Universitätsmedizin Rostock

Genexpressionsanalyse

31460 QB1 Epidemiologie, Med. Biometrie u. Med. Informatik

Dr.rer.nat. Israel Barrantes

Institut für Biostatistik und Informatik in Medizin und Alternsforschung
Universitätsmedizin Rostock
Ernst-Heydemann-Str. 8, 18057 Rostock, Deutschland

israel.barrantes@uni-rostock.de

November 2025

Zentrale klinische Frage

„Was sind die molekularen Unterschiede zwischen Krebszellen und normalem menschlichem Gewebe, und wie können wir diese Unterschiede nutzen, um neue therapeutische Ziele zu identifizieren?"

Warum das wichtig ist:

- Krebs-Heterogenität erfordert personalisierte Behandlung
- Medikamentenresistenz erfordert neue Strategien
- Präzisionsmedizin benötigt molekulare Biomarker
- Patientenergebnisse hängen von zielgerichteten Therapien ab

Lernkontext:

- Dauer: 90-minütiges praktisches Seminar
- Plattform: Google Colab praktische Analyse
- Ziel: Brücke zwischen computergestützter Biologie und klinischer Medizin

Datensatz - Echte menschliche Krebsdaten

UHR = Krebsproben

- 10 verschiedene menschliche Krebszelllinien
- Typen: Brust, Leber, Gebärmutterhals, Hoden, Gehirn, Haut + Immunzellen
- **Klinische Relevanz:** Pan-Krebs molekulare Merkmale

HBR = Normale Kontrollen

- Gesundes Hirngewebe von 23 Personen
- Alter: 60-80 Jahre alt
- **Klinische Relevanz:** Baseline normaler Zellfunktion

Technische Spezifikationen

- **Probengröße:** 6 Proben (3 Krebs + 3 normale Replikate)
- **Datentyp:** Paired-End-RNA-Sequenzierung
- **Umfang:** Chromosom 22 Teilmenge
- **Quelle:** Griffith et al. (2015)

Qualitätskontrolle:

- ERCC-Spike-ins: 92 Kontrolltranskripte
- Validierung der RNA-seq-Genauigkeit

Warum Genexpressionsanalyse?

QB1-Modul Kontext: „Epidemiologie, Med. Biometrie u. Med. Informatik"

- Bioinformatik-Grundlagen → Angewendet in RNA-seq-Pipeline
- Biomarker-Kandidaten → Identifiziert durch differentielle Expression
- Individualisierte Medizin → Präzisionsonkologie-Anwendungen
- Genexpressionsdaten-Visualisierung

Essenziell für zukünftige Ärzte:

- Medizininformatik in der Praxis
- Evidenzbasierte Medizin
- Vorbereitung auf Präzisionsmedizin
- Karrierebereitschaft: Viele Methoden in der medizinischen Forschung und Diagnostik beinhalten Bioinformatik

Professionelle Kompetenzen:

- Kritische Bewertung genomischer Studien
- Verständnis personalisierter Medizin
- Grundlage für akademische Medizin

Lernziele

Hypothese

Krebszellen zeigen systematische Genexpressionsänderungen, die Folgendes offenbaren:

1. **Onkogene** (krebsfördernd), die überexprimiert sind
2. **Tumorsuppressoren** (krebsverhindernd), die stillgelegt sind
3. **Metabolische und Signalwege**, die verändert sind, um Krebswachstum zu unterstützen
4. **Wirkstoffziele**, die Krebszellen selektiv abtöten könnten

Technische Fertigkeiten:

- Genexpressionsanalyse
- Pathway-Anreicherung
- Wirkstoff-Repositionierung

Medizinisches Verständnis:

- Krebs-Gensignaturen
- Biomarker-Identifizierung
- Wirkstoff-Repositionierungsstrategien

Klinische Translation:

- Präzisionsmedizin-Ansätze
- Therapeutische Zielbewertung
- Wirkstoffkandidaten-Bewertung

Illumina-Sequenzierungs-Workflow

1. **Library-Vorbereitung:** RNA-Extraktion → cDNA-Synthese → Adapter-Ligation
2. **Cluster-Amplifikation:** Bridge-PCR erzeugt klonale Cluster auf Flow Cell
3. **Sequenzierung-durch-Synthese:** Fluoreszent markierte Nukleotide werden zyklisch hinzugefügt
4. **Base-Calling:** Kamera detektiert fluoreszierende Signale → Qualitätswerte zugewiesen

