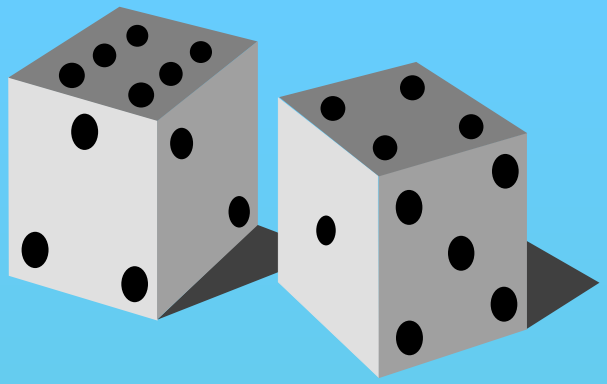


Statistische Analyse von Genexpressionsdaten

Dr. A. Yassouridis

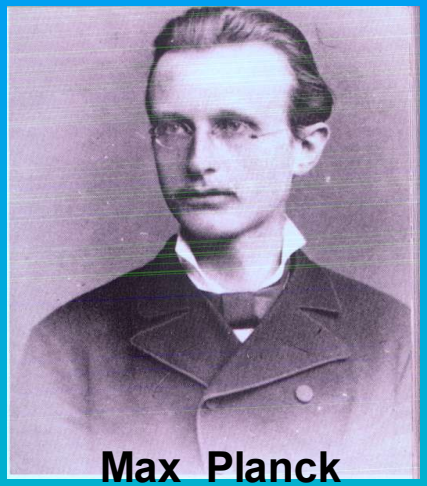
Max-Planck-Institut für Psychiatrie

AG: Statistik

Die exakten Wissenschaften

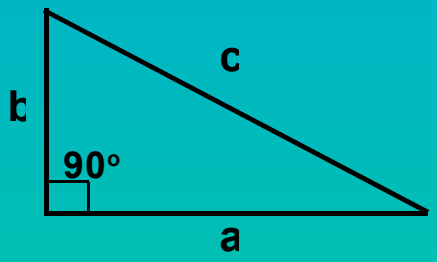
Die schönste Forschungsaufgabe in einer Wissenschaft ist, eine Gesetzmäßigkeit zu finden, die bis in alle Ewigkeit Gültigkeit hat, ... die Suche m.a.W. nach etwas Absolutem, Unvergänglichem, Allgemeingültigem.....

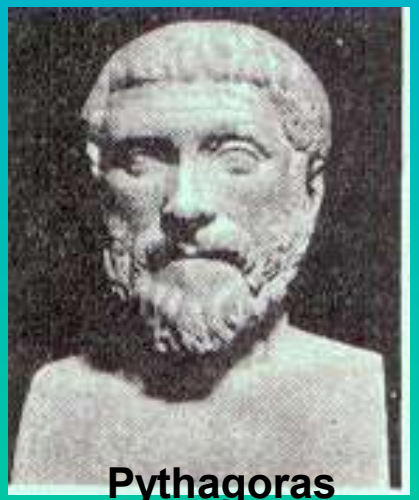
Strahlungsgesetz: $\varepsilon = h.\nu$



Max Planck
1858 -1947

Satz von Pythagoras:
$a^2+b^2=c^2$





Pythagoras
570 - 500 v. Chr.

