Microarray Normalization

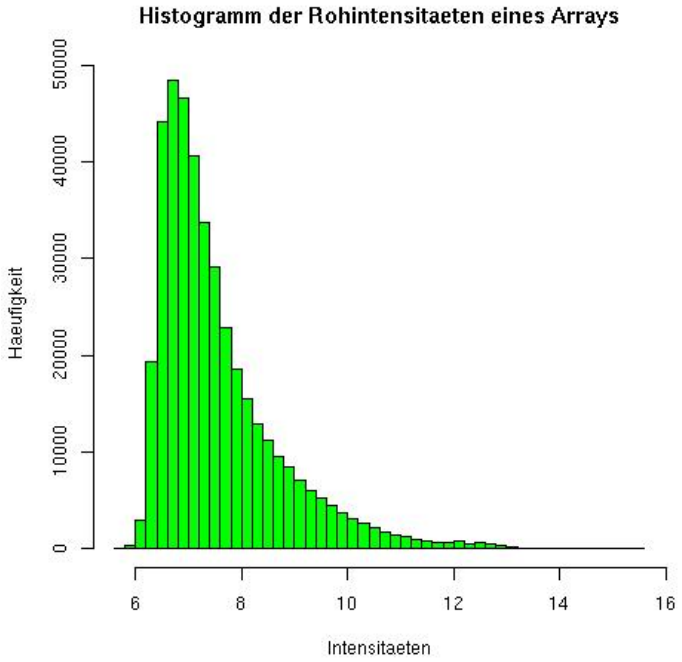
Methods Course: Gene Expression Data Analysis

-Day Two –

Rainer Spang

Basic Statistics

Histogram of log gene expression levels of the same sample across 30.000 genes

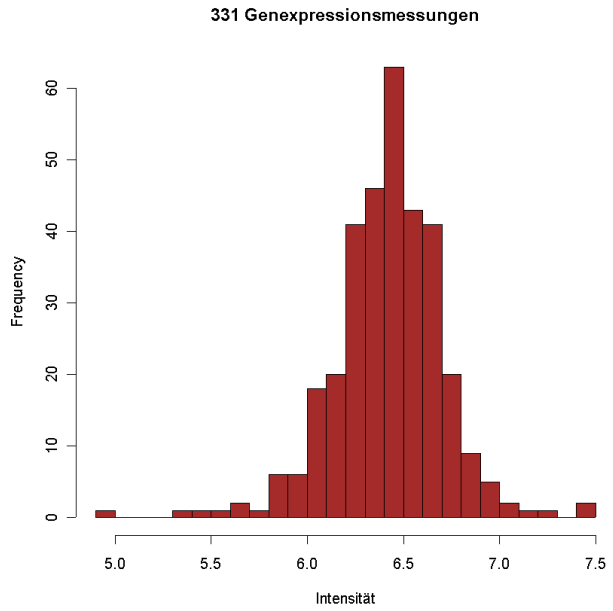


x-axis: Intensities in bins

e.g. [8.1-8.3]

y-axis: number of genes with $\log_2$ expression levels in this bin

Histogram of log expression of the same gene across 331 samples



Normal Distribution

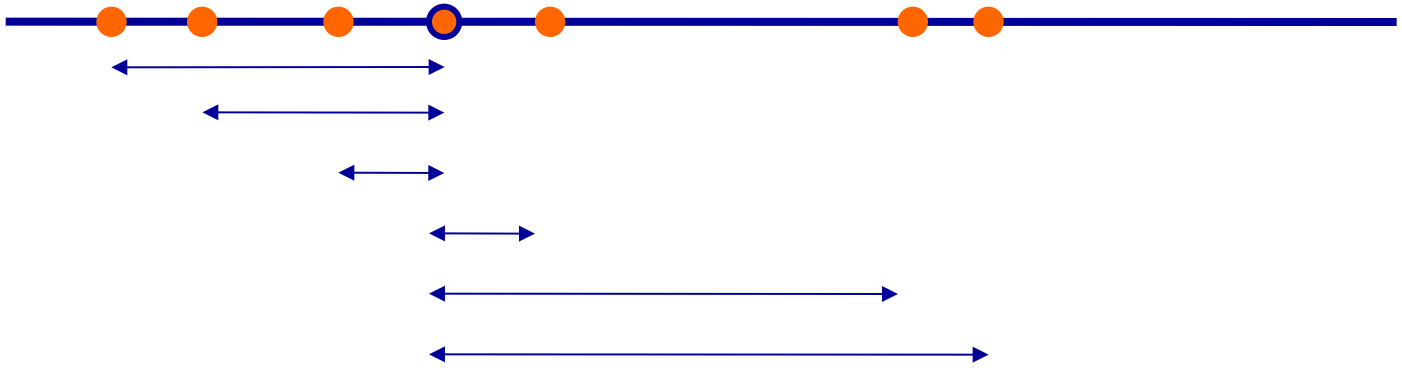
Superposition of various random effects in the measurements tend to produce a normal distribution

Central Limit Theorem

How can we summarize this data?

The Median

The median is the point in the middle of data. There is always the same number of data points to its left and to its right.

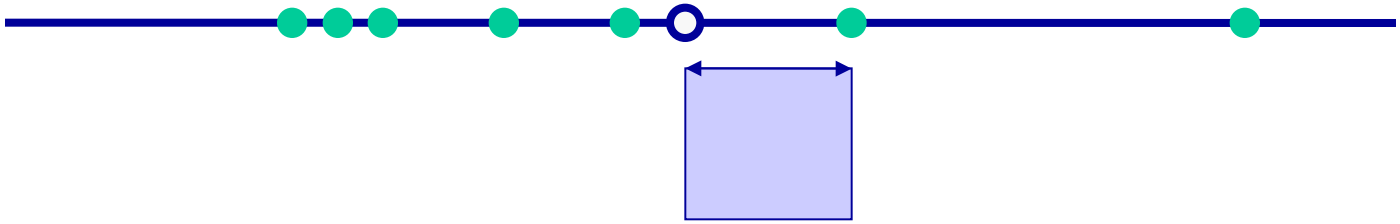


There is no point with a smaller sum of absolute differences to the data points

The Mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

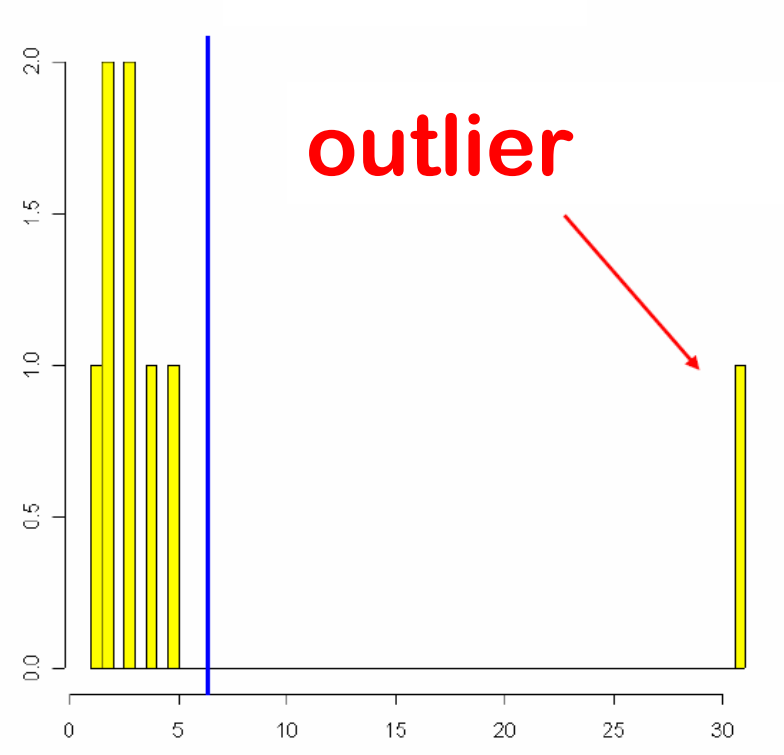
No point has a smaller sum of quadratic distances to the data points



Outlier

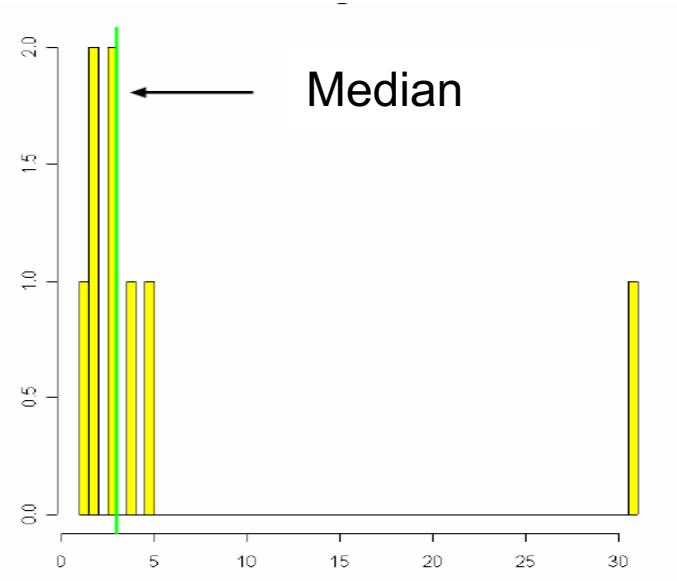
The mean is not in the middle of the data

Cause: the outlier



Robustness

The median is doing better

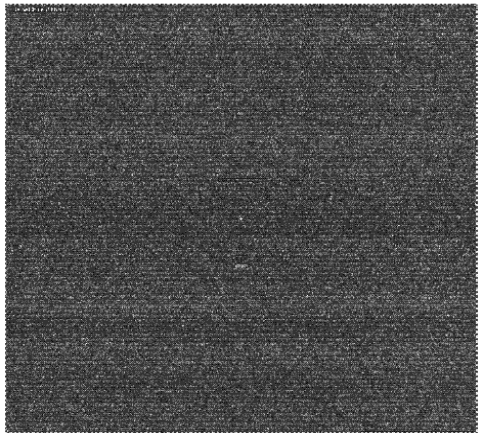


Wer viel misst, misst viel Mist

30.000 measurements

If there are outliers, you will probably not notice them at once

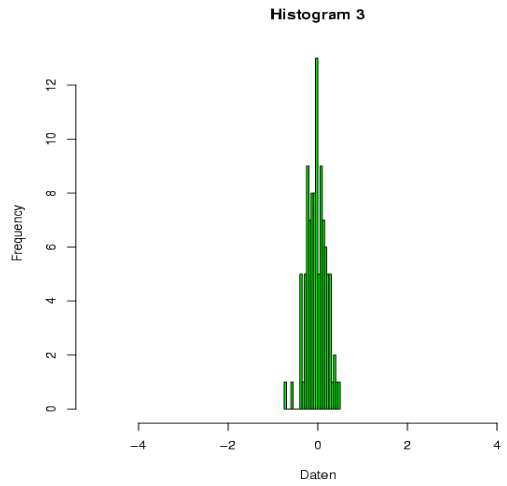
A single outlier can screw up your analysis and it might take a long time until you find the reason

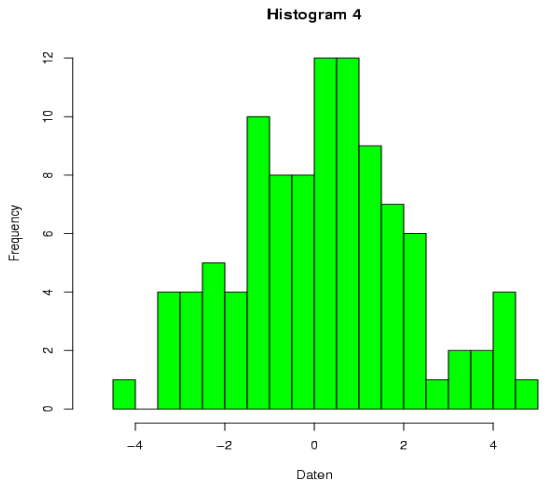


In microarray analysis:

Always prefer robust statistical measure

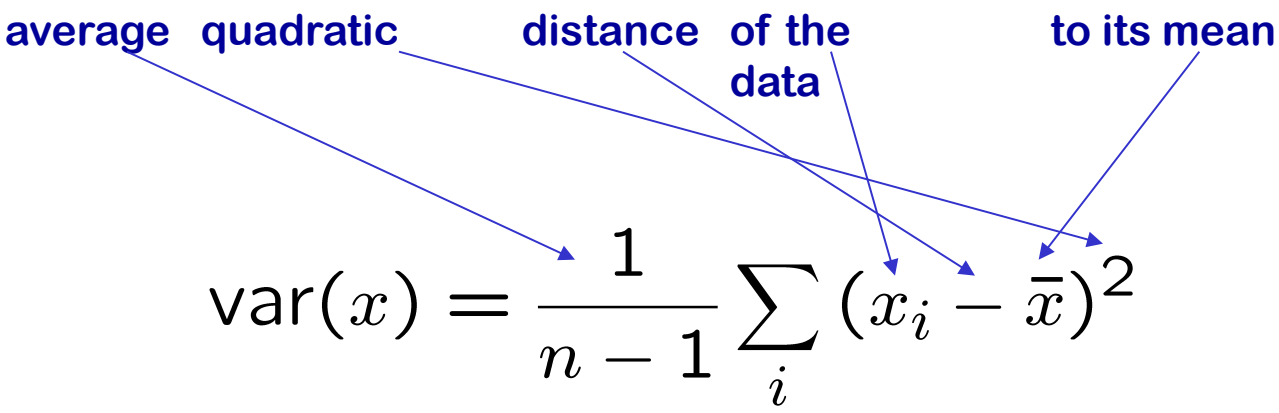
Variability





How can we quantify variability?

The Variance



Certainly not robust !

The Standard Deviation

„Wenn die Zahl der Kinder in einem Haushalt untersucht wird, so ist die Einheit der Varianz ein Quadratkind, die Einheit der Standardabweichung aber wieder ein Kind“

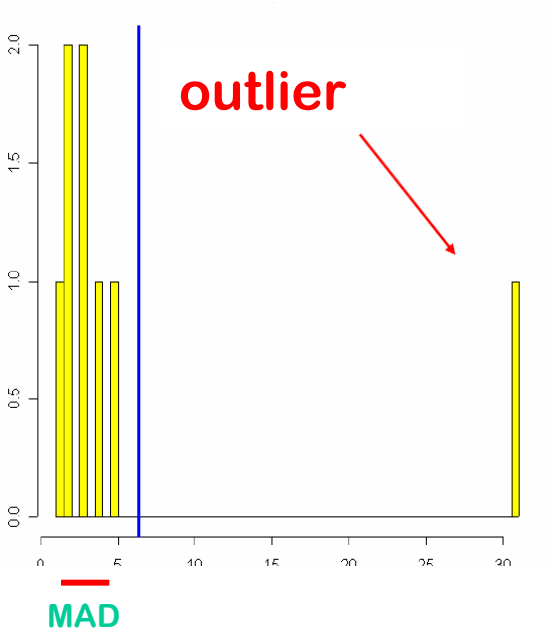
$$\sigma(D) = \sqrt{\mathrm{var}(D)}$$

The Median Absolute Deviation

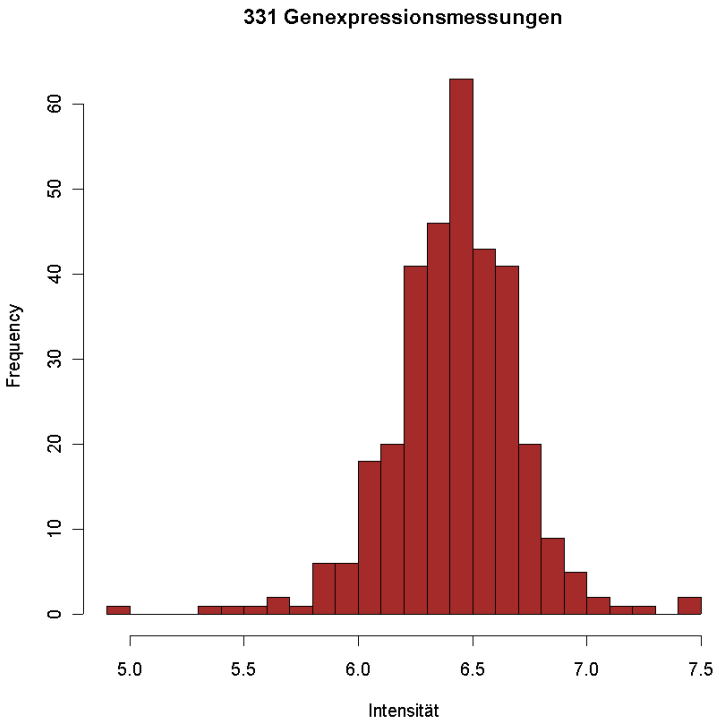
$$\mathrm{MAD} = \mathrm{med}|x_i - \mathrm{med}(x_i)|$$

A robust measure of variability

Practical advantage when dealing with large data sets



Units of Gene Expression

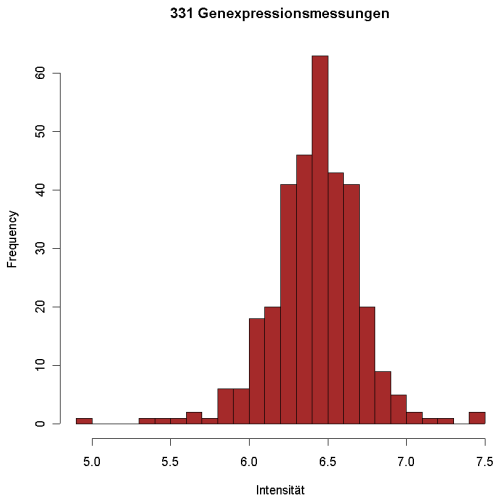


The intensities have no well defined units

A statement like: “A gene has an expression level of 5.8” does not tell us a lot

However, the histogram reveals that 5.8 is a small expression level

The data defines its own units



5.8 lies 2.16 standard deviations below the mean

Standard units: Number of standard deviations above or below the mean

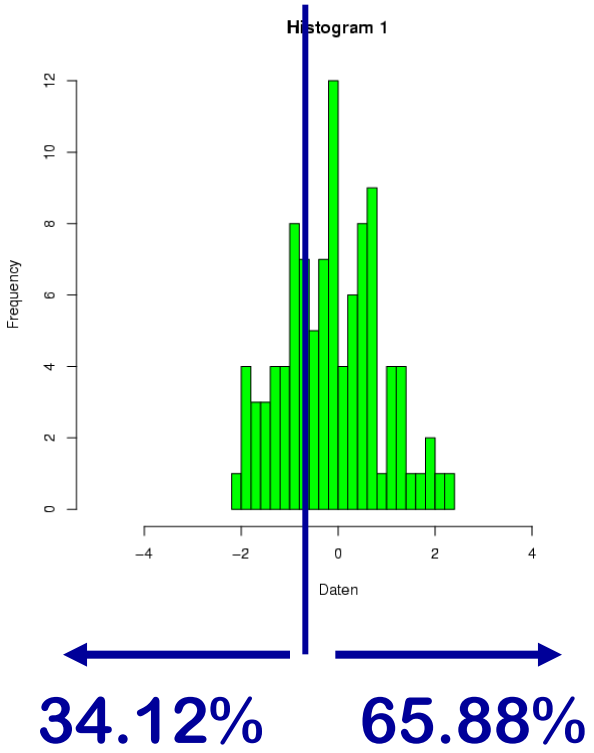
Transformation to standard units:

1. Calculate the mean and subtract it from every data point
2. Calculate the standard deviation and divide the data points by it

Quantiles

The 34.12% quantile:

A point, where 34.12% of the data points lie to the left and 65.88 % to the right



