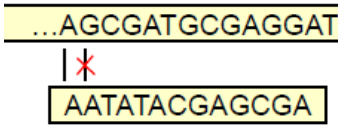
WS2017/2018 MSc Modul 7A

„Genomforschung und Sequenzanalyse

- Einführung in Methoden der Bioinformatik- “

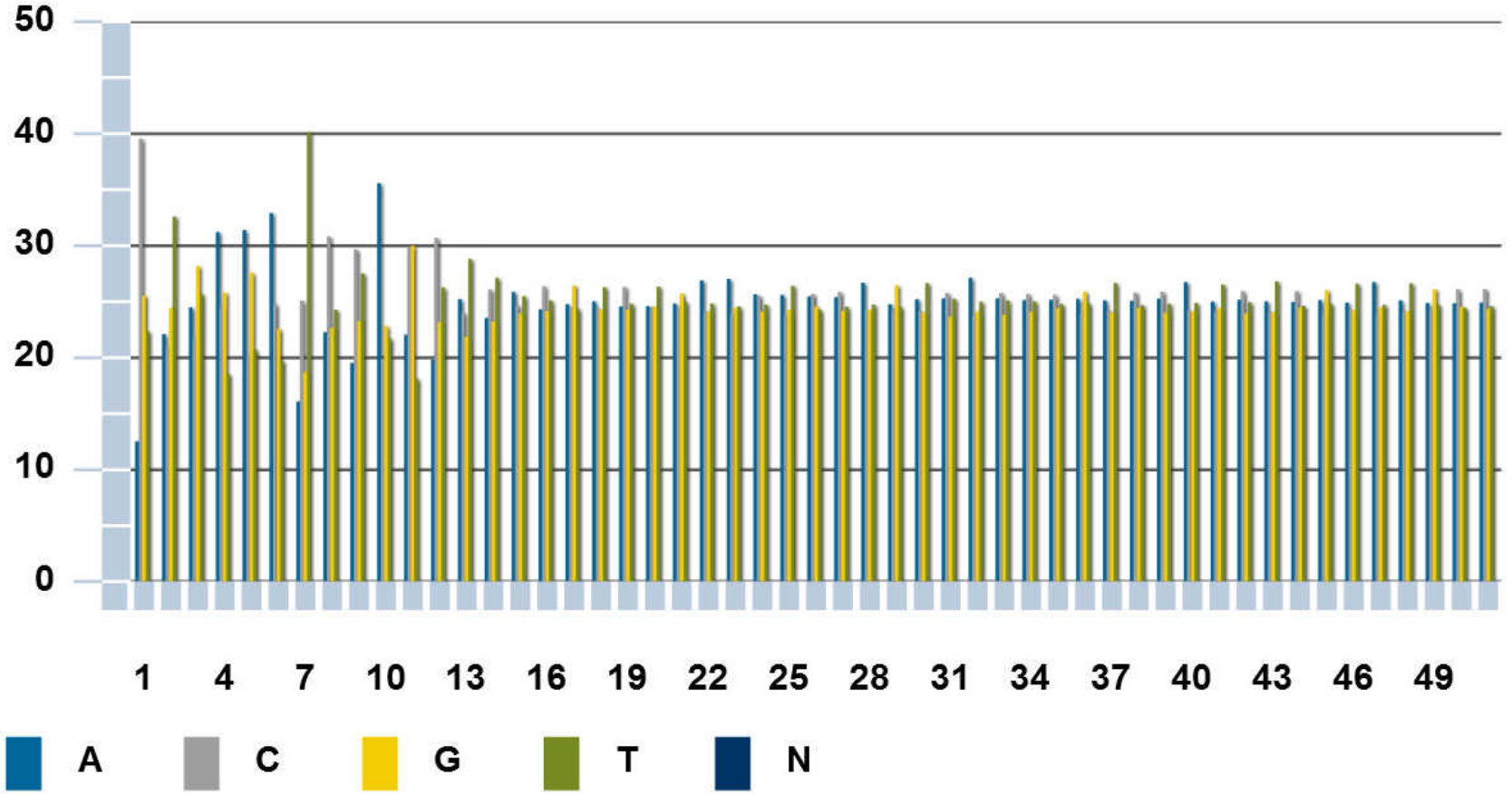
AG Hankeln

Methoden der Genomsequenzierung:

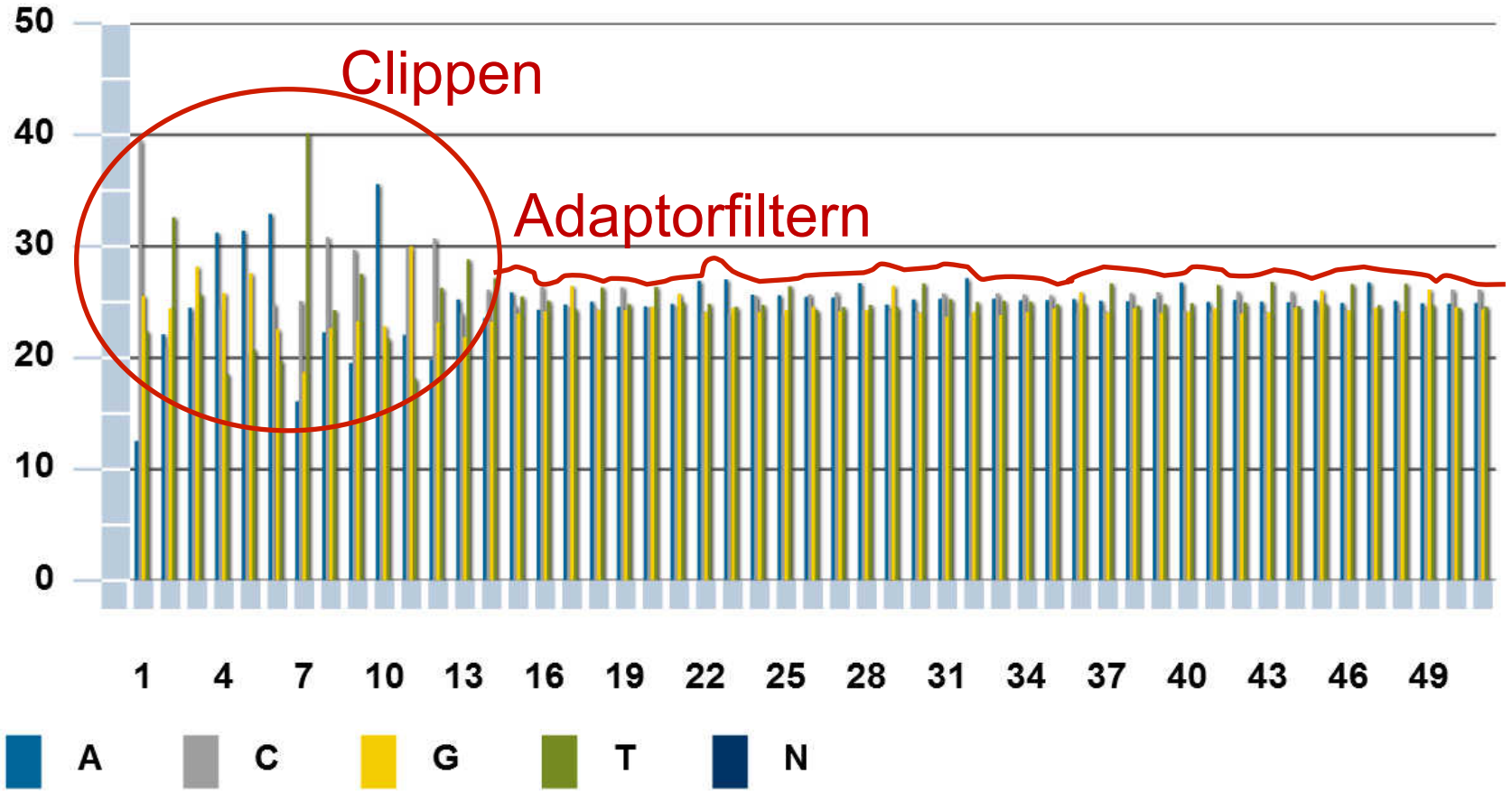


Mapping-Strategien

Sind alle sequenzierten Reads qualitativ hochwertig? Was fällt noch auf?



Qualitätsprozessierung: Nukleotidverteilung in Rohdaten



FASTQ-Format

Format zum Speichern der Sequenzschnipsel

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

33; 39; 39; 42; 40 usw.

Header

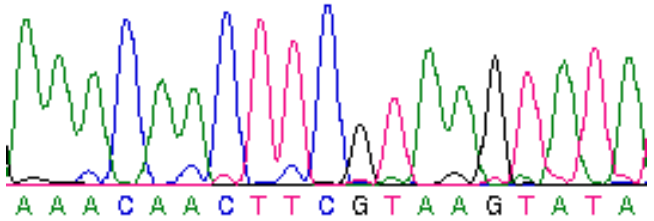
Sequenz

Header2

Phred-kodierte Qualitätswerte

Vor dem Mapping: Aussortieren

Qualitätswerte



Sanger-Sequenzierung
- Chromatogramme
- Phred-scores

Phred-scores:

Q = -10 Log10 p
p = 10^(-Q/10)
p: Wahrscheinlichkeit, dass der Basecall falsch ist

NGS
- Phred ähnliche Qualitätswerte
- Werden im FASTQ und SAM-format umformatiert um weniger Speicherplatz zu belegen
- Q wird mit X addiert, der entsprechende Eintrag der Ascii-Tabelle steht für die Qualität
- X = 33 bei Sanger, bei Illumina lange Zeit 64, jetzt auch 33

ASCII-Codetabelle

+	0	1	2	3	4	5	6	7	8	9
30				!	"	#	$	%	&	'
40	(	)	*	+	,	-	.	/	0	1
50	2	3	4	5	6	7	8	9	:	;
60	<	=	>	?	@	A	B	C	D	E
70	F	G	H	I	J	K	L	M	N	O
80	P	Q	R	S	T	U	V	W	X	Y
90	Z	[	\	]	^	_	`	a	b	c
100	d	e	f	g	h	i	j	k	l	m
110	n	o	p	q	r	s	t	u	v	w
120	x	y	z	{	|	}	~			

Sanger Beispiel:
Ascii: 73 → minus X (33) → Q = 40 → p = 10^(-40/10) = 0,0001

Berechne den Quality score zu p=0,05

$$a = b^{x}$$
$$x = \log_{b} a$$

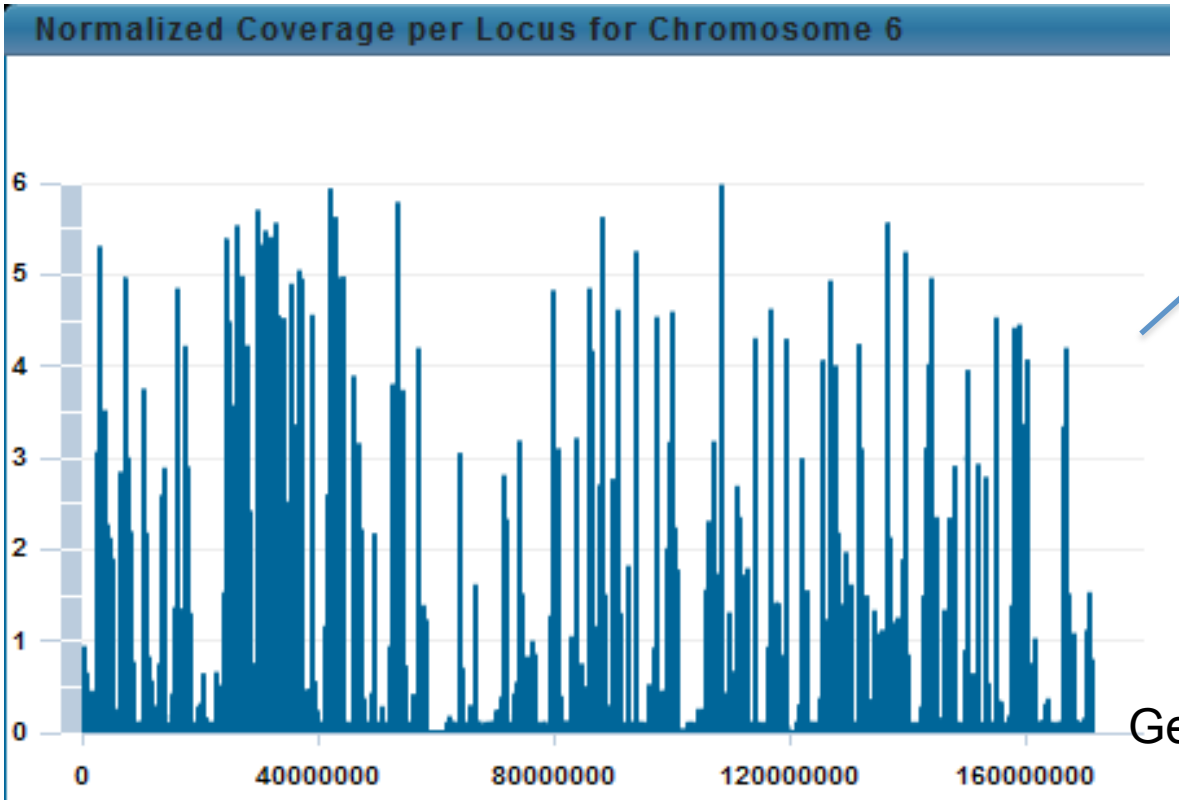
Berechne den Quality score zu p=0,05

- $p = 10^{Q/-10}$
- $Q/-10 = \log_{10} p$
- $Q = -10 \log_{10} p$
- $-10 \log_{10} 0{,}05 = 13$
- → Cutoff bei einem Quality-Score von 13

$a = b^x$
$x = \log_b a$

Mapping

- Alignierung der generierten Reads an eine Referenz

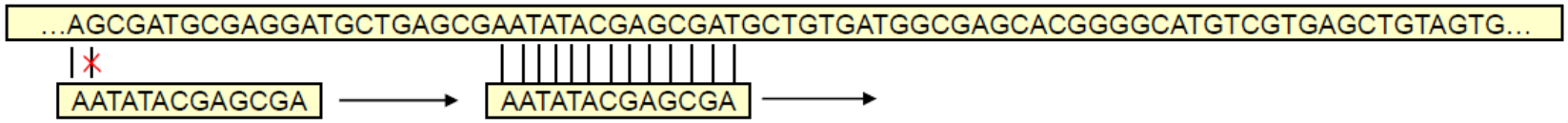


Quantitative Information!

Mapping - aber wie?

Challenge

Map billions of short reads back to a reference sequence



Scanning every position of the human genome for every sequence read will take years!

Mapping approaches must involve smart pre-processing:

Hashing (reference sequence or sequence reads)

Burrows-Wheeler transformation (reference sequence)

Mapping-Algorithmus I: Burrows-Wheeler-Transformation

- Umstrukturierung der Daten
- Verringerter Speicherbedarf
- Verkürzte Zugriffszeiten
- Schnelleres Alignieren

Mapping-Algorithmus I: Burrows-Wheeler-Transformation

- Umstrukturierung der Daten
- Verringerter Speicherbedarf
- Verkürzte Zugriffszeiten
- Schnelleres Alignieren

Beispiel: Mississippi

mississippi$

Mapping-Algorithmus I: Burrows-Wheeler-Transformation

Alle möglichen Rotationen erzeugen.

m	i	s	s	i	s	s	i	p	p	i	$
i	s	s	i	s	s	i	p	p	i	$	m
s	s	i	s	s	i	p	p	i	$	m	i
s	i	s	s	i	p	p	i	$	m	i	s
i	s	s	i	p	p	i	$	m	i	s	s
s	s	i	p	p	i	$	m	i	s	s	i
s	i	p	p	i	$	m	i	s	s	i	s
i	p	p	i	$	m	i	s	s	i	s	s
p	p	i	$	m	i	s	s	i	s	s	i
p	i	$	m	i	s	s	i	s	s	i	p
i	$	m	i	s	s	i	s	s	i	p	p
$	m	i	s	s	i	s	s	i	p	p	i

Mapping-Algorithmus I: Burrows-Wheeler-Transformation

Sortierung der Zeilen in alphabetischer Reihenfolge.

$	m	i	s	s	i	s	s	i	p	p	i
i	$	m	i	s	s	i	s	s	i	p	p
i	p	p	i	$	m	i	s	s	i	s	s
i	s	s	i	p	p	i	$	m	i	s	s
i	s	s	i	s	s	i	p	p	i	$	m
m	i	s	s	i	s	s	i	p	p	i	$
p	i	$	m	i	s	s	i	s	s	i	p
p	p	i	$	m	i	s	s	i	s	s	i
s	i	p	p	i	$	m	i	s	s	i	s
s	i	s	s	i	p	p	i	$	m	i	s
s	s	i	p	p	i	$	m	i	s	s	i
s	s	i	s	s	i	p	p	i	$	m	i

Mapping-Algorithmus I: Burrows-Wheeler-Transformation

Ausgabe der letzten Spalte.

mississippi$

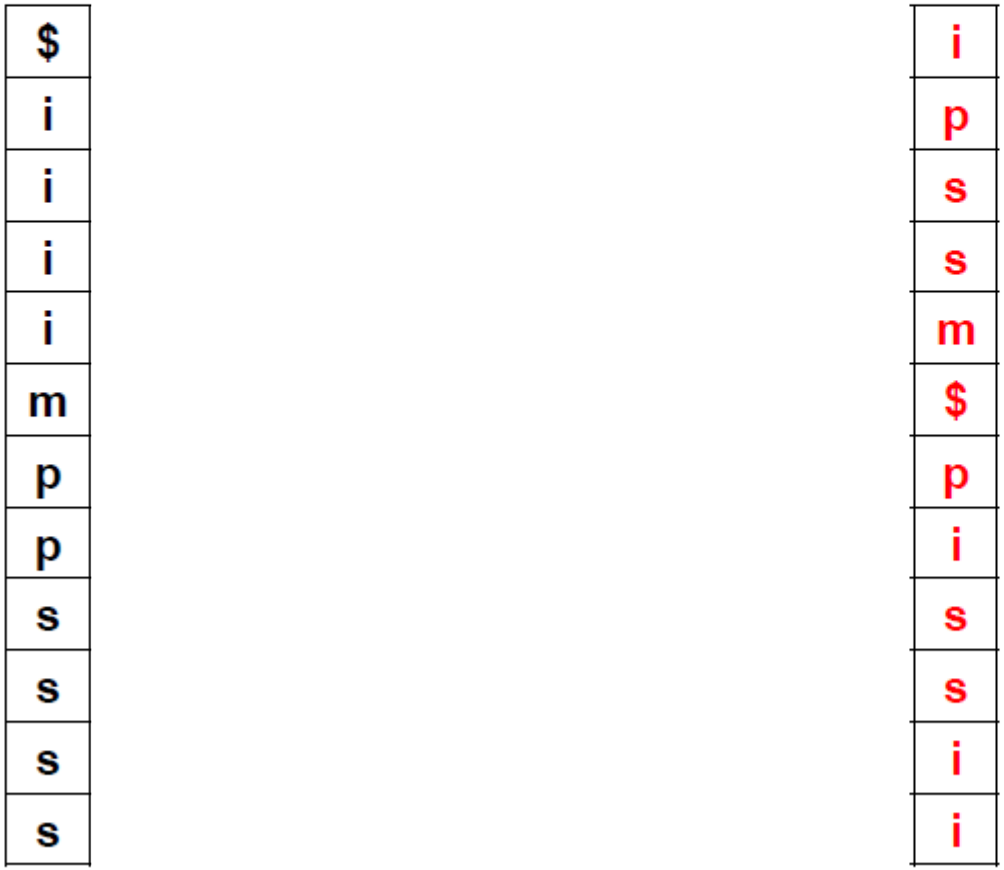
ipssm$pissii

(4 i, 1 m, 2 p, 4 s)

$	m	i	s	s	i	s	s	i	p	p	i
i	$	m	i	s	s	i	s	s	i	p	p
i	p	p	i	$	m	i	s	s	i	s	s
i	s	s	i	p	p	i	$	m	i	s	s
i	s	s	i	s	s	i	p	p	i	$	m
m	i	s	s	i	s	s	i	p	p	i	$
p	i	$	m	i	s	s	i	s	s	i	p
p	p	i	$	m	i	s	s	i	s	s	i
s	i	p	p	i	$	m	i	s	s	i	s
s	i	s	s	i	p	p	i	$	m	i	s
s	s	i	p	p	i	$	m	i	s	s	i
s	s	i	s	s	i	p	p	i	$	m	i

Mapping-Algorithmus I: Burrows-Wheeler-Transformation

- Durch die Information (4 i, 1 m, 2 p, 4 s) kann die erste Spalte rekonstruiert werden.
- Durch die Beziehung der beiden Spalten kann die Sequenz jeder Zeile rekonstruiert werden.
- Für jede Teilsequenz (read) können durch alphabetische Überprüfung schnell alle möglichen passenden Positionen gefunden werden.



