V8 Genexpression - Microarrays

- **Idee:** analysiere die Ko-Expression von mehreren Genen um auf funktionelle Ähnlichkeiten zu schließen

- **wichtige Fragen:**

  (1) wie wird Genexpression reguliert?

  (2) was wird mit MicroArray-Chips gemessen?

  (3) wie analysiert man Daten aus MicroArray-Experimenten?

  (4) was bedeutet Ko-Expression funktionell?

- **Inhalt V8:**

  (1) Hintergrund zu Transkription und Genregulationsnetzwerken

  (2) Micro-Arrays

  (3) Übung: analysiere selbst Daten aus einem MicroArray-Experiment

das Transkriptom

Als **Transkriptom** kennzeichnet man den jeweiligen Level an transkribierter messenger RNA (mRNA) für alle Gene des Genoms.

Dies beinhaltet Protein-kodierenden Gene und RNA-kodierende Gene,

die nicht in Protein translatiert werden.

An die eigentliche Transkription in **pre-mRNA** schließen sich noch viele Prozessierungsschritte zur eigentlichen mRNA an, wie

- die Anheftung eines ca. 250 nt-langen **PolyA-Schwanzes**,
- evtl. Editing (Austausch von Nukleotidbasen), sowie
- Spleißen.

Heute werden wir uns auf den reinen Prozess der DNA-Transkription beschränken.

Transkription durch RNA Polymerase II

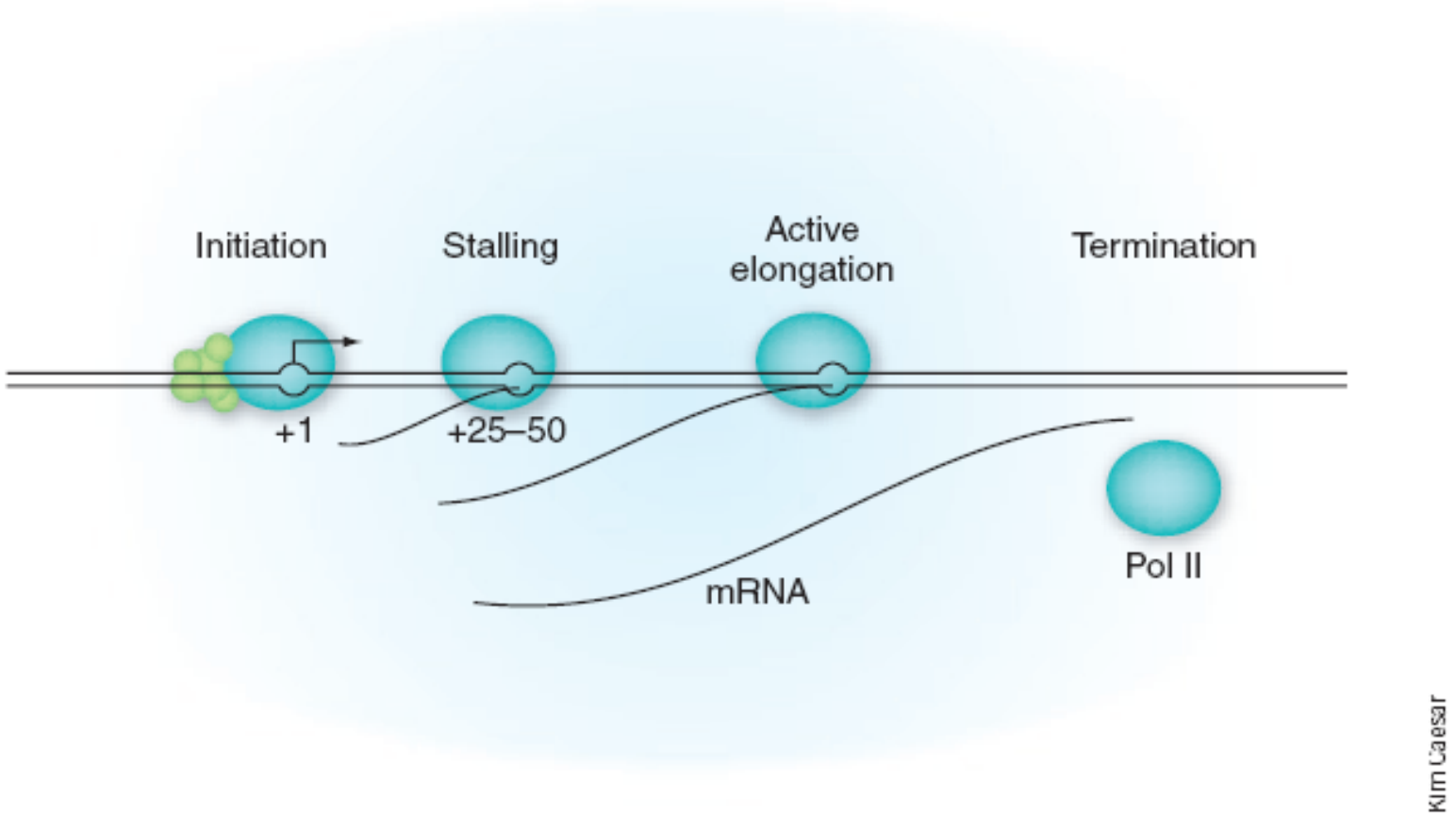


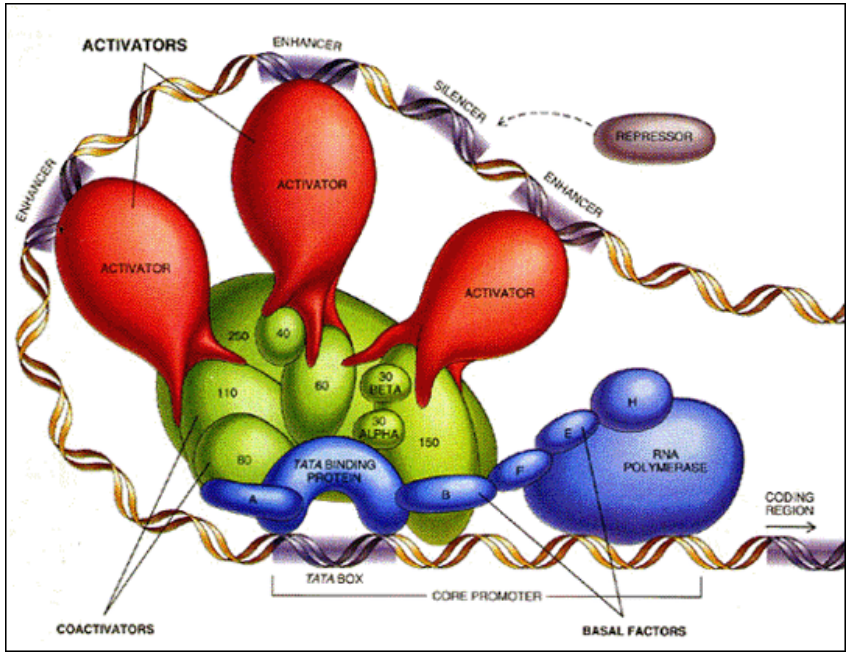
Figure 1 Transcription by RNA polymerase II. Eukaryotic transcription involves a cycle of highly regulated events[1]. After clearing the promoter, RNA polymerase II may pause or stall 25–50 base pairs downstream of the transcription start site before transcribing the body of the gene. Pausing is subject to both positive and negative regulation.

Tamkun J. Nat. Gen. 39, 1421 (2007)

Transkriptions – Gen-Regulationsnetzwerke

Die **Maschine,** die ein Gen transkribiert, besteht aus etwa 50 Proteinen, einschließlich der **RNA Polymerase**. Dies ist ein Enzym, das DNA code in RNA code übersetzt.

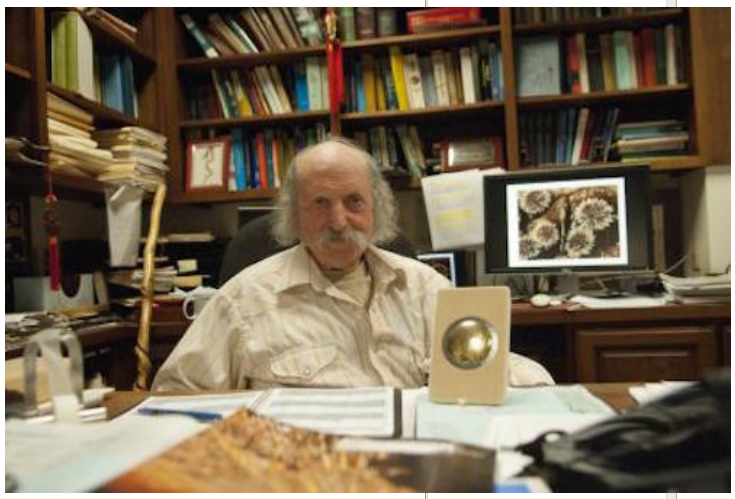
Eine Gruppe von **Transkriptions-faktoren** bindet an die DNA gerade oberhalb der Stelle des Kern-**Promoters**, während assoziierte Aktivatoren an Enhancer-Regionen weiter oberhalb der Stelle binden.



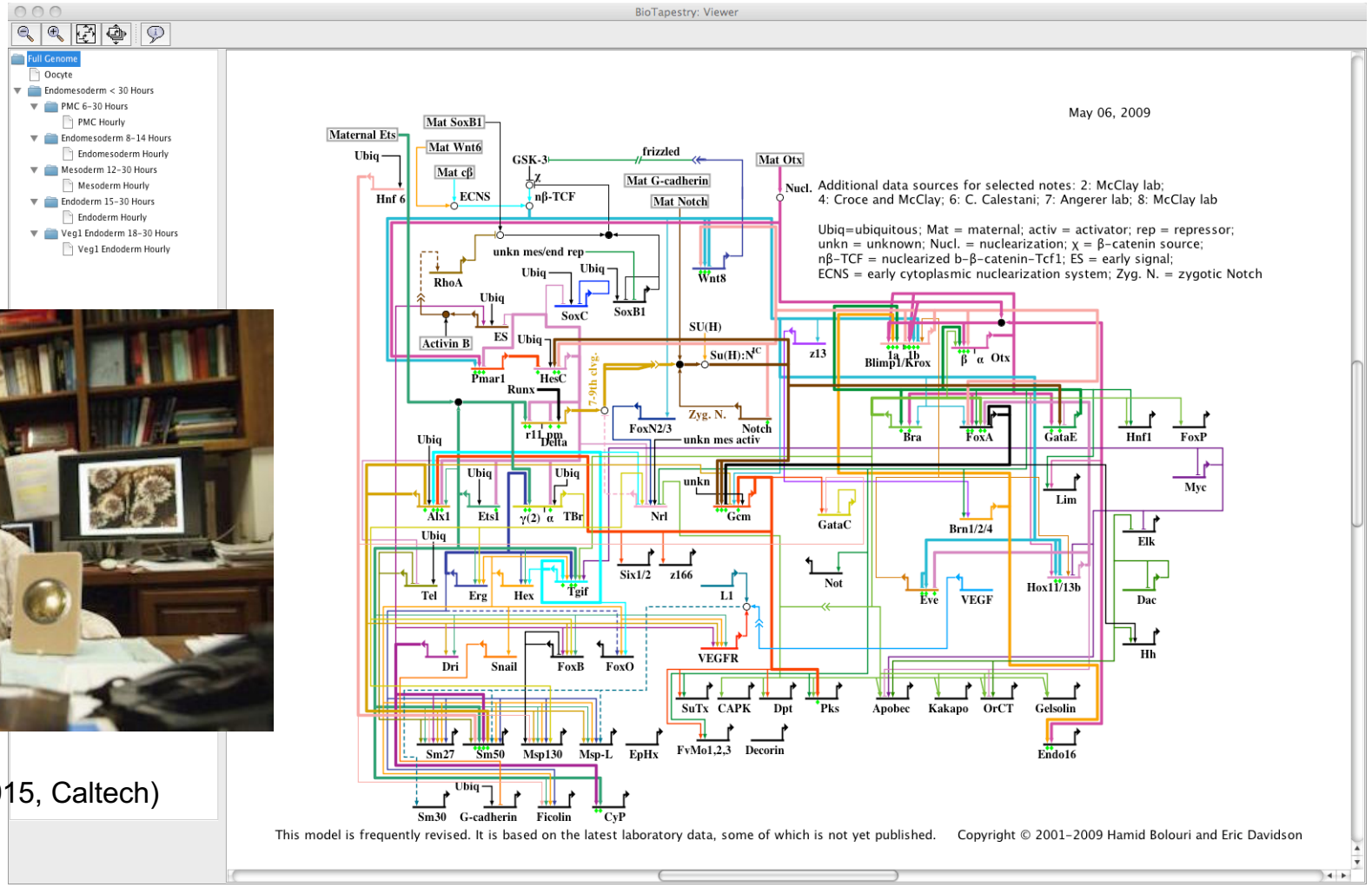
Roger Kornberg
(Stanford Univ)
Noble prize chemistry 2006
„for his studies of the molecular basis of eukaryotic transcription“

http://www.berkeley.edu/news/features/1999/12/09_nogales.html
http://www.osti.gov/

