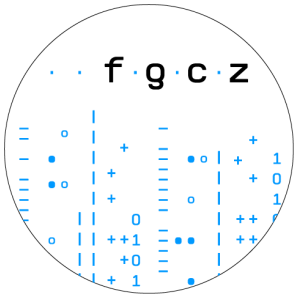
Assessing Differential Expression

Dr. Hubert Rehrauer

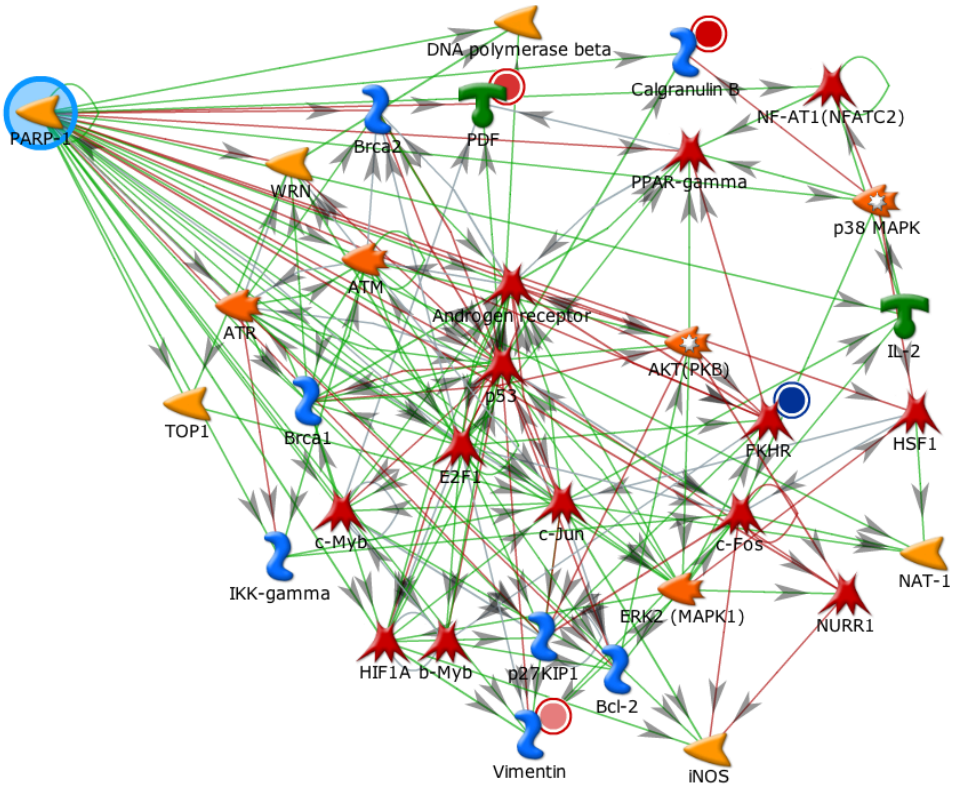


Hubert.Rehrauer@fgcz.ethz.ch

Gene Expression Study: Gene Function

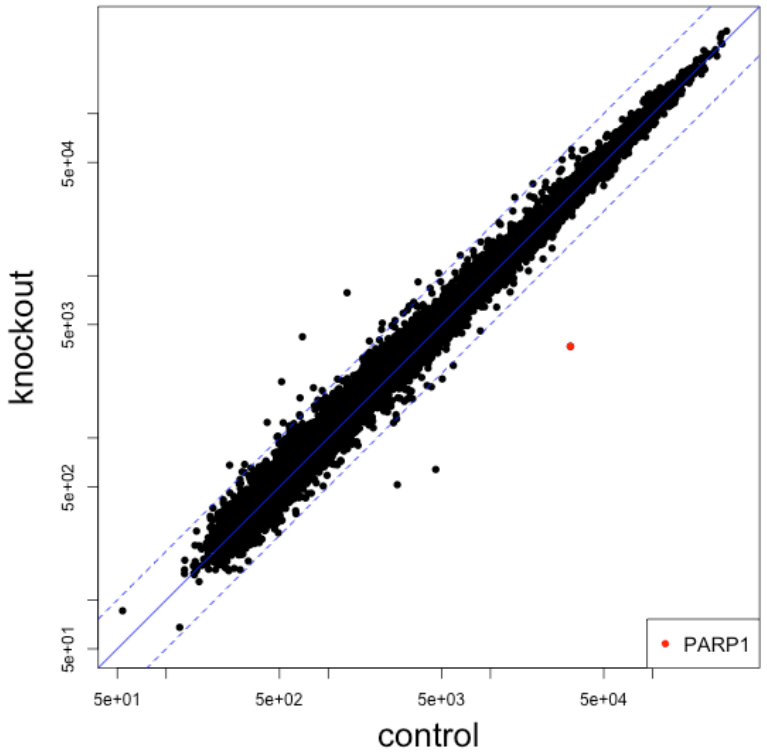
- PARP1 is an enzyme that is known to be involved in differentiation, proliferation but also in the recovery from DNA damage
- Question: What is its role in leukemia cells? Which cell activities are controlled by PARP1
- Experiment: Take leukemia cells and knockout PARP1 gene. Measure gene expression with and without knockout

PARP1 known interactions:



Gene Expression Study: Gene Function

- Result: Knockout decreases PARP1 abundance by a factor of 8
- ~ 10 other genes show also high expression changes
- Is it reproducible?
- Are these genes known interactors? Are they newly identified interactors?

