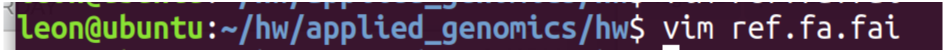
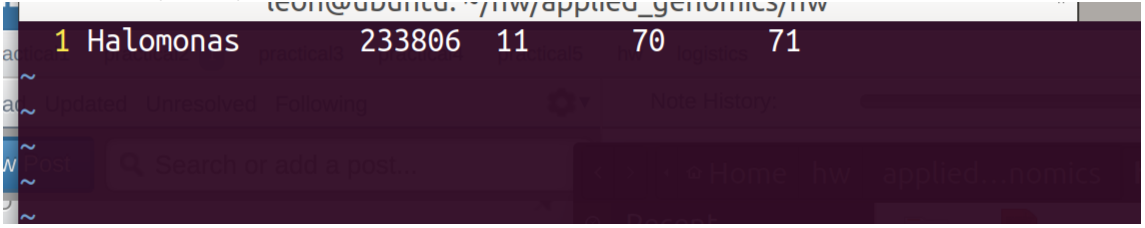
- Question 1a. How long is the reference genome? [Hint: Try `samtools faidx`]

After samtools faidx ref.fa, we could get a file with .fai extension. Try following command:



```
leon@ubuntu:~/hw/applied_genomics/hw$ vim ref.fa.fai
```

