Homework

1. Create a directory named “hw2".

2. Move to “hw2" directory, and download this file (SRR25653406.fasta.tar.gz) using wget command:

https://github.com/UeenHuynh/MGMA_2024/raw/main/lecture2/2.%20basic%20Linux%20commands%20part%202/SRR25653406.fasta.tar.gz

3. Decompressing this file SRR25653406.fasta.tar.gz using tar command.

Additional information about FASTA format (The file has the extension .fasta, .faa, .fna, etc.)

FASTA format is a text-based format, that contains two lines:

- First line: is the comment (description) line.
  + Always start with the “>” sign (This information is useful for the exercise).
  + Give basic information about the sequence (nucleotide or amino acid).
- Second line: The actual sequence of the first line description, using a standard one-letter character string.

Example:

```
>M35309.1 E.coli 16S rRNA fragment
GGCATGAAGACACACTGCTAACTCCGAATACGCACAAGCCCGTAATGGAGCGACGGTGGGCCTTGTTCCC
GTGCCCCGATGTGGGGTGGAGGTGACTGTGGGTTGTGATATTCGGGGAGGCAAAAGAAGTAGCGAGTCTA
ACCTTGCTTACCACTTTGCCTAATACGGGAAACG
```

(https://www.ncbi.nlm.nih.gov/nuccore/M35309.1?report=fasta)

=> Let's call this single-sequence FASTA format (https://en.wikipedia.org/wiki/FASTA_format), so if a text file contains a single-sequence FASTA format, this file is called a single-sequence FASTA file.

If a text file contains two or more single-sequence FASTA format, this file is called a multiple-sequence FASTA file or multi-FASTA file.

An example of the multi-FASTA format :

(https://www.researchgate.net/figure/A-sample-of-the-Multi-FASTA-file_fig1_309134977)

Additional information about this file: SRR25653406.fasta (after decompressing using tar command).

- This is a multi-FASTA file. It is converted from fastq to fasta format file (Don't worry about fastq format in this exercise), and the original fastq format file is obtained from the sequencing machine.
- Each single-sequence FASTA in this file is information about a read - a piece of DNA sequence obtained from the sequencing machine:
    + First line: some information about the sequencing platform of this file, that's all, don't worry about the details.
    + Second line: DNA sequence of that read, a character string consisting of 4 types of letters: A, T, G, C representing 4 types of nucleotides. In addition, there is the letter N, which means 1 of 4 possible types of nucleotides, because the sequencing machine cannot identify which type of nucleotide it is.

A single-sequence FASTA in this file:

```
>SRR25653406.lite.3.8 M02606:41:000000000-J3RLF:1:1101:8131:2344 length=200
TTTGGTCCAGCAGGCTATTGCTGGGAATTGTTAGCTTTTCGGGATCCATTTTATTGGGCGGTGTGCTCATCTCTACCATTTCCAACATTATAGAGCGGCGTGTGG
ATGTGGTAAACACGGGGCGAATGACTTACCGCAACATAACTCAACACTATGTATTGATAGGCTTCAATGAACTAACTATCAATATGATACGCGAA
```

(The line containing the DNA sequence has 2 lines because the screen is not enough to display 1 line, so it is split into 2 lines).

4. How many lines does this file have? Using: wc.

5. How many reads does this file have? Using: grep for the search “^>” means the line starts with the character ">", and -c option.

6. How many reads do not have any N? Using: grep with -v, -c option; pipe.

7. Create a file named "id_read.txt" containing only the first line (the line contains information) of all reads. Using: grep with “^v"; output redirection.

8. Create a file named "part_of_id_read.txt" containing a part of the first line (the line contains information) of all reads. Using: grep; pipe; cut; output redirection.

Two lines of 1 read:

```
>SRR25653406.lite.3.8 M02606:41:000000000-J3RLF:1:1101:8131:2344 length=200
TTTGGTCCAGCAGGCTATTGCTGGGAATTGTTAGCTTTTCGGGATCCATTTTATTGGGCGGTGTGCTCATCTCTACCATTTCCAACATTATAGAGCGGCGTGTGG
ATGTGGTAAACACGGGGCGAATGACTTACCGCAACATAACTCAACACTATGTATTGATAGGCTTCAATGAACTAACTATCAATATGATACGCGAA
```

(The line containing the DNA sequence has 2 lines because the screen is not enough to display 1 line, so it is split into 2 lines).

a part of the first line of 1 reads:

```
M02606:41:000000000-J3RLF:1:1101:8131:2344
```

Answers will be uploaded to GitHub, on 25/05/2024