V7 Genexpression - Microarrays

- **Idee:** analysiere 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 V7:**

  (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 Level an transkribierter messenger RNA (mRNA) für alle Gene des Genoms.

Heutzutage gilt dies sowohl für die Protein-kodierenden Gene als auch für 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

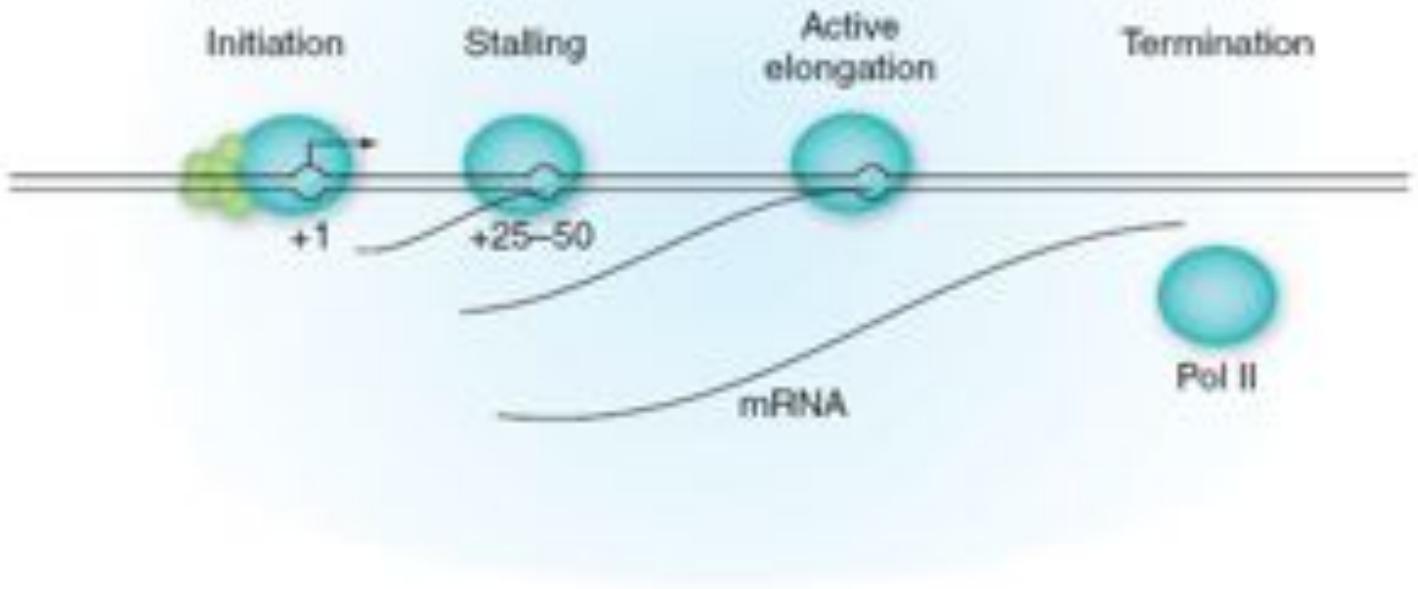


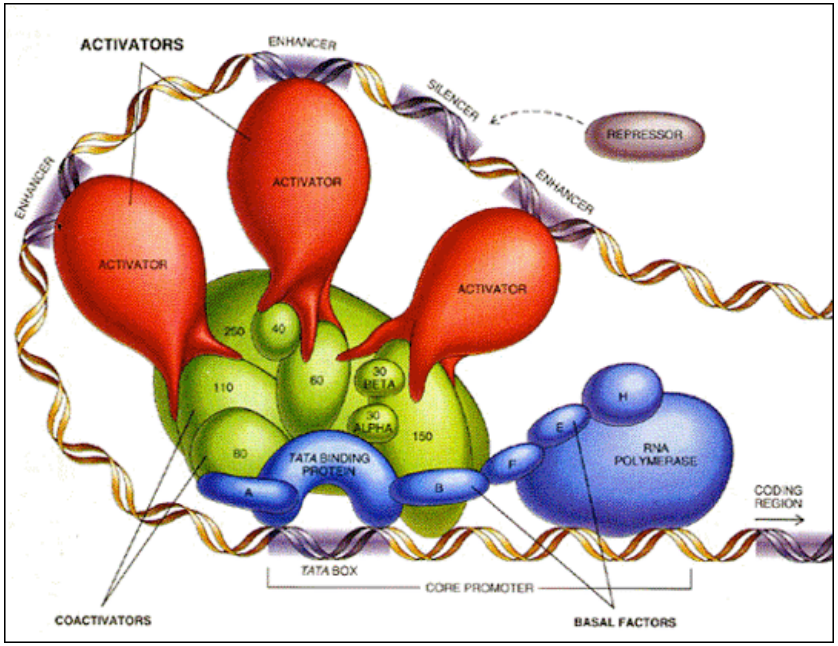
Figure 1 Transcription by RNA polymerase II. Eukaryotic transcription involves a cycle of highly regulated events[1]. After clearing the promoter, RNA polymerase II may pause or stall 25–50 base pairs downstream of the transcription start site before transcribing the body of the gene. Pausing is subject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

