

BIN508: Next Generation Sequence Analysis & Informatics

Assignment 01

By: Taha Ahmad

Student ID : 2546125

Instructor: Dr. **Yesim Aydin Son**

Before Trimming (cutadapt)

Basic Statistics

| Measure | Value |
|-----------------------------------|-------------------------|
| Filename | data01 |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 812 |
| Total Bases | 240.3 kbp |
| Sequences flagged as poor quality | 0 |
| Sequence length | 296 |
| %GC | 44 |

After trimming (cutadapt)

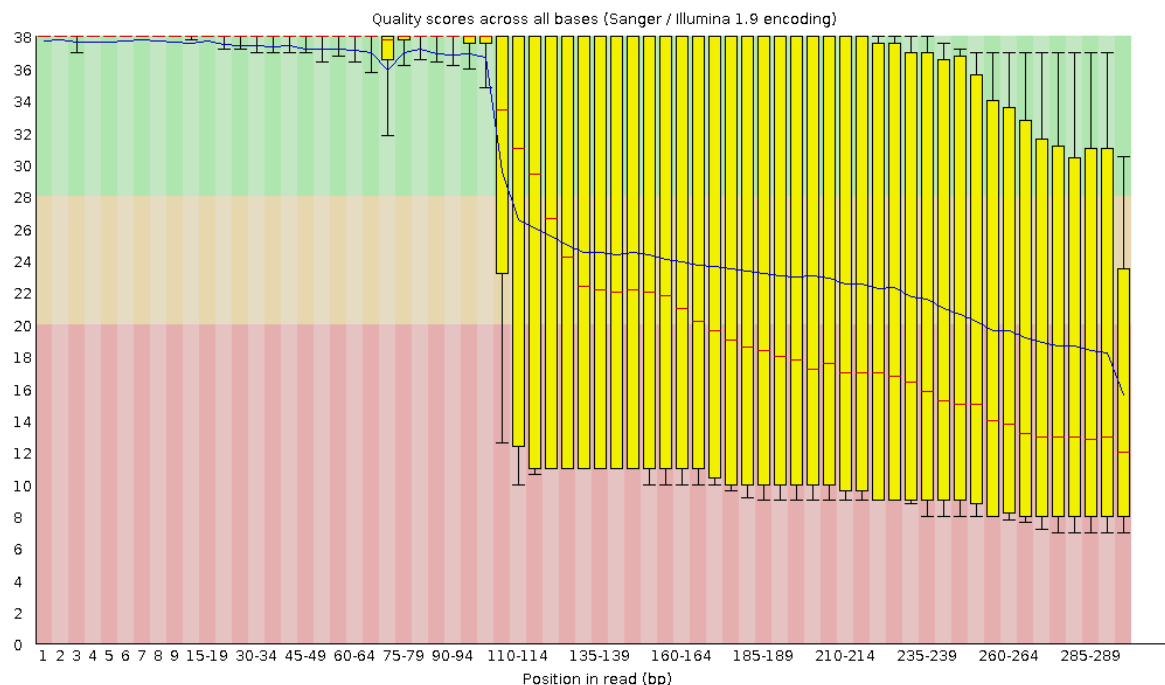
Basic Statistics

| Measure | Value |
|-----------------------------------|-----------------------------------|
| Filename | Cutadapt on data 1_ Read 1 Output |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 812 |
| Total Bases | 116.4 kbp |
| Sequences flagged as poor quality | 0 |
| Sequence length | 26-296 |
| %GC | 54 |

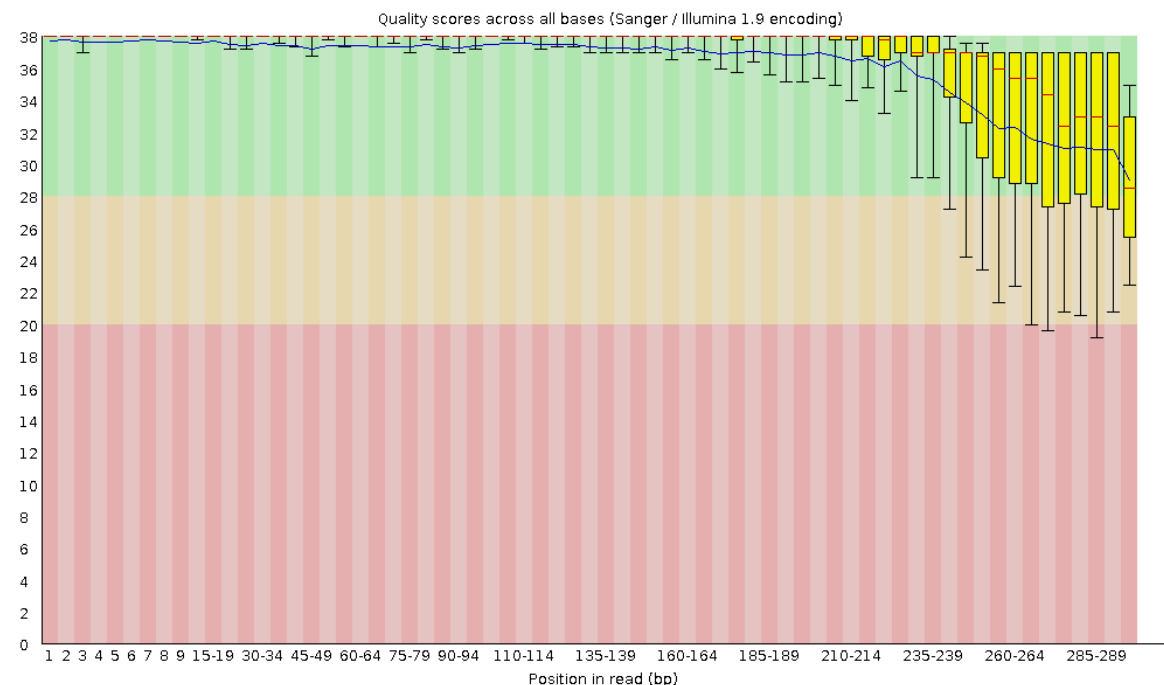
The FASTQC results show that before trimming, all reads were 296 bases long. After trimming, lengths varied (26-296 bases) because cutadapt removed low-quality bases and adapters. The total bases decreased, but the number of reads stayed the same, meaning some reads got shorter. GC content increased by 10%, likely because the removed adapters or low-quality regions were AT-rich, leaving more GC-rich sequences.

