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 RNAseq analysis to compute the gene expression for two experiments. We will use data from this paper on light induction in bacterium Prochlorococcus

See this paper:

- Thompson et al. PLoS One 2016; 11(10): e0165375 doi: 10.1371/journal.pone.0165375
- Here is a view of the sequencing data. All data are on cluster in this folder `/bigdata/gen220/shared/projects/HW4/hw4_hyphaltip`

Also you can get it from the web directly yourself. https://www.ebi.ac.uk/ena/data/view/PRJNA315575

The data for one timepoint: 0hrs light

Go and get the fastq files to process - see the download script

Light 0hr

- 0hr_1.fastq.gz
- 0hr_2.fastq.gz

The data for another timepoints: 4hrs light

The fastq files to process

- 4hr_1.fastq.gz
- 4hr_2.fastq.gz

Here is the Prochlorococcus marinus subsp. pastoris str. CCMP1986 genome: GCF_000011465.1. The RefSeq for this genome is at that link.

Can be obtained from this link

And the GFF

- Use Hisat2, samtools, and stringtie to generate a table of expression for these two experiments
- Write a python script that will generate a report with 5 columns of data. Gene Name, Gene Location, Gene Length, FPKM exp1, FPKM exp 2
- Using R or other tools, make these plots
  - Plot gene expression from exp1 vs gene expression of exp 2
  - Plot gene length vs expr1 expression (FPKM). Is there a relationship?
  - Plot gene location vs expr2 (FPKM). Is there any relationship?

- print a list of genes which are > 2 fold higher in the 4hr time point?