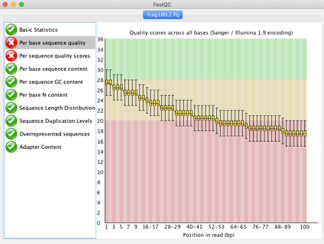
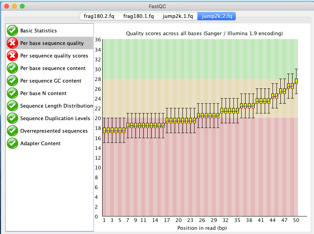
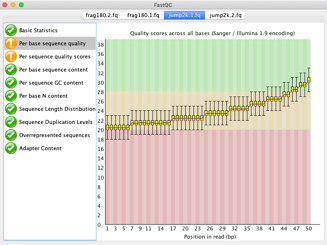
Question 1. Coverage Analysis

Download the reads and reference genome from: <https://github.com/bxlab/cmdb-lab/blob/gh-pages/2020/assignment1/asm.tgz?raw=true>

Note I have provided both paired-end and mate-pairs reads (see included README for details). Make sure to look at all of the reads for the coverage analysis and kmer analysis, as well as in the assembly.

* Question 1a. How long is the reference genome? [Hint: Try samtools faidx]
  + 233806
* Question 1b. How many reads are provided and how long are they? Make sure to measure each file separately [Hint: Try FastQC]
  + Frag180.2
    - Number of sequences: 35178
    - Sequence length: 100
  + Frag180.1
    - 35178
    - 100
  + jump2k.1
    - 70355
    - 50
  + jumpk2k.2
    - 70355
    - 50
* Question 1c. How much coverage do you expect to have? [Hint: A little arthmetic]
  + Total reads \* read length / genome
  + Frag180.2
    - 15x
  + Frag180.1
    - 15x
  + jump2k.1
    - 15x
  + jumpk2k.2
    - 15**x**
* Question 1d. Plot the average quality value across the length of the reads [We want a screenshot from FastQC]
  + Frag180.2

Question 2. Kmer Analysis

Use Jellyfish to count the 21-mers in the reads data. Make sure to use the “-C” flag to count cannonical kmers, otherwise your analysis will not correctly account for the fact that your reads come from either strand of DNA.

* Question 2a. How many kmers occur exactly 50 times? [Hint: try jellyfish histo]
  + 1091
* Question 2b. What are the top 10 most frequently occurring kmers [Hint: try jellyfish dump along with sort and head]

GCCCACTAATTAGTGGGCGCC 105

CGCCCACTAATTAGTGGGCGC 104

CCCACTAATTAGTGGGCGCCG 104

ACGGCGCCCACTAATTAGTGG 101

CAGGCCAGCTTATAAGCTGGC 98

AACAGGCCAGCTTATAAGCTG 98

ACAGGCCAGCTTATAAGCTGG 97

AGGCCAGCTTATAAGCTGGCC 95

AGCATCGCCCACATGTGGGCG 83

GCATCGCCCACATGTGGGCGA 82

* 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]
  + 233,468 bp
* Question 2d. How well does the GenomeScope genome size estimate compare to the reference genome? [Hint: In a sentence or two]
  + The GenomeScope length is slightly shorter than my reference sequence.
  + (233468- 233806)/233806 \* 100 = 0.14%

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]
  + 4
* Question 3b. What is the total length of the contigs? [Hint: try samtools faidx, plus a short script if necessary]
  + Total = NODE1 + NODE2 + NODE3 + NODE4 = 234467

NODE\_1\_length\_105831\_cov\_20.671371 105831 36 60 61

NODE\_2\_length\_47861\_cov\_20.231319 47861 107666 60 61

NODE\_3\_length\_41352\_cov\_20.588756 41352 156360 60 61

NODE\_4\_length\_39423\_cov\_20.384723 39423 198437 60 61

* Question 3c. What is the size of your largest contig? [Hint: check samtools faidx plus sort -n]
  + NODE\_1\_length\_105831\_cov\_20.671371
* Question 3d. What is the contig N50 size? [Hint: Write a short script if necessary]
  + N50 = 47861
    - 234469/2 =11724.5
    - Base pair 11724.4 would be in NODE\_2\_length\_47861\_cov\_20.231319

Question 4. Whole Genome Alignment

Use MUMmer for whole genome alignment.

* Question 4a. What is the average identity of your assembly compared to the reference? [Hint: try dnadiff]
  + 100%
* [Alignments]
* 1-to-1 5 5
* TotalLength 233755 233755
* AvgLength 46751.00 46751.00
* AvgIdentity 100.00 100.00
* Question 4b. What is the length of the longest alignment [Hint: try nucmer and show-coords]
  + 105831
* Question 4c. How many insertions and deletions are in the assembly? [Hint: try dnadiff]
  + There are 1 insertion in the query
  + There’s 5 insertions in the reference
* Insertions 1 5
* InsertionSum 712 51
* InsertionAvg 712.00 10.20

Question 5. Decoding the insertion

We need you to wget an updated script. Please wget https://raw.githubusercontent.com/bxlab/qbb2020/master/week1/ported\_decoder.py now.

* Question 5a. What is the position of the insertion in your assembly? Provide the corresponding position in the reference. [Hint: try show-coords]
  + 26789
* [S1] [E1] | [S2] [E2] | [LEN 1] [LEN 2] | [% IDY] | [LEN R] [LEN Q] | [TAGS]
* ==========================================================================================================
* 1 105831 | 233795 127965 | 105831 105831 | 99.99 | 105831 233806 | NODE\_1\_length\_105831\_cov\_20.671371 Halomonas
* 1 39423 | 127954 88532 | 39423 39423 | 100.00 | 39423 233806 | NODE\_4\_length\_39423\_cov\_20.384723 Halomonas
* 1 47861 | 88511 40651 | 47861 47861 | 100.00 | 47861 233806 | NODE\_2\_length\_47861\_cov\_20.231319 Halomonas
* 1 13853 | 40642 26790 | 13853 13853 | 100.00 | 41352 233806 | NODE\_3\_length\_41352\_cov\_20.588756 Halomonas
* 14566 41352 | 26789 3 | 26787 26787 | 100.00 | 41352 233806 | NODE\_3\_length\_41352\_cov\_20.588756 Halomonas
* Question 5b. How long is the novel insertion? [Hint: try show-coords]
  + 14565-13853 =712
* Question 5c. What is the DNA sequence of the encoded message? [Hint: try samtools faidx to extract the insertion]
  + (see desired\_seq.txt)
* Question 5d. What is the secret message? [Hint: Run the provided script ported\_decoder.py to decode the string from 5c.]
  + (see secret\_message.txt)