Per base sequence quality

Untrimmed



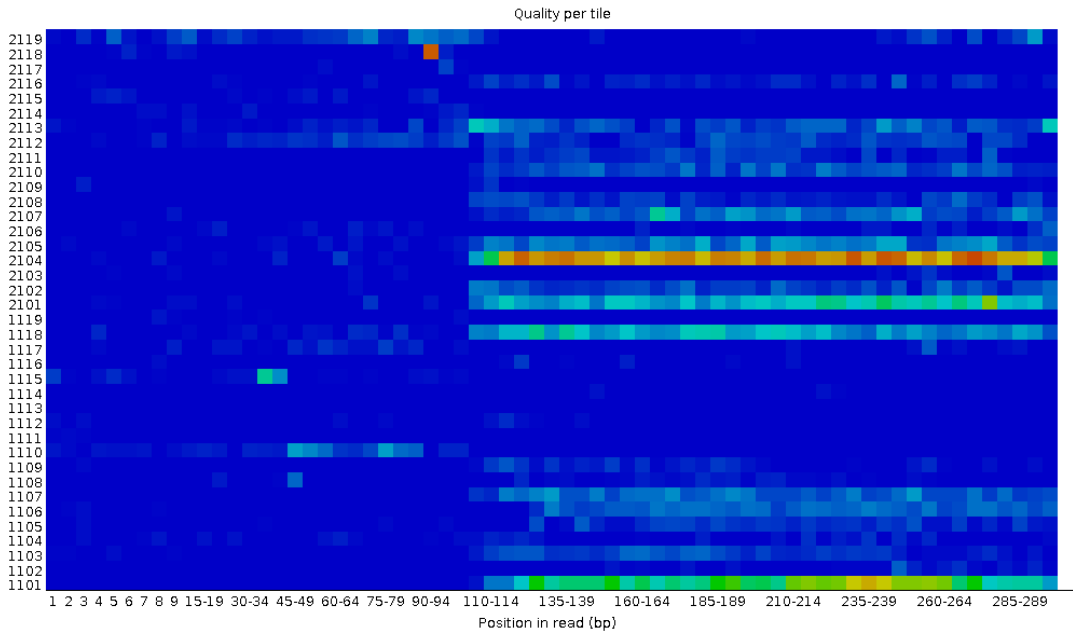
Trimmed



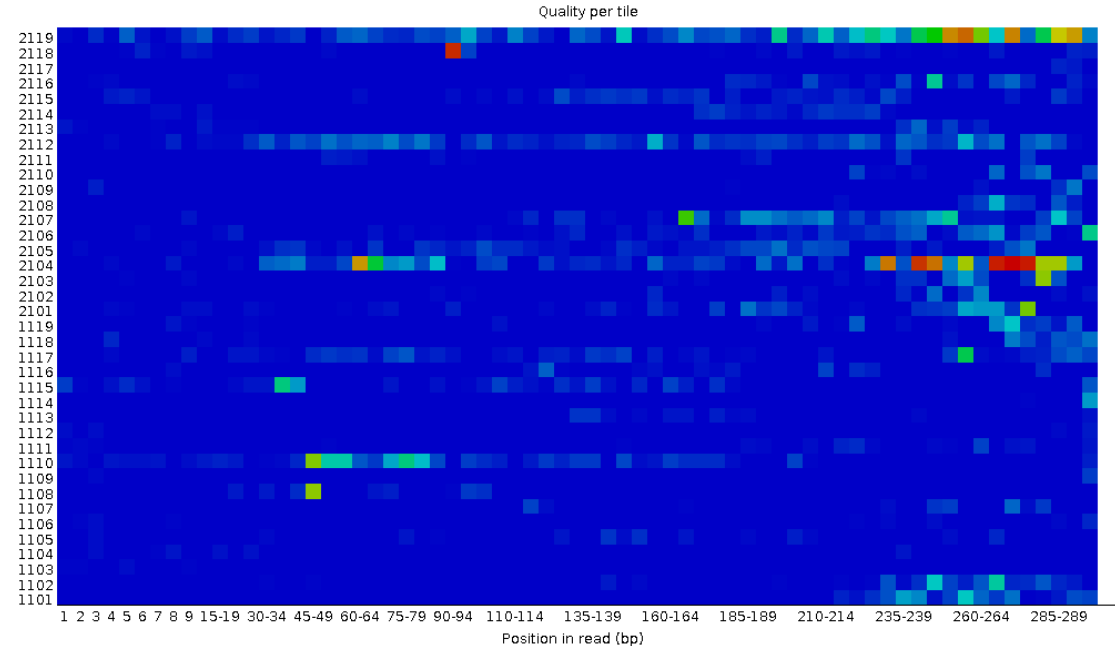
This graph shows the Phred quality scores for each base in the sequence (296 bases for us). The yellow box represents the middle 50% of reads (25th to 75th percentile), and the whiskers show the 10th to 90th percentile. The red line is the median, and the blue/black line is the mean. In the raw (untrimmed) data, the quality starts high for the first 100 bases but drops sharply toward the end. This is common in Illumina data because of signal decay or phasing issues during sequencing. The raw data failed quality checks because after around base 110, the median and 25th percentile scores drop below 12, which is really bad. After trimming, though, the quality improves a lot, with most reads having scores above 20, meaning the data is now good.

Per tile sequence quality

untrimmed



trimmed

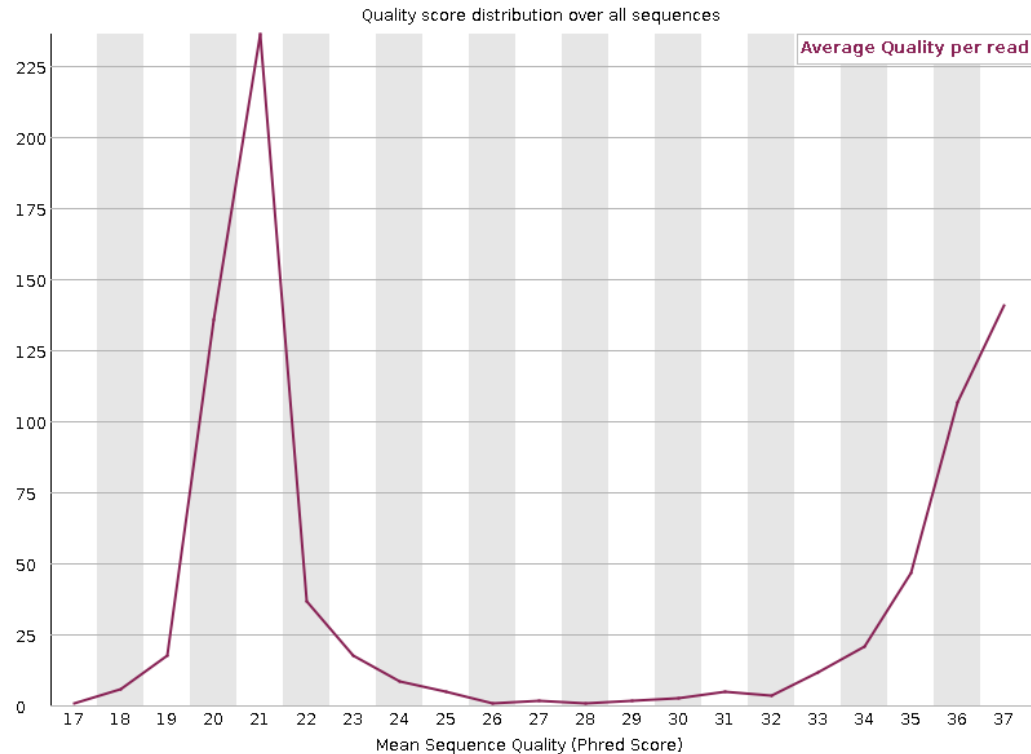


This graph shows how well the flow cell performed during sequencing and how accurate the base calling was. If the quality is bad, the tiles turn red, but if it's good, they stay blue. Ideally, we'd want the whole graph to be blue.

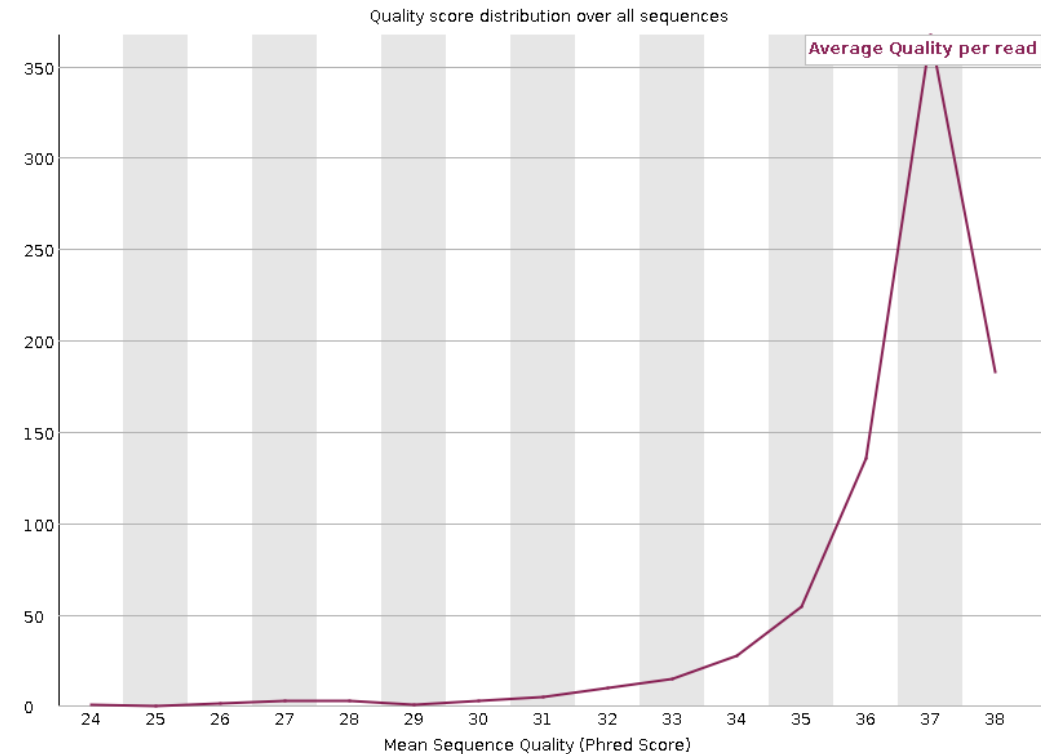
In our data, the untrimmed sequences gave a warning, meaning the quality wasn't perfect but still decent. Surprisingly, the trimmed data failed, which is unusual since trimming usually improves quality. This might mean something went wrong during trimming or the data had issues that trimming couldn't fix.

