QAA

Vandana Reddy

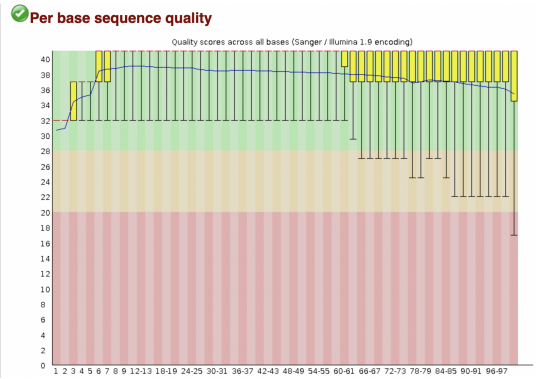
9/3/2021

QUESTION 1

Undetermined_S0_L008_R1_001.fastq.gz

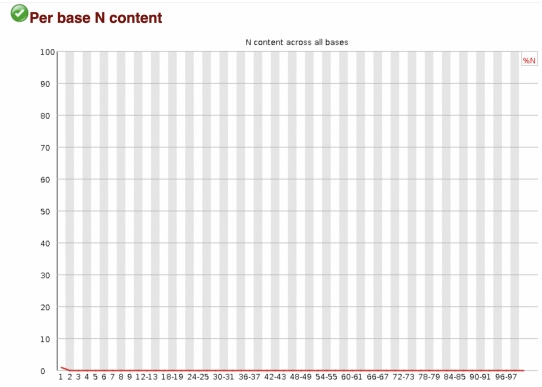
Quality Score Distribution for Undetermined Read 1

```
knitr::include_graphics("/Users/vandanareddy/Bioinformatics/Bi623/QAA/und_r1-1.png")
```



Per base N Distribution for Undetermined Read 1

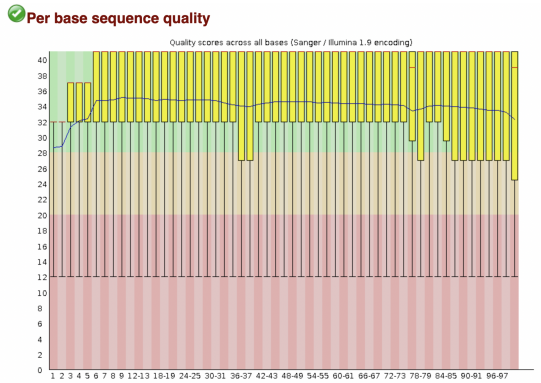
```
knitr::include_graphics("/Users/vandanareddy/Bioinformatics/Bi623/QAA/und_r1-2.png")
```



Undetermined_S0_L008_R2_001.fastq.gz

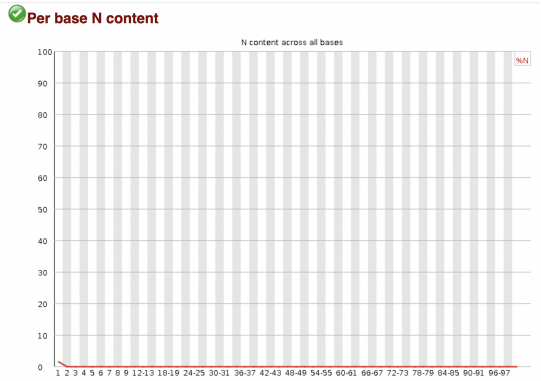
Quality Score Distribution for Undetermined Read 2

```
knitr::include_graphics("/Users/vandanareddy/Bioinformatics/Bi623/QAA/und_r2-1.png")
```



Per base N Distribution for Undetermined Read 2

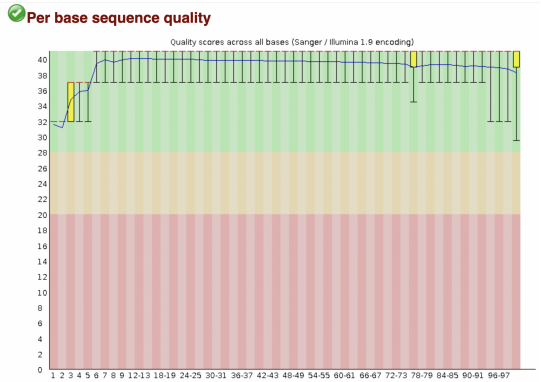
```
knitr::include_graphics("/Users/vandanareddy/Bioinformatics/Bi623/QAA/und_r2-2.png")
```