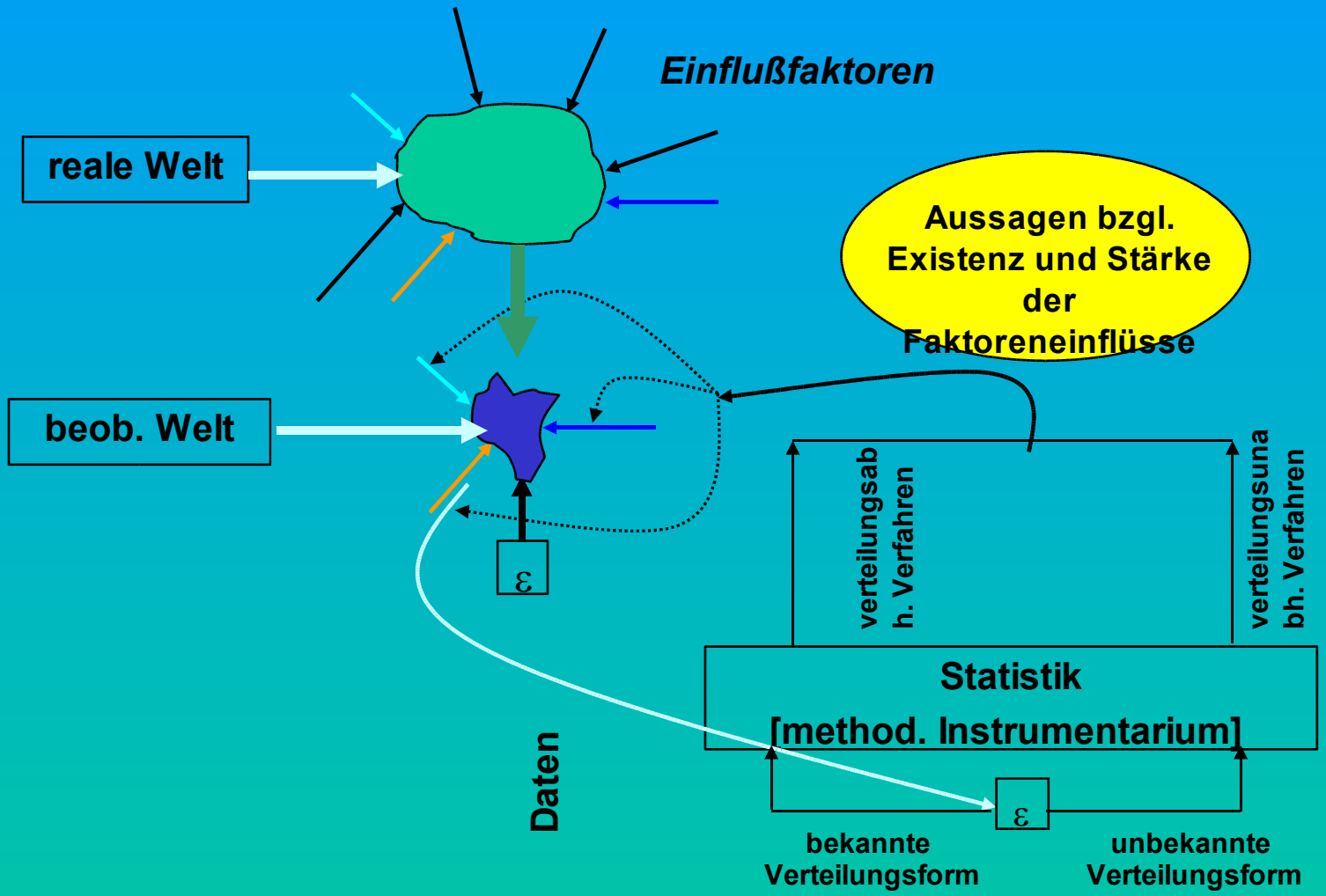
Die empirischen Wissenschaften

Die Phänomene hinter den empirischen Wissenschaften beinhalten i.d.R. das Verhalten biologischer Wesen und zeichnen sich somit durch eine hohe Komplexität und eine starke Veränderungsplastizität aus. Demzufolge können sie nicht exakt erklärt und prognostiziert werden

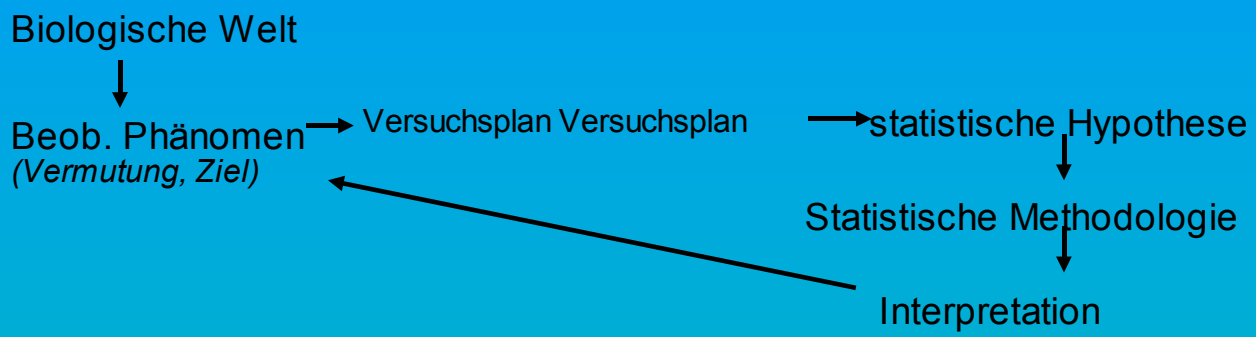
Die Komplexität und Veränderungsplastizität der zu erklärenden Phänomene bei den empirischen Wissenschaften ist vorwiegend auf die große Anzahl der internen und externen Einflußfaktoren zurückzuführen, die das Verhalten biologischer Wesen, insbesondere intelligenter Wesen, von Zeit zu Zeit und von Ort zu Ort unterschiedlich zu beeinflussen vermögen

Fazit: **Die empirischen Wissenschaften, wie z. B. Psychiatrie, Biologie, Wirtschaft u.a. können keine absoluten, universellen und in aller Zeiten geltenden Gesetzmäßigkeiten vorweisen**

Die Bedeutung der Statistik für die empirischen Wissenschaften



Die zwei wichtigen Prinzipien bei der Versuchsplanung



Postulat 1

Die Zielsetzung hinter einem Experiment lässt sich nur durch die Erstellung des richtigen Versuchsplanes erreichen

Postulat 2

Bei jedem Versuchsplan soll darauf geachtet werden, dass die zwei entgegengesetzten Prinzipien (Prinzip der *Vergleichbarkeit* und Prinzip der *Verallgemeinerungsfähigkeit*) aufeinander abgestimmt sind.