Per sequence quality scores

untrimmed



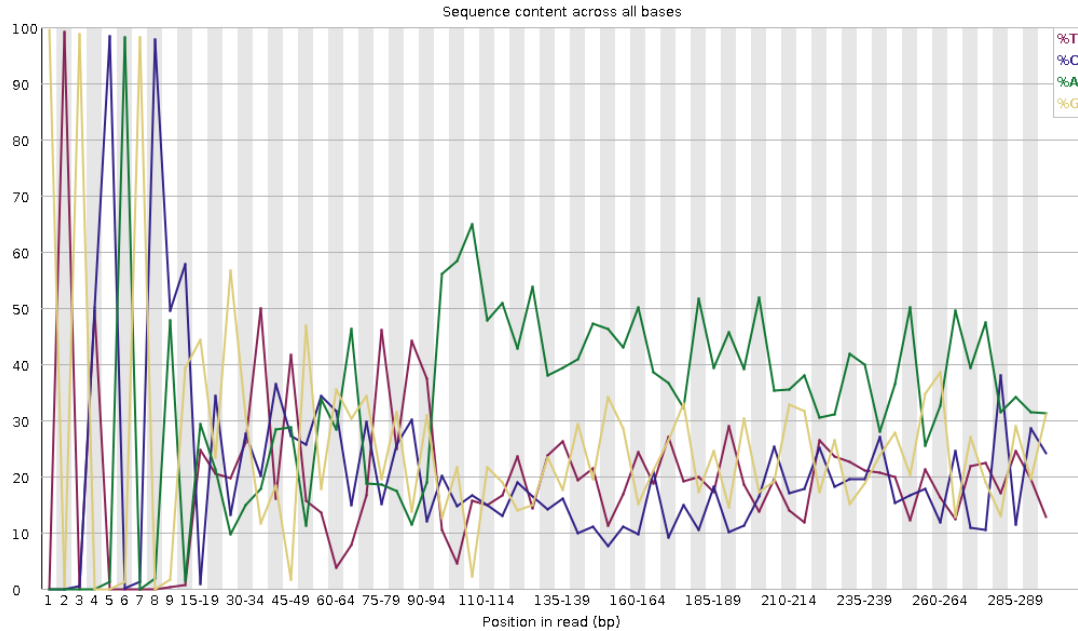
trimmed



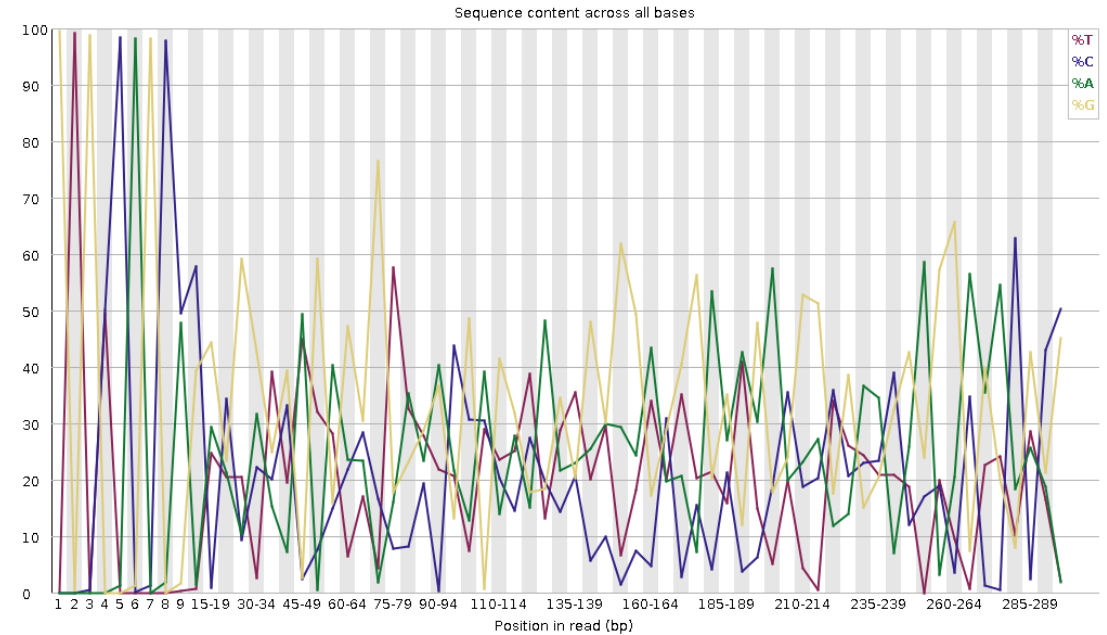
This graph shows the quality scores for all sequences, with the y-axis as the number of reads and the x-axis as their mean quality score. A good result has peaks on the upper right, meaning most reads have high scores. In the untrimmed data, the main peak is around 21, which isn't great, but there's a smaller peak at the high end where about 130 reads score 37 or higher. The trimmed data is much better, with over 350 reads clustered around the highest score of 37, showing a big improvement in quality.

Per base sequence content

untrimmed



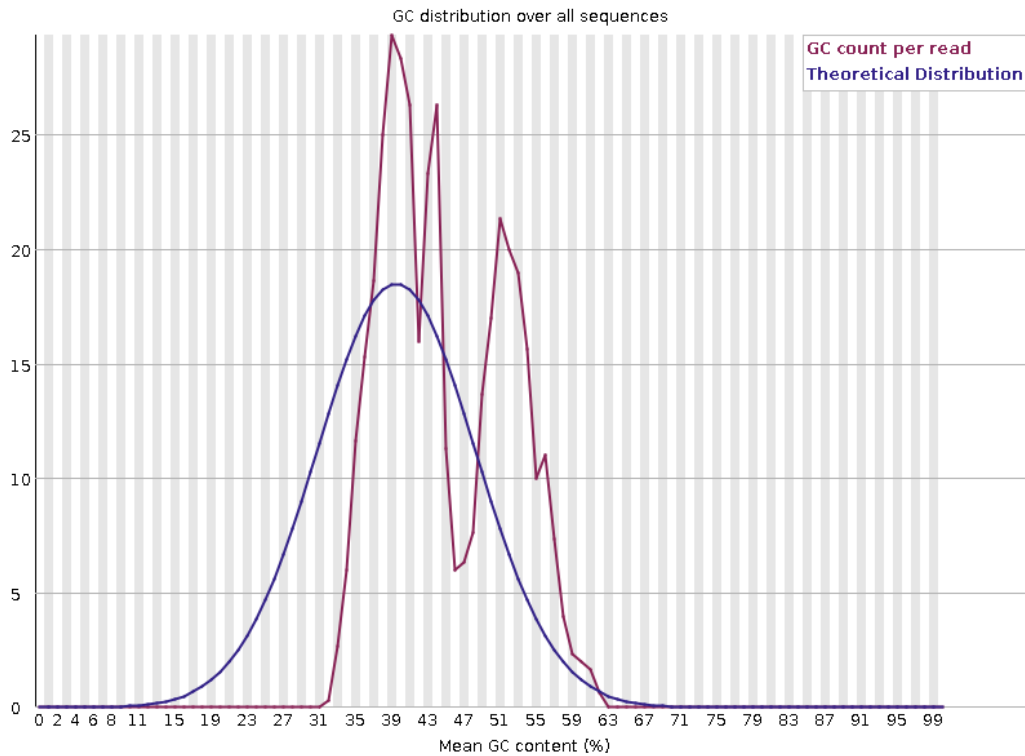
trimmed



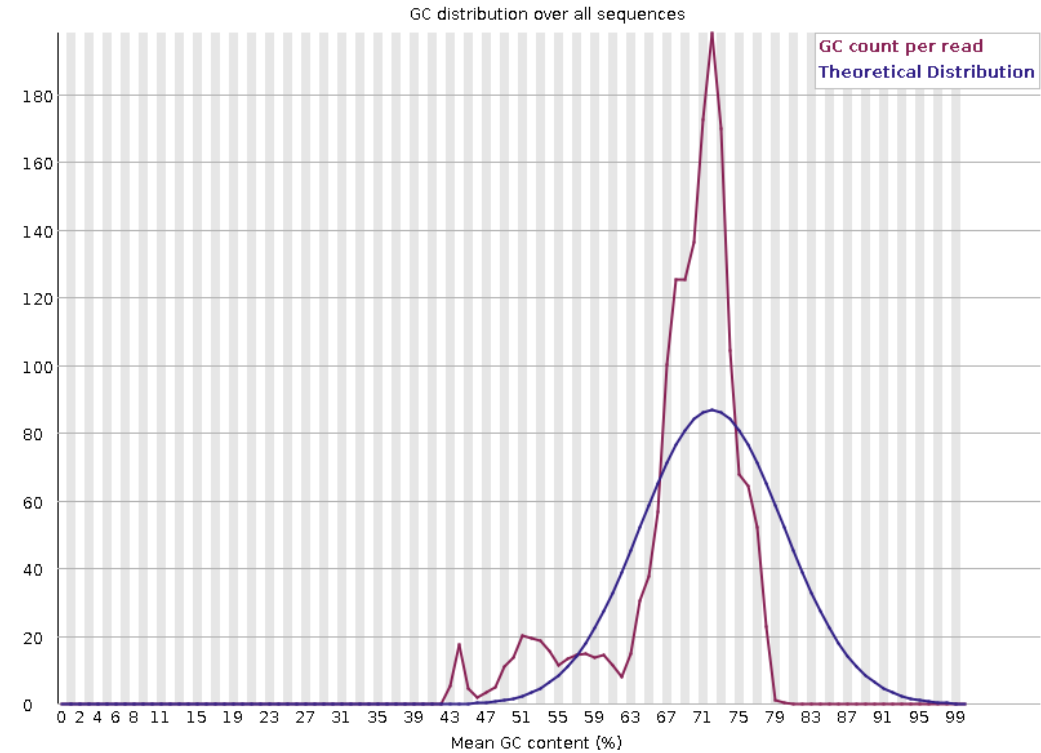
This graph shows the percentage of A, T, C, and G at each position in the reads. Ideally, A and T should be roughly equal, and so should G and C. Big peaks or dips might mean contamination, like from PCR. Since our data is 16S amplicon (PCR-amplified), there's expected bias, and adapters can also cause bias in the first few bases. Both the untrimmed and trimmed graphs show many positions where A&T or G&C differ by more than 20%, which is why FASTQC gives a failure for base content.

