Nhi Nguyen



995: download fastp from github readme

996: make the downloaded file executable

997: move fastp to Desktop

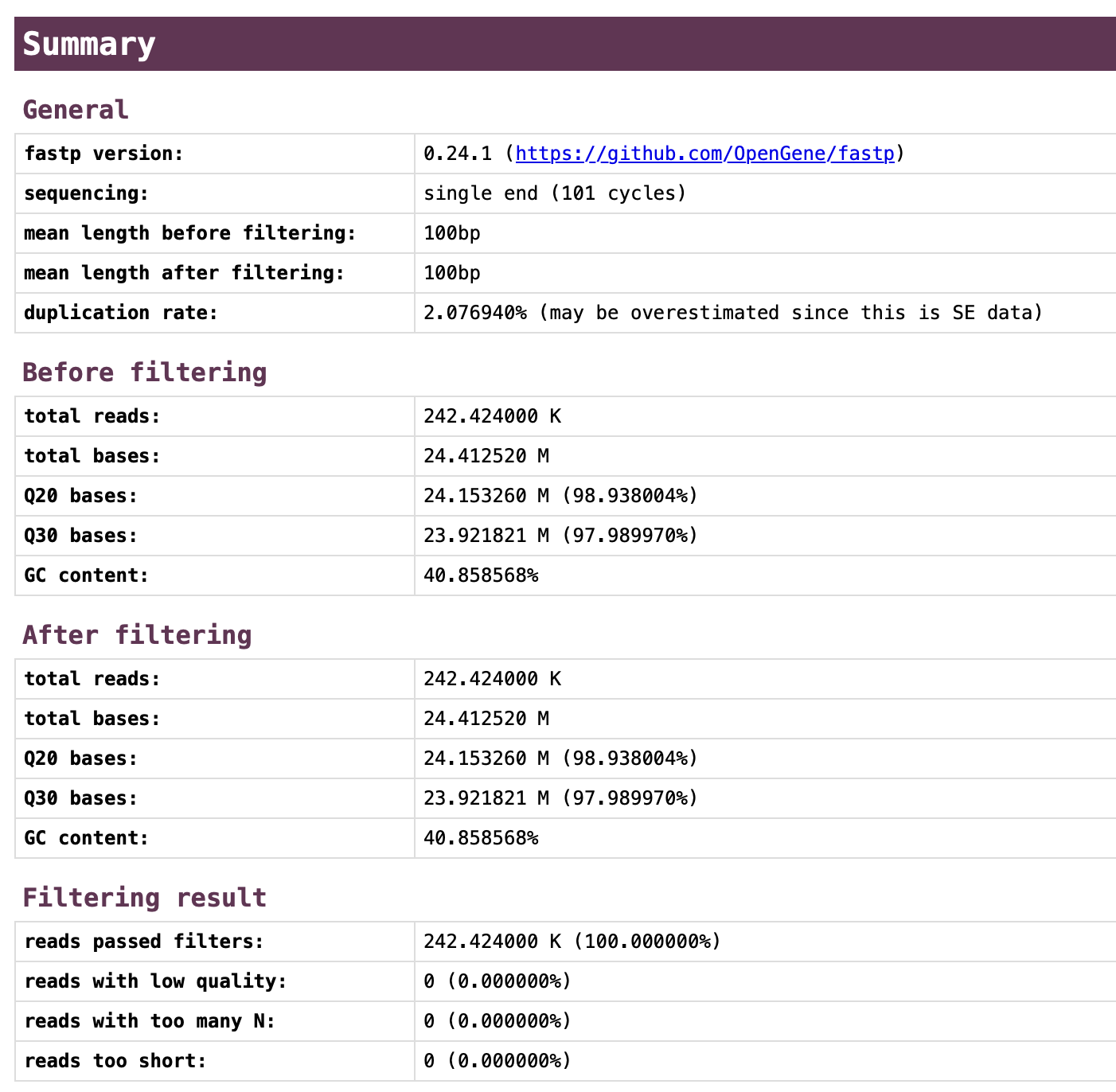
998: try to run it to mke sure it’s working

1000: single-end mode on the reverse read file, -h is to have html

Comment:

On canvas, only the R1 file was given : out.insub732\_2\_R1\_fastp.fastq.  
I looked for the matching R2 file on NCBI and ENA using the sample name insub732\_2, but I couldn’t find it !!!  
Since I didn’t want to use the wrong file or make one up, I ran fastp in single-end mode using just the R1 file.  
This still let me check and clean the reads, and create the HTML and JSON reports for quality.