Die zwei wichtigen Prinzipien bei der Versuchsplanung

Prinzipien der Versuchsplanung

Prinzip der Vergleichbarkeit —erfordert→ homogenes Material (impliziert geringe biol. Varianz)

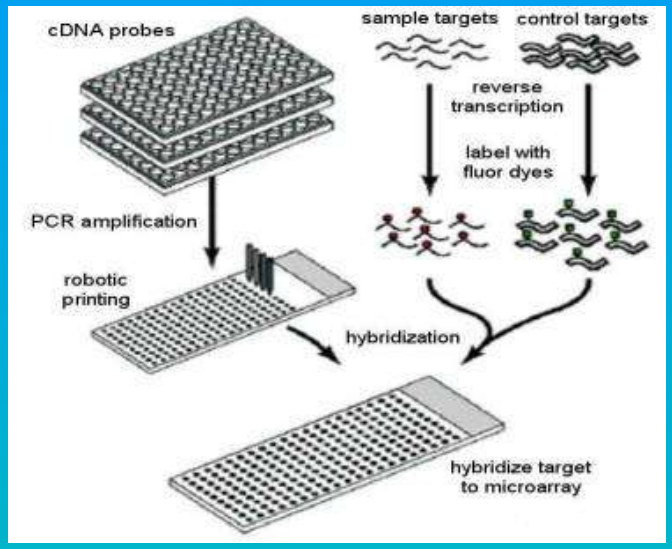
Prinzip der Verallgemeinerungsfähigkeit —erfordert→ heterogenes Material (impliziert große biol. Varianz)

Statistische Analyse von Genexpressionsdaten

biologisches Phänomen

<u>Die wesentlichen Schritte der Genexpressionsanalyse</u>

- Erstellung der Microarrays (Glasträger) mit den Sonden
- Auswahl der zu untersuchenden Zellpopulationen
- Extraktion der mRNA und reverse Transkription
- Markierung
- Hybridisierung
- Scannen der Microarrays
- Datenanalyse der Fluoreszenzintensitäten



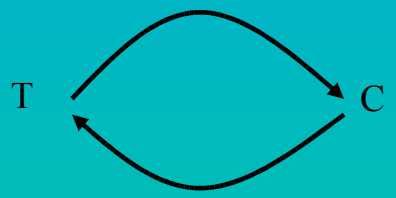
biologische Ziele

- *Identifizierung regulierender (exprimierender) Gene [Datennormalisierung und komparative Analysen]*
- *Suche nach funktionellen Zusammenhängen unter den Genen*
- *Suche nach biologischen Markern (Gene), die bestimmte Verhaltensweisen bzw. Krankheiten zu charakterisieren vermögen*

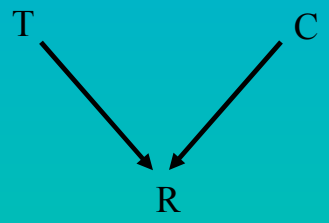
Statistische Analyse von Genexpressionsdaten

Genidentifizierung über Datennormalisierung und komparativen Analysen

Die 2-Farben-Hybridisierung und die nachfolgende Normalisierung der Fluoreszenz-intensitäten bei der Genexpression ermöglichen die Genidentifizierung sowohl direkt *(direkte Vergleiche oder Designs)* als auch indirekt *(indirekte Vergleiche oder Designs).*



direkte Vergleiche
(gleiche Hybridisierung)



indirekte Vergleiche
(unters. Hybridisierung

Statistische Analyse von Genexpressionsdaten

Bei direkten Vergleichen (gleiche Hybridisierung) wird das Zwei-Stichproben-Problem in ein Ein-Stichproben-Problem transformiert.

Direkte Vergleiche sind empfehlenswert, wenn das Vergleichbarkeitsprinzip im Vordergrund des Interesses liegt (verbundene Stichproben; biol. replicates type I).

Indirekte Vergleiche über unterschiedliche (referenzbezogene) Hybridisierungen empfehlen sich dort, wo die Verallgemeinerung der Ergebnisse im Vordergrund des Interesses steht (unverbundene Stichproben; biol. replicates type II).

Die Transformation des Zwei- in Ein-Stichproben-Problem

X_A und X_B sind FI eines bestimmten Gens in zwei Hirnregionen A und B.

direkter Vergleich [gleiche Hybridisierung: A (rot) vs B (grün)]

b) $Y=\log_2(X_A/X_B)$;

c) Datennormalisierung $\longrightarrow$ $Y \sim N(\mu, \sigma)$

c) $H_0: \mu_A = \mu_B$ gegen $\mu_A \neq \mu_B$ $\longrightarrow$ $H_0: \mu_Y = 0$ gegen $\mu_Y \neq 0$

$\downarrow$

Ein-Stichproben t-Test $t := \dfrac{|\bar{y}|}{\frac{s}{\sqrt{n}}}$ $\longrightarrow$ *t*-verteilt mit (n-1) FG

Statistische Analyse von Genexpressionsdaten

Das Zwei-Stichproben-Problem

X_A und X_B sind FI eines bestimmten Gens bei einer mutanten bzw. Kontrollmaus.

