MICROARRAYS

- Was sind Microarrays?
- Welche Technologieplattformen gibt es?
- Beispiel: Rot-Grün Chip
    - Wie wird ein Chip hergestellt (Film) ?
- Welche Fragen kann man mit Chips beantworten ?
- Datenfluß:
    - Experiment-Design
    - Image Processing
    - Preprocessing
    - Normalisierung
    - Analyse
- Biologische Verfikation

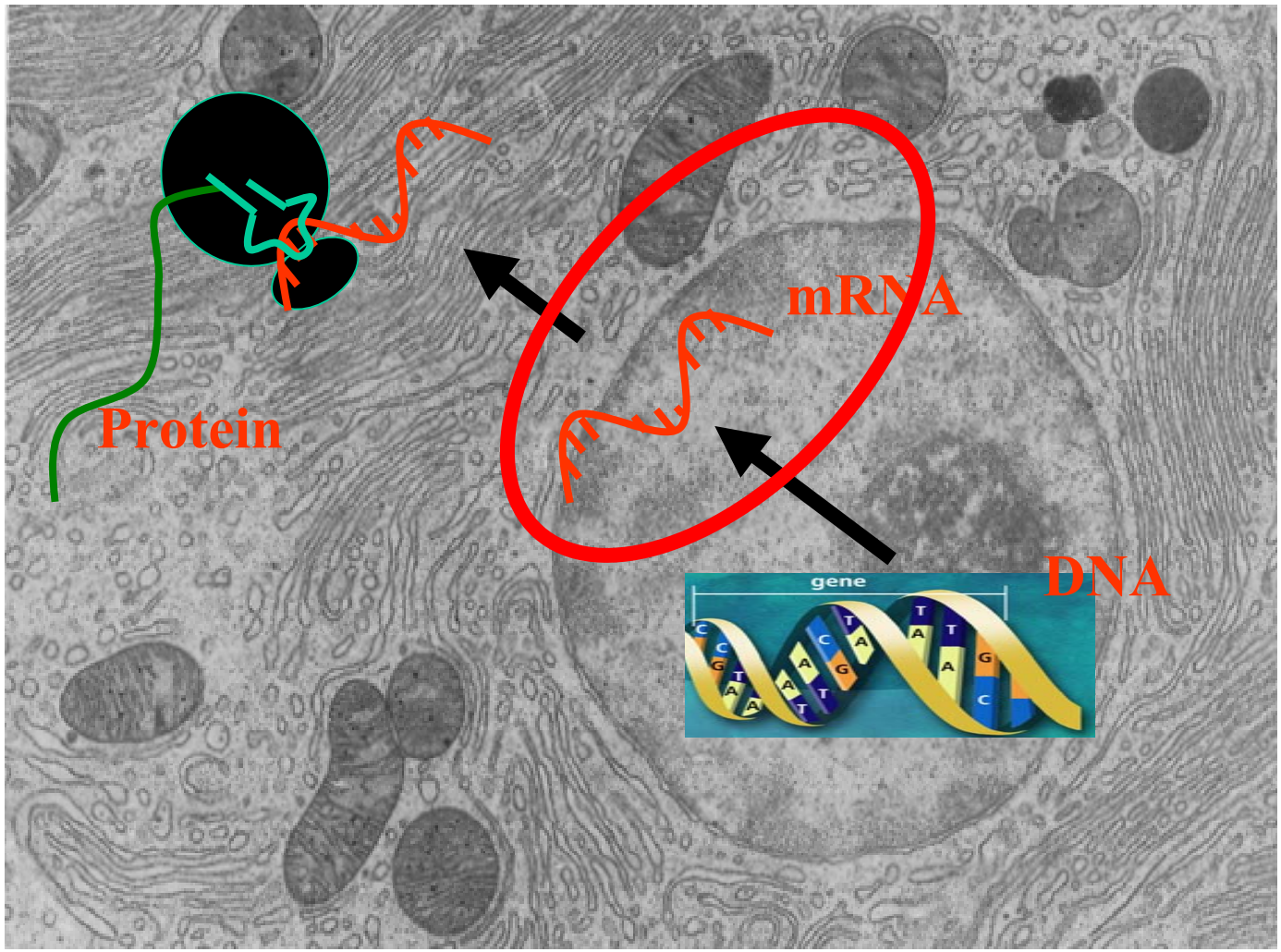
Was sind MICROARRAYS ?

Microarrays sind Technologieplattformen zur Messung der Aktivität einer großen Anzahl von Genen.

Dabei werden ihre Produkte (idR mRNA) quantifiziert.

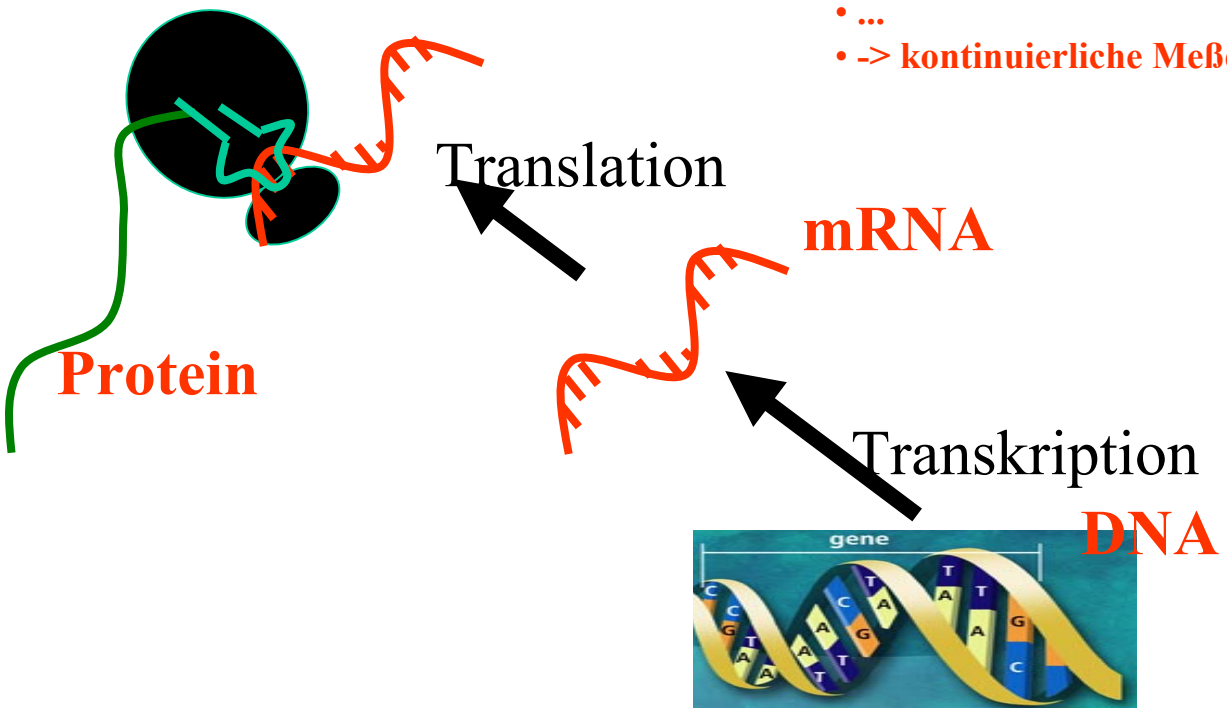
Hierzu werden DNA Sequenzen verwendet, die auf einer Oberfläche (je nach Plattform verschiedene) immobilisiert werden.

Was sind MICROARRAYS ?



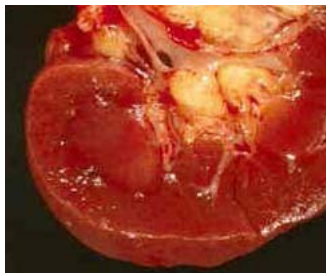
Was sind MICROARRAYS ?

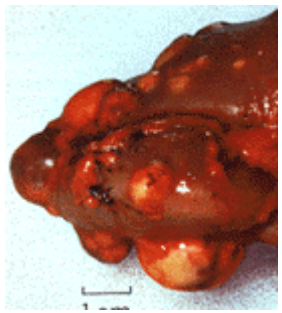
- Wieviel von Gen x ?
- Ist überhaupt etwas von Gen x exprimiert ?
- Ist mehr oder weniger als in einem anderen Patienten da ?
- ...
- -> kontinuierliche Meßdaten, keine binären Daten!



Wie kann man die Aktivität messen ?

