Assignment\_QAA lab notebook:

Date:9/12/2023

1. Environment: conda create -n fastqcQAA

2. Module loading fastqc (which is already installed in talapas you are just accessing it): module load fastqc/0.11.5 (to look at modules in your environment type ml

   but conda list to look at stuff involved in your conda environment)

3. cd /projects/bgmp/shared/Bi623/QAA\_data\_assignments.txt: Surbhi  2\_2B\_control\_S2\_L008    Undetermined\_S0\_L008

4. Get four reads from this folder: cd /projects/bgmp/shared/2017\_sequencing/demultiplexed/ Read 1 and Read 2

5. fastqc -o /projects/bgmp/surbhin/bioinfo/Bi623/QAA/fastqc\_output/ /projects/bgmp/shared/2017\_sequencing/demultiplexed/Undetermined\_S0\_L008\_R2\_001.fastq.gz

Then went into assignment the first to run the file wiht the code I generated. Python script used q\_scoreavg.py and batch script q\_scoreavg.batch

6. conda create -n QAA1 (QAA1 and not QAA since I installed trimmomatic first and cutadapt didn't like it so I created a new environment - I'm also workin on login node so a bit scared it might crash talapas)

 ----- conda install -c bioconda cutadapt (cutadapt --version 2.6)

 ----- conda install -c bioconda trimmomatic (trimmomatic -version 0.39)

 Documentation for cutadapt https://cutadapt.readthedocs.io/en/stable/

 Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads.

 Documentation for trimmomatic https://usermanual.wiki/Document/TrimmomaticManualV032.1972804677

 Trimmomatic is a fast, multithreaded command line tool that can be used to trim and crop Illumina(FASTQ) data as well as to remove adapters.

 These adapters can pose a real problem depending on the library preparationand downstream application.

7. Format for cutadapt for a pairend read

   For paired-end reads:

    cutadapt -a ADAPT1 -A ADAPT2 [options] -o out1.fastq -p out2.fastq in1.fastq in2.fastq

   For the 2\_2B\_control reads:

   cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

   -o v1\_2\_2B\_control\_R1.fastq.gz -p v1\_2\_2B\_control\_R2.fastq.gz

   /projects/bgmp/shared/2017\_sequencing/demultiplexed/2\_2B\_control\_S2\_L008\_R1\_001.fastq.gz

   /projects/bgmp/shared/2017\_sequencing/demultiplexed/2\_2B\_control\_S2\_L008\_R2\_001.fastq.gz > outputcutadapt

   For the undertermined reads: cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

   -o Undetermined\_S0\_L008\_R1.fastq.gz -p Undetermined\_S0\_L008\_R2.fastq.gz

   /projects/bgmp/shared/2017\_sequencing/demultiplexed/Undetermined\_S0\_L008\_R1\_001.fastq.gz

   /projects/bgmp/shared/2017\_sequencing/demultiplexed/Undetermined\_S0\_L008\_R2\_001.fastq.gz > outputcutadapt\_undertermined

   time taken: [----=8      ] 00:16:52    14,760,166 reads  @     68.6 µs/read;   0.87 M reads/minute

8. Trimmomatic

trimmomatic PE -threads 8 /projects/bgmp/surbhin/bioinfo/Bi623/QAA/v1\_2\_2B\_control\_R1.fastq.gz /projects/bgmp/surbhin/bioinfo/Bi623/QAA/v1\_2\_2B\_control\_R2.fastq.gz paired.V1\_2\_2B\_cont

rol\_R1.fastq.gz unpaired.V1\_2\_2B\_control\_R1.fastq.gz paired.V1\_2\_2B\_control\_R2.fastq.gz unpaired.V1\_2\_2B\_control\_R2.fastq.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:35

Quality encoding detected as phred33

Input Read Pairs: 5830665 Both Surviving: 5652541 (96.95%) Forward Only Surviving: 133562 (2.29%) Reverse Only Surviving: 4595 (0.08%) Dropped: 39967 (0.69%)

