9/2/25:

Data

I was assigned two RNA-seq files from the PRJNA1005245 and PRJNA1005244 projects on NCBI.

SRR25630303 = Campylomormyrus rhynchophorus, young (6cm)

SRR25630398 = Campylomormyrus rhynchophorus, adult

Downloading data and creating environment:

To download both files from NCBI I ran these commands:

```
prefetch SRR25630303
prefetch SRR25630398

fasterq-dump --split-files SRR25630303/
fasterq-dump --split-files SRR25630398/
```

To set up the environment for this project, I ran these commands:

```
conda create --name QAA
conda activate QAA

conda install bioconda::fastqc
conda install bioconda::cutadapt
conda install bioconda::trimmomatic
```

Analysis:

I first ran the fastqs through fastqc and generated the html files for each, alsong with many plots

```
fastqc SRR25630398_1.fastq SRR25630398_2.fastq
fastqc SRR25630303_1.fastq SRR25630303_2.fastq
```

Command being timed: "fastqc SRR25630398_1.fastq SRR25630398_2.fastq"
User time (seconds): 295.88
System time (seconds): 12.38
Percent of CPU this job got: 98%
Elapsed (wall clock) time (h:mm:ss or m:ss): 5:13.30

Then I ran the files through my own histogram script from demultiplexing: Demultiplexing

Note: I also edited the script slightly to use argparse and changed the read length to 150

Command being timed: "./nucl_mean_dist.py -i SRR25630303_1.fastq -o SRR25630303_1_hist.png"
User time (seconds): 925.22
System time (seconds): 10.13
Percent of CPU this job got: 99%
Elapsed (wall clock) time (h:mm:ss or m:ss): 15:37.75

9/3/25:

Adapter Trimming:

```
# To trim adapters:
/usr/bin/time -v cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT  -o trimmed_SRR25630398_1.fastq -p
trimmed_SRR25630398_2.fastq SRR25630398_1.fastq SRR25630398_2.fastq
```

SRR25630398

Command being timed: "cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o trimmed_SRR25630398_1.fastq -p trimmed_SRR25630398_2.fastq SRR25630398_1.fastq SRR25630398_2.fastq"
User time (seconds): 185.54
System time (seconds): 6.71
Percent of CPU this job got: 98%
Elapsed (wall clock) time (h:mm:ss or m:ss): 3:14.23

Total read pairs processed: 34,878,180
Read 1 with adapter: 4,844,719 (13.9%)
Read 2 with adapter: 4,978,371 (14.3%)
Pairs written (passing filters): 34,878,180 (100.0%)

Total basepairs processed: 10,463,454,000 bp
Read 1: 5,231,727,000 bp
Read 2: 5,231,727,000 bp
Total written (filtered): 10,264,089,024 bp (98.1%)
Read 1: 5,128,839,415 bp
Read 2: 5,135,249,609 bp

SRR25630303

Command being timed: "cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o trimmed_SRR25630303_1.fastq -p trimmed_SRR25630303_2.fastq SRR25630303_1.fastq SRR25630303_2.fastq"
User time (seconds): 179.38
System time (seconds): 10.08
Percent of CPU this job got: 98%
Elapsed (wall clock) time (h:mm:ss or m:ss): 3:12.05

Total read pairs processed: 41,934,422
Read 1 with adapter: 1,767,108 (4.2%)
Read 2 with adapter: 2,066,337 (4.9%)
Pairs written (passing filters): 41,934,422 (100.0%)

Total basepairs processed: 12,580,326,600 bp
Read 1: 6,290,163,300 bp
Read 2: 6,290,163,300 bp
Total written (filtered): 12,536,648,265 bp (99.7%)
Read 1: 6,268,302,186 bp
Read 2: 6,268,346,079 bp

Sanity check:

```
awk 'NR % 4 == 2' SRR25630303_1.fastq | grep -c
'AGATCGGAAGAGCACACGTCTGAACTCCAGTCA'
129695

wk 'NR % 4 == 2' SRR25630303_1.fastq | grep -c
'TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT'
0

# Good! this makes sense that the r1 would have lots of the r1 adapter and
very little of the reverse compliment.
# This indicates it is the correct direction for this file
```

```
# This was repeated for each of the fastq files and has similar results
```

9/4/25

Ran trimmomatic commands:

```
/usr/bin/time -v trimmomatic PE -threads 8 trimmed_SRR25630303_1.fastq
trimmed_SRR25630303_2.fastq \
  qualtrimmed_paired_SRR25630303_1.fastq.gz
qualtrimmed_unpaired_SRR25630303_1.fastq.gz \
  qualtrimmed_paired_SRR25630303_2.fastq.gz
qualtrimmed_unpaired_SRR25630303_2.fastq.gz \
  LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:35

Timing:
User time (seconds): 220.29
System time (seconds): 33.53
Percent of CPU this job got: 276%
Elapsed (wall clock) time (h:mm:ss or m:ss): 1:31.95

/usr/bin/time -v trimmomatic PE -threads 8 trimmed_SRR25630398_1.fastq
trimmed_SRR25630398_2.fastq \
  qualtrimmed_paired_SRR25630398_1.fastq.gz
qualtrimmed_unpaired_SRR25630398_1.fastq.gz \
  qualtrimmed_paired_SRR25630398_2.fastq.gz
qualtrimmed_unpaired_SRR25630398_2.fastq.gz \
  LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:35

Timing:
User time (seconds): 192.80
System time (seconds): 23.15
Percent of CPU this job got: 305%
Elapsed (wall clock) time (h:mm:ss or m:ss): 1:10.76
```

