Markdown file created by Joel Swift for RNAseq homework for (SLU)BCB5200

1. Download the required fastq file and run fastQC

Download

```
fastq-dump SRR2567795
fastq-dump SRR2567786
```

FastQC

```
fastqc SRR2567795.fastq
unzip SRR2567795_fastqc.zip
cd SRR2567795_fastqc/
less fastqc_data.txt
```

A. Results

```
>>Per base sequence quality     pass
#Base   Mean    Median  Lower Quartile  Upper Quartile  10th Percentile 90th Percentile
1       32.61830937634032       34.0    31.0    34.0    31.0    34.0
2       32.76358148901108       34.0    31.0    34.0    31.0    34.0
3       32.85438385669965       34.0    31.0    34.0    31.0    34.0
4       36.29484318341469       37.0    37.0    37.0    35.0    37.0
5       36.129828934292924      37.0    35.0    37.0    35.0    37.0
6       36.07655709672344       37.0    35.0    37.0    35.0    37.0
7       36.076784371043544      37.0    35.0    37.0    35.0    37.0
8       36.11765685798143       37.0    36.0    37.0    35.0    37.0
9       37.87177184400032       39.0    38.0    39.0    35.0    39.0
10-14   38.18013047902665       39.4    38.2    39.4    35.2    39.4
15-19   39.21341152617363       40.6    39.0    41.0    36.0    41.0
20-24   39.00362805598665       40.0    38.4    41.0    35.4    41.0
25-29   38.97166848077938       40.0    38.4    41.0    35.0    41.0
30-34   38.90321370201526       40.0    38.0    41.0    35.0    41.0
35-39   38.65456082414552       40.0    38.0    41.0    34.2    41.0
40-44   38.33118902207272       40.0    38.0    41.0    33.6    41.0
45-49   38.12031297160288       40.0    37.0    41.0    33.0    41.0
50-54   37.65930857005945       39.4    36.4    41.0    32.6    41.0
55-59   36.886755329066794      38.4    35.2    40.4    31.6    41.0
60-64   35.97182879735841       36.8    35.0    39.6    31.0    41.0
65-69   34.947880045060536      35.4    34.0    38.2    30.0    40.4
70-74   33.94976444260529       35.0    33.6    36.6    29.2    39.0
75-79   32.9069587607627        34.8    32.4    35.2    28.4    37.2
80-84   32.58564200836563       35.0    33.0    35.0    27.8    36.0
```

```
85-89   31.850809973792213      35.0    32.8    35.0    26.4    35.4
90-94   32.8899968569756        34.8    32.4    34.8    30.2    34.8
95-99   36.44330745671564       37.4    37.2    37.4    35.0    37.4
100-104 38.09498247461609       39.4    38.2    39.4    35.2    39.4
105-109 39.17699467873108       41.0    39.0    41.0    36.0    41.0
110-114 39.07978590158321       41.0    39.0    41.0    35.8    41.0
115-119 38.85918676149805       40.2    38.4    41.0    35.0    41.0
120-124 38.59160258739991       40.0    38.0    41.0    34.2    41.0
125-129 38.31340480267376       40.0    38.0    41.0    33.6    41.0
130-134 38.08754019948232       40.0    38.0    41.0    33.0    41.0
135-139 38.28364462059552       40.0    38.0    41.0    33.2    41.0
140-144 37.69917080685618       39.6    37.2    40.6    32.8    41.0
145-149 37.59765279769605       39.6    36.8    41.0    32.6    41.0
150-154 36.793472321091556      38.6    35.6    40.6    31.8    41.0
155-159 35.750308545491336      36.8    35.0    39.2    31.0    41.0
160-164 34.66280970361996       35.4    34.4    37.6    30.0    39.4
165-169 33.732835624906954      35.0    34.0    36.2    29.6    38.0
170-174 33.040653545752804      35.0    34.0    35.2    29.0    36.6
175-179 32.46505940362325       35.0    34.0    35.0    29.0    36.0
180     32.08374923917784       35.0    33.0    35.0    27.0    35.0
```

B. What does the "Yellow Box" in the "Per base seqeunce quality" represent?

```
The yellow box represents the inter-quartile range (25-75%).
```

2. Retreiving the Reference Genome and annotation file.

```
cp /public/ahnt/courses/bcb5200/HW7/Schizosaccharomyces_pombe_all_chromosomes.fa ./
cp /public/ahnt/courses/bcb5200/HW7/schizosaccharomyces_pombe.genome.gff3 ./
```

A. Provide the citation for the genome assembly.

```
- From .gff3 file I extracted #!genome-build-accession GCA_000002945.2
- Searched GCA_000002945.2 on NCBI
- Clicked on related information tab Assembly
- Clicked on related information tab PubMed

Wood, V., et al. "The genome sequence of Schizosaccharomyces pombe." Nature 415.6874
(2002): 871-880
```