The total length of genome is 233806 in bases



```
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`]

Use the commands shown on github:
$ fastqc /path/to/reads.fq
For frag180.1.fq, there are 35217 100bp reads.
For frag180.2.fq, there are 35217 100bp reads.
For jump2k.1.fq, there are 70435 50bp reads.
For jump2k.2.fq, there are 70435 50bp reads.

Basic Statistics

Measure	Value
Filename	frag180.1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	35217
Sequences flagged as poor quality	0
Sequence length	100
%GC	54

Basic Statistics

Measure	Value
Filename	frag180.2.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	35217
Sequences flagged as poor quality	0
Sequence length	100
%GC	54

Basic Statistics

Measure	Value
Filename	jump2k.1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	70435
Sequences flagged as poor quality	0
Sequence length	50
%GC	54

Basic Statistics

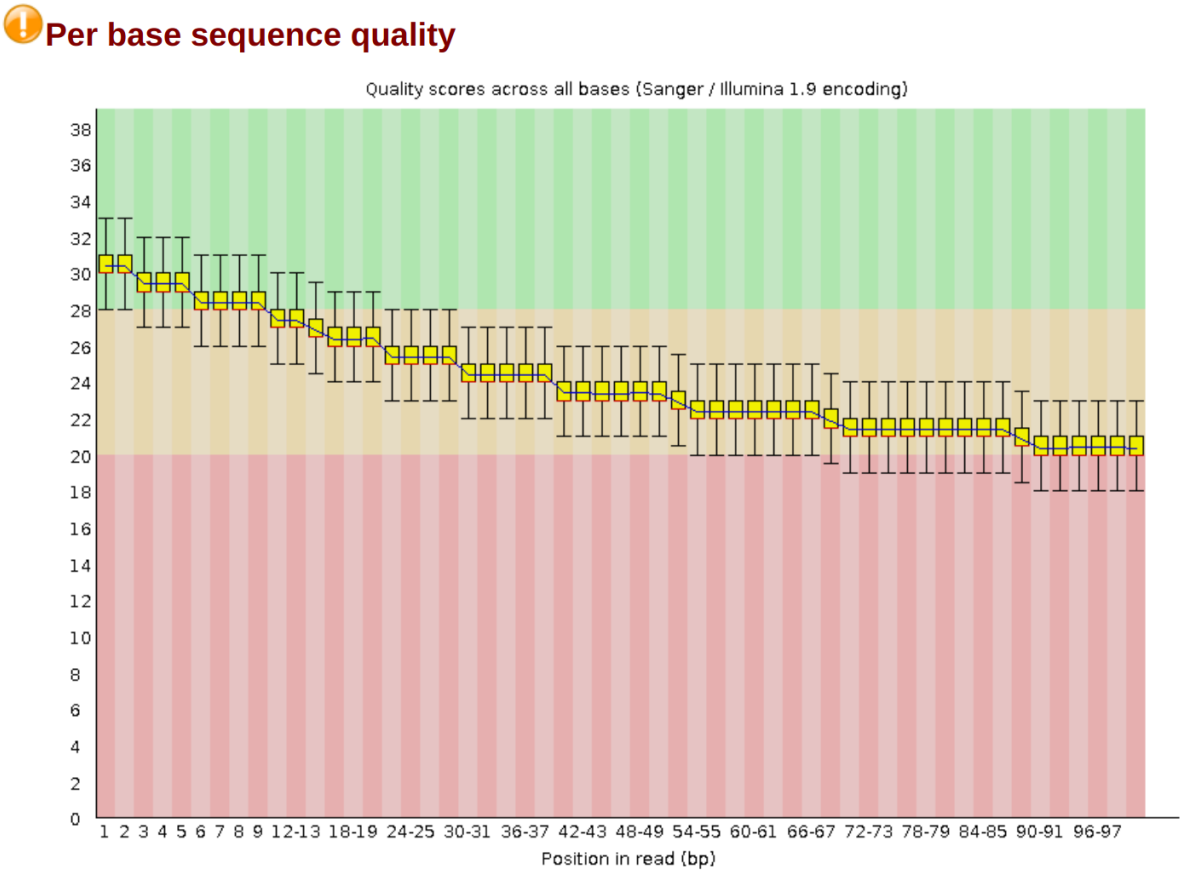
Measure	Value
Filename	jump2k.2.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	70435
Sequences flagged as poor quality	0
Sequence length	50
%GC	54

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

We expect 15x coverage.((length of reads * number of reads)/ total genome size)

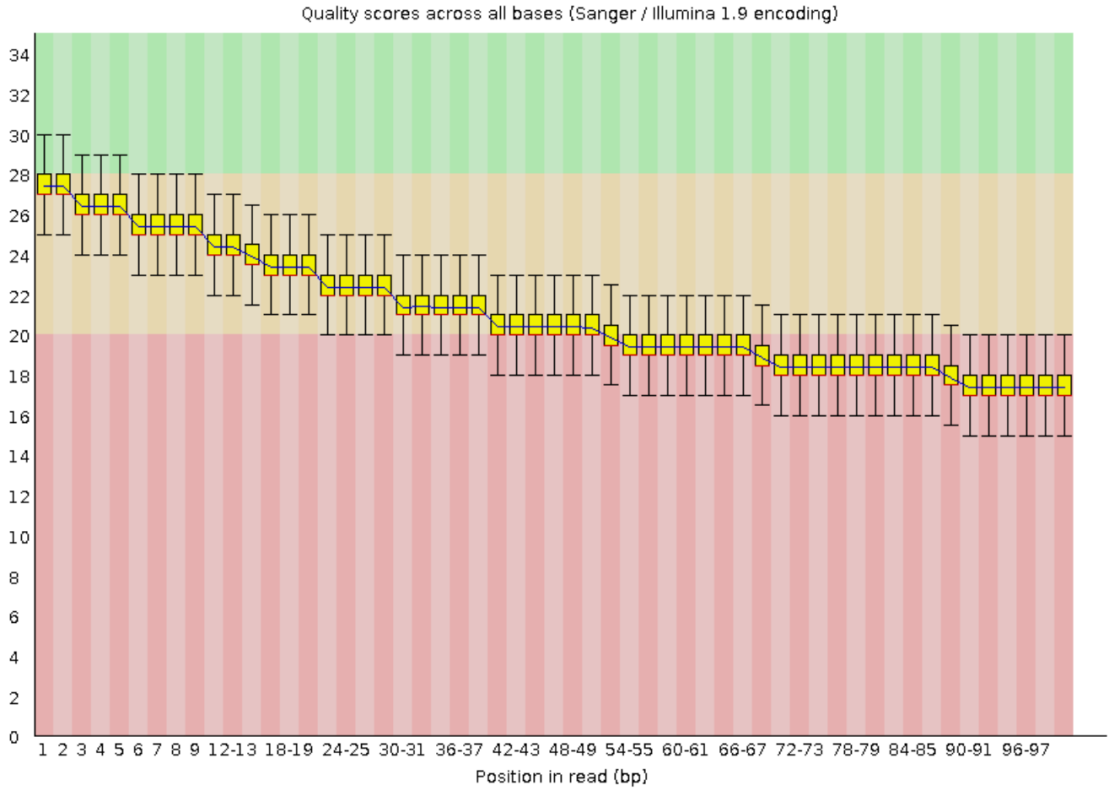
- Question 1d. Plot the average quality value across the length of the reads [Hint: Screenshot from `FastQC`]

