Readme

frag180.1.fq Read 1 of Illumina 2x100 reads from 180 +/- 20 bp fragments

frag180.2.fq Read 2 of Illumina 2x100 reads from 180 +/- 20 bp fragments

jump2k.1.fq Read 1 of Illumina 2x50 reads from 2000 +/- 200 bp fragments

jump2k.2.fq Read 2 of Illumina 2x50 reads from 2000 +/- 200 bp fragments

Brad Olinger

Quant Bio

Week 1

9/10/21

Commands

* Question 1a. How long is the reference genome? [Hint: Try samtools faidx]

Command: samtools faidx ref.fa

Output: Halomonas 233806 11 70 71

233806 is the length of the refernce genome

* Question 1b. How many reads are provided and how long are they? Make sure to measure each file separately [Hint: Try FastQC]

Command

FastQC frag180.1.fq

Open frag180.1\_fastqc.html

Sequences: 35178

Length: 100

Command

FastQC frag180.2.fq

Open frag180.2\_fastqc.html

Sequences: 35178

Length: 100

Command

FastQC jump2k.1.fq

Open jump2k.1\_fastqc.html

Sequences: 70355

Length: 50

Command

FastQC jump2k.2.fq

Open jump2k.2\_fastqc.html

Sequences: 70355

Length: 50

* Question 1c. How much coverage do you expect to have? [Hint: A little arthmetic]

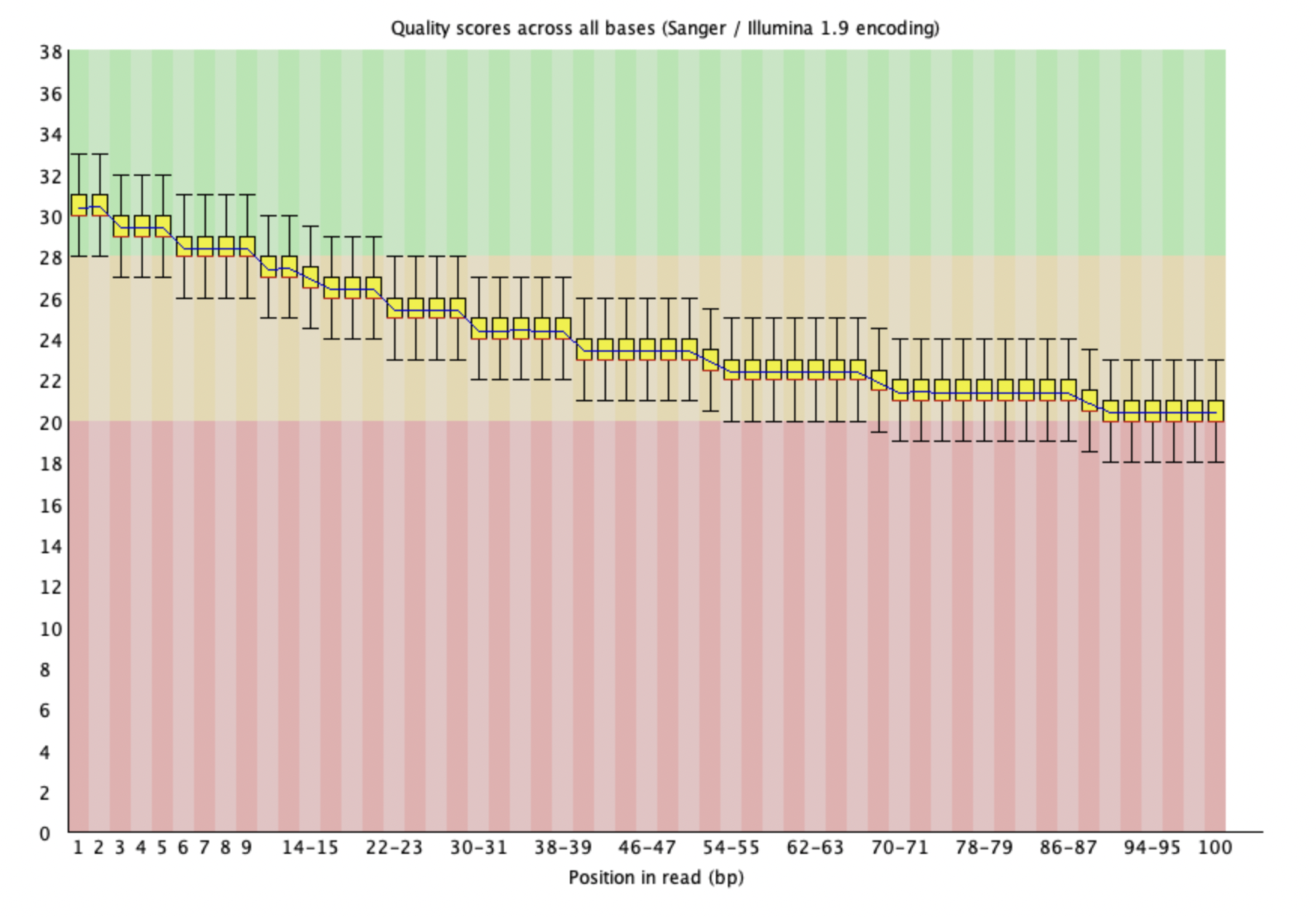
Frag180.1 and frag180.2 = 3,517,800, x 2