Fastp : to filtering and trimming



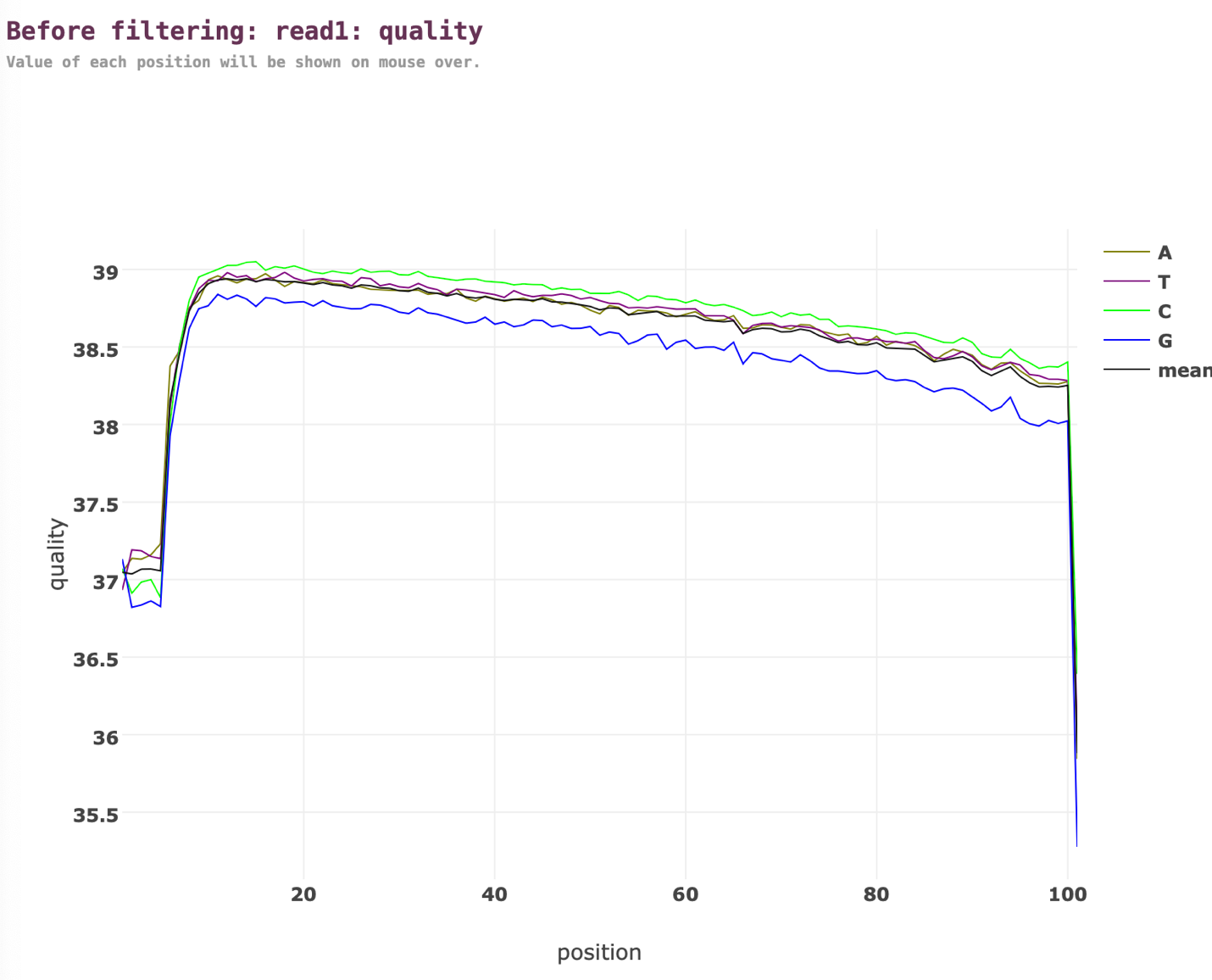
Total read : 242424000 K