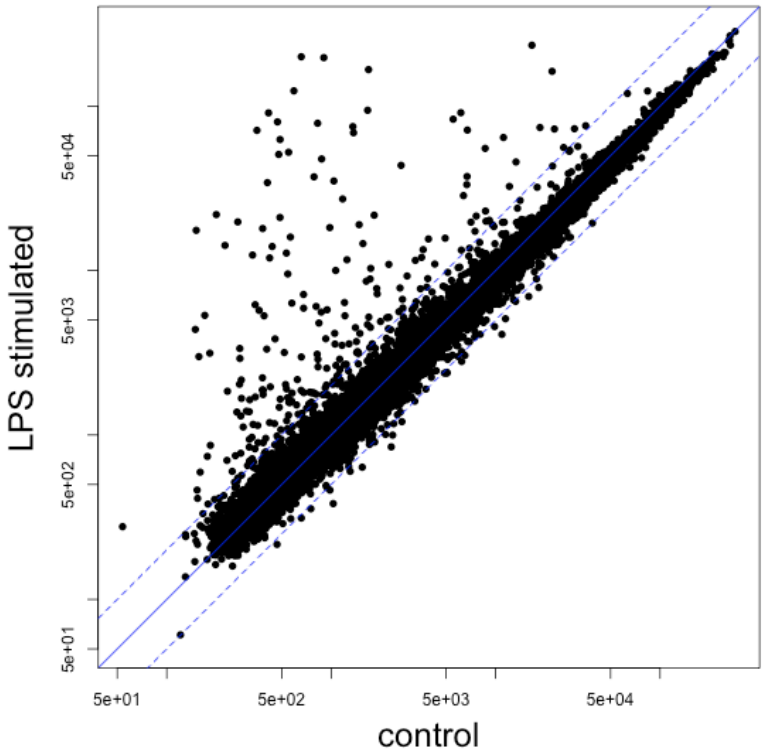


Gene Expression Study: Immune System Processes

- Bacteria have lipo-polisaccharides (LPS) on their surface
- Eukariotic cells can detect those and start immune defense processes
- Question: Which processes are started? Which genes are involved?
- Experiment: Add LPS to a cell line and measure expression response

Gene Expression Study: Immune System Processes

- LPS stimulation strongly upregulates (10-fold) a set of ~100 genes
- Many other genes show minor effects (~2-fold) up or down



Exploratory Studies

- Measuring the gene expression under different conditions:
  - healthy vs diseased
  - wild type vs mutant
  - treated vs untreated
  - different tissue types
- Goal:
  - Determine gene functions (which gene plays a role in a disease?)
  - Finding gene interactions (which genes change expression after knocking out gene X?)
  - …
- Approach:
  - Compute differentially expressed genes
  - …

Differential Expression

- Differential expression is always reported as relative change
- Factor by which the expression is increased or decreased (fold-change)
- Reasons for considering multiplicative changes:
  - In biology relative changes are more important than absolute changes
  - Microarrays anyway do not allow measuring absolute changes

- Example: Gene xyz is 1.5 times higher expressed in sick patients relative to healthy humans
- Since we consider only multiplicative changes we always work on the log expression values
- This treats up- and down-regulation symmetrically

- Problems with multiplicative changes:
  - genes not expressed in one of the conditions

