Einführung in die Angewandte Bioinformatik: Analyse und Design von DNA Microarrays 08.07.2010

Prof. Dr. Sven Rahmann

Transcript – Omics (Genexpressionsanalyse)

Zum Verständnis von lebenden Organismen untersucht man u.a. ihr
- Genom
- Transkriptom
- Proteom
- Metabolom,

sowie Wechselwirkungen zwischen diesen Ebenen.

Das Genom wird durch DNA-Sequenzierung in Genomprojekten bestimmt.
Das Proteom und Metabolom z.B. mit Massenspektrometrie.

Transcript – Omics (Genexpressionsanalyse)

Das Transkriptom (Gesamtheit der in RNA transkribierten DNA) wird bestimmt mittels

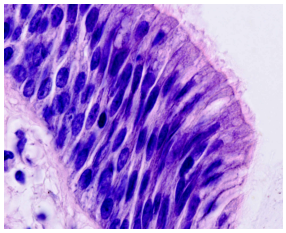
- aktuell: DNA-Microarrays,
- Zukunft: Hochdurchsatz-Sequenzierung von mRNA.

Transkription variiert

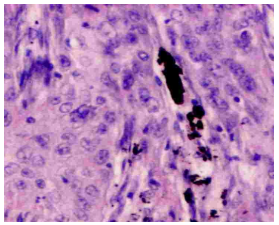
- von Gewebe zu Gewebe,
- abhängig von äußeren und inneren Einflüssen,
- als Antwort auf Signale anderer Zellen, ...

Genexpressionsanalyse: Motivation
Beispiel Bronchialkarzinom

Phänotyp „gesund“



~ Proteinexpression ~ Genexpression



~ Differenzielle Proteinexpression ~ Differenzielle Genexpression

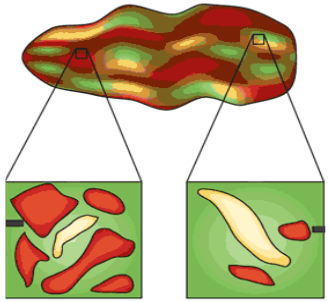
Phänotyp „Bronchialkarzinom“

Kandidaten-Gen: differenziell exprimiert in Tumorzellen.
Identifikation wichtig für Diagnostik und Therapie.

(Alte) cDNA-Array-Technologie

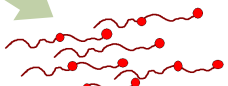
2. Vorbereitung der Proben (samples)

Zellen aus Gewebe

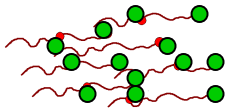


mRNA-Extraktion & Amplifikation

Markierung mit Fluoreszens-Farbstoffen



3. Hybridisierung

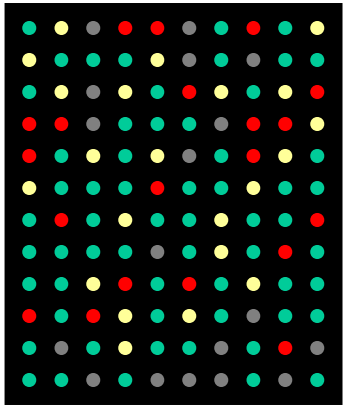


1. Vorbereitung des Microarrays

PCR-Amplifikation der zu untersuchenden Gene (pro Gen ein Arbeitsschritt)

Spotten der cDNA auf definierten Plätzen

Microarray



4. Bild- und Datenanalyse

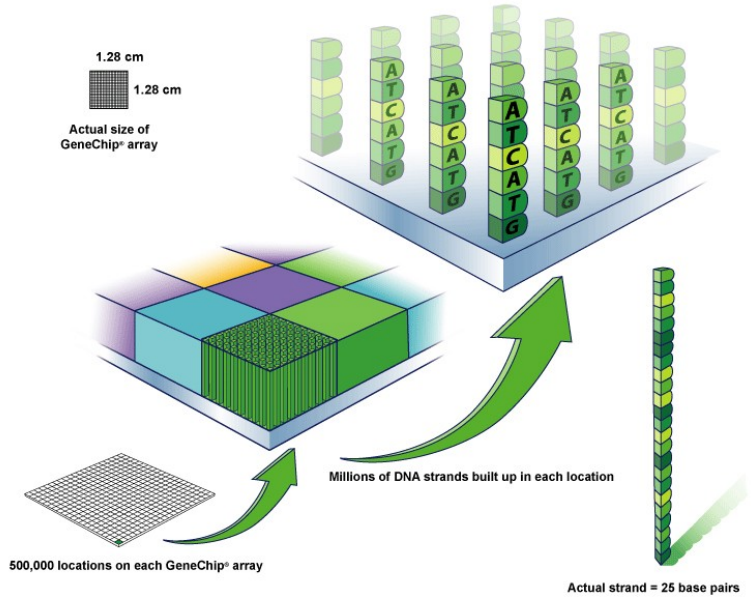
Rot/Grün-Verhältnis
Hintergrundkorrektur

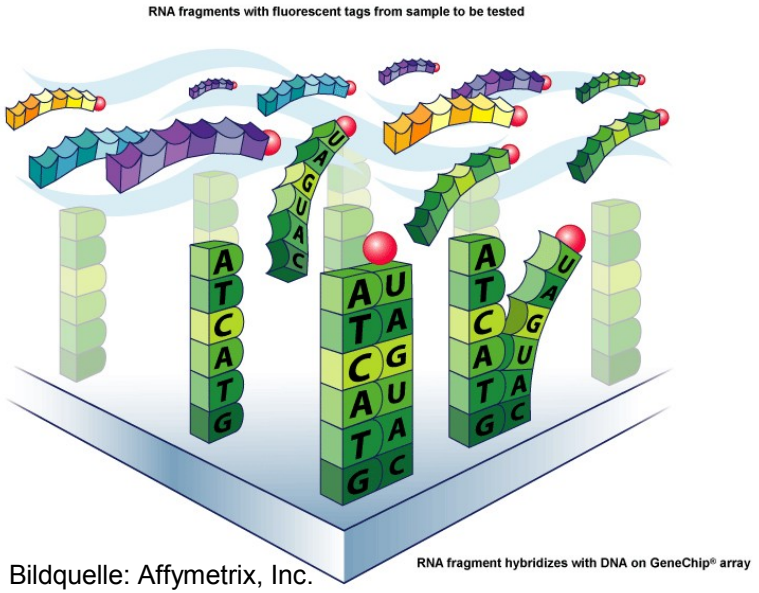
(Aktuelle) Oligonuktleotid-Array-Technologie

Oligonukleotid-Microarrays bestehen aus Mio. spots oder features.

- Spot enthält Mio. Kopien eines bestimmten Oligonukleotids.
- Oligonukleotid hybridisiert (spezifisch) an mRNA / cDNA (Watson-Crick-Paarung).
- mRNA/cDNA ist fluoreszierend markiert.
- Ungebundene RNA wird abgewaschen.
- Fluoreszens-Signal auf Spot erlaubt Rückschluss auf Genexpression.

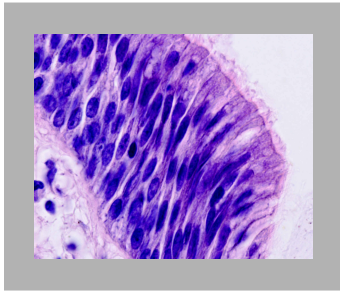
Wichtig: Oligonukleotid-Sonden **gen-spezifisch**.



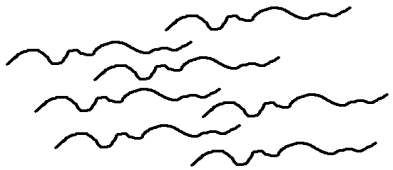


Bildquelle: Affymetrix, Inc.