Jump2k.1 and Jumk2k.2 = 3,517,750 x 2

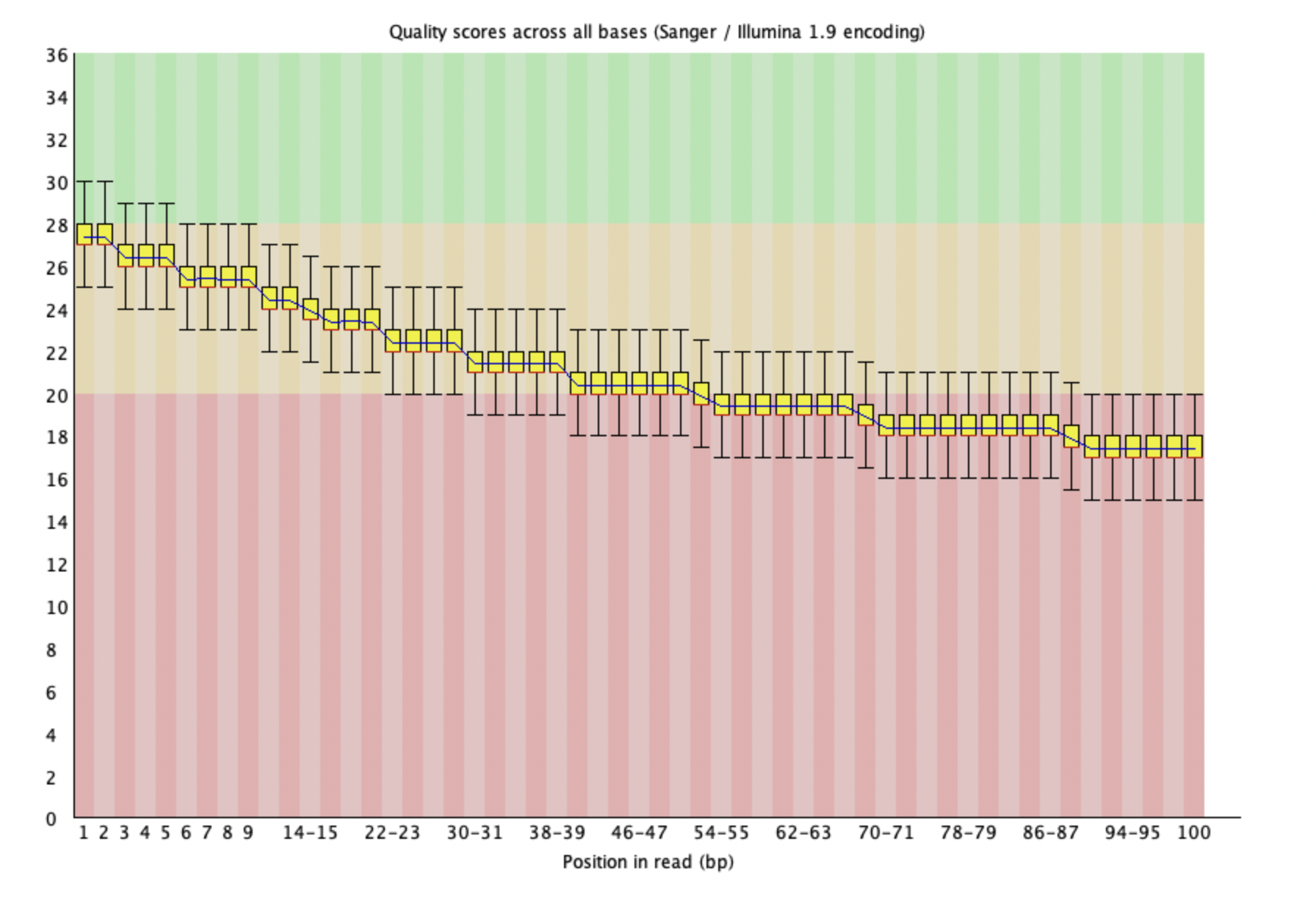
Total = 14,071,100 / 233,806 = 60.18 coverage