Gen-Regulationsnetzwerk der Seegurke



Eric Davidson (1937 – 2015, Caltech)

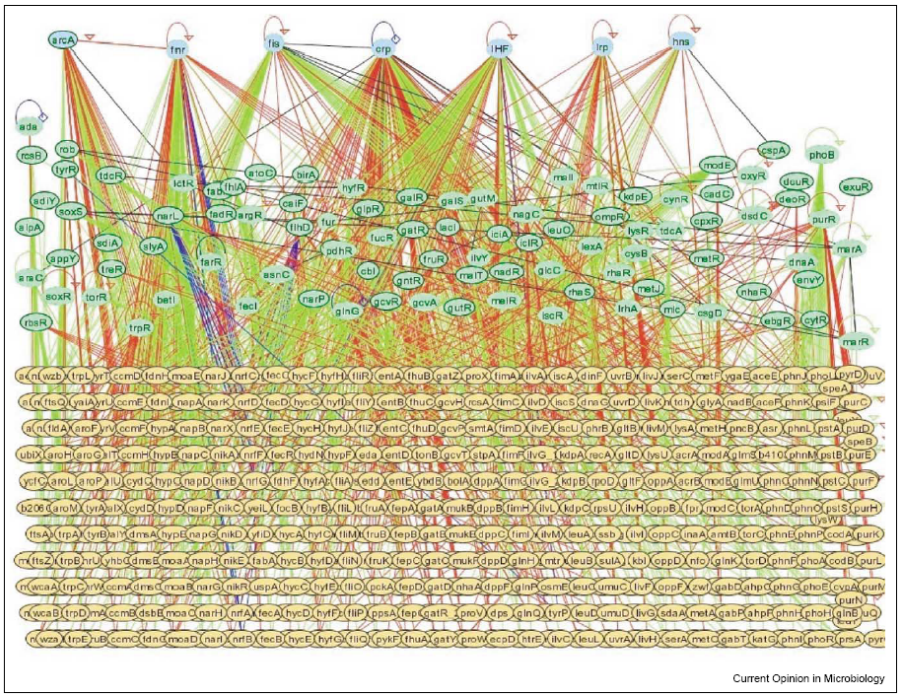


http://sugp.caltech.edu/endomes
http://www.evolutionnews.org/

regulatorisches Netwerk von *E. coli*

RegulonDB: Datenbank mit Information zur transkriptionellen Regulation in *E.coli*; 167 Transkriptionsfaktoren steuern Tausende von Genen.

Durch den hierarchischen Aufbau reichen 7 regulatorische Proteine (CRP, FNR, IHF, FIS, ArcA, NarL and Lrp) aus um die Expression von mehr als der Hälfte aller *E.coli* Gene zu modulieren.



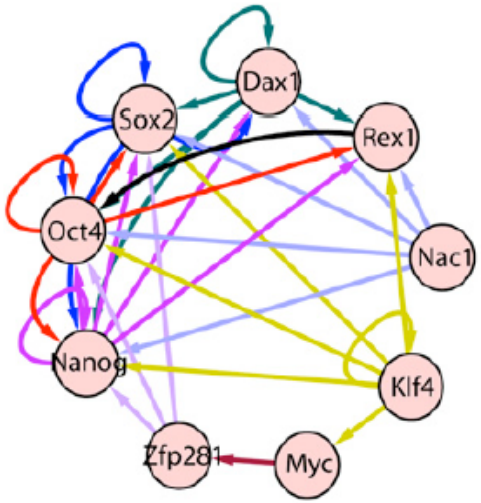
Julio Collado-Vides,
UNAM Mexico-City

Martinez-Antonio, Collado-Vides, Curr Opin Microbiol 6, 482 (2003)

Genregulationsnetzwerk in ESCs um Oct4

Ein eng verwobenes Netzwerk aus neun Transkriptionsfaktoren hält embryonale Stammzellen (ESC) im pluripotenten Zustand.

Der Masterregulator Oct4 sowie Sox2 und Dax1 haben autoregulatorische Feed-Forward Feedback-Schleifen.



Kim et al. Cell 132, 1049 (2008)

veränderte Genregulation bei Krankheiten etc.

Ausgangspunkt: bestimmte Krankheiten (Krebs ?) entstehen anscheinend durch die veränderte Expression einer Anzahl von Genen, nicht eines einzelnen Gens.

Wie kann man alle Gene identifizieren, die für diese Veränderung des Phänotyps verantwortlich sind?

Am besten müsste man z.B. die Expression aller Gene in den Zellen von gesunden Menschen und von Krebspatienten bestimmen.

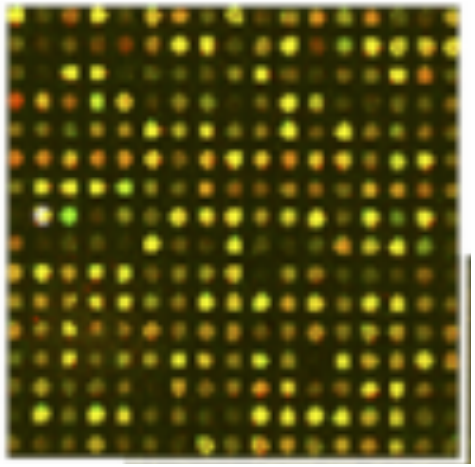
Dann möchte man herausfinden, worin die Unterschiede bestehen.

Genau dies ermöglicht die Methode der **Microarrays**.

Microarrays messen die Expression „aller“ Gene zu einem bestimmten Moment im Zellzyklus unter bestimmten Umgebungsbedingungen.

Was wird mit Microarrays gemessen?

Microarrays enthalten eine Menge an DNA-Proben, die an definierten Positionen an eine feste Oberfläche, z.B. eine Glas-schicht gebunden sind.



Die Proben sind üblicherweise Oligo-Nukleotide, die mit einem “Tintenstrahldrucker” auf Schichten (Agilent) gedruckt wurden oder *in situ* synthetisiert wurden (*Affymetrix*) wurden.

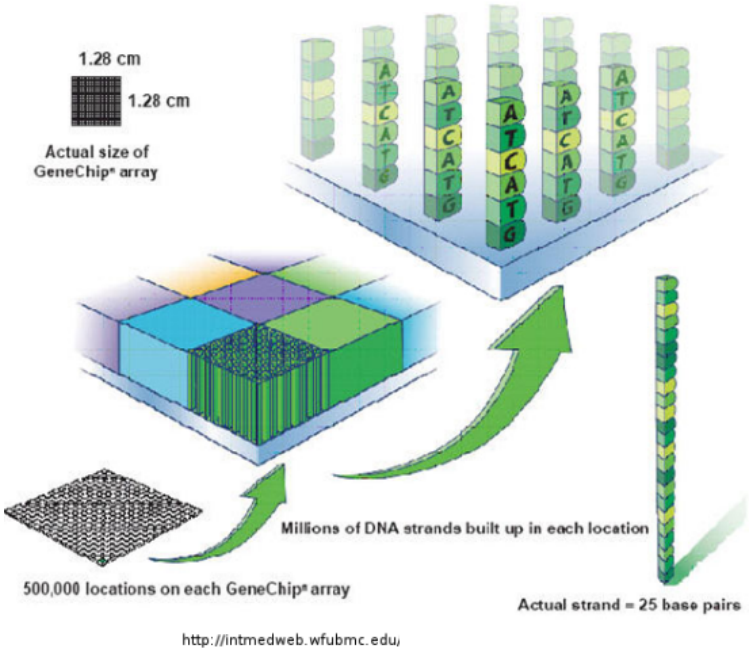
Gelabelte einzelsträngige DNA oder antisense *RNA* Fragmente aus einer Probe werden an den DNA-Microarray **hybridisiert**.

Die Menge an Hybridisierung für eine bestimmte Probe ist **proportional** zur Menge an Nukleotid-Fragmenten in der Probe.

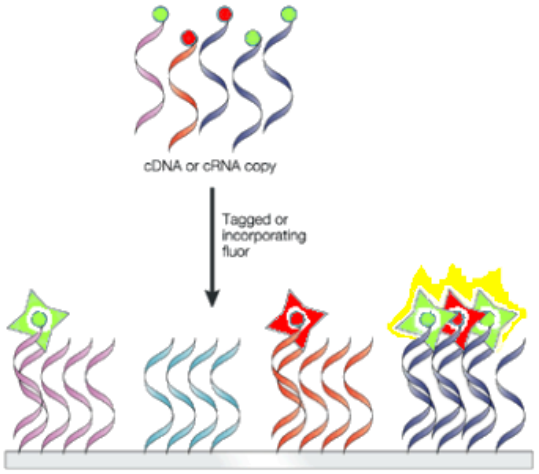
http://www.ebi.ac.uk/training/online/course/
functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

Experimentelles Vorgehen

Aufbringen eines zellulären cDNA-Gemischs auf die einzelnen Zellen des Arrays.



Jede Zelle enthält eine komplementäre Probe für ein Gen, die an die Oberfläche funktionalisiert wurde (typisch 45-60 nt lang).



changed from:
A. Butte, Nature Reviews Drug Discovery 1, 951-960, 2002

Jede **Zelle** misst daher die Expression eines **einzelnen Gens**.

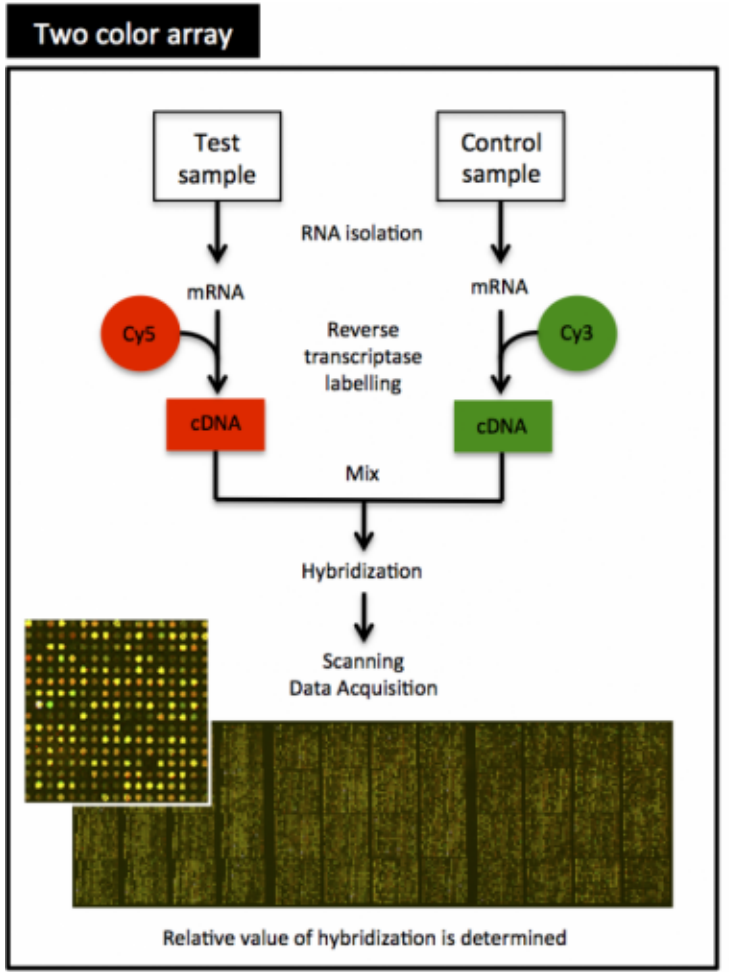
pgrc.ipk-gatersleben.de

2-Farben Microarrays

In 2-Farben Microarrays werden 2 biologische Proben mit zwei verschiedenen Fluoreszenz-farbstoffen **gelabelt**, üblicherweise Cyanin 3 (Cy3) und Cyanin 5 (Cy5).