Die jeweils gleichen Buchstaben sind in der ersten und in der letzten Spalte gleich sortiert! Und zwar rückwärts!

Mapping-Algorithmus I: Burrows-Wheeler-Transformation

- Anzahl nötiger Schritte nur von read-Länge und Anzahl möglicher Zustände und nicht von der Länge der Referenzsequenz abhängig
- Extrem nützlich für Mapping an große Genome

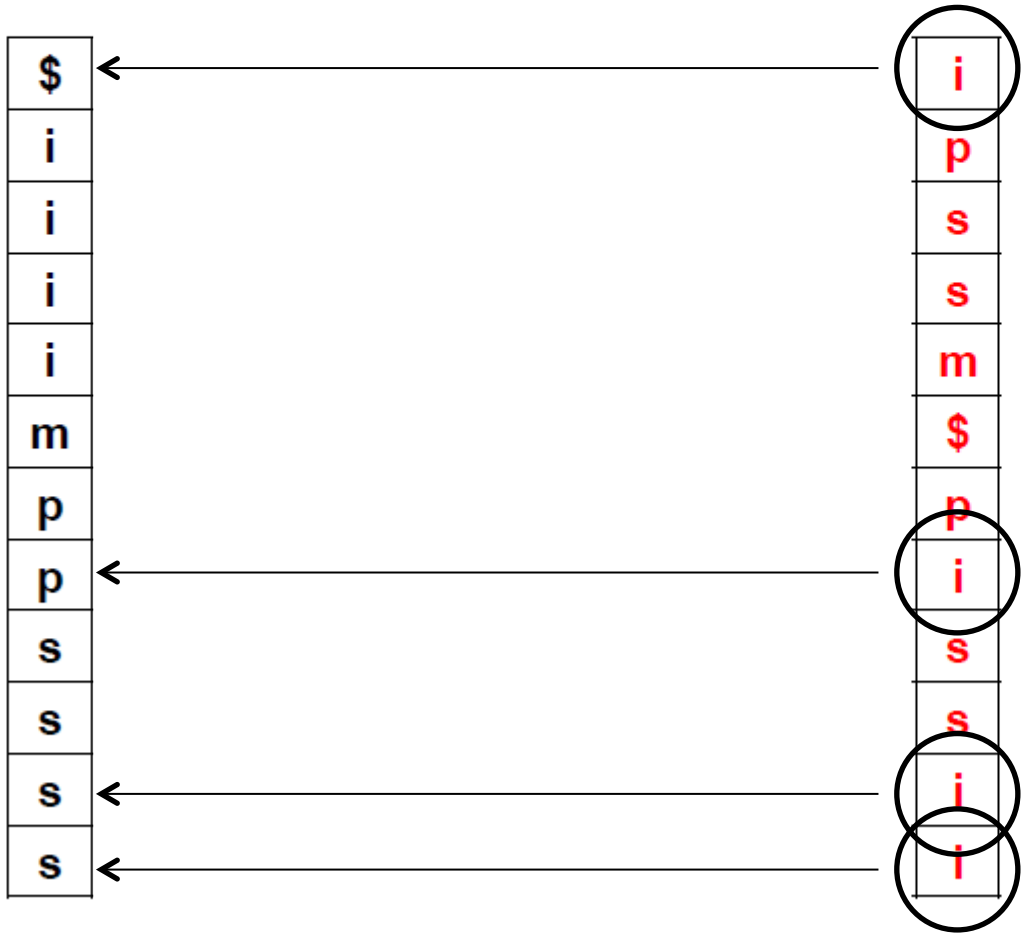
$	m	i	s	s	i	s	s	i	p	p	i
i	$	m	i	s	s	i	s	s	i	p	p
i	p	p	i	$	m	i	s	s	i	s	s
i	s	s	i	p	p	i	$	m	i	s	s
i	s	s	i	s	s	i	p	p	i	$	m
m	i	s	s	i	s	s	i	p	p	i	$
p	i	$	m	i	s	s	i	s	s	i	p
p	p	i	$	m	i	s	s	i	s	s	i
s	i	p	p	i	$	m	i	s	s	i	s
s	i	s	s	i	p	p	i	$	m	i	s
s	s	i	p	p	i	$	m	i	s	s	i
s	s	i	s	s	i	p	p	i	$	m	i

Mapping-Algorithmus I: Burrows-Wheeler-Transformation

Beispiel: Steckt die Zeichenabfolge „ims“ in der Ausgangssequenz?

- Normale Vorgehensweise: positionsweise „ims“ an Ausgangssequenz entlangschieben
- BWT: hinter „i“ kann „$“, „p“ oder „s“ kommen

➡ „ims“ ist nicht in der Sequenz



Mapping-Algorithmus I: Burrows-Wheeler-Transformation

- Temporäre Rekonstruktion relevanter Bereiche
- Für „ims“ bereits in zweiter Zeile fertig (p im Alphabet nach m)

$	m	i	s	s	i	s	s	i	p	p	i
i	$	m	i	s	s	i	s	s	i	p	p
i	p	p	i	$	m	i	s	s	i	s	s
i	s	s	i	p	p	i	$	m	i	s	s
i	s	s	i	s	s	i	p	p	i	$	m
m	i	s	s	i	s	s	i	p	p	i	$
p	i	$	m	i	s	s	i	s	s	i	p
p	p	i	$	m	i	s	s	i	s	s	i
s	i	p	p	i	$	m	i	s	s	i	s
s	i	s	s	i	p	p	i	$	m	i	s
s	s	i	p	p	i	$	m	i	s	s	i
s	s	i	s	s	i	p	p	i	$	m	i

Mapping-Algorithmus I: Burrows-Wheeler-Transformation

- Temporäre Rekonstruktion relevanter Bereiche
- Für „ims“ bereits in zweiter Zeile fertig (p im Alphabet nach m)

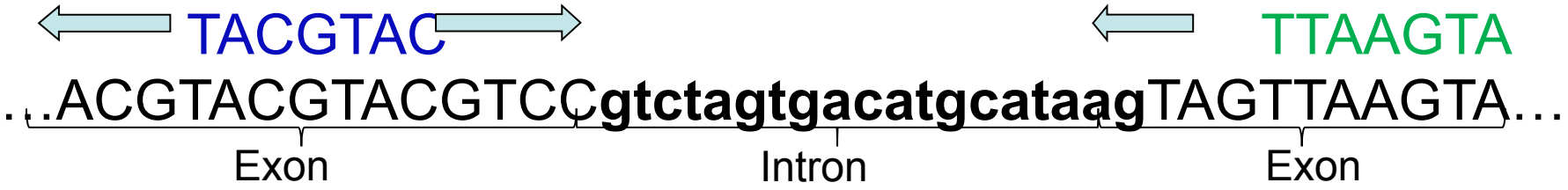
$	m	i	s	s	i	s	s	i	p	p	i
i	$	m	i	s	s	i	s	s	i	p	p
i	p	p	i	$	m	i	s	s	i	s	s
i	s	s	i	p	p	i	$	m	i	s	s
i	s	s	i	s	s	i	p	p	i	$	m
m	i	s	s	i	s	s	i	p	p	i	$
p	i	$	m	i	s	s	i	s	s	i	p
p	p	i	$	m	i	s	s	i	s	s	i
s	i	p	p	i	$	m	i	s	s	i	s
s	i	s	s	i	p	p	i	$	m	i	s
s	s	i	p	p	i	$	m	i	s	s	i
s	s	i	s	s	i	p	p	i	$	m	i

Mapping-Algorithmus II: Seed-and-Extend

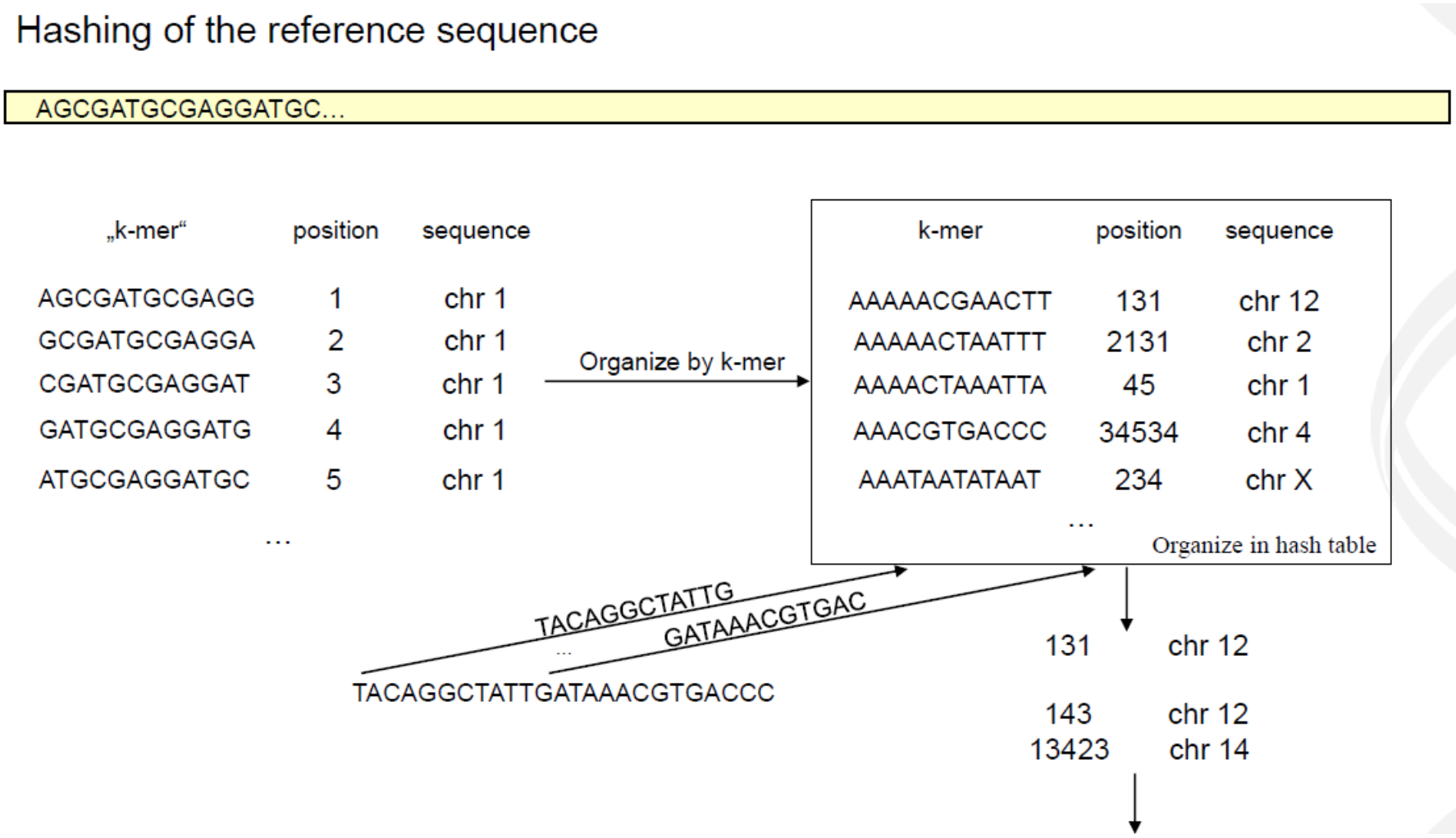
Spliced Read:
AATGTCGTACGTACGTCCTAGTTAAGTA

Seeds: AATGTCG, TACGTAC, GTCCTAG, TTAAGTA

Seed-and-extend:



Mapping-Algorithmus II: Seed-and-Extend

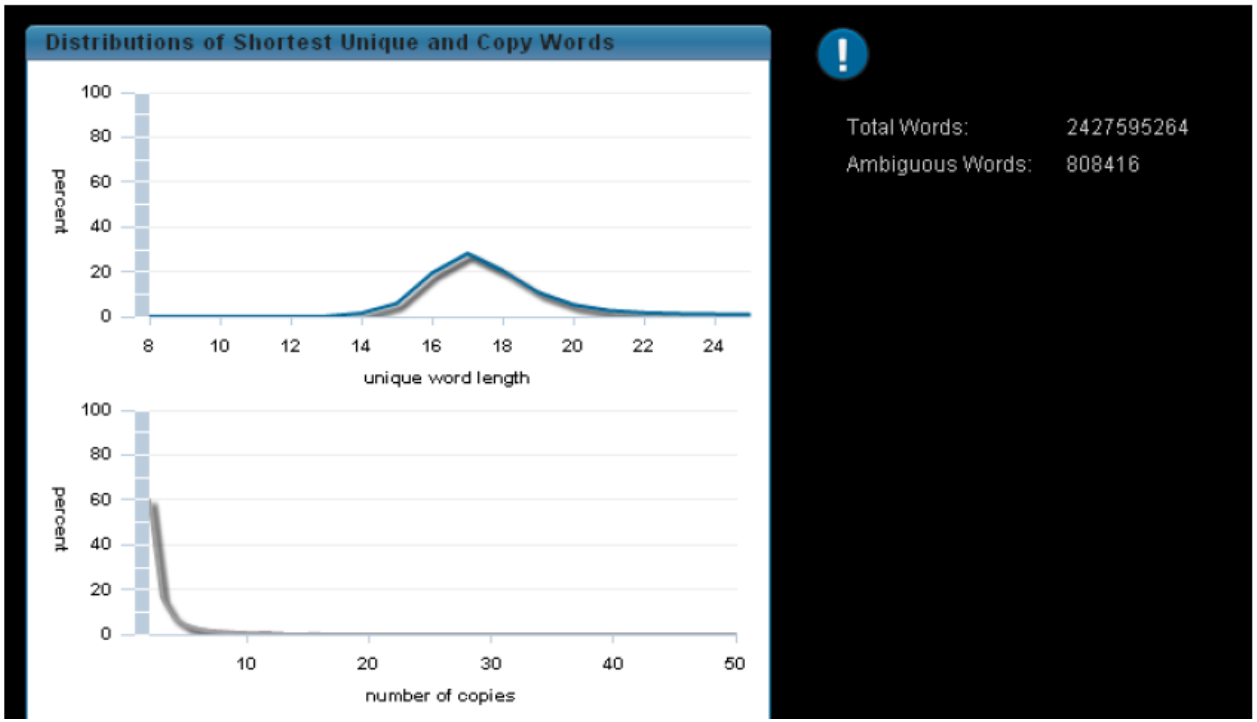


Optimal length of k-mer?

Mapping-Algorithmus II: Seed-and-Extend

Wie lang sollte ein Seed mindestens sein um „unique“ zu mappen?

Hashing: Human genome NCBI build 37 – SUS distribution

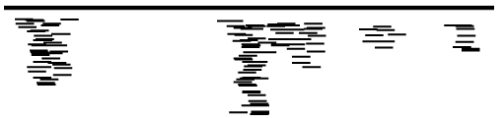


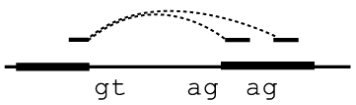
Mapping-Algorithmus II: Seed-and-Extend

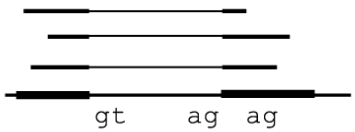
Reads, die Introns überspannen (spliced reads) können nicht gemappt werden

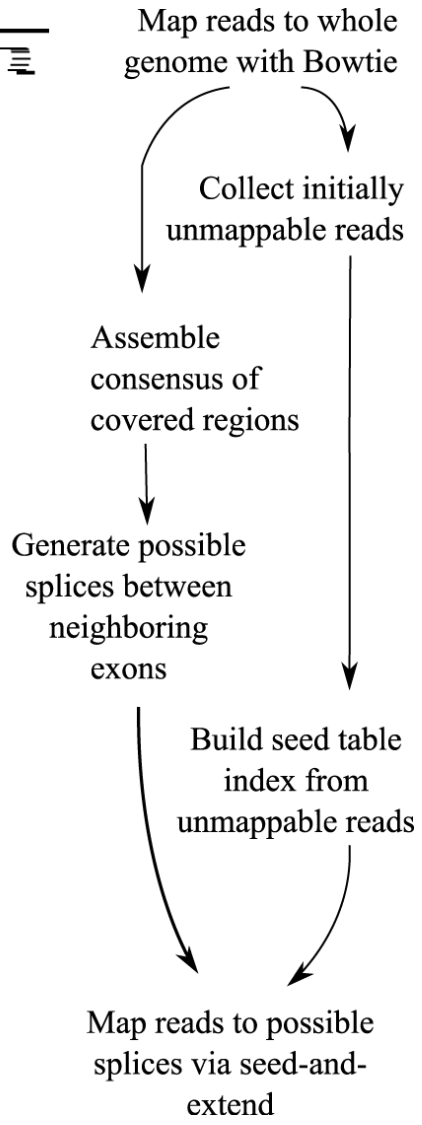
Seed table:
Nicht gemappte Reads werden in kürzere Sequenzen (seeds) unterteilt

Seed-and-extend:
Seeds werden gegen die Exongrenzen gemappt und bei einem Match zu beiden Seiten erweitert









