Isabella Fregoso
Proff. Van Laar
BIOL 4810-001
2 May 2025

NOTES: Transcriptome Analysis M. leprae

activate environment
```
Conda activate transcriptomics
```

navigate to external hard drive and folder
```
sudo mount -t drvfs D: /mnt/d
cd /mnt/d

cd Final
```

zip the files
```
gunzip fastq/*.gz
```

create index files
```
hisat2-build mleprae.fna mlepr
```

create sam files for each pair
```
hisat2 -q -x mlepr -U fastq/old_01.fastq -S old_01.sam

hisat2 -q -x mlepr -U fastq/old_02.fastq -S old_02.sam

hisat2 -q -x mlepr -U fastq/young_01.fastq -S young_01.sam

hisat2 -q -x mlepr -U fastq/young_02.fastq -S young_02.sam
```

create bam files for each
```
samtools view -bS old_01.sam > old_01.bam

samtools view -bS old_02.sam > old_02.bam

samtools view -bS young_01.sam > young_01.bam

samtools view -bS young_02.sam > young_02.bam
```

sort to organize in correct chronological order of chromosome

```
samtools sort old_01.bam -o old_01.sorted.bam

samtools sort old_02.bam -o old_02.sorted.bam

samtools sort young_01.bam -o young_01.sorted.bam

samtools sort young_02.bam -o young_02.sorted.bam
```

assemble into transcripts
```
stringtie old_01.sorted.bam -G mleprae.gff -o
stringtie/old_01.transcripts.gtf

stringtie old_02.sorted.bam -G mleprae.gff -o
stringtie/old_02.transcripts.gtf

stringtie young_01.sorted.bam -G mleprae.gff -o
stringtie/young_01.transcripts.gtf

stringtie young_02.sorted.bam -G mleprae.gff -o
stringtie/young_02.transcripts.gtf
```

create .txt file with all the pathways to the assembled files (figure out the next 3)
```
pico assemblies.txt

stringtie/old_01.transcripts.gtf

stringtie/old_02.transcripts.gtf

stringtie/young_01.transcripts.gtf

stringtie/young_02.transcripts.gtf
```

merge all transcripts to do comparisons
```
stringtie --merge -G mleprae.gff -o stringtie_merged.gtf
assemblies.txt
```

count the transcripts
```
cat stringtie_merged.gtf | grep -v "^#" | awk
'$3=="transcript" {print}' | wc -l
```

Q. How many transcripts are there?
There are 3120 transcripts

comparing the annotated transcripts to the known transcripts
```
gffcompare -r mleprae.gff -G -o merged stringtie_merged.gtf
```

3051 reference transcriptions loaded
3120 query transfrags loaded

determine the quality
```
cat merged.stats
```

```
#= Summary for dataset: stringtie_merged.gtf
#     Query mRNAs :    3120 in    2742 loci  (15 multi-exon transcripts
)
#           (10 multi-transcript loci, ~1.1 transcripts per locus)
# Reference mRNAs :    3051 in    2722 loci  (0 multi-exon)
# Super-loci w/ reference transcripts:     2688
#-----------------| Sensitivity | Precision |
        Base level:    100.0     |    95.4   |
        Exon level:     86.8     |    84.6   |
   Transcript level:    88.8     |    86.8   |
        Locus level:    97.7     |    95.9   |

    Matching intron chains:       0
       Matching transcripts:    2708
             Matching loci:    2659

       Missed exons:       0/3051    (  0.0%)
        Novel exons:      57/3133    (  1.8%)
       Novel introns:     15/15      (100.0%)
        Missed loci:       0/2722    (  0.0%)
         Novel loci:      54/2742    (  2.0%)

 Total union super-loci across all input datasets: 2742
3120 out of 3120 consensus transcripts written in merged.annotated.gtf
(0 discarded as redundant)
(transcriptomics) isabellaelisa@Isabellas-Laptop:/mnt/d/Final$
```


prepare files to use in R
```
stringtie -e -B -p 8 -G stringtie_merged.gtf -o
ballgown/old_01/old_01.gtf old_01.sorted.bam
```


```
stringtie -e -B -p 8 -G stringtie_merged.gtf -o
ballgown/old_02/old_02.gtf old_02.sorted.bam
```


```
stringtie -e -B -p 8 -G stringtie_merged.gtf -o
ballgown/young_01/young_01.gtf young_01.sorted.bam
```

```
stringtie -e -B -p 8 -G stringtie_merged.gtf -o
ballgown/young_02/young_02.gtf young_02.sorted.bam
```

open file "TranscriptomeDemo.rmd" in R by clicking Session<Set working directory< Choose directory

Complete the R template

this is how I pushed my code at the end of the github upload

- Do this after `git add .`

```
git branch -M main
git push -u origin main
```