Gleiche Mengen an gelabelter cDNA warden dann gleichzeitig auf denselben Microarray-Chip **hybridisiert**.

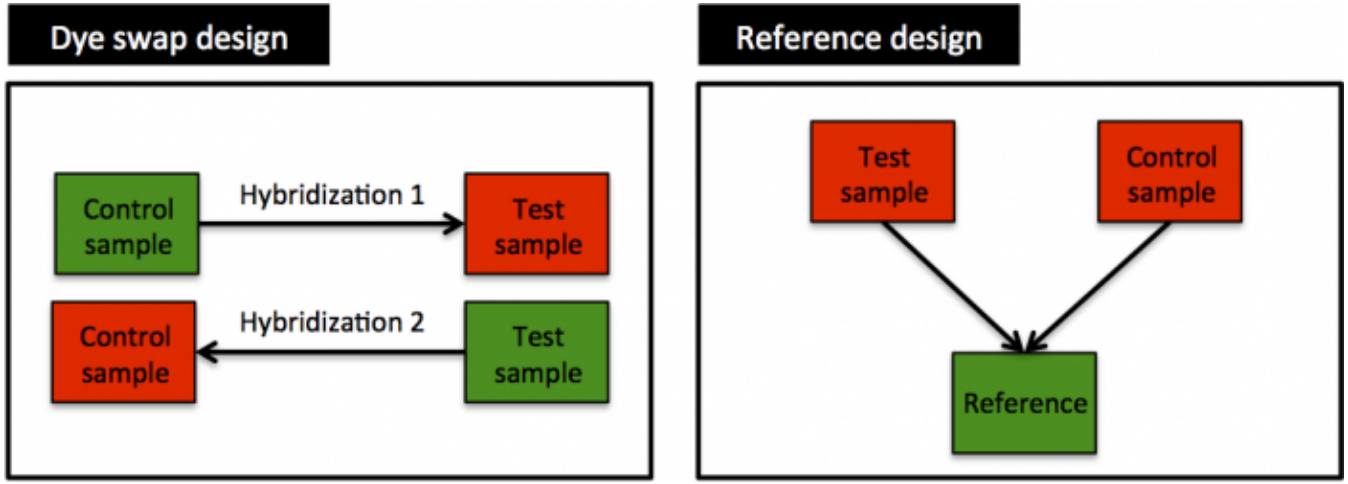
Dann wird die Fluoreszenz für jeden Farbstoff separat gemessen.
Dies repräsentiert die Menge jedes Gens in der Testprobe (Cy5) relativ zur Kontrollprobe (Cy3).



http://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

Bias-Korrektur

Bei Zweifarben-Microarrays können aufgrund der etwas unterschiedlichen **Photochemie** der beiden Farbstoffe Verschiebungen (Biases) auftreten.
Dieser Effekt kann mit 2 unterschiedlichen Methoden **korrigiert** werden.



In einem **Farbstoff-Austausch-Design** werden beide Proben zweimal miteinander verglichen, wobei die Zuordnung der Farbstoffe bei der zweiten Hybridisierung vertauscht wird.

Am häufigsten verwendet man das **Referenzdesign**, wo jede experimentelle Probe gegen eine einheitliche Referenzprobe hybridisiert wird.

http://www.ebi.ac.uk/training/online/course/
functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

Einstellung des Gleichgewichts

Die Gesamtzahl an gebundenen DNA-Strängen zu einer Zeit t sei $\boldsymbol{n_c(t)}$.

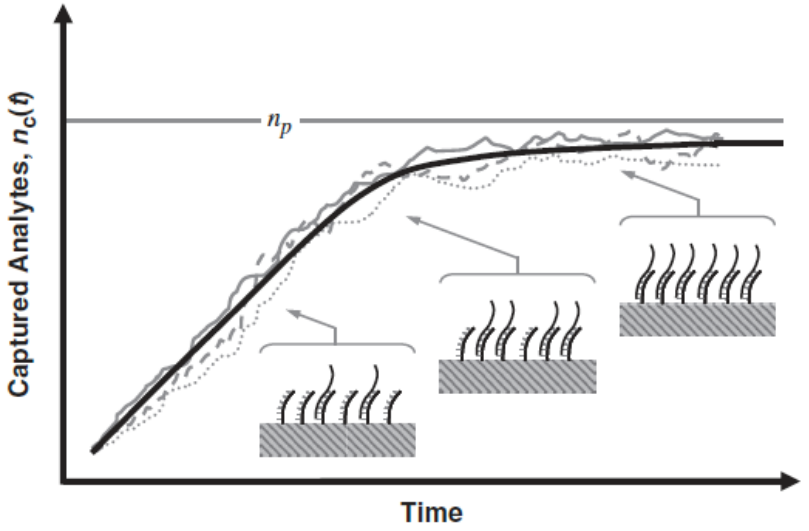
Dann kann man den erwarteten Mittelwert $\langle n_c(t) \rangle$ nach dieser Zeit t durch eine Ratengleichung ausdrücken:

$$\frac{\mathrm{d}\langle n_c(t)\rangle}{\mathrm{d}t} = k_1^* \left(\frac{n_p - \langle n_c(t)\rangle}{n_p} \right) (n_t - \langle n_c(t)\rangle) - k_{-1} \langle n_c(t)\rangle .$$

k_1^* und k_{-1} : Assoziations- und Dissoziationsraten, mit der die DNA-Stränge der Probe an den Microarray binden,

n_p : Gesamtzahl an Bindungsplätzen auf der Microarray-Oberfläche

n_t : Gesamtzahl an DNA-Strängen in der Probe



Einstellung des Gleichgewichts muss im MA-Experiment abgewartet werden!

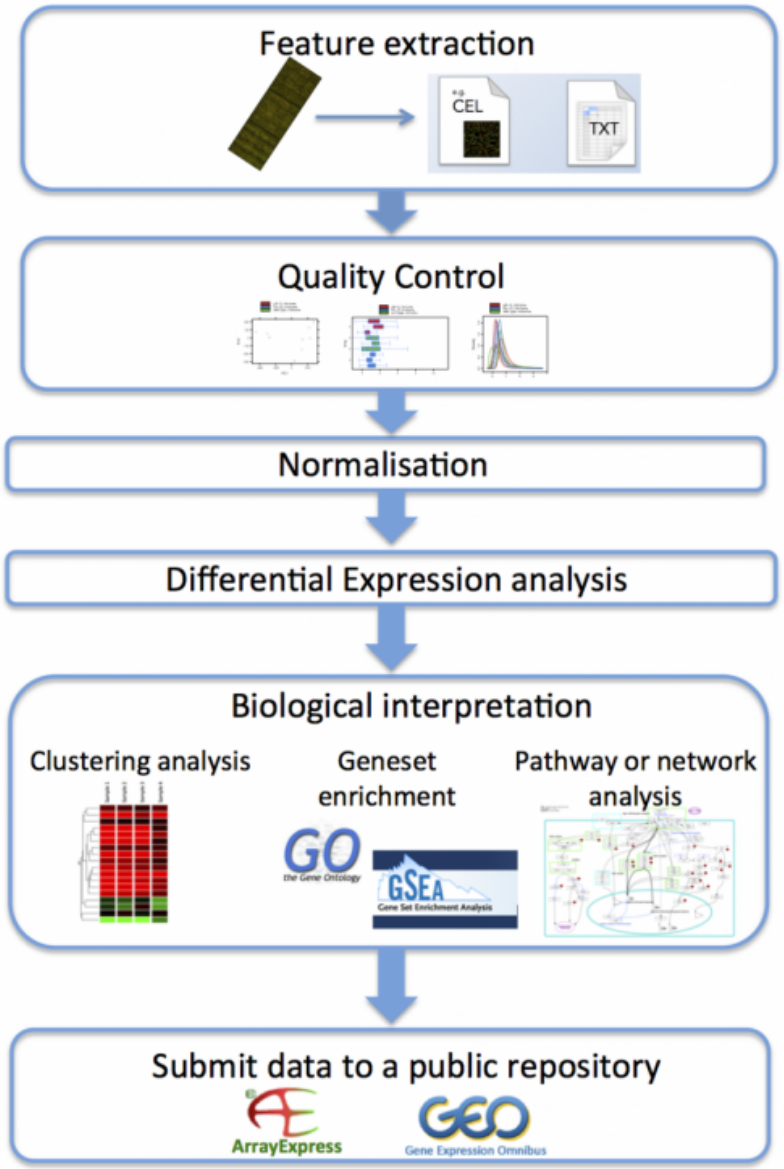
Hassibi et al., Nucl. Ac. Res. 37, e132 (2009)

Analyse von Microarray-Daten: workflow

Microarrays können für sehr unterschiedliche Experimente benutzt werden, z.B.

- Messung der Genexpression
- Messung der Translation
- Genotypisierung,
- Epigenetik.

Genexpression profiling ist die weitaus häufigste Anwendung.

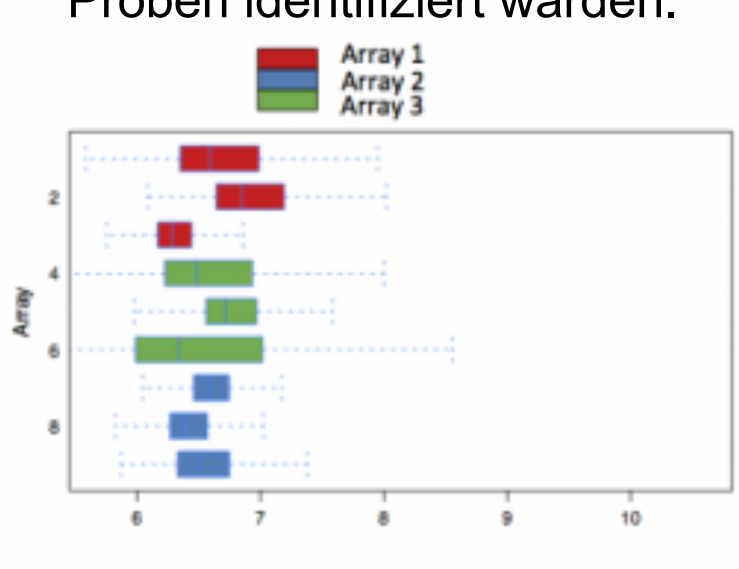


http://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

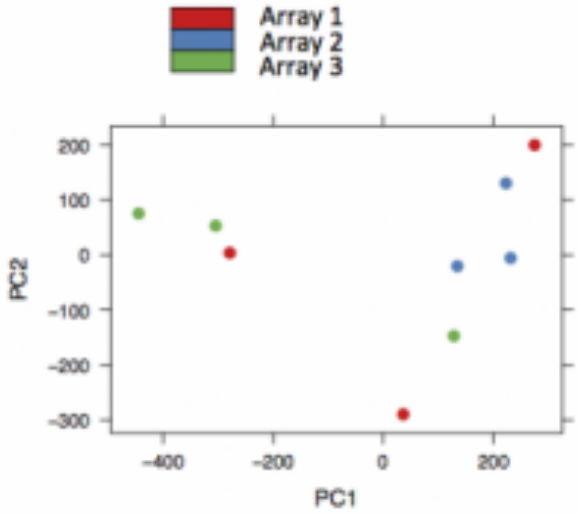
Qualitätskontrolle (QC)

QC von Microarray-Daten beginnt mit der **visuellen Überprüfung** der eingescannten Microarray-Bilder um sicherzustellen, dass es keine offensichtlichen Kratzer oder leere Regionen gibt.

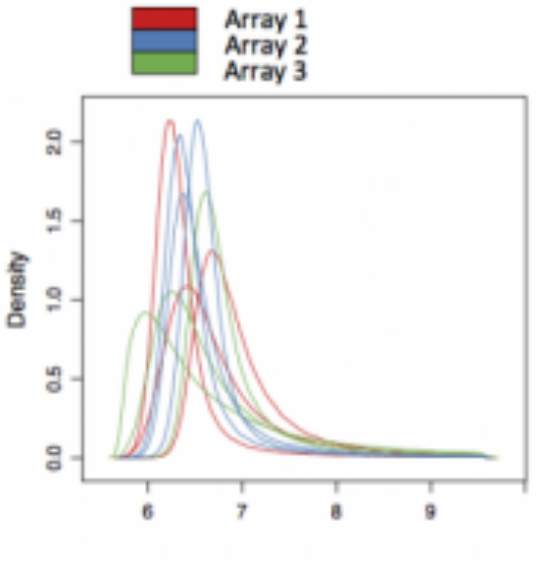
Datenanalyse-Programmpakete produzieren dann verschiedene diagnostische Plots, z.B. des Hintergrundsignals, der mittleren Intensitäten sowie wieviele Gene über dem Hintergrundsignal liegen. Dadurch können problematische Arrays und Proben identifiziert warden.