TrimmomaticPE: Completed successfully

trimmomatic PE /projects/bgmp/surbhin/bioinfo/Bi623/QAA/Undetermined\_S0\_L008\_R1.fastq.gz /projects/bgmp/surbhin/bioinfo/Bi623/QAA/Undetermined\_S0\_L008\_R2.fastq.gz pairedundertermined\_S0\_L008\_R1.fastq.gz unpaired.Undetermined\_S0\_L008\_R1.fastq.gz paired.Undetermined\_S0\_L008\_R2.fastq.gz unpaired.Undetermined\_S0\_L008\_R2.fastq.gz.fastq.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:5:

15 MINLEN:35

Input Read Pairs: 14760166 Both Surviving: 12160073 (82.38%) Forward Only Surviving: 2511278 (17.01%) Reverse Only Surviving: 31172 (0.21%) Dropped: 57643 (0.39%)

TrimmomaticPE: Completed successfully

numpy                     1.21.5(installed in anaconda)

star                      2.7.10b

conda install htseq

conda install star

conda install matplotlib

conda install -c anaconda numpy

 wget https://ftp.ensembl.org/pub/release-110/fasta/mus\_musculus/dna/Mus\_musculus.GRCm39.dna.primary\_assembly.fa.gz

 wget https://ftp.ensembl.org/pub/release-110/gtf/mus\_musculus/Mus\_musculus.GRCm39.110.gtf.gz

Created two scripts: first for the mouse database alignment and second the mouse alignment against the trimmomatic alignments

sbatch ./starbatchscript.batch (chmod it prior) [memory was throwing error hence requested 32 mbs of memories]

error that I was recieving on my end Detected 1 oom\_kill event in StepId=70315.batch. Some of the step tasks have been OOM Killed

JobID for a) undertermined: jobalign\_70327.err b) 2\_2B: jobalign\_70342.err

sbatch ./star\_database\_mouse.batch (chmod it prio)

./mapped\_vs\_unmapped.py -f alignmentfile2\_2BAligned.out.sam

Mapped reads: 11078806

Unmapped reads: 226276

./mapped\_vs\_unmapped.py -f alignmentfileAligned.out.sam

Mapped reads: 15584495

Unmapped reads: 8735651

htseq-count manual https://htseq.readthedocs.io/en/master/htseqcount.html

Command being timed: "htseq-count -c 2\_2bstranded.tsv --stranded=yes alignmentfile2\_2BAligned.out.sam Mus\_musculus.GRCm39.110.gtf"

        User time (seconds): 430.06

        System time (seconds): 8.92

        Percent of CPU this job got: 98%

        Elapsed (wall clock) time (h:mm:ss or m:ss): 7:27.34

        Average shared text size (kbytes): 0

        Average unshared data size (kbytes): 0

        Average stack size (kbytes): 0

        Average total size (kbytes): 0

        Maximum resident set size (kbytes): 177404

        Average resident set size (kbytes): 0

        Major (requiring I/O) page faults: 0

        Minor (reclaiming a frame) page faults: 898328

        Voluntary context switches: 1879

        Involuntary context switches: 3656

        Swaps: 0

        File system inputs: 0

        File system outputs: 8

        Socket messages sent: 0

        Socket messages received: 0

        Signals delivered: 0

        Page size (bytes): 4096

        Exit status: 0

Command being timed: "htseq-count -c 2\_2breverse.tsv --stranded=reverse alignmentfile2\_2BAligned.out.sam Mus\_musculus.GRCm39.110.gtf"

        User time (seconds): 437.47

        System time (seconds): 3.90

        Percent of CPU this job got: 99%

        Elapsed (wall clock) time (h:mm:ss or m:ss): 7:25.75

        Average shared text size (kbytes): 0

        Average unshared data size (kbytes): 0

        Average stack size (kbytes): 0

        Average total size (kbytes): 0

        Maximum resident set size (kbytes): 177752

        Average resident set size (kbytes): 0

        Major (requiring I/O) page faults: 0

        Minor (reclaiming a frame) page faults: 73285

        Voluntary context switches: 752

        Involuntary context switches: 3720

        Swaps: 0

        File system inputs: 0

        File system outputs: 8

        Socket messages sent: 0

        Socket messages received: 0

        Signals delivered: 0

        Page size (bytes): 4096

        Exit status: 0

Command being timed: "htseq-count -c undertermined\_reverse.tsv --stranded=reverse alignmentfileAligned.out.sam Mus\_musculus.GRCm39.110.gtf"