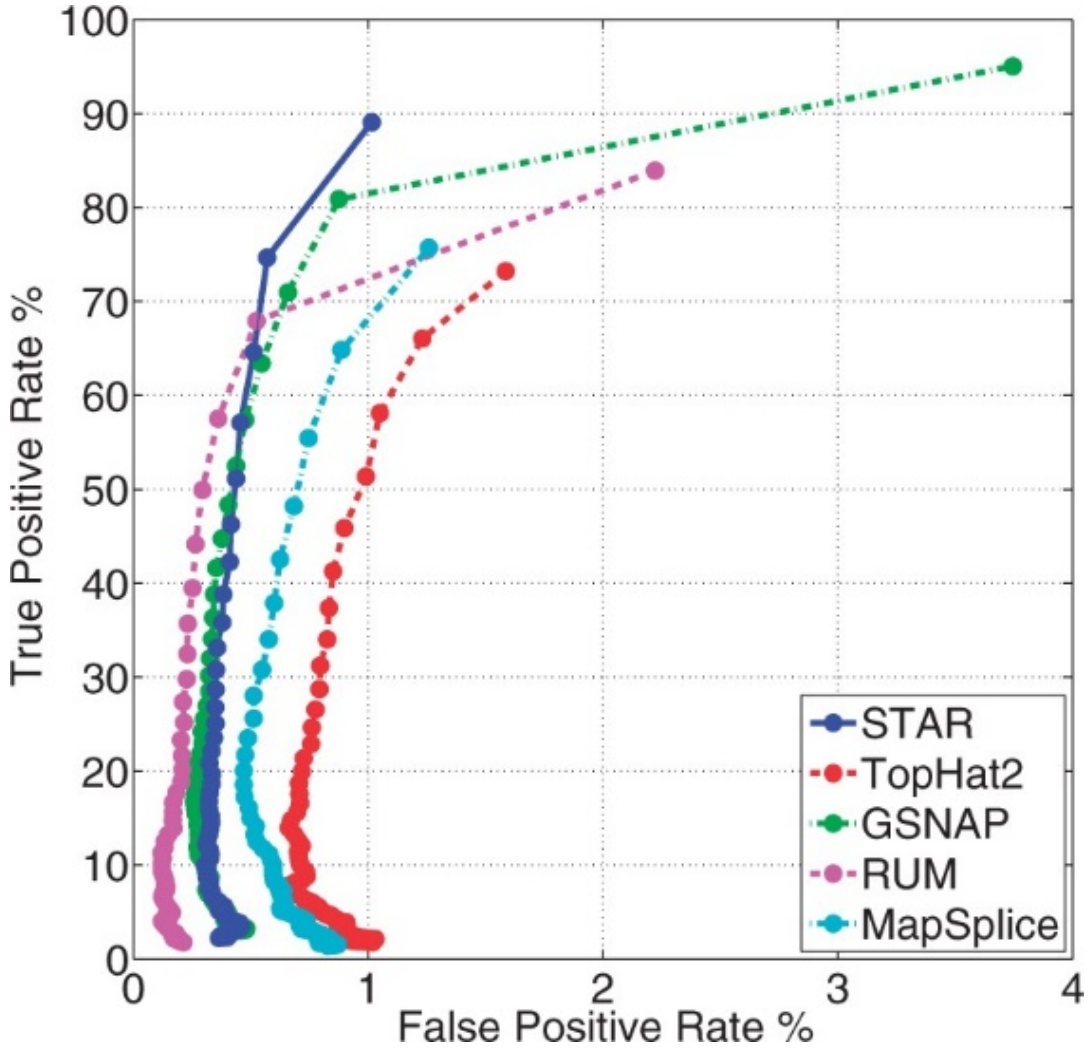
Tools: BWT vs. Hash

Overview available Methods

Tool	link	method
CloudBurst	sourceforge.net/apps/mediawiki/couldburst-bio/index.php?title=CloudBurst	Hash reads
Eland	-	Hash reads
Maq	maq.sourceforge.net	Hash reads
RMAP	rulai.cshl.edu/map	Hash reads
SeqMap	biogibbs.standford.edu/~jingah/SeqMap	Hash reads
SHRiMP	compbio.cs.toronto.edu/shrimp	Hash reads
ZOOM	www.bioinfor.com	Hash reads
BFAST	sourceforge.net/projects/bfast/files	Hash reference
MOM	mom.csb.vcu.edu	Hash reference
Mosaik	bioinformatics.bv.edu/marthlab/Mosaik	Hash reference
SSAHA2	www.sanger.ac.uk/resources/software/ssaha2	Hash reference
NovoAlign	www.novocraft.com	Hash reference
PASS	pass.cribi.unipd.it	Hash reference
PerM	code.google.com/p/perm	Hash reference
ProbeMatch	pages.cs.wisc.edu/~jignesh/probematch	Hash reference
Bowtie	bowtie.cbcb.umd.edu	BWT reference
BWA	bio-bwa.sourceforge.net	BWT reference
SOAP2	Soap.genomics.org.cn	BWT reference

Journal of Human Genetics (2011) **56, 406–414; doi:10.1038/jhg.2011.43; published online 28 April 2011**

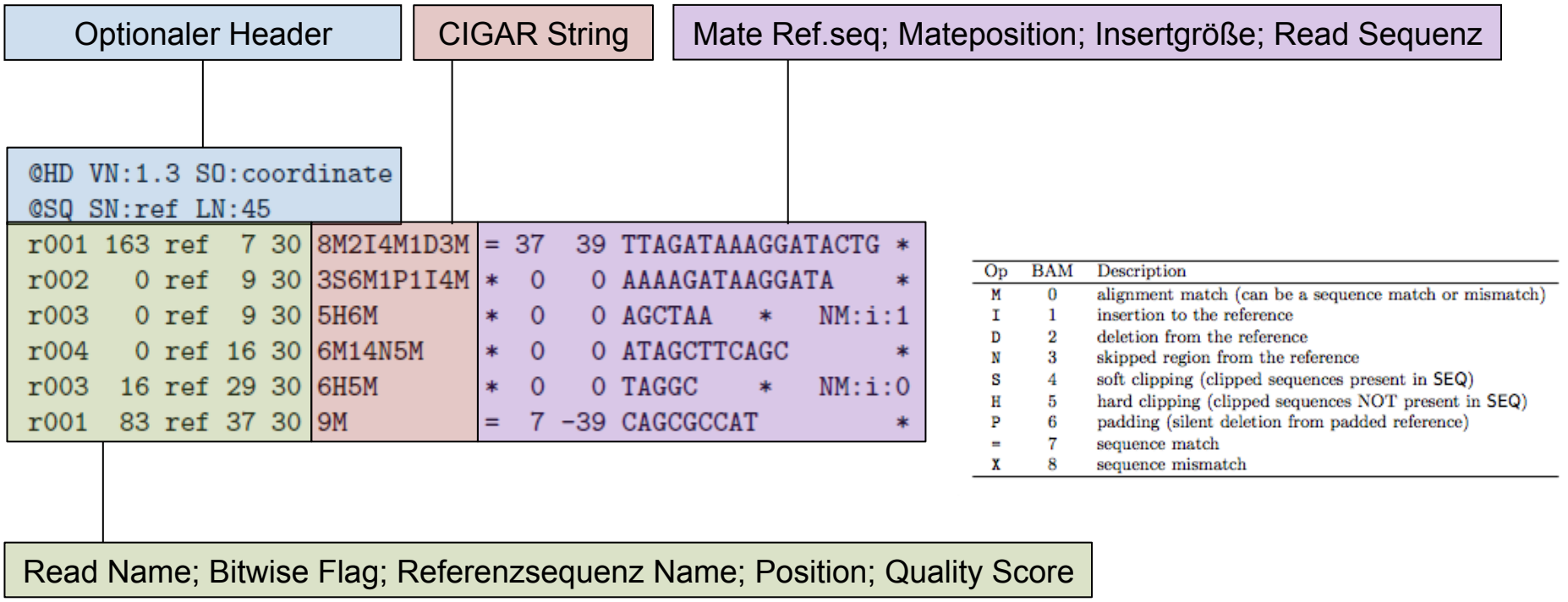
Aber bitte ohne falsch-positive Mapping-Treffer!!



Dobin et al. 2013

Mapping-Output

Das SAM-Format (binär: BAM) ist das wichtigste Dateiformat für Mapping-Daten



Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Was besagt die *Bitwise Flag* des Mapping-Outputs?