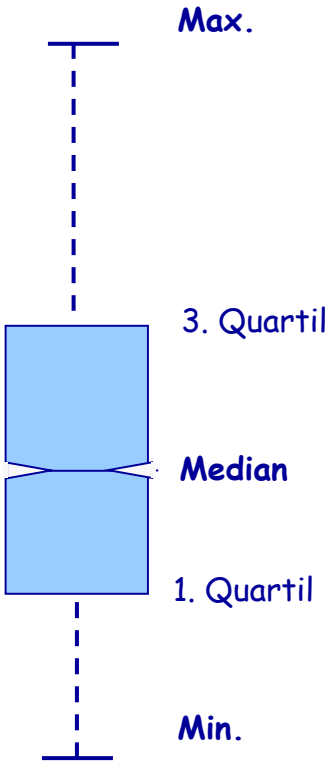
1 quartile = 25% quantile

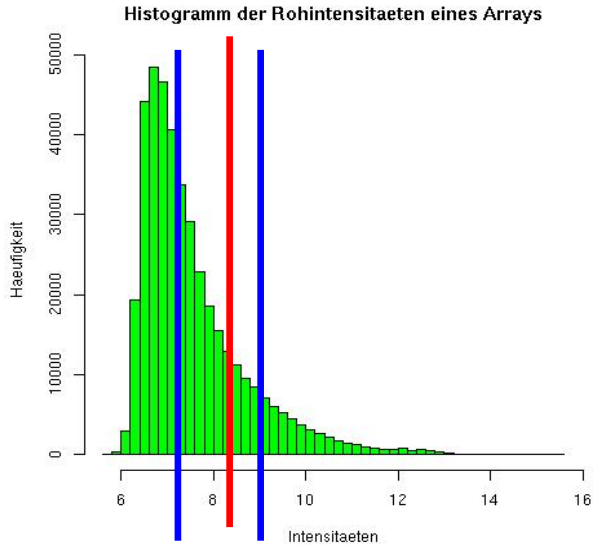
median = 50% quantile

3 quartile = 75% quantile

Box plots

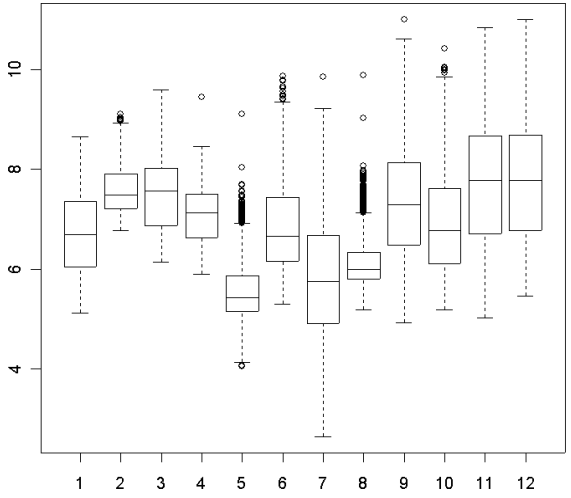
Boxplot





Median

1.Quartil 3.Quartil



Normalization

Bias and Variance

var high

Bias = systematic error

The mean is incorrect

var low

Variance = random fluctuation

The mean is correct

bias high

bias low

Repeat the experiment 5 times and take the mean

Individual meassure-ment

Mean over repeated measure ments

bias high

bias low

The variance is reduced by averaging, the bias is not

Standard Deviation and Standard Error

Standard Deviation (SD): **Variability of the measurement**

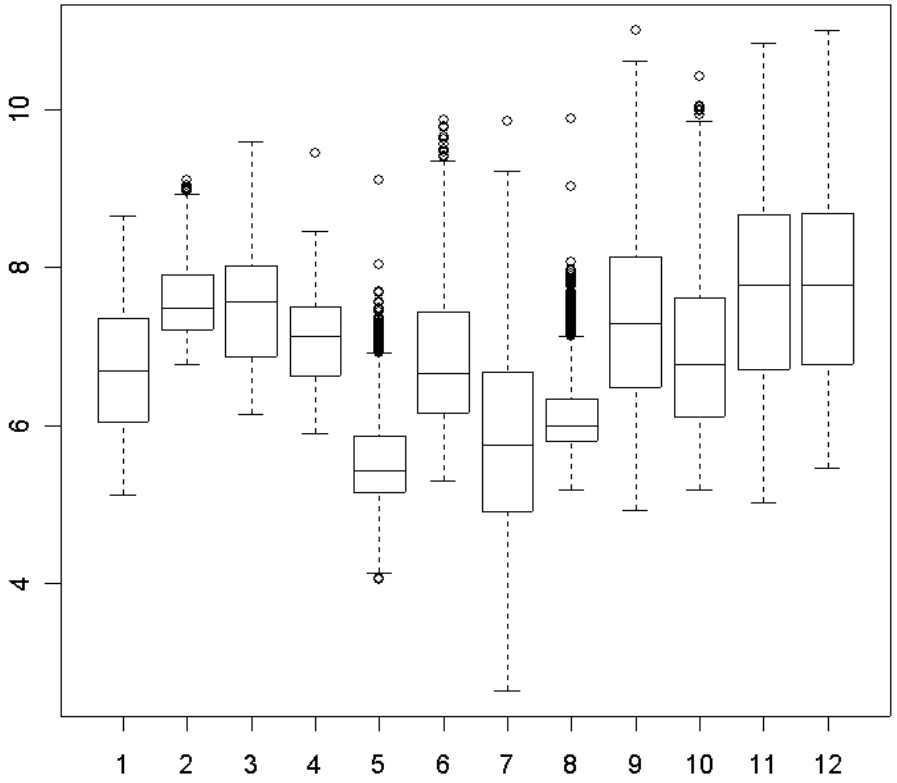
Standard Error (SE): **Variability of the mean of several measurements**

n Replications

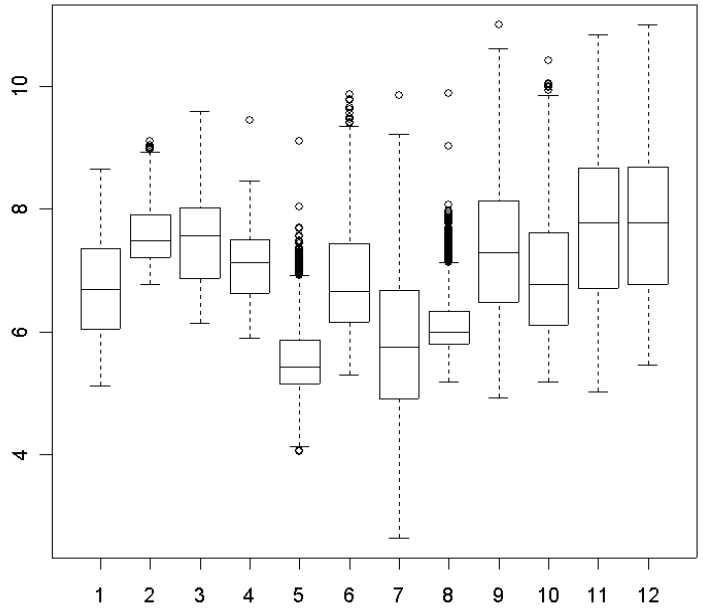
Normal Distributed Data:

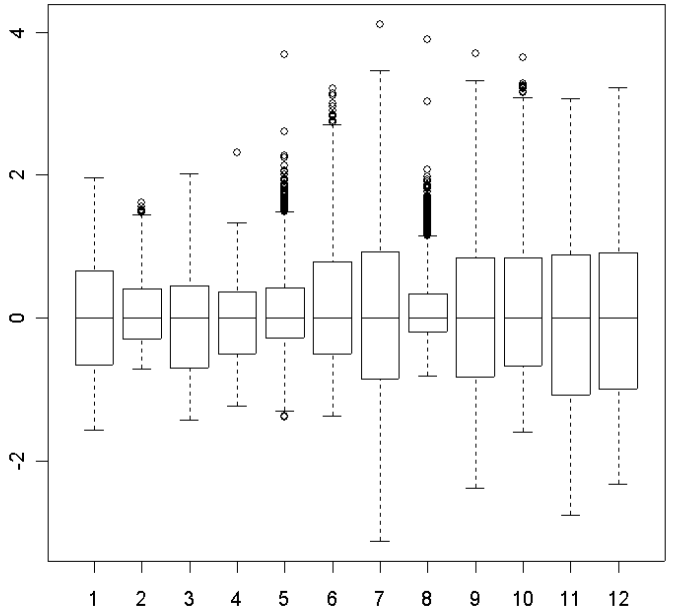
$$SE = \frac{1}{\sqrt{n}} SD$$

Why is this data biased?



Subtract the Median





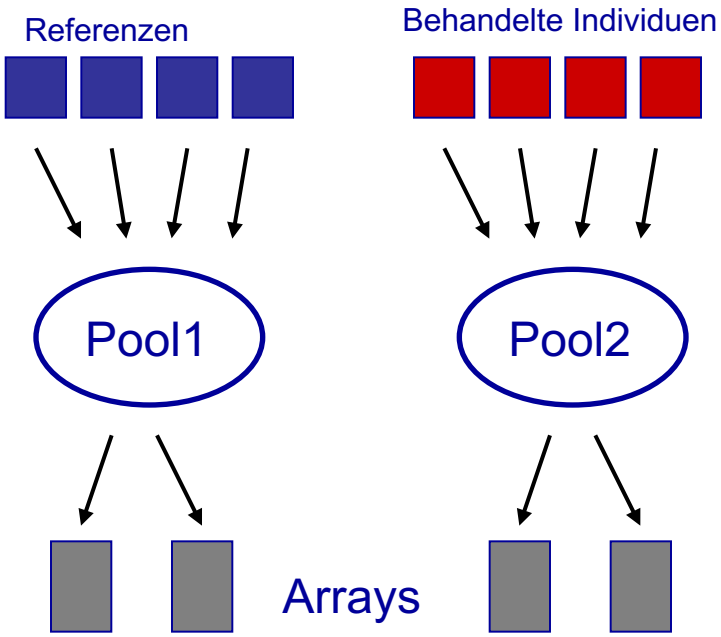
Normalization

Fighting errors

You can fight variance by averaging …
but not by normalization

You can fight bias by normalization …
but not by averaging

Pooling



Pooling reduces biological variability

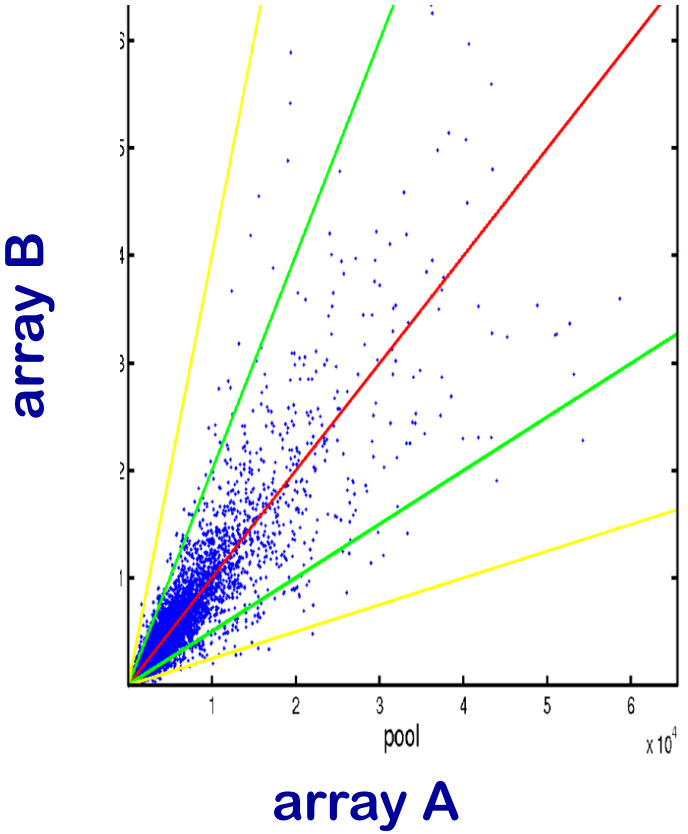
We reduce variance but introduce a bias with respect to the variability of life

If pooling can be avoided it should be avoided

You do not want to reduce natural variability !

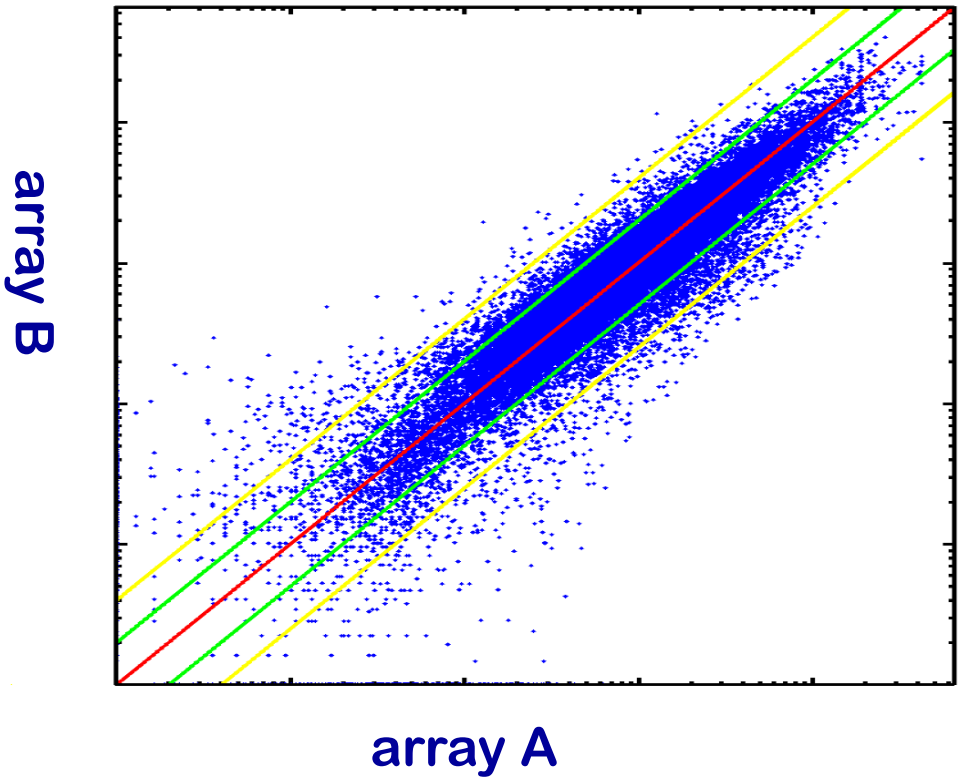
You want to know it !