7_2E_fox_S6_L008_R1_001.fastq.gz

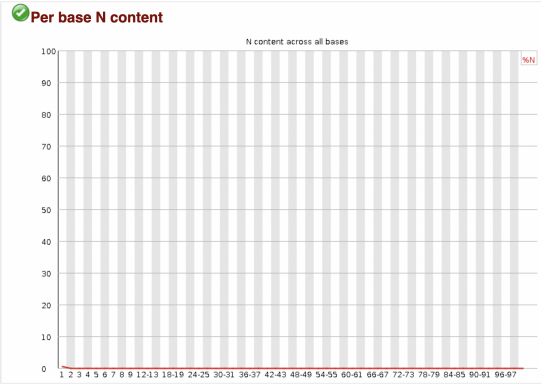
Quality Score Distribution for 7_2E Read 1

```
knitr::include_graphics("/Users/vandanareddy/Bioinformatics/Bi623/QAA/72e_r1-1.png")
```



Per base N Distribution for 7_2E Read 1

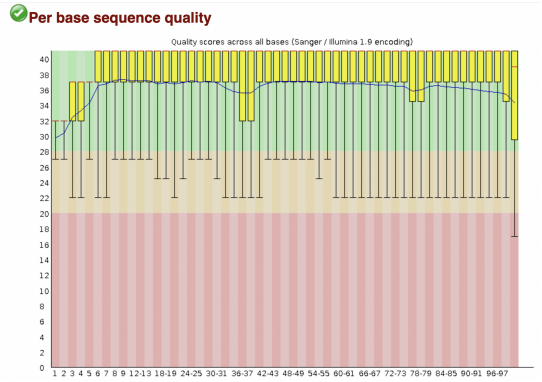
```
knitr::include_graphics("/Users/vandanareddy/Bioinformatics/Bi623/QAA/72e_r1-2.png")
```



7_2E_fox_S6_L008_R2_001.fastq.gz

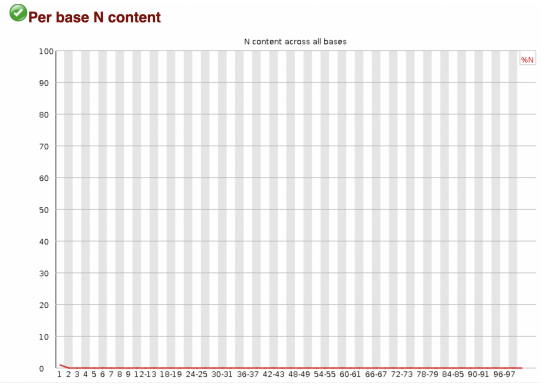
Quality Score Distribution for 7_2E Read 2

```
knitr::include_graphics("/Users/vandanareddy/Bioinformatics/Bi623/QAA/72e_r2-1.png")
```



Per base N Distribution for 7_2E Read 2

```
knitr::include_graphics("/Users/vandanareddy/Bioinformatics/Bi623/QAA/72e_r2-2.png")
```

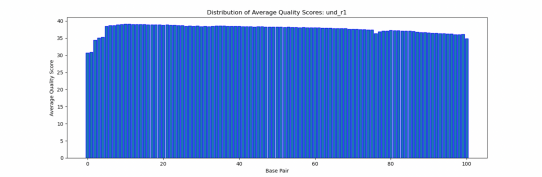


All of the per N distributions show a very low percentage of N's in all 4 files. Read 1 of the undetermined file has quality scores with a dip at the end. Read 2 has lower quality scores but the program still gave it a green check mark indicating that the scores are good. Read 1 of the 7_2E file has really good quality scores throughout the reads. Read 2 has lower quality but it is still high enough for the green check mark.