For frag180.1:



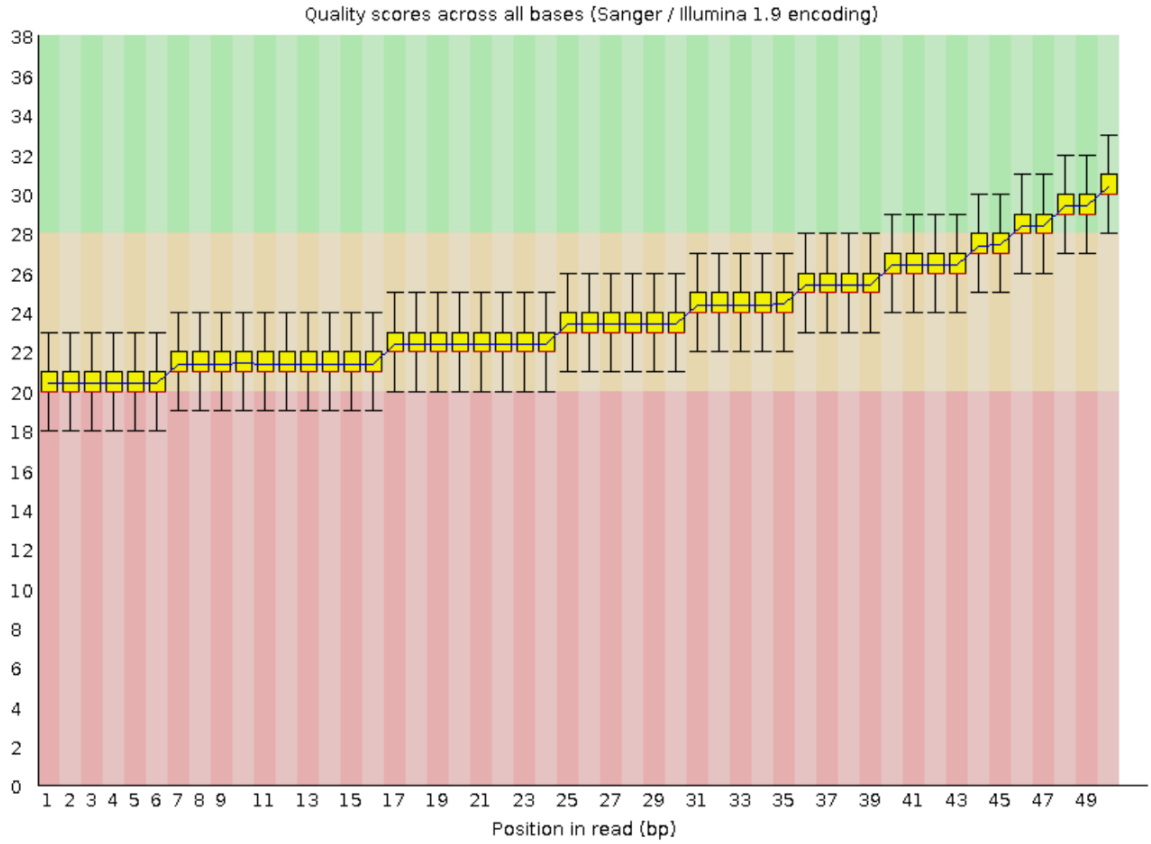
For frag180.2:

Per base sequence quality



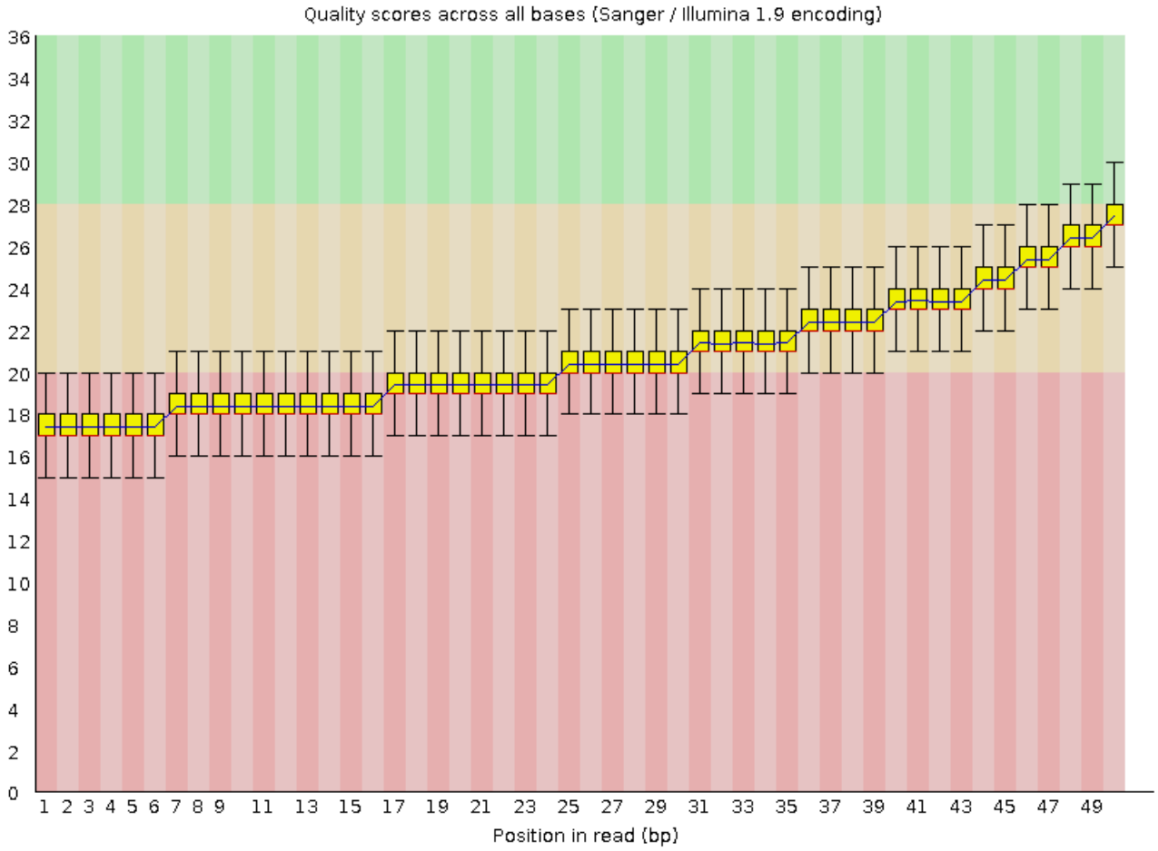
For jump2k.1:

Per base sequence quality



For jump2k.2:

Per base sequence quality

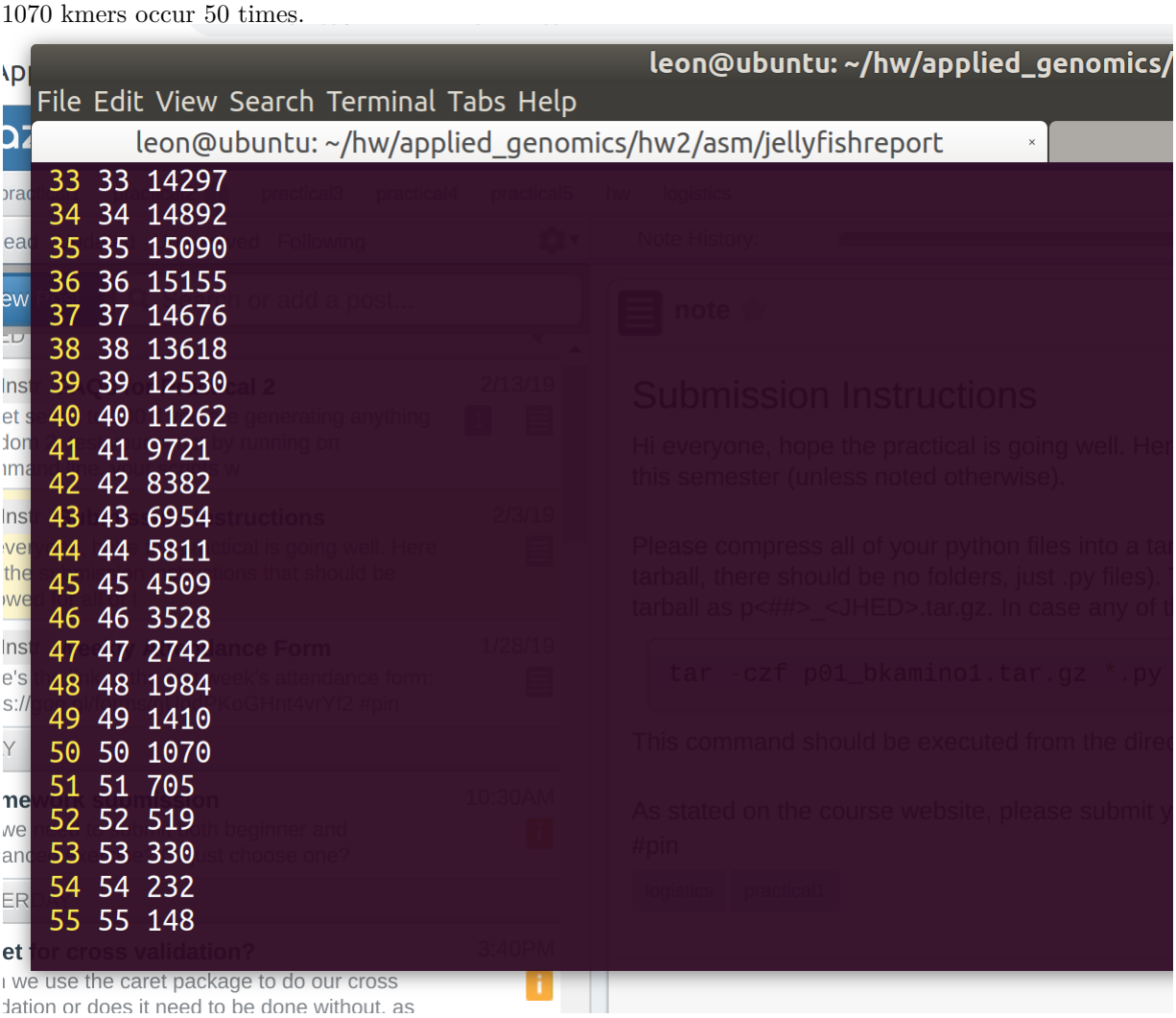


- Question 2a. How many kmers occur exactly 50 times? [Hint: try jellyfish histo]

After following codes:

$ jellyfish count -m 21 -C -s 1000000 *.fq
$ jellyfish histo mer_counts.jf > reads.histo
$ vim reads.histo

1070 kmers occur 50 times.



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

Though following commands, we could get the top 10 most frequently occurring kmers.

