V8 Genexpression - Microarrays

- **Idee 1:** bestimme die Expressionslevel ausgewählter Gene/miRNAs als Biomarker für Krankheitsentstehung.
- **Idee 2:** analysiere die Ko-Expression von mehreren Genen um auf funktionelle Ähnlichkeiten der Gene 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

8. Vorlesung WS 2020/21 Softwarewerkzeuge 1

Im dritten Teil der Vorlesung werden wir uns mit der Analyse von Daten aus biologischen Hochdurchsatzexperimenten beschäftigen. Eine der wichtigsten und am besten etablierten Methoden ist die Transkriptomanalyse. Stephen Fodor und Kollegen stellten 1991 die **Microarray-**Methode vor (https://science.sciencemag.org/content/251/4995/767.long). 1993 gründete Fodor die Firma Affymetrix, die heute zu Thermo Fisher gehört. Microarrays werden heute noch oft verwendet. Interessant ist vor allem die Möglichkeit, spezielle Chips z.B. für bestimmte diagnostische Zwecke herzustellen. Heutzutage wird die Microarray-Technologie allerdings zunehmend von **RNAseq**-Methoden verdrängt.

das Transkriptom

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

Dies beinhaltet Protein-kodierende Gene und RNA-kodierende Gene, die nicht in Protein translatiert werden.

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

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

Heute werden wir uns auf die Detektion des Transkriptoms beschränken und die Prozessierungsschritte von der DNA zur reifen mRNA ignorieren.

Wikipedia schreibt hierzu:

"Das **Transkriptom** ist die Summe aller zu einem bestimmten Zeitpunkt in einer Zelle transkribierten, das heißt von der DNA in RNA umgeschriebenen Gene, also die Gesamtheit aller in einer Zelle hergestellten RNA-Moleküle. Der Begriff ist vergleichbar mit dem **Proteom**, der Gesamtheit der zu einem bestimmten Zeitpunkt in einer Zelle vorliegenden Proteine. Da aber nicht jede nach der Transkription vorliegende RNA, wie z. B. rRNA oder die RNA von Ribonukleoproteinen, in ein Protein übersetzt (translatiert) wird und mRNAs noch prozessiert werden können, sind Proteom und Transkriptom einer Zelle nicht identisch."

Oft ignoriert man in der Bioinformatik diese Unterschiede von Transkriptom und Proteom.

veränderte Genregulation bei Krankheiten etc.

Ausgangspunkt: bestimmte Krankheiten (Krebs ?) führen zur veränderten Expression einer Anzahl von Genen, nicht der 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 in einer Probe (Anzahl von homogenen Zellen bzw. Gemisch) unter bestimmten Umgebungsbedingungen.

Zunächst einmal fragen wir, was der Zweck einer Transkriptom-Analyse sein soll. Schließlich sind solche Analysen teuer. Eine aktuelle Preisliste der Boston University (http://www.bumc.bu.edu/microarray/pricing/) listet folgende Preise: Menschlicher Affymetrix-Microarray ($300-$700), RNAseq ($2000-$7000).

Die Hauptanwendungen kommen natürlich aus der Medizin. Das Hauptinteresse liegt darin, schnelle und präzise Diagnosen zu erstellen, damit den Patienten schnell geholfen werden kann (und sie das Krankenhausbett möglichst schnell wieder verlassen) und damit sie von Beginn an zielgerichtet mit der optimalen Therapie behandelt werden. Wo liegt das Potential von Transkriptom-Analysen?

Wikipedia:

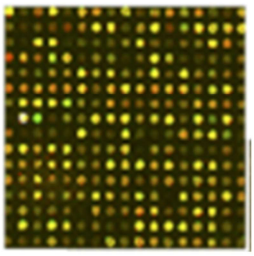
"Eine **monogenetische Erkrankung**, auch als monogene Erkrankung bezeichnet, ist eine Krankheit, die durch einen Defekt in einem einzelnen Gen (= mono-gen) hervorgerufen wird. Meistens handelt es sich dabei um ererbte Erkrankungen. Charakteristisch für monogenetische Krankheiten ist, dass sie in ihrem Vererbungsmuster den mendelschen Regeln folgen. Sie treten häufig schon in der frühen Kindheit auf und zeigen in vielen Fällen einen schwerwiegenden chronischen oder sogar tödlichen Verlauf. Da prinzipiell alle Gene anfällig dafür sind von Zeit zu Zeit zu mutieren, ist die Anzahl unterschiedlicher monogenetischer Erkrankungen sehr groß."
Die Online-Datenbank OMIM (Online Mendelian Inheritance in Man, **www.omim.org**) listet Tausende an monogenetischen Krankheiten.

Monogenetische Erkrankungen kann man mit einem **PCR-Test** des betreffenden Gens diagnostizieren.

Im Gegenzug dazu gibt es polygenetische Krankheiten wie Krebs. https://naturwissenschaften.ch/ schreibt dazu: **Polygenetische** Krankheiten werden durch eine Vielzahl an Mutationen im Genom ausgelöst. Somit ist bei diesen Krankheiten nicht nur ein einzelnes Gen verantwortlich, sondern das Zusammenspiel verschiedener Veränderungen.

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*).

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

8. Vorlesung WS 2020/21 Softwarewerkzeuge 4

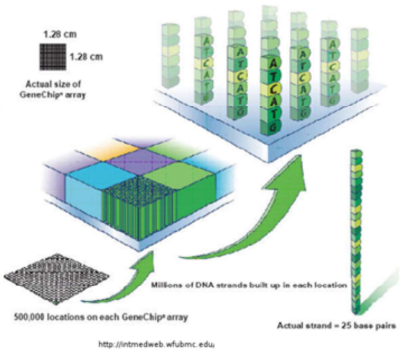
Die Abbildung zeigt einen kleinen Ausschnitt aus einem Mikroarray. Manche der Fingernagel-grossen Microarrays enthalten bis zu 30.000 „spots". Die Auswertung geschieht optisch und detektiert die Signale von fluoreszierenden Substanzen (rot / grün /gelb). Dazu kommen wir gleich. Man hybridisiert anstelle von mRNA meist die viel robustere cDNA, die mit dem Enzym reverse Transkriptase aus der zellulären mRNA umgeschrieben wird. Jedes Testfeld („spot") enthält Tausende identische Kopien eines Oligo-Nukleotids, an die die cDNA eines Gens hybridisieren soll.

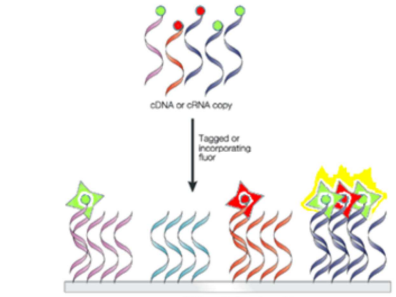
Experimentelles Vorgehen

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

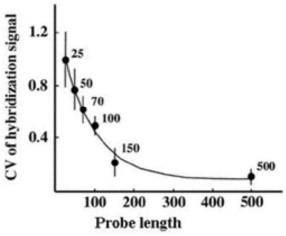
Jede Zelle enthält eine komplementäre Probe für ein Gen, die an die Oberfläche funktionalisiert wurde (typisch 20-70 nt lang).