QUESTION 2

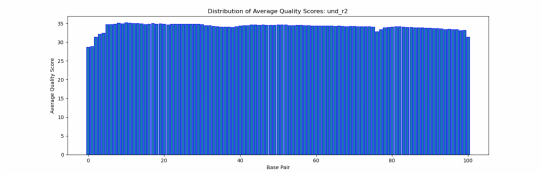
Quality Score Distribution for Undetermined Read 1

```
knitr::include_graphics("/Users/vandanareddy/Bioinformatics/Bi623/QAA/undr1.png")
```



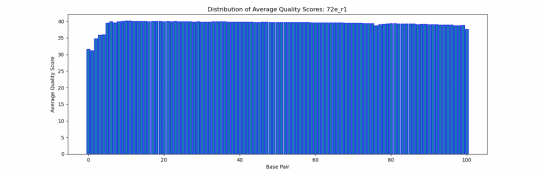
Quality Score Distribution for Undetermined Read 2

```
knitr::include_graphics("/Users/vandanareddy/Bioinformatics/Bi623/QAA/undr2.png")
```



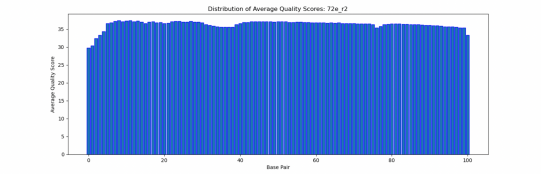
Quality Score Distribution for 7_2E Read 1

```
knitr::include_graphics("/Users/vandanareddy/Bioinformatics/Bi623/QAA/72r1.png")
```



Quality Score Distribution for 7_2E Read 2

```
knitr::include_graphics("/Users/vandanareddy/Bioinformatics/Bi623/QAA/72r2.png")
```



All of distribution plots I made look very similar to the FastQC plots. The read 1 distributions has higher quality scores than the read 2 distributions. Additionally, the end of the reads have lower quality scores than the rest of the read. Using FastQC, the undetermined files look 1:30 each and the 7_2E files took :30 each. Using my plotting program, undetermined files look 10:30 each and the 7_2E files took 4:00 each. FastQC has multitreading capability but my program does not which may be why the FastQC program runs so much faster.

QUESTION 3

Overall, the 7_2E file has better data quality than the Undetermined file because the quality scores are higher. The read 1 files had better quality scores than the read 2 files.

QUESTION 4

```
conda create --name QAA
conda install python=3.9.5
conda install -c bioconda cutadapt
conda install -c bioconda trimmomatic

cutadapt --version (3.4)
trimmomatic -version (0.39)
```

QUESTION 5

```
#!/bin/bash
#SBATCH --partition=bgmp                ### Partition
#SBATCH --job-name=cutadapt             ### Job Name
#SBATCH --output=cutadapt.out_%j        ### File in which to store job output
#SBATCH --time=0-02:00:00               ### Wall clock time limit in Days-HH:MM:SS
#SBATCH --nodes=1                       ### Number of nodes needed for the job
#SBATCH --ntasks-per-node=8             ### Number of tasks to be launched per Node
#SBATCH --account=bgmp                  ### Account used for job submission

/usr/bin/time -v cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
-j 0 -o Undetermined_R1.fastq. -p Undetermined_R2.fastq /projects/bgmp/shared/2017_sequencing/demul
tiplexed/Undetermined_S0_L008_R1_001.fastq.gz /projects/bgmp/shared/2017_sequencing/demultiplexed/U
ndetermined_S0_L008_R2_001.fastq.gz

/usr/bin/time -v cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
-j 0 -o 7_2E_fox_R1.fastq -p 7_2E_fox_R2.fastq /projects/bgmp/shared/2017_sequencing/demultiplexed/
7_2E_fox_S6_L008_R1_001.fastq.gz /projects/bgmp/shared/2017_sequencing/demultiplexed/7_2E_fox_S6_L0
08_R2_001.fastq.gz
```

> Cutadapt trimmed 14,760,166 read pairs from the undetermined files. 543,021 read 1s were adapter trimmed and 607,660 read 2s were adapter trimmed. Cutadapt trimmed 5,278,425 read pairs from the 7_2E files. 173,473 read 1s were adapter trimmed and 212,512 read 2s were adapter trimmed.

```
SANITY CHECK:
forward adapters
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/Undetermined_S0_L008_R1_001.fastq.gz | gre
p 'AGATCGGAAGAGCACACGTCTGAACTCCAGTCA'
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/Undetermined_S0_L008_R2_001.fastq.gz | gre
p 'AGATCGGAAGAGCACACGTCTGAACTCCAGTCA' --> nothing

zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/7_2E_fox_S6_L008_R1_001.fastq.gz | grep 'A
GATCGGAAGAGCACACGTCTGAACTCCAGTCA'
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/7_2E_fox_S6_L008_R2_001.fastq.gz | grep 'A
GATCGGAAGAGCACACGTCTGAACTCCAGTCA' --> nothing

reverse adapters
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/Undetermined_S0_L008_R1_001.fastq.gz | gre
p 'AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT' --> nothing
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/Undetermined_S0_L008_R2_001.fastq.gz | gre
p 'AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT'

zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/7_2E_fox_S6_L008_R1_001.fastq.gz | grep 'A
GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT' --> nothing
zcat /projects/bgmp/shared/2017_sequencing/demultiplexed/7_2E_fox_S6_L008_R2_001.fastq.gz | grep 'A
GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT'
```

