Freie Universität Berlin

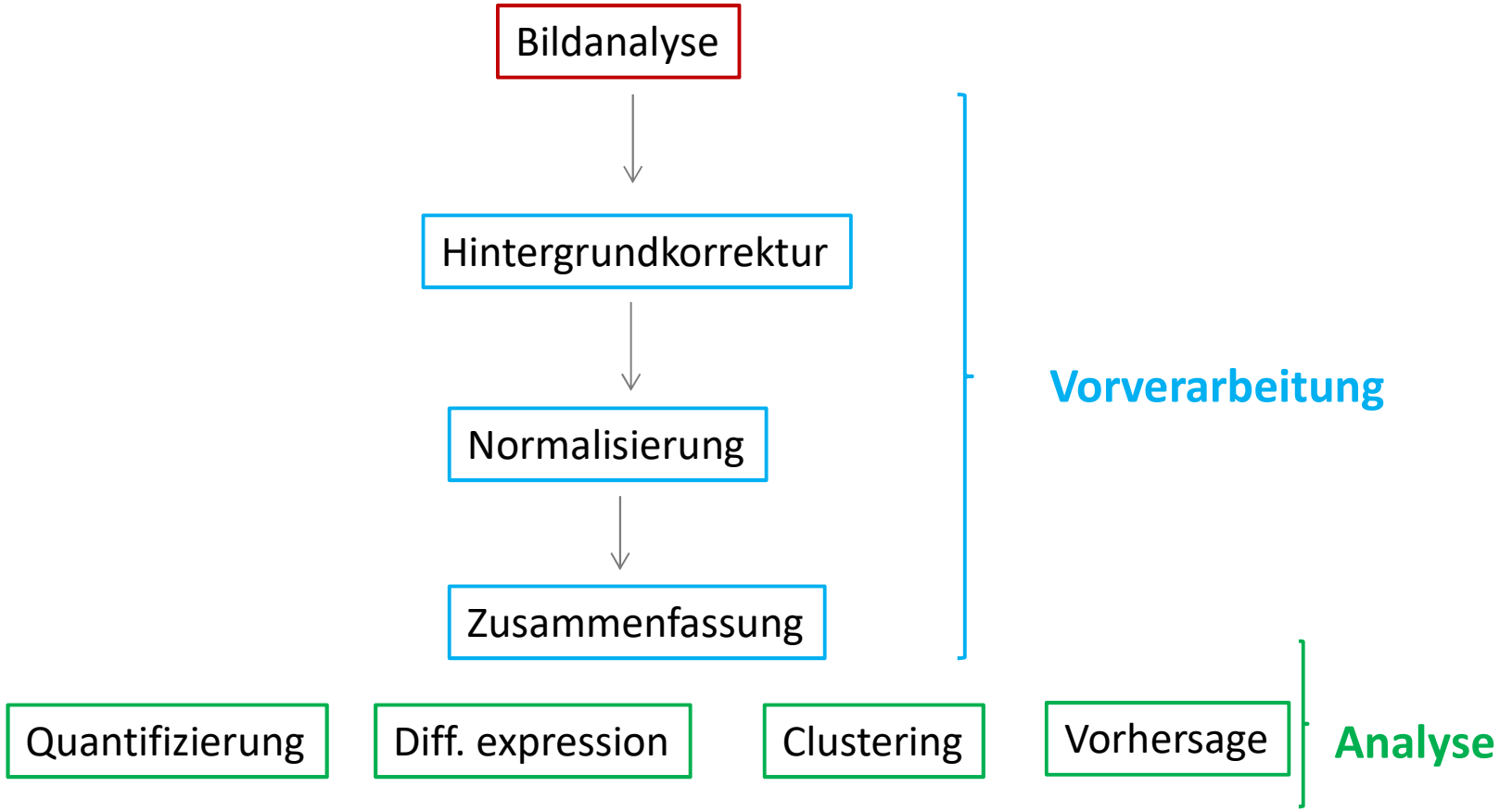
MAX-PLANCK-GESELLSCHAFT

AlgoBio WS 16/17
Differenzielle Genexpression

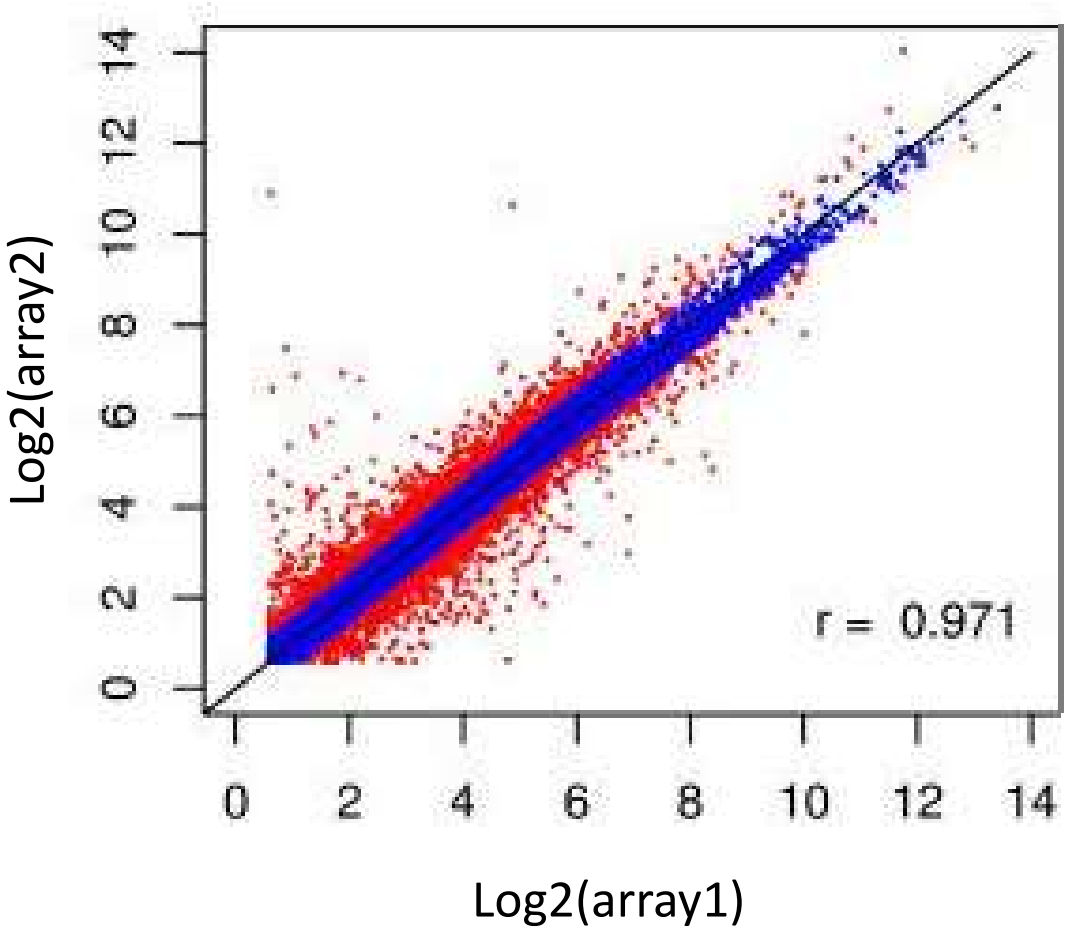
Annalisa Marsico

04.01.2017

Pipeline für die Mikroarray-Analyse



Microarrays re-cap



Differentielle Genexpression

Beim Vergleich von zwei Bedingungen (ca. 20000 Genmessungen für jedes ‚Sample') wollen wir vor allem wissen:

"Welche Gene sind differentiell exprimiert zwischen den beiden Bedingungen?"

- Wie finden wir differenziell exprimierte Gene?
  - Wir brauchen eine Zahl zur Quantifizierung des Differentialausdrucks

- Fühlen wir uns konfident genug, um diesen Unterschied zu berichten? (z.B. in einem Paper)

Differentielle Genexpression

Beim Vergleich von zwei Bedingungen (ca. 20000 Genmessungen für jedes ‚Sample') wollen wir vor allem wissen:

"Welche Gene sind differentiell exprimiert zwischen den beiden Bedingungen?"