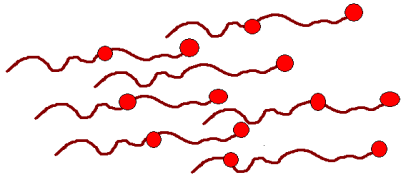
Ablauf eines einzelnen Microarray-Experiments



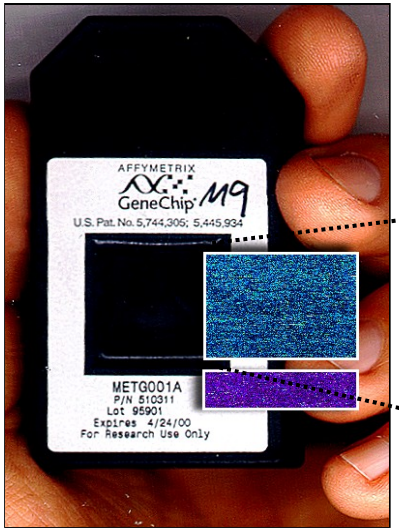
Extraktion der poly-A-RNA

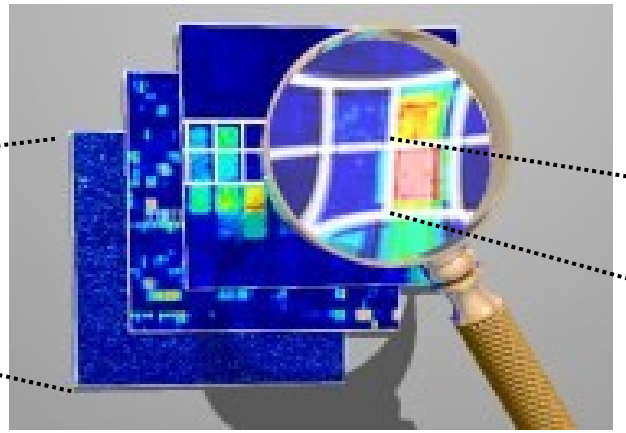


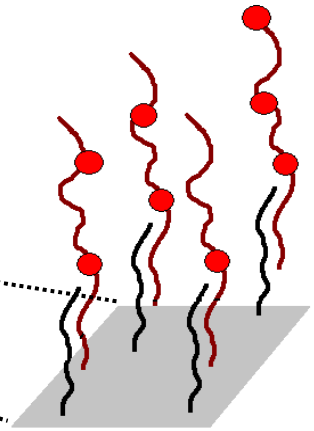
Amplifikation und Markierung der RNA



Fragmentierung
Hybridisierung







metaGen

Schritte einer Microarray-Studie

Fragestellung

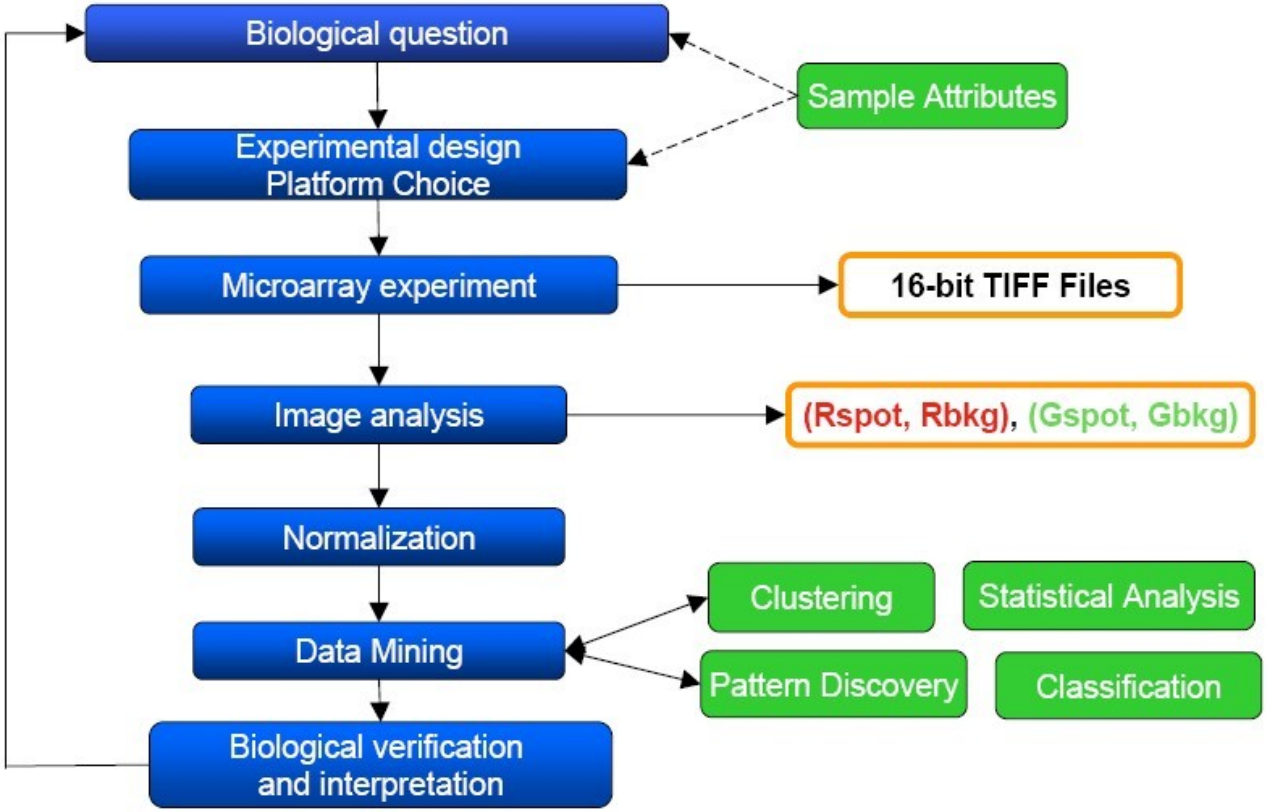
Experiment-Design, Auswahl der Technologie