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

Transkriptions – Gen-Regulationsnetzwerke

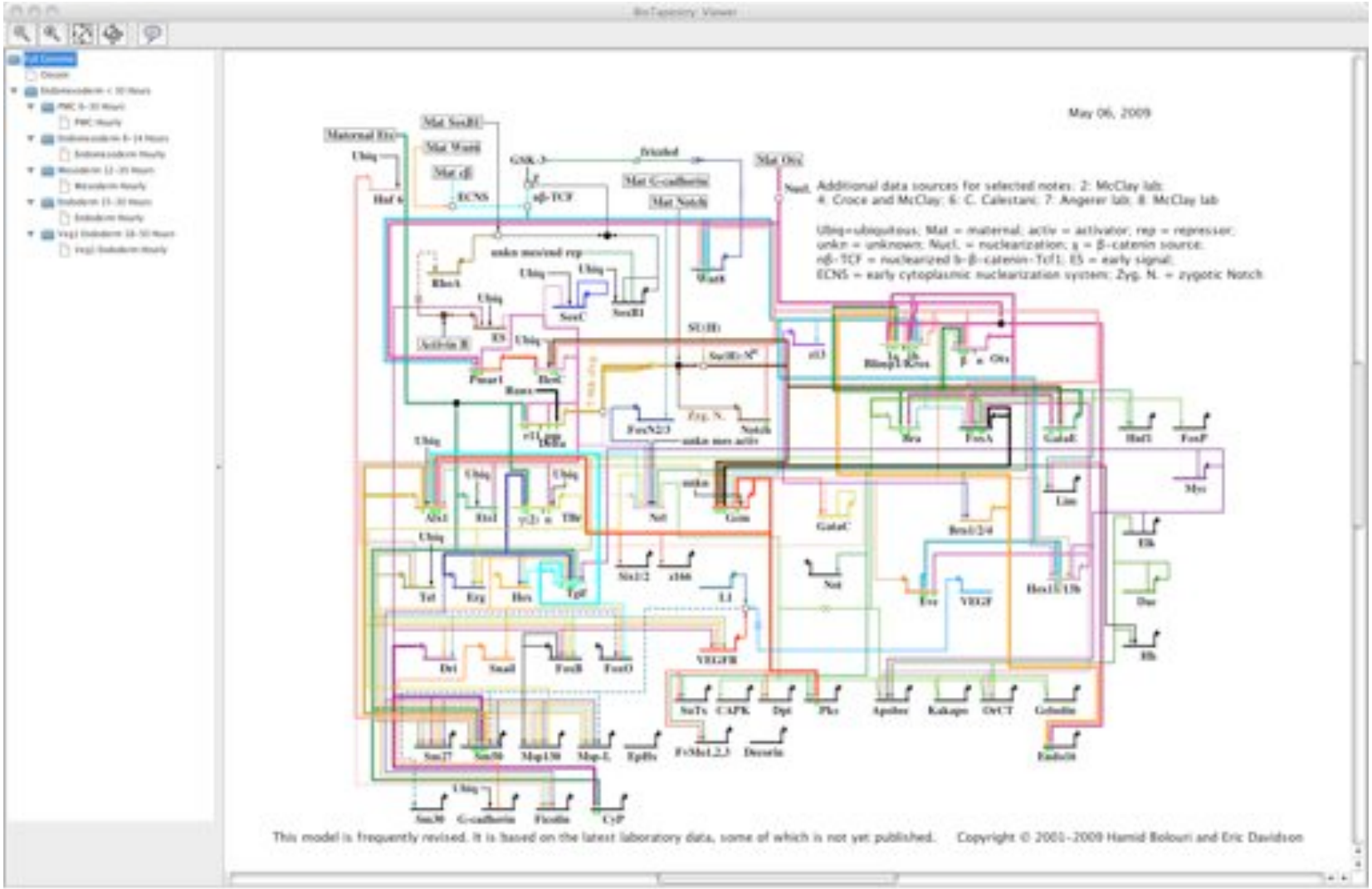
Die **Maschine,** die ein Gen transkribiert, besteht aus etwa 50 Proteinen, einschließlich der **RNA Polymerase**. Dies ist ein Enzym, das DNA code in RNA code übersetzt.

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.



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

endomesodermales Gen-Regulationsnetzwerk der Seegurke

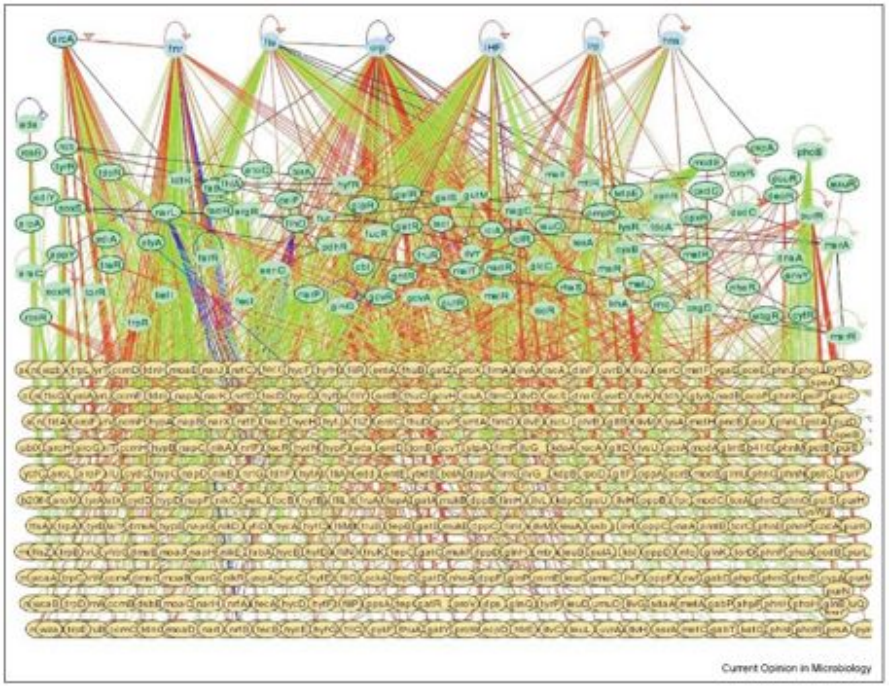


http://sugp.caltech.edu/endomes

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.

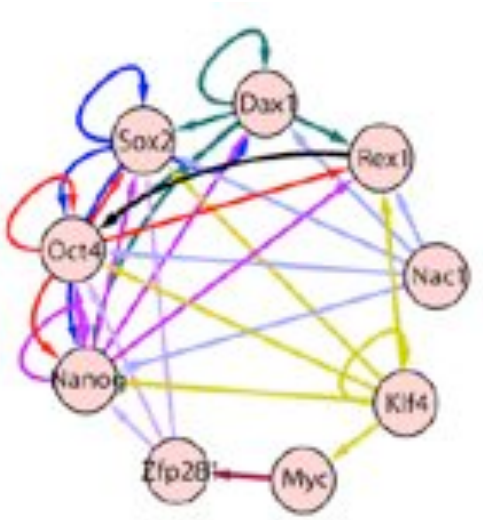


Martinez-Antonio, Collado-Vides, Curr Opin Microbiol 6, 482 (2003)

