

Freitag
15. April 2005
9:15

Bioinformatik Ringvorlesung Sommersemester 2005

TP3 – INF 580

t.beissbarth@dkfz.de

a.tresch@dkfz.de

Dr. Tim Beißbarth + Dr. Achim Tresch
Deutsches Krebsforschungszentrum
Molekulare Genomanalyse
Bioinformatik und Datenanalyse



MGA

Molekulare Genomanalyse -
Bioinformatik und Datenanalyse

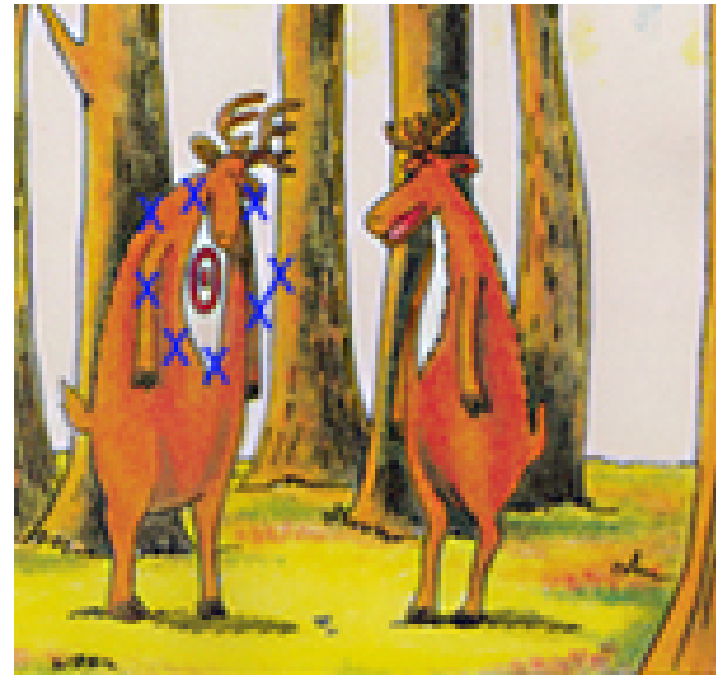
dkfz.

DEUTSCHES
KREBSFORSCHUNGSZENTRUM
IN DER HELMHOLTZ-GEMEINSCHAFT

Überblick

- Differentielle Gene finden
 - T-test
 - Wilcoxon Test (signed Rank Test, Man-Whitney test)
 - Multiple Testing
 - Permutationstests (SAM)
 - Modifizierter T-test
 - Empirical Bayes
- Komplexes Experiment-Design
 - Lineare Modelle
 - Gutes Design
- Normalisierung
 - Varianz -Stabilisierung

Statistik:



Differentielle Gene finden

- Wir haben in n Experimenten jeweils die differentielle Genexpression zwischen zwei Bedingungen gemessen, z.B. WT/KO.

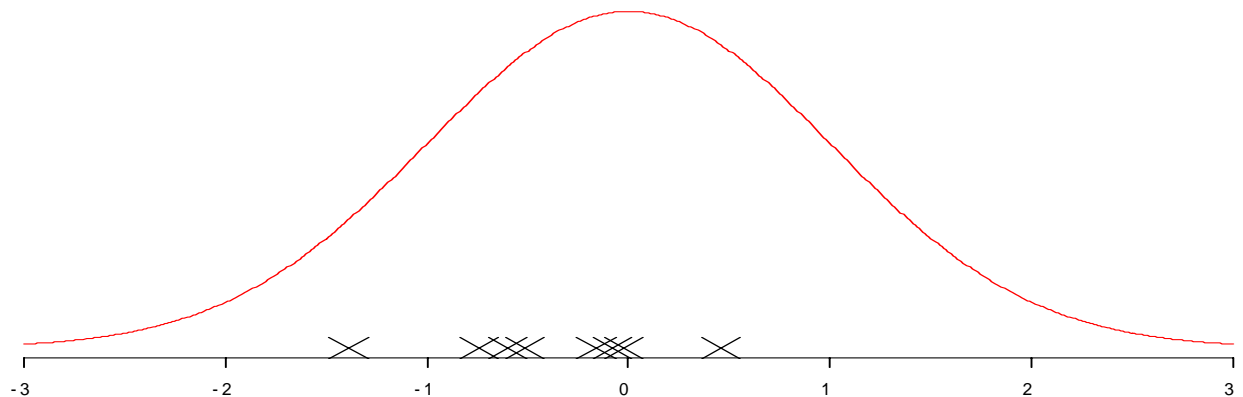
- Messwerte M-values für Gen g : x_1, \dots, x_n × × × × × × × ×

- Hypothese H_0 - die Gene sind nicht differentiell:

$$E(x) = 0$$

- Annahme – die Meßwerte sind Normalverteilt $N(\mu_0, \sigma_0)$.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Differentielle Gene finden

- Bester Schätzer für μ : (Mittelwert)

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

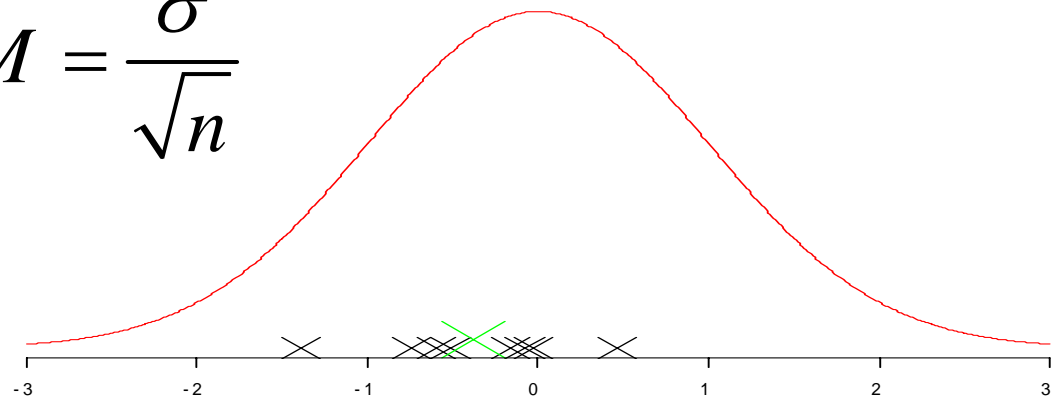
- Bester Schätzer für σ : (Standardabweichung)

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

- Erwartete Abweichung von \bar{x} von μ :

Standard error of the mean

$$SEM = \frac{\sigma}{\sqrt{n}}$$



T-test

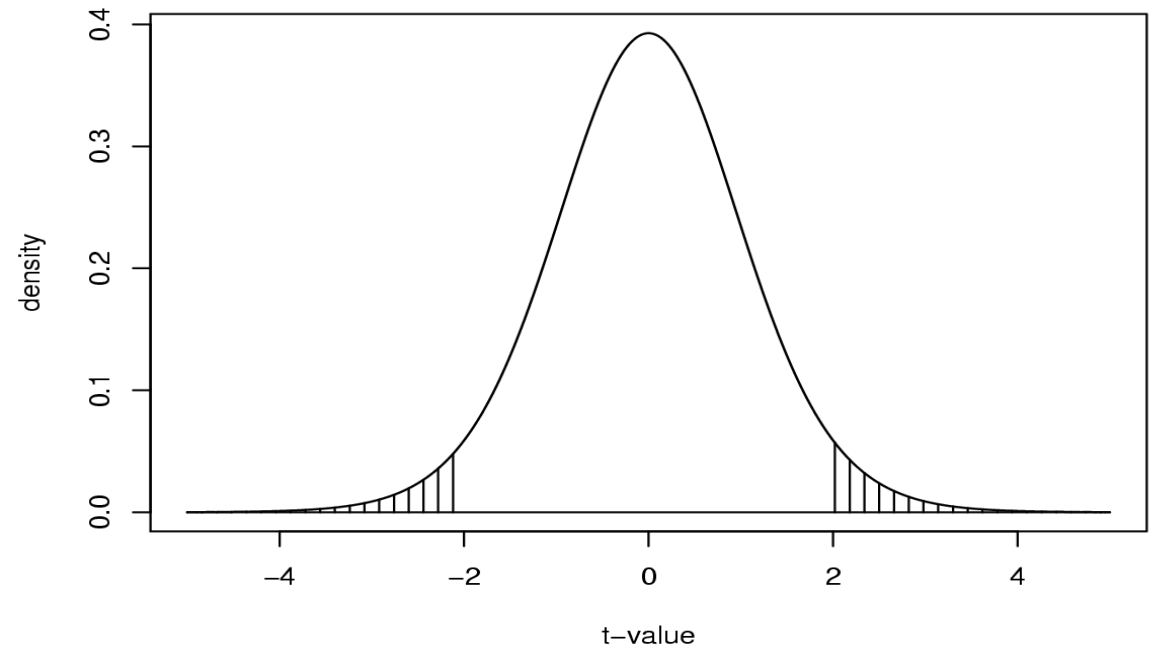
- Vergleiche beobachtete Abweichung von μ_0 mit erwarteter:

$$T(x) = \frac{\bar{x} - \mu_0}{SEM}$$

- t-Verteilung mit n-1 Freiheitsgraden.

- Berechne p-value.

- Wenn $T(x)$ außerhalb des Akzeptanzbereichs lehne H_0 ab.



Beispiel 2 – T-test 2

- Affymetrix Daten von Golub et al, 1998
- 38 Tumor Gewebe:
 - 27 acute lymphoblastic leukemia (ALL)
 - 11 acute myeloid leukemia (AML)
- 6817 Gene, 3051 nach filtern
- Expressionswerte von Gen g in ALL x_1, \dots, x_n und in AML y_1, \dots, y_n .
- Berechne gemeinsame Varianz:

$$s^2 = \frac{1}{n_x + n_y - 2} \left(\sum_{i=1}^{n_x} (x_i - \bar{x})^2 + \sum_{i=1}^{n_y} (y_i - \bar{y})^2 \right)$$

- 2 Sample T-test:

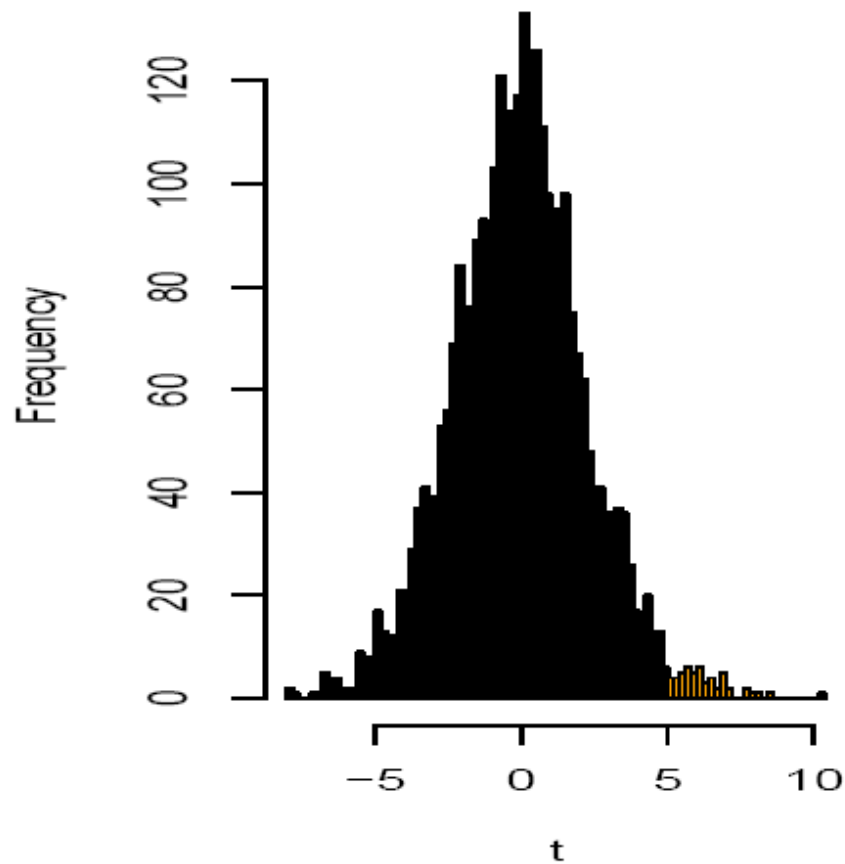
$$T(x, y) = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

t-Verteilung mit $n_x + n_y - 2$ Freiheitsgraden.