> I looked for the forward primer in the read 1 files and in the read 2 files. The adapter did not appear in the read 2 files but on every line of the read 1 files. The reverse primer was the same way; it only appeared on every line of the read 2 files and not in the read 1 files.

QUESTION 6

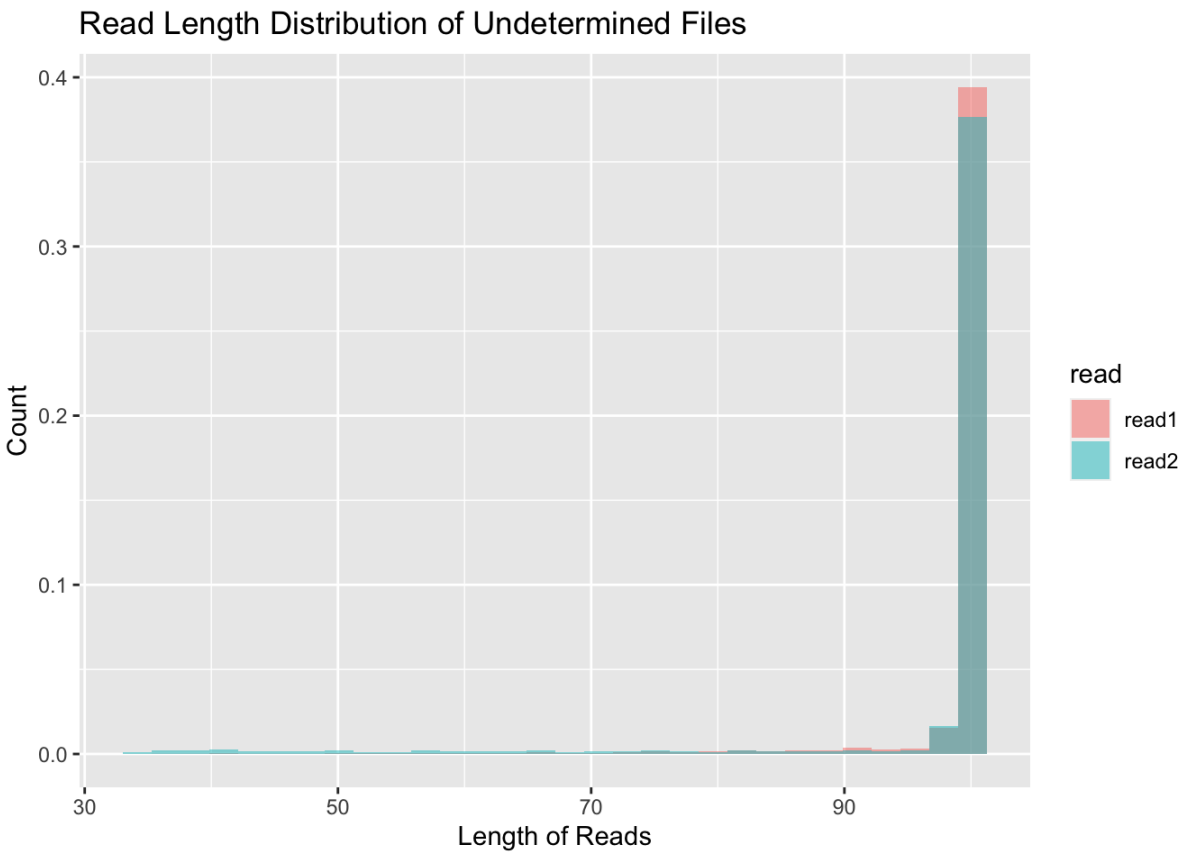
```
#!/bin/bash
#SBATCH --partition=bgmp        ### Partition (like a queue in PBS)
#SBATCH --job-name=trimmomatic      ### Job Name
#SBATCH --output=Trimmomatic.out_%j       ### File in which to store job output
#SBATCH --time=0-04:00:00       ### Wall clock time limit in Days-HH:MM:SS
#SBATCH --nodes=1               ### Number of nodes needed for the job
#SBATCH --ntasks-per-node=4     ### Number of tasks to be launched per Node
#SBATCH --account=bgmp          ### Account used for job submission

/usr/bin/time -v trimmomatic PE -threads 4 Undetermined_R1.fastq Undetermined_R2.fastq Undetermined
_R1.trimmed.fastq.gz Undetermined_R1un.trimmed.fastq.gz Undetermined_R2.trimmed.fastq.gz Undetermin
ed_R2un.trimmed.fastq.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:35

/usr/bin/time -v trimmomatic PE -threads 4 7_2E_fox_R1.fastq 7_2E_fox_R2.fastq 7_2E_fox_R1.trimmed.
fastq.gz 7_2E_fox_R1un.trimmed.fastq.gz 7_2E_fox_R2.trimmed.fastq.gz 7_2E_fox_R2un.trimmed.fastq.gz
LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:35
```

