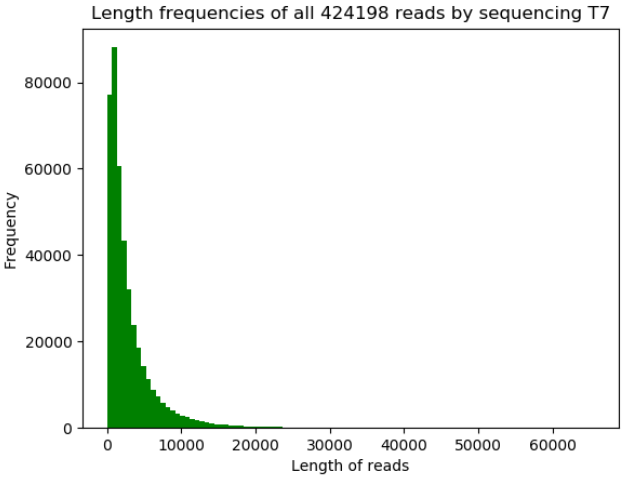
There are three scripts in /home/proj/biocluster/projekte/PHAGES/scripts:

1. plots_for_fastq.py FOR .FASTQ

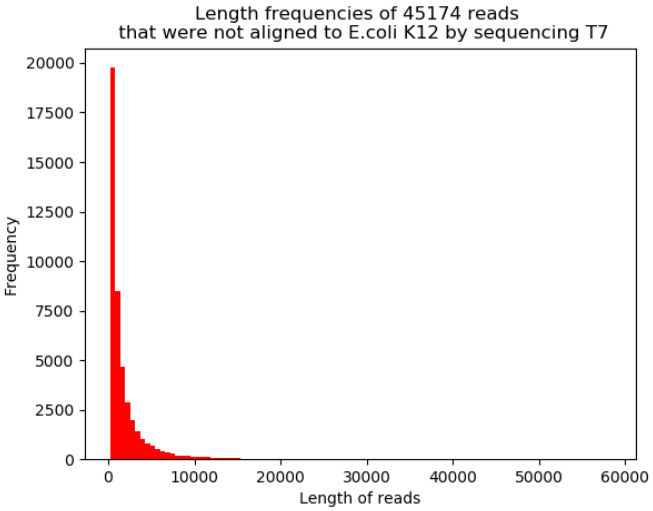
Meant to visualise “distribution of read length in .fastq file”, so by using this script you will see how often which read length appeared.

Type python3 plots_for_fastq.py to see the detailed description:
Please enter 3 parameters:
1. name of the phage (for example: T7)
2. path to .fastq file with all reads
3. path to .fastq file with reads that were left after aligning to E.coli K12
***if you do not have a file mensioned in 2. or 3. just type "no" instead of 2. or 3. parameter respectively



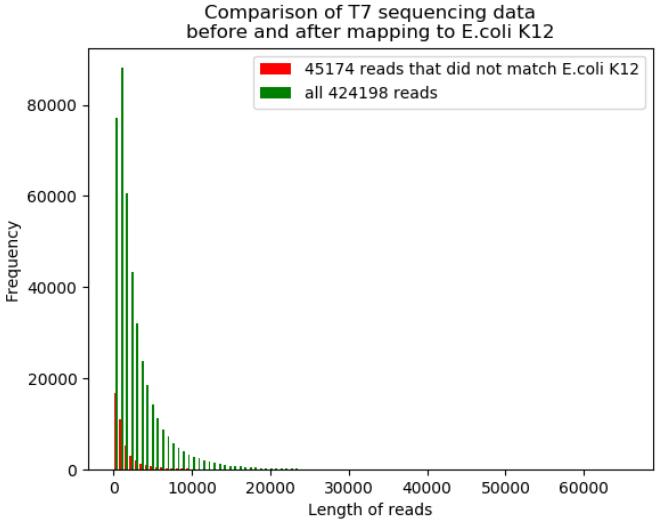
Here you see that there were 424198 reads in t7.fastq and the most of them are somewhere between 0 and ~5000 bp.

Here you can see that in t7notAligned.fastq were only 45174 reads.



But the point of writing this script was to compare those two.

