Bio Data Management

Kapitel 6

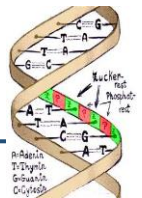
Genexpressionsanalyse

Wintersemester 2014/15
Dr. Anika Groß
Universität Leipzig, Institut für Informatik, Abteilung Datenbanken
http://dbs.uni-leipzig.de

UNIVERSITÄT LEIPZIG

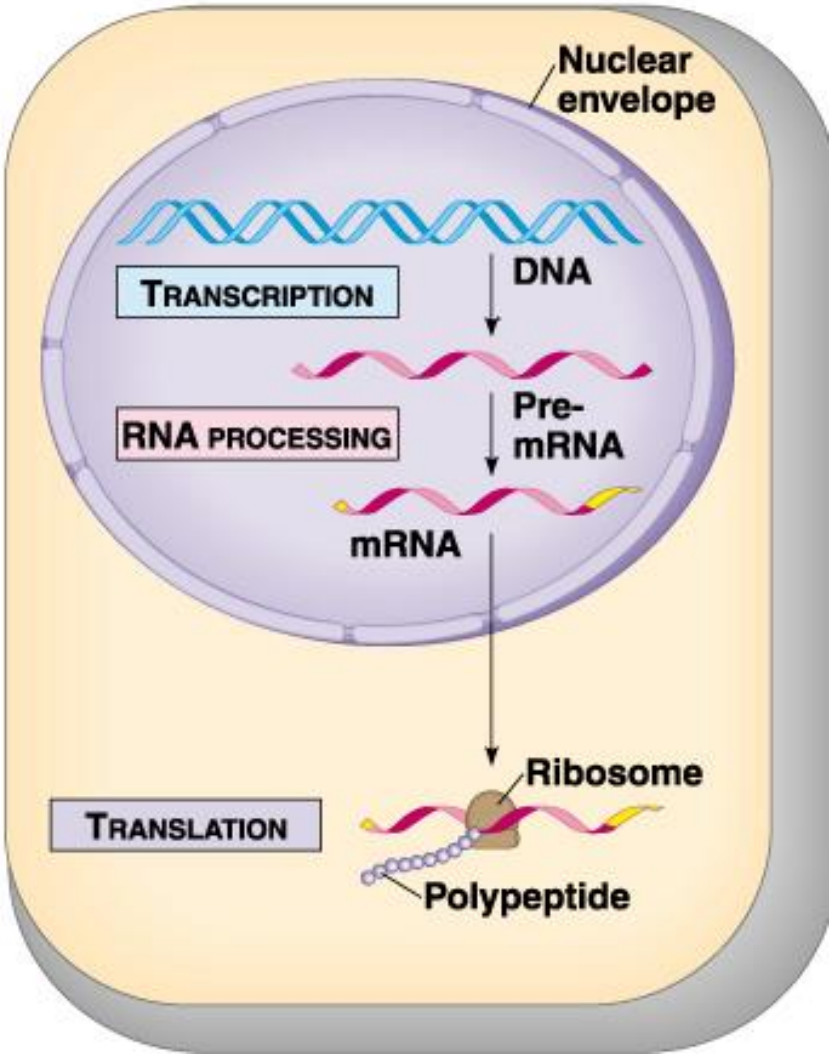
Vorläufiges Inhaltsverzeichnis

1. Motivation und Grundlagen
2. Bio-Datenbanken
3. Datenmodelle und Anfragesprachen
4. Modellierung von Bio-Datenbanken
5. Sequenzierung und Alignments
6. Genexpressionsanalyse
7. Annotationen
8. Matching
9. Datenintegration: Ansätze und Systeme
10. Versionierung von Datenbeständen
11. Neue Ansätze



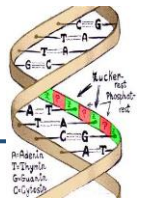
Genexpression

- Was ist Genexpression?
  - Aktivierung der Gentranskription durch endogene und exogene Einflüsse
  - Ausbildung der einem Gen inhärenten Eigenschaften
- Ziel der Genexpressionsanalyse
  - Charakterisierung der Funktion von Genen, deren Abhängigkeiten, Interaktionen und Einfluss in verschiedenen Netzwerken (metabolische N., regulatorische N. etc.)
- Annahme: Genexpression korreliert mit Proteinsynthese



Messung der Genexpression

- Messung der mRNA Konzentration in Zellen unter verschiedenen Bedingungen
  - Gesundes vs. krankes Gewebe
  - Verschiedene Zell-/Gewebetypen
  - Entwicklungsstadium: Embryo, Kind, Erwachsener …
  - Umwelt: Einfluss von Hitze, Ernährung, Therapie …
  - Subtypen einer Krankheit (z.B. teilweise Chemotherapieresistenz)
- Suche nach Genen mit gleicher Expression (Koexpression, →) bzw. differenzieller Expression (↑, ↓)
- Co-Regulation
  - Ähnliches Genmuster → ähnliche Funktion?
  - Ähnliches Genmuster → ähnliche Regulation?
- Techniken: Northern Blotting, SAGE, Microarray …

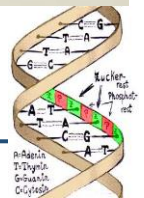


Von der objekt- zur genomweiten Messung

- "Gene expression profiling"
  - Idee: Messung der Aktivität von Genen / Transkripten unter verschiedenen Bedingungen
  - Chip-Plattformen: Gen ~, Exon ~, Tiling arrays, …
- "Mutation profiling"
  - Idee: Messung der genetischen Variabilität der Genomsequenz
  - Chip-Plattformen: Matrix-CGH ~, SNP arrays, …
- "Sequence profiling"
  - Idee: Sequenzierung des Genomes eines Probanden
  - Chip-Plattformen: Roche 454, Illumina Solexa, …

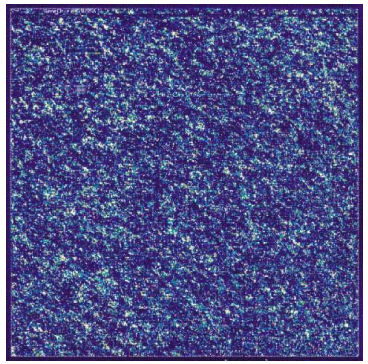
Affymetrix gene expression microarray

Bildquelle: http://www.affymetrix.com



Microarrays

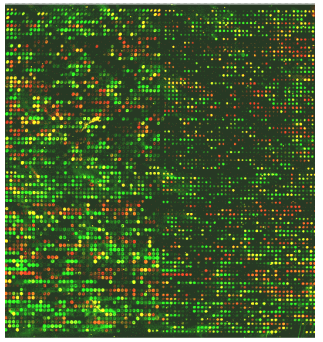
- Parallele Analyse von mehreren tausend Einzelnachweisen/Genen
- Nur geringe Menge an biologischem Probenmaterial nötig
- Auch Gen-/Biochip genannt (ähnlich Computerchip)
- cDNA Arrays, Oligo Arrays
  - Einzelstränge
  - Unterscheidung nach Sequenzart, Sequenzlänge
- Sonden werden an definierten Positionen eines Rasters z.B. auf einem Glasträger aufgebracht werden
- Hersteller: Affymetrix, Agilent, Rosetta Inpharmics etc.



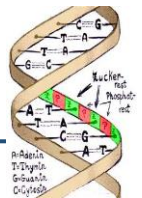
Oligonucleotide Microarrays, einfarbig

Verteilung von Extrakten aus Untersuchungsgewebe und Kontrolle auf einem oder mehreren Chips

Bildquellen:
← http.//www.affymetrics.com

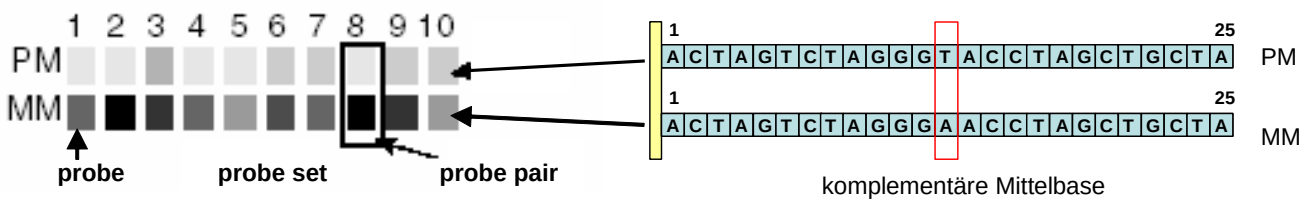


DNA Microarrays, zweifarbig



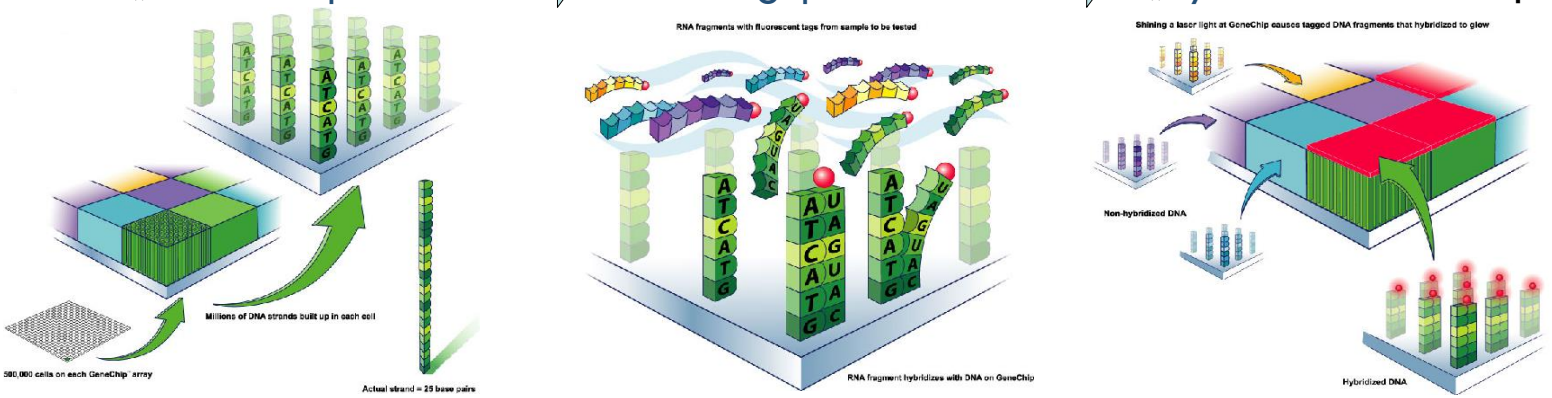
Affymetrix GeneChip Technologie

- Verschiedene Hersteller und Chiptypen
  - Abbildung unterschiedlicher Spezies, Transkriptregionen
- Terminologie:



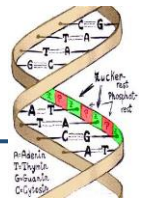
- Hybridisierungsprozess (stark vereinfacht)

„Roh"-Chip → Bindungsprozess → „hybridisierter" Chip



Bildquelle: Affymetrix, http://www.affymetrix.com

1) **2)** **3)**



Hybridisierungsprozess

- **Hybridisierung** ist der Prozess zwei komplementäre DNA- oder RNA-Einzelstränge zu einem Doppelstrang zusammenzuführen/ zu vereinigen.

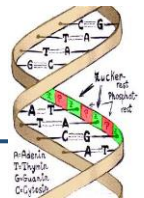
1) 500000 Zellen (Bereiche) auf jedem Gen-Chip-Array

Millionen von DNA Strängen in jeder Zelle – Redundanz!!!

Eigentlicher Strang ca. 25bp

2) RNA Fragmente mit fluoreszenz-markierten Tags aus dem zu testenden Sample → RNA Fragment hybridisiert mit DNA auf Gen-Chip

3) Bestrahlung des Gen-Chip durch Laser
→ hybridisierte, fluoreszenz-markierte DNA-Fragmente beginnen zu leuchten



Ablauf

Probe

Sample

Reporter, Auf Chip fixiert

RNA-Hybridisierung

Auswaschen des Chips

Scannen

TIFF-Bild

Bilderkennung

Rohdaten

Analyse

Sample Vorbereitung, Labeling

gelabelte Targets
→ „Zu untersuchendes Material"

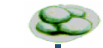
Sample-cDNA/RNA hybridisiert mit Probe-cDNA

Entfernen nicht gebundener Sample-cDNA/RNA

Scannen des Arrays mittels Laserstrahl

Erkennung der Lichtintensitäten und Bildeinteilung/ -segmentation

Normalisierung und Datenanalyse (Data Mining)

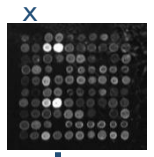


sample

mRNA

labeling

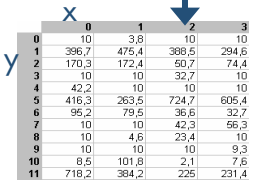
chip



x

y

array image

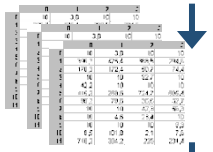


x

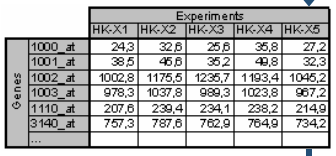
y

	0	1	2	3
0	10	3,8	10	10
1	396,7	475,4	388,5	294,6
2	170,3	172,4	50,7	74,4
3	10	10	32,7	10
4	42,2	10	10	10
5	416,3	263,5	724,7	605,4
6	95,2	79,5	36,8	32,7
7	10	10	42,3	56,3
8	10	4,6	23,4	10
9	10	10	10	9,3
10	8,5	101,8	2,1	7,6
11	718,2	384,2	225	231,4

arrayspot intensities

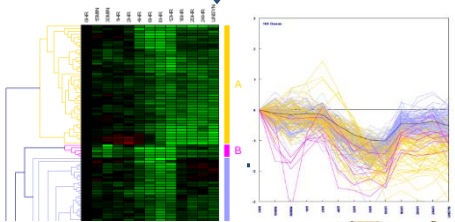


experiment series



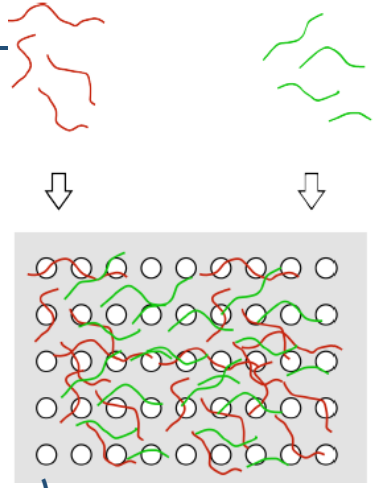
		Experiments				
		HK-X1	HK-X2	HK-X3	HK-X4	HK-X5
Genes	1000_at	24,3	32,6	25,6	35,8	27,2
	1001_at	38,5	46,6	35,2	49,8	32,3
	1002_at	1002,8	1175,5	1235,7	1193,4	1045,2
	1003_at	978,3	1037,8	989,3	1023,8	967,2
	1110_at	207,6	239,4	234,1	238,2	214,9
	3140_at	757,3	787,6	762,9	764,9	734,2
	...					

gene expression matrix



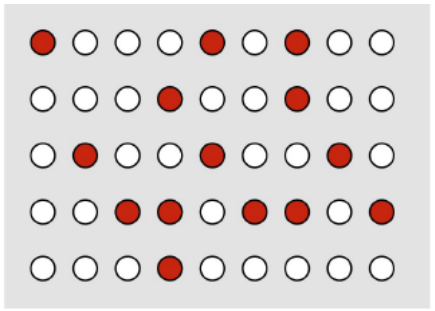
Teilweise aus „Analysis of gene expression data" (Ulf Leser, Philip Thomas)

DNA-Microarrays

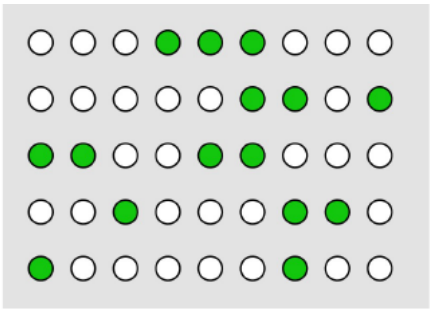


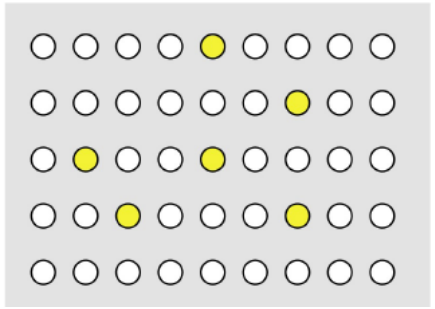
- Zugabe von rot (Cy5) und grün (Cy3) markierten (Fluoreszenzfarbstoff) Untersuchungsproben (sample)
  - Binden bei komplementärer Basenabfolge an die DNA im Chip
  - Kontrolle versus „Stress“ (z.B. gesundes versus krankes Gewebe)

Stress



Kontrolle

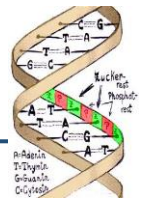




Gelb (Grün&Rot)

→Exprimiert in Kontrolle und Stress

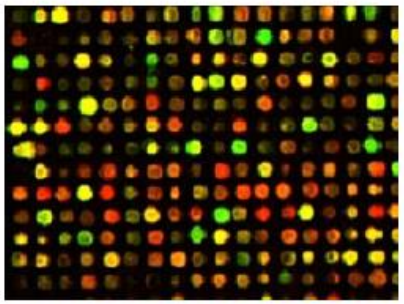
Quelle: Dion Whitehead, Division of Bioinformatics,Westfälische WilhelmsUniversität Münster, http://dionamago.net/content/courses/2006.06.12.BioinformatikII.Microarrays.Improved.GERMAN.pdf



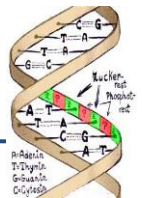
DNA-Microarrays

- rot:grün-Verhältnis entspricht der „relativen Expression“
- Beispiel: der Wert 2,5 von Gen G bedeutet, dass dessen Expression im untersuchten „Stress“-Fall 2,5 mal höher ist als unter normalen Bedingungen

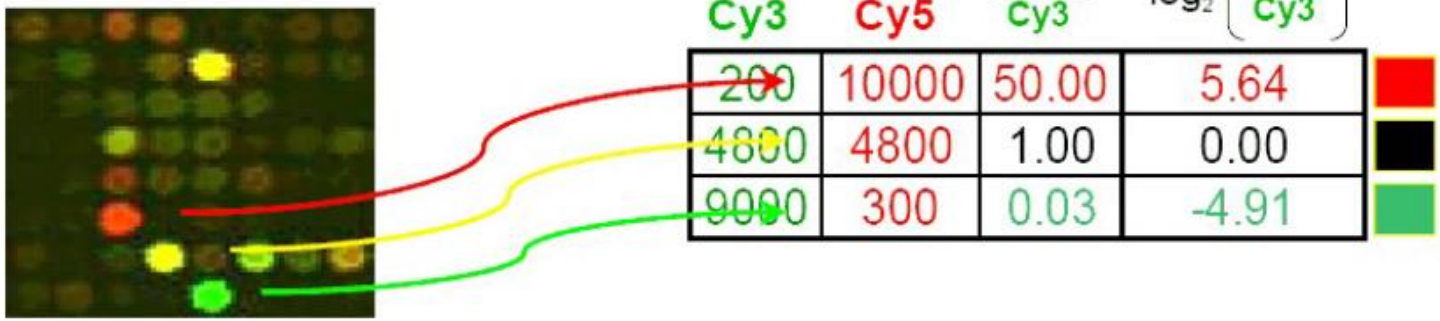
Gene Name	Rot:Grün Verhältnis
A	1,2
B	0,1
C	-3,2
D	2,3
E	-0,5
F	8,3
G	2,5
F	8,3
G	2,5
.	.
.	.
.	.