```
leon@ubuntu:~/hw/applied_genomics/hw2/asm/jellyfishreport$ jellyfish dump -c mer_counts.jf | sort -n -r -k 2 -t " "|head -n 10
GCATCGCCCACATGTGGGCGA 84
AGCATCGCCCACATGTGGGCG 83
CAAAACGGCCCCTTAAGGGGC 82
AAAACGGCCCCTTAAGGGGCC 81
AACGGCCCCTTAAGGGGCCGT 80
AAACGGCCCCTTAAGGGGCCG 80
ATCGCCCACATGTGGGCGATG 79
AGGCCAGCTTATAAGCTGGCC 75
AATTGAACCTGCGACCTTCGG 75
CGCCCACTAATTAGTGGGCGC 73
```

- Question 2c. What is the estimated genome size based on the kmer frequencies?

The min estimated Genome Size is 233876 bp. It probably has something to do with kmer coverage.

Results

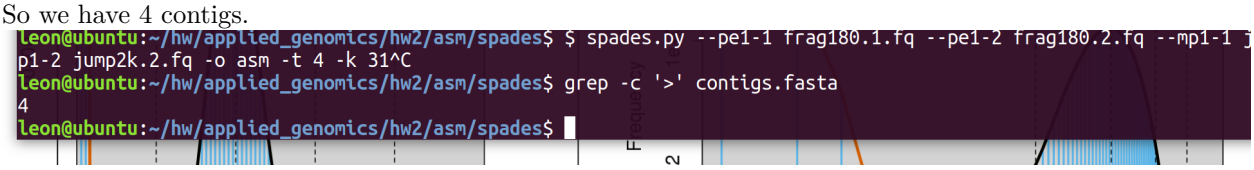
```
GenomeScope version 1.0
k = 21