Per sequence GC content

untrimmed



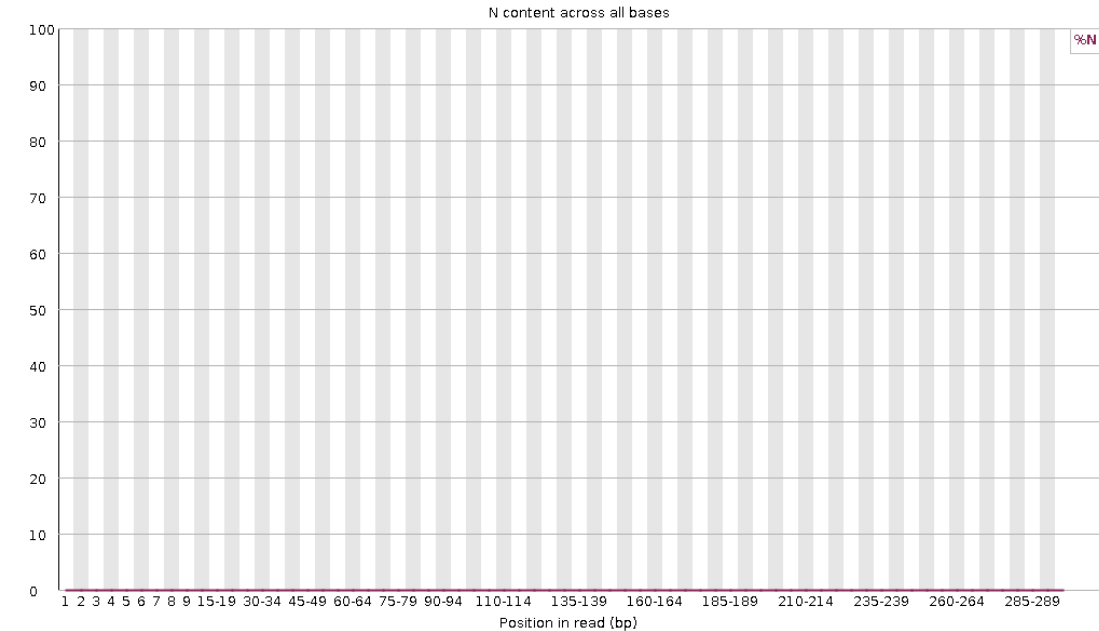
trimmed



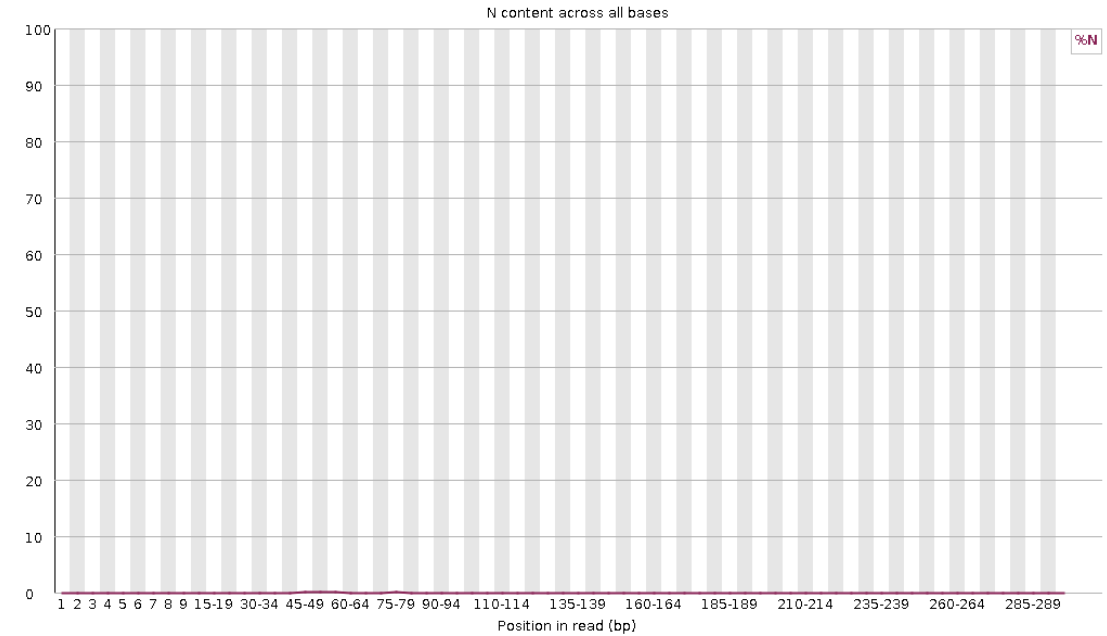
This graph shows the number of reads against their GC percentage. For whole genome sequencing (WGS), we'd expect a normal distribution peaking at the genome's GC content, like 40-44% for humans. If the genome's GC content isn't known, FASTQC uses the most common GC value from the data to create a theoretical distribution. In our untrimmed data, there's a double peak in GC content, which could mean contamination from adapters or overrepresented sequences. The trimmed data shows less of a double peak, suggesting adapter removal helped. However, since over 30% of reads deviate from the expected distribution, both untrimmed and trimmed data fail the GC content check.

Per base N content

Untrimmed



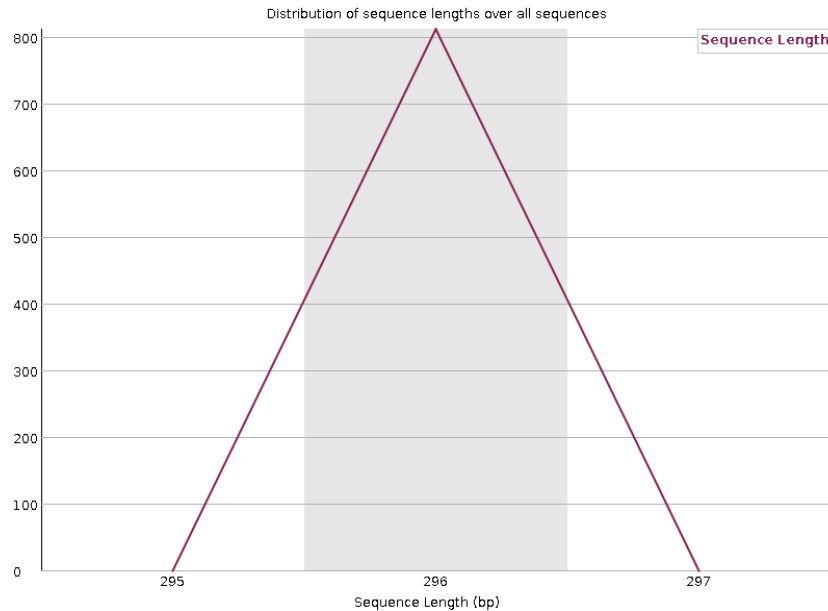
trimmed



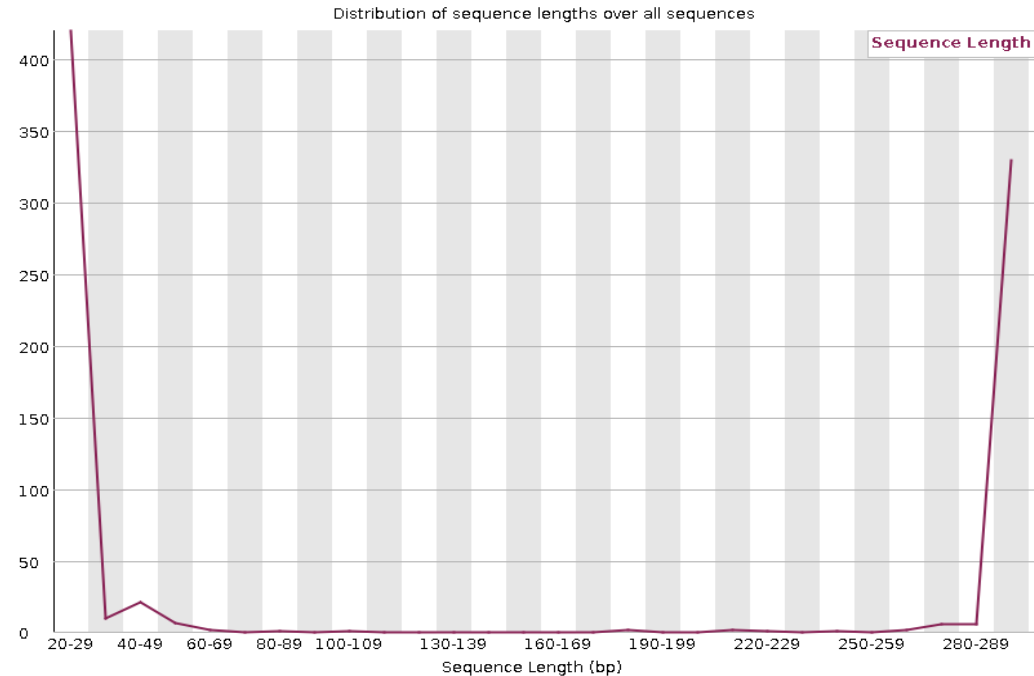
This graph shows the percentage of ambiguous bases (N) in the reads, which occur when the sequencer can't determine the nucleotide due to issues like noise. A good read should have no N content. In our data, both before and after trimming, there are no Ns, so this part of the analysis passes as expected.