3. Use Tophat2 to map files to reference, guided by the gff3 file.

```
#Correcting the fasta genome file

sed -i 's/chromosome_1/I/g' spombe.fa
sed -i 's/chromosome_2/II/g' spombe.fa
sed -i 's/chromosome_3/III/g' spombe.fa

#Remove ab and mating type sequences from fasta genome file
vi spombe.fa

#Build the index
bowtie2-build spombe.fa spombe

#Align
tophat2 -p 8 -o SRR2567786/ -G spombe.gff spombe ../data/SRR2567786_1.fastq \
../data/SRR2567786_2.fastq
tophat2 -p 8 -o SRR2567795/ -G spombe.gff spombe ../data/SRR2567795_1.fastq \
../data/SRR2567795_2.fastq
```

4. What are the overall mapping rates for SRR2567786 & SRR2567795?

```
SRR2567786

Left reads:
          Input     :  13112183
           Mapped   :  12801875 (97.6% of input)
            of these:    701758 ( 5.5%) have multiple alignments (127 have >20)
Right reads:
          Input     :  13112183
           Mapped   :  12284963 (93.7% of input)
            of these:    658511 ( 5.4%) have multiple alignments (128 have >20)
95.7% overall read mapping rate.

Aligned pairs:  12138221
     of these:    650677 ( 5.4%) have multiple alignments
                    3060 ( 0.0%) are discordant alignments
92.5% concordant pair alignment rate

SRR2567795

Left reads:
          Input     :  12984309
           Mapped   :  12714799 (97.9% of input)
```

```
            of these:    951056 ( 7.5%) have multiple alignments (183 have >20)
Right reads:
          Input     :  12984309
           Mapped   :  12634825 (97.3% of input)
            of these:    941243 ( 7.4%) have multiple alignments (206 have >20)
97.6% overall read mapping rate.

Aligned pairs:  12452673
     of these:    926951 ( 7.4%) have multiple alignments
                    4725 ( 0.0%) are discordant alignments
95.9% concordant pair alignment rate.
```

5. Use Cuflinks to assemble transcriptomes from the two alignment files, report the number of genes and transcripts assemblied for each RNA library.

```
#SRR7725686
mv accepted_hits.bam SRR7725686.bam
mkdir cufflinks
cufflinks -o SRR2567786/cufflinks/ -p 8 -u -g spombe.gff -b spombe.fa \
SRR2567786/SRR7725686.bam

#SRR7725695
mv accepted_hits.bam SRR7725695.bam
mkdir cufflinks
cufflinks -o SRR2567795/cufflinks/ -p 8 -u -g spombe.gff -b spombe.fa \
SRR2567795/SRR7725695.bam
```

```
#Number of genes

###SRR7725686
head -n -1 genes.fpkm_tracking | wc -l
#6160

###SRR7725695
head -n -1 genes.fpkm_tracking | wc -l
#5897

#Number of Transcripts

###SRR7725686
cut -f3 transcripts.gtf | grep "transcript" -c
#7437

###SRR7725695
```

```
cut -f3 transcripts.gtf | grep "transcript" -c
#7298
```

6. Use Cuffmerge to merge the transcripts generated by cufflinks for both files

```
#make assembly file that contains the path to both transcript.gtf files

echo SRR2567786/cufflinks/transcripts.gtf > assemblies.txt
echo SRR2567795/cufflinks/transcripts.gtf >> assemblies.txt

#Cuffmerge
cuffmerge -s spombe.fa assemblies.txt


#Find number of transcripts produced after cuffmerge
wc -l merged_asm/merged.gtf
#11359
```