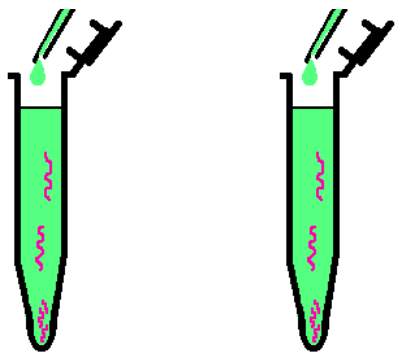
Normale Niere





Tumor (Niere)

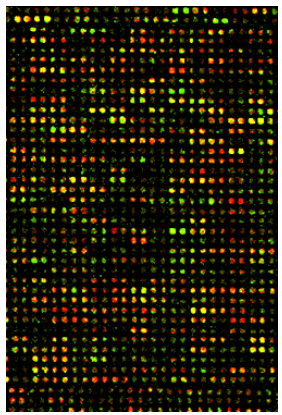
RNA-Präparation



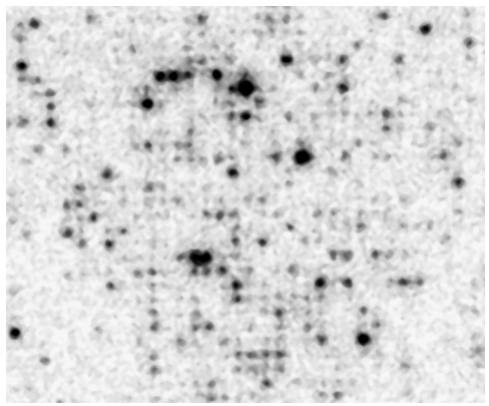
MESSUNG ?!

was unterscheidet
“Tumor” von “Normal” ?

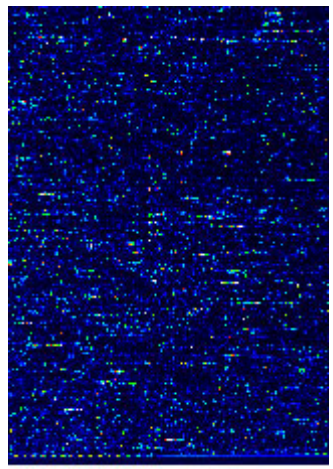
Welche Technologieplattformen gibt es?



Rot Grün



Radioaktiv



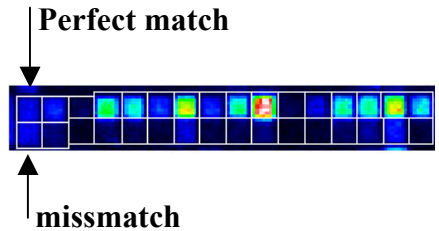
Affymetrix

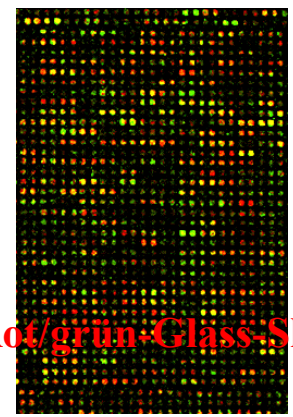
Welche Technologieplattformen gibt es?

- Schena M, Schalon D, David RW, Brown PO Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **1995**

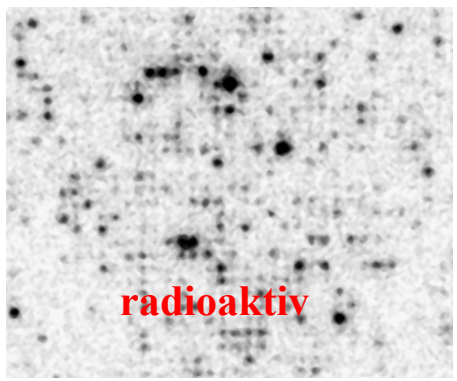
Erste Publikation:

- Lennon GG & Lehrach HH. Hybridization analyses of arrayed cDNA libraries. Trends Genet. **1991**





Rot/grün-Glass-Slides



radioaktiv

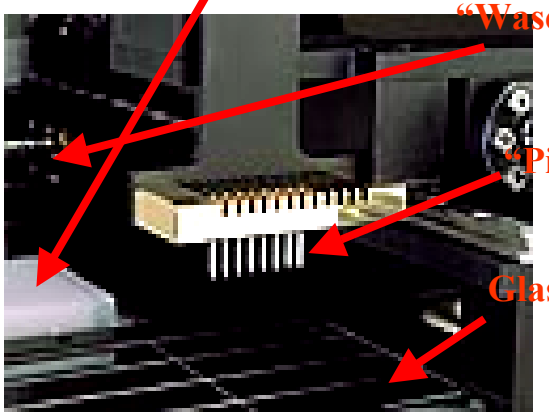
Welche Technologieplattformen gibt es?

Rot Grün

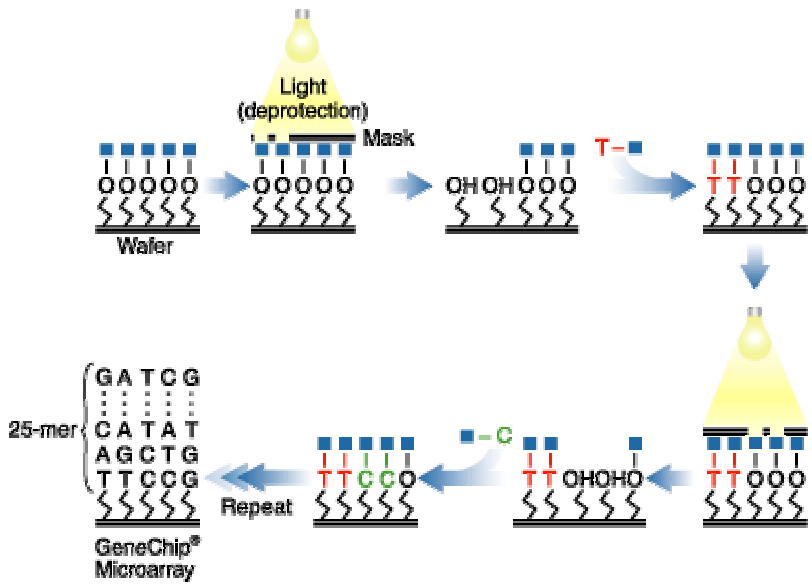
Auswahl “interessanter Sequenzen”