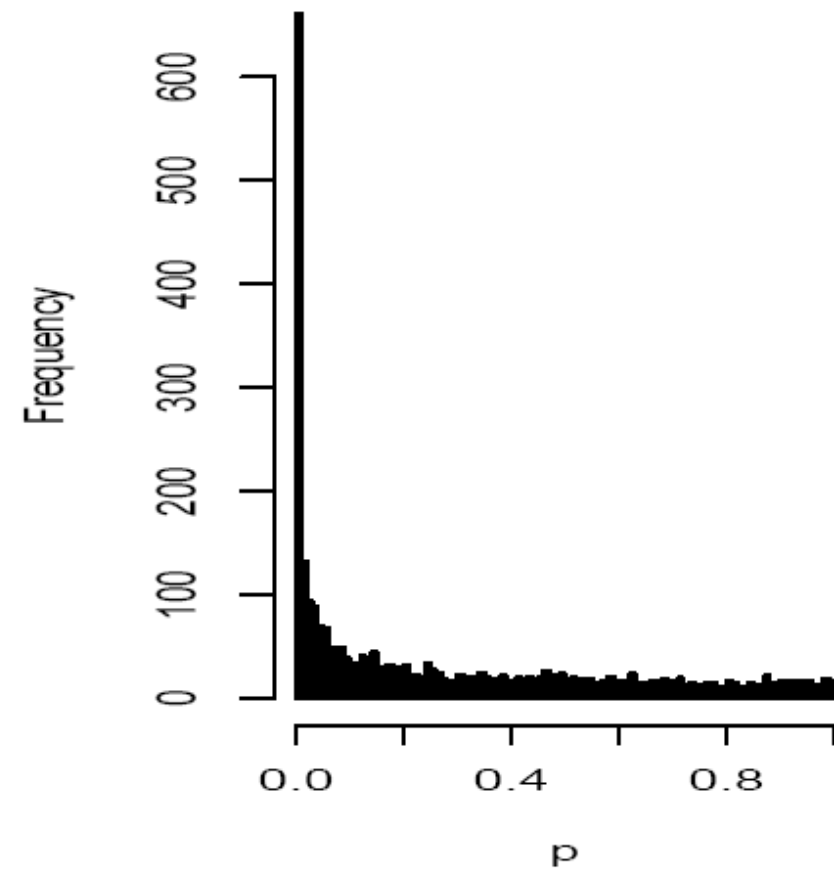
Beispiel

Golub data, 27 ALL vs. 11 AML samples, 3,051 genes.

Histogram of t



histogram of p -values



t -test: 1045 genes with $p < 0.05$.

Bedeutung des p-Wertes: Typ I Fehler

	# non-rejected hypotheses	# rejected hypotheses	
# true null hypotheses (non-diff. genes)	U	V Type I error	m_0
# false null hypotheses (diff. genes)	T Type II error	S	m_1
	$m - R$	R	m

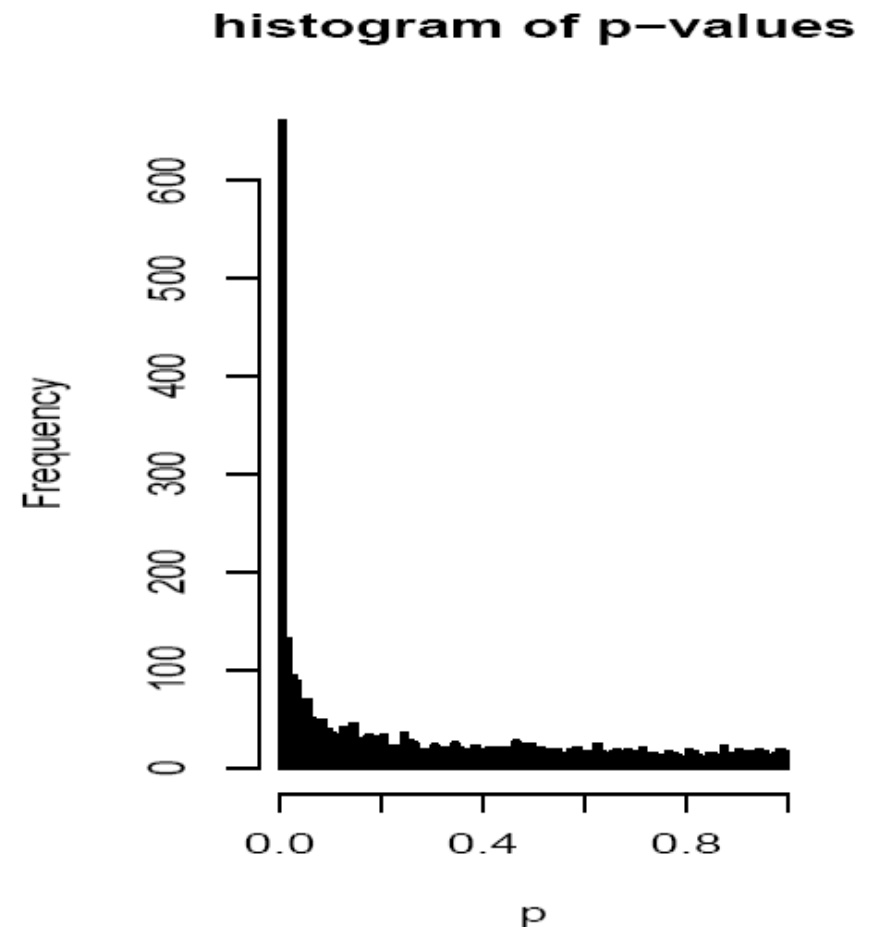
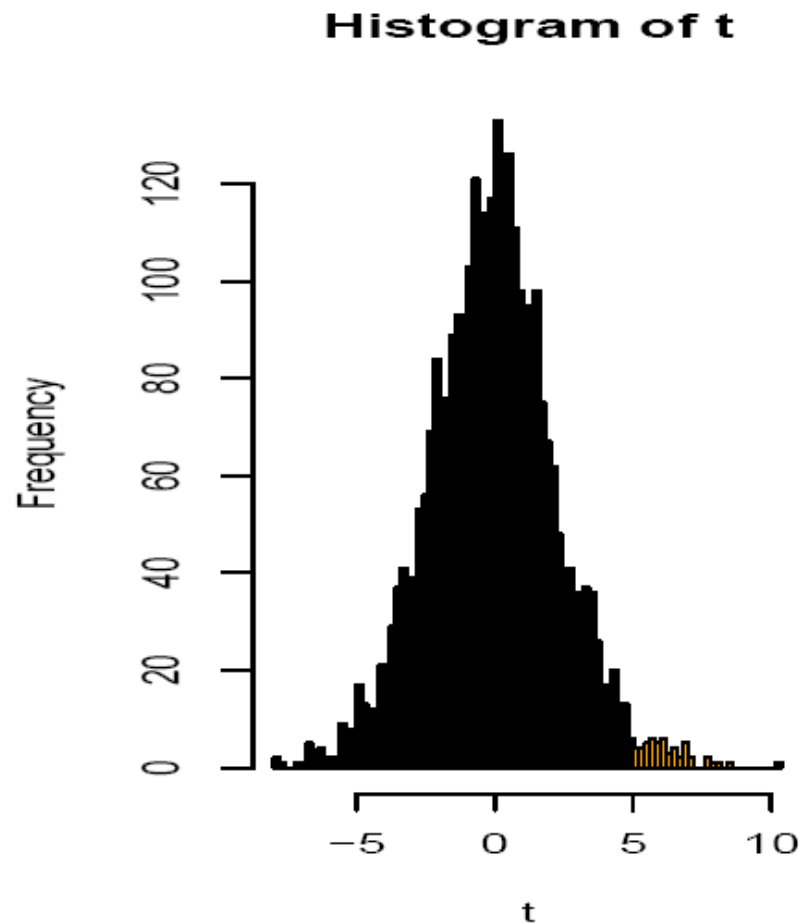
From Benjamini & Hochberg (1995).

Multiples Testen

- Problem: Tausende von Hypothesen werden gleichzeitig getestet.
- Beispiel: Bei 10000 Genen auf dem Chip und einem Cutoff des p-Wertes von 0.01 erwarte ich, daß $10000 \times 0.01 = 100$ Gene einen signifikanten p-Wert $p < 0.01$ haben.
- Resultat: Ein einzelner p-Wert von 0.01 indiziert nicht mehr unbedingt ein signifikantes Gen. Es gibt eine erhöhte Chance falsch-positive Gene zu finden.
- Lösung: Man muß die p-Werte für multiples Testen korrigieren.
- Methode: Einfachstes Verfahren von Bonferroni
Multipliziere alle p-Werte mit der Anzahl der Tests.
- Mehr dazu am Montag 18.4.

Beispiel

Golub data, 27 ALL vs. 11 AML samples, 3,051 genes.



98 genes with Bonferroni-adjusted $\tilde{p}_g < 0.05 \Leftrightarrow p_g < 0.000016$
(t-test)

T-test: Variante von Welch

- Erlaube verschiedene Varianzen in den beiden verschiedenen Stichproben.
- Wir nehmen an $x \sim N(\mu_1, \sigma_1)$ und $y \sim N(\mu_2, \sigma_2)$.
- Testen Hypothese $H_0: \mu_1 = \mu_2$.

$$T(x, y) = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

Mehr Tests: Wilcoxon-Test (auch Man-Whitney Test)

- Nicht-parametrischer Test (keine Normalverteilungsannahme) zum Vergleichen von zwei empirischen Verteilungen.
- Berechne die Ränge der Werte aus beiden Messreihen:

Observations:	0.3	0.5	0.8	0.9	1.3	2.4
Ranks:	1	2	3	4	5	6
Groups:	1	1	1	2	2	2

- Die Teststatistik wird aus der Summe der Ränge berechnet: $R_1=6$
- Für kleine Stichprobengrößen kann die Verteilung der Teststatistik exakt berechnet werden (i.e. alle Möglichkeiten), für große Stichproben kann eine Approximation durch Normalverteilung benutzt werden.
- **Vorteile:** Nicht-parametrisch, robust gegen Ausreißer.
- **Nachteile:** weniger Mächtig da keine Verteilungsannahmen.

