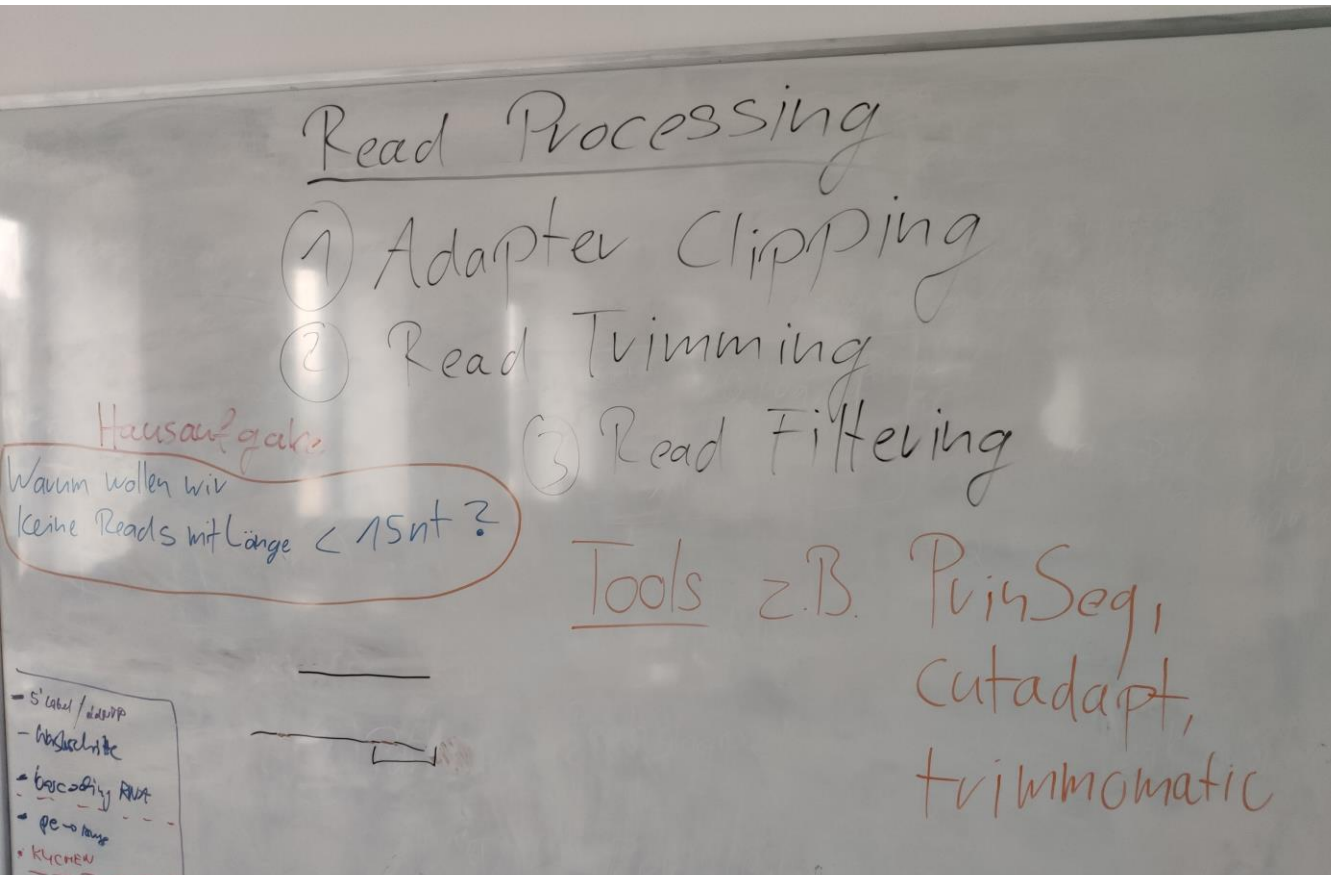


# Methoden der Hochdurchsatz- sequenzierung

– *Die klassische Genexpressionsanalyse Pipeline –*  
*Part II*

## Schritt 6 – Read Processing

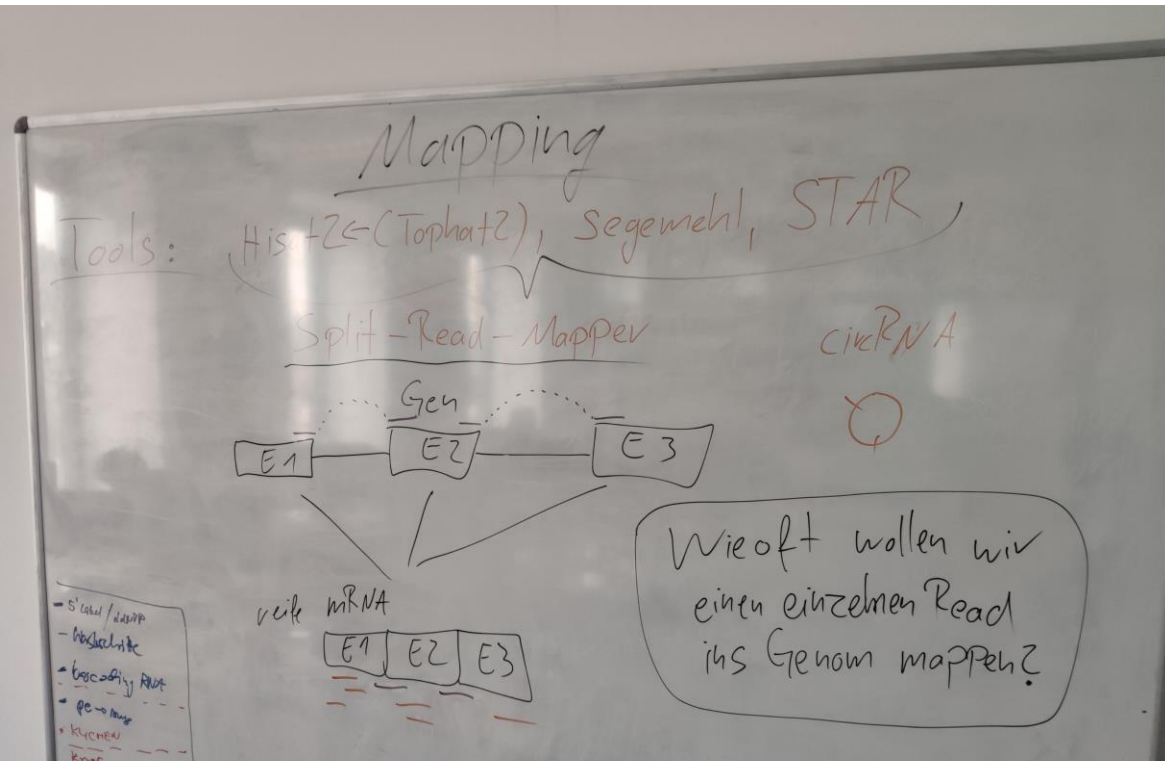
- Nachdem wir im Schritt 5 die Qualität der Read Rohdaten bestimmt bzw. analysiert haben, folgt nun die anschließende Prozessierung um einzelne Basen oder ganze Reads mit schlechten Quality Scores aus unserem Datensatz zu entfernen.  
→ Damit verhindern wir falsches Mapping von Reads im nächsten Schritt.
- Übliche Schritte:
  - Entfernen von Adaptersequenzen (und evtl. Barcode-sequenzen) aus dem Library Preparation Schritt.
  - Abschneiden (Trimmen) von einzelnen Basen am 5'- und 3'-Ende der Reads, deren Quality Score zu niedrig ist.
  - Entfernen von ganzen Reads, wenn z. B. deren durchschnittlicher Quality Score zu niedrig ist, deren Länge zu gering, ...



- Es gibt verschiedene Tools um Reads zu prozessieren z. B. PrinSeq, cutadapt, trimmomatic.
- Diese Tools unterscheiden sich hauptsächlich in der Auswahl der Trimming- und Filtering-Optionen, also der Art und Weise wie man bestimmte Quality-Grenzwerte festlegen kann, aber vom Prinzip her können alle in etwa das gleiche leisten.

# Schritt 7 – Mapping

- Allgemeines zu Mapping: siehe die vorangegangenen Vorlesungen
- Aktuelle Mapping-Tools für NGS RNA-Seq Daten: [Hisat2](#), [TopHat2](#), [segemehl](#), [STAR](#)