Paired-End-Sequenzierung

- Liest beide Enden von DNA-Fragmenten (R1 und R2)
- Bessere Mapping-Genauigkeit
- Verbesserte Splice-Junction-Detektion
- Erhöhte Quantifizierungsgenauigkeit

Wichtige Spezifikationen

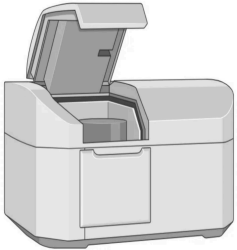
- Read-Länge: 50-150 bp pro Read
- Tiefe: Millionen von Reads pro Probe
- Qualität: Phred-Scores (Q30 = 99,9% Genauigkeit)

Genexpression

qPCR

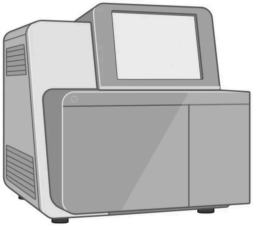
- Specific gene primers required
- Low throughput (1-100 genes)
- High precision quantification

Microarrays



- Predefined gene probes needed
- Medium throughput (~20K genes)
- Known transcripts only

RNA-seq



- **No probes required**
- **High throughput (all genes)**
- **Novel transcript discovery**

Genexpression und Quantifizierungsmethoden

Genexpressions-Grundlagen

- Zentrales Dogma: DNA → RNA → Protein
- RNA-seq misst mRNA-Abundanz
- Dynamikbereich: 0 bis >10.000 Reads

Quantifizierungs-Herausforderungen:

- Multi-Mapping-Reads
- Isoform-Komplexität
- Bias-Korrektur erforderlich

Evolution der Quantifizierungsmethoden

- **Traditionell:** Reads alignieren → pro Gen zählen
- **Neuere:** Leichtgewichtiges Mapping → probabilistische Zuordnung
- **Vorteile:** Schneller, behandelt Multi-Mapping, korrigiert Bias

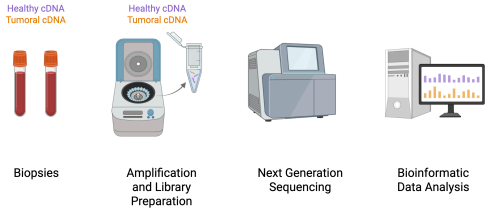
Expressionseinheiten:

- **Raw Counts:** Für statistische Analyse
- **TPM:** Transcripts Per Million (normalisiert)
- **FPKM/RPKM:** Ältere Methoden

Was ist RNA-seq?

RNA-seq-Prozess:
RNA-seq bestimmt, wie viele RNA-Moleküle (Genexpressionsniveau) in jeder Probe für jedes Gen vorhanden waren. Dieser Prozess umfasst:

1. **Mapping von Reads**: Bestimmen, von welchem Gen/Transkript jeder Sequenzierungs-Read stammt
2. **Zählen**: Aufaddieren, wie viele Reads zu jedem Gen mappen
3. **Normalisierung**: Anpassung für Sequenzierungstiefe und Genlängenunterschiede



Die vollständige Analyse-Pipeline

Praktisch

Teil 1: Setup und Grundlagen **(20 Min)**

Teil 2: Differentielle Expressionsanalyse mit DESeq2 **(20 Min)**

Teil 3: Pathway-Analyse **(15 Min)**

Teil 4: Drug Repositioning **(15 Min)**

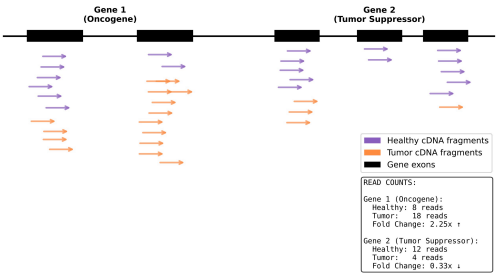
Lernansatz:

- Plattform: Google Colab (keine Installation erforderlich)
- Daten: Echte menschliche Krebs- vs. Normalproben
- Methoden: Industriestandard-Bioinformatik-Tools

Warum Count-Daten verwenden?

