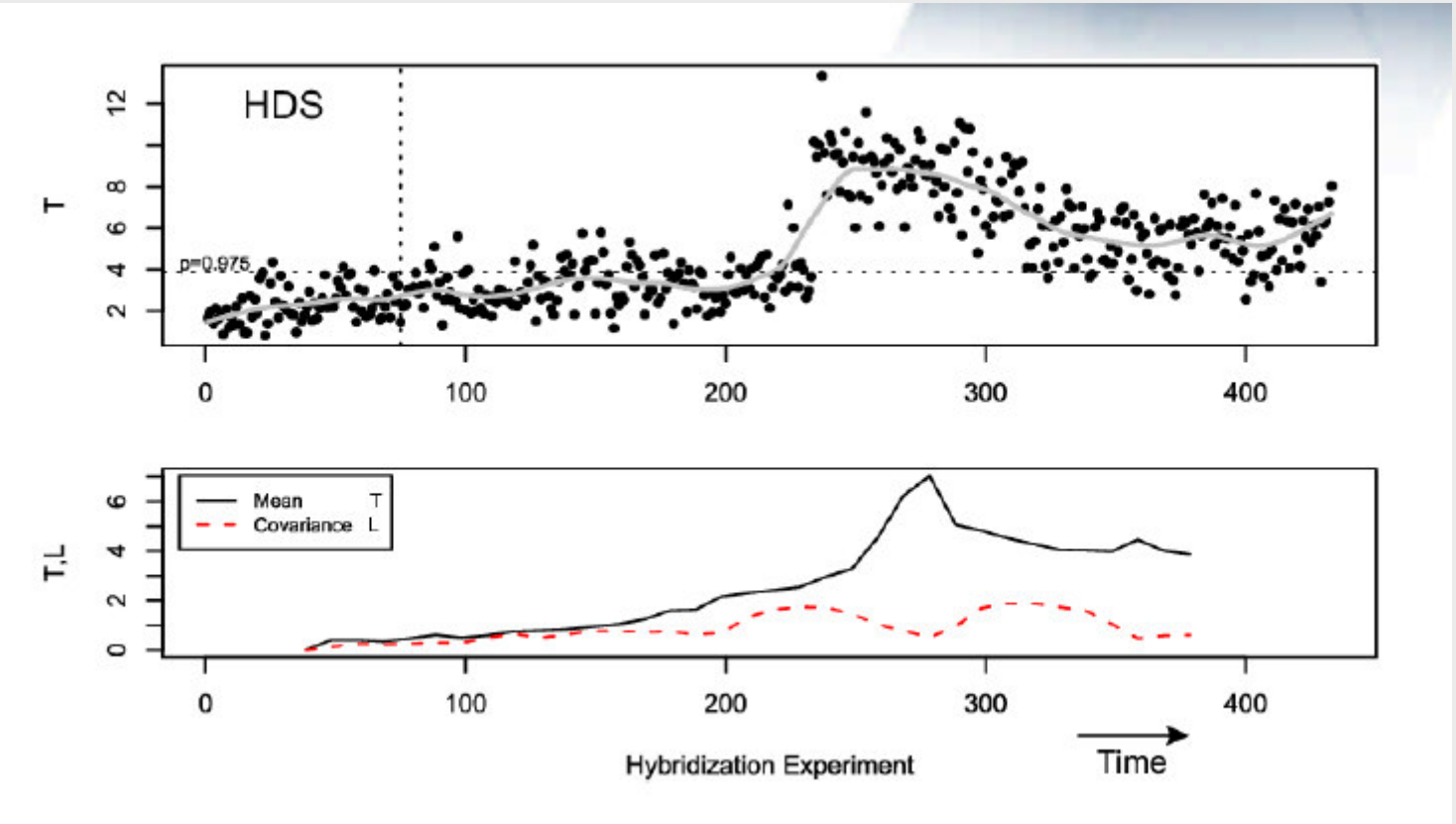


W. Huber

Process Control for large scale experiments

- Model F. et al. (2002) Bioinformatics, 18:S155-163
- HDS: historical data set, variables of a process for some time under perfect working conditions. CDS: current data set.
- How far is the current state of the process away from the perfect state?
- Distance measure: Hotelling's $T^2 = (m-\mu)^t S^{-1} (m-\mu)^t$
  parameter μ and S estimated from the HDS
- Additional tasks:
  1. outlier treatment with robust Principle Component Analysis (rPCA)
     The estimates μ and S are not robust against outliers
  2. For fewer arrays than number of gene expression, the sample covariance matrix S is singular and not invertible.
     PCA is used to reduce dimensionality of the measurement space.
- Calculate an upper control limit to initiate interventions in the ongoing process.
- In order to see whether an observed change in T^2 comes from a simple translation, it is of interest to compare the two sample covariances between HDS and CDS.
  A LRT for different covariances is used, calculates statistic L (H_0: L=0)

Process Control for large scale experiments



T^2 control chart of ALL/AML study. Over the course of the experiment a total of 46 oligomeres for 35 different CpG positins had to be re-synthesized.

They all talked at once, their voices insistent and contradictory and impatient, making of unreality a possibility, then a probability, than an incontrovertible fact, as people will when their desires become words.

Willian Faulkner, *The sound and the fury*, 1929