Angewandte Bioinformatik

Abgabe: 22. Mai, SS 2024

Aufgabe 1 FastQC Qualitätskontrolle

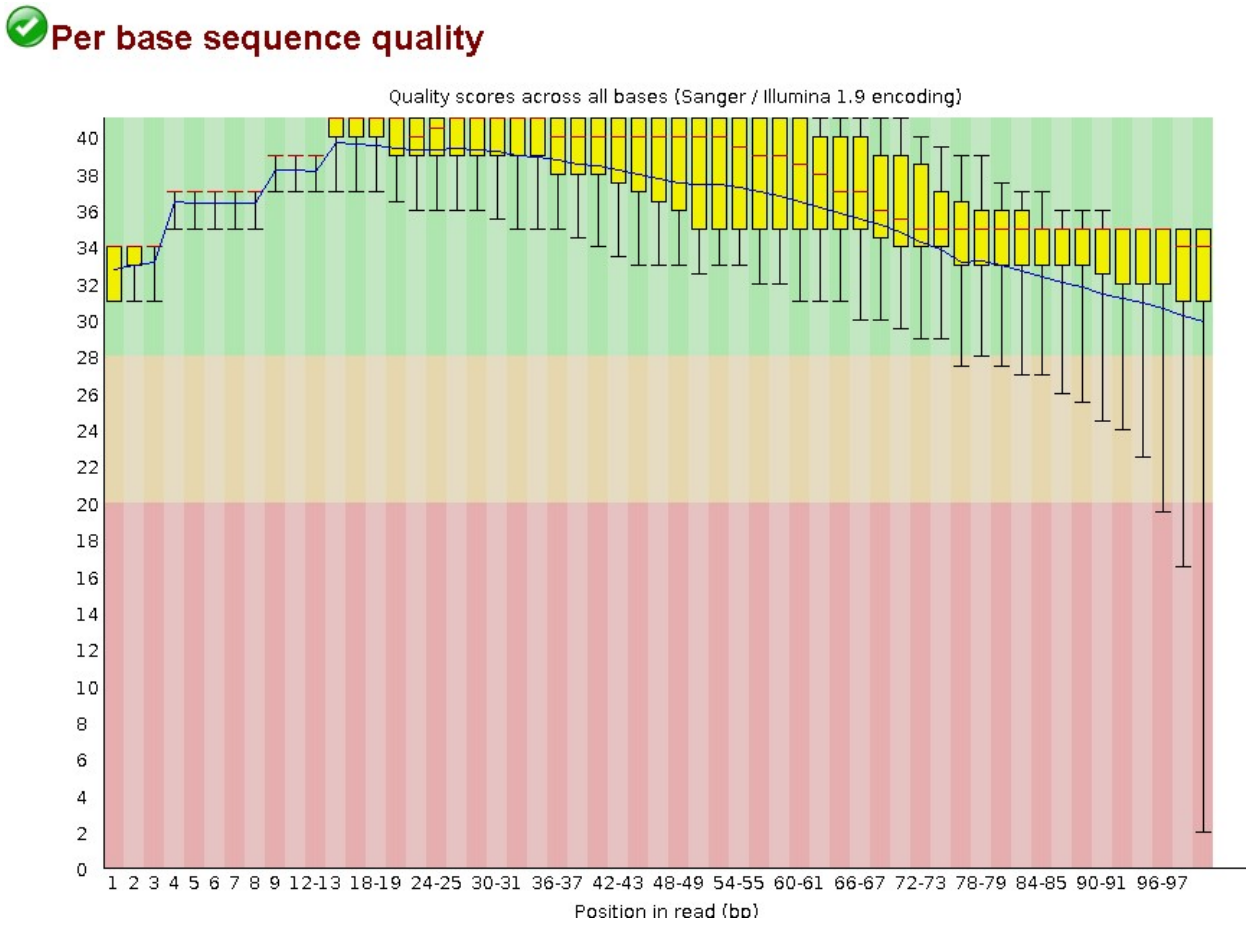
Die Mock-Experiment Datei wurde anhand des *fastq-dum* Befehls aus den *sra-tools* heruntergeladen.

fastq-dump –gzip SRR1952808

Für die Erzeugung der entsprechenden FastQC.html Datei wurde die Tool FastQC verwendet. Folgend werden die jeweiligen Kategorien der FastQC Datei kurz zusammengefasst.