indirekter Vergleich [zwei untersch. Hybridisierungen: i) A (rot) vs refer. mRNA R (grün); ii) B (rot) vs R (grün)]

e) $Y=\log_2(X_A/X_R) - \log_2(X_B/X_R) = Y_1 - Y_2$;

f) Datennormalisierung $\longrightarrow$ $\mathbf{Y \sim N(\mu, \sigma)}$

g) $H_0: \mu_A = \mu_B$ gegen $\mu_A \neq \mu_B$

$\downarrow$

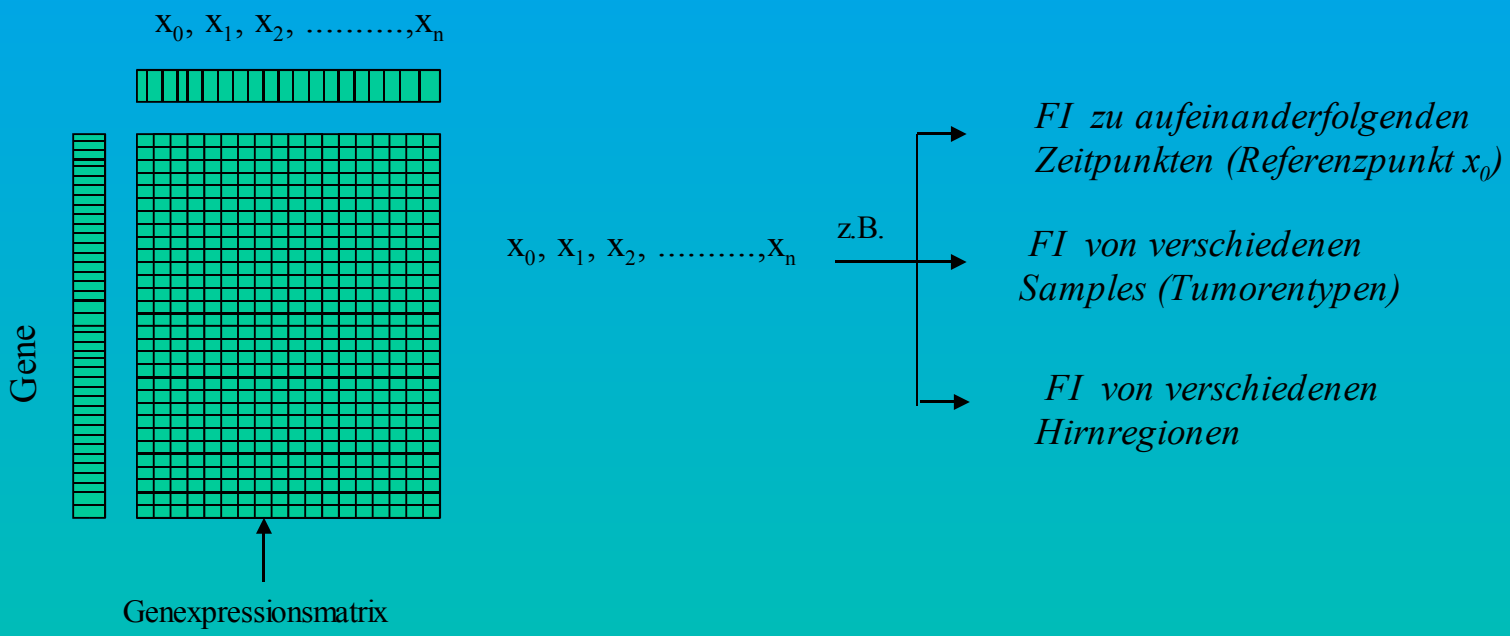
Zwei-Stichproben t-Test $t := \dfrac{(|\bar{y}1 - \bar{y}2|)}{\frac{s}{\sqrt{n}}}$ $\longrightarrow$ t-verteilt mit (n-1) FG

Schlussfolgerung

Nimmt t (nicht absolut gesehen) große positive (bzw. große negative) Werte, weist es auf signifikant große Exprimierung (bzw. Reprimierung) von A im Vergleich zu B hin.

Statistische Analyse von Genexpressionsdaten

Suche nach funktionellen Zusammenhängen unter den Genen

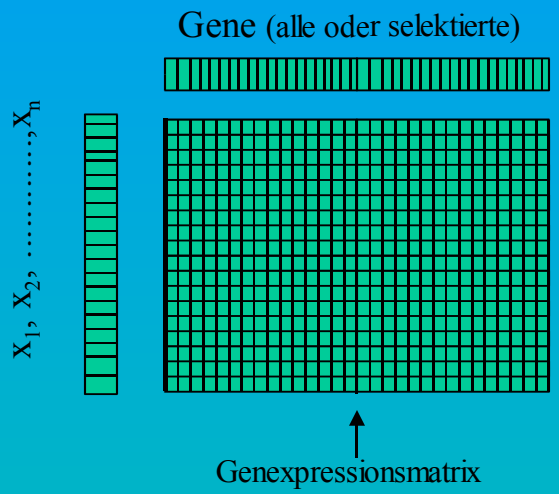


Zielsetzung: Objekte (hier Gene) mit ähnlichen Expressionseigenschaften in Gruppen zusammenzufassen, die in sich ähnlich aber untereinander unähnlich sind

(Clusteranalysis, unsupervised analysis)