Sequence Length Distribution

untrimmed



trimmed



This graph shows the distribution of read lengths, with the x-axis representing the length and the y-axis showing the number of reads. For the untrimmed data, all reads are 296 bases long, which is okay. However, the trimmed data gives a warning because the lengths vary, with two peaks—one at the start (shorter reads) and one at the end (full-length reads). This happens because trimming removes low-quality bases and adapters, leaving reads of different lengths. The gap in between likely represents reads that were completely removed during trimming.

Overrepresented sequences

Untrimmed

| Sequence | Count | Percentage | Possible Source |
|---|-------|--------------------|-----------------|
| GTGTCAGCCGCCGCGGTAGTCCGACGTGGCTGTCTCTTATACACATCTCC | 46 | 5.665024630541872 | No Hit |
| GTGCCAGCAGCCGCGGTAGTCCGACGTGGCTGTCTCTTATACACATCTCC | 41 | 5.0492610837438425 | No Hit |
| GTGCCAGCCGCCGCGGTAGTCCGACGTGGCTGTCTCTTATACACATCTCC | 39 | 4.80295566502463 | No Hit |
| GTGTCAGCAGCCGCGGTAGTCCGACGTGGCTGTCTCTTATACACATCTCC | 33 | 4.064039408866995 | No Hit |
| GTGCCAGCAGCCGCGGTAGTCCGACGTGGCTGTCTCTTATACACATCTCCG | 23 | 2.832512315270936 | No Hit |
| GTGTCAGCCGCCGCGGTAGTCCGACGTGGCTGTCTCTTATACACATCTCCG | 23 | 2.832512315270936 | No Hit |
| GTGCCAGCAGCCGCGGTAATACGAGGGTGCAAGCGTTAATCGGAATTAC | 17 | 2.0935960591133003 | No Hit |
| GTGCCAGCCGCCGCGTAGTCCGACGTGGCTGTCTCTTATACACATCTCCG | 16 | 1.9704433497536946 | No Hit |
| GTGTCAGCAGCCGCGGTAATACGTAGTGGAAGCGTTATCCGGAATTAT | 16 | 1.9704433497536946 | No Hit |
| GTGTCAGCCGCCGAGTAGTCCGACGTGGCTGTCTCTTATACACATCTCCG | 15 | 1.8472906403940887 | No Hit |
| GTGCCAGCCGCCGAGTAGTCCGACGTGGCTGTCTCTTATACACATCTCCG | 13 | 1.600985221674877 | No Hit |
| GTGTCAGCAGCCGCGGTAATACGAGGGTGCAAGCGTTATCCGATTAT | 13 | 1.600985221674877 | No Hit |
| GTGTCAGCCGCCGCGGTAATACGAGGGTGCAAGCGTTATCCGATTAT | 13 | 1.600985221674877 | No Hit |
| GTGTCAGCCGCCGCGGTAATACGTAGTGGAAGCGTTATCCGGAATTAT | 12 | 1.477832512315271 | No Hit |
| GTGCCAGCAGCCGCGGTAATACGAGGGTGCAAGCGTTATCCGATTAT | 12 | 1.477832512315271 | No Hit |
| GTGCCAGCCGCCGCGGTAATACGAGGGTGCAAGCGTTATCCGATTAT | 12 | 1.477832512315271 | No Hit |
| GTGTCAGCAGCCGCGTAGTCCGACGTGGCTGTCTCTTATACACATCTCC | 11 | 1.354679802955665 | No Hit |
| GTGTCAGCCGCCGCGTAGTCCGACGTGGCTGTCTCTTATACACATCTCC | 11 | 1.354679802955665 | No Hit |
| GTGCCAGCCGCCGCGTAGTCCGACGTGGCTGTCTCTTATACACATCTCC | 10 | 1.2315270935960592 | No Hit |

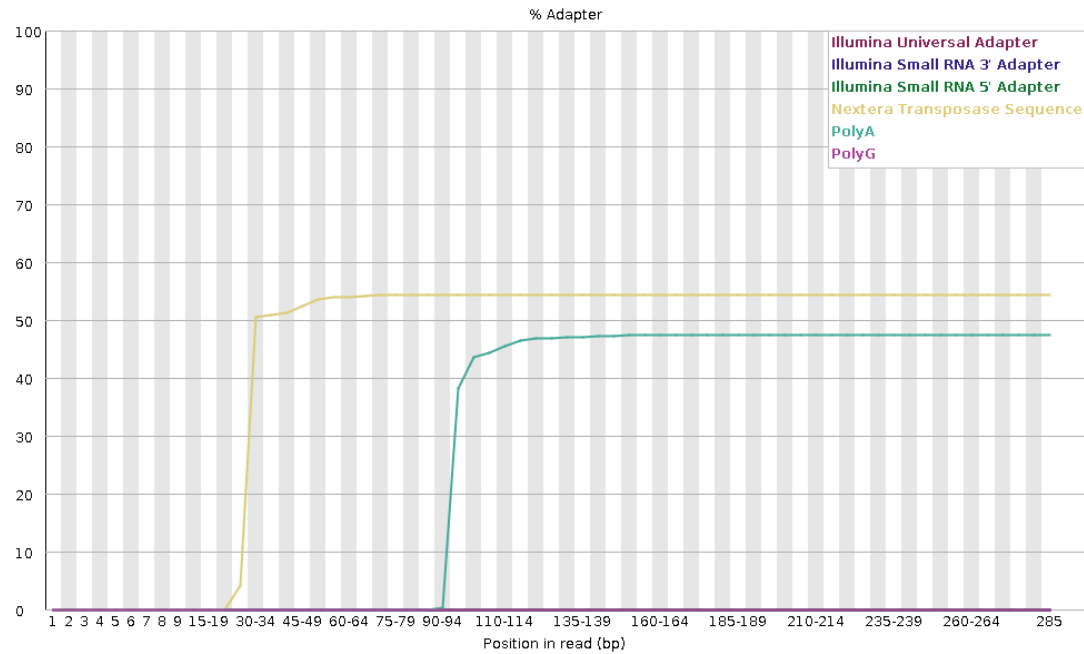
trimmed

| Sequence | Count | Percentage | Possible Source |
|---|-------|--------------------|-----------------|
| GTGTCAGCCGCCGCGTAGTCCGACGTGG | 47 | 5.788177339901478 | No Hit |
| GTGCCAGCAGCCGCGTAGTCCGACGTGG | 44 | 5.41871921182266 | No Hit |
| GTGCCAGCCGCCGCGTAGTCCGACGTGG | 42 | 5.172413793103448 | No Hit |
| GTGTCAGCAGCCGCGTAGTCCGACGTGG | 35 | 4.310344827586207 | No Hit |
| GTGCCAGCAGCCGCGTAGTCCGACGTGG | 25 | 3.0788177339901477 | No Hit |
| GTGTCAGCCGCCGCGTAGTCCGACGTGG | 24 | 2.955665024630542 | No Hit |
| GTGCCAGCCGCCGCGTAGTCCGACGTGG | 17 | 2.0935960591133003 | No Hit |
| GTGCCAGCAGCCGCGTAATACGGAGGGTGCAAGCGTTAATCGGAATTAC | 17 | 2.0935960591133003 | No Hit |
| GTGTCAGCAGCCGCGTAATACGTAGTGGCAAGCGTTATCCGGAATTAT | 16 | 1.9704433497536946 | No Hit |
| GTGTCAGCCGCCGAGTAGTCCGACGTGG | 15 | 1.8472906403940887 | No Hit |
| GTGCCAGCCGCCGAGTAGTCCGACGTGG | 14 | 1.7241379310344827 | No Hit |
| GTGTCAGCAGCCGCGTAATACGGAGGGTGCAAGCGTTATCCGGATTAT | 13 | 1.600985221674877 | No Hit |
| GTGTCAGCCGCCGCGTAATACGGAGGGTGCAAGCGTTATCCGGATTAT | 13 | 1.600985221674877 | No Hit |
| GTGTCAGCCGCCGCGTAATACGTAGTGGCAAGCGTTATCCGGAATTAT | 12 | 1.477832512315271 | No Hit |
| GTGTCAGCCGCCGAGTAGTCCGACGTGG | 12 | 1.477832512315271 | No Hit |
| GTGCCAGCAGCCGCGTAATACGGAGGGTGCAAGCGTTATCCGGATTAT | 12 | 1.477832512315271 | No Hit |
| GTGTCAGCAGCCGAGTAGTCCGACGTGG | 12 | 1.477832512315271 | No Hit |
| GTGCCAGCCGCCGCGTAATACGGAGGGTGCAAGCGTTATCCGGATTAT | 12 | 1.477832512315271 | No Hit |
| GTGCCAGCCGCCGAGTAGTCCGACGTGG | 10 | 1.2315270935960592 | No Hit |