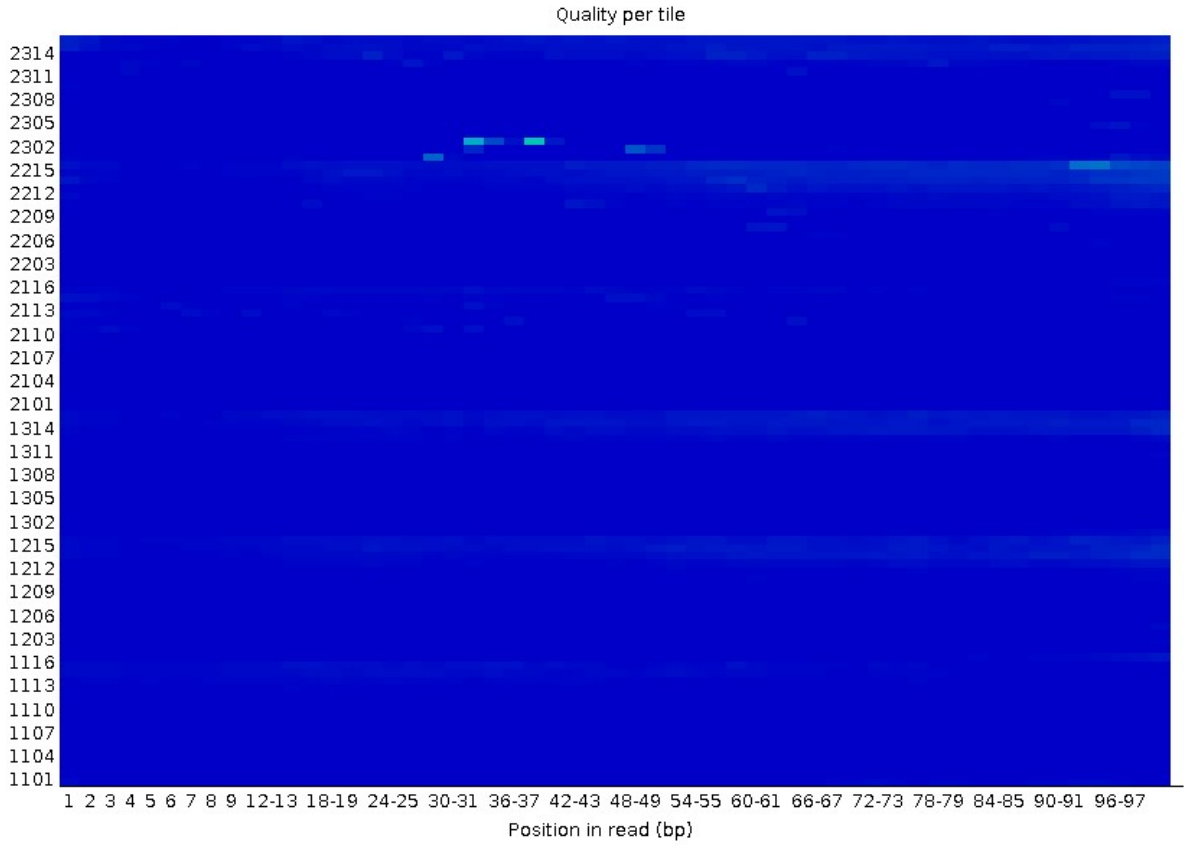
Basic Statistics

Measure	Value
Filename	SRR1952808.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	26656102
Total Bases	2.6 Gbp
Sequences flagged as poor quality	0
Sequence length	101
%GC	48

