# Comparison of a low quality and a good quality sample based on fastp reports

To better understand what distinguishes a good quality sample from a low quality sample, we compared two fastp HTML reports:

- ERR13827992 → good quality (more than 30,000 reads after filtering)
- ERR13828009 → low quality (less than 30,000 reads after filtering)

We took the higher number of reads and the lowest to well see the differences.

We analysed both the summary tables and the interactive visualizations generated by fastp.

## 1. Statistics comparison

This section gives a quick overview of both samples before and after filtering. It includes useful information like the total number of reads, the quality of bases and the duplication rate.

| Metric | ERR13827992 (high quality) | ERR13828009 (low quality) |
|---|---|---|
| Total reads (before) | 1,023,766 | 3 ,530 |
| Total reads (after) | 1,018,146 | 3,490 |
| Reads passed filter | 99.45% | 98.87% |
| Q20 rate (after) | 93.99% | 93.01% |
| Q30 rate (after) | 90.94% | 89.64% |
| GC content (after) | 52.00% | 51.95% |
| Duplication rate | 48.76% | 13.14% |
| Insert size peak | 440 bp | 465 bp |

When comparing the high-quality sample and the low-quality sample, we observe that both have similar Q20 and Q30 rates. This means that the overall base quality is high in both cases, and the low-quality sample is not necessarily bad in terms of sequencing accuracy.

The GC content, which tells us how the percentage of bases that are either Guanine or Cytosine, is also very similar for both cases. That is normal and doesn't suggest any problem as for human derived samples like gut microbiota this percentage have to be between 45 and 60% to be considered as normal. If the GC content were too high, it could indicate contamination sequencing bias.

The percentage of reads passed filter shows the proportion of reads that were kept after fastp cleaned the dataset. Here again, both samples have very high values, meaning that the initial reads were already of good enough quality to be kept.

However, the main difference is in the total number of reads after filtering as we have a bit more than 1 million reads for the high-quality sample against less than 3,500 reads for the low-quality one.
This big gap is critical because bioinformatics tools like DADA2 need a minimum amount of data to work reliably.

Another interesting metric is the duplication rate, which tells us how many sequences are repeated. A high duplication rate might suggest that some sequences are over-represented, and many reads are identical. While a lower rate could mean more sequence diversity, which is good for microbiome analysis.

But this metric must always be interpreted in relation to the total number of reads. In samples with very few reads, a low duplication rate doesn't necessarily indicate high diversity. On the other hand, in samples with high sequencing depth, a moderate or even high duplication rate is less problematic since we still have enough unique reads to characterize the microbial composition.

So, a high duplication rate is not ideal, but it is less of a problem when the sample has a lot of reads. With enough total reads, we can still get a good variety of sequences even if some are repeated.

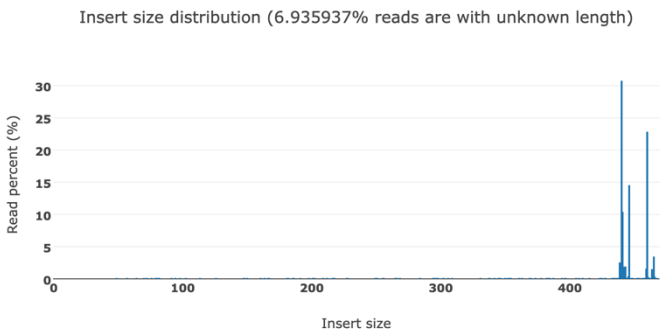
Insert size peak refers to the most common length of DNA fragments sequenced between the two paired-end reads. The values for the two samples are normal and consistent with what we except from 16S sequencing. O this metric is not concerning either.

Although Q scores and other metrics like GC content or duplication rate look acceptable, the total number of reads is the key factor that impacts downstream analysis. Tools like DADA2 require enough input data to perform well. That is why we decided to keep only the samples with more than 30,000 reads after filtering.

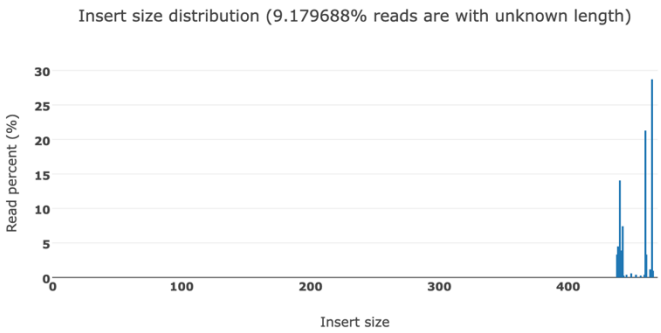
## 2. Insert size estimation

This section shows the distribution of insert sizes, which is the length of the DNA fragment between the two paired-end reads. It helps us check if the sequencing worked well.

In a good sequencing run, most fragments should have a similar size. This creates a clear peak in the insert size plot. It means that the DNA was properly fragmented and sequenced.




High quality




Low quality

When comparing the two samples:

- In both the high quality and low-quality samples, we can see a peak around 440-460 bp. This suggests that in both cases, the DNA fragments were sequenced with similar insert sizes.
- The general shape of the distribution is also quite similar between the two graphs, with a narrow peak on the right and almost no values elsewhere. This indicates that the fragmentation process was consistent.

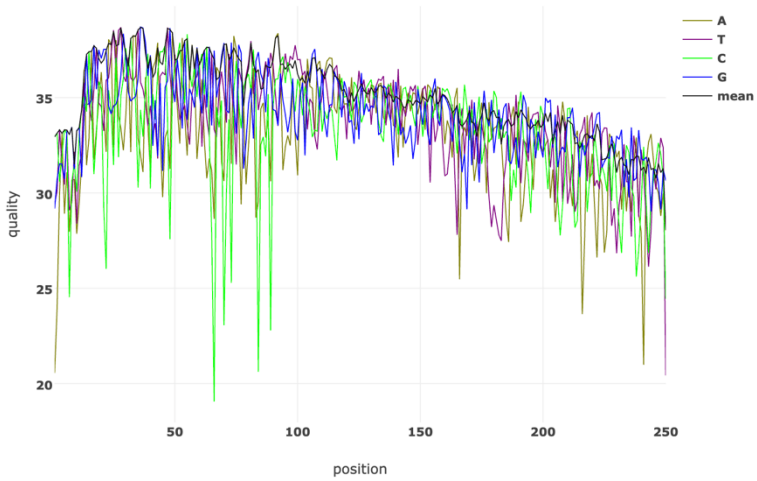
However, even though the low-quality sample seems to have a peak, it is based on a very small number of reads. This means the distribution is less reliable, and the peak may not represent the true distribution of fragment sizes.