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

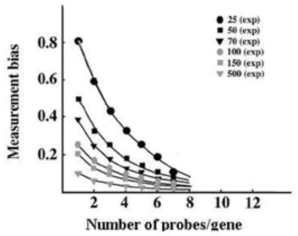




changed from:
A. Butte, Nature Reviews Drug Discovery 1, 951-960, 2002



(I) Probe length v.s. hybridization signal



(II) Probe set size v.s. measurement bias

pgrc.ipk-gatersleben.de
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3902802/

Wie lange sollen Mikroarray-Proben sein? https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3902802/ schreibt dazu: Currently, probes used in major commercial platforms can be either short (20-30 mers) or long (50-70 mers) oligonucleotides."

Eine untere Grenze ergibt sich durch die Anforderung der Spezifität. Man möchte ja schließlich ein ganz bestimmtes Gen messen. Ein guter Anhaltspunkt ist die Länge von siRNA-Molekülen (short interfering RNAs). Diese hybridisieren auf der gesamten Länge von 21nt mit einer mRNA und "silencen" diese dadurch. Das menschliche Genom hat 3 Milliarden Basen (3 x 10^9). Es gibt 4^21 Oligonukleotide mit 21 Basen. 2^10 = 1024 ≅ 10^3. Aneinandergehängt wäre deren Länge 4^21 = 2^42 = 4 x (2^10)^4 = 4 x 10^12, also 4 Trillionen Basen. In dem menschlichen Genom mit 3 Milliarden Basen kommt jedes 21nt-Oligo daher 1/1330 mal zufällig vor. Eine Länge von 21nt ist daher bereits sehr spezifisch für ein Gen.

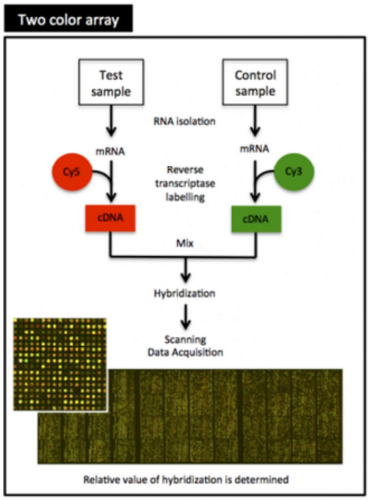
Wie in der Abbildung unten rechts gezeigt wird, erreicht man mit längeren Proben jedoch eine kleinere Streuung (coefficient of variation, CV) der Messungen (links) und einen kleineren Bias (rechts) als mit kurzen Proben. Um verschiedene Transkripte auflösen zu können und um robustere Ergebnisse zu erhalten, verwendet man meist mehrere Proben pro Gen. Von diesen Proben verwendet man dann normalerweise den **Median** als den Expressionslevel des Gens/Transkripts. Generell braucht man bei längeren Proben weniger Proben pro Gen für eine ähnliche genaue Messung (rechts).

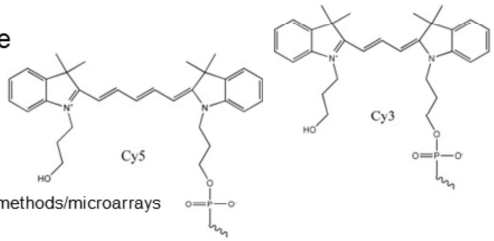
2-Farben Microarrays

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

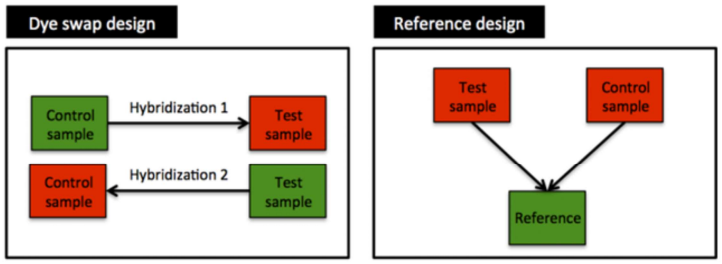
8. Vorlesung WS 2020/21 Softwarewerkzeuge 6

Beim Umschreiben in cDNA stellt man in der Lösung gelabelte Nukleotide bereit, an die entweder ein rot fluoreszierender Farbstoff (Cy5) oder ein grün fluoreszierender Farbstoff (Cy3) angeheftet ist. Diese werden dann in die synthesierte cDNA eingebaut.

Durch die Verwendung zweier Farbstoffe, die bei unterschiedlichen Wellenlängen fluoreszieren, kann man beim Vergleich zweier Proben feststellen, in welcher Probe mehr von einer bestimmten cDNA/mRNA vorhanden ist.

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

Kein Kommentar.

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 der Zeit t durch eine Ratengleichung ausdrücken:

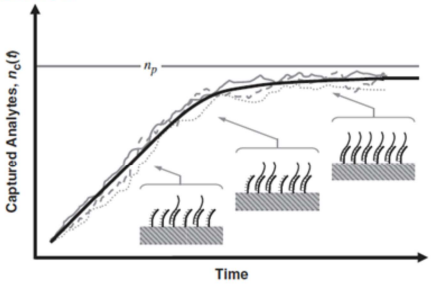
$$\frac{\mathrm{d}\langle n_c(t)\rangle}{\mathrm{d}t} = k_1^* \left(\frac{n_p - \langle n_c(t)\rangle}{n_p} \right) (n_t - \langle n_c(t)\rangle) - k_{-1}\langle n_c(t)\rangle.$$

Bindung **Dissoziation**

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)

Die Hybridisierung der cDNA-Stränge an die Proben des Mikroarrays kann man als dynamischen Prozess auffassen, der wie in der Abbildung gezeigt, nach einer ausreichenden Zeit in die Sättigung läuft. Das ganze ist eine Balance zwischen Bindung und Dissoziation. Die Dissoziation geschieht mit der Dissoziations-Ratenkonstante k-1 multipliziert mit der Anzahl an gebundenen DNA-Strängen zur Zeit t.

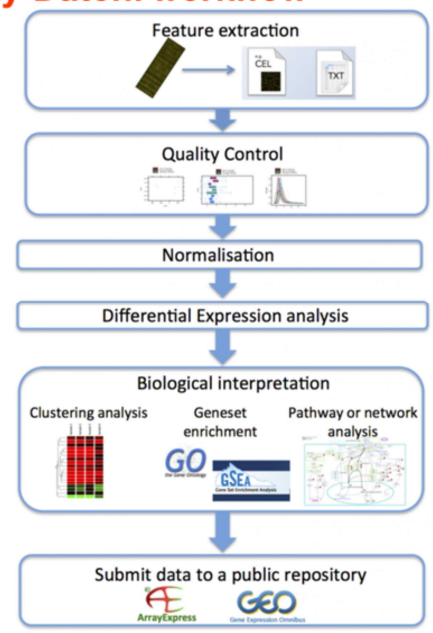
Die Bindung geschieht mit der Assoziations-Ratenkonstante k1 multipliziert mit der Anzahl an DNA-Strängen in der Lösung (erster Klammerausdruck – Gesamtzahl minus gebundene DNA-Stränge) und multipliziert mit der Anzahl an freien Bindungsplätzen. Man kann sich das in etwa wie die Suche von Autos nach einem freien Parkplatz in einer Grossstadt vorstellen.

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.