Read Name; Bitwise Flag; Referenzsequenz Name; Position; Quality Score

```
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Frage 0: Read paired?
Frage 1: Read mapped in proper pair?
Frage 2: Read unmapped?
Frage 3: Mate unmapped?
Frage 4: Read reverse strand?
Frage 5: Mate reverse strand?
Frage 6: First in pair?
Frage 7: Second in pair?
Frage 8: Not primary alignment?
Frage 9: Read fails platform/vendor quality checks?
Frage 10: Read is PCR or optical duplicate?
Frage 11: Supplementary alignment?

Ja = 1
Nein =0

https://broadinstitute.github.io/picard/explain-flags.html

Bitwise Flag

Ja=1; Nein=0

Frage 0: Read paired?	Ja	2^0=1	*1=1
Frage 1: Read mapped in proper pair?	Ja	2^1=2	*1=2
Frage 2: Read unmapped?	Nein	2^2=4	*0=0
Frage 3: Mate unmapped?	Nein	2^3=8	*0=0
Frage 4: Read reverse strand?	Ja	2^4=16	*1=16
Frage 5: Mate reverse strand?	Ja	2^5=32	*1=32
Frage 6: First in pair?	Nein	2^6=64	*0=0
Frage 7: Second in pair?	Ja	2^7=128	*1=128
Frage 8: Not primary alignment?	Nein	2^8=256	*0=0
Frage 9: Read fails platform/vendor quality checks?	Nein	2^9=512	*0=0
Frage 10: Read is PCR or optical duplicate?	Nein	2^10=1024	*0=0
Frage 11: Supplementary alignment?	Nein	2^11=2048	*0=0

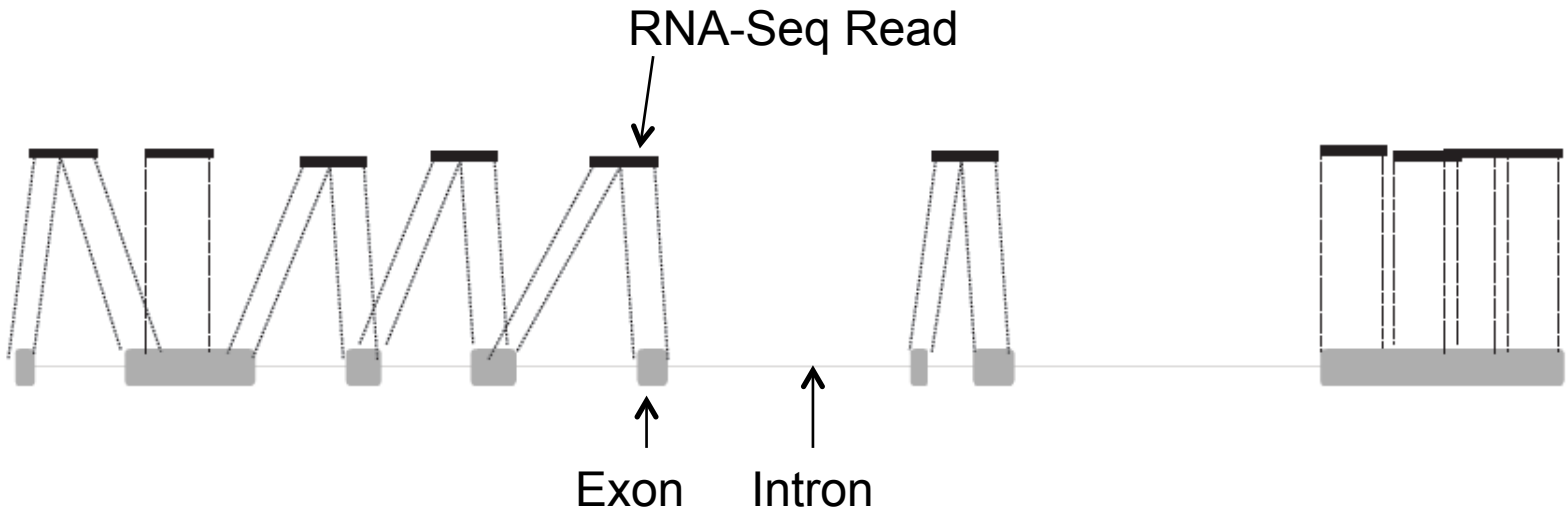
SUMME: 179

Alle obigen Informationen stecken in der bitwise flag 179!

https://broadinstitute.github.io/picard/explain-flags.html

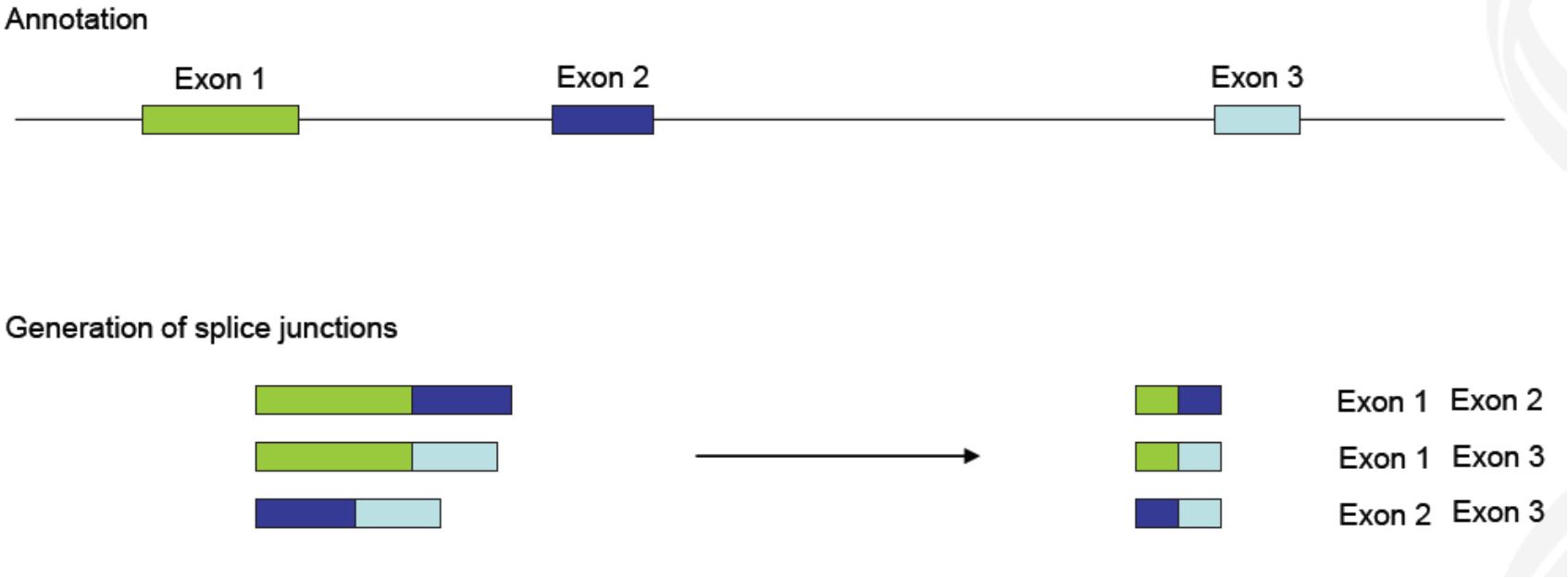
Ein Sonderfall: Mapping bei RNA-Seq

Mapping of RNASeq data to genome requires “gapped alignment”

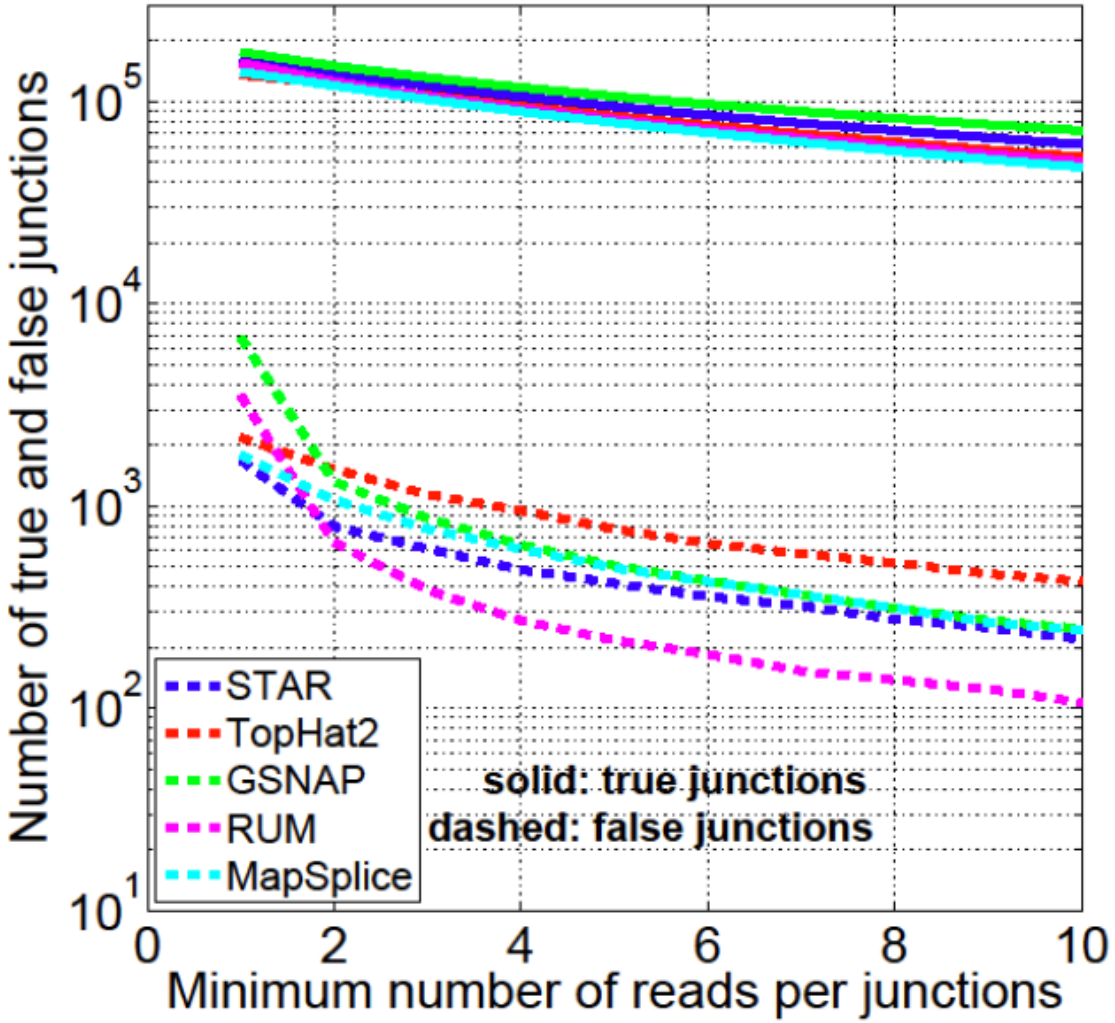


RNA-Seq Mapping

Lösung:
Verwendung von „Splice-aware“ Mappern (STAR, TopHat, CLC usw.)



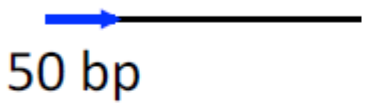
Splice Junction Mapping: Algorithmen im Vergleich



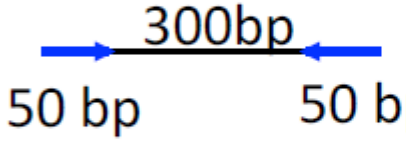
Dobin et al. 2013

Paired-end libraries

Illumina/SOLiD

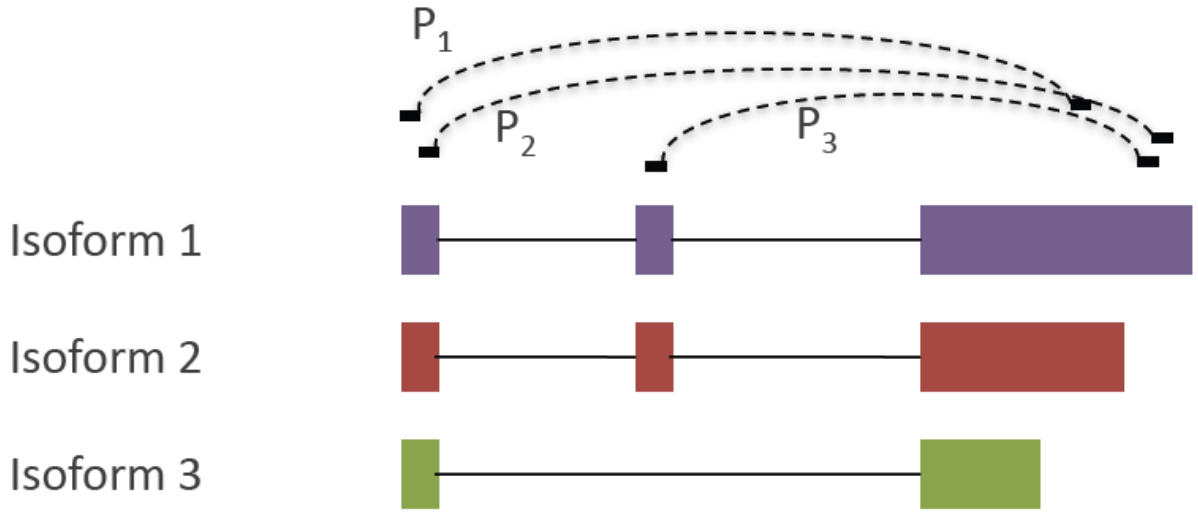


Single end (SE)



Paired-end (PE), short fragment ends

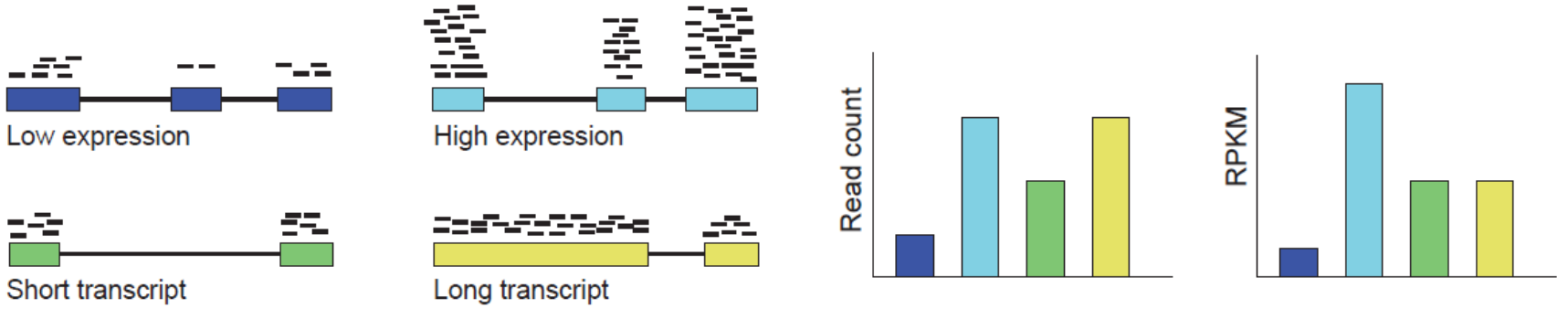
Paired-end: Vorteile



Paired ends increase isoform deconvolution confidence

- P_1 originates from isoform 1 or 2 but not 3.
- P_2 and P_3 originate from isoform 1

Nach dem Mapping:
Quantifizierung und Normalisierung der Expressionsstärke

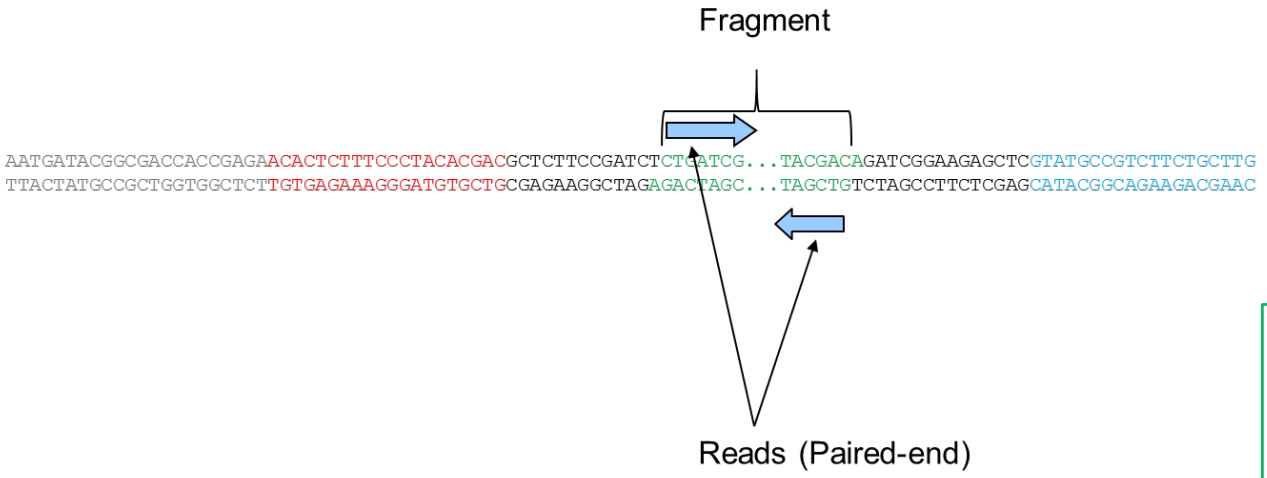


Fragmentation of transcripts results in length bias: longer transcripts have higher counts
Different experiments have different yields. Normalization is required for cross lane comparisons:

$$\textbf{RPKM} = \frac{\text{Gemappte reads innerhalb des Transkripts}}{\text{Länge des Transkripts (Kb) x Gemappte reads insgesamt (Mio.)}}$$

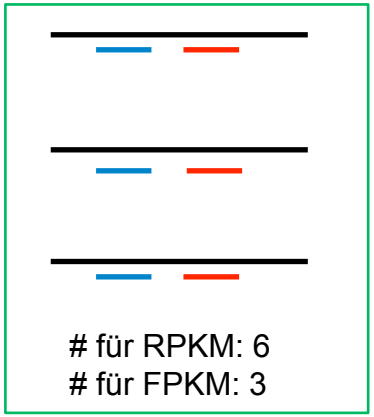
Reads per kilobase of exonic sequence per million mapped reads
(*Mortazavi* et al Nature methods 2008)

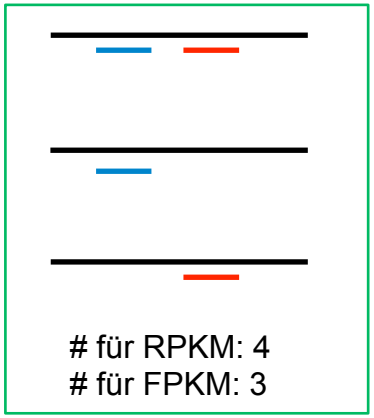
RPKM vs. FPKM



- **Fragment**: physischer DNA-Schnipsel aus einem Transkript
- **Read**: bioinformatischer Sequenzschnipsel

- Relevantester Unterschied: broken pairs
  - → FPKM: Fragment gefunden und voll gewertet
  - → RPKM: ein Teil gefunden und gewertet





Normalisierung der Expressionsstärke: TPM statt RPKM

(Transcripts per Kilobase per Million)

- RPKM ist kein „Kuchenstück“: die Summe aller RPKMs ≠100%
  → gibt nicht den Anteil des Transkripts pro eingesetzter mRNA wieder

- RPKMs teilweise schlecht untereinander vergleichbar wenn Transkripte generell länger oder kürzer sind

- → anderer „Scaling factor“ sinnvoll:

$$\mathrm{TPM} = \frac{r_g \times \mathrm{rl} \times 10^6}{\mathrm{fl}_g \times T} \quad \textbf{statt} \quad \mathrm{RPKM}_g = \frac{r_g \times 10^9}{\mathrm{fl}_g \times R}$$

$$T = \sum_{g \in G} \frac{r_g \times \mathrm{rl}}{\mathrm{fl}_g}$$

r_g: number of reads mapped to g
rl: Readlänge
fl_g: feature (Exon/Transkript) Länge von g
T: Summe aller „transcripts sampled“
R: Summe aller gemappten Reads

- → TPM ist proportional der relativen molaren Konzentration

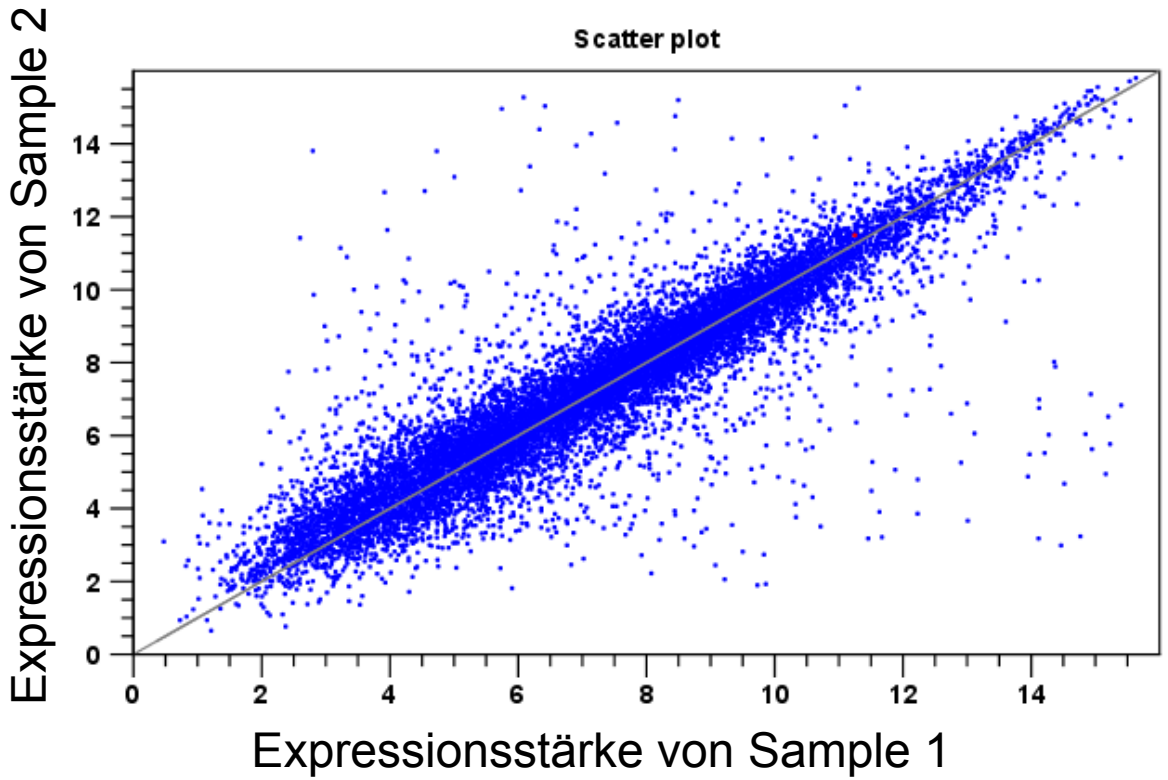
$$\mathrm{RPKM}_g = \frac{T \times 10^3}{R \times \mathrm{rl}} \times \mathrm{TPM}_g$$

Differenziell regulierte Gene der RNA-Seq Analyse

Fold Change	ID	Symbol
-5,683	83896	KRTAP3-1
-5,464	4151	MB
-4,784	9235	IL32
-4,779	7805	LAPTM5
-4,755	254228	FAM26E
-4,736	90853	SPOCD1
-4,694	8530	CST7
-4,470	6317	SERPINB3
-4,363	79148	MMP28
-4,260	6347	CCL2
-4,215	8091	HMGA2
-4,174	2267	FGL1
-4,108	140628	GATA5
-3,893	4856	NOV
-3,874	8740	TNFSF14
-3,852	54538	ROBO4
-3,767	115701	ALPK2
-3,756	286	ANK1
-3,755	159963	SLC5A12