So now hopeful everybody agrees, that the sequencing went not the way everyone expected.



You can also generate only plots for one .fastq file.
python3 plots_for_fastq.py T7 t7.fastq no

To generate these three plots use:
python3 plots_for_fastq.py T7 t7.fastq t7notAligned.fastq

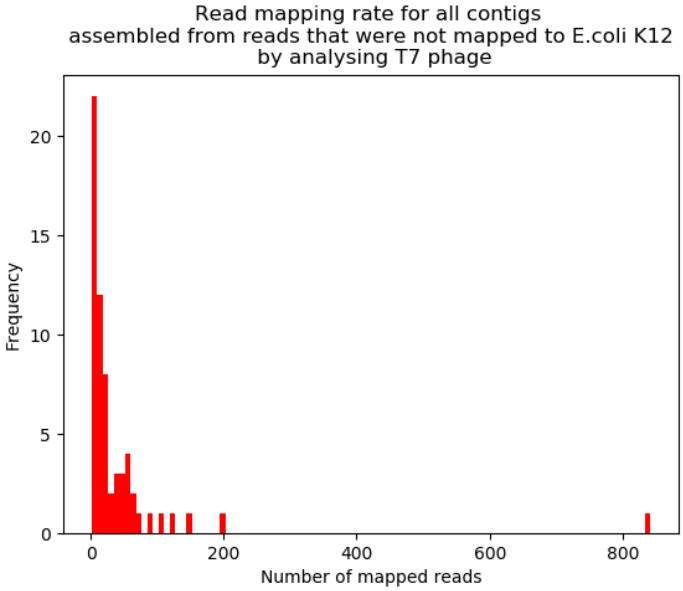
2. plots_for_contigs_fasta.py FOR .CONTIGS.FASTA
This script is for visualising:

- contig length distribution (how often which contig length appeared)
- number of reads contigs were made of (how often which amount of reads were used to construct contigs)

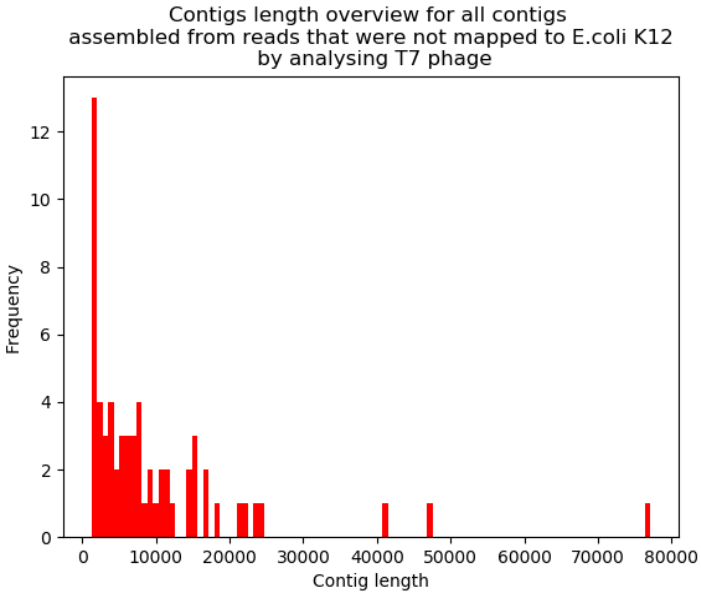
Type python3 plots_for_contigs_fasta.py to see the detailed description:
Please enter 3 parameters:
1. name of the phage (for example: T7)
2. path to .fasta file with contigs assembled from all reads
3. path to .fasta file with contigs assembled from reads that were left after aligning to E.coli K12
***if you do not have a file mensioned in 2. or 3. just type "no" instead of 2. or 3. parameter respectively

Plots below were made by calling:
python3 plots_for_contigs_fasta.py T7 no t7notAligned.contigs.fasta

The whole idea of writing this script was kind of because T7 and this one contig that was constructed out of more then 800 reads:

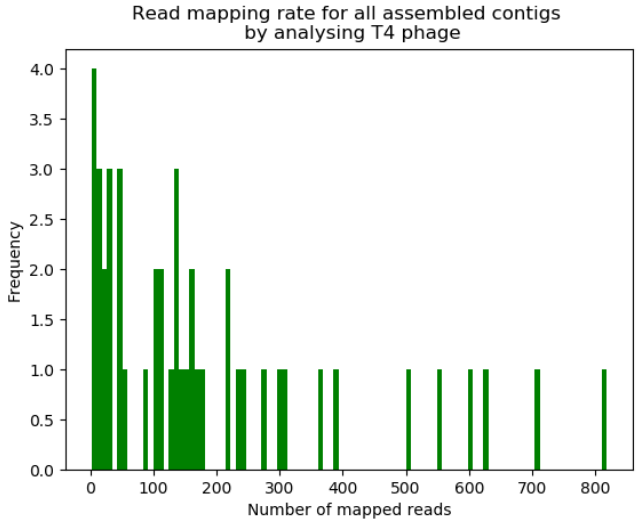


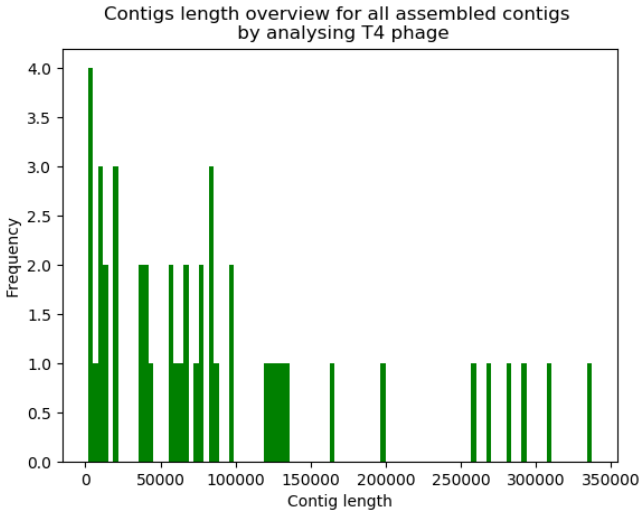
Which is also the longest (len=77151) contig in this file:
Another thing is that there are also 41007 and 46825 bp long contigs. If I remember correctly they were *E.coli*. And the 41007 was made out of 41 reads and 46825 bp one out of 37. Quite short in comparison to the longest contig with 842 reads.