- Dies sind alles so genannte Split-Read-Aligner, d. h. sie splitten einen Read gegebenenfalls in zwei (oder mehr) Fragmente auf um diese dann auf das übergebene Referenzgenom zu mappen. (Frage: Wollen wir das? → Meistens schon!)
- Weitere Fragen: Wie oft wollen wir einen einzelnen Read auf das Referenzgenom mappen, wenn es mehrere **gleich gute** Mapping-Positionen gibt? → Im Idealfall an alle diese möglichen Stellen, aber das wirkt sich sehr negativ auf die Laufzeit aus.



- Standardmäßig geben die oben genannten Mapping-Tools nur die 5 bis 10 ersten gefundenen Mapping-Positionen für jeden Read wieder.
- Ein Kompromiss wäre die Anzahl an zu erwartenden paralogen Gene + Pseudogene zu berücksichtigen → Das ist allerdings oft nicht leicht zu bestimmen.

# Schritt 7 – Mapping – Das SAM-Format

- Input für ein Mapping-Tool sind die (prozessierten) Fastq Read-Files, sowie das gewünschte Referenzgenom.
- Output eines Mapping-Tools sind die gemappten (und ungemappten) Reads im SAM-Format:
  - Steht für Sequenz Alignment Map Format und ist wieder ein reines ASCII Textformat.
  - Besteht aus Headerzeilen (beginnen mit einem '@'-Symbol) und Alignmentzeilen
  - Erstere beschreiben das Referenzgenom und den Mapping-Aufruf, welcher das SAM-File erzeugt hat.
  - Zweitere geben detaillierte Informationen über das Mapping der einzelnen Reads.