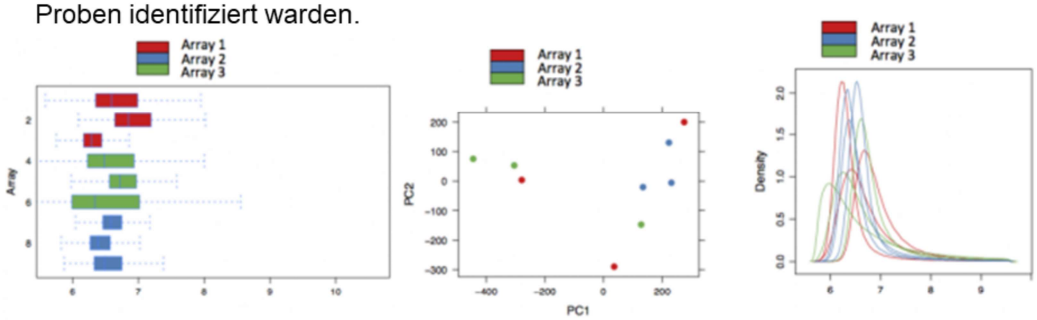
http://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays

Diese Folie bietet eine Überblick über die verschiedene Schritte der Microarray-Analyse. Die einzelnen Schritte werden im Folgenden erklärt werden.

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

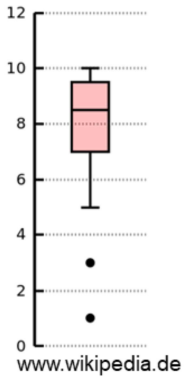
Box-Plots (links), PCA (Hauptkomponentenanalyse, Mitte) und Dichteverteilung (rechts) sind unterschiedliche Arten, die Verteilung der Datenpunkte in den einzelnen Proben darzustellen. In dem hier gezeigten Beispiel gibt es keine klar ersichtlichen Ausreißer (outlier).

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

8. Vorlesung WS 2020/21 Softwarewerkzeuge 11

Die Boxplot-Darstellung ist weitverbreitet. Wikipedia schreibt hierzu unter https://de.wikipedia.org/wiki/Box-Plot

„Ein Box-Plot besteht immer aus einem Rechteck, genannt **Box**, und zwei Linien, die dieses Rechteck verlängern. Diese Linien werden als „Antenne“ oder seltener als „Fühler“ oder „**Whisker**“ bezeichnet und werden durch einen Strich abgeschlossen. In der Regel repräsentiert der Strich in der Box den Median der Verteilung. Die **Box** entspricht dem Bereich, in dem die mittleren 50 % der Daten liegen. Sie wird also durch das obere und das untere **Quartil** begrenzt, und die Länge der Box entspricht dem Interquartilsabstand (englisch interquartile range, IQR). Dieser ist ein Maß der Streuung der Daten, welches durch die Differenz des oberen und unteren Quartils bestimmt wird. Des Weiteren wird der Median als durchgehender Strich in der Box eingezeichnet. Dieser Strich teilt das gesamte Diagramm in zwei Bereiche, in denen jeweils 50 % der Daten liegen. Durch seine Lage innerhalb der Box bekommt man also einen Eindruck von der Schiefe der den Daten zugrunde liegenden Verteilung vermittelt. Ist der Median im linken Teil der Box, so ist die Verteilung rechtsschief, und umgekehrt.“

Allerdings gilt

„Durch die Antennen werden die außerhalb der Box liegenden Werte dargestellt. Im Gegensatz zur Definition der Box ist die Definition der Antennen nicht einheitlich.“

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.

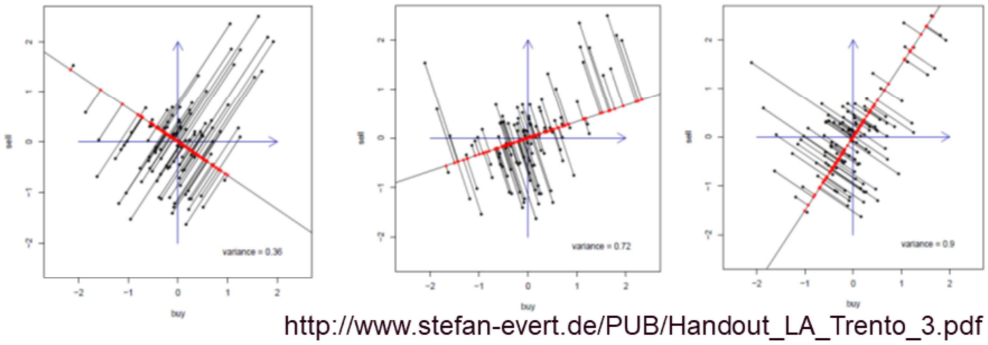
Wir betrachten eine Datenmatrix **X** für *I* Beobachtungen und *J* Variablen.

Ihre Elemente sind x_{ij}.

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

Die zwei Arten von Variablen in einem wissenschaftlichen Experiment sind die unabhängige und die abhängige Variable.
Eine unabhängige Variable wird in dem Experiment geändert oder kontrolliert um die Auswirkung auf die abhängige Variable zu testen. Die abhängige Variable ist diejenige, die im Experiment gemessen wird.
In unserem Fall ist die abhängige Variable das Ergebnis des DNA-Microarray-Experiments.
Die unabhängige Variable könnte z.B. das Alter der Patienten sein, ob sie mit Diabetes infiziert sind oder nicht, in welchem Labor die Analysen gemacht wurden etc.
Die Frage wäre dann, ob die Genexpressionswerte eine Funktion solcher unabhängiger Variablen sind.

Ziel der PCA



http://www.stefan-evert.de/PUB/Handout_LA_Trento_3.pdf

Das Ziel der PCA-Analyse ist, einen Satz von zueinander orthogonalen Vektoren (principal components = Hauptkomponenten) zu konstruieren, die in Richtung der größten Varianz in der Datenwolke zeigen.

Hier ist dies für PC1 gezeigt. Die 3 Abbildungen enthalten die identischen Datenpunkte. Es sind drei unterschiedliche PC-Vektoren dadurch gelegt. Die linke Abbildung generiert die größten quadratischen Abweichungen. Der Vektor beschreibt die kleinste Varianz in den Daten. Der optimale Vektor ist rechts gezeigt.

Die **geometrische Konstruktion** von PC-Vektoren ist zwar im Prinzip möglich, wird aber in der Praxis nicht verwendet. PC2 würde dann senkrecht auf PC1 stehen und zeigt in Richtung der größten dann verbleibenden Varianz etc.

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**.

Wenn man die Daten nicht zentriert, erhält man unterschiedliche Ergebnisse, die schwieriger zu interpretieren sind.

PCA- Präprozessierung der Werte

(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* statt durch *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?

15 8. Vorlesung WS 2020/21 Softwarewerkzeuge 15

Wenn Variablen mit unterschiedlichen Einheiten gemeinsam analysiert werden, ist es wichtig, die Daten vorher zu standardisieren oder zu normalisieren.
Im Gegensatz zur geometrischen Konstruktion (siehe Folie 13) verwendet man meist die sogenannte SVD-Zerlegung der Datenmatrix X. Wir werden die mathematischen Details überspringen, welche Matrizen eine solche SVD-Zerlegung besitzen.

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**.

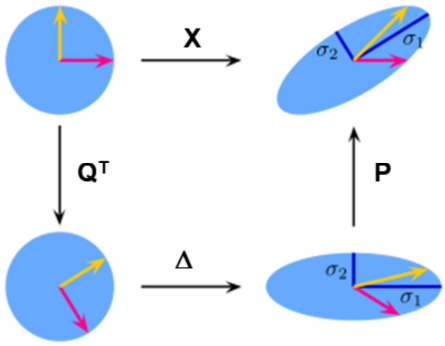
Δ : ist die Diagonalmatrix der *singulären Werte*. Diese sind die Quadratwurzeln der Eigenwerte der Matrix $\mathbf{X}\,\boldsymbol{X}^T$ (entsprechen denen von $\boldsymbol{X}^T\mathbf{X}$).