↓

PCR-Amplifikation + Aufreinigung



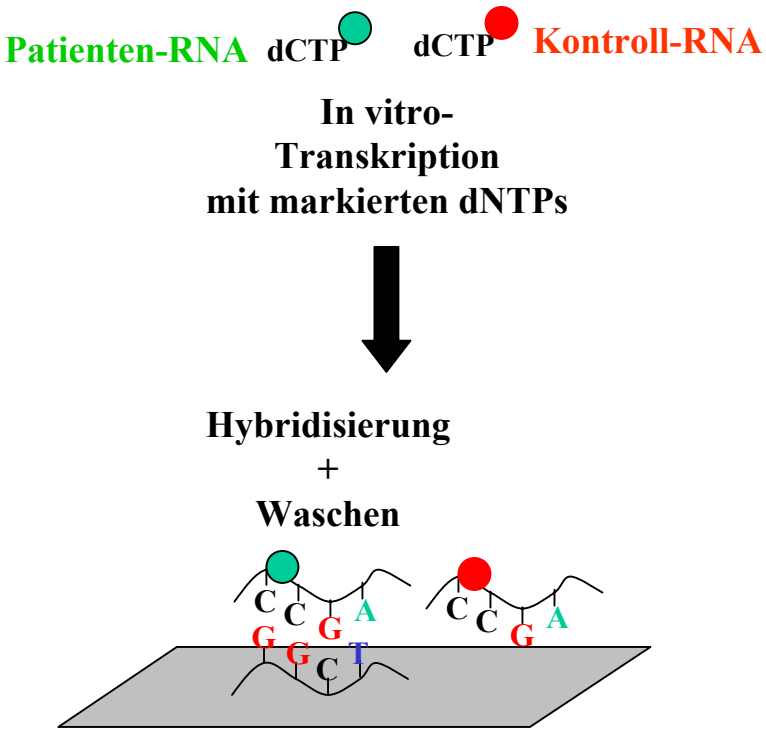
Affymetrix



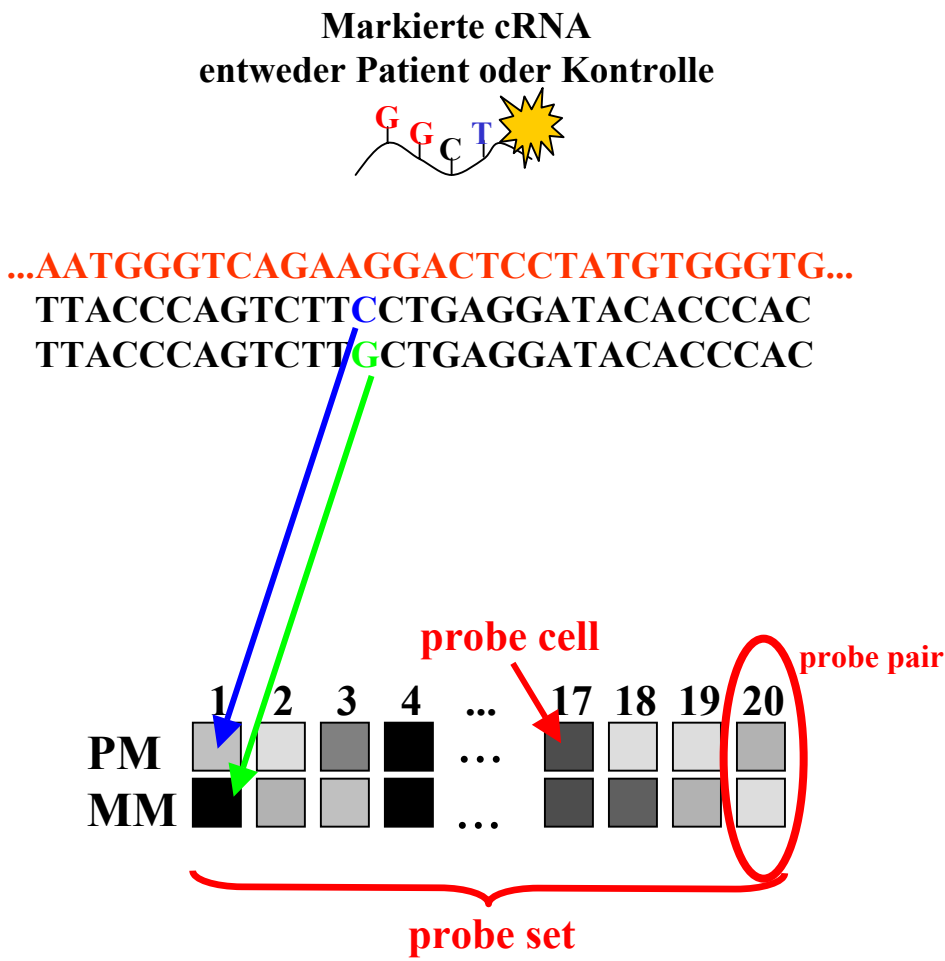
Welche Technologieplattformen gibt es?

Hybridisierung

Rot Grün

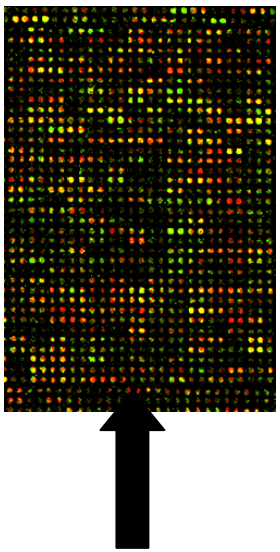


Affymetrix



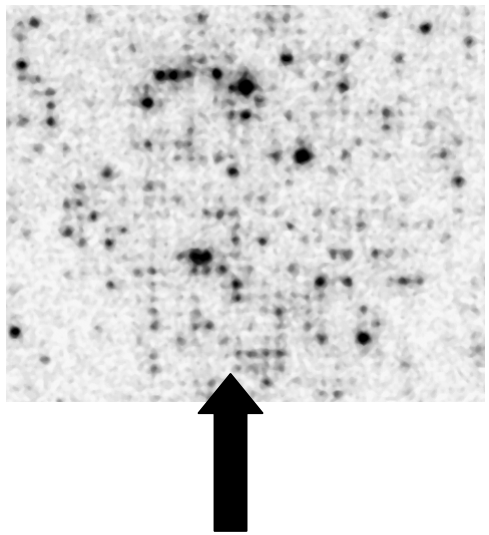
Welche Technologieplattformen gibt es?

Rot Grün



- **Glas Träger**
- **rote und grüne Probe**
- **Floureszenz Signal**

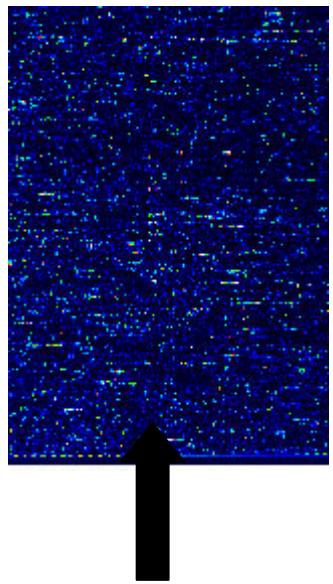
- bis ~ 20000 Spots möglich
- gleichzeitiges Hybridisieren von Probe und Kontrolle (rot/grün)



- **Nylon Filter**
- **eine Probe**
- **radioaktives Signal**

- viele Spots möglich
- große Fläche / lokale Effekte
- Überstrahlen
- nur eine Probe pro Hybridisierungsvorgang

Affymetrix



- **Chip**
- **eine Probe bestehend aus 16-20 Wdh. und zugehörigen Mismatches**

- kommerzieller Chip
- gute reproduzierbare Daten
- nur eine Probe pro Hybridisierungsvorgang

Wie wird ein Chip hergestellt ?

Film: **DKFZ Heidelberg**
W. Huber; G. Sawitzki; H. Sültmann

cDNA Microarrays for Gene Expression Analysis

http://www.dkfz-heidelberg.de/mga/whuber

Welche Fragen kann man mit Chips beantworten ?

Drei Beispiele:

(1) Was bringt eine Stammzelle dazu, eine differenzierte Zelle zu werden ?

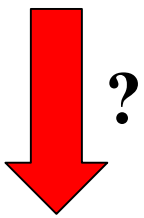
(2) Welche Gene unterscheiden einen Tumor vom normalen Gewebe ?

(3) Welche Faktoren begünstigen einen raschen Tumorprogress ?

Welche Fragen kann man mit Chips beantworten ?

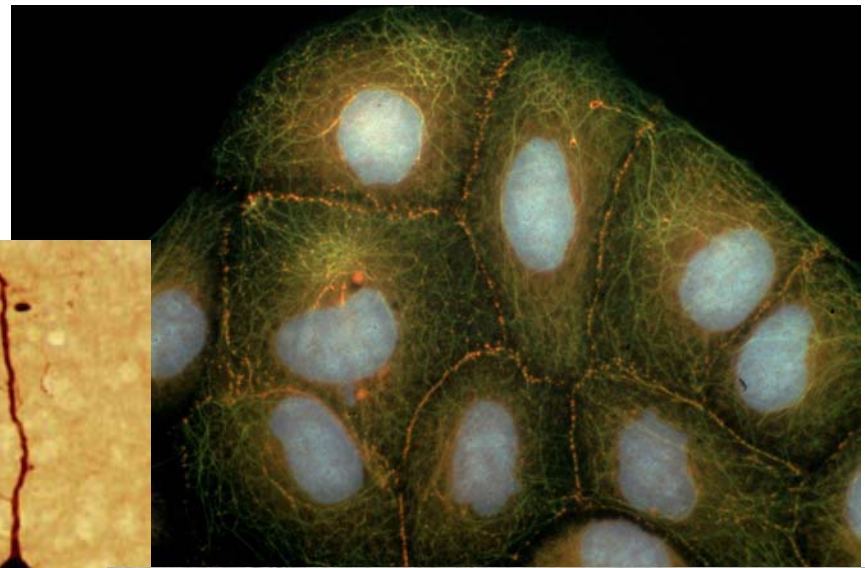
Was bringt eine Stammzelle dazu, eine differenzierte Zelle zu werden?

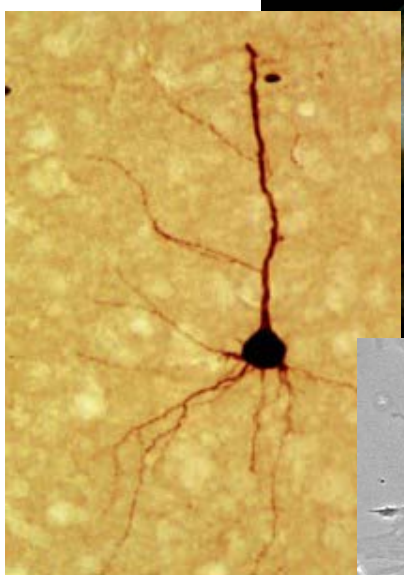
Knochenmark-Stammzelle

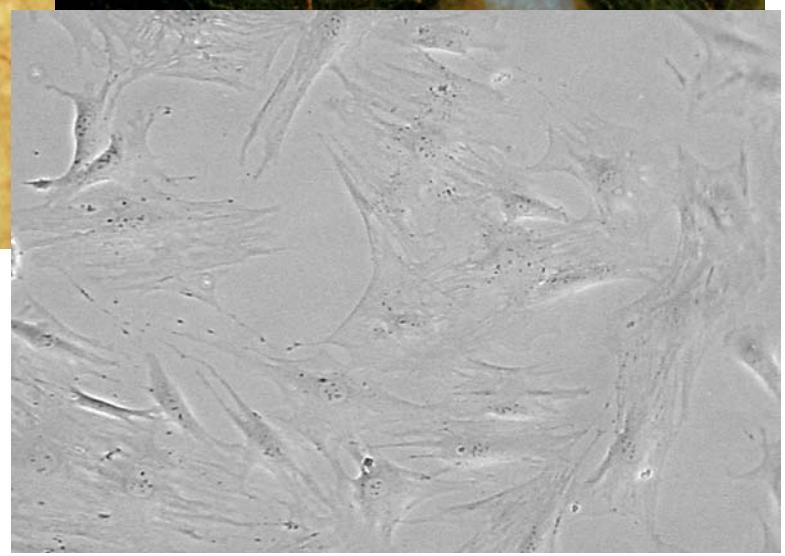


?