## Mapping

Output von einem Mapper erhält man ein File im SAM-Format

- Seq. Alg. Map Format

- reines ASCII Textformat

- besteht aus Headerzeilen (@) und Alignmentzeilen

→ Spaltenweise Tabgetrennt und bestehen mind. aus 11 Einträgen  
+ optionale Einträge die Mapper spezifisch sind

- S-label / alignment  
- next read

# Schritt 7 – Mapping – Das SAM-Format

- Die Mappinginformationen der Reads sind in (mind. 11) Spalten organisiert.
- Die genaue SAM-Format Dokumentation gibt es [hier](#).

Col	Field	Type	Regex/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2 <sup>16</sup> -1]	bitwise FLAG
3	RNAME	String	\*  [!-( )+-<>-~] [!-~]*	Reference sequence NAME
4	POS	Int	[0,2 <sup>31</sup> -1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2 <sup>8</sup> -1]	MAPping Quality
6	CIGAR	String	\*  ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	\* =  [!-( )+-<>-~] [!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2 <sup>31</sup> -1]	Position of the mate/next read
9	TLEN	Int	[-2 <sup>31</sup> +1,2 <sup>31</sup> -1]	observed Template LENgth
10	SEQ	String	\*  [A-Za-z=.]+	segment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Bit	Description
1	0x1 template having multiple segments in sequencing
2	0x2 each segment properly aligned according to the aligner
4	0x4 segment unmapped
8	0x8 next segment in the template unmapped
16	0x10 SEQ being reverse complemented
32	0x20 SEQ of the next segment in the template being reverse complemented
64	0x40 the first segment in the template
128	0x80 the last segment in the template
256	0x100 secondary alignment
512	0x200 not passing filters, such as platform/vendor quality controls
1024	0x400 PCR or optical duplicate
2048	0x800 supplementary alignment

<http://broadinstitute.github.io/picard/explain-flags.html>

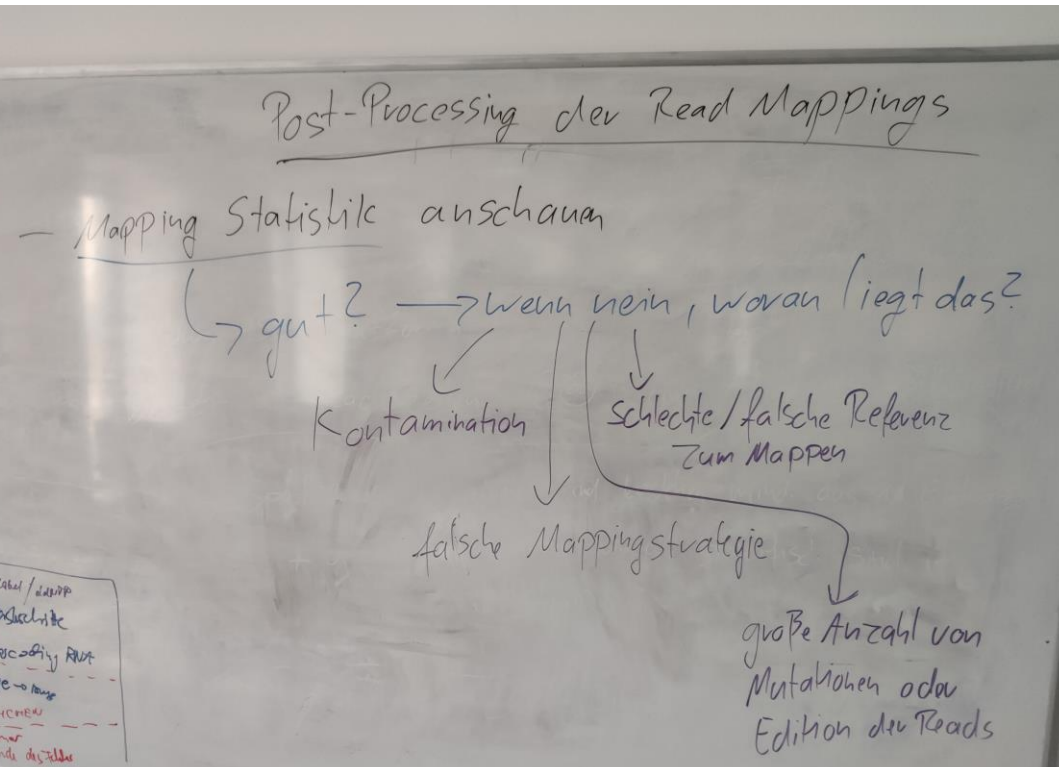


# Schritt 8 – Post-Processing der Read Mappings

- Die meisten Mapping-Tools generieren aus einem SAM-File auch direkt ein paar Mapping-Statistiken, mit denen man die Güte des Mappings einschätzen kann:

```
17285886 reads; of these:  
 17285886 (100.00%) were unpaired; of these:  
   904284 (5.23%) aligned 0 times  
  12098691 (69.99%) aligned exactly 1 time  
   4282911 (24.78%) aligned >1 times  
94.77% overall alignment rate
```