QUESTION 7

```
und_r1 = read.table(file = "/Users/vandanareddy/Bioinformatics/Bi623/QAA/und_r1",
    sep = "\n", header = FALSE)
colnames(und_r1) = c("length")
und_r2 = read.table(file = "/Users/vandanareddy/Bioinformatics/Bi623/QAA/und_r2",
    sep = "\n", header = FALSE)
colnames(und_r2) = c("length")

other_r1 = read.table(file = "/Users/vandanareddy/Bioinformatics/Bi623/QAA/7_2e_r1",
    sep = "\n", header = FALSE)
colnames(other_r1) = c("length")
other_r2 = read.table(file = "/Users/vandanareddy/Bioinformatics/Bi623/QAA/7_2e_r2",
    sep = "\n", header = FALSE)
colnames(other_r2) = c("length")
```

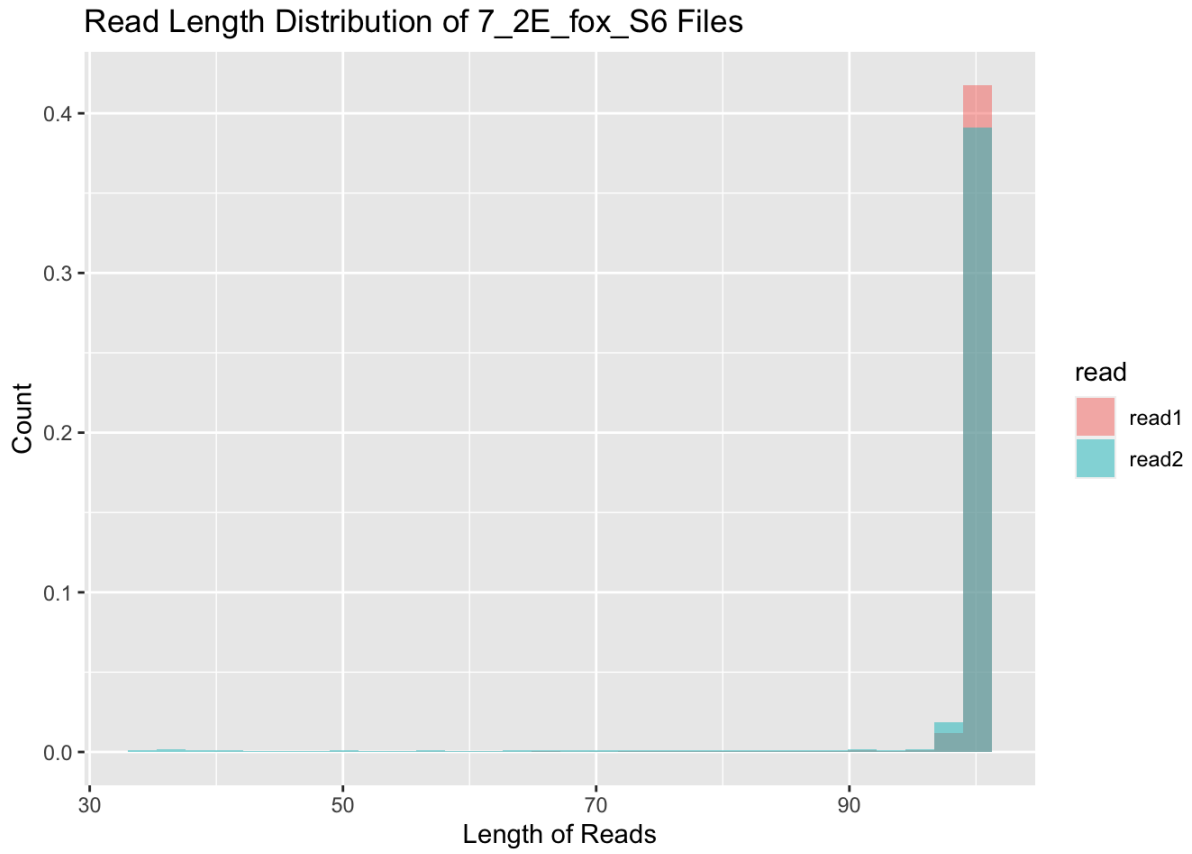
```
library(ggplot2)
und_r1$read <- "read1"
und_r2$read <- "read2"
other_r1$read <- "read1"
other_r2$read <- "read2"

und = rbind(und_r1, und_r2)
other = rbind(other_r1, other_r2)

und_plot = ggplot(und, aes(length, fill = read)) + geom_histogram(alpha = 0.5, aes(y = ..densit
y..),
    position = "identity") + ggtitle("Read Length Distribution of Undetermined Files") +
    xlab("Length of Reads") + ylab("Count")
und_plot
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
other_plot = ggplot(other, aes(length, fill = read)) + geom_histogram(alpha = 0.5,
    aes(y = ..density..), position = "identity") + ggtitle("Read Length Distribution of 7_2E_fox_S6
Files") +
    xlab("Length of Reads") + ylab("Count")
other_plot
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In both files, the read 2 files' reads are shorter than than the read 1 files' reads. Read 2 is blue on the graph and the read length is lower than that of the read 1 files.

QUESTION 8