* Question 1d. Plot the average quality value across the length of the reads [We want a screenshot from FastQC]

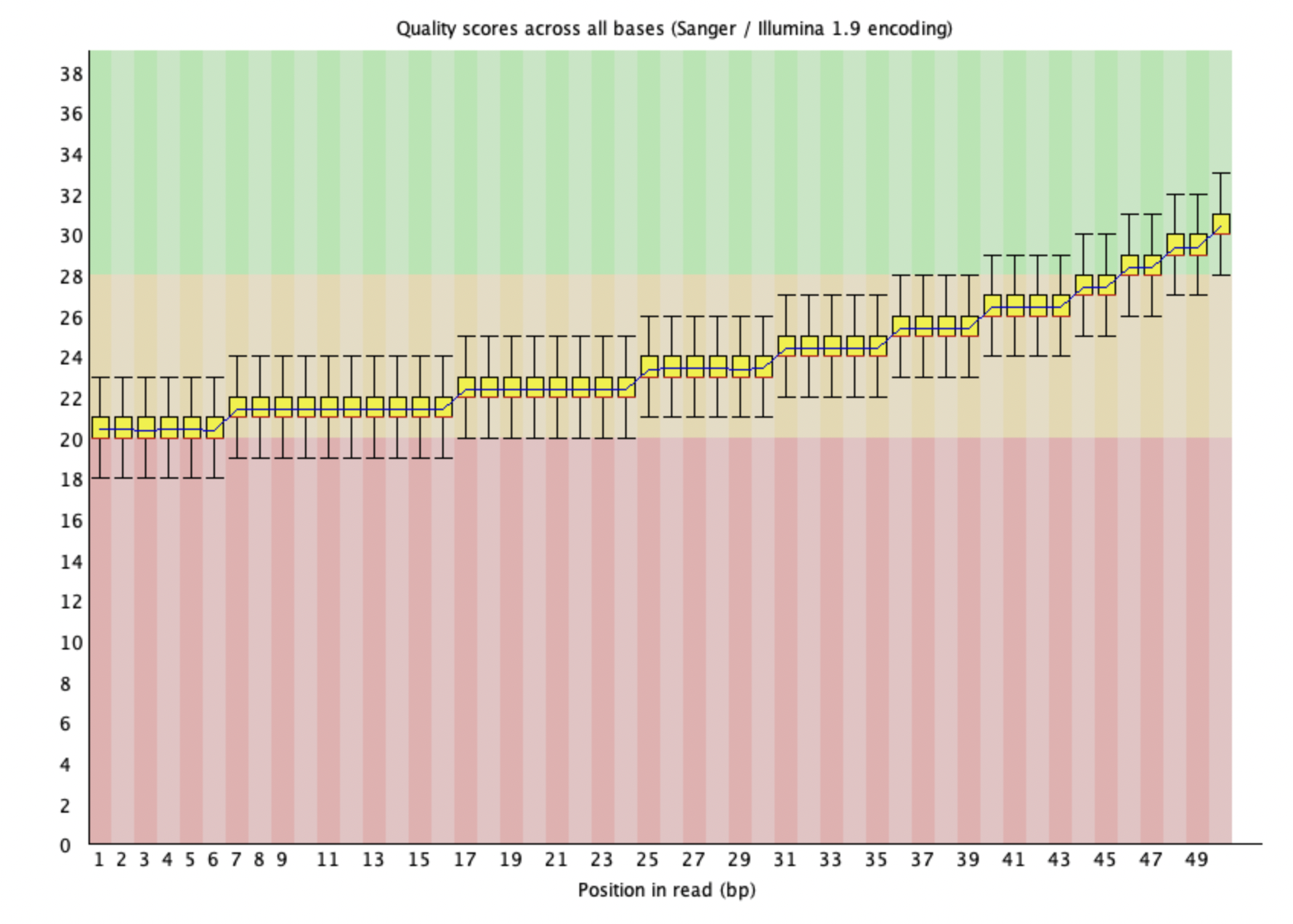
Frag180.1