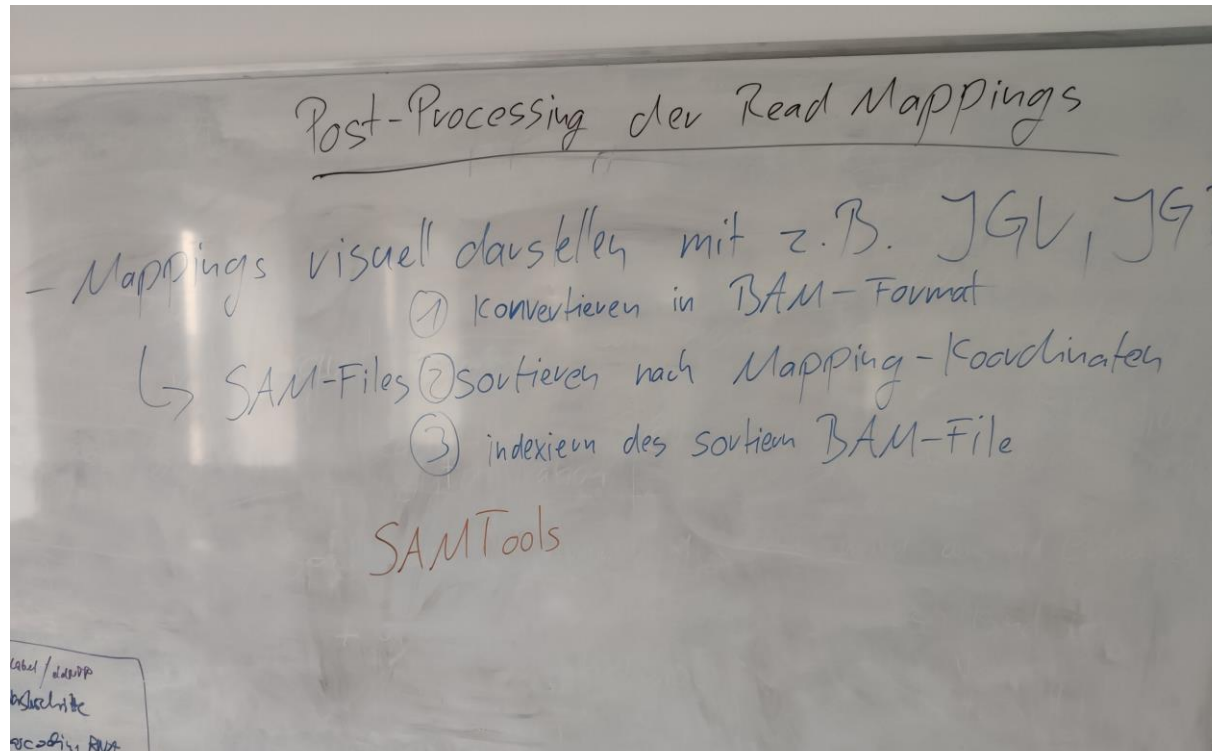
- Sollten die Mapping-Statistiken nicht den Erwartungen entsprechen (z. B. deutlich weniger gemappte Reads als erwartet), muss man sich fragen woran das liegt:



- Gab es eine Kontamination in der sequenzierten Probe?
- Wurde ein falsches oder schlechtes Referenzgenom/-transkriptom gewählt?
- Hat man sich für eine ungeeignete Mapping-Strategie entschieden (falsche Parameterwahl)?
- Gibt es einen möglichen biologischen Grund, wie z. B. erhöhte Mutationsrate oder Editierungsrate innerhalb der sequenzierten Transkripte im Vergleich zur Referenz?

