

HW 8 due Thursday, March 7

This is a continuation of the problem we have discussed in class. From Content, download the files “XI1_ATCACG_L001_R1_001.fastq” and “XI1_ATCACG_L001_R2_001.fastq”. These are abbreviated next generation sequencing files from a paired-end run. Each read is 250 base pairs long.

1. Write a function that for each of the 250 positions will compute the fraction of reads in a file with quality scores greater than or equal to 30 at that position. We expect the sequence quality to be poorer at the beginning and the end of the reads.
2. Write another function that for k from 0 to 250 will count the number of reads in the file with exactly k positions with quality scores greater than or equal to 30. This function will enable us to make a histogram showing the number of good and bad reads.

Make a plot of the output of the first function for the two files. Make a histogram of the output of the second function for the two files. Turn in your code.