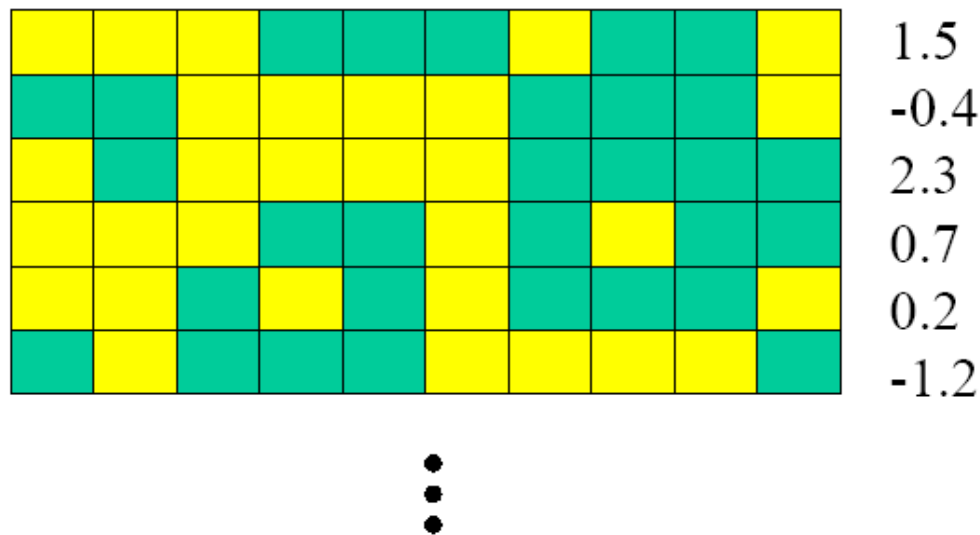
Permutationstests

true class labels:

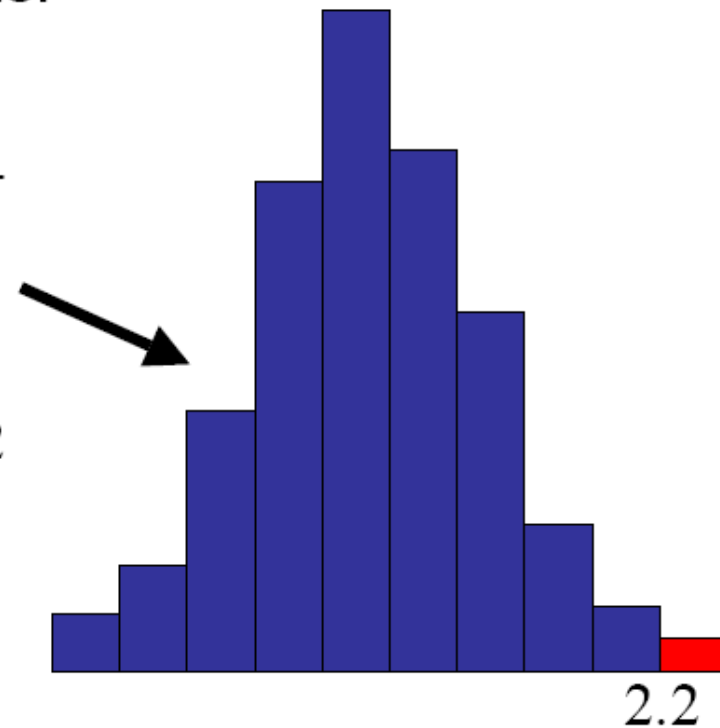


test statistic

(random) permutations of class labels:



null distribution of
test statistic



- **Vorteile:** kein statistisches Modell notwendig.
- **Nachteile:** evtl. sehr rechenintensiv.

Problem beim T-test

- **Problem:**
 - Es gibt sehr viele Gene (Tests) meistens aber nur sehr wenige Wiederholungen.
 - Als Schätzer für σ^2 habe ich s^2 benutzt → evtl. zufällige Fehler bei s^2 .
- **Beispiel:** Gen g wird mit den M-Werten 0.0011, 0.0012 und -0.0009 gemessen → $T \sim 12$, $p \sim 0.007$.
- **Merke:** Bei sehr kleiner gemessener Varianz wird der Wert für T sehr groß.
- **Fazit:** Bei Microarray Experimenten nach Möglichkeit keinen Standard-T-test verwenden.
- **Folgerung:** Modifizierte T-Statistik nötig.

moderated T-statistics

- Beim T-test schätzen wir die Varianz für jedes Gen s_g^2 einzeln. Dies ist evtl. unstabil.
- Stattdessen versuchen wir nun die Varianz über alle Gene s_0^2 oder Subgruppen von Genen zu schätzen und damit die Varianzschätzung zu korrigieren.
- Dieser „Fudge Factor“ kann noch durch Faktoren (α, β) unterschiedlich gewichtet werden.

$$T(x, y) = \frac{\bar{x} - \bar{y}}{\sqrt{\alpha s^2 + \beta s_0^2}}$$

- Referenz: Efron/Tibshirani, Genet. Epidemiol., 2000
- Software: R/Bioconductor Pakete – limma, siggenes

Empirical Bayes

- Methode um über viele Gene moderierte Statistiken zu Berechnen unter Annahme von Prior-Distributions.
- Verschiedene Varianten existieren. Siehe Efron et al 2001, Lönsted/Speed 2002, ...
- Beispiel:

Für große n :

$\approx t = M. / s$

$$B = \text{const} + \log \left(\frac{\frac{2a}{n} + s^2 + M_{\bullet}^2}{\frac{2a}{n} + s^2 + \frac{M_{\bullet}^2}{1 + nc}} \right)$$

Komplexere Designs

Zuordnung der Samples zu verschiedenen Slides

A Typen von Samplen

- Replication – technische, biologische.
- Gepoolt vs individuelle Samples.
- Gepoolt vs amplifizierte Samples.

B Verschiedene Design Layouts

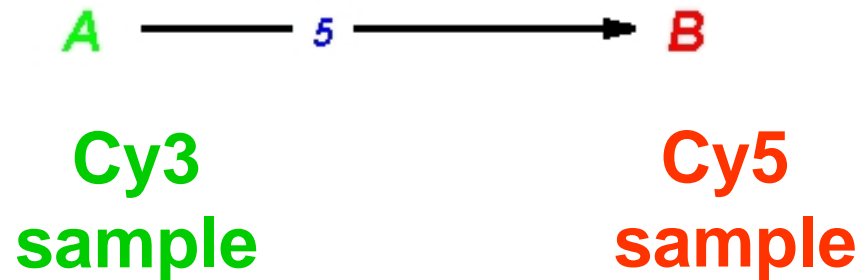
- Ziel des Experiment.
- Robustheit.
- Erweiterbarkeit.
- Effizienz.

Berücksichtigen von physikalischen und Kosten-limitationen:

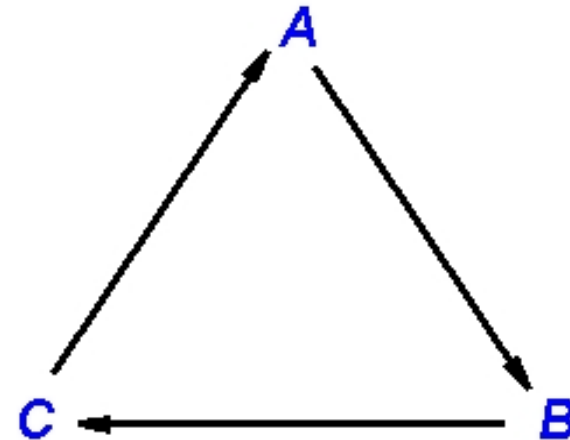
- Anzahl der Slides.
- Menge an verbrauchtem biologischen Material.

Graphische Representation

Knoten: mRNA Samples;
Kanten: Hybridisierung;
Richtung: dye assignment.



(a)

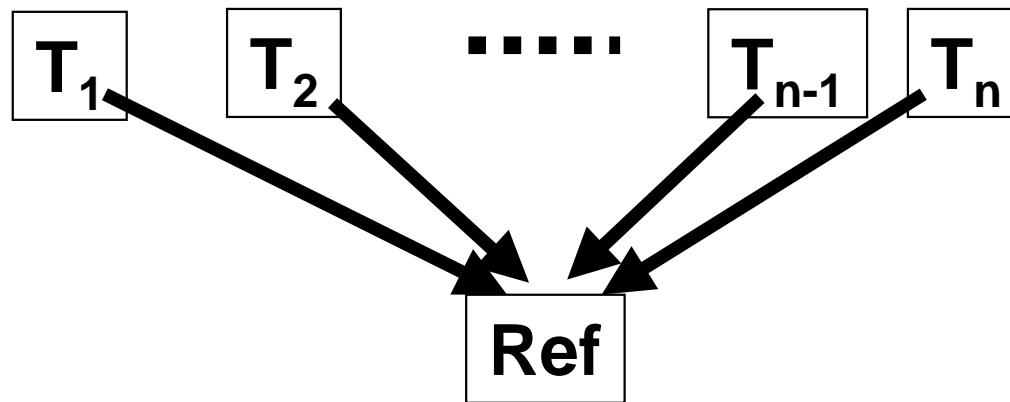


(b)

Graphische Representation

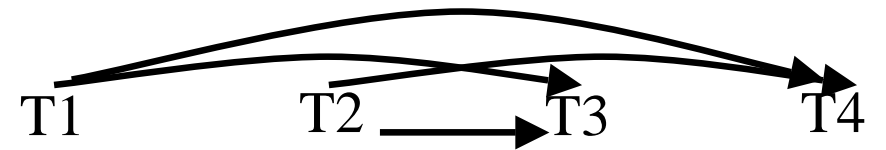
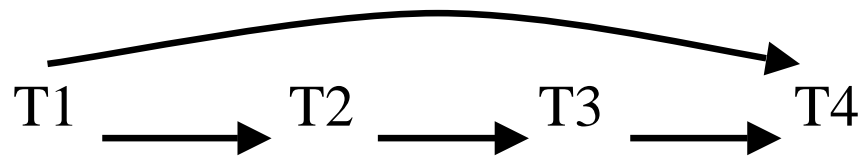
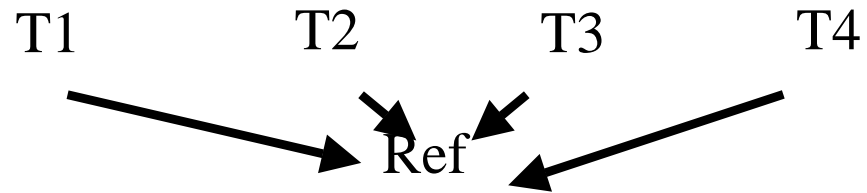
- Die Struktur des Graphen legt fest, welche Effekte geschätzt werden können und wie präzise die Schätzungen sind.
 - Zwei mRNA Samples können nur verglichen werden, wenn es einen **Pfad** gibt, welcher die zugehörigen Knoten verbindet.
 - Die **Präzision** der geschätzten Kontraste hängt direkt von der **Anzahl der Pfade**, welche die Knoten verbinden, ab und ist invers proportional zur **Länge dieser Pfade**.
- Direkte Vergleiche auf demselben Slide geben präzisere Messungen als indirekte Vergleiche.

