Genexpressionsanalyse Projekt 2 (Emmelie E. Tessema)

Aufgabe 0

Zusammenfassend beziehen sich die Daten auf die Genexpression in Saccharomyces cerevisiae unter verschiedenen Bedingungen, und die Studie vergleicht die Ergebnisse der RNA-Seq- und Mikroarray-Analysen, wobei verschiedene analytische Schritte und Methoden im Prozess bewertet werden.

Überschrift	Beschreibung
Art der Daten	• Daten: transkriptomische Daten → RNA • *Generierung der Daten*: RNA-Seq-Analyse und Mikroarray-Analyse • *Transkriptom-Analyse:* Enthält den vollständigen Satz von RNA-Transkripten, die zu einem bestimmten Zeitpunkt vom Genom produziert werden.
Organismus	• *Untersuchter Organismus:* Saccharomyces cerevisiae, Stamm CEN.PK 113-7D • Es wird die *Saccharomyces cerevisiae* als Bäckerhefe und Modellorganismus in der biologischen Forschung verwendet

Vergleich	• Studie vergleicht Ergebnisse zwischen RNA-Seq-Analyse und Mikroarray-Analyse
	• Für die Arbeit wuchs der Saccharomyces cerevisiae Stamm unter zwei verschiedenen Bedingungen (Batch und Chemostat)
	  ○ Chemostat: kontinuierliche Zufuhr von einem Nährmedium und Entfernen von verbrauchtem Medium
	  ○ Batch: diskontinuierliche Kultur. Medium bleibt gleich und wird weder ergänzt noch aufgefüllt oder ausgewechselt
	• Es werden die analytische Schritte bei RNA-Seq-Datenanalyse mit Illumina-Plattform bewertet und ein Vergleich basierend auf Affymetrix-Mikroarrays durchgeführt

Analytische Schritte	• Studie zur *Bewertung des Einflusses genetischer Variationen* auf die Schätzung der Genexpression • Verwendung von drei verschiedenen Read-Mapping-Alignern (Gsnap, Stampy und TopHat) auf dem S288c-Genom • Untersuchung der Fähigkeiten von fünf verschiedenen statistischen Methoden zur Erkennung differentieller Genexpression
Ergebnisse	• Hohe Reproduzierbarkeit zwischen biologischen Replikaten • Hohe Konsistenz zwischen den beiden Plattformen für die Genexpressionsanalyse (Korrelation ≥ 0,91) • Gute Übereinstimmung bei der Identifizierung differentieller Genexpression durch verschiedene statistische Methoden
Fazit	• Die Studie vergleicht RNA-Seq und Mikroarrays für die Genexpressionsanalyse. • Es wird untersucht, wie verschiedene Schritte die Analyse von RNA-Seq-Daten beeinflussen.

Aufgabe 1

```
fastq -dump--gzip --split-files SRR4535xx
```

- fastq-dump: SRA-Daten (Sequence Read Archive) in das FASTQ-Format zu konvertieren
- --gzip: komprimiert die Ausgabedateien im GZIP-Format
- --split-files: teilt die Reads in separate Dateien auf
    - i.d.R. Forward- & Reverse-Reads

Befehl: *fastqc *.fastq.gz*

- Fastqc: Ruft das Programm FastQc auf
- *.fastq.gz: Wendet das Programm auch alle Files an

Aufgabe 2

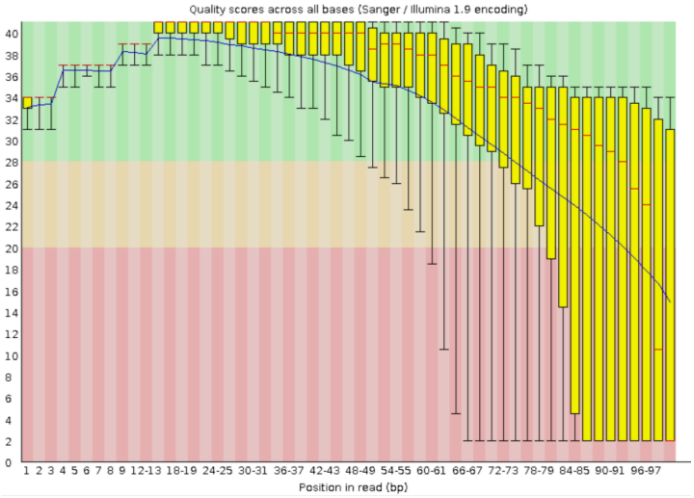
Probe SRR453570_2

Basic Statistics

Measure	Value
Filename	SRR453570_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6745975
Total Bases	681.3 Mbp
Sequences flagged as poor quality	0
Sequence length	101
%GC	42

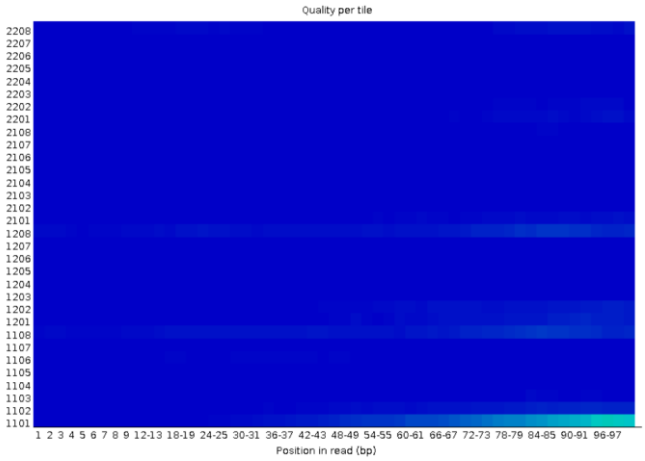
- Hier sind die allgemeinen Daten über die Sequenzierte RNA zusammengefasst
- Die Quality-Score-Kodierung basiert auf Sanger/Illumina 1.9
- Insgesamt wurden 681.3 Mbp sequenziert
- Die Read-Länge beträgt zudem 101 bp

Per base sequence quality



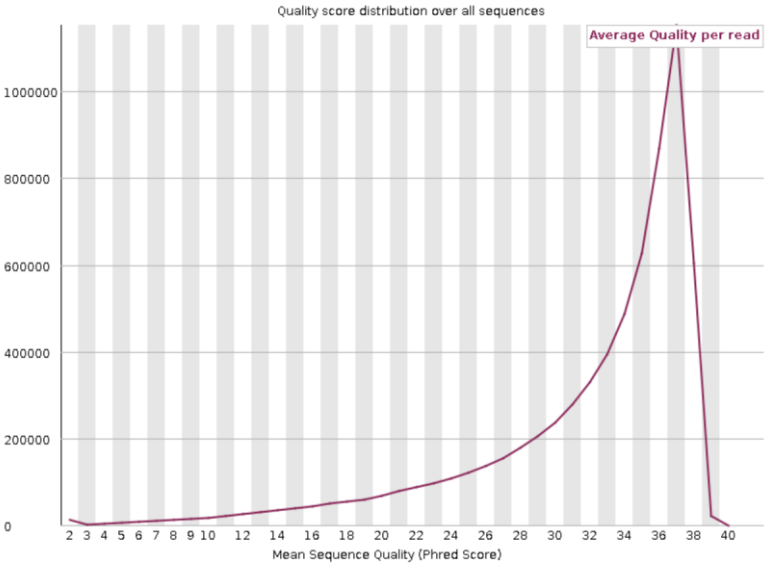
- Mit zunehmender Read-Länge ist eine deutliche Abnahme der Qualität der sequenzierten Basen zu erkennen
- Es erfolgt im Bereich der ersten Basenpaare (ca.1bp-12bp) ein Anstieg des Quality Scores
- Im Bereich von ca. 13-19bp sind die Basen, mit den höchsten Quality Scores
- Danach folgt eine deutliche Abnahme der Qualität, was man anhand der abnehmenden Quality Scores erkennt

Per title sequence quality



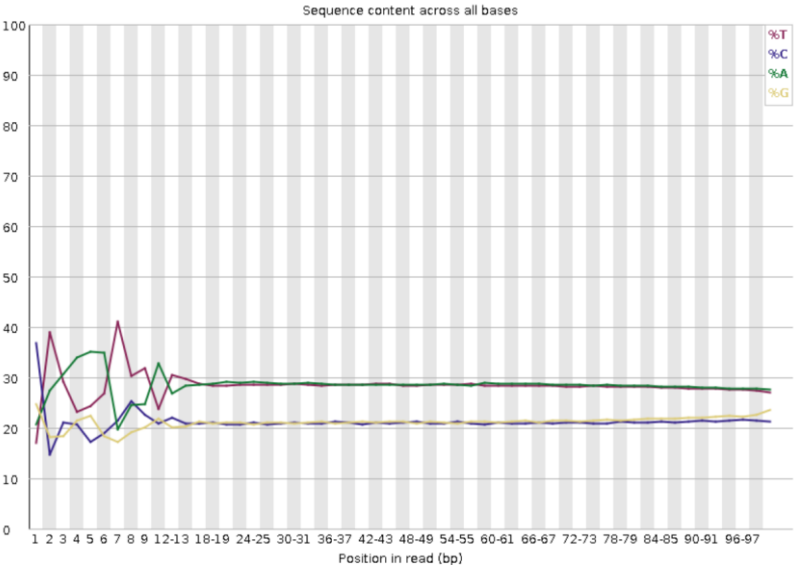
- Es ist erkennbar, dass nur minimale Fehler in der Sequenzierung von Illumina selbst vorlagen
- Die Signale sind alle überwiegend eindeutig
- Mit Zunehmender Readlänge, ist das Signal nur leicht qualitativ schlechter