This plot is useful to confirm that insert sizes are consistent, but in low quality samples with very few reads, we cannot trust the shape of the curve as much. This again supports our decision to only keep samples with enough total reads for further analysis.

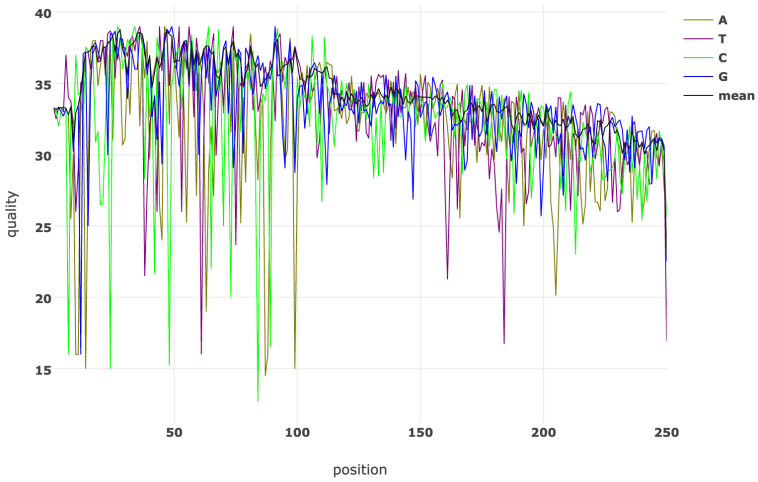
So, if we only had these graphs and no other information (like the number of reads), it would be very hard to tell which sample is high quality and which one is low quality. Both curves show a consistent insert size and a similar shape. That is why these graphs are helpful to check if the sequencing worked well, but they are not enough on their own to decide if sample should be kept or removed. For that, we also need to look at how many reads we have and how good their quality is.

### 3. Read 1 before filtering

#### a) Quality




High quality




Low quality

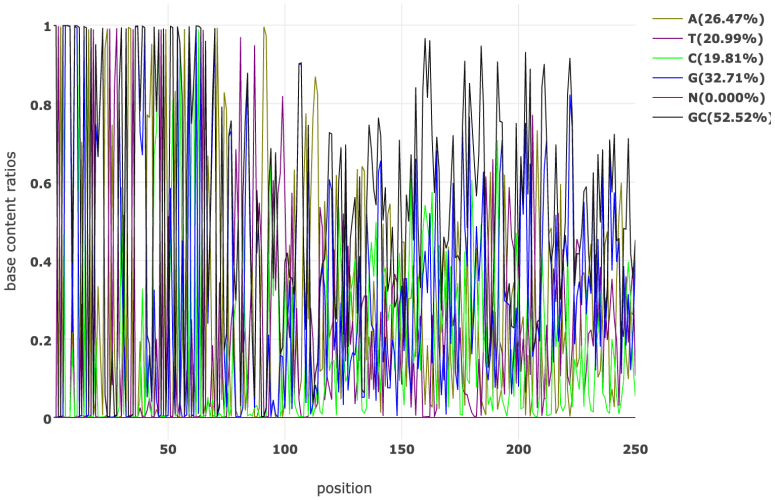
These plots show the quality of each base (A, T, C, G) at every position in the read, before any filtering. In the samples, we see a common pattern: the quality tends to be higher at the beginning and then it decreases until the end. This is normal and expected in most sequencing runs.

If we compare the two samples:

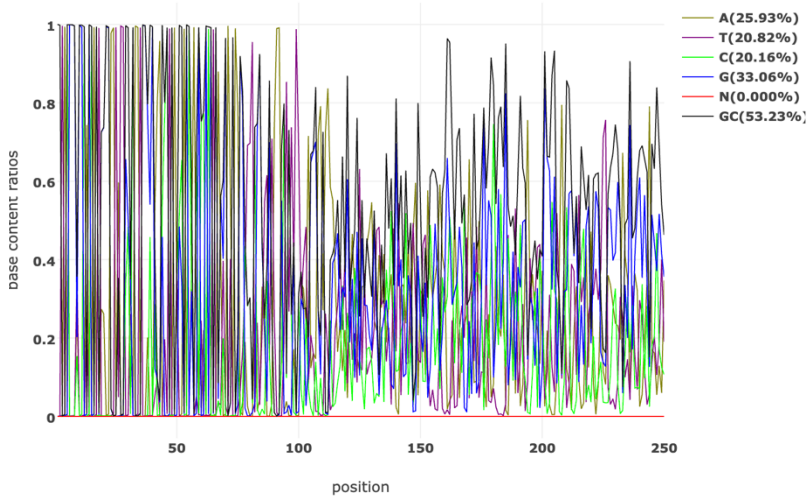
- The high-quality sample shows fewer drops in quality. Most base positions are above 30 in quality scores, which is very good.
- The low-quality sample has more irregular lines and more frequent drops below 30, especially in the second half of the read. This means the reads are a bit noisier.

Even though both samples follow the same general trend, the low-quality sample shows more fluctuations, which might affect downstream analysis. This visualization helps explain why some reads are removed during filtering.

