Question 1. Coverage Analysis

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

233806 bps

~/desktop/asm/$samtools faidx ref.fa

$cat ref.fa.fai

Halomonas 233806 11 70 71

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

~/desktop/asm/$FastQC \*.fq  
 $open \*.html

4 reads provided:

|  |  |
| --- | --- |
| Filename | frag180.1.fq |
| Total Sequences | 35178 |
| Sequence length | 100 |

|  |  |
| --- | --- |
| Filename | frag180.2.fq |
| Total Sequences | 35178 |
| Sequence length | 100 |

|  |  |
| --- | --- |
| Filename | jump2k.1.fq |
| Total Sequences | 70355 |
| Sequence length | 50 |

|  |  |
| --- | --- |
| Filename | jump2k.2.fq |
| Total Sequences | 70355 |
| Sequence length | 50 |

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

We expect to have ~60x coverage.

Frag180.1, 180.2  
Total sequences (35178) \* sequence length (100) \* 2 (files) = 7,035,600

Jump2k.1, 2k.2   
Total sequences (70355) \* sequence length (50) \* 2( files) = 7,035,500

Total bases sequenced = 7,035,600 + 7,035,500 = 14,071,110 bps

Total bases sequenced (14,071,110 bps) / length of ref seq (233806 bps) = 60.1828011

**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 2. Kmer Analysis

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

1030 kmers occur exactly 50 times

~/desktop/asm/$jellyfish count -m 21 -C -s 1000000 \*.fq  
-m = Length of kmer, -C = count both strand? -s = initial hash size

$jellyfish histo mer\_counts.jf > reads.histo

$cat reads.histo

…  
50 1030  
…

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

~/desktop/asm/$jellyfish dump -c mer\_counts.jf | sort -n -r -k 2 | head -n 10

Jellyfish -c = column format  
sort -n = sort by number, -r = sort by reverse order, -k 2 = sort by second field

Head -n 10 = print first 10 lines (by number)

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]**

The estimated genome size is range 233,510 – 233799 bps

min max

Genome Haploid Length 233,510 bp 233,799 bp

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

The GenomeScope estimated genome size gave a range of 233,510 – 233799 bps and the reference genome is 233806 bps long. The estimation is close to the reference genome, however the reference genome is outside of it’s range.

Question 3. De novo assembly

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

4 contigs were produced.

~/desktop/asm/$ spades.py --pe1-1 frag180.1.fq --pe1-2 frag180.2.fq --mp1-1 jump2k.1.fq --mp1-2 jump2k.2.fq -o contigdata -t 4 -k 3

\*\*\*created new output folder “contigdata” in current directory\*\*\*

~/desktop/asm/$cd contigdata  
~/desktop/asm/asm/$grep -c '>' contigs.fasta  
4

Grep -c = count number of lines

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

The total length of the contigs is 234,446 bps

~/desktop/asm/contigdata/$samtools faidx contigs.fasta

~/desktop/asm/contigdata/$awk '{sum += $2} END {print sum}' contigs.fasta.fai

234467

Awk ‘script’ filenames

Awk /pattern/ {actions}  
sum += $2 = sum second column

**Question 3c. What is the size of your largest contig? [Hint: check samtools faidx plus sort -n]**

The largest contif is 105,830 bps long. It is in Node 1.

~/desktop/asm/contigdata/$cat contigs.fasta.fai | sort -n -r -k 2 | head -n 1 | awk '{print $2}'

105830

Take contigs.fasta.fai data | sort 2nd field numerically in reverse order (hi > lo) | take first line | print second field of first line

**Question 3d. What is the contig N50 size? [Hint: Write a short script if necessary]**

The contig N50 size is the size of Node 2, which is 47860 bps.

Question 4. Whole Genome Alignment

**Question 4a. What is the average identity of your assembly compared to the reference? [Hint: try dnadiff]**Average identity of assembly compared to reference = 99.998%

~/desktop/asm/$cp ref.fa contigdata/  
#copy ref.fa file into contigdata directory

~/desktop/asm/contigdata/$dnadiff ref.fa contigs.fasta

~/desktop/asm/contigdata/$nucmer ref.fa contigs.fasta

~/desktop/asm/contigdata/$show-coords out.delta

[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

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

The longest alignment is Node\_1, which is 105,803 bps long (see table above)

**Question 4c. How many insertions and deletions are in the assembly? [Hint: try dnadiff]**There is 1 insertion in the query.  
No indels… unsure how to find deletions?

~/desktop/asm/contigdata/$cat out.report

NUCMER

[REF] [QRY]

[Sequences]

Insertions 5 1

InsertionSum 51 712

InsertionAvg 10.20 712.00

**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]**

~/desktop/asm/contigdata/$cat out.qdiff

NODE\_3\_length\_41351\_cov\_20.528098 GAP 26788 27499 712 0 712

The insertion spans position 26788 – 27499 in Node 3 of the query assembly.

~/desktop/asm/contigdata/$cat out.rdiff

Halomonas GAP 26790 26789 0 712 -712

The insertion spans position 26790– 26789 in the reference.

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

The novel insertion is 712 bps.

~/desktop/asm/contigdata/$cat out.qdiff

NODE\_3\_length\_41351\_cov\_20.528098 GAP 26788 27499 712 0 712

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

~/desktop/asm/contigdata/$samtools faidx contigs.fasta NODE\_3\_length\_41351\_cov\_20.528098:'26788-27499' > final.fasta

~/desktop/asm/contigdata/$cat final.fasta

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.]**

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

~/desktop/asm/$cp dna-decode.py contigdata

~/desktop/asm/$cd contigdata

~/desktop/asm/contigdata/$python3 dna-decode.py -h

-d, --decode use this flag if you want to decode a message from a DNA Sequence

--input INPUT\_FM If --decode flag used, then this is the FASTA file which

has a single DNA sequence with the encoded message;

alternatively, if the --encode flag, this is a single line

text file with the message to encode in DNA sequence form

~/desktop/asm/contigdata/$python3 dna-decode.py -d --input final.fasta

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