Genome Analysis
Unit 2 Deliverable A
Working with Sequencing Reads

1. First I obtained the data using the wget command as shown below

```
[ali.may@login-01 Assignments]$ wget https://raw.githubusercontent.com/BayLab/MarineGenomicsData/main/week4_semester.tar.gz
```

2. Next I uncompressed the files using the tar command

```
[ali.may@login-01 Assignments]$ tar -xzvf week4_semester.tar.gz
```

3. Then I checked if the following modules were available on Discovery: Samtools, bowtie2, catadapt, fastqc

```
[ali.may@login-01 U2]$ module avail sam
samos/2023-05-17  samos/2023-07-03  samtools/1.10     samtools/1.18     samtools/1.9
[ali.may@login-01 U2]$ module avail bowtie

------------------------------------------------------------------------------ /shared/centos7/modulefiles ---------------
bowtie/1.3.0   bowtie/2.3.5.1 bowtie/2.5.2
[ali.may@login-01 U2]$ module avail cutadapt
[ali.may@login-01 U2]$ module avail fastqc

------------------------------------------------------------------------------ /shared/centos7/modulefiles ---------------
fastqc/0.11.8 fastqc/0.11.9
[ali.may@login-01 U2]$ ls
```

   -cut adapt was not available and needed to be accessed through shared environment as we will see later

4. Next the gunzip command was used to unzip the files ending in .fastq.gz

```
[ali.may@login-01 U2]$ gunzip SRR6805880.tiny.fastq.gz
```

5. The head command was used to examine the contents of the files

```
[ali.may@login-01 U2]$ head -n50 SRR6805880.tiny.fastq
```

6. The number of sequences for the file was determined using the grep command

```
[ali.may@login-01 U2]$ grep -c '^@' SRR6805880.tiny.fastq
1000
```

- SRR6805880.tiny.fastq: 1000
- SRR6805881.tiny.fastq: 1248
- SRR6805882.tiny.fastq: 1104
- SRR6805883.tiny.fastq: 1134
- SRR6805884.tiny.fastq: 1173
- SRR6805885.tiny.fastq: 1258

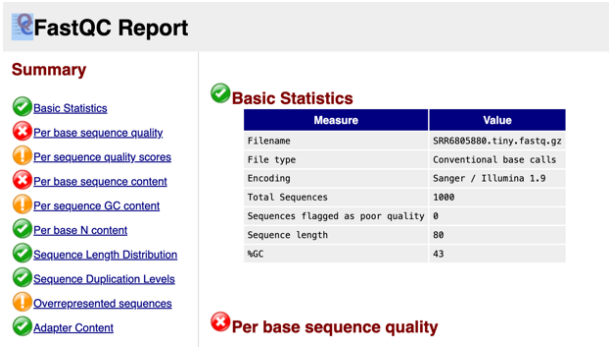
7. Next I ran the quality reports on the files. To do this I loaded Open/JDK and fastqc

```
[ali.may@c0281 Day]$ module load OpenJDK/19.0.1
[ali.may@c0281 Day]$ module load fastqc/0.11.9
```

8. Then I used the following command which generated html files with the content of the report.

```
fastqc *.fastq
```

9. This is a sample of one of the reports



10. Now I began to trim the reads, using the head command I noticed the following adapter sequence: TGCAG

11. To trim I needed to access cut adapt, in order to do that I needed to log onto a computing node using the following command

```
[ali.may@login-01 week4]$ srun --pty /bin/bash
srun: job 40819973 queued and waiting for resources
srun: job 40819973 has been allocated resources
```

12. I loaded the module anaconda which is a necessary precursor to using cutadapt

13. Then finally to access the cutadapt tool I needed to activate from the shared environment.

```
[ali.may@c0325 week4]$ source activate /courses/BIOL3411.202430/shared/cutadapt_env
```

14. Now to actually trim a single read file I used the cutadapt tool

```
(/courses/BIOL3411.202430/shared/cutadapt_env) [ali.may@c0325 week4]$ cutadapt -g TGCAG SRR6805880.tiny.fastq.gz -o SRR6805880.tiny_trimmed.fastq.gz
```

15. To do this to all of the files I wrote a shell script. I also needed to make the script executable using chmod +x

```
  GNU nano 2.3.1                                                File: trim.sh

for filename in *.tiny.fastq.gz
do

  base=$(basename $filename .tiny.fastq.gz)
  echo ${base}

  cutadapt -g TGCAG ${base}.tiny.fastq.gz -o ${base}.tiny_trimmed.fastq.gz

done
```

16. Now I began the process of indexing the genome using the bowtie tool.

```
(/courses/BIOL3411.202430/shared/cutadapt_env) [ali.may@c0325 week4]$ module load bowtie/2.3.5.1
(/courses/BIOL3411.202430/shared/cutadapt_env) [ali.may@c0325 week4]$ bowtie2-build Ppar_tinygenome.fna.gz Ppar_tinygenome
```