Als Ergebnis der SVD erhält man die Zerlegung der Datenmatrix X in ein Produkt dreier Matrizen P, Delta und Q.
Delta ist hierbei eine Diagonalmatrix, sie enthält nur auf der Diagonalen Einträge ungleich Null. Diese Diagonalwerte sind die Quadratwurzeln des Matrixprodukts von X mit seiner transponierten Form X_transponiert (dabei wird die Matrix an der Diagonale von links oben nach rechts unten gespiegelt).
Man betrachtet dieses Matrixprodukt, da die Datenmatrix X normalerweise eine Rechteckmatrix ist, wohingegen nur quadratische Matrizen eine Eigenvektorzerlegung besitzen. Das Matrixprodukt von X mit seiner transponierten Form X_transponiert ist jedoch quadratisch. Je nachdem, ob man X^T x X bildet, oder X x X^T, ist das Quadrat unterschiedlich gross.
Q und P enthalten die normierten Eigenvektoren der zwei möglichen Matrixprodukte.

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

Diese Abbildung (aus Wikipedia) illustriert, dass man sich das Matrixprodukt P x Delta x Q^T als Hintereinanderausführung von 3 Operationen vorstellen kann.

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.

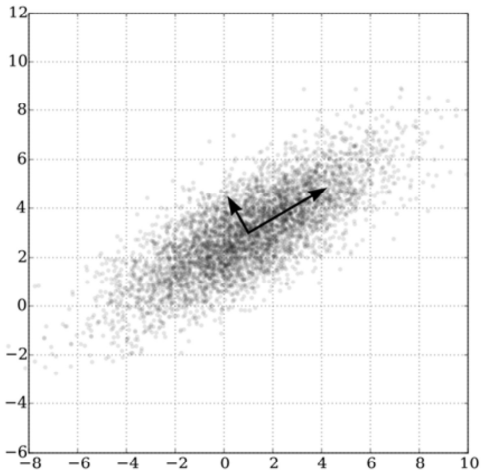
PC1 ist der Eigenvektor von $\mathbf{X}^T\mathbf{X}$ mit dem größten Eigenwert (siehe $\mathbf{Q}$) usw.

PCA ist eine der weit verbreitetsten Methoden der Datenanalyse und dient allgemein zur Dimensionsreduktion. Solche Methoden werden verwendet, um durch geeignete Projektion der Daten in einen niedrigdimensionalen Raum einen guten Überblick über die Verteilung der Daten zu bekommen.

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

Diese Wolke von Datenpunkten wurde vor der PCA-Analyse nicht im Nullpunkt zentriert, sondern hat ihren Mittelwert bei x=1, y=3. Die x-Achse und y-Achse sind die Originalkoordinaten. Die beiden schwarz eingezeichneten Vektoren sind PC1 und der dazu orthogonale PC2-Vektor. Offensichtlich zeigt PC1 in Richtung der größten Varianz (bzw. Standardabweichung).

Konstruktion der PC-Vektoren

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

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

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

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

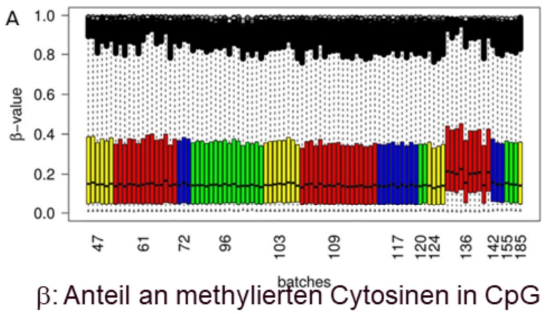
$$\mathbf{F} = \mathbf{P}\Delta = \mathbf{P}\Delta\mathbf{Q}^T\mathbf{Q} = \mathbf{X}\mathbf{Q}$$

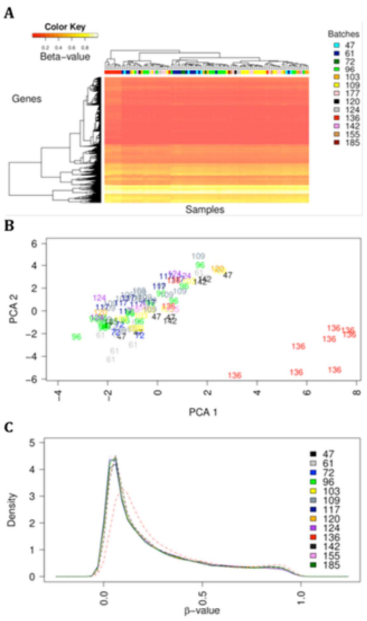
F kann daher als eine **Projektionsmatrix** interpretiert werden.
Die Multiplikation von **X** mit **Q** entspricht der Projektion der Beobachtungen **X** auf die principal components **Q**.

Kein Kommentar.

Ausreißer-Datenpunkte?

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





β: Anteil an methylierten Cytosinen in CpG

Links: box-plot

Rechts/oben: hierarchisches Clustering

Rechts/Mitte: PCA

Rechts/unten: Dichteverteilung

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4999208/

Das ist ein Beispiel für die „explorative“ Analyse der Rohdaten. Man schaut sich zunächst einmal an, was für Daten vorhanden sind.

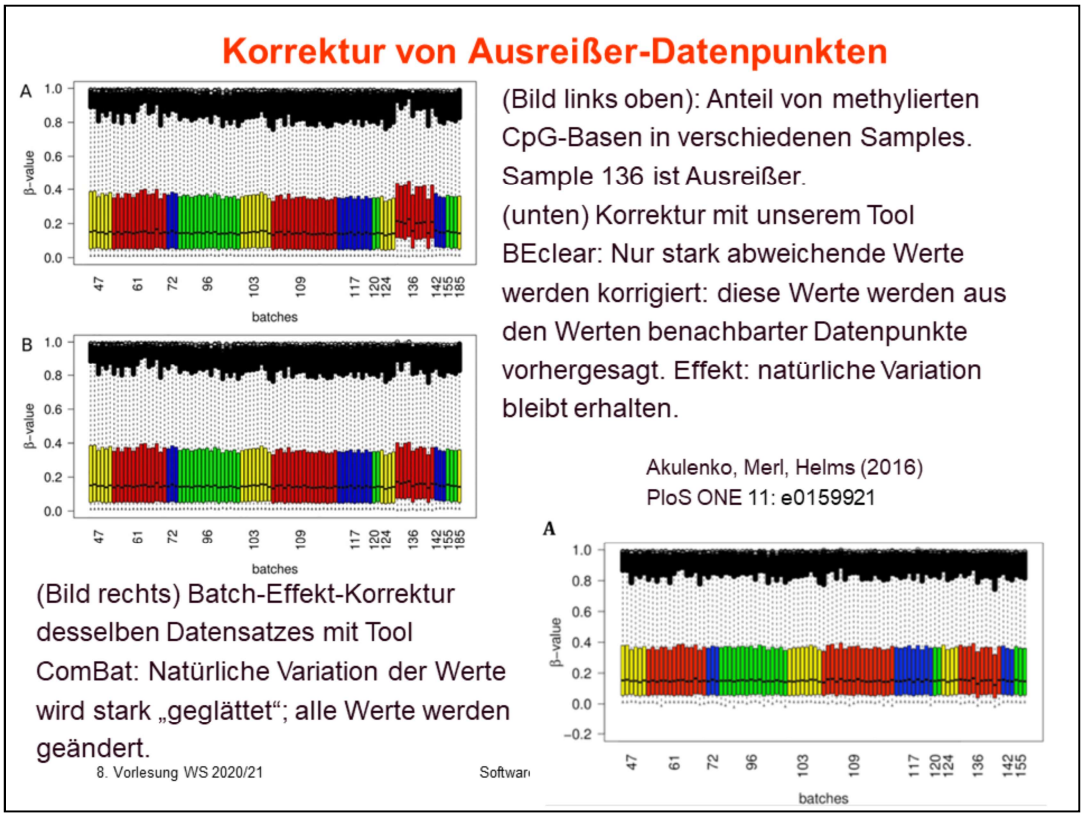
Hier sind DNA-Methylierungsdaten aus dem TCGA (The Cancer Genome Atlas)-Datensatz für Brustkrebs gezeigt.