Durchführung Experiment

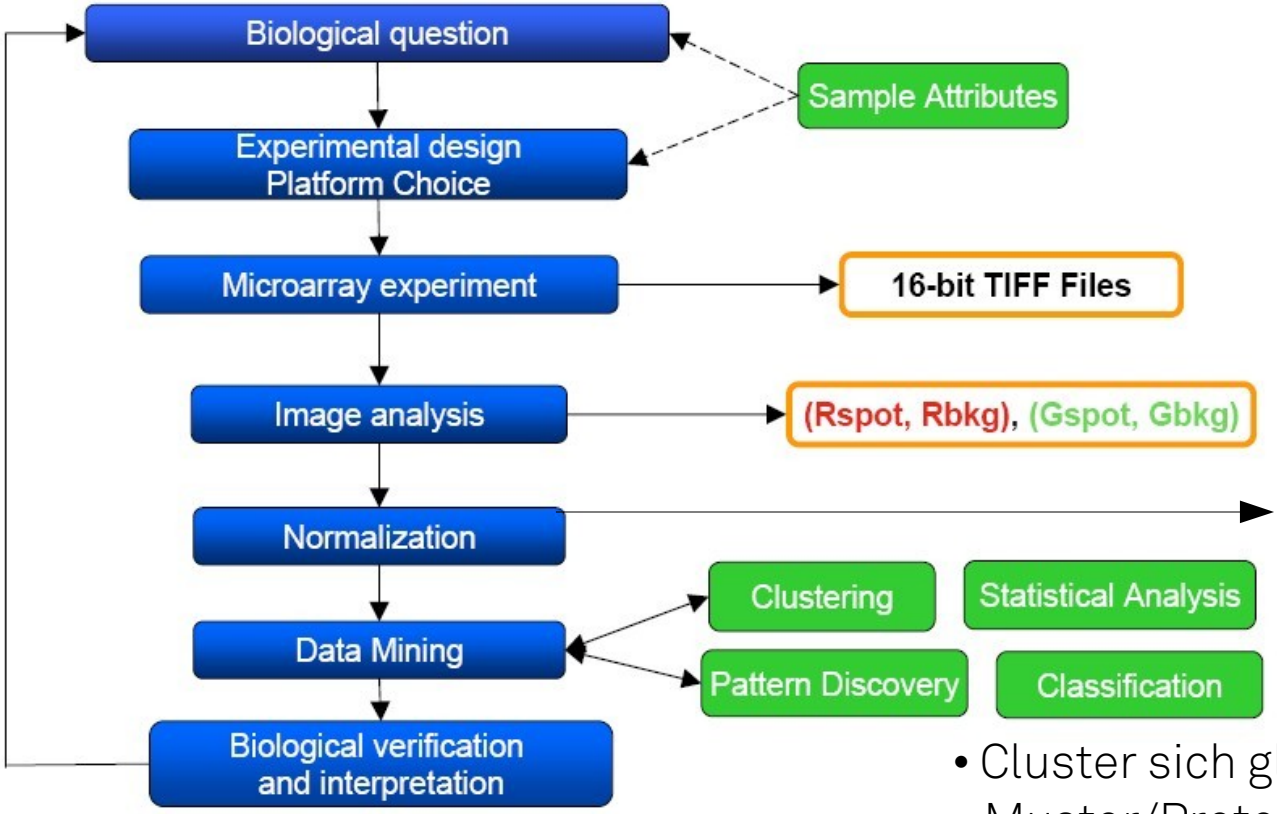
Bildanalyse
Normalisierung

Gewinnung von Information aus den Rohdaten

Verifikation, Interpretation



Resultate einer Microarray-Studie



Ergebnisse aus Daten-Sicht:

- Rohdaten (Bild-Dateien).
- „Halbgare“ Daten (Roh-Farb-Intensitäten).
- Normalisierte Expressionswerte pro Gen.
- Cluster sich gleich verhaltender Gene, Muster/Prototypen, Einteilung in Klassen.

Ergebnisse aus Wissens-Sicht:

- Unterstützung oder Falsifikation der initialen Hypothese; neue Hypothese(n).
- wenn keine Hypothese vorlag: Generierung von Hypothesen.

Auswertung einer Microarray-Studie

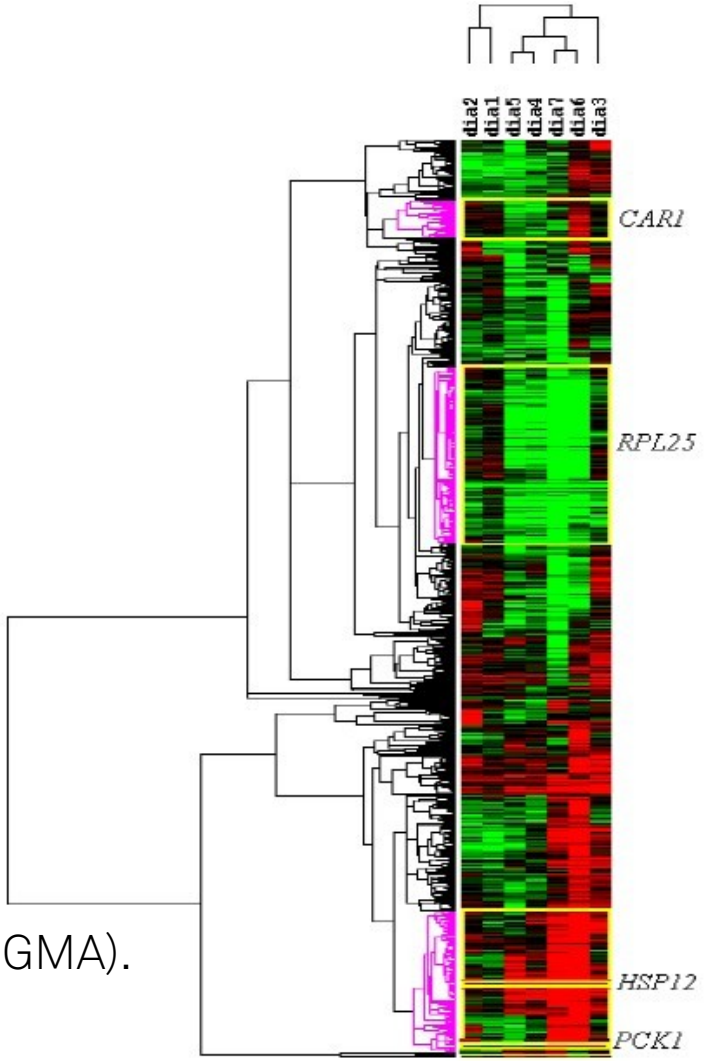
Typische Visualisierung einer Microarray-Studie:
hierarchisches (Bi-)Clustering

- Zeilen: Gene
- Spalten: Experimente (z.B. Zeitpunkte)

Rot: Gen unterexprimiert gegenüber Standard.
Grün: Gen überexprimiert gegenüber Standard.

Gene, die sich über viele Experimente gleich verhalten, bilden Gen-Cluster
Experimente, in denen sich viele Gene gleich verhalten, bilden Experiment-Cluster.

Anordnung der Gene und Experimente:
Ähnliche möglichst nebeneinander.
Berechnung wie bei phylogenetischen Bäumen (UPGMA).



Die einzelnen Analyseschritte

Low-Level-Analyse:
- Bildanalyse
- Hintergrundkorrektur
- Normalisierung (Herausrechnen von systematischen Fehlern)
- Schätzung der absoluten oder relativen Genexpression pro Gen

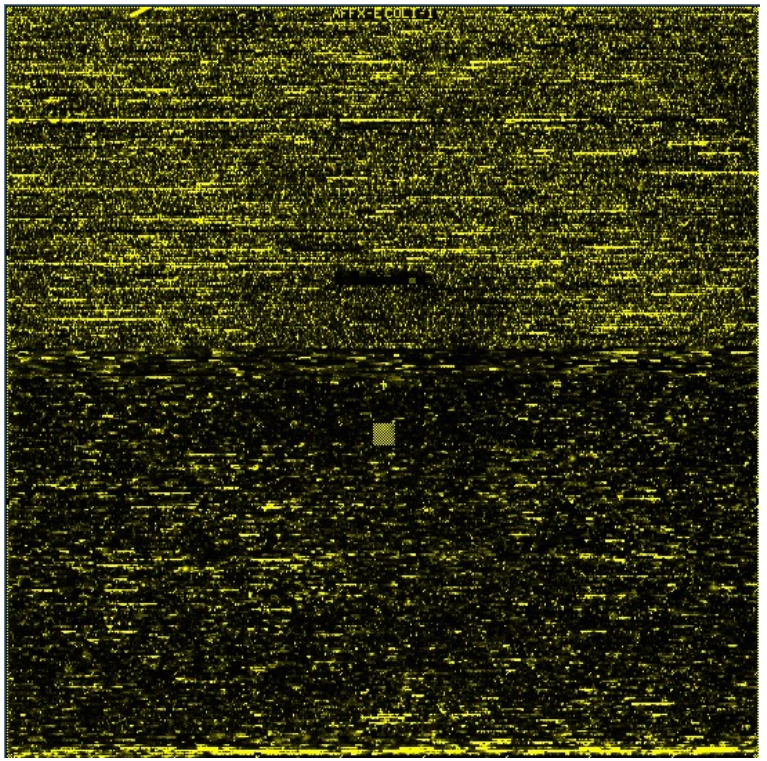
High-Level-Analyse:
- Identifizierung differenziell exprimierter Gene
- Clustering und Klassifikation von Genen und Experimenten

Low-Level: Bildanalyse und Hintergrundkorrektur

Eingabe:
- fotographierte Bilddaten (ein Intensitätswert pro Bildpunkt)

Vorgehen:
- Identifikation der Lage der Spots
- Berechnung einer Gesamtintensität pro Spot aus Bildpunkten, die Spot bilden
- Abzug des „Hintergrunds“
- Transformation (z.B. Logarithmierung)

Ausgabe:
- ein Intensitätswert pro Spot/Feature



Low-Level: Normalisierung

Problem:

- Intensitätswerte pro Spot nicht direkt vergleichbar für mehrere Arrays
- Grund z.B.: mehr Farbstoff; längere Belichtung auf zweitem Array
- Daher z.B. alle Werte auf Array 2 doppelt so groß wie auf Array 1, aber Gene nicht doppelt so aktiv.

