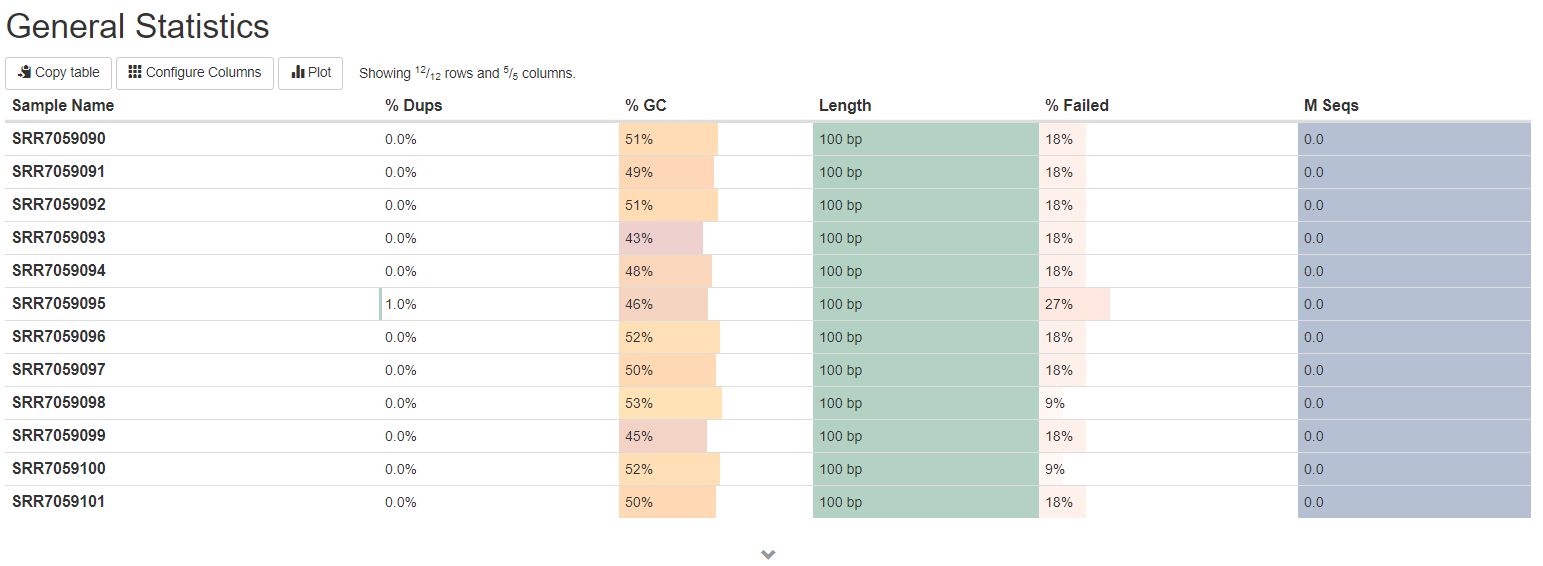
P. falciparum and P. yoelii Sporozoites

NCBI Project: PRJNA453153

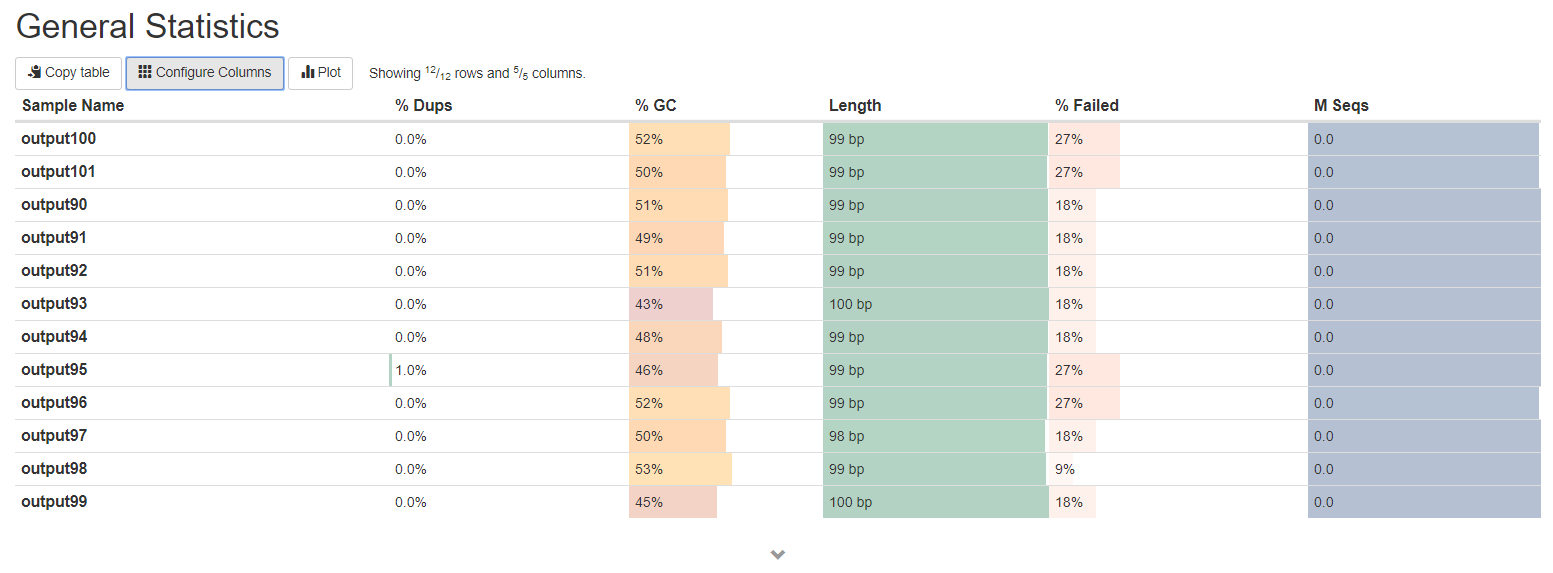
NCBI Biosample: SAMN08975569

Search results: 12 runs (all single ends) - 12 SRRs

MultiQC results - (before trimming adapters)



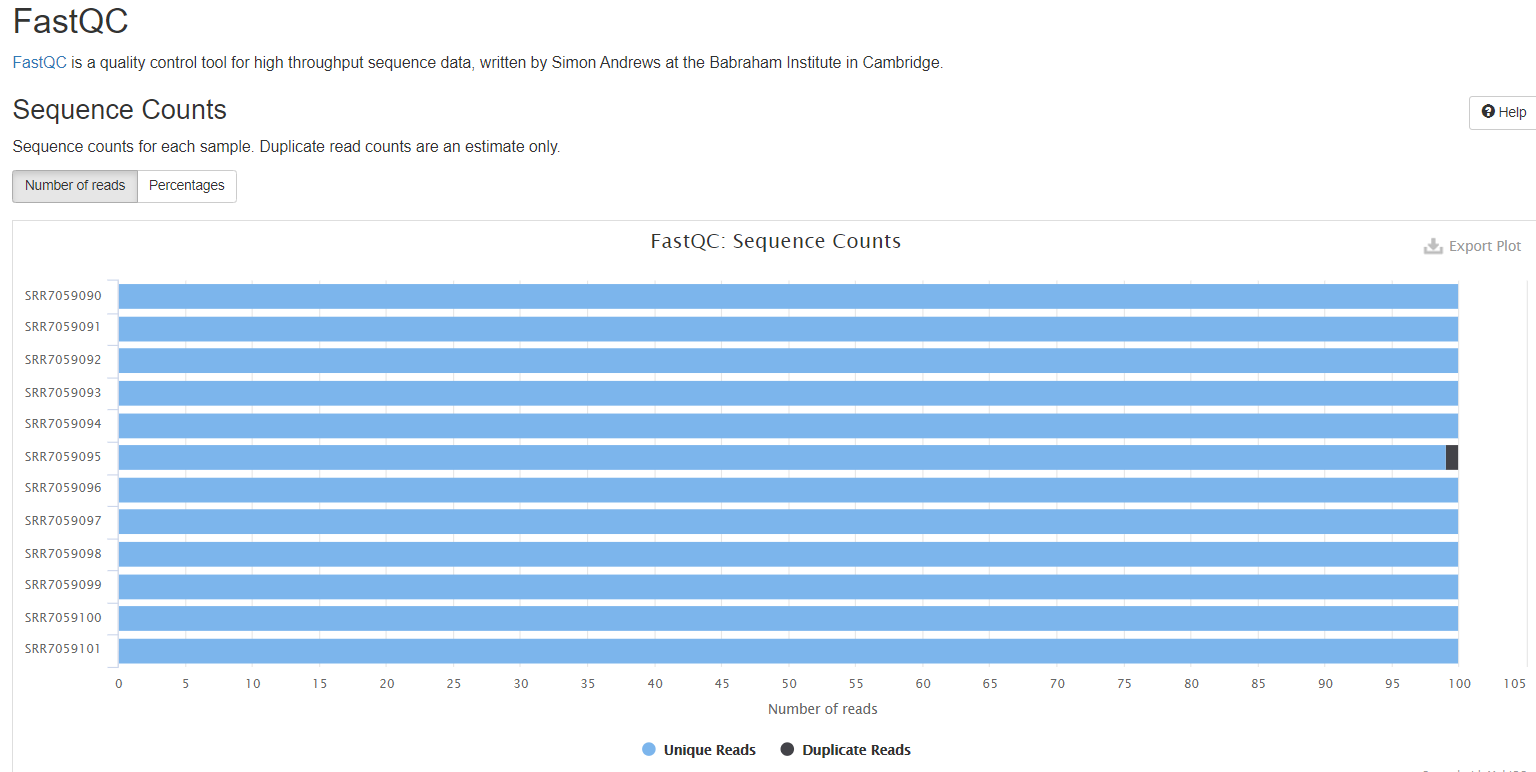
MultiQC results – (after trimming adapters)