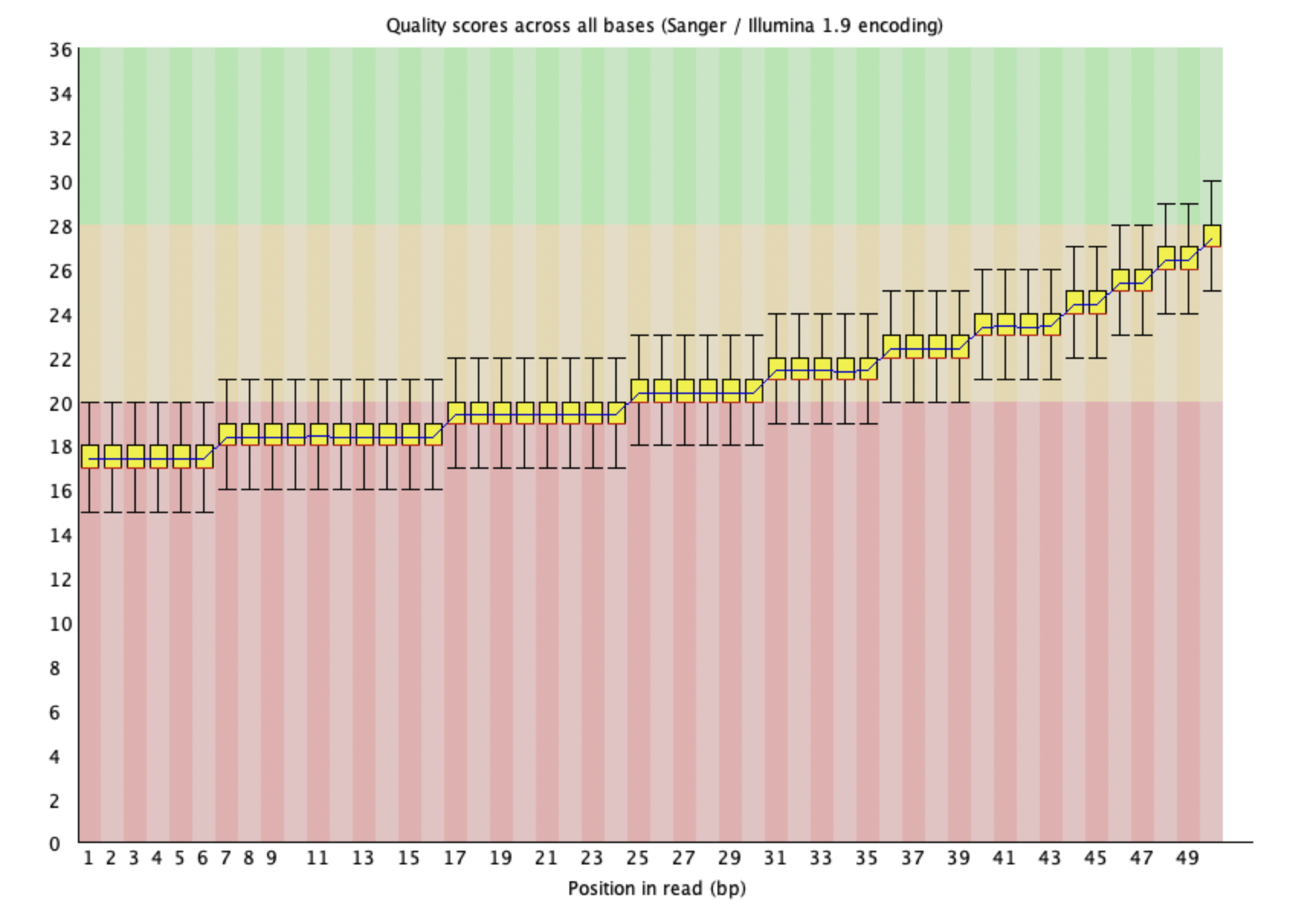
Frag180.2



Jump2k.1



Jump2k.2



* Question 2a. How many kmers occur exactly 50 times? [Hint: try jellyfish histo]

