QAA Assignment

9-6-2023

Make sure to run user bin time on fastqc
Copied github repo “QAA” to Talapas and to my local under Bi623/QAA

The files I will be working with are located in
`/projects/bgmp/shared/2017_sequencing/demultiplexed/`
And are the following:

```
/projects/bgmp/shared/2017_sequencing/demultiplexed/27_4C_mbnl_S19_L008_R1_001.fastq.gz
```

```
27_4C_mbnl_S19_L008_R1_001.fastq.gz
27_4C_mbnl_S19_L008_R2_001.fastq.gz
28_4D_mbnl_S20_L008_R1_001.fastq.gz
28_4D_mbnl_S20_L008_R2_001.fastq.gz
```

```
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/27_4C_mbnl_S19_L008_R1_001.fastq.gz | head
```

Head to make sure my files are as expected

Getting fastqc module loaded:
Identify where on Talapas the module is located
```
module spider fastqc
```
Load module
```
module load fastqc/0.11.5
```

Made a fastqc directory with subdirs of my four runs to hold my fastqc results.

Ran runfastq.srun to generate fastqc files.
slurm-49730.out for information on run times.

Unzipped the fastqc Images folder to retrieve graphs.
Note: Downloaded html viewer extension to view the html file, which is just like the GUI. Nice!

9-7-2023

Scp’d image files for quality score per base and N’s per base from Talapas onto my local folder.

For running my python script qscore_dist_2.py

Checked the read length on w bash command:

```
zcat
/projects/bgmp/shared/2017_sequencing/demultiplexed/27_4C_mbnl_S19_L008_R1_001.fastq.g
z | wc -L
>>>101
```

Read length is 101.
Made an srun for sbatch called mydists.srun
slurm-49717.out for output files.

Made 4 histogram output files that I then scp'd down to my local from Talapas

***

Resources for fastqc:

- FastQC_Manual.pdf
- Quality control: Assessing FASTQC results | Introduction to RNA-Seq using high-performance computing - ARCHIVED
- FastQC/Configuration/adapter_list.txt at master · s-andrews/FastQC

Resources for Illumina:
Illumina adapter sequences

- https://dnatech.genomecenter.ucdavis.edu/wp-content/uploads/2019/03/illumina-adapter-sequences-2019-1000000002694-10.pdf
- https://wikis.utexas.edu/display/CoreNGSTools/Pre-processing+raw+sequences#Preprocessingrawsequences-Adaptertrimmingwithcutadapt

Illumina header format

- https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/FileFormat_FASTQ-files_swBS.htm

Environment
Making the rna_seq environment with python 3.9 (required for cutadapt)

```
conda create -n rna_qc cutadapt python=3.9
```

Then installed trimmomatic from inside the environment with

```
conda install -c bioconda trimmomatic
```

```
trimmomatic -version
>>> 0.39
cutadapt --version
>>> 4.4
```

Cutadapt user guide: https://cutadapt.readthedocs.io/en/stable/guide.html

9-8-2023
Adapters:

Full sequence identified in fastqc:
`GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCGTGGTATCTCGTAT`

Illumina documentation says the following:
R1: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
R2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
https://knowledge.illumina.com/library-preparation/general/library-preparation-general-reference_material-list/000001314

R1 sequence found in R1 and not in R2; and vice versa.

```
zcat
/projects/bgmp/shared/2017_sequencing/demultiplexed/27_4C_mbnl_S19_L008_R1_001.
fastq.gz | grep -c "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA"
>>> 132794

zcat
/projects/bgmp/shared/2017_sequencing/demultiplexed/27_4C_mbnl_S19_L008_R2_001.
fastq.gz | grep -
c "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT"
>>> 135250

zcat
/projects/bgmp/shared/2017_sequencing/demultiplexed/28_4D_mbnl_S20_L008_R1_001.
fastq.gz | grep -
c "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA"
>>> 56758

zcat
/projects/bgmp/shared/2017_sequencing/demultiplexed/28_4D_mbnl_S20_L008_R2_001.
fastq.gz | grep -
c "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT"
>>> 57626
```