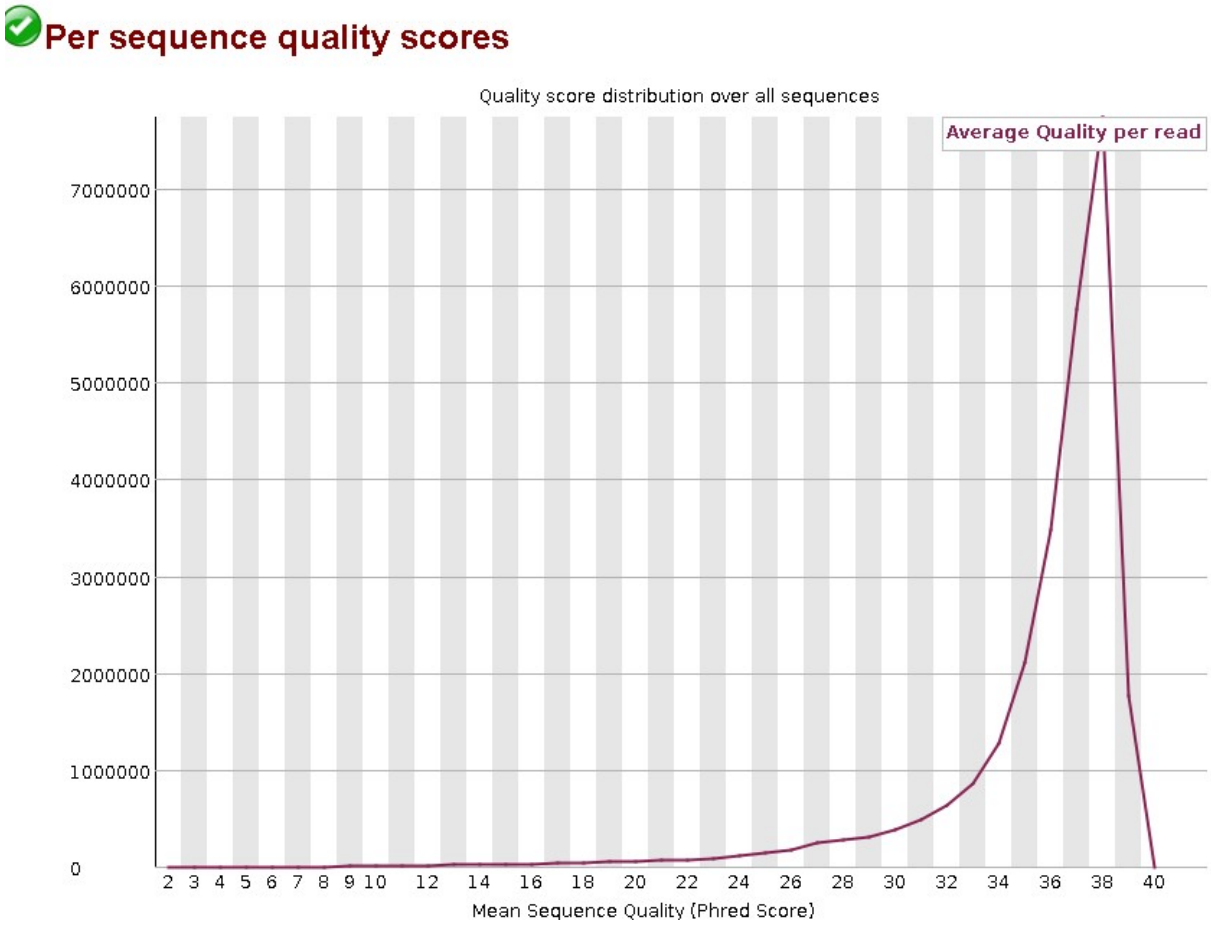


Die y-Achse des Diagramms zeigt die Qualitätswerte und teilt sie in sehr gute (grün), akzeptable (orange) und in schlechte (rot) Qualität ein. In dieser Abbildung sehen wir, dass der Interquartileabstand, wie auch der Mittelwert der Qualität (blaue Linie) der Boxplots im grünen Feld liegen folglich also sehr gute Qualitätswerte widerspieglt.
Das wird nochmal bestätigt, da die Qualität der Aufrufe auf den meisten Plattformen im Verlauf abnimmt und daher üblicherweise Basenaufrufe gegen Ende eines Reads im orangefarbenen Bereich zu sehen sind. Hier sind unsere Basenaufrufe dennoch im grünen Bereich.