Scatter Plots



We can only see what is going on with the most highly expressed genes

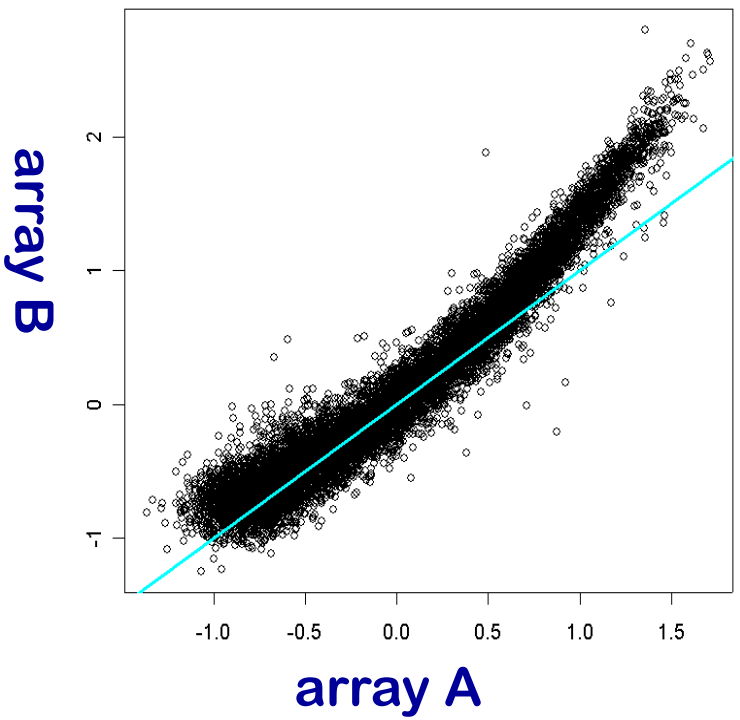
Taking the log



Fold changes become absolute differences

log(ab)=log(a)+log(b)

The major source of technical bias are problems with the dynamic spectrum of the array

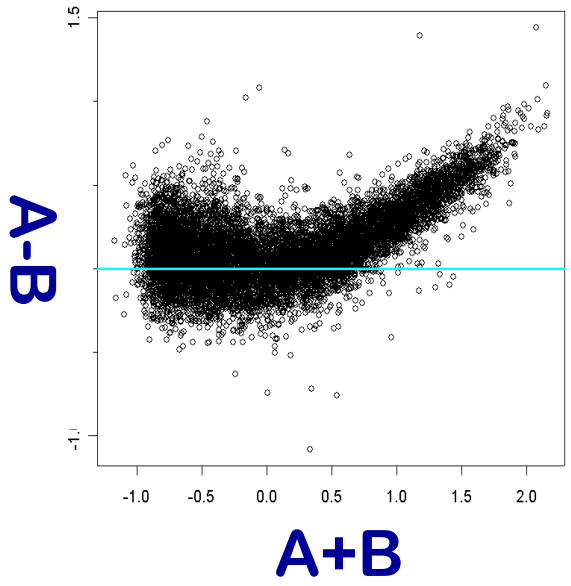


Array B systematically over estimates expression of the most highly expressed genes …

… or array A under estimates them

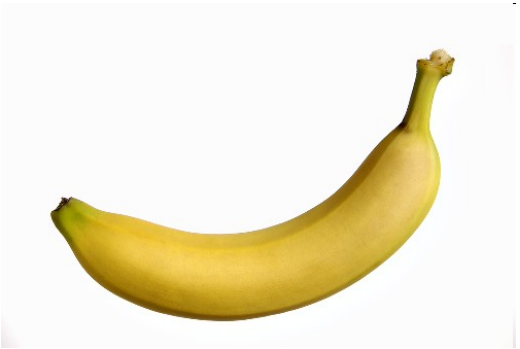
… or the problem is with the low end of the dynamic spectrum

M vs. A plots



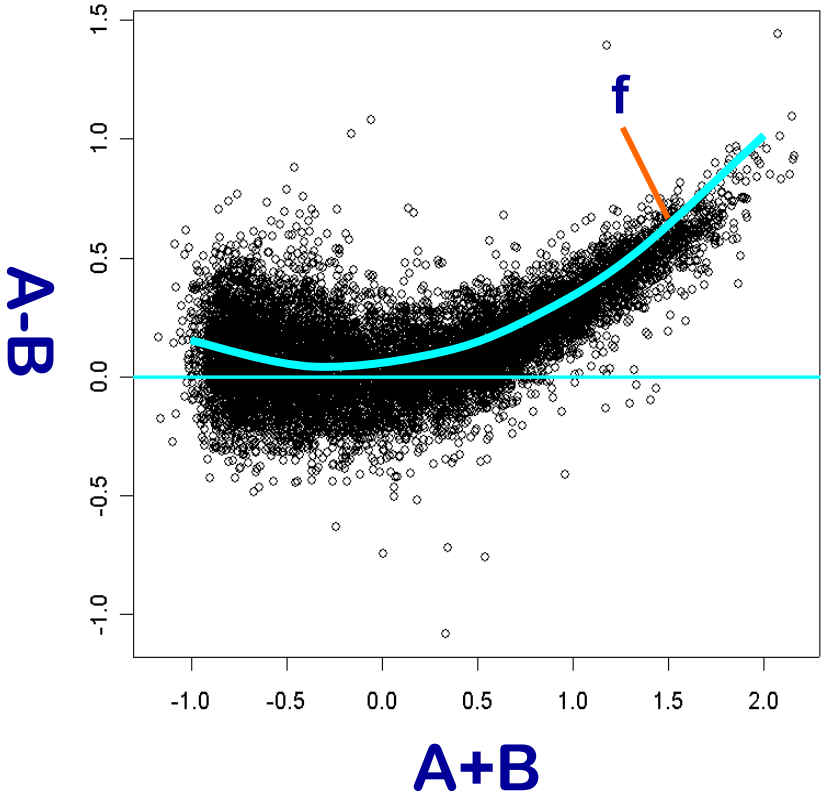
x-axis: dynamic spectrum

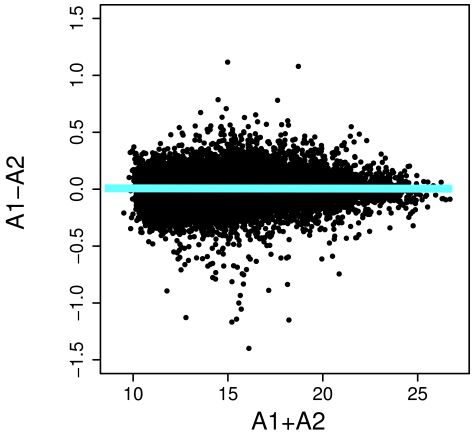
y-axis: difference



The technical bias makes the plot look like a banana (banana plot)

Straitening out bananas

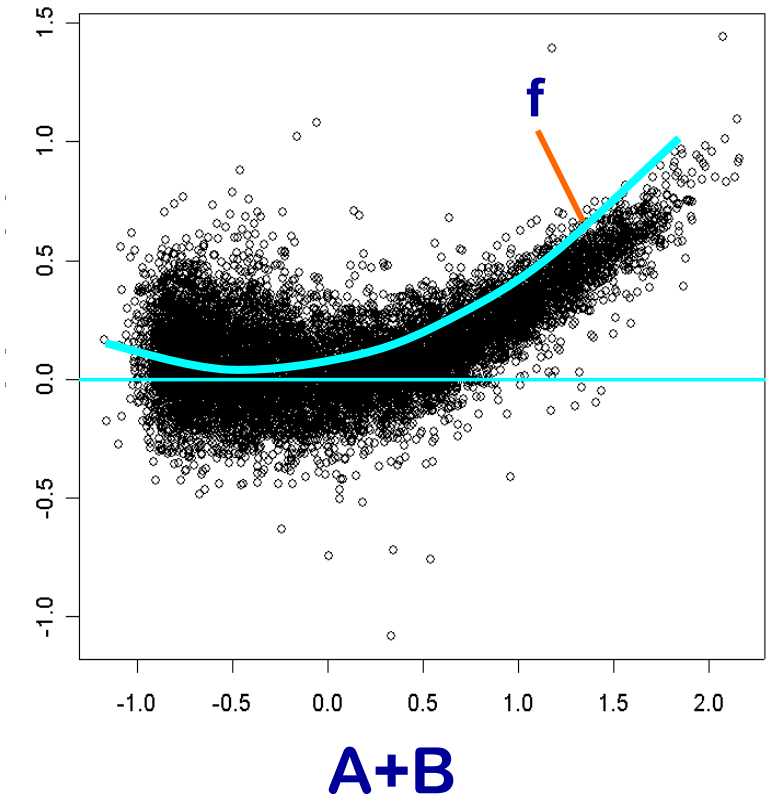




Lo(w)es-Normalization:

Fit a smooth line through the M vs. A plot and subtract it

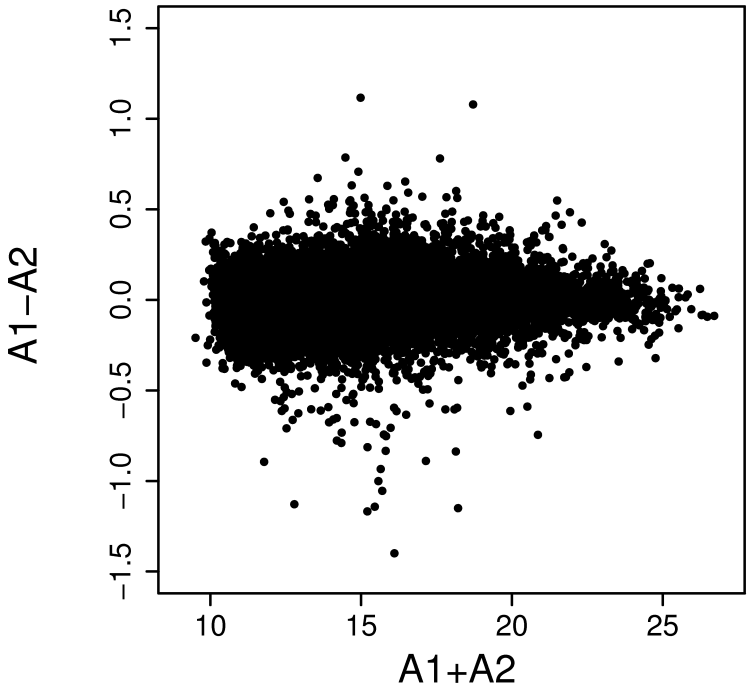
Cyclic Loess Normalization



What if we have 119 arrays?

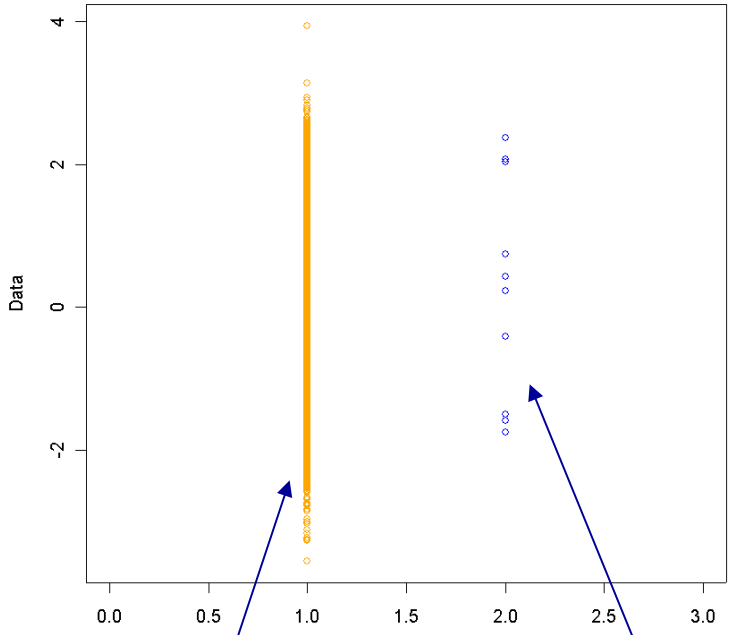
repeat until convergence
  for i=1 to 119
    for j=i+1 to 119
      A(i) ≈ norm (A(i) with A(j))
    end for
  end for
end repeat

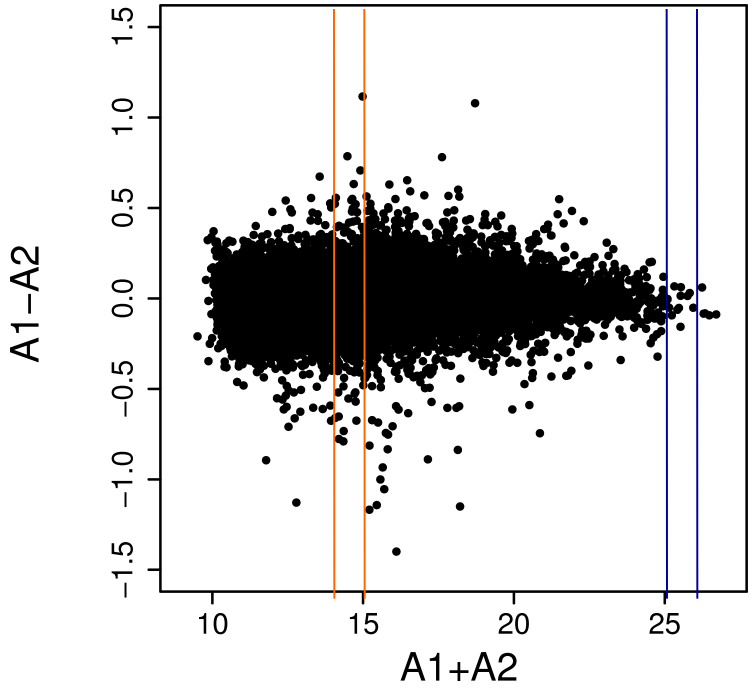
Is there a problem with the variance ?