Grundannahme der Normalisierung:

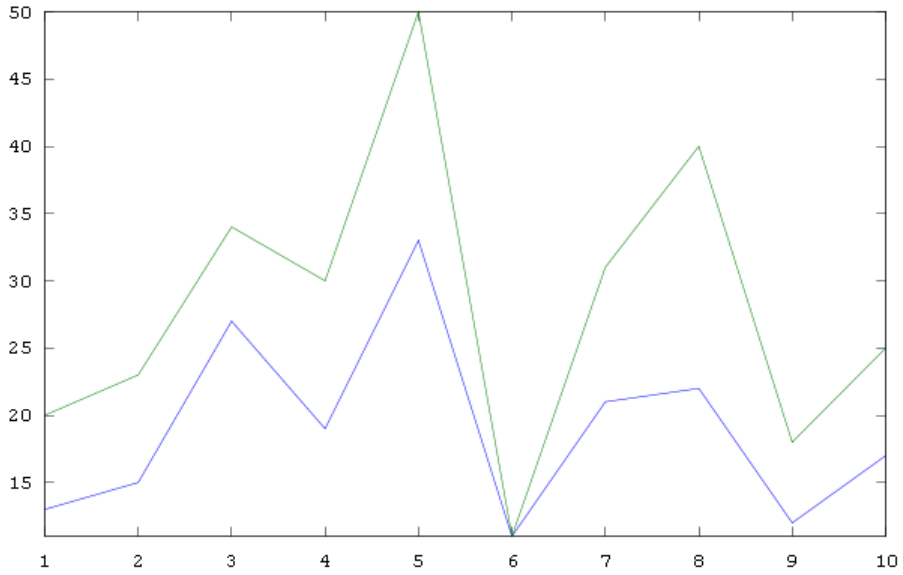
Zwischen zwei Experimenten (Arrays)

- ändert sich die Expression einzelner Gene
- ändern sich nicht globale Eigenschaften der Verteilung der Expressionswerte

Daher:
Angleichung der Histogramme möglich und sinnvoll.

Beispiel zur Normalisierung

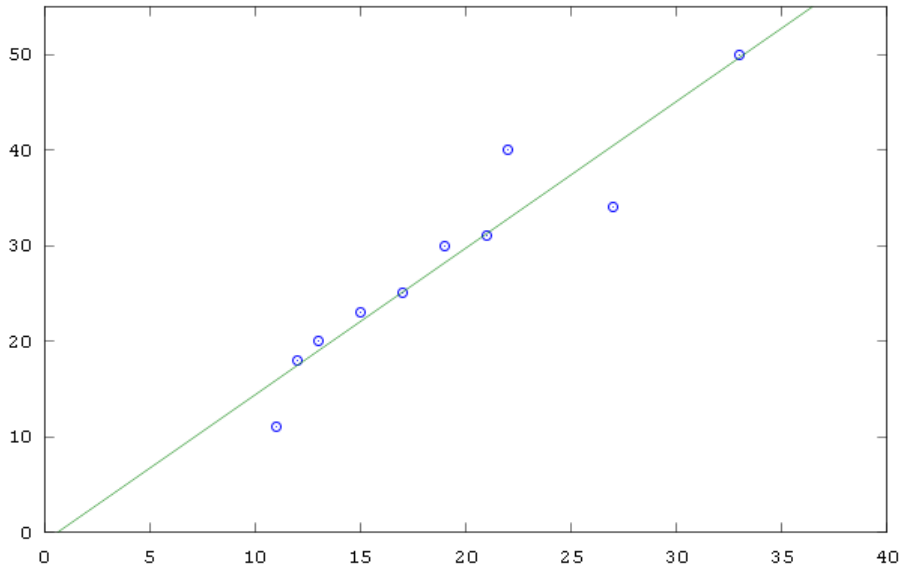
Zwei „Mini“-Microarrays mit nur 10 Spots:
Array 1 in blau: (13, 15, 27, 19, 33, 11, 21, 22, 12, 17)
Array 2 in grün: (20, 23, 34, 30, 50, 11, 31, 40, 18, 25)



Alle Werte in Array 2 scheinen höher.
Widerspricht Grundannahme, also: angleichen = normalisieren.

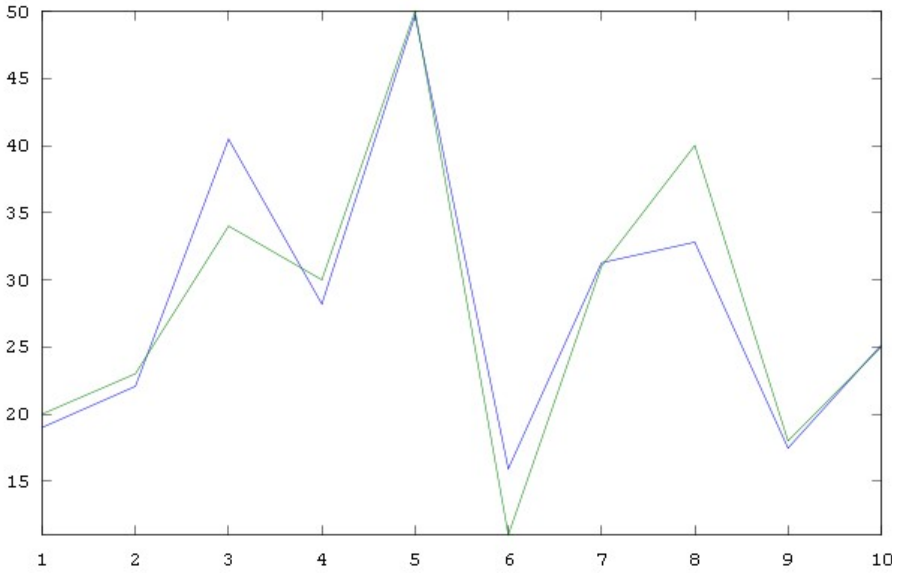
Affin-Lineare Normalisierung

Annahme:
Es gilt ein affiner Zusammenhang $y = ax + b$
mit x: Wert aus Array 1, y: entsprechender Wert aus Array 2.
Scatterplot von x gegen y (blau) und
Berechnung einer Ausgleichsgeraden (grün): $y = 1.53394\,x - 0.94480$.