Common Reference Design

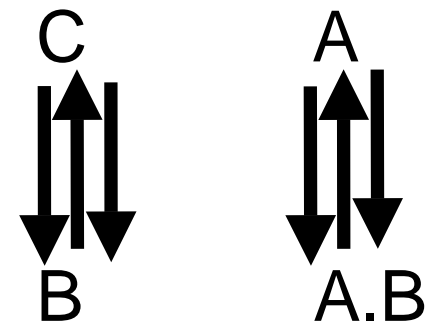
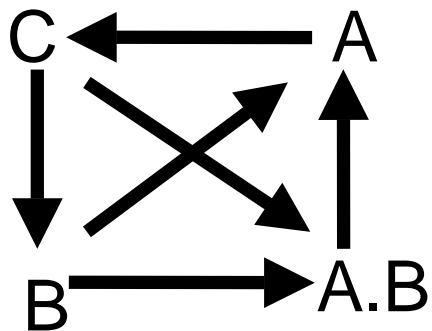
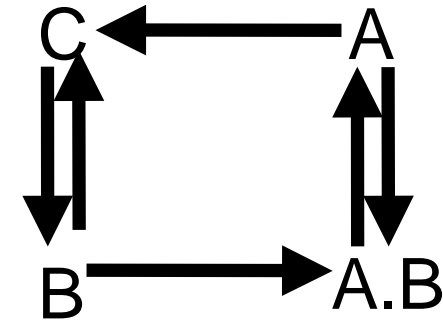
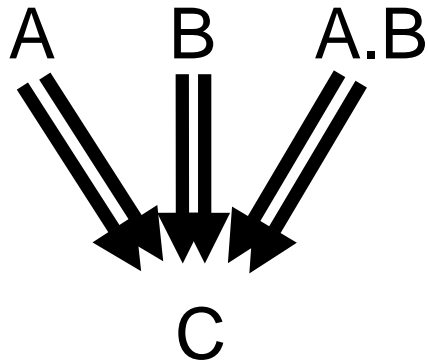


- Experimente bei denen ein Common Reference Design angebracht ist:
 - Wenn es eine aussagekräftige biologische Kontrolle gibt.
 - Vergleiche mit vielen verschiedenen Bedingungen.
- Vorteile:
 - Leichte Interpretation.
 - Erweiterbarkeit.

Zeitreihen



Verschiedene Designs – 4 Bedingungen



Direkte und Indirekte Vergleiche

Zwei Samples (A vs B)
e.g. KO vs. WT oder Mutante vs. WT

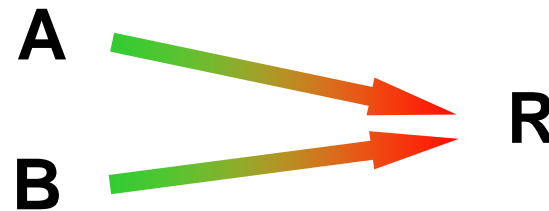
Direct



Mittelwert ($\log (A/B)$)

$$\sigma^2 / 2$$

Indirekt



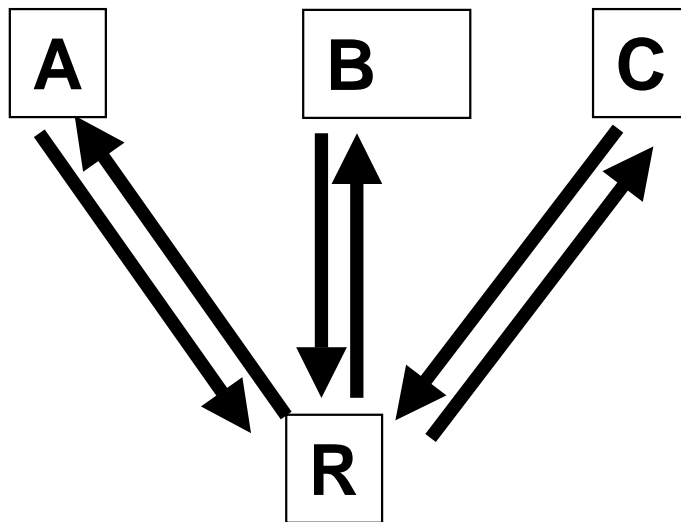
$\log (A / R) - \log (B / R)$

$$2\sigma^2$$

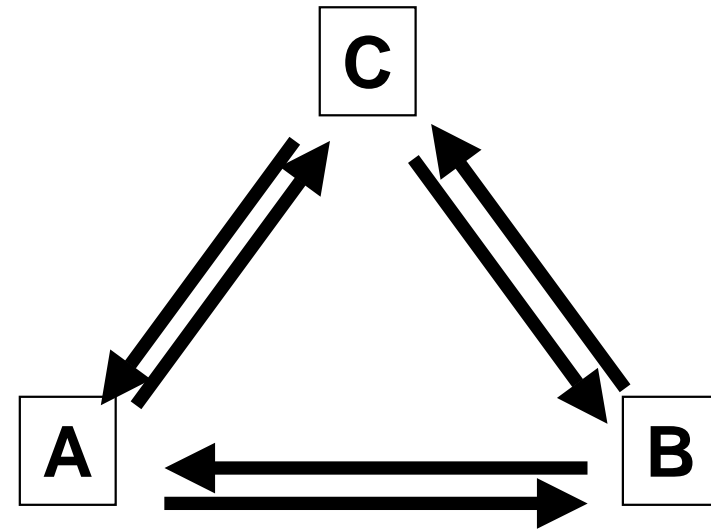
Diese Berechnungen nehmen Unabhängigkeit zwischen den Replikaten an:
die Wirklichkeit ist schwieriger.

K Bedingungen Vergleichen

(i) Indirektes Design



(ii) Direktes Design

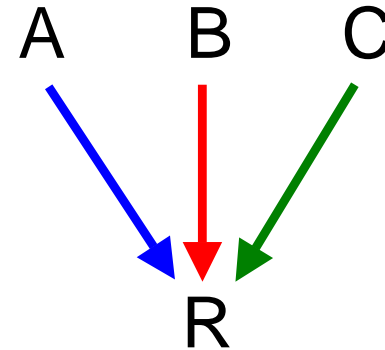


Frage: Mit welchem Design kann ich die Vergleiche A-B, A-C, und B-C am genauesten berechnen?

Common reference design

Log ratios: y

Model: $E(y_1) = A - R$
 $E(y_2) = B - R$
 $E(y_3) = C - R$



Frage 1: Berechne die log-ratios zwischen Sample A and B (= $a-b$)

Berechne $y_1 - y_2$ für jedes Gen.

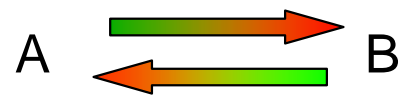
Frage 2: Which genes are differentially expressed in *at least one* of the samples relative to R .

Berechne die **T-statistik** für jedes Gen.

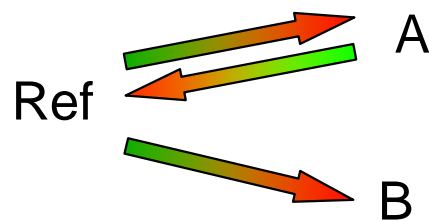
Lineare Modelle um differentielle Expression zu messen



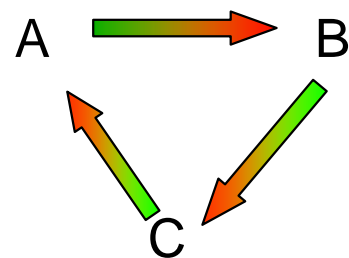
$$y = \log_2(R) - \log_2(G) \equiv B - A$$



$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \beta \quad \beta \equiv B - A$$



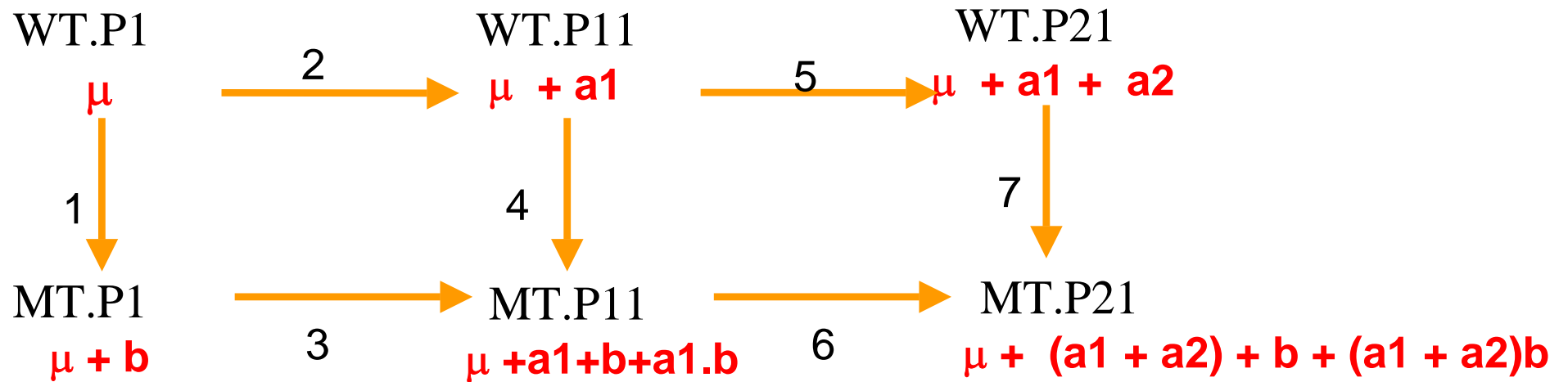
$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \begin{aligned} \beta_1 &\equiv A - \text{Ref} \\ \beta_2 &\equiv B - A \end{aligned}$$



$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \begin{aligned} \beta_1 &\equiv B - A \\ \beta_2 &\equiv C - A \end{aligned}$$

Erlaubt, daß alle Vergleiche simultan ausgewertet werden.

Ein etwas größeres Beispiel:



$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{pmatrix} a_1 \\ a_2 \\ b \\ a_1 b \\ a_2 b \end{pmatrix}$$

Lineare Modelle berechnen

Entwerfe lineares Modell für jedes Gen g

$$E(\underline{y}_g) = X \underline{\beta}_g \quad \text{var}(\underline{y}_g) = W_g^{-1} \sigma_g^2$$

Schätze Modell durch **robuste Regression**, **least squares** oder **generalized least squares** (in R Funktionen lm, rlm, glm) und erhalte

coefficients

$$\hat{\beta}_{gj}$$

standard deviations

$$s_g$$

standard errors

$$\text{se}(\hat{\beta}_{gj})^2 = c_{gj} s_g^2$$

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**. *Biostatistics*.
- 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 expression 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

Lineare Modell Analyse

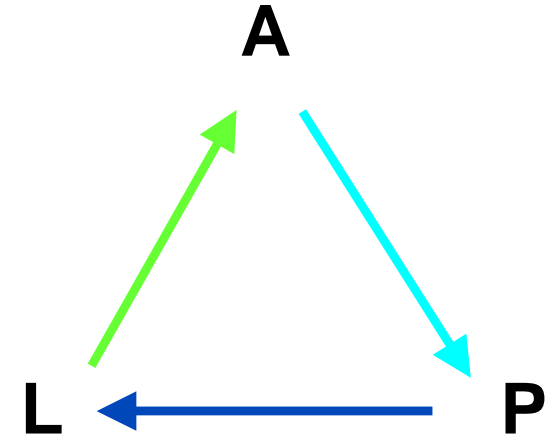
Log ratios: y

Parameters: $\beta = (a-p, l-p)$, where
 $a = \log_2 A$, $p = \log_2 P$ and $l = \log_2 L$

Model:

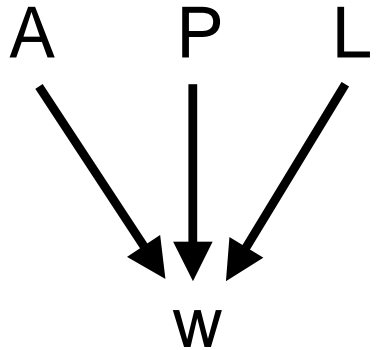
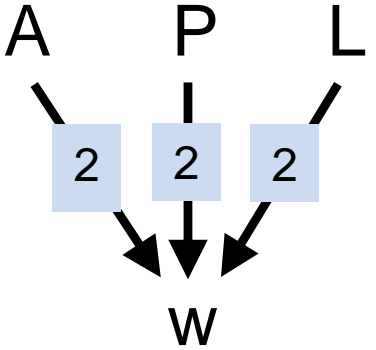
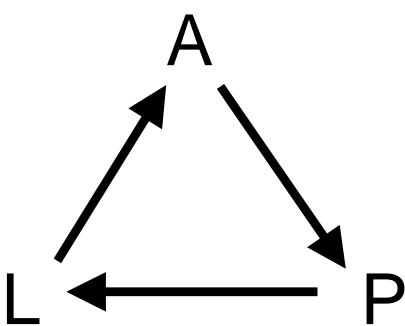
$$E \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} a-p \\ l-p \end{pmatrix}$$

$$\Sigma = \text{Cov} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \sigma^2 & \chi_1 & \chi_1 \\ \chi_1 & \sigma^2 & \chi_1 \\ \chi_1 & \chi_1 & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$