- Wie finden wir differenziell exprimierte Gene?
  - Wir brauchen ein Zahl zur Quantifizierung des Differentialausdrucks
  - Fold changes sind die bevorzugte Quantifizierung des Differentialausdruck. Fold changes sind grundsätzlich Ratios
- Fühlen wir uns konfident genug, um diesen Unterschied zu berichten? (z.B. in einem Paper)

Warum logs ratios?

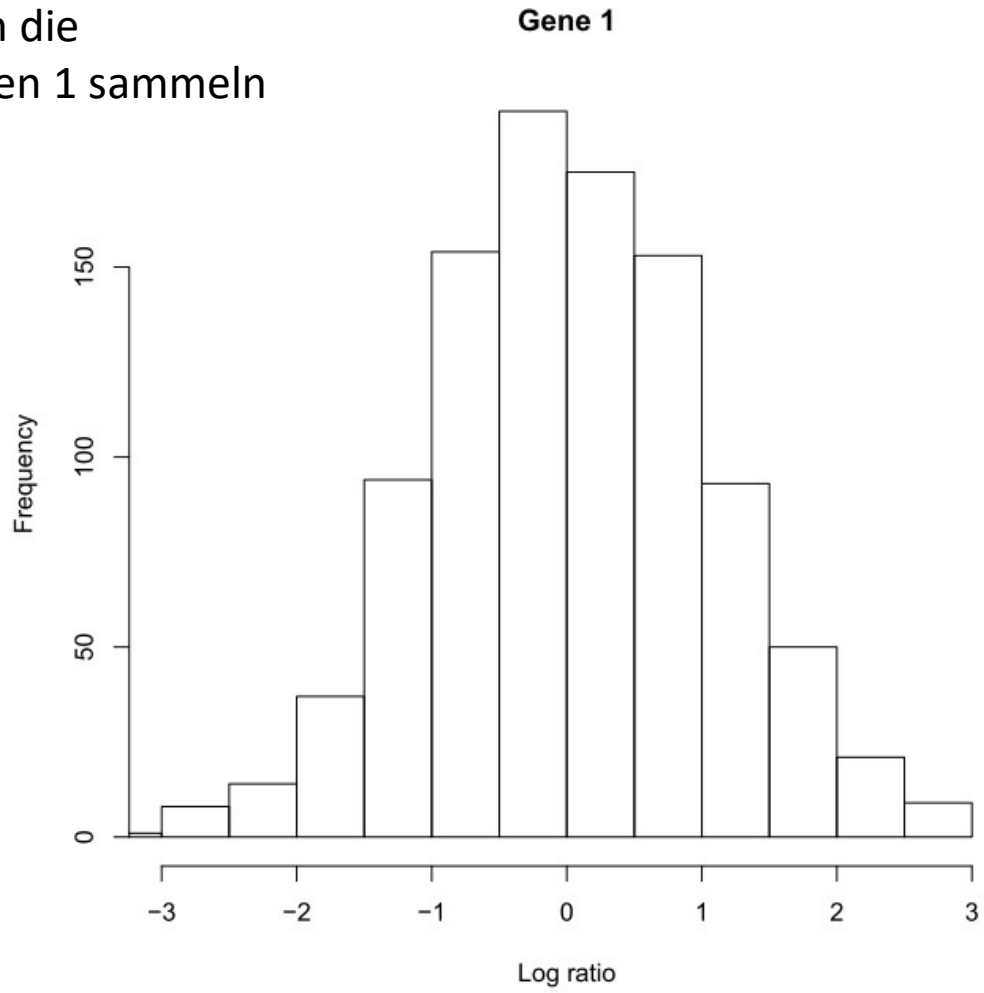
- Wie finden wir differenziell exprimierte Gene?
  - Zahl zur Quantifizierung des Differentialausdrucks
  - **Fold changes (Ratios)** werden verwendet, um den differentiellen Ausdruck zu quantifizieren
    - Ratios $\frac{\exp\_array1}{\exp\_array2}$ sind nicht um 1 symmetrisch. Dies macht die Interpretation schwieriger
    - Wir werden also $log_2\left(\frac{\exp\_array1}{\exp\_array2}\right)$ verwenden

Differentielle Expressionaufgabe

- Wie orden / repräsentieren wir Gene nach “*level of interest*”?
- Ein Zahl fur Differentiellexpression definieren (***Fold change***):
  Große Werte (positiv oder negativ) werden als *interessant* betrachtet
- Wie interessant sind denn die interessantesten Gene? **Statistical testing** hilft uns zu entscheiden.