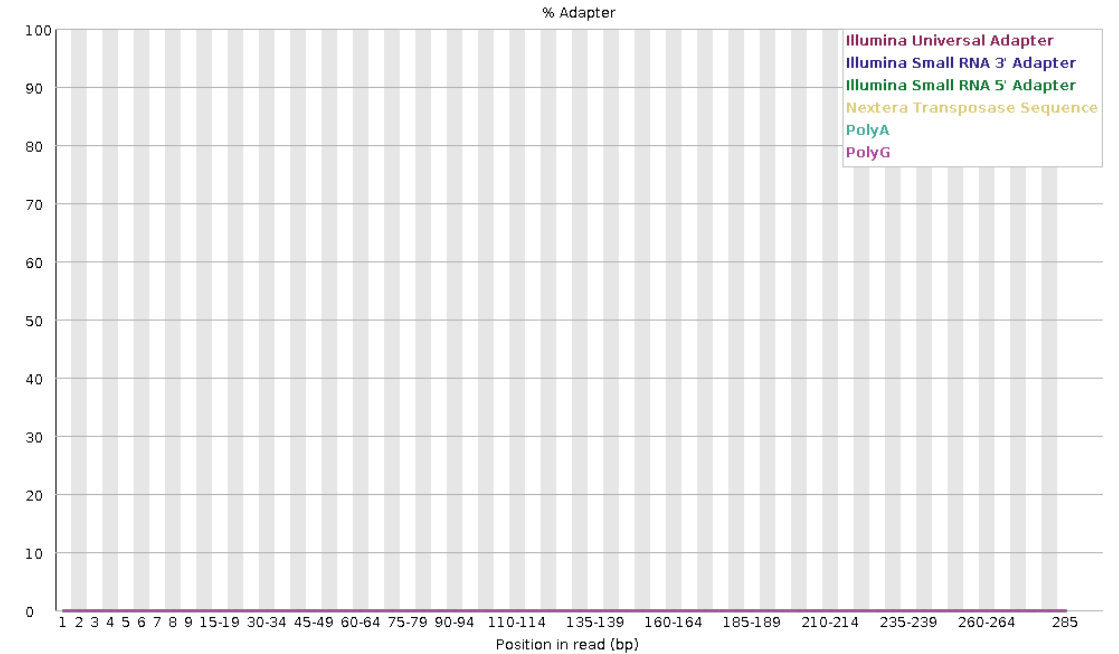
This table lists sequences that make up 0.1% or more of the library. For example, the first sequence appears in 5.66% of untrimmed reads and 5.78% of trimmed reads. Since none of these overrepresented sequences match any known sources in the FastQC database (likely due to the long list not being fully shown), they could be adapters, contaminants, or highly abundant biological sequences not in the database. Because some sequences make up more than 1% of the reads, both the trimmed and untrimmed data fail this check.

Adapter Content

untrimmed

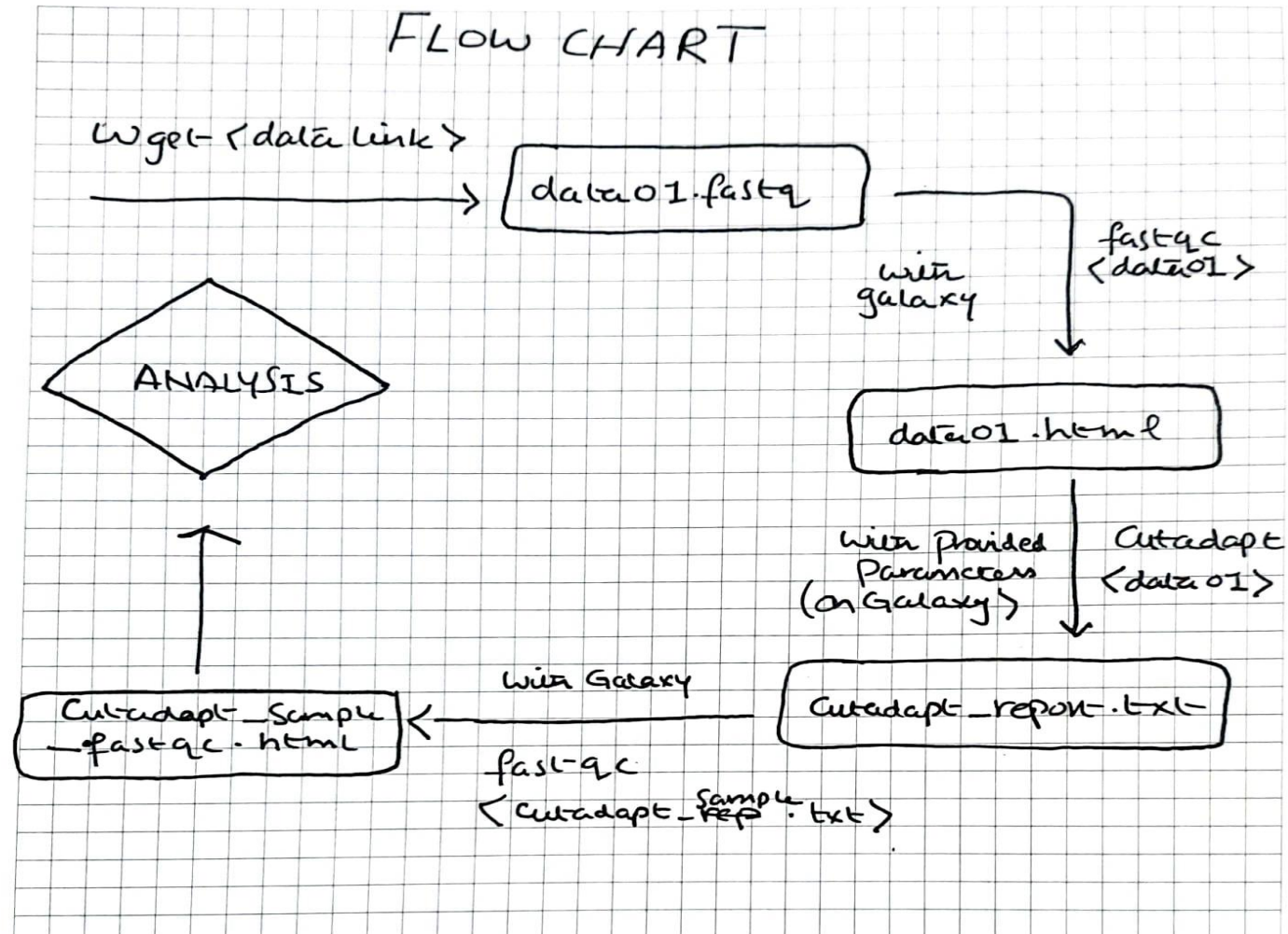


trimmed








This graph shows whether adapters are present in the sequencing data. The y-axis is the percentage of reads with an adapter, and the x-axis shows the position in the read. In the raw data, the Nextera transposase sequence and Illumina Small RNA 5' adapter are present, but after trimming, they're completely gone. This helps decide if trimming is needed. FASTQC fails the check if over 10% of reads have adapters, which happened with our raw data. After trimming, the adapter issue is resolved

Flow Chart for Question 1 :



Question 2 :

General Statistics

|  Copied! |  Configure columns |  Scatter plot |  Violin plot | Export as CSV... | Showing $2\frac{1}{2}$ rows and $6\frac{1}{6}$ columns. |  Summarize table |
|---|---|--|---|------------------|---|---|
| Sample Name | Dups ▾ | GC | Avg len | Median len | Failed | Seqs |
| GSM461178_untreat_paired_subset_1_fastq | 8.8% | 53.0% | 37 bp | 37 bp | 0% | 0.1 M |
| GSM461178_untreat_paired_subset_2_fastq | 7.3% | 53.0% | 37 bp | 37 bp | 9% | 0.1 M |

This is the overall MultiQC summary for our untrimmed paired-end data, where subset 1 is the forward read and subset 2 is the reverse. At first glance, the forward and reverse reads look similar, except the forward read has 1.5% more duplication. However, the reverse read failed 9% of the FASTQC modules, which suggests some quality issues.