```
conda install -c bioconda star
conda install -c anaconda numpy
conda install -c bioconda pysam
conda install -c conda-forge matplotlib
pip install HTSeq
```

QUESTION 9

create the mouse alignment database

```
#!/bin/bash
#SBATCH --partition=bgmp        ### Partition (like a queue in PBS)
#SBATCH --job-name=star      ### Job Name
#SBATCH --output=star.out_%j       ### File in which to store job output
#SBATCH --time=0-04:00:00       ### Wall clock time limit in Days-HH:MM:SS
#SBATCH --nodes=1               ### Number of nodes needed for the job
#SBATCH --ntasks-per-node=8    ### Number of tasks to be launched per Node
#SBATCH --account=bgmp          ### Account used for job submission

/usr/bin/time -v STAR --runMode genomeGenerate --runThreadN 8 --genomeDir /projects/bgmp/vandanar/B
ioinformatics/Bi623/QAA/Mus_musculus.GRCm39.104.STAR_2.7.9a --genomeFastaFiles /projects/bgmp/vanda
nar/Bioinformatics/Bi623/QAA/Mus_musculus.GRCm39.dna.primary_assembly.fa --sjdbGTFfile /projects/bg
mp/vandanar/Bioinformatics/Bi623/QAA/Mus_musculus.GRCm39.104.gtf
```

Align the undertermined filed to the alignment database

```
#!/bin/bash
#SBATCH --partition=bgmp        ### Partition (like a queue in PBS)
#SBATCH --job-name=star      ### Job Name
#SBATCH --output=star_un.out_%j       ### File in which to store job output
#SBATCH --time=0-04:00:00       ### Wall clock time limit in Days-HH:MM:SS
#SBATCH --nodes=1               ### Number of nodes needed for the job
#SBATCH --ntasks-per-node=8    ### Number of tasks to be launched per Node
#SBATCH --account=bgmp          ### Account used for job submission