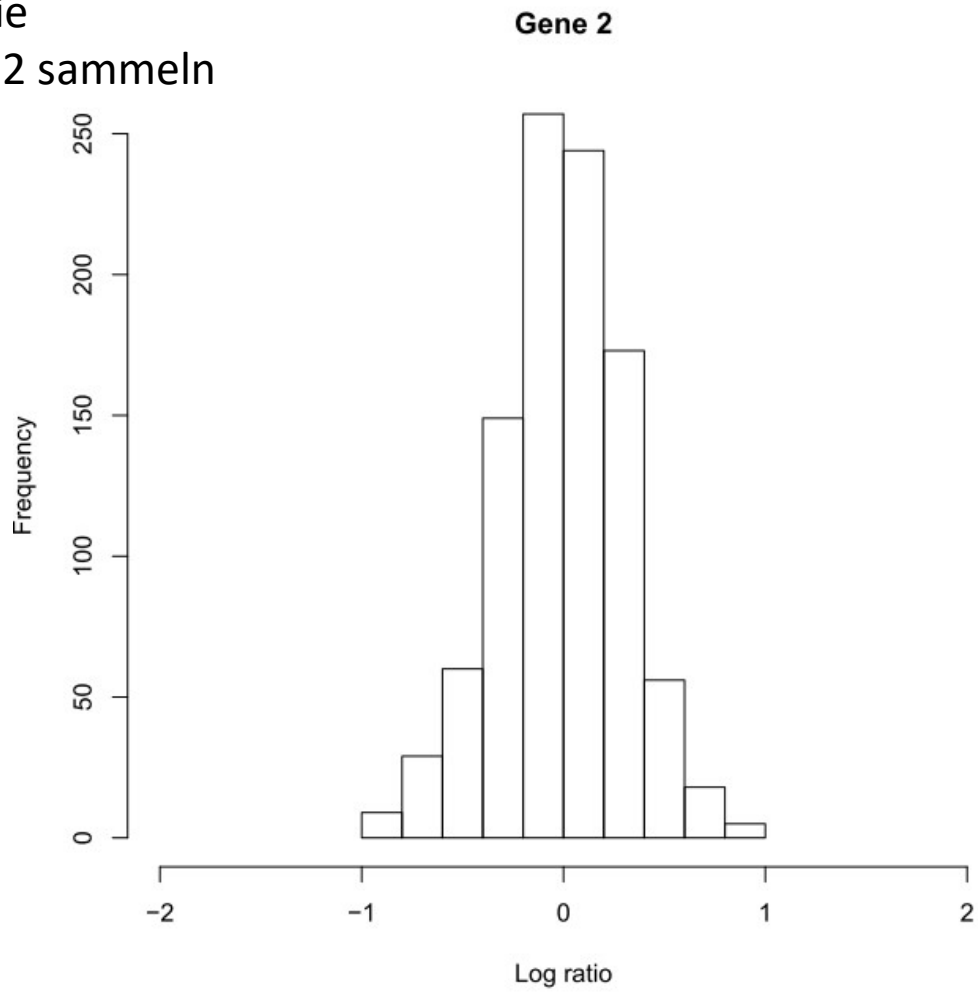
Wiederholtes Experiment

Wiederholen des Experiments
1000 mal und dann die
Fold changes für Gen 1 sammeln

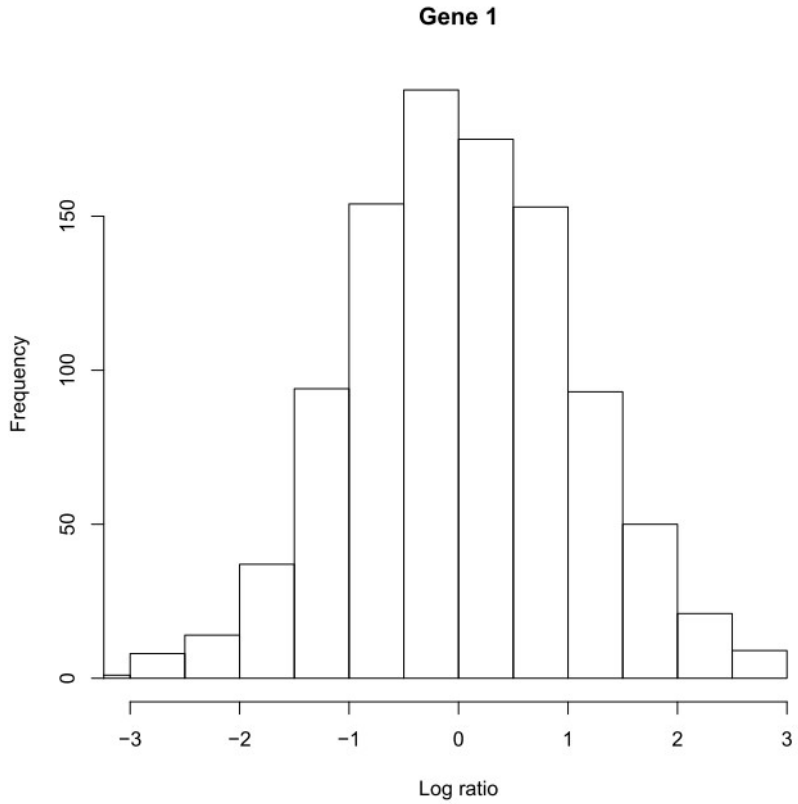


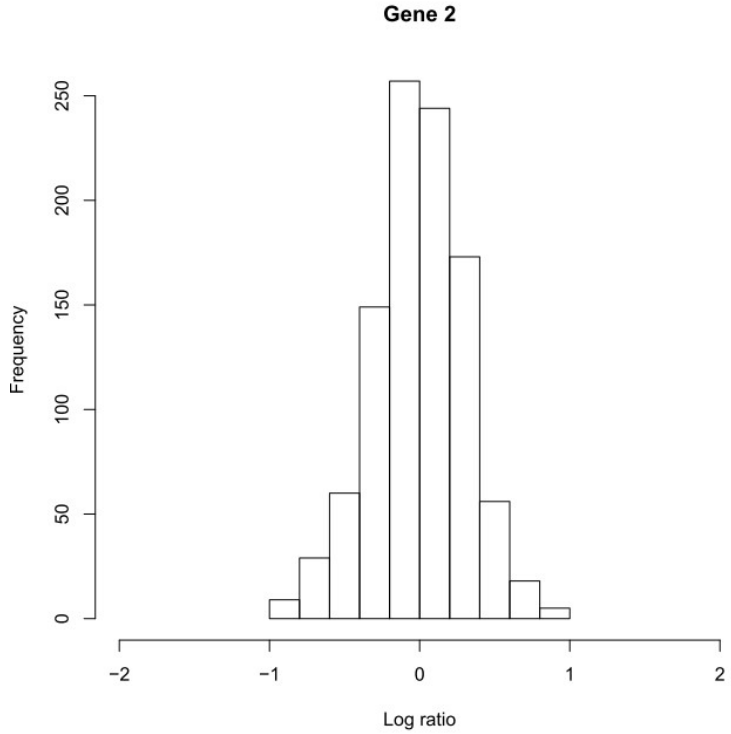
Wiederholtes Experiment

Wiederholen des Experiments
1000 mal und dann die
Fold changes für Gen 2 sammeln



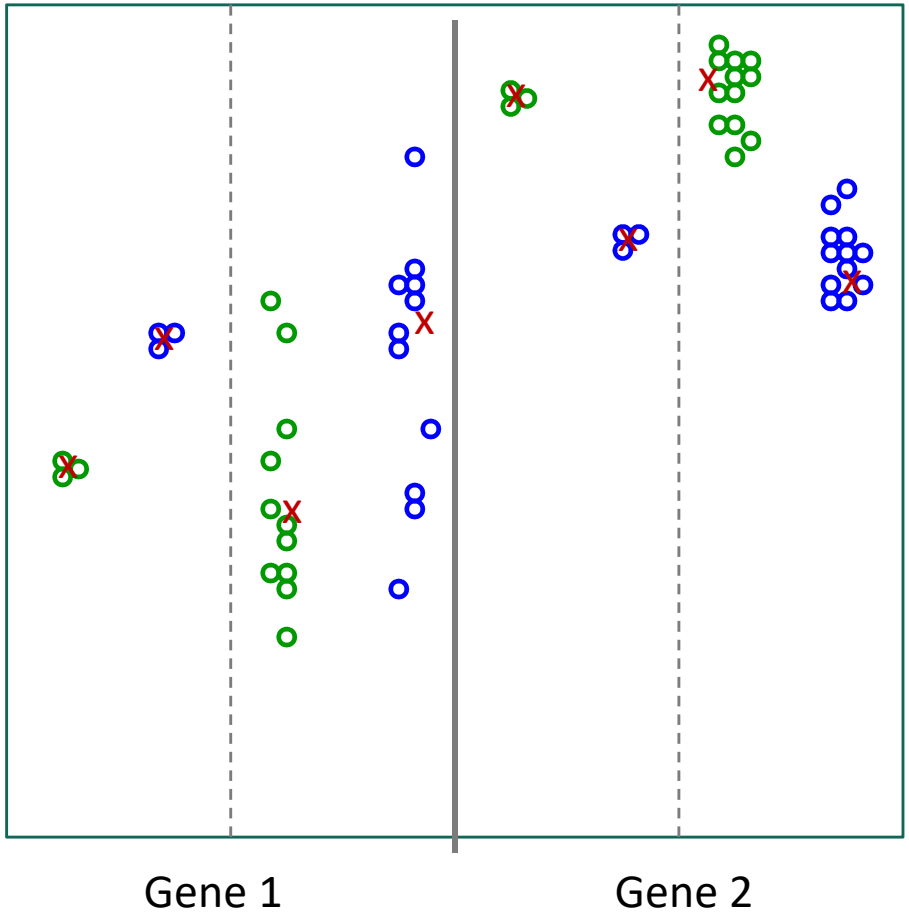
Wiederholtes Experiment





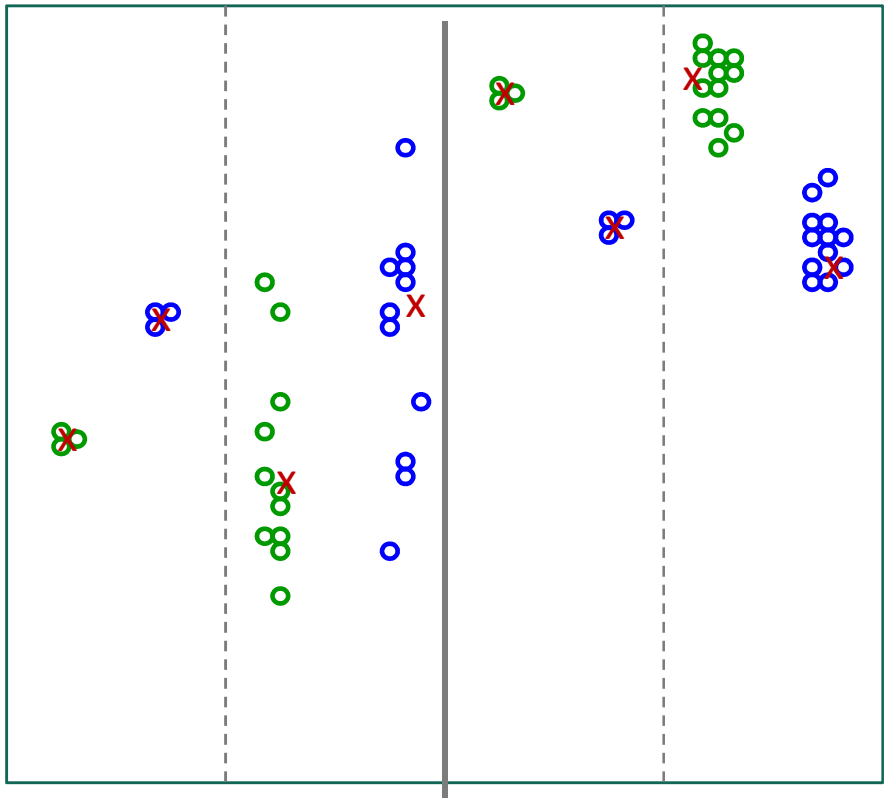
Ein *Fold change* von 1 scheint für Gen 2 signifikanter als für Gen 1.
Ein *Fold change* könnte auch nur wegen natürlicher Variabilität beobachtet werden (Gen 1)

Technische und biologische Variabilität



Welches dieser Gene wird reproduziert, wenn ein anderes Labor sie testet?