Neurale Zelle

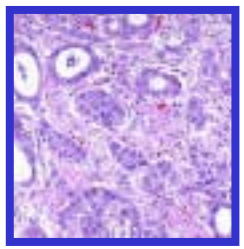




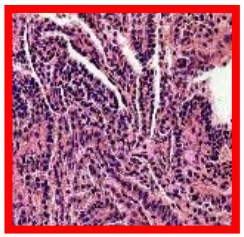


Welche Fragen kann man mit Chips beantworten ?

Welche Gene unterscheiden einen Tumor vom normalen Gewebe ?



Gesund



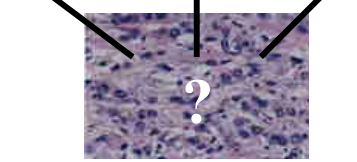
Krank

S#1 Mean	S#1 S.Dev	S#1 Area
1964.028	682.7736	113
2149.386	769.6178	91
906.1724	420.9323	74
3588.557	1168.349	89
60317.82	11562	153
54301.75	20957.93	135
771.2751	409.6172	73
662.4827	309.9964	73
1245.646	923.4761	52
488.5027	297.9345	31
5783.04	1924.275	125
1961.644	1296.955	76
2838.966	964.7534	82

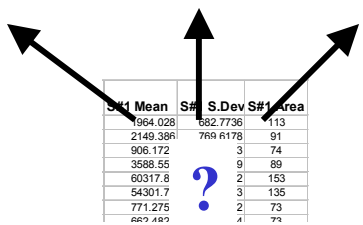
Gesund

S#1 Mean	S#1 S.Dev	S#1 Area
1964.028	682.7736	113
2149.386	769.6178	91
906.1724	420.9323	74
3588.557	1168.349	89
60317.82	11562	153
54301.75	20957.93	135
771.2751	409.6172	73
662.4827	309.9964	73
1245.646	923.4761	52
488.5027	297.9345	31
5783.04	1924.275	125
1961.644	1296.955	76
2838.966	964.7534	82

Krank



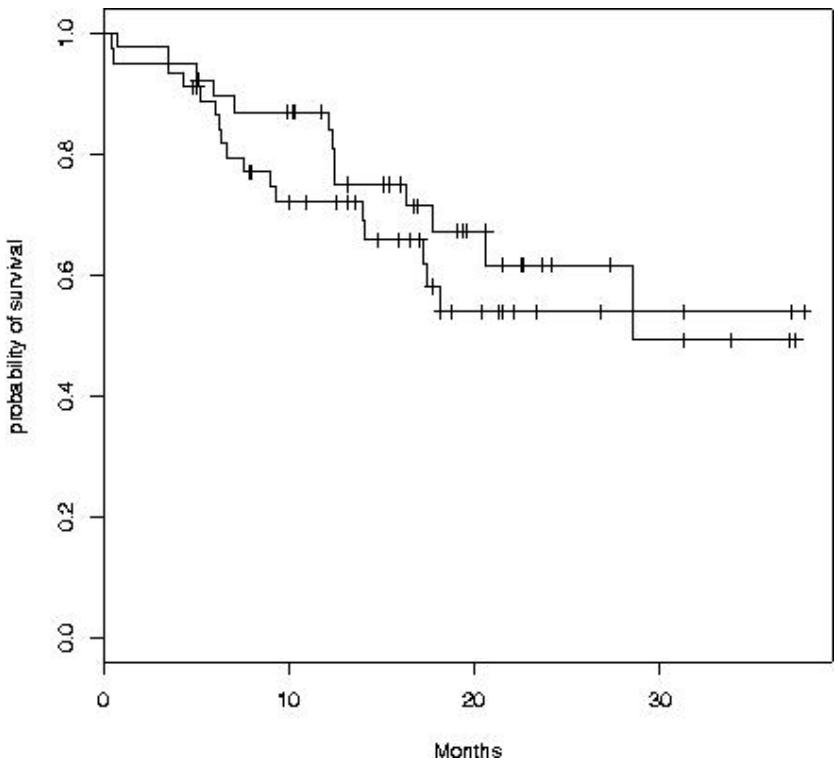
Neuer Patient

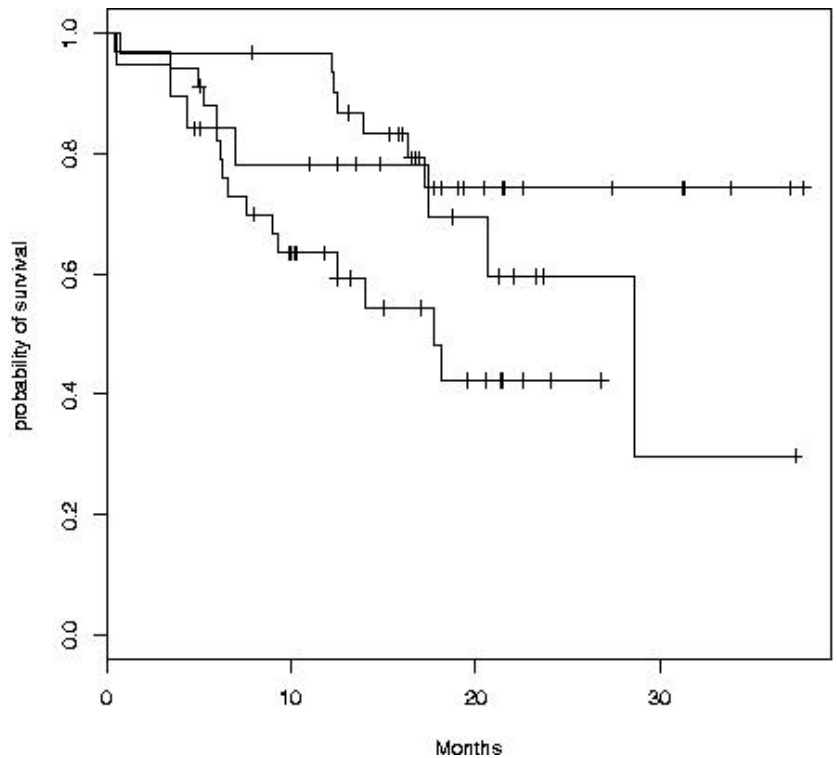


Neuer Patient

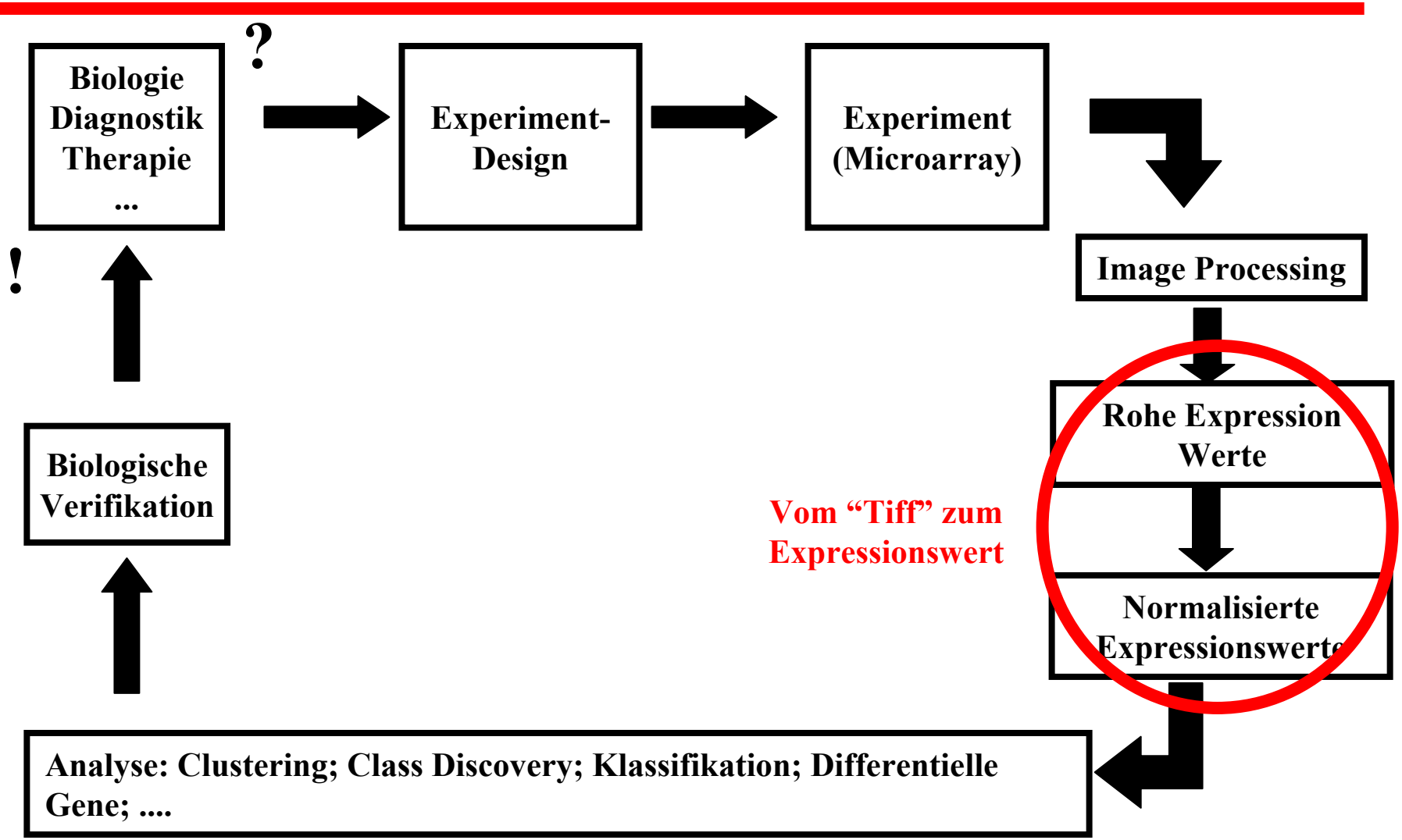
Welche Fragen kann man mit Chips beantworten ?

Welche Faktoren begünstigen einen raschen Tumorprogress ?



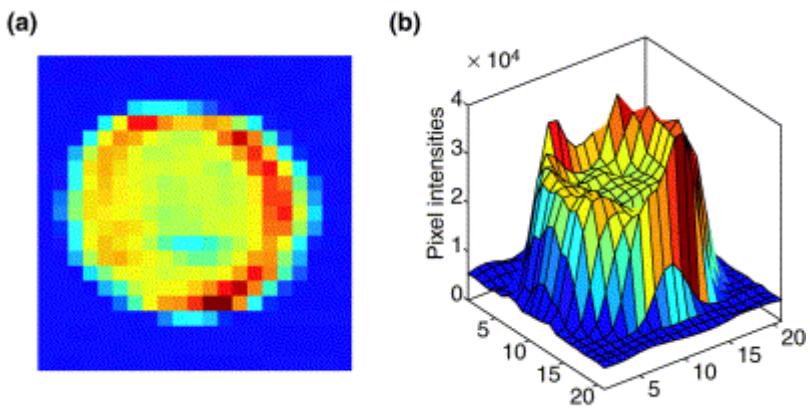


Datenfluß


Biologie
Diagnostik
Therapie
...
?
Experiment-
Design
Experiment
(Microarray)
Image Processing
!
Biologische
Verifikation
Rohe Expression
Werte
Vom "Tiff" zum
Expressionswert
Normalisierte
Expressionswerte
Analyse: Clustering; Class Discovery; Klassifikation; Differentielle
Gene; ....

Was brauchen wir ?

- Intensitäten
- Hintergrund
- Pixel
- Standardabweichung
- Position
- Annotation



Trends in Biotech
Hess et al, 19(11),2001

Beispiel: Affymetrix Experiment

.CEL File:

```
[CEL]
Version=3

[HEADER]
Cols=640
Rows=640
TotalX=640
TotalY=640
OffsetX=0
OffsetY=0
GridCornerUL=232 233
GridCornerUR=4490 220
GridCornerLR=4495 4484
GridCornerLL=238 4498
Axis-invertX=0
AxisInvertY=0
swapXY=0
DatHeader=[0..46139]  2353t99hpp_av08:CLS=4733 RWS=4733 XIN=3  YIN=3  VE=17
2.0 10/12/00 15:29:25 HPB4   ^T  ^T HG_U95A.1sq ^T  ^T  ^T  ^T  ^T  ^T  ^T  ^T  ^T 6
Algorithm=Percentile
AlgorithmParameters=Percentile:75;CellMargin:2;OutlierHigh:1.500;OutlierLow:1.004

[INTENSITY]
NumberCells=409600
CellHeader=X    Y      MEAN    STDV    NPIXELS
 0      0     175.0   31.2     25
 1      0     7935.5  1296.9   20
 2      0     175.0   31.3     25
 3      0     7979.0  1439.2   25
 4      0     83.3    15.2     20 ….
….
```

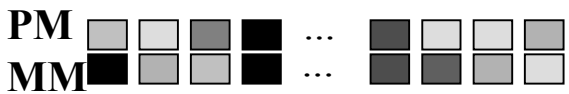
.CDF File:

```
[CDF]
Version=GC3.0

[Chip]
Name=HG_U95E
Rows=640
Cols=640
NumberOfUnits=12639
MaxUnit=12672
NumQCUnits=13
ChipReference=

[QC1]
Type=10
NumberCells=300
CellHeader=X    Y     PROBE   PLEN    ATOM    INDEX   MATCH   BG
Cell1=167       80      N       20      1       51367   0       0
Cell2=167       81      N       20      1       52007   1       0
Cell3=167       82      N       20      1       52647   0       0
Cell4=167       83      N       20      1       53287   0       0
Cell5=167       84      N       1       1       53927   -1      1
Cell6=168       80      N       20      2       51368   0       0
```

.CIF File:

```
[Chip]
Rows=640
Cols=640
CellMargin=2
CellMarginDefault=2
XOrigin=-7100
YOrigin=8140
Width=14200
Height=14200
FocusXOrigin=-7200
FocusYOrigin=8040
FocusWidth=14400
FocusHeight=14400
PixelSize=300
Wavelengths=570
NScans=2

[HP]
XOrigin=-7100
YOrigin=8140
Width=14200
Height=14200
FocusXOrigin=-7200
FocusYOrigin=8040
FocusWidth=14400
FocusHeight=14400
PixelSize=300
Wavelengths=570
NScans=2
```

Preprocessing



- **Hintergrund**
- **20x“PM“; 20x“MM“ (~20000 mal)**
- **einige „MM“ sind größer als die zugehörigen „PM“‘s !**
- **aus den 20+20 Werten soll ein Expressionswert abgeleitet werden**
- **systematische Fehler und ungleiche Varianzen**

Preprocessing: ein Lösungsvorschlag

MAS 5.0

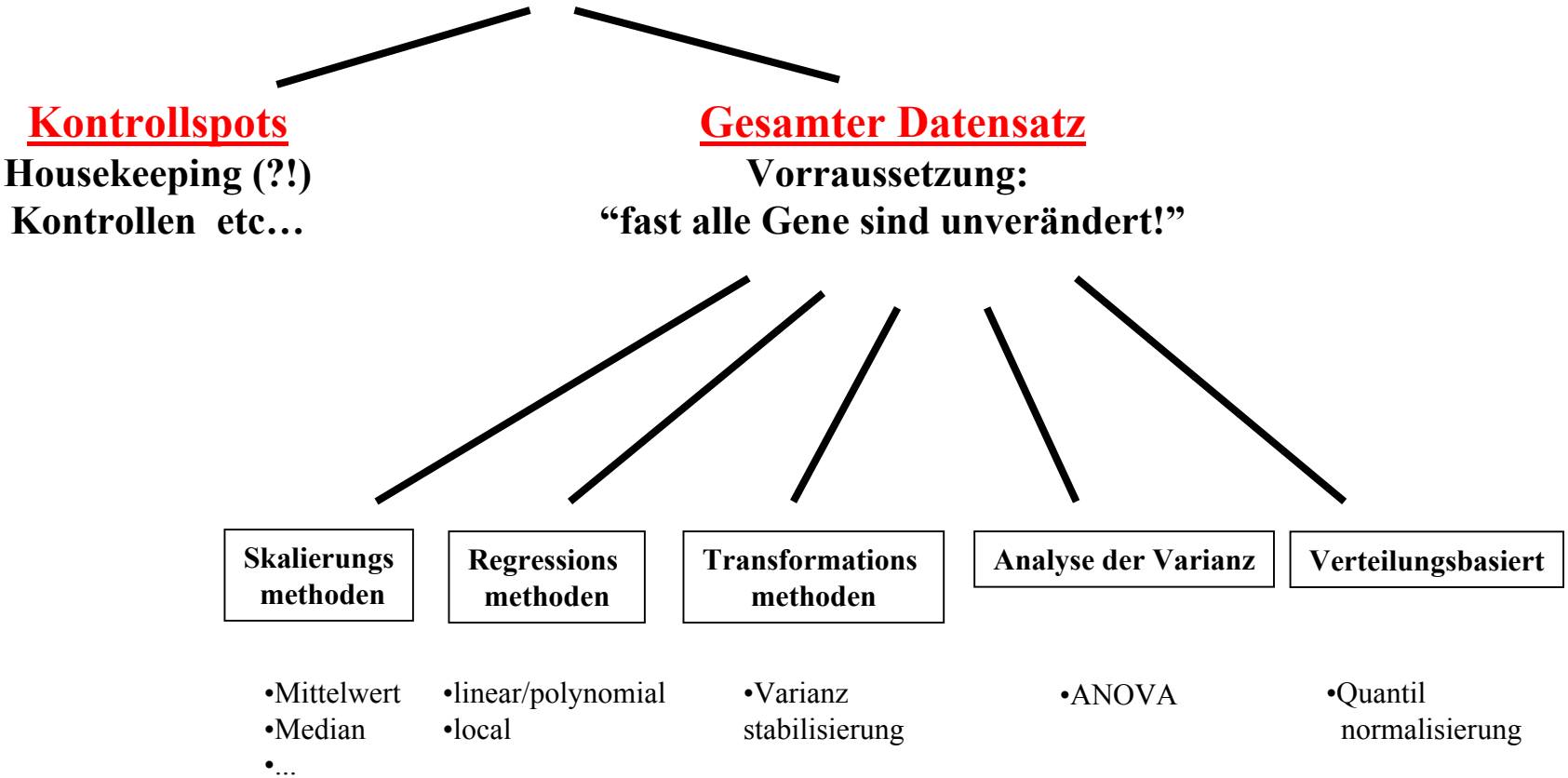
(1) Was ist Hintergrund ?

(2) Wie behandeln wir
„PM“ und „MM“ ?

(3) Wie sollte man summieren ?

		PM	MM	…
Gen 1	Cell1	23	913	
	Cell2	627	32	
	Cell3	367	34	
	Cell4	276	43	
	Cell5	748	90	
	Cell6	278	38	
	Cell7	672	39	
	Cell8	9	263	
	Cell9	1002	373	
	Cell10	2019	43	
	Cell11	378	578	
	Cell12	278	303	
	Cell13	378	20	
	Cell14	298	32	
	Cell15	389	12	
	Cell16	803	...	
	Cell17	289		
	Cell18	...		
	Cell19			
	Cell20			
Gen 2	Cell1			
	Cell2			
	Cell3			
	Cell4			
	...			