Is the variance systematically higher at the left end of the dynamic spectrum ?

What the eye detects is the range not the variance

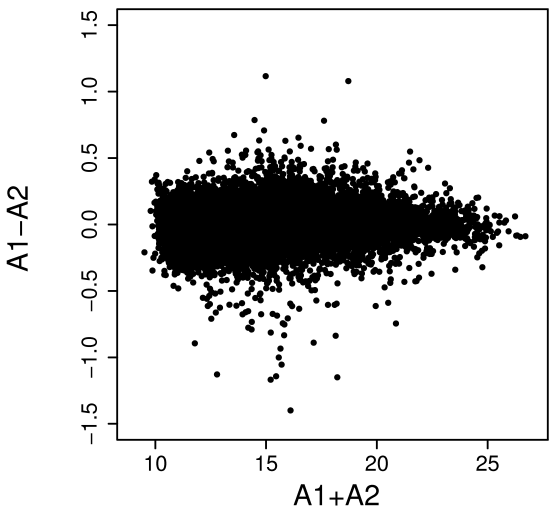


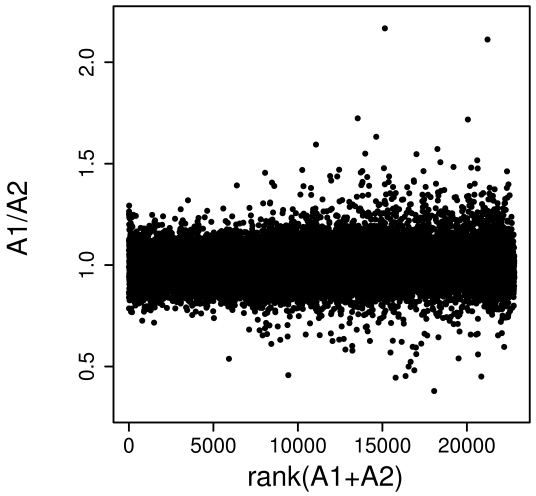


5000 N(0,1) distributed points

10 N(0,1.5) distributed points

The rank(M) vs A plot

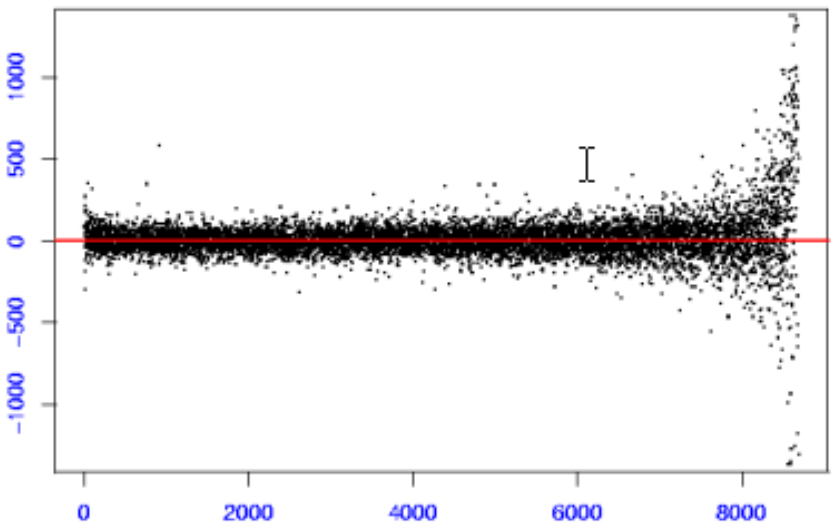


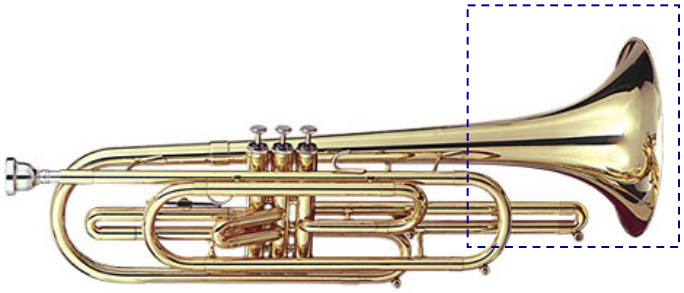


With the same number of points in every vertical stripe of the plot the variance appears stable over the entire dynamic spectrum …

… but this is nicely normalized data …

Trumpet plots

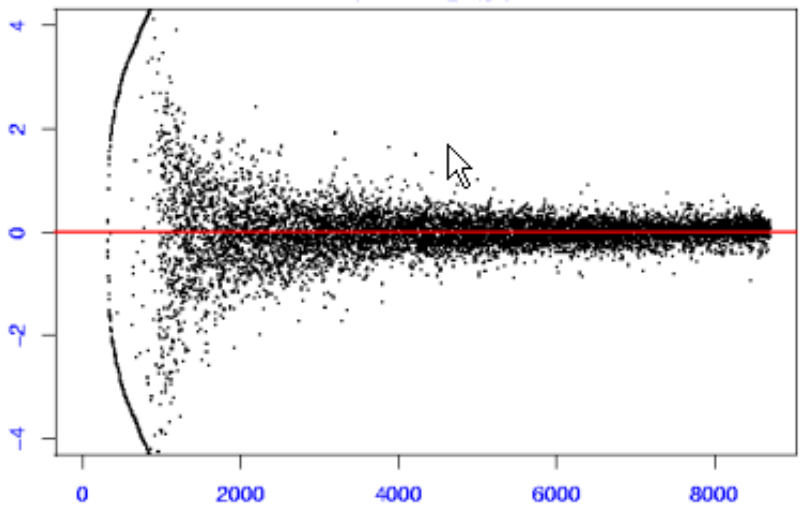


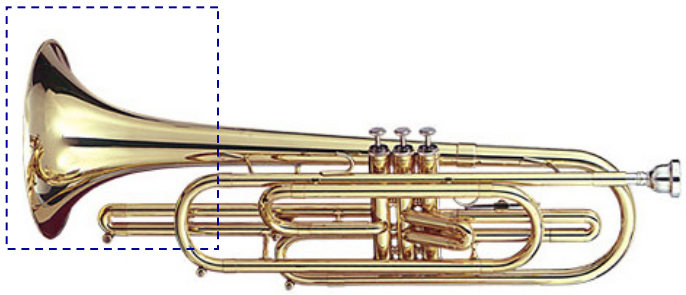


A tenfold up regulation of a low abundance transcription factor can correspond to the same absolute increase as a permille change of actin abundance …

… that is why biologists always use fold changes. They correct for this bias … do they?

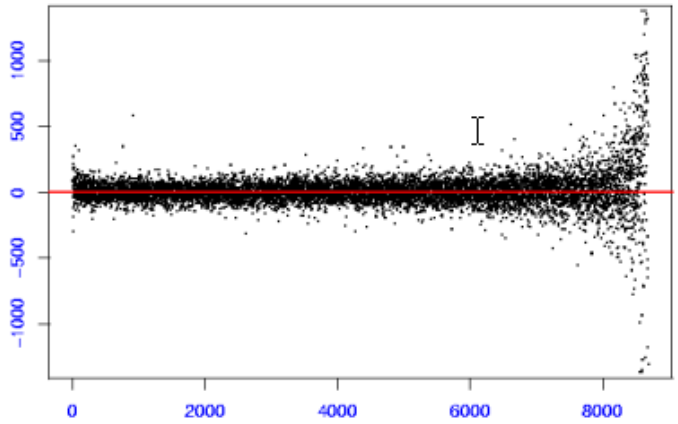
The trumpet after taking the log





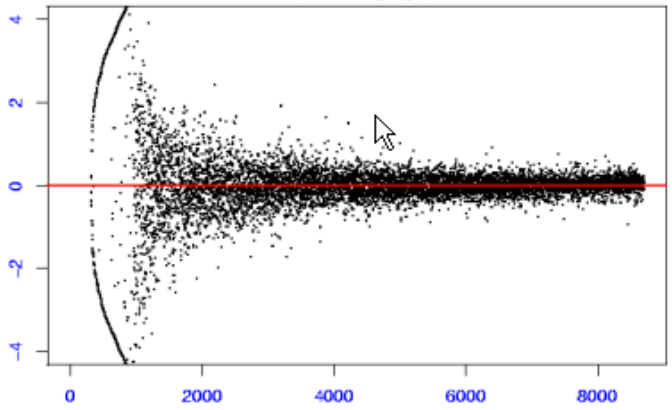
If we meassure two genes that are not present at all, we will get two small numbers. It can easily happen that one of them is ten times higher than the other

Multiplicative and additive noise



The absolute intensities are dominated by the biological variance of the highly expressed genes

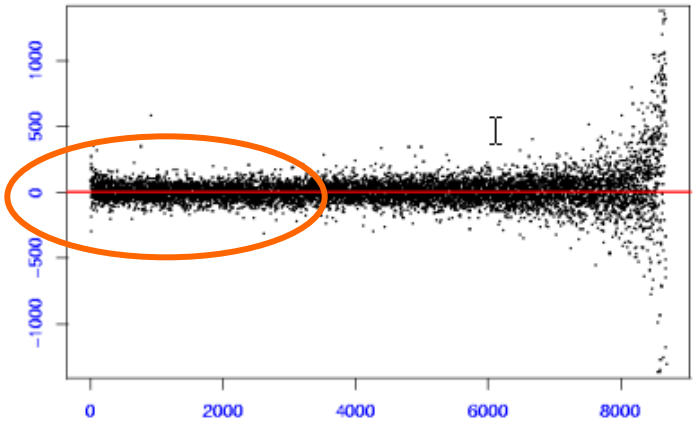
- multiplicative noise



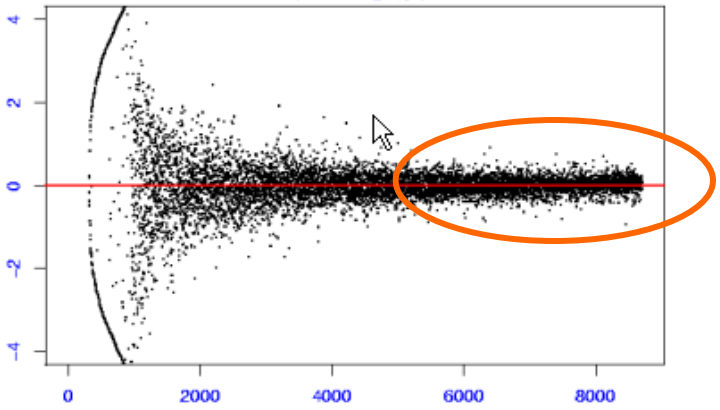
The log intensities are dominated by the technical variance of the lowly expressed genes