Technische und biologische Variabilität



Gene 2 ist konsistent, so dass wir berichten würden: die Mittelwerte der beiden Populationen sind sehr verschieden und innerhalb der Population gibt es nicht so viel Variabilität. Der p-Wert aus einem T-Test gibt uns eine formale Art um diese Entscheidung zu treffen.

Review der Statistical Inference

- Seien $Y - X$ unsere Messungen, die differentielle Genexpression repräsentieren (a log(fold_change))
- Was ist die typische Nullhypothese?
- P-value ist die Wahrscheinlichkeit von $Y - X$ größer oder gleich als der beobachtete Wert unter der Nullhypothese zu sein. Das ist ein Weg zusammenzufassen, wie *interessant* ein Gen ist.
- Annahme: Unter der Nullhypothese folgt $Y - X$ einer Normalverteilung mit dem Mittelwert 0 und der Standardabweichung σ
- Ohne σ kennen wir nicht den p-Value
- Wir können σ durch das Nehmen eines Samples und unter Nutzung der *Sample Standardabweichung s* schätzen.

Sample Summaries

Beobachtungen: $X_1, \ldots\ldots\ldots X_M$ $\qquad\qquad$ $Y_1, \ldots\ldots\ldots Y_M$

Mittelwerte: $\bar{X} = \frac{1}{M}\sum_{i=1}^{M} X_i$ $\qquad\qquad$ $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$

SD^2 (Statistiker nennen diese $Varianzen$):

$s_x^2 = \frac{1}{M-1}\sum_{i=1}^{M}(X_i - \bar{X})^2;$ $\qquad\qquad$ $s_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$

Die t-Statistik

T-Statistik:

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{s_y^2}{N} + \frac{s_x^2}{M}}}$$

Multiple Testing Problem

http://xkcd.com/882/

Multiple Testing Problem

- Messungen von 10,000 Genen
- 10,000 p-values berechnen
- Nenne Gene “signifikant” wenn p-value < 0.05

- Erwartete Anzahl von falschen Positiven:

10,000 x 0.05 = 500 false positives

Multiple Comparison Error Rates

- **Family-wise error rate**

$$P(\#\text{ False Positives} = 1)$$

- **False Discovery rate**

$$E\left[\frac{\#\ False\ Positives}{\#\ Discoveries}\right]$$

Multiple Comparison Error Rate

- Annahme: 50 von 10,000 Genen sind signifikant mit einem 0.05 Signifikanzniveau

Keine Korrektur

Erwarte 0.05*10,000=500 false positives

False Discovery Rate

Erwarte 0.05*50=2.5 false positives

Family Wise Error Rate

Die Wahrscheinlichkeit von mindestens einem false positive < 0.05

Kontrolle der Error Rates

Bonferroni Korrektur

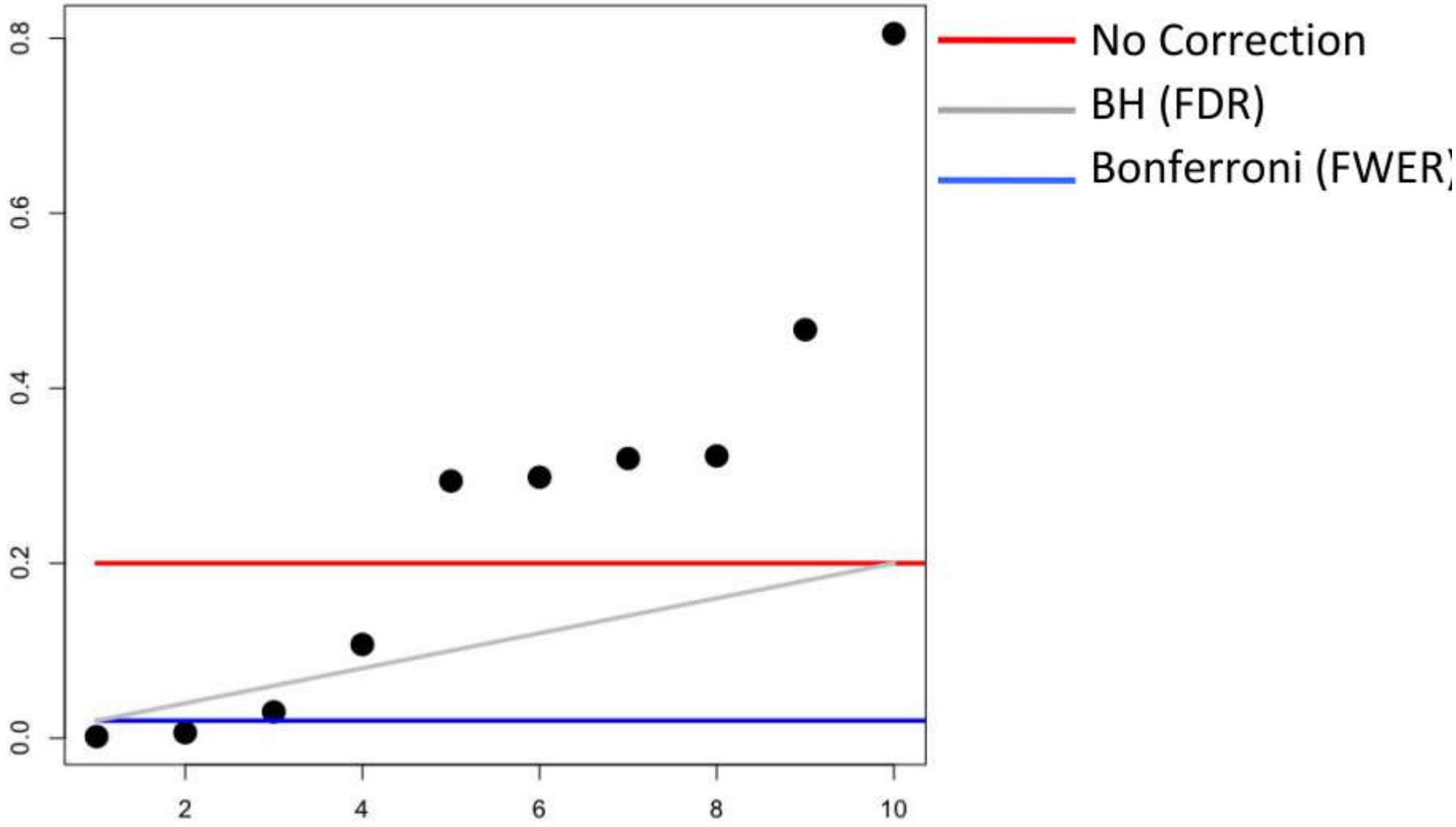
P-values kleiner als α/m sind signifikant

Benjamini-Hochberg Korrektur

Ordne die p-values: $p_{(1)}, \ldots\ldots p_{(m)}$

Wenn $p_{(i)} \leq \alpha \times i/m$ dann ist es signifikant

Beispiel mit 10 p-values



High-throughput RNA Sequencing als Alternative zu Microarrays um Genexpression zu messen