-3,711	725	C4BPB
-3,702	125704	FAM69C
-3,693	8632	DNAH17
-3,666	84659	RNASE7
-3,647	400950	C2orf91
-3,644	23092	ARHGAP26
-3,636	54210	TREM1
-3,623	7850	IL1R2
-3,611	4118	MAL

Genliste mit „fold-change“-Werten, die eine differenzielle Regulation auf mRNA-Ebene zeigen.

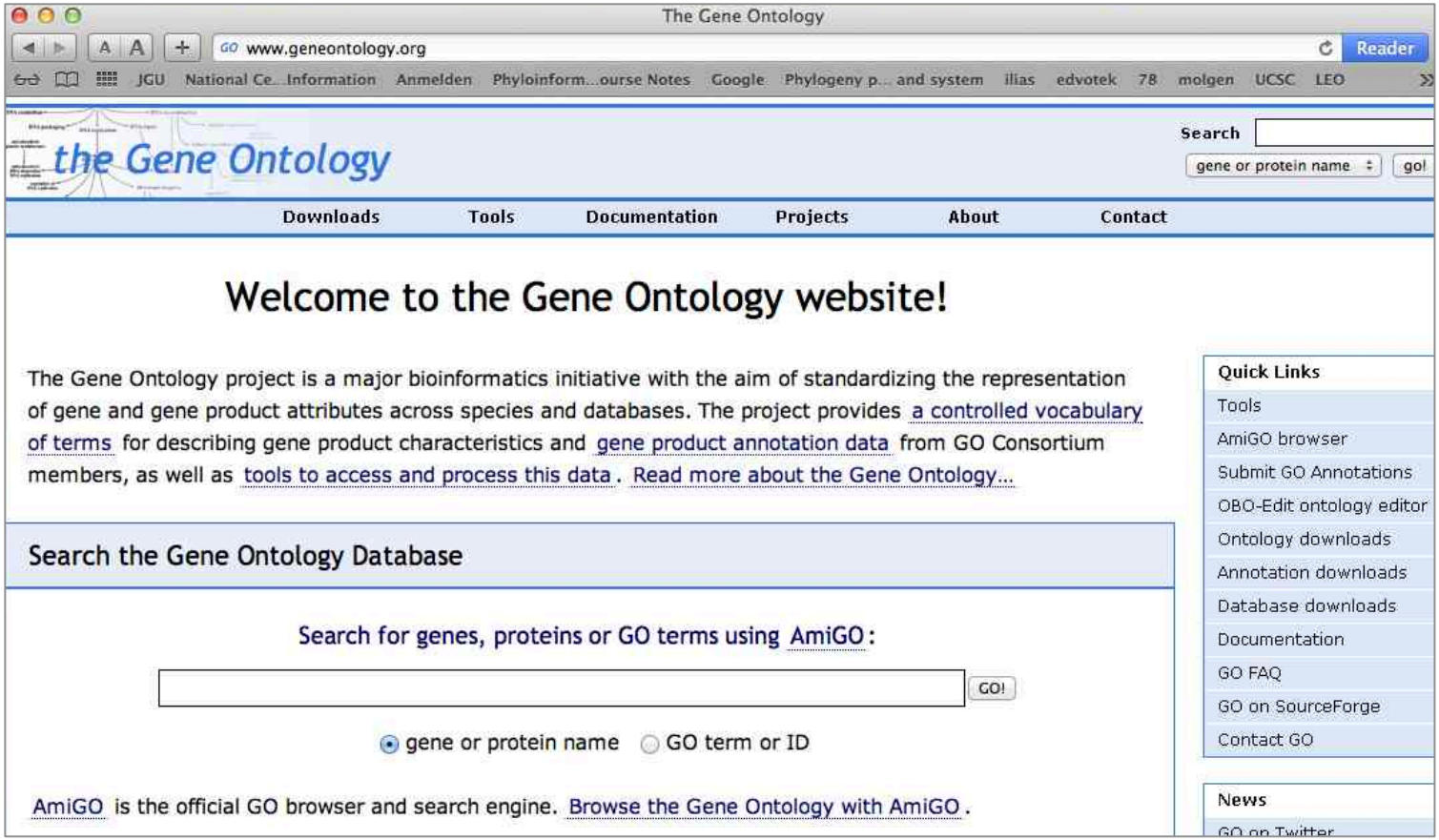
Scatter Plot:
Visuelle Darstellung differenziell regulierter Gene der RNA-Seq Analyse



Biologische Interpretation der differenziell regulierten Gene

- Gibt es eine gemeinsame Assoziation mit bestimmten Zellkomponenten?
- Gibt es eine gemeinsame Assoziation mit bestimmten Funktionen, z.B. Schutz vor Sauerstoffradikalen?
- Gibt es Pathways, in denen viele der Gene vorkommen?
- Gibt es gemeinsame Regulatoren der Gene, z.B. stressinduzierbare Transkriptionsfaktoren?

Biologische Interpretation: Gene Ontology Annotation



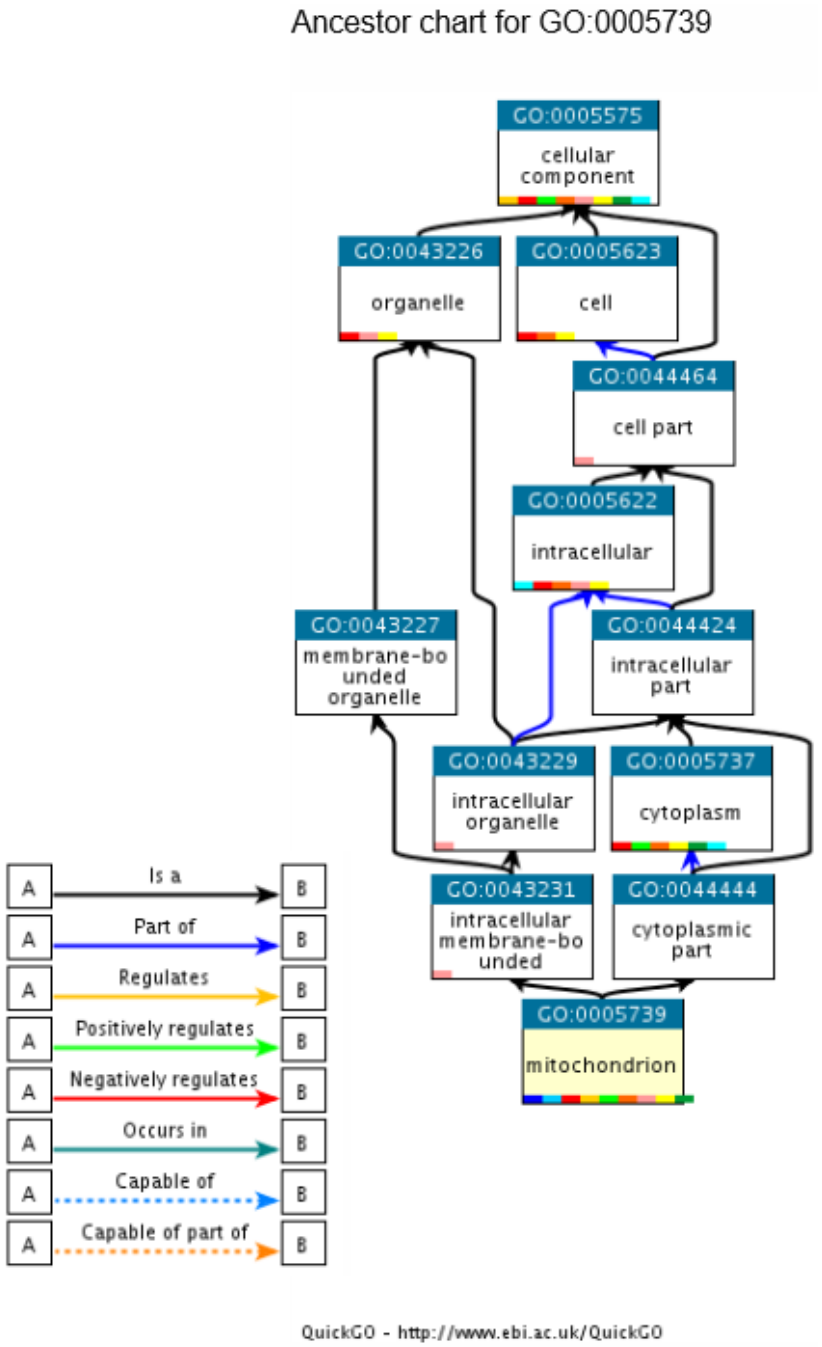
Aufspüren von Assoziation mit bestimmten biologischen Prozessen, Zell-Komponenten, und molekularen Funktionen

Gene Ontology Annotation

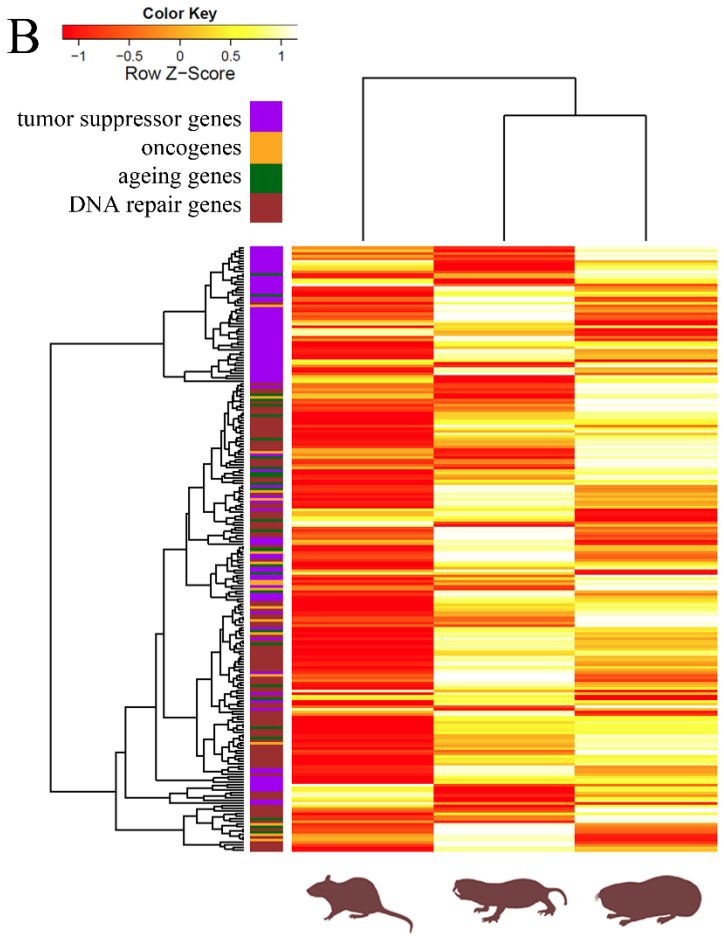
- Jedes Gen wird in der Gene Ontology-Datenbank mit mehreren Schlagwörtern (=terms) versehen

- Diese sind vernetzt und stehen in verschiedenen Beziehungen zueinander (z.B. „part of“, „is a“, „regulates“)

→ Zusammenfassung und Vernetzung von Genen zu biologisch sinnvollen Gruppen (~25 000 Eigenschaften)



Enrichment von funktionell annotierten Genen in Datensätzen



Viele Tumorsuppressorgene des Datensatzes sind angereichert, aber handelt es dabei sich um eine signifikante Anreicherung?

→ Fisher‘s Exact test:

https://www.youtube.com/watch?v=udyAvvaMjfM

Schmidt et al. 2017

Enrichment von funktionell annotierten Genen in Datensätzen

To determine whether any GO terms annotate a specified list of genes at a frequency greater than that would be expected by chance, GO::TermFinder calculates a *P*-value using the hypergeometric distribution:

Hypergeometrische Verteilung

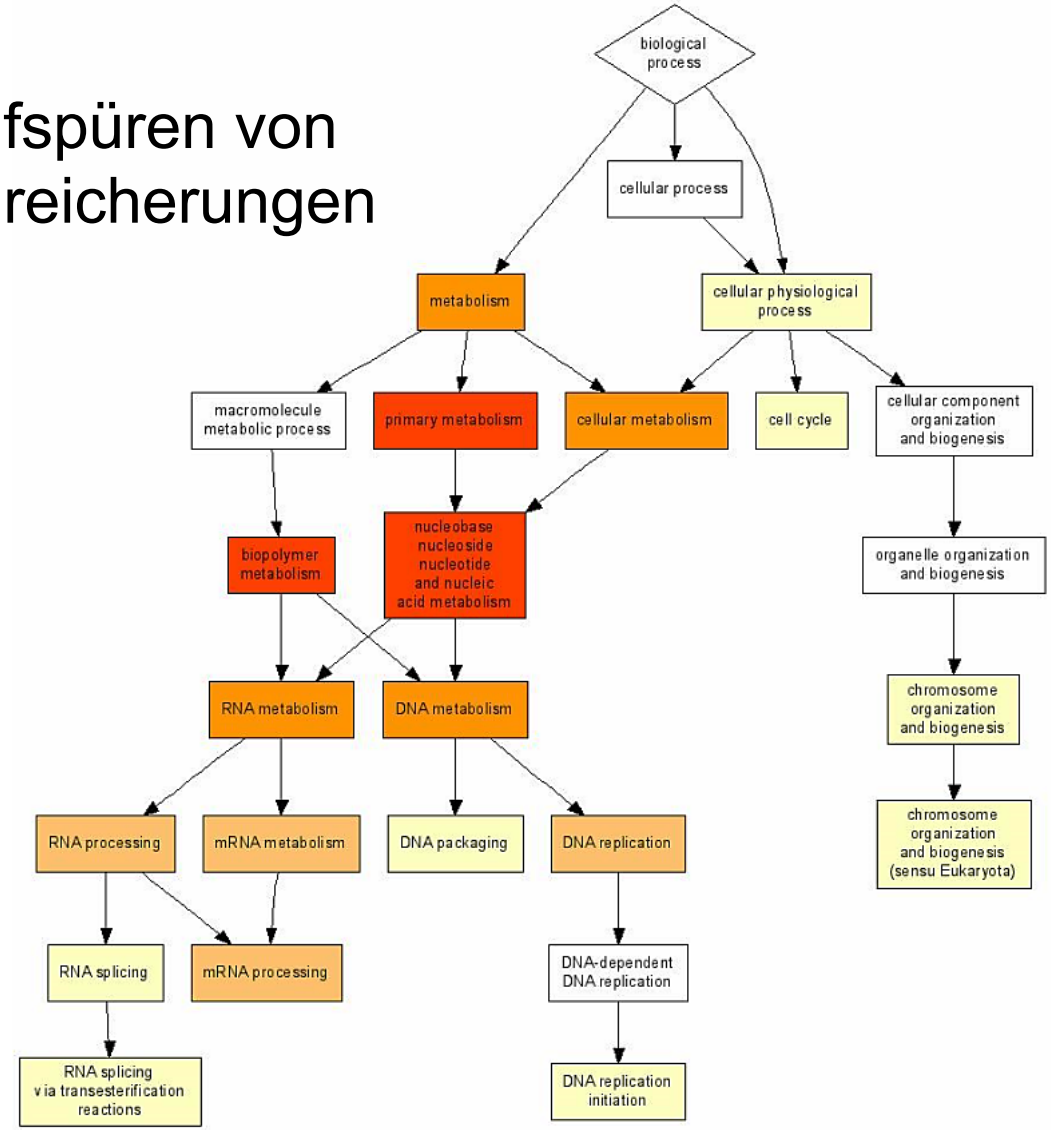
$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}}.$$

In this equation, N is the total number of genes in the background distribution, M is the number of genes within that distribution that are annotated (either directly or indirectly) to the node of interest, n is the size of the list of genes of interest and k is the number of genes within that list which are annotated to the node. The background distribution by default is all the genes within a given annotation file, though the software also allows a user-defined background distribution, such that biases in the sampling population (e.g. the genes represented on a microarray) can be accounted for correctly. The hypergeometric distribution is sampling without replacement—for instance, consider a bag with 500 red and 500 green beads. If 20 beads were selected randomly, and beads were not replaced after each selection, and 17 were green, we would use the hypergeometric distribution to calculate the *P*-value as the probability of picking 17, or more, green beads from 20, given that there are 500 of each in the background distribution.

Boyle et al. 2004

Gene Ontology Analysen

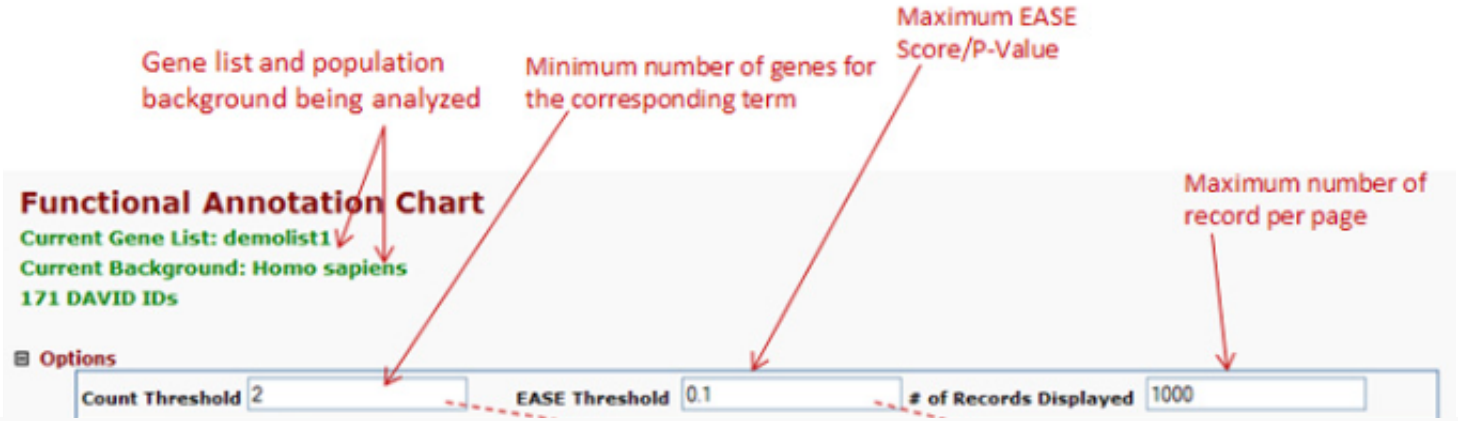
Aufspüren von Anreicherungen



GORILLA

http://cbl-gorilla.cs.technion.ac.il/help.html

Gene Ontology-Listen in DAVID



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	GOTERM_CC_FAT	cytosol	RT		45	2,1	5,5E-17	1,2E-14
☐	GOTERM_CC_FAT	cytosolic part	RT		16	0,7	1,7E-12	1,9E-10
☐	GOTERM_CC_FAT	melanosome	RT		10	0,5	4,8E-8	3,6E-6
☐	GOTERM_CC_FAT	pigment granule	RT		10	0,5	4,8E-8	3,6E-6
☐	GOTERM_CC_FAT	chaperonin-containing T-complex	RT		5	0,2	1,5E-7	8,5E-6
☐	GOTERM_CC_FAT	cytosolic ribosome	RT		9	0,4	3,3E-7	1,5E-5
☐	GOTERM_CC_FAT	small ribosomal subunit	RT		8	0,4	8,3E-7	3,1E-5
☐	GOTERM_CC_FAT	cytosolic small ribosomal subunit	RT		7	0,3	8,6E-7	2,7E-5
☐	GOTERM_CC_FAT	ribonucleoprotein complex	RT		18	0,8	1,0E-6	2,9E-5
☐	GOTERM_CC_FAT	ribosomal subunit	RT		10	0,5	1,1E-6	2,7E-5
☐	GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	RT		44	2,0	1,4E-6	3,1E-5
☐	GOTERM_CC_FAT	non-membrane-bounded organelle	RT		44	2,0	1,4E-6	3,1E-5
☐	GOTERM_CC_FAT	ribosome	RT		12	0,6	1,6E-6	3,3E-5
☐	GOTERM_CC_FAT	chromatin	RT		8	0,4	1,4E-3	2,5E-2
☐	GOTERM_CC_FAT	organelle lumen	RT		28	1,3	1,5E-3	2,6E-2
☐	GOTERM_CC_FAT	proteasome complex	RT		5	0,2	1,6E-3	2,5E-2

RT = related term

Percentage, e.g. 14/171=8.2% (involved genes/total genes)