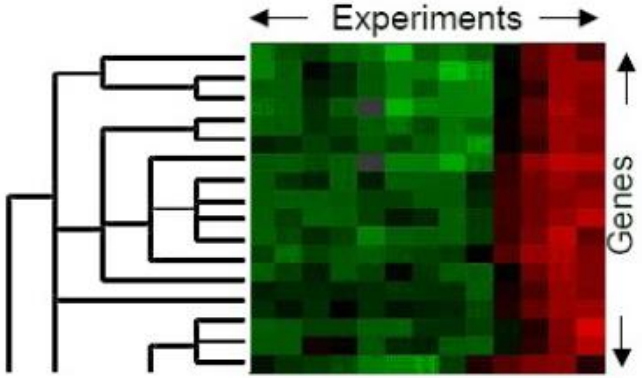
Quelle: Dion Whitehead, Division of Bioinformatics,Westfälische WilhelmsUniversität Münster, http://dionamago.net/content/courses/2006.06.12.BioinformatikII.Microarrays.Improved.GERMAN.pdf

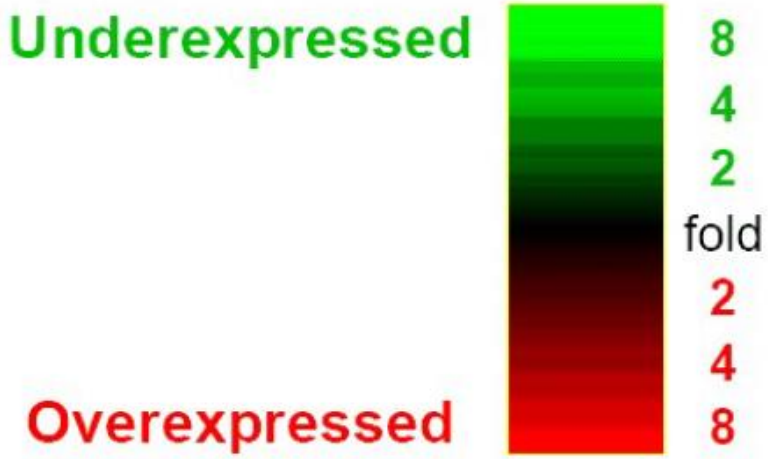


Analyse und Visualisierung



Cy3	Cy5	$\frac{Cy5}{Cy3}$	$\log_2 \frac{Cy5}{Cy3}$
~~200~~	10000	50.00	5.64
~~4800~~	4800	1.00	0.00
~~9000~~	300	0.03	-4.91





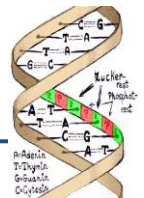
Quelle: Dion Whitehead, Division of Bioinformatics,Westfälische Wilhelms-Universität Münster, http://dionamago.net/content/courses/2006.06.12.BioinformatikII.Microarrays.Improved.GERMAN.pdf

Genexpressionsanalyse: Rohdatentransformation

- Ziel: Generierung von Expressionswerten per Gen auf Basis bereinigter, normalisierter Messwerte von Probenintensitäten

- **Grundannahme: Mehrheit der Gene ist nicht differentiell exprimiert**

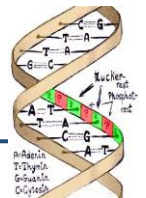
- Kein Standard für Datentransformationsschritte (z.B. verschiedene Normalisierungsverfahren)
  → Speicherung der Rohdaten notwendig



Messfehler

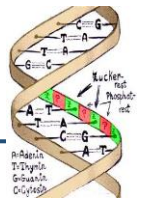
- Systematische Fehlerkomponente
  - Temperatureinfluss, falsch geeichte Messinstrumente, …
- Zufällige Fehlerkomponente
  - Nicht beherrschbare Einflüsse aus der Umgebung, uneinheitliche Ablesung von einer Skala (Winkel) …
  - Abschätzung aus genügend großer Anzahl von Einzelmesswerten (Wiederholung der Messung, Messreihen)

- Additiver Fehler
  - Konstant-systematische Messfehler: Offset
  - Hintergrundbereinigung (Background subtraction): Bereinigung der Probenintensität um messtechnisches Signallevel
- Multiplikativer Fehler
  - Ansteigend/abfallende Messfehler: Trend (bias), Drift
  - Normalisierung: Ausgleich von Niveauunterschieden zwischen Daten verschiedener Experimente



Normalisierung von Microarray-Daten

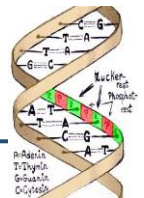
- Ein ideales Experiment bräuchte keine Normalisierung
- Normalisierung von Microarray-Daten innerhalb eines Arrays zur Kontrolle eines systematischen Bias
  - Trend, Messung hat „Schlagseite“
  - Fehlerquellen: Neigung/Winkel bei der Bildaufnahme (Laser), Farbstoffe fluoreszieren unterschiedlich stark, unterschiedliche Intensität bei ungleichmäßiger Array-Beschichtung, Hybridisierungseffizienz, …
- Messwerte zwischen zwei Experimenten sind nicht direkt vergleichbar
- Einfluss systematischer Fehler verringern
- Minimieren der Variationen
  → Finden der tatsächlichen biologischen Unterschiede
- Ist Normalisierung notwendig?
  - Einfache Möglichkeit: Grün gegen rot plotten → Anstieg sollte 1 sein
  - MA-Plot



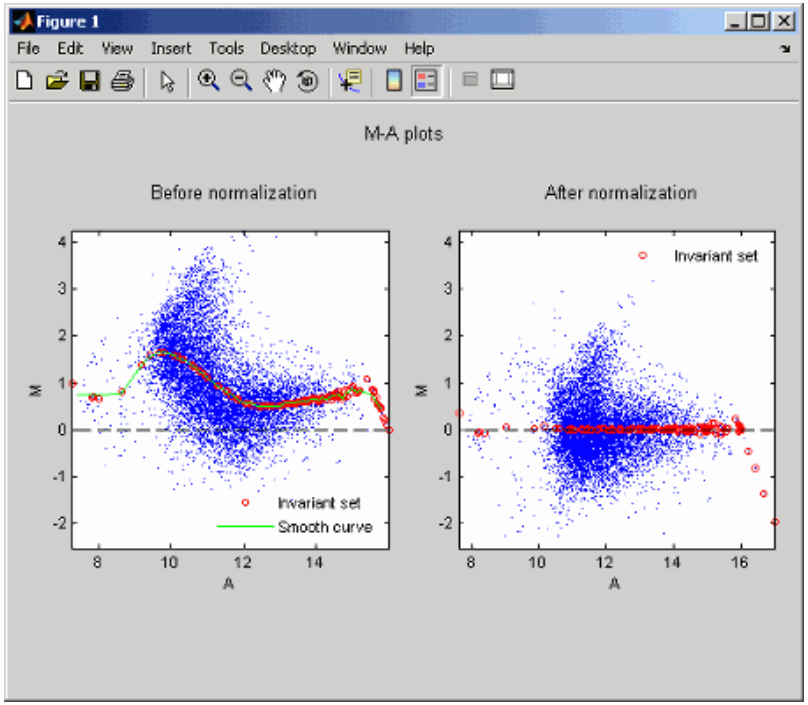
M/A Plot

- Qualitätskontrolle - gibt es einen Bias?

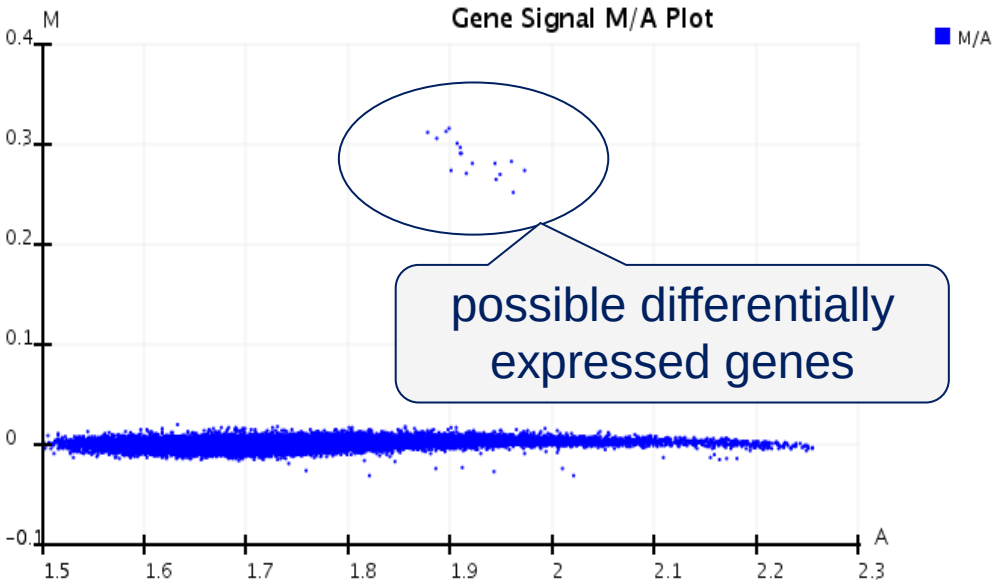
- M (y-Achse) geplottet gegen A (x-Achse)
- R= Rot-Intensität, G=Grün-Intensität
- $M = log_2\left(\frac{R}{G}\right)$
  - Unterschied/Differenz zwischen log-Intensitäten
- $A = \frac{1}{2}(log_2(R) + log_2(G))$
  - Durchschnittsintensität eines Punktes im Plot
- Meist keine Änderung: $M = \log(1) = 0$
  - Viele Werte nahe 0 (y-Achse)



