**Question 1**

**Question 2**

**Question 3**

**Part b**

grep -v ">" chr22.fa | grep "AAGCTT" | wc -l

4128

**Part c**

grep -v ">" chr22.fa | grep "AAGCTT\|AGGCTT\|ATGCTT\|ACGCTT" | wc -l

14800

**Part d**

**Part e**

First we design the input pattern file containing the patterns of ATTCCGAATCAGGGT allowing one error:

cat ATTCCGAATCAGGGT.txt

.TTCCGAATCAGGGT

A.TCCGAATCAGGGT

AT.CCGAATCAGGGT

ATT.CGAATCAGGGT

ATTC.GAATCAGGGT

ATTCC.AATCAGGGT

ATTCCG.ATCAGGGT

ATTCCGA.TCAGGGT

ATTCCGAA.CAGGGT

ATTCCGAAT.AGGGT

ATTCCGAATC.GGGT

ATTCCGAATCA.GGT

ATTCCGAATCAG.GT

ATTCCGAATCAGG.T

Then we search the FASTA file for the desired pattern by the following Linux command:

grep -f ATTCCGAATCAGGGT.txt chr22.fa | wc -l

where ATTCCGAATCAGGGT.txt is the text file solving the pattern of ATTCCGAATCAGGGT with one mismatch tolerance.

**Part f**

cat chr22.fa | tr "actg" "ACTG" | grep -f <(cat chr22.fa | tr "actg" "ACTG" | grep -oE "(CAG)+" | sort | tail -1 | head)

**CACAGCTGCAGCTCCAACAACAGCAACAGCAGCAGCAGCAGCAGCAGCAG**

**TGGCTCGAATGGTGAGTGCACTGCAGCAGCAGCAGCAGCAGCAGCAGAGG**

The longest consecutive simple repeat of CAG is **8** “CAG” s.

**Part g**

grep -v ">" chr22.fa | grep -o -n "N" | uniq -c | wc -l

233222

Excluding the “N” strings at the very beginning and the very end, the total number of gaps is:

233222 – 2 = 233220

**Question 4**

**Task 1: Compute length-k substrings of s are high-GC**

The number of high-GC k-length strings is **378292**