# COMP 3353 Assignment 2

## Name: Chan Tsai Hor

## UID: 3035376355

## Email: [howardchanth@gmail.com](mailto:howardchanth@gmail.com)

## Question 1

cat ERR2337147\_1.fastq | awk 'NR%4 == 2 {print}' | wc -l

184803

cat ERR2337147\_2.fastq | awk 'NR%4 == 2 {print}' | wc -l

184803

There are 184803 sequences resulted from this run.

## Question 2

cat ERR2337147\_1.fastq | awk 'NR%4 == 2 {print}' | wc -c

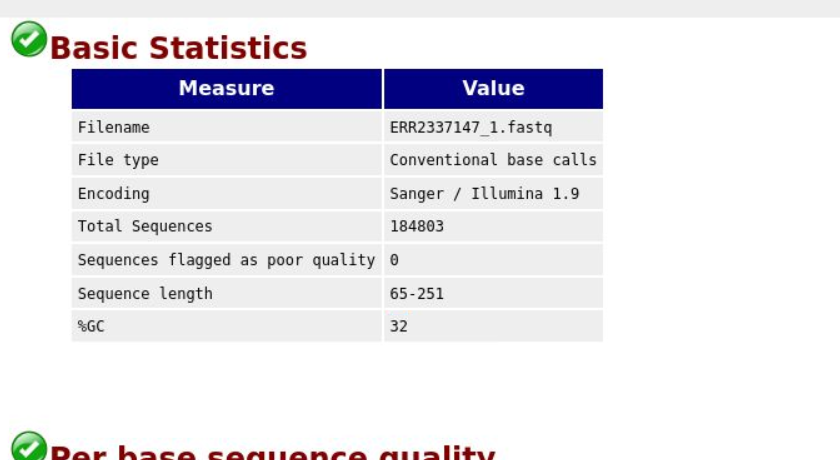
46118961

cat ERR2337147\_2.fastq | awk 'NR%4 == 2 {print}' | wc -c

46118961

There are 46118961 nucleotides sequenced for this run

## Question 3



The GC content is **32%**

## Question 4

seqtk fqchk -q0 ERR2337147\_1.fastq | awk '{print $8}' | head

distinct

avgQ

**37.9**

33.4

33.7

33.8

33.8

33.7

36.3

seqtk fqchk -q0 ERR2337147\_1.fastq | awk '{print $8}' | tail -1

**31.9**

The average Phred quality score is larger for the 1st position than the last position. AS the DNA is not stiff and could easily be affected by the external factors (gravity). The last sequenced position usually subject to more influences, which will subsequently lower its quality.

## Question 5

cat ERR2337147\_1.fastq | awk 'NR%4 == 2 {print}' | awk ' {print length($1)}' | uniq | head

251

250

251

250

251

250

248

250

251

248

We can see that there are sequences of different lengths, hence we can claim that the length of every sequence in the FASTQ files is not the same.

## Question 6

The script of computation is in Q6.py

Sample Input: (input.txt)

GAGCCTACTAACGGGAT

CATCGTAATGACGGCCT

Sample Output:

7

## Question 7

The script of computation is in Q7.py

Input:

* input.txt: Storing the input protein string (sample: SKADYEK)
* amino\_acid\_mass.csv: Storing the mapping of monoisotopic mass.

Output:

* The total weight of PP: 821.39192

## Question 8

The script of computation is in Q8.py