Ungraded exercise

This exercise is not going to count toward your grade and is only made available for those of you who wish to practice your Bash skills

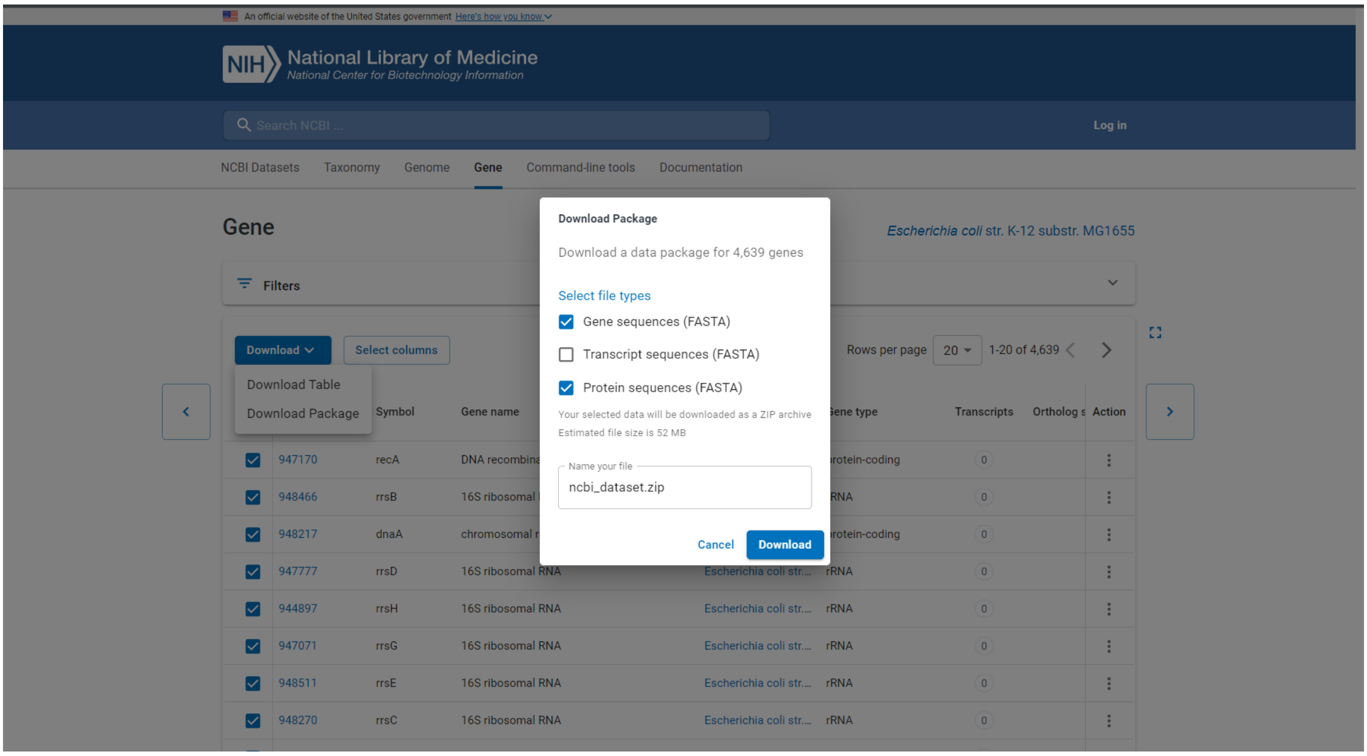
Premise

For week 3, I provided you with a directory full of gene and protein sequence files so that you could use regexes to sort them out. In order to provide you with that dataset, I needed to first generate it. That process used many of the Bash concepts we have covered in class.

This exercise is to figure out how to generate a dataset similar to that I gave you starting with the same inputs that I did.

Setup

You must first download the gene and protein sequences used in the assignment. I retrieved them from the NCBI Gene database at this page. To download the desired data, select all entries by clicking the checkboax in the table header (you should see all checkboxes become checked). Next click download, and then "Download Package" as shown in the below image



Escherichia coli str. K-12 substr. MG1655

Download Package

Download a data package for 4,639 genes

Select file types

☑ Gene sequences (FASTA)

□ Transcript sequences (FASTA)

☑ Protein sequences (FASTA)

Your selected data will be downloaded as a ZIP archive

Estimated file size is 52 MB

Name your file: ncbi_dataset.zip

		Symbol	Gene name		ene type	Transcripts	Ortholog s	Action
☑	947170	recA	DNA recombina		rotein-coding	0		⋮
☑	948466	rrsB	16S ribosomal		RNA	0		⋮
☑	948217	dnaA	chromosomal r		rotein-coding	0		⋮
☑	947777	rrsD	16S ribosomal RNA	Escherichia coli str....	rRNA	0		⋮
☑	944897	rrsH	16S ribosomal RNA	Escherichia coli str....	rRNA	0		⋮
☑	947071	rrsG	16S ribosomal RNA	Escherichia coli str....	rRNA	0		⋮
☑	948511	rrsE	16S ribosomal RNA	Escherichia coli str....	rRNA	0		⋮
☑	948270	rrsC	16S ribosomal RNA	Escherichia coli str....	rRNA	0		⋮

Once you have downloaded and unzipped your files you should have a directory structure like the following:

```
$ tree
.
└── ncbi_dataset
    └── data
        ├── data_report.jsonl
        ├── data_table.tsv
```

```
          ├── dataset_catalog.json
          ├── gene.fna
          └── protein.faa

2 directories, 5 files
```

Of the downloaded files, you only want the gene.fna and protein.faa files. Those contain your sequences.

Steps

To generate the dataset you received, I performed the following steps:

1. Split the gene and protein multifasta files into one file per gene and one file per protein. Name each file according to the gene or protein name contained in the existing header.
2. Change protein filenames to begin with an uppercase letter.
3. Delete any files that do not follow the gene/protein name convention of 4 characters like dnaA or DnaA.
4. Change the extension of half or either your gene or protein names to .fasta. Change the other half to .fa
5. For each gene or protein file changed, make the opposite change to the corresponding gene or protein file (i.e., if you changed dnaA.fna to dnaA.fasta, change DnaA.faa to DnaA.fa)
6. Copy all modified gene and protein files into a single directory

The reason for the renaming so that file extensions of corresponding files are different is to deal with a Windows filesystem quirk. Ubuntu files are case sensetive. You can therefore have both DnaA.fasta and dnaA.fasta on Ubuntu, but on Windows, those files are considered the same. To get around those filename clashes, I changed the extensions. That way, no two files would have the same name.

As a bonus challenge, see if you can come up with a way to choose half of your files so that you won't have an obvious repeating pattern of file extensions. See if you can make it look like filenames were randomly changed.