        User time (seconds): 726.96

        System time (seconds): 6.29

        Percent of CPU this job got: 99%

        Elapsed (wall clock) time (h:mm:ss or m:ss): 12:18.80

        Average shared text size (kbytes): 0

        Average unshared data size (kbytes): 0

        Average stack size (kbytes): 0

        Average total size (kbytes): 0

        Maximum resident set size (kbytes): 177728

        Average resident set size (kbytes): 0

        Major (requiring I/O) page faults: 0

        Minor (reclaiming a frame) page faults: 154134

        Voluntary context switches: 763

        Involuntary context switches: 6184

        Swaps: 0

        File system inputs: 0

        File system outputs: 8

        Socket messages sent: 0

        Socket messages received: 0

        Signals delivered: 0

        Page size (bytes): 4096

        Exit status: 0

Command being timed: "htseq-count -c undertermined\_stranded.tsv --stranded=yes alignmentfileAligned.out.sam Mus\_musculus.GRCm39.110.gtf"

        User time (seconds): 708.01

        System time (seconds): 12.21

        Percent of CPU this job got: 99%

        Elapsed (wall clock) time (h:mm:ss or m:ss): 12:04.11

        Average shared text size (kbytes): 0

        Average unshared data size (kbytes): 0

        Average stack size (kbytes): 0

        Average total size (kbytes): 0

        Maximum resident set size (kbytes): 177100

        Average resident set size (kbytes): 0

        Major (requiring I/O) page faults: 0

        Minor (reclaiming a frame) page faults: 1171495

        Voluntary context switches: 687

        Involuntary context switches: 6065

        Swaps: 0

        File system inputs: 0

        File system outputs: 8

        Socket messages sent: 0

        Socket messages received: 0

        Signals delivered: 0

        Page size (bytes): 4096

        Exit status: 0

tail undertermined\_reverse.tsv

tail undertermined\_stranded.tsv

tail 2\_2bstranded.tsv

tail 2\_2breverse.tsv

QAA1) [surbhin@n0349 QAA]$ tail 2\_2breverse.tsv

ENSMUSG00002076988      0

ENSMUSG00002076989      0

ENSMUSG00002076990      0

ENSMUSG00002076991      0

ENSMUSG00002076992      0

\_\_no\_feature    293950

\_\_ambiguous     89681

\_\_too\_low\_aQual 4760

\_\_not\_aligned   110518

\_\_alignment\_not\_unique  347951

(QAA1) [surbhin@n0349 QAA]$ tail 2\_2bstranded.tsv

ENSMUSG00002076988      0

ENSMUSG00002076989      0

ENSMUSG00002076990      0

ENSMUSG00002076991      0

ENSMUSG00002076992      0

\_\_no\_feature    4960672

\_\_ambiguous     8445

\_\_too\_low\_aQual 4760

\_\_not\_aligned   110518

\_\_alignment\_not\_unique  347951

(QAA1) [surbhin@n0349 QAA]$ tail undertermined\_stranded.tsv

ENSMUSG00002076988      0

ENSMUSG00002076989      0

ENSMUSG00002076990      0

ENSMUSG00002076991      0

ENSMUSG00002076992      0

\_\_no\_feature    7064396

\_\_ambiguous     5958

\_\_too\_low\_aQual 79476

\_\_not\_aligned   4325481

\_\_alignment\_not\_unique  378431

(QAA1) [surbhin@n0349 QAA]$ tail undertermined\_reverse.tsv

ENSMUSG00002076988      0

ENSMUSG00002076989      0

ENSMUSG00002076990      0

ENSMUSG00002076991      0

ENSMUSG00002076992      0

\_\_no\_feature    630007

\_\_ambiguous     129419

\_\_too\_low\_aQual 79476

\_\_not\_aligned   4325481

\_\_alignment\_not\_unique  378431