Assignment 2

Due Thursday, September 3 before midnight (California time)

Four FASTQ files are placed on Blackboard:

XI1_ATCACG_L001_R1_001.fastq
XI1_ATCACG_L001_R2_001.fastq
RETT-1_S1_L001_R1_001.fastq
RETT-1_S1_L001_R2_001.fastq

These are next generation sequencing files from a paired-end run. The XI1 reads are mostly 250 base pairs long, the RETT reads are mostly 150 base pairs long. Complicating the assignment, there are a few reads that are a few bases shorter or longer than 250 or 150.

1. Write a Python function that counts the number of reads with the same, shorter and longer length than expected. For each of the four files print the out. Compare the results and describe your finding [3pt].

2. Write a Python function that for each of the positions 1 to 250 (or 1 to 150) will compute the fraction of reads in the file with quality scores greater than or equal to 30 at that position. For each of the four files plot the output. Compare the results and describe your finding [5pt].

3. Write another function that for k from 0 to 250 (or from 0 to 150) will count the number of reads in the file with exactly k positions with quality scores greater than or equal to 30. For each of the four files plot the output [Bonus 2pt].

For the plots, please use the package matplotlib.pyplot mentioned in the class.

Turn in the code for the two (or three) Python functions and the answers into one file in Jupyter Notebook format (.ipynb). Use the “Alignment 2” link on Blackboard/Assignments/Assignment 2 to submit this file.