Began writing histogram script:
found at /projects/bgmp/epea/bioinfo/Bi623/PS/QAA/make_hist_r1_r2.py

Had many small annoying errors.
Note: want to log transform y axis, but numpy didn't work. Instead I used plt.yscale('log')

Alignments:

1. Made database:

```
STAR --runMode genomeGenerate --genomeDir ./ --genomeFastaFiles
./campylomormyrus.fasta --sjdbGTFfile campylomormyrus.gtf
```

2. Aligned both files to the db:

```
/usr/bin/time -v STAR --runThreadN 8 --runMode alignReads \
 --outFilterMultimapNmax 3 \
 --outSAMunmapped Within KeepPairs \
 --alignIntronMax 1000000 --alignMatesGapMax 1000000 \
 --readFilesCommand zcat \
 --readFilesIn
/projects/bgmp/epea/bioinfo/Bi623/PS/QAA/qualtrimmed_paired_SRR25630303_1.fast
q.gz
/projects/bgmp/epea/bioinfo/Bi623/PS/QAA/qualtrimmed_paired_SRR25630303_2.fast
q.gz \
 --genomeDir /projects/bgmp/epea/bioinfo/Bi623/PS/QAA/align \
 --outFileNamePrefix aligned_SRR25630303_

/usr/bin/time -v STAR --runThreadN 8 --runMode alignReads \
 --outFilterMultimapNmax 3 \
 --outSAMunmapped Within KeepPairs \
 --alignIntronMax 1000000 --alignMatesGapMax 1000000 \
 --readFilesCommand zcat \
 --readFilesIn
/projects/bgmp/epea/bioinfo/Bi623/PS/QAA/qualtrimmed_paired_SRR25630398_1.fast
q.gz
/projects/bgmp/epea/bioinfo/Bi623/PS/QAA/qualtrimmed_paired_SRR25630398_2.fast
q.gz \
 --genomeDir /projects/bgmp/epea/bioinfo/Bi623/PS/QAA/align \
 --outFileNamePrefix aligned_SRR25630398_
```

Command being timed: "STAR --runThreadN 8 --runMode alignReads --outFilterMultimapNmax 3 --outSAMunmapped Within KeepPairs --alignIntronMax 1000000 --alignMatesGapMax 1000000 --readFilesCommand zcat --readFilesIn /projects/bgmp/epea/bioinfo/Bi623/PS/QAA/qualtrimmed*paired_SRR25630398_1.fastq.gz /projects/bgmp/epea/bioinfo/Bi623/PS/QAA/qualtrimmed_paired_SRR25630398_2.fastq.gz --genomeDir /projects/bgmp/epea/bioinfo/Bi623/PS/QAA/align --outFileNamePrefix aligned_SRR25630398*"

User time (seconds): 4829.68
System time (seconds): 17.55
Percent of CPU this job got: 755%
Elapsed (wall clock) time (h:mm:ss or m:ss): 10:41.66

3. Converted to bams and sorted:

```
samtools view -b aligned_SRR25630303_Aligned.out.sam >
aligned_SRR25630303_Aligned.out.bam
samtools sort aligned_SRR25630303_Aligned.out.bam >
aligned_SRR25630303_Aligned.sorted.out.bam

samtools view -b aligned_SRR25630398_Aligned.out.sam >
aligned_SRR25630398_Aligned.out.bam
samtools sort aligned_SRR25630398_Aligned.out.bam >
aligned_SRR25630398_Aligned.sorted.out.bam
```

Command being timed: "samtools view -b aligned_SRR25630303_Aligned.out.sam"
User time (seconds): 403.35
System time (seconds): 3.39
Percent of CPU this job got: 99%
Elapsed (wall clock) time (h:mm:ss or m:ss): 6:48.57

Command being timed: "samtools sort aligned_SRR25630303_Aligned.out.bam"
User time (seconds): 532.87
System time (seconds): 2.30
Percent of CPU this job got: 99%
Elapsed (wall clock) time (h:mm:ss or m:ss): 9:00.44