Per sequence quality scores



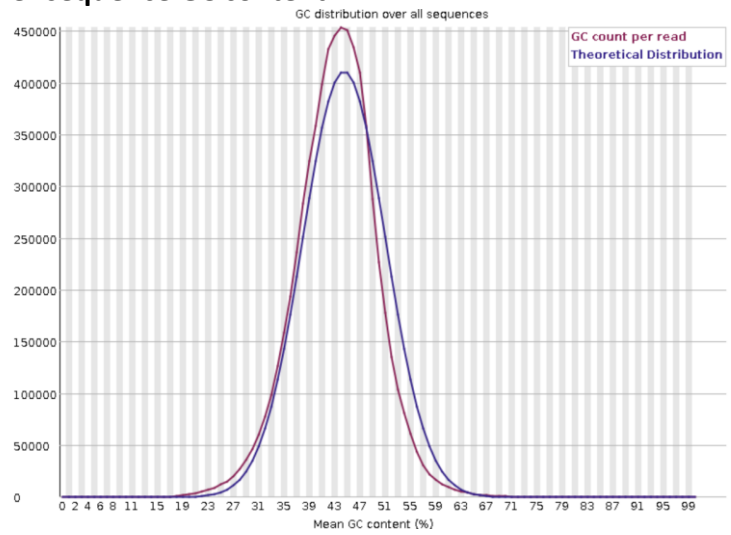
- Der größte Teil der gemessenen Basen haben durchschnittlich gesehen einen hohen Quality Score
- Es sind keine weiteren Peaks im unteren Bereich erkennbar
- Der Anstieg beginnt jedoch bereits im niedrigeren Bereich des Phred Scores (bei ca. 12- 14)
- Ein erkennbarer Anteil der Reads besitzt einen niedrigen Phred Score

Per base sequence content



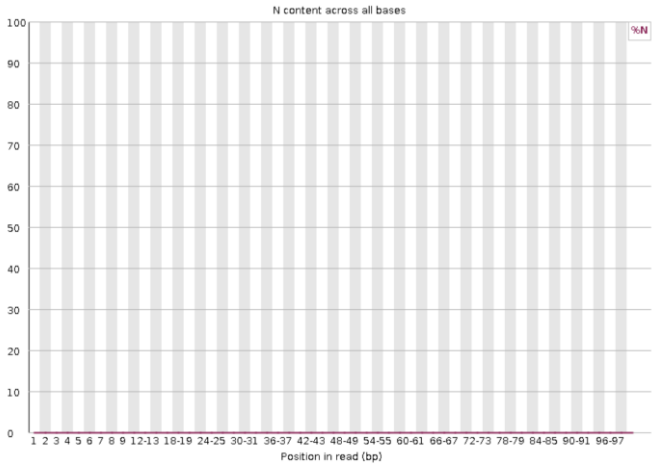
- Hier ist anfangs eine sehr schlechte Qualität (Bereich 1 bis 12bp) und danach eine relativ gute Qualität der einzelnen Basen pro Position über alle Reads zu erkennen
- In der ersten Phase sind hohe Schwankungen jedoch zu erwarten, da der Sequenzierer sich da noch in einer Art „Findungsphase“ befindet

Per sequence GC content



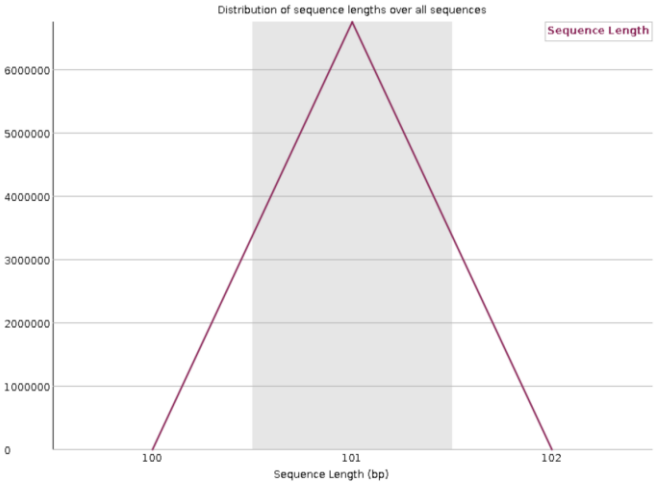
- An sich ist eine Überlappung des Peaks bezüglich des relativen durchschnittlichen GC-Gehalts im Bereich des mittleren GC-Gehalts erkennbar
- Die Daten sind zudem auch normalverteilt, was zu einer guten Bewertung der Qualität der Daten beiträgt, da der GC-Gehalt ausgewogen vorliegt
- Bei den Gemessenen Reads ist der GC-Gehalt aber bei deutlich mehr Reads aufgetreten als in dem theoretischen Vergleichsmessung
- Positiv zu bewerten ist ebenfalls, dass es keine weiteren Peaks gibt

Per base N content



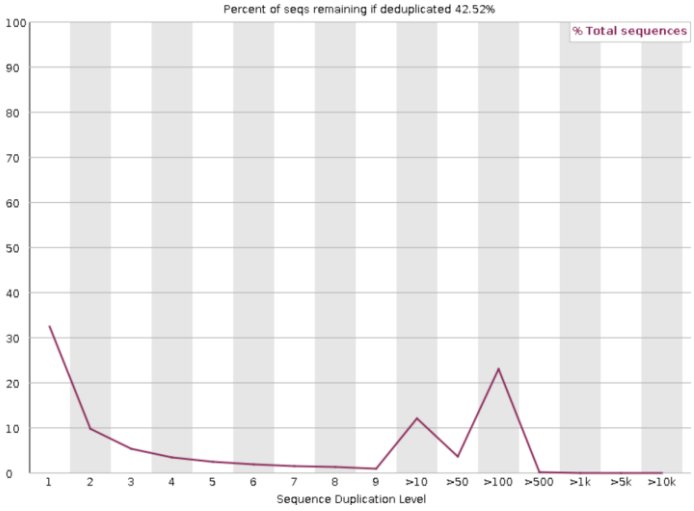
- Keine uneindeutige-Determinierung der gemessenen Basen graphisch erkennbar
- Daraus lässt sich schließen, dass die gemessenen Basen eindeutig erkannt worden sind

Sequence Length Distribution



- Die Länge der Sequenzen sind überwiegend gleich, da ein eindeutiger Peak erkennbar ist
- Der Peak liegt ungefähr bei ca. 101bp
    - Im Intervall von 100 und 102 bp
- Es ist erkennbar, dass Fragmente nur von gleicher Länge generiert worden sind
- Dies trägt zu einer positiven Bewertung der Datenqualität bei

Sequence Duplication Levels



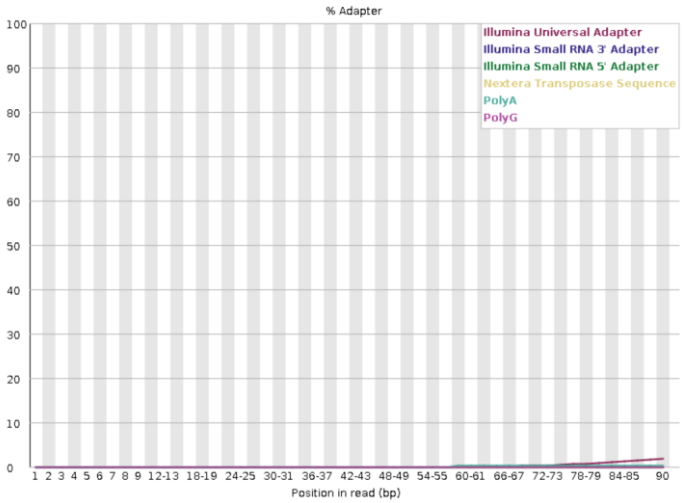
- Ein Teil der Reads, haben eine erkennbare hohe Duplikationsrate
- Die Peaks liegen im Bereich von 9 bis >500 bezüglich des Duplikationslevels
- Es sind 2 große Peaks erkennbar
    - Der größere befindet sich auf einem Duplikationslevel von ca. >100
    - Der kleinere Peak befindet sich bei ca. >10
- Da es sich um Transkriptionsdaten handelt, waren jedoch auch Duplikate zu erwarten

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCG	11924	0.1767572515462924	Illumina Single End PCR Primer 1 (100% over 50bp)

- Da nur eine Sequenz gelistet ist, spricht dies für eine relativ gute Qualität der gemessenen Daten

Adapter Content



- Der Illumina Universal Adapter zeigt einen leichten Anstieg im Bereich von ca. ab 78-79 bp
- Es scheinen somit viele Sequenzen am Ende einen leichten Anteil der Adaptersequenz zu enthalten

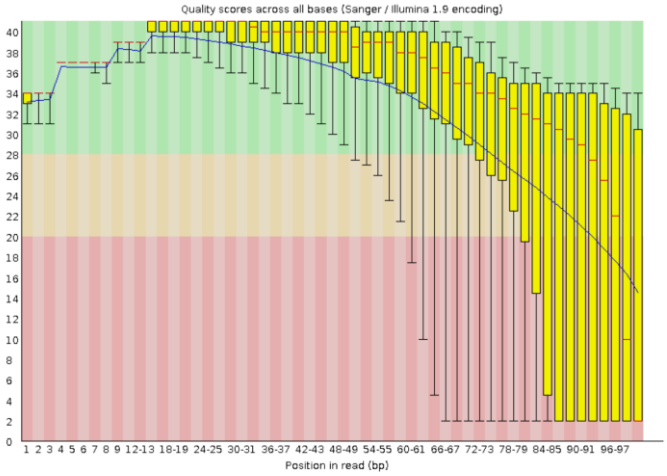
Probe SRR453571_2

Basic Statistics

Measure	Value
Filename	SRR453571_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6163396
Total Bases	622.5 Mbp
Sequences flagged as poor quality	0
Sequence length	101
%GC	41