Statistische Analyse von Genexpressionsdaten

Suche nach biologischen Markern



unterschiedliche Einheiten oder Individuals

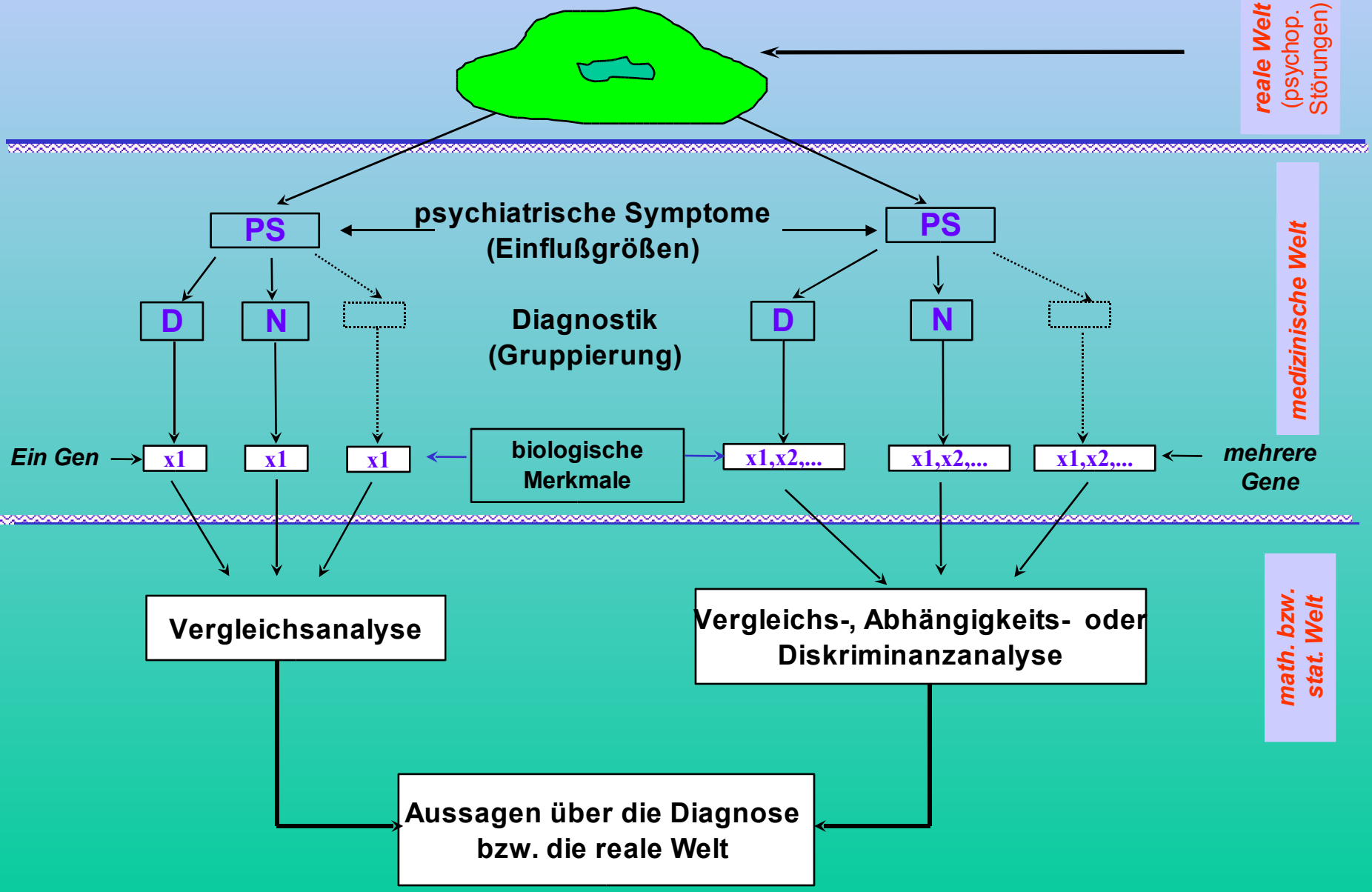
sie können unterschiedlichen Gruppen angehören

sie können alle oder teilweise unterschiedlichen Bedingungen unterliegen (Faktoreneinflüsse)

Mögliche Fragestellungen bzw. Zielsetzungen

- welche Gene charakterisieren die einzelnen Gruppen bzw. tragen zur Gruppeneinteilung und Unterscheidung bei? *(Klassifikationsanalysen)*
- wie kann man auf der Basis der gewonnenen Erkenntnisse Voraussagen über die Gruppenzugehörigkeit einer neuen Einheit erzielen? *(Prognose mittels supervised Analysen)*
- auf welche Gene sind Faktoreneinflüsse zu verzeichnen und wo sind sie am stärksten? *(MANOVA auf faktorielle Designs; multiple Vergleiche)*

DER KLASSISCHE WEG: von der Psychiatrie zur Biologie über die Gene



DER KLASSISCHE WEG: von der Psychiatrie zur Biologie

Typische Fragestellungen

Ein Gen

Unterscheiden sich die Gruppen voneinander bzgl. des beobachteten Gens?

Beeinflussen die PS das Gen und wie?

Läßt sich die Variabilität des Gens durch die PS bzw. durch die Diagnose erklären?

Welcher Zusammenhang besteht zwischen Gen und Diagnose?