M/A Plot



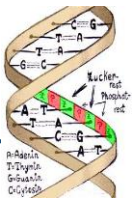
http://www.mathworks.com/help/toolbox/bioinfo/ref/function_mainvarsetnorm_plot2.gif



Normalisierung *

- Annahme: mRNA-Konzentration ist in jeder Zelle gleich
  - Messung der Gesamt-mRNA Menge
  - Dividiere Intensitäten durch diesen Wert
- Annahme: ähnliche Expression von Referenzgenen über alle Gewebe
  - Auswahl der „Housekeeping“ Gene
  - Dividiere Intensitäten durch diesen Wert
- Nicht lineare Methoden
  - LOWESS (LOESS, *locally weighted scatterplot smoothing*): lokale, gewichtete, polynomielle Regression
  - Quantil-Normalisierung
  - …

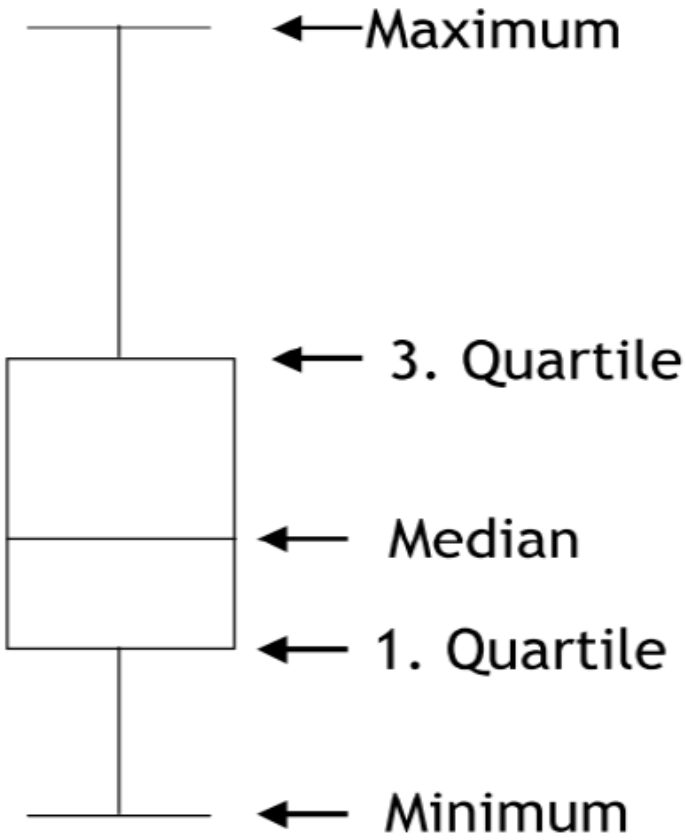
* Folien zur Normalisierung „Analysis of gene expression data“ von Ulf Leser und Philip Thomas



Quantil-Normalisierung

Erinnerung/Grundlage

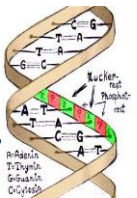
- Quantile
  - zur Einteilung der Verteilung aller Werte in gleich große Abschnitte
  - Beispiele
    - Terzil: unteres, mittleres und oberes Drittel
    - Quartil: jeweils $\frac{1}{4}$ der Werte in einem Bereich
    - Median: 50% aller Werte liegen rechts bzw. links vom Median



Quantil-Normalisierung

- Annahme: die Verteilung der Genexpression ist bei allen Experimentwiederholungen ungefähr gleich
- Ziel: gleiche empirische Verteilung in jedem (zu vergleichenden) Array
- Meist besser als lineare Methoden

1) Matrix X: p x n
   - jedes Array ist eine Spalte
   - jedes Transkript / Gen ist eine Zeile
2) Sortiere jede Spalte von X separat $\rightarrow X_{sort}$
3) Weise jedem Feld in einer Zeile den jeweiligen arithmetischen Mittelwert der Zeile zu (X'_{sort})
4) Bestimme X_n durch Reorganisieren der Spalten von X'_{sort} um die gleiche Sortierung wie im Inputvektor zu erhalten



Quantil-Normalisierung

Sort by column

	Array 1	Array 2	Array 3
Gene1	1	6	8
Gene2	2	5	9
Gene 3	3	4	7

	Array 1	Array 2	Array 3
Gene1	1	4	7
Gene2	2	5	8
Gene 3	3	6	9

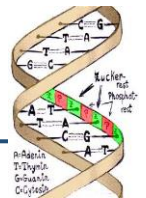
Row wise mean

	Array 1	Array 2	Array 3
Gene1	4	4	4
Gene2	5	5	5
Gene 3	6	6	6

Rearrange

	Array 1	Array 2	Array 3
Gene1	4	6	5
Gene2	5	5	6
Gene 3	6	4	4

Rearrange



Quantil-Normalisierung

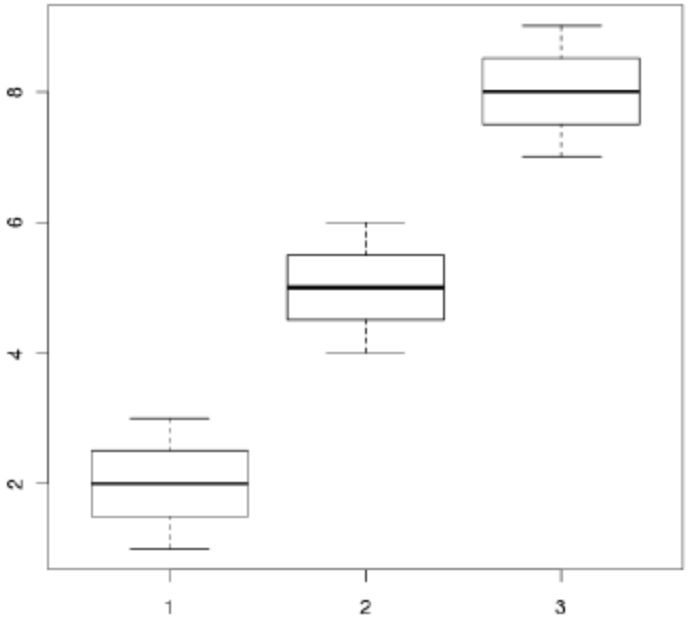
Before

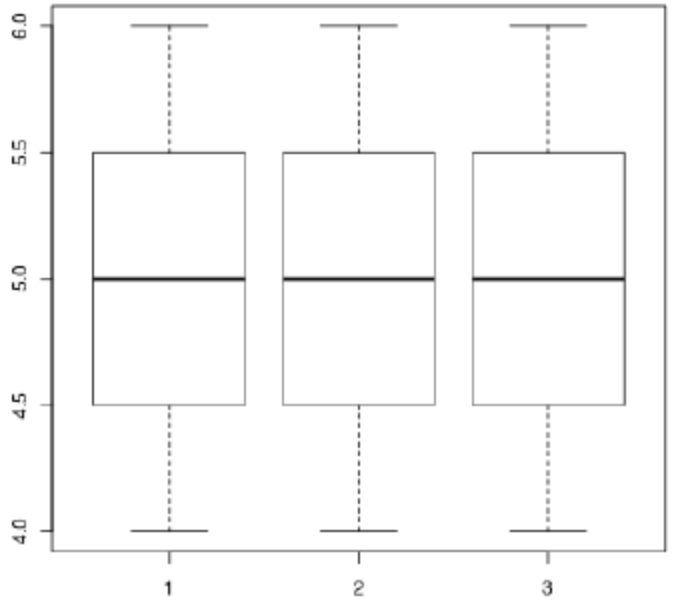
	Array 1	Array 2	Array 3
Gene1	1	6	8
Gene2	2	5	9
Gene 3	3	4	7

After

	Array 1	Array 2	Array 3
Gene1	4	6	5
Gene2	5	5	6
Gene 3	6	4	4

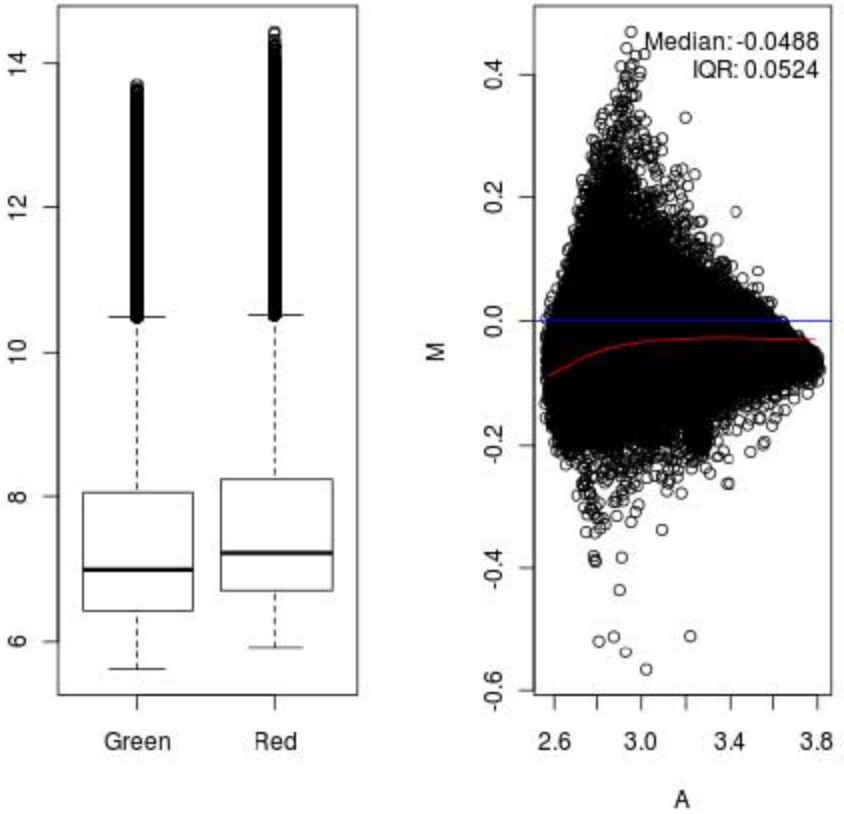
Jeder normalisierte Wert liegt in dem Vektor aus Mittelwerten im selben Quantil wie der ursprüngliche (nicht normalisierte) Wert.



