Department of Computing and Information Systems
COMP 90016
Workshop 2

The lectures introduced how DNA is read into strings by sequencing machines.
This workshop is going to explore the kind of data coming out of sequencing machines in a bit more detail. The folder /home/subjects/comp90016/tutorials/week2 on digitalis contains a FASTQ file of read data.

1. Take a glance at the data by calling *head reads.fastq* on the command line. You should be able to recognise the format introduced in the lecture.
2. Assess the read length in the data. Write a short Python script that reads in the input data and then finishes with a few lines of text commenting on the length of the read sequences.
   Pysam offers functionality to read FASTQ data and further analyse the reads.
   What are the different read lengths observed in the data? Discuss in the lab.
3. Extend your program to further investigate the quality scores of the reads. The lecture mentioned that the quality sequence assigns a quality value to each of the bases in the read, which reflects on the confidence of the sequencing machine that this base is the correct call. Quality values are encoded via the so-called Phred scale as the log of a probability:
   $$Q = -10\ log_{10}P$$
   With *P* the probability of the base being correct. See also https://en.wikipedia.org/wiki/Phred_quality_score. A range of quality scores from 0 to typically 40-50 (40 is equal to 99.99% confidence) is mapped onto the ASCII table to represent the score with a single character. For Illumina quality score this mapping is done by adding 33 to the quality score and representing the score with the respective ASCII character (see https://en.wikipedia.org/wiki/ASCII).
   Take a look at the read data again and observe the varying characters encoding base quality.
   Write a Python script that reads in all of the quality strings and summarises the distribution of base quality values for each position in the reads. That is, what is the minimum, maximum, average… etc. quality for each of the reads' first position? For the second? … etc. I recommend using Matplotlib, which is installed on digitalis, to plot a boxplot (one box for each of the positions).
   What can you observe about differences in quality from one position to the other in the data? How does this relate to read length of sequencing data?