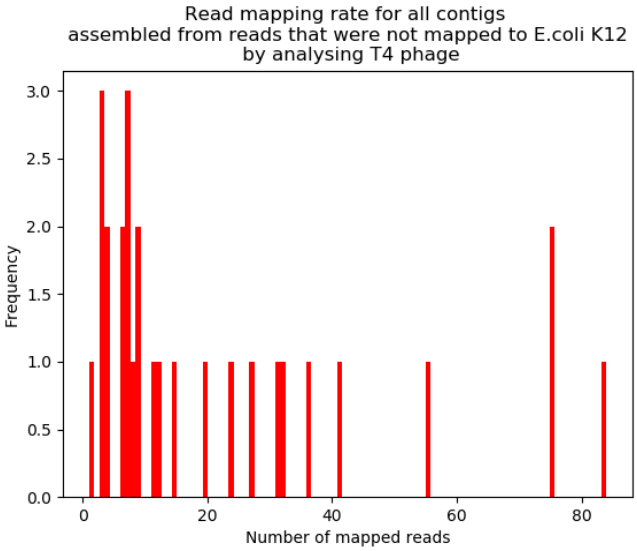


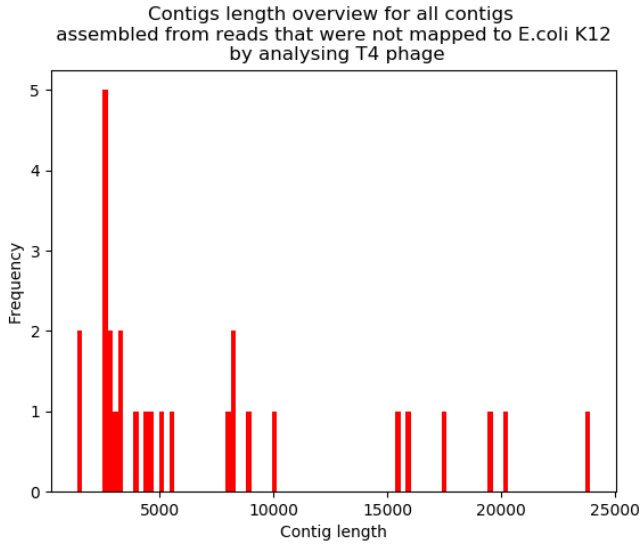
Here is also a possibility to enter two files (contig file before and after mapping to *E.coli*, if you enter some other contig files then you should change the plot title and legend labels in the code. It's very easy: plt.title is for the title and plt.hist(…, labels) is for labels)

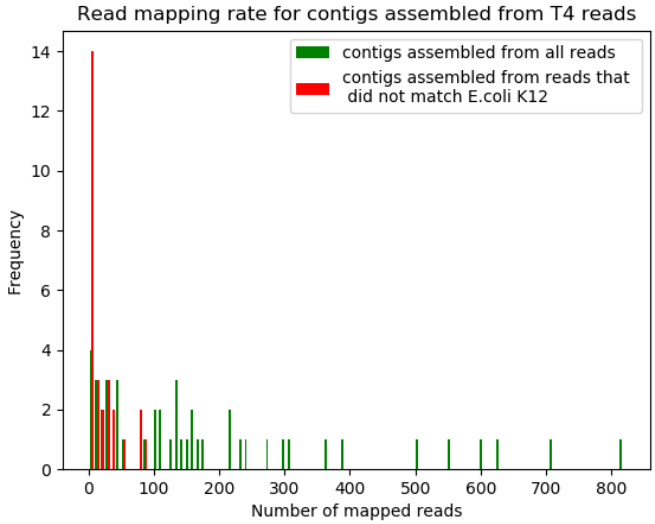
All 6 plots below were made by calling:
python3 plots_for_contigs_fasta.py T4 t4phage.contigs.fasta t4notAligned.contigs.fasta
Don't see the point by calling this script for T4. Data is bad. But maybe you will see something…