Differential Expression

- Example: Treatment-control study
- Expression change: Ratio r

Linear scale: $y_{treated} = y_{control} \cdot r \cdot \varepsilon$

Transform: $x = \log(y)$

Logarithmic scale: $x_{treated} = x_{control} + \log(r) + \varepsilon$

M-A Representation

- Logarithm of the ratio

$$M = \log_2(\text{ratio}) = \log 2(\frac{y_{treated}}{y_{control}})$$

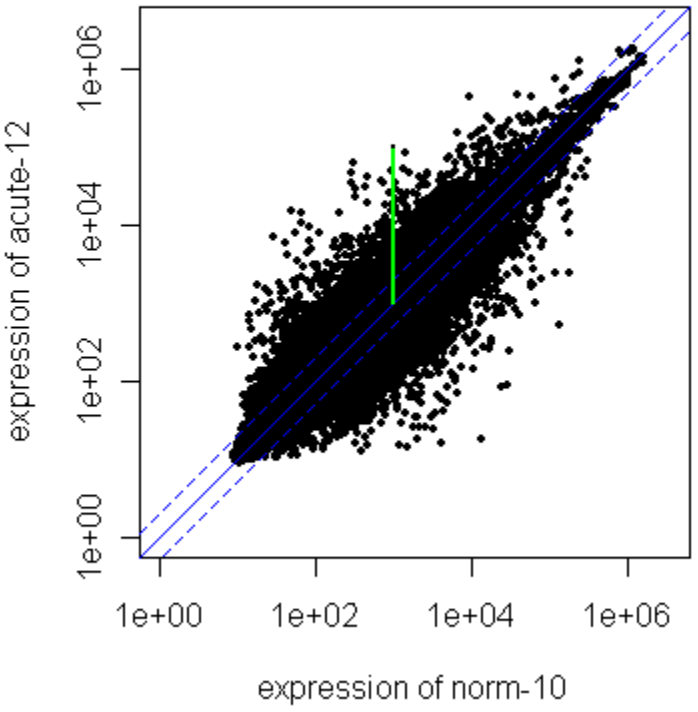
- Average expression

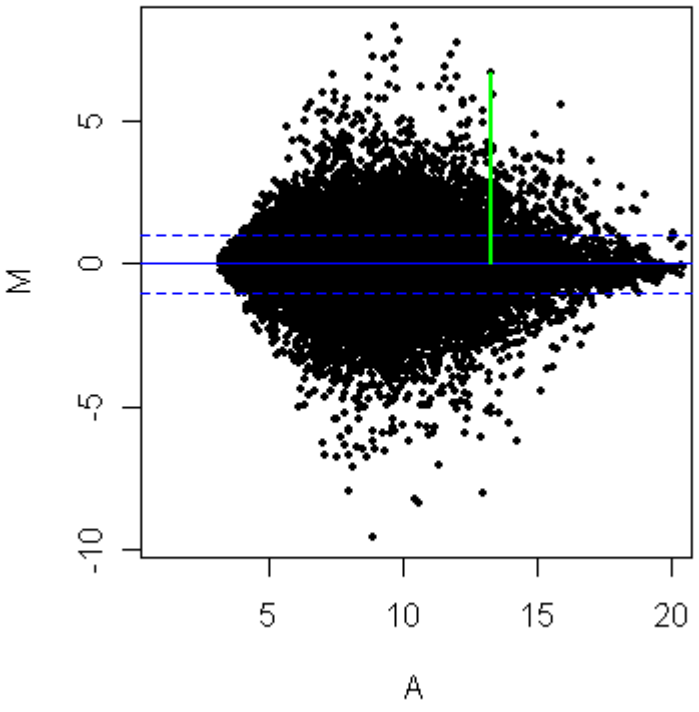
$$A = \frac{\log 2(y_{control}) + \log 2(y_{treated})}{2}$$

- Applicable to pairs of samples
- .. and to pairs of condition with multiple replicates
- In the case of conditions the average M value is computed from the logarithms of the intensities.

M-A Plots

green: log-ratio





Is there something remarkable about the M distribution?