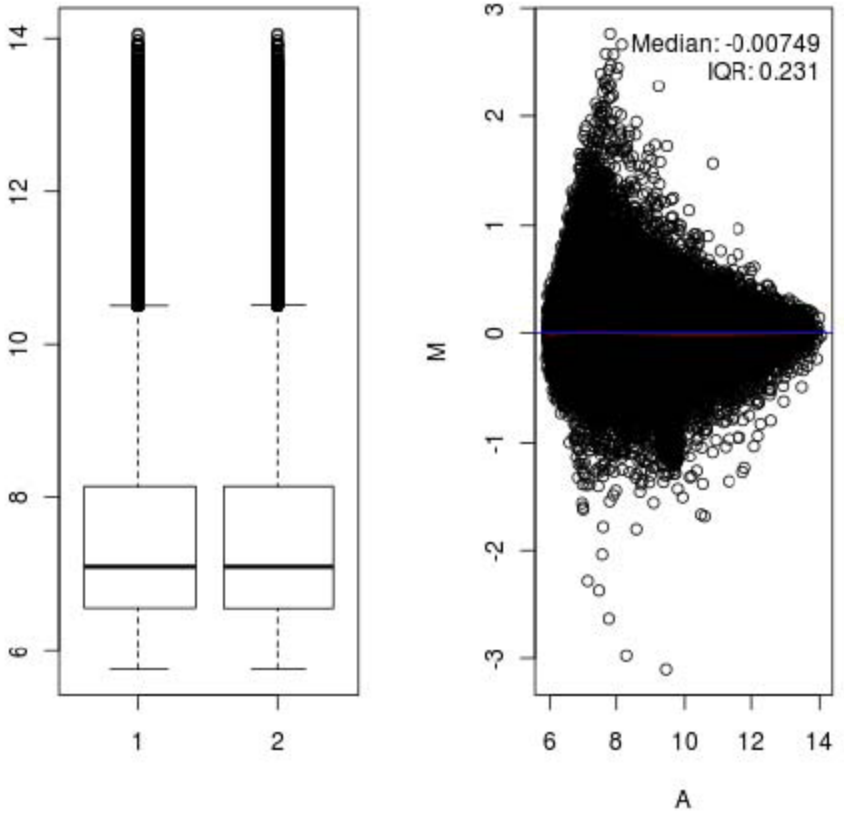


Quantil-Normalisierung

Before

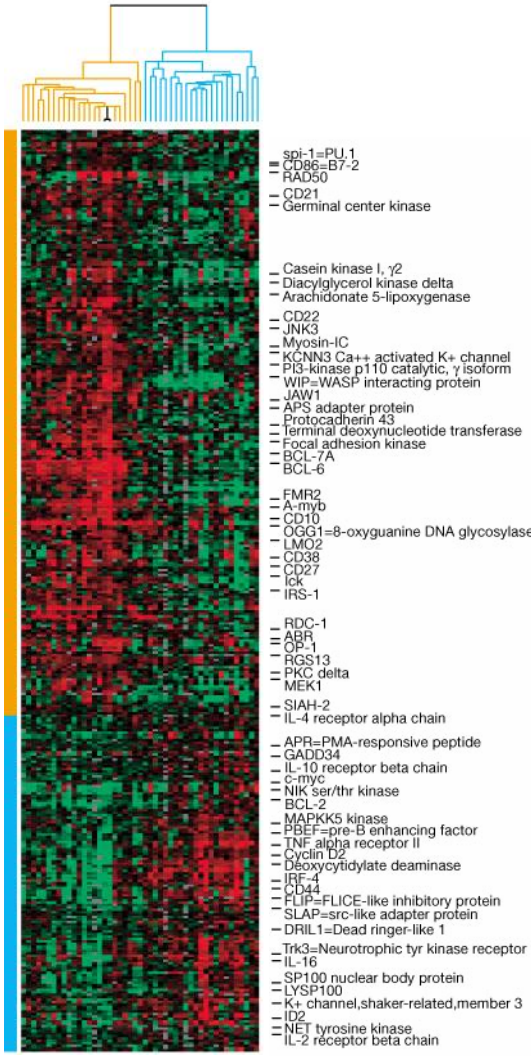


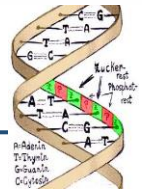
After



Genexpressionsanalyse und Visualisierung

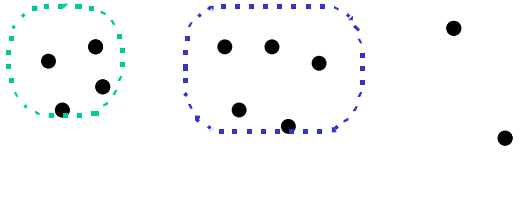
- Integrierte Berichte, parametrisierte Abfragen
  - Weit verbreitete Anwendung zur oberflächlichen Datensichtung (Überblick, Filterung)
- Großes Spektrum an statistischen Verfahren und Data Mining Ansätzen
  - Hauptkomponenten~, Korrespondenzanalyse
  - Clustering, Klassifikation
- Visualisierung
  - Tabellen, Genexpressionsmatrix
  - Dendrogramm, Heatmaps
  - Kombination mit Pathways
- Analyseintegration
  - Dateibasierter Datenaustausch
  - API basierter DB-Zugriff
  - DB-Integration (user defined functions, stored procedures)



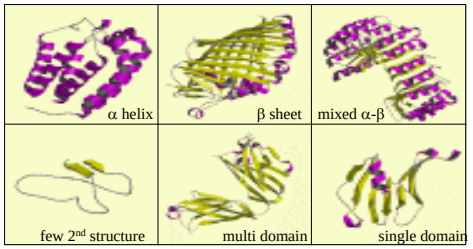


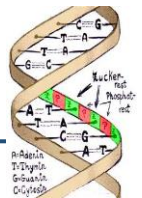
Data Mining-Techniken (Life Sciences)

- Clusteranalyse
  - Objekte (z.B. Proteine) werden aufgrund von Ähnlichkeiten in Klassen eingeteilt (Segmentierung)



- Assoziationsregeln
  - z.B. Genexpression vom Grad x bei Gen y => Hinweis auf Erkrankung z
  - Sonderformen zur Berücksichtigung von Dimensionshierarchien (z.B. Gengruppen), quantitativen Attributen, zeitlichen Beziehungen (sequence mining)
- Klassifikation
  - Zuordnung von Objekten (z.B. Proteinen) zu Gruppen/Klassen mit gemeinsamen Eigenschaften bzw. Vorhersage von Attributwerten
  - Explizite Erstellung von Klassifikationsregeln (z.B. "wenn Teilsequenz T dann Proteingruppe P")
  - Verwendung von Stichproben (Trainingsdaten)
  - Ansätze: Entscheidungsbaum-Verfahren, statistische Auswertungen (z.B. Maximum Likelihood-Schätzung / Bayes-Schätzer), neuronale Netze
- Weitere Ansätze:
  - Genetische Algorithmen (multivariate Optimierungsprobleme, z.B. beim Proteindesign)
  - Regressionsanalyse zur Vorhersage numerischer Attribute . . .