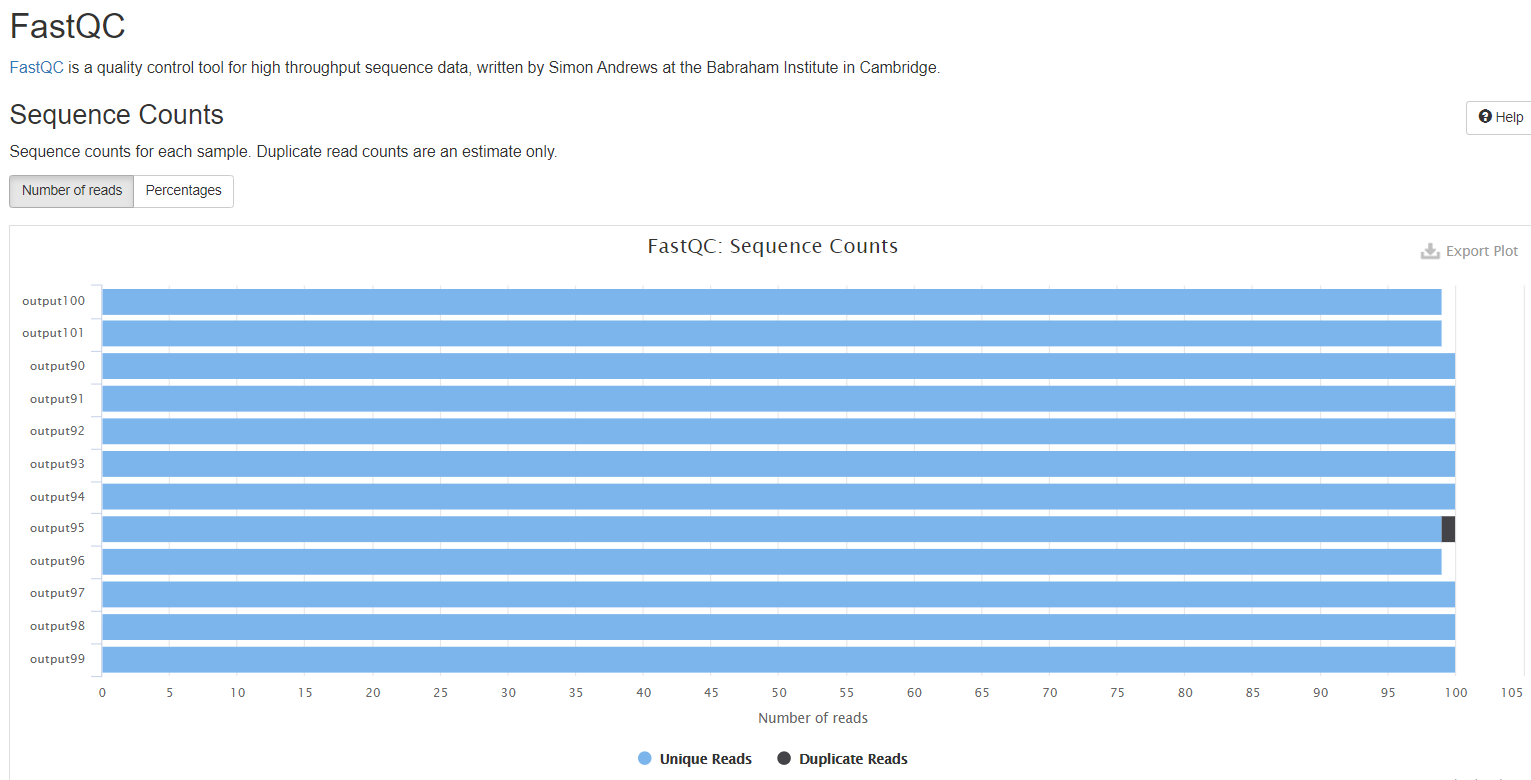
Conclusion:

After trimming adapters, the numbers of higher failure rate increased.

MultiQC results - (before trimming adapters)

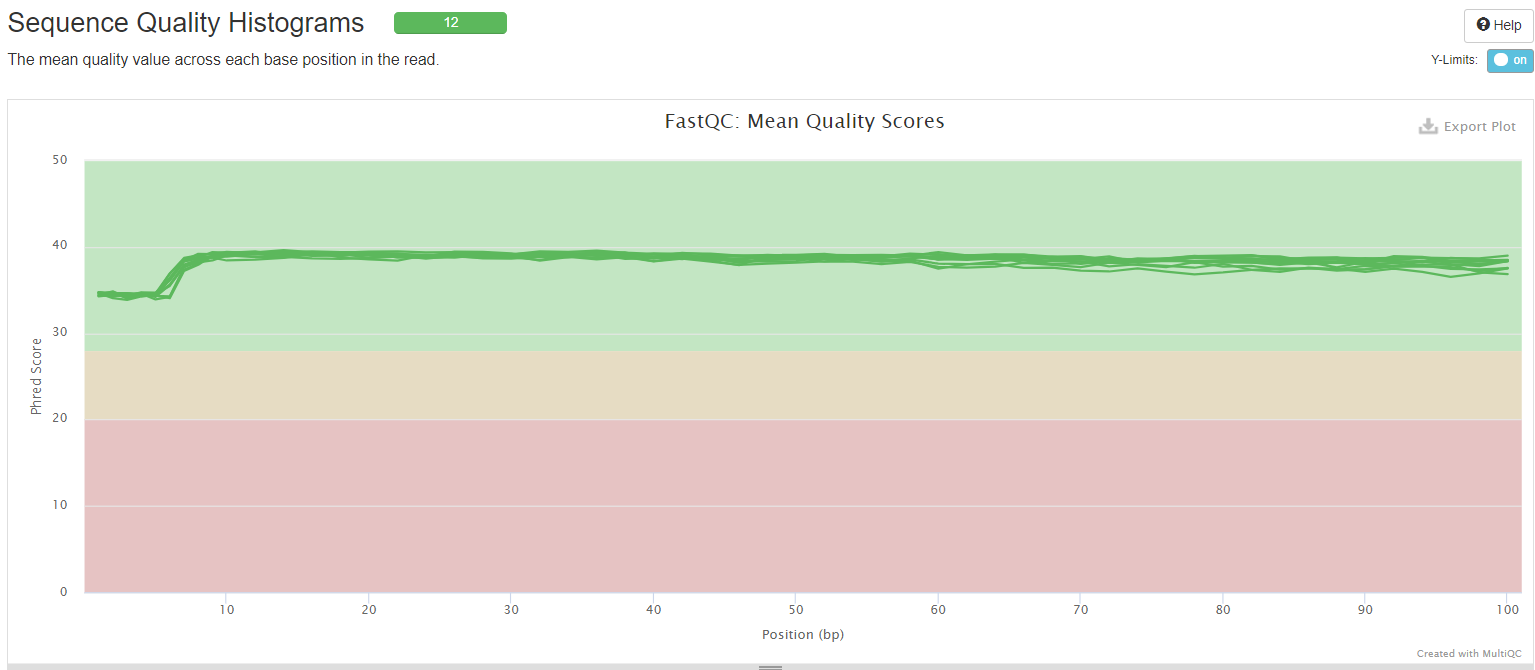


MultiQC results – (after trimming adapters)

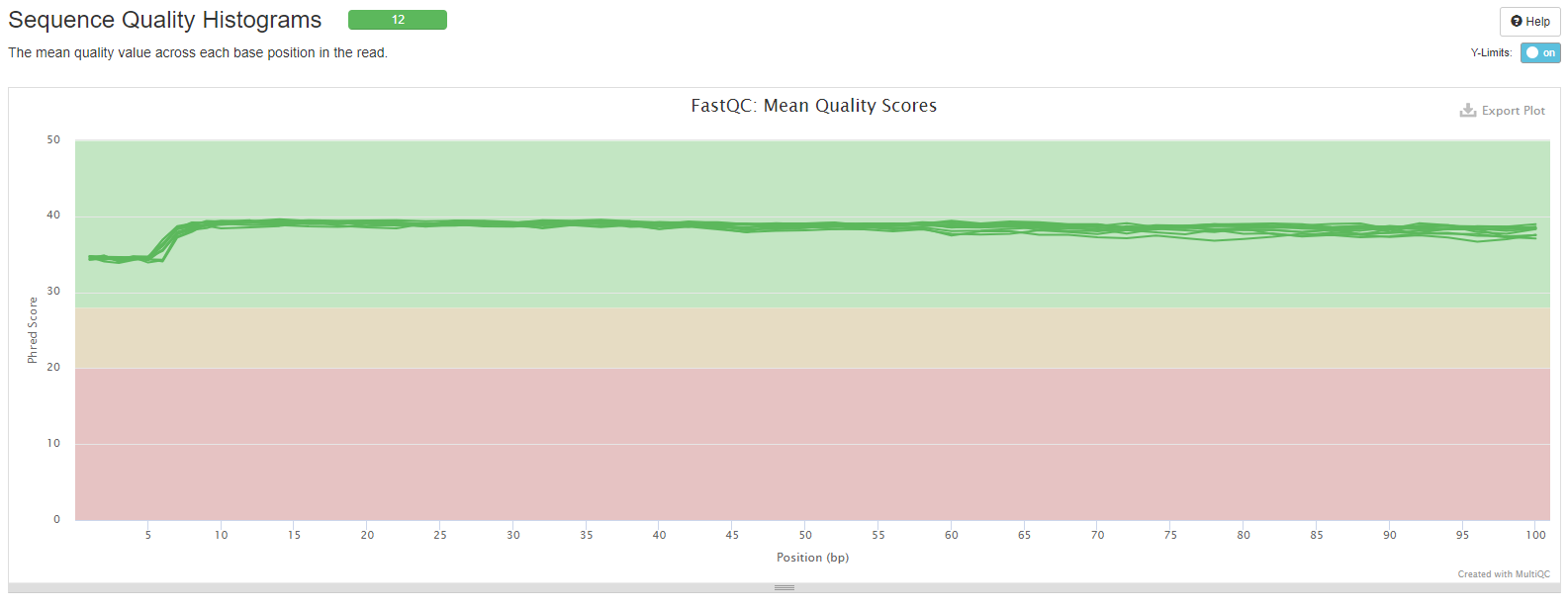


Conclusion: about the same

MultiQC results - (before trimming adapters)