### b) Base contents




High quality




Low quality

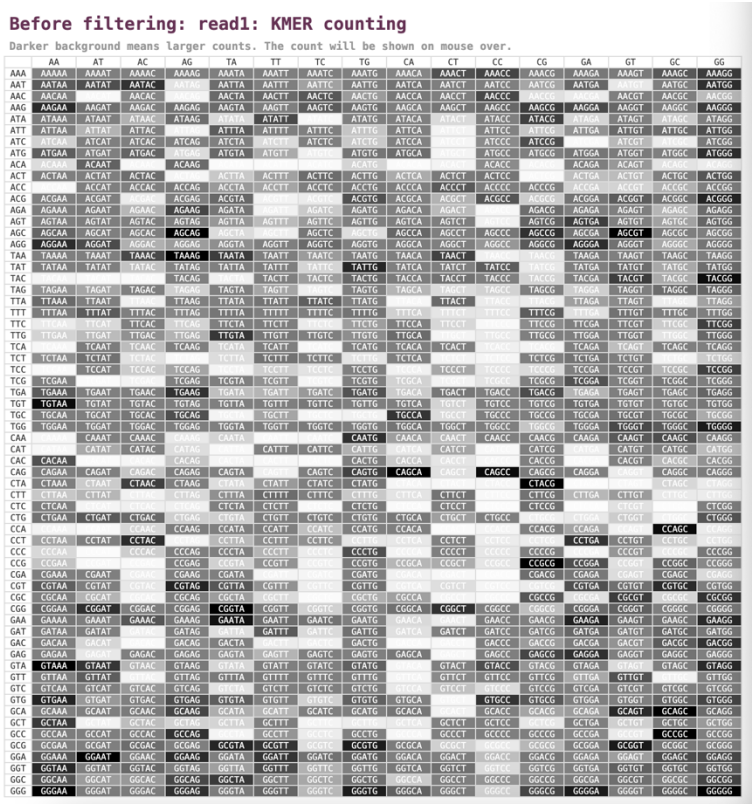
These plots show the base content across all positions in the reads before filtering. It tells us the proportion of each base (A, T, C, G) at every position from start to end of the sequences. It also shows the N content (red line) which represents undetermined bases, and the GC content (black line) which combines the percentage of G and C.

In both samples, we see a lot of variation in the beginning (first 80 positions). This is expected in amplicon sequencing because primers and barcodes can introduce sequence bias at the start. After this region, the curves become more stable.

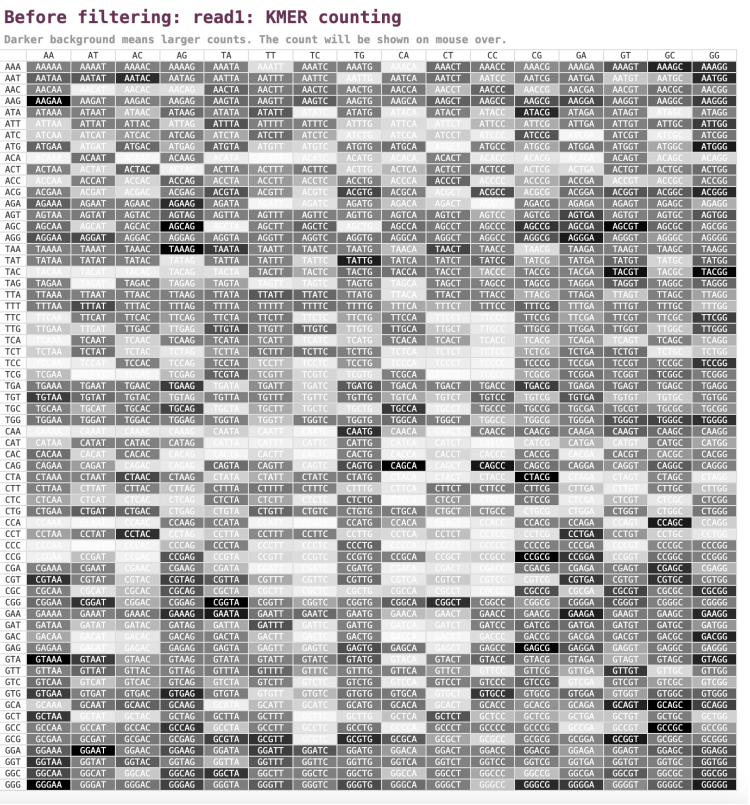
Also, the GC content is similar in both samples (52% and 53%), which is what we expect for human microbiome data. This means the base composition is normal and doesn't show any strange pattern.

Even if the low-quality sample has fewer reads, the base content still looks okay and similar to the high quality one. So again, this is more a problem about the number of reads and not the base composition.

### c) KMER counting




High quality




Low quality

This section shows the K-mer distribution, where a K-mer is a short sequence of K nucleotides. Here we have K=5 so each box in the figure corresponds to a specific 5-mer and the background color indicates how often that sequence was found in the reads:

- Dark background means that the 5-mer appeared very frequently
- Light background means that it was less common.

The goal of this plot is to help detect overrepresented sequences that may signal technical problems. For example: adapters or primers not fully trimmed, contaminating sequences, low diversity due to amplification bias.

In normal microbiome samples (especially after PCR amplification), it is expected to see some repeated sequences, but too much repetition or very dark isolated K-mers may be a sign of artifacts.

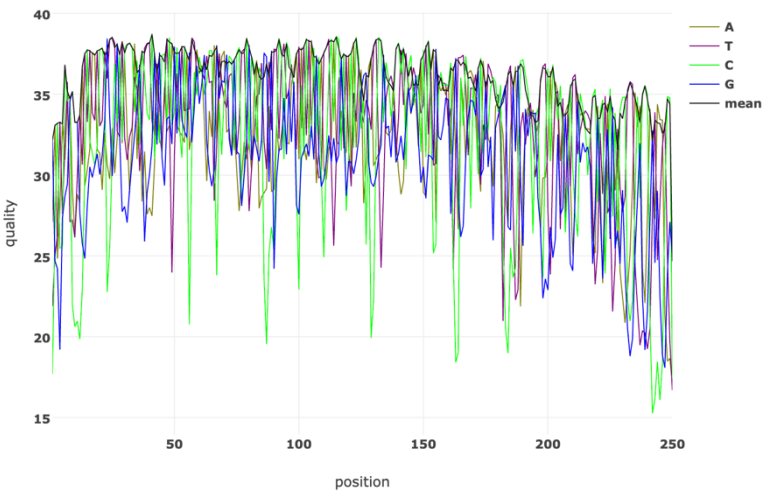
Here, when we compare the low-quality sample and the high quality one, we observe that both plots have some dark boxes, which means that a few K-mers are repeated often. And these over-represented sequences are common in amplicon based 16S sequencing and may reflect conserved regions (like the V3-V4 region being targeted), and not necessarily poor quality.