Genregulationsnetzwerk um Oct4

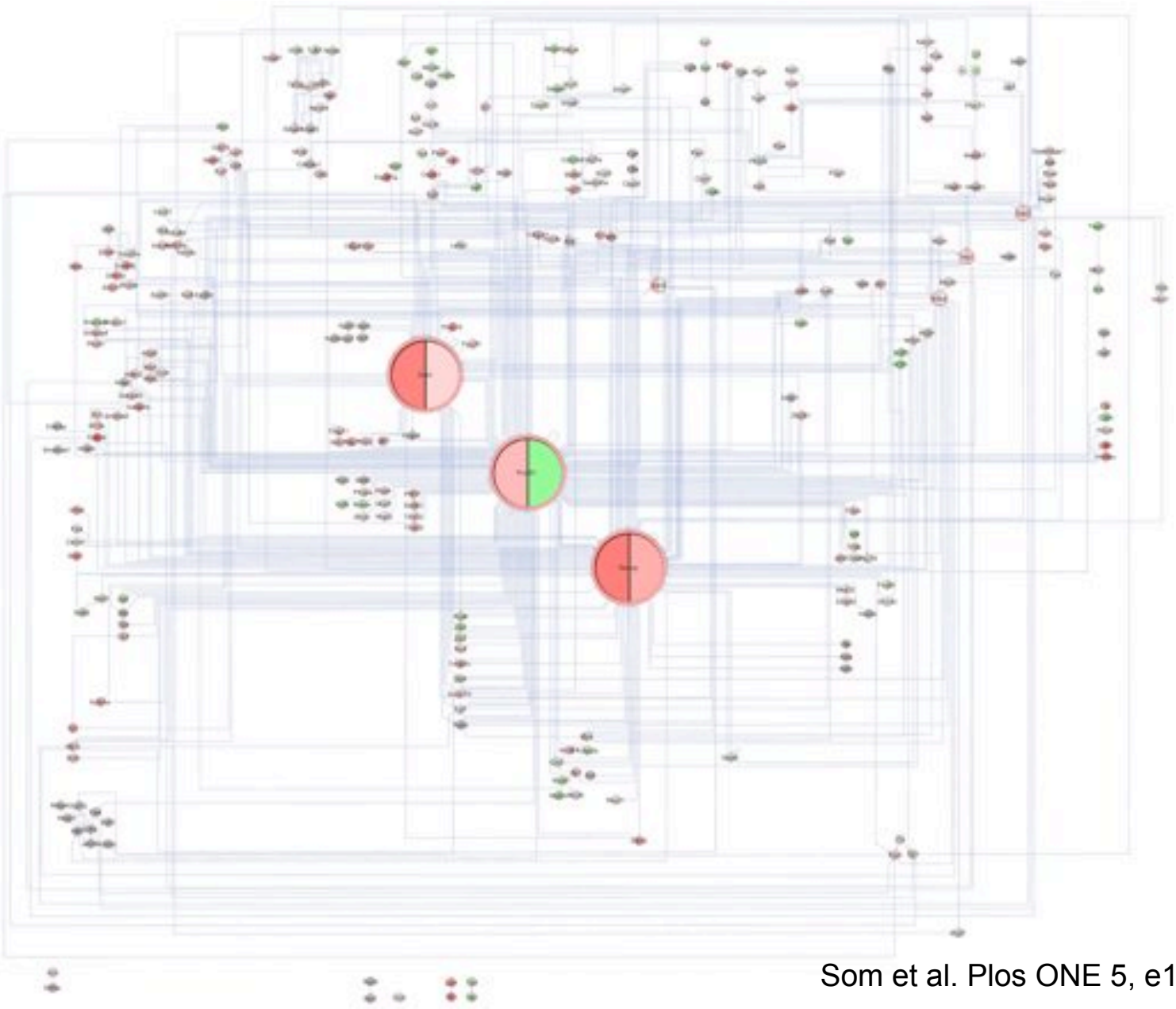
Ein eng verwobenes Netzwerk aus neun Transkriptionsfaktoren hält embryonale Stammzellen im pluripotenten Zustand.

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



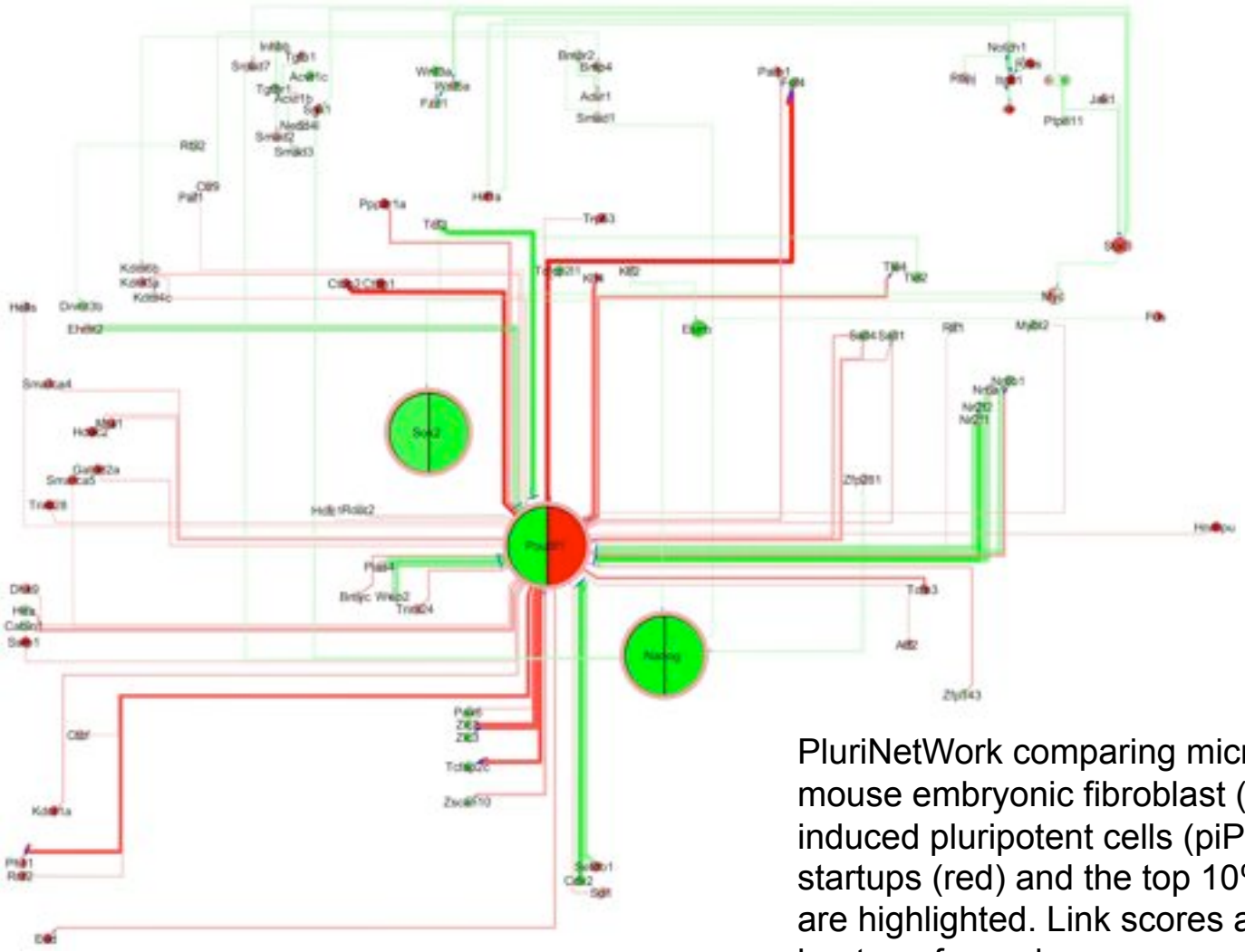
Kim et al. Cell 132, 1049 (2008)

PluriNetwork in Maus



Som et al. Plos ONE 5, e15165 (2010)