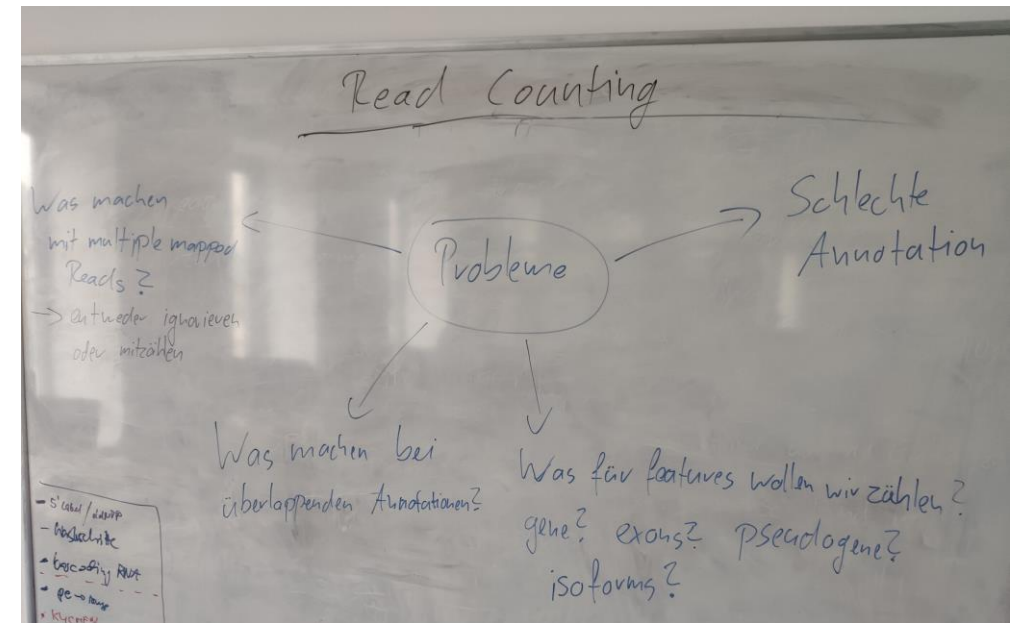
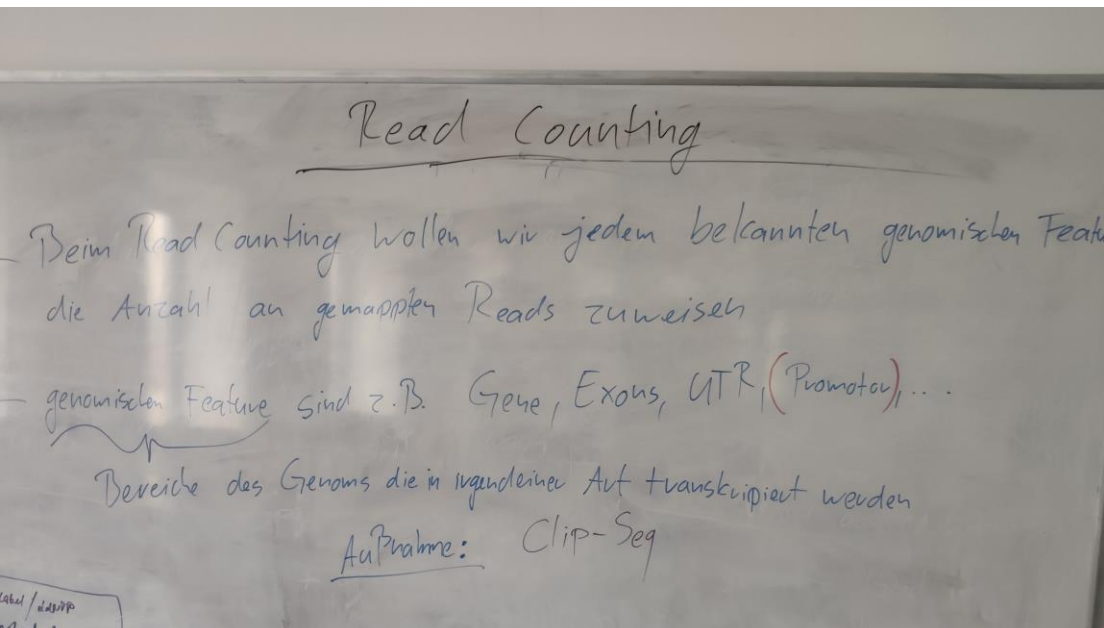
# Schritt 8 – Post-Processing der Read Mappings

- Mappings lassen sich auch visuell darstellen mit z. B. [IGV](#), [IGB](#), [Tablet](#).
- Dafür müssen i. d. R. die SAM-Files vorher konvertiert werden:
  1. Konvertieren in das binäre, aber inhaltlich identische, BAM-Format.
  2. Sortieren der gemappten Reads nicht anhand der Read-Namen, sondern der Mapping-Koordinaten.
  3. Indexieren des sortierten BAM-Files.
- Üblicherweise kann man für diese Post-Processing Aufgaben (und viele weitere) das Programm [SamTools](#) nutzen.



# Schritt 9 – Read Counting


- Ziel des Read Countings ist es jedem (bekannten) genomischem Feature die Anzahl an gemappten Reads zuzuweisen.
- Dabei sind genomische Features i.d.R. definiert als alle genomischen Regionen die in irgendeiner Art und Weise exprimiert werden. (Aber es gibt auch Ausnahmen, wie z. B. Promotorregionen u. Ä. → Kann man mit Clip-Seq Techniken analysieren)
- Wichtige Fragen:
  - Ist die zugrunde liegende Annotation vollständig (genug)?
  - Was für Features wollen wir zählen? (Gen? Exon? UTR? Pseudogen? ncRNAs? Isoformen?)
  - Wie gehen wir mit mehrfach gemappten Reads um? (Ignorieren? Mehrfach zählen? Nur anteilig zählen?)
  - Wie zählt man Reads die auf überlappende Features mappen?
  - Muss man die Read-Orientierung (Forward vs. Reverse) beachten?