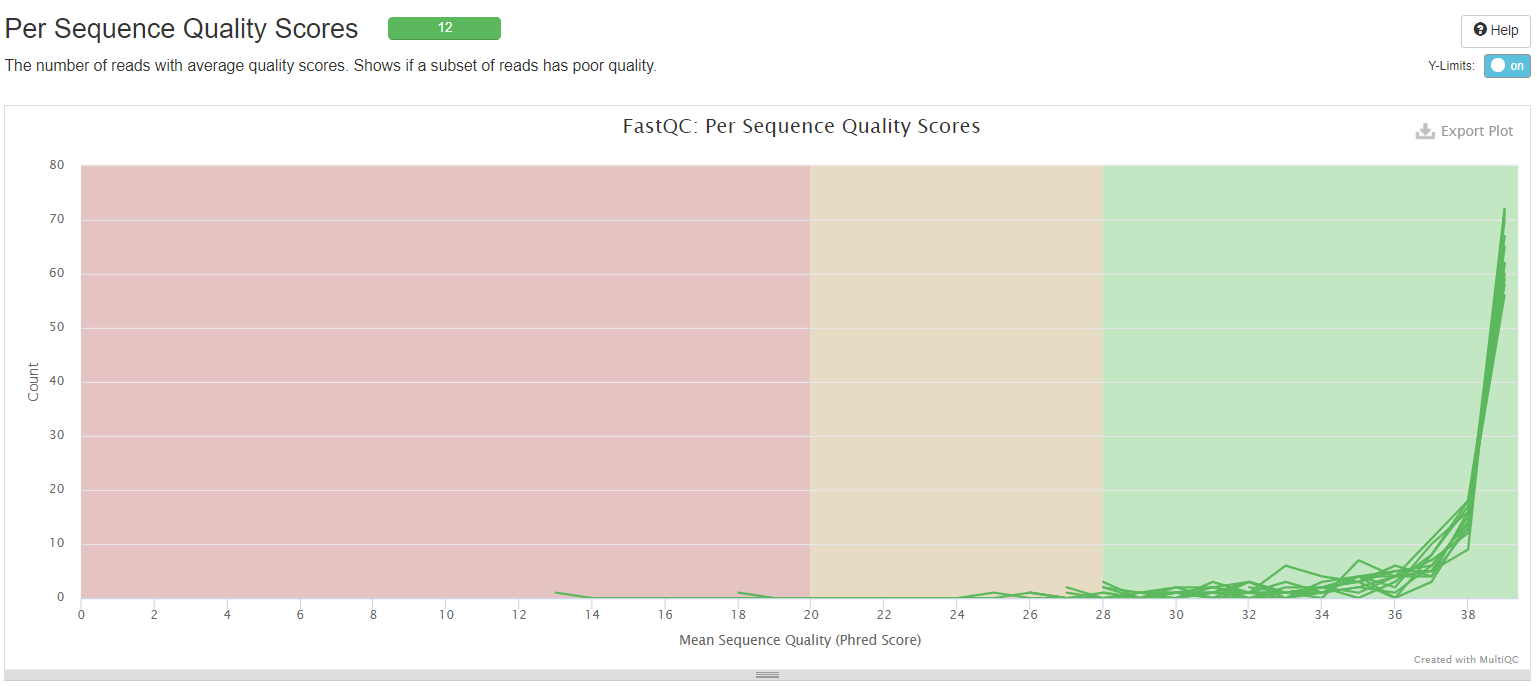


MultiQC results – (after trimming adapters)

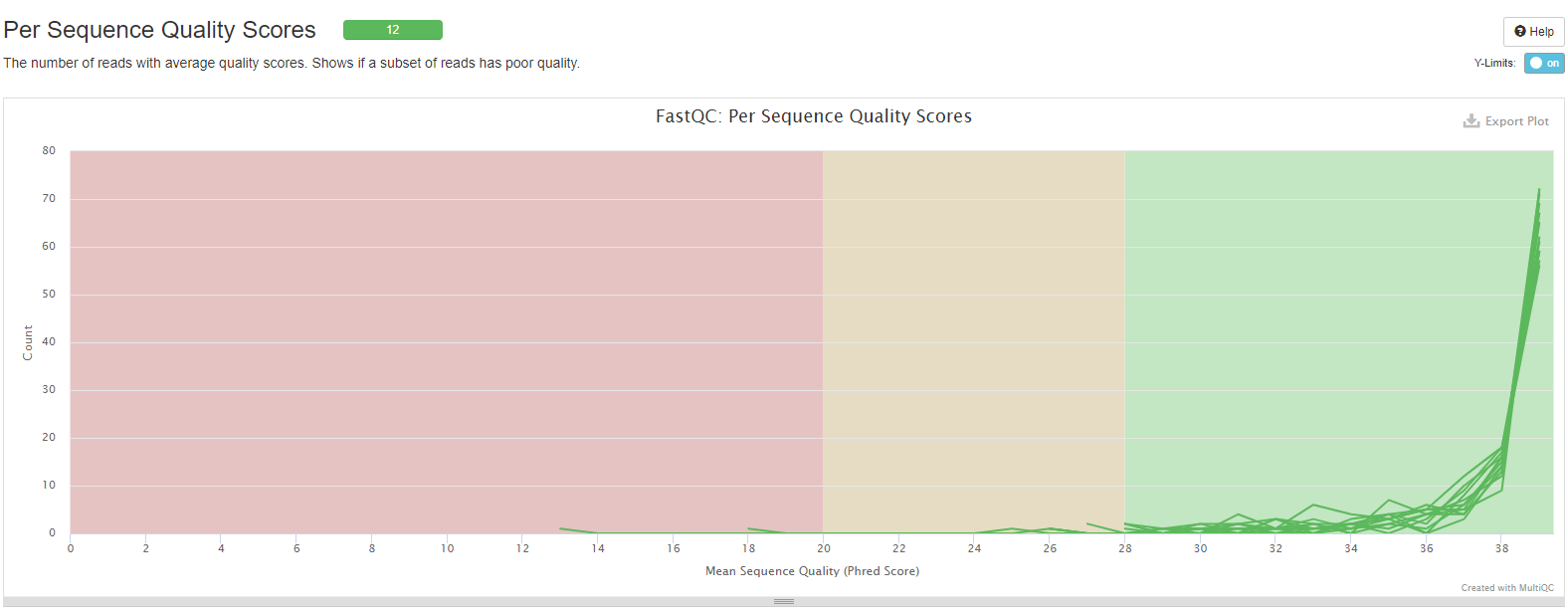


Conclusion: about the same, good quality

MultiQC results - (before trimming adapters)