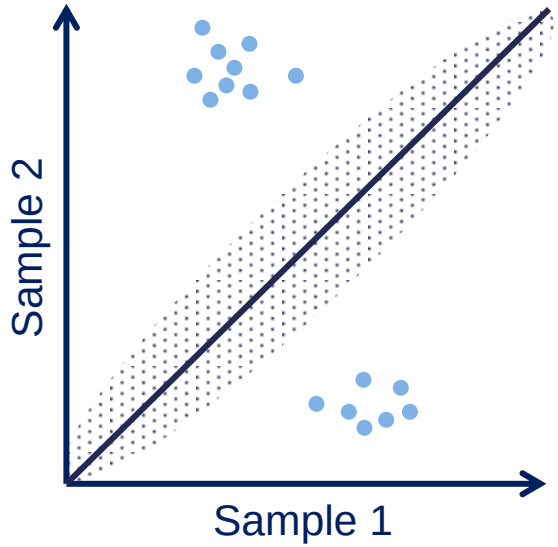




Differentielle Expression

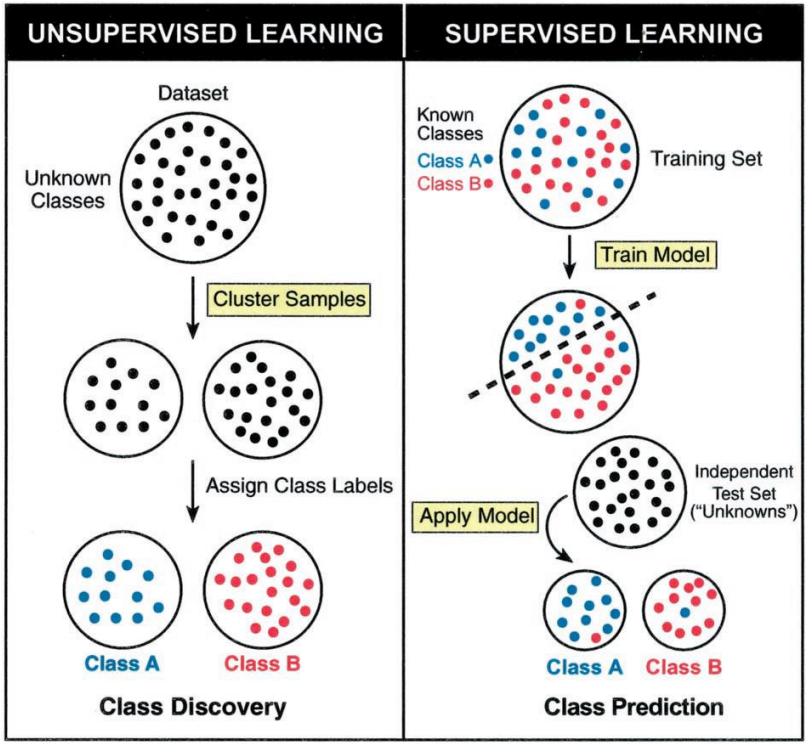
- Suche nach Genen, die signifikante Unterschiede in ihrer Expression aufweisen
- Vergleiche die Werte eines Gens in zwei verschiedenen Gruppen (z.B. krankes vs. gesundes Gewebe, zwei verschiedene Organismen, …)
- Fold change, T-test …

- Finde „Ausreißer“, welche nach Normalisierung (Ausschließen zufälliger und systematischer Fehler) noch vorhanden sind

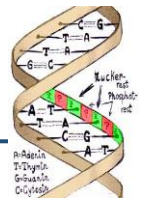


Klassifikation und Clustering

- Welche (Sub)gruppen von Genen repräsentieren ein biologisches Phänomen?
- Sind sie z.B. co-reguliert / co-exprimiert?
- Nicht überwachtes Lernen
  - *Zum Clustern*
  - SOM (self-organizing map)
  - K-means
  - Hierarchisches Clustern
  - Entscheidungsbäume
  - …
- Überwachtes Lernen
  - *Zur Klassifikation*
  - Training mithilfe von Eingabedaten + erwartete Ausgabewerte
  - SVM (support vector machine)
  - Bayes Klassifikator
  - KNN (K-Nearest-Neighbor)
  - …

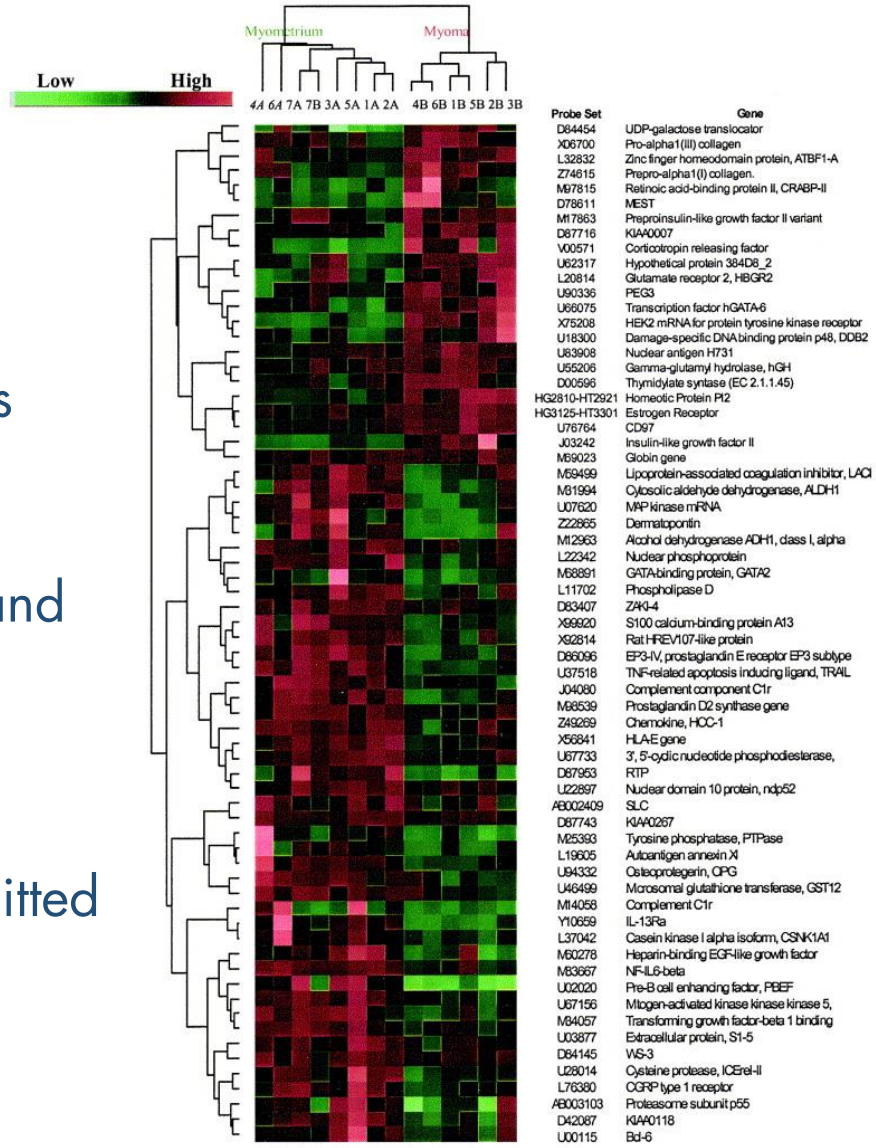


Ramaswamy, Golub: **DNA Microarrays in Clinical Oncology, 2002**

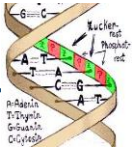


Beispiel: Hierarchisches Clustern

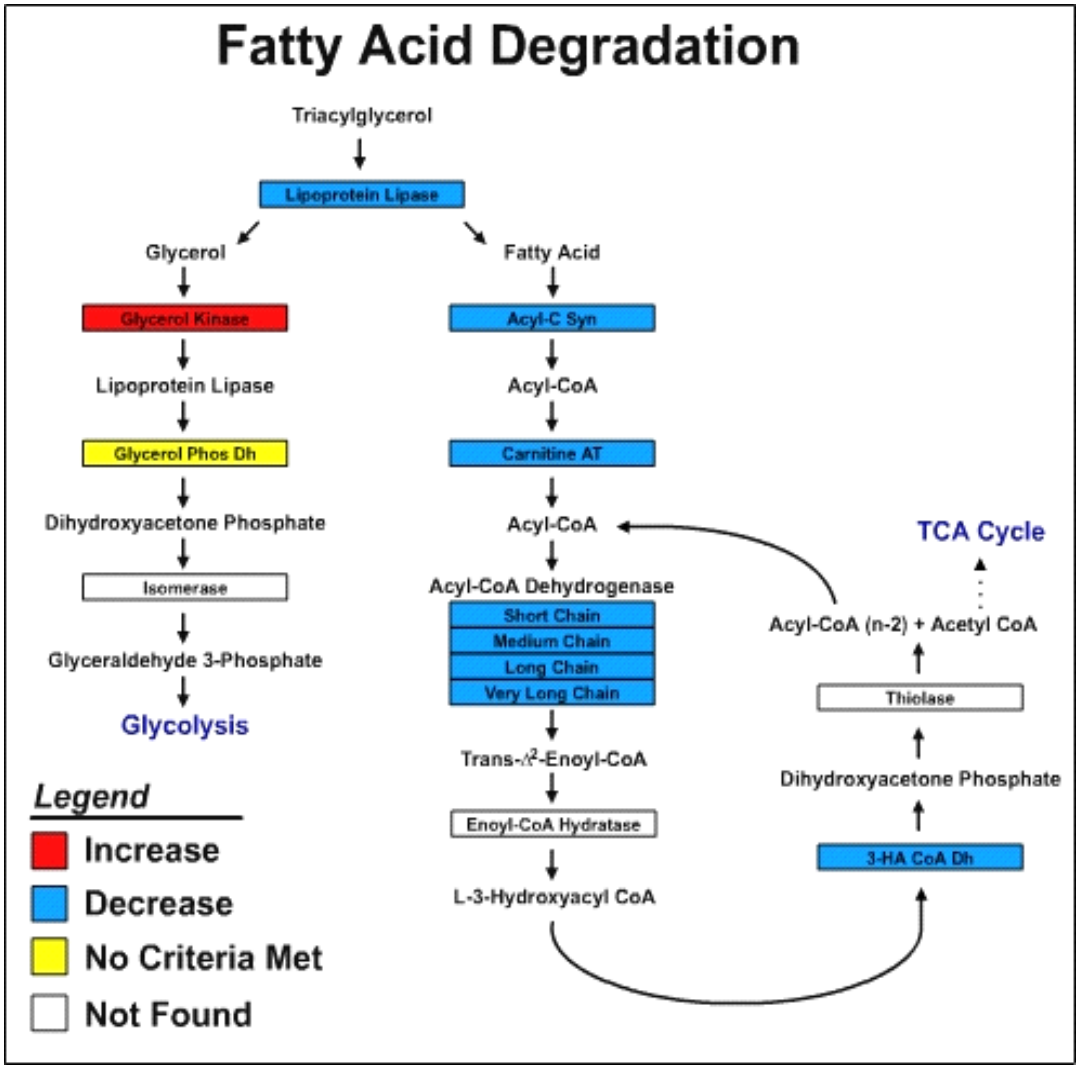
- Nutze z.B. euklidische Distanz
- Berechne alle paarweisen Distanzen (Ähnlichkeitsmatrix)
- Bottom-up, agglomerativ
  - jedes Gen ist zunächst ein eigenes Cluster
  - schrittweises Zusammenfassen der Gene/Cluster mit kürzestem Abstand
- Top-Down, divisiv
  - Zunächst sind alle Gene in einem Cluster
  - In jedem Schritt wird Cluster gesplitted
- Ergebnis z.B. **Dendrogram**