Oben rechts ist ein Boxplot der Daten aus verschiedenen „Batches“ gezeigt. Beta-Werte variieren zwischen 0 (nicht methyliert) und 1 (alle CpGs vollständig methyliert).

Rechts oben ein hierarchisches Clustering, in der Mitte ein PCA-Plot und unten eine Dichteverteilung der β-Werte.

In allen Plots sind man deutlich, dass die β-Werte in Batch 136 hin zu höheren Werten verschoben sind.

Der Boxplot zeigt, dass dies nicht nur an einer Probe liegt, sondern an allen (bis auf 2) Proben aus diesem Batch.



Bevor man Rohdaten verarbeitet, sollte man Ausreißerdatenpunkte entweder löschen bzw. korrigieren.

In der linken Abbildung wird gezeigt, wie man dies mit unserem Tool BEclear (http://bioconductor.org/packages/release/bioc/html/BEclear.html) tun kann. Dieses Tool korrigiert nur die von einem „batch effect“ betroffenen Ausreißer-Datenpunkte. Alle anderen Werte bleiben unverändert. Ein anderes, weit verbreitetes Tool namens ComBat korrigiert dagegen alle Datenpunkte (unten rechts).

Sie mögen sich wundern, ob so etwas nicht eine „Manipulation“ der Daten darstellt. Dies ist zweifelsohne der Fall. Die nächste Frage ist, ob dies zulässig ist. Ja! Denn sonst würden die Ergebnisse der anschließenden Analysen durch die Ausreißerdatenpunkte verfälscht. Wichtig ist allerdings immer, dass man in einer Publikation, einer Abschlussarbeit, oder einem Praktikumsbericht solche Korrekturen deutlich beschreibt und kennzeichnet.

Wenn man „genügend“ Proben hat, wie dies oft bei Expressionsanalysen der Fall ist, ist es vermutlich einfacher, von „batch“ Effekten betroffene Proben einfach wegzulassen.

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

Normalisierung ist essentiell wichtig bei der Analyse von Microarray-Daten.

Die Hersteller der Microarray-Chips empfehlen meistens eine bestimmte Normalisierungsstrategie, die für die Analyse der mit diesem Gerät generierten Daten am besten geeignet ist.

Es ist meistens am einfachsten, diesen Empfehlungen zu folgen. Dies vermeidet Diskussionen mit den Gutachtern Ihrer Manuskripte / Abschlussarbeiten.

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

Ordne jede Spalte nach Größe

→

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

Bilde Mittelwert jeder Reihe

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.

Quantile Normalisierung ist eine weit verbreitete Normalisierungsmethode. Dabei werden die Datenpunkte in allen Proben der Größe nach geordnet und dann die Mittelwerte der größten, zweitgrößten, etc Werte berechnet. Anschließend werden alle Datenpunkte durch diese Mittelwerte ersetzt. Man erreicht dadurch, dass alle Proben hinterher (bis auf doppelte Datenpunkte) die identischen Werte enthalten und damit natürlich auch dieselbe **statistische Werteverteilung** besitzen. Dies ist sehr vorteilhaft für die statistische Bewertung von Abweichungen.

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 oder normalisierte Intensitätswerte verwenden (bei denen z.B. der Median des Hintergrund abgezogen wurde).

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

In einem Zweifarben-Microarray vergleicht man stets die Transkriptionslevel eines Gens in den beiden Proben miteinander, bzw. das von den entsprechenden gelabelten cDNA-Proben emittierte Fluoreszenzsignal. Man interessiert sich weniger für die absoluten Werte, da diese stark von den Bedingungen abhängen, sondern für die relativen Unterschiede (**fold-change**).

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

Man möchte Hoch- und Runterregulation von Genen gleich wichtig betrachten. Dies wird dadurch erschwert, dass die relative Hochregulation beliebig große Werte annehmen kann, die relative Runterregulation jedoch nur Werte zwischen 0 und 1.

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).

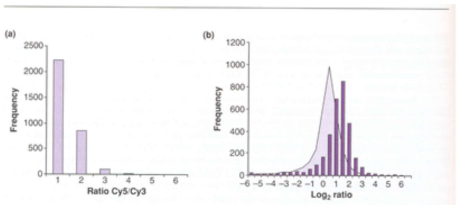


Figure 14.3
(a) Frequency histogram produced by plotting the distribution of the intensity ratios of the Cy5 and Cy3 channels per spot. (b) The same Cy5 and Cy3 spot ratio data (shaded distribution) compared to a second set of array data (bar distribution). Both sets of data are converted to $\log_2$.

M. Madan Babu, An Introduction to Microarray Data Analysis

Orengo-Buch

Durch Betrachtung des Logarithmus des fold changes haben Hoch- und Runterregulation dann denselben Wertebereich. Außerdem ähneln **$\log_2$(fold change)-**Verteilungen oft einer Normalverteilung (Bild unten rechts), so dass ein einfacher t-Test angewendet werden kann.

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.

Oft interessiert man sich dafür, ob 2 oder mehrere Gene ähnliche (korrelierte) Änderungen ihrer Transkriptionslevel zeigen, also z.B. in bestimmten Bedingungen gemeinsam hoch und in anderen Bedingungen gemeinsam runterreguliert. So etwas bezeichnet man als **Koexpression**. Dies deutet auf einen funktionellen Zusammenhang dieser Gene hin.