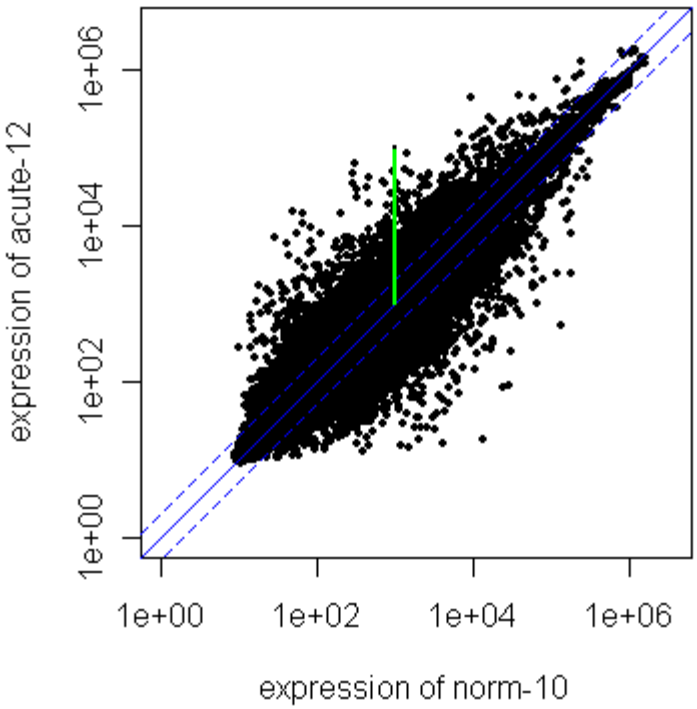
M-A Plots

Observations

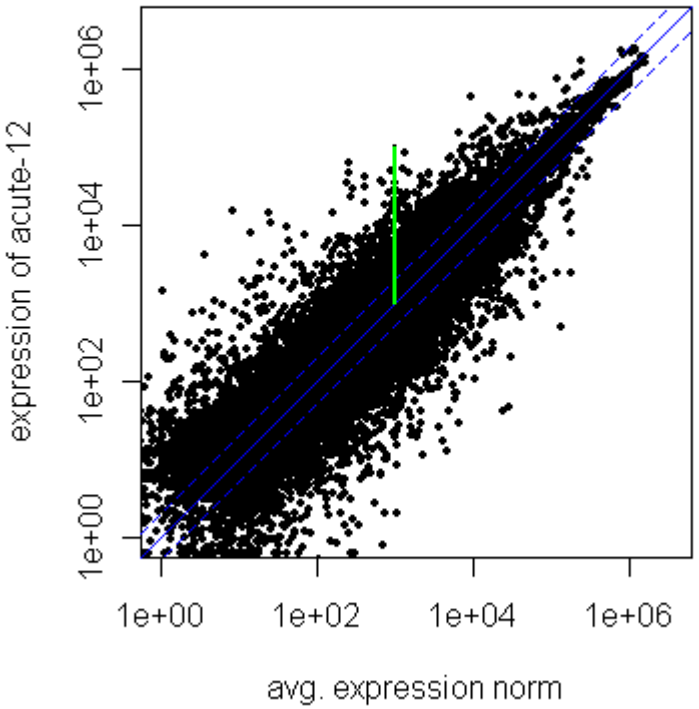
- Many genes not expressed at all but no genes with a measurement signal of 0
  → Low signal values are biased; additive background
- Large M-values are only observed for genes with medium A-value
  → Microarrays underestimate expression changes for high-signal genes
  → If there is an additive background component then the expression changes are also underestimated for low signal genes

Background effect und expression ratios

Standard RMA



Additional background subtraction (BG=40)



- Low-signal genes do now also get large ratios
- This is not wanted, because low signal genes are considered unreliable

Significance Tests

- Up to now: Only effect size
- Now: How significant are the measured expression changes?
- Compute for each gene if it is significantly differentially expressed
- Hypothesis test:

$$x_{treated} = x_{control} + \log(r) + \varepsilon$$

$$\mathrm{H}_0 : \log(r) = 0$$

- Statistical problem: Frequently only 3 to 5 replicates
  - Computation of standard deviations is not precise
- Technical problem in R
  - many tests (~ 25 000)
  - loop is slow
  - special functions

Significance Tests: Student's t-test

- Given a gene
  - log expression control group: $x_{c,i}, i = 1,\ldots,n_c$
  - log expression treatment group: $x_{t,i}, i = 1,\ldots,n_t$
- Assumption
  - Normally distributed (log values!)
  - Same variance
- Problem:
  - Many studies only with 3-5 replicates
  - Assumptions cannot be verified
  - But: Gaussian assumption is the best if one does not know anything about the distribution
- Result for each gene
  - t-statistics $\sim \frac{\mu_c - \mu_t}{\sigma}$
  - p-value

Precision of Estimated Standard Deviation

- The estimate of the standard deviation is unreliable if only a few replicates are available!

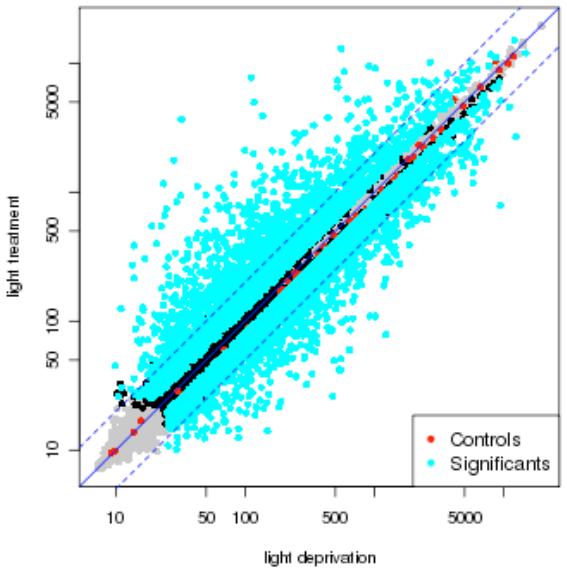
Number of Replicates	Percentage where s over-/underestimates σ by a factor of 2
3	25%
5	9%
10	0.9%

Example: t-test Comparison

- Comparison of 3 against 3 samples with genes significant at p=0.01 highlighted in cyan

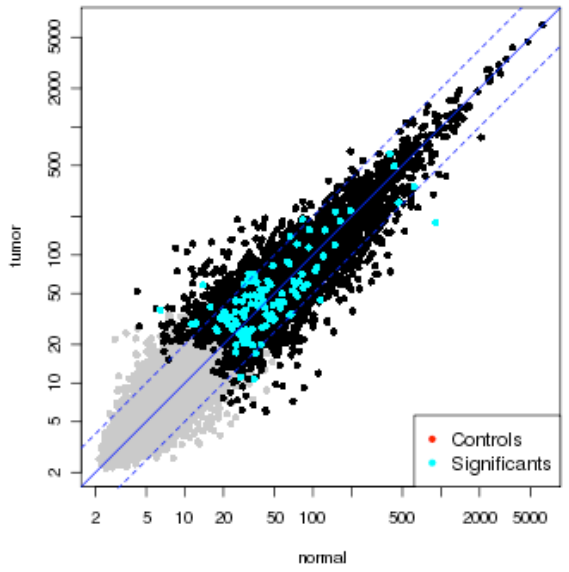
Study A:

- highly consistent replicates; high fold-change genes are significant



Study B:

- inconsistent replicates; only few genes significant; mainly small fold-changes



Comparison of methods

- Notation:
  - avg. log expression control group: μ_c
  - avg. log expression treatment group : μ_t
  - standard deviation: σ
- Statistics used:
  - average log ratio: $\mu_c - \mu_t$
  - t-Test $\frac{\mu_c - \mu_t}{\sigma}$
  - SAM $\frac{\mu_c - \mu_t}{\sigma + \sigma_0}$
  - Bayesian posterior: $\frac{\mu_c - \mu_t}{\sqrt{\sigma^2 + \sigma_1{}^2}}$

SAM: Significance Analysis of Microarrays

- Assumption:
  - few replicates, σ is very imprecise
  - genes with small σ may be statistical significant but those are probably biologically not relevant if the magnitude of the expression change is small
- Ad hoc Solution:
  - Test-statistic for gene i:

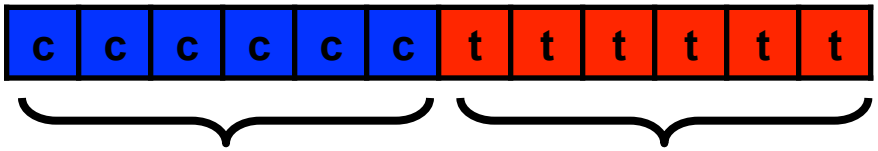
$$d_i = \frac{\mu_{c,i} - \mu_{t,i}}{\sigma_i + \sigma_0}$$

  - Problem: How to decide on significance? No theoretical solution available
  - Solution: Permutation

- Tusher, V. G., Tibshirani, R., and Chu, G. (2001). *Significance analysis of microarrays applied to the ionizing radiation response*. PNAS 98, 5116–5121.

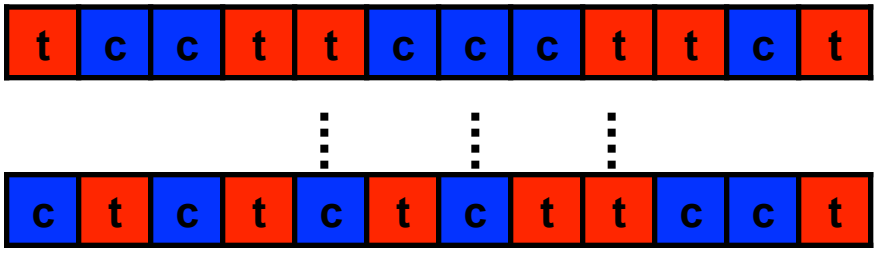
SAM: Permutation analysis

- true test statistic for gene i is computed using:



$$\mu_{ci} \qquad \mu_{ti} \longrightarrow d_i$$

- permuted test-statistic for gene i is computed using:



$$\longrightarrow d_i^1$$

$$\longrightarrow d_i^k$$

- The distribution of the test statistics computed from the permuted data can be used as an approximation of the null distribution.

SAM: Significance Level

- Which threshold to choose for the score d?
- Approach: Call a score significant if

$$d_i > \hat{d} \vee d_i < -\hat{d}$$

- Compute the number of significants
- Compute the number of significants from the permuted data set (as an estimate of the number of false positives)