4. Ran Picard:

```
picard MarkDuplicates INPUT=aligned_SRR25630303_Aligned.sorted.out.bam
OUTPUT=dedup_aligned_SRR25630303_Aligned.sorted.out.bam
METRICS_FILE=SRR25630303_dup.metrics REMOVE_DUPLICATES=TRUE
VALIDATION_STRINGENCY=LENIENT

samtools view -h dedup_aligned_SRR25630303_Aligned.sorted.out.bam >
dedup_aligned_SRR25630303_Aligned.sorted.out.sam

picard MarkDuplicates INPUT=aligned_SRR25630398_Aligned.sorted.out.bam
OUTPUT=dedup_aligned_SRR25630398_Aligned.sorted.out.bam
METRICS_FILE=SRR25630398_dup.metrics REMOVE_DUPLICATES=TRUE
VALIDATION_STRINGENCY=LENIENT
```

```
samtools view -h dedup_aligned_SRR25630398_Aligned.sorted.out.bam >
dedup_aligned_SRR25630398_Aligned.sorted.out.sam
```

Picard didn't work, not sure why. Its late, so I'll pick this up tomorrow.

To do:

1. rerun converting alignment to bams for SRR25630398, job died because srun ended (actually it was fine)
2. figure out picard error (fixed)
3. Continue from there

9/5/25:

Needed to change the version of picard to 2.18

Command being timed: "picard MarkDuplicates
INPUT=aligned_SRR25630303_Aligned.sorted.out.bam
OUTPUT=dedup_aligned_SRR25630303_Aligned.sorted.out.bam
METRICS_FILE=SRR25630303_dup.metrics REMOVE_DUPLICATES=TRUE
VALIDATION_STRINGENCY=LENIENT"
User time (seconds): 589.73
System time (seconds): 8.46
Percent of CPU this job got: 119%
Elapsed (wall clock) time (h:mm:ss or m:ss): 8:22.42

Command being timed: "picard MarkDuplicates
INPUT=aligned_SRR25630398_Aligned.sorted.out.bam
OUTPUT=dedup_aligned_SRR25630398_Aligned.sorted.out.bam
METRICS_FILE=SRR25630398_dup.metrics REMOVE_DUPLICATES=TRUE
VALIDATION_STRINGENCY=LENIENT"
User time (seconds): 578.89
System time (seconds): 10.04
Percent of CPU this job got: 116%
Elapsed (wall clock) time (h:mm:ss or m:ss): 8:25.62

Converted bams to dedup sams:

/usr/bin/time -v samtools view -h dedup_aligned_SRR25630303_Aligned.sorted.out.bam >
dedup_aligned_SRR25630303_Aligned.sorted.out.sam

Command being timed: "samtools view -h
dedup_aligned_SRR25630303_Aligned.sorted.out.bam"
User time (seconds): 11.20
System time (seconds): 0.82
Percent of CPU this job got: 94%
Elapsed (wall clock) time (h:mm:ss or m:ss): 0:12.68

NOTE FOR FUTURE SELF: consider making one long submit script for each project so that all jobs are in one file in order. Just comment out the ones you've already used. There are so many submit scripts in the ps2 folder

Counting Mapped and unmapped reads:

Running /projects/bgmp/epea/bioinfo/Bi623/PS/QAA/mapped_unmapped_count.py with both samples

File proccessed: dedup_aligned_SRR25630303_Aligned.sorted.out.sam
Number of reads mapped: 18484340
Number of reads unmapped: 3651623
Command being timed: "python mapped_unmapped_count.py -i
dedup_aligned_SRR25630303_Aligned.sorted.out.sam"
User time (seconds): 30.83
System time (seconds): 3.81
Percent of CPU this job got: 97%
Elapsed (wall clock) time (h:mm:ss or m:ss): 0:35.36

File proccessed: dedup_aligned_SRR25630398_Aligned.sorted.out.sam
Number of reads mapped: 32532195
Number of reads unmapped: 4777469
Command being timed: "python mapped_unmapped_count.py -i
dedup_aligned_SRR25630398_Aligned.sorted.out.sam"
User time (seconds): 50.96
System time (seconds): 5.89
Percent of CPU this job got: 99%
Elapsed (wall clock) time (h:mm:ss or m:ss): 0:57.07

Running htseq-count:

You should run htseq-count twice: once with `--stranded=yes` and again with `--stranded=reverse`. Use default parameters otherwise. You may need to use the `-i` parameter for this run.

```
/usr/bin/time -v htseq-count --stranded=yes -i gene_id
dedup_aligned_SRR25630398_Aligned.sorted.out.sam ./align/campylomormyrus.gtf >
counts_stranded_SRR25630398

/usr/bin/time -v htseq-count --stranded=reverse -i gene_id
dedup_aligned_SRR25630398_Aligned.sorted.out.sam ./align/campylomormyrus.gtf >
counts_reverse_SRR25630398

awk '$1 !~ /^__/ {sum += $2} END {print sum}' counts_stranded_SRR25630398
awk '$1 !~ /^__/ {sum += $2} END {print sum}' counts_reverse_SRR25630398

# Done again for SRR25630303
```

SRR25630398:

Stranded: 899068
Reverse: 19458807

SRR25630303:

Stranded: 486475
Reverse: 11196054