Between the two samples we don't see a big difference, they look quite similar and there are no strange or unexpected patters that appear in the low-quality sample.

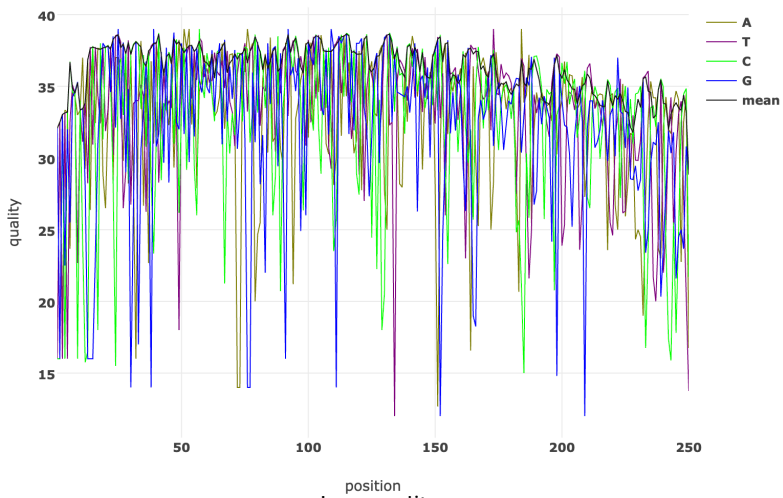
So, the K-mer plots do not explain the quality difference between the two samples. The low number of reads in the low-quality sample is probably due to other factors, but not to unusual or problematic sequence repetition.

## 4. Read 2 before filtering

### a) Quality




High quality



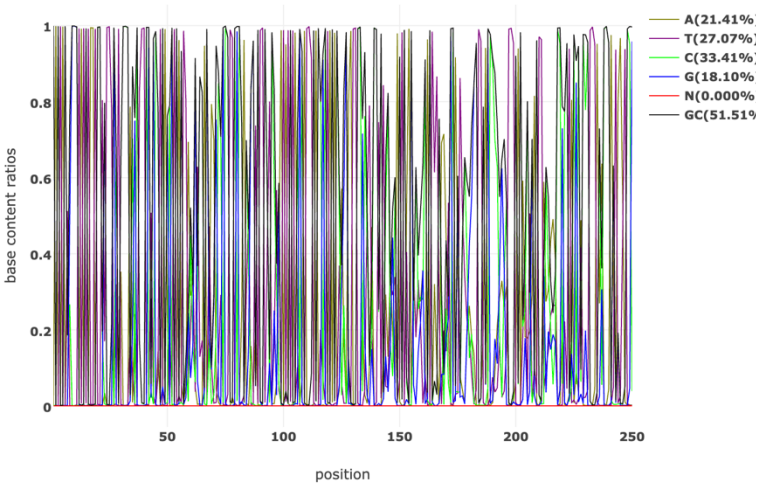

Low quality

In the high-quality sample, the curves are more regular and show fewer sharp drops in quality below 25. The average base quality remains high, between 35 and 38. This indicates that the sequencing was accurate across almost all positions.

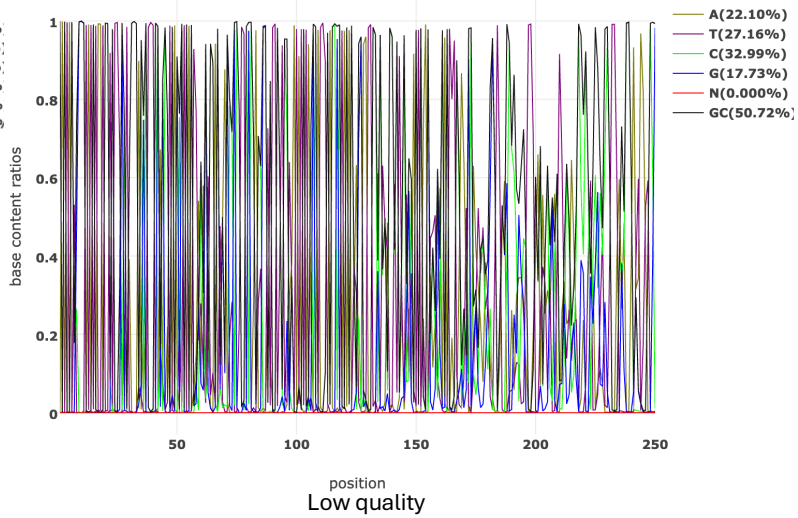
In the low-quality sample, the quality is more unstable. Several positions drop below 25 especially near the start and end of the reads. The overall average is still decent but there are many more fluctuations and low peaks than in the high-quality sample.

This differences in stability and consistency combined with the very low number of reads in this sample can justify why it was marked as a low quality and should be removed before downstream analysis.