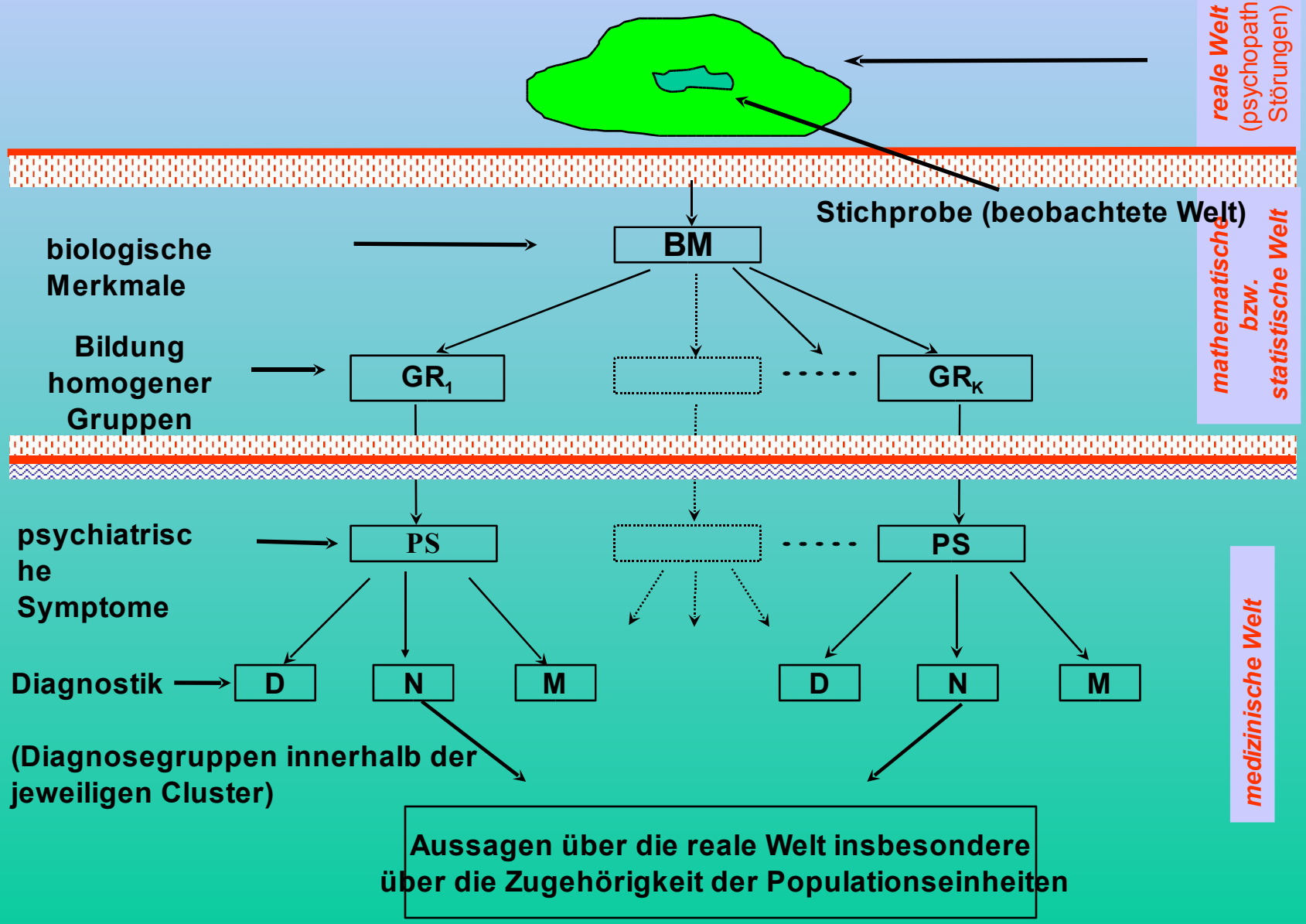
Mehrere Gene

Bei welchen Genen - wenn überhaupt - läßt sich ein Einfluß von den PS oder der Diagnose feststellen?

Welche Assoziationsstruktur besteht zwi-schen PS bzw. Diagnose und Genen?

Welche der Gene tragen am stärksten zur Differenzierung innerhalb der Diagnosen-gruppen bei?

DER ALTERNATIVE WEG: von der Biologie der Gene zur Psychiatrie



DER ALTERNATIVE WEG: von der Biologie der Gene zur Psychiatrie

Typische Fragestellungen

Welche und wie viele unterschiedliche biologische Muster (Gengruppen) gibt es und wie verteilen sich die Diagnosegruppen auf diese Muster?

Haben die verschiedenen biologischen Muster Vorhersagewert z.B. für den Krankheitsverlauf, die Therapieansprechbarkeit und die Rückfall-Wahrscheinlichkeit?

Erhofftes Resultat

In manchen der nach den Genen (BM) gebildeten Gruppen dominiert die Symptomatik der einen oder anderen Diagnosegruppe

Folgerung

Die Populationseinheiten (z.B. Patienten) werden nach Feststellung ihrer Zugehörigkeit zu einer durch die BM definierten Gruppe zusätzlich mit Wahrscheinlichkeitsangaben den jeweiligen (psychiatrischen) Diagnosengruppen zugeordnet.