For paired-end reads:
cutadapt -a ADAPT1 -A ADAPT2 [options] -o out1.fastq -p out2.fastq in1.fastq in2.fastq

Run cutadapt (4.4 with Python 3.9.18) (with an sbatch, results in dir “cutadapt”), this is the script in it (but for both samples):

```
cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o
28_4D_mbnl_S20_L008_R1_001.fastq.gz_trimmed.fastq.gz -p
28_4D_mbnl_S20_L008_R2_001_trimmed.fastq.gz
```

```
/projects/bgmp/shared/2017_sequencing/demultiplexed/28_4D_mbnl_S20_L0
08_R1_001.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/28_4D_mbnl_S20_L0
08_R2_001.fastq.gz
```

Cutadapt summary:

Sample 27

Finished in 195.373 s (27.036 µs/read; 2.22 M reads/minute).

=== Summary ===

Total read pairs processed: 7,226,430
Read 1 with adapter: 751,117 (10.4%)
Read 2 with adapter: 803,568 (11.1%)
Pairs written (passing filters): 7,226,430 (100.0%)

Total basepairs processed: 1,459,738,860 bp
Read 1: 729,869,430 bp
Read 2: 729,869,430 bp
Total written (filtered): 1,429,426,877 bp (97.9%)
Read 1: 714,826,948 bp
Read 2: 714,599,929 bp

=== First read: Adapter 1 ===

Sequence: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA; Type: regular 3';
Length: 33; Trimmed: 751117 times

=== Second read: Adapter 2 ===

Sequence: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT; Type: regular 3';
Length: 33; Trimmed: 803568 times

Sample 28

=== Summary ===

Total read pairs processed: 12,428,766
Read 1 with adapter: 743,440 (6.0%)
Read 2 with adapter: 841,389 (6.8%)
Pairs written (passing filters): 12,428,766 (100.0%)

Total basepairs processed: 2,510,610,732 bp
Read 1: 1,255,305,366 bp

Read 2: 1,255,305,366 bp
Total written (filtered): 2,489,647,234 bp (99.2%)
Read 1: 1,245,001,943 bp
Read 2: 1,244,645,291 bp

=== First read: Adapter 1 ===

Sequence: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA; Type: regular 3';
Length: 33; Trimmed: 743440 times

=== Second read: Adapter 2 ===

Sequence: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT; Type: regular 3';
Length: 33; Trimmed: 841389 times

Exit code 0 for both.

Now trimming with trimmomatic
(See manual at http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf)

Usage for paired-end:

trimmomatic PE [-threads <threads] [-phred33 | -phred64] [-trimlog <logFile>] >] [-basein <inputBase> | <input 1> <input 2>] [-baseout <outputBase> | <unpaired output 1> <paired output 2> <unpaired output 2> <step 1> ...

- LEADING: quality of 3
- TRAILING: quality of 3
- SLIDING WINDOW: window size of 5 and required quality of 15
- MINLENGTH: 35 bases

```
trimmomatic PE -phred33 $inputFile1 $inputFile2 $outputFile1P
$outputFile1U $outputFile2P $outputFile2U LEADING:3 TRAILING:3
SLIDINGWINDOW:5:15 MINLEN:35
```

Sample 27:

```
/usr/bin/time -v trimmomatic PE -phred33
/projects/bgmp/shared/2017_sequencing/demultiplexed/27_4C_mbnl_
S19_L008_R1_001.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/27_4C_mbnl_
S19_L008_R2_001.fastq.gz
```

```
/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/27R1_P.fastq.gz
/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/27R1_U.fastq.gz
/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/27R2_P.fastq.gz
/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/27R2_U.fastq.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:35
```

Sample 27:

Input Read Pairs: 7226430 Both Surviving: 6891402 (95.36%) Forward Only Surviving: 326732 (4.52%) Reverse Only Surviving: 5500 (0.08%) Dropped: 2796 (0.04%)

Sample 28:

```
/usr/bin/time -v trimmomatic PE -phred33
/projects/bgmp/shared/2017_sequencing/demultiplexed/28_4D_mbnl_S20_L008_R1_001.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/28_4D_mbnl_S20_L008_R2_001.fastq.gz
/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/28R1_P.fastq.gz
/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/28R1_U.fastq.gz
/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/28R2_P.fastq.gz
/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/28R2_U.fastq.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:35
```

Sample 28:

Input Read Pairs: 12428766 Both Surviving: 11736976 (94.43%) Forward Only Surviving: 677966 (5.45%) Reverse Only Surviving: 8735 (0.07%) Dropped: 5089 (0.04%)

Distributions of read lengths after trimming using command like this for each paired file:

```
zcat 27R1_P.fastq.gz | grep -A 1 "^@" --no-group-separator | grep -v "@" | awk '{print length
 ($0)}' | sort -n | uniq -c > 27R1_P_dist.txt
```

Scp'd distribution files from Talapas to my computer
27R1_P_dist.txt, 27R2_P_dist.txt, 28R1_P_dist.txt, 28R2_P_dist.txt

/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/trim_dists

9-10-2023

Graphs in R for frequencies of length distributions for both reads in both samples

9-11-2023

Make QAA environment

- star
- numpy
- matplotlib
- htseq

```
conda create --name qaa

conda activate qaa

conda install star -c bioconda

conda install numpy -c conda-forge

conda install -c conda-forge matplotlib

conda install -c bioconda htseq
```

Htseq is raising issues for required matplotlib version being not available and for python not being the right version. Idk how to fix.

warning libmamba Problem type not implemented SOLVER_RULE_STRICT_REPO_PRIORITY failed

```
LibMambaUnsatisfiableError: Encountered problems while solving:
  - nothing provides matplotlib 1.2.1 needed by htseq-0.6.1-np17py27_0
Could not solve for environment specs
The following packages are incompatible
├─ htseq is installable with the potential options
│  ├─ htseq [0.11.0|0.11.1|...|0.9.1] would require
│  │  └─ python [2.7* |>=2.7,<2.8.0a0 ], which can be installed;
│  ├─ htseq [0.11.0|0.11.1|0.11.2|0.7.2|0.9.1] would require
│  │  └─ python [3.5* |>=3.5,<3.6.0a0 ], which can be installed;
│  ├─ htseq [0.11.0|0.11.1|...|0.9.1] would require
│  │  └─ python >=3.6,<3.7.0a0 , which can be installed;
│  ├─ htseq [0.11.2|0.11.3] would require
│  │  └─ python >=3.7,<3.8.0a0 , which can be installed;
│  ├─ htseq [0.12.3|0.12.4|0.13.5|0.7.2|0.9.1] would require
│  │  └─ python_abi 3.6.* *_cp36m, which can be installed;
│  ├─ htseq [0.12.3|0.12.4|...|2.0.3] would require
│  │  └─ python_abi 3.7.* *_cp37m, which can be installed;
│  ├─ htseq [0.12.4|0.13.5|...|2.0.3] would require
│  │  └─ python_abi 3.8.* *_cp38, which can be installed;
│  ├─ htseq [0.13.5|1.99.2|2.0.1|2.0.2|2.0.3] would require
│  │  └─ python_abi 3.9.* *_cp39, which can be installed;
│  ├─ htseq 0.6.1 would require
│  │  └─ matplotlib 1.2.1 , which does not exist (perhaps a missing channel);
│  ├─ htseq [0.6.1|0.6.1.post1|0.7.2|0.9.1] would require
│  │  └─ python_abi 2.7.* *_cp27mu, which can be installed;
│  ├─ htseq 0.7.2 would require
```

```
|   |   └─ python 3.4* , which can be installed;
|   ├─ htseq [0.7.2|0.9.1] would require
|   |   └─ python 3.6* , which can be installed;
|   └─ htseq [2.0.2|2.0.3] would require
|       └─ python_abi 3.10.* *_cp310, which can be installed;
└─ python 3.11**  is not installable because there are no viable options
    ├─ python [3.11.0|3.11.1|...|3.11.5] would require
    |   └─ python_abi 3.11.* *_cp311, which conflicts with any installable versions
previously reported;
    └─ python [3.11.0|3.11.2|3.11.3|3.11.4|3.11.5] conflicts with any installable
versions previously reported.
```