### b) Base contents




High quality



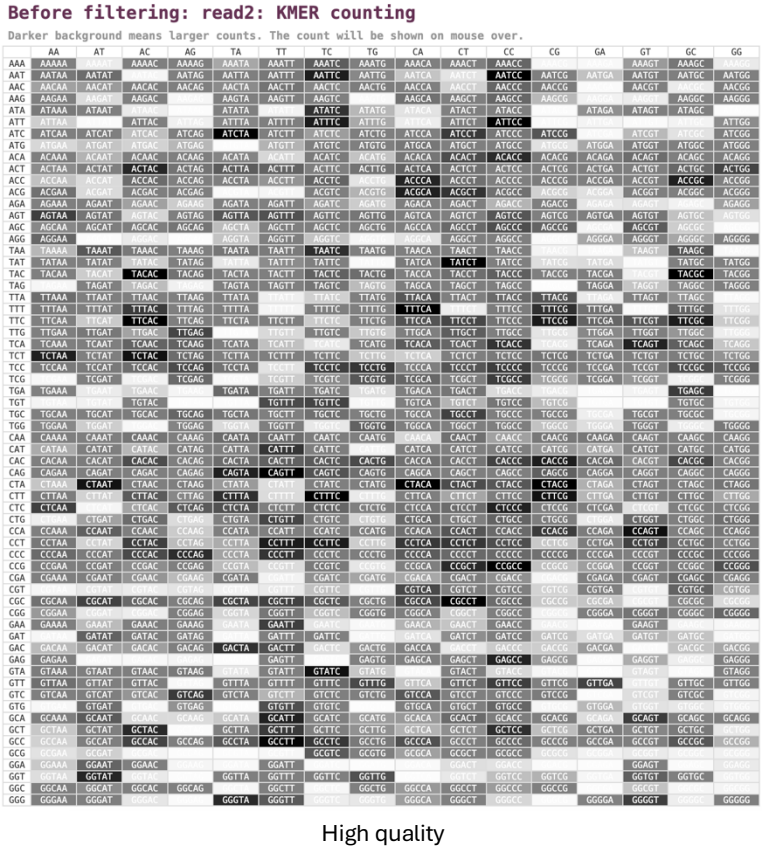

Low quality

Here, we observe the same pattern as in read 1: strong variation at the start of the sequences, especially in the first 100 positions.

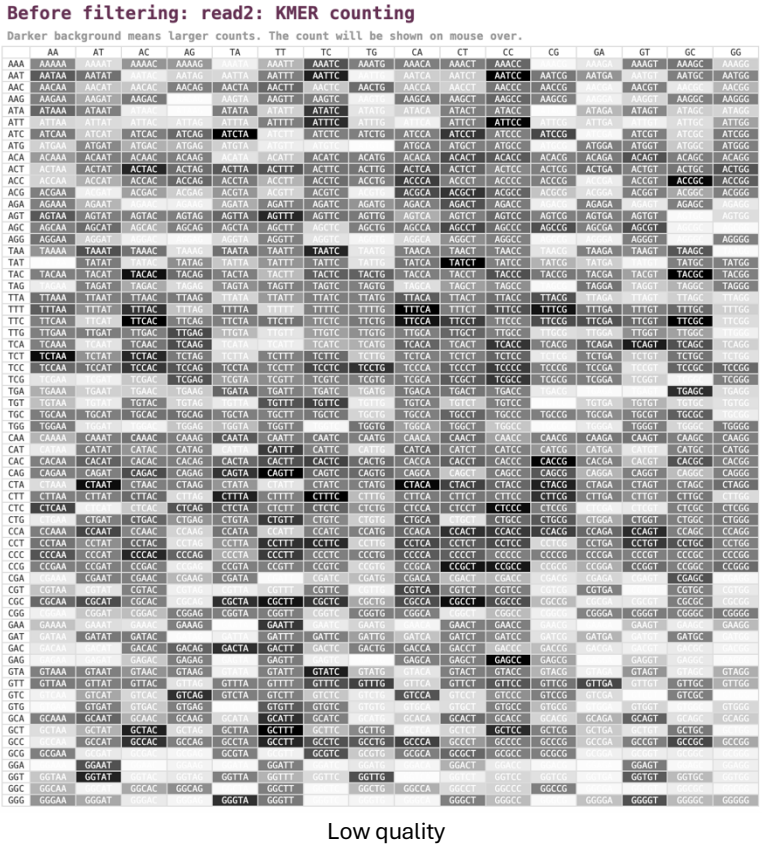
The GC content is again similar in both samples (50 and 51%). This means the overall base composition is normal.

Even though one sample is labelled “low quality”, its base content looks very similar to the high quality one. This again confirms that the issue is not with the base composition but with the number of reads.

### c) KMER counting




High quality



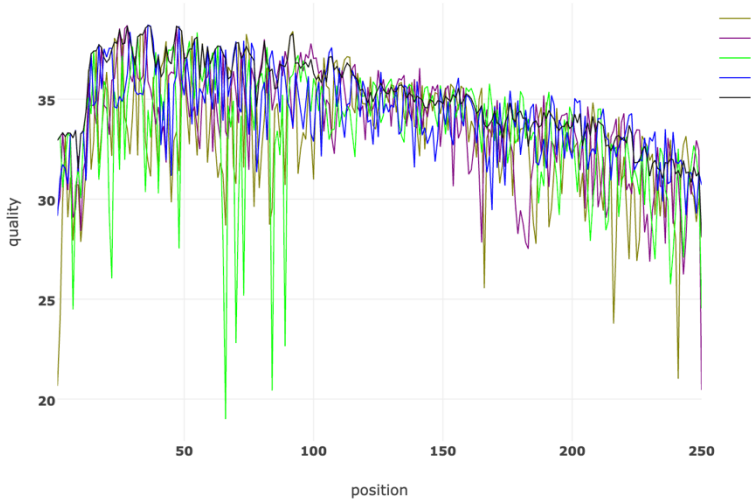

Low quality

When comparing the low-quality sample with the high-quality one for read 2, we see that both plots show a similar pattern: several dark boxes scattered across the table. This means that certain K-mers are repeated quite often in both samples.

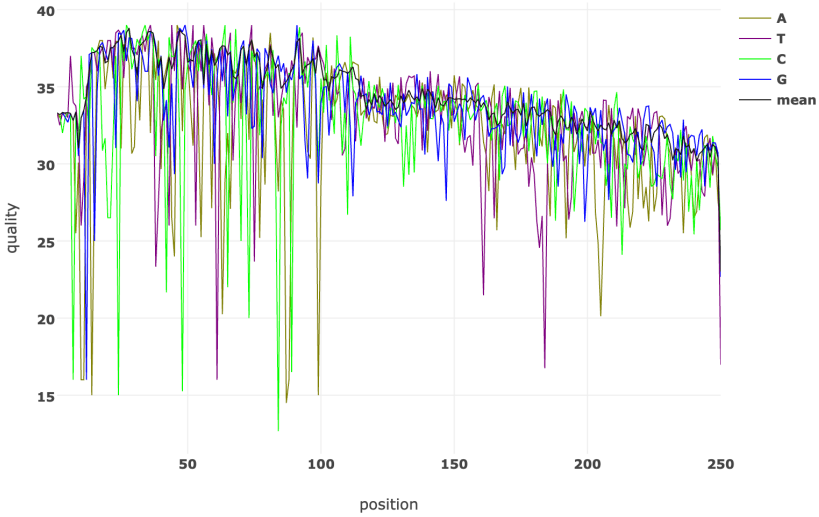
Just like in read 1, there is no obvious difference between the two samples based on this plot alone.

## 5. Read 1 after filtering

### a) Quality




High quality

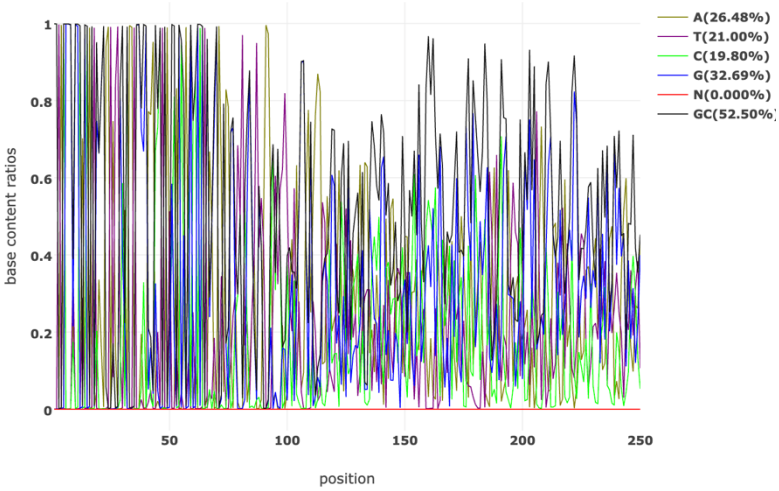



Low quality

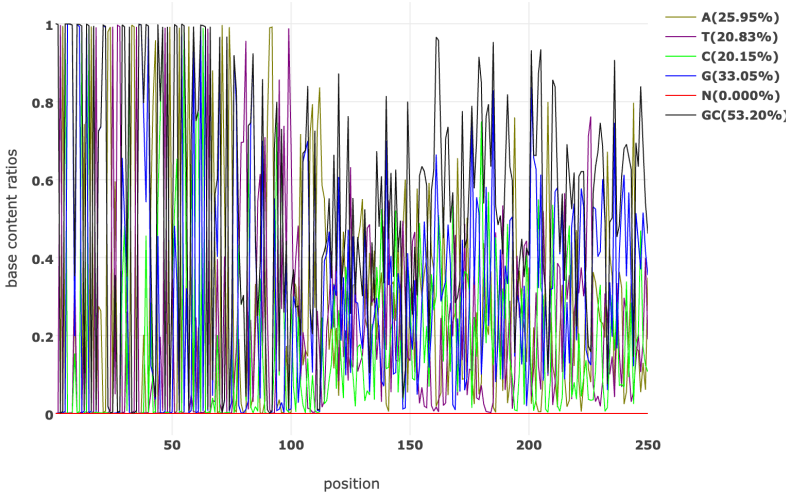
After filtering, the quality curves for read 1 in the high-quality sample remain high and consistent across almost all positions. The average quality is above 30 which is excellent. There are few fluctuations confirming that most of the low-quality bases were successfully removed during filtering.

In contrast, the low-quality sample still shows more irregularities and lower peaks, although some of the worst regions were cleaned. The curve is slightly smoother compared to before filtering, but due to the extremely low number of reads, the quality data is less reliable and more prone to noise. This supports our decision to exclude such samples from downstream analysis.