Besteht eine besonders enge Beziehung (Sonderfall: Deckungsgleichheit) zwischen einem nach BM ermittelten Cluster und einer diagnostischen Gruppe, dann können diese BM als *biologische Tests* ("biological marker") für die diesbezügliche Diagnose angesehen werden

Statistische Analyse von Genexpressionsdaten

Die Tücken der multiplen Vergleiche

- Kontrastbezogener Fehler (α_K)
- Familienbezogener Fehler (α_F)
- Experimentbezogener Fehler (α_E)

$(\alpha_F) \leq (\alpha_F) \leq (\alpha_F)$

Schwaches bzw. strenges Kontrollieren des Fehlers 1. Art

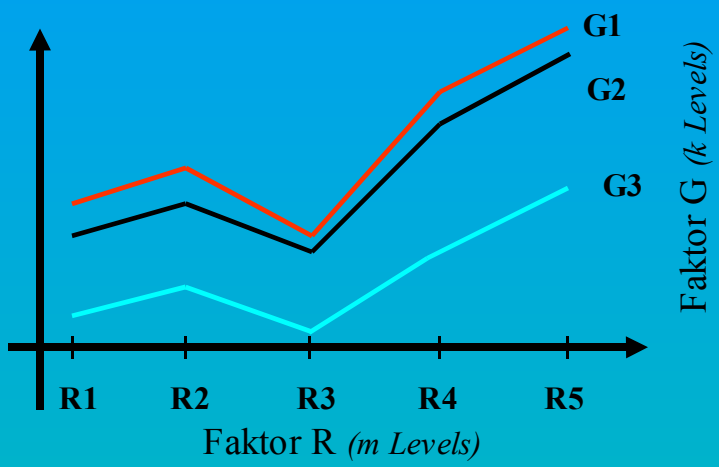
- Dunn‘s oder Bonferroni Korrektur
- Dunn-Sidak Procedur
- Holm‘s sequentielle Test-Prozedur
- Benjamini/Hochberg
- .
- .

Postulat

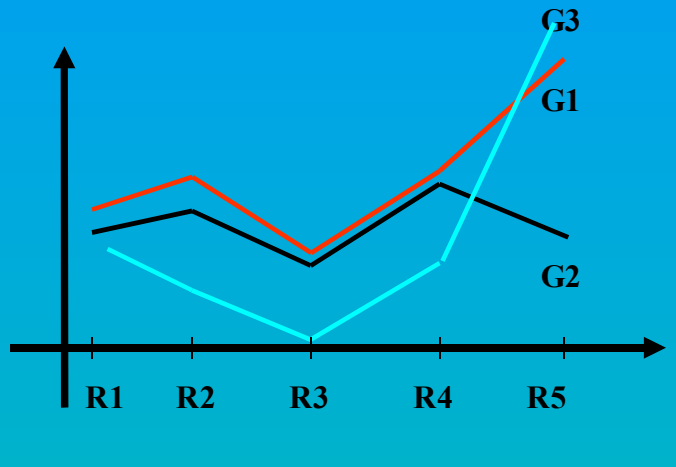
Der Einfluss von mehreren Faktoren oder von einem mehrstufigen Faktor auf Genexpressionsdaten ist inferentiell-statistisch nur dann möglich, wenn man sich nur auf eine sehr geringe Anzahl von Genen (Kandidatengene) einschränkt.

Statistische Analyse von Genexpressionsdaten

Die Tücken der multiplen Vergleiche



Interaktion R x G nicht signifikant



Interaktion R x G signifikant

Anzahl familienbez. Kontraste

für k=3, m=5

Faktor G $\longrightarrow \binom{k}{2}$ =3

Faktor R $\longrightarrow \binom{m}{2}$ =10

Anzahl experimentalbez. Kontraste

$r\binom{k}{2} + k\binom{m}{2}$ =15 + 30=45

Fazit: Bei sign. Interaktionen und gleichzeitiger Betrachtung von mehreren Genen kann die Anzahl der Kontraste schnell ins Übermäßige wachsen

Statistische Analyse von Genexpressionsdaten

Ein Beispiel:

Sachproblem

Bei den Genexpressionsanalysen unter zwei Gruppen von Tieren (mutants vc controls), die je einer unterschiedlichen Zell-Linie angehören, haben sich 5 Kanditatengene für unterschiedliche Expression herauskristallisiert.

Fragestellungen

- Unterscheiden sich die zwei Tiertypen jeder Zell-Linie in dem Expressionslevel dieser Gene?
- Falls ja, bei welchem der Gene sind die festgestellten Unterschiede am prägnantesten?
- Ist das Profil der Gruppenunterschiede in den Genen zwischen den zwei Zell-Linien ähnlich?