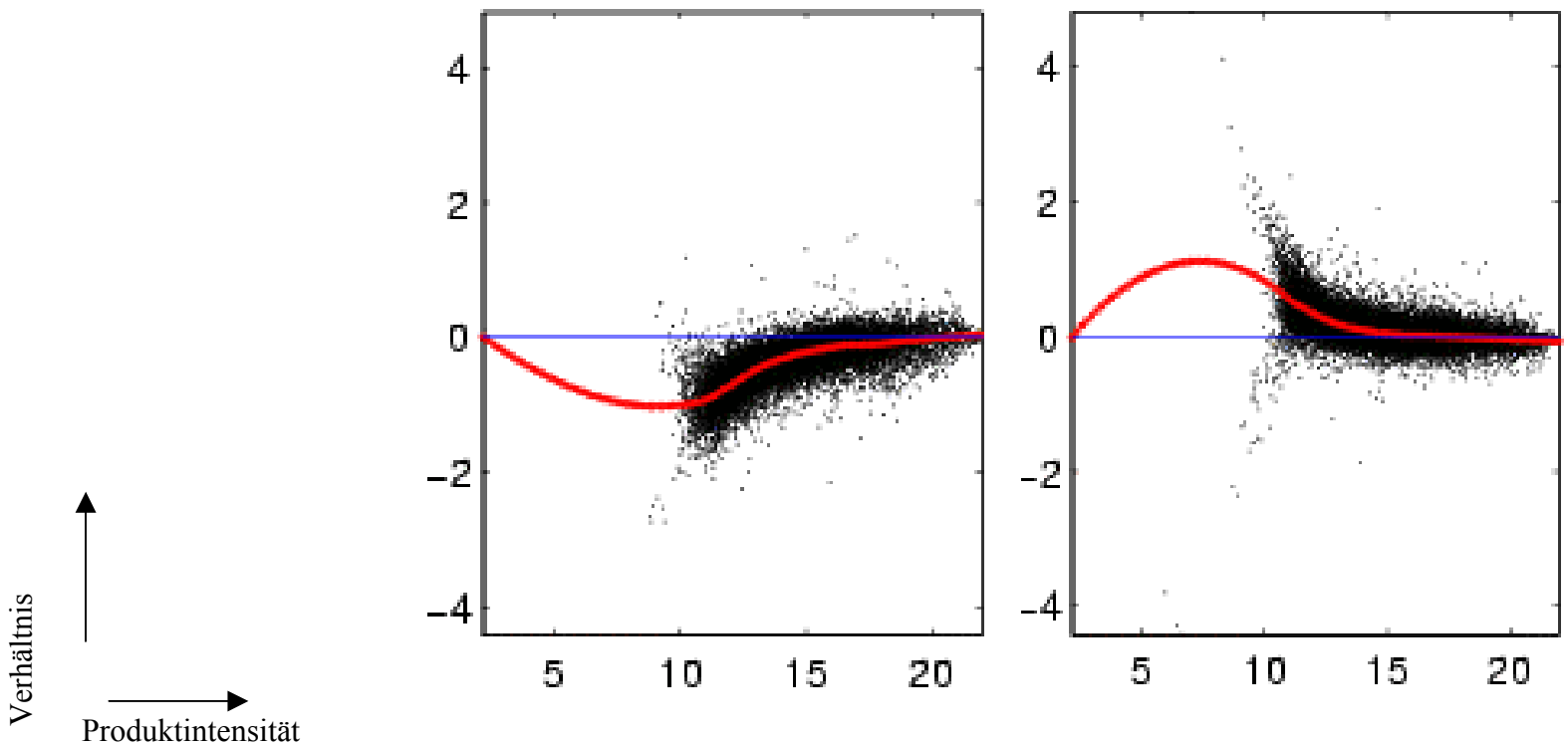
Normalisierung



> http://www.bioconductor.org

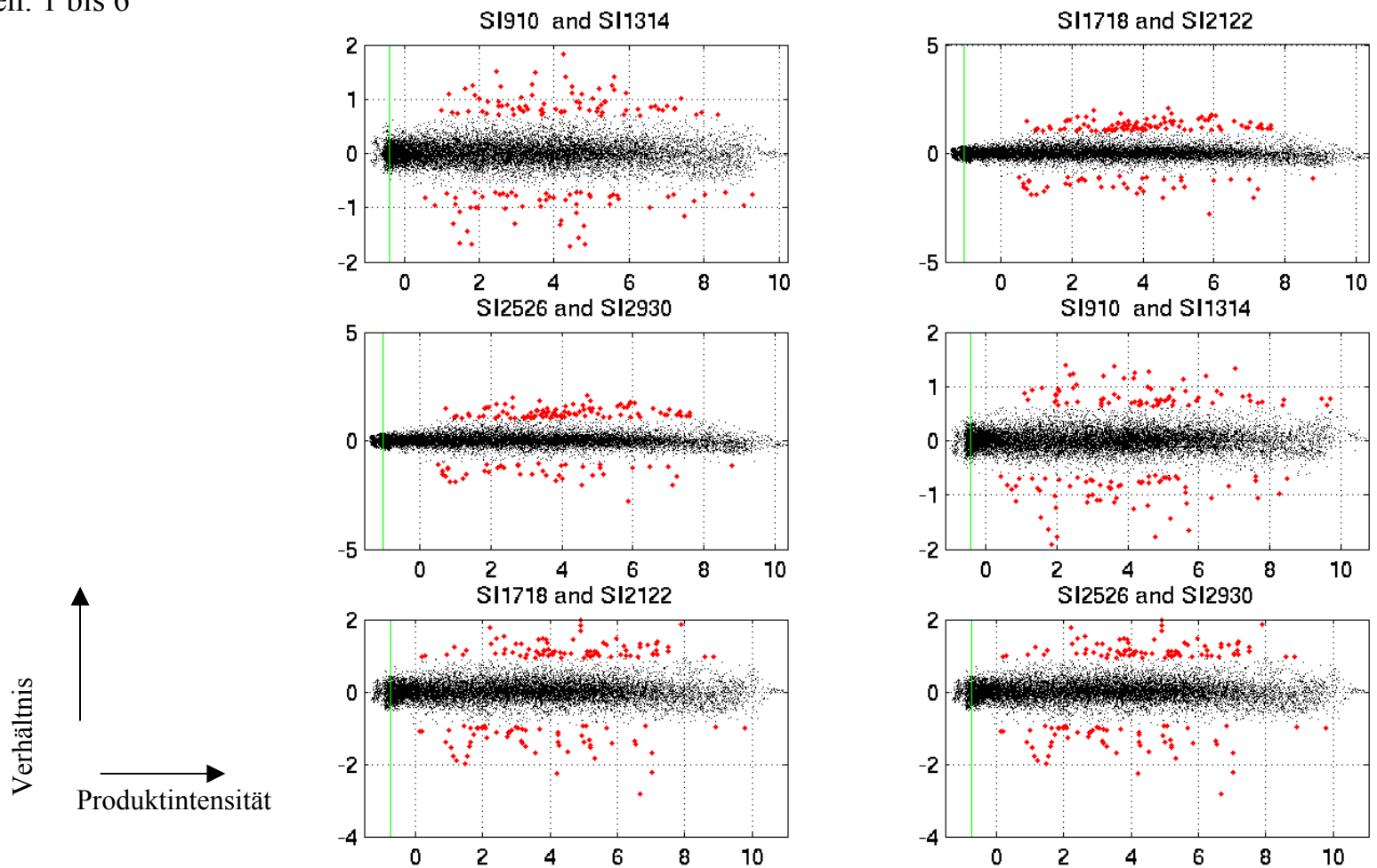
Normalisierung: ein Lösungsvorschlag

Loess / lokale Regression



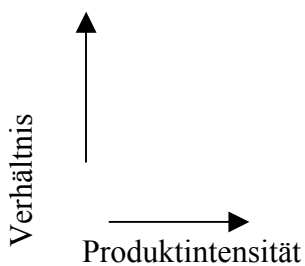
Differentielle Gene

Patienten: 1 bis 6



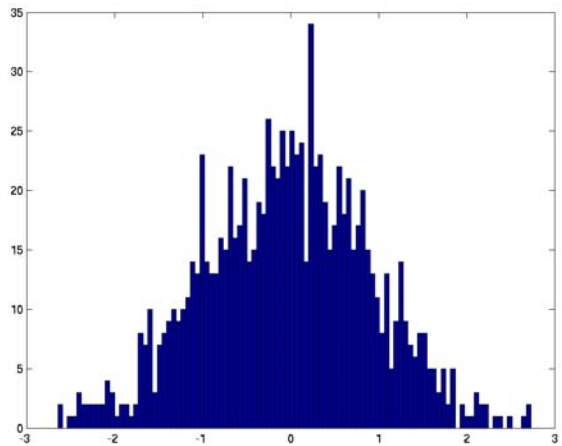
Differentielle Gene

- **<u>einfachste Methode</u>: suche alle Gene mit mind. “twofold change”**
  - **nicht statistisch**
  - **willkürliche Wahl**
  - **je nach Experiment und Varianz unterschiedlich “gut”**
  - **je nach Normalisierungsmethode starker “Bias”**
  - **wenn keine Varianzstabilisierung vorgenommen wurde, tendieren niedrig exprimierte Gene zu hoher Streuung**



Differentielle Gene

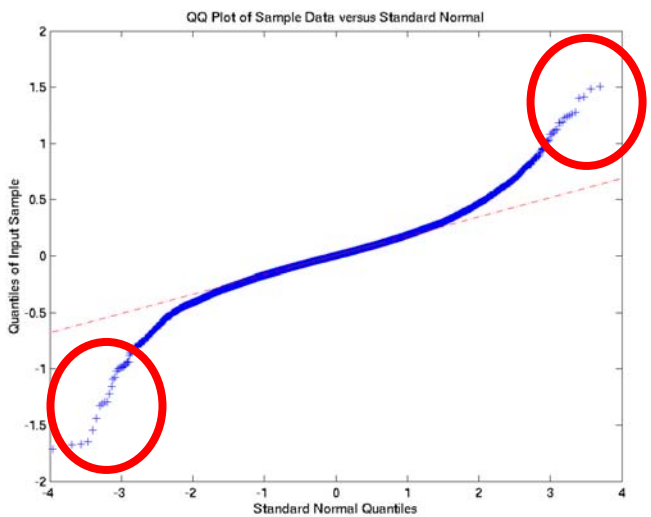
- **Methode: t-Test und Modifikationen**



t Statistik für jedes Gen:

$$\frac{\bar{x}_{kontrolle} - \bar{x}_{patient}}{\sqrt{[(1/n_{kontrolle})SD_{kontrolle}^2 + (1/n_{patient})SD_{patient}^2]}}$$

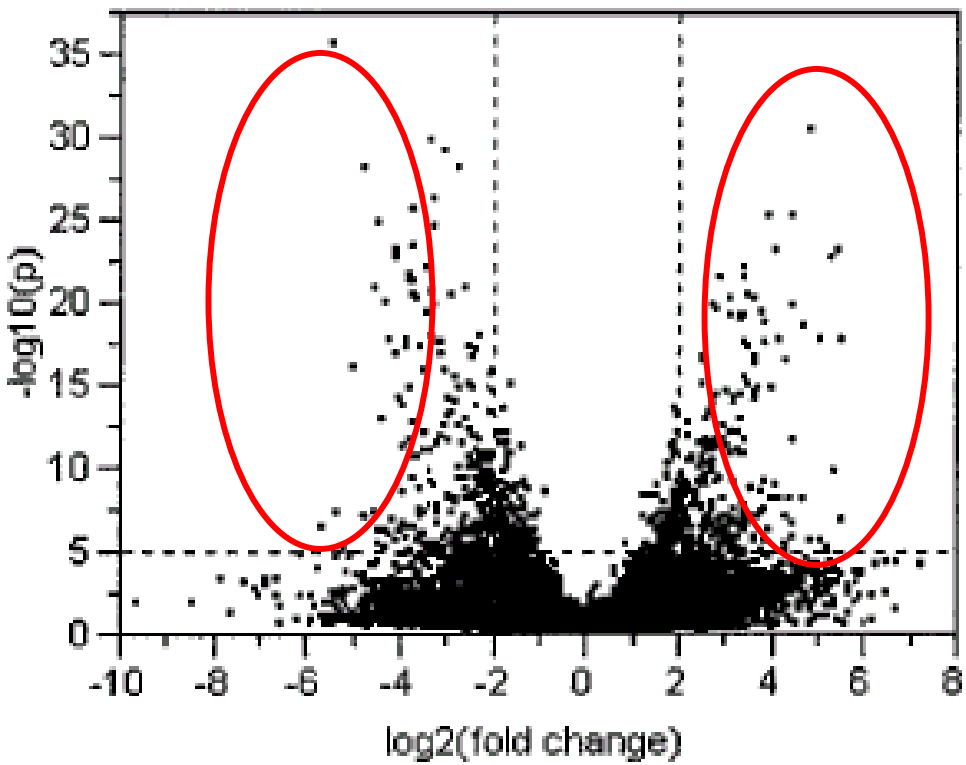
QQ-Plot zur Visualisierung:



Differentielle Gene

- **ttest:**
  - **p = tcdf(t,Freiheitsgrade)**
  - **Problem: Multiples Testen (einige tausend mal !)**
- **ttest mit Bonferroni Adjustierung:**
  - **Multipliziere die p-Werte mit der Anzahl der Tests.**
  - **sehr konservativ**
- **SAM (Significance analysis of microarrays):**
  - **Ziel: Gene mit kleinen "fold changes" werden nicht signifikant**
  - **addiere eine Konstante c (90% Quantil des Standarderrors) im Nenner**
- **Regularisierter ttest:**
  - **$S = ratio / \sqrt{(cSE^2 + (n-1)\,SE^2) / (c+n-2)}$**
- **B-Statistik:**
  - **Logarithmus eines Wahrscheinlichkeiten-Ratios**
  - **Zähler: Wahrscheinlichkeit, daß ein Gen differentiell ist**
  - **Nenner: Wahrscheinlichkeit, daß ein Gen nicht differentiell ist**