→ $\hat{\beta} = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} y$

$$\text{Cov}(\hat{\beta}) = (X' \Sigma^{-1} X)^{-1}$$

	I (a) Common reference	I (b) Common reference	II Direct comparison
			
Anzahl der Slides	N = 3	N=6	N=3
Mittlere Varianz	2		0.67
Verbrauchtes Material	A = P = L = 1	A = P = L = 2	A = P = L = 2
Mittlere Variance		1	0.67

Für $k = 3$, Effizienzrate (Design I(b) / Design II) = 1.5
Generell, Effizienzrate = $k / (k-1)$

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

Acknowledgements – Slides geborgt von

- Anja von Heydebreck
- Jean Yang
- Terry Speed
- Wolfgang Huber

4/15/2005

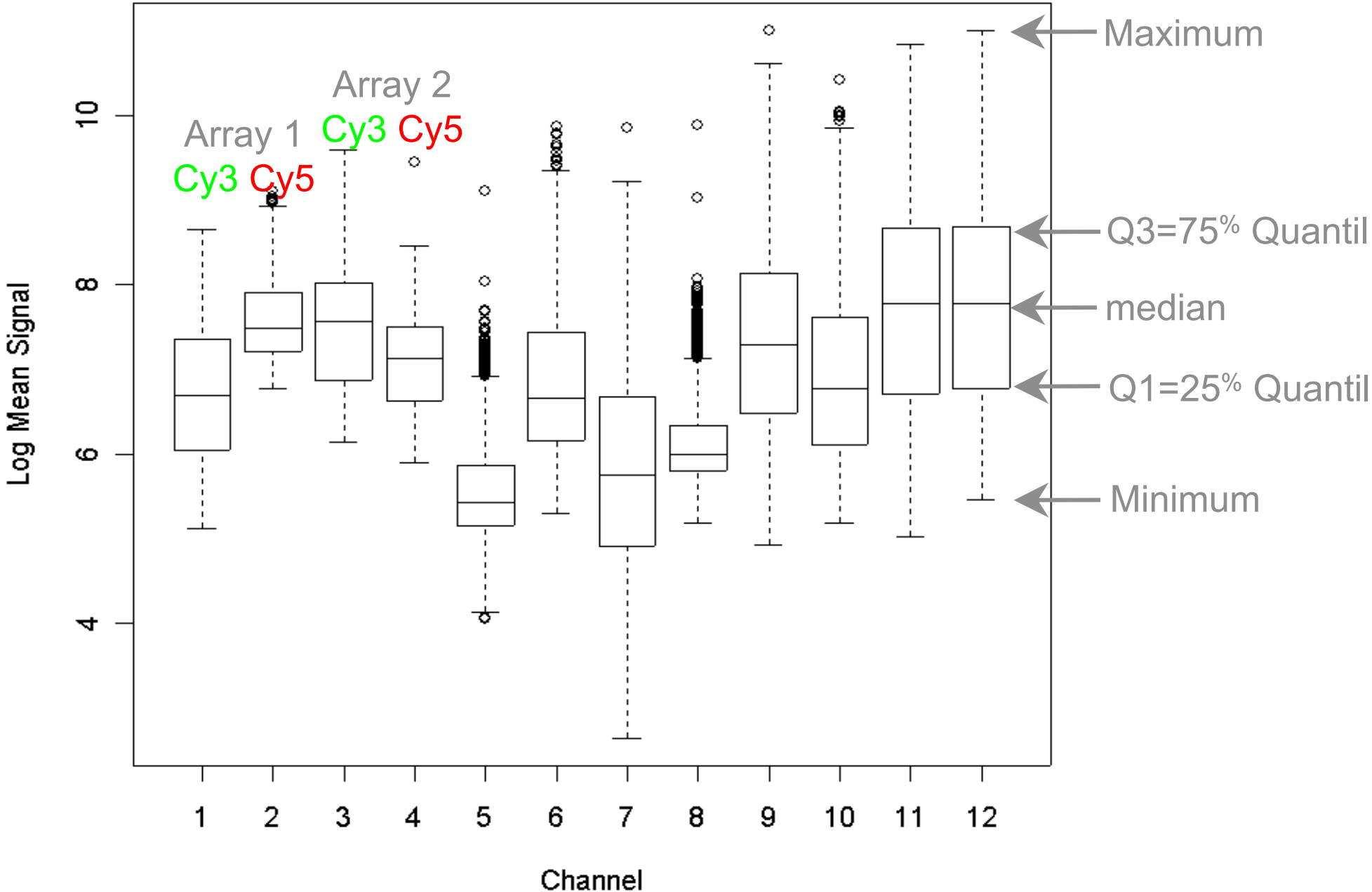
Normalisierungsverfahren: Quantilnormalisierung, Varianzstabilisierung

Achim Tresch

dkfz.

GERMAN
CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION

Quellen der Variabilität bei Microarray-Messungen



Quellen der Variabilität bei Microarray-Messungen

- RNA-Menge in der Gewebeprobe/Zellprobe
- RNA-Extraktionseffizienz
- RNA-Amplifikationseffizienz
- Labelingeffizienz
- Fluoreszenzdetektion

Systematisch

- Ähnliche Effekte auf alle (viele) Messungen
- Effekte können aus den Messungen heraus geschätzt werden



Normalisierung

- Reinheit der DNA-Sonden
- Spottingeffizienz
- Kreuzhybridisierung / unspezifische Hybridisierung
- Streusignale

Stochastisch

- Effekte, die einzelne Spots betreffen
- Effekte einzeln nicht berechenbar, „Rauschen“



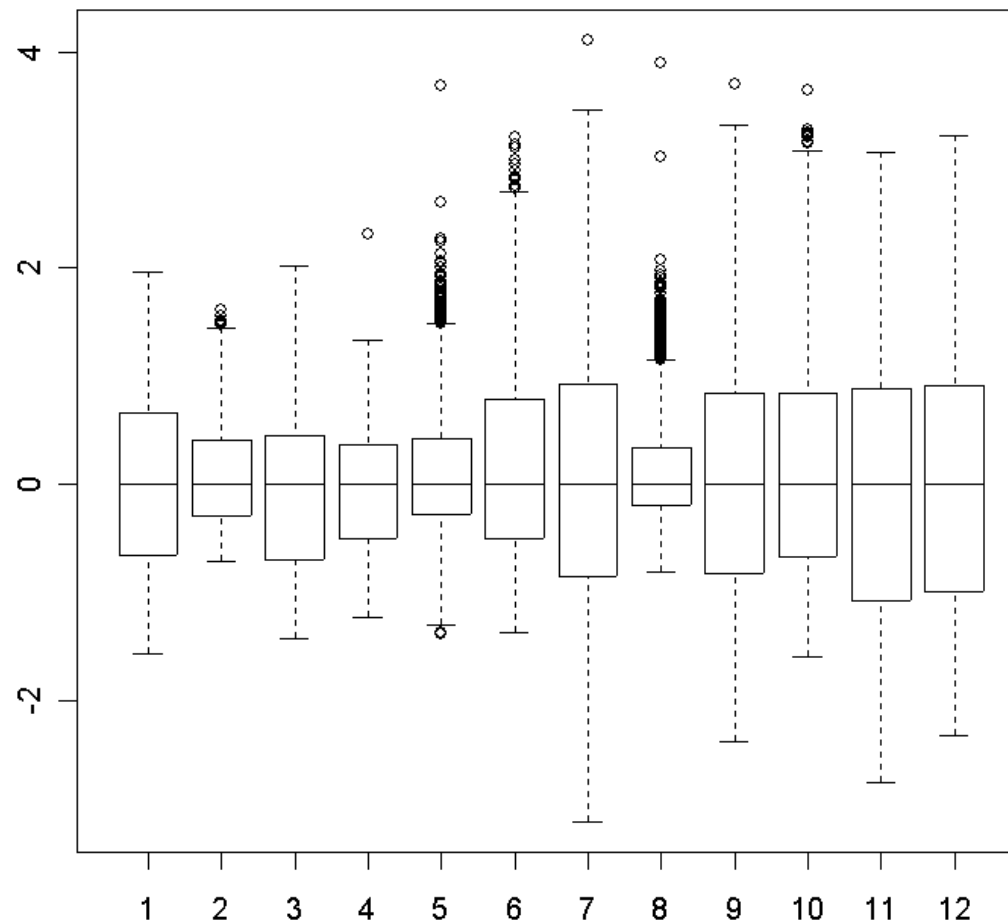
Fehlermodell

Medianzentrierung

Eine der einfachsten Strategien besteht darin, die „Zentren“ aller Arraydaten auf das gleiche Niveau zu bringen, da man annimmt, dass diese bei allen Microarrays in Wirklichkeit in etwa gleich sind. Als robustes Maß für das Zentrum eines Datensatzes wird gerne der Median verwendet. Hieraus ergibt sich folgende Normalisierungsvorschrift (Medianzentrierung):

Teile jeweils alle Expressionswerte eines Arrays durch den Median seiner Expressionswerte.