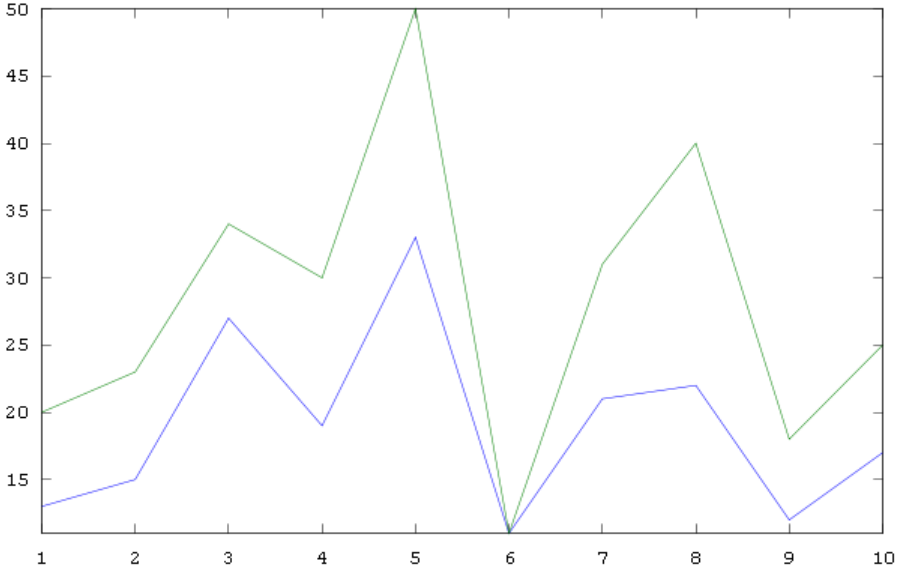


Affin-Lineare Normalisierung

Zwei „Mini“-Microarrays mit nur 10 Spots:
Array 1: (13, 15, 27, 19, 33, 11, 21, 22, 12, 17)
angleichen mittels $y = 1.53394\,x - 0.94480$.
Array 1' in blau: (19.0, 22.0, 40.5, 28.2, 49.7, 15.9, 31.3, 32.8, 17.5, 25.1)
Array 2 in grün: (20, 23, 34, 30, 50, 11, 31, 40, 18, 25)



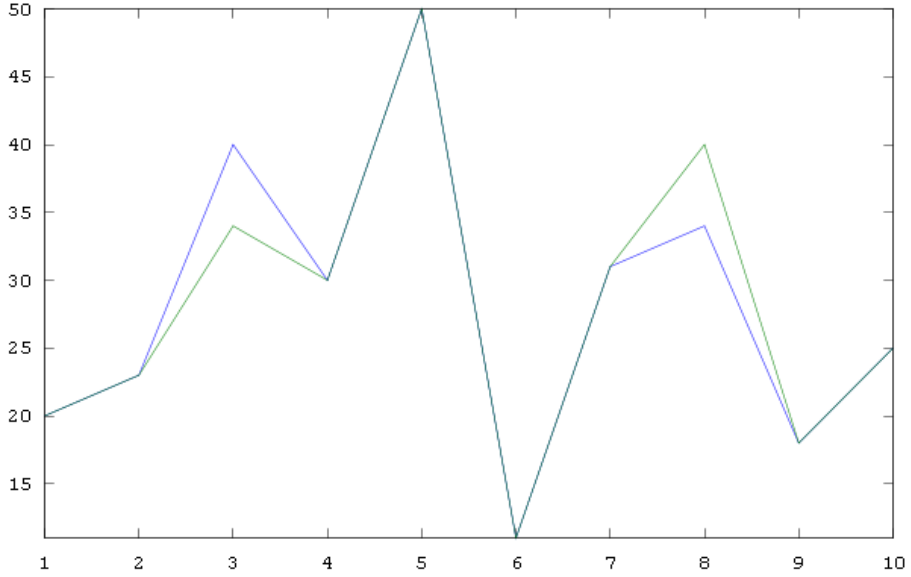
(nach Normalisierung)

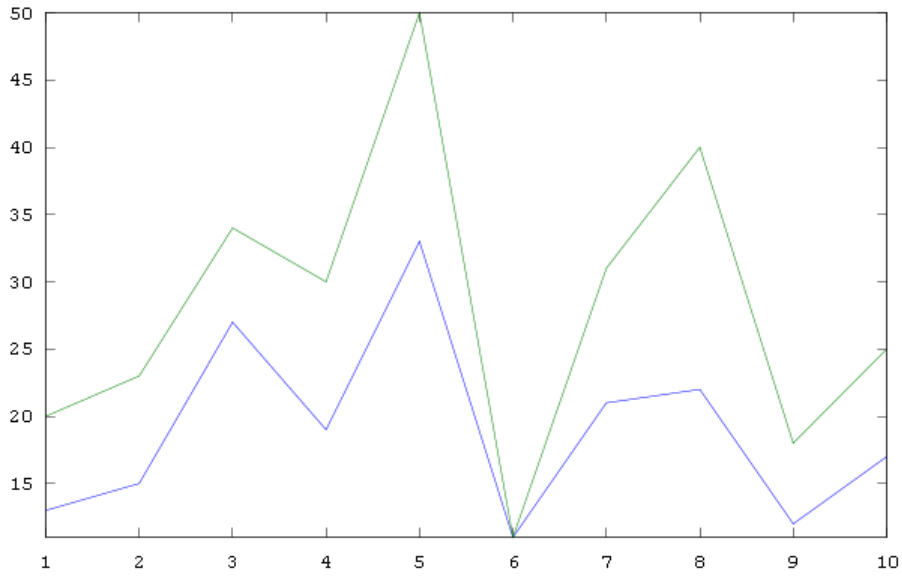


(vor Normalisierung)

Quantil-Normalisierung (RMA-Verfahren)

Annahme: Sortiert man die Intensitäten beider Arrays, sind die Werte gleich.
D.h. Alle Quantile der Intensitätsverteilungen sind gleich (und gleich Referenz).
Array 1: (13, 15, 27, 19, 33, 11, 21, 22, 12, 17) # x
Array 2: (20, 23, 34, 30, 50, 11, 31, 40, 18, 25) # y
Array 1': (20, 23, 40, 30, 50, 11, 31, 34, 18, 25) # x normalisiert
Array 1' enthält die Werte aus Array 2, aber Sortierung aus Array 1.
in R: `x[order(x)] = referenz  # referenz: sortierte y-Werte`





Schätzung eines Expressionswerts pro Gen

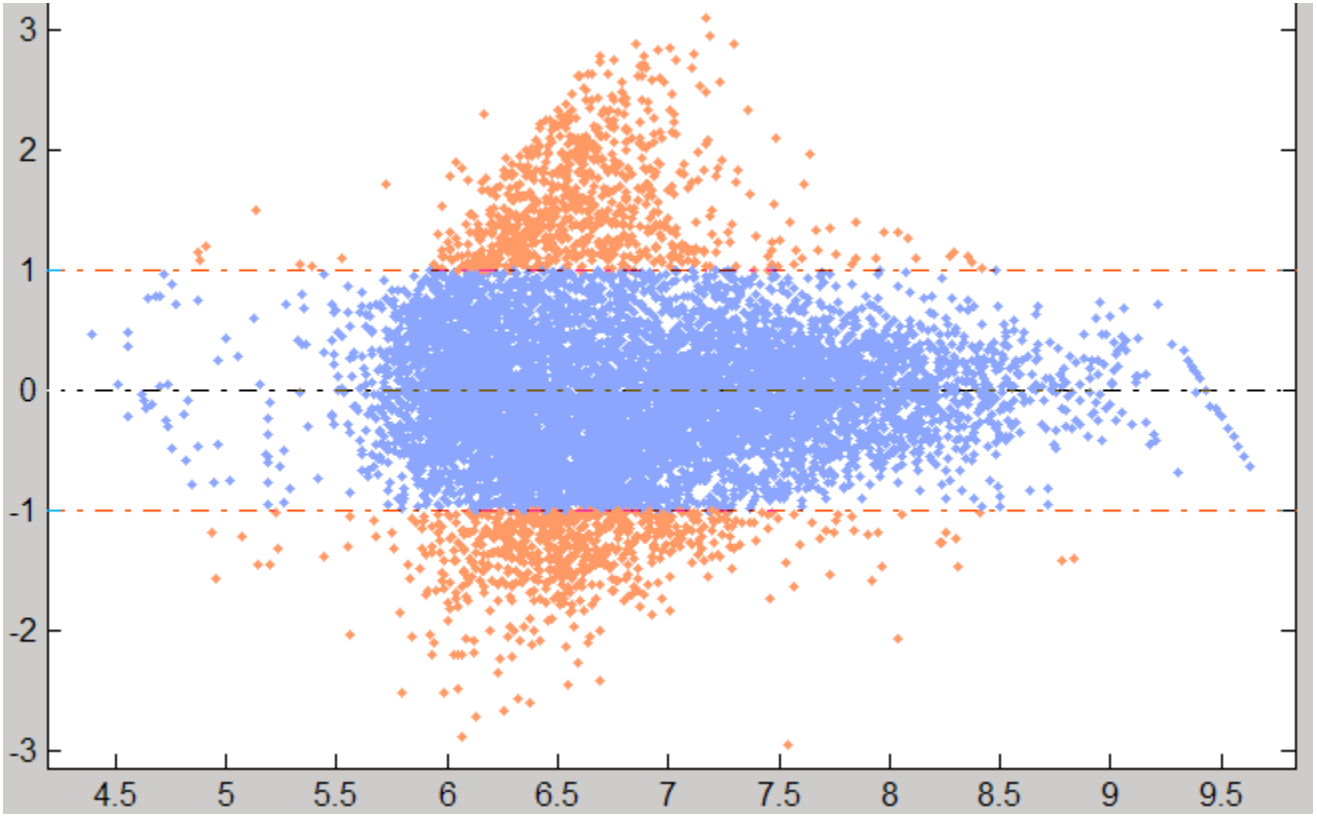
Jedes Gen wird durch mehrere (z.B. 11) Sonden / Spots gemessen.
Idealerweise liefern alle nach Normalisierung denselben Wert.
Das ist aber in der Praxis nicht so.
Die 11 Intensitätswerte werden zu einem Expressionswert zusammengefasst.