Box plot



PCA



Density plot

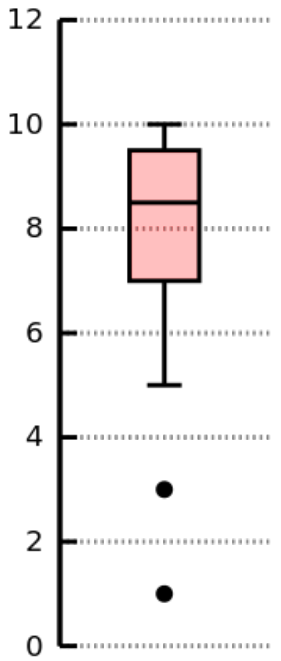
http://www.ebi.ac.uk/training/online/course/
functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

Boxplot

Die Boxplot-Darstellung erlaubt es, schnell einen Überblick über die Werteverteilung in einem Datensatz zu erhalten. Beispiel:

Datenpunkt	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Wert (unsortiert)	9	6	7	7	3	9	10	1	8	7	9	9	8	10	5	10	10	9	10	8
Wert (sortiert)	1	3	5	6	7	7	7	8	8	8	9	9	9	9	9	10	10	10	10	10

Kennwert	**Beschreibung**	**Lage im Boxplot**
Minimum	Kleinster Datenwert des Datensatzes	Ende eines Whiskers oder entferntester Ausreißer
Unteres Quartil	Die kleinsten 25% der Datenwerte sind kleiner oder gleich diesem Wert	Beginn der Box
Median	Die kleinsten 50% der Datenwerte sind kleiner oder gleich diesem Kennwert	Strich innerhalb dieser Box
Oberes Quartil	Die kleinsten 75% der Datenwerte sind kleiner oder gleich diesem Kennwert	Ende der Box
Maximum	Größter Datenwert des Datensatzes	Ende eines Whiskers oder entferntester Ausreißer



www.wikipedia.de

PCA- intro

PCA analysiert eine Datenmatrix **X** für Werte aus Beobachtungen, die durch mehrere abhängige Variablen beschrieben werden und die üblicherweise miteinander korreliert sind.

Das Ziel der PCA ist es, wichtige Informationen aus der Datenmatrix zu extrahieren und diese Information mit Hilfe einer Menge an orthogonalen Variablen, den **principal components** (Hauptkomponenten) darzustellen.

Wir betrachten eine Datenmatrix **X** für I Beobachtungen und J Variablen.

Ihre Elemente sind x_{ij}.

Die Matrix **X** hat den Rang L, wobei $L \leq \min [I,J]$.

PCA- Präprozessierung der Werte

Üblicherweise werden die Einträge der Matrix vor der PCA-Analyse präprozessiert.

Die Spalten von **X** werden **zentriert**, so dass der **Mittelwert** jeder Spalte 0 ist:

$$x_{ij} \rightarrow x_{ij} - \mu_j$$

(Fall 1) Wenn zusätzlich jedes Feld von **X** durch $\sqrt{I}$ oder $\sqrt{I-1}$ geteilt wird,
wird die Matrix $\Sigma = \mathbf{X}^T\mathbf{X}$ zu einer Kovarianzmatrix,

$$\Sigma = \left[(\mathbf{X} - \mu)^T(\mathbf{X} - \mu)\right]$$

Man nennt die Analyse dann **Kovarianz-PCA**.

PCA- preprocessing data entries

(Fall 2) Wenn die Variablen verschiedene Einheiten haben, ist es üblich, die Variablen (nach der Zentrierung) stattdessen zu **standardisieren**.

Dazu teilt man jede Variable durch ihre Norm $\sqrt{\frac{1}{n}\sum_i (x_i)^2}$.
Dies entspricht der Division durch die Standardabweichung der Variable (ausser dass durch *n* stattdurch *n*-1 geteilt wird).

In diesem Fall nennt man die Analyse **Korrelations-PCA**, da die Matrix $\mathbf{X}^T\mathbf{X}$ nun eine Korrelationsmatrix ist.

Wir benutzen nun die Tatsache, dass die Matrix **X** eine **singular value decomposition (SVD,** Singulärwertzerlegung**)** besitzt:

$$\mathbf{X} = \mathbf{P}\Delta\mathbf{Q}^T$$

Was ist eine SVD?

Singular Value Decomposition (SVD)

SVD zerlegt eine rechteckige Matrix **X** in drei einfache Matrizen:
zwei orthogonale Matrizen **P** und **Q** und eine Diagonalmatrix Δ.

$$\mathbf{X} = \mathbf{P}\Delta\mathbf{Q}^T$$

P : enthält die normierten Eigenvektoren der Matrix $\mathbf{X}\,\mathbf{X}^T$. (d.h. $\mathbf{P}^T\mathbf{P} = \mathbf{1}$)
Die Spalten von **P** nennt man *linke singulare Vektoren* von **X**.

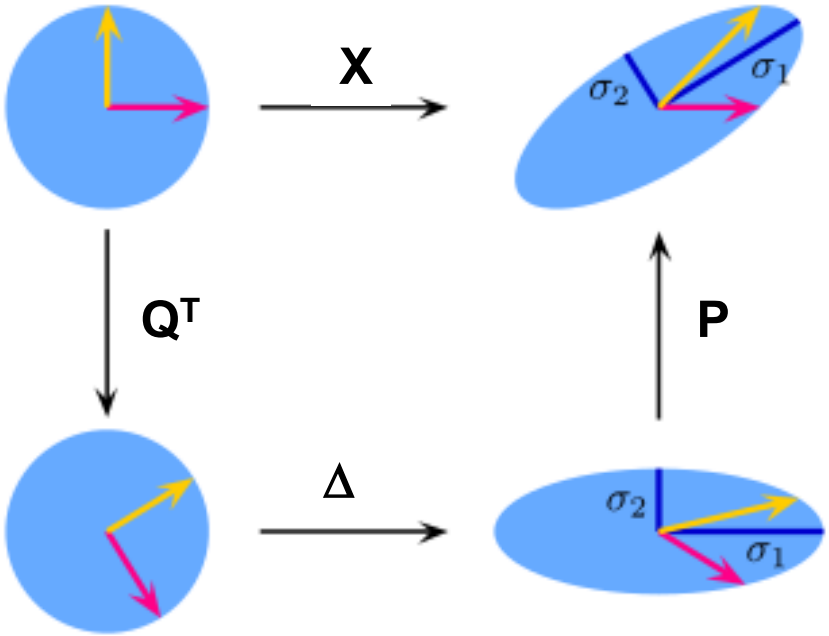
Q : enthält die normierten Eigenvektoren der Matrix $\boldsymbol{X}^T\mathbf{X}$. (d.h. $\mathbf{Q}^T\mathbf{Q} = \mathbf{1}$)
Die Spalten von **Q** nennt man *rechte singulare Vektoren* von **X**.

Δ : ist die Diagonalmatrix der *singulären Werte*. Diese sind die Quadratwurzeln der Eigenwerte der Matrix $\mathbf{X}\,\boldsymbol{X}^T$ (entsprechen denen von $\boldsymbol{X}^T\mathbf{X}$).

Interpretation der SVD

In dem (gebräuchlichen) Spezialfall, dass **X** eine $m \times m$ reelle Quadratmatrix mit positiver Determinante ist, sind **P**, **Q**, und Δ ebenfalls reelle $m \times m$ Matrizen.

Δ kann dann als Skalierungsmatrix aufgefasst werden und **P** und **Q** als Rotationsmatrizen.



$$X = P \Delta Q^T$$

www.wikipedia.org

Ziele der PCA

(1) Extrahiere die wichtigsten Informationen aus der Datenmatrix

→ PC1 soll die Richtung beschreiben entlang welcher die Daten die größte Varianz enthalten, PC2 ist orthogonal zu PC1 und beschreibt die Richtung der größten verbleibenden Varianz etc

(2) Komprimiere und vereinfache den Datensatz auf diese wichtigen Informationen.

(3) Analysiere die Struktur der Beobachtungen und Variablen.

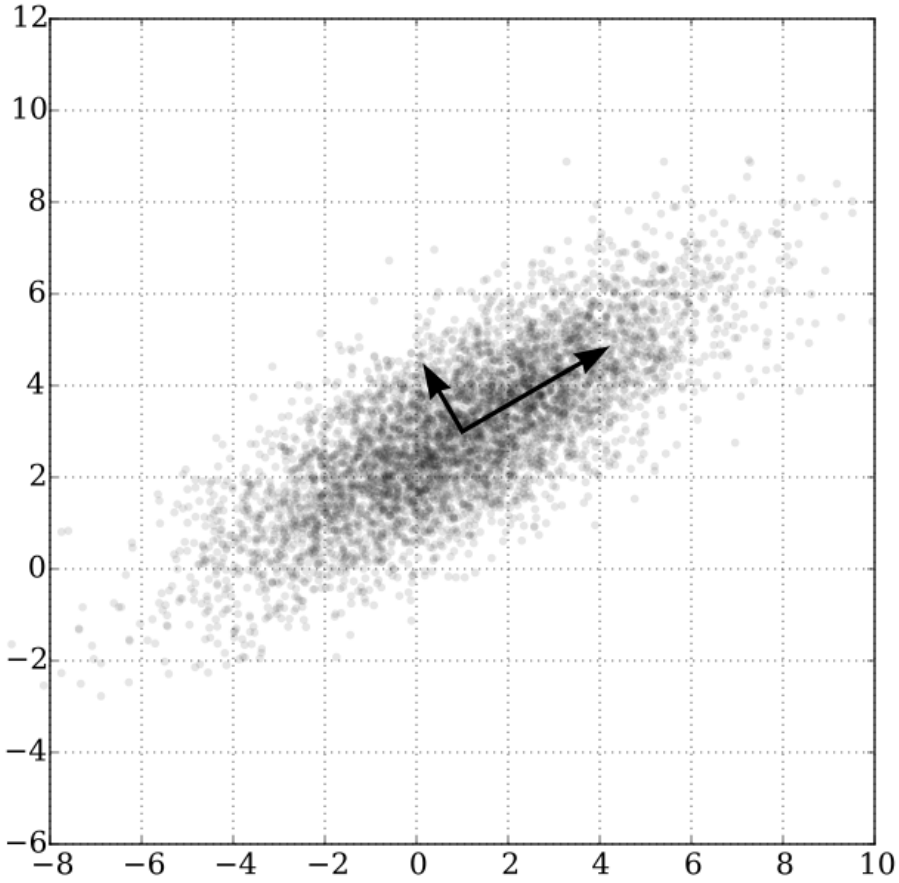
Um diese Ziele zu erreichen, konstruiert PCA neue Variablen – principal components (PCs) – als lineare Kombinationen der Originalvariablen.

PC1 ist der Eigenvektor von $\mathbf{X}^T\mathbf{X}$ mit dem größten Eigenwert usw.

PCA Beispiel

PCA einer multivariaten Gauß-Verteilung **X,** die bei (1,3) zentriert ist und entlang der Richtung (0.866, 0.5) eine Standardabweichung von 3 hat und $\sigma = 1$ in die dazu orthogonale Richtung.

Die zwei eingezeichneten PCA Vektoren sind die Eigenvektoren der Kovarianzmatrix $\mathbf{X}^T\mathbf{X}$, die mit den Quadratwurzeln der zugehörigen Eigenwerte skaliert wurden und verschoben wurden, so dass ihr Endpunkt auf dem Mittelwert liegt.



Note that shown here is the data along the original coordinates. In a PCA plot, the data is projected onto two PCs, usually PC1 and PC2.

www.wikipedia.org

Konstruktion der PC-Vektoren

Die Hauptkomponenten enthält man aus der SVD von **X**,

$$\mathbf{X} = \mathbf{P}\Delta\mathbf{Q}^T$$

Q enthält die principal components (normierte Eigenvektoren von $\mathbf{X}^T\mathbf{X}$).

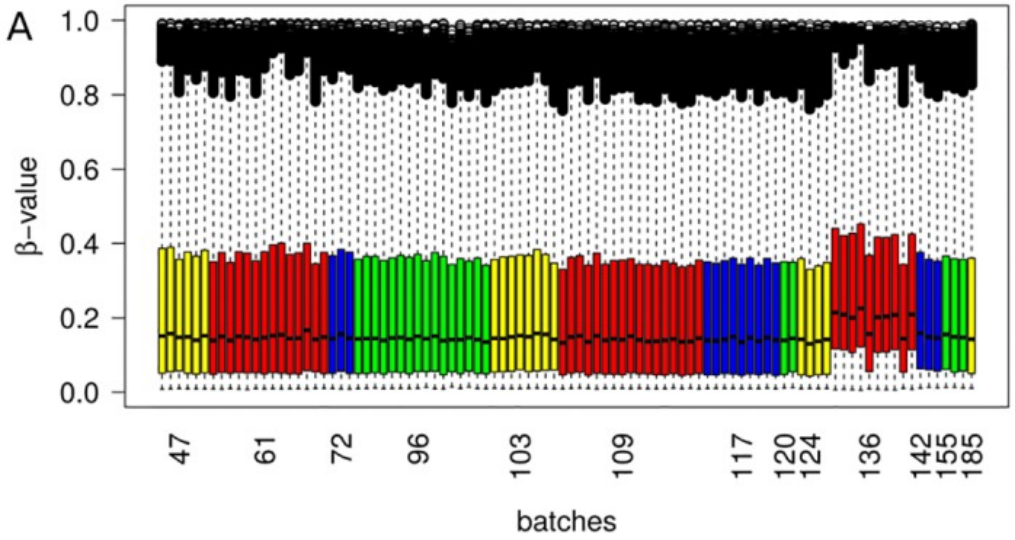
Die *I* x *L* Matrix der **Faktoren F** enthält man durch

$$\mathbf{F} = \mathbf{P}\Delta = \mathbf{P}\Delta\mathbf{Q}^T\mathbf{Q} = \mathbf{X}\mathbf{Q}$$

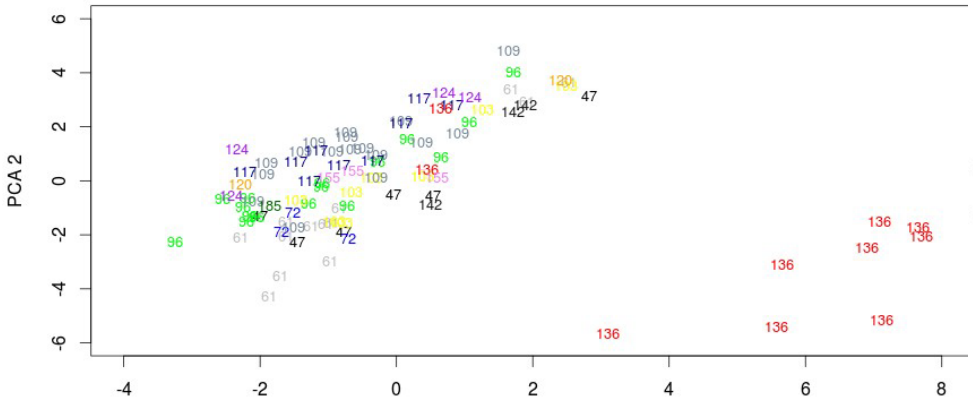
F kann daher als eine **Projektionsmatrix** interpretiert werden.
Die Multiplikation von **X** mit **Q** entspricht der Projektion der Beobachtungen **X** auf die principal components **Q**.

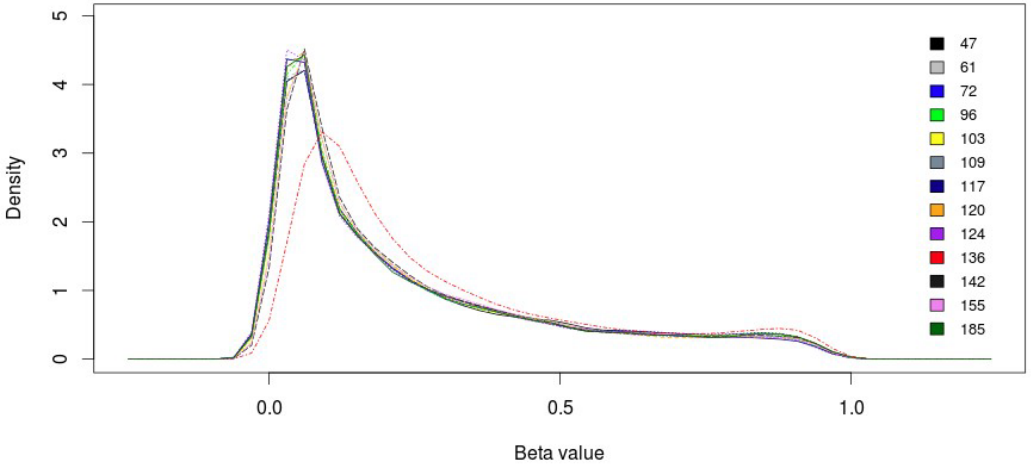
Ausreißer-Datenpunkte?

Datensatz 136 in diesen DNA-Methylierungsdaten (Boxplot-Darstellung) verhält sich anders als die anderen Datensätze.



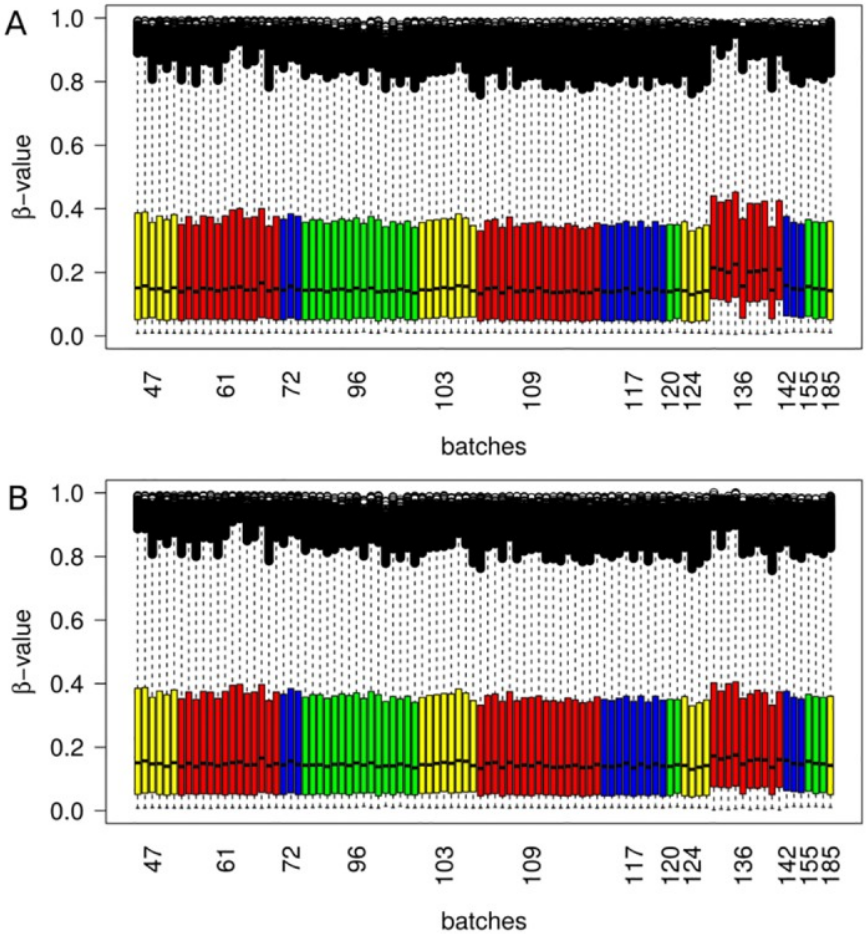
Dies sieht man auch im PCA-Plot (unten links) bzw. im Plot der Werteverteilung (unten rechts).





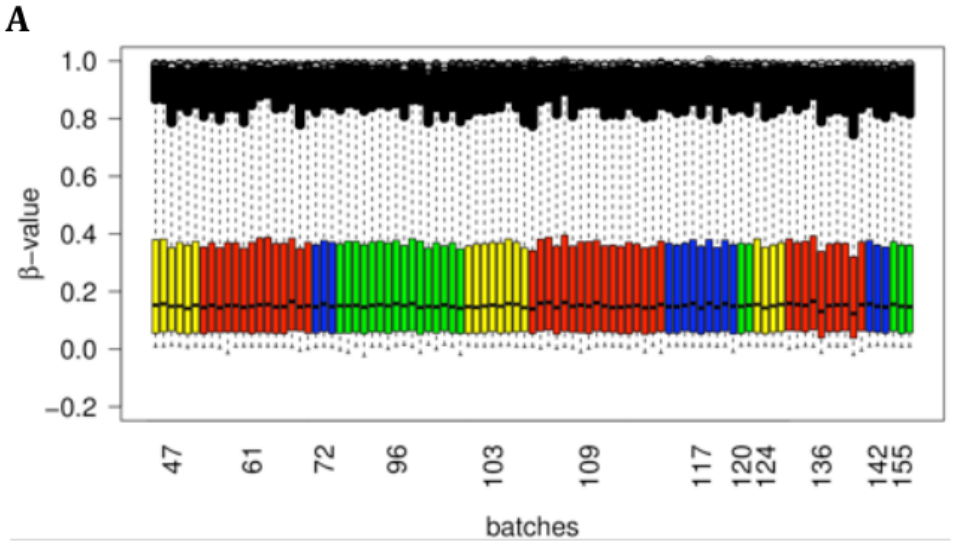
PCA: principle component analysis;
Projektion der Daten auf PC1 und PC2

Korrektur von Ausreißer-Datenpunkten



(Bild links oben): Anteil von methylierten CpG-Basen in verschiedenen Samples. Sample 136 ist Ausreißer.
(unten) Korrektur mit unserem Tool BEclear: Nur stark abweichende Werte werden korrigiert: diese Werte werden aus den Werten benachbarter Datenpunkte vorhergesagt. Effekt: natürliche Variation bleibt erhalten.

Akulenko, Merl, Helms (2016)
PloS ONE 11: e0159921



(Bild rechts) Batch-Effekt-Korrektur desselben Datensatzes mit Tool ComBat: Natürliche Variation der Werte wird stark „geglättet“; alle Werte werden geändert.

Normalisierung

Mit Normalisierungsverfahren **kontrolliert** man die **technische Variation** zwischen einzelnen Assays, wobei die **biologische Variation** erhalten bleibt.

Es gibt viele Verfahren zur Normalisierung der Daten, abhängig von :
- dem verwendeten Array;
- dem Design des Experiments;
- Annahme über die Verteilung der Daten;
- der verwendeten Software.

Für den **Expression Atlas** am EBI werden Affymetrix-Microarray Daten mit der 'Robust Multi-Array Average' (RMA) Methode im 'oligo‘ Programm normalisiert.

Agilent-Microarray-Daten warden mit dem 'limma' Programm normalisiert:
'quantile normalisierung' für Ein-Farben Microarray-Daten;
'Loess normalisierung für Zwei-Farben Microarray-Daten.

http://www.ebi.ac.uk/training/online/course/
functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

Quantile Normalisierung

Gegeben: 3 Messungen von 4 Variablen A – D.

Ziel: alle Messungen sollen eine identische Werte-Verteilung bekommen

A	5	4	3
B	2	1	4
C	3	4	6
D	4	2	8

Originaldaten

→

A	iv	iii	i
B	i	i	ii
C	ii	iii	iii
D	iii	ii	iv

Bestimme in jeder Spalte den Rang jedes Wertes

A	2	Rang i
B	3	Rang ii
C	4.67	Rang iii
D	5.67	Rang iv

Bilde Mittelwert jeder Reihe

→

A	2	1	3
B	3	2	4
C	4	4	6
D	5	4	8

Ordne jede Spalte nach Größe

A	5.67	4.67	2
B	2	2	3
C	3	4.67	4.67
D	4.67	3	5.67

Ersetze die Originalwerte durch die Mittelwerte entsprechend dem Rang des Datenfeldes.
Nun enthalten alle Spalte dieselben Werte (bis auf doppelte Datenpunkte) und können leicht miteinander verglichen werden.

Expressionsverhältnis

Der relative Expressions-Wert eines Gens kann als Menge an rotem oder grünen Licht gemessen werden, die nach Anregung ausgestrahlt wird.

Man drückt diese Information meist als **Expressionsverhältnis** T_k aus:

$$T_k = \frac{R_k}{G_k}$$

Für jedes Gen *k* auf dem Array ist hier R_k der Wert für die Spot-Intensität für die Test-Probe und G_k ist die Spot-Intensität für die Referenz-Probe.

Man kann entweder absolute Intensitätswerte verwenden, oder solche, die um den mittleren Hintergrund (Median) korrigiert wurden (siehe vorige Folie).

In letzterem Fall lautet das Expressionsverhältnis für einen Spot:

$$T_{median} = \frac{R^{spot}_{median} - R^{background}_{median}}{G^{spot}_{median} - G^{background}_{median}}$$

M. Madan Babu, An Introduction to Microarray Data Analysis

Bereich der Expressionsverhältnisse

Das Expressionsverhältnis (**fold change**) stellt auf intuitive Art die Änderung von Expressions-Werten dar. Gene, für die sich nichts ändert, erhalten den Wert 1.

Allerdings ist die Darstellung von Hoch- und Runterregulation nicht balanciert.

Wenn ein Gen um den Faktor 4 hochreguliert ist, ergibt sich ein Verhältnis von 4.

$$R/G = 4G/G = 4$$

Wenn ein Gen jedoch um den Faktor 4 runterreguliert ist, ist das Verhältnis 0.25.

$$R/G = R/4R = 1/4.$$

D.h. Hochregulation wird aufgebläht und nimmt Werte zwischen 1 und unendlich an, während die Runterregulation komprimiert wird und lediglich Werte zwischen 0 und 1 annimmt.

M. Madan Babu, An Introduction to Microarray Data Analysis

Logarithmische Transformation

Eine bessere Methode zur Transformation ist, den Logarithmus zur Basis 2 zu verwenden.

d.h. $\log_2$(Expressionsverhältnis)

Dies hat den großen Vorteil, dass Hochregulation und Runterregulation gleich behandelt werden und auf ein kontinuierliches Intervall abgebildet werden.

Für ein Expressionsverhältnis von 1 ist $\log_2(1) = 0$, das keine Änderung bedeutet.

Für ein Expressionsverhältnis von 4 ist $\log_2(4) = 2$,

für ein Expressionsverhältnis von 1/4 ist $\log_2(1/4) = -2$.

Für die **logarithmierten Daten** ähneln die Expressionsraten dann oft einer **Normalverteilung** (Glockenkurve).

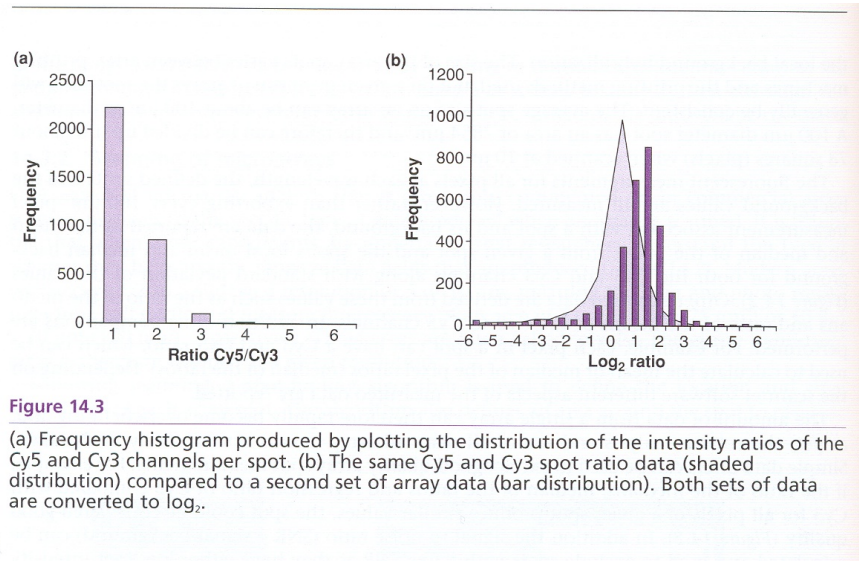


Figure 14.3

(a) Frequency histogram produced by plotting the distribution of the intensity ratios of the Cy5 and Cy3 channels per spot. (b) The same Cy5 and Cy3 spot ratio data (shaded distribution) compared to a second set of array data (bar distribution). Both sets of data are converted to $\log_2$.

M. Madan Babu, An Introduction to Microarray Data Analysis

Orengo-Buch

Daten-Interpretation von Expressionsdaten

Annahme:

Funktionell zusammenhängende Gene sind oft ko-exprimiert.

Z.B. sind in den 3 Situationen

X ➔	Y	(Transkriptionsfaktor X aktiviert Gen Y)
Y ➔	X	(Transkriptionsfaktor Y aktiviert Gen X)
Z ➔	X, Y	(Transkriptionsfaktor Z aktiviert Gene X und Y)

die Gene X und Y ko-exprimiert.

Durch Analyse der Ko-Expression (beide Gene an bzw. beide Gene aus) kann man also funktionelle Zusammenhänge im zellulären Netzwerk entschlüsseln.

Allerdings nicht die kausalen Zusammenhänge, welches Gen das andere reguliert.

4.a Hierarchisches Clustering zur Analyse von Ko-Expression

Man unterscheidet beim Clustering zwischen anhäufenden Verfahren (**agglomerative clustering**) und teilenden Verfahren (**divisive clustering**).

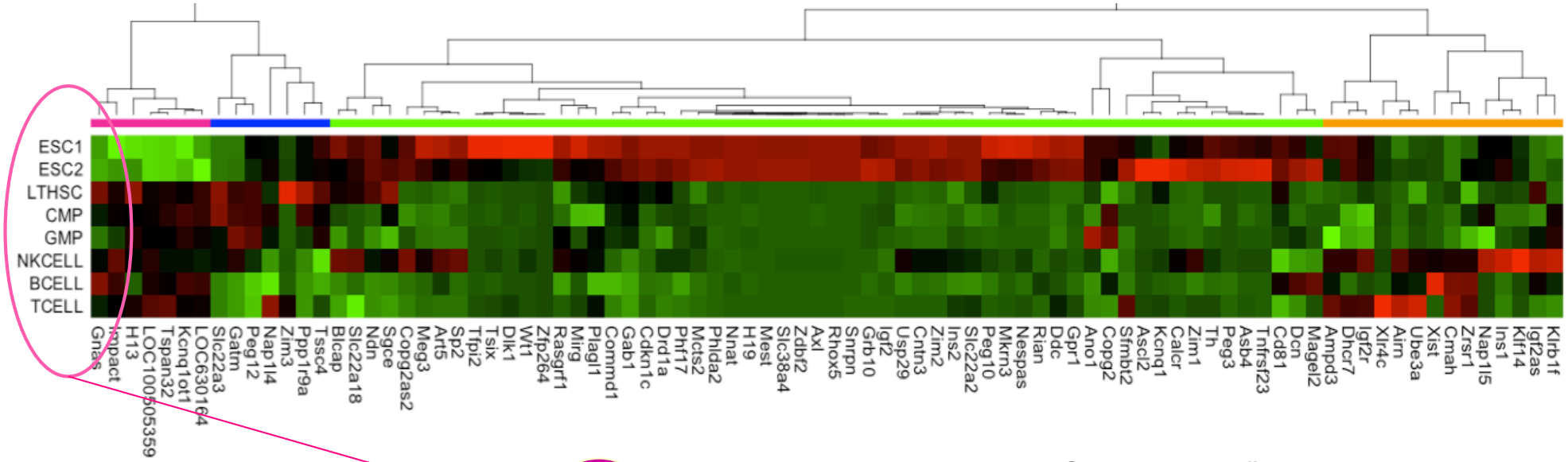
Bei den anhäufenden Verfahren, die in der Praxis häufiger eingesetzt werden, werden schrittweise einzelne Objekte zu Clustern und diese zu größeren Gruppen zusammengefasst, während bei den teilenden Verfahren größere Gruppen schrittweise immer feiner unterteilt werden.

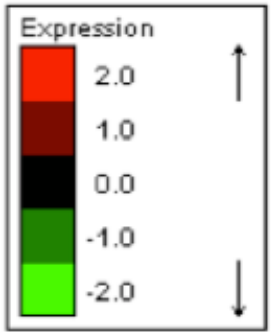
Beim Anhäufen der Cluster wird zunächst jedes Objekt als ein eigener Cluster mit einem Element aufgefasst.

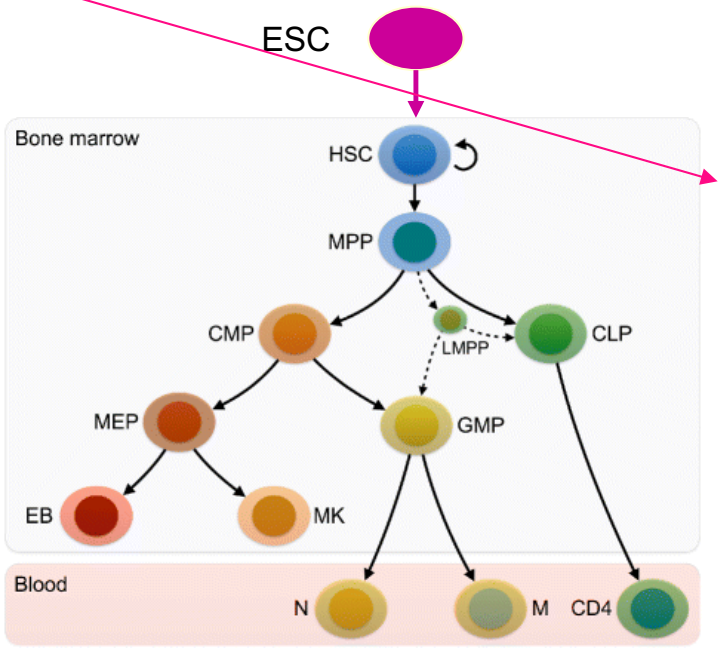
Nun werden in jedem Schritt die jeweils einander nächsten Cluster zu einem Cluster zusammengefasst.

Das Verfahren kann beendet werden, wenn alle Cluster eine bestimmte Distanz zueinander überschreiten oder wenn eine genügend kleine Zahl von Clustern ermittelt worden ist.

Hierarchisches Clustering







Frage: welche Gene zeigen ähnliche Expressions-Muster entlang der Entwicklung von Blutzellen?

ESC: embryonale Stammzelle
HSC: hämatopoetische Stammzelle
CMP: Common Myeloid Progenitor (Vorläufer)
GMP: Granulocyte-Macrophage-Progenitor

Clustering ergibt 4 Gruppen von Genen, die jeweils ähnliches Verhalten zeigen (pink, blau, grün, orange)

Hamed M et al. (2017)
PLoS ONE 12(1): e0166852.

k-means Clustern

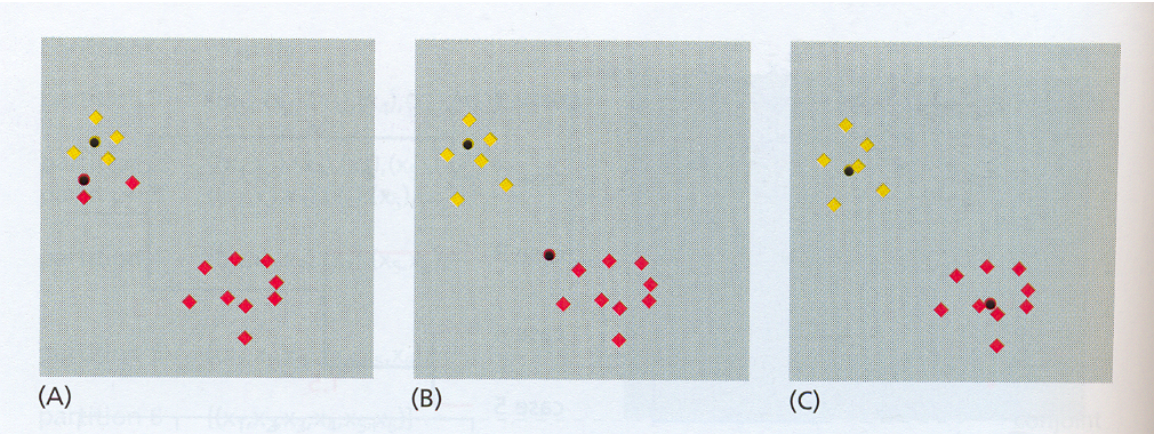
Ein Durchlauf der *k*-means Clustering Methode erzeugt eine Auftrennung der Datenpunkte in *k* Cluster. Gewöhnlich wird der Wert von *k* vorgegeben.

Zu Beginn wählt der Algorithmus *k* Datenpunkte als Centroide der *k* Cluster. Anschließend wird jeder weitere Datenpunkt dem nächsten Cluster zugeordnet.

