Jeffrey Liao
9/17/2020
Quantitative Biology Lab

Assignment 1: Genome Assembly

Question 1a:

Commands:

```
cd /Users/cmdb/qbb2020-answers/assignment1/asm
samtools faidx ref.fa
```

Answer:

233806bp. The first two fields of the index file are the reference name and length of the fasta file.

Question 1b:

Commands:

```
fastQC frag180.1.fq
fastQC frag180.2.fq
fastQC jump2k.1.fq
fastQC jump2k.2.fq
```

Answer:

Both frag180.1.fq and frag180.2.fq have 35178 reads with 100bp per read
Both jump2k.1.fq and jump2k.2.fq have 70355 reads with 50bp per read

Question 1c:

Answer:

For the frag180 files, disregarding the 20bp sequence overlap between the two reads for 180bp sized fragment:

180* 35178 / 233806 bp = 27.08X coverage

For the jump2k files:

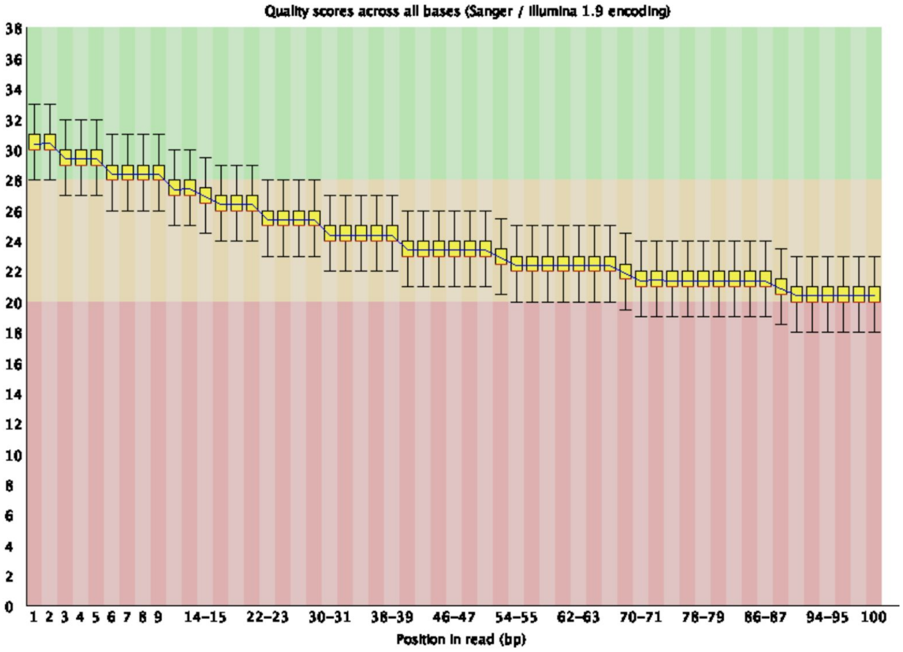
(50*2)*70335 / 233806 = 30.08X coverage

Jeffrey Liao
9/17/2020
Quantitative Biology Lab

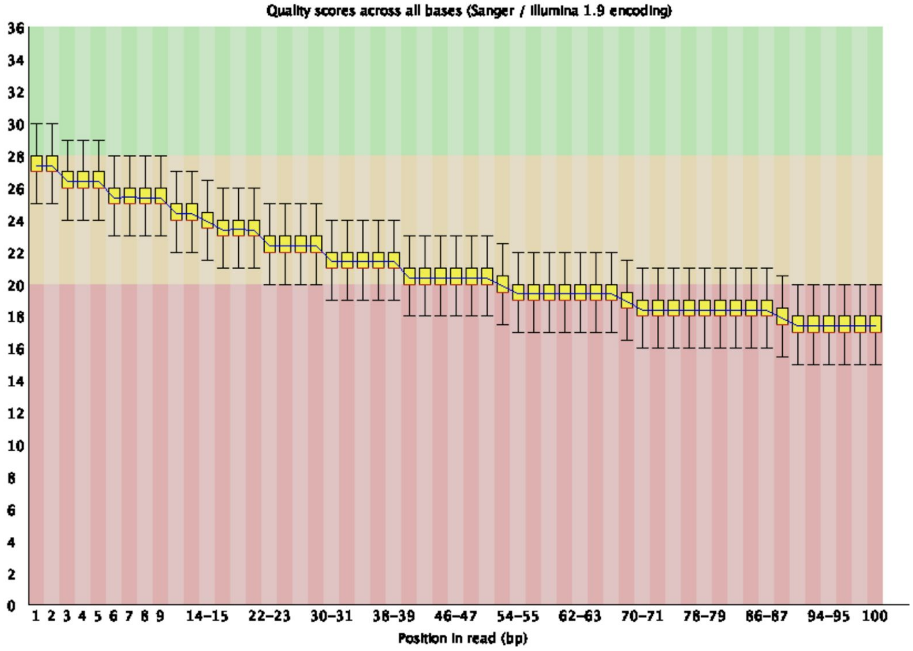
Question 1d:

- Opened the html files generated from fastQC

frag180.1.fq

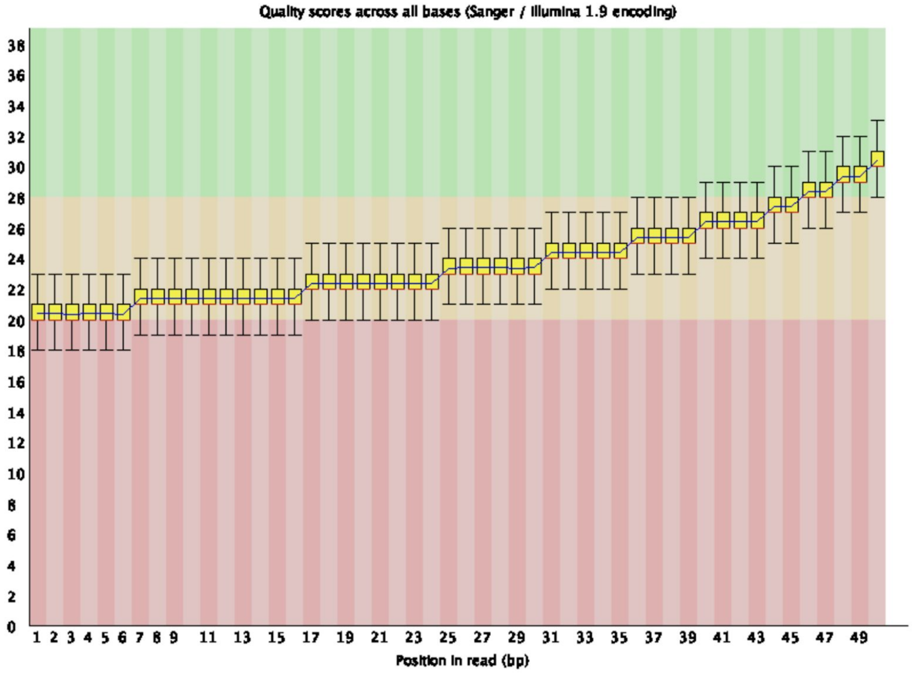


frag180.2.fq

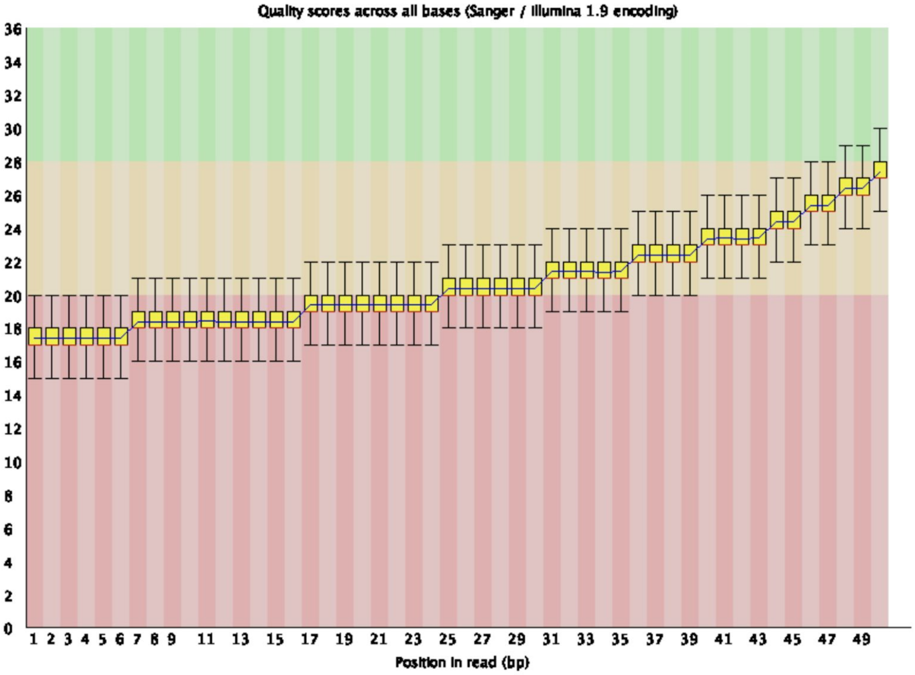


Jeffrey Liao
9/17/2020
Quantitative Biology Lab

jump2k.1.fq



jump2k.2.fq



Jeffrey Liao
9/17/2020
Quantitative Biology Lab

Question 2a:

Commands:

```
# generate the output file from jellyfish
jellyfish count -m 21 -C -s 1000000 /Users/cmdb/qbb2020-answers/assignment1/asm/*fq

# generate the histograms from the jellyfish output files
jellyfish histo mer_counts.jf > reads.histo
```

Answer:

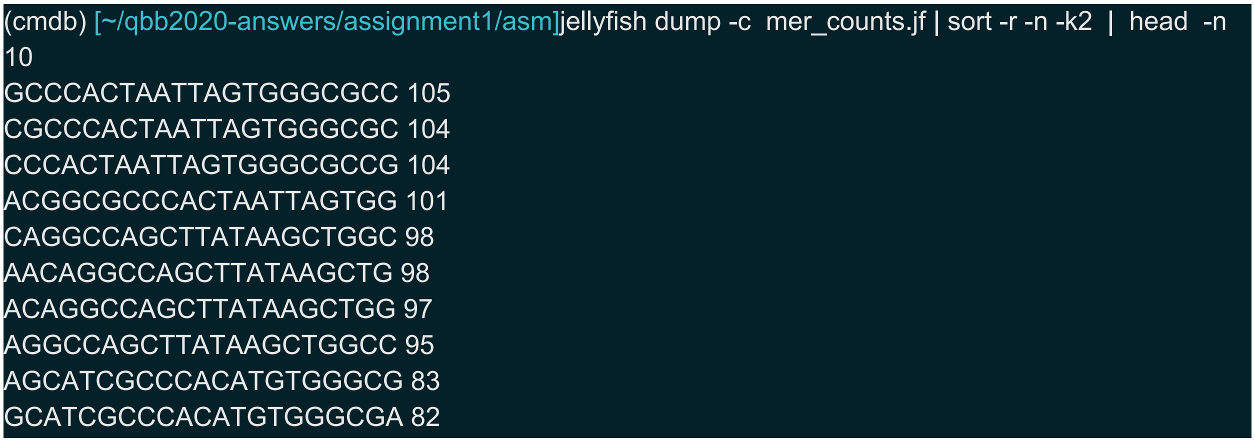
1091 different kmers that occur exactly 50 times.

Question 2b:

Commands:

```
jellyfish dump -c mer_counts.jf | sort -r -n -k2 | head -n 10
```

Answer:



```
(cmdb) [~/qbb2020-answers/assignment1/asm]jellyfish dump -c  mer_counts.jf | sort -r -n -k2  |  head  -n
10
GCCCACTAATTAGTGGGCGCC 105
CGCCCACTAATTAGTGGGCGC 104
CCCACTAATTAGTGGGCGCCG 104
ACGGCGCCCACTAATTAGTGG 101
CAGGCCAGCTTATAAGCTGGC 98
AACAGGCCAGCTTATAAGCTG 98
ACAGGCCAGCTTATAAGCTGG 97
AGGCCAGCTTATAAGCTGGCC 95
AGCATCGCCCACATGTGGGCG 83
GCATCGCCCACATGTGGGCGA 82
```

Question 2c:

Answer: The min genome haploid length is 233468bp.

Question 2d:

Answer: The GenomeScope estimation is relatively accurate with a difference of 338bp compared to the true reference length of 233806bp.

Jeffrey Liao
9/17/2020
Quantitative Biology Lab

Question 3a:

Commands:

```
#alignment using spades
spades.py --pe1-1 frag180.1.fq --pe1-2 frag180.2.fq --mp1-1 jump2k.1.fq --mp1-2 jump2k.2.fq -o asm -t 4 -k 31

cd ~/Users/cmdb/qbb2020-answers/assignment1/asm/asm

#count number of contigs

grep -c ‘>’ contigs.fasta
```

Answer: 4 different contigs

Question 3b:

Commands:

```
samtools faidx contigs.fasta

#visually inspect index file, size of contigs is the 2nd field in each line
vim contig.fasta.fai
```

Answer:

Contig 1 is 105831 bp
Contig 2 is 47861 bp
Contig 3 is 41352 bp
Contig 4 is 38423 bp

The total length of the contigs is 233467 bp

Question 3c:

Commands:

```
sort -r -n -k2 contigs.fasta.fai
```

Answer: The length of the longest contig is 105831

Question 3d:

Wrote some code in jupyter notebook name Assignment1_Code and uploaded to GitHub
Answer: The N50 is: 39423

Jeffrey Liao
9/17/2020
Quantitative Biology Lab

Question 4a:

Commands:

```
#perform alignment
dnadiff ../ref.fa contigs.fasta

#read alignment report
vim out.report
```

Answer:

The average identity for 1-1 alignments between the reference and the query was 100%

Question 4b:

Commands:

```
nucmer ../ref.fa ./contigs.fasta
show-coords out.delta
```

Answer: The length of the longest alignment is 108531

Question 4c:

Commands:

```
vim out.report
```

Answer: Looking at the out.report file generated in question 4a, it looks like there is a single insertion in the query sequence/contigs with length 712 bp. It doesn't seem like there are any deletions.

Question 5a:

Commands:

```
show-coords out.delta
```

Answer:

The insertion appears to be in the NODE_3_length_41352_cov_20.588756 contig from base 13853 to 14566. The insertion corresponding to the reference would be between bases 26789 and 26790.

Jeffrey Liao
9/17/2020
Quantitative Biology Lab

Question 5b:

Commands:

```
show-coords out.delta
```

Answer:

Subtracting 14566-13853 gives 713bp. Since we are looking at insertions, we need to subtract the difference by 1 to get the true length of the insert = 712 bp.

Question 5c:

Commands:

```
samtools faidx contigs.fasta NODE_3_length_41352_cov_20.588756:13854-14565 >
extracted.fa

cat extracted.fa
```

Answer:

```
>NODE_3_length_41352_cov_20.588756:13854-14565
TAACGATTTACATCGGGAAAGCTTAATGCAATTCACGCAGATATTCAGCTTAGAAGGTAC
GCAGCGGTGACGGGGTGCGGTCCATAATCTATGAAGCTATGAATTCGTACCTCAAGTAAT
GTTTTCTTCGCTGCAGTTCAGAAGTGATAAAGGTATCCCGCTTAGCCTGGCATACTTTGT
GCGTTCGTACCGCCCAGCATTAATGACTTGTGTAGGCAAGTAATGAACGACTCTTCTACG
CCGCGCCTAACCTCCGCACATAATGGCAGCATGTGGTAGTTACATACGCACAGAAGTGGT
TCGGTTTTAACTATAGTCAGATATGAATAAGCTGCGTGTGTCGTTGTGTCGGCGTGTCGT
ACTTACCTCCTGACATAGGTGAATTTCAGCCTACTGTAAGTTTGGAGTCGCGCTCTTTTC
TTATTATATTCTTTGGTATGTGTGTGATGGGTTCGGGCGTGTATTGATGTCTCTAAGGCT
CATGTTAGTGTTTATTTGGTCAGTTATGACGGTGTTCCTGTCGTACGTGTTGGCTTAGCG
GACTTGTAGACGGGATCAAGGTTGTCTGACCCTCCGGTCGACCGTGGGTCGGCCGTCCCG
GCCAGAATACAAGCCGCTTAGACTTTCGAAAGAGGGTAAGTTACTACGCGCGAACGTTAT
ACCTCGTTTCAGTATGCACTCCCTTAAGTCACTCAGAAAAGACTAAGGGGCT
```

Question 5d:

Commands:

```
mv extracted.fa ..
cd ..
python ported_decoder.py --decode --rev_comp --input extracted.fa
```

Answer:

```
The decoded message :  Congratulations to the 2020 CMDB @ JHU class!  Keep on looking for little green aliens...
```