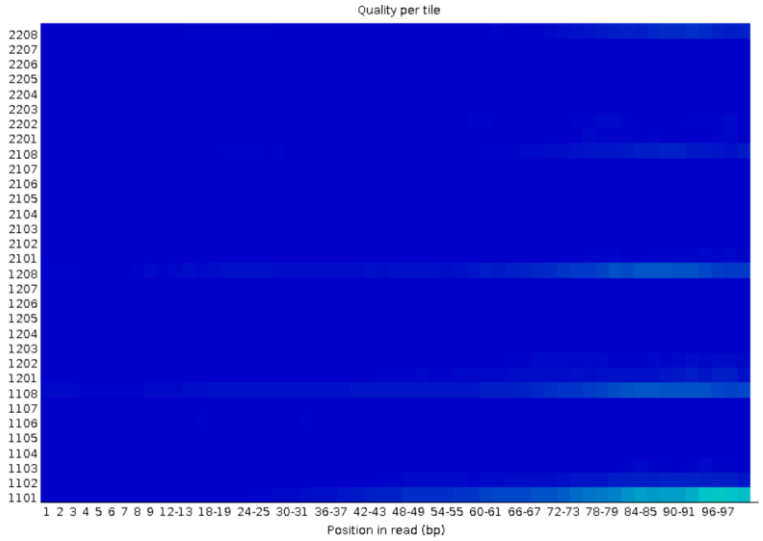
- Hier sind die allgemeinen Daten über die Sequenzierte RNA zusammengefasst
- Die Quality-Score-Kodierung basiert auf Sanger/Illumina 1.9
- Insgesamt wurden 622.5 Mbp sequenziert
- Die Read-Länge beträgt zudem 101 bp

Per base sequence quality



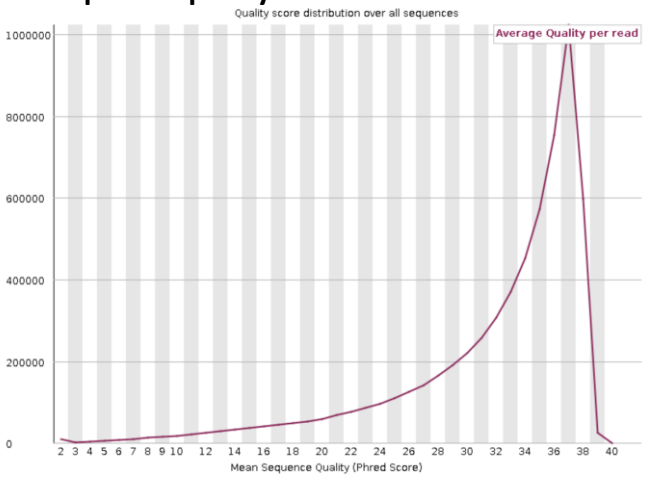
- Mit zunehmender Read-Länge ist auch hier eine deutliche Abnahme der Qualität der sequenzierten Basen zu erkennen
- Es erfolgt im Bereich der ersten Basenpaare (ca.1bp-12bp) ein Anstieg des Quality Scores
- Im Bereich von ca. 13-19bp sind die Basen, mit den höchsten Quality Scores
- Danach folgt eine starke Abnahme der Qualität, was man anhand der kleiner werdenden Quality Scores erkennt

Per title sequence quality



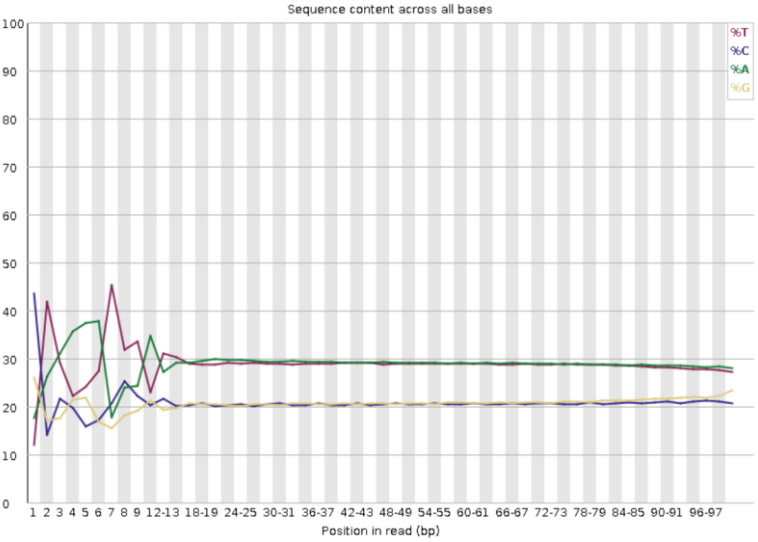
- Es ist erkennbar, dass nur minimale Fehler in der Sequenzierung von Illumina selbst vorlagen
- Die Signale sind alle überwiegend eindeutig
- Mit Zunehmender Readlänge, ist das Signal nur leicht qualitativ schlechter

Per sequence quality scores



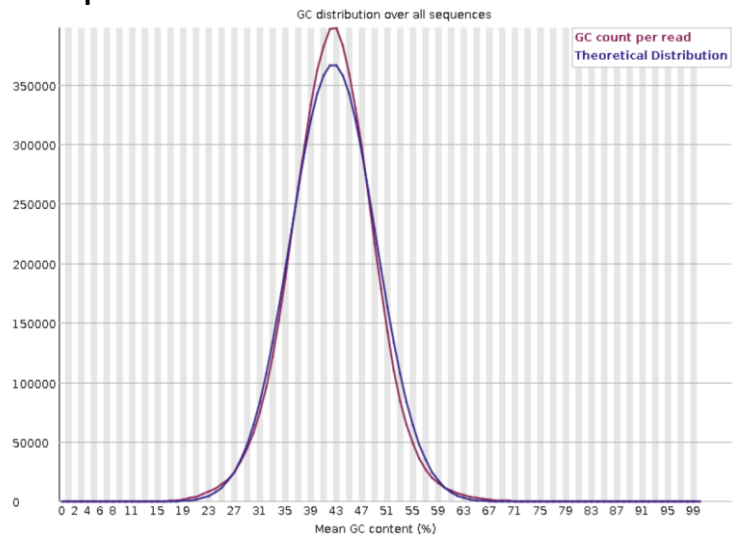
- Der größte Teil der gemessenen Basen haben durchschnittlich gesehen einen hohen Quality Score
- Es sind keine weiteren Peaks im unteren Bereich erkennbar
- Der Anstieg beginnt jedoch bereits im niedrigeren Bereich des Phred Scores (bei ca. 12- 14)
- Ein erkennbarer Anteil der Reads besitzt einen niedrigen Phred Score

Per base sequence content



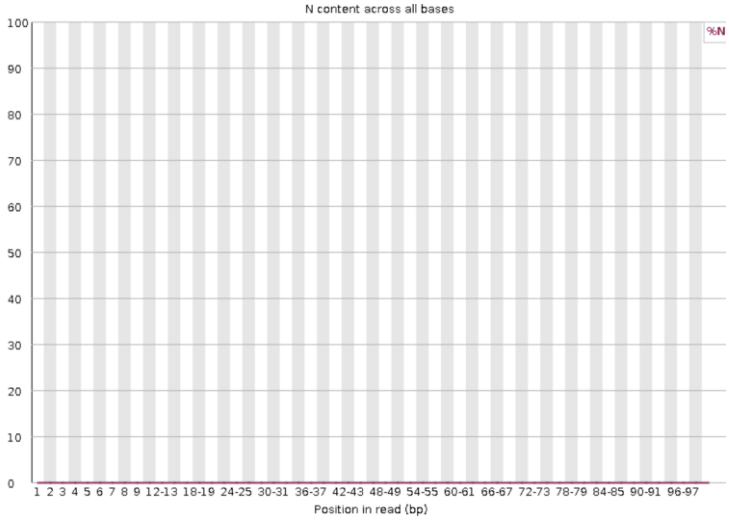
- Hier ist anfangs eine sehr schlechte Qualität (Bereich 1 bis 12bp) und danach eine relativ gute Qualität der einzelnen Basen pro Position über alle Reads zu erkennen
- In der ersten Phase sind hohe Schwankungen jedoch zu erwarten, da der Sequenzierer sich da noch in einer Art „Findungsphase“ befindet