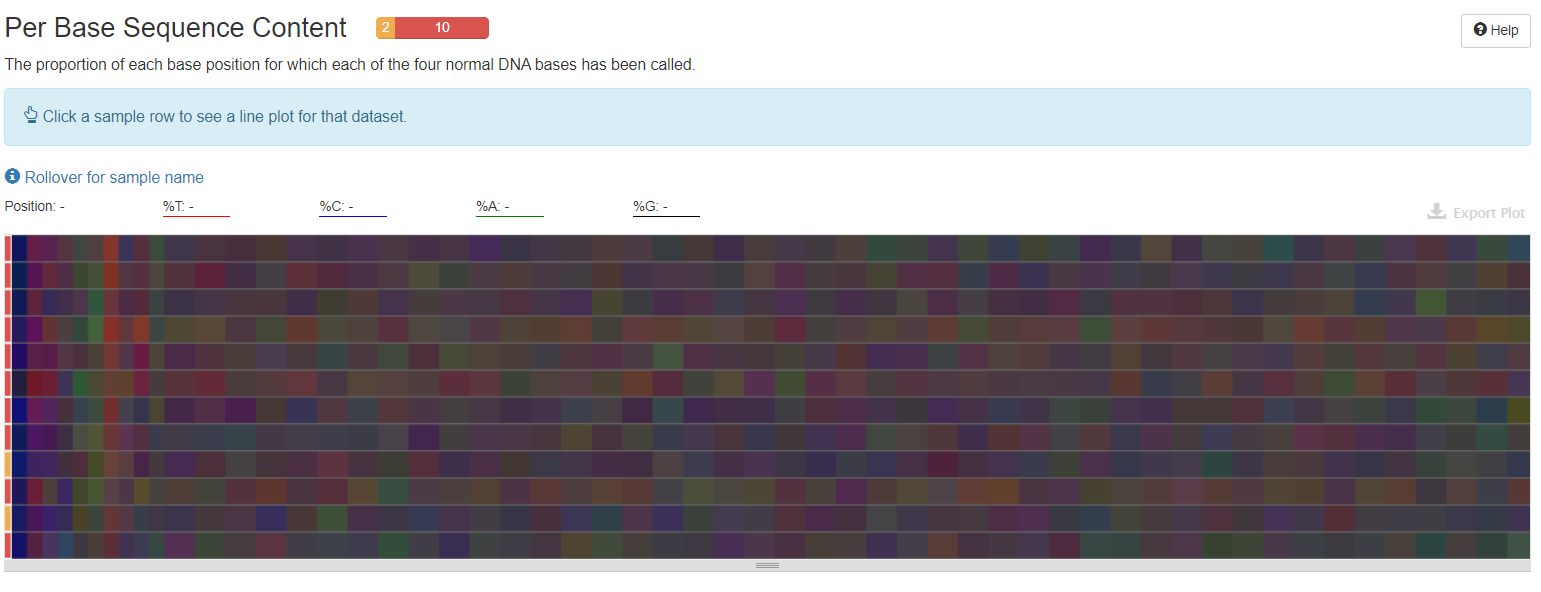


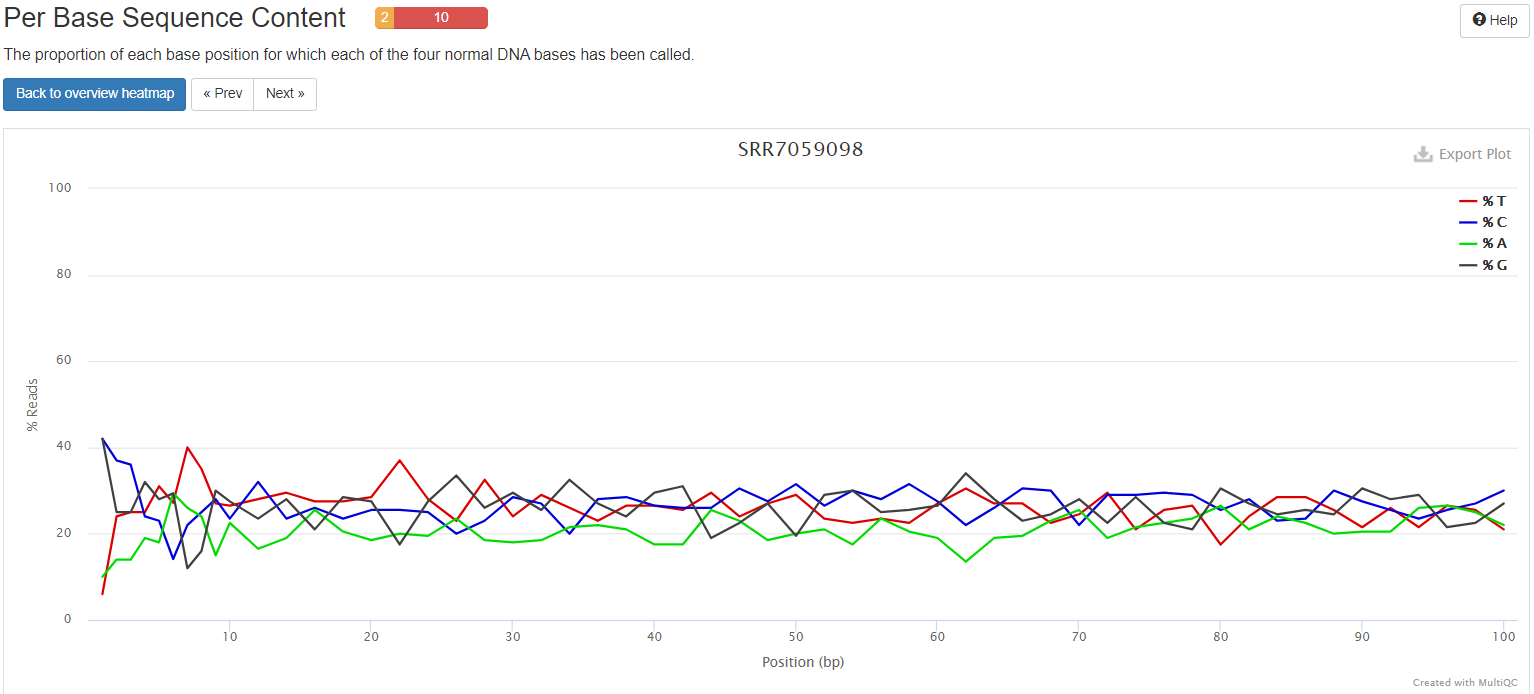
MultiQC results – (after trimming adapters)

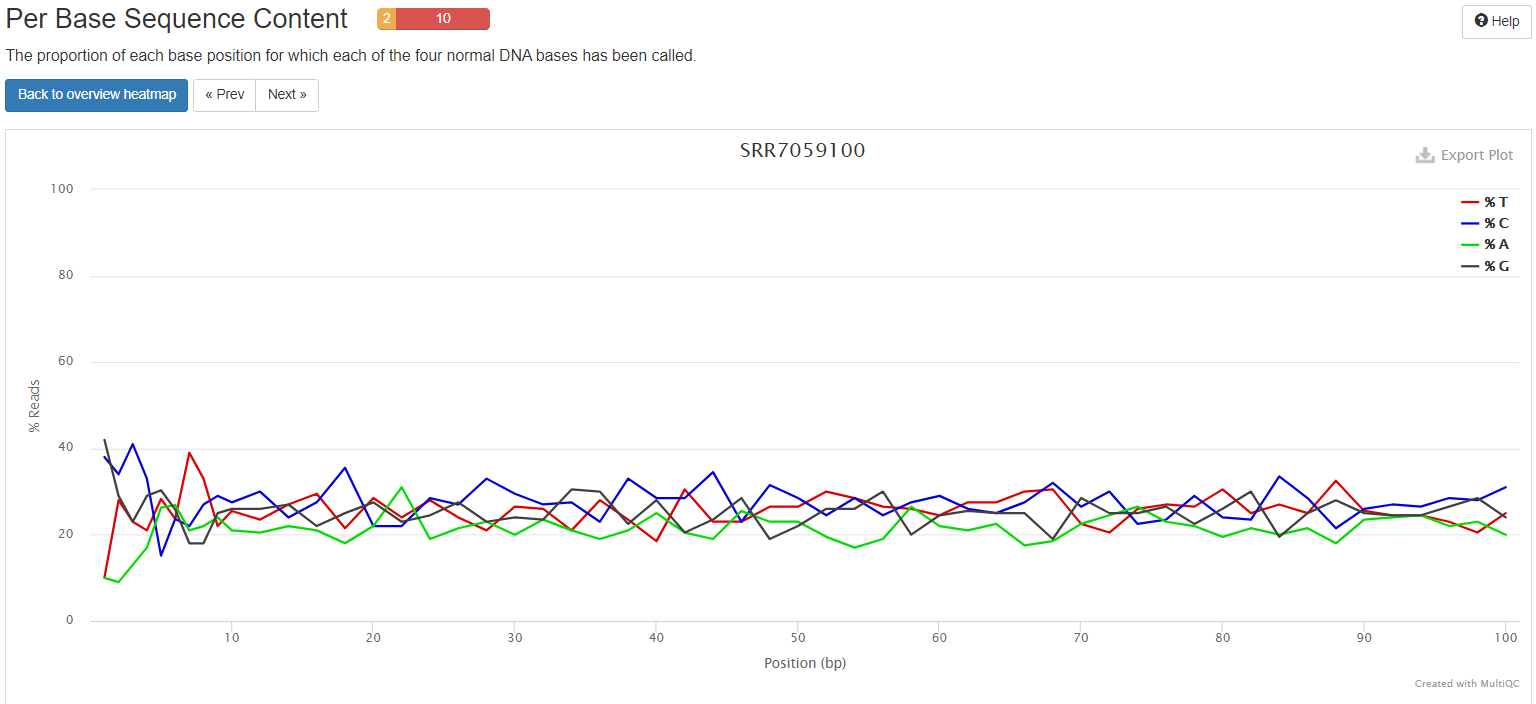


Conclusion: about the same, good quality

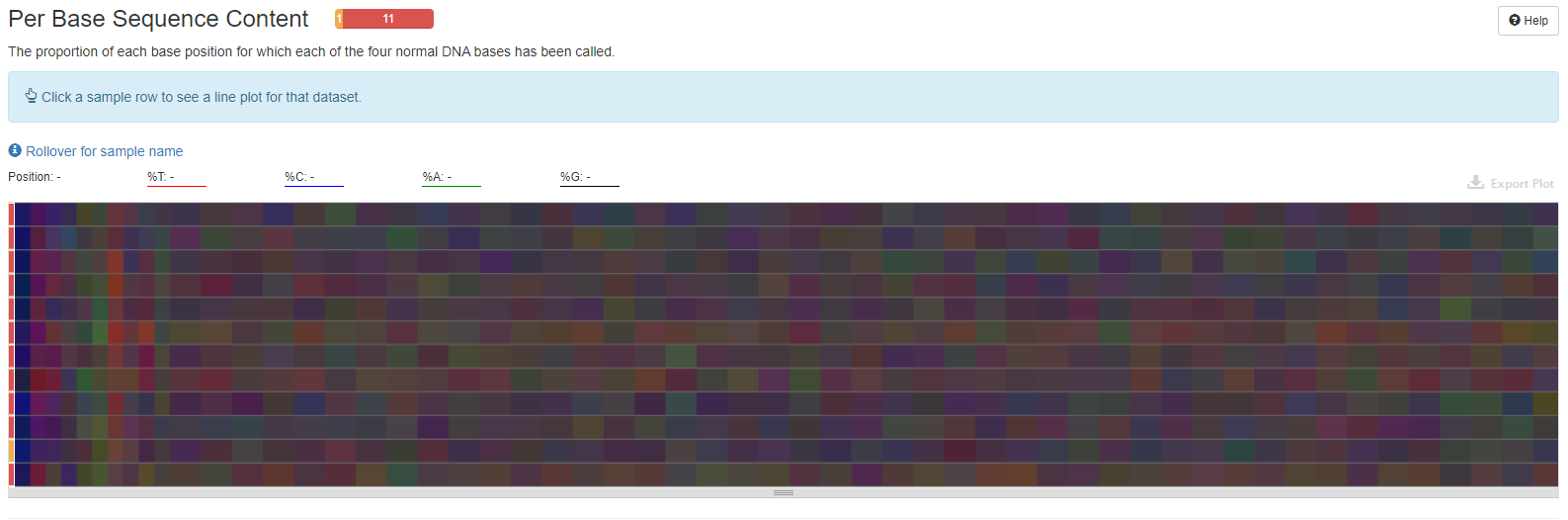
MultiQC results - (before trimming adapters)