- additive noise

Scale and variance

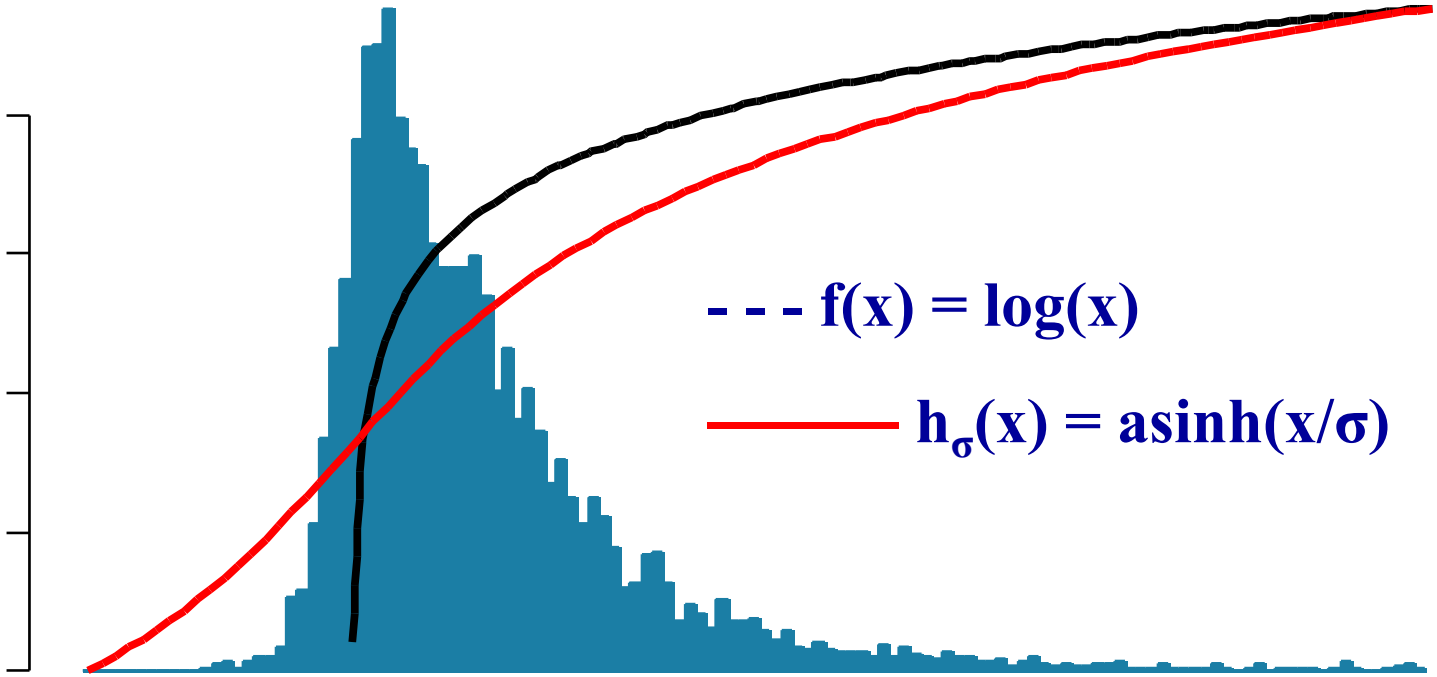


At the left end of the dynamic spectrum we obtain stable variance if we scale the data linearly

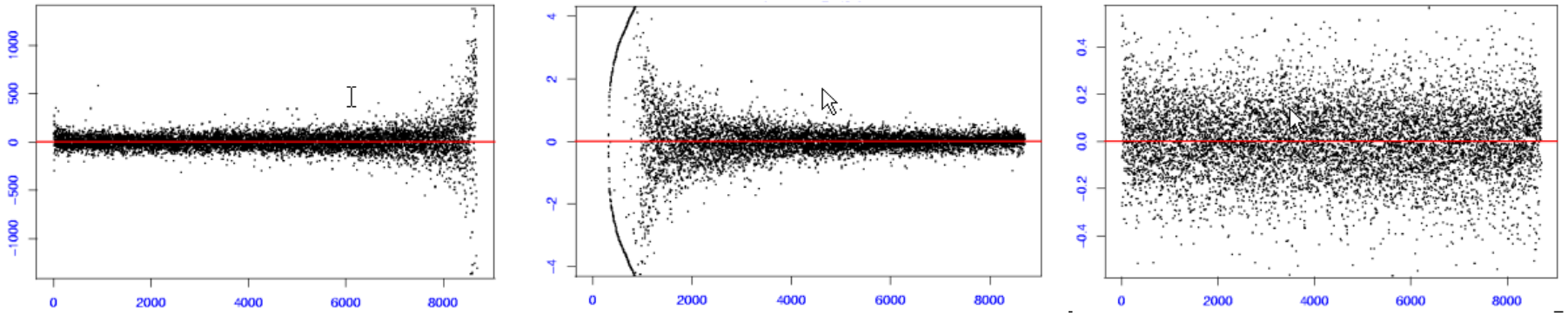


At the right end of the dynamic spectrum we obtain stable variance if we scale the data logarithmically

asinh-Transformation



Variance Stabilization



linear logarithmic asinh

The vsn package of Wolfgang Huber combines variance stabilization with background correction and chip to chip normalization

Ranks and Quantiles

Chip with 1000 genes

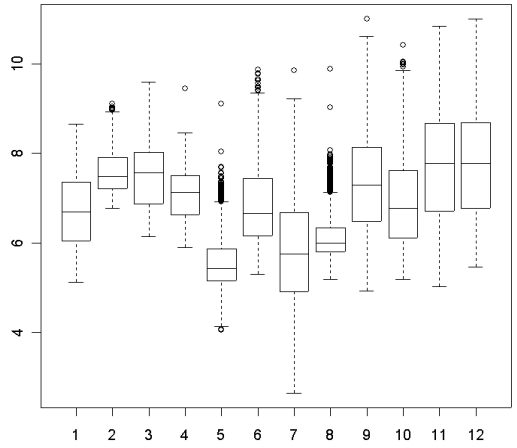
The value of the top ranking gene is the 100%-quantile

The second ranking gene gives the 99.9%-quantile

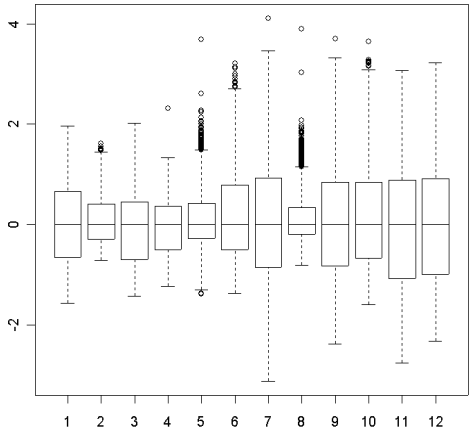
The median is given by the gene in the middle of the list