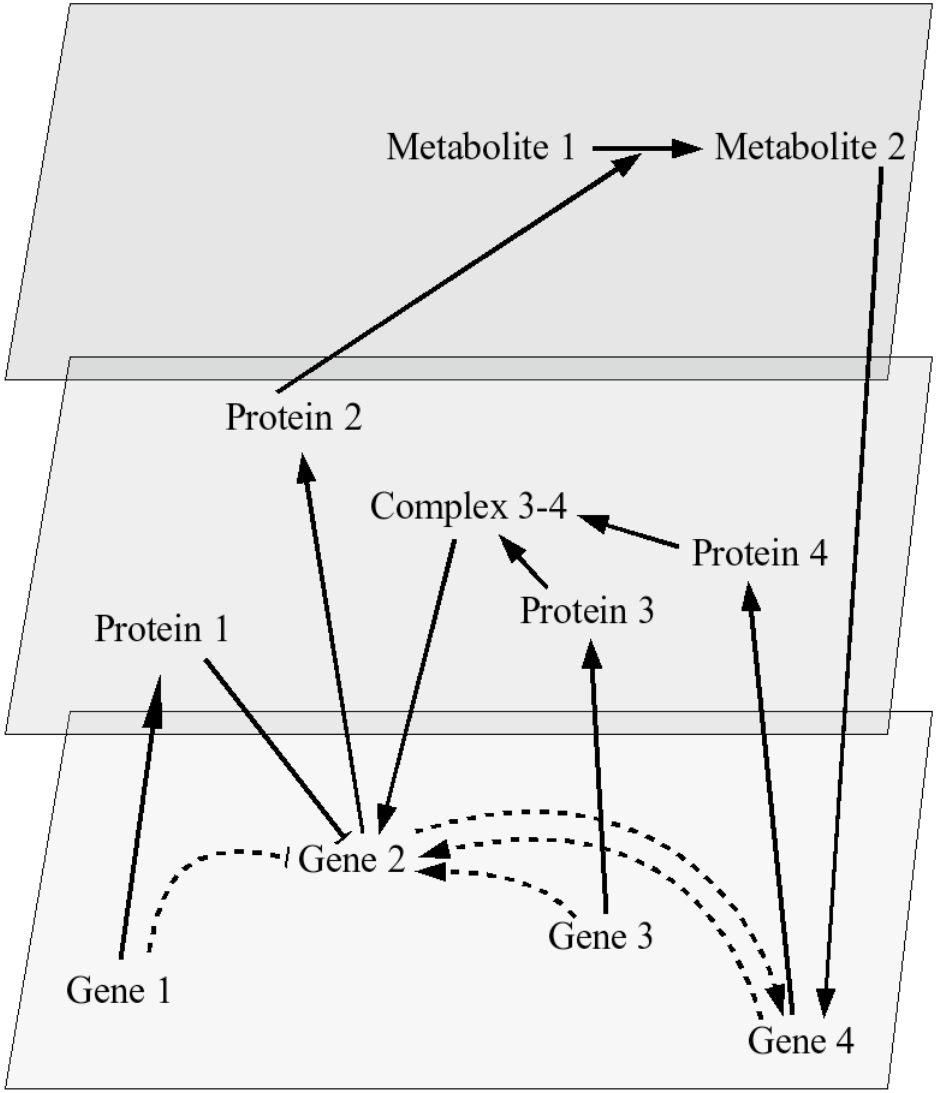
PluriNetwork in Maus



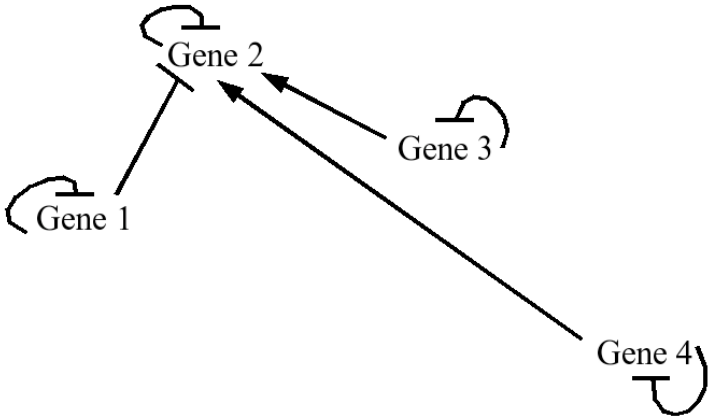
PluriNetWork comparing microarray data from mouse embryonic fibroblast (MEF) and partially induced pluripotent cells (piPS).The top 10% startups (red) and the top 10% shutdowns (green) are highlighted. Link scores are based on log-transformed gene expression intensities, corrected for variance.

Som et al. Plos ONE 5, e15165 (2010)

integrierte zelluläre Netzwerke



Statt des komplexen zellulären Netzwerks (links) stellen Genregulationsnetzwerke nur die Projektion auf die Gen-Ebene dar (unten).



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 mißt man mit Microarrays?

Häufig verwendet werden **Zweifarben-MicroAssays**:

Sample A: rot

Sample B: grün

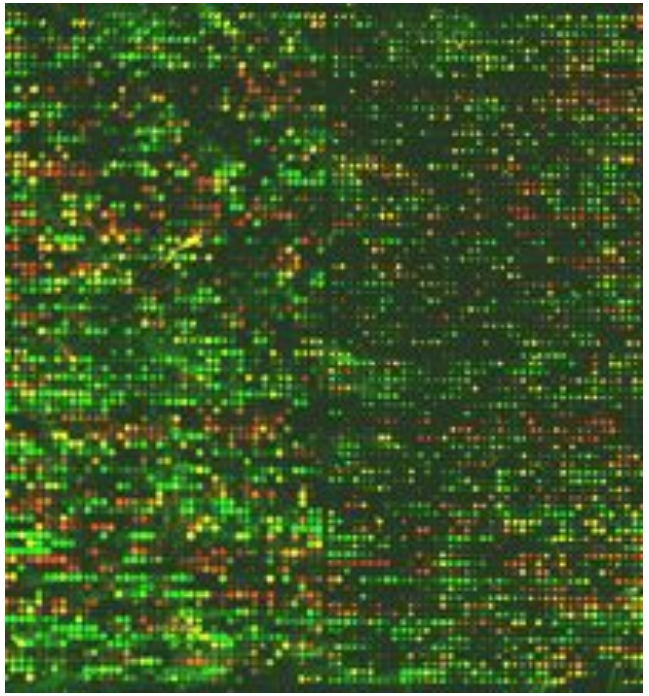
Ziel: bestimme das Verhältnis rot/grün

dunkel: Gen weder in A noch B exprimiert

rot: Gen nur in A exprimiert (bzw. viel stärker)

grün: Gen nur in B exprimiert

gelb: Gen in A und in B exprimiert.



Das Licht wird von zwei Farbstoffen (roter Cy5 und grüner Cy3) erzeugt, die an die cDNA angeheftet wurden (die cDNA wurde „gelabelt“) und die unter Laserlicht fluoreszieren.

pgrc.ipk-gatersleben.de

Experimentelles Vorgehen

Isolierung einer Zelle im Zustand X

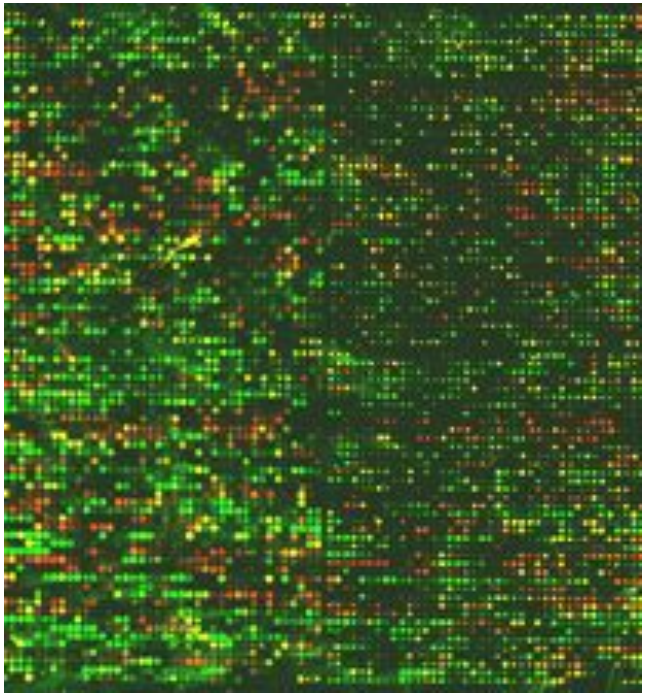
Extraktion aller RNA

Umwandlung in cDNA

Markierung mit Farbstoff (rot oder grün)

Pipette enthält markierte cDNA aller in der Zelle exprimierten Gene.