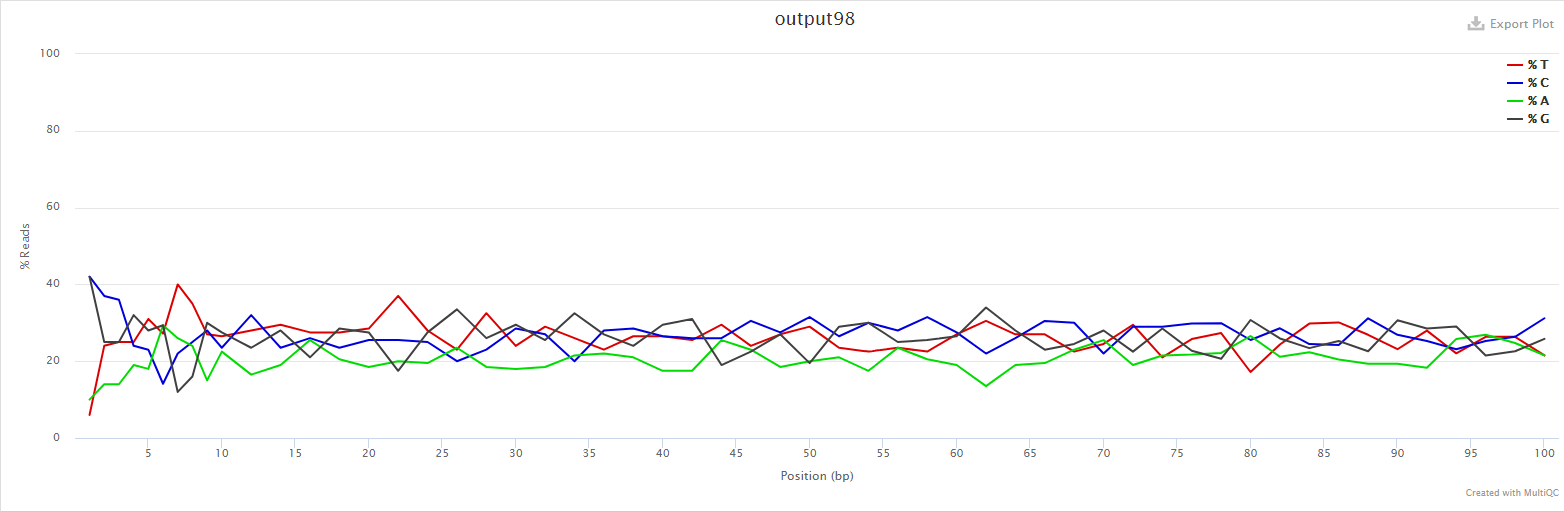






MultiQC results – (after trimming adapters)

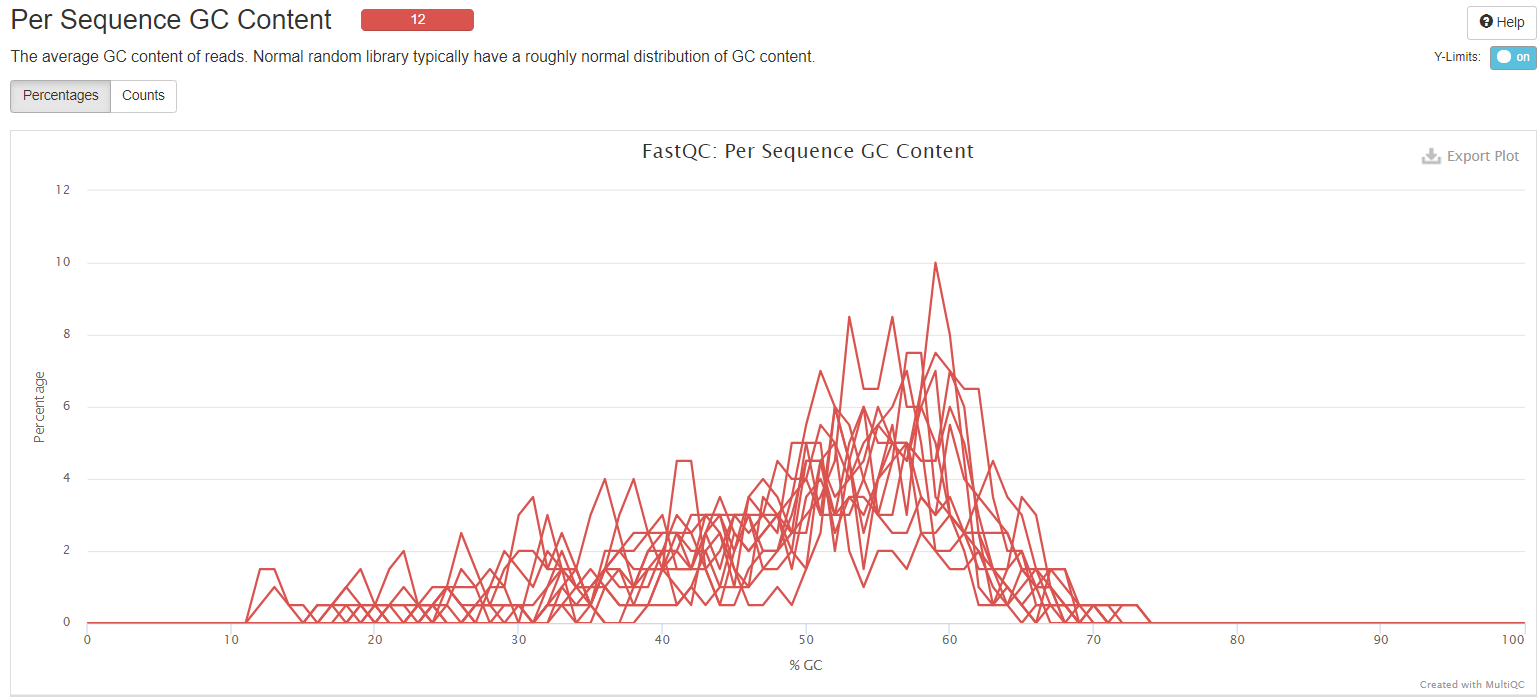




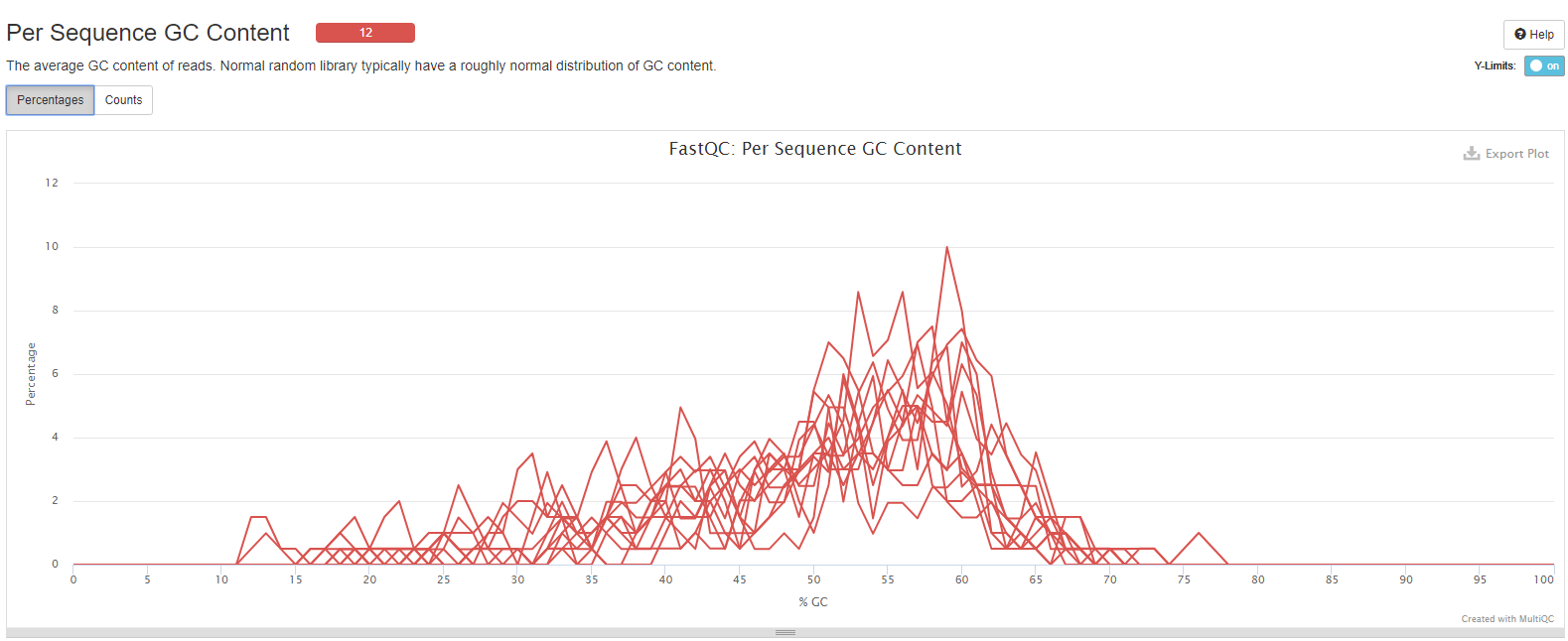
Conclusion –

After trimming, this part of the report did seem to be a little “worse” – before trimming, 10 failed, 2 got warning; but after trimming, 11 failed and 1 got warning. Failure means the difference between A and T, or G and C is greater than 20% in any position. And warning means the difference between A and T, or G and C is greater than 10% in any position. One of the reasons that might cause this is the overrepresented sequence in the sample. These sequences may bias the overall composition. Later, in the report, I found that before trimming, there are only one overrepresented sequences that failed (SRR7059095), but after trimming, the number of overrepresented sequences failure went up to four (SRR7059095, SRR7059096, SRR7059100, SRR7059101)

MultiQC results - (before trimming adapters)



