**BIOMI 609 Computational Genomics and Bioinformatics**

**Spring 2022**

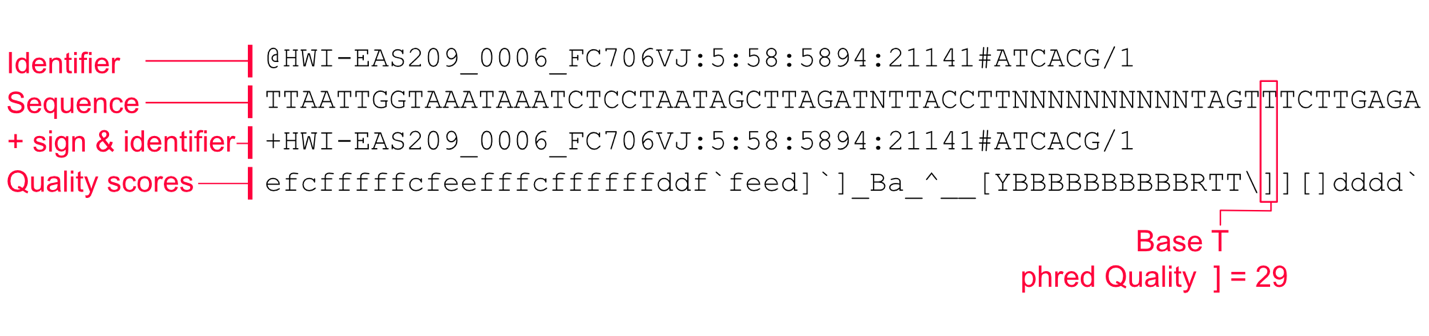
**Dr. Arun Sethuraman**

**Assignment 1 (20 points)**

**Due on Friday, 2/4/2022 before class at 8 AM – you will upload two things - your code, and output from the provided FASTQ file to Canvas.**

1) You will write a program in Python/R/C (just pick a language you like – I’m listing ones here that I prefer) that will take as input a FASTQ file and print the distribution of quality scores across all reads. You can summarize the distribution of Q scores at each base with a statistic of your choice (e.g. mean, mode, median, quantile distribution). If you’d like, you can also plot the distribution of Q scores as a box plot much like what’s generated by FASTQC. You will then run your program on the provided FASTQ file, and obtain the output from it.

Note that a FASTQ file has the following format:



This format is repeated for each read.

Hints: The Illumina PHRED quality score encoding can be found here: <https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm>

The idea is real simple; for each character in the quality score line, the ASCII value of that character - 33 = Q. Thereon, Q = -10log10Pe, where Pe is the probability of error in calling that nucleotide base.

Here are functions in various languages to convert to the ASCII encoding:

Python: ord()

R: iconv()

C: When you scanf() the character, you scanf() with a %c, which automatically converts it into its ASCII encoding