Homework 4: Python and data analyses

1. Using python write a tool which generates a random subsampling tool for sequences. Given a FASTA sequence database file, which has 100,000 sequences, generate a new file which is a random subset these sequences selecting only 10% of them. Make this 10% an option in the program so it is easy to change to 20%, etc.

See the script `rand_shuffle_seqs.py` in the homework template

2. Run RNA-Seq analysis of this Bacteria light induction project. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA315575. Only focus on 2 time points, the data can be downloaded with `download_bacteria_light_fastq.sh` script

- present a table of the genes with 2x upregulation in the 4hrs (SRR3234522) vs the 1hr (SRR3234519) timepoint
- Use Kallisto as easiest way to get expression computed for the timepoints
- can also explore GSNAP or Hisat2