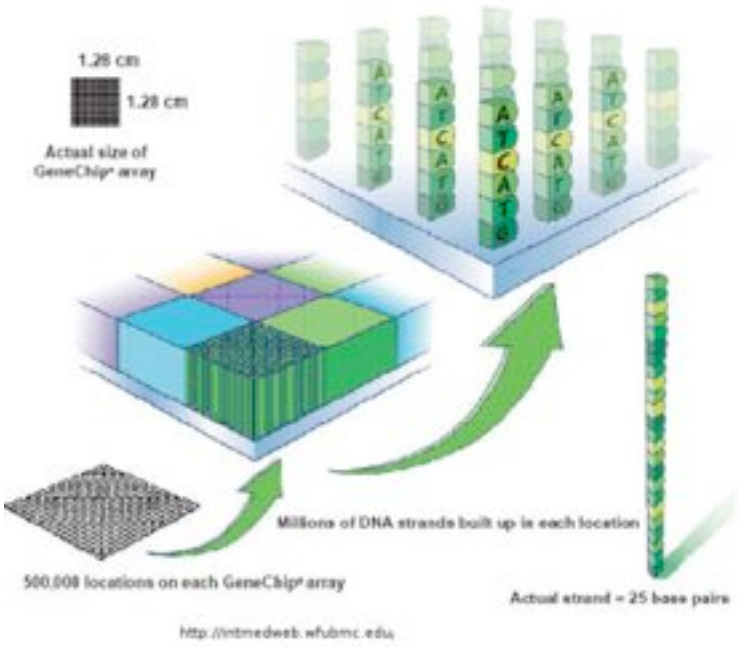
Man bringt nacheinander die cDNA aus zwei verschiedenen Zellpräparationen auf, die unterschiedlich (rot/grün) gelabelt wurden.



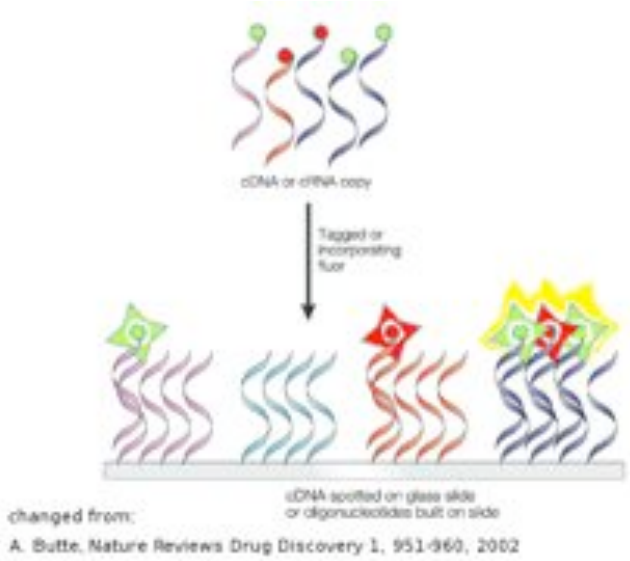
pgrc.ipk-gatersleben.de

Experimentelles Vorgehen

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



Jede Zelle enthält an die Oberfläche funktionalisiert einen cDNA-Klon aus einer cDNA-Bibliothek.



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

pgrc.ipk-gatersleben.de

Einstellung des Gleichgewichts

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

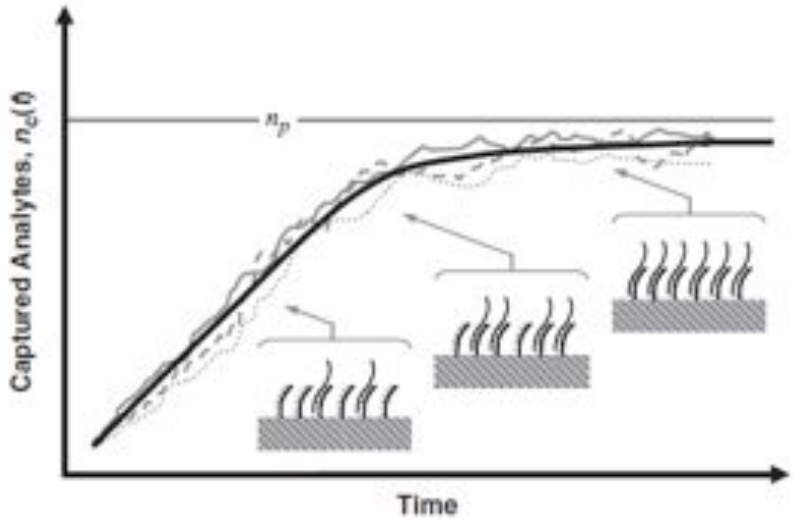
Dann kann man den erwarteten Mittelwert $\langle n_c(t) \rangle$ nach dieser Zeit t durch eine Ratengleichung ausdrücken:

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

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

n_p : Gesamtzahl an immobilisierten DNA-Strängen auf der Microarray-Oberfläche

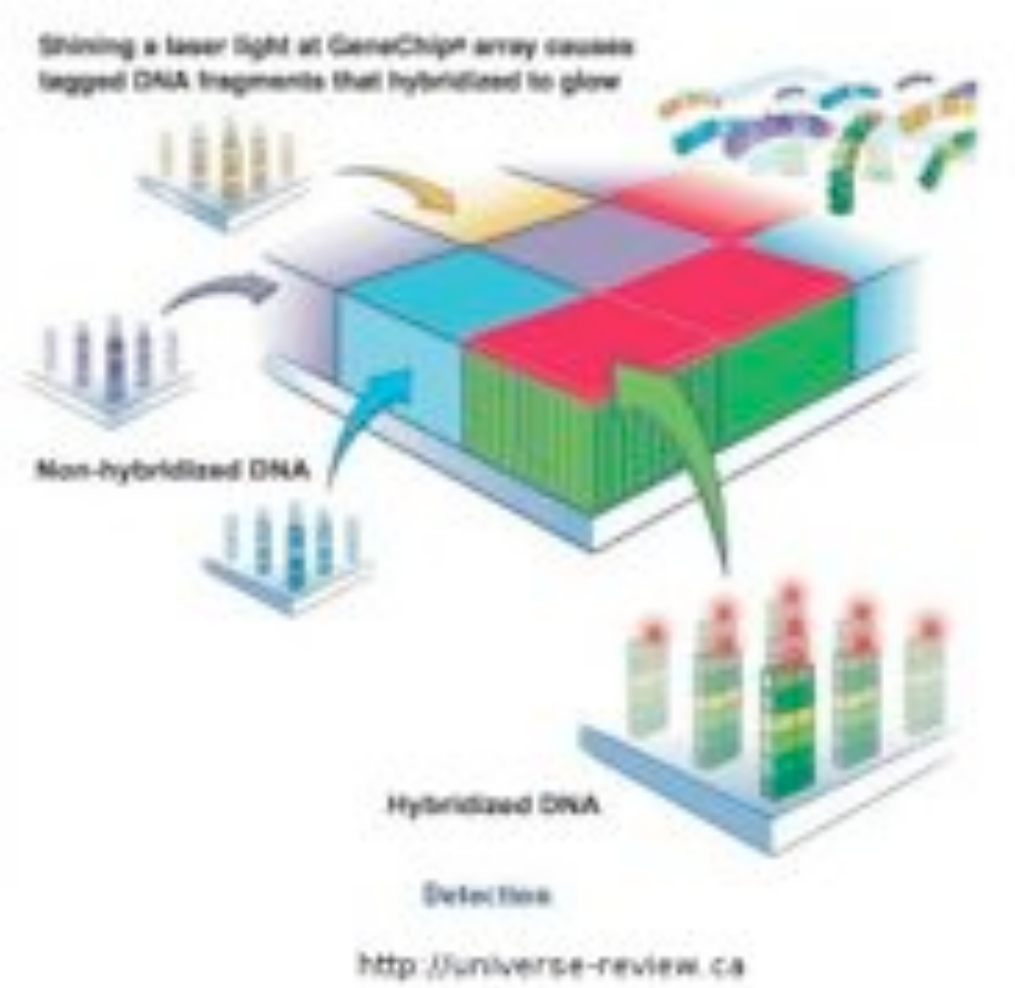
n_t : Gesamtzahl an DNA-Strängen in der Probe



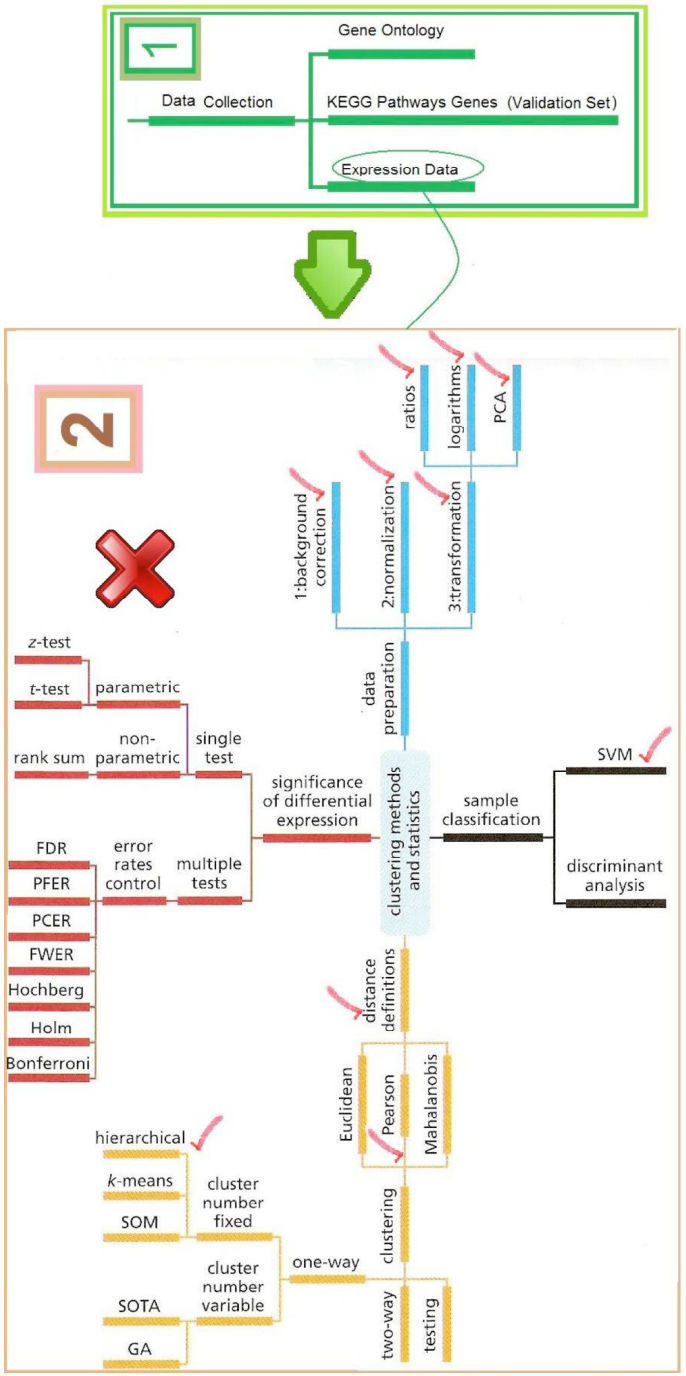
Hassibi et al., Nucl. Ac. Res. 37, e132 (2009)

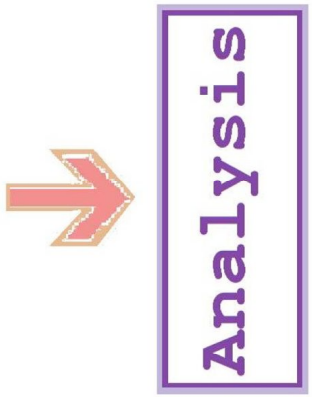
Auslesen der Probe: Laserlicht

Man stimuliert sowohl die Fluoreszenz bei der roten als auch bei der grünen Wellenlänge.