### b) Base contents




High quality

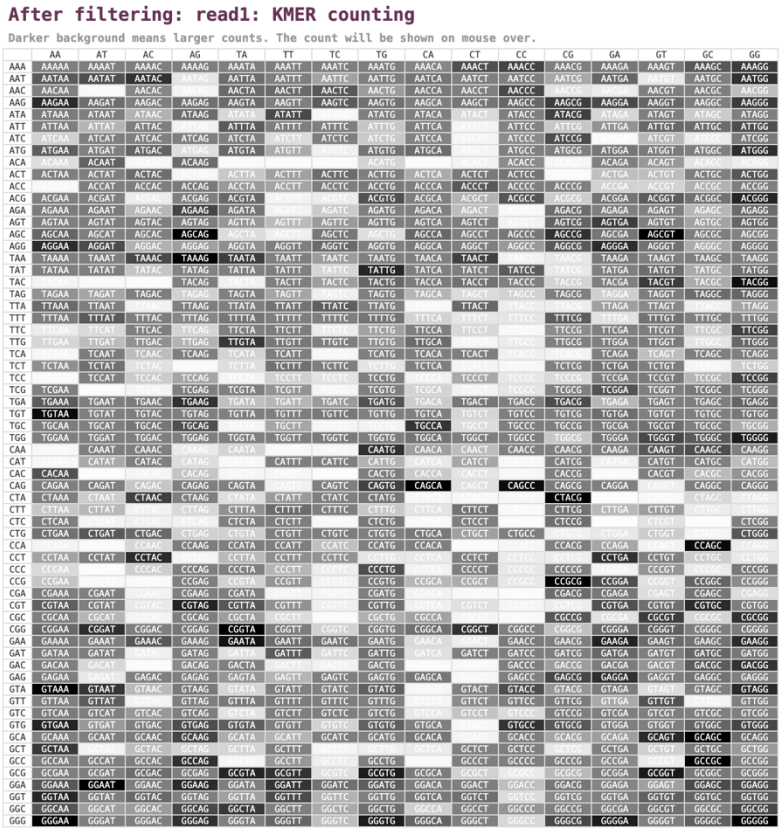



Low quality

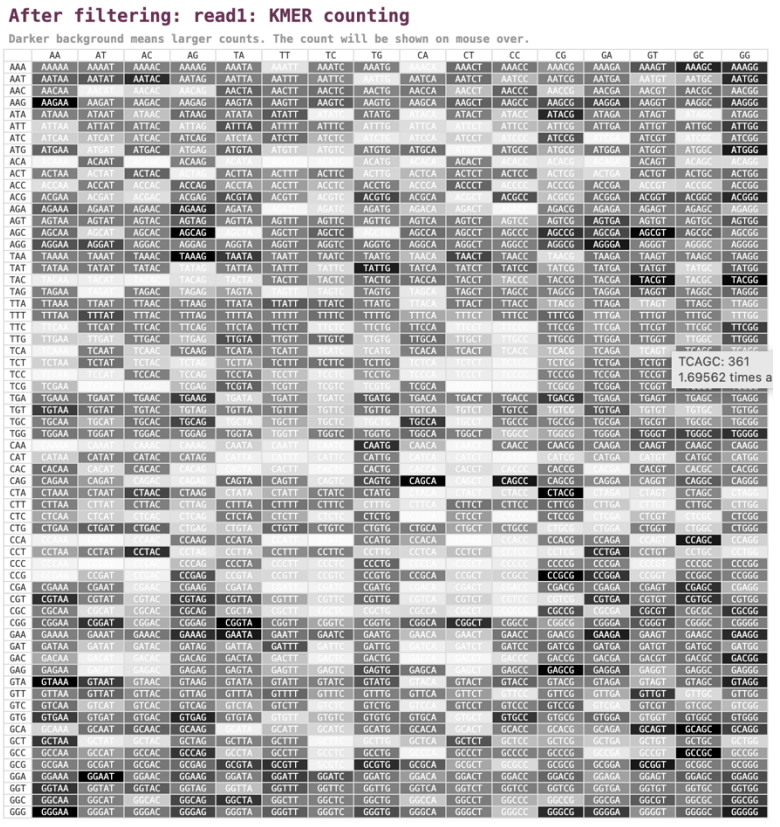
The high-quality sample shows a more balanced and stabilized base composition across positions. The initial fluctuations seen before filtering are still visible but slightly reduced, and the GC content remains stable at around 52%.

In the low-quality sample, the base content pattern is noisier and less consistent, likely due to the very small read count. However, the GC content remains within the expected range and similar to the high-quality sample. This confirms again that the issue is in the quantity rather than the composition of the reads.

### c) KMER counting




High quality




Low quality

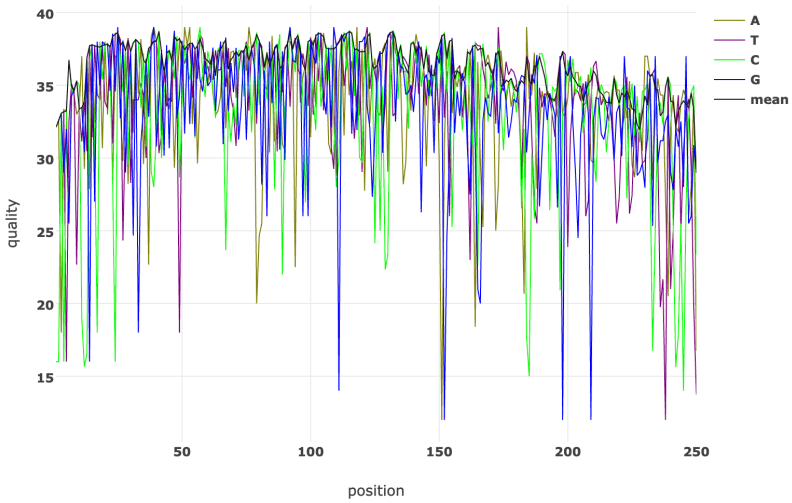
In both heatmaps we still observe several dark boxes, which correspond to repeated 5-mers. This repetition is expected in 16S amplicon sequencing and usually reflects conserved regions rather than technical issues.

There is no clear difference between the two samples based on this plot. And this confirms one time again that the main issue is the insufficient number of reads rather than sequence composition or technical artifacts.

## 6. Read 2 after filtering

### a) Quality

High quality

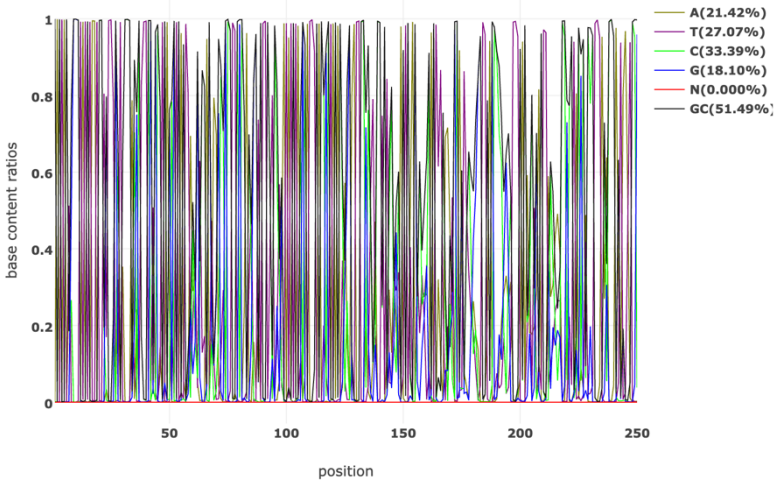


Low quality

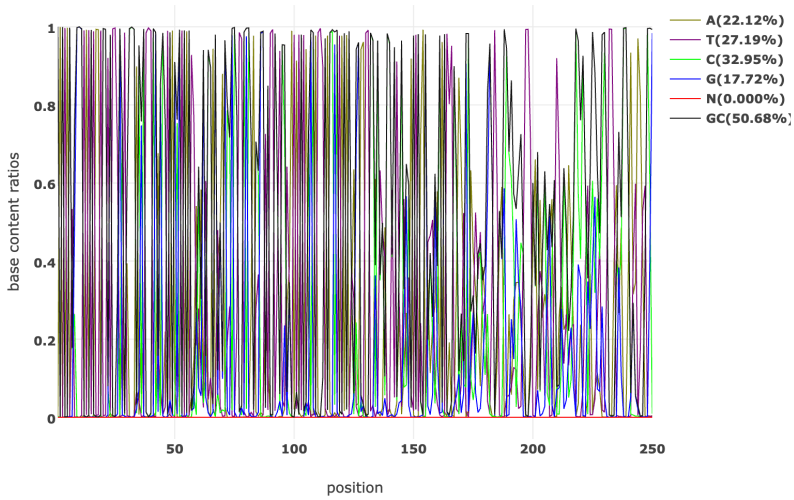
The quality profile for read 2 after filtering in the high-quality sample remains strong and regular. Quality values stay above 30 for most of the sequence.