$$n = \sum_i d_i > \hat{d} \vee d_i < -\hat{d}$$

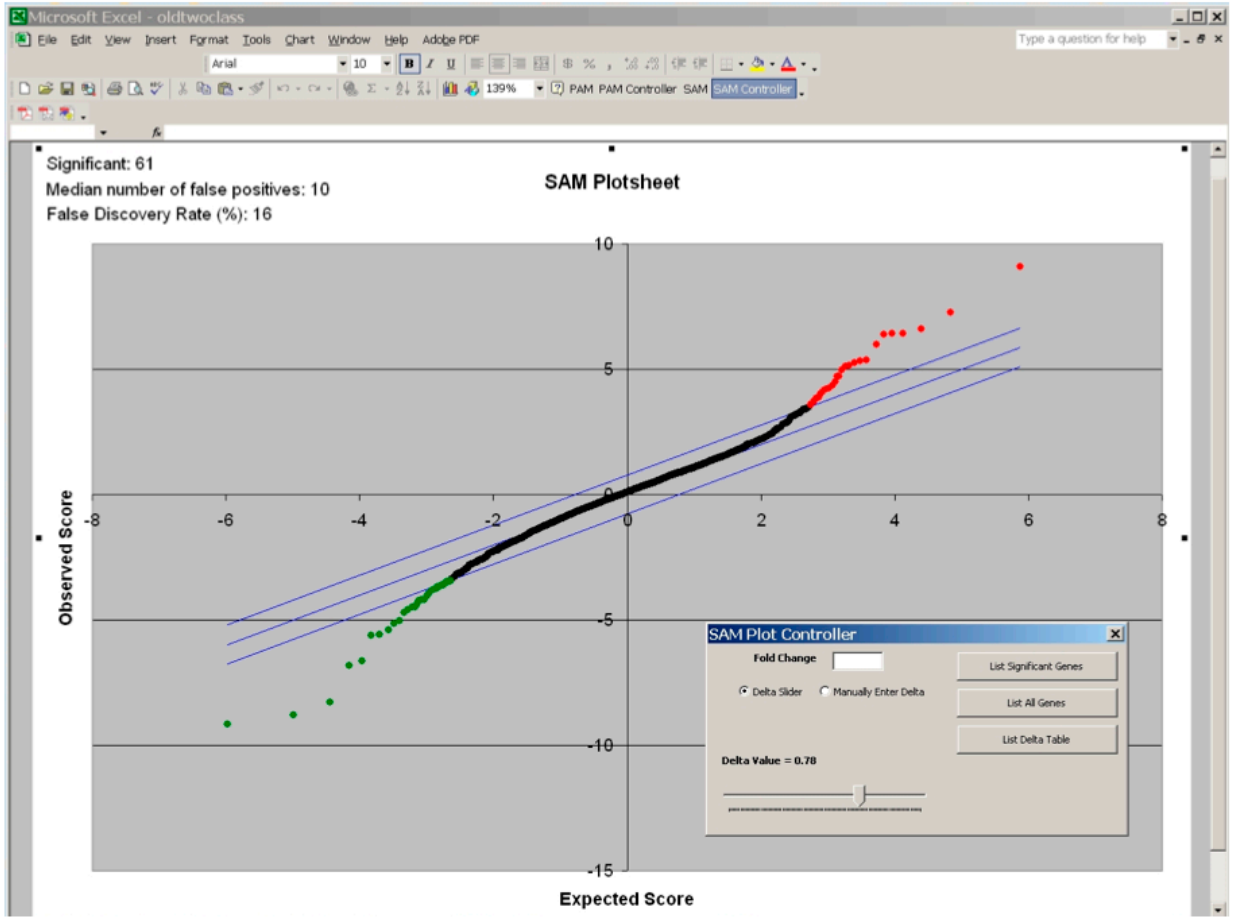
$$m = \frac{1}{K} \sum_{ik} d_i^k > \hat{d} \vee d_i^k < -\hat{d}$$

- False Discovery Rate:

$$\frac{m}{n}$$

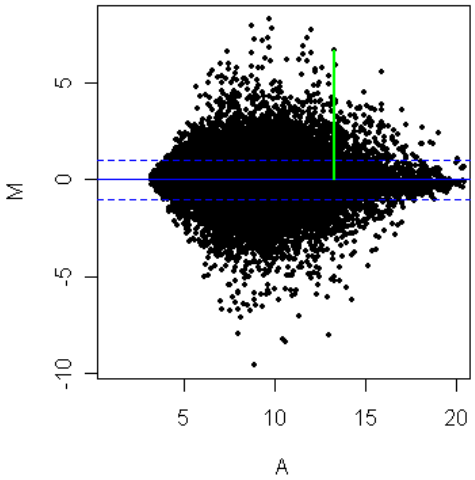
- Computing the p-value (False Positive Rate) is not possible

Quantile-Quantile Plot of Expected and Observed Scores



Variance Stabilization

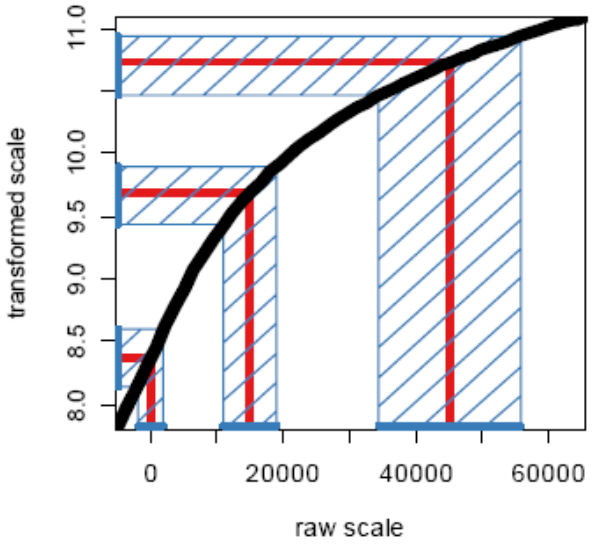
- Up to now: SAM approach with improved estimates of the variance in order to get better estimates of the test statistics