Modified Fisher Exact P-Value, EASE Score. The smaller, the more enriched.

Biologische Interpretation der differenziell regulierten Gene

- Gibt es eine gemeinsame Assoziation mit bestimmten Zellkomponenten?
- Gibt es eine gemeinsame Assoziation mit bestimmten Funktionen, z.B. Schutz vor Sauerstoffradikalen?
- Gibt es Pathways, in denen viele der Gene vorkommen?
- Gibt es gemeinsame Regulatoren der Gene, z.B. stressinduzierbare Transkriptionsfaktoren?

Enrichment von Genen in Pathways

fatty acid metabolism predicted to be decreased (z-score -2,807). Overlap p-value **1,75E-04**

22 of 38 genes have expression direction consistent with decreases in fatty acid metabolism.

ADD TO MY PATHWAY | ADD TO MY LIST | CUSTOMIZE TABLE | CREATE DATASET

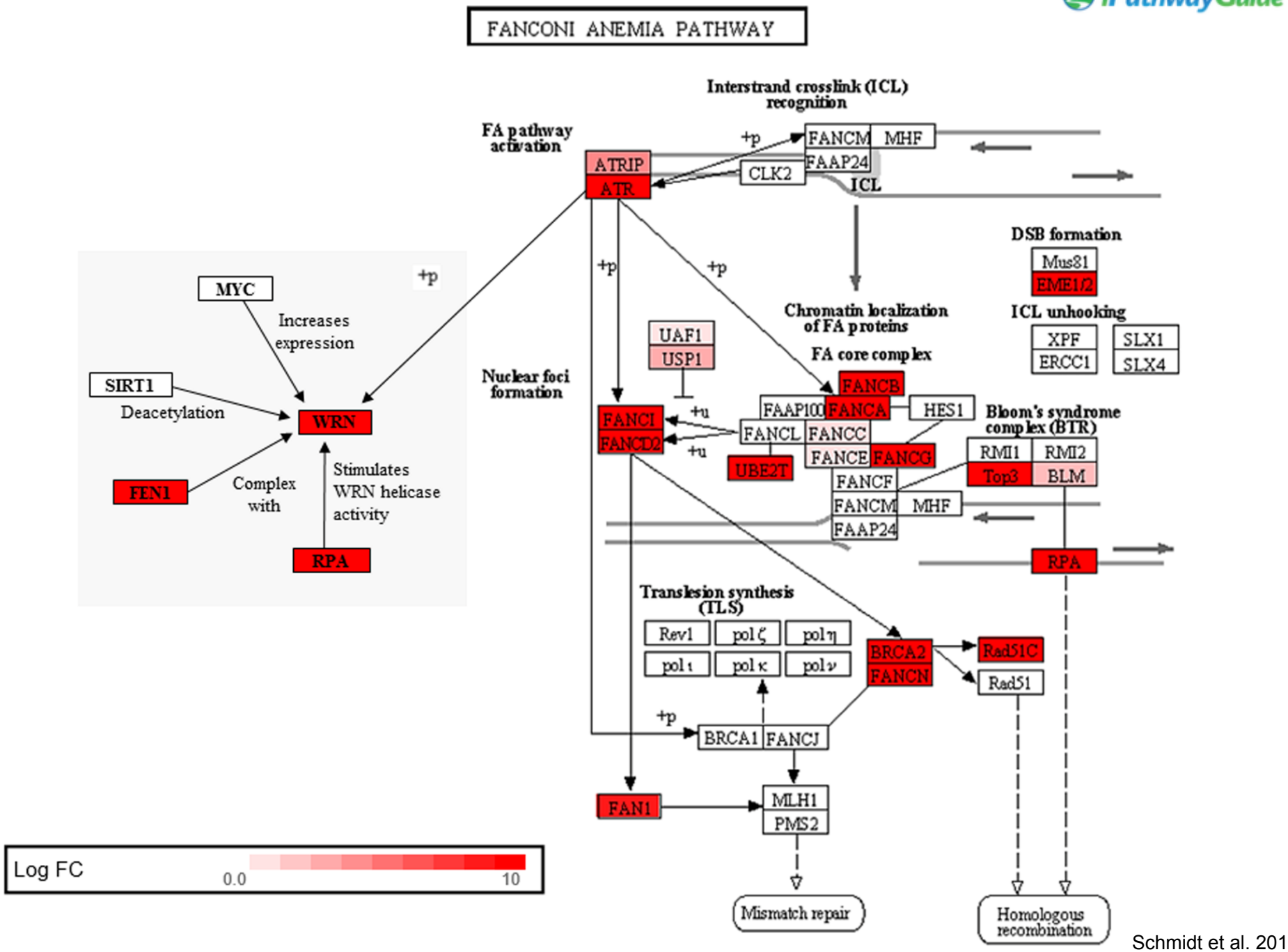
ID	Genes in dataset	Prediction (based on expression direction)	Fold Change	Findings
5130	PCYT1A	Decreased	-2,710	Increases (2)
3569	IL6	Decreased	-2,197	Increases (5)
301	ANXA1	Decreased	-1,019	Increases (1)
654	BMP6	Decreased	-1,493	Increases (2)
54677	CROT	Decreased	-1,323	Increases (5)
7357	UGCG	Decreased	-1,545	Increases (11)
3606	IL18 (includes EG:16173)	Decreased	-1,175	Increases (7)
355	FAS	Decreased	-1,085	Increases (11)
2172	FABP6	Decreased	-2,340	Increases (1)
2571	GAD1 (includes EG:100006588)	Decreased	-2,880	Increases (4)
7431	VIM	Decreased	1,530	Decreases (1)
9235	IL32	Decreased	-4,784	Increases (1)
9415	FADS2	Decreased	-1,106	Increases (6)
857	CAV1	Decreased	-1,063	Increases (7)
7124	TNF	Decreased	-3,305	Increases (114)
6288	SAA1	Decreased	-2,312	Increases (2)
6505	SLC1A1	Decreased	-1,640	Increases (1)
177	AGER	Decreased	2,848	Decreases (1)
7097	TLR2	Decreased	-1,546	Increases (4)
3675	ITGA3	Decreased	-1,075	Increases (1)
3992	FADS1	Decreased	-1,171	Increases (3)
627	BDNF	Decreased	-1,185	Increases (8)
4973	OLR1	Increased	-1,872	Decreases (1)
3603	IL16	Increased	1,482	Increases (10)
2687	GGT5	Increased	-1,559	Decreases (2)
2919	CXCL1	Increased	-2,607	Decreases (2)
53947	A4GALT	Increased	-1,651	Decreases (1)

“Voraussagen” eines aktivierten Fettsäuremetabolismus **Input Genliste mit fold changes**

Abgleich der “vorausgesagten Aktivität” von Genen eines aktivierten Fettsäure-Metabolismus (auf Literatur basierend) mit der eigenen Genliste und deren Regulationsrichtungen:

Die Regulationsrichtung differenziell exprimierter Gene widerspricht meistens genau der Regulationsrichtung, die hier eine Verstärkung des Fettsäuremetabolismus anzeigen würde

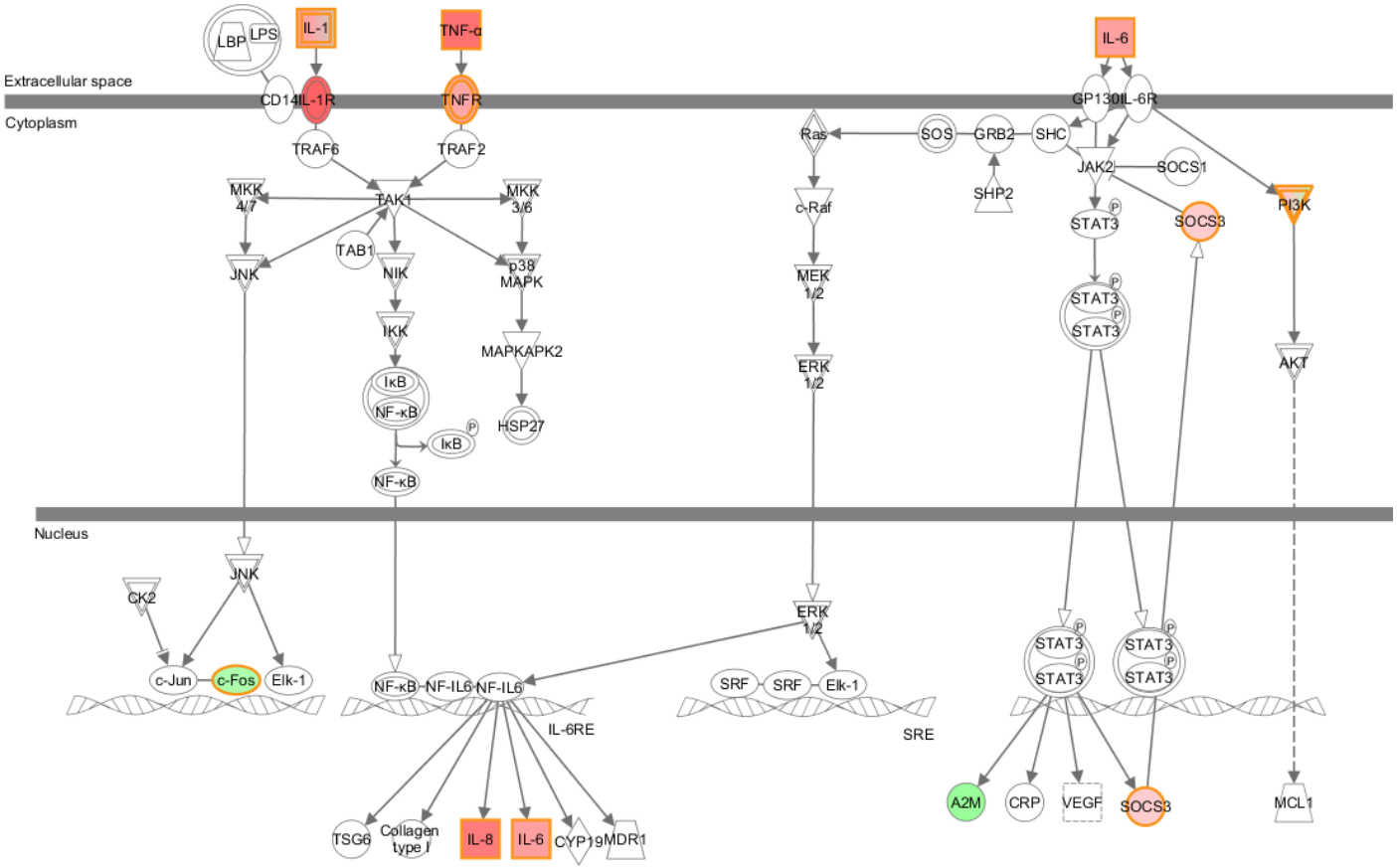
→ Der Fettsäuremetabolismus ist signifikant herabreguliert!



Schmidt et al. 2017

Enrichment von Genen in Pathways

Canonical pathway: IL6 signaling



Grün: upregulated gene in MB-
Rot: downregulated gene in MB-

Detektion gemeinsamer *Upstream regulators*

INGENUITY SYSTEMS

Upstream Regulator	Fold Change	Predicted Activati...	Bias-corrected z-sc...	p-value of overlap	Target molecules in dataset
TP53 (includes EG:22059)		Inhibited	-2,456	1,07E-04	↓ALDH4A1, ↓ANXA1, ↓BIRC3, ↓BMP1, ↓CAV1, ↑CLU, ↑COL3A1, ↓COL4A1, ↓CRIP2, ↑CRYAB
RELA		Inhibited	-2,380	1,94E-07	↓BIRC3, ↓CAV1, ↓CCL2, ↓CXCL1, ↓F3, ↓FABP6, ↓FAS, ↑FOS, ↓ICAM1, ↓IL32
EGR1			-1,957	1,00E-05	↓CCL2, ↑CLU, ↓F3, ↓FAS, ↓FLT1, ↓FOSL1, ↓GAD1 (includes EG:100006588), ↓GADD45A, ↓ICAM1, ↓IL8
BRCA1			-1,756	4,79E-01	↓FAS, ↓GADD45A, ↓SERPINE2, ↓TNF, ↓TNFAIP2
SMAD2			-1,701	9,48E-03	↓DAPK1, ↓FLT1, ↓IL6, ↓MMP2, ↓SERPINE1, ↓THBS1, ↓TPM1 (includes EG:22003)
STAT4			-1,610	1,21E-02	↑C13orf15, ↓CCL2, ↓ENO2, ↓FSCN1, ↓GBE1, ↓IL6, ↓JAG2, ↓PYGL, ↓SCG5, ↓SERPINB5
NFKBIB			-1,502	1,36E-03	↓CCL2, ↓CXCL1, ↓ICAM1, ↓IL6, ↓IL8, ↓SAA1, ↓TNF
ETS1	↓-1,415		-1,490	7,10E-03	↑BCL11A, ↓CAV1, ↓CCL2, ↓ETS1, ↓FLT1, ↓ICAM1, ↓MMP2, ↓MMP7, ↑PEG10, ↓RHOBTB3
STAT6			-1,426	2,31E-03	↓ACP5, ↑BCL6, ↓CST7, ↓GNA15, ↓IL4R, ↓IL6, ↓MMP2, ↑MYB, ↓NEDD9, ↓PDGFA
ID3			-1,418	3,72E-04	↓CXCL1, ↓ELOVL6, ↓ICAM1, ↓IL6, ↓IL7R, ↓IL8, ↓MMP2
SP3			-1,395	8,05E-06	↑DNM1, ↑F2R, ↓F3, ↓FLT1, ↑FOS, ↓FOSL1, ↓ITGA2, ↓LAMA1, ↓MMP2, ↓PCYT1A
FOS	↑1,121		-1,302	3,01E-07	↓ACP5, ↓AKR1C3, ↓BDNF, ↑CASZ1, ↓CCL2, ↑CLU, ↓COL16A1, ↑CYCS, ↑EEA1, ↓ELOVL6
NOTCH1			-1,299	1,80E-08	↓ANKRD1, ↓BIRC3, ↓ENO2, ↓FLT1, ↑FOS, ↓FOSL1, ↓ICAM1, ↓ID1, ↓IGFBP2, ↓IL6
SMARCA4			-1,259	1,11E-05	↓BIN1, ↓CCL2, ↓CD74, ↓FADS3, ↑FOS, ↓GADD45A, ↓HKDC1, ↑IGFBP5, ↓IL6, ↓ITGA3
STAT1			-1,229	4,08E-02	↓AXL, ↓CCL2, ↓FAS, ↑FOS, ↓ICAM1, ↓IL6, ↓IL8, ↓PDGFA, ↓SAMHD1, ↓SOCS3
RELB	↓-1,049		-1,182	4,86E-03	↓IL6, ↓IL7, ↑MYB, ↓PRDM1, ↓RELB, ↓TNF
ATF2			-1,162	2,63E-02	↓GADD45A, ↓IL6, ↓IL8, ↓SERPINB5, ↓TGFB2, ↓TNF