For the low-quality sample, the curve is less noisy than before but still displays more variation than the high quality one. The filtered reads still contain several low-quality regions, which may impact downstream analysis.

## b) Base contents




High quality

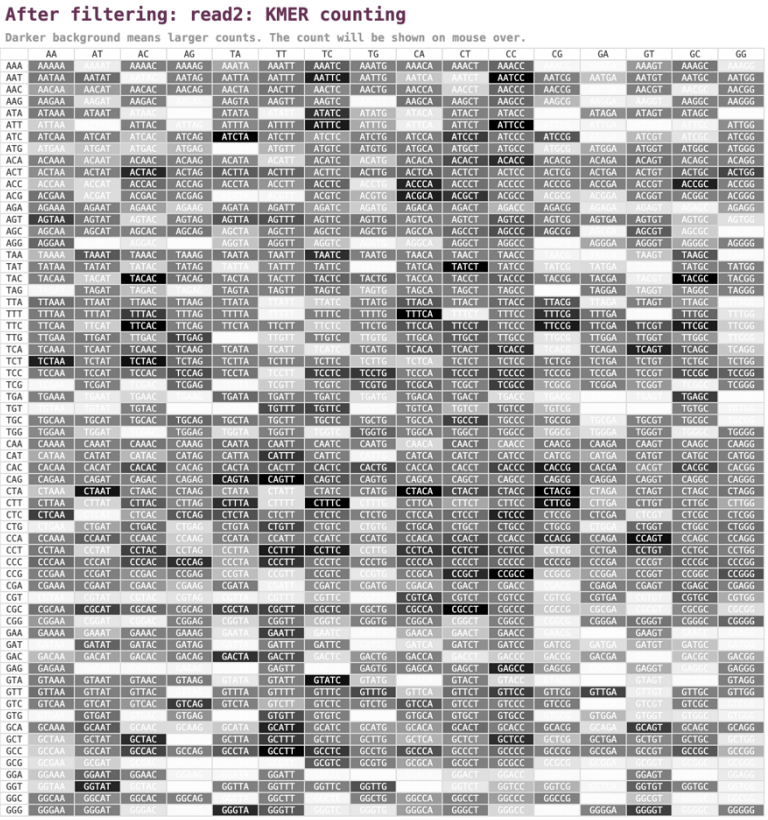



Low quality

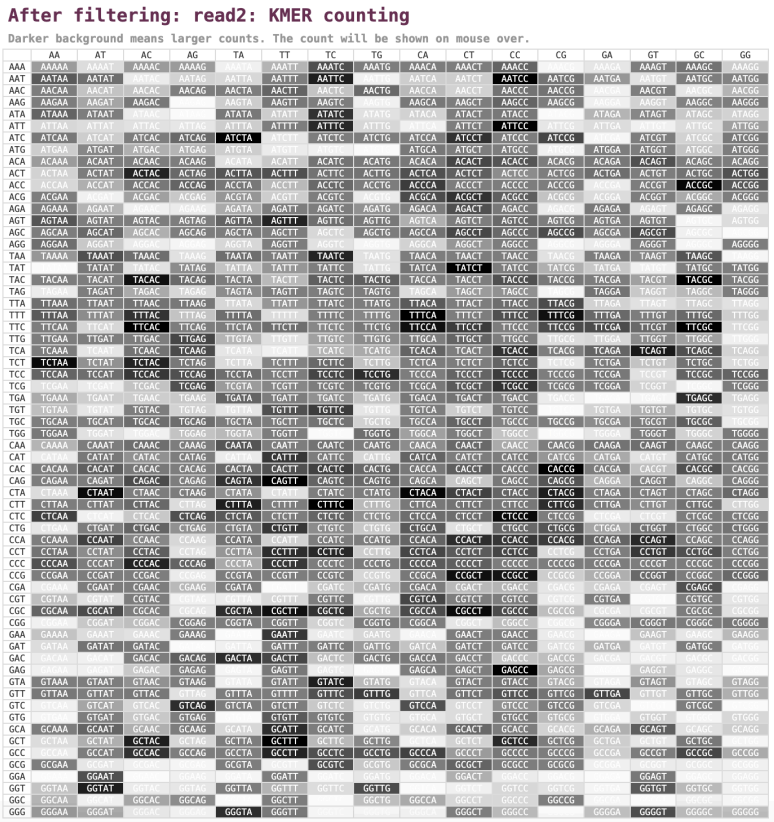
After filtering, both samples show very similar base content profiles for read 2. As expected in amplicon sequencing, there is some strong fluctuations across positions.

GC content is around 51% in both cases, which is normal. There is no sign of unusual base composition, so the quality issue is clearly related to read quantity rather than sequence bias.

## c) KMER counting




High quality




Low quality

After filtering, both the high quality and low-quality samples show similar K-mer patterns as before. We still observe several dark boxes (which correspond to repeated 5-mers and this repetition is expected).

There is no clear difference between the two samples based on this plot.

## 7. Conclusion

To conclude, this comparison between a high-quality and a low-quality sample using fastp reports highlights the importance of read count in microbiome analysis workflows. Although both samples showed acceptable quality metrics, the low-quality sample had too few reads to be considered reliable for downstream analysis with tools like DADA2.

Most visual indicators did not reveal major technical issues in the low-quality sample. However, their interpretation is limited by the very small number of reads which leads to noisy patterns and low statistical confidence.

Therefore, the decision to remove samples with fewer than 30,000 reads is justified and necessary to ensure accurate bioinformatics results. This filtering step is essential to avoid biased microbial diversity estimation and false conclusions in downstream analysis.