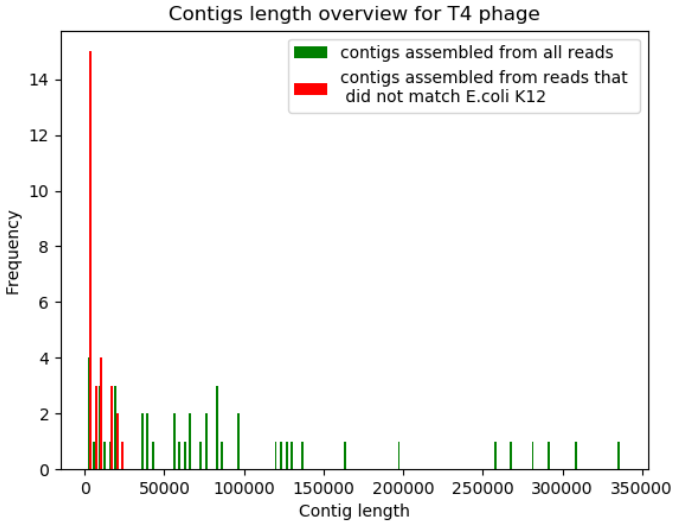








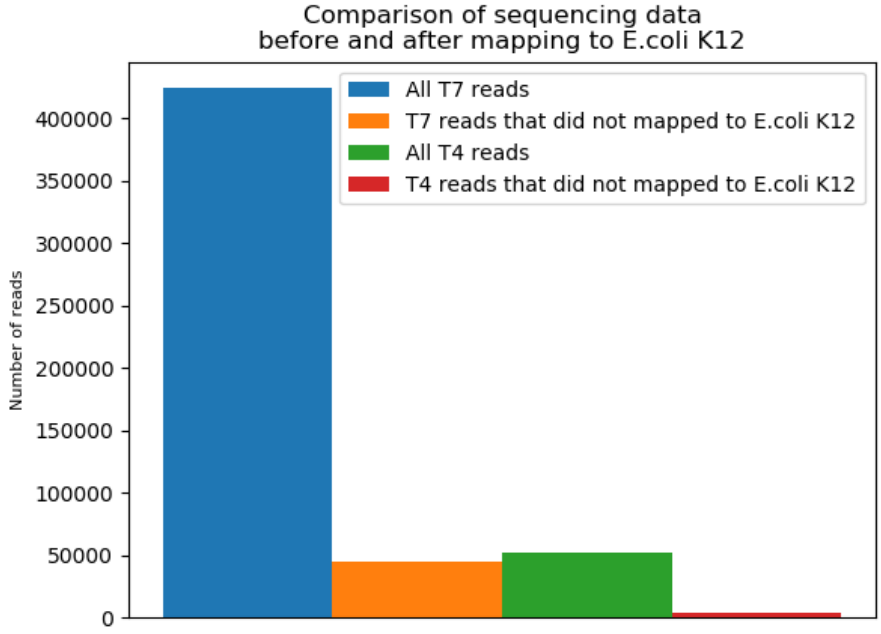




3. compare_fastqs.py FOR A LOT OF FASTQ FILES
The idea of this script was to compare the number of reads between different sequencing data (T4, T7, NES, 3S, FFP) and/or sequencing data after mapping after mapping to *E.coli*. By sequencing data I mean .fastq files.

Type python3 compare_fastqs.py to see the detailed description:
Please enter AT LEAST ONE .fastq read file and lable with the explanation of the file for the plot in this format:
label1-path_file1, label2-path_file2, label3-path_file3,... (Please don't use "-" anywhere else)
Example:
All T7 reads -T7/t7.fastq, Reads that did not mapped to E.coli K12 - T7/t7notAligned.fastq, All T4 reads -T4/t4.fastq Reads, Reads that did not mapped to E.coli K12 - T4/t4notAligned.fastq

Since I am in the middle of nowhere and have almost no internet connection I compared only T4 and T7:
python3 compare_fastqs.py All T7 reads -t7.fastq, T7 reads that did not mapped to E.coli K12 - t7notAligned.fastq, All T4 reads -t4.fastq, T4 reads that did not mapped to E.coli K12 - t4notAligned.fastq



Could be interesting to compare “our” sequencing (T7, T4) with NES, 3S, etc. And sorry, but if you wish to change the title please go into code and change it in plt.title().

Would be great if you could try this on the other data… With love, Rita <3

python3 plots_for_fastq.py T7 t7.fastq t7notAligned.fastq
python3 plots_for_fastq.py T4 t4.fastq t4notAligned.fastq
python3 plots_for_fastq.py NES nes.fastq no *(If I remember correctly here was no mapping to E.coli needed)*
python3 plots_for_fastq.py 3S 3S.fastq 3snotAligned.fastq
python3 plots_for_fastq.py FFP tffp.fastq no *(If I remember correctly here was no mapping to E.coli needed)*

python3 plots_for_contigs_fasta.py T7 t7phage.contigs.fasta t7notAligned.contigs.fasta
python3 plots_for_contigs_fasta.py T4 t4phage.contigs.fasta t4notAligned.contigs.fasta
python3 plots_for_contigs_fasta.py NES t4phage.contigs.fasta no *(If I remember correctly here was no mapping to E.coli needed)*
python3 plots_for_contigs_fasta.py 3S t4phage.contigs.fasta t4notAligned.contigs.fasta
python3 plots_for_contigs_fasta.py FFP t4phage.contigs.fasta no *(If I remember correctly here was no mapping to E.coli needed)*

python3 compare_fastqs.py All T7 reads -t7.fastq, T7 reads that did not mapped to E.coli K12 - t7notAligned.fastq, All T4 reads -t4.fastq, T4 reads that did not mapped to E.coli K12 - t4notAligned.fastq, All NES reads - nes.fastq, All 3S reads - 3s.fastq, All FFP reads - ffp.fastq

= 11 + 22 + 1 = 34 plots
I would definitely do:
1. the last one (compare_fastqs.py) because it's only one plot. And if the comparison is interesting it is worth it.
2. plots_for_contigs_fasta for T7 (I have already started, see before)
3. plots_for_fastq for everything…