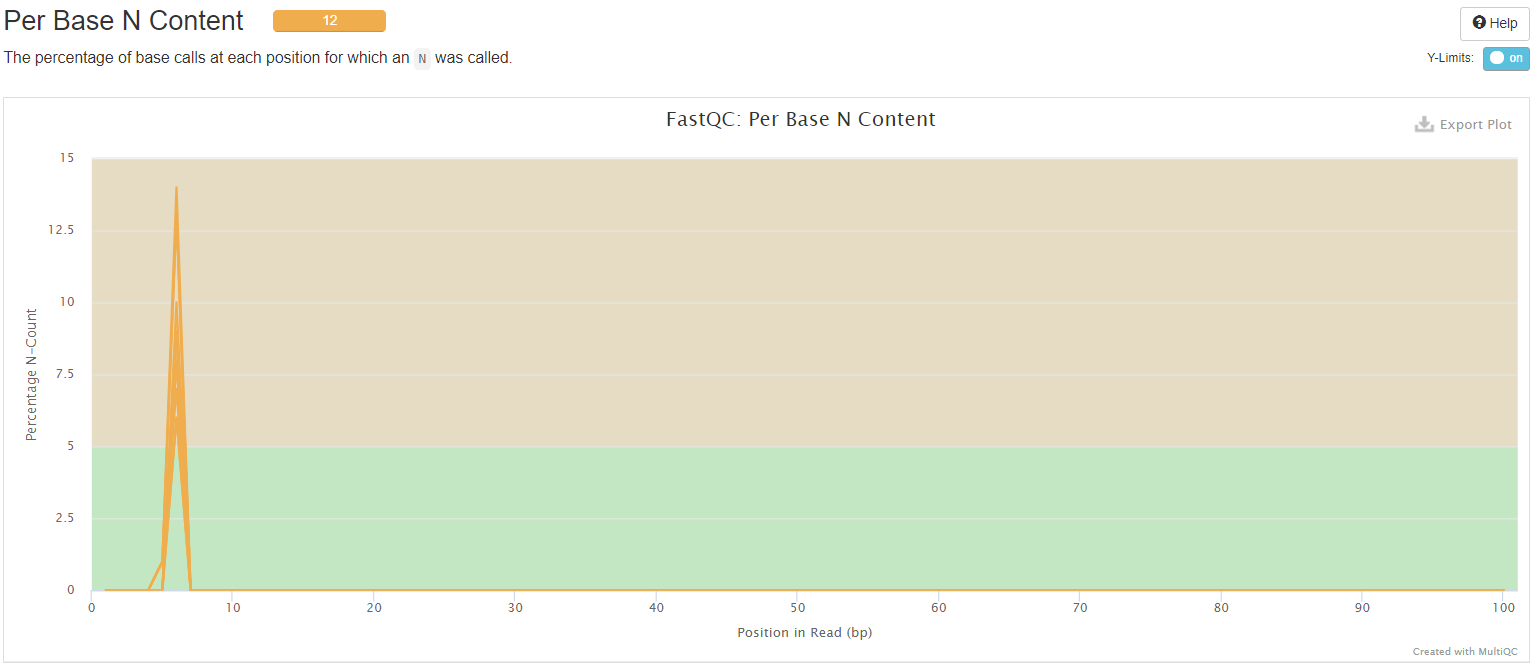
MultiQC results – (after trimming adapters)



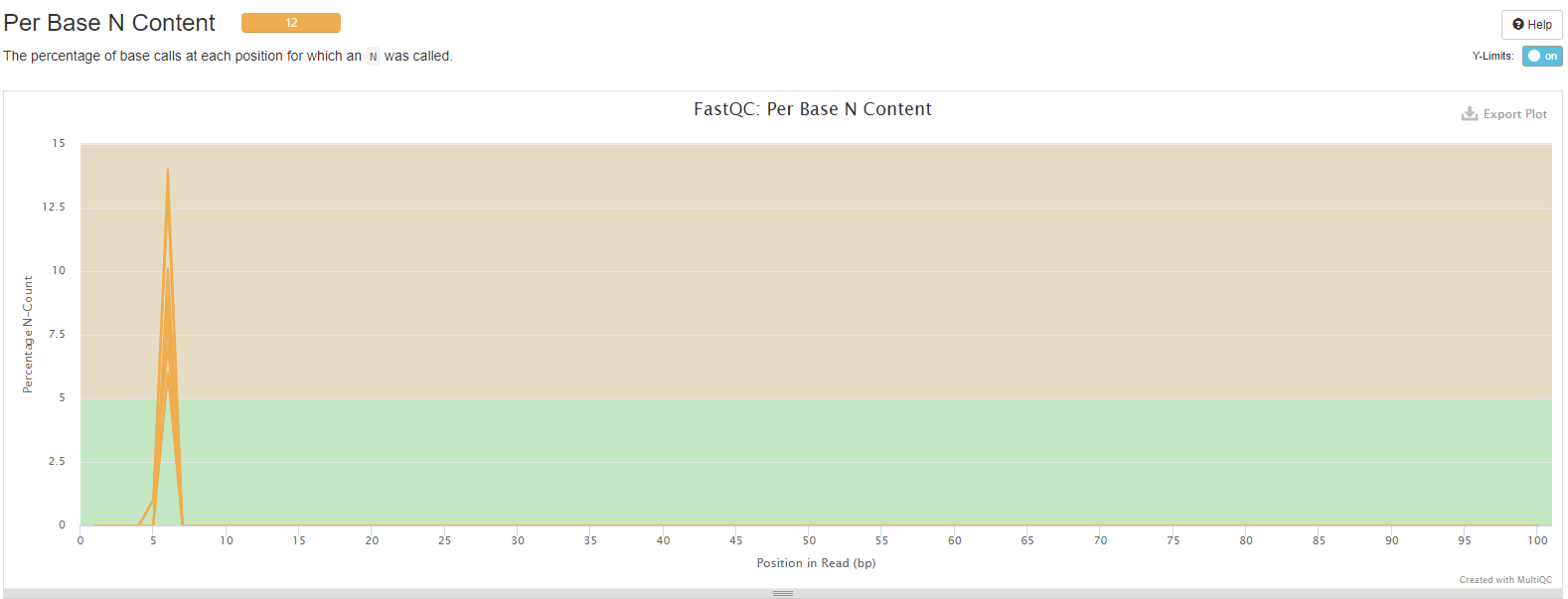
Conclusion –

About the same, all 12 runs fail because the sum of the deviations from the normal distribution represents more than 30% of the reads. I think there might be contamination in the library that cause this. Will further investigate.

MultiQC results - (before trimming adapters)



MultiQC results – (after trimming adapters)

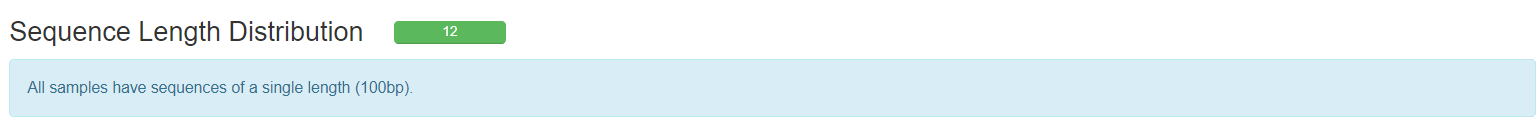


Conclusions-

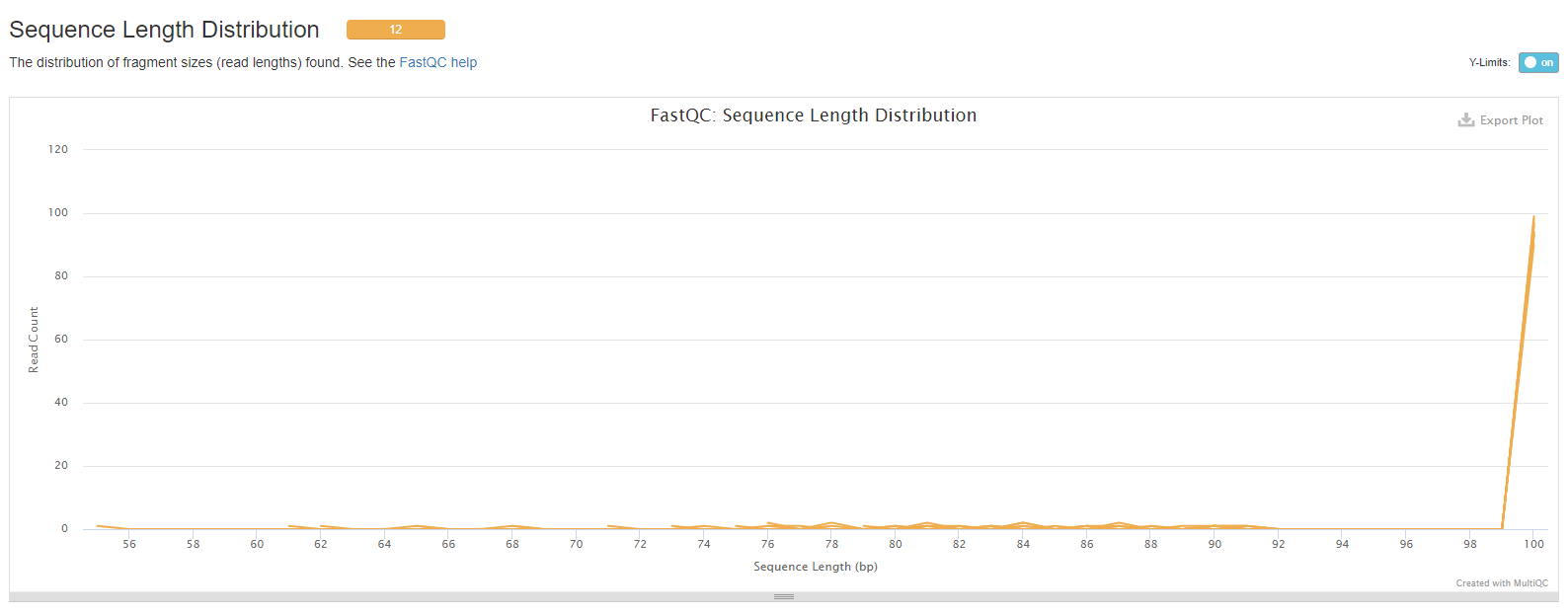
About the same, all 12 are flagged as warning because the position shows N content is above 5%.