Per sequence GC content



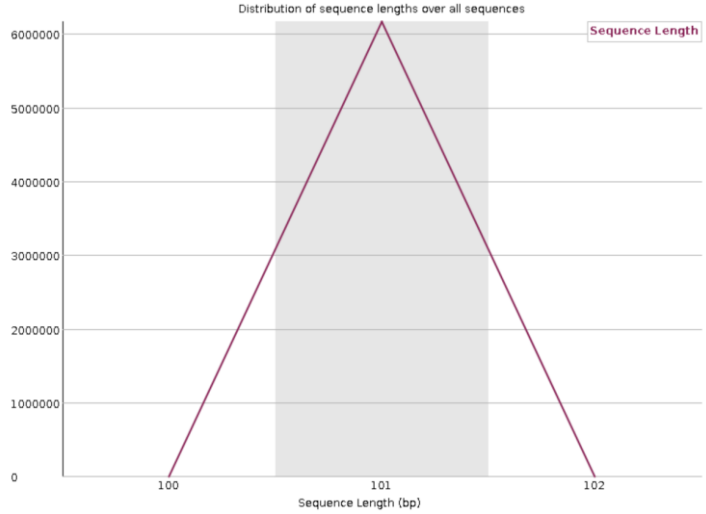
- An sich ist eine Überlappung des Peaks bezüglich des relativen durchschnittlichen GC-Gehalts im Bereich des mittleren GC-Gehalts erkennbar
- Die Daten sind zudem auch normalverteilt, was zu einer guten Bewertung der Qualität der Daten beiträgt, da der GC-Gehalt ausgewogen vorliegt
- Bei den Gemessenen Reads ist der GC-Gehalt etwas mehr Reads, als in dem theoretischen Vergleichsmessung
- Positiv zu bewerten ist ebenfalls, dass es keine weiteren Peaks gibt

Per base N content



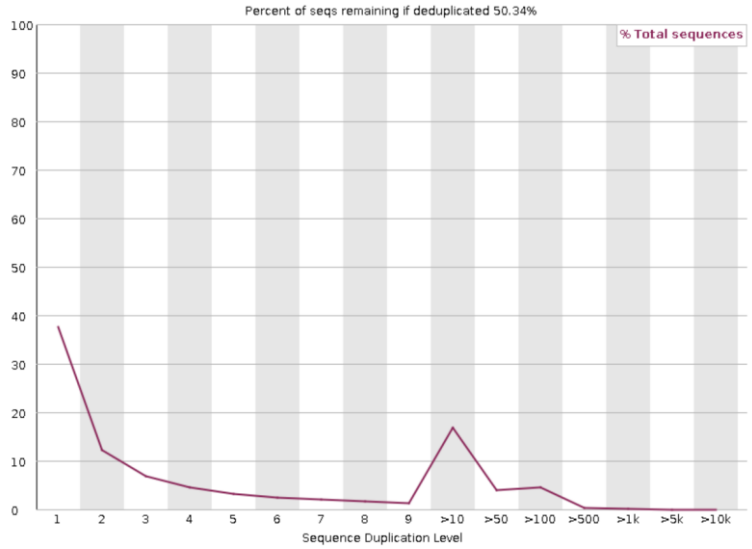
- Keine uneindeutige-Determinierung der gemessenen Basen graphisch erkennbar
- Daraus lässt sich schließen, dass die gemessenen Basen eindeutig erkannt worden sind

Sequence Length Distribution



- Die Länge der Sequenzen sind überwiegend gleich, da ein eindeutiger Peak erkennbar ist
- Der Peak liegt ungefähr bei ca. 101bp
    - Im Intervall von 100 und 102 bp
- Es ist erkennbar, dass Fragmente nur von gleicher Länge generiert worden sind
- Dies trägt zu einer positiven Bewertung der Datenqualität bei

Sequence Duplication Levels

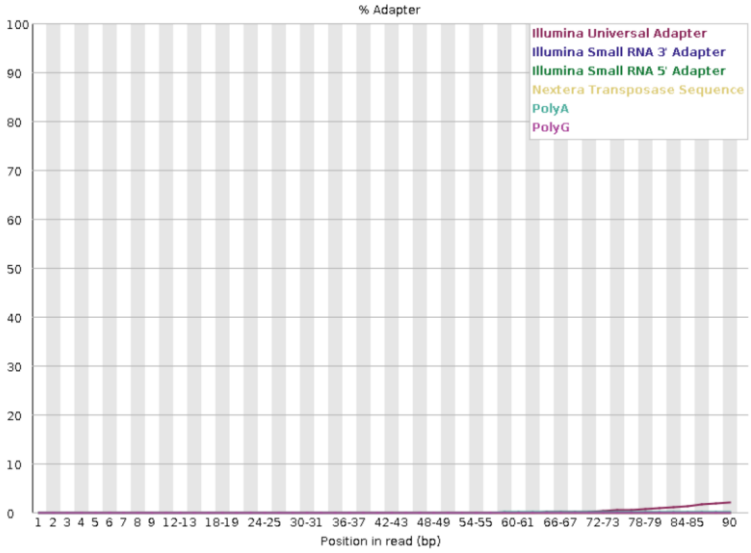


- Ein Teil der Reads, haben eine erkennbare hohe Duplikationsrate
- Es gibt einen großen Peak bei ca. >10 Duplikationslevel und ein etwas kleinerer Peak bei ca. >100 Dublikationslevel
- Die Peaks liegen im Bereich von 9 bis >500 bezüglich des Duplikation Levels
- Da es sich um Transkriptionsdaten handelt, waren jedoch auch Duplikate zu erwarten

Overrepresented sequences

No overrepresented sequences

Adapter Content



- Da keine überrepräsentierte Reads erkannt worden sind, spricht dies zusätzlich von einer guten Qualität der Daten

- Der Illumina Universal Adapter zeigt einen leichten Anstieg im Bereich von ca. ab 78-79 bp
- Es scheinen somit viele Sequenzen am Ende einen leichten Anteil der Adaptersequenz zu enthalten

Aufgabe 3

Cutadapt

- Das Programm ist in der Lage die Adaptersequenzen aus den Sequenzen zu entfernen

Trim -Galore

- Dies ist eine “Wrapper-Software“, welche verschiedene Werkzeuge für das Trimmen von den Reads kombiniert
- Einschließlich Cutadapt
- Es erkennt die Adaptersequenzen und führt das Trimmen basierend auf den Qualitätsschwellenwerten durch
- Zudem ist das Tool in der Lage Reads von zu niedriger Qualität und nicht ausreichende Länge zu entfernen

Durchgeführter Befehl:

```
(cutadaptenv) →  ~ trim_galore -j 4 --paired
```

- Trim_galore: ruft das Program Trim Galor auf
- -j 4: gibt an, dass der Befehl 4 Prozesse parallel ausführen soll
  - dies kann die Geschwindigkeit der Datenverarbeitung verbessern
- --paired: diese Option gibt an, dass es sich um gepaarte (Paired-End) Reads handelt
  - Sie enthalten somit sowohl den Vorwärts- als auch den Rückwärtsstrang

Warum ist es wichtig, dass die gepaarten Reads zusammen beschnitten werden?

- es ist wichtig, da sie aus derselben DNA-Fragmentsequenz stammen und daher miteinander verbunden sind
- es dient dem Sicherstellen von konsistenten beschnittenen Reads
  - sonst können Unterschiede in der Datenqualität auftreten
- zudem kann man durch das gleiche Trimming der Readds ein genaueres Mapping der Reads auf das Referenzgenom ausführen
- Zudem können durch die Verbindung von Informationen aus beiden Read-Paaren, Fehler in der Sequenzierung besser erkannt werden

Vergleich zwischen den Rohdaten und den getrimmten Daten

→ SRR453566_1

Rohdaten

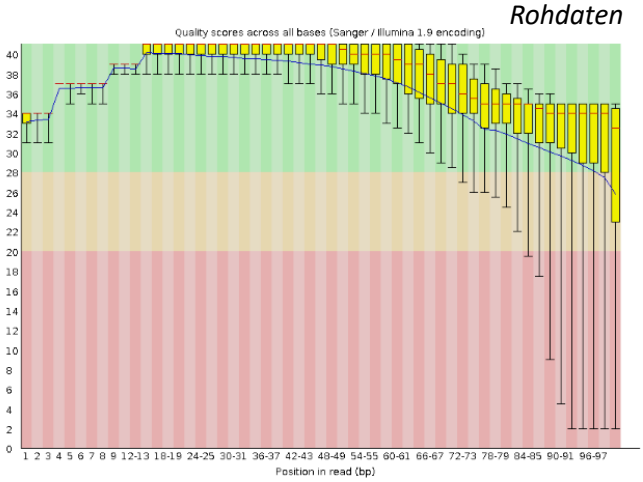
Measure	Value
Filename	SRR453566_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	5725730
Total Bases	578.2 Mbp
Sequences flagged as poor quality	0
Sequence length	101
%GC	41

Getrimmte Daten

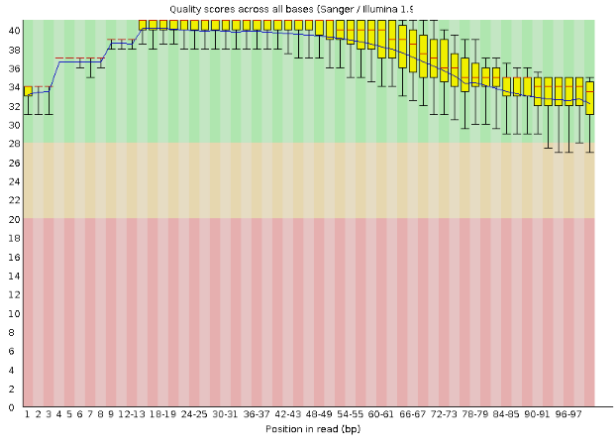
Measure	Value
Filename	SRR453566_1_val_1.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	5650497
Total Bases	540.4 Mbp
Sequences flagged as poor quality	0
Sequence length	20-101
%GC	41

- In den Basic Statistics ist eine deutliche Abnahme der Gesamtbasen-Anzahl, der Gesamtzahl der Sequenzen sowie der Sequenzlänge erkennbar