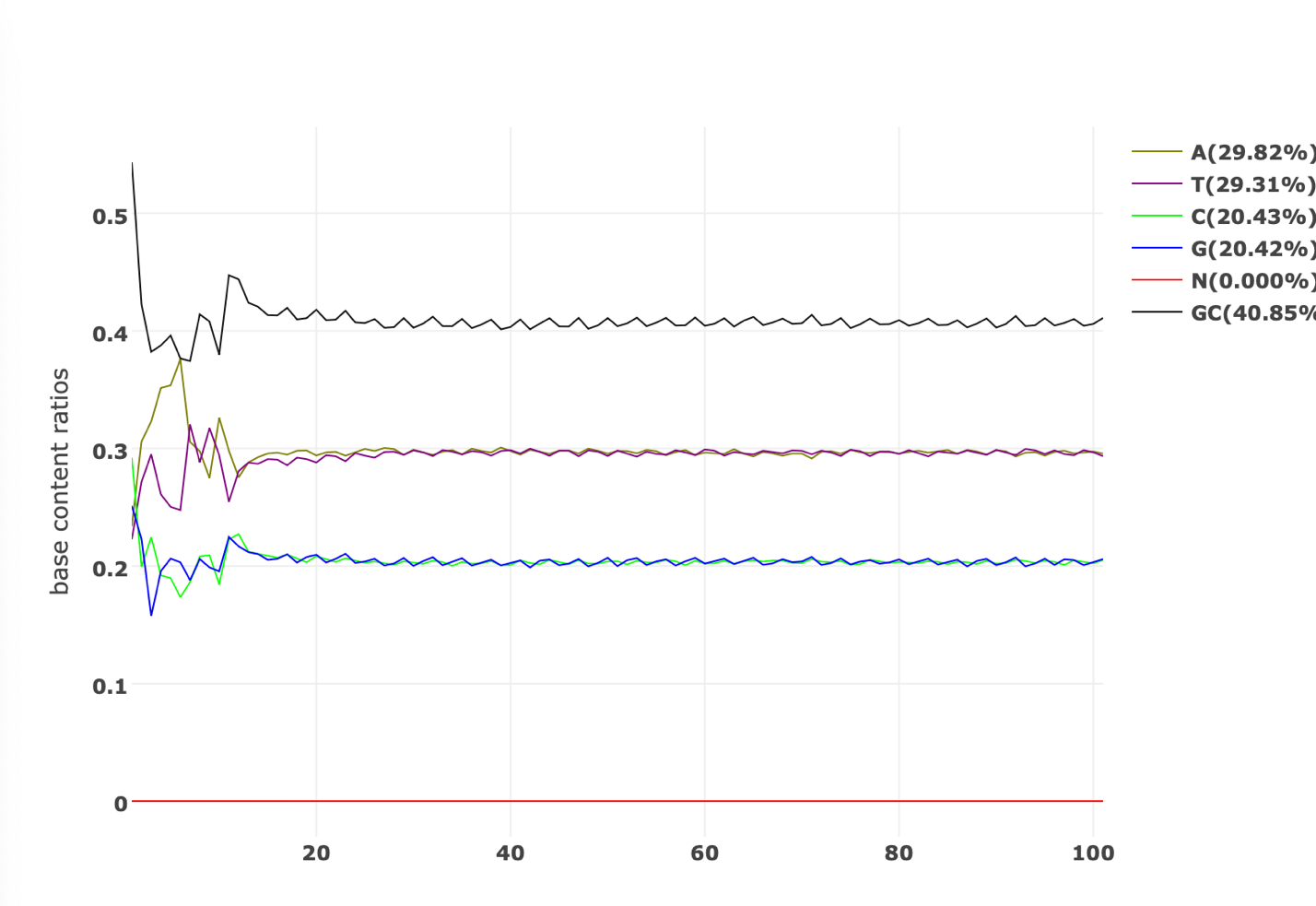
Read kept after filter: 242424000K remain the same