Solution: I just have to install htseq first since it uses python 3.7

Made new environment called htseq_only with only htseq installed.
Made another new environment called qaa_2 with all installed.

```
conda create --name qaa_2 htseq -c bioconda

conda activate qaa_2

conda install star -c bioconda

conda install numpy -c conda-forge

conda install -c conda-forge matplotlib
```

Versions:

htseq	2.0.3
star	2.7.10b
matplotlib	3.5.3
numpy	1.21.6

Pull mouse genome fasta and gtf files from ensembl (release 110)
Fasta:
https://ftp.ensembl.org/pub/release-110/fasta/mus_musculus/
Mus_musculus.GRCm39.dna.primary_assembly.fa.gz
GTF:
https://ftp.ensembl.org/pub/release-110/gtf/mus_musculus/ Mus_musculus.GRCm39.110.gtf.gz

Saved in dir “QAA/ensembl_files” on Talapas

Talapas is running slow so use interactive node and use reservation bgmp-res.

#SBATCH --reservation=bgmp-res
#SBATCH --partition=interactive

Aligning
https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/lecture_notes/STARmanual.pdf

1. Made a directory to contain STAR database: Mus_musculus.GRCm39.dna.ens110.STAR_ 2.7.10b
2. Build STAR database using runstar.srun using unzipped files (slurm 55904)

```
STAR --runThreadN 8 \
--runMode genomeGenerate \
--genomeDir
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/Mus_musculus.GRCm39.dna.ens1
10.STAR_2.7.10b \
--genomeFastaFiles
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/ensembl_files/Mus_musculus.G
RCm39.dna.primary_assembly.fa \
--sjdbGTFfile
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/ensembl_files/Mus_musculus.G
RCm39.110.gtf \
```

3. Run STAR to to align the reads to the reference genome using staralign.srun (slurm 56061, 56066) for sample 27 and sample 28

```
STAR --runThreadN 8 --runMode alignReads \
--outFilterMultimapNmax 3 \
--outSAMunmapped Within KeepPairs \
--alignIntronMax 1000000 --alignMatesGapMax 1000000 \
--readFilesCommand zcat \
--readFilesIn
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/27R1_P.fastq.gz
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/27R2_P.fastq.gz \
--genomeDir
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/Mus_musculus.GRCm39.dna.ens1
10.STAR_2.7.10b \
--outFileNamePrefix alignsamp27_Mus_musculus.GRCm39.dna.ens110.STAR_2.7.10b

STAR --runThreadN 8 --runMode alignReads \
--outFilterMultimapNmax 3 \
--outSAMunmapped Within KeepPairs \
--alignIntronMax 1000000 --alignMatesGapMax 1000000 \
--readFilesCommand zcat \
--readFilesIn
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/28R1_P.fastq.gz
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/trimmomatic/28R2_P.fastq.gz \
--genomeDir
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/Mus_musculus.GRCm39.dna.ens1
10.STAR_2.7.10b \
--outFileNamePrefix alignsamp27_Mus_musculus.GRCm39.dna.ens110.STAR_2.7.10b
```

4. Mapped and unmapped reads using samparse_2.py
    a. Sample 27
        i. Mapped: 12720381, Notmapped: 1062423
    b. Sample 28
        i. Mapped: 22250477, Notmapped: 1223475

HTSeq count

https://htseq.readthedocs.io/en/release_0.11.1/count.html

Usage: htseq-count [options] <alignment_files> <gff_file>

In htseq_only environment.