The lowest expressed gene gives the 0.1%-quantile

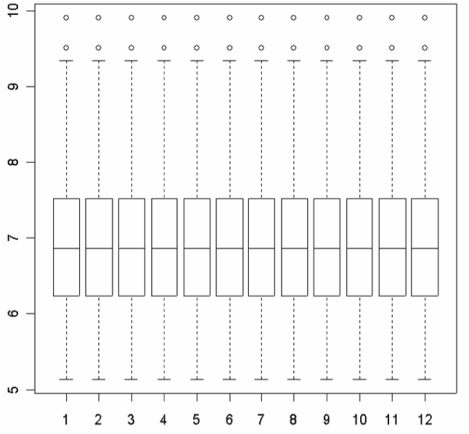
Adjusting all Quantiles



not normalized

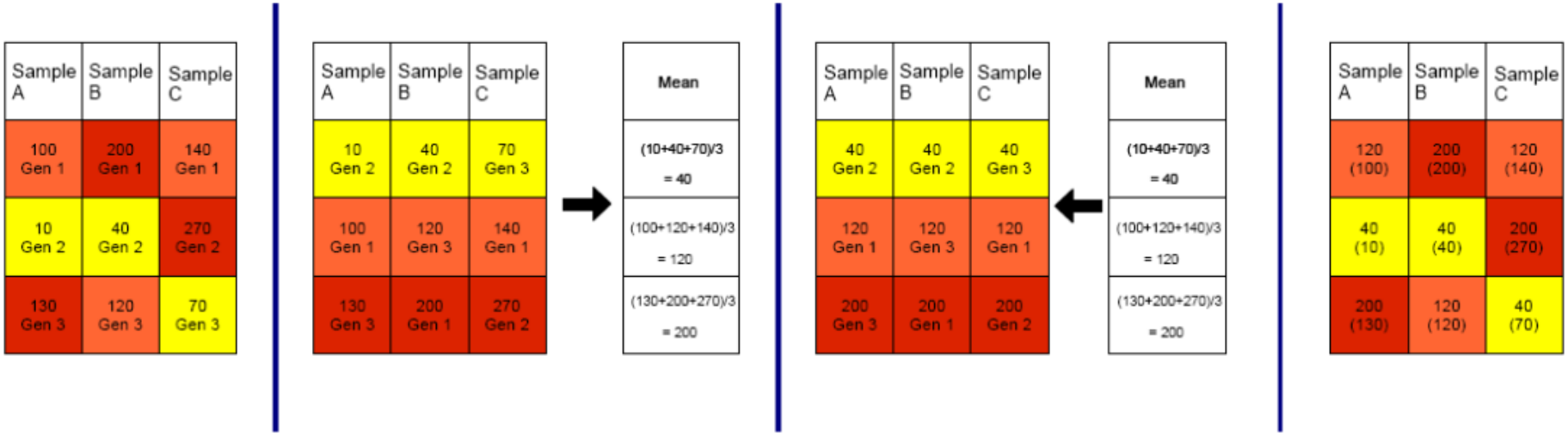


adjusted for the
50%-quantile only



adjusted for all quantiles

Quantile Normalization



not normalized

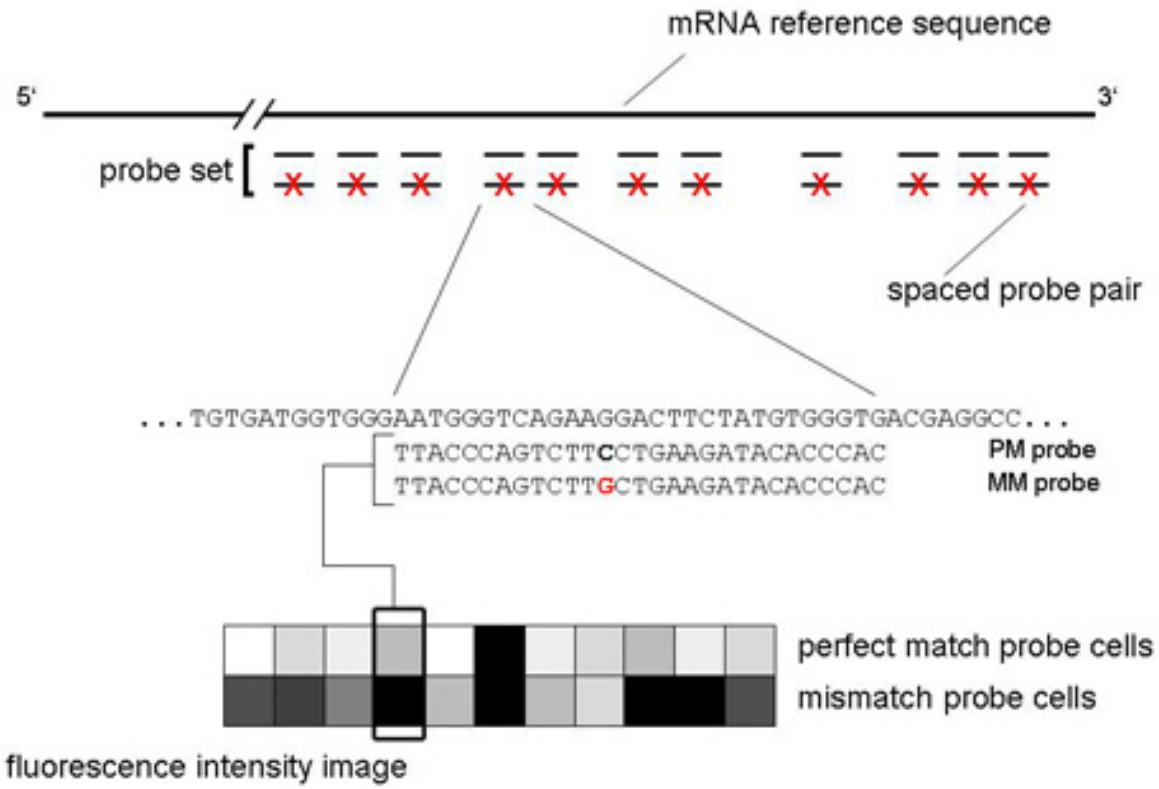
sort and calculate mean of the ranks

set expression to mean of the corresponding rank

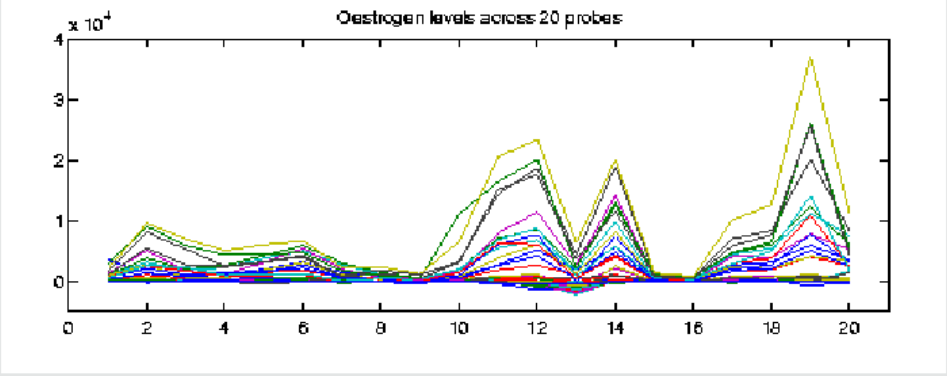
normalized

All profiles have exactly the same values just distributed differently over the genes

The Affymetrix Design



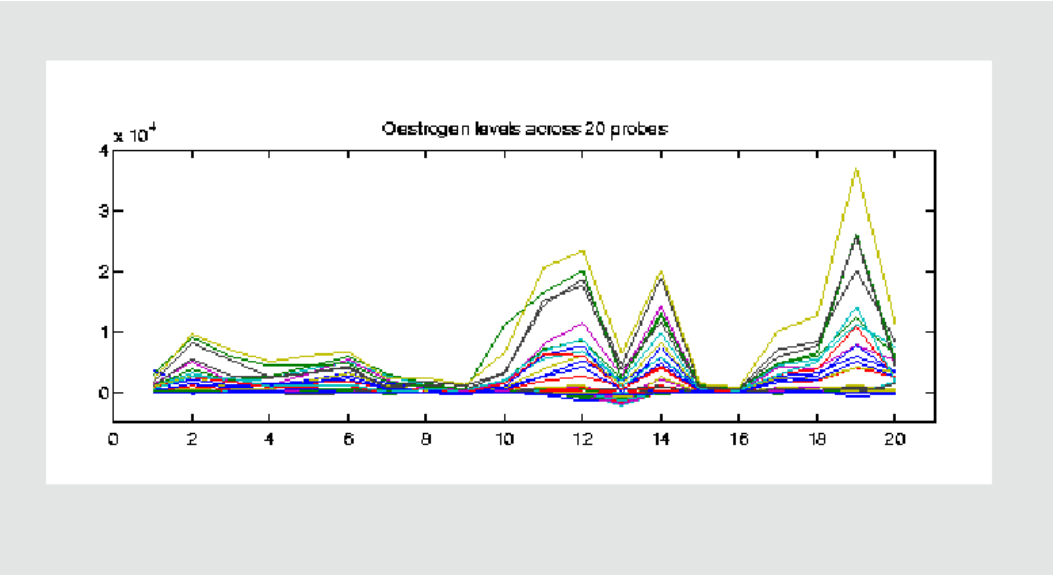
A calibration problem



Some probes are more sensitive to changes in molecule abundance than others …

… and some do not work at all

Binding affinities vary, because hybridization conditions do not vary

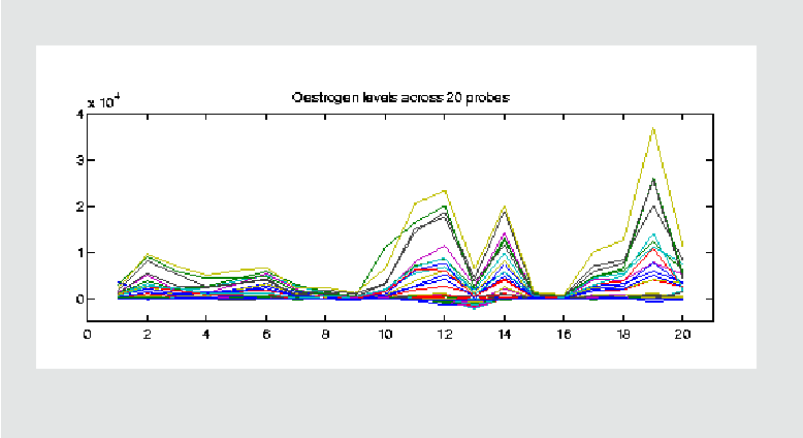


Hybridization affinities depend on temperature and base composition of sequences.
On the chip we have one temperature for all genes but different base compositions

Microarrays are no measurement devices because they are not calibrated and can not be calibrated

… at least not very well

Summarization

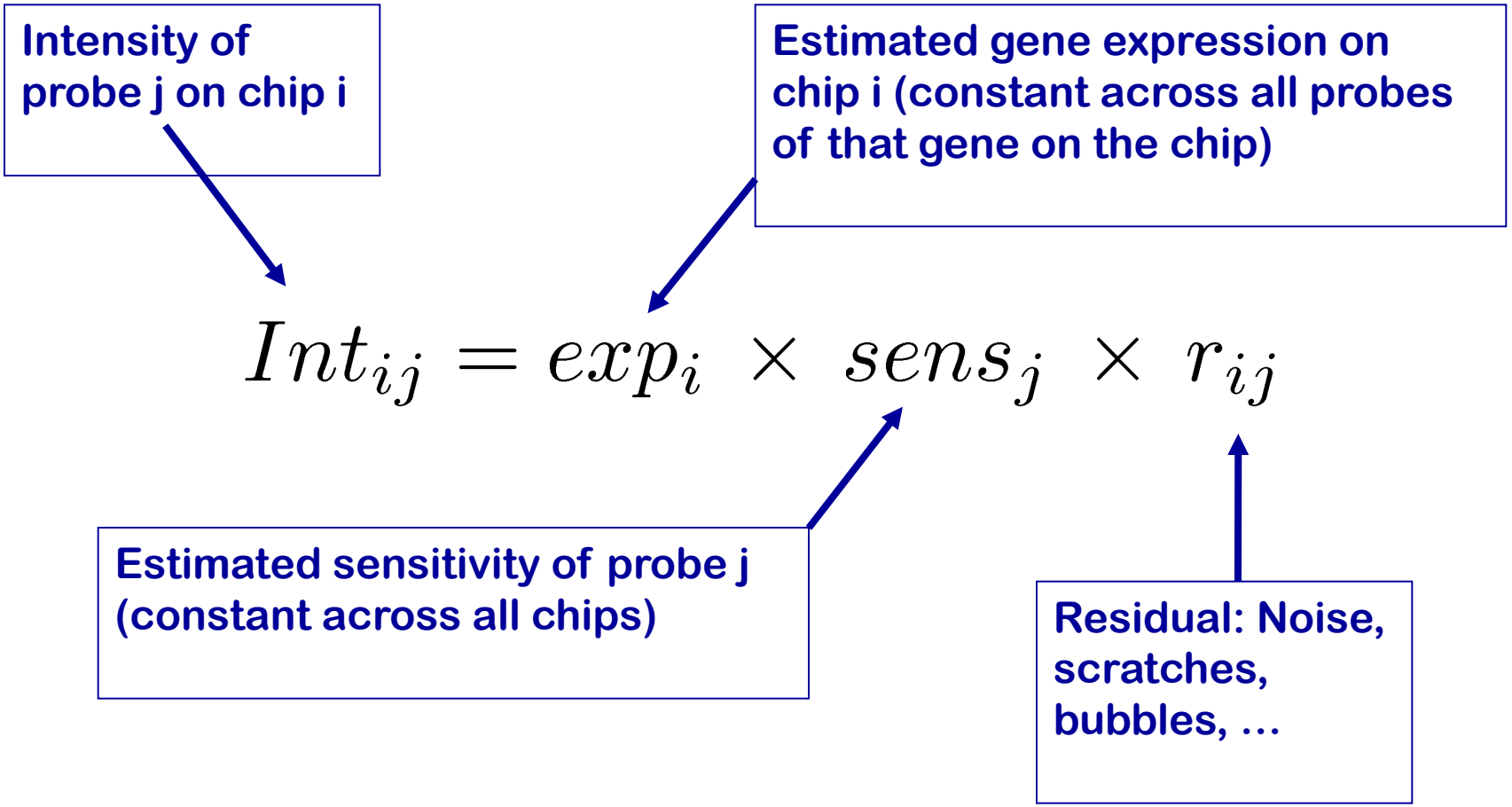


We still need to condense the probe intensities to a single expression level for the gene

Averaging ?

We can do better …

Intensity Model



Taking the log (RMA)

Log-intensity of probe j on chip i

Estimated log gene expression on chip i (constant across all probes of that gene on the chip)

$$Y_{ij} = E_i + S_j + R_{ij}$$

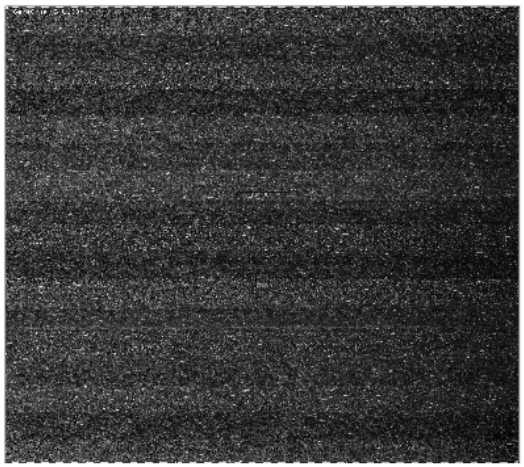
Estimated log sensitivity of probe j (constant across all chips)