Rohdaten

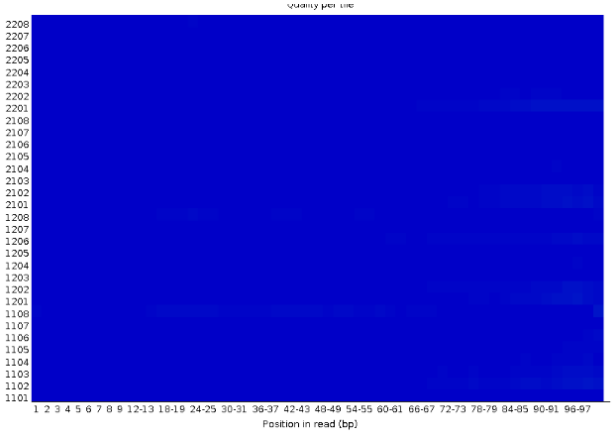


Getrimmte Daten

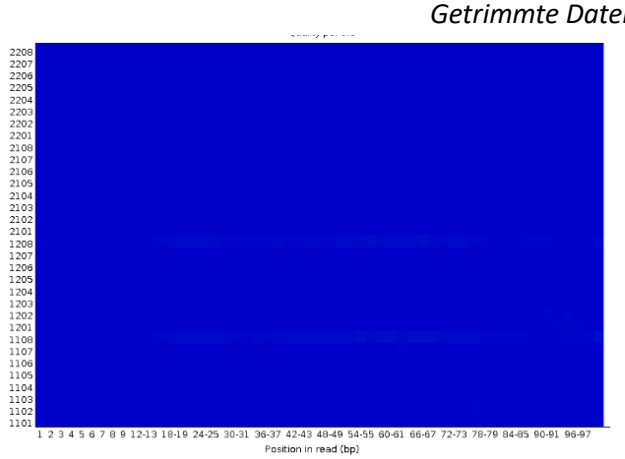


- Es ist deutlich erkennbar, dass Reads mit einem niedrigen Qualitätsscore entfernt worden sind, sodass nur noch Daten mit einem guten bis mäßigen Qualitätsscore erhalten worden sind

Rohdaten

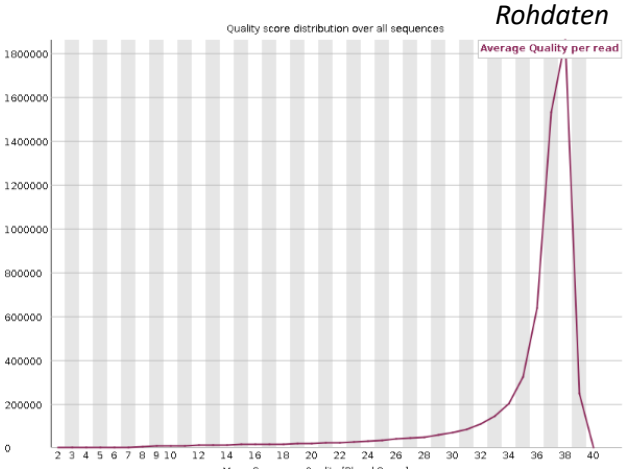


Getrimmte Daten

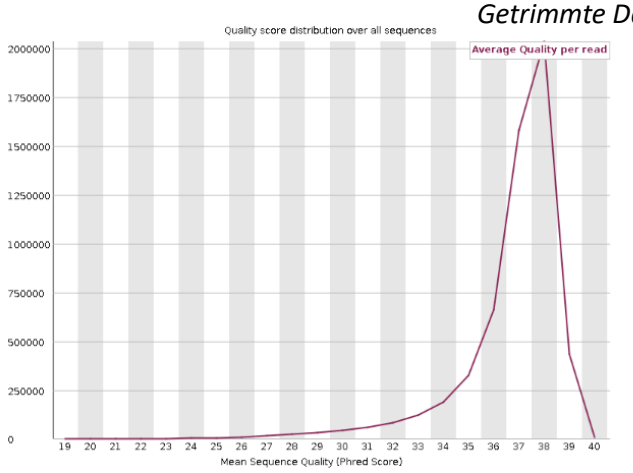


- Bei genauerem Hinschauen ist erkennbar, dass auch hier das Reads oder Teile eines Rads entfernt worden sind, dessen Signal nicht eindeutig genug ausgefallen sind

Rohdaten



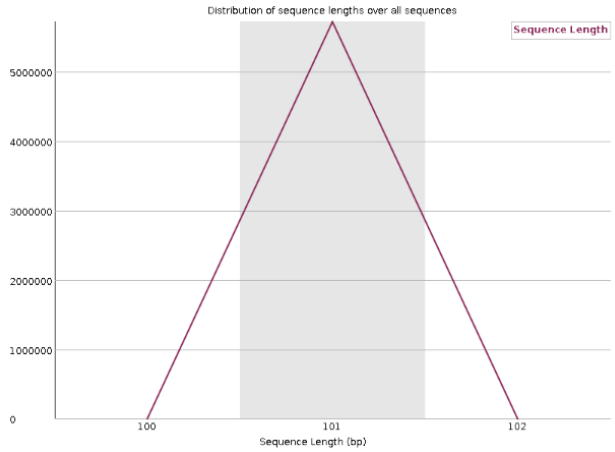
Getrimmte Daten



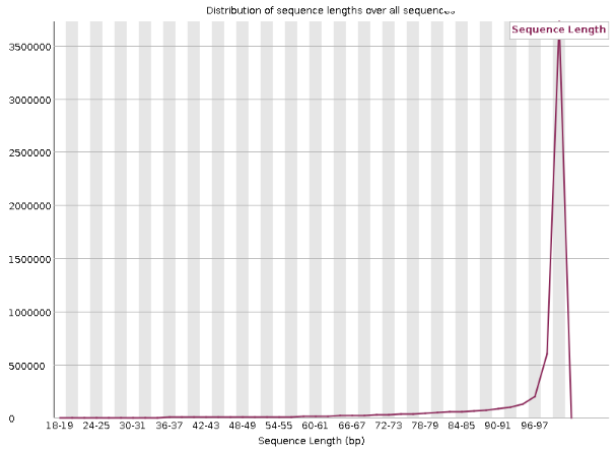
- Da alle Reads mit zu niedrigen Qualitätsscore entfernt worden sind (was man am abgeschnittenen Index erkennt), ist verhältnismäßig erkennbar, dass der Anteil an Reads mit hoher Qualität höher als vorher ist

- Beim Durchschnittlichen GC-Gehalt, beim Basengehalt pro Sequenz, beim Duplikationslevel der Sequenzen und beim N-Gehalt pro Sequenz, waren keine graphischen Unterschiede erkennbar, weshalb sie hier auch nicht mehr aufgeführt werden

Rohdaten



Getrimmte Daten



- Durch das Trimmen sind unterschiedlich lange Read-Längen entstanden, wodurch sich hier kein symmetrischer Verlauf wie bei den Rohdaten ergibt
- Jedoch bleibt der allergrößte Teil auf einer Sequenzlänge von 96-97 bis 102 erhalten

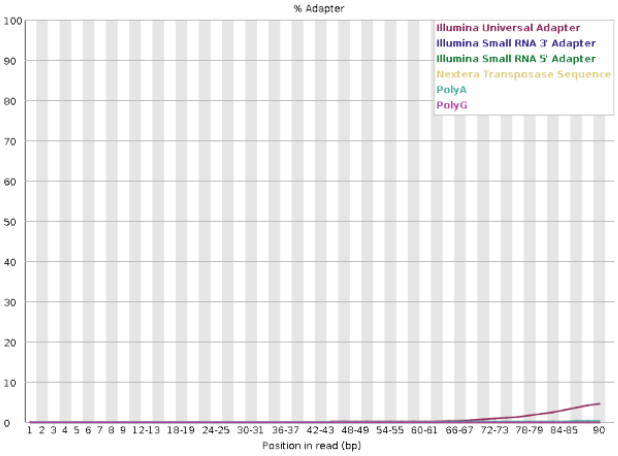
Spaltennan

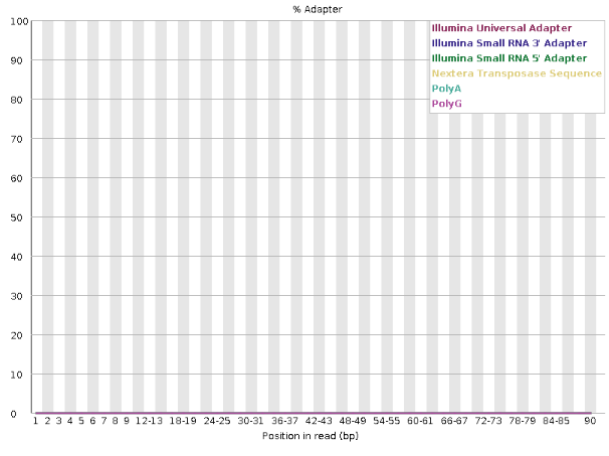
Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	13810	0.24119195281649677	TruSeq Adapter, Index 2 (100% over 50bp)

Getrimmte Daten

No overrepresented sequences

- Durch das Trimming wurden auch die Adaptersequenzen entfernt, die vorher zu den überpräsenten Sequenzen gehörten





- Auch hier ist zu erkennen, dass die Adaptersequenzen aus den Reads entfernt worden sind
- Im Gegensatz zu den Rohdaten (wo die Illumina Universal Adaptersequenz noch deutlich in den Reads vorhanden war), erkennt man das anhand der Nulllinie entlang der x-Achse