property                      min               max
Heterozygosity                -0.00147547%      0.0136271%
Genome Haploid Length         233,876 bp        234,164 bp
Genome Repeat Length          -576 bp           -577 bp
Genome Unique Length          234,453 bp        234,742 bp
Model Fit                     98.9817%          NA%
Read Error Rate               0.800338%         0.800338%
```

- 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 only 70bp longer than the reference genome. It does pretty well.

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

After running the Spades command:
$ 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
We could find the number of contigs:
grep -c '>' contigs.fasta
So we have 4 contigs.



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

After samtools faidx contigs.fa, we could use following command to get the total length of contigs:
awk '{print $2}' contigs.fasta.fai|paste -s -d +|bc

```
File Edit View Search Terminal Tabs Help
leon@ubuntu: ~/hw/applied_genomics/hw
#!/bin/bash
awk '{print $2}' contigs.fasta.fai|paste -s -d +|bc
sort -n -k 2 -t \t  contigs.fasta.fai
```

We could get the total length is 234743

```
leon@ubuntu:~/hw/applied_genomics/hw/asm$ ./q3
234743
NODE_1_length_105841_cov_20.480749      105841  36      60      61
NODE_2_length_47856_cov_20.556299       47856   107677  60      61
NODE_3_length_41610_cov_20.667909       41610   156366  60      61
NODE_4_length_39436_cov_20.425707       39436   198705  60      61
leon@ubuntu:~/hw/applied_genomics/hw/asm$
```

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

As shown in Question 3b's bash file:
sort -n -k 2 -t ĉontigs.fasta.fai
We could get the largest contig is 105841

```
leon@ubuntu:~/hw/applied_genomics/hw/asm$ ./q3
234743
NODE_1_length_105841_cov_20.480749	105841	36	60	61
NODE_2_length_47856_cov_20.556299	47856	107677	60	61
NODE_3_length_41610_cov_20.667909	41610	156366	60	61
NODE_4_length_39436_cov_20.425707	39436	198705	60	61
leon@ubuntu:~/hw/applied_genomics/hw/asm$
```

- Question 3d. What is the contig N50 size? [Hint: Write a short script, or use excel]

Use the codes as shown below:

```
#!/bin/bash
awk '{print $2}' contigs.fasta.fai|paste -s -d +|bc
sort -n -k 2 -t \t  contigs.fasta.fai

sort -n contigs.fasta.fai|cut -f2 | awk '{len[i++]=$1;sum+=$1} END {for (j=0;j<i+1;j++) {csum+=len[j]; if (csum>sum/2) {print len[j];break}}}'
```

According to the defination of contig N50, the size of contig N50 should be 47856.

```
leon@ubuntu:~/hw/applied_genomics/hw/asm$ ./q3
234743
NODE_1_length_105841_cov_20.480749      105841  36      60      61
NODE_2_length_47856_cov_20.556299       47856   107677  60      61
NODE_3_length_41610_cov_20.667909       41610   156366  60      61
NODE_4_length_39436_cov_20.425707       39436   198705  60      61
47856
leon@ubuntu:~/hw/applied_genomics/hw/asm$
```

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

Try following commands:
$ dnadiff ref.fa contigs.fasta

```
leon@ubuntu:~/hw/applied_genomics/hw/dnadiff$ ll
total 48
drwxr-xr-x 2 leon leon 4096 Feb 13 16:03 ./
drwxr-xr-x 5 leon leon 4096 Feb 13 18:40 ../
-rw-r--r-- 1 leon leon  549 Feb 13 15:29 .1coords
-rw-r--r-- 1 leon leon  478 Feb 13 15:29 .1delta
-rw-r--r-- 1 leon leon  478 Feb 13 15:29 .delta
-rw-r--r-- 1 leon leon  549 Feb 13 15:29 .mcoords
-rw-r--r-- 1 leon leon  478 Feb 13 15:29 .mdelta
-rw-r--r-- 1 leon leon   60 Feb 13 15:29 .qdiff
-rw-r--r-- 1 leon leon  380 Feb 13 15:29 .rdiff
-rw-r--r-- 1 leon leon 4173 Feb 13 15:29 .report
-rw-r--r-- 1 leon leon   92 Feb 13 15:29 .snps
leon@ubuntu:~/hw/applied_genomics/hw/dnadiff$ vim .report
```

we could see average identity of the assembly is 100:

```
leon@ubuntu: ~/hw/applied_genomics/hw/dnadiff
UnalignedBases          39(0.02%)          976(0.42%)

[Alignments]
1-to-1                          5                  5
TotalLength                233767             233767
AvgLength                46753.40           46753.40
AvgIdentity                100.00             100.00

M-to-M                          5                  5
TotalLength                233767             233767
AvgLength                46753.40           46753.40
AvgIdentity                100.00             100.00

[Feature Estimates]
Breakpoints                    10                  2
Relocations                     0                  0
Translocations                  3                  0
Inversions                      0                  0

Insertions                      5                  1
InsertionSum                   39                976
InsertionAvg                 7.80             976.00
```

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

After trying the commands shown on github:
$ nucmer /path/to/ref.fa /path/to/qry.fa
Through following commands:

```
leon@ubuntu:~/hw/applied_genomics/hw/nucmer$ ll
total 12
drwxr-xr-x 2 leon leon 4096 Feb 12 00:51 ./
drwxr-xr-x 5 leon leon 4096 Feb 13 18:40 ../
-rw-r--r-- 1 leon leon  478 Feb 12 00:51 .delta
leon@ubuntu:~/hw/applied_genomics/hw/nucmer$ show-coords -r .delta
```

We could see the longest alignment's length is 105841:

```
leon@ubuntu:~/hw/applied_genomics/hw/nucmer$ show-coords -r .delta
/home/leon/hw/applied_genomics/hw/ref.fa /home/leon/hw/applied_genomics/hw/asm/contigs.fasta
NUCMER

    [S1]     [E1]  |     [S2]     [E2]  |  [LEN 1]  [LEN 2]  |  [% IDY]  | [TAGS]
==========================================================================================
       4    26789  |        1    26786  |    26786    26786  |    100.00  | Halomonas    NODE_3_length_41610_cov_20.667909
   26790    40637  |    27763    41610  |    13848    13848  |    100.00  | Halomonas    NODE_3_length_41610_cov_20.667909
   40654    88509  |        1    47856  |    47856    47856  |    100.00  | Halomonas    NODE_2_length_47856_cov_20.556299
   88516   127951  |        1    39436  |    39436    39436  |    100.00  | Halomonas    NODE_4_length_39436_cov_20.425707
  127963   233803  |        1   105841  |   105841   105841  |     99.99  | Halomonas    NODE_1_length_105841_cov_20.480749
leon@ubuntu:~/hw/applied_genomics/hw/nucmer$
```

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

After trying commands shown on github:
$ dnadiff /path/to/ref.fa /path/to/qry.fa
Through following command:

```
leon@ubuntu:~/hw/applied_genomics/hw/dnadiff$ ll
total 48
drwxr-xr-x 2 leon leon 4096 Feb 13 20:21 ./
drwxr-xr-x 5 leon leon 4096 Feb 13 18:40 ../
-rw-r--r-- 1 leon leon  549 Feb 13 15:29 .1coords
-rw-r--r-- 1 leon leon  478 Feb 13 15:29 .1delta
-rw-r--r-- 1 leon leon  478 Feb 13 15:29 .delta
-rw-r--r-- 1 leon leon  549 Feb 13 15:29 .mcoords
-rw-r--r-- 1 leon leon  478 Feb 13 15:29 .mdelta
-rw-r--r-- 1 leon leon   60 Feb 13 15:29 .qdiff
-rw-r--r-- 1 leon leon  380 Feb 13 15:29 .rdiff
-rw-r--r-- 1 leon leon 4173 Feb 13 15:29 .report
-rw-r--r-- 1 leon leon   92 Feb 13 15:29 .snps
leon@ubuntu:~/hw/applied_genomics/hw/dnadiff$ vim .report
```

We could find there are 5 insertions in the reference 1 insertion in the assembly, which means 5 deletions in the assembly.

```
leon@ubuntu: ~/hw/applied_genomics/hw/dnadiff
AvgLength                46753.40            46753.40
AvgIdentity                100.00              100.00

M-to-M                          5                   5
TotalLength                233767              233767
AvgLength                46753.40            46753.40
AvgIdentity                100.00              100.00

[Feature Estimates]
Breakpoints                    10                   2
Relocations                     0                   0
Translocations                  3                   0
Inversions                      0                   0

Insertions                      5                   1
InsertionSum                   39                 976
InsertionAvg                 7.80              976.00

TandemIns                       0                   0
TandemInsSum                    0                   0
TandemInsAvg                 0.00                0.00

[SNPs]
```

- Question 5a. What is the position of the insertion on the reference? [Hint: try show-coords]

As shown in snapshot:

```
leon@ubuntu:~/hw/applied_genomics/hw/dnadiff$ show-coords -r .delta
/home/leon/hw/applied_genomics/hw/ref.fa /home/leon/hw/applied_genomics/hw/asm/contigs.fasta
NUCMER

    [S1]     [E1]  |     [S2]     [E2]  |  [LEN 1]  [LEN 2]  |  [% IDY]  | [TAGS]
==========================================================================
       4    26789  |        1    26786  |    26786    26786  |    100.00  | Halomonas    NODE_3_length_41610_cov_20.667909
   26790    40637  |    27763    41610  |    13848    13848  |    100.00  | Halomonas    NODE_3_length_41610_cov_20.667909
   40654    88509  |        1    47856  |    47856    47856  |    100.00  | Halomonas    NODE_2_length_47856_cov_20.556299
   88516   127951  |        1    39436  |    39436    39436  |    100.00  | Halomonas    NODE_4_length_39436_cov_20.425707
  127963   233803  |        1   105841  |   105841   105841  |     99.99  | Halomonas    NODE_1_length_105841_cov_20.480749
```

We could infer the insertion happens on the position of bases: 26789-26790

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

