Session 7 Recitation

1) Log into the MSU HPCC and a dev-node. Either in your home directory or scratch, make a new directory CMSE890_Sec304/Session7.

2) Look at /mnt/research/CMSE-bioinformatics/week3/fastq/SRR2012208_1.fastq. This file contains next-generation sequencing reads in fastq format, which means the first line for each data point starts with '@' and is a unique identifier. The second line is the sequence of the read. The third line is a '+' sign. The fourth line is a quality score. Write a Python script that will go through this file and convert it to a fasta file, which is where there is a unique identifier with a '>' at the front of the first line and the second line is the sequence. Do not print the third or fourth lines in the output file.

3) Copy the files 20180416_trna_seqs_uniqueIDs.txt and 20180416_trna_seqs.fasta from D2L to either the HPCC or to your laptop, depending on where you want to work.

4) Imagine that you downloaded a series of sequences from img.jgi.doe.gov (20180416_trna_seqs.fasta) and ran them through a software package that identified which sequences were of the most interest for a given project. But the software only reported the unique IDs for each sequence (20180416_trna_seqs_uniqueIDs.txt), not the sequence itself or complete metadata. Write a Python script to read in the unique IDs and save them to a data structure. Then add code to read the full sequence file and write a new file containing the metadata and sequence data if the unique identifier is in the list of unique IDs.