→ SRR453566_2

Rohdaten

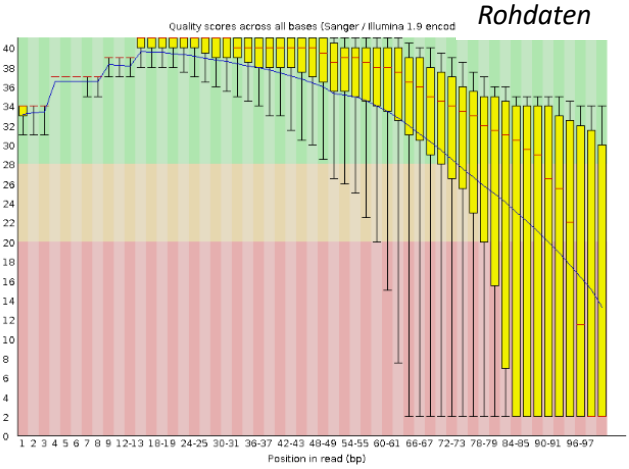
Measure	Value
Filename	SRR453566_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	5725730
Total Bases	578.2 Mbp
Sequences flagged as poor quality	0
Sequence length	101
%GC	42

Getrimmte Daten

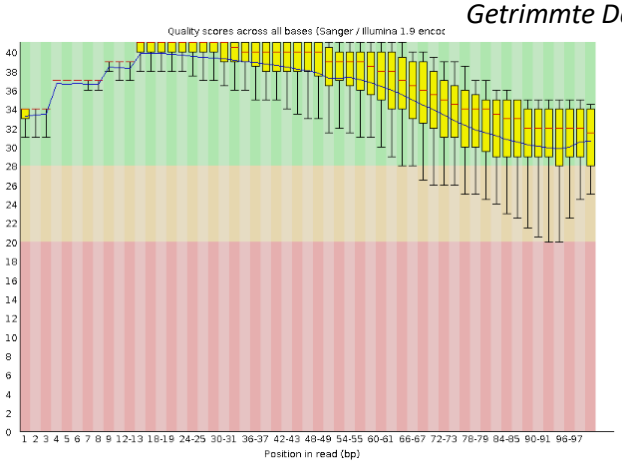
Measure	Value
Filename	SRR453566_2_val_2.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	5650497
Total Bases	496.7 Mbp
Sequences flagged as poor quality	0
Sequence length	20-101
%GC	41

- In den Basic Statistics ist eine deutliche Abnahme der Gesamtbasen-Anzahl, der Gesamtzahl der Sequenzen sowie der Sequenzlänge erkennbar

Rohdaten

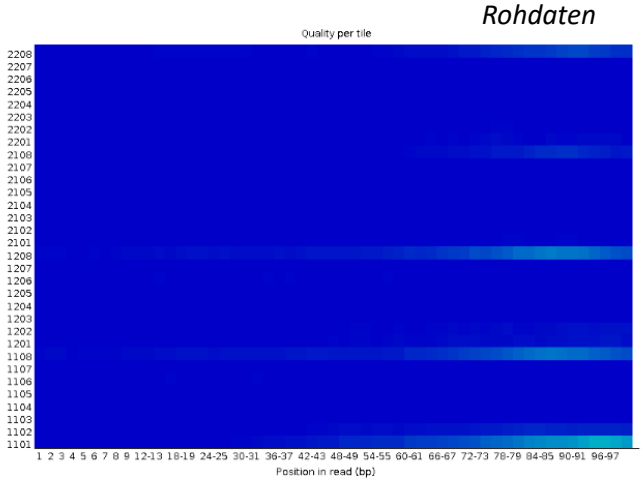


Getrimmte Daten

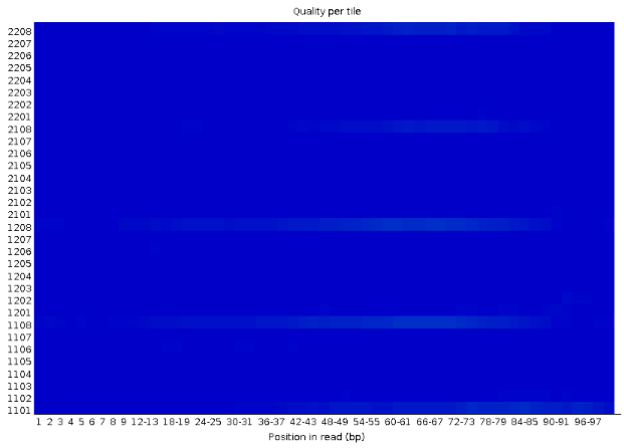


- Es ist deutlich erkennbar, dass Reads mit einem niedrigen Qualitätsscore entfernt worden sind, sodass nur noch Daten mit einem guten bis mäßigen Qualitätsscore erhalten worden sind
- Der niedrigste Qualitätsscore beträgt hier 20

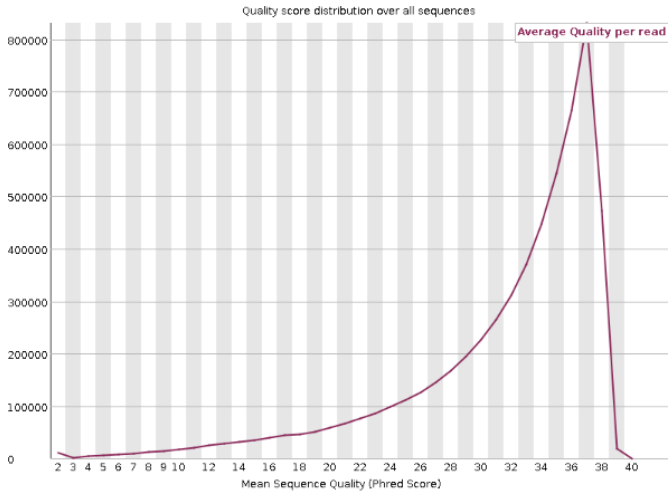
Rohdaten

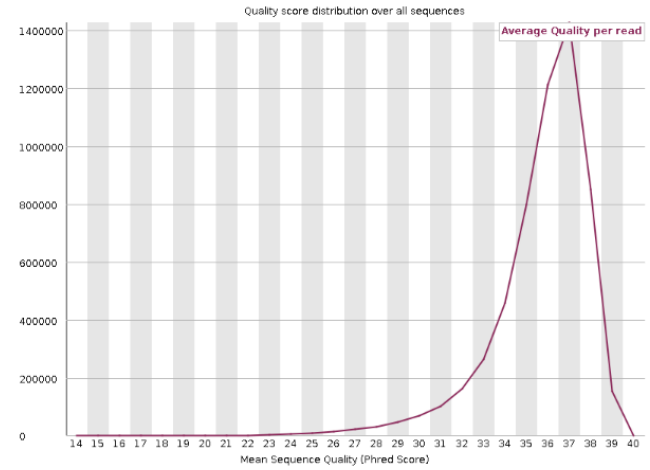


Getrimmte Daten



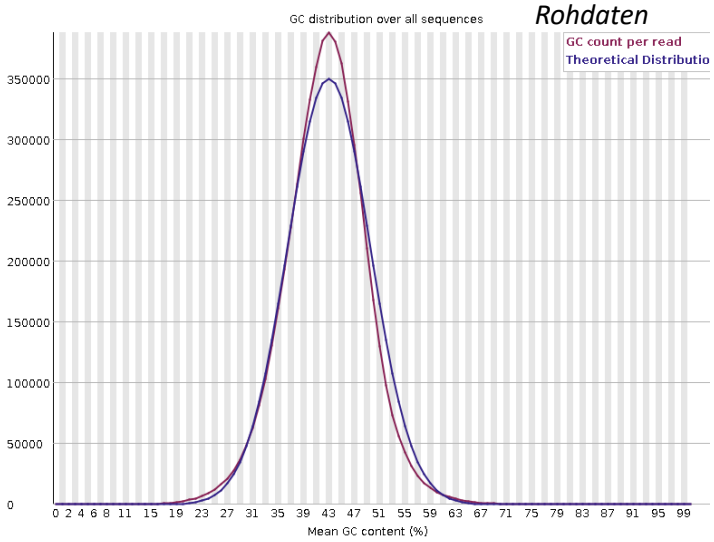
- Es ist deutlich erkennbar, dass auch hier das Reads oder Teile eines Rads entfernt worden sind, dessen Signal nicht eindeutig genug ausgefallen sind



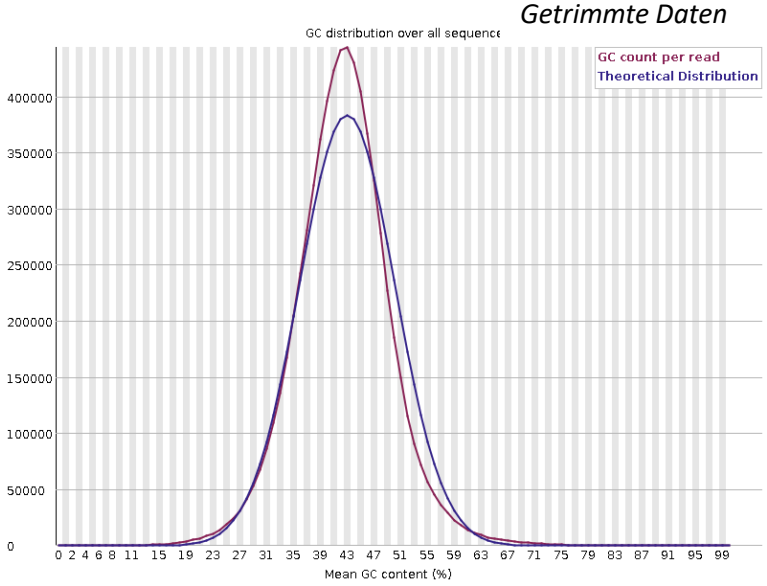


- Da alle Reads mit zu niedrigen Qualitätsscore entfernt worden sind (was man am abgeschnittenen Index erkennt), ist verhältnismäßig erkennbar, dass der Anteil an Reads mit hoher Qualität höher als vorher ist