4.a Hierarchisches Clustering zur Analyse von Ko-Expression

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

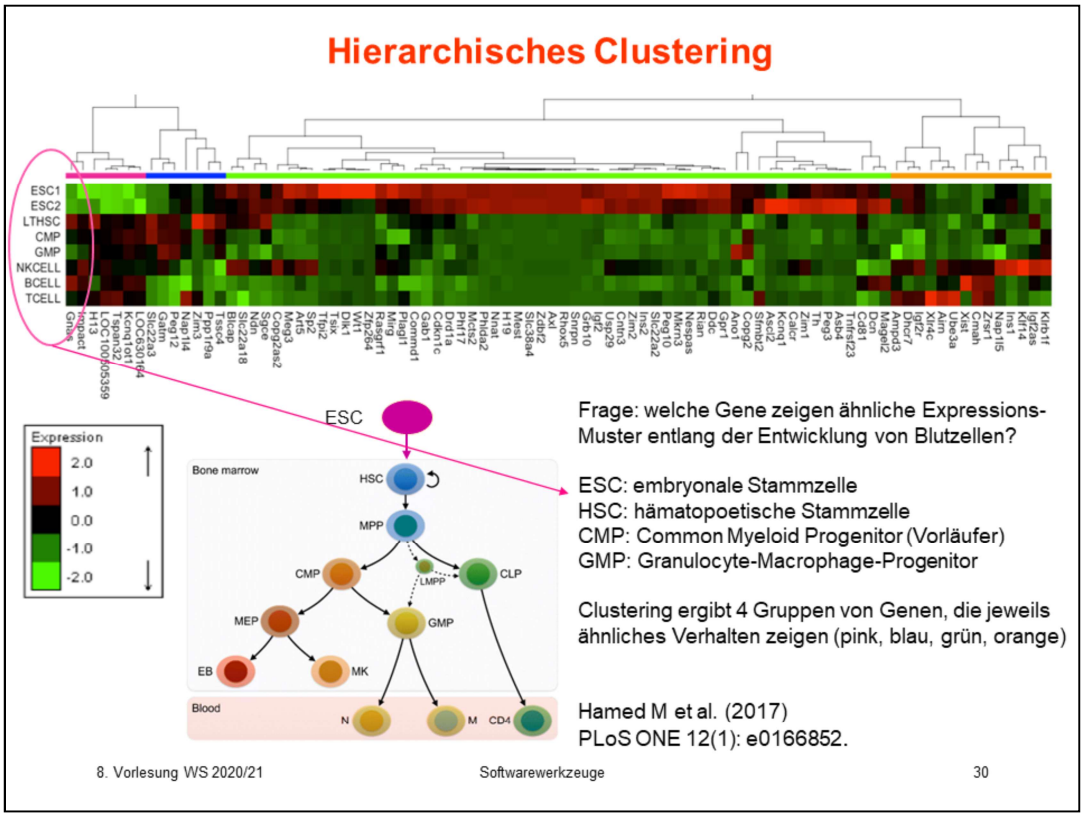
Bei den anhäufenden Verfahren, die in der Praxis häufiger eingesetzt werden, werden schrittweise einzelne Objekte zu Clustern und diese zu größeren Gruppen zusammengefasst, während bei den teilenden Verfahren größere Gruppen schrittweise immer feiner unterteilt werden.

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.

Um einen einfachen Überblick zu erhalten, verwendet man oft **Clustering** der Daten.



Gezeigt ist eine „**Heatmap**“ der logarithmierten Expressionslevel verschiedener Gene (auf x-Achse aufgetragen) entlang verschiedener Stadien der Hämatopoese (Reifung von Blutzellen) (auf y-Achse aufgetragen). In der Abbildung sind nur die Gene gezeigt, die eine nennenswerte Änderung aufweisen. Durch Clustern fasst man diejenigen Gene zusammen, die ähnliche Intensitätsverläufe zeigen.

k-means Clustern

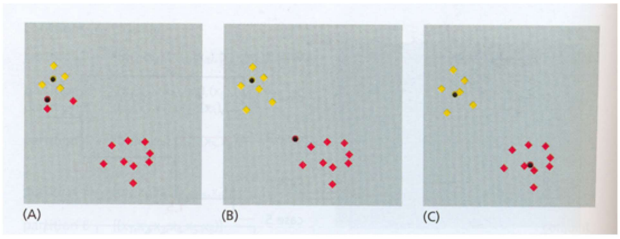
Ein Durchlauf der *k*-means Clustering Methode erzeugt eine Auftrennung der Datenpunkte in *k* Cluster. Gewöhnlich wird der Wert von *k* vorgegeben.

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



Es gibt eine riesengroße Anzahl an Clustering-Methoden, die sich in der Komplexität und der Eignung für verschiedene Datentypen unterscheiden. Der **K-means** Cluster-Algorithmus ist besonders einfach zu erklären und verstehen. Allerdings muss der Anwender vor dem Clustern die gewünschte Anzahl an Clustern vorgeben. Da dies meist zunächst nicht bekannt ist, empfiehlt es sich, den Algorithmus für unterschiedliche Vorgaben von k anzuwenden.

4.b Abschätzung der Signifikanz

Cancer Research

Lipid Metabolism Signatures in NASH-Associated HCC—Letter

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

Cancer Res Published OnlineFirst April 28, 2014.

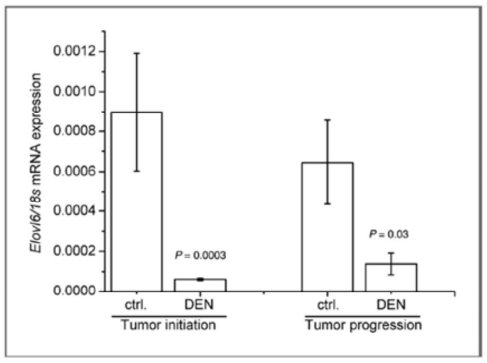


Figure 2. Wild-type mice were treated with the carcinogen DEN at the age of 2 weeks. Livers were analyzed after 24 weeks to assess the tumor initiation state. Analyses in the tumor progression stadium were done after 36 weeks. *Elovl6* mRNA expression as determined by real-time reverse transcriptase 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.

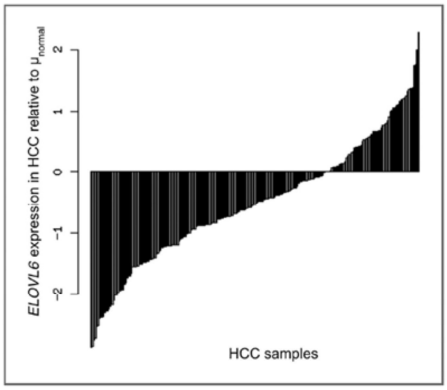


Figure 1. mRNA levels of *ELOVL6* in 247 human HCC samples relative to the mean of 239 nontumor liver tissue (μ_{normal}). Samples of dataset GSE14520 [$\log_2$ (expression) values from GEO after Robust Multi-array Average normalization] were mapped to hgu133a.db using bioconductor. Significance values: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

8. Vorlesung WS 2020/21 Softwarewerkzeuge 32

Ganz wichtig ist es, den Grad an **differentieller Expression** statistisch zu bewerten. In der linken Abbildung sind die mRNA-Expressionslevel des Gens Elovl6 in Kontrollbedingungen und in Mäusen gezeigt, bei denen durch die Chemikalie Di-Ethyl-Nitrosamin eine Tumorbildung initiiert wurde. Sowohl nach 24 Wochen (links) als auch nach 36 Wochen (rechts) war der mRNA-Expressionslevel von Elovl6 signifikant reduziert. Dies wurde mit dem Mann-Whitney U-Test bewertet.

Ist dies im Mensch genauso? Dort kann man aus ethischen Gründen keine analogen Experimenten durchführen. Man kann jedoch analysieren, ob in HCC-Patienten (Hepatozelluläres Karzinom) dieselbe Veränderung von ELOVL6 gegenüber Kontrollgewebe vorliegt. Dies ist in der Tat in den meisten Patienten der Fall (rechte Abbildung). Dann viel mehr menschliche Daten vorlagen (247 bzw. 239) als für die Mäuse (8-18) ist die Signifikanz rechts deutlich größer.

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 nicht den **Konfidenzlevel** bewerten, ob diese Gene wirklich differentiell exprimiert sind oder nicht.

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

Der Fold change ist kein statistischer Test.

Standardfehler

Die Standardabweichung σ

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(a_i - \bar{a}\right)^2}$$

gibt die „Standard“ abweichung aller Messwerte an.

Meist interessieren wir uns aber mehr für die Std.abw. des Mittelwerts.

Diese wird als Standardfehler des Mittelwerts **(SEM)** bezeichnet:

$$SEM = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(a_i - \bar{a}\right)^2}}{\sqrt{n}}$$

Immer wenn man eine Population durch eine zufällige Stichprobe abschätzt, enthält der Schätzwert wahrscheinlich einen Fehler.

SEM gibt eine Abschätzung für diesen Fehler.