/usr/bin/time -v STAR --runThreadN 8 \
--runMode alignReads \
--outFilterMultimapNmax 3 \
--outSAMunmapped Within KeepPairs \
--alignIntronMax 1000000 \
--alignMatesGapMax 1000000 \
--readFilesCommand zcat \
--readFilesIn /projects/bgmp/vandanar/Bioinformatics/Bi623/QAA/Undetermined_R1.trimmed.fastq.gz /pr
ojects/bgmp/vandanar/Bioinformatics/Bi623/QAA/Undetermined_R2.trimmed.fastq.gz \
--genomeDir /projects/bgmp/vandanar/Bioinformatics/Bi623/QAA/Mus_musculus.GRCm39.104.STAR_2.7.9a \
```

Align the 7_2E filed to the alignment database

```
#!/bin/bash
#SBATCH --partition=bgmp        ### Partition (like a queue in PBS)
#SBATCH --job-name=star      ### Job Name
#SBATCH --output=star72.out_%j       ### File in which to store job output
#SBATCH --time=0-04:00:00       ### Wall clock time limit in Days-HH:MM:SS
#SBATCH --nodes=1               ### Number of nodes needed for the job
#SBATCH --ntasks-per-node=8    ### Number of tasks to be launched per Node
#SBATCH --account=bgmp          ### Account used for job submission


/usr/bin/time -v STAR --runThreadN 8 \
--runMode alignReads \
--outFilterMultimapNmax 3 \
--outSAMunmapped Within KeepPairs \
--alignIntronMax 1000000 \
--alignMatesGapMax 1000000 \
--readFilesCommand zcat \
--readFilesIn /projects/bgmp/vandanar/Bioinformatics/Bi623/QAA/7_2E_fox_R1.trimmed.fastq.gz /projec
ts/bgmp/vandanar/Bioinformatics/Bi623/QAA/7_2E_fox_R2.trimmed.fastq.gz \
--genomeDir /projects/bgmp/vandanar/Bioinformatics/Bi623/QAA/Mus_musculus.GRCm39.104.STAR_2.7.9a \
```

QUESTION 10

> The undetermined SAM file had 15584518 mapped reads and 8735628 unmapped reads. The 7_2E SAM had 9424763 mapped reads and 340643 unmapped reads.

QUESTION 11

```
#!/bin/bash

#SBATCH --partition=bgmp        ### Partition (like a queue in PBS)
#SBATCH --job-name=htseq      ### Job Name
#SBATCH --output=htseq.out_%j       ### File in which to store job output
#SBATCH --time=0-04:00:00       ### Wall clock time limit in Days-HH:MM:SS
#SBATCH --nodes=1               ### Number of nodes needed for the job
#SBATCH --ntasks-per-node=1     ### Number of tasks to be launched per Node
#SBATCH --account=bgmp          ### Account used for job submission

/usr/bin/time -v htseq-count -f sam --stranded=yes -r name sorted_Aligned.72e.out.sam Mus_musculus.
GRCm39.104.gtf > 72e_stranded.txt
/usr/bin/time -v htseq-count -f sam --stranded=no -r name --samout=72e_notstranded sorted_Aligned.7
2e.out.sam Mus_musculus.GRCm39.104.gtf > 72e_notstranded.txt

/usr/bin/time -v htseq-count -f sam --stranded=yes -r name --samout=und_stranded sorted_Aligned.un
d.out.sam Mus_musculus.GRCm39.104.gtf > und_stranded.txt
/usr/bin/time -v htseq-count -f sam --stranded=no -r name --samout=und_notstranded sorted_Aligned.u
nd.out.sam Mus_musculus.GRCm39.104.gtf > und_notstranded.txt
```

> The ht-seq files have been uploaded to github. They were 55,000 lines long.

QUESTION 12

I believe that this data came from a unstranded library prep. For the 7_2E file, the number of mapped and unmapped reads were the same for both the stranded and unstranded options. However, the stranded option returned 4316124 reads not mapped to a feature, and the unstranded option returned 371746 reads not mapped to a feature. For the undetermined file, the stranded option returned 7071214 reads not mapped to a feature and 583519 reads not mapped to a feature with the unstranded option. In a stranded kit, you would expect that all of the reads map to the genome in the same direction. With an unstranded kit, you would expect it to map in both directions, thus increasing the numbers of reads mapped to a feature. Because the number of genes mapping to a feature is so much higher with the unstranded option compared to the stranded option, I would conclude that the kit was unstranded.