I think the reason causes this is still the biased sequence. Will further investigate.

MultiQC results - (before trimming adapters)



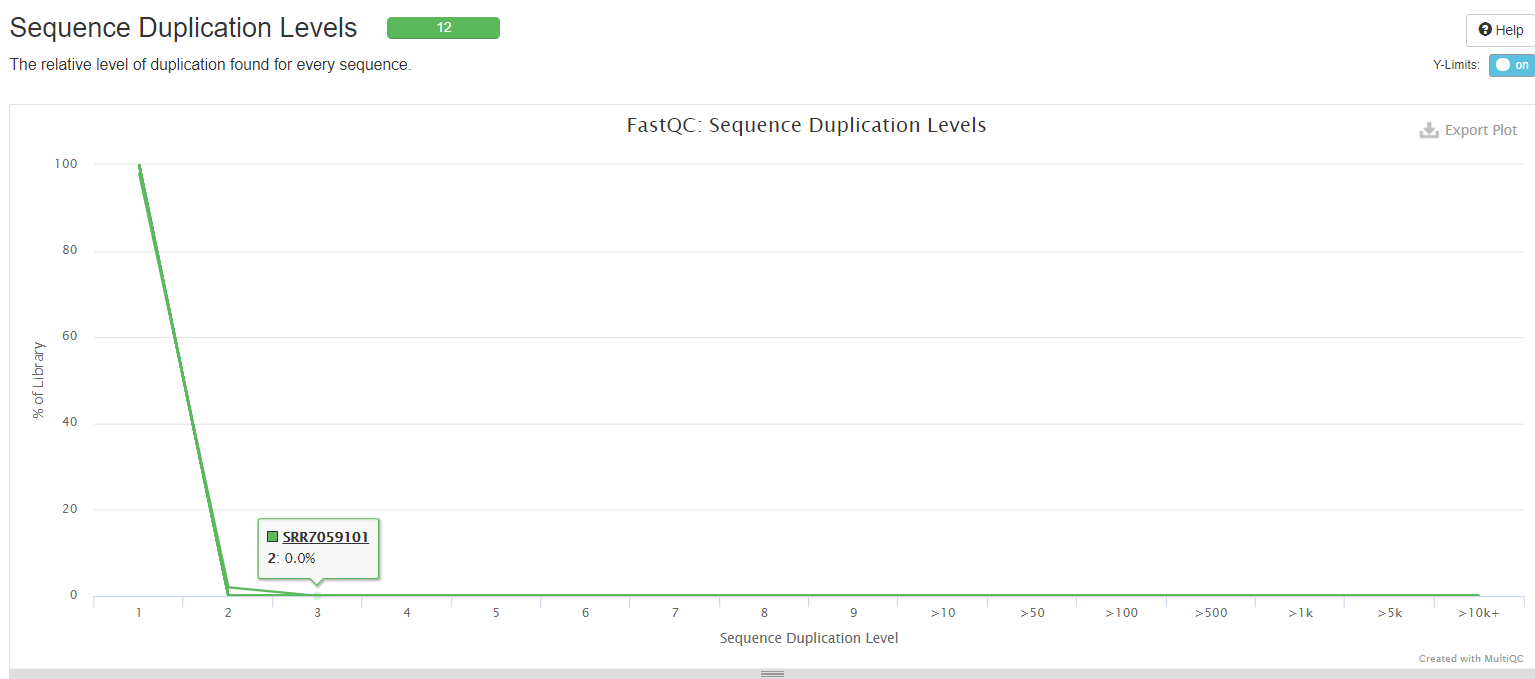
MultiQC results – (after trimming adapters)



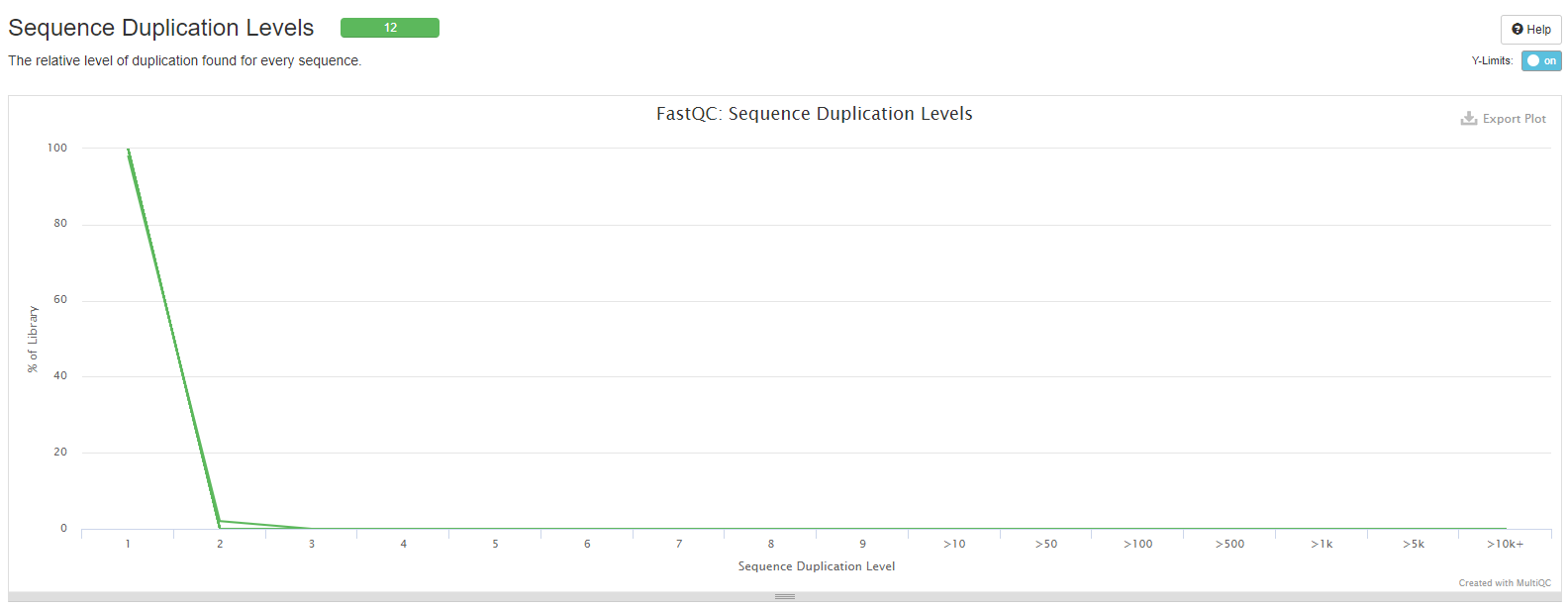
Conclusion –

After trimming adapters, it seems “worse”, but I think since this graph just shows us the sequence length distribution and I trimmed the adapters, then it’s normal to see that now all sequences are not the same length. This can actually be seen in the first graph (all the way on top), the green shaded column where we can see before trimming they all have the same length but after trimming, some got shorter therefore they differ in length.

MultiQC results - (before trimming adapters)



MultiQC results – (after trimming adapters)



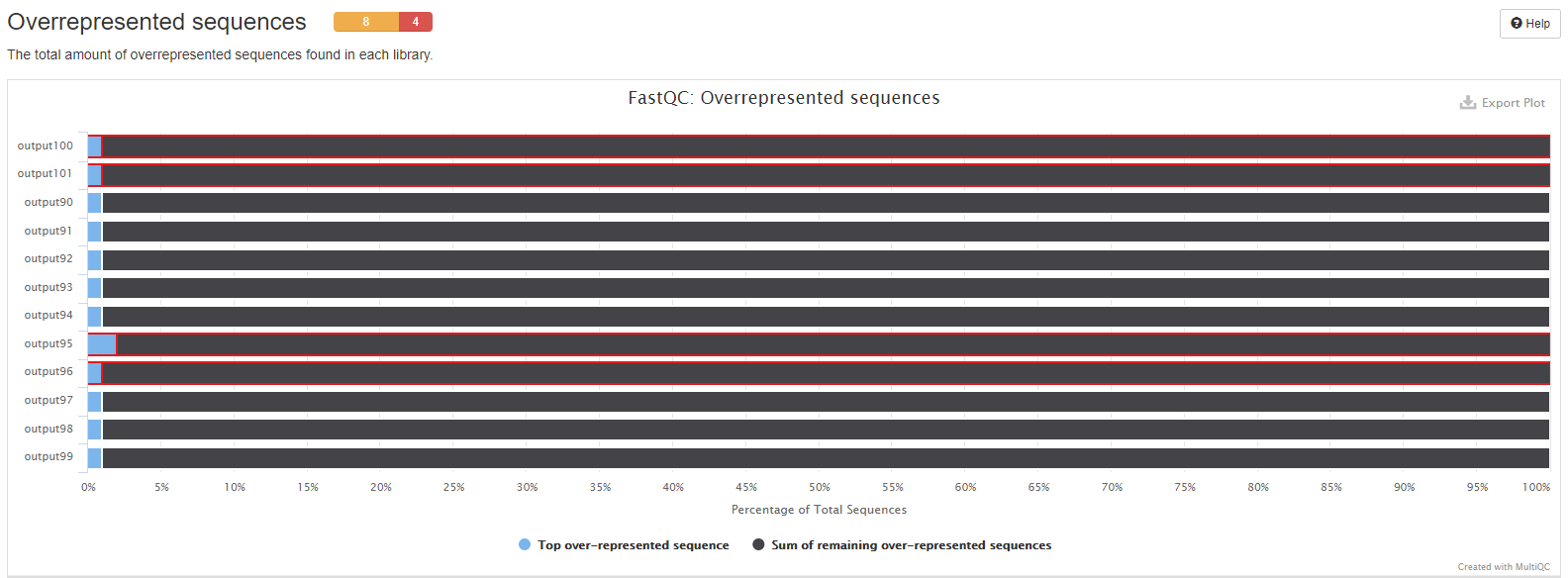
Conclusion –

Interestingly they are about the same, same as good which I though would have some flagged as warning because of biased sequence.