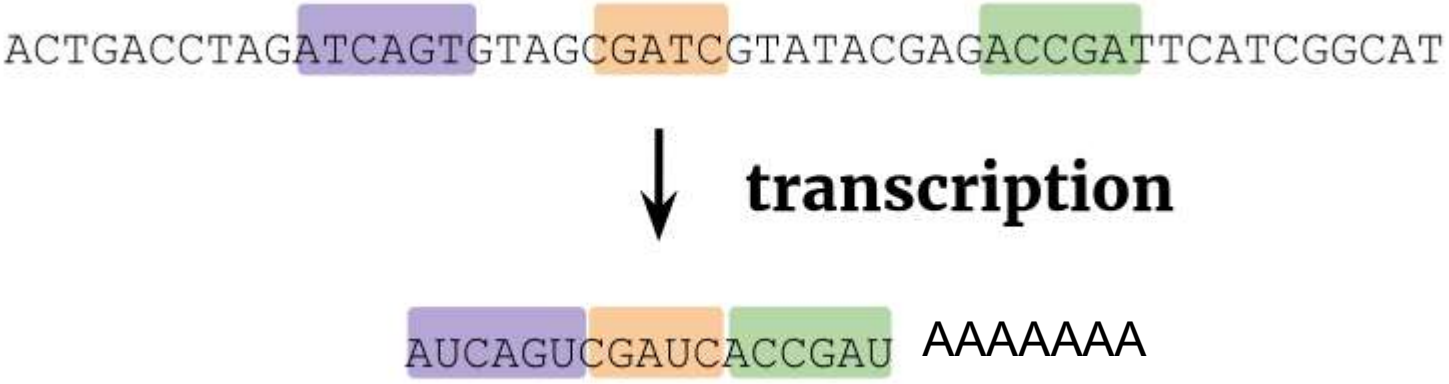
RNAs

Polyadenyliert (kodierende) RNAs, “Gene”
Kurze nicht-kodierende RNAs (ncRNAs), z.B. microRNAs
Lange nicht-kodierende RNAs
Ribosomal RNA

Total RNA

Größter Teil der RNA in der Zelle

Anreicherung
polyA capture
ribominus



RNA-Seq

ACTGACCTAGATCAGTGTAGCGATCGTATACGAGACCGATTCATCGGCAT

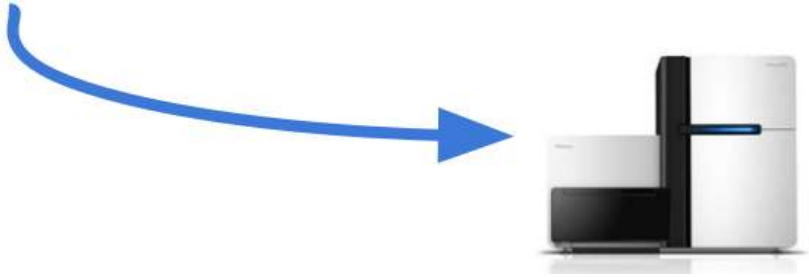
↓ **transcription**

AUCAGUCGAUCACCGAU AAAAAAA

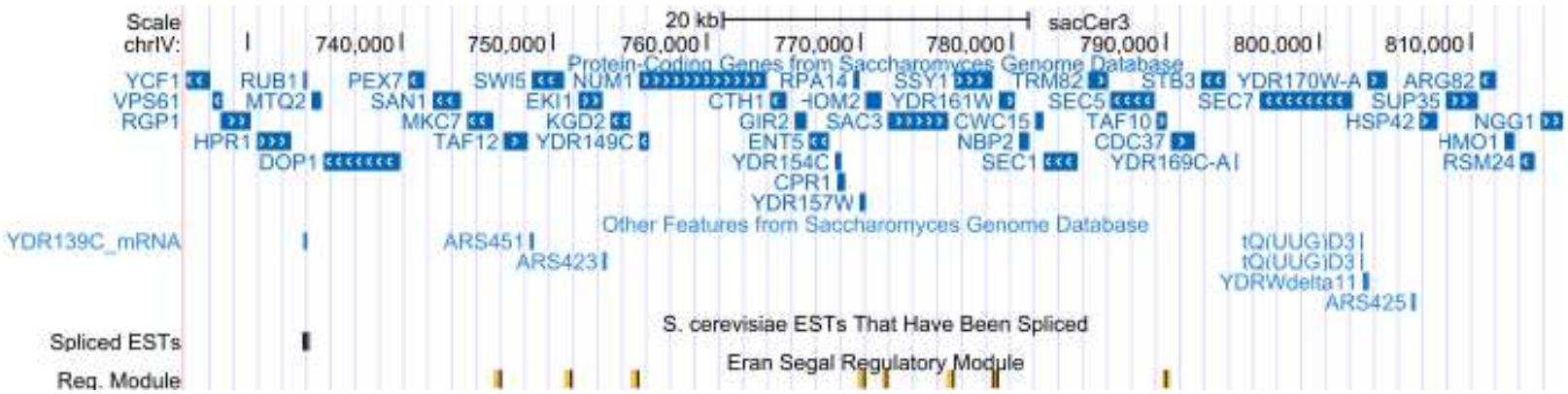
Erfassen der mature RNA durch den poly(A) tail

TAGTCAGCTAGTTGGCTA

Invertierte Transkription in eine Complementary DNA (cDNA)

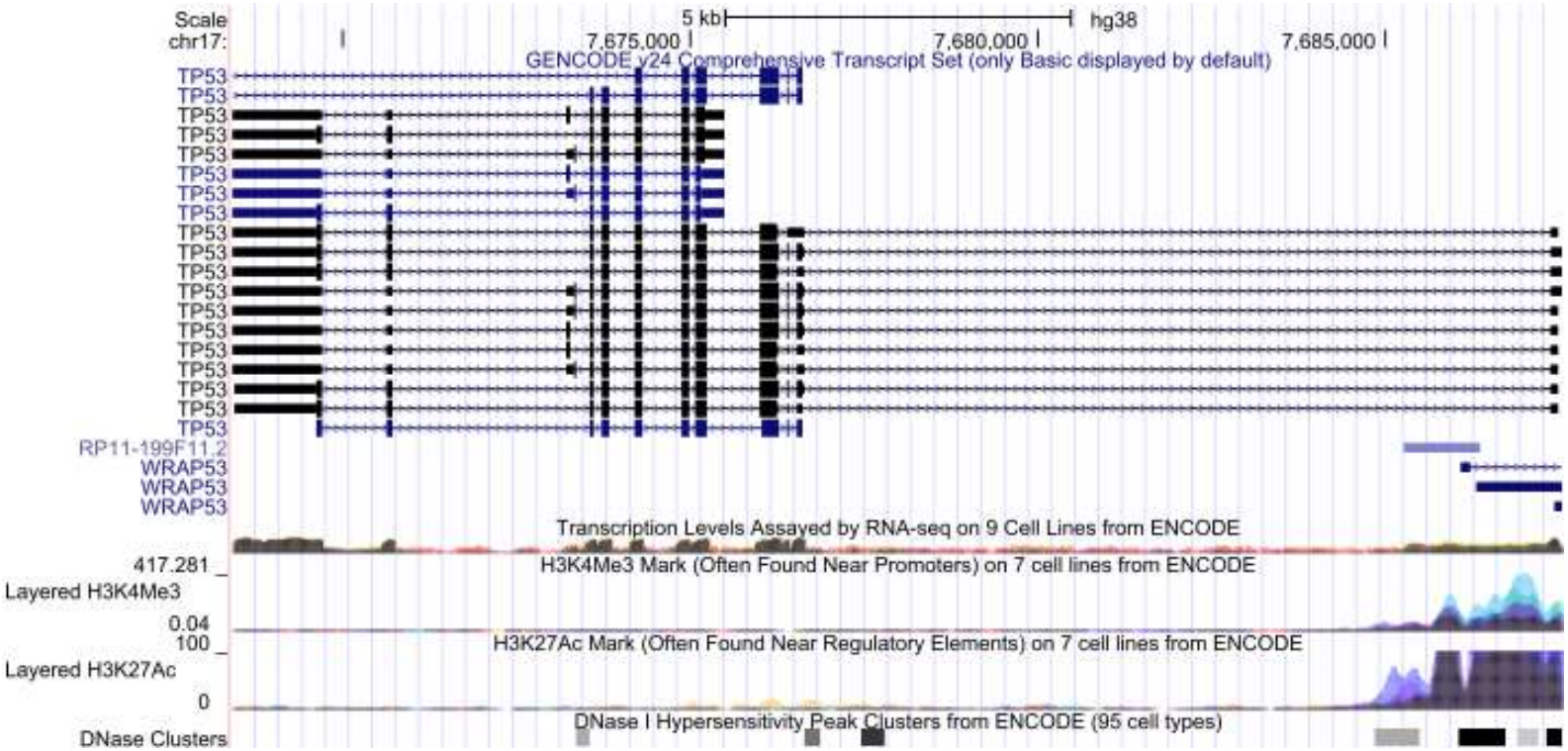


Yeast



Nur ein Transkript pro Gen
Kein (wenig) Splicing
Überlappende Gene
Wenig Raum zwischen den Genen