CTNNB1			-1,158	2,64E-10	↓ALDH1A1, ↓ANXA1, ↓MMP1, ↓CD34, ↑CLU, ↓COL4A1, ↓COL4A2, ↓CYP4F12, ↓DKK1, ↓ECM1 (includes EG:100332249)
EPAS1			-1,077	1,62E-02	↓BIRC3, ↓ENO2, ↓FLT1, ↑FOS, ↓GBE1, ↑IGFBP5, ↓IL6, ↓LOX, ↓NRN1, ↓SERPINE1
MLL2			-1,059	6,43E-03	↓CRIP2, ↓DKK1, ↓LAMB3, ↓LOXL1
REL			-1,052	2,41E-03	↓F3, ↑FOS, ↓ICAM1, ↓IL18 (includes EG:16173), ↓IL6, ↓IL8, ↓RELB, ↓TNF, ↓TNFAIP3, ↑VIM
GABPA			-1,037	5,59E-04	↓FAS, ↑IL16, ↓IL7R, ↓ROBO4, ↓TNC, ↓UTRN
CEBPA			-0,997	3,35E-04	↓AKR1B1, ↓ANXA1, ↓BIN1, ↑CYCS, ↓FLT1, ↑FOS, ↓GADD45A, ↓ICAM1, ↓ID1, ↓IL6
FOSL2			-0,972	8,98E-05	↓F3, ↓FAS, ↓FOSL1, ↓IL8, ↓MMP2, ↓RELB, ↓TIMP1
HOXA10			-0,972	1,36E-03	↓ALDH1A1, ↓COL15A1, ↑COL3A1, ↓DKK1, ↓FLT1, ↓GAS6, ↓ID1
CEBPB (includes EG:1051)			-0,904	1,49E-04	↓ALDH1A1, ↓ARL6, ↓BDNF, ↓CCL2, ↓DAPK1, ↓FAS, ↑FOS, ↓GA
MTPN			-0,899	1,32E-03	↓ANXA1, ↓FAS, ↑FOS, ↓GAS6, ↓IL6, ↓PLAT, ↓SERPINE1, ↓TAGLN, ↓TGFB2, ↓TNF
NFATC2			-0,870	1,91E-02	↓ACP5, ↓FAS, ↓PLD1, ↑RGS2 (includes EG:19735), ↓S100A3, ↓TNF
AR					↓ABCC4 (includes EG:10257), ↓CAV1, ↓CLDN11, ↑COL3A1, ↓ENO2, ↑IGFBP5, ↓IL6, ↓INPP4B, ↓ITGA2, ↓MAOA
EZH2					↓BIRC3, ↓C4BPB, ↑CDK6, ↓CXCL1, ↓DENND2A, ↓DKK1, ↓ICAM1, ↓IL6, ↓IL8, ↓MMP7
ARNT2					↑AGER, ↓AKR1B1, ↓AOX1, ↓CALB2, ↓COL12A1, ↓GALNT4, ↓ICAM1, ↓LAMB3, ↓LCP1, ↑MDM4
HIF1A					↑BACE1, ↓ENO2, ↓ETS1, ↓FLT1, ↑FOS, ↓GBE1, ↓IGFBP2, ↑IGFBP5, ↓IL6, ↓IL8
ESR2					↓CAV1, ↓CCL2, ↓IL8, ↓ITGA2, ↓MMP2, ↓NEDD9, ↓PDGFA, ↓PDZK1, ↓SOCS3, ↓TGM2
SP1			-0,622	3,75E-07	↑AGER, ↑BACE1, ↓BDNF, ↓CAV1, ↓CCL2, ↑CDK6, ↑CRYAB, ↑DNM1, ↑F2R, ↓F3
SMAD4			-0,600	1,04E-03	↑C13orf15, ↓DAPK1, ↑FOS, ↓GADD45A, ↓ICAM1, ↓ID1, ↓JAG1, ↓JAG2, ↓SERPINE1, ↓TGFB2
CREBBP			-0,532	2,73E-03	↓ADCYAP1 (includes EG:11516), ↑FOS, ↓IL6, ↑KRT14, ↑RGS2 (includes EG:19735), ↑RHOA, ↓SOCS3, ↑SREBF1, ↓TAGLN, ↓TLR2
MKL1			-0,520	2,45E-03	↑FOS, ↓ICAM1, ↓MYLK, ↓TAGLN, ↓TNC
HMGB1			-0,485	5,82E-05	↑AGER, ↓CCL2, ↓ICAM1, ↓IL6, ↓IL8, ↑MOK, ↓RELB, ↓TLR2, ↓TNF
ETS2			-0,478	1,03E-02	↓CD34, ↓FLT1, ↑FOS, ↓ICAM1, ↓PCSK6, ↓TNF

Die mRNA des Transkriptionsfaktors selber muss nicht notwendigerweise hochreguliert sein, auch das Zusammenspiel mit interagierenden Kinasen etc. kann zu deren Aktivität beitragen.

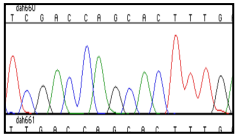
Regulationsrichtung der Zielgene des Transkriptionsfaktors

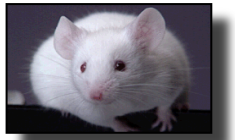
RNA-Seq Praxisteil

Screenshots und Einstellungen

Anwendungsbeispiel: RNA Seq

Genom
& Variom



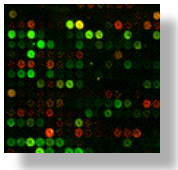


Gen-Knock-out

Gene identifizieren,
Funktion bestimmen!

⇩

Transkriptom



DNA-Chip
EST-Sequenzierung
RNASeq

Wann und wie stark
sind Gene aktiv?

⇩

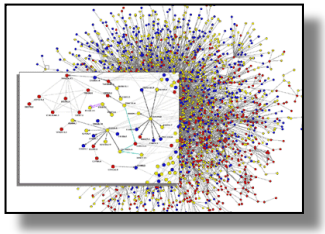
Proteom

Gleiches Genom, unterschiedliches Proteom

25 000 Gene, aber
> 500 000 Proteine?

⇩

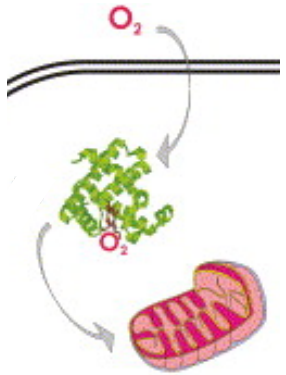
Interaktom

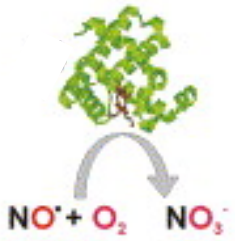


Wie arbeiten die
Proteine zusammen?

Myoglobin in seiner klassischen Rolle

- Exprimiert im Zytoplasma von Herzmuskeln und quergestreiften Muskeln
  - Zuständig für den O_2 Transport
  - Dient als Kurzzeit O_2 Speicher

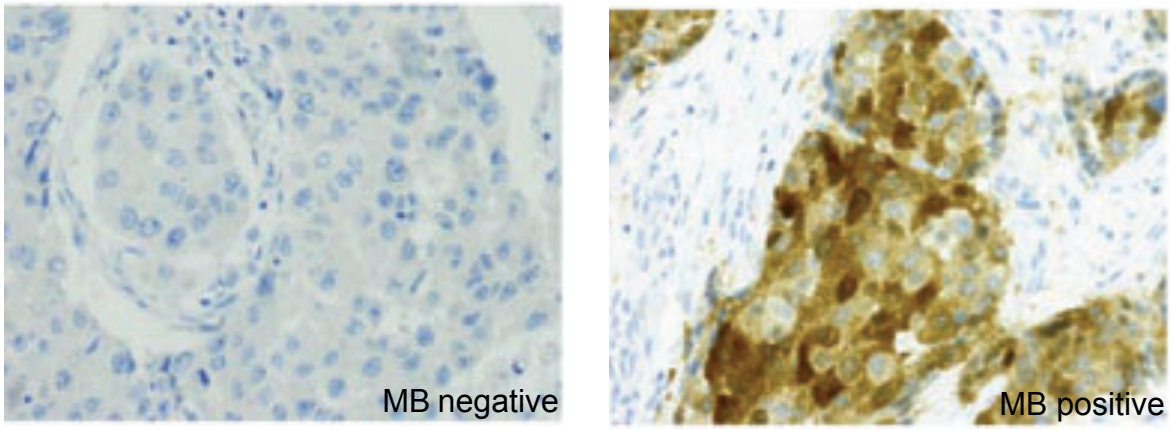




  - Detoxifiziert als Dioxygenase ROS und RNS
  - Kann unter hypoxischen Bedingungen NO produzieren

Myoglobin in Brustkrebs

MB immunostaining on breast tumors

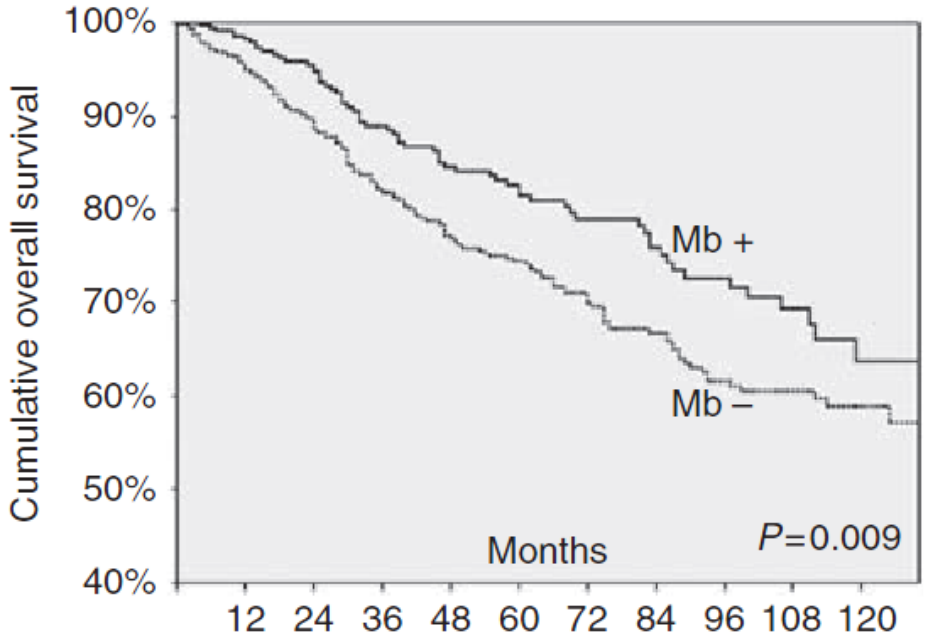


- ~ 40% invasiver luminaler Brustkrebs-Tumore exprimieren MB endogen
- ~ 350 mal mehr MB in Brustkrebs-Tumoren als in gesundem Brustepithel

Gorr et al, 2011

MB in Brustkrebs und die Folgen

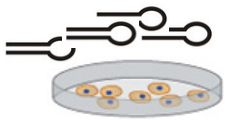
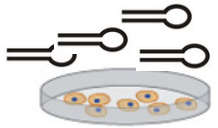
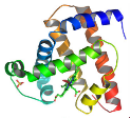
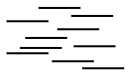
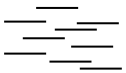
- Kaplan Meier Analyse:
  Survival Function von 917 primären Brustkrebs-Erkrankungen



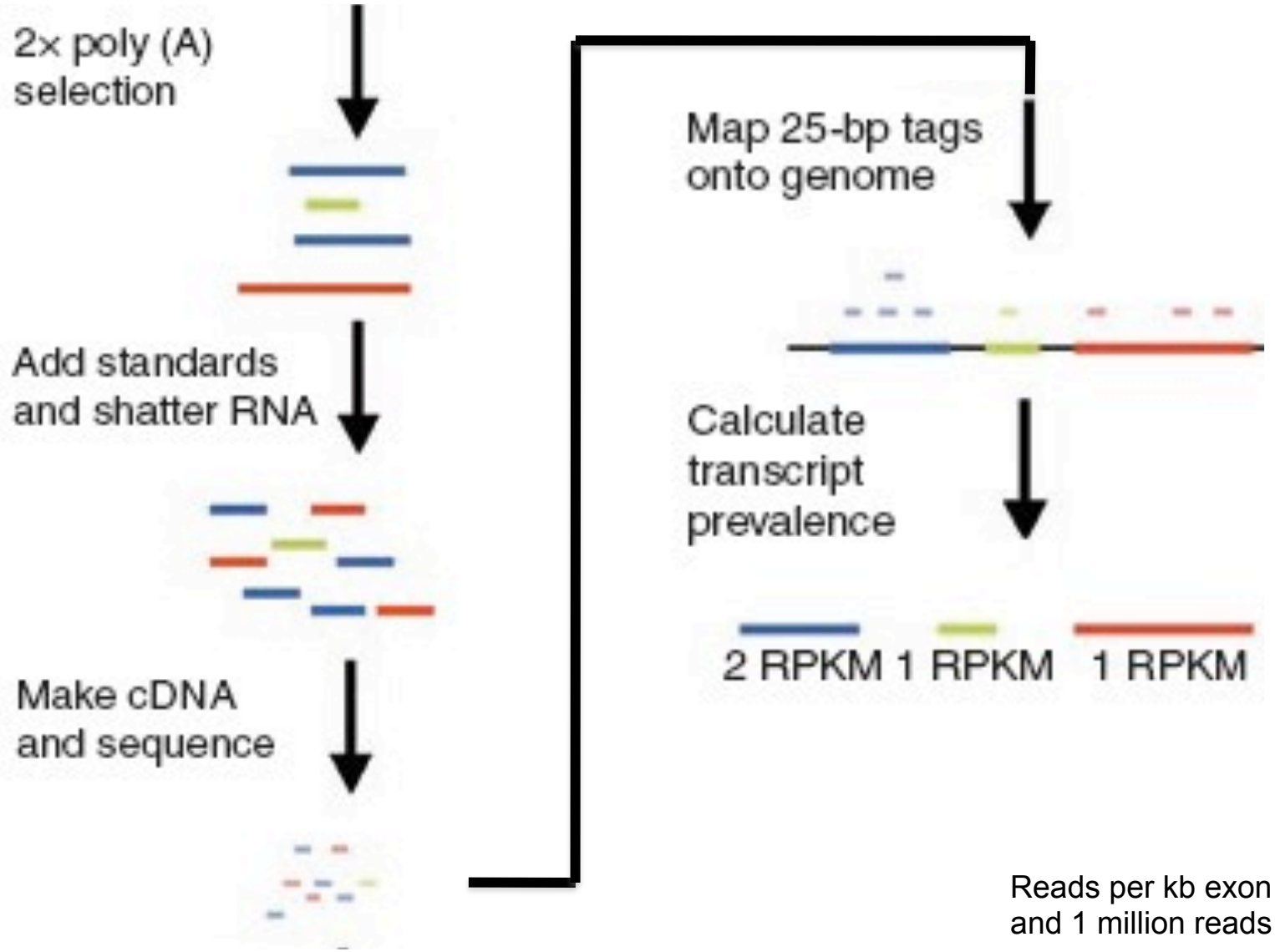
→ MB-Expression korreliert mit einer milderen Prognose für betroffene Patienten

Kristiansen et al. 2010

Analyse von Brustkrebs-transkiptomen mit und ohne MB

Mb Knockdown	Kontrollzellen
• Transfection von MDA-MB468 Zellen mit **siRNA gegen *MB*** 	• Transfection von MDA-MB468 Zellen mit **Kontroll-siRNA** 
• Wachstum; **Herabregulation der MB-mRNAs** 	• Wachstum; **MB-Expression** 
• RNA-Isolation • Preparation von Sequenzier-Libraries • Illumina-Sequenzierung beider Transkriptome 	
→ 34 Mio. Sequenz-Reads 	→ 29 Mio. Sequenz-Reads 

Zur Erinnerung: RNA-Seq Überblick



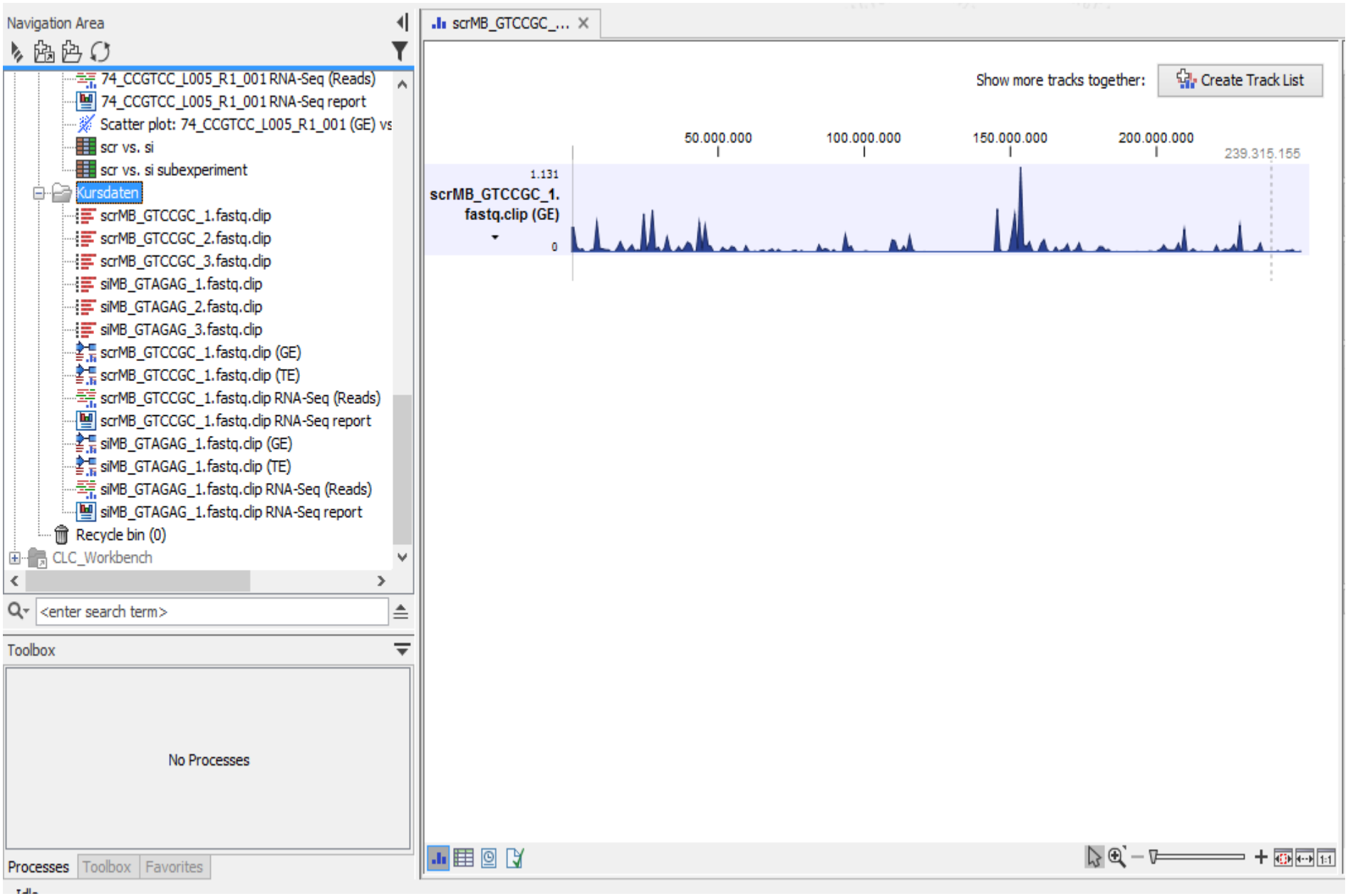