Nachdem alle Datenpunkte eingeteilt wurden, wird für jedes Cluster das Centroid als Schwerpunkt der in ihm enthaltenen Punkte neu berechnet.

Diese Prozedur (Auswahl der Centroide - Datenpunkte zuordnen) wird so lange wiederholt bis die Mitgliedschaft aller Cluster stabil bleibt.

Dann stoppt der Algorithmus.



4.b Abschätzung der Signifikanz

Cancer Research

AACR

Lipid Metabolism Signatures in NASH-Associated HCC—Letter

Sonja M. Kessler, Stephan Laggai, Ahmad Barghash, et al.

Cancer Res Published OnlineFirst April 28, 2014.

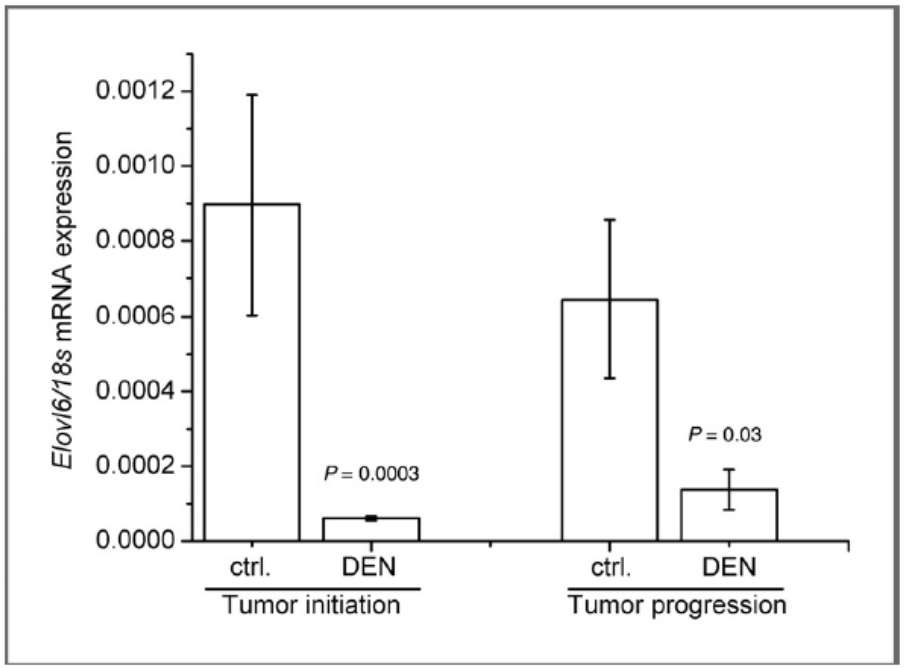


Figure 2. Wild-type mice were treated with the carcinogen DEN at the age of 2 weeks. Livers were analyzed after 24 weeks to assess the tumor initiation state. Analyses in the tumor progression stadium were done after 36 weeks. *Elovl6* mRNA expression as determined by real-time reverse transcriptase PCR with $n = 8–18$ per group. Data were normalized to *18S*. Statistical differences compared with untreated animals of the same age (ctrl.) were calculated by Mann–Whitney *U* test.

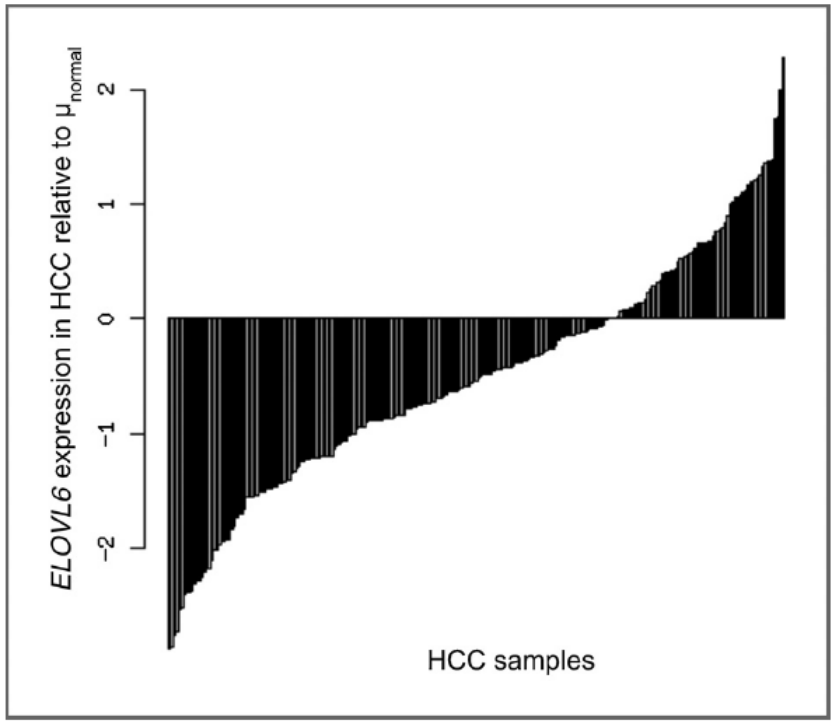


Figure 1. mRNA levels of *ELOVL6* in 247 human HCC samples relative to the mean of 239 nontumor liver tissue (μ_{normal}). Samples of dataset GSE14520 [$\log_2$ (expression) values from GEO after Robust Multi-array Average normalization] were mapped to hgu133a.db using bioconductor. Significance values: $P = 3.8E{-}11$, Kolmogorov–Smirnov test; $P = 6.7E{-}11$, *t* test; 5.1E−11, Mann–Whitney *U* test.

Differentielle Expressionsanalyse: Fold change

Die einfachste Methode um differenziell exprimierte (DE) Gene zu identifizieren ist, das **log Verhältnis** zwischen zwei Bedingungen zu bilden (oder das mittlere Verhältnis, wenn es Replikate gibt).

Alle Gene, die sich stärker als ein willkürlicher **cut-off value** unterscheiden, werden als differentiell exprimiert angesehen.

Ein typischer cut-off Wert kann **zweifacher** (**two-fold)** Unterschied zwischen den beiden Bedingungen sein.

Dieser **'fold' change** Test ist jedoch kein statistischer Test.

→ man kann damit den **Konfidenzlevel** nicht bewerten, ob diese Gene wirklich differentiell exprimiert sind oder nicht.

Cui & Churchill, Genome Biol. 2003; 4(4): 210.

Differentielle Expressionsanalyse: *t*-test

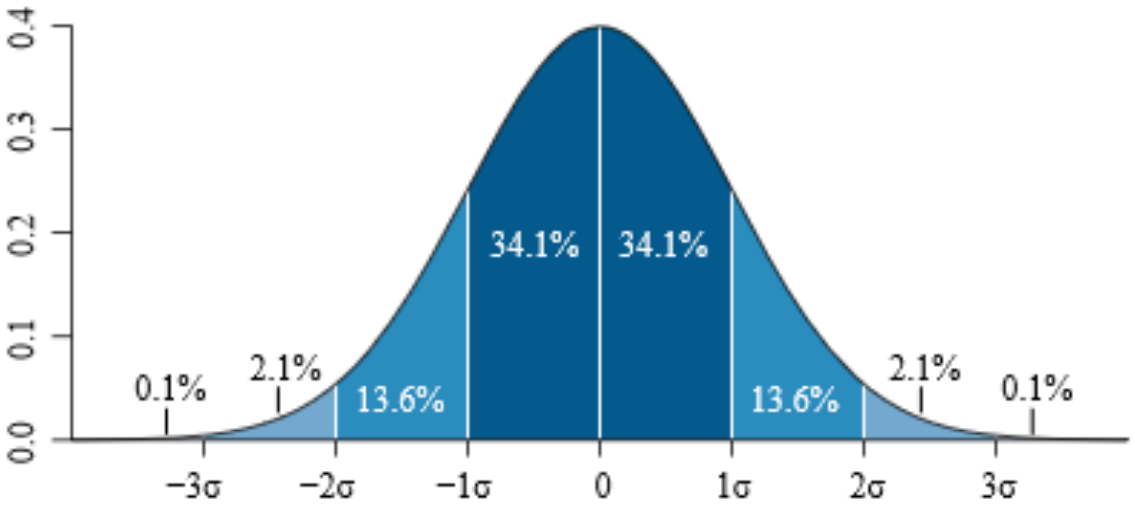
Der *t* Test ist eine einfache statistische Methode um DE-Gene zu identifizieren.

R_g : mittleres log Verhältnis der Expressionslevel für ein Gen *g* = “der Effekt”
SE : Standardfehler (erhalten durch Kombination der Daten für alle Gene = “die Variation in den Daten”)

Globale t-test Statistik : $t = \frac{R_g}{SE}$

Standardfehler: Standardabweichung der gesampelten Verteilung einer Statistik.

Falls ein Wert mit einem normal-verteilten Fehler gesampelt wird, zeigt die Abb. den Anteil an Proben, die 0, 1, 2, und 3 Standardabweichungen oberhalb und unterhalb des tatsächlichen Werts liegen.



Cui & Churchill, Genome Biol. 2003; 4(4): 210;
www.wikipedia.org (M.M. Thoews)

Differentielle Expressionsanalyse: t-test

SE_g : Standardfehler eines Gens *g* (aus Replikat-Experimenten)

Gen-spezifische T-test Statistik: $t = \frac{R_g}{SE_g}$

Falls Replikat-Experimente vorliegen, kann man daraus SE_g für jedes Gen berechnen und den *t*-Test durchführen.

Mit der resultierenden **Gen-spezifischen *t*-Statistik** kann man DE-Gene bestimmen.

Vorteil: Mit diesem Verfahren vermeidet man die unterschiedliche Varianz einzelner Gene. Man nutzt jedes Mal nur die Information für ein Gen.

Nachteil: Allerdings kann das Verfahren geringe statistische Aussagekraft haben, da die Menge an Proben für jede Bedingung üblicherweise klein ist.

Falls die für ein Gen abgeschätzte Varianz aus Zufall sehr klein ist, ergeben sich große t-Werte auch dann, wenn der entsprechende fold change-Wert klein ist.

Cui & Churchill, Genome Biol. 2003; 4(4): 210.

Differentielle Expressionsanalyse: SAM

Falls nur wenige Proben vorliegen, ist die Abschätzung der Varianz der Gen-spezifischen t-Statistik schwierig. Es kann **erratische Fluktuationen** geben.

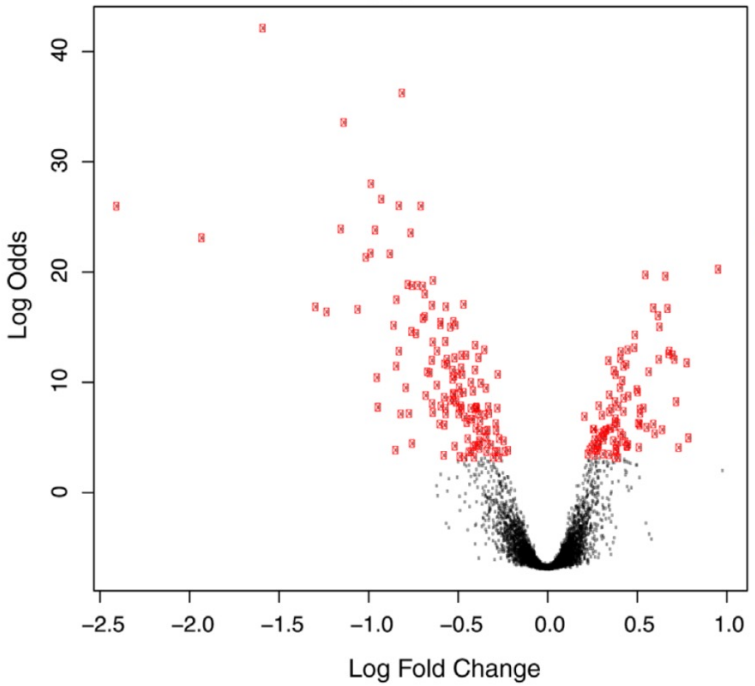
Die **'significance analysis of microarrays' (SAM)-**Methode ist eine Variante des *t* Tests. Dort addiert man eine kleine positive Konstante *c* im Zähler des Gen-spezifischen *t* Tests.

Significance analysis of microarrays (SAM): $S = \frac{R_g}{c+SE_g}$

Durch diese Modifikation werden Gene mit kleinen fold changes *(R_g)* nicht als signifikant ausgewählt.
Die SAM-Methode liefert daher deutlich robustere Ergebnisse.