Command:

* jellyfish count -m 21 -s 100M -t 10 -C frag180.1.fq frag180.2.fq jump2k.1.fq jump2k.2.fq
* jellyfish histo mer\_counts.jf

Output: 1030 k-mers that occurred exactly 50 times

* Question 2b. What are the top 10 most frequently occurring kmers [Hint: try jellyfish dump along with sort and head]

Command:

* jellyfish dump -c mer\_counts.jf | sort -n -r -k 2 | head

Output:

GCCCACTAATTAGTGGGCGCC 104

CGCCCACTAATTAGTGGGCGC 104

CCCACTAATTAGTGGGCGCCG 104

ACGGCGCCCACTAATTAGTGG 102

AACAGGCCAGCTTATAAGCTG 100

ACAGGCCAGCTTATAAGCTGG 99

CAGGCCAGCTTATAAGCTGGC 96

AGGCCAGCTTATAAGCTGGCC 94

AGCATCGCCCACATGTGGGCG 82

* GCATCGCCCACATGTGGGCGA 80
* Question 2c. What is the estimated genome size based on the kmer frequencies? [Hint: upload the jellyfish histogram to [GenomeScope](http://genomescope.org/) and report the min “Genome Haploid Length” in the “Results” section]

Min haploid length: 233,510 bp

* Question 2d. How well does the GenomeScope genome size estimate compare to the reference genome? [Hint: In a sentence or two]

The GenomeScope genome size is quite close to the actual genome size (233,806) but it is not in the min to max range for the haploid genome length. (233,510 – 233,799).

Question 3. De novo assembly

Assemble the reads using Spades. Spades will *not* run on Windows you must use a linux or mac environment.

* Question 3a. How many contigs were produced? [Hint: try grep -c '>' contigs.fasta]

Command: spades.py -1 frag180.1.fq -1 jump2k.1.fq -2 frag180.2.fq -2 jump2k.2.fq -o ~/qbb2021-answers/QuantBioLab/Week1/

Retrying commands: spades.py --pe1-1 frag180.1.fq --pe1-2 frag180.2.fq --mp1-1 jump2k.1.fq --mp1-2 jump2k.2.fq -o ~/qbb2021-answers/QuantBioLab/Week1/ -t 4 -k 31

grep -c '>' contigs.fasta

Output: 4

* Question 3b. What is the total length of the contigs? [Hint: try samtools faidx, plus a short script if necessary]

Commands: samtools faidx contigs.fasta

Cat contigs.fasta..fai

awk '{ sum += $2 } END { print sum }' contigs.fasta.fai

Output:

NODE\_1\_length\_105830\_cov\_20.649108 105830 36 60 61

NODE\_2\_length\_47860\_cov\_20.367392 47860 107665 60 61

NODE\_3\_length\_41351\_cov\_20.528098 41351 156358 60 61

NODE\_4\_length\_39426\_cov\_20.336388 39426 198434 60 61

Total length: 234467

* Question 3c. What is the size of your largest contig? [Hint: check samtools faidx plus sort -n]
* Command: cat contigs.fasta.fai | sort -n -r -k 2 | head -n 1

Output: NODE\_1\_length\_105830-cov-20.649108 , length 105830

* Question 3d. What is the contig N50 size? [Hint: Write a short script if necessary]
  + N50 size = 47860 ( the size of the contig that incorporates at least half of the genome)

Question 4. Whole Genome Alignment

Use MUMmer for whole genome alignment.

* Question 4a. What is the average identify of your assembly compared to the reference? [Hint: try dnadiff]

Commands:

dnadiff /path/to/ref.fa /path/to/qry.fa

$ nucmer /path/to/ref.fa /path/to/qry.fa

$ show-coords out.delta

dnadiff ~/qbb2021-answers/QuantBioLab/Week1/ref.fa ~/qbb2021-answers/QuantBioLab/Week1/contigs.fasta

nucmer ~/qbb2021-answers/QuantBioLab/Week1/ref.fa ~/qbb2021-answers/QuantBioLab/Week1/contigs.fasta

show-coords out.delta

average identity: 100%

[S1] [E1] | [S2] [E2] | [LEN 1] [LEN 2] | [% IDY] | [TAGS]

=====================================================================================

127965 233794 | 1 105830 | 105830 105830 | 99.99 | Halomonas NODE\_1\_length\_105830\_cov\_20.649108

40651 88510 | 1 47860 | 47860 47860 | 100.00 | Halomonas NODE\_2\_length\_47860\_cov\_20.367392

3 26789 | 1 26787 | 26787 26787 | 100.00 | Halomonas NODE\_3\_length\_41351\_cov\_20.528098

26790 40641 | 27500 41351 | 13852 13852 | 100.00 | Halomonas NODE\_3\_length\_41351\_cov\_20.528098

88532 127957 | 1 39426 | 39426 39426 | 100.00 | Halomonas NODE\_4\_length\_39426\_cov\_20.336388

* Question 4b. What is the length of the longest alignment [Hint: try nucmer and show-coords]

Longest length: 105830

* Question 4c. How many insertions and deletions are in the assembly? [Hint: try dnadiff]

Command: cat out.report

Insertions in Qry = 1

Insertions sum = 712

Question 5. Decoding the insertion

* Question 5a. What is the position of the insertion in your assembly? Provide the corresponding position in the reference. [Hint: try show-coords]

Command: cat out.qdiff

Output: GAP 26788 27499

Command: out.rdiff

Output: GAP 26790-26789

* Question 5b. How long is the novel insertion? [Hint: try show-coords]

Answer: 712

* Question 5c. What is the DNA sequence of the encoded message? [Hint: try samtools faidx to extract the insertion]

Command: samtools faidx ~/qbb2021-answers/QuantBioLab/Week1/contigs.fasta NODE\_3\_length\_41351\_cov\_20.528098:'26788-27499'

Sequence: CGCCCATGCGTAGGGGCTTCTTTAATTACTTGATTGACGCATGCCCCTCGTTCTACATGT

CTAGCTTCGTAACTGCCCCGATTTATACAGGAGCATATGCGTTTCGTAGTGCCGGGAATG

CATACCAAAGGGCTCACGGCGGGTACGCCACAATGGCTCAAGTCGAAAATGAATCGAAGA

CAACAAGGAATACCGTACCCAATTACTCAAGGACCTCATACACCATCCCATGCTACTTAT

CTACAGACATACACGCCAGCACCCAGCAACCAAAGCACACCGACGATAAGACTACAATCG

CGATAAGCACAACTTACATTAGGAGGCCCGGCAAATCTTGACGGCGTTAAGTCCGACACG

AATACCCCCCGACAAAAGCCTCGTATTCCGAGAGTACGAGAGTGCACAAAGCACCAAGGC

GGGGCTTCGGTACATCCACCAGTAGTCCCGTCGTGGCGGATTTTCGTCGCGGATGATCCG

AGGATTTCCTGCCTTGCCGAACACCTTACGTCATTCGGGGATGTCATAAAGCCAAACTTA

GGCAAGTAGAAGATGGAGCACGGTCTAAAGGATTAAAGTCCTCGAATAACAAAGGACTGG

AGTGCCTCAGGCATCTCTGCCGATCTGATTGCAAGAAAAAATGACAATATTAGTAAATTA

GCCTATGAATAGCGGCTTTAAGTTAATGCCGAGGTCAATATTGACATCGGTA

* Question 5d. What is the secret message? [Hint: Run the provided script dna-decode.py to decode the string from 5c.]

Save the output to fasta file:;;

Command: samtools faidx ~/qbb2021-answers/QuantBioLab/Week1/contigs.fasta NODE\_3\_length\_41351\_cov\_20.528098:'26788-27499' > final.fasta

Decode:

Command:

python3 dna-decode.py -d --input final.fasta

Output:

Congratulations to the 2021 CMDB @ JHU class! Keep on looking for little green aliens...