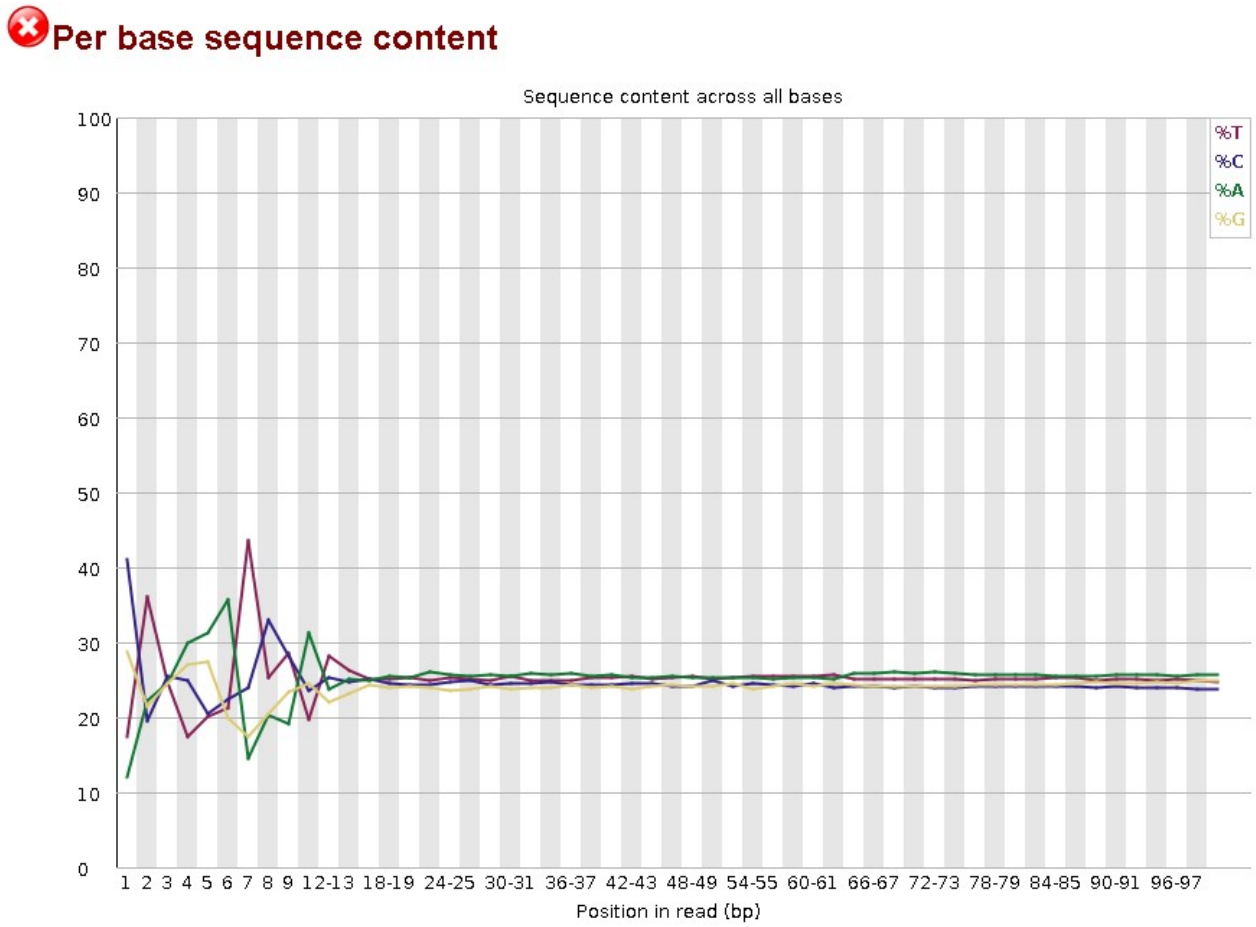
Per tile sequence quality



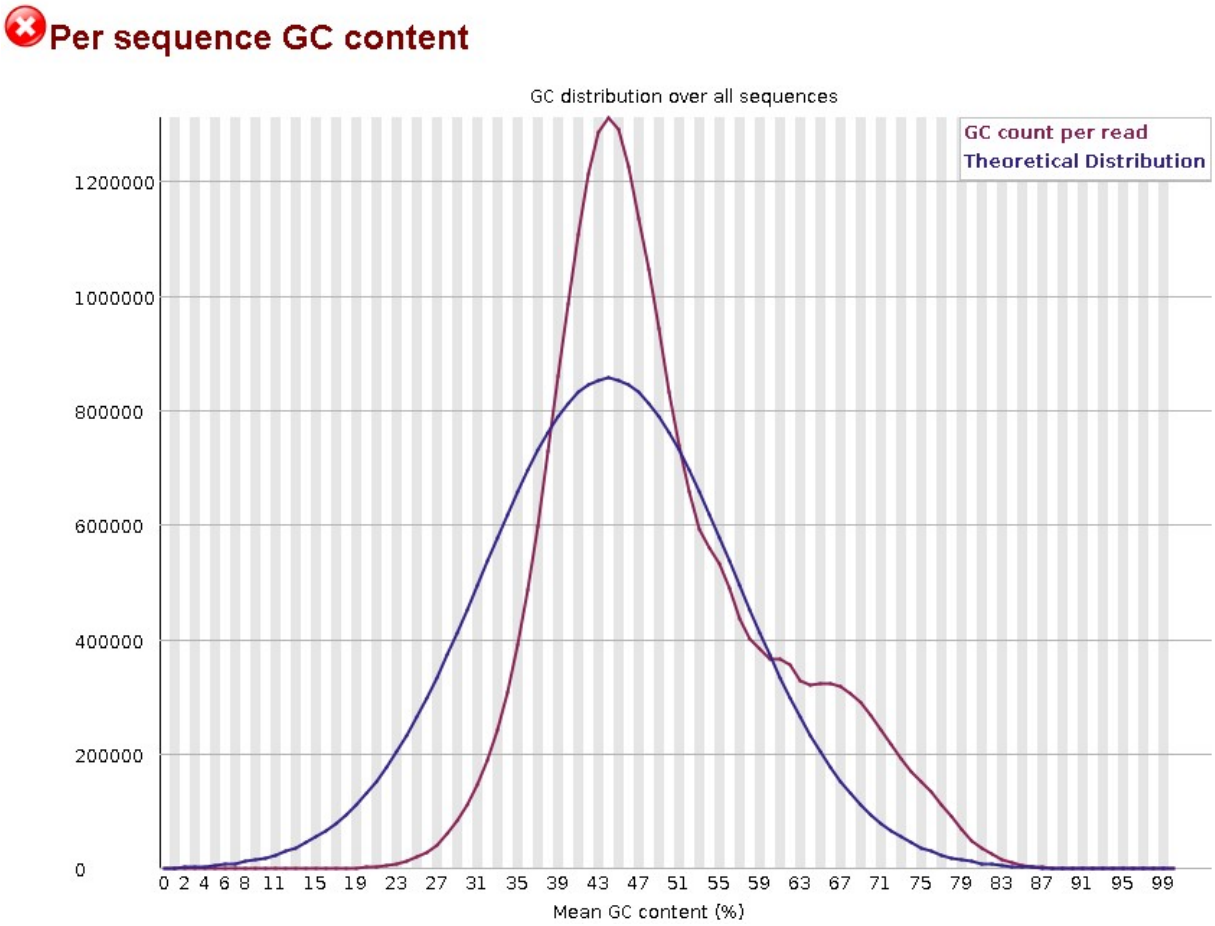
Das Diagramm zeigt die Abweichung der Qualität vom Durchschnitt. Blau bedeutet diese liegt über dem Durchschnitt, rot bedeutet die Qualität liegt unter dem Durchschnitt. Da unsere Graphik komplett blau ausfällt sehen wir auch hier die exzellente Qualität unserer Werte.