Stranded Sample 27

```
/usr/bin/time -v htseq-count --stranded=yes \
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/alignsamp27_Mus_musculus.GRCm39.dna
.ens110.STAR_2.7.10bAligned.out.sam \
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/ensembl_files/Mus_musculus.GRCm39.1
10.gtf \
-o 27_out.sam
```

Stranded Sample 28

```
/usr/bin/time -v htseq-count --stranded=yes \
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/alignsamp28_Mus_musculus.GRCm39.dna
.ens110.STAR_2.7.10bAligned.out.sam \
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/ensembl_files/Mus_musculus.GRCm39.1
10.gtf \
-o 28_out.sam
```

Reverse Sample 27

```
/usr/bin/time -v htseq-count --stranded=reverse \
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/alignsamp27_Mus_musculus.GRCm39.dna
.ens110.STAR_2.7.10bAligned.out.sam \
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/ensembl_files/Mus_musculus.GRCm39.1
10.gtf \
-o 27_out_rev.sam
```

Reverse Sample 28

```
/usr/bin/time -v htseq-count --stranded=reverse \
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/alignsamp28_Mus_musculus.GRCm39.dna
.ens110.STAR_2.7.10bAligned.out.sam \
/gpfs/projects/bgmp/rubenl/bioinfo/Bi623/QAA/align/ensembl_files/Mus_musculus.GRCm39.1
10.gtf \
-o 28_out_rev.sam
```

Out recorded: slurm-56591.out

sum of all genes that mapped to a feature:

```
awk '$1 ~ "ENSMUS" {sum+=$2} END {print sum}'  <genecountfile>
```

Sum of all reads:

```
awk '{sum+=$2} END {print sum}' <genecountfile>
```

Report percent reads and summary with

```
awk '{total+=$2} $1 ~ "ENSMUS" {mapread+=$2} END {print "For file: " FILENAME "
Total reads: " total", " "Number of reads mapped: " mapread", " "Percent of
reads mapped: " mapread/total*100"%"}' <genecount file>
```

Sample 27 stranded:

For file: 27out.tsv Total reads: 6891402, Number of reads mapped: 252094, Percent of reads mapped: 3.65809%
Mapped: 252094
Total: 6891402
__no_feature 5809524
__ambiguous 5415
__too_low_aQual 11921
__not_aligned 524865
__alignment_not_unique 287583

Sample 27 reverse:

For file: 27revout.tsv Total reads: 6891402, Number of reads mapped: 5453279, Percent of reads mapped: 79.1316%
Mapped: 5453279
Total: 6891402
__no_feature 505626
__ambiguous 108128
__too_low_aQual 11921
__not_aligned 524865
__alignment_not_unique 287583

Sample 28 stranded:

For file: 28out.tsv Total reads: 11736976, Number of reads mapped: 417513, Percent of reads mapped: 3.55725%
Mapped: 417513
Total: 11736976
__no_feature 10164741
__ambiguous 8432
__too_low_aQual 22879
__not_aligned 599390
__alignment_not_unique 524021

Sample 28 reverse:

For file: 28revout.tsv Total reads: 11736976, Number of reads mapped: 9537088, Percent of reads mapped: 81.2568%
Mapped: 9537088
Total: 11736976
__no_feature 866263
__ambiguous 187335
__too_low_aQual 22879
__not_aligned 599390
__alignment_not_unique 524021

It is a stranded library because it's not a 50/50 split of mapped reads between forward and reverse reads. Only about 3% of reads map on the forward reads, and about 80% of reads mapped on the reverse reads. This means it likely was a stranded library.
Confirmed by looking at the methods; libraries were prepared with the KAPA Stranded mRNA-Seq kit.

Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap