Joshua Derrick September 18, 2020

Quant Bio Lab 1

Question 1

1a

Command: (in the asm folder). samtools faidx ref.fa. We then took the second column of the outputted text file and got 233806 bp

1b

Commands: fastQC frag180.1.fq, fastQC frag180.2.fq,fastQC jump2k.1.fq.fq fastQC jump2k.2.fq. Open the HTML files that are provided. 35,178 reads in frag180.1.fq and frag180.2.fq , each 100bp. 70,355 sequences, each 50 bp for the jump2k files.

1c

This should be $\frac{reads*readlength}{genomesize}$. In our case this is equal to $\frac{35{,}178*100}{233{,}806}$ which is 15x coverage for each file. If you include all files, then this would be multiplied by 4, for 60x.

1d

.

Question 2

2a

Commands: jellyfish count -m 21 -C -s 1000000 *.fq, jellyfish histo mer_counts.jf ¿ reads.histo. This gives you a value of 1091 kmers which are seen 50 times in the file.

Figure 1: frag180.1

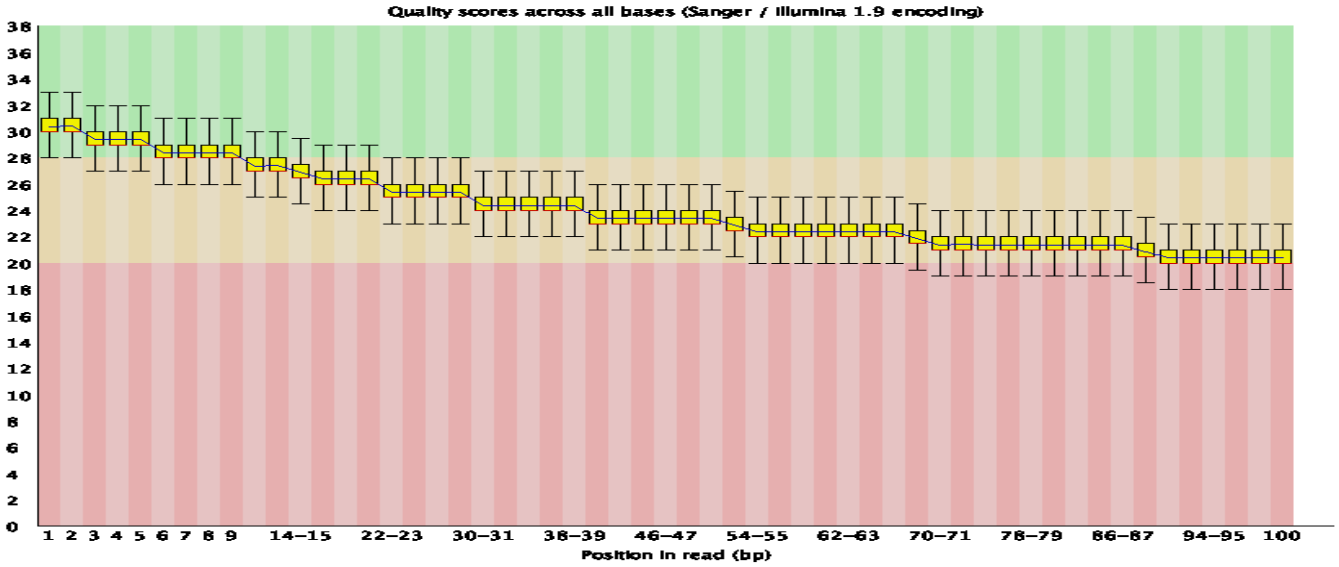


Figure 2: frag180.2

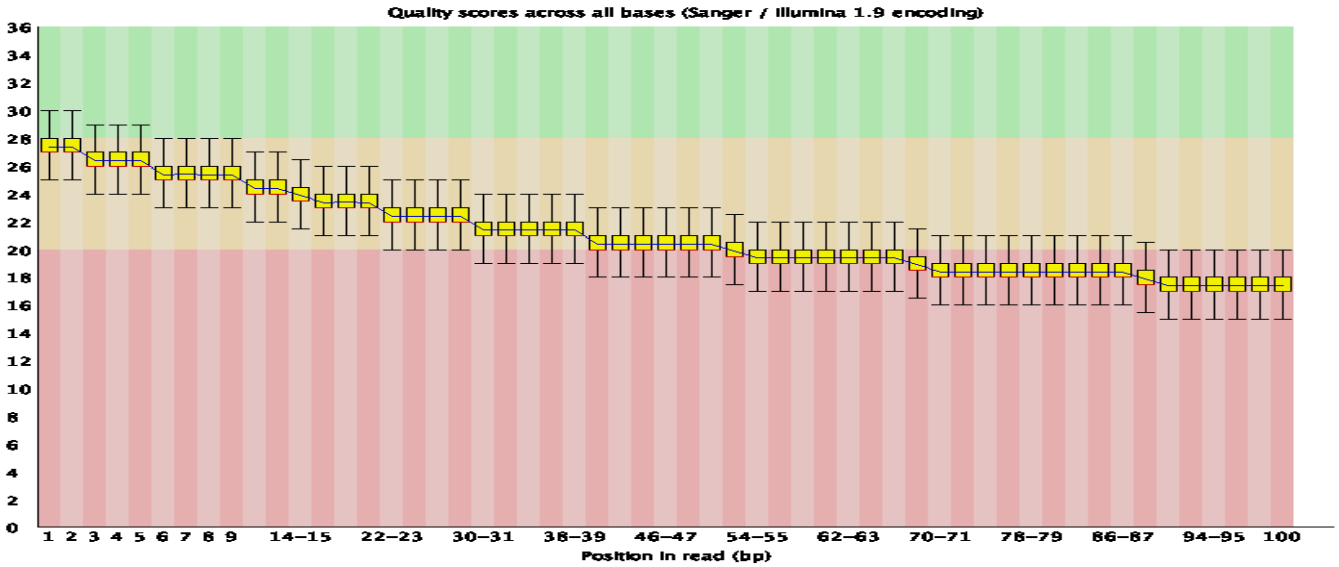


Figure 3: jump2k.1

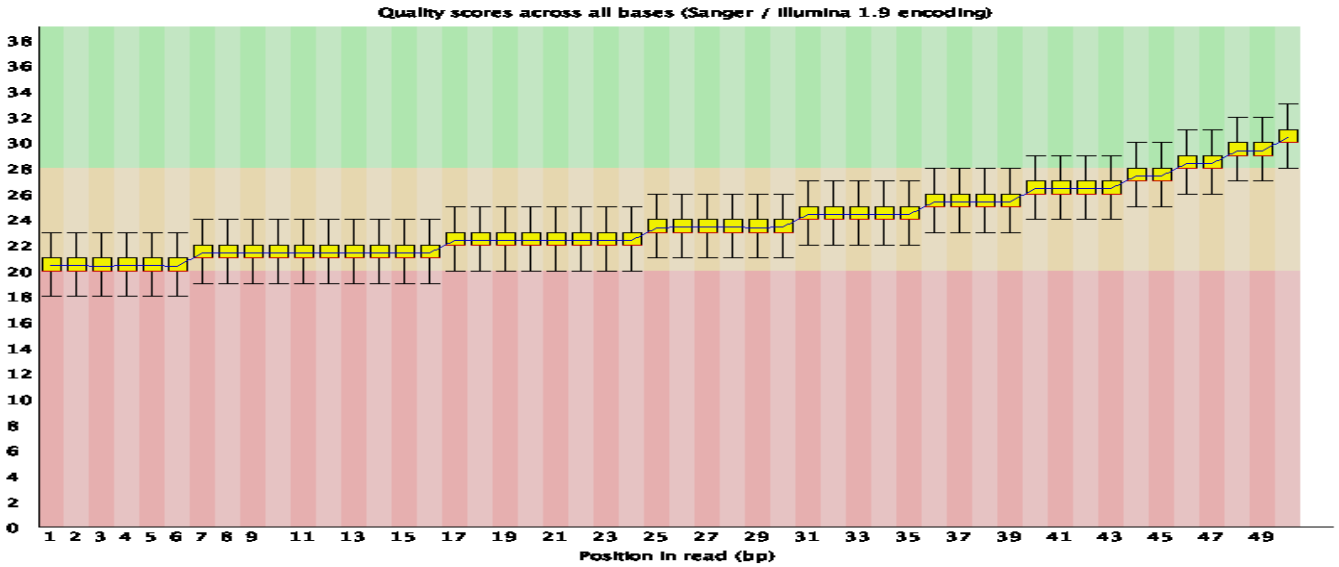
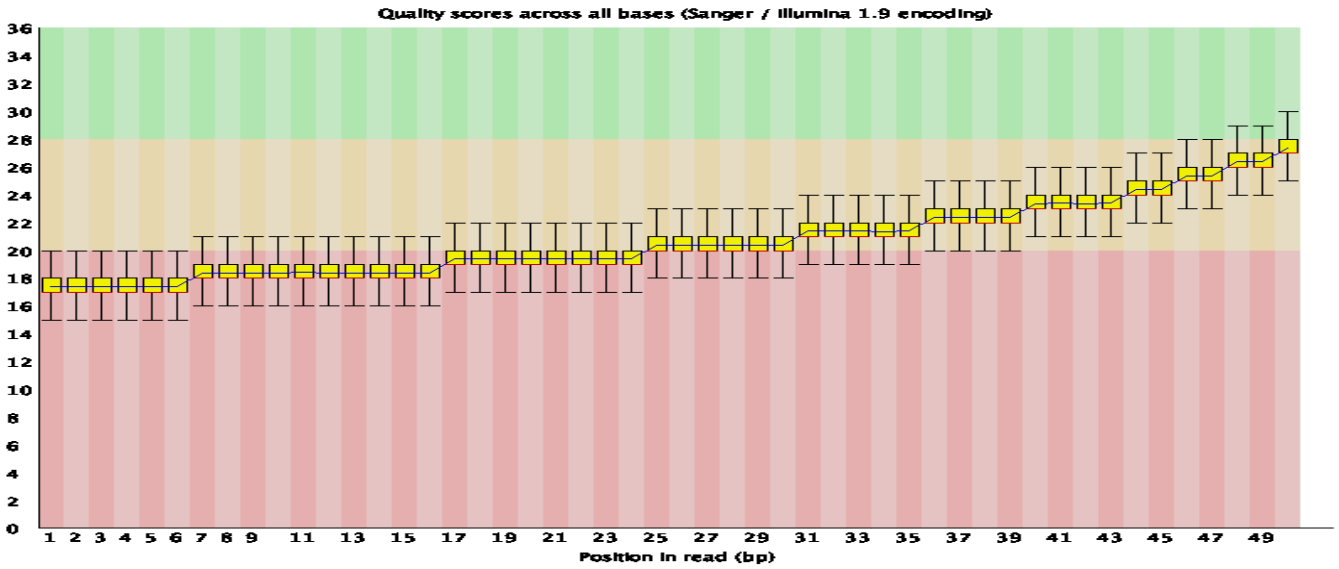


Figure 4: jump2k.2



2b

Commands: jellfish dump -c mer_counts.jf > kmers.txt, sort kmers.txt -k 2 -rn | head
Kmers:
GCCCACTAATTAGTGGGCGCC 105
CGCCCACTAATTAGTGGGCGC 104
CCCACTAATTAGTGGGCGCCG 104
ACGGCGCCCACTAATTAGTGG 101
CAGGCCAGCTTATAAGCTGGC 98
AACAGGCCAGCTTATAAGCTG 98
ACAGGCCAGCTTATAAGCTGG 97
AGGCCAGCTTATAAGCTGGCC 95
AGCATCGCCCACATGTGGGCG 83
GCATCGCCCACATGTGGGCGA 82

2c

233,468

2d

It is extremely close, although undershot by a few hundred bp. This is undershot because the genome coverage is not enough to guarantee 100 percent coverage.

Question 3

3a

Command: grep -c > contigs.fasta. This gives us a contig count of 4.

3b

105831+47861+41352+39423= 234467, which is longer than the genome. This makes sense, because contigs probably contain some shared regions.

3c

105831

3d

47861 is the N50 size because $105831 + 47861 > \frac{233806}{2}$

Problem 4

4a

Command: dnadiff ref.fa asm/contigs.fa. Both alignments are 100 percent.

4b

Commands: nucmer ref.fa asm/contigs.fa, show-coord out.delta. This gives us four alignments, the longest of which is the full first contig which is 105831 long.

4c

There is one insertion into the third contig, denoted by the gap between the third and fourth alignments.

Problem 5

5a

The insert is at the 13853 position of the third contig. In the reference its at 26789.

5b

$14565 - 13854 = 711$

5c

Command: samtools faidx contigs.fasta NODE_3_length_41352_cov_20.588756:13854-14565>NODE_3_length_41352_cov_20.588756:13854-14565

TAACGATTTACATCGGGAAAGCTTAATGCAATTCACGCAGATATTCAGCTTAGAAGGTA
CGCAGCGGTGACGGGGTGCGGTCCATAATCTATGAAGCTATGAATTCGTACCTCAAG-
TAA TGTTTTCTTCGCTGCAGTTCAGAAGTGATAAAGGTATCCCGCTTAGCCTGGCAT-
ACTTTG TGCGTTCGTACCGCCCAGCATTAATGACTTGTGTAGGCAAGTAATGAAC-
GACTCTTCTAC GCCGCGCCTAACCTCCGCACATAATGGCAGCATGTGGTAGTTA-
CATACGCACAGAAGTGG TTCGGTTTTAACTATAGTCAGATATGAATAAGCTGCGT-
GTGTCGTTGTGTCGGCGTGTCG TACTTACCTCCTGACATAGGTGAATTTCAGCC-
TACTGTAAGTTTGGAGTCGCGCTCTTTT CTTATTATATTCTTTGGTATGTGTGTGATGGGTTC
GTATTGATGTCTCTAAGGC TCATGTTAGTGTTTATTTGGTCAGTTATGACGGTGTTC-
CTGTCGTACGTGTTGGCTTAGC GGACTTGTAGACGGGATCAAGGTTGTCTGAC-
CCTCCGGTCGACCGTGGGTCGGCCGTCCC GGCCAGAATACAAGCCGCTTAGACTTTC-
GAAAGAGGGTAAGTTACTACGCGCGAACGTTA TACCTCGTTTCAGTATGCACTC-
CCTTAAGTCACTCAGAAAAGACTAAGGGGCT

5c

Command: python ported_decoder.py -d –input hidden.txt –rev_comp
Message: Congratulations to the 2020 CMDB @ JHU class! Keep on looking for little green aliens...