## Schritt 9 – Read Counting – Ein Beispiel Count-File

- Ziel des Read Countings ist es jedem (bekannten) genomischem Feature die Anzahl an gemappten Reads zuzuweisen.

Feature ID                      Count-Wert



ENSG00000194717_3P	104
ENSG00000194717_5P	3
ENSG00000198972_3P	186
ENSG00000198972_5P	17058
ENSG00000198973_3P	296
ENSG00000198973_5P	0
ENSG00000198974_3P	5151
ENSG00000198974_5P	63783
ENSG00000198975_3P	1
ENSG00000198975_5P	245739
ENSG00000198976_3P	6
ENSG00000198976_5P	0
ENSG00000198982_3P	103
ENSG00000198982_5P	2
ENSG00000198983_3P	0
ENSG00000198983_5P	22
ENSG00000198984_3P	2
ENSG00000198984_5P	692
ENSG00000198987_3P	967
ENSG00000198987_5P	85015
ENSG00000198995_3P	121
ENSG00000198995_5P	7561
ENSG00000198997_3P	8503

# Schritt 10 – Differentielle Expressionsanalyse

- Im Prinzip wollen wir den *Foldchange* also das Verhältnis der Expressionsänderung eines Gens in unterschiedlichen Bedingungen (z. B. krank/gesund, mutiert/wildtyp, etc.) ermitteln. → Wir vergleichen die Countwerte der Gene zwischen den untersuchten Bedingungen.
- Genauer: Wir untersuchen die aus den biologischen Replikaten stammenden gemittelten Countwerte der Gene zwischen den unterschiedlichen Bedingungen, da der gemittelte Countwert (hoffentlich) näher am wahren Expressionswert eines Gens liegt, als die tatsächlich gemessenen Countwerte.
- Zusätzlich wollen wir bestimmen ob die Änderung in der Expressionsstärke (also in den gemittelten Countwerten) statistisch signifikant ist, oder durch biologische Variation (Rauschen) erklärt werden kann.

- Vereinfachtes Beispiel:

Sample 1	Sind die Unterschiede zwischen den Countwerten der Gene schon signifikant?	Sample 2
Gen A 3		Gen A 9
Gen B 7		Gen B 20
Gen C 15	←————→	Gen C 42
Gen D 900		Gen D 2518
Gen E 1700		Gen E 5000
<hr/> Σ 2625	←————→ Unterschiedliche Sequenziertiefe!	<hr/> Σ 7589

# Schritt 10 – Differentielle Expressionsanalyse

- Um die Countwerte zweier Samples überhaupt miteinander vergleichen zu können, müssen diese vorher erst normalisiert werden.
- Ein einfaches normalisieren anhand der RNA-Seq Library-Größen (also der Anzahl der sequenzierten Reads der entsprechenden Fastq-Files) ist nicht möglich, da dies einen Bias für sehr lange Transkripte einführen würde (siehe cDNA Amplifizierungsschritt in der Library Preparation).
- D. h. wir müssen nicht nur nach Library-Größe, sondern auch nach Transkriptlänge normalisieren.
- Probleme:
  - Was ist die richtige Library-Größe? (Anzahl sequenzierter Reads? Anzahl gemappter Reads? Anzahl gecounteter Reads?)
  - Was ist die korrekte Transkriptlänge? (Genlänge? Aufsummierte Exonlänge? Längste/kürzeste/mittlere Isoformlänge?)

→ Die Auswahl der entsprechenden Parameter hat sehr großen Einfluss auf die Normalisierung.
- Mögliche Normalisierungen: RPKM, FPKM, CPM, TPM.

# Schritt 10 – Differentielle Expressionsanalyse

- Als ein Beispiel schauen wir uns die TPM (Transcripts Per Million) Normalisierung an:
  1. Für jedes Gen  $i$ , teile seinen Countwert durch dessen Genlänge  $L_i$  und durch 1000  
→ Wir erhalten den RPK-Wert (Reads per kilobase) für jedes Gen.
  2. Summiere alle RPKs auf und teile diese Summe durch  $10^6$   
→ Das ist unser Skalierungsfaktor  $T$  für unser Sample.
  3. Für jedes Gen  $i$ , teile dessen RPK-Wert durch den Skalierungsfaktor  $T$   
→ Wir erhalten den TPM-Wert.

$$TPM_i = \left( \frac{X_i}{L_i} \right) * 10^3 * \left( \frac{1}{\sum_{\forall Gene\ j} \frac{X_j}{L_j}} \right) * 10^6,$$

wobei  $X_i$  der Countwert von Gen  $i$  ist.

- Aber: TPM (genauso wie RPKM, FPKM, CPM) ist nur ein **relatives Anteilsmaß**, d. h. es lassen sich keine absoluten normalisierten Countwerte berechnen. Es eignet sich daher eher für Sample-interne Vergleiche und zur Bestimmung von Genaktivitätsgrenzwerten.