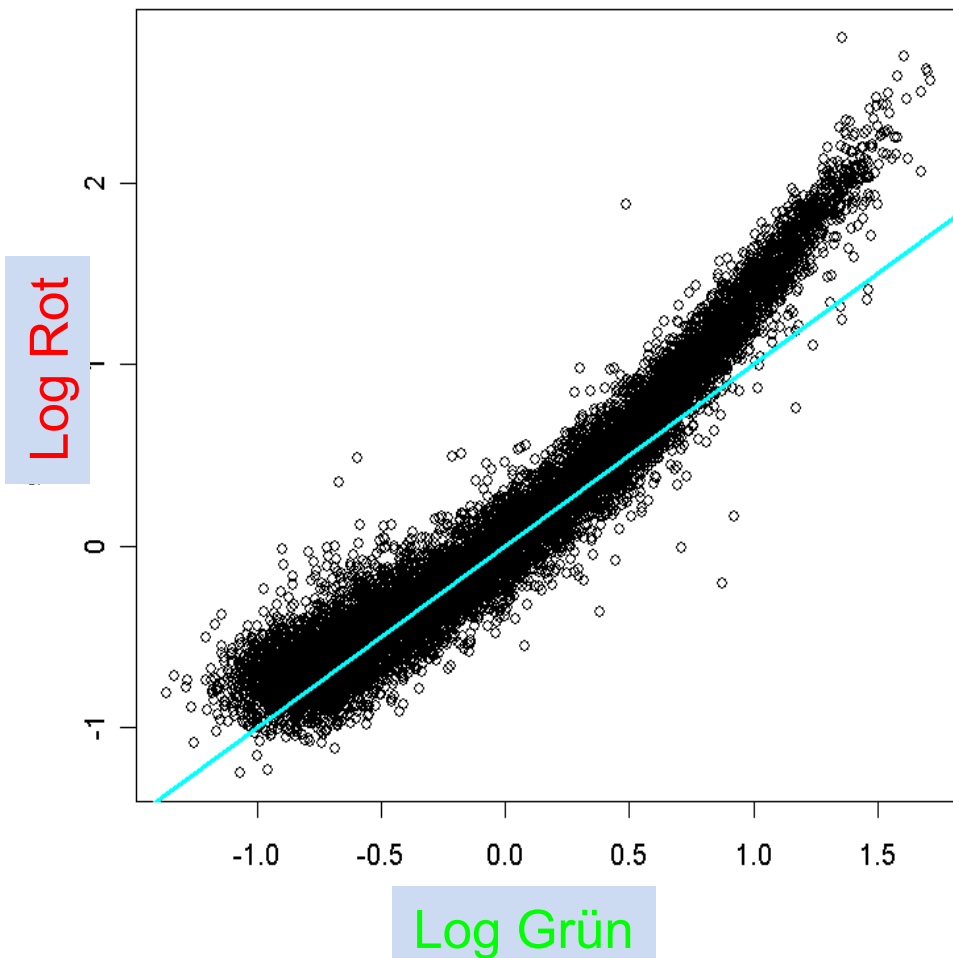
Log Signal, zentriert bei 0



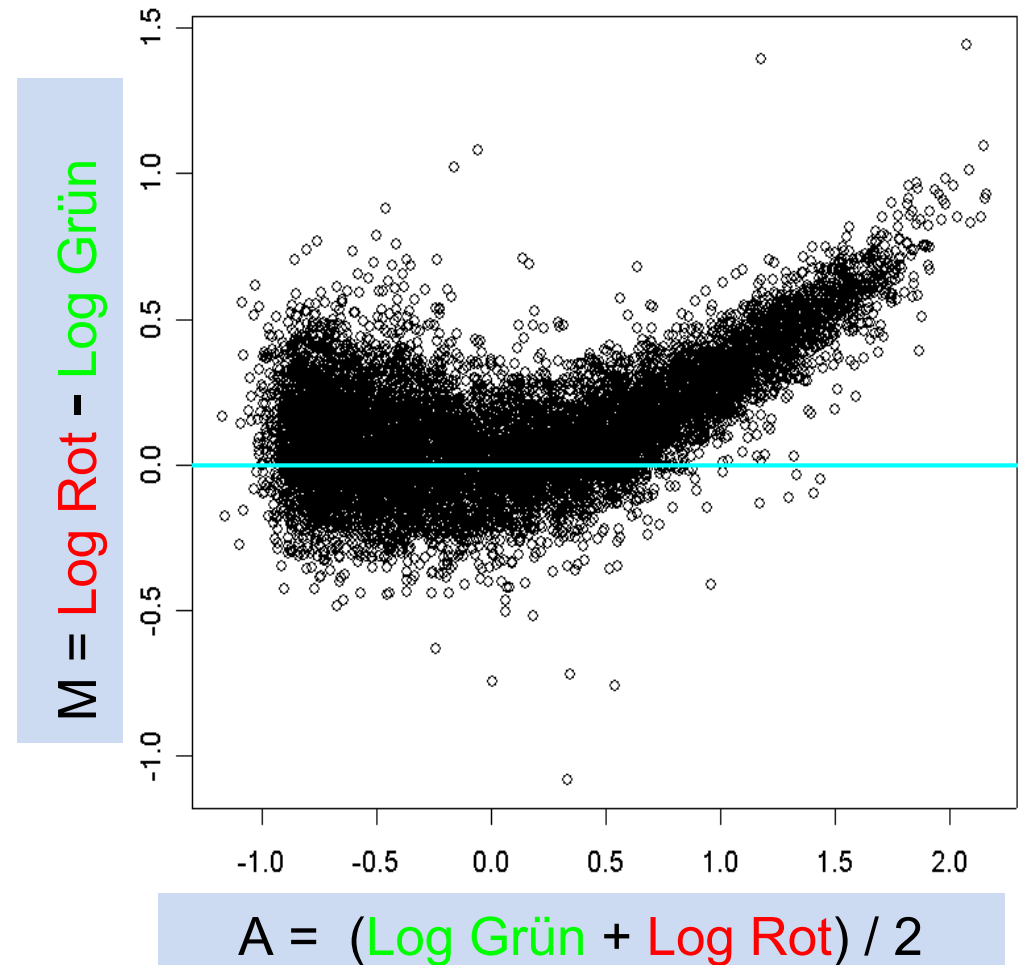
Probleme bei der Medianzentrierung

Die Medianzentrierung ist eine **globale Methode**: Es wird pro Array nur ein Parameter (der Median) aus den Daten geschätzt und zur Adjustierung benutzt. Oft gibt es jedoch intensitätsabhängige Effekte:

Scatterplot der log-Signale nach Medianzentrierung



M-A Plot der gleichen Daten



Quantilnormalisierung

Die Grundidee der Quantilnormalisierung ist bestechend einfach:

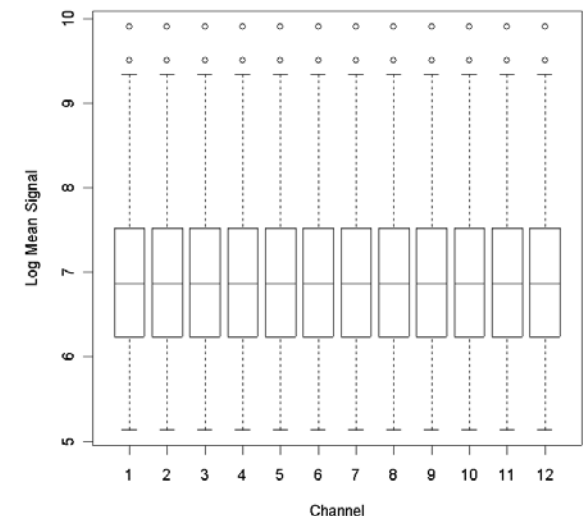
„Die Histogramme aller Microarrays sehen gleich aus“

Dies ist eine Verschärfung der Hypothese, die der Medianzentrierung zu Grunde lag. Nicht nur das 50%-Quantil=der Median soll in allen Arrays (in etwa) gleich sein, sondern *alle* Quantile.

Der Algorithmus lautet:

- Ordne die Gene eines jeden Arrays der Größe nach.
- Sei M_n der Mittelwert der Gene mit der n -t höchsten Expression. Ersetze den Messwert dieser Gene jeweils durch M_n .
- Verfahre so für alle Positionen n .

Boxplot nach
Quantilnormali-
sierung



Nachteil der Quantilnormalisierung: In einzelnen Arrays differentiell exprimierte Gene am unteren und am oberen Ende der Messskala werden nivelliert.

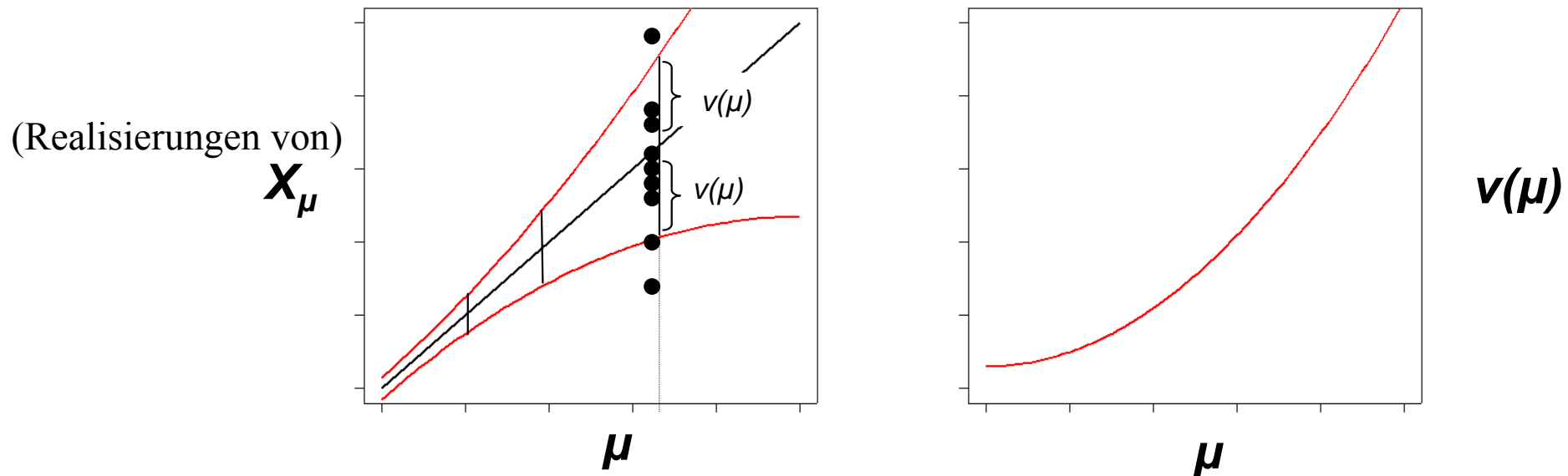
Varianzstabilisierende Transformationen

Gegeben sei eine Familie von Zufallsvariablen X_μ , $\mu \in [a,b]$, mit Erwartungswert

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

Die Varianz dieser Zufallsvariablen sei eine Funktion von μ ,

$$\text{Var}(X_\mu) = v(\mu).$$



Gesucht ist eine Transformation $T: \mathbb{R} \rightarrow \mathbb{R}$ derart, dass

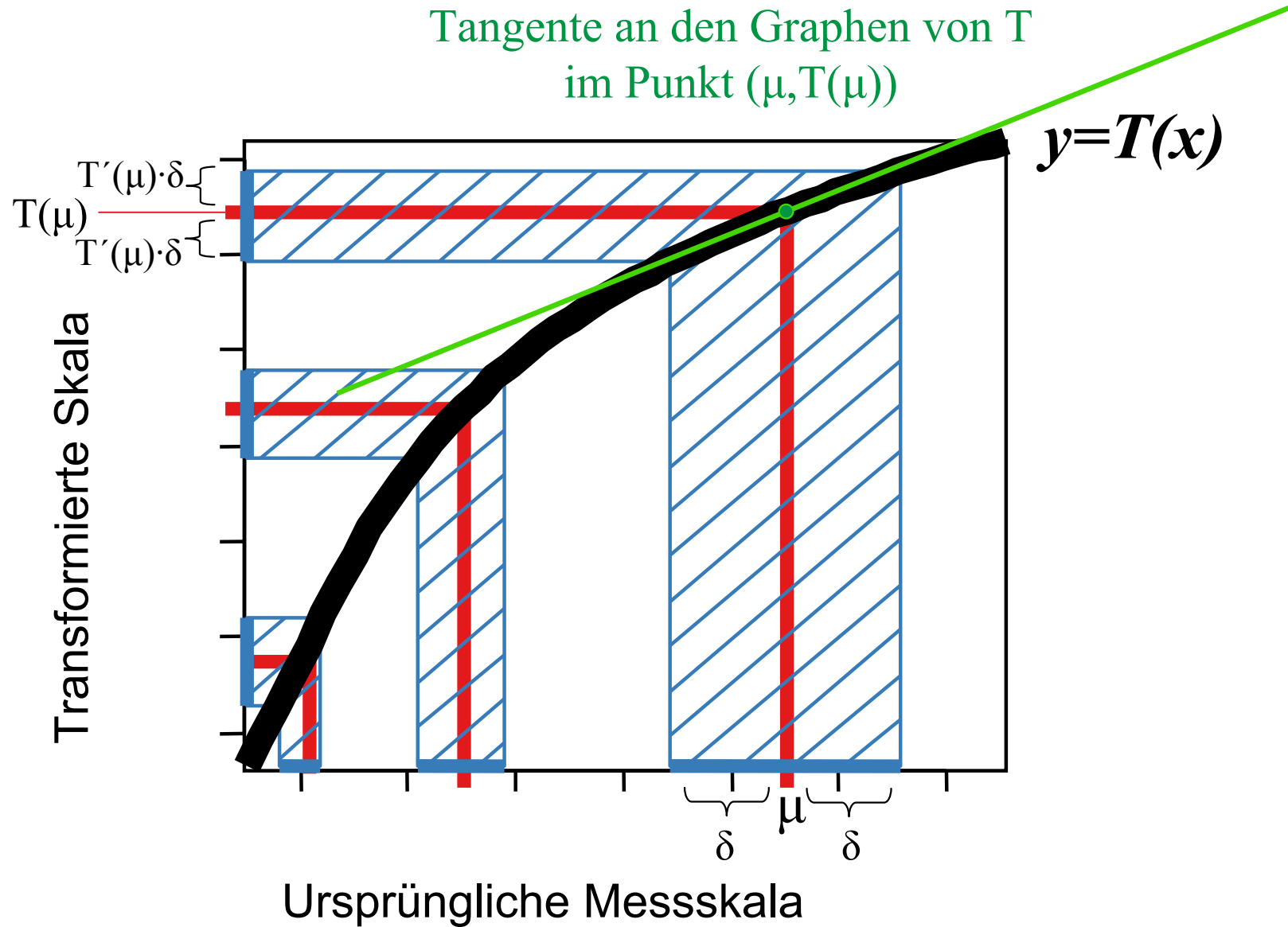
$$\text{Var}(T(X_\mu)) \approx \text{const.}$$

Wozu dient eine varianzstabilisierende Transformation?