GC ~40.85%

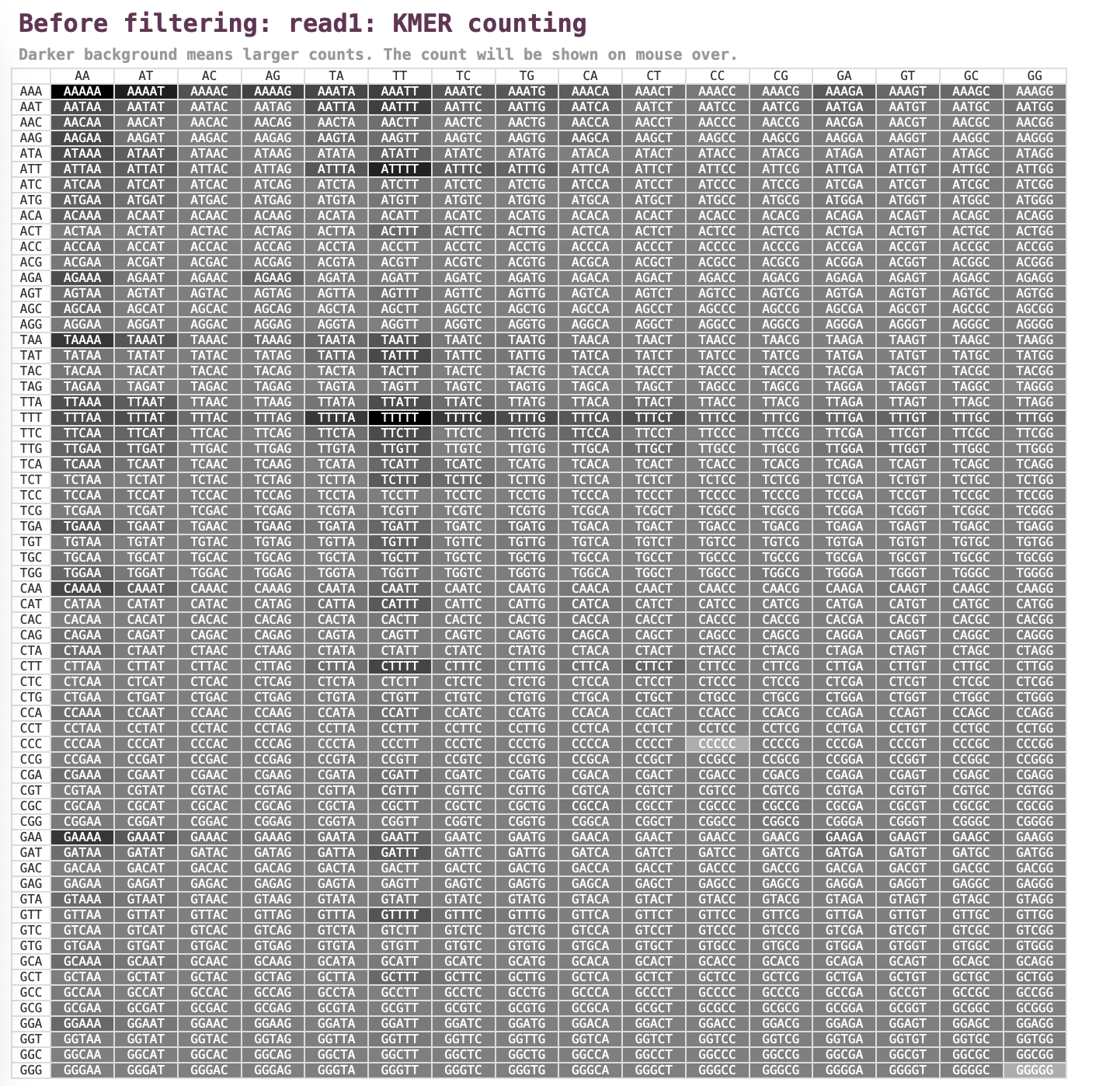
Q30 before ~24% , Q30 after ~ 24 % stay the same