7. Use Cuffcompare to compare the merged annotation file to the reference annotation. Report the stats output and the number of novel transcripts with novel isoforms (include a description).

```
mkdir cuffcompare
cuffcompare -r spombe.gff merged_asm/merged.gtf


# Summary for dataset: merged_asm/merged.gtf :
#     Query mRNAs :    5756 in    5622 loci  (2524 multi-exon transcripts)
#            (110 multi-transcript loci, ~1.0 transcripts per locus)
# Reference mRNAs :    6886 in    6197 loci  (2531 multi-exon)
# Super-loci w/ reference transcripts:     5362
#--------------------|   Sn   |  Sp   |  fSn |  fSp
#        Base level:    94.9  97.8    -     -
#        Exon level:    78.3  85.4  78.4  85.6
#      Intron level:    98.3  95.4  98.5  95.6
# Intron chain level:    84.2  84.5 100.0 100.0
#  Transcript level:    62.2  74.5  62.5  74.8
#       Locus level:    75.8  83.5  80.8  87.1
#
#     Matching intron chains:    2132
#              Matching loci:    4700
#
#          Missed exons:     447/12220 (  3.7%)
#           Novel exons:     251/11192 (  2.2%)
#        Missed introns:      92/5333  (  1.7%)
#         Novel introns:     172/5490  (  3.1%)
#           Missed loci:     323/6197  (  5.2%)
#            Novel loci:     166/5622  (  3.0%)
#
# Total union super-loci across all input datasets: 5622

#Finding all novel transcripts
cut -f4 cuffcmp.tracking | grep -c "j"
#285

#I was able to find the number of novel transcipts by first searching
#the Class codes that are assigned to each record in the cuffcmp.tracking
#file, these are located in column 4. Per the documentation j represents
#"Potentially novel isoform (fragment): at least one splice junction is shared with a
# reference transcript". So a cut and grep command has the ability to count these.
```

8. Run htseq-count to convert mapped results to counts.

```
#Sort the bam files by name
samtools sort -n SRR7725695.bam -o SRR7725695_sorted.bam
samtools sort -n SRR7725686.bam -o SRR7725686_sorted.bam

#Convert gff3 to gtf
gffread spombe.gff -T -o spombe.gtf

#Run htseq-count with both sorted BAM files.
htseq-count -f bam -r name SRR2567795/SRR7725695_sorted.bam \
SRR2567786/SRR7725686_sorted.bam spombe.gtf > counts.txt
```

```
getwd()
setwd("/home/kenizzer/Downloads")
library(edgeR)

#Reading and tidying data
count_table <- read.delim("/home/kenizzer/Downloads/counts.txt", header = FALSE,\
 row.names = 1, sep = "\t")
colnames(count_table) <- c("sample_1","sample_2")
y = DGEList(count_table, group=1:2)

y$samples

#Filter out lowly expressed genes
keep <- rowSums(cpm(y)>1) >= 2
```

```r
y <- y[keep, , keep.lib.sizes=FALSE]

#Normalize for differnces in library size
y <- calcNormFactors(y)
y$samples

#Accounting for the lack of replicates within the data
#must set the biological variability coef. by hand
#I selected 0.2 because this study compared wild type
#fission yeasts. This likely means that there is some variability
#between samples of the same organism but likely not as much as is
#common in a study with humans.

bcv <- 0.2
et <- exactTest(y, dispersion=bcv^2)
FDR <- p.adjust(et$table$PValue, method="BH")
sum(FDR < 0.05)
plotMD(et)
abline(h=c(-1,1), col="blue")


#Result 427 genes that are deferentially expressed between the samples
```

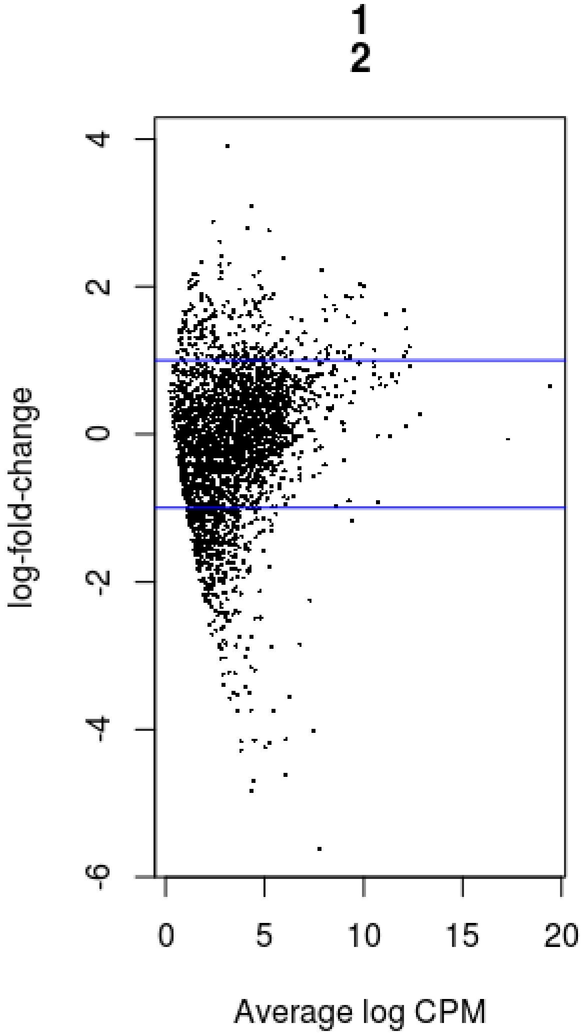

Figure 1: plotSmear plot generated using the edgeR package.