Nach Varianzstabilisierung mit T sind die Daten **homoskedastisch**, d.h. die Varianz der betrachteten Zufallsvariablen $T(X_\mu)$, $\mu \in [a,b]$, ist (annähernd) konstant (Gegenteil: **heteroskedastisch**. Betrachtet man die Replikatmessungen der Expressionswerte eines Gens mit mittlerer Expression μ als Realisierungen einer Zufallsvariable X_μ , so sind die X_μ , $\mu \in [a,b]$, heteroskedastisch).

Homoskedastische Daten ermöglichen die Anwendung zuverlässigerer statistischer Tests. Beispielsweise sind die Voraussetzungen zur Anwendung des t-Tests auf differentielle Expression bei den transformierten Daten besser erfüllt als bei den untransformierten Daten.

Herleitung der varianzstabilisierenden Transformation



Herleitung der Varianzstabilisierenden Transformation

Erinnerung:

- Für reellwertige Zufallsvariablen X und reelle Zahlen a, b gilt

$$\text{Var}(aX+b) = a^2 \text{Var}(X)$$

- Eine differenzierbare Funktion $T: \mathbb{R} \rightarrow \mathbb{R}$ wird in der Umgebung eines Punktes μ in erster Näherung (= linear) approximiert durch

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

Ist T eine gegebene Transformation, so kann man daher in erster Näherung schreiben:

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

Somit hat man eine Näherung für die Varianz von $T(X_\mu)$:

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

Herleitung der varianzstabilisierenden Transformation

Der Rest ist reine Technik. Alles, was noch zu tun bleibt, ist, sich die Konstanz von $\text{Var}(T(X_\mu))$ zu wünschen und die entstehende „Differentialgleichung“ nach T aufzulösen

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

$$\rightarrow \text{löse } T'(\mu) = 1/\sqrt{v(\mu)}$$

$$T(\mu) = \int_0^\mu 1/\sqrt{v(t)} dt$$

(T ist bis auf eine additive Konstante eindeutig bestimmt, wir legen die Konstante durch die Wahl der unteren Integrationsgrenze willkürlich fest).

Die varianzstabilisierende Transformation T hängt nur noch von v ab und somit vom gewählten Fehlermodell.

Das Zweikomponenten-Fehlermodell

μ : „wahre“ Genexpression

X_μ : gemessene Genexpression

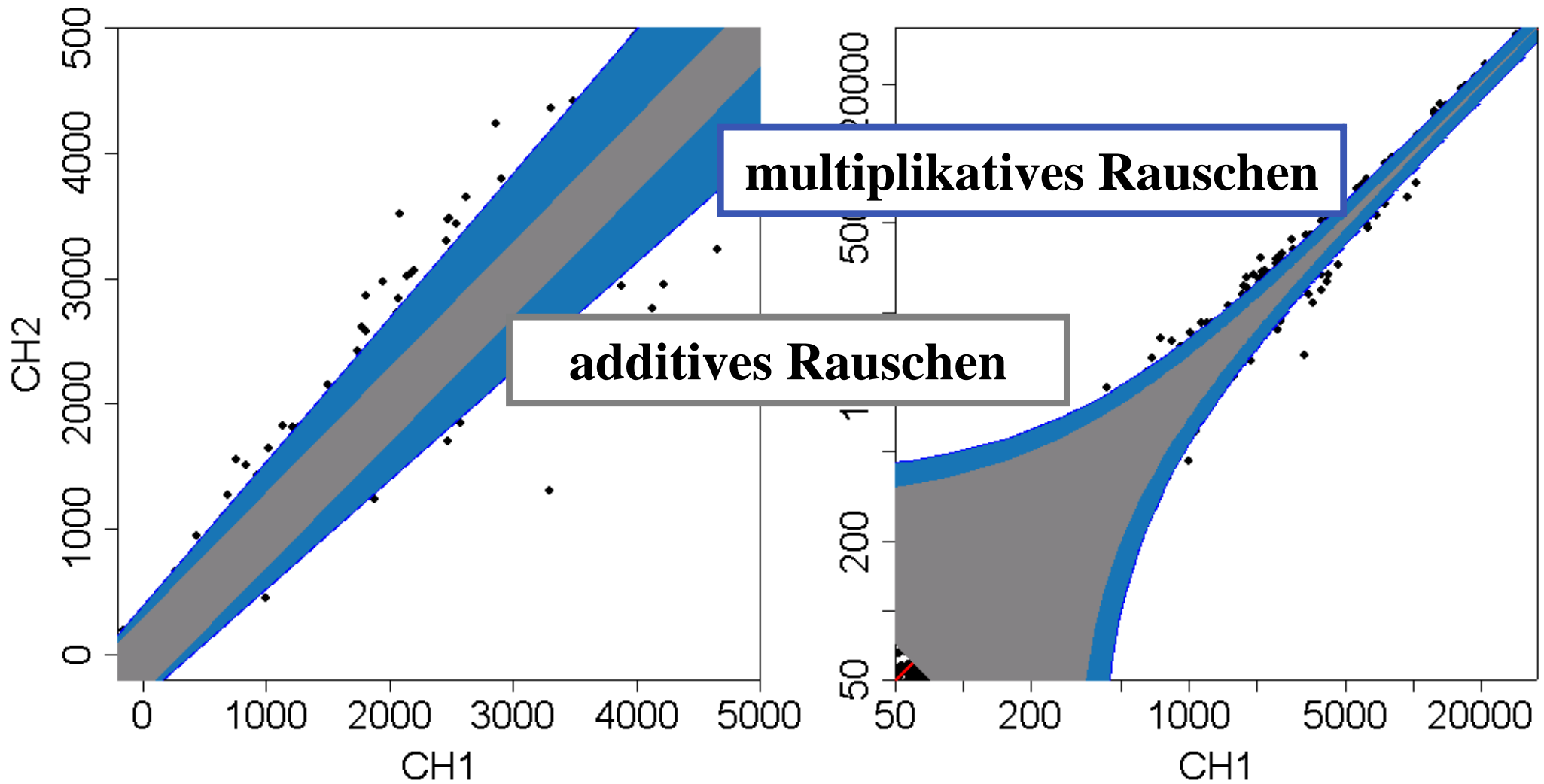
$$X_\mu = a + \varepsilon + b \cdot \mu \cdot (1 + \eta)$$

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

Für kleine η sind die beiden Varianten praktisch äquivalent

a konstanter Hintergrund	Konstant für alle Sonden eines Arrays und einer Farbe, variiert mit Array und Farbe(Cy5/Cy3)
ε zufälliger Hintergrund	iid für jeden Spot
b konstanter Verstärkungsfaktor	Konstant für alle Sonden eines Arrays und einer Farbe, variiert mit Array und Farbe(Cy5/Cy3)
η zufällige Verstärkungsschwankungen	iid für jeden Spot

Das Zweikomponenten-Fehlermodell



Originalskala

Logarithmische Skala

B. Durbin, D. Rocke, JCB 2001

Berechnung der varianzstabilisierenden Transformation für verschiedene Fehlermodelle

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

$$\varepsilon \sim N(0, \sigma^2) \quad , \quad \eta \sim N(0, \tau^2)$$

a) Kein multiplikativer Fehler ($\tau = 0$) :

$$\begin{aligned} v(\mu) &= \text{Var}(X_{\mu}) = \text{Var}(a + \varepsilon + b \cdot \mu) \\ &= \text{Var}(\varepsilon) = \sigma^2 \end{aligned}$$

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

***T* ist lediglich eine proportionale Umskalierung.**

Berechnung der varianzstabilisierenden Transformation für verschiedene Fehlermodelle

b) Kein additiver Fehler ($\sigma = 0$) :

$$\begin{aligned}v(\mu) &= \text{Var}(X_\mu) = \text{Var}(a + b \cdot \mu \cdot (1 + \eta)) \\ &= b^2 \mu^2 \text{Var}(\eta) = b^2 \mu^2 \tau^2\end{aligned}$$

$$\begin{aligned}\Rightarrow T(\mu) &= \int_1^\mu \frac{1}{\sqrt{v(t)}} dt = \int_1^\mu \frac{1}{bt\tau} dt \\ &= \frac{\log(b\tau\mu)}{b\tau} + \text{const.} = \frac{\log(\mu)}{b\tau} + \text{const}'.\end{aligned}$$

T ist im wesentlichen (= bis auf affin-lineare Transformationen) ein Logarithmus.

Berechnung der varianzstabilisierenden Transformation für verschiedene Fehlermodelle

c) Allgemeines Modell :