TP53 (Menschliches Gen)



Fragen, die wir mit RNA-seq beantworten können

- Was ist die Struktur von bekannten und unbekannten Transkripts?
- Veränderungen im Splicing
- Genexpression messen
- Transcriptexpression messen

RNA-Seq Standard Protokoll

- Extrahierung von RNA/ polyA Anreicherung
- Fragmentierung von RNA
- Inverse Transkription von RNA zu cDNA (durch random hex.)
- Ligation von Adapters
- Größenselektion ~200 bp (~300 bp)
- PCR Amplifikation
- Sequenzierung

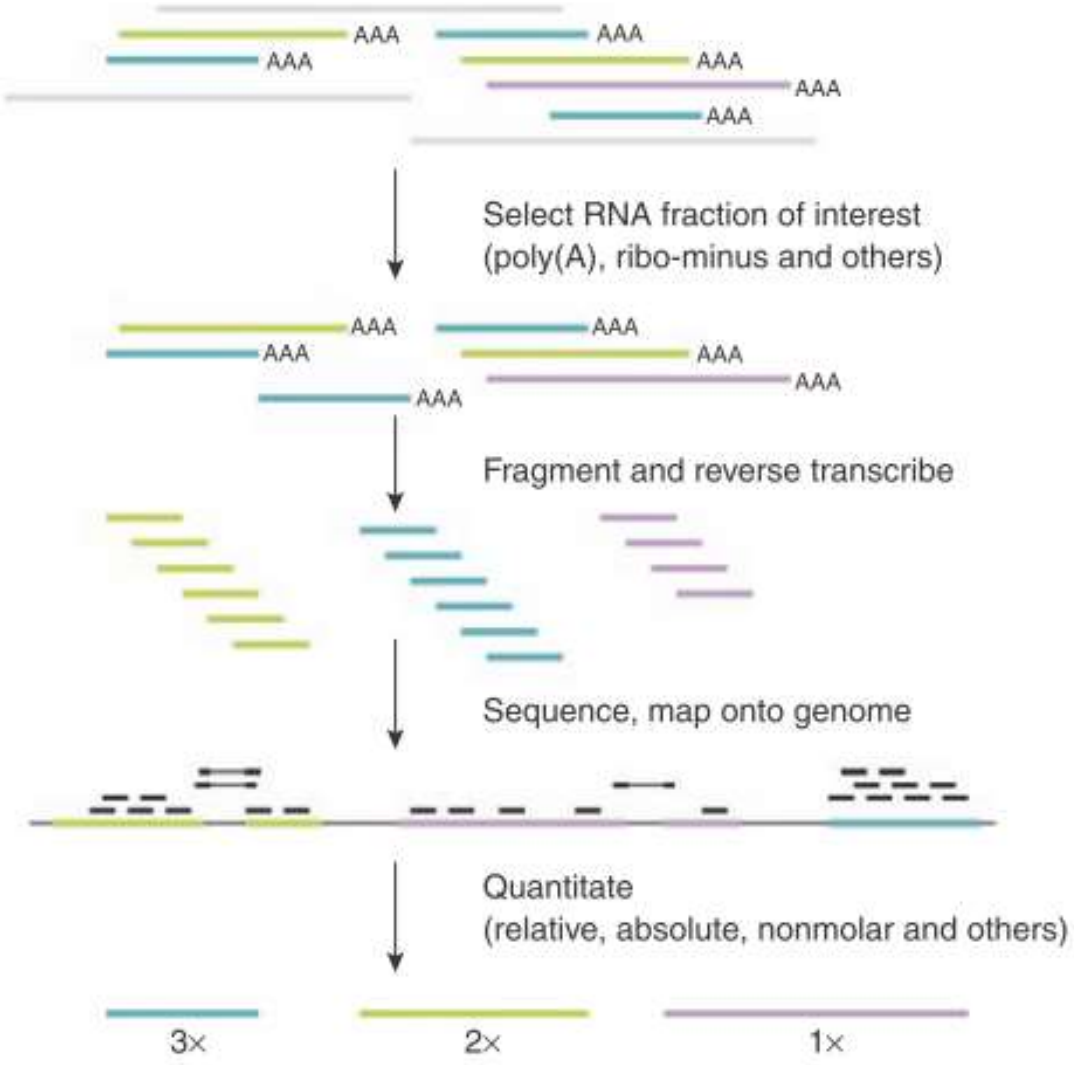
Das produziert reads von Polyadenylierter RNA ohne Stranginformation.

<u>Varianten</u>:

Ribominus statt polyA Anreicherung

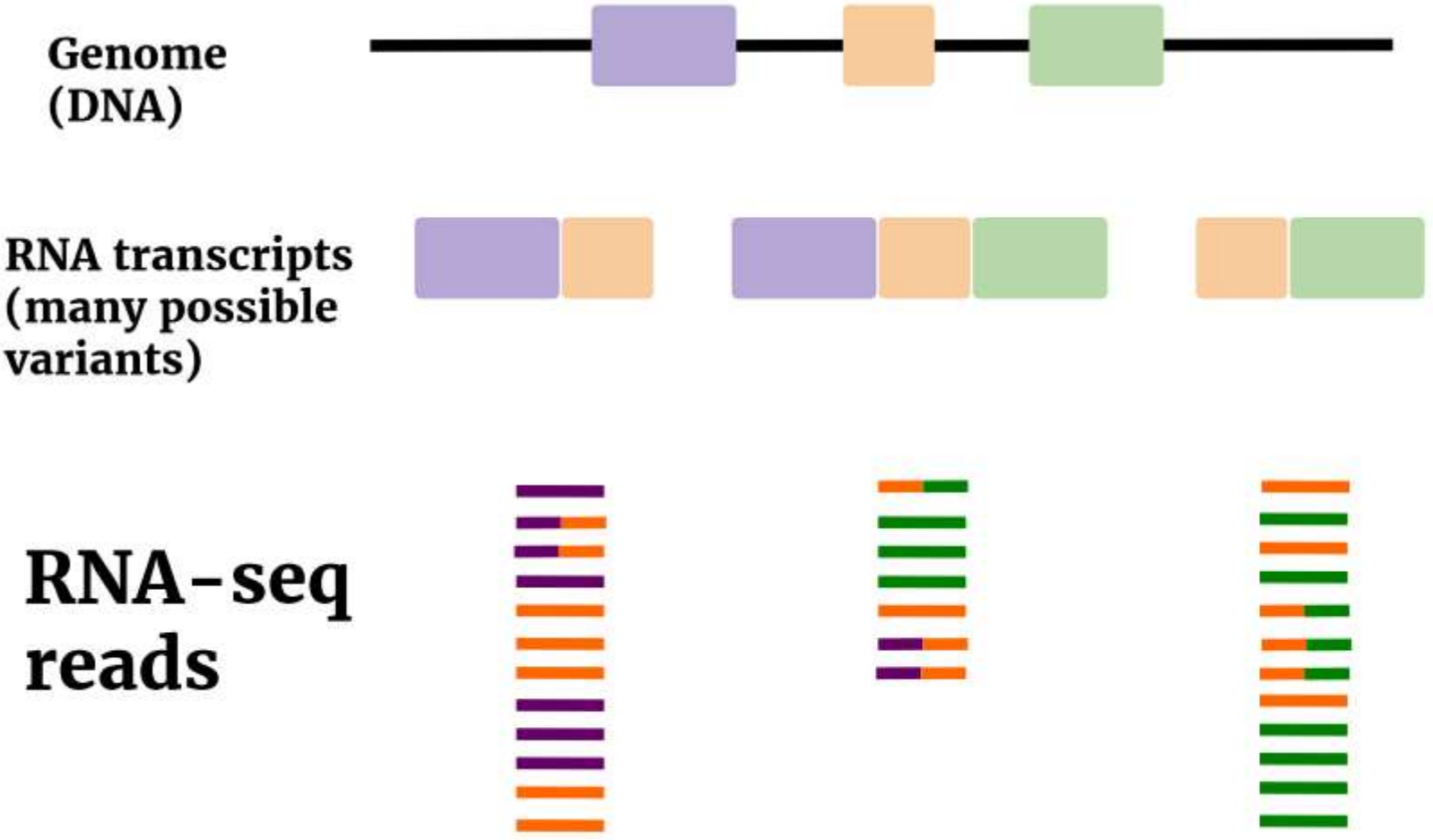
Strangspezifität

Überblick RNA-Seq



Pepke et al. 2009, Nature Methods

Überblick RNA-Seq

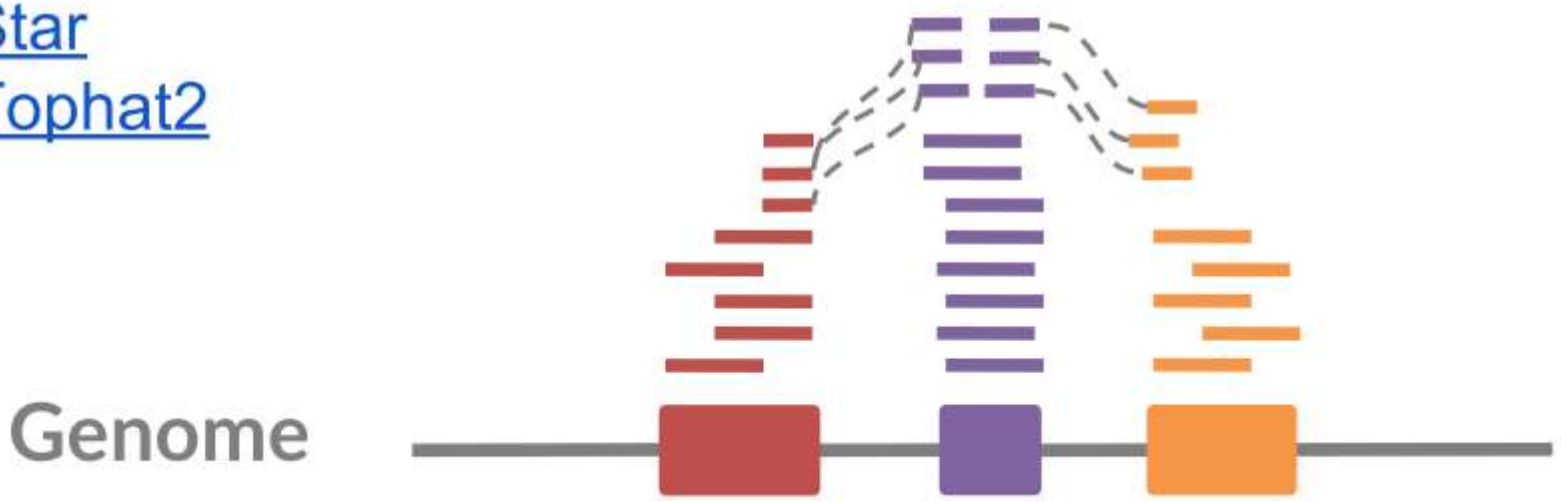


Überblick RNA-Seq

Step 1: Align

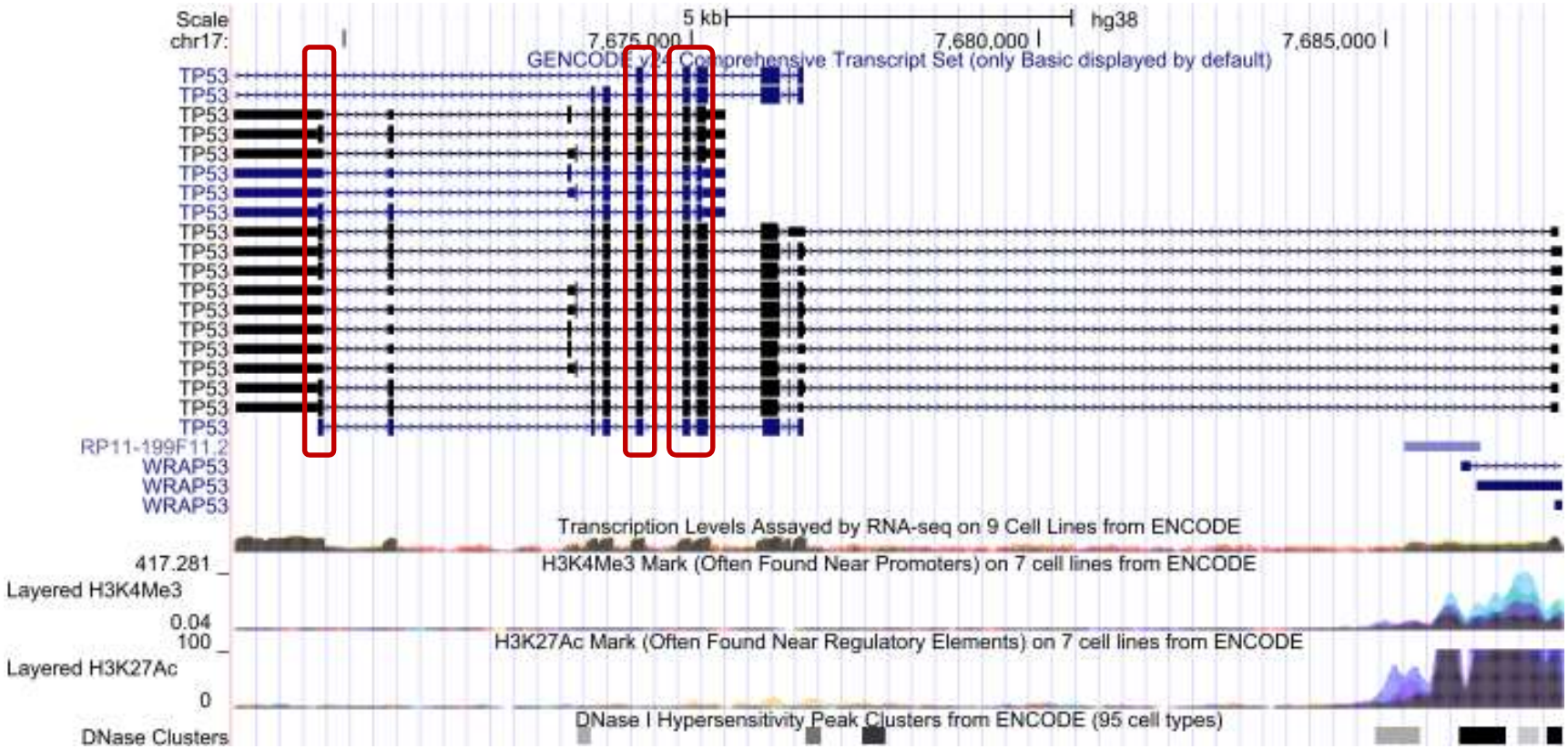
Software:

- HiSat
- Rail
- Star
- Tophat2



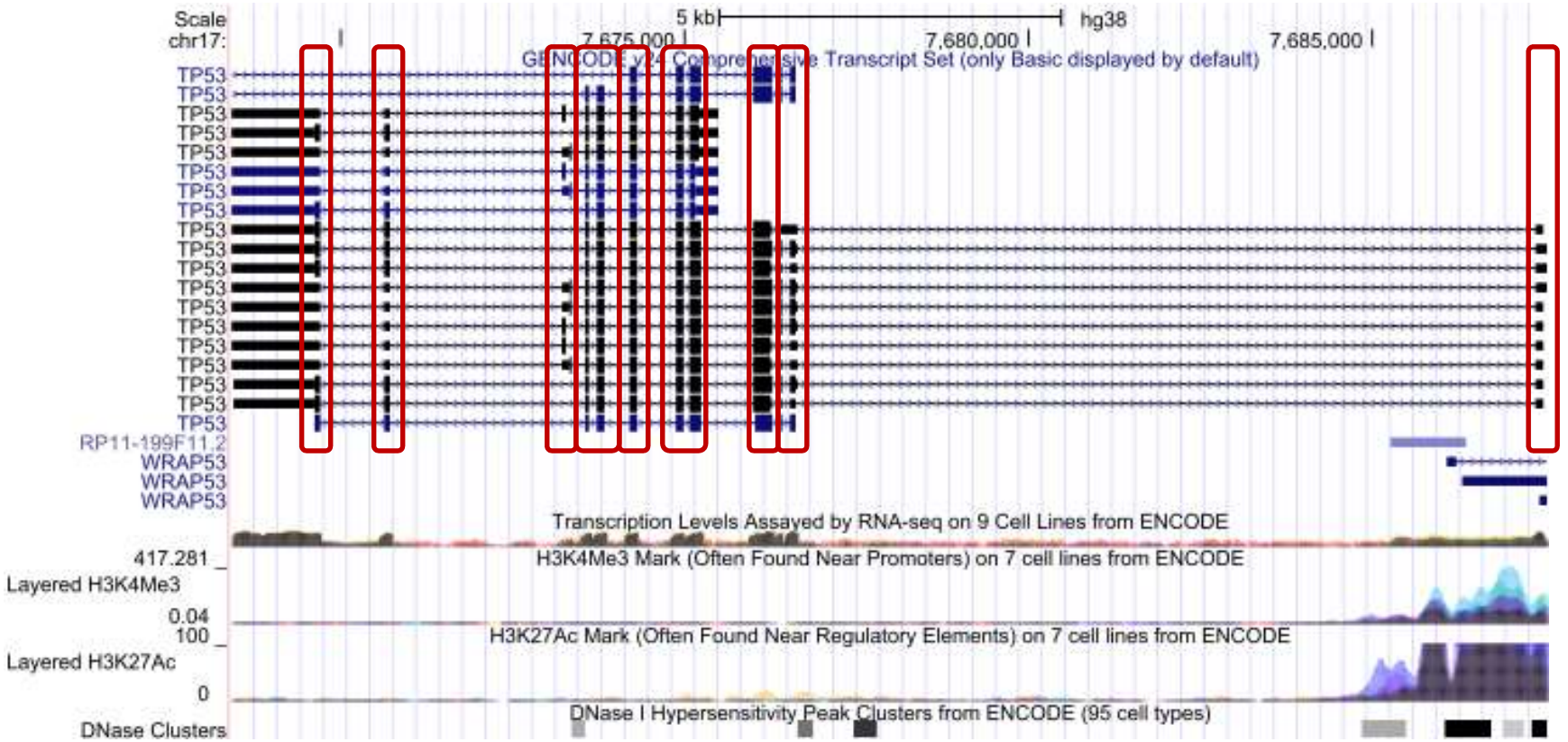
Danach muss ich die reads in jedem Transcript zählen (in den Transcript exons)
Es gibt zwei Wege die Gene-Level Quantifikation zu erreichen.

Gen Modelle



"Intersection Model" – Zähle reads, die in exons fallen, die zwischen isoforms geteilt sind

Gen Modelle



"Union Model" – Zähle nur reads, die in die union der Exons fallen

Gen Level Daten

- Gene Model + Überlappungsregel = Gen x Sample Matrix (wie Microarrays)
- Wir wollen wieder die Gen-Veränderungen zwischen den Bedingungen schätzen (Differentielle Genexpression)
- Count data
- Wie modellieren wir biologische Variabilität?

RPKM

- Wir müssen die Sequenzierungstiefe und die Gen-Länge kontrollieren

- Mortazavi et al. (2008), Nature Methods führt “RPKM” ein:

$$RPKM(g,i) = \frac{X(g,i) \cdot 10^9}{L(g)N(i)}$$

Gene count

Gene length

Sequencing depth

Poisson Verteilung für read counts

Marioni et al. zeigte das die technische Variabilität über RNA-seq Replikate durch die Nutzung eines Poisson Modells modelliert werden kann

Poisson Verteilung ist geeignet für count Daten

$$X(g,i) \sim Poisson(\lambda_g N(i))$$

Gen g

Library i

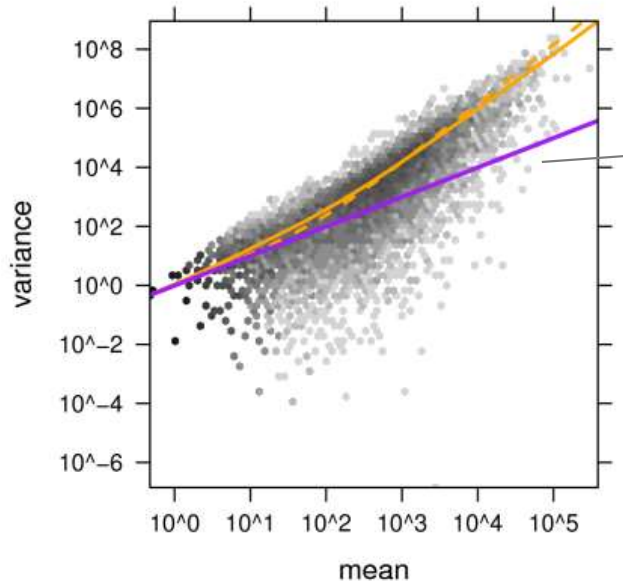
Erwartete Zahl der reads
An einer bestimmten Position
in dem Transcript

Für die Poisson-verteilten Daten ist der Mittelwert (λ) = Varianz!

Marioni et al. Genome Research 2008

Varianz hängt stark vom Mittelwert ab

Für biologische Replikanten ist der Mittelwert ≠ Varianz -> Der read count kann nicht mit der Poisson Verteilung modelliert werden



Wenn eine Poisson Verteilung eingepasst würde…

$$X(g,i) \sim NB(\theta(g), N(i))$$

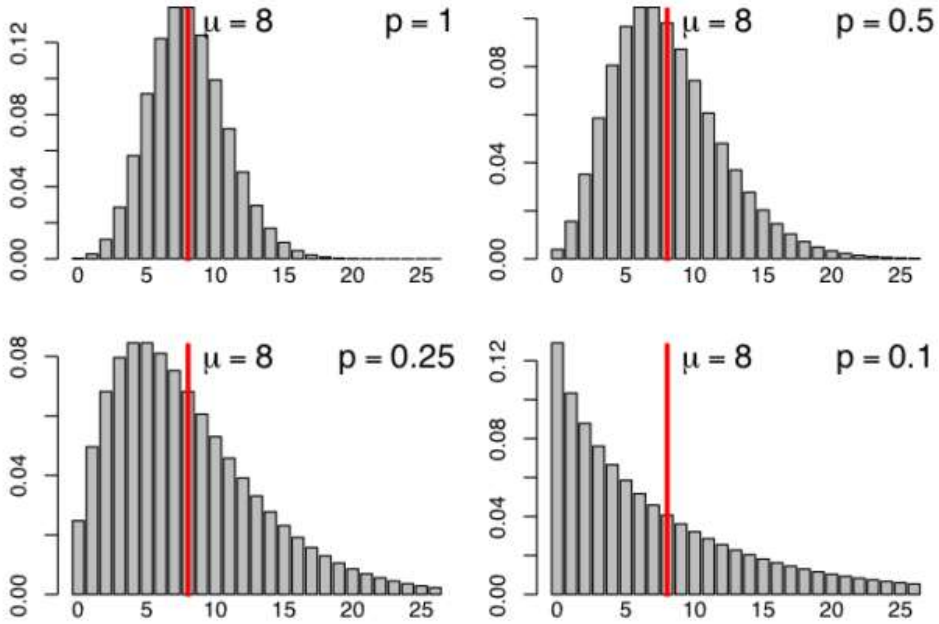
Expression von Gen g

Gesamtanzahl der reads in Sample i

Dieser Effekt heißt “Overdispersion”

Die negative binomial Verteilung

A commonly used generalization of the Poisson distribution with *two* parameters



Das DESeq package implementiert die negative binomial Verteilung für jedes Gen und schätzt die Dispersion der Daten (datengetriebene Beziehung zwischen Mittelwert und Varianz). Es stellt auch einen statistischen Test für Differentielle Genexpression bereit.