# Schritt 10 – Differentielle Expressionsanalyse

- Zur Berechnung differentiell exprimierter Gene gibt es verschiedene Tools wie z. B. [DESeq2](#), [edgeR](#), [limma](#).
- Sie alle versuchen die Expressionswerte eines Gens durch **diskrete Wahrscheinlichkeitsverteilungen** zu modellieren (genauer: durch Poisson- oder Negativ Binomialverteilungen).
- Diese Verteilungen werden durch die Countwerte der Gene geschätzt → Je mehr Replikate, desto besser sind diese Verteilungsschätzungen.
- Am Ende einer differentiellen Expressionsanalyse erhalten wir für jedes Gen seinen Foldchange zwischen den untersuchten Bedingungen und einen p-Wert für die Signifikanz dieses Foldchanges (und noch ein paar weitere statistische Werte).
- Der p-Wert gibt, grob gesagt, die Wahrscheinlichkeit an, dass ein gemessener Expressionsunterschied von einem Gen nur durch Zufall entstanden ist. (Eigentlich stimmt das so nicht genau, aber eine genauere Erklärung würde an dieser Stelle zu weit führen)
  - Für uns wichtig: Je **kleiner** der p-Wert, desto **wahrscheinlicher** ist die Expressionsänderung eines Gens auf eine biologische Ursache zurückzuführen.