Methoden:

- Mittelwert
- Median
- Modelle, die die Hybridisierungsstärke der Sonden berücksichtigen.

Identifikation differenziell exprimierter Gene (Kandidaten-Gene)

MA-Plot („Minus gegen Average“):
Differenz gegen Durchschnitt aller Genexpressionswerte in zwei Experimenten.



in Experiment 2 mindestens doppelt so stark exprimiert

nicht differenziell

MIAME - Standard

MIAME: Minimal Information About a Microarray Experiment
Richtlinien der MGED (Microarray and Gene Expression Data) Society
URL: http://www.mged.org/Workgroups/MIAME/miame_2.0.html

Ziel: Informationen zu Microarray Studien in Datenbanken
sollen langfristig nutzbar und interpretierbar bleiben.

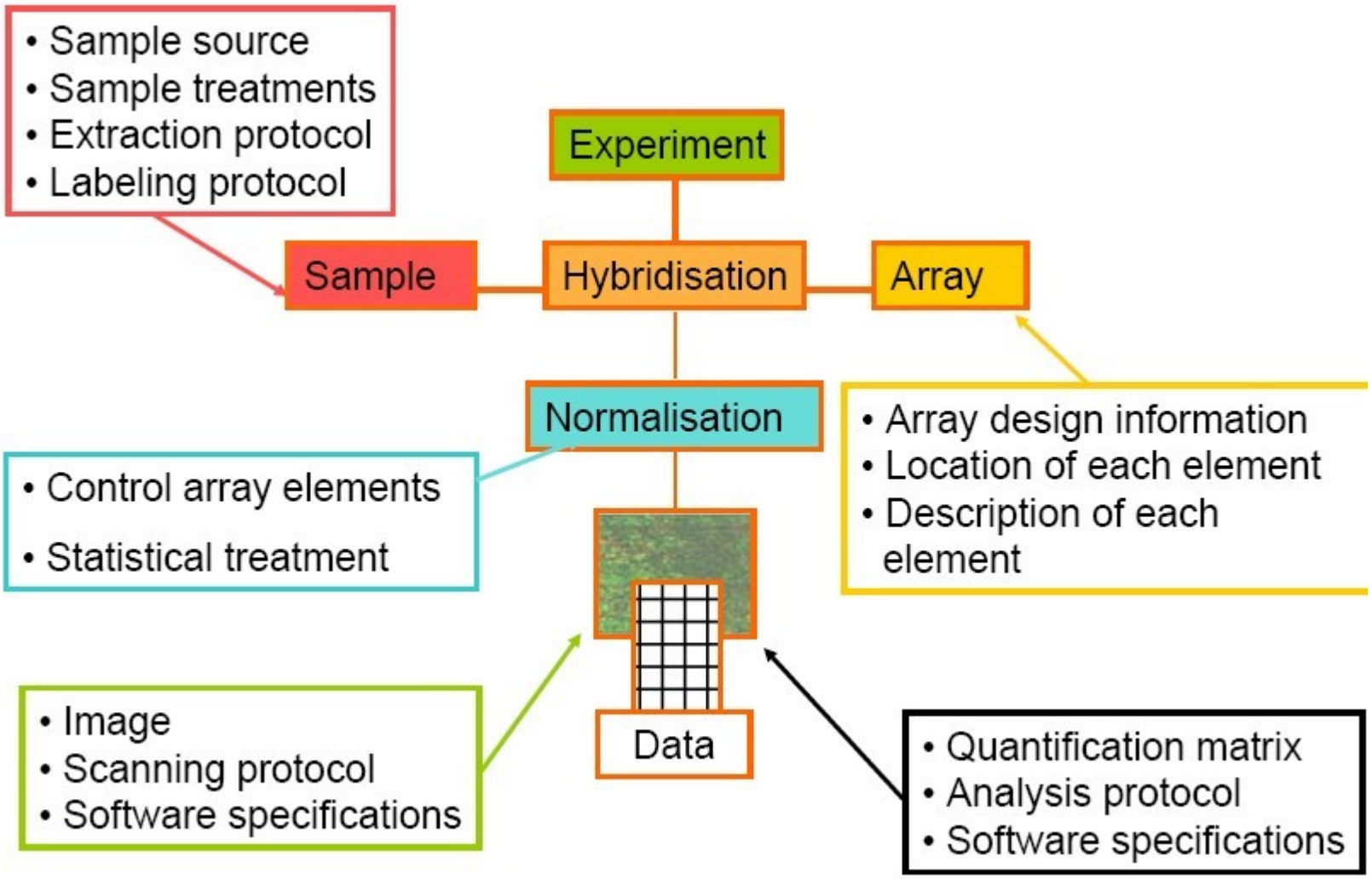
Verlangt werden Informationen zu 6 Bereichen.

Nach MIAME verlangte Informationen

6 Bereiche:

1. Rohdaten (Bild-Daten, unbearbeitet).
2. Prozessierte Daten (Matrix mit Expressionswerten pro Gen und Experiment).
3. Daten zu den Proben (samples):
   Woher stammen die Proben (z.B. Tumorgewebe / gesundes Gewebe);
   wie wurden in jedem Experiment die Proben behandelt?
4. Daten zum experimentellen Design:
   Welche Beziehungen bestehen zwischen den Experimenten einer Studie,
   zwischen verschiedenen samples ?
   Welche Rohdaten gehören zu welchem sample?
5. Daten zum Array-Design:
   Entweder welches Array von welchem Hersteller,
   oder bei Eigenentwicklungen Liste der DNA-Sequenzen aller Sonden.
6. Protokolle, sowohl experimentell, als auch zur Datenanalyse.