Sequence Counts

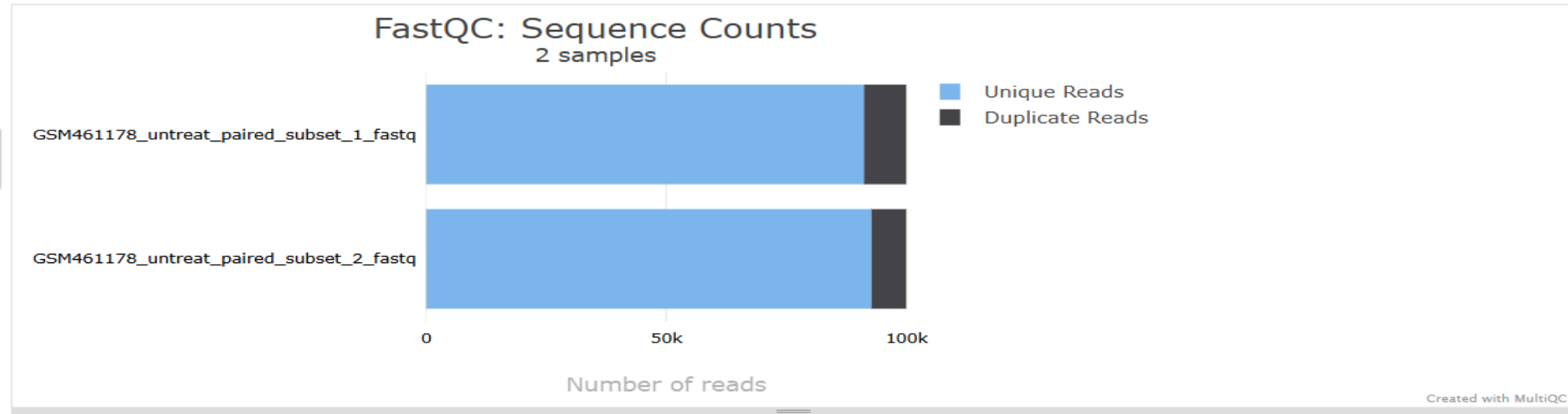
[Help](#)

Sequence counts for each sample. Duplicate read counts are an estimate only.

Percentages

Summarize plot

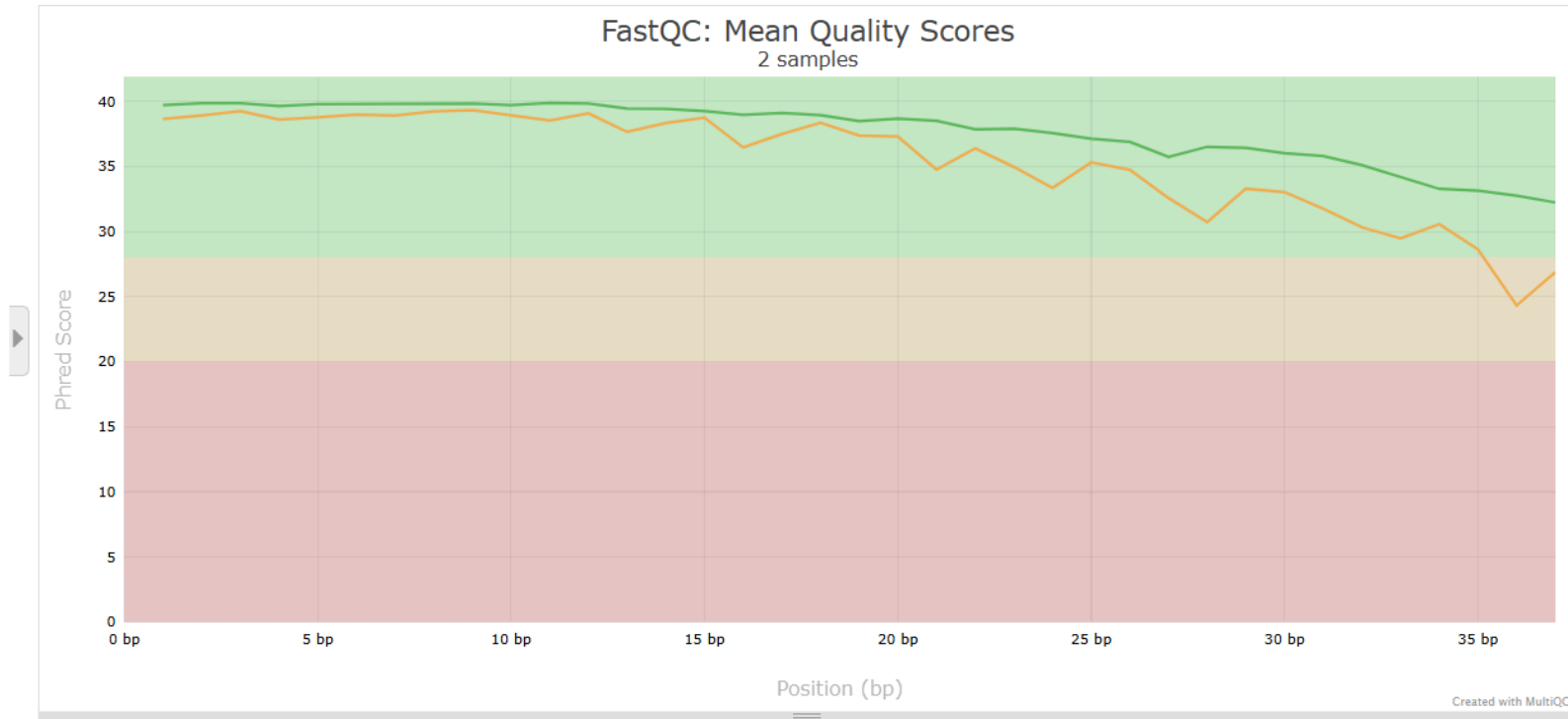
Export...



This graph shows the total number of reads, split into unique and duplicate reads. There's a small difference between the forward and reverse sequences, with the reverse having slightly fewer duplicates. This suggests the reverse reads might have less redundancy compared to the forward reads.

Sequence Quality Histograms

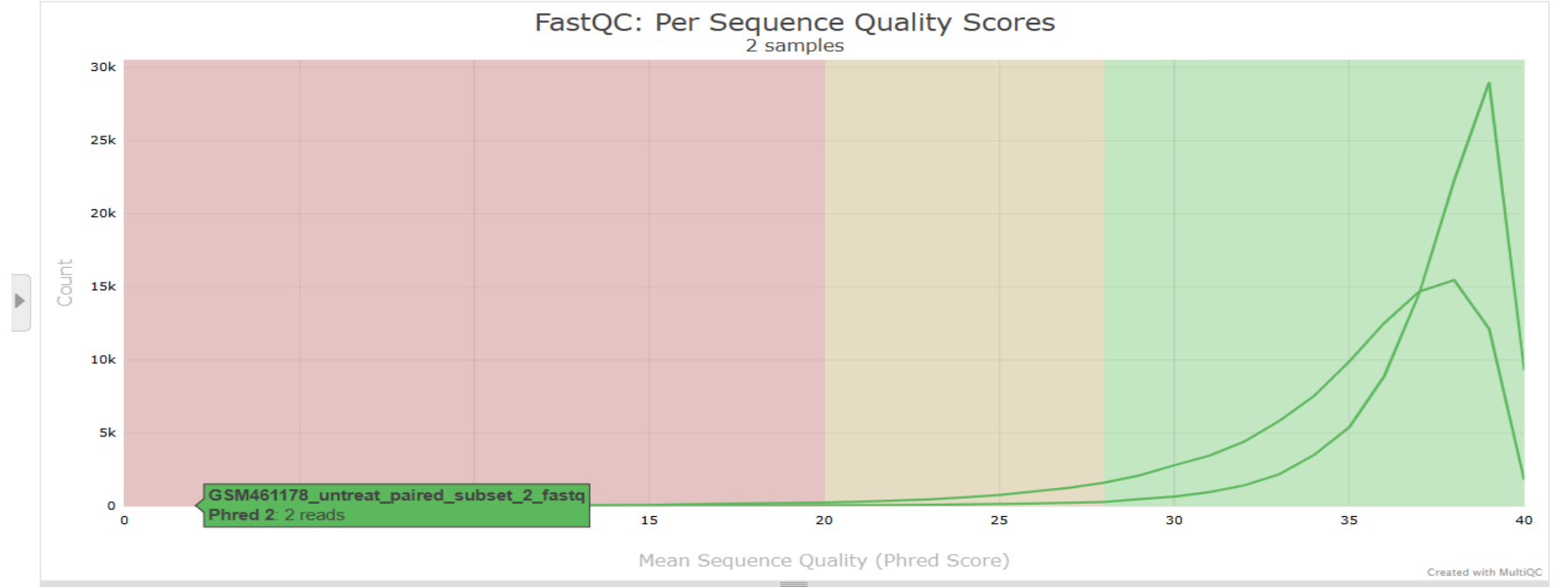
The mean quality value across each base position in the read.

[Help](#)[Summarize plot](#)[Export...](#)

This graph compares the mean quality scores of the forward (R1, green line) and reverse (R2, orange line) sequences at each base position. Both sequences have decent overall scores, but the reverse sequence lags behind, especially toward the end. While it's normal for quality to drop at the ends, the reverse sequence falls into the "reasonable" range (below 29, orange area) by the end, showing a more significant drop in quality compared to the forward read.

Per Sequence Quality Scores

The number of reads with average quality scores. Shows if a subset of reads has poor quality.