17. Using the head tool to examine the contents of the new files revealed the following

```
(/courses/BIOL3411.202430/shared/cutadapt_env) [ali.may@c0325 week4]$ head -20 Ppar_tinygenome.1.bt2
?x[
?????NX?
 5?H&JB ??O%D6o????1??
                    ?S4??4?????I?R6?
                                  ?F?2?\?˜
                                          \????X?i\D?u?????z.`$+?%X?T??
?\H?D??<???p[,??y79?PC??nZ3?W?M*0??␣sy???=?5?~? {?g???k??E?s(?@^?\??        ?a?~c?2?x?,??VA?}|)vZ?XS2j
        ????/?}_??&w?
~W[??                ?#knl????
3 ;Ie?ᴍM??}-{"Y?C*"A??  ???yi;^&?>R????,b???5Ş???N
??d?N^u?]o?
```

18. After indexing the genome I began to map the reads to see where they would align. To do this I wrote another shell script. I again had to make it executable using chmod +x

```
  GNU nano 2.3.1                                                File: map.sh

for filename in *.tiny_trimmed.fastq.gz
do

  base=$(basename $filename .tiny_trimmed.fastq.gz)
  echo ${base}

  bowtie2 -x Ppar_tinygenome -U ${base}.tiny_trimmed.fastq.gz -S ${base}.sam

done
```

19. This generate new “sam” files. Using the head command I inspected the beginning lines of one of the new files

```
[[ali.may@login-01 week4]$ head -20 SRR6805881.sam
@HD     VN:1.0  SO:unsorted
@SQ     SN:KN893585.1   LN:22606
@SQ     SN:KN897506.1   LN:3832
@SQ     SN:JXUT01146130.1       LN:3328
@SQ     SN:KN897010.1   LN:3247
@SQ     SN:KN894258.1   LN:13593
@SQ     SN:KN887772.1   LN:84168
@SQ     SN:KN882209.1   LN:477734
@SQ     SN:JXUT01150820.1       LN:2370
@SQ     SN:JXUT01148685.1       LN:1169
@SQ     SN:KN882212.1   LN:364294
@SQ     SN:KN885770.1   LN:75087
@SQ     SN:KN896765.1   LN:13892
@SQ     SN:KN882215.1   LN:458863
@SQ     SN:KN885329.1   LN:98487
@SQ     SN:KN885697.1   LN:49645
@SQ     SN:KN888763.1   LN:56113
@SQ     SN:JXUT01146289.1       LN:3264
@SQ     SN:KN891677.1   LN:21450
@SQ     SN:KN885380.1   LN:53812
```

20. Next I converted these new sam files to compressed bam files using a shell script. Make sure to have Samtools loaded!

```
  GNU nano 2.3.1                                                File: bam.sh

for filename in *.sam
do

  base=$(basename $filename .sam)
  echo ${base}

  samtools view -bhS ${base}.sam | samtools sort -o ${base}.bam

done
```

21. Converting these files to bam format allows me to now use the ANGSD tool to “call” the genotypes (estimating genotypes). To access this tool I needed to activate it from the shared environment.

```
source activate /courses/BIOL3411.202430/shared/angsd_env/
```

22. Once activated the angsd tool was utilized through the following command

/courses/BIOL3411.202430/shared/angsd_env/angsd/angsd -bam bam.filelist -GL 1 -out genotype_likelihoods -doMaf 2 -SNP_pval 1e-2 -doMajorMinor 1

```
-bam bam.filelist -GL 1 -out genotype_likelihoods -doMaf 2 -SNP_pval 1e-2 -doMajorMinor 1
```

23. This generated two new files

```
genotype_likelihoods.arg
genotype_likelihoods.mafs.gz
```

24. Examining the contents of the new mafs.gz file was done using gunzip and cat

```
(/courses/BIOL3411.202430/shared/cutadapt_env) [ali.may@c0325 week4]$ gunzip genotype_likelihoods.mafs.gz
(/courses/BIOL3411.202430/shared/cutadapt_env) [ali.may@c0325 week4]$ cat *.mafs
chromo  position        major   minor   unknownEM       pu-EM   nInd
KN882277.1      41498   G       T       0.332737        3.127339e-03    3
KN885472.1      10712   C       G       0.126253        1.118604e-03    6
KN885472.1      10741   T       A       0.205533        2.729806e-03    6
KN885472.1      10746   C       T       0.113382        1.394211e-03    6
KN894013.1      22082   T       C       0.098327        3.551274e-03    2
KN894013.1      22084   C       T       0.106562        3.241062e-03    2
KN883616.1      31041   C       A       0.422659        2.070393e-03    3
KN883616.1      31042   T       G       0.424129        1.269827e-03    3
KN883758.1      179190  A       T       0.336645        3.103740e-03    3
```