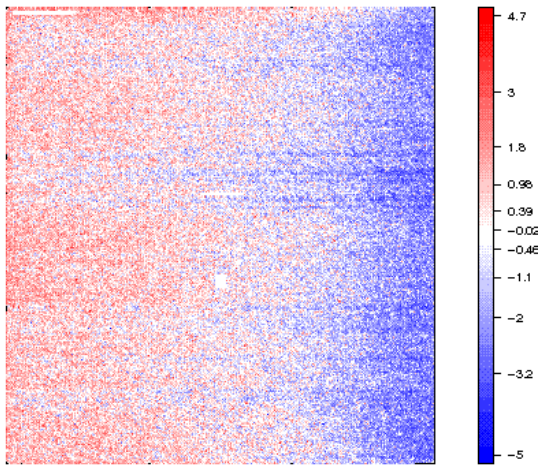
Residual: Noise, scratches, bubbles, …

E_i und S_j can be estimated using Tuckey's robust median polish algorithm

Inspecting the Residuum

$$Y_{ij} = E_i + S_j + R_{ij}$$

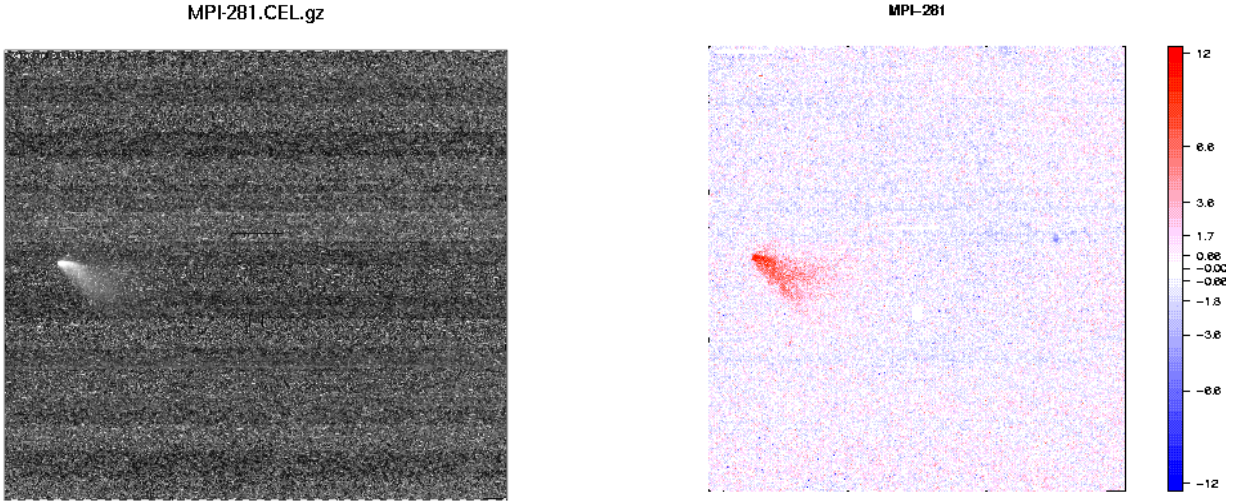




Global spatial gradient … array should be removed

Outliers are automatically corrected for

$$E_i = Y_{ij} - S_j - R_{ij}$$



Local artifact … array does not need to be removed

Single Chip vs. Multiple Chip Normalization

VSN, RMA, Cyclic Loess, Quantile Normalization normalize batches of chips. They borrow information across chips and do not work on single chips.

This can be a practical problem !

Assume you add one more experiment:

-You need to normalize all chips again

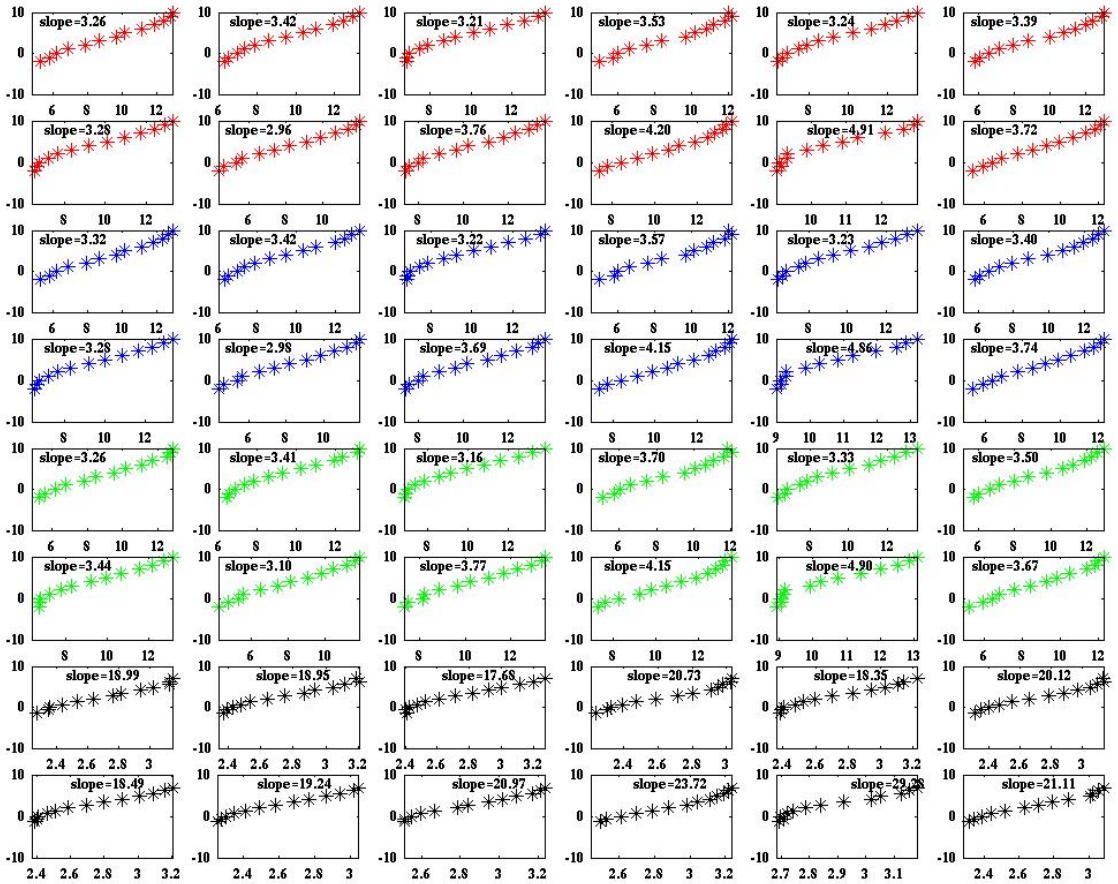
-Old results might change

Remedies:

Single chip normalization methods (Affy's MAS 5)

Add on normalization

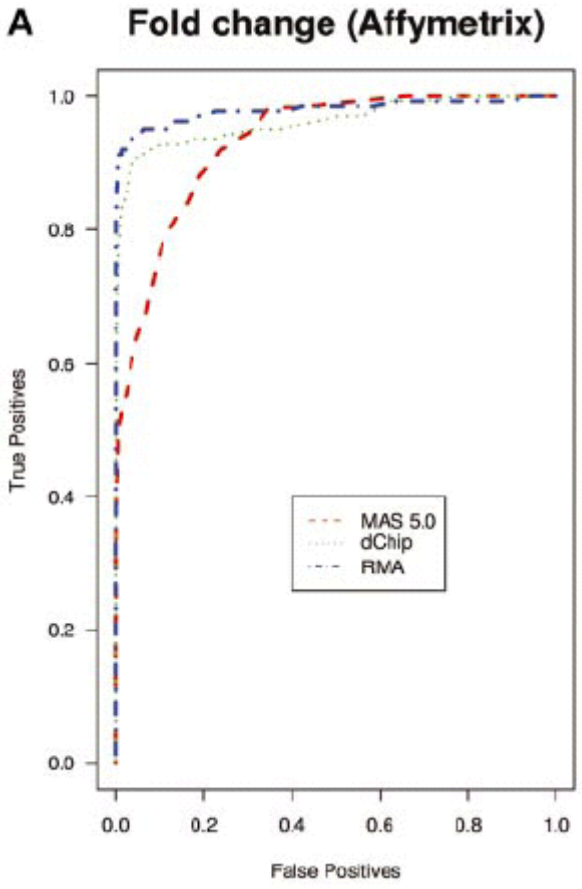
Calibration of fold changes



Generating known fold changes by spiking in genes with known concentrations

It is possible to calibrate the array to fold changes of the same gene across chips

Putting a needle in the hey stack and searching it again



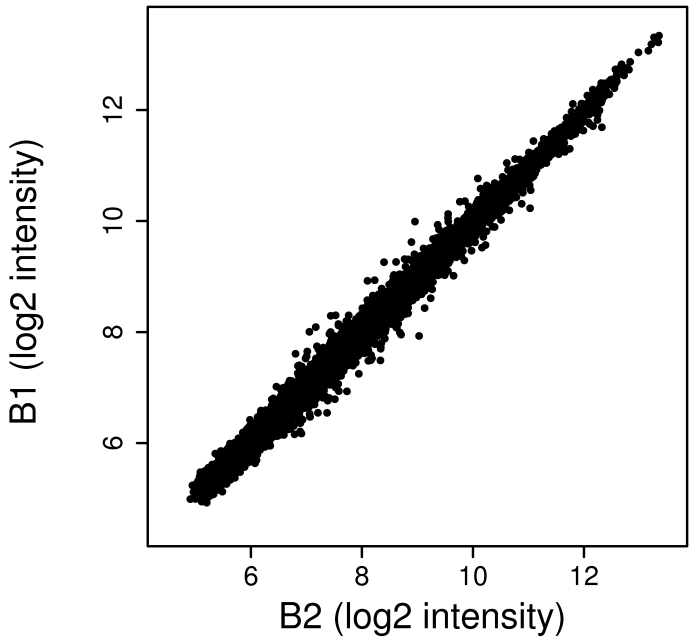
Spike in concentrations on an identical background (From small fold changes to large ones)

Only the spike in genes are differentially expressed

Try to find them

Receiving operator curves (ROC)

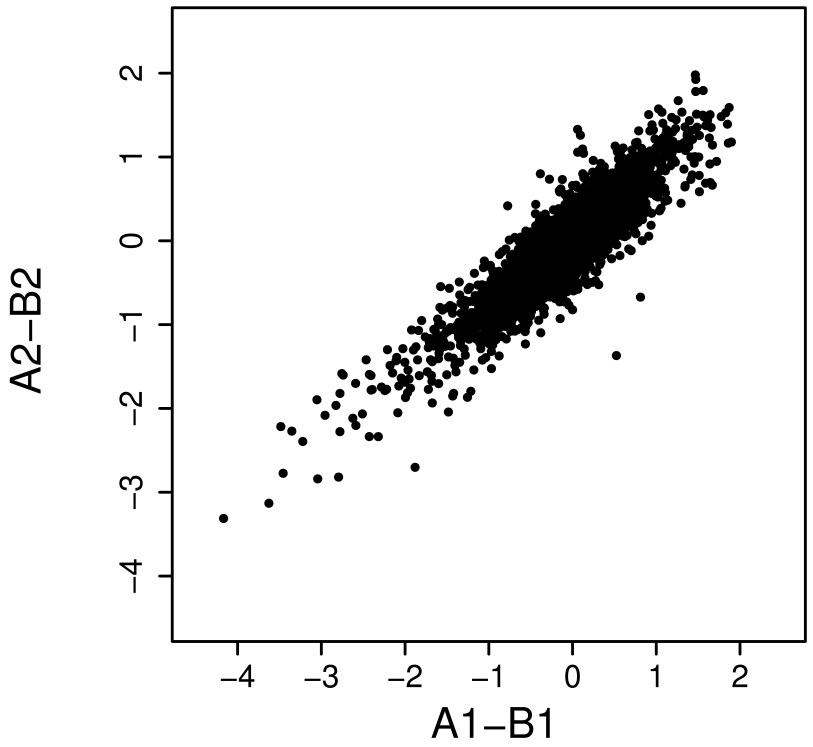
Reproducibility of absolute expression



Do the same experiment twice and make a scatter plot

We can reproduce that actin is more highly expressed then myc, so what ?

We need reproducibility of expression changes



Do two different experiments twice and make a scatter plot of the differences !

Does not look as nice, but that is the type of reproducibility you need !

Reproducibility highly depends on normalization

A good normalization …

… corrects for technical bias (global shifts, background, bananas, trumpets, spatial gradients)

… makes the data robust with respect to outliers

… yields reproducible expression differences

… leaves the intensities sensitive to true expression changes

… calibrates the array to reproduce fold changes of the same gene across chips

… helps us detect differentially expressed genes

Calibration Summary

Absolute calibration of arrays is hard …

(GC-RMA, Hook-Statistics,…)

… and still not reliable

Present or absent calls depend on absolute measurement … and are notoriously unreliable

Focus your research questions on relative measurement

Acknowledgement

Ideas, slides and images borrowed from:

Wolfgang Huber

Terry Speed

Achim Tresch

Tim Beissbarth

Christine Steinhoff

Tobias Müller

Julia Engelmann

Affymetrix

Questions

?