Bei der differentiellen Expressionsanalyse müssen wir SEM für die Differenz der Mittelwerte zweier Proben berechnen → 2-sample t-test.

Die **Standardabweichung** misst die typische Abweichung eines einzelnen Datenpunkts vom Mittelwert. Was ist mit der Standardabweichung des Mittelwerts selbst? Diese misst man mit dem **Standardfehler des Mittelwerts** (SEM). Man erhält ihn, wenn man die Standardabweichung durch die Wurzel aus der Anzahl an Datenpunkten dividiert.

T-Tests

t-Wert: um wieviele Standardfehler unterscheidet sich eine Differenz von 0?

Es gibt 3 verschiedene Arten von t-Tests:

Ungepaarter t-Test

$$t = \frac{Mittelwert\ von\ Stichprobe\ 1\ - Mittelwert\ von\ Stichprobe\ 2}{SEM\ für\ die\ Differenz\ der\ Mittelwerte}$$

Gepaarter t-Test

$$t = \frac{Mittelwert\ der\ paarweisen\ Differenzen\ - Referenzwert}{SEM\ der\ Differenzen\ der\ gepaarten\ Mittelwerte}$$

1-sample t-Test

$$t = \frac{Mittelwert\ derStichprobe - Referenzwert}{SEM\ der\ Stichprobe}$$

https://matheguru.com/stochastik/t-test.html

8. Vorlesung WS 2020/21 Softwarewerkzeuge 35

Der student **t-Test** vergleicht die Stärke des Effekts (z.B. wie stark unterscheiden sich die Mittelwerte in zwei Proben voneinander) mit dem Standardfehler des Mittelwerts.

Zweistichproben t-Test

Ungepaarter t-Test

$$t = \frac{Mittelwert\ von\ Stichprobe\ 1\ -\ Mittelwert\ von\ Stichprobe\ 2}{SEM\ für\ die\ Differenz\ der\ Mittelwerte}$$

Annahme hierbei: beide Stichproben sind annähernd normalverteilt und haben (nach der Normierung) die gleiche Standardabweichung.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\left(\sum X_1^2 - \frac{\left(\sum X_1\right)^2}{n_1} + \right)^2 + \left(\sum X_2^2 - \frac{\left(\sum X_2\right)^2}{n_2} + \right)^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Korrektur des SEM

geschätzte Varianz von X_1

Anzahl an Freiheits-Graden (Form der t-Verteilung)

geschätzte Varianz von X_2

Falls 2 Zufallsvariablen X and Y voneinander unabhängig sind, ist die Varianz ihrer Summe gleich der Summe der einzelnen Varianzen
$V(X+Y)=V(X)+V(Y)$

https://matheguru.com/stochastik/t-test.html

Der 2-sample (Zweistichproben) t-Test vergleicht die Mittelwerte zweier Verteilungen.

Differentielle Expressionsanalyse: *t*-Test

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

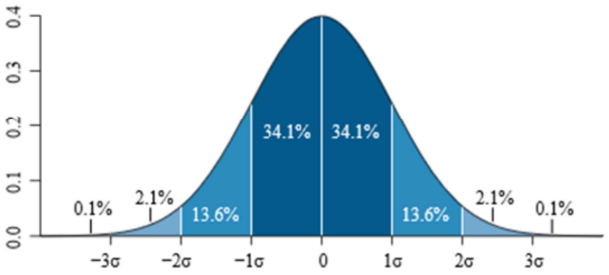
R_g : mittleres log Verhältnis der Expressionslevel für ein Gen *g* = "der Effekt"
SE : Standardfehler (erhalten durch Kombination der Daten für alle Gene = "die Variation in den Daten")

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

$$R_g = \overline{\log\left(\frac{x_i^{(1)}}{x_i^{(2)}}\right)} = \overline{log\left(x_i^{(1)}\right) - log\left(x_i^{(2)}\right)}$$

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)

Wann wird ein Gen nun als differentiell experimiert angesehen? Die Normalverteilung unten zeigt, dass nur knapp 5% der Kurve außerhalb des Intervalls [-2σ,+2σ] liegen. Die Hälfte davon links und die Hälfte davon rechts. Da man 5% üblicherweise als Signifikanzgrenze ansieht, kann man damit ungefähr abschätzen, wie groß eine Abweichung sein muss um als differentiell exprimiert angesehen werden. Allerdings verwendet der t-Test die sogenannte t-Statistik, die etwas flacher als die Normalverteilung verläuft.

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 ausreichend 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.

Hier wird eine Variante gezeigt, die nicht den Standardfehler aller Gene verwendet, sondern den Standardfehler des betreffenden Gens allein. Die Variation für dieses Gen könnte sich ja stark von dem mittleren Verhalten aller Gene unterscheiden.

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): $S = \frac{R_g}{c+SE_g}$

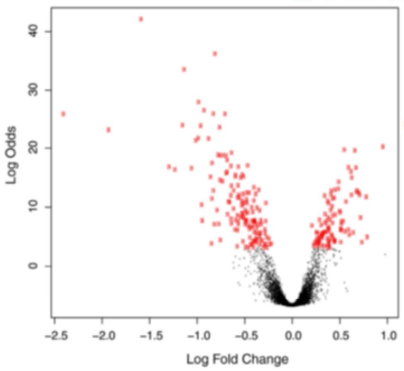
Durch diese Modifikation werden Gene mit kleinen fold changes (R_g) nicht als signifikant ausgewählt.

Die SAM-Methode liefert daher deutlich robustere Ergebnisse.

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

Die **SAM-Methode** ist eine sehr einfache Modifikation des genspezifischen t-Tests. Sie unterdrückt Zufallstreffer, die bei kleinem Rg und sehr kleinem Seg auftreten können.

Limma Paket: 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.

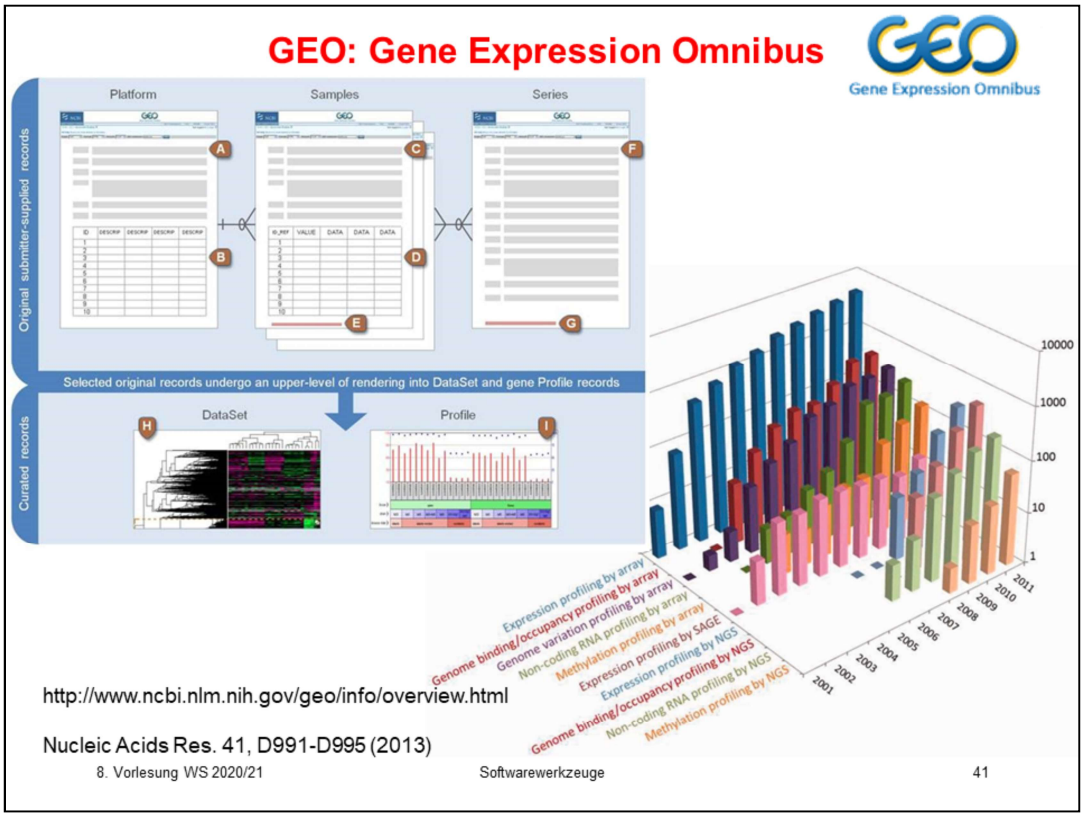
In dieser Abb ist dies der schwarz/rot-Übergang.