```
leon@ubuntu:~/hw/applied_genomics/hw/dnadiff$ show-coords -r .delta
/home/leon/hw/applied_genomics/hw/ref.fa /home/leon/hw/applied_genomics/hw/asm/contigs.fasta
NUCMER

    [S1]     [E1]  |     [S2]     [E2]  |  [LEN 1]  [LEN 2]  |  [% IDY]  | [TAGS]
==========================================================================================
       4    26789  |        1    26786  |    26786    26786  |   100.00  | Halomonas    NODE_3_length_41610_cov_20.667909
   26790    40637  |    27763    41610  |    13848    13848  |   100.00  | Halomonas    NODE_3_length_41610_cov_20.667909
   40654    88509  |        1    47856  |    47856    47856  |   100.00  | Halomonas    NODE_2_length_47856_cov_20.556299
   88516   127951  |        1    39436  |    39436    39436  |   100.00  | Halomonas    NODE_4_length_39436_cov_20.425707
  127963   233803  |        1   105841  |   105841   105841  |    99.99  | Halomonas    NODE_1_length_105841_cov_20.480749
```

We could see the length is (27763 - 26786 -1) = 976

- Question 5c. What is the DNA sequence of the encoded message? [Hint: try samtools faidx to extract the insertion]
  Through the commands shown below, we could get the sequence:

```
leon@ubuntu:~/hw/applied_genomics/hw/asm$ samtools faidx contigs.fasta NODE_3_length_41610_cov_20.667909:26787-27762
>NODE_3_length_41610_cov_20.667909:26787-27762
CTAAACGGATGATGTGATGCTTGCCTGCCGGGCTGTAATCAGGCACCGATTTAGCCCGTT
CGCGCTTATGACCGGACACGCTTGAGCCATGATACTAGTATGTTAGTCGTTCATTTCCTT
AATCACCCATGTCTAACGGCTGTTCCTACCAACTGGAGCCCGTCTCCAATGACGCTAATA
CCACAGCGCAGTATGTAAACCGGGCCGAAGGCGACGATTCTTGCCTGAAGGTCATAAAAC
ACTGCATAAATGCCCACAGGCACGACTTGCCGAAGAAACCCTCCGCGACTAAGACCCTCT
ATAGCCTAAAAACTAAACATCTTTCCCCAGTTCACCATTCTTACCGGATACGAGGAATAT
CTGACTCCACGACAACAGACCTGGATGACTCTATTCTGTCATTCTGGGCTTCTGAGAGTC
GCCGCTGACATTCGTGACGTCATACACACTTCCATGCGTCTTGTCGGGCGAGAGGTCCTA
CGTTACTGCGGACTCGCATCGTTCAAGCTGGAACGAGGTCCCTACAAACAGACCCAAGAT
TCAGCGGATGGTCTGTCGCTCCGCACAAAGGGAGGTCGGCTACTCGTCTGAACGGCTGAA
CATCCACAATTATACCAGTCCCCGATGTAGGCAGTCCTAGCCTACACCCGGGCGCAATGA
GGTTAAGACAAACGGAGAGTCGTACGCTCTGACGCTAGGTACCCCCGAACAACTGGCCTT
CGTCCTCGCTTCCCATAGGTAAGAAGGCACGGCGGAGCCCCTTCGCCTCGGAGTTACTGA
CGTTCAGAACCCATTACGGAATGCTGGTAGGGCCTCACTAAAACCGGCACAGAGTTACGA
ATTTCCGTATGCATCTCTTCTTGCAGGCGAATAGGACAGGACTAACCCCGTAACGCCTGC
AACTCTGTACTGCTGAATCGATGCATCAAAGAACACCTGATTAAAGGATCATATTCATGA
CTTAATCTAATCAACG
leon@ubuntu:~/hw/applied_genomics/hw/asm$
```

- Question 5d. What is the secret message? [Hint: run dna-encode.pl -d to decode the string from 5c]

```
leon@ubuntu:~/hw/applied_genomics/hw/asm$ samtools faidx contigs.fasta NODE_3_length_41610_cov_20.667909:26787-27762|./dna-enco
de.pl -d
>NODE_3_length_41610_cov_20.667909:26787-27762
Congratulations to the Spring 2019 JHU Applied Genomics course...  You will have to keep searching for arsenic based life!
leon@ubuntu:~/hw/applied_genomics/hw/asm$
```