Rohdaten

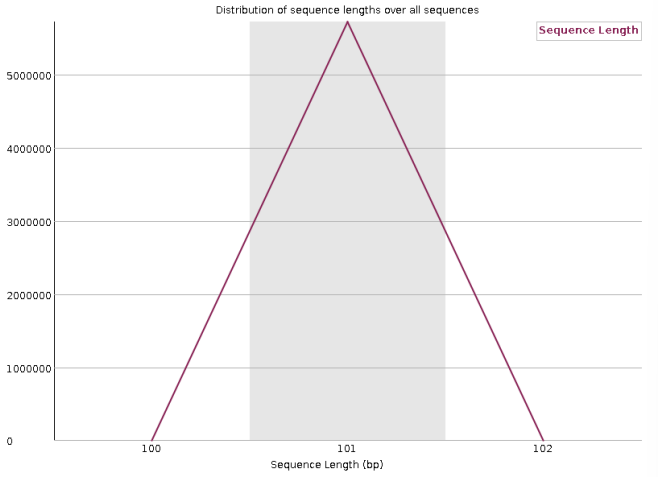


Getrimmte Daten

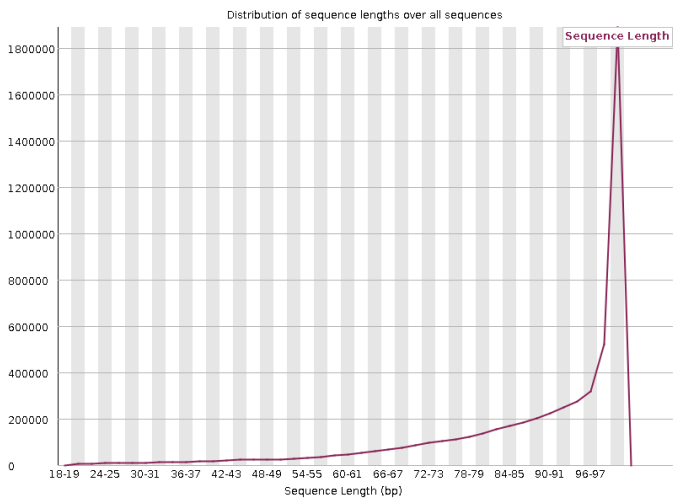


- Da ein großer Anteil der Reads entfernt worden sind, ist nach dem Trimming auch der relative Gehalt bzw. der Durchschnittliche GC-Gehalt gestiegen, was man an dem höher liegendem Peak erkennt

Rohdaten



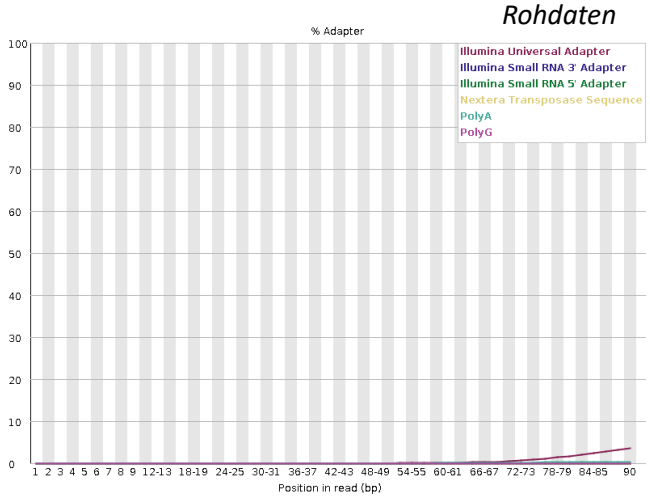
Getrimmte Daten



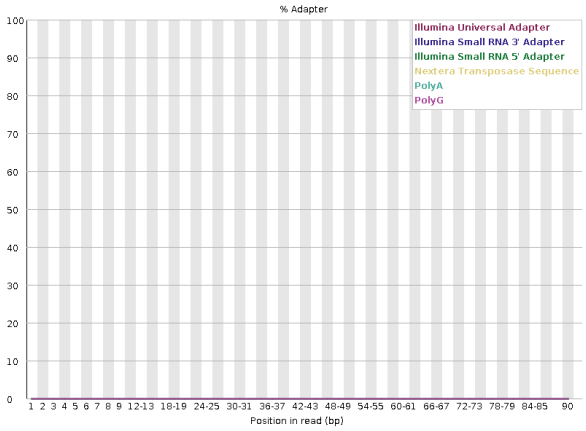
- Durch das Trimmen sind unterschiedlich lange Read-Längen entstanden, wodurch sich hier kein symmetrischer Verlauf wie bei den Rohdaten ergibt
- Jedoch bleibt der allergrößte Teil auf einer Sequenzlänge von 96-97 bis 102 erhalten

- Beim Basengehalt pro Sequenz, bei der Angabe der Überpräsentierten Sequenzen, beim Duplikationslevel der Sequenzen und beim N-Gehalt pro Sequenz, waren keine graphischen Unterschiede erkennbar, weshalb sie hier auch nicht mehr aufgeführt werden

Rohdaten



Getrimmte Daten



- Auch hier ist zu erkennen, dass die Adaptersequenzen aus den Reads entfernt worden sind
- Im Gegensatz zu den Rohdaten (wo die Illumina Universal Adaptersequenz noch deutlich in den Reads vorhanden war), erkennt man das anhand der Nulllinie entlang der x-Achse

Aufgabe 4

Was ist ein Index?

- Bowtie-built erstellt einen Bowtie-Index aus einer Reihe von DNA-Sequenzen
- Es wird ein Satz mit 6 Dateien rausgegeben mit folgenden Suffixen:
    - `.1.ebwt, .2.ebwt, .3.ebwt, .4.ebwt, .rev.1.ebwt, .rev.2.ebwt.`
- Diese Dateien bilden zusammen den Index
- Sie sind alles, was benötigt wird, um die Reads am Referenzgenom auszurichten
    - Sie bildet eine vorgefertigte Datenstruktur, wo effizient nach bestimmten Mustersequenzen gesucht werden, kann
- Die Orginalsequenzdateien werden von Bowtie nicht mehr verwendet sobald er Index erstellt worden ist

Quelle: https://bowtie-bio.sourceforge.net/manual.shtml#the-bowtie-build-indexer

Durchgeführter Befehl:

```
GenExProjekt2 bowtie2-build ./index/Saccharomyces_cerevisiae.R64-1-1.dna.toplevel.fa.gz ./index/yeast
```

- Bowtie2-build: erstellt einen Index für eine Referenzsequenz, sodass das Bowtie2-Programm schneller nach Übereinstimmungen zwischen kurzen DNA-Sequenzen und dieser Referenzsequenz suchen kann.
- ./index/Saccharomyces_cercisiae.R64-1-1.dna.toplevel.fa.gz: Anwendung auf die im Verzeichnis './index/' liegende Datei und im gz-komprimierten FASTA-Format vorliegend
- ./index/yeast: Dies ist der Basisname für den Index, der erstellt wird. Bei diesem Befehl ist es './index/yeast'

Durchgeführter Befehl:

```
bowtie2 -p 6 -x ./index/yeast -1 SRR453566_1_val_1.fq.gz -2 SRR453566_2_val_2.fq.gz -S SRR453566_mapped.sam
```

- Bowtie2: ruft das Bowtie2-Programm auf, um Alignments durchzuführen
- -p 6: Anzahl an Threads oder Prozessorkerne, die Bowtie verwenden soll (hier: 6 Kerne)
- -x ./index/yeast: Pfad zum Bowtie2-Index wird hier angegeben
- -1 SRR4535xx_1_cutadapt_val_1.fq.gz: Die erste jeweilige Eingabedatei.
- -2 SRR4535xx_2_val_2.fq.gz: Die jeweilige zweite Eingabedatei
- -S SRR4535xx_mapped.sam: Gibt an, dass die Ausgabe in einer SAM-Datei gespeichert werden soll

Was sind SAM-Dateien?

- = "Sequence Alignment Map"
- Ein spezielles Dateiformat, das Informationen über das Alignment von DNA- oder RNA-Sequenzen gegen eine Referenzsequenz enthält
- besteht aus einem Kopf [Header] + dem Alignment-Abschnitt

Woraus bestehen sie?