Quality drop after position 60





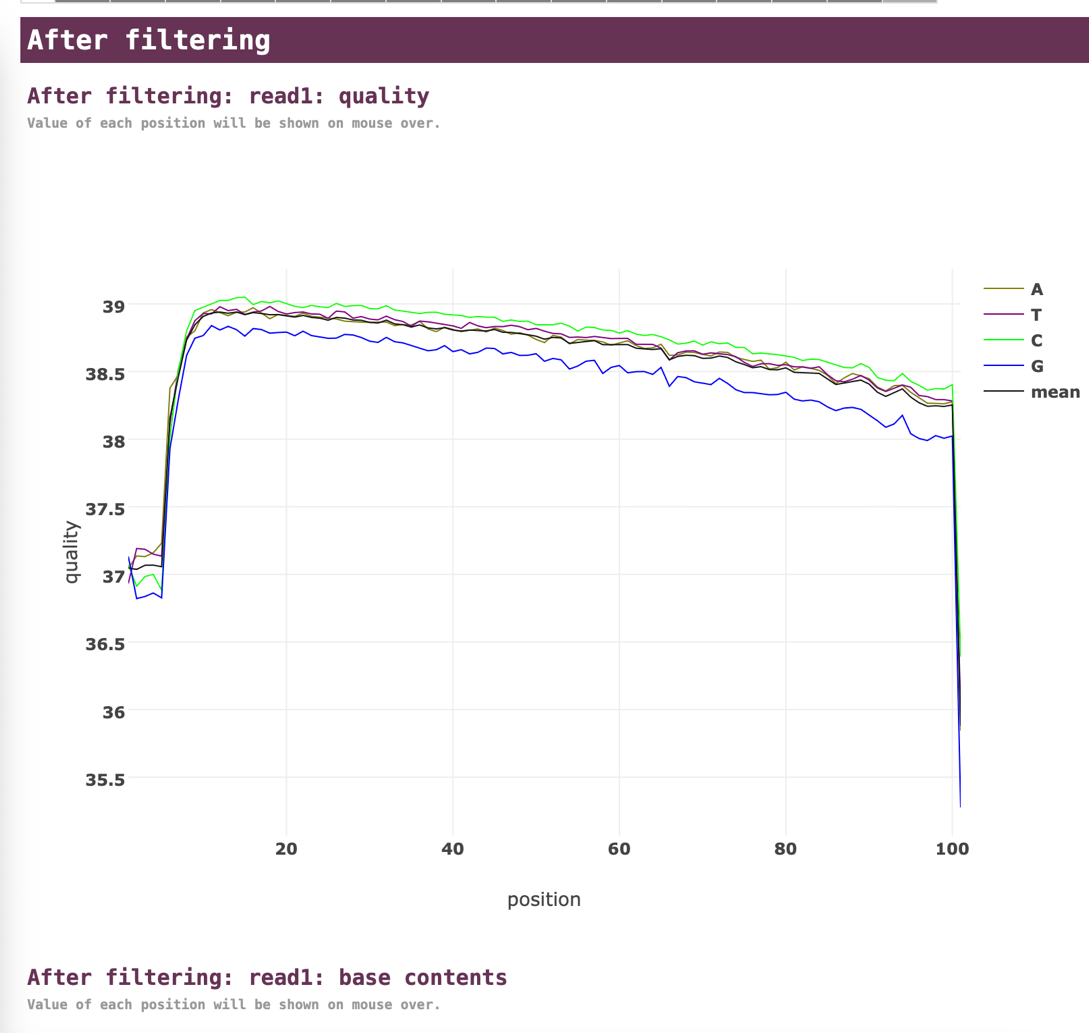
GC ~40.85%

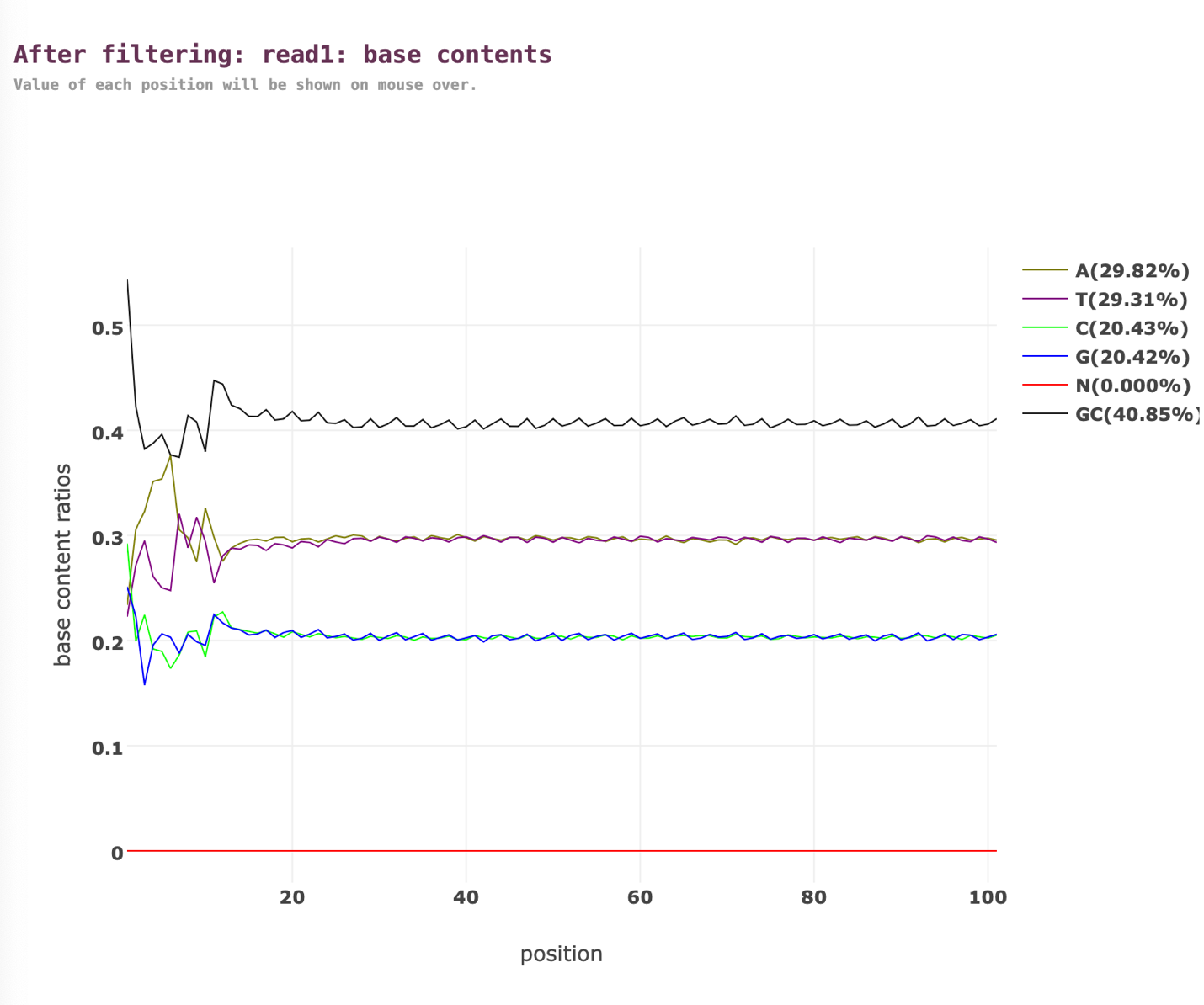


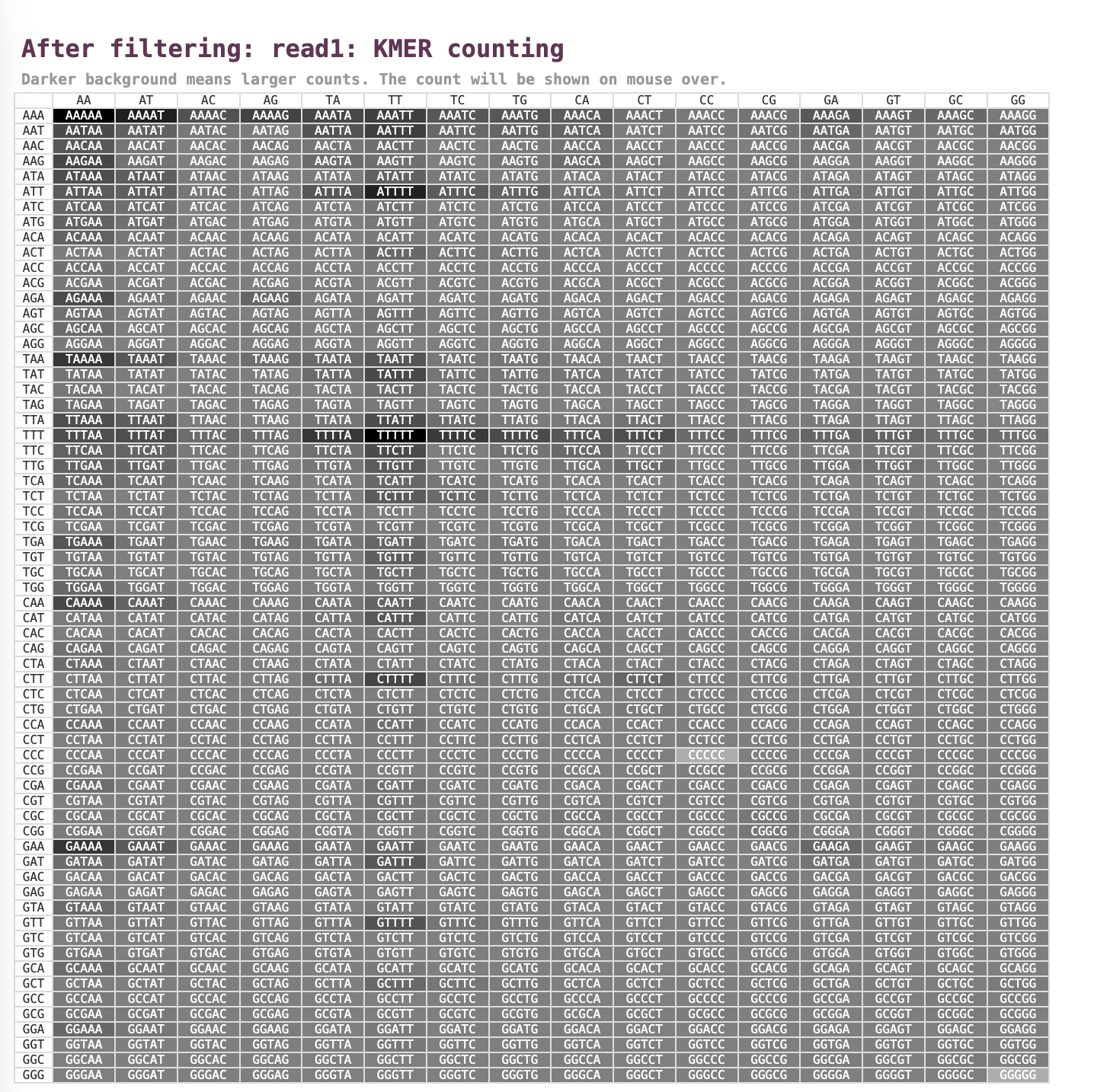
k-mer= 5

darker is more frequent

After filtering :

I don’t see any strong differences.

GC 40.8%, nothing change 



Look the same after,

But if I have 2 file, if it look the same does that mean that the data is stable?

Kaiju Virus Database Analysis

While working on this part of the assignment, I followed the instructions from the Kaiju README to install the program and analyze a metagenomic dataset using the virus database. I tried many different ways, but I kept getting errors and couldn’t get it to work.

First, I installed Kaiju from GitHub using the command line and successfully ran make to build the program. Then I tried to download the virus database file called kaiju\_db\_viruses.tgz. I started with the official FTP link, but it wasn’t working. After that, I searched for other ways to get the file and found HTTPS mirror links. I downloaded the file multiple times using both curl and wget, but each time it was either incomplete or corrupted.

When I tried to extract the file, I kept getting an error that said: "Error opening archive: Unrecognized archive format." Because of this, Kaiju couldn’t find the files it needs to run properly, like nodes.dmp, names.dmp, and the .fmi file. That meant I couldn’t run the classification step.

I really did my best and tried. Please give me some advice and I can keep working on this assignment .