Bildquelle: Hongbo Wang, MD et al.: Distinctive proliferative phase differences in gene expression in human myometrium and leiomyomata, 2003, http://ars.els-cdn.com/content/image/1-s2.0-S0015028203007301-gr2.jpg

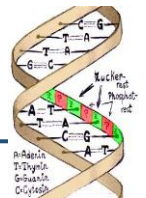


Visualisierung der GE in einem Pathway



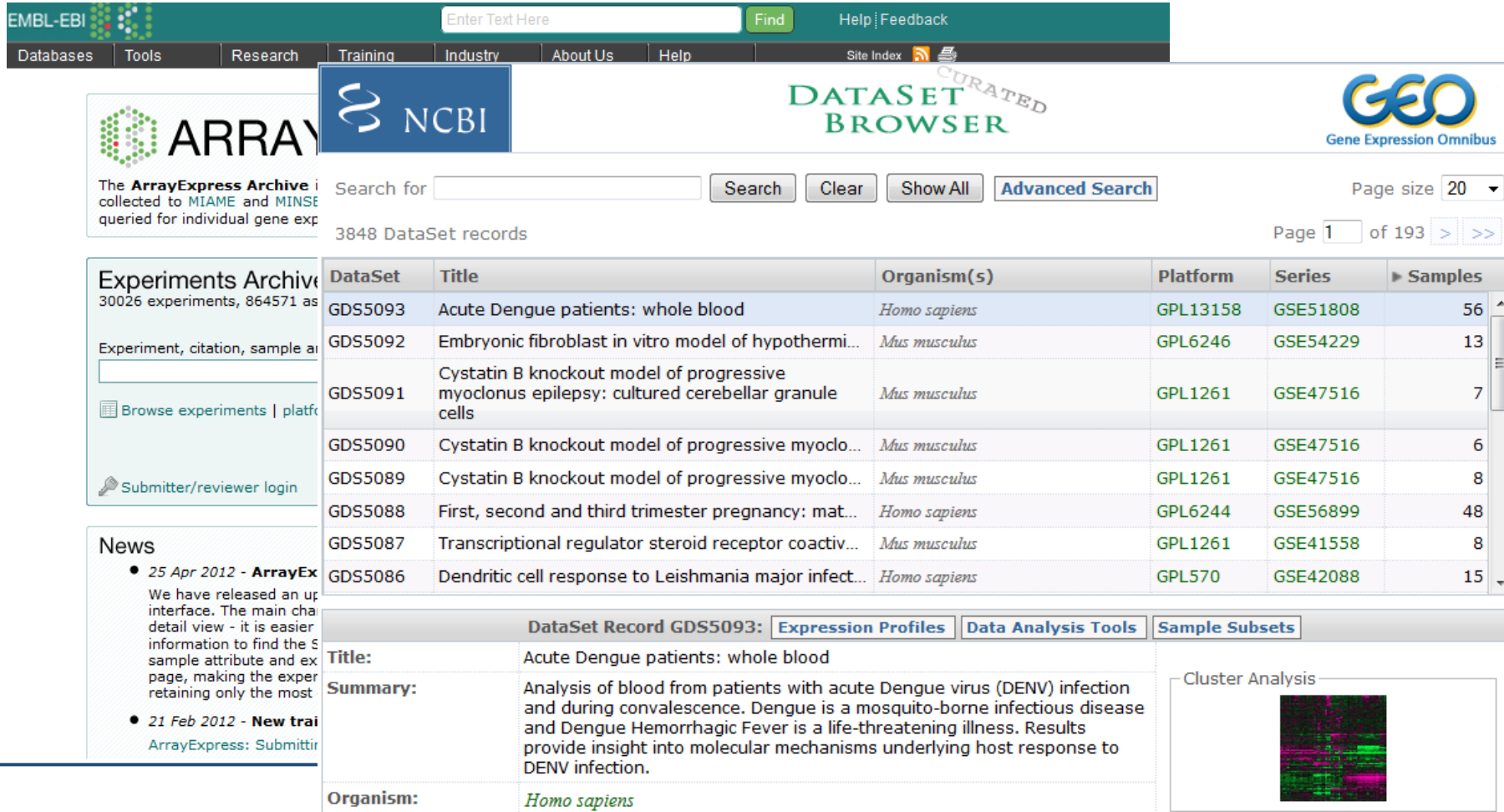
Genexpressionsdatenbanken

- Speicherung verschiedener Daten zur Genexpressionsanalyse
  - Management und Transformation von Expressionsdaten
  - Integration und Management beschreibender Daten (Annotationsdaten)
- Integration/Kopplung von Methoden und Verfahren zur Genexpressionsanalyse
- Datenformat und -modellierung
  - Rohdaten
    - Binärdateien (Bilder) + numerische Daten
    - Herstellerspezifische Formate: CEL-Dateien bei Affymetrix
    - Speicherung meist im Dateisystem
    - Seltene Re-prozessierung → Archivierung
  - Expressionsdaten
    - Speicherung in RDBMS
    - Eignung der multidimensionalen Modellierung (Data Warehouse)



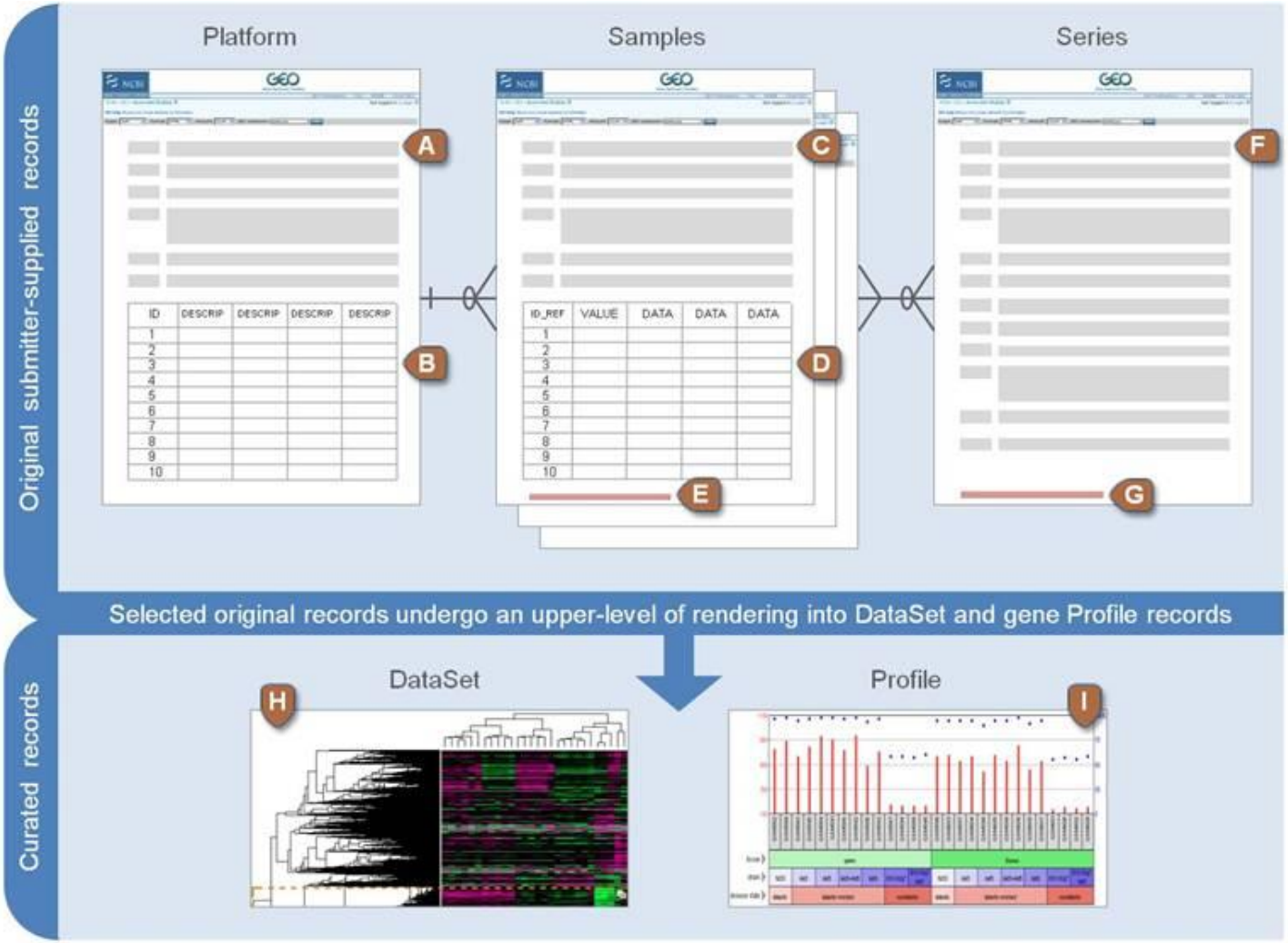
Beispiele

- Diverse Plattformen zur Genexpressionsanalyse: ArrayExpress (EBI), Gene Expression Omnibus (GEO NCBI), Stanford Microarray Database, …



GEO = Gene Expression Omnibus (NCBI)

- Data Organization:

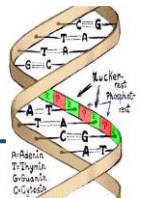


A	Text description of the array or sequencer
B	Text tab-delimited table of the array template
C	Text description of the biological sample and protocols to which it was subjected
D	Text tab-delimited table of processed hybridization result(may optionally include raw data columns)
E	Original raw data file, or processed sequence data file
F	Text description of the overall experiment
G	Tar archive of original raw data files, or processed sequence data files

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

GEO Data Organization

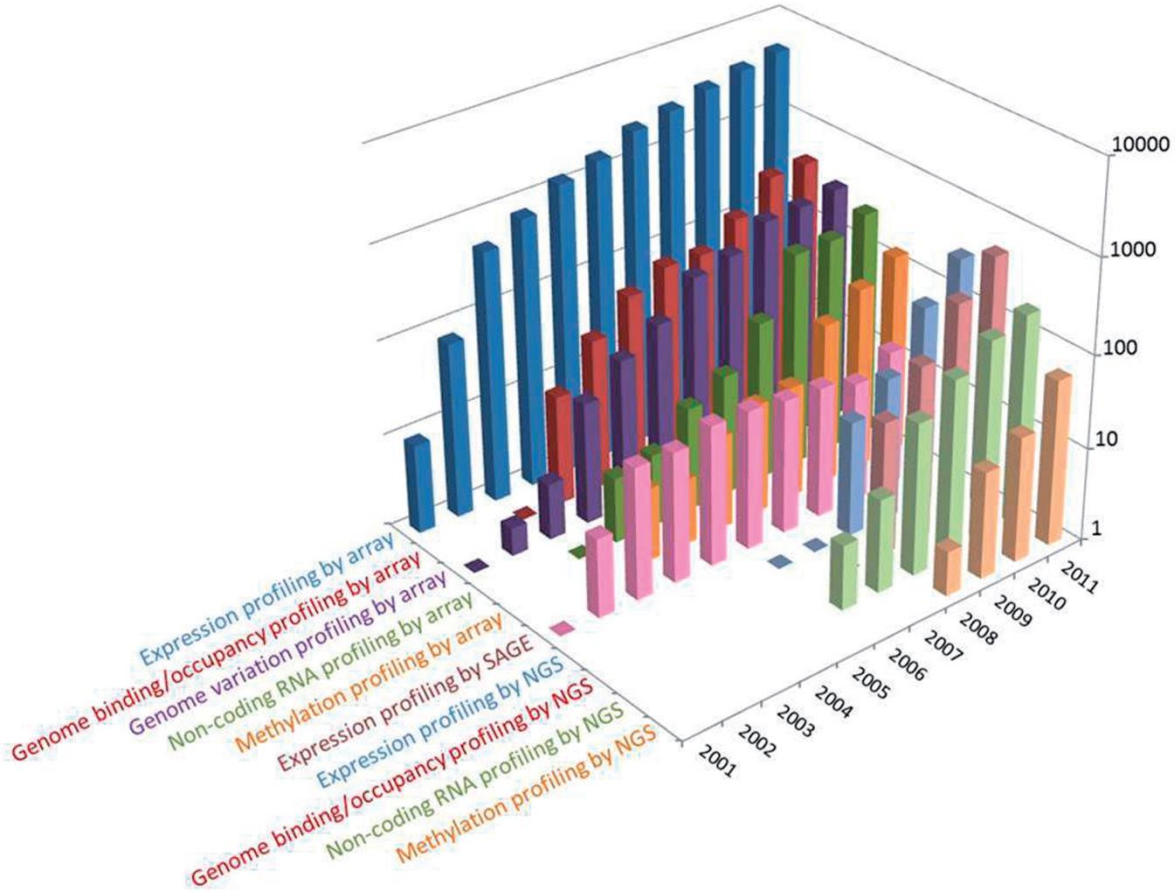
- ***‚Platform record'*:** zusammenfassende Beschreibung von Array (o. Sequenzierer) + Tabelle zur Definition des Array Templates
  - Referenziert mehrere Samples verschiedener Einreichungen
- ***‚Sample record'*:** Beschreibung der Bedingungen ‚unter welchen ein Sample untersucht wurde (Manipulationen etc.), Erfassung der Messwerte
  - Referenziert genau eine ‚*Platform'* und ist in mehrere ‚*Series'* involviert
- ***‚Series record'*:** Zusammenfassung der ganzen Studie / des Experiments, Verknüpfung der zusammengehörigen Samples; enthält auch extrahierte Daten, zusammenfassende Erkenntnisse, etc.
- ***‚Dataset records'*:** Durch GEO Kuratoren neu zusammengestellte *Series records*; kuratierte Sammlung biologisch und statistisch vergleichbarer GEO Samples
  - Samples in einem ‚*DataSet*' gehören zur gleichen ‚*Platform*', sie haben also Array-Elemente gemeinsam (Werteberechnung in Dataset muss konsistent erfolgt sein, z.B. Verwendung des gleichen Normalisierungsverfahrens)
- ***‚Profile'*:** Messung der Expression für individuelle Gene über alle Samples in einem ‚*Dataset'*



GEO Daten

- Verteilung der Anzahl und Typen ausgewählter Studien pro Jahr
- Query-Beispiel:

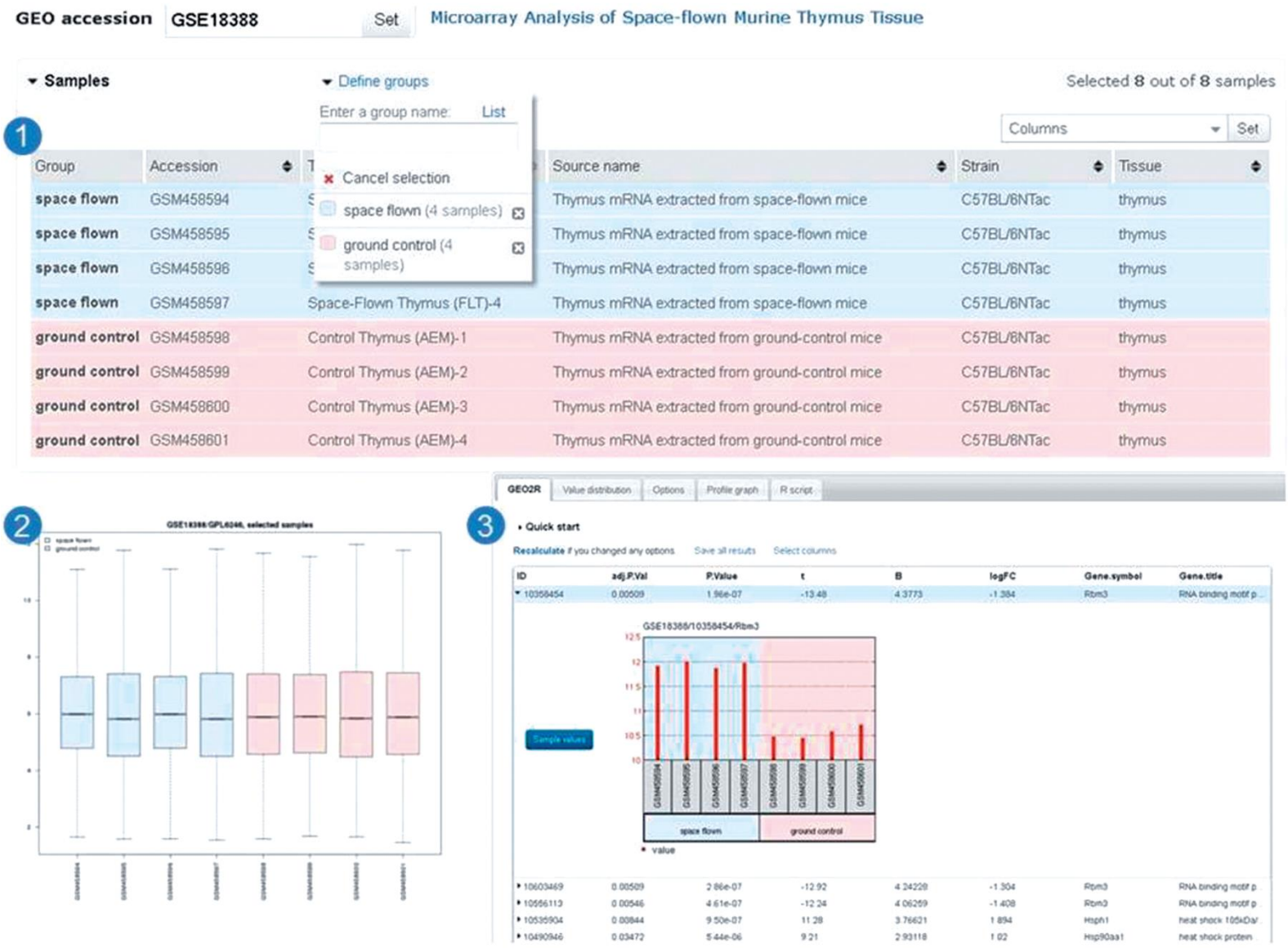
```
Expression profiling by array[DataSet Type] AND 2014[Publication Date]
```



- 3848 data sets
- 53187 series
- 13714 platforms
- 1.297.649 samples

Barrett et al.: NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5.

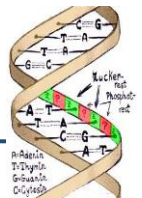
GEO Ausgabe / Analyse



Barrett et al.: NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5.

Zusammenfassung: Genexpressionsanalyse

- Microarray
  - populäre Technik zur Genexpressionsanalyse
  - Generierung eines enormen Datenvolumens
- Visualisierung
- Normalisierung von Rohdaten
- Data Mining Techniken zur Analyse
- Datenbanken zur Unterstützung der Genexpressionsanalyse
  - Fehlende/unzureichende Integration von Experiment & Sample-, Genannotation
  - Analyseintegration (oft beschränkt auf eine Analysesoftware)



Fragen ?