[Help](#)[Summarize plot](#)[Export...](#)

This graph shows the number of reads (y-axis) at each quality score (x-axis). The forward sequence (R1) has higher overall quality, with a larger peak pushed toward higher scores, while the reverse sequence (R2) has lower quality. For example, around 28k reads in R1 have a score of about 37, compared to only 15k in R2 with the same score. This shows R1 has better quality overall.

Per Sequence GC Content

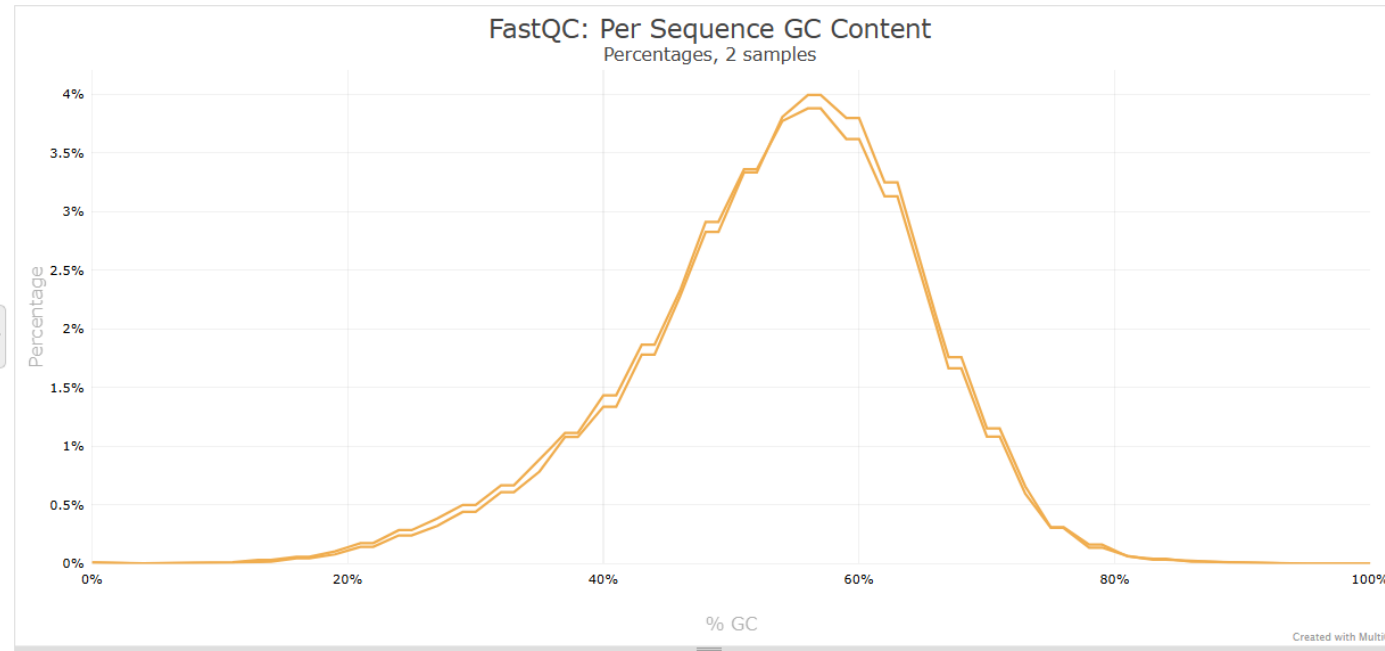
[Help](#)

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content.

Percentages Counts

Summarize plot

Export...

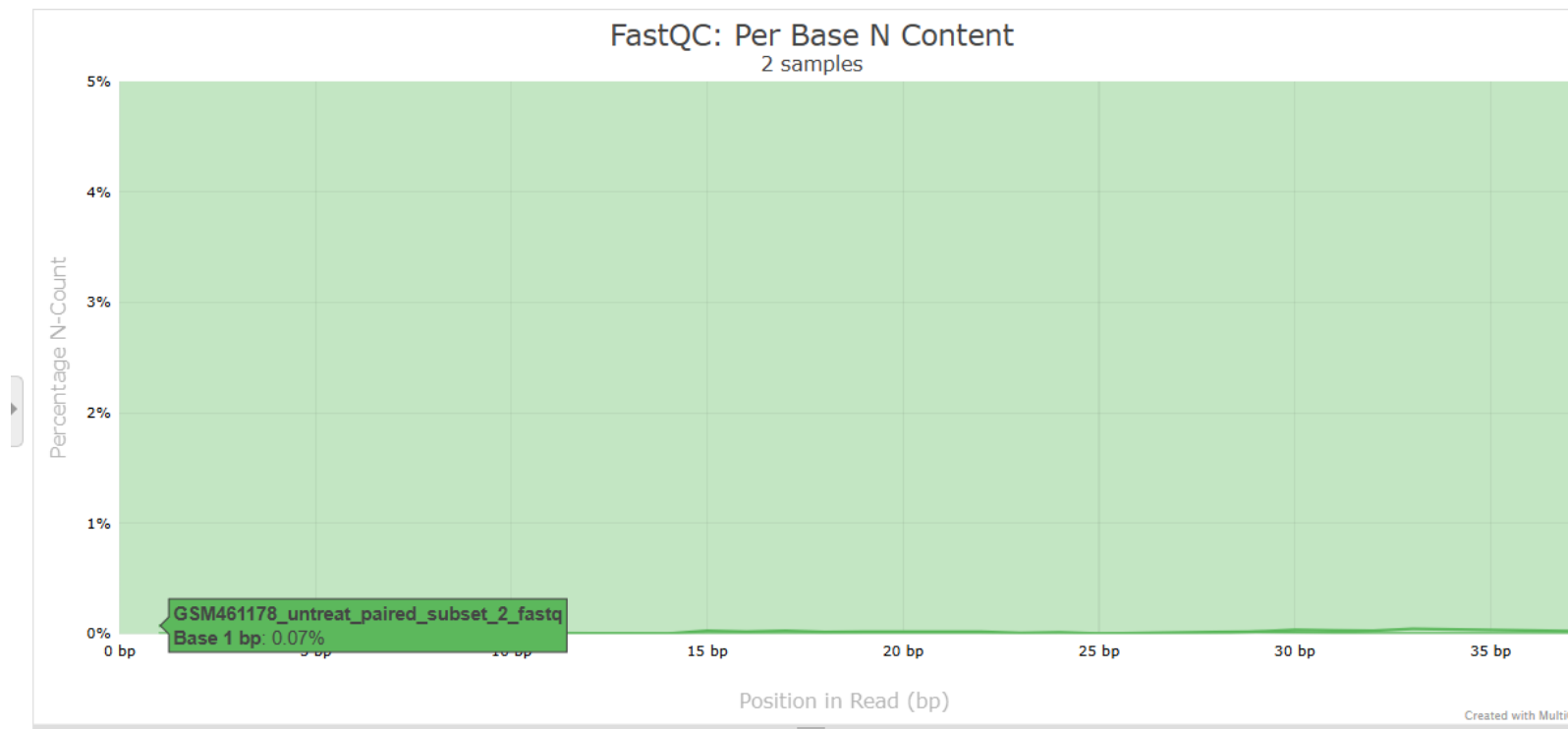


This graph compares the average GC content of the forward (R1) and reverse (R2) reads. Both sequences have very similar GC content, but R1 is slightly higher, ranging around 56-63%, while R2 is just a bit lower. This shows the GC content is fairly consistent between the two.

Per Base N Content


[Help](#)

The percentage of base calls at each position for which an **N** was called.

[Summarize plot](#)[Export...](#)

For N content and sequence length, R1 and R2 are almost identical. The only small difference is that R2 has a 0.07% N count at one base position, while R1 has none. Other than that, both sequences show no significant differences in these aspects.

Adapter Content

 Help

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

No samples found with any adapter contamination > 0.1%

In terms of adapter content, there's no difference between the sequences, and no adapter sequence was detected in either R1 or R2

```
This is cutadapt 5.0 with Python 3.12.8
Command line parameters: -j=8 --error-rate=0.1 --times=1 --overlap=3 --action=trim --quality-cutoff=20 --minimum-length=20 -o out1.fq
-p out2.fq GSM461178_untreat_paired_subset_1_fastq.fq GSM461178_untreat_paired_subset_2_fastq.fq
Processing paired-end reads on 8 cores ...

=== Summary ===

Total read pairs processed:          100,000

== Read fate breakdown ==
Pairs that were too short:           1,376 (1.4%)
Pairs written (passing filters):     98,624 (98.6%)

Total basepairs processed:          7,400,000 bp
  Read 1:      3,700,000 bp
  Read 2:      3,700,000 bp
Quality-trimmed:                    182,802 bp (2.5%)
  Read 1:      44,164 bp
  Read 2:      138,638 bp
Total written (filtered):           7,159,132 bp (96.7%)
  Read 1:      3,616,660 bp
  Read 2:      3,542,472 bp
```

The cutadapt analysis shows that 182,802 bp were removed due to low quality, with R2 reads contributing much more (138,638 bp) compared to R1 (44,164 bp). This matches the lower quality of R2 seen in the MultiQC report. Also, 1,376 read pairs (1.4% of the total) were removed because they were too short after trimming.

Flow Chart for Question 2 :