Differentielle Gene: Grafische Darstellung



JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 8, Number 6, 2001

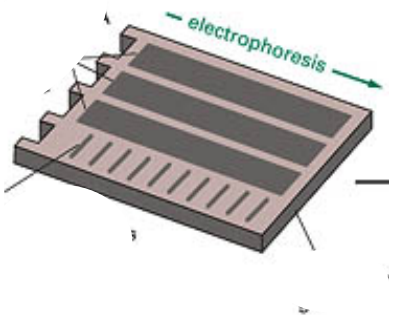
Wolfinger et al

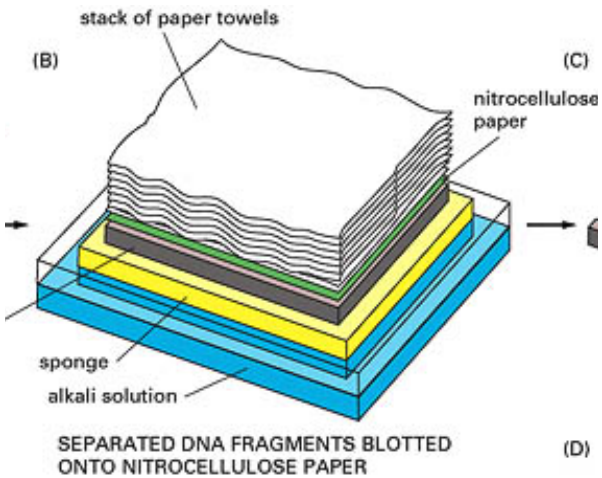
Biologische Verifikation

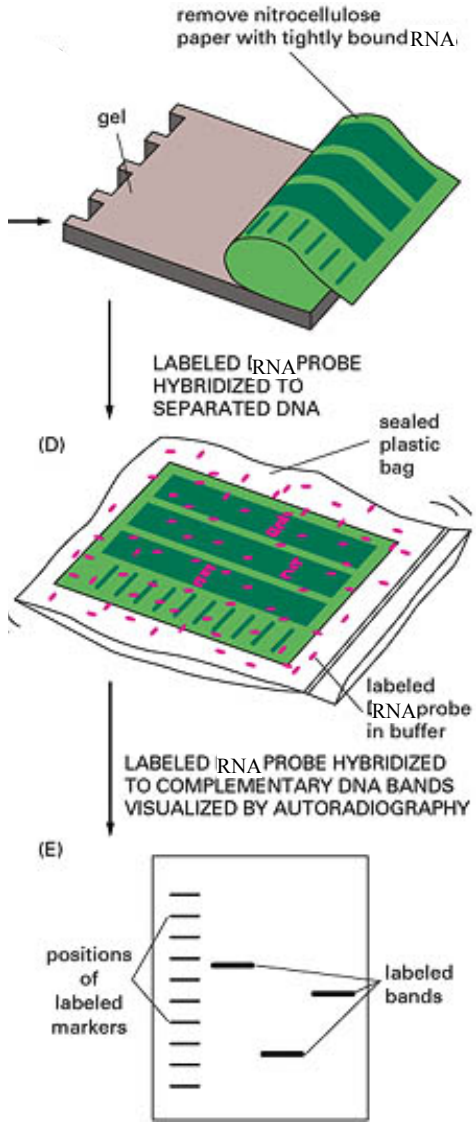
Welche unabhängigen Methoden zur Verifikation der Microarray-Expressions Ergebnisse sind möglich?

- Northern Blot
- RT PCR
- SAGE
- quantifizierbare Kontrollen

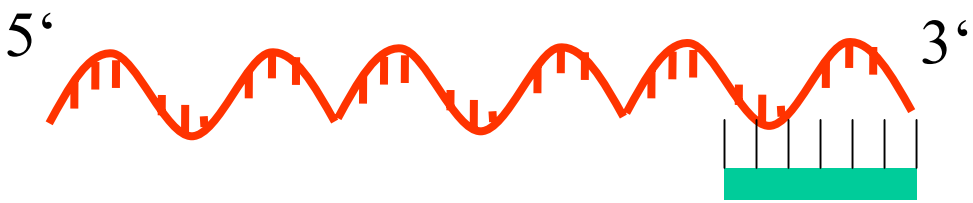
Biologische Verifikation: Northern Blot



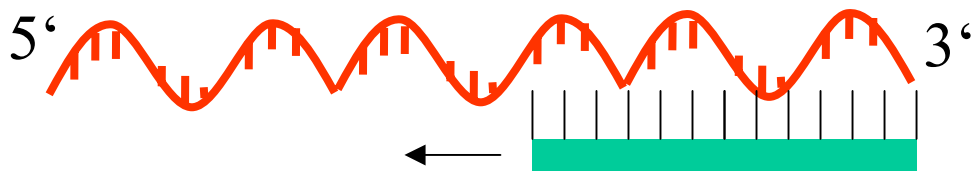




Biologische Verifikation: RT PCR

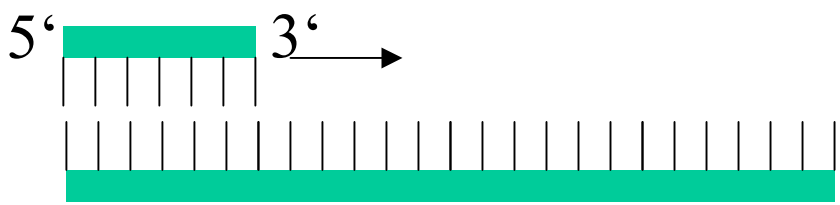


RNA

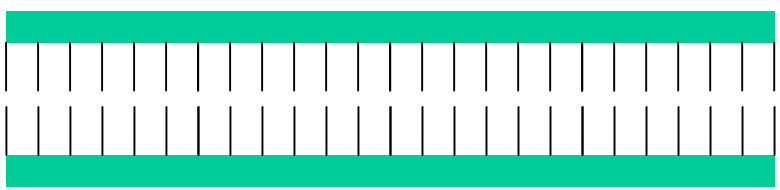


RNA

cDNA



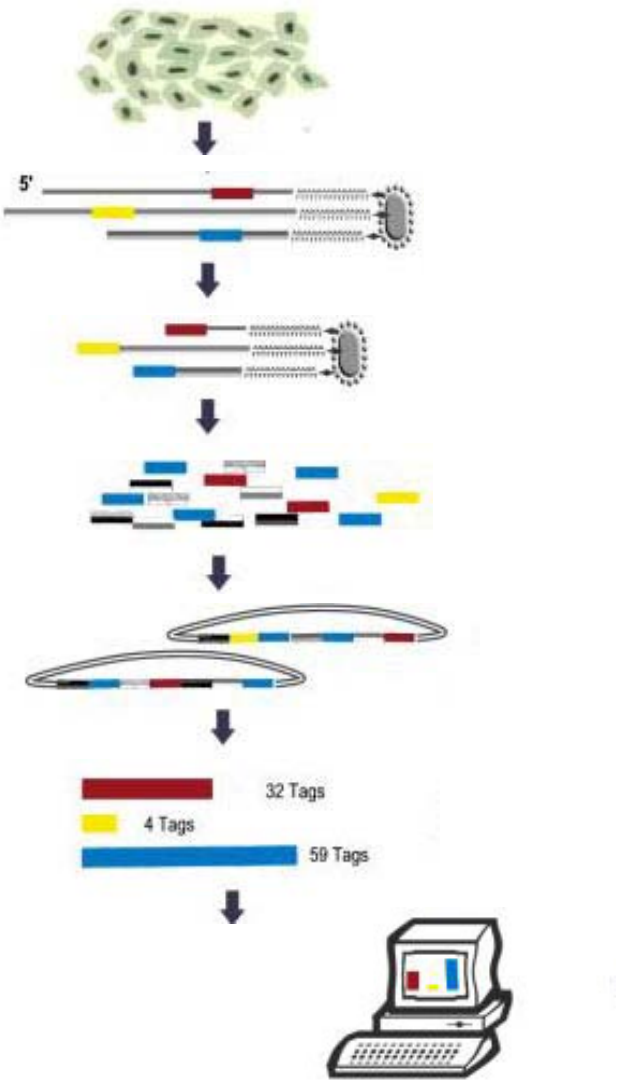
cDNA



dsDNA

- **Da RNA durch PCR nicht direkt amplifiziert werden kann, muß sie zunächst in cDNA umgeschrieben werden (revers transkribiert, RT)**
- **Zur Quantifizierung sind zwei Ansätze möglich:**
- **1 Interner endogener Standard (zB Housekeeping gene)**
- **2 Kompetitive RT PCR: Zugabe von sog Mimic Fragmenten, die der Reaktion zugegeben werden und zusammen mit der eigentlichen Zielsequenz amplifiziert werden**

Biologische Verifikation: SAGE
Serial Analysis of Gene Expression



Zellen isolieren

mRNA isolieren und cDNA synthetisieren

Transkript mit Anchor Enzym schneiden

„Taggen“

Ligieren der Tags

Sequenzierung

Quantifizierung