PARSING A FASTQ AND COMPUTING QUALITY STATISTICS

check_fastq.ipynb download fastq from aula.esci.upf (inside reads/unknown_illumina_2024.fastq)

1. Use SeqIO.parse to print the first record in the fastq
2. What's the read length?
3. How many reads are stored in this file?
4. Can you print the quality score? (check https://biopython.org/wiki/SeqRecord)
5. Plot the mean quality at every position in the reads: start with mean quality at position 1 across all reads, then at position 2, and so on until N (that is the length of the reads).
6. Show lines with the mean quality score and the 95% conficence interval (2 s.d.)
7. Convert the qualities to error probabilities using the Phred Quality Score equations. Plot them, at which positions is higher? what's the expected error rate of them?
8. Practical Assessment: Identify the reads origin. (How would you find out from which genome come these reads? To which species they belong? Please describe the method used and the reliability of the results)

Consulting materials: https://en.wikipedia.org/wiki/FASTQ_format

https://biopython.org/wiki/SeqIO

https://biopython.org/wiki/SeqRecord