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

8. Vorlesung WS 2020/21 Softwarewerkzeuge 40

Der **Vulcano-Plot** ist eine sehr häufig verwendete Analysemethode. Man identifiziert damit Gene, die sowohl eine statistische signifikante Expressionsänderungen zeigen (y-Achse), die aber auch stark genug ist (x-Achse). Ein verwandtes Kriterium ist Cohen‘s d (https://de.wikipedia.org/wiki/Effektstärke). Bei einer großen Anzahl an Proben, fällt der p-Wert manchmal stark signifikant aus, obwohl die eigentliche Änderung gar nicht sehr groß ist.



In der GEO-Datenbank sind sehr viele Expressionsdatensätze öffentlich verfügbar. Von Forschern, die ein Manuskript mit Expressionsdaten bei einer Zeitschrift zur Veröffentlichung einreichen, wird erwartet, dass die Daten ebenfalls bei GEO eingereicht werden. Dies dient dazu, Forschung reproduzierbar zu machen und um die Dopplung von Experimenten zu reduzieren.

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

8. Vorlesung WS 2020/21 Softwarewerkzeuge 42

Ein weiterer Signifikanztest neben dem t-Test ist der **Mann-Whitney-Rangsummentest**. Der Vorteil gegenüber dem t-Test ist, dass diese Methode auch für Datensätze angewendet werden kann, die nicht normalverteilt sind. Wie in der unteren Tabelle gezeigt wird, bildet man eine gemeinsame Rangreihe der Datenpunkte. Falls die beiden Stichproben aus derselben Verteilung stammen (bzw. nur geringfügig variiert), würde man erwarten, dass sich die Daten aus beiden Stichproben in etwa abwechseln. Eine Anhäufung einer Stichproben entweder am unteren oder am oberen Rand der Rangreihe deutet dagegen auf einen systematischen Unterschied der Verteilungen hin.

Bewertung von Signifikanz: Mann Whitney Text

Die Rangsumme T_1 für Schulklasse A ist die Summe aller Rangplätze von Werten für Schulklasse A:

$T_1 = \sum_{m1=1}^{n_1} R_{m1}$ =2+5+8+9+11+12+13+15+16+19+22 = 132

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

Dies ergibt $U_1 = 55$

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

Für die Summe aller Rangplätze gilt (addiere das erste Element mit Rangplatz 1 und das letzte mit Rangplatz *n*, das zweite Element und *n* - 1 … das sind *n*/2 Terme jeweils mit der Summe *n*+1):

$$\sum_{m=1}^{n} R_m = \frac{n \cdot (n+1)}{2} = T_1 + T_2$$

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

Man vergleicht nun die tatsächlichen Rangsummen beider Datensätze mit einer zufällig verteilten Anordnung.

Bewertung von Signifikanz: Mann Whitney Text

Die Teststatistik U überprüft nun die Gleichmässigkeit der Verteilung der Rangplätze in der gemeinsamen Rangreihe.
U gibt die Summe der Rangplatzüberschreitungen an.

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 $U_1 + U_2 = n_1 n_2$

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

Man vergleicht nun die tatsächlichen Rangsummen beider Datensätze mit einer zufällig verteilten Anordnung. U ist die Summe der Rangplatzüberschreitungen.

Bewertung von Signifikanz: Mann Whitney Text

Als Prüfgrösse wird immer der kleinere der beiden Werte verwendet, hier also 55.

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

Das beobachtete Ungleichgewicht wird mit der U-Verteilung verglichen. Diese drückt aus, wie oft solch ein Ungleichgewicht zufällig auftreten kann.

Differentielle Expression aus RNA-seq Daten

Man bestimmt mit RNA-seq read counts für jedes Gen. Daraus muss man durch Abschätzung der Verteilung die vermutliche tatsächliche Anzahl jeder mRNA abschätzen. Bei der Abschätzung verwendet man meist die negative Binomialverteilung und schätzt deren Mittelwert und Varianz aus den beobachteten Daten. Hier geht die coverage = Sequenziertiefe ein.

Für ein bestimmtes Gen erhält man dann:

	Bedingung 1	**Bedingung 2**	**gesamt**
Gen *i*	n_{11}	n_{12}	$n_{11}+n_{12}$
restliche Gene	n_{21}	n_{22}	$n_{21}+n_{22}$
gesamt	$n_{11}+n_{21}$	$n_{22}+n_{22}$	n

Mit dem exakten Fisher-Test berechnet man dann den p-Wert, ob die Daten mit der Hypothese gleicher Expression in den Bedingungen 1 und 2 vereinbar sind:

$$p(\text{read count} \geq n_{11}) = \sum_{k=n_{11}}^{n_{11}+n_{12}} \frac{\binom{k+n_{12}}{k}\binom{n_{21}+n_{22}}{n_{21}}}{\binom{n}{k+n_{21}}}$$

http://www.mi.fu-berlin.de/wiki/pub/ABI/GenomicsLecture13Materials/rnaseq2.pdf

8. Vorlesung WS 2020/21 Softwarewerkzeuge 46

Zum Abschluss folgt noch eine Folie zu der Bewertung von differentieller Expression anhand von RNAseq-Datensätzen.

Die dort beobachteten Daten (reads) sind ein Abschätzung für die tatsächlichen Expressionslevel der Gene. Man nimmt an, dass die Abdeckung der experimierten Bereiche durch reads im Wesentlichen ein stochastischer Prozess ist. Dies würde man üblicherweise durch eine Poisson-Verteilung modellieren. Allerdings ist die Poisson-Verteilung etwas zu unflexibel, da bei ihr sowohl Mittelwert als auch Varianz gleich dem Parameter λ sind. Stattdessen verwendet man häufig die negative Binomialverteilung.

Aus den abgeschätzten tatsächlichen mRNA-Anzahlen berechnet man dann mit dem exakten Fisher-Text den p-Wert.

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.

Beim Vergleich von gesunden Gewebeproben mit Tumorproben findet man oft Hunderte bis ein paar Tausende an differentiell exprimierten Genen. Es ist sehr mühsam, diese Listen an Genen zu durchsuchen, um eine biologische Bedeutung in den Ergebnissen zu „lesen“. In der nächsten Vorlesung #9 werden wir uns daher mit Methoden zur funktionellen Annotation der differentiell exprimierten Gene beschäftigen. Damit bekommt man rasch einen Überblick, welche biologischen Prozesse und Pfade im Tumorgewebe relativ zu gesundem Gewebe hoch- bzw. runterreguliert sind.