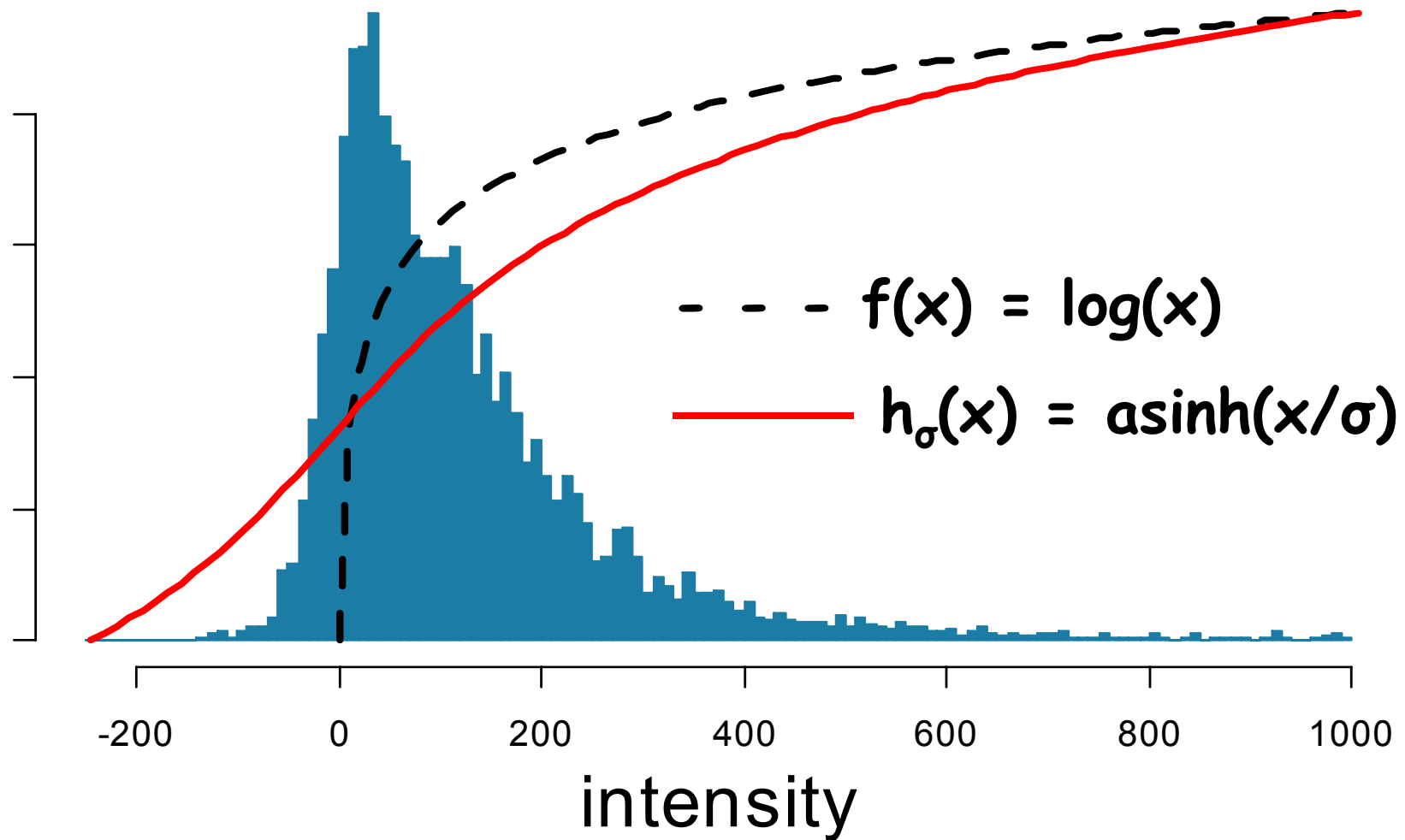
$$v(\mu) = \text{Var}(a + \varepsilon + b \cdot \mu \cdot (1 + \eta)) = \sigma^2 + b^2 \mu^2 \tau^2$$

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

$$\begin{array}{l} \text{Im wesentlichen} \\ = \end{array} \operatorname{asinh}\left(\frac{\mu}{\sigma}\right)$$

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

Die „glog“-Transformation



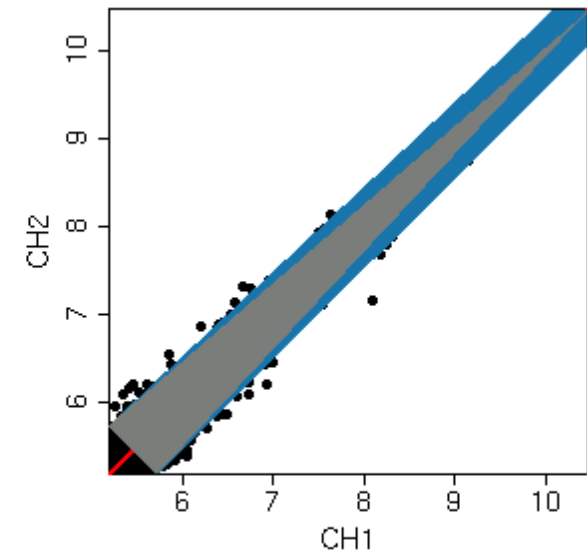
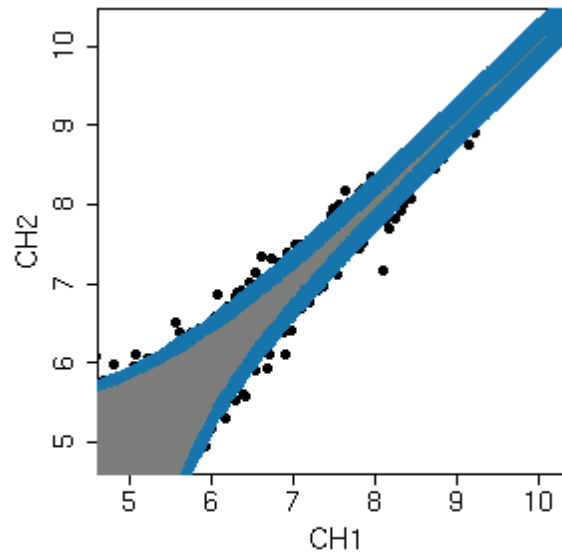
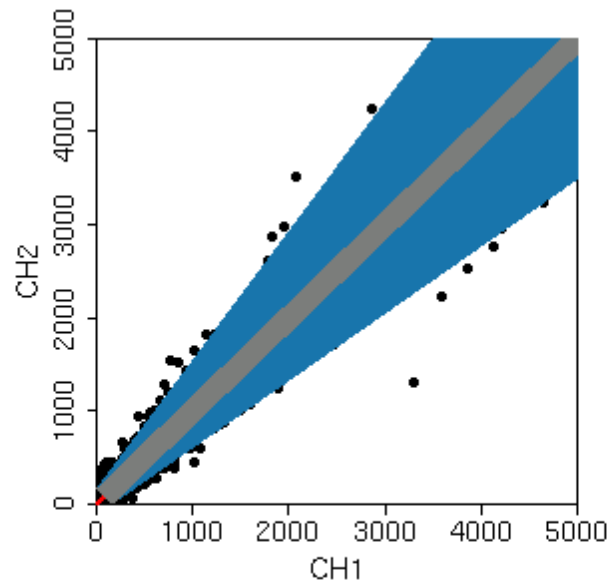
$$\lim_{x \rightarrow \infty} (\operatorname{asinh}(x) - \log(x)) \longrightarrow \log(2)$$

P. Munson, 2001

D. Rocke & B. Durbin,
ISMB 2002

W. Huber et al., ISMB
2002

Die „glog“-Transformation



Varianz:



Additiver Anteil

Multiplikativer Anteil

P. Munson, 2001

D. Rocke & B. Durbin,
ISMB 2002

W. Huber et al., ISMB
2002

Schätzung der Parameter des Fehlermodells

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”

Fold change Schätzung: Bias-Variance tradeoff

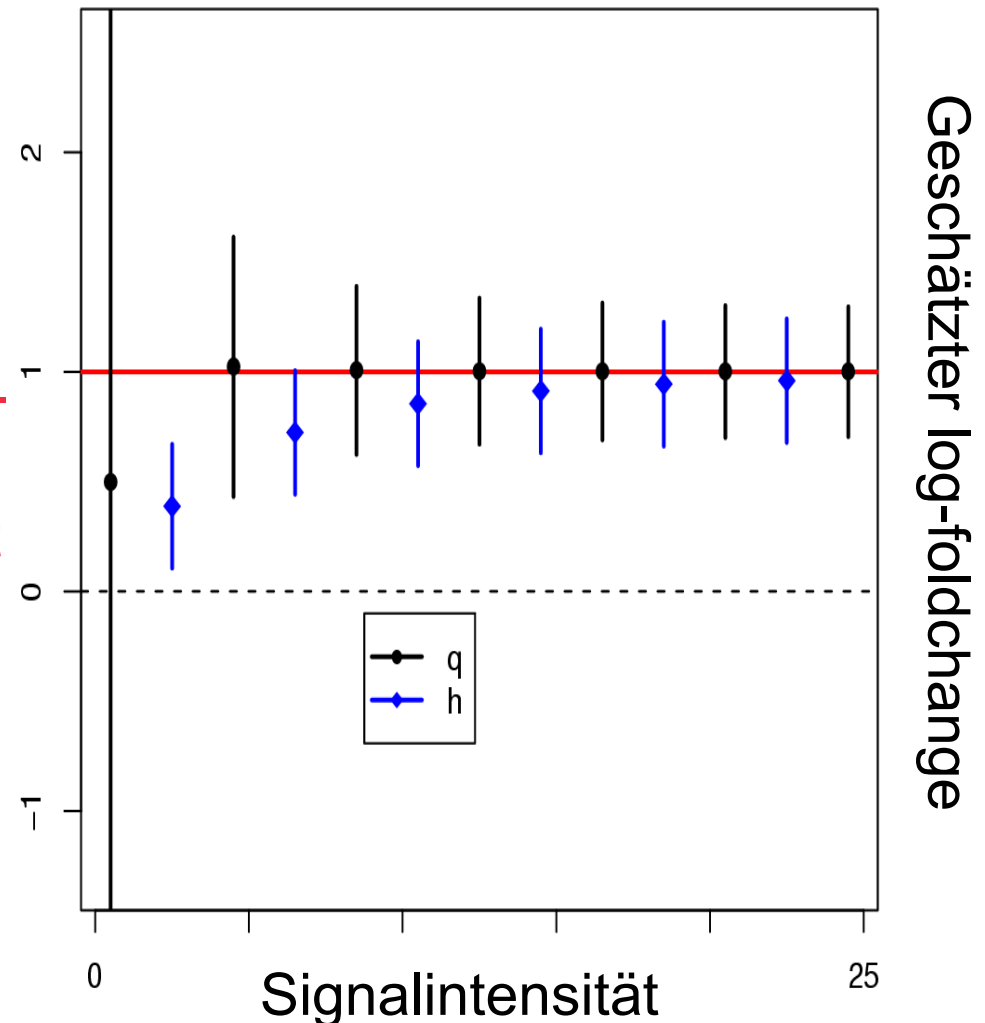
An die Stelle der log-ratio $q = \log \frac{x_1}{x_x}$ tritt die „glog“-ratio

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

(c_1, c_2 experimentspezifische Parameter)

Die glog-ratio ist ein sog. **Shrinkage-Schätzer**: Im Tausch gegen einen erhöhten Bias Richtung Null erreicht man eine **Verkleinerung der Varianz des Schätzers**.

Ein solcher Schätzer ist besonders nützlich, wenn aufgrund weniger Replikate mit einer hohen Varianz gerechnet werden muss.



Danksagung

Wolfgang Huber (etliche Folien), David M. Rocke, Blythe Durbin

Bolstad, et al. (2003, *Bioinformatics* 19 2:185-193) propose *quantile normalization* for microarray data

Literatur, Links

- **Bioconductor vignette for vsn.** W. Huber <http://www.maths.lth.se/help/R/.R/library/vsn/doc/vsn.pdf>
- **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** Bolstad, B., et al. *Bioinformatics* 19 2:185-193 (2003)
- **A model for measurement error for gene expression analysis.** D. Rocke, B. Durbin. *Journal of Computational Biology*, 8:557-569, 2001
- **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, M. Vingron. *Bioinformatics* 18 suppl. 1 (2002), S96-S104 (ISMB 2002).
- **Parameter estimation for the calibration and variance stabilization of microarray data.** W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, M. Vingron. *Statistical Applications in Genetics and Molecular Biology* 2003 Vol. 2: No. 1, Article 3
- **Error models for microarray intensities.** W. Huber, A. von Heydebreck, and M. Vingron. to appear in: *Encyclopedia of Genomics, Proteomics and Bioinformatics*. John Wiley & sons (2004).
- **Interpretability and Data Transformations for Gene Expression Microarray Data.** D. M. Rocke, W. Huber, B. Durbin, A. von Heydebreck, M. Vingron. submitted (2004).
- **Statistical methods for identifying differentially expressed genes in microarray experiments.** S. Dudoit, Y.H. Yang, T.P. Speed, and M.J. Callow. *Statistica Sinica*, 12:111-139, 2002.