**Genomes mini-project**

I. Create a repository on GitHub for your Genome mini-project

In a README.md file, give basic information on the genome, including:

- Size of genome
- Structure (such as, linear or circular, how many chromosomes, how many plasmids)
- Estimated number of protein-coding genes
- Where to access the genome sequence and at what coverage it was sequenced
- Any other interesting facts about the genome that turn up

On your own, search for sites from which you can download the genome sequence of your organism and make a note of them for later (paste the url in your lab notebook, for example). Here are the accession numbers for the genome sequences you should use (depending on your group's organism):

Borrelia: GCA_000008685.2_ASM868v2
C. elegans: GCF_000002985.6
Ambystoma: GCA_002915635.3

II. Download the ncbi_datasets tools. You'll need these in order to download your genome data. Log onto Discovery and cd to your "scratch" account:

```
cd /scratch/<username>
```

Download the NCBI command line tools (datasets and dataformat)

```
curl -o datasets https://ftp.ncbi.nlm.nih.gov/pub/datasets/command-
line/v2/linux-amd64/datasets
```

Then:

```
curl -o dataformat https://ftp.ncbi.nlm.nih.gov/pub/datasets/command-
line/v2/linux-amd64/dataformat
```

Make them executable:

```
chmod +x datasets dataformat
```

***Use the `ls` command to check that they are there (should see "datasets" and "dataformat" in the directory).***

III. Download the genome sequence to your Discovery Cluster "scratch" account (`/scratch/<username>`) using good practices for how you store the files (in other words, set up folders and subfolders that have a logical structure).

Download your genome using the accession number for your group's genome (for example, GCF000002985.6) by entering:

```
./datasets download genome accession <accession-number>
```

After the download, use `ls` to see that an NCBI zip file is now in the directory. Unzip it by entering:

```
unzip ncbi_dataset.zip
```

You'll see some files being "inflated". When the command prompt returns, use the `ls` command again and you should see a directory 'ncbi_dataset' that contains the unzipped files.

***

When you use `ls` to see what's in a "room", add the -l (lowercase L) flag so that you can distinguish between files and directories. For example, below is what I see as I travel down the subdirectories within my "borrelia_dataset". I used `ls` until I got down to what looked like the sequence file (GCF_000008685.2) then I used `ls -l` and found out it is actually another directory:

```
[jhenzy@login-00 borrelia_dataset]$ ls
data_bor
[jhenzy@login-00 borrelia_dataset]$ cd data_bor/
[jhenzy@login-00 data_bor]$ ls
assembly_data_report.jsonl  dataset_catalog.json  GCF_000008685.2
[jhenzy@login-00 data_bor]$ ls -l
total 4
-rw------- 1 jhenzy users 2855 Jan 27 13:34 assembly_data_report.jsonl
-rw------- 1 jhenzy users  339 Jan 27 13:34 dataset_catalog.json
drwxr-xr-x 2 jhenzy users 4096 Jan 28 17:55 GCF_000008685.2
[jhenzy@login-00 data_bor]$
```

This makes it clear that GCF_000008685.2 is a directory and NOT the sequence file. To get to the sequence file, I'll need to change into that directory:

```
[jhenzy@login-00 data_cel]$ cd GCF_000002985.6/
[jhenzy@login-00 GCF_000002985.6]$ ls -l
total 99161
-rw------- 1 jhenzy users 101540352 Jan 26 12:13 GCF_000002985.6_WBcel235_genomic.fna
```

Within the GCF_000008685.2 directory, I see the actual sequence file, `GCF_000002985.6_WBcel235_genomic.fna`

***

IV. Analysis

To analyze your organism's genome sequence, you'll use two programs that are available as modules on Discovery: **seqtk** and **emboss**.

Remember that you need to load the modules into your workspace (scratch) to use them.

```
module load emboss
```

```
module load seqtk
```

When you download your organism's genome sequence, multiple fasta sequences are usually included in one file. You can use the emboss command '`infoseq`' to see what a file contains:

```
infoseq <filename>
```

Note: if your organism is Ambystoma, the sequence is too large for this task to be performed without submitting it as a job. Hold off for now – we'll go over some workarounds in class.

One item on your deliverables list is to generate the **reverse complement** of a genomic sequence from your organism. First you'll need to extract just one fasta sequence from the whole batch. You can use a seqtk command to extract a sequence. First you need to put the sequence identifier (usually the accession number) into a file. You can do this quickly using echo.

```
echo '<identifier>'  >  <make-up-a-filename>
```

For example, it might look like this:

```
echo 'NC_008524.2' > list.txt
```

This will create a file called `list.txt` that contains just one line: NC_008524.2.

Then use the command that tells seqtk to extract that sequence from the larger file and store it in a new file by itself:

```
seqtk subseq <file-with-many-sequences> list.txt > NC_008524.2.txt
```

Now we can generate the reverse complement of this sequence, putting it into a new file:

```
revseq NC_008524.2.txt NC_008524.2.rev
```

***Remember to keep using tab-complete, and check after commands that the new file is added to your directory by using `ls`. ***