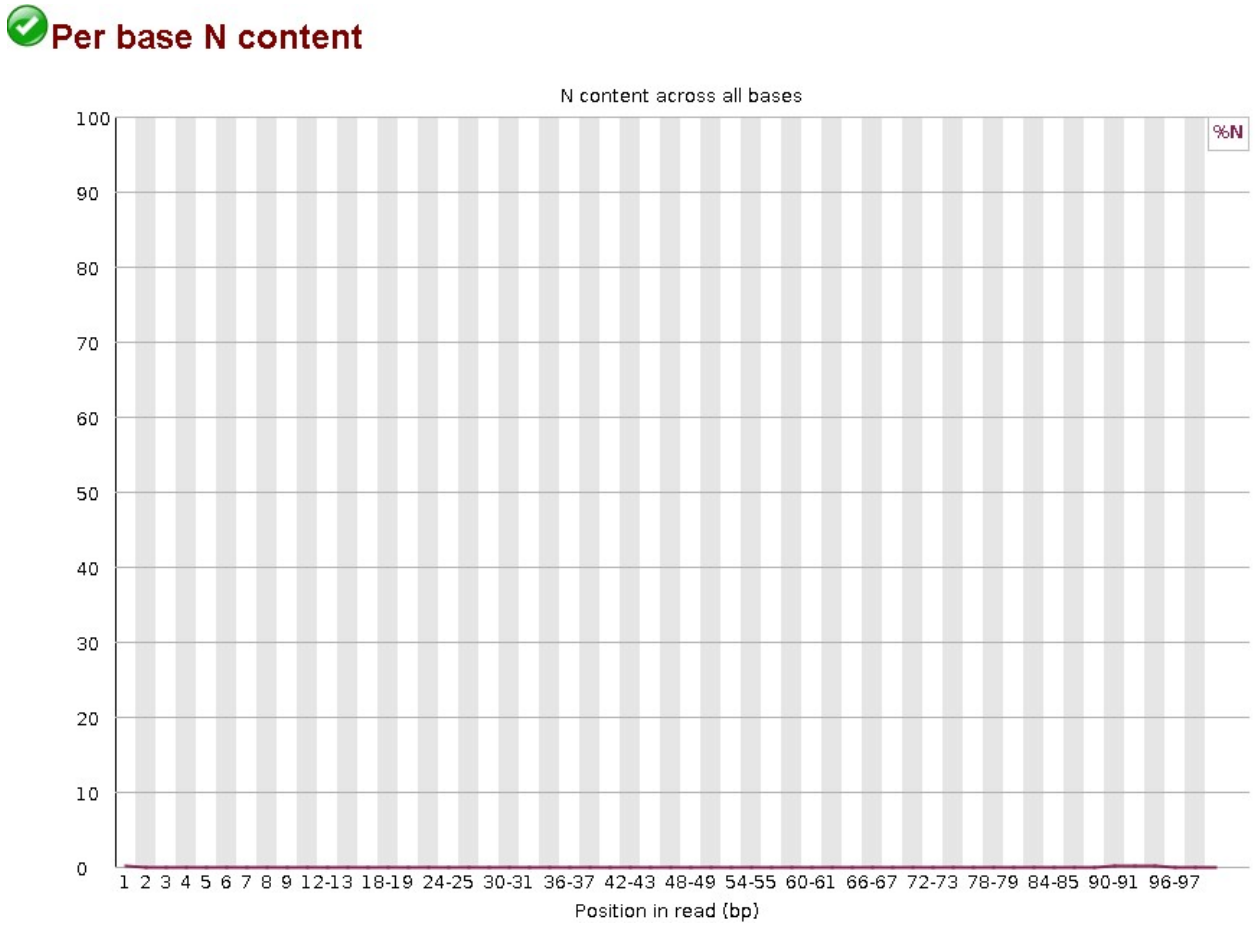
Die Qualitätswerte pro Sequenz zeigen, ob eine Teilmenge der Sequenzen niedrige Qualitätswerte aufweist. Es ist oft der Fall, dass eine Teilmenge von Sequenzen schlechte Qualität aufweist, oft weil sie schlecht abgebildet sind. Da wir jedoch in unserer Graphik nur einen sehr hohen Peak im Bereich des sehr ohnen Mittelwertes der Sequenzqualität sehen, wird deutlich dass alle Werte einer hohen Qualität pro Read entsprechen.