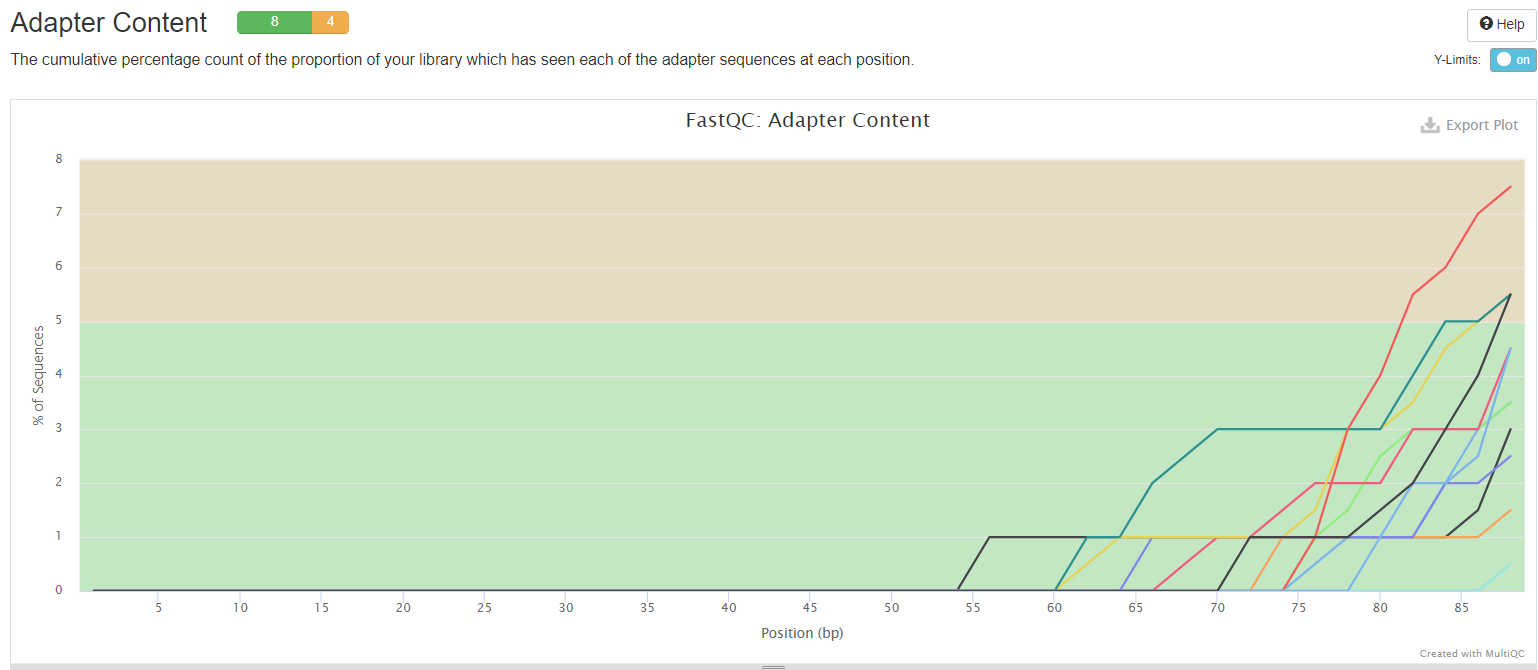
MultiQC results - (before trimming adapters)



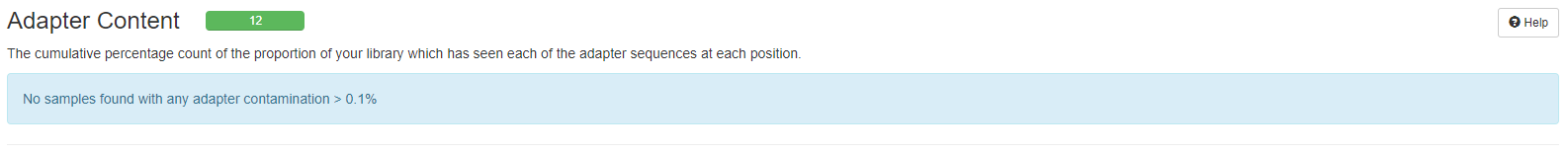
MultiQC results – (after trimming adapters)



MultiQC results - (before trimming adapters)



MultiQC results – (after trimming adapters)



Thoughts –

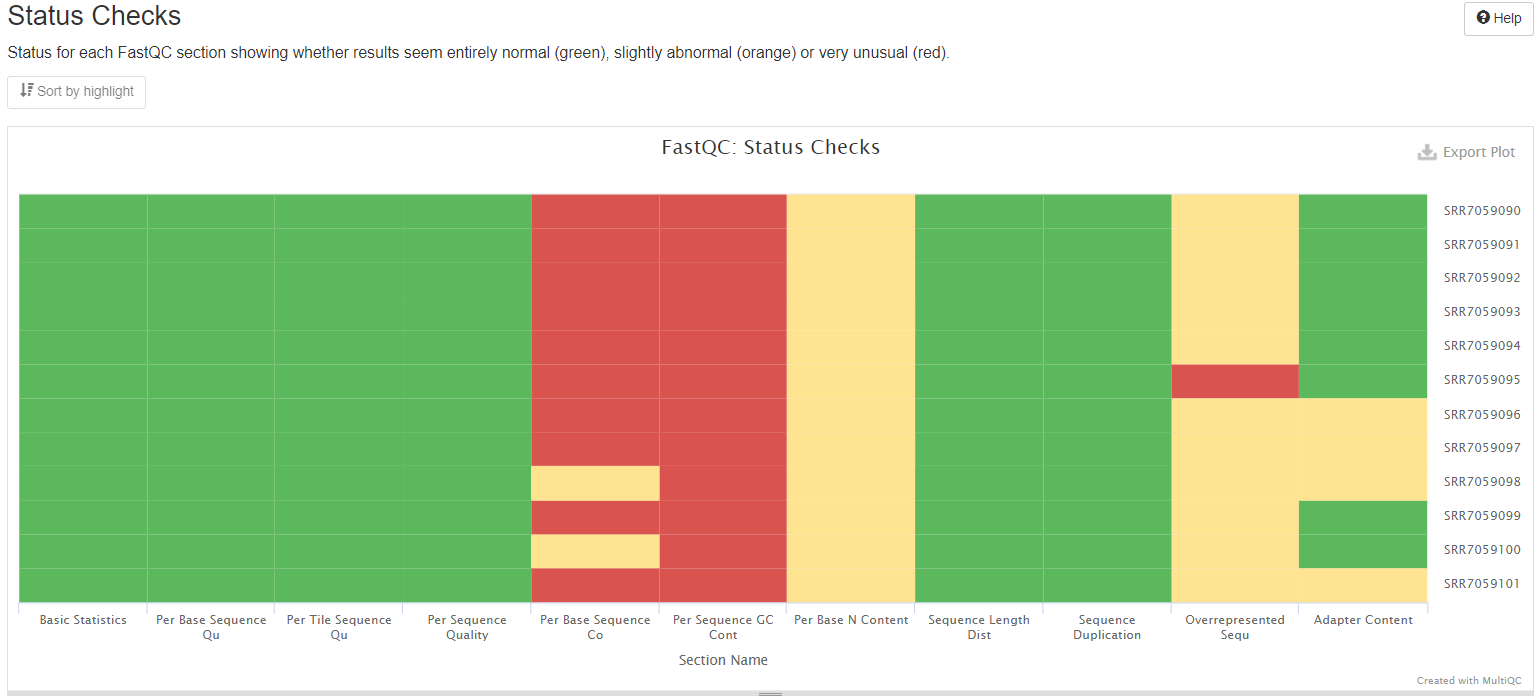
Adapter is actually the part the I spent a lot of time thinking, I kept asking myself questions such as “What type of adapters are being used in this project?”, “Do I even need to trim adapters?”, “If so what would the data look like? Better or worse?” I think knowing the type of adapter being used for the project is important because it’s important to input the right adapter into adapter.fa otherwise it will mess up trimming later. I went to the NCBI page that has all information about this project. I found out that the instrument used is illumina HiSeq 2500, then I went to illumina’s website trying to figure out the type of adapter being used here. I couldn’t. Finally, the way I solved this problem is by running fastqc on these runs and I scrolled down to the bottom of the report where it shows % adapters. I saw one red line in each which indicates the type of adapter being used here is illumina universal adapter.

Is trimming adapters necessary? To answer this question of mine, I did some investigation. I read an article from <https://www.ecseq.com/support/ngs/trimming-adapter-sequences-is-it-necessary> , It says if it’s small RNA sequencing, then adapter trimming is highly necessary whereas if it’s transcriptome sequencing, whole genome sequencing, etc., then trimming is not necessary because adapter contamination can be expected to be so small (due to an appropriate size selection) In this project, I think our source is transcriptomic (found on NCBI project page), but just curious, I want to see how trimming is going to affect the data, I did it.

Conclusions –

After comparing the report before and after trimming, I’d say trimming isn’t necessary going to make the data/report “better” or “worse”, it depends on what our goal is or what we are looking for. Now looking at the report, since there’s only four runs that received warning, what I can do next is to trim only these four and compare again. Also, determining the adapter type and whether they need to be trimmed or not is also crucial.

MultiQC results - (before trimming adapters)



MultiQC results – (after trimming adapters)