Auswertung von Microarray-Experimenten





Korrekte Reihenfolge bei Prozessierung der Daten:

0. Löschen fehlerhafter Daten
1. Background-correction

(Details werden hier nicht behandelt)

1. Normalisierung
2. Transformation (z.B. Log)

Dann

Clustering

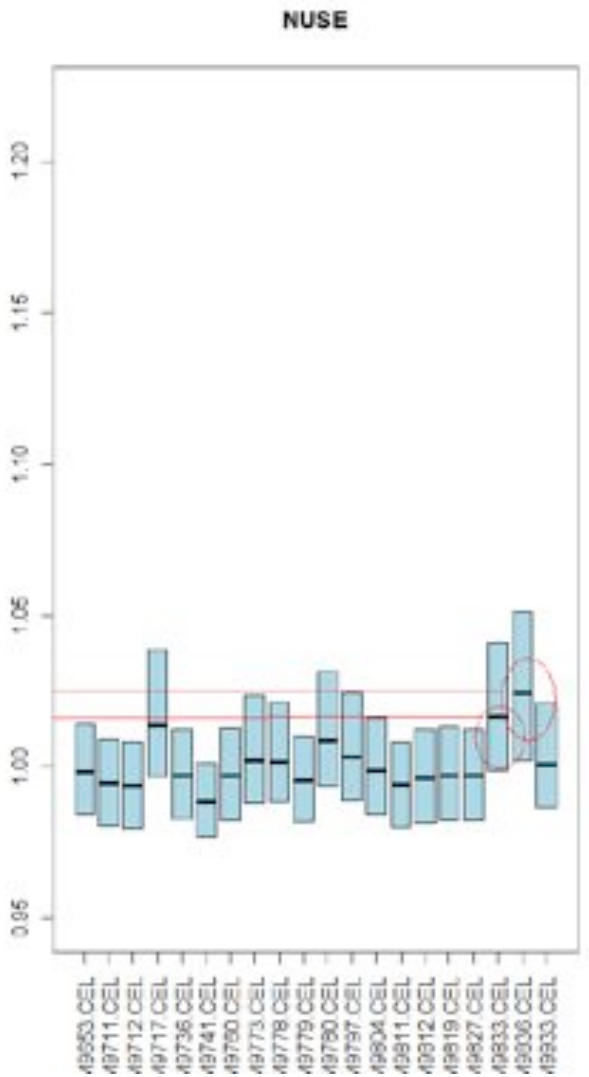
bzw.

Analyse der Signifikanz

0. Entfernen problematischer Datensätze: NUSE-Test

Man schätzt den unskalierten Standardfehler für die Expression jedes Gen auf jedem Array ab und mittelt über alle Arrays und normiert dann über alle Arrays. Danach sollte der normierte unskalierte Standardfehler für alle Gene ungefähr 1 sein.

In diesem Beispiel weicht der NUSE-Wert für GSM9833 und GSM9836 signifikant von 1 ab. Daher werden diese Datensätze als Ausreißer betrachtet und entfernt.



MSc thesis Siba Ismael

1. Normalisierung von Arrays

Wie alle anderen biologischen Experimente zeigen auch Microarrays **zufällige** und **systematische Abweichungen**.

Zufällige Schwankungen treten auf

- in der absoluten Menge an mRNA, die eingesetzt wird,
- in der Hybridisierungs-Technik und
- in Waschschritten.

Systematische Unterschiede gibt es z.B. bei den physikalischen Fluoreszenz-eigenschaften der beiden Farbstoffmoleküle bzw. durch inhomogen hergestellte Microarray-Chips.

Um diese systematischen Abweichungen der Genexpressionslevel zwischen zwei Proben zu unterdrücken, verwendet man **Normalisierungsmethoden**.

Normalisierung

1. Schritt einer Normalisierung: wähle einen Satz von Genen, für die ein Expressionsverhältnis von 1 erwartet wird (z.B. House keeping-Gene).

2. Berechne eine Normierungsfaktor aus der beobachteten Variabilität der Expression für diese Gene.

3. Wende diesen Normierungsfaktor auf die Expressionswerte der anderen Gene an.

Wichtig: durch diese Normierung werden die Daten verändert.

Normalisierung

Man nimmt grundsätzlich an, dass die absolute Menge an RNA in beiden Messungen (bzw. Proben) dieselbe ist und dass die selbe Anzahl an RNA-Molekülen in beiden Messungen mit dem Microarray hybridisieren.

Dann sollten auch die Hybridisierungsintensitäten der beiden Gen-Mengen gleich sein.

Berechne Normierungsfaktor

$$N_{total} = \frac{\sum_{k=1}^{N_{gene-set}} R_k}{\sum_{k=1}^{N_{gene-set}} G_k}$$

Reskaliere die Intensitäten, so dass

$$G_k^{'} = G_k \times N_{total}$$

und

$$R_k^{'} = R_k$$

D.h. nur die Grünwerte werte skaliert, die Rotwerte bleiben unverändert.

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.

In letzterem Fall lautet das Expressionsverhältnis für einen Spot:

$$T_{median} = \frac{R_{median}^{spot} - R_{median}^{background}}{G_{median}^{spot} - G_{median}^{background}}$$

M. Madan Babu, An Introduction to Microarray Data Analysis

Ausgefeiltere Normalisierung: Tukey's median polish

Probe / GeneChip	1	2	3	4	5	Row Medians
1	18	11	8	21	4	**11**
2	13	7	5	16	7	7
3	15	6	7	16	6	7
4	19	15	12	18	5	**15**

Reihenmedian von jeder Reihe abziehen. (Bei ungerader Anzahl von Werten ist der Wert in der Mitte, so dass die Hälfte der Werte kleine ist, die Hälfte größer.)

Probe / GeneChip	1	2	3	4	5
1	7	0	-3	10	-7
2	6	0	-2	9	0
3	8	-1	0	9	-1
4	4	0	-3	3	-10
Column Meedians	**6.5**	**0**	**-2.5**	**9**	**-4**

Spaltenmedian von jeder Spalte abziehen. (Bei gerader Anzahl von Werten gibt es zwei mittlere Werte. Man nimmt oft deren arithmetischen Mittelwert.)

Probe / GeneChip	1	2	3	4	5	Row Medians
1	0.5	0	-0.5	1	-3	**0**
0.52	-0.5	0	0.5	0	4	**0**
3	1.5	-1	2.5	0	3	**1.5**
4	-2.5	0	-0.5	-6	-6	**-2.5**

Orengo-Buch

Normalisierung: Tukey's median polish (II)

Probe / GeneChip	1	2	3	4	5	Row Medians
1	0.5	0	-0.5	1	-3	**0**
0.52	-0.5	0	0.5	0	4	**0**
3	1.5	-1	2.5	0	3	**1.5**
4	-2.5	0	-0.5	-6	-6	**-2.5**

Nochmal Reihenmedian von jeder Reihe abziehen.

Probe / GeneChip	1	2	3	4	5
1	0.5	0	-0.5	1	-3
2	-0.5	0	0.5	0	4
3	0	-2.5	1	-1.5	1.5
4	0	2.5	2	-3.5	-3.5
Column Meedians	**0**	**0**	**0.75**	**-0.75**	**-0.75**

Nochmal Spaltenmedian von jeder Spalte abziehen.

Probe / GeneChip	1	2	3	4	5	Row Medians
1	0.5	0	-1.25	1.75	-2.25	**0**
2	-0.5	0	-0.25	0.75	4.75	**0**
3	0	-2.5	0.25	-0.75	2.25	**0**
4	0	2.5	1.25	-2.75	-2.75	**0**
Column Meedians	**0**	**0**	**0**	**0**	**0**	

Normalisierung: Tukey's median polish (II)

Probe GeneChip	1	2	3	4	5	Row Medians
1	0.5	0	-1.25	1.75	-2.25	**0**
2	-0.5	0	-0.25	0.75	4.75	**0**
3	0	-2.5	0.25	-0.75	2.25	**0**
4	0	2.5	1.25	-2.75	-2.75	**0**
Column Meedians	**0**	**0**	**0**	**0**	**0**	

Ziehe diese Matrix von der Originalmatrix

Probe GeneChip	1	2	3	4	5	Row Medians
1	18	11	8	21	4	**11**
2	13	7	5	16	7	**7**
3	15	6	7	16	6	**7**
4	19	15	12	18	5	**15**

ab:

Fitted Values Probe GeneChip	1	2	3	4	5
1	17.5	11	9.25	19.25	6.25
2	13.5	7	5.25	15.25	2.25
3	15	8.5	6.75	16.75	3.75
4	19	12.5	10.75	20.75	7.75

Bereich der Expressionsverhältnisse

Das Expressionsverhältnis 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).