Spaltenname	Beschreibung

Header	• beginnt mit “@” + Identifier
	• indiziert den Typen + Subtypen der Heder Lin
	
	`@SQ SN:chr14 LN:107349540`
	`@PG ID:bwa PN:bwa VN:0.7.7-r441 CL:bwa mem ref/seq.fa r1.fastq r2`
	
	• @SQ: gibt den Referenzort zurück
	• SN: hier Chr 14
	• LN: gibt die Länge des Referenzort zurück
	• @PG: beschreibtdas Programm, womit der SAM-File generiert worden ist
	• PN: gibt den Namen des Programms zurück
	• CL: gibt eine Kopie, des Befehls, welches aufgeführt wurde zurück

QNAME	• = eindeutiger Identifier für den Read • ***Ausnahme:*** Gepaarte Enden haben denselben QNAME, da sie von derselben DNA-Sequenz stammen • Unterscheidung von gepaarten Reads erfolgt anhand ihrer Ausrichtung (= bestimmt durch den FLAG-Wert)

FLAG	• eine dezimale (Basis-10) Zahl
	  ○ verwendet, um eine binäre (Basis-2) Zahl darzustellen
	  ○ Ziffern repräsentieren verschiedene Wahrheits-/Falschaussagen bezüglich der Ausrichtung des Gelesenen
	  ▪ 0= false
	  ▪ 1=true
	• Bitfeld, das Informationen über das Mapping des Reads enthält
	  ○ über versch. Ausssagen

Decimal	Binary	Exp.	Meaning
1	1	2^0	This is a paired read
2	10	2^1	This read is part of a pair that aligned properly*
4	100	2^2	This read was not aligned
8	1000	2^3	This read is part of a pair and its mate was not aligned
16	10000	2^4	This read aligned in the reverse direction**
32	100000	2^5	This read is part of a pair and its mate aligned in the reverse direction*
64	1000000	2^6	This read is the first in the pair (read 1)
128	10000000	2^7	This read is the second in pair (read 2)
256	100000000	2^8	The given alignment is a secondary alignment***
512	1000000000	2^9	Read failed quality check (such as Illumina quality filtering)
1024	10000000000	2^{10}	Read was flagged as a duplicate (such as a PCR duplicate)
2048	100000000000	2^{11}	Supplementary alignment (Exact meaning varies by aligner)

* Proper alignment indicates both reads in a pair are oriented towards one another (one forward, one reverse), are both on the same contig, and are within the expected distance from one another.

** Direction is relative to the reference sequence used for alignment

*** The read had multiple potential alignments; this was one of them, but not the first choice from among t

Beispiel:

- FLAG value → 99
  - 64+32+2+1
- Paired Reads kreiieren immer ungerade zahlen

RNAME

- Name bzw. Kennung des referenzierten DNA-Segments
- erscheint auch im Header

POS

- Startposition der gelesenen DNA-Sequenz auf dem Referenzgenom (leftmost)
- 1 → die erste Postiion im Referenzgenom

MAPQ

- Mapping-Qualität des Reads
- Score der angibt, wie wahrscheinlich es ist, dass die Sequenz richtig oder falsch gemappt worden ist
- 255 → keine Wahrscheinlichkeit gegeben, und wird als Platzhalter Wert benutzt
- **Formel:** -10log10(e)
  - e= *Wahrscheinlichkeit, dass die mapping Position falsch ist*
  - insg. gerundet auf einen Integer
- Abhängig von
  - wie einzigartig die alignte Region im Genom ist
  - Länge des Alignments
  - Anzahl Mismatches & Gaps

CIGAR

- Beschreibung der Alignment-Operationen
- = Concise Idiosyncratic Gapped Alignment Report String.

	• Sequenz von Zahlen & Buchstaben, die Kontinuitäten oder Diskontinuitäten in der Ausrichtung anzeigen
	  ○ die durch eingefügte/gelöschte Basen (oder andere Gründe für Diskontinuität) verursacht werden
	
	Beispiel:
	
	• 5M2D5M
	
	• Die Buchstaben in der CIGAR-Sequenz haben folgende Bedeutungen:
	  ○ M: Match (Übereinstimmung)
	  ▪ **CAVE:** es zählt auch als Match, wenn die Base nicht übereinstimmen aber zueinander gemapp sind!
	  ○ D: Deletion (Löschung)
	  ○ I: Insertion (Einfügung)
	  ○ S: Soft clip (Teilweise Abschneiden)
	  ○ H: Hard clip (Komplettes Abschneiden)
	  ○ N: Skipped region (Übersprungener Bereich)
	• Die Zahlen in der CIGAR-Sequenz geben die Anzahl der aufeinanderfolgenden Vorkommen des jeweiligen Buchstabens an.

RNEXT	• Name bzw. Kennung des nächsten referenzierten Segments
	• entspricht Feld 3 (Referenzname)+ folgt denselben Regeln
	  ○ Ausnahme: beschreibt das gepaarte End-Mate des Reads (wenn vorhanden)
	  ○ Wert ist "=", wenn es identisch mit dem Referenznamenwert ist (Platz sparen)

PNEXT	• entspricht Feld 4 + die gleichen Regeln

TLEN	• gibt die Länge der Referenzsequenz an, auf die sich das Read abbildet.
	  ○ wird manchmal mit der Read-Länge verwechselt, ist aber oft gleichwertig
	  ○ Ein Read mit mehreren Einfügungen kann eine kleinere Referenzlänge haben als die Read-Länge
	  ○ Ein Read mit mehreren Deletionen kann eine längere Referenzlänge haben als die Read-Länge
	  ○ *Merke:* Wenn RNA oder cDNA auf genomische DNA abgebildet wird, kann die Referenzlänge aufgrund eines Introns bei einem kurzen Read mehrere Zehntausend Basen betragen

SEQ	• die Read Sequenz
QUAL	• Qualitätswerte der Basen in SEQ • Phred-scaled • vom FASTQ-File generiert

Quelle: https://www.zymoresearch.de/blogs/blog/what-are-sam-and-bam-files

Aufgabe 5: Von Alignments zu Genen

Tool: Samtools

- = eine Reihe von Programmen für die Interkation und Nachbearbeitung kurzen DNA-Read-Alignments im SAM (oder BAM, CRAM) Format
- Es konvertiert zwischen den Formaten, führt Sortierung, Zusammenführung du Indizierungen und kann Lesevorgänge in beliebigen Regionen schnell abrufen

Durchgeführter Befehl:

```
(cutadaptenv) → GenExProjekt2 samtools view -bS SRR453566_mapped.sam > SR
R453566_mapped.bam
samtools sort SRR453566_mapped.bam > SRR453566_mapped_sorted.bam
samtools index SRR453566_mapped_sorted.bam
```

- `samtools view -bS SRR4535XX_mapped.sam > SRR4535XX_mapped.bam`
  - Der Befehl konvertiert die übergebene SAM-Datei in das BAM Format
  - -b: gibt an, dass die Ausgabe im BAM-Format sein soll
  - -s: gibt an, dass die Eingabedatei im SAM-Format vorliegt
- `samtools sort SRR4535XX_mapped.bam > SRR4535XX_mapped_sorted.bam`
  - Der Befehl sortiert die BAM-Datei und speichert die sortierte Ausgabe in der Datei SRR4535XX_mapped_sorted.bam
  - Durch das Sortieren wird die BAM-Datei nach genomischer Position sortiert
- `samtools index SRR4535XX_mapped_sorted.bam`
  - index: Der Befehl erstellt einen Index für sortierte BAM-Datei
  - der Index erleichtert einen schnellen Zugriff auf bestimmte Bereiche des Genoms in der BAM Datei

Ausgeführter Befehl:

```
GenExProjekt2 sed -i 's/^chr//g' yeast_genes.bed
GenExProjekt2
```

- mit diesem Befehl wird "chr" am Anfang jeder Zeile in der yeast_genes.bed entfernt und speichert die Änderungen direkt in der Datei

Tool: Bedtools

- ein Toolkit, welches mehrere Operationen ermöglicht, um sie auf genomischen Intervallen (BED-Dateien) auszuführen
- Hauptfunktionen: genomische Intervalle aus mehreren Dateien überschneiden, zusammenzuführen, zu zählen, zu ergänzen und zu mischen

Quelle: https://bedtools.readthedocs.io/en/latest/

Ausgeführter Befehl:

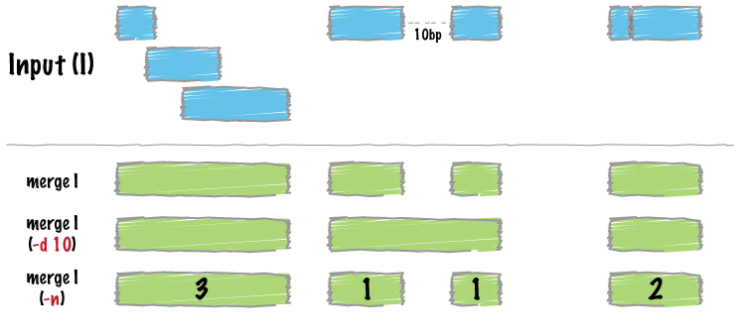
```
(cutadaptenv) → GenExProjekt2 bedtools multicov -bams SRR453566_mapped_sorted.bam SRR453567_mapped_sorted.bam SRR453568_mapped_sorted.bam SRR453569_mapped_
sorted.bam SRR453570_mapped_sorted.bam SRR453571_mapped_sorted.bam -bed yeast_genes.bed >Bedtools_out_all.bed
```

- bedtools multicov: Anzahl der Reads bestimmen, die sich in bestimmten genomischen Bereichen befinden
  - Bereiche zwischen einer .bed-Datei und BAM-Dateien (die Alignments der Reads) analysieren
- -bams SRR453566_mapped_sorted.bam SRR453567_mapped_sorted.bam SRR453568_mapped_sorted.bam SRR453569_mapped_sorted.bam

SRR453570_mapped_sorted.bam SRR453571_mapped_sorted.bam: gibt die BAM-Dateien an, die für die Überlappungsberechnung verwendet werden sollen

- -bed yeast_genes.bed: Gibt die .bed-Datei an, die die genomischen Regionen definiert, über die die Überlappungen berechnet werden sollen
- > Bedtools_out_all.bed: Leitet die Ausgabe des Befehls in eine Datei namens Bedtools_out_all.bed

Graphische Darstellung



Quelle: https://bioweb.pasteur.fr/docs/modules/bedtools/2.19.1/_images/intersect-glyph.png