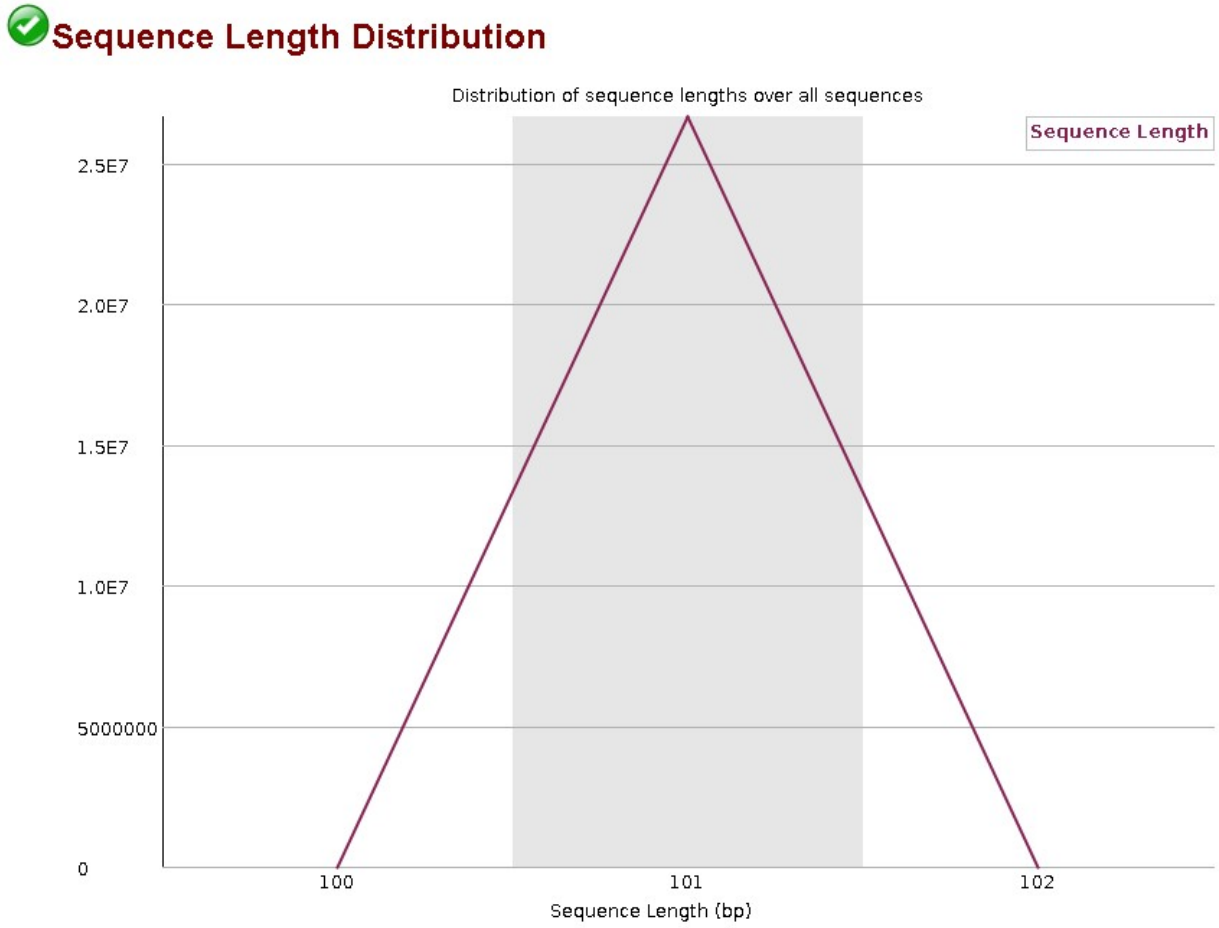
Die Graphik stellt die Proportion der vier Basen dar. Zu erwarten wäre, dass es wenig bis gar keinen Unterschied zwischen den verschiedenen Basen eines Sequenzlaufs gibt, daher sollten die Linien in diesem Diagramm parallel zueinander verlaufen. Jedoch sehen wir in unserer Graphik dass die Graphen am Anfang bis zur 13. Position der Reads stark schwanken. Dies deutet in der Regel darauf hin, dass die ursprüngliche Bibliothek sequenzvoreingenommen war oder dass es während der Sequenzierung der Bibliothek ein systematisches Problem gab.



