Genome Miniproject Instructions

Reverse Sequencing

1) cd into scratch directory which contains genomic data
```
cd /scratch/ali.may
```
2) load the modules you will need to analyze sequence and generate the reverse sequence. These modules are emboss/6.6.0 and seqtk/1.3 (remember to use tab complete to avoid spelling errors)
```
module load emboss/6.6.0
module load seqtk/1.3
```
3) We are going to reverse sequence only one chromosome so we need to isolate this file from the rest of those downloaded with the entire genome. Use the infoseq command to see what's in the genome file and see the different chromosomes
```
[ali.may@login-01 ali.may]$ infoseq GCF_000002985.6_WBcel235_genomic.fna
Display basic information about sequences
USA                     Database  Name          Accession      Type Length %GC   Organism          Description
fasta::GCF_000002985.6_WBcel235_genomic.fna:NC_003279.8 -        NC_003279.8   -          N    15072434 35.75                 Caenorhabditis elegans chromosome I
fasta::GCF_000002985.6_WBcel235_genomic.fna:NC_003280.10 -       NC_003280.10  -          N    15279421 36.20                 Caenorhabditis elegans chromosome II
fasta::GCF_000002985.6_WBcel235_genomic.fna:NC_003281.10 -       NC_003281.10  -          N    13783801 35.66                 Caenorhabditis elegans chromosome III
fasta::GCF_000002985.6_WBcel235_genomic.fna:NC_003282.8 -        NC_003282.8   -          N    17493829 34.59                 Caenorhabditis elegans chromosome IV
fasta::GCF_000002985.6_WBcel235_genomic.fna:NC_003283.11 -       NC_003283.11  -          N    20924180 35.43                 Caenorhabditis elegans chromosome V
fasta::GCF_000002985.6_WBcel235_genomic.fna:NC_003284.9 -        NC_003284.9   -          N    17718942 35.20                 Caenorhabditis elegans chromosome X
fasta::GCF_000002985.6_WBcel235_genomic.fna:NC_001328.1 -        NC_001328.1   -          N    13794  23.78                 Caenorhabditis elegans mitochondrion, complete genome
```
4) We will use to the echo command to take one chromosome and make it into a new separate file.
```
echo NC_003280.10 > ch_2
```
5) Now we will use the subseq command which will take the contacts of the file we wish to analyze and store it in the new file we created.
```
seqtk subseq GCF_000002985.6_WBcel235_genomic.fna > ch_2 > NC_003280.1
```
6) Next we can reverse sequence the data with the revseq command
```
[[ali.may@login-01 ali.may]$ revseq NC_003280.1 ch_2.txt
 Reverse and complement a nucleotide sequence
```

Amino Acid Sequence

1) To do this we will create a job. This requires writing a script with the following structure shown below. Use Nano to write the script.
```
#!/bin/bash
#SBATCH --partition=short
#SBATCH --job-name=3a0
#SBATCH --time=24:00:00
#SBATCH --nodes=1
#SBATCH --cpus-per-task=2
#SBATCH --mem=256G
#SBATCH --output=%j.output
#SBATCH --error=%j.error
cd /scratch/ali.maya/ncbi_data/data/GCF_000002985.3
grep '>' GCF_000002985.6_WBcel235_genomic.fna
```

2) Then we will run the job by doing sbatch <filename>

```
[ali.may@login-00 GCF_000002985.6]$ sbatch job.bash
```

3) Then we can check the status of the job we submitted by doing squeue -u<username>

```
[ali.may@login-00 GCF_000002985.6]$ squeue -u ali.may
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          40578438     short    Diego  ali.may CG       0:02      1 d0142
```

4) Once the job is complete, you can use the ls command to see two new files in your directory. They should be an output and error file from the job just run. You can open the contents of these files with the cat command.

```
[ali.may@login-00 GCF_000002985.6]$ ls
40576715.error  40576715.output  ch1.rev  GCF_000002985.6_WBcel235_genomic.fna  job.bash  NC_003279.8
```

```
[ali.may@login-00 GCF_000002985.6]$ cat 40578438.output
>NC_003279.8 Caenorhabditis elegans chromosome I
>NC_003280.10 Caenorhabditis elegans chromosome II
>NC_003281.10 Caenorhabditis elegans chromosome III
>NC_003282.8 Caenorhabditis elegans chromosome IV
>NC_003283.11 Caenorhabditis elegans chromosome V
>NC_003284.9 Caenorhabditis elegans chromosome X
>NC_001328.1 Caenorhabditis elegans mitochondrion, complete genome
```

Determine GC content

1) cd into scratch directory which contains genomic data

```
cd /scratch/ali.may
```

2) load emboss

```
[[ali.may@login-00 ali.may]$ module load emboss/6.6.0
```

3) use infoseq command and then the file name with the data you want to analyze, followed by -only and -pgc, as shown below.

```
[ali.may@login-00 ali.may]$ infoseq GCF_000002985.6_WBcel235_genomic.fna -only -pgc
Display basic information about sequences
%GC
35.75
36.20
35.66
34.59
35.43
35.20
23.78
[ali.may@login-00 ali.may]$
```