Cui & Churchill, Genome Biol. 2003; 4(4): 210.

Limma Package: Volcano Plot



Der 'volcano plot' ist eine einfach interpretierbare Darstellung, die fold-change und *t*-test Kriterium zusammenfasst.

Jedes Symbol (hier: Kreuz) steht für ein Gen. Aufgetragen sind negative $\log_{10}$-transformierte *p*-Werte des Gen-spezifischen *t*-Tests gegen $\log_2$–transformierte old change Werte.

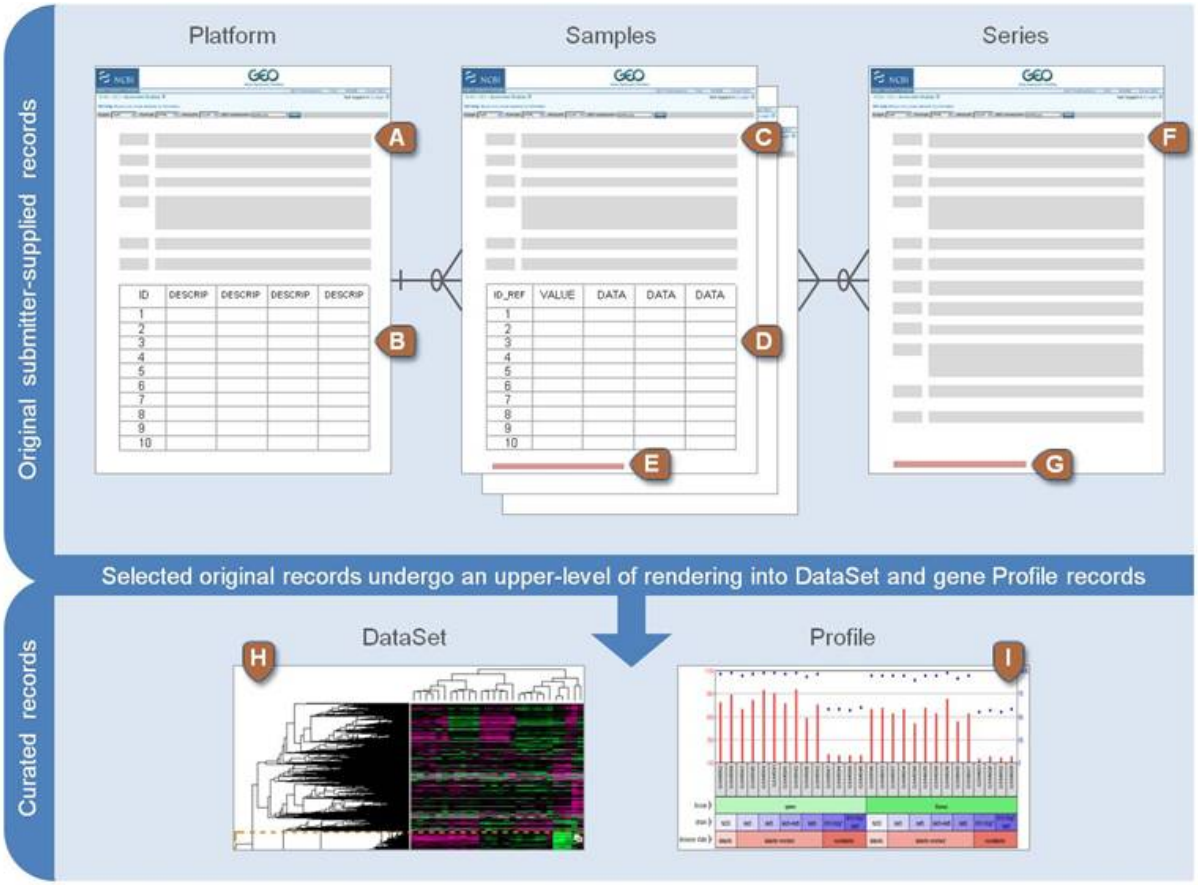
Gene mit einer statistisch signifikanten differentiellen Expression (gemäß dem Gen-spezifischen *t*-Test) liegen oberhalb einer horizontalen Schranke.

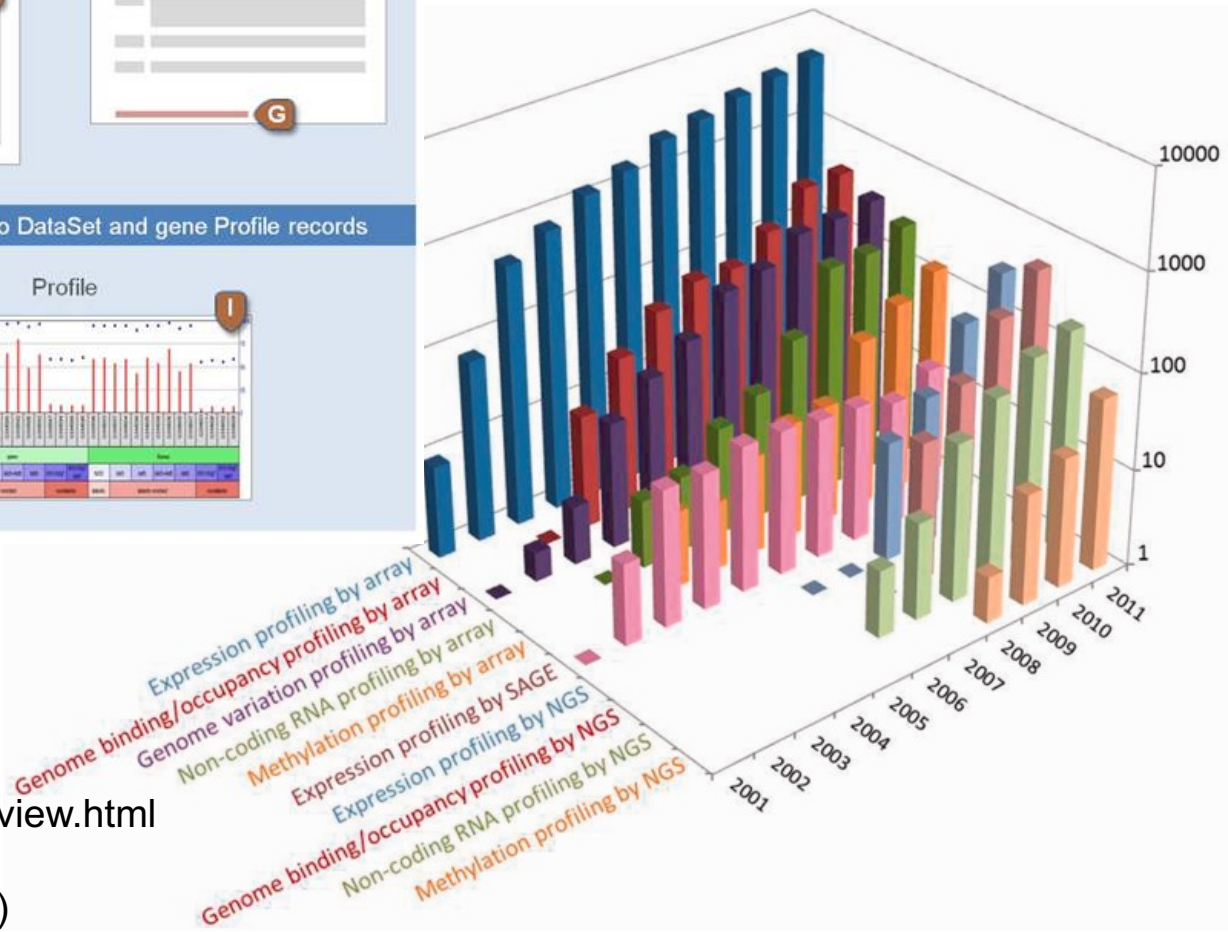
In dieser Abb ist dies der schwarz/rot-Übergang.

Gene mit einem großen fold-change Wert liegen außerhalb von vertikalen Schranken. Signifikante Gene liegen in den Regionen oben links bzw. oben rechts.

Rapaport et al. (2013) Genome Biol. 14: R95
Cui & Churchill, Genome Biol. 2003; 4(4): 210

GEO: Gene Expression Omnibus





http://www.ncbi.nlm.nih.gov/geo/info/overview.html

Nucleic Acids Res. 41, D991-D995 (2013)

Bewertung von Signifikanz: Mann Whitney Text

Im Gegensatz zum *t*-Test ist dies ist ein nicht-parametrischer Test. Die abhängige Variable muss NICHT normalverteilt sein.

Beispiel: durchschnittliche Noten der Schüler in 2 Schulklassen.

	Schulnoten											Median
Schulklasse A	4.2	6	4.5	4.9	3.9	5	3.6	4.7	5.5	4.3	4.6	4.6
Schulklasse B	4.8	5.8	5.9	4	5.4	3.5	3.8	3.7	5.3	4.4	4.1	4.4

Median : Schüler in Klasse A bessere Noten (Schweiz: 1 bis 6 (am besten).

Ist der Unterschied statistisch signifikant?

Bilde eine gemeinsame Rangreihe:

Schulnoten	3.5	3.6	3.7	3.8	3.9	4	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5	5.3	5.4	5.5	5.8	5.9	6
Schulklasse	B	A	B	B	A	B	B	A	A	B	A	A	A	B	A	A	B	B	A	B	B	A
Gemeinsamer Rangplatz	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22

Bei 2 Stichproben mit identischer zentraler Tendenz würden sich die Rangplätze der beiden Stichproben gleichmässig verteilen und z.B. folgende Muster ergeben:

ABABABABABAB oder AABBBBAA

www.methodenberatung.uzh.ch/datenanalyse/unterschiede/zentral/mann.html

Bewertung von Signifikanz: Mann Whitney Text

Die Teststatistik U überprüft nun die Gleichmässigkeit der Verteilung der Rangplätze in der gemeinsamen Rangreihe.

Für die erste Stichprobe (Schulklasse A) lautet die Teststatistik

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - T_1$$

mit n_k = Stichprobengrösse der Stichprobe k

T_1 = Rangsumme der Stichprobe 1

Entsprechend gilt für die zweite Stichprobe

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - T_2$$

Zwischen beiden Werten besteht folgender Zusammenhang $U_1 + U_2 = n_1 n_2$

Die Rangsumme T_1 für Schulklasse A ist die Summe aller Rangplätze von Werten für Schulklasse A: 2+5+8+9+11+12+13+15+16+19+22 = 132

Dies ergibt $U_1 = 55$

Für Schulklasse B gilt $T_2 = 121$, $U_2 = 66$

Schulnoten	3.5	3.6	3.7	3.8	3.9	4	4.1	4.2
Schulklasse	B	A	B	B	A	B	B	A
Gemeinsamer Rangplatz	1	2	3	4	5	6	7	8

www.methodenberatung.uzh.ch/datenanalyse/unterschiede/zentral/mann.html

Bewertung von Signifikanz: Mann Whitney Text

Als Prüfgrösse wird immer der kleinere der beiden Werte verwendet, hier also 55.
U gibt die Summe der Rangplatzüberschreitungen an.

Die Frage ist daher, wie oft ein solches Ungleichgewicht der Rangplätze zufällig auftreten kann.

Dazu vergleicht man den kleineren U-Wert mit dem kritischen Wert auf der theoretischen U-Verteilung.

Im konkreten Beispiel ergibt dies eine Signifikant (p-Wert) von 0.718.
Daher liegt kein statistisch signifikanter Unterschied der zentralen Tendenz zwischen den Klassen vor.

Genauso geht man vor, wenn man den Unterschied der Expression eines bestimmten Gens zwischen zwei Mengen von Proben bewerten möchte.

www.methodenberatung.uzh.ch/datenanalyse/unterschiede/zentral/mann.html

Zusammenfassung

Die Methode der Microarrays erlaubt es, die Expression aller möglichen kodierenden DNA-Abschnitte eines Genoms experimentell zu testen.

Die **Zwei-Farben-Methode** ist weit verbreitet um differentielle Expression zu untersuchen.

Aufgrund der natürlichen biologischen Schwankungen müssen die Rohdaten **prozessiert** und *normalisiert* werden.

Durch **Clustering** von Experimenten unter verschiedenen Bedingungen erhält man Gruppen von **ko-exprimierten Genen**.

Diese haben vermutlich **funktionell** miteinander zu tun.

Die Signifikanz der unterschiedliche Expression in zwei Gruppen von Proben bewertet man mit statistischen Testverfahren.