Homework 1

Create a hw1 repository on github and make it within the biodataprog organization

Write a bash / shell script to accomplish the following tasks. Remeber you can make a shell script as a series of commands just as you would type on the command line.

1. **Getting data**
    - Use a cmdline downloading program (e.g. curl) to download this BLAST report from the web: Ecoli-vs-Yersinia.BLASTP.tab.gz or https://biodataprog.github.io/2018_programming-intro/data/Ecoli-vs-Yersinia.BLASTP.tab.gz
    - Print out how big this file is (in kilobytes)?
2. **Compressing and uncompressing**
    - Uncmpress this file with gunzip.
    - How big is the uncompressed file (in kilobases)?
3. **Counting and viewing** (Using the same BLAST report)
    - Print out the first 25 lines of the file
    - Print out the last 3 lines of the file
    - Print the total number lines in the file
4. **Sorting**
    - Obtain the data file Nc3H.expr.tab. This file contains gene expression value assigned to each gene in the *Neurospora crassa* (a fungus) genome. https://biodataprog.github.io/2018_programming-intro/data/Nc3H.expr.tab
    - Sort the file based on the FPKM column (which is the gene expression) (write out to a new file called Nc3H.expr.sorted.tab). Remember the in-class introduction to sorting which ignores the header. Note your sorting will be more complicated than what is depicted here.

```
(head -n 1 <filename> && tail -n +2 <filename> | sort) > newfile
```

    - Print out a list of the top 10 most highly expressed genes based on FPKM.
5. **Finding and Counting**
    - Report the number of CDS features in this genbank file - see for example this explanation of a genbank file if you are not familiar. https://biodataprog.github.io/2018_programming-intro/data/D_mel.63B12.gbk
    - Print how many sequence alignments are 100% identical in the previously downloaded file `Ecoli-vs-Yersinia.BLASTP.tab`.

- Print how many sequence alignments are 90% identical or better in the previously downloaded file `Ecoli-vs-Yersinia.BLASTP.tab`. (Hint, review the options in awk for filtering or processing column delimited data).

6. **Sort and Uniq**

   - Obtain the file listing the standard codons and amino acids translations. codon_table.txt. Column 1 is the codon, Column 2 is the amino acid, and Column 3 is the Amino acid written out. https://biodataprog.github.io/2018_programming-intro/data/codon_table.txt
   - Print out the name or symbol of each amino acid and how many codons encode that amino acid. (e.g. Lysine/K is encoded by 2 codons)