Python 5 Problem Set

1. Write a script to do the following to Python_05.txt

- Open and read the contents.

- Uppercase each line
- Print each line to the STDOUT

2. Modifiy the script in the previous problem to write the contents to a new file called "Python_05_uc.txt"
3. Open and print the reverse complement of each sequence in Python_05.fasta. Make sure to print the output in fasta format including the sequence name and a note in the description that this is the reverse complement. Print to STDOUT and capture the output into a file with a command line redirect '>'.
4. Open the FASTQ file Python_05.fastq and go through each line of the file. Count the number of lines and the number of characters per line. Have your program report the:
    - total number of lines
    - total number of characters
    - average line length
5. You are going to generate a couple of gene list that are saved in files, add their contents to sets, and compare them.

Generate Gene Lists:

Get all genes:

1. Go to Ensembl Biomart.
2. In dropdown box, select "Ensembl Genes 90"
3. In dropdown box, select "Alpaca Genes"
4. On the left, click Attributes
5. Expand GENE:
6. Deselect "transcript stable ID".
7. Click Results (top left)
8. Export all results to "File" "TSV" –> GO
9. Rename the file to "alpaca_all_genes.tsv"

Get genes that have been labeled with Gene Ontology term stem cell proliferation

10. Click "Filters"
11. Under "Gene Ontology", check "Go term name" and enter "stem cell proliferation"
12. Click Results (top left)
13. Export all results to "File" "TSV" –> GO
14. Rename the file to "alpaca_stemcellproliferation_genes.tsv"

Get genes that have been labeled with Gene Ontology term pigmentation

15. Click "Filters"
16. Under "Gene Ontology", check "Go term name" and enter "pigmentation"
17. Click Results (top left)
18. Export all results to "File" "TSV" –> GO
19. Rename the file to "alpaca_stemcellproliferation_genes.tsv"

Open each of the three files and add the geneIDs to a Set. One Set per file.

A. Find all the genes that are not cell proliferation genes.
B. Find all genes that are both stem cell proliferation genes and pigment genes.

Note Make sure to NOT add the header to your set.

Now, let do it again with transciption factors.

1. Go back to your Ensembl Biomart window
2. Deselect the "GO Term Name"
3. Select "GO Term Accession"
4. Enter these two accessions IDs which in most organisms will be all the transcription factors

- GO:0006355 is "regulation of transcription, DNA-dependent".
- GO:0003677 is "DNA binding"

5. Click Results (top left)
6. Export all results to "File" "TSV" –> GO
7. Rename the file to "alpaca_transcriptionFactors.tsv"

Open these two files: 1) the transcription factor gene list file and 2) the cell proliferation gene list file. Add each to a Set, One Set per file

A. Find all the genes that are transcription factors for cell proliferation

Now do the same on the command line with `comm` command. You might need to `sort` each file first.