Vorteile von Count-Daten:

- **Raw Counts**: Repräsentieren die tatsächliche Anzahl von Sequenzierungs-Reads pro Gen
- **Statistische Anforderungen**: Count-basierte statistische Modelle (wie DESeq2) benötigen ganzzahlige Counts
- **Vergleichbarkeit**: Counts können über Proben hinweg normalisiert werden für fairen Vergleich



Beispiel-Count-Tabelle

		CANCER SAMPLES			NORMAL SAMPLES		
Gene ID	**Gene Name**	**UHR_Rep1**	**UHR_Rep2**	**UHR_Rep3**	**HBR_Rep1**	**HBR_Rep2**	**HBR_Rep3**
ENSG00000070371	MAPK1	902	1148	1070	206	171	288
ENSG00000128191	DGCR2	70	152	171	814	930	687
ENSG00000100296	MIF	372	351	370	401	350	436
ENSG00000093010	COMT	1093	1185	991	287	120	260
ENSG00000070831	CDC45	107	71	138	648	658	769
ENSG00000128218	VPREB1	516	451	498	498	483	498
ENSG00000100292	HMOX1	850	1163	854	163	230	150
ENSG00000186092	BCR	184	70	122	766	873	987
ENSG00000128274	A4GALT	397	351	346	427	390	339
ENSG00000100253	APOL1	1139	891	1166	287	298	271
ENSG00000100281	HMGB1	57	84	130	649	959	987
ENSG00000128228	SDF2L1	256	304	254	304	343	313

Technische Grundlagen - Dateiformate verstehen

FASTQ-Dateien (Rohdaten)

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAA
+
!''*((((****+))%%%++)(%%%%).1***
```

- 4 Zeilen pro Read
- Sequenz + Qualitätswerte
- Paired-End: R1 und R2 Dateien

GTF-Dateien (Annotationen)

```
chr22 gene 10736170 10736283
  gene_id "ENSG00000099"
  gene_name "POTEH"
```

FASTA-Dateien (Referenz)