# Schritt 10 – Differentielle Expressionsanalyse – Beispiel Ergebnis von DESeq2

```
"GeneID", "baseMean", "log2FoldChange", "lfcSE", "stat", "pvalue", "padj"  
"ENSMUSG00000015451", 1761.64217193547, 1.33086755088502, 0.163174102766743, 8.15611992539954, 3.45958943542573e-16, 9.50764368643699e-12  
"ENSMUSG00000073418", 3594.46422903302, 1.32298535779046, 0.177513142979177, 7.45288678678655, 9.13196440572806e-14, 1.25482322899109e-09  
"ENSMUSG00000030789", 28.2877011989067, 1.54210513936855, 0.231330271144389, 6.6662487868093, 2.62424269172036e-11, 2.40398125512863e-07  
"ENSMUSG00000041596", 87.4146693089831, 1.21431693856719, 0.189832180333327, 6.39679182125478, 1.58675425526946e-10, 1.09017951108289e-06  
"ENSMUSG00000093574", 2419.08615728259, 0.648509483062987, 0.111485904860376, 5.81696389220839, 5.99260799842489e-09, 3.29377706025426e-05  
"ENSMUSG0000004707", 35.3165528949023, 1.31893198816917, 0.23496577696612, 5.61329400902219, 1.98510883561927e-08, 8.80307576110921e-05  
"ENSMUSG00000037708", 167.604439033623, 0.5509539853438935, 0.0985220679123839, 5.5921886041704, 2.24225057593205e-08, 8.80307576110921e-05  
"ENSMUSG000000100586", 28.3335833512869, 1.20528116165182, 0.222053195117405, 5.42789380271945, 5.70229499661192e-08, 0.000195888088871111  
"ENSMUSG00000026822", 50.3958524174659, 2.35438837952453, 0.439942013560285, 5.35158795240161, 8.71857440798708e-08, 0.000266226513200334  
"ENSMUSG00000023452", 4470.69020097332, 0.691714021834066, 0.130987695990486, 5.28075569696491, 1.28652140914396e-07, 0.000353561813660943  
"ENSMUSG00000023795", 2213.05774737149, 0.948774014402168, 0.189944816548647, 4.99499818758769, 5.8836319959848e-07, 0.0014699452228514  
"ENSMUSG00000082286", 4631.95764785386, 1.05493567634811, 0.213506418736887, 4.94100216091466, 7.77220469432399e-07, 0.0017799644117451  
"ENSMUSG000000109006", 1562.30697808852, 0.305013630314216, 0.0626627808828839, 4.86754060411528, 1.12995620812712e-06, 0.00238872742398073  
"ENSMUSG00000068129", 11.3549182443493, 1.82911237367788, 0.384487877563493, 4.75726929355745, 1.96229319613872e-06, 0.00380503776493873  
"ENSMUSG00000030087", 927.728801555626, 0.366573055717492, 0.077247729072128, 4.74542177641512, 2.08072312988896e-06, 0.00380503776493873  
"ENSMUSG0000002831", 532.888795730417, 1.33905546171239, 0.282935595405675, 4.73272180473595, 2.21529016225237e-06, 0.00380503776493873  
"ENSMUSG000000105647", 473.913950345178, 0.856996064796981, 0.184263590190722, 4.65092460159897, 3.30450135892671e-06, 0.00504523924144578  
"ENSMUSG00000046805", 1651.29062594595, 0.5067122739285, 0.108774987013179, 4.65835287911465, 3.18749482019057e-06, 0.00504523924144578  
"ENSMUSG00000010461", 213.692008155519, 0.476428887950654, 0.103035572181746, 4.62392626024608, 3.76543772924972e-06, 0.00517408798376204  
"ENSMUSG00000020932", 10135.3508055692, 0.567951124418037, 0.122794344840299, 4.62522215625392, 3.74197109291657e-06, 0.00517408798376204  
"ENSMUSG00000034957", 369.10248486197, 0.366414345350399, 0.079827861938868, 4.59005585832924, 4.43127407478794e-06, 0.00579906067253915  
"ENSMUSG00000030577", 28.452249904693, 0.959548266765001, 0.214319264426292, 4.47719092977293, 7.56316052454217e-06, 0.00903916515658742  
"ENSMUSG00000051439", 98.9330669925931, 0.655680333896406, 0.146450725545825, 4.47713953927281, 7.56498066376212e-06, 0.00903916515658742  
"ENSMUSG00000084898", 110.33009606118, -0.743639675193526, 0.167416366458679, -4.44185769243216, 8.91855387188604e-06, 0.0102124873961322  
"ENSMUSG00000020805", 522.774551078191, 0.308368664583893, 0.0695949267135556, 4.43090723915986, 9.38374516359292e-06, 0.0103153633834344  
"ENSMUSG00000033383", 34.921528749485, 1.11995614148214, 0.254355937908422, 4.40310594158556, 1.06712022520825e-05, 0.0112794607804512  
"ENSMUSG00000039457", 119.845405883906, 0.697079571779013, 0.158752034393279, 4.39099614970683, 1.12832538877497e-05, 0.0114846808645606  
"ENSMUSG00000056501", 164.700153599586, 0.478267361675202, 0.10920895877105, 4.37937845994725, 1.19018278635452e-05, 0.0116816440480697  
"ENSMUSG00000038642", 2880.2588724084, 0.402471420216725, 0.0924089285247843, 4.35533044957643, 1.32866326495977e-05, 0.0125911461543533  
"ENSMUSG00000030256", 1836.68203857127, 0.347648150662383, 0.0801427256004281, 4.33786283231333, 1.43874896130778e-05, 0.0131798996515535  
"ENSMUSG00000034855", 20.0157780182822, 1.44637781984155, 0.334763219552592, 4.3205956220585, 1.55605827622562e-05, 0.0137947075958814  
"ENSMUSG00000073633", 88.9777805005758, -0.55619146183011, 0.129642834813111, -4.29018281366657, 1.78526097420577e-05, 0.0153320444041009  
"ENSMUSG00000048752", 46.7899574320399, 0.749865355743415, 0.17612643470608, 4.25754008473964, 2.0668860902341e-05, 0.0172127768278223  
"ENSMUSG00000030124", 216.295495581968, 0.744246057013352, 0.176743119741108, 4.21089125338241, 2.54365176265982e-05, 0.020560187571005  
"ENSMUSG00000051486", 295.126716768672, 0.667073453924681, 0.159938752010023, 4.1708056711789, 3.03524618201952e-05, 0.0238327530212173  
"ENSMUSG00000024197", 951.664755725161, 0.317359676510197, 0.077640434889545, 4.0875566573228, 4.35940231601044e-05, 0.0332791929023886  
"ENSMUSG00000026821", 4053.20257505357, 0.246772877692466, 0.0607440446684563, 4.06250323038848, 4.85492756437685e-05, 0.0357104299521225  
"ENSMUSG00000070527", 31.2448337621453, -0.850662554766151, 0.209597477023808, -4.05855340839586, 4.93776412990559e-05, 0.0357104299521225  
"ENSMUSG00000051678", 235.237311811284, 0.778994114660096, 0.194944170404103, 3.99598568680102, 6.44256265290433e-05, 0.0453985914941325  
"ENSMUSG00000051085", 428.279733290893, 0.457675055855458, 0.114800948886179, 3.98668356225197, 6.7003293034799e-05, 0.0460346124795587  
"ENSMUSG00000041828", 867.724124998753, 0.58523578944522, 0.147085258831175, 3.97888812309162, 6.92383101009283e-05, 0.046409932638871  
"ENSMUSG00000034906", 40.5264370189149, -0.716063076792112, 0.182626545559704, -3.92091453407039, 8.82135475907843e-05, 0.057721064602365  
"ENSMUSG00000055027", 130.312918583319, -0.595770735195821, 0.152563565652965, -3.90506562065425, 9.41998203669213e-05, 0.0602046386819472  
"ENSMUSG00000023943", 4.57317069494152, 2.41194731404386, 0.618576960474481, 3.89918711520354, 9.65161685106428e-05, 0.0602831214320338  
"ENSMUSG00000040552", 122.991743663786, 0.48184312664299, 0.124398783650984, 3.87337490368763, 0.000107338514358667, 0.0655528233689976  
"ENSMUSG00000027956", 803.01553285771, 0.329368987985159, 0.0854178102041425, 3.85897555355248, 0.000115269007725301, 0.0688657145718854
```