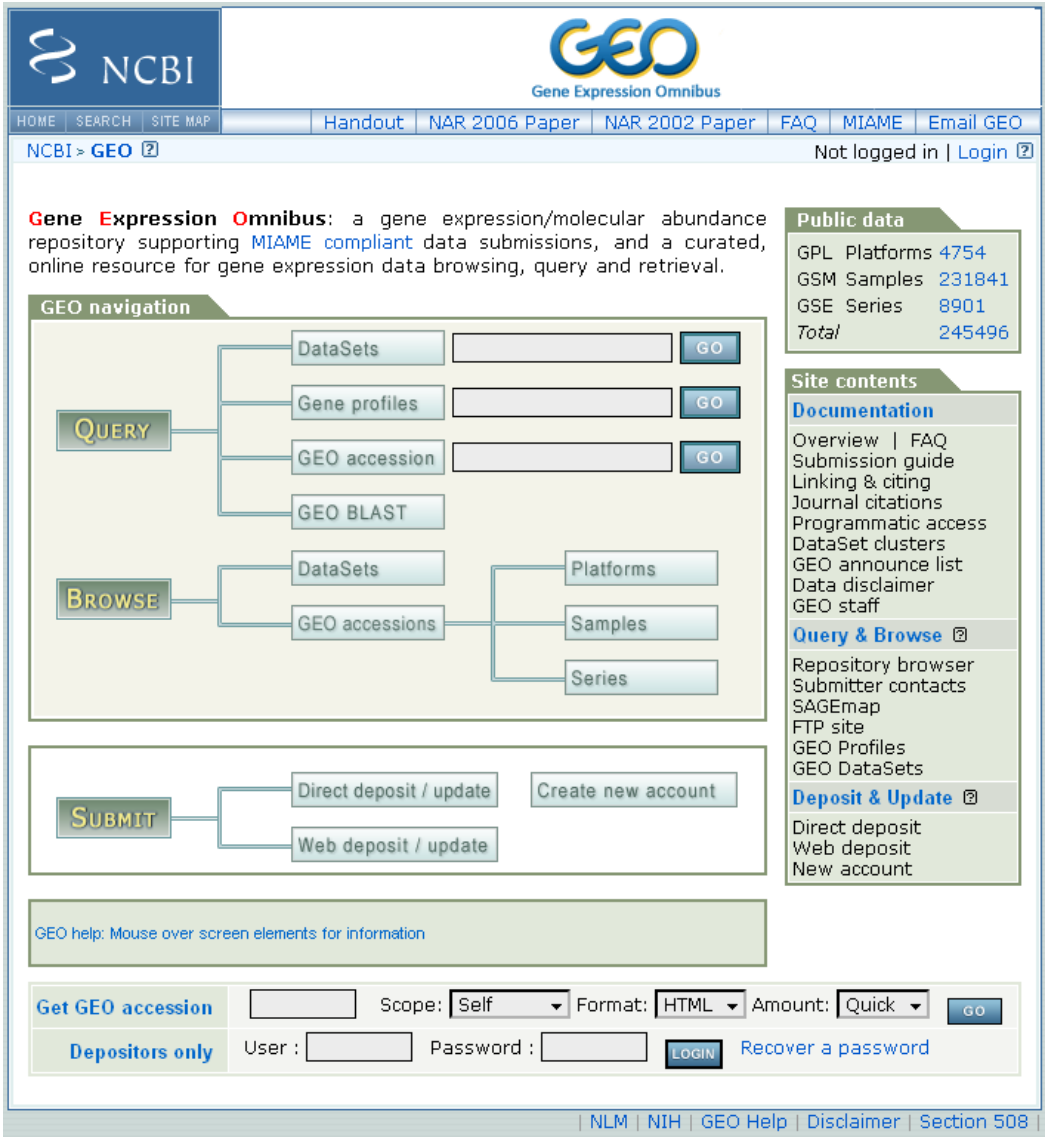
Nach MIAME verlangte Informationen



GEO

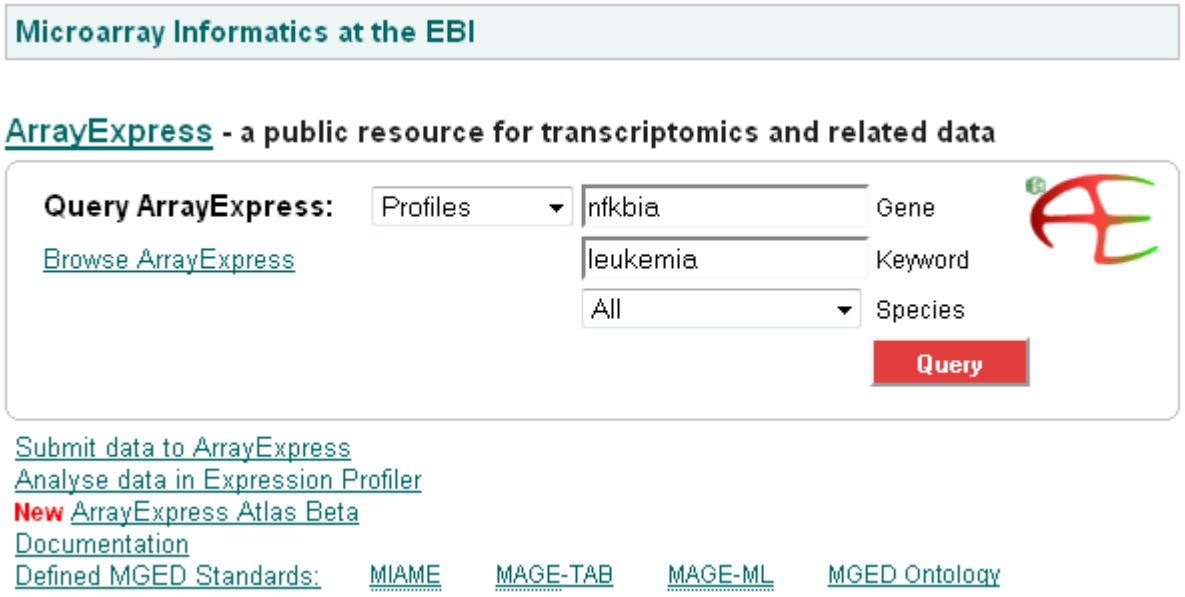
Gene Expression Omnibus

- MIAME-konforme Microarray-Datenbank
- am NCBI
- http://www.ncbi.nlm.nih.gov/geo/



ArrayExpress

- MIAME-konforme Microarray-Datenbank
- am EBI
- http://www.ebi.ac.uk/microarray/



Ein GEO-Beispiel

Accession Number GSM120719
URL: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM120719

Standard Affymetrix Experiment; Homo sapiens; Muskelgewebe.
Prozessierte Daten, ein Intensitätswert pro Transkript: einfache Textdatei.
3 Spalten: Transkript-ID; Intensitätswert; Present/Absent - Entscheidung

ID_REF	VALUE	ABS_CALL
200000_s_at	7844.43408203125	P
200001_at	31178.318359375	P
200002_at	65324.34375	P
200003_s_at	60674.9921875	P
200004_at	28636.283203125	P
200005_at	16733.904296875	P
200006_at	39892.4765625	P
200007_at	23791.220703125	P
200008_s_at	10948.0546875	P
200009_at	18589.80078125	P
200010_at	48697.41015625	P
200011_s_at	3656.510009765625	P
...		

Software für die Analyse von Microarray-Experimenten

Anbieter von Analyse-Software

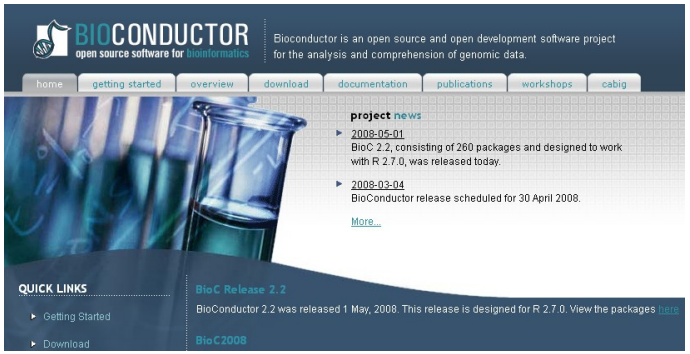
- Gerätehersteller
- Kommerzielle Anbieter (spezialisierte Firmen)
- Forschergruppen, oft freie Software

Was ist Freie Software?

- „free as in speech“
- „free as in beer“

Bioconductor für R: Paketsammlung, Freie Software.

- URL: http://www.bioconductor.org/
- u.a. Funktionen für Microarrays



Erinnerung zu R: Test auf Normalität

Gegeben: Daten (Vektor) x
Frage: Stammt x aus einer Normalverteilung?

