Genome Analysis

Unit 2 Deliverable A

Working with Sequencing Reads

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



1. Next I uncompressed the files using the tar command
2. Then I checked if the following modules were available on Discovery: Samtools, bowtie2, catadapt, fastqcA screen shot of a black screen

   Description automatically generated

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

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



1. The head command was used to examine the contents of the files
2. The number of sequences for the file was determined using the grep command 

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

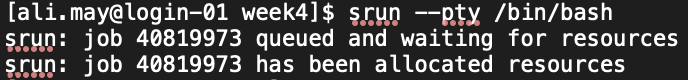
1. Next I ran the quality reports on the files. To do this I loaded Open/JDK and fastqc
2. Then I used the following command which generated html files with the content of the report.



1. This is a sample of one of the reports

A screenshot of a report

Description automatically generated

1. Now I began to trim the reads, using the head command I noticed the following adapter sequence: TGCAG
2. 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
3. I loaded the module anaconda which is a necessary precursor to using cutadapt
4. Then finally to access the cutadapt tool I needed to activate from the shared environment. 
5. Now to actually trim a single read file I used the cutadapt tool



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

A computer screen shot of a black screen

Description automatically generated

1. Now I began the process of indexing the genome using the bowtie tool. 
2. Using the head tool to examine the contents of the new files revealed the followingA computer screen with white text

   Description automatically generated
3. 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 A screen shot of a computer code

   Description automatically generated
4. This generate new “sam” files. Using the head command I inspected the beginning lines of one of the new files

A computer screen shot of a computer

Description automatically generated

1. Next I converted these new sam files to compressed bam files using a shell script. Make sure to have Samtools loaded!A black screen with white text

   Description automatically generated
2. 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/

1. 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



1. This generated two new files

A black and white text

Description automatically generated

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

A screen shot of a computer

Description automatically generated