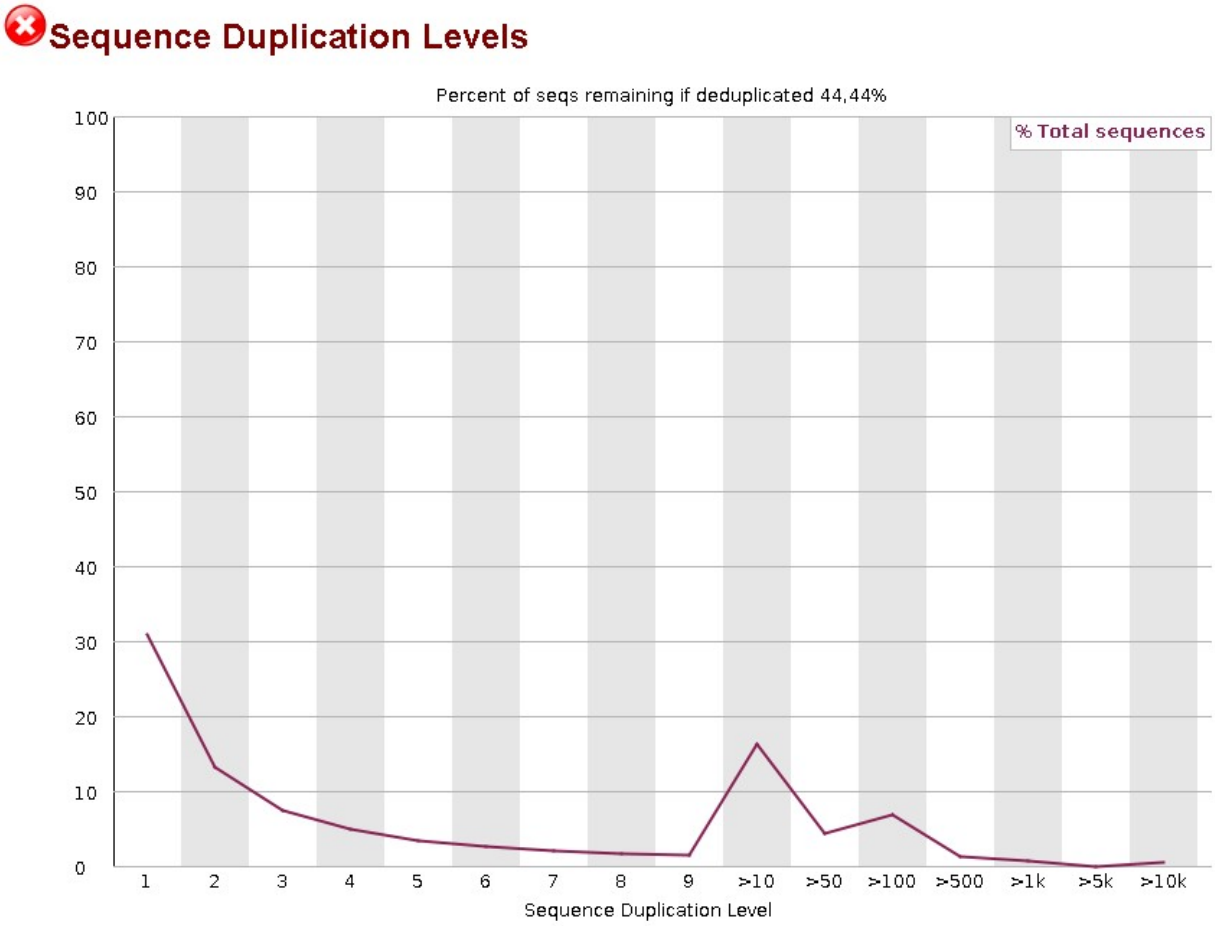
In einer normalen Bibliothek würde man eine ungefähr normale Verteilung des GC-Gehalts erwarten, wobei der zentrale Peak dem Gesamt-GC-Gehalt des Genoms entspricht. Unsere Daten zeigen jedoch keine übereinstimmung der theoretischen Verteilung und der berechnete GC Gehalt. Dies deutet erneut auf eine kontaminierte Bibliothek oder andere Arten von voreingenommenen Teilgruppen hin.



Wenn man nicht in der Lage ist, einen Basenaufruf mit ausreichendem Vertrauen zu machen, wird normalerweise ein N anstelle eines herkömmlichen Basenaufrufs einsetzen. Dieses Graphik zeigt den Prozentsatz der Basenaufrufe an jeder Position, für die ein N aufgerufen wurde. Hier tritt der Fall auf, dass die Linie in diesem Diagramm horizontal und nahezu Werte von N = 0 zeigt. Das heißt unsere Daten konnten sehr gut interpretiert werden, um gültige Basenaufrufe zu erhalten.



In diese Diagramm wird die Länge der Sequnzen dargestellt. Diese liegt hier einheitlich zwischen 100 und 102 bp.