Kann Folgendes tun:

```
y=rnorm(10000)
qqplot(x,y)   # Gerade?
```

Einfacher:

keine Zufallszahlen ziehen, sondern mit theoretischen Quantilen vergleichen

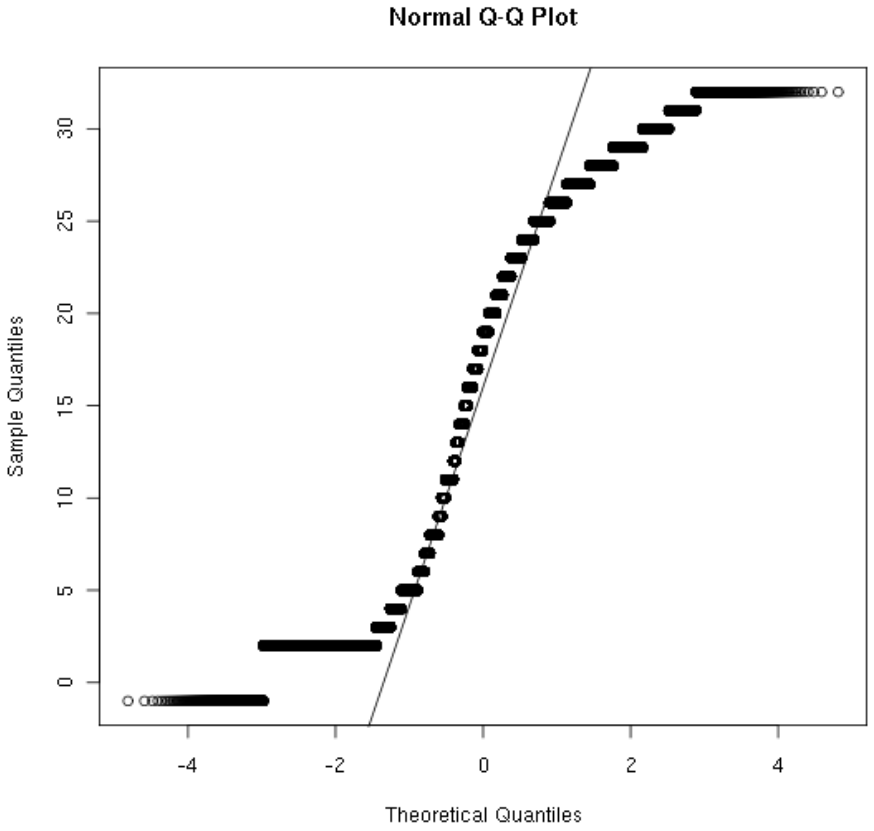
```
qqnorm(x)     # Gerade ??
qqline(x)     # Beste Gerade dazu einzeichnen!
```

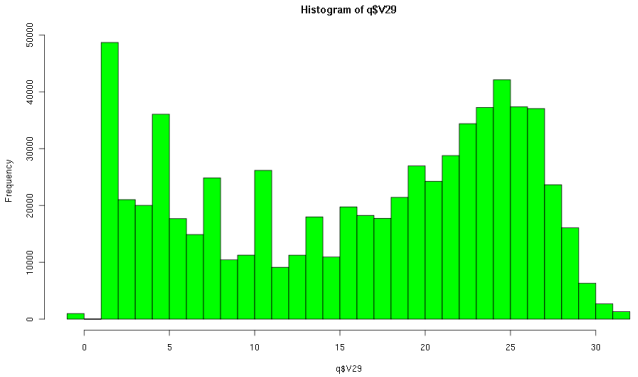
Beispiel:

Sequenzierdaten; Qualitätswerte an Position 29 normalverteilt?

Test auf Normalität mit qqnorm()

```
qqnorm(q$V29) # Gerade??
qqline(q$V29) # beste Gerade dazu
```





Warum nicht Histogramm mit Dichte dnorm() vergleichen?

- qqplot zeigt Abweichungen besser; invariant bei Verschiebung, Skalierung

Tests: Vergleich von normalverteilten Stichproben (Vektoren)

Fragestellung:
Gegeben zwei Datensätze (Vektoren),
haben die Verteilungen, aus denen sie stammen, gleichen Mittelwert?

Erste Idee und Problem:
Vergleiche empirische Mittelwerte (Durchschnitte) der beiden Vektoren.
Aber Durchschnitte sind nie exakt gleich!
(Auch nicht, wenn aus exakt der gleichen Verteilung gezogen wurde.)

Experiment:
Ziehe zweimal je 10 Werte aus Standard-Normalverteilung (rnorm(10)).
Berechne Durchschnitt (mean). Differenz der Durchschnitte ungleich Null!

Also:
Gewisse (kleine) Unterschiede des Mittelwerts zwischen Stichproben
aus derselben Verteilung sind kein Indiz für verschiedene Mittelwerte!

Test auf gleichen Mittelwert (bei Normalverteilung)

Nullhypothese:
Stichproben stammen aus Normalverteilungen mit gleichem Mittelwert.

Alternative:
Stichproben stammen aus Normalverteilungen mit verschiedenen Mittelwerten.

Frage, die die Statistik beantworten kann:
Angenommen, die Nullhypothese trifft zu.
Wie wahrscheinlich ist es, dass sich die beobachteten Mittelwerte
um mindestens so viel wie die beobachtete Differenz unterscheiden?
Diese Wahrscheinlichkeit nennt man p-value (p-Wert).
(Für nicht normalverteilte Daten wird hier nichts ausgesagt!)

Anwendung:
Man gibt eine Grenze (Signifikanzniveau) vor (z.B. 0.05 oder 0.01).
Ist der p-Wert <= 0.05, sagt man: „Der Unterschied ist signifikant“.
Ist der p-Wert <= 0.01, sagt man: „Der Unterschied ist hoch signifikant“.
(Wahrscheinlichkeit für einen so großen Unterschied ist bei Nullhypothese klein!)

Der t-Test

Vergleich von zwei normalverteilten Stichproben *x*, *y* heißt
Zwei-Stichproben-t-Test.

Man „darf“ diesen Test nur auf (approximativ) normalverteilte Daten anwenden.
(Wenn man das nicht beachtet, ist das Ergebnis bedeutungslos.)

In R:
Zuerst x,y auf Normalverteilung prüfen (z.B. mit qqnorm).
Dann: `t.test(x,y)`

... liefert viele Informationen; wichtig ist der p-value.

(Aufgabe der mathematischen Statistik und Wahrscheinlichkeitsrechnung
ist Erfinden solcher Tests und exakte Berechnung der p-values.)

Der t-Test bei Microarrays

Situation:
Microarray-Experimente von 10 Tumor-Proben und 100 Kontroll-Proben von ca. 25,000 Genen liegen vor.
Betrachte Genexpression in beiden Klassen (10 und 100 Werte).
Gibt es einen signifikanten Unterschied zwischen den Klassen?

Lösung:
t-Test für jedes Gen (sofern die 10 und 100 Werte normalverteilt sind).

Aber:
Wahrscheinlichkeit, dass ein Gen-p-Wert <= *x* ist,
obwohl kein Unterschied besteht, ist *x* (nach Definition des p-Werts).
Wir testen viele Gene (~25,000).
Da ist „durch Zufall“ schon ein p-Wert <= 1/25000.
Das bedeutet noch nichts.

Korrektur für multiples Testen

Wahrscheinlichkeit, dass ein Gen-p-Wert <= *x* ist,
obwohl kein Unterschied besteht, ist *x* (nach Definition des p-Werts).
Wir testen viele Gene (~25,000).
Da ist „durch Zufall“ schon ein p-Wert <= 1 / 25000.
Das bedeutet noch nichts.

Bonferroni-Korrektur:
Multipliziere p-Werte mit Anzahl der Gene.
Betrachte dann p-Werte <=0.05 (signifikant) bzw. 0.01 (hoch signifikant).