```
>chr22
NNNNNNNNNNNNNNNNGATCACAGGTC
TATCACCCTATTAACCACTCACGGGAG
```

Klinische Relevanz: Das Verständnis von Datenformaten ist essenziell für die Interpretation genomischer Medizin

Quiz: RNA-seq-Grundlagen

Aussage	**R**	**F**
RNA-seq misst direkt die Protein-Abundanz in Zellen	R	F
Paired-End-Sequenzierung liest beide Enden von DNA-Fragmenten für verbesserte Genauigkeit	R	F
Höhere Phred-Qualitätswerte (z.B. Q30) zeigen niedrigere Sequenzierungsgenauigkeit an	R	F
FASTQ-Dateien enthalten sowohl die Nukleotidsequenz als auch Qualitätswerte für jeden Read	R	F
RNA-seq kann alternatives Spleißen und Gen-Isoformen detektieren	R	F

Differentielle Expressionsanalyse mit DESeq2

DESeq2 Statistisches Framework

- Modelliert Count-Daten mit negativer Binomialverteilung
- Berücksichtigt biologische Variabilität
- Schätzt Dispersionsparameter
- Führt Hypothesentests durch

Wichtige Output-Metriken:

- Log2 Fold Change (Effektgröße)
- P-Wert (statistische Signifikanz)
- Adjustierter p-Wert (Korrektur für multiples Testen)
- Base-Mean-Expressionsniveau

Analyse-Workflow

1. DESeq2-Datensatz aus Count-Matrix erstellen
2. Normales Gewebe als Referenz festlegen
3. Gene mit niedrigen Counts filtern
4. Differentielle Expression durchführen
5. Onkogene und Tumorsuppressoren identifizieren

Signifikanz-Schwellenwerte:

- Adjustierter p-Wert $< 0{,}05$
- $|\text{Log2 Fold Change}| > 1$
- Entspricht 2-facher Änderung

Visualisierung molekularer Signaturen

Volcano-Plot

- X-Achse: Log2 Fold Change
- Y-Achse: -Log10 p-Wert
- Rot: Onkogene (hoch)
- Blau: Tumorsuppressoren (runter)

PCA-Plot

- Zeigt Proben-Clustering
- PC1 & PC2 erklären Varianz
- Klare Trennung = starkes Signal
- Validiert Versuchsdesign

Heatmap

- z.B. Top 20 Krebs-Gene
- Zeilen-skalierte Expression
- Konsistente Muster
- Potenzielle Biomarker

Klinische Interpretation:

- PCA: Klare Trennung zeigt robuste molekulare Unterschiede
- Heatmap: Konsistente Muster über Replikate hinweg deuten auf zuverlässige Krebs-Biomarker hin

Ergebnisse - Identifizierte Krebs-Gensignaturen

Onkogene (Überexprimiert bei Krebs)

Eigenschaften:

- Positiver Log2 Fold Change (> 1)
- Statistisch signifikant (adj. $p < 0{,}05$)
- Fördern Zellproliferation
- Potenzielle Wirkstoffziele zur Hemmung

Erwartete Funktionen, z.B.:

- Zellzyklus-Progression
- Resistenz gegen Apoptose

Tumorsuppressoren (Unterexprimiert)

Eigenschaften:

- Negativer Log2 Fold Change (< -1)
- Statistisch signifikant (adj. $p < 0{,}05$)
- Hemmen Krebsprogression
- Ziele für Wiederherstellungstherapie

Erwartete Funktionen, z.B.:

- Zellzyklus-Checkpoints
- Apoptose-Induktion

Klinische Bedeutung: Diese Gensignaturen können als diagnostische Biomarker, prognostische Indikatoren und therapeutische Ziele dienen

Quiz: Grundlagen differentieller Expression

Aussage	**R**	**F**
DESeq2 verwendet eine negative Binomialverteilung zur Modellierung von RNA-seq-Count-Daten	R	F
Log2 Fold Change von $+2$ bedeutet, dass das Gen 4-mal stärker bei Krebs vs. normal exprimiert ist	R	F
Adjustierte p-Werte (padj) korrigieren für multiples Hypothesentesten	R	F
Gene mit padj $< 0{,}05$ und $\lvert\text{log2FC}\rvert > 1$ gelten als signifikant differentiell exprimiert	R	F
Der Base-Mean-Expressionswert repräsentiert den durchschnittlichen normalisierten Count über alle Proben	R	F

Pathway-Analyse verstehen

Pathway-Analyse identifiziert Stoffwechsel- und Signalwege, die beim Vergleich verschiedener Bedingungen signifikant verändert sind.

Erwartete hochregulierte Wege, z.B.:

- Zellzyklus-Progression
- Angiogenese (Blutgefäßbildung)
- Zellproliferationssignale

Klinische Bedeutung:

- Therapeutische Vulnerabilitäten identifizieren
- Krebsbiologie verstehen
- Kombinationstherapien leiten

Erwartete herunterregulierte Wege, z.B.:

- Apoptose (programmierter Zelltod)
- DNA-Reparaturmechanismen
- Zellzyklus-Checkpoints

Tools:

- gprofiler, DAVID, Enrichr
- GO (Gene Ontology) Datenbanken
- KEGG, Reactome Pathways

Drug-Repositioning

Was ist Drug-Repositioning ?

Der Prozess der Identifizierung neuer therapeutischer Anwendungen für bestehende FDA-zugelassene Medikamente, wodurch die Wirkstoffentwicklung beschleunigt und Kosten reduziert werden.

Vorteile:

- **Schneller:** 5-10 vs. 15-20 Jahre
- **Sicherer:** Bekannte Sicherheitsprofile (reduz. Risiko)
- **Günstiger:** Niedrigere Entwicklungskosten
- **Zugänglich:** Wirkstoffe bereits verfügbar

Computergestützter Ansatz:

- Krankheits-Gensignaturen identifizieren
- Mit Wirkstoff-induzierten Expressionssignaturen abgleichen
- Wirkstoffe identifizieren, die Krankheitsmuster umkehren oder imitieren

L1000CDS2-Datenbank:

- Genexpressionssignaturen für 1000+ Wirkstoffe
- Wirkstoffeffekte auf Zelllinien
- Ermöglicht Signatur-Matching
- Prognostiziert therapeutisches Potenzial

Drug-Repositioning Ergebnisse - Behandlungskandidaten

Bekannte Krebs-Wirkstoffe

Validierung des Ansatzes:

- Doxorubicin
- Paclitaxel
- Cisplatin
- Tamoxifen
- Imatinib

Das Erscheinen dieser Wirkstoffe in den Ergebnissen validiert unsere computergestützte Methode

Repositionierte Wirkstoffe

Neue Möglichkeiten:

- **Metformin** (Diabetes)
- **Aspirin** (entzündungshemmend)
- **Statine** (Cholesterin)
- **Rapamycin** (Immunsuppressivum)

Bereits in klinischen Studien oder für Krebs zugelassen

Natürliche Verbindungen

Präventives Potenzial:

- Curcumin
- Resveratrol
- Quercetin

Diätetische Interventionen und adjuvante Therapie

Klinische Translation: Genexpressionsanalyse → Wirkstoff-Vorhersage → Labor-Validierung → Klinische Studien → Patientenbehandlung

Quiz: Grundlagen der Drug-Repositioning

Aussage	**R**	**F**
Wirkstoff-Repositionierung beinhaltet das Finden neuer therapeutischer Anwendungen für bestehende FDA-zugelassene Medikamente	R	F
Computergestützte Wirkstoff-Repositionierung ist schneller und weniger teuer als die Entwicklung völlig neuer Wirkstoffe	R	F
Wirkstoff-Repositionierung erfordert das Starten klinischer Studien ab Phase-I-Sicherheitsstudien	R	F
Die L1000CDS2-Datenbank enthält Genexpressionssignaturen für Tausende von Wirkstoffbehandlungen	R	F
Signatur-Matching bei Wirkstoff-Repositionierung zielt darauf ab, Wirkstoffe zu finden, die Krankheits-Genmuster umkehren	R	F

Klinische Translation - Präzisionsmedizin-Anwendungen

Biomarker-Anwendungen

Diagnostische Marker:

- Früherkennung von Krebs
- Molekulare Klassifizierung
- Krebs vs. benigne Unterscheidung

Prognostische Marker:

- Patientenergebnisse vorhersagen
- Risikostratifizierung
- Langzeitüberwachung

Prädiktive Marker:

- Responsive Patienten auswählen
- Ineffektive Therapien vermeiden
- Kombinationsstrategien leiten

Klinisches Entscheidungsframework

1. Tumor-RNA-seq-Profiling
2. Vergleich mit normaler Gewebe-Baseline
3. Dominante Krebs-Pathways identifizieren
4. Abgleich mit zielgerichteten Therapien
5. Behandlungsantwort überwachen

Behandlungsauswahl:

- Hohe Onkogen-Expression $\rightarrow$ Zielgerichtete Inhibitoren
- Niedriger Tumorsuppressor $\rightarrow$ Wiederherstellungstherapie
- Pathway-Aktivierung $\rightarrow$ Pathway-Wirkstoffe

Limitierungen und zukünftige Richtungen

Aktuelle Limitierungen

Technisch:

- Chr22-Teilmenge vs. vollständiges Transkriptom
- Tumor-Heterogenitäts-Herausforderungen
- Validierungslücke: Berechnung zu Klinik
- Kostenbarrieren für Routinenutzung

Klinisch:

- Zeitliche Evolution während der Behandlung
- Gewebespezifitäts-Anforderungen
- Bevölkerungsvielfalt-Überlegungen
- Workflow-Integrations-Herausforderungen

Zukünftige Richtungen

Technologisch, z.B.:

- Multi-Omics-Integration
- Einzelzell-RNA-seq-Analyse

Computergestützt, z.B.:

- Machine-Learning/KI-Ansätze
- Genregulatorische Netzwerkanalyse

Klinisch:

- Integration in Behandlungsleitlinien
- Ärzteausbildungsprogramme
- Kosten-Wirksamkeits-Studien

Schlussfolgerungen und klinische Auswirkungen

Technische Erfolge

- ✓ Vollständige RNA-seq-Pipeline: FASTQ → Therapeutika
- ✓ Krebs-Biomarker-Identifizierung mit statistischer Strenge
- ✓ Wirkstoff-Repositionierungskandidaten computergestützt entdeckt
- ✓ Professioneller Bioinformatik-Workflow gemeistert
- ✓ Klinische Anwendungen verstanden und bewertet

Wichtige Lernergebnisse

Für zukünftige Ärzte:

- Transkriptom-Kompetenz
- Dateninterpretation
- Klinische Integration
- Forschungskompetenz

Professionelle Fähigkeiten:

- Bioinformatik-Workflow-Management
- Statistische Analyse-Interpretation
- Wissenschaftliche Kommunikation
- Computergestützte Problemlösung