Reads per kb exon model and 1 million reads

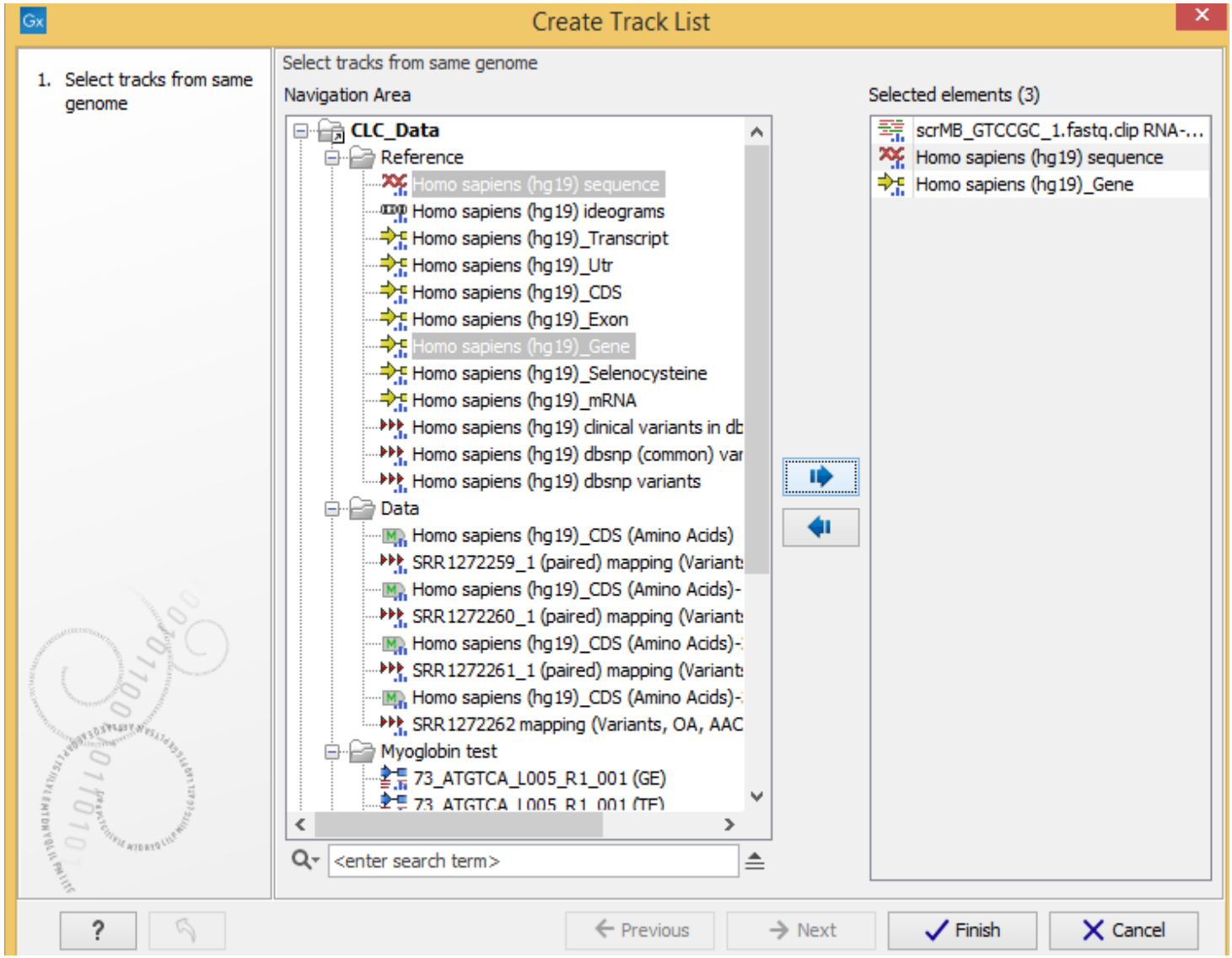
Kurs RNA-Seq

Der Einfluss der Myoglobinexpression auf Krebszellen:

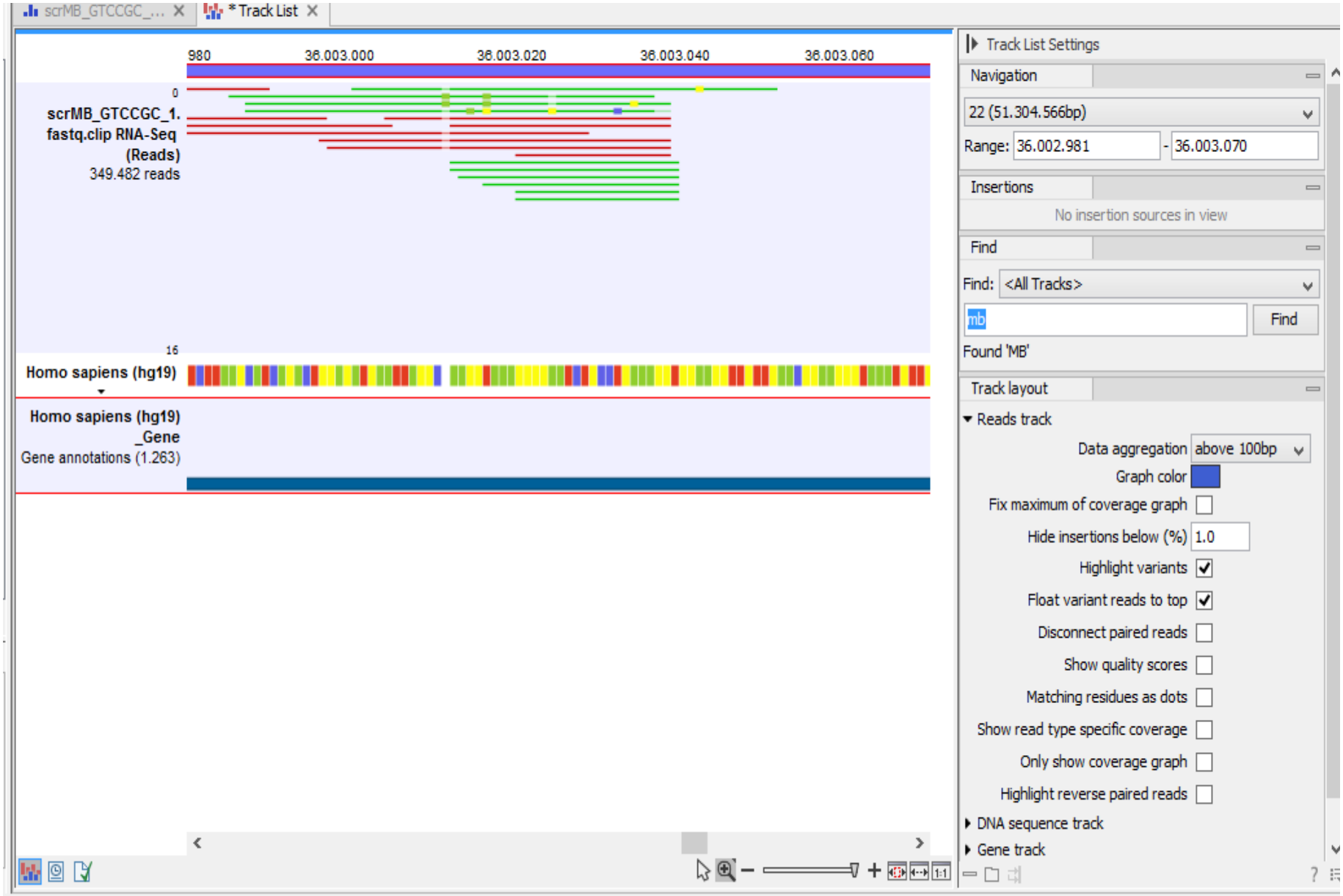
- Qualitätsbewertung und Nachbearbeitung von Illumina-Rohdaten
- Erstellung von Qualitätsstatistiken
- Importieren der Transkriptom-Datensätzen in die CLC-Workbench
- Importieren eines annotierten Hsa-Referenzgenoms
- Mapping der Transkriptomreads an das annotierte Hsa-Genom
- Statistische Analyse, Identifizierung differenziell exprimierter Gene
- Biologische Interpretation der RNA-Seq Analyse

Ansicht unseres Mappings

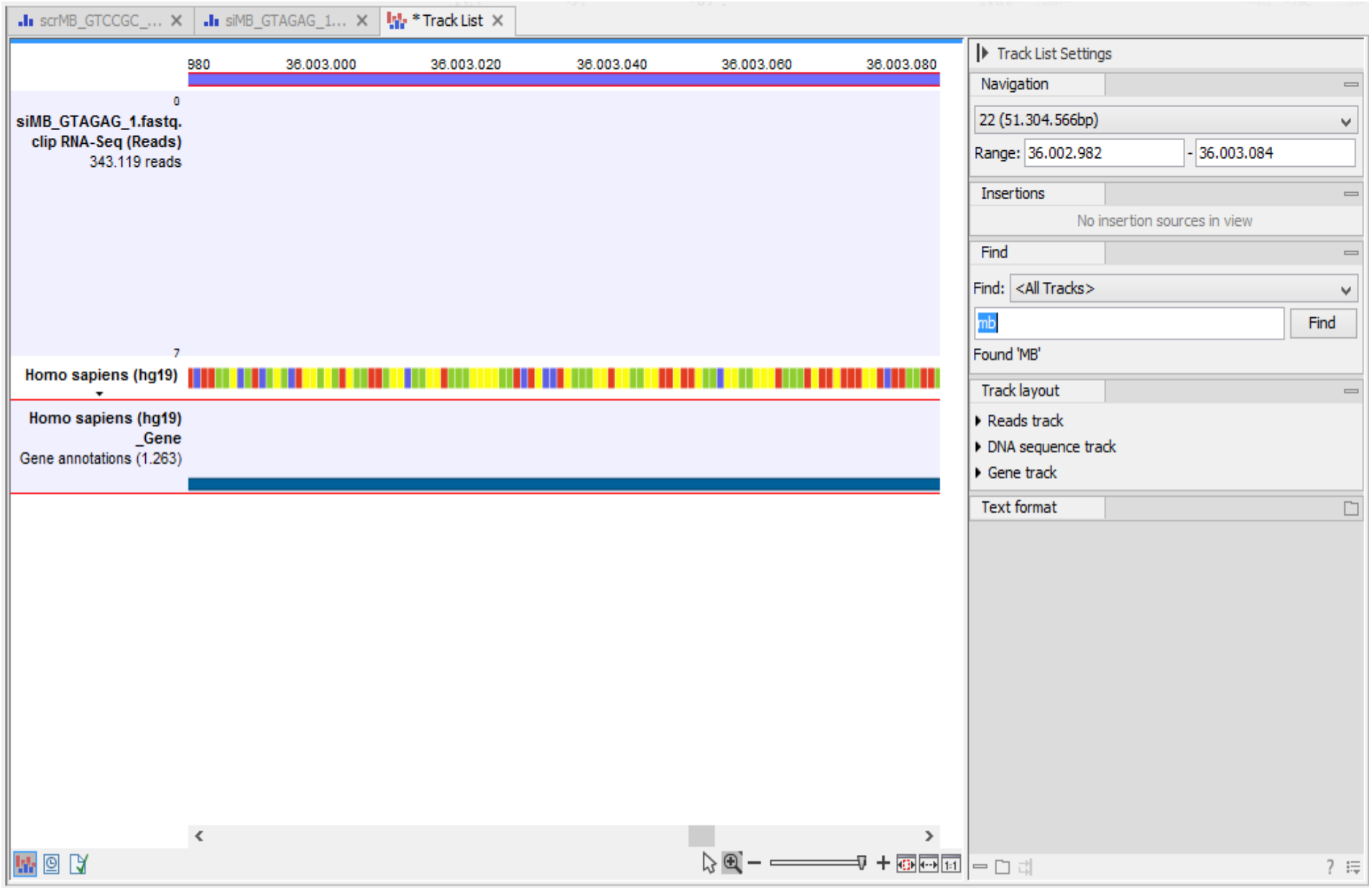


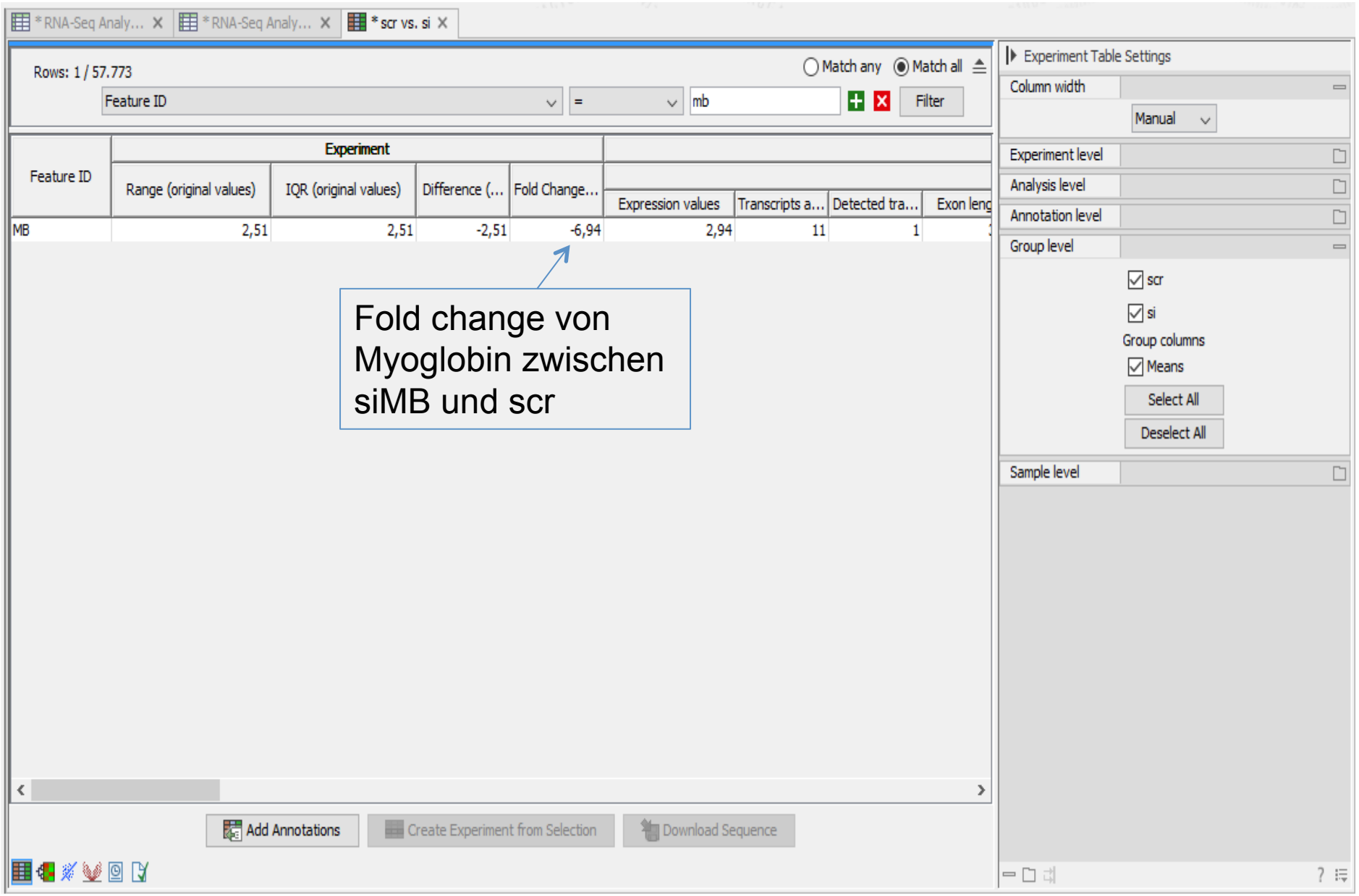

Create Track List
1. Select tracks from same genome
Select tracks from same genome
Navigation Area
CLC_Data
Reference
Homo sapiens (hg19) sequence
Homo sapiens (hg19) ideograms
Homo sapiens (hg19)_Transcript
Homo sapiens (hg19)_Utr
Homo sapiens (hg19)_CDS
Homo sapiens (hg19)_Exon
Homo sapiens (hg19)_Gene
Homo sapiens (hg19)_Selenocysteine
Homo sapiens (hg19)_mRNA
Homo sapiens (hg19) clinical variants in db
Homo sapiens (hg19) dbsnp (common) var
Homo sapiens (hg19) dbsnp variants
Data
Homo sapiens (hg19)_CDS (Amino Acids)
SRR1272259_1 (paired) mapping (Variants
Homo sapiens (hg19)_CDS (Amino Acids)-
SRR1272260_1 (paired) mapping (Variants
Homo sapiens (hg19)_CDS (Amino Acids)-
SRR1272261_1 (paired) mapping (Variants
Homo sapiens (hg19)_CDS (Amino Acids)-
SRR1272262 mapping (Variants, OA, AAC
Myoglobin test
73_ATGTCA_L005_R1_001 (GE)
73_ATGTCA_L005_R1_001 (TE)
<enter search term>
Selected elements (3)
scrMB_GTCCGC_1.fastq.clip RNA-...
Homo sapiens (hg19) sequence
Homo sapiens (hg19)_Gene
?
Previous
Next
Finish
Cancel

Scr-Myoglobinausschnitt von Mapping, Referenzgenom und Genannotationen

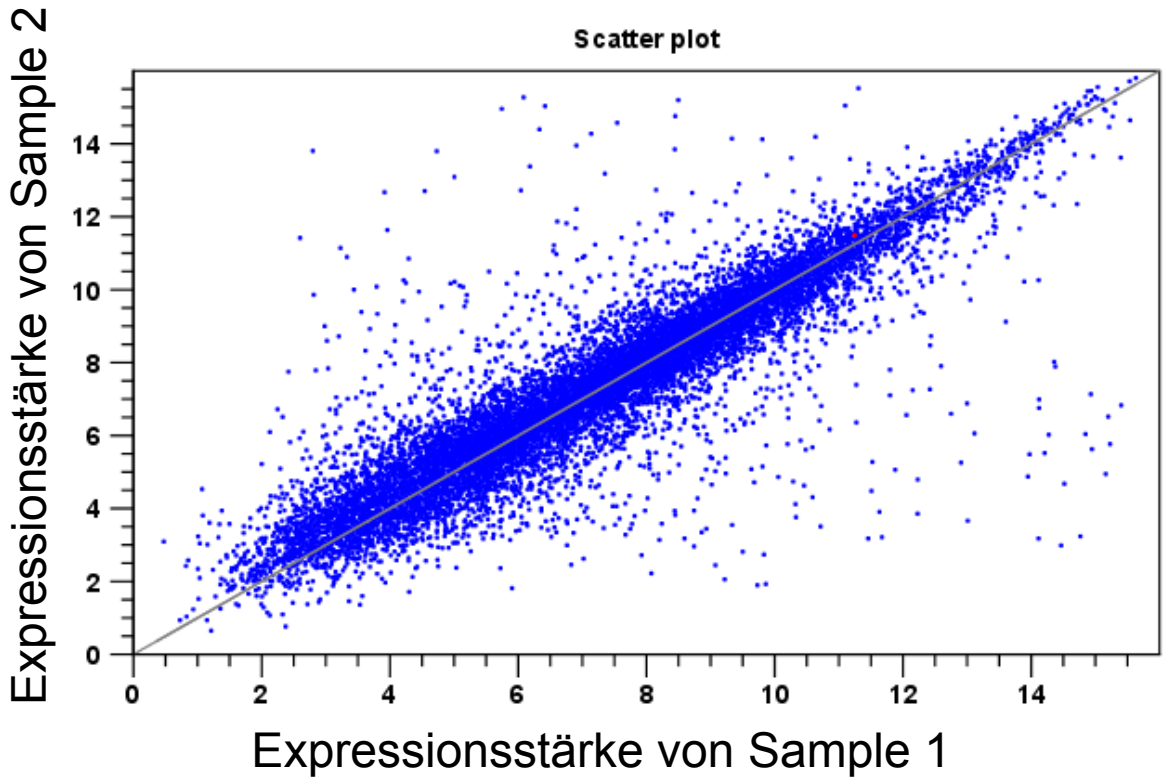


siMB- Myoglobinauschnitt von Mapping, Referenzgenom und Genannotationen

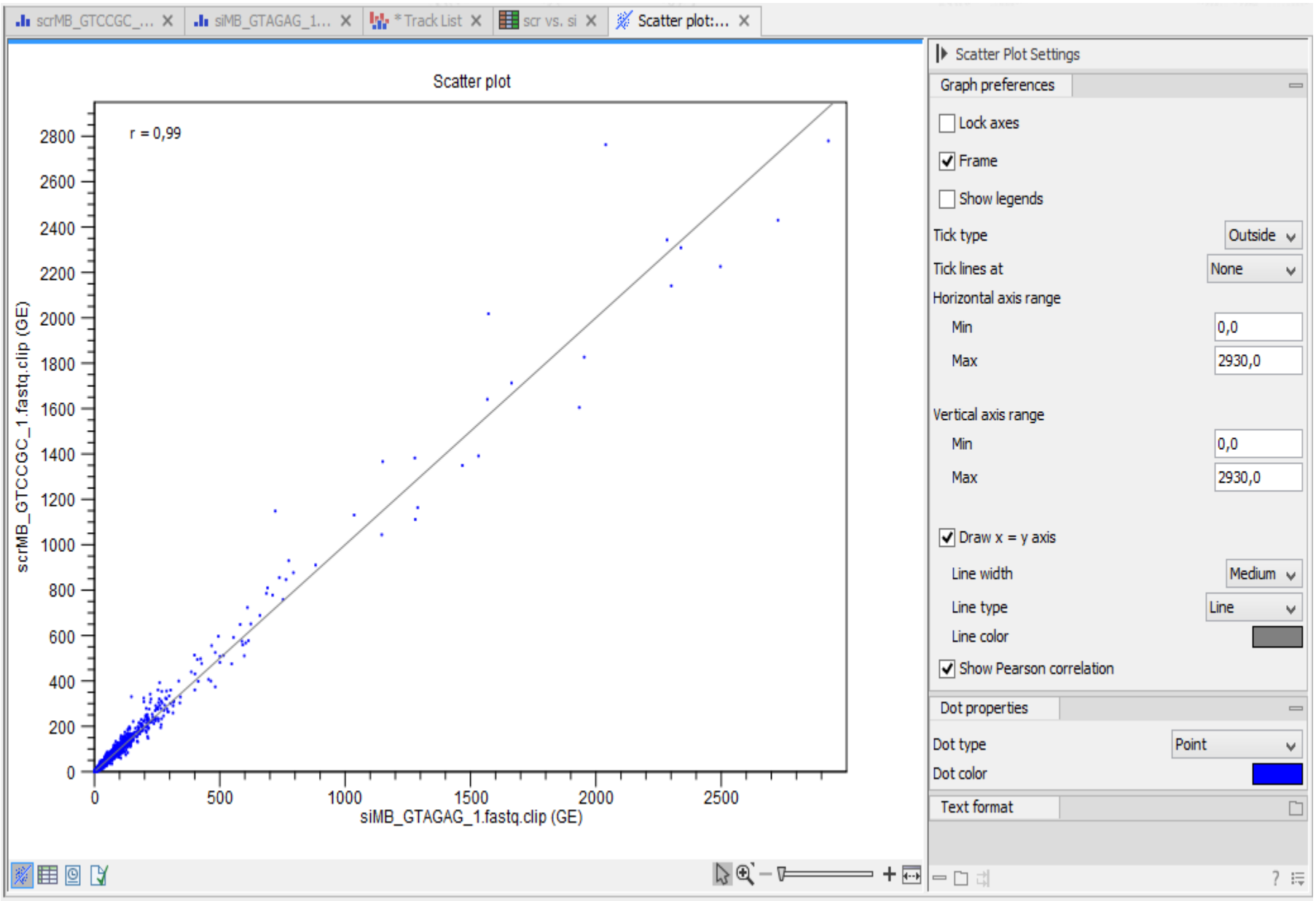



* RNA-Seq Analy...
* RNA-Seq Analy...
* scr vs. si
Rows: 1 / 57.773
Match any
Match all
Feature ID
=
mb
Filter
Feature ID
Experiment
Range (original values)
IQR (original values)
Difference (...
Fold Change...
Expression values
Transcripts a...
Detected tra...
Exon leng
MB
2,51
2,51
-2,51
-6,94
2,94
11
1
Fold change von Myoglobin zwischen siMB und scr
Add Annotations
Create Experiment from Selection
Download Sequence
Experiment Table Settings
Column width
Manual
Experiment level
Analysis level
Annotation level
Group level
scr
si
Group columns
Means
Select All
Deselect All
Sample level

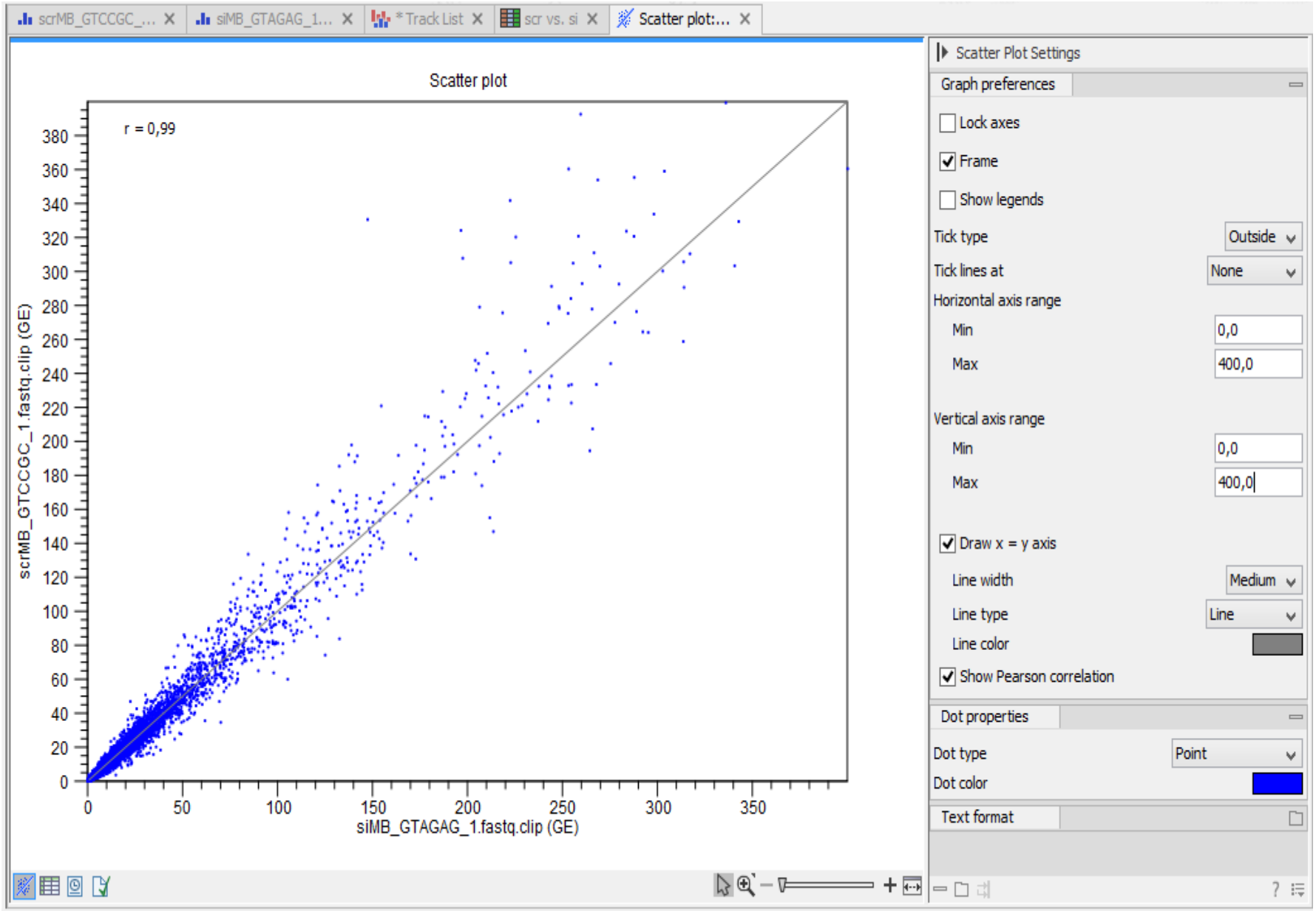
Statistische Auswertung: Scatter Plot



Unser Scatter Plot



Unser Scatter Plot



Statistik zur Expressionsstärke: Z-Test

Kal et al.'s test (Z-test): Vergleich einzelner Proben gegeneinander (n=1)

- **Basierend auf der "**Approximation of the binomial distribution by the normal distribution" [Kal et al., 1999]
- Proportions-basiert statt "raw count"-basiert, darum auch geeignet, wenn ein Sample insgesamt viel höhere "Sum-of-counts" hat
- 'Proportions difference' für ganze Gruppen berechnet
- Zweiseitiger 'P-value', optional mit FDR und Bonferroni-Correction

„Korrektur“ der Irrtumswahrscheinlichkeiten

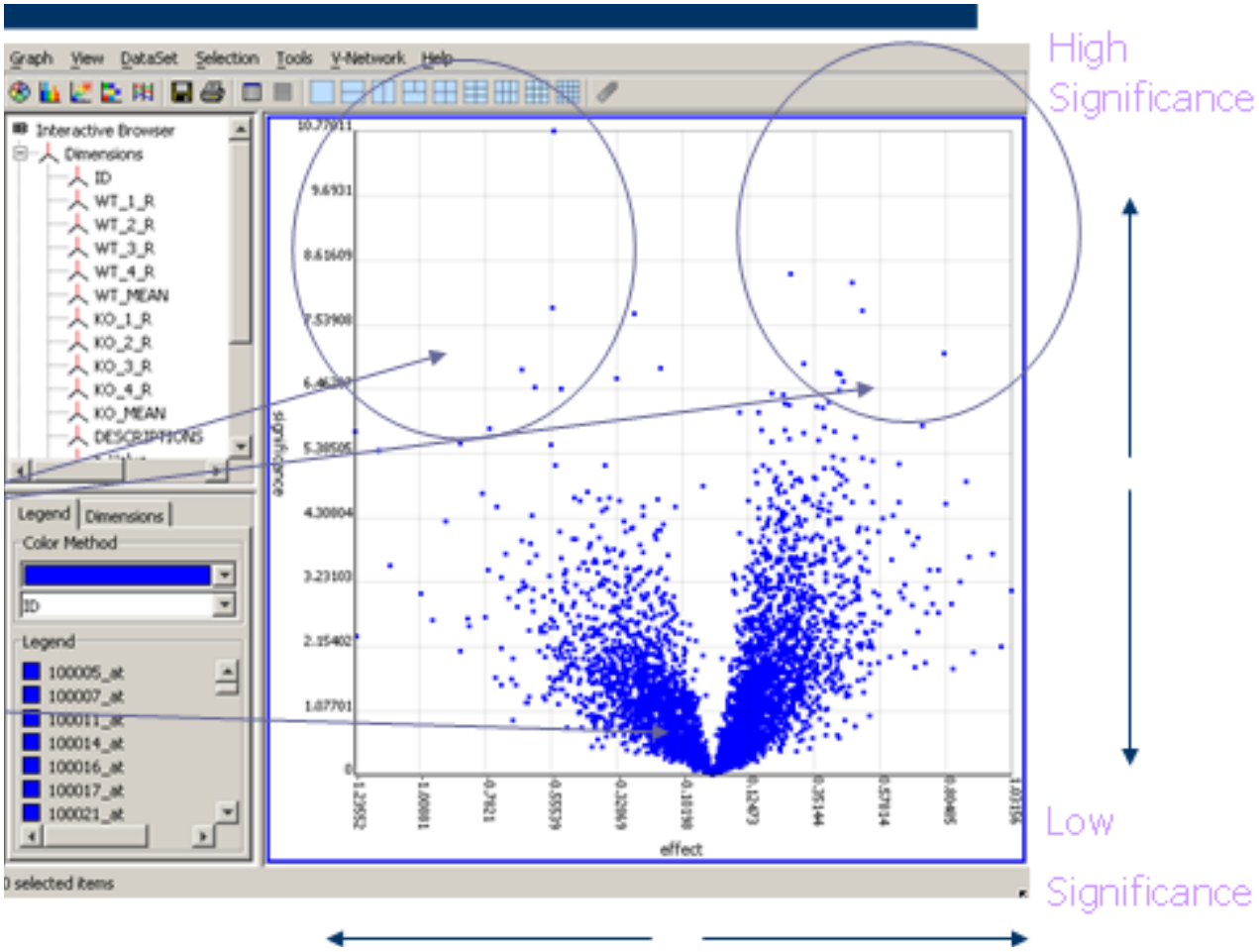
Bonferroni corrected:
The Bonferroni corrected p-values handle the multiple testing problem by controlling the 'family-wise error rate': the probability of making at least one false positive call. They are calculated by multiplying the original p-values by the number of tests performed. The probability of having at least one false positive among the set of features with Bonferroni corrected p-values below 0.05, is less than 5%. The Bonferroni correction is conservative: there may be many genes that are differentially expressed among the genes with Bonferroni corrected p-values above 0.05, that will be missed if this correction is applied.

FDR (False discovery rate) corrected:
Instead of controlling the family-wise error rate we can control the false discovery rate: FDR. The false discovery rate is the proportion of false positives among all those declared positive. We expect 5 % of the features with FDR corrected p-values below 0.05 to be false positive [Benjamini and Hochberg, 1995].

Statistische Auswertung: Vulcano Plot

Hohe Differenz der Expressionsstärken + hohe Signifikanz

Ähnlich stark exprimierte Gene/ Transkripte + niedrige Signifikanz



Unser Vulcano Plot