- Variance depends on signal intensity
- However: Differential expression (average log-ratio) is computed independent of the intensity
- Alternative approach: Transform the data so that variances are the same independent of the intensities
- Huber, W., Heydebreck, von, A., Sültmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**, S96–104 (2002).
- Bioconductor package: vsn

Variance Stabilization

- Apply a transform such that the variance is independent of the intensity: Var[h(Y_k)] independent of Mean[h(Y_k)]
- Linear expansion of h in the neighborhood of u=E[Y_k] gives:
  h(Y) = h(u) + h‘(u)·(Y-u)
- Approximately this gives Var[h(Y)] = Var[Y]· h‘(u)².
- We want a transformation that makes Var[h(Y)] constant, independent of u
- The transformation ν(u) can be derived from the differential equation:



$$h'(u) = c / \sqrt{v(u)}$$

$$h(y) = \int_0^y c / \sqrt{v(u)}\, \mathrm{d}u$$

Variance Stabilization

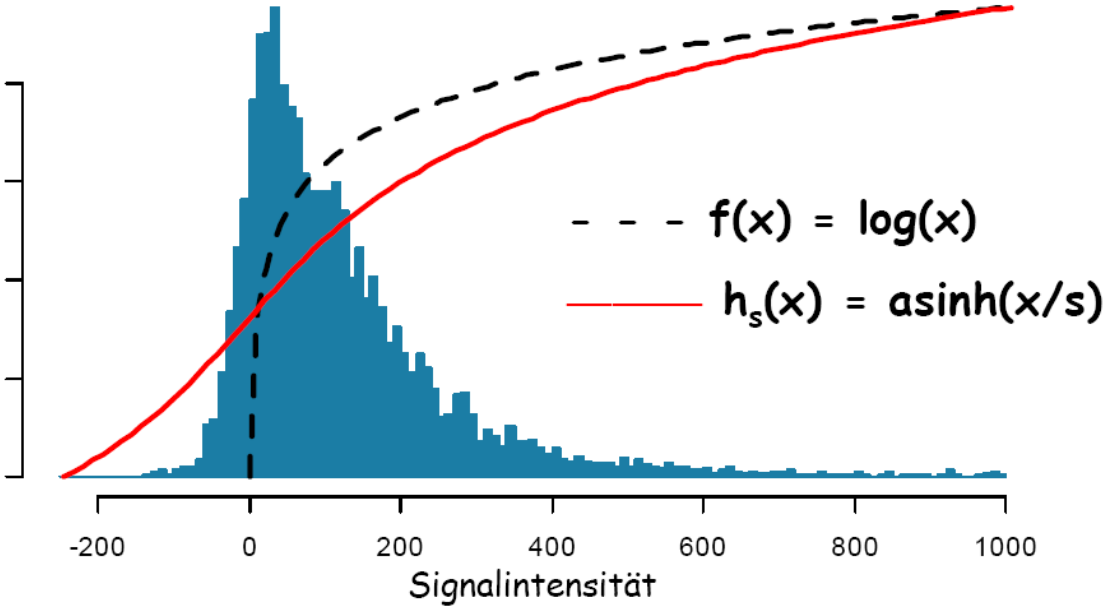
- With

$$v(u) \propto (u + u_0)^2 + s^2$$

- we get the solution:

$$h \propto \operatorname{arsinh} \frac{u + u_0}{s}$$

Variance Stabilization

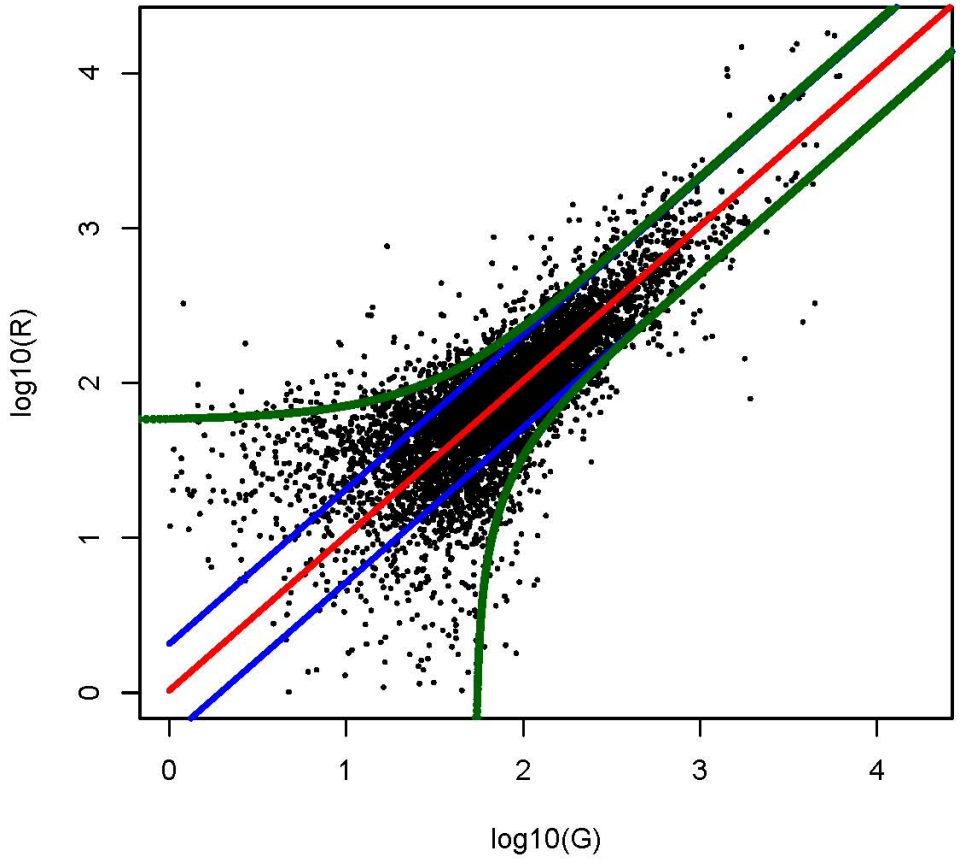


Variance Stabilization

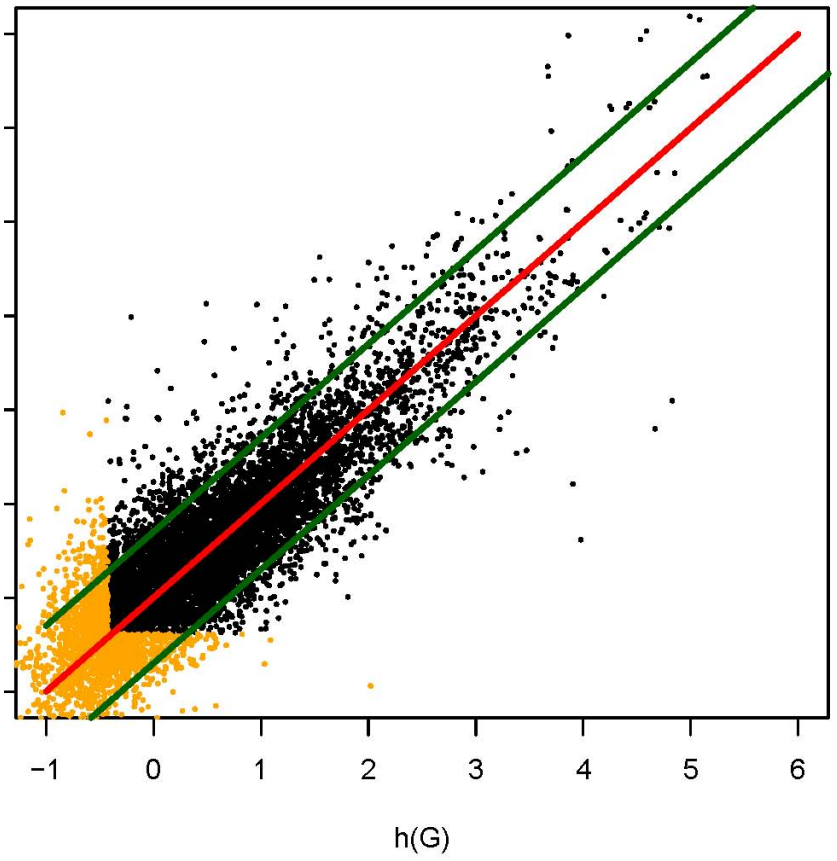
- Parameter estimation:
- Use only the genes that are not differentially expressed

- Model:
  - Robust ML-estimator (least trimmed squares)
  - Differentially expressed genes will behave like outliers and will be removed by the trimming. They do not have an influence on the result as long as the majority is not differentially expressed

Variance Stabilization



log scale



generalized log scale

Robust Tests

- Wilcoxon rank-sum test
  - non-parametric test
  - use the ranks instead of the expression values

- Advantage:
  - Robust with respect to deviations from normality
- Disadvantage:
  - Needs more replicates

- Recommended as alternative to t-test if there are many replicates (e. g. >7)
- Was more important in the beginning of microarrays when microarray measurements were rather error-prone with many outliers

Generalizations

- More than two conditions
- Example: 3 tissue types
  - normal, sick, acute
- Approaches:
  - ANOVA
  - linear model
- Both formalisms are equivalent

- More than one factor
- Example:
  - Tissue type: normal, sick, acute
  - Patient: 02, … , 15
- Approaches:
  - n-way ANOVA
  - linear model

Summary

- Differential expression is computed to get the genes involved in the difference of two conditions
- The computed list of significant genes has to be treated with care, since we compute many tests (20 000) there is the risk of many false positives

- Special methods take into account the specific properties of microarray data
  - Signal distribution
  - Large number of genes that can be considered independent
  - Low number of replicates