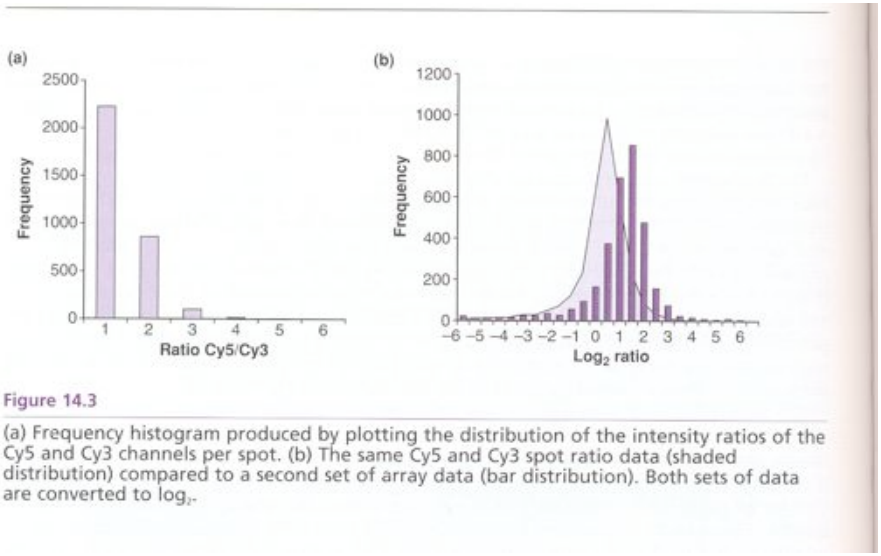


Figure 14.3

(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.

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

(a)

Genes	Arrays I	II	III	IV	V	VI
1 ■	–3	–3	–1	0	2	3
2 ●	–2	–2	0	1	2	2
3 ▲	–3	–2	0	1	2	3
4 ★	3	2	0	–1	–2	–3
5 □	2	2	1	0	–2	–3
6 ○	3	2	1	0	–2	–3
7 △	2	2	2	2	2	2
8 ☆	–2	–2	–2	–2	–2	–2

(b)

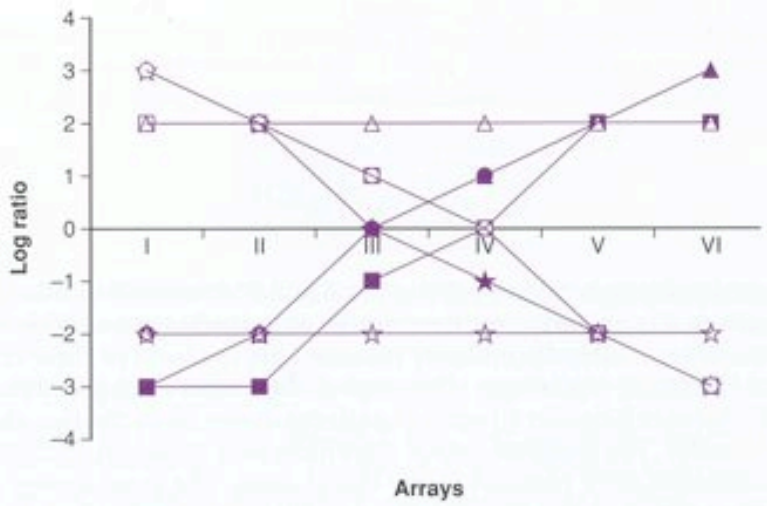


Figure 15.1

(a) Gene names are in column 1 (gene 1 to 8) and each gene is coded per row, with respect to (b). Arrays (I to VI) are in columns (2–7) with each feature being a gene expression $\log_2$ ratio value (see Chapter 14). (b) A graphical representation of the data for eight genes measured across six arrays corresponding the gene matrix in (a).

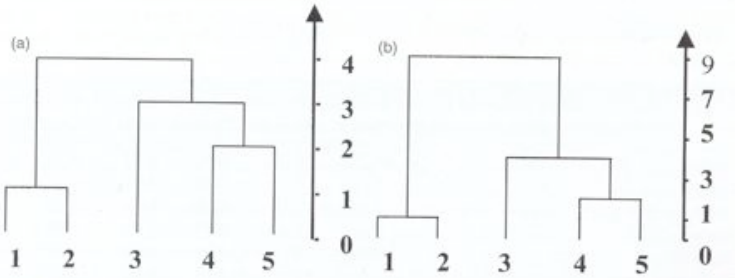


Figure 15.2

Dendrograms from single-linkage (a) and complete-linkage clustering (b)

Orengo-Buch

k-means Clustern

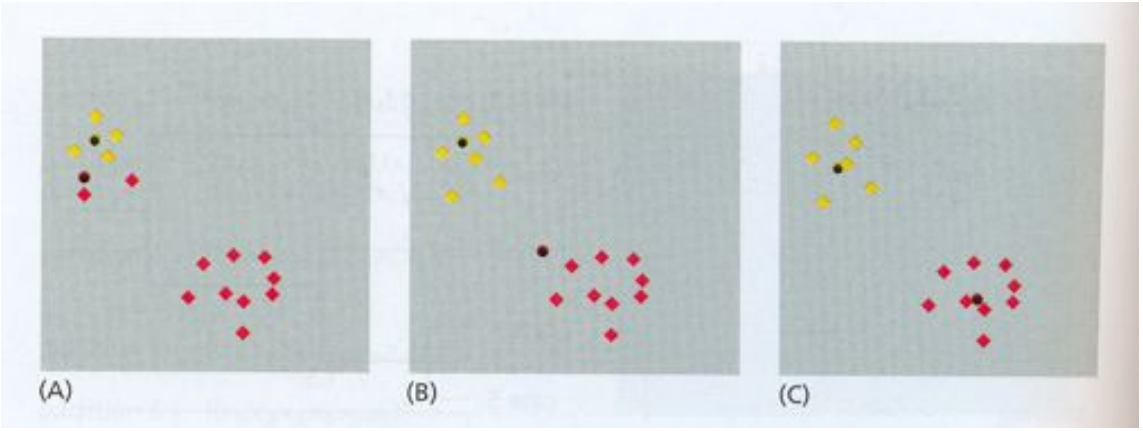
Ein Durchlauf der *k*-means Clustering Methode erzeugt eine Auftrennung der Datenpunkte in *k* Cluster. Gewöhnlich wird der Wert von *k* vorgegeben.

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



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.