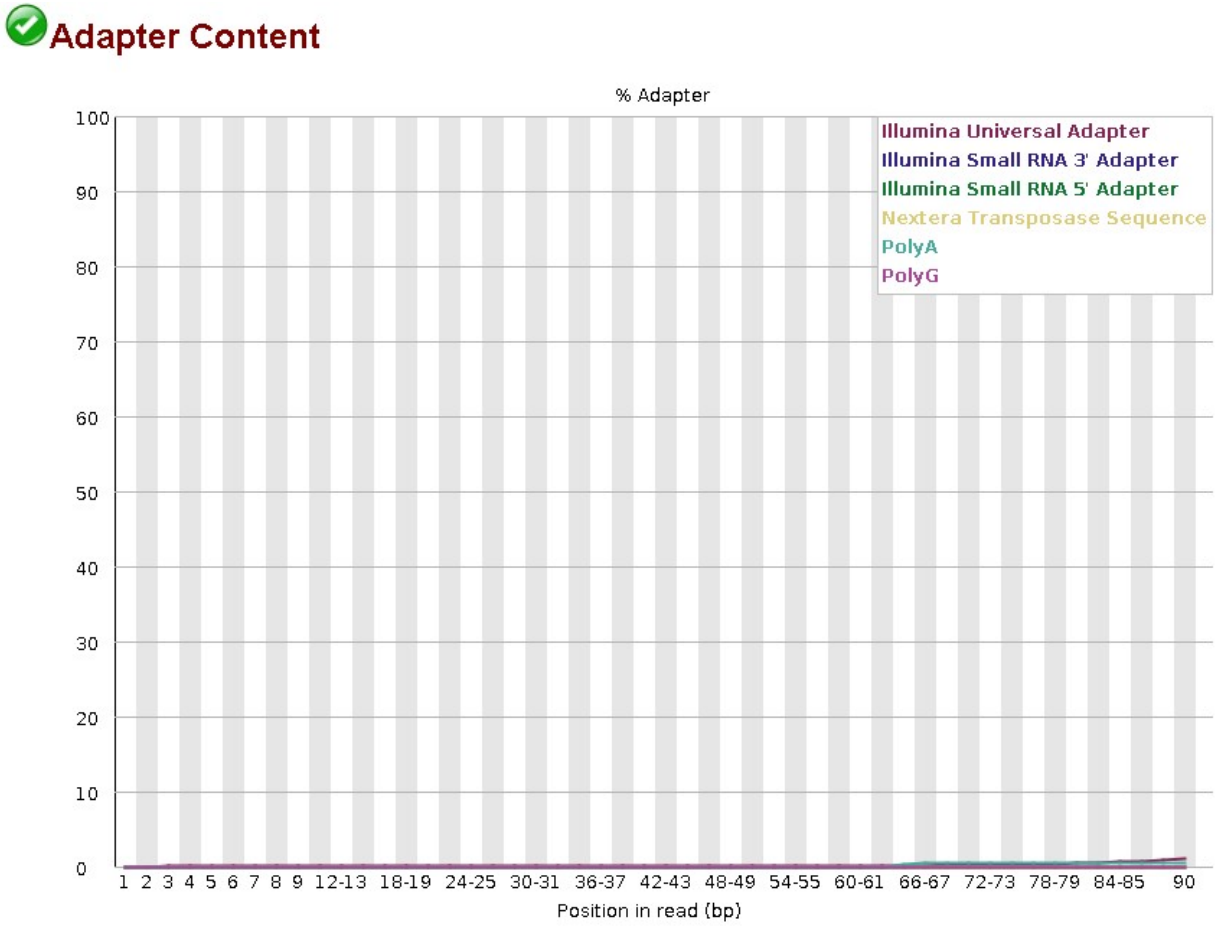


Diese Graphik stellt die relative Anzahl der Sequenzen mit unterschiedlichen Duplikationsgraden dar. Normalerweise erwarten wir einen exponentiell fallenden Graphen. In unserer Graphik sehen wir jedoch nur einen geringen exponentiellen Fall und zwei Peaks bei 10 und 100 Dupliktionslevel. Ein niedriger Grad an Duplikation kann auf eine sehr hohe Abdeckung der Zielsequenz hinweisen.

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGC	145086	0.5442881333512304	TruSeq Adapter, Index 7 (100% over 50bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATG	45814	0.17187059083132258	TruSeq Adapter, Index 7 (100% over 49bp)

Wenn eine Sequenz sehr überrepräsentiert ist, bedeutet dies entweder, dass sie biologisch hochsignifikant ist, oder dass die Bibliothek kontaminiert ist oder nicht so vielfältig wie erwartet. Die Tabelle listet alle Sequenzen auf, die mehr als 0,1% des Gesamten ausmachen.



Hier werden Sequenzen abgebildet die keine gleichmäßige Abdeckung über die Länge Ihrer Reads haben. Dies kann verschiedene Quellen für Verzerrungen in der Bibliothek aufdecken, einschließlich der Anwesenheit von Read-Through-Adaptersequenzen, die sich am Ende ansammeln, was bei uns nicht der Fall ist, da wir konstante horizonale Graphen haben die fast an der x-Achse verlaufen.