Statistische Analyse von Genexpressionsdaten

Gewonnene Daten aus der Expressionanalyse

				Quotienten der FI bzgl. einer Referenz-mRNA				
tier_nr	group	cel line	genotyp	Gen A	Gen B	Gen C	Gen D	Gen E
1	1	1	+/+	0,683	0,055	2,677	0,367	0,300
2	1	1	+/+	0,708	0,050	2,962	0,310	0,273
3	1	1	+/+	1,024	0,034	2,697	0,326	0,212
4	1	1	+/+	0,963	0,046	3,096	0,335	0,210
5	1	1	+/+	0,667	0,040	3,439	0,272	0,219
6	1	1	+/+	0,697	0,031	3,273	0,310	0,225
7	1	1	+/+	0,782	0,030	2,462	0,344	0,184
8	2	1	-/-	1,143	0,048	2,021	0,205	0,185
9	2	1	-/-	1,110	0,037	2,198	0,207	0,183
10	2	1	-/-	1,021	0,038	1,854	0,241	0,157
11	2	1	-/-	0,901	0,051	2,472	0,221	0,172
12	2	1	-/-	0,636	0,046	2,534	0,182	0,196
13	2	1	-/-	1,010	0,034	1,588	0,259	0,182
14	2	1	-/-	1,515	0,036	1,444	0,190	0,165
15	1	2	+/+	0,601	0,034	0,960	0,559	0,261
16	1	2	+/+	1,029	0,033	0,925	0,522	0,264
17	1	2	+/+	2,868	0,044	1,022	0,672	0,287
18	1	2	+/+	0,882	0,038	1,238	0,611	0,322
19	1	2	+/+	0,543	0,049	1,727	0,492	0,354
20	1	2	+/+	1,054	0,039	1,140	0,515	0,270
21	1	2	+/+	0,955	0,036	1,306	0,491	0,326
22	2	2	-/-	0,714	0,044	1,425	0,653	0,394
23	2	2	-/-	0,710	0,060	1,715	0,663	0,411
24	2	2	-/-	0,616	0,066	1,230	0,675	0,435
25	2	2	-/-	0,457	0,063	2,278	0,731	0,418
26	2	2	-/-	0,402	0,054	2,449	0,686	0,353
27	2	2	-/-	0,401	0,052	2,401	0,709	0,364
28	2	2	-/-	0,390	0,056	2,169	0,759	0,376

Statistische Analyse von Genexpressionsdaten

A] Deskriptive Statistik

Table: Means±SEMs of the quotients of the fluorescence intensity in the two groups (for each cel line)

	cel lines					
	1,00			2,00		
	Mean	SEM	N	Mean	SEM	N
GROUP						
+/+						
Gen A	,789	,055	7	1,133	,299	7
Gen B	,041	,004	7	,039	,002	7
Gen C	2,944	,133	7	1,188	,104	7
Gen D	,324	,011	7	,552	,026	7
Gen E	,232	,015	7	,298	,014	7
-/-						
Gen A	1,048	,101	7	,527	,056	7
Gen B	,041	,003	7	,057	,003	7
Gen C	2,016	,158	7	1,952	,186	7
Gen D	,215	,010	7	,697	,014	7
Gen E	,177	,005	7	,393	,011	7

Statistische Analyse von Genexpressionsdaten

B] Inferentielle Statistik

```
MANOVA gen_a gen_b gen_c gen_d gen_e  BY group(1 2)
  /PRINT SIGNIF(MULT UNIV)/NOPRINT PARAM(ESTIM)
  /METHOD=UNIQUE/ERROR WITHIN+RESIDUAL/DESIGN.
```

CEL LINE: 1,00

14 cases accepted.

EFFECT .. GROUP

Multivariate Tests of Significance (S = 1, M = 1 1/2, N = 3)

Test Name	Value	Exact F	Hypoth. DF	Error DF	Sig. of F
Wilks	,03743	41,14208	5,00	8,00	,000

Univariate F-tests with (1;12) D. F.

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	F	Sig. of F
GEN A	,23441	,55125	,23441	,04594	5,10283	,043
GEN B	,00000	,00086	,00000	,00007	,01247	,913
GEN C	3,01343	1,78683	3,01343	,14890	20,23767	,001
GEN D	,04118	,01007	,04118	,00084	49,07043	,000
GEN E	,01061	,01077	,01061	,00090	11,81598	,005

Statistische Analyse von Genexpressionsdaten

CEL LINE: 2,00

14 cases accepted.

EFFECT .. GROUP
Multivariate Tests of Significance (S = 1, M = 1 1/2, N = 3)

Test Name	Value	Exact F	Hypoth. DF	Error DF	Sig. of F
Wilks	,10181	14,11490	5,00	8,00	,001

Univariate F-tests with (1;12) D. F.

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	F	Sig. of F
GEN A	1,28501	3,88614	1,28501	,32385	3,96798	,070
GEN B	,00109	,00053	,00109	,00004	24,47970	,000
GEN C	2,04322	1,91537	2,04322	,15961	12,80097	,004
GEN D	,07348	,03627	,07348	,00302	24,31187	,000
GEN E	,03177	,01329	,03177	,00111	28,69152	,000

Statistische Analyse von Genexpressionsdaten

Resümee

In beiden Zell-Linien weisen die mutanten Mäuse im Vergleich zu den Kontrolltieren signifikante Unterschiede in den Expressionen der untersuchten Gene auf [Wilks multivariate tests of significance; effect of group for Timpl: $F(5,8)=41.14$, sig of $F<0.0001$; effect of group for Stenzl: $F(5,8)=14.11$, sig of $F=0.001$]. Zu den Gruppenunterschieden bei Zell-Linie 1 tragen am stärksten die Gene C, D und E bei, während bei Zell-Linie 2 alle 5 Gene abgesehen vom Gen A signifikant dazu beitragen (univariate F-Tests, $p<0.05$; *Bonferroni korrigiert*).

Wenn man die zwei Zell-Linien unter die Lupe nimmt, stellt man fest, dass bei beiden Linien die Expressionen der drei Gene C, D und E bei den Mutanten signifikant unterschiedlich als bei den Kontrollen sind; allerdings weisen diese Unterschiede gegensinnige Richtung in der zwei Zell-Linien auf.