HORT 59000
Lab 4 Exercises

1. Copy the following files from /home/kvarala/Files/ to your working directory.
    a. North_of_Boston.txt
    b. GSE49418_series_matrix.txt
    c. SRR6473489.fastq
    d. At_Promoters_1KB.fasta
    e. bZIP1_TargetIDs.txt

```
$ cp /home/kvarala/Files/North_of_Boston.txt /scratch/scholar/kvarala/IDAB/Week4/

$ cp /home/kvarala/Files/ GSE49418_series_matrix.txt
/scratch/scholar/kvarala/IDAB/Week4/

$ cp /home/kvarala/Files/ SRR6473489.fastq /scratch/scholar/kvarala/IDAB/Week4/

$ cp /home/kvarala/Files/At_Promoters_1KB.fasta /scratch/scholar/kvarala/IDAB/Week4/

$ cp /home/kvarala/Files/bZIP1_TargetIDs.txt /scratch/scholar/kvarala/IDAB/Week4/
```

2. Count the number of lines in each file.

```
$ cd /scratch/scholar/kvarala/IDAB/Week4/
$ wc -l *
```

North of Boston

3. Find all lines in North_of_Boston.txt that start with the word ‘I’
   NOTE: The grep command should be invoked with the `-E` option to allow use of regular expressions.

```
$ grep -E '^\s*I' North_of_Boston.txt
```

4. Replace all occurrences of the word ‘My’ with ‘Our’ in the file North_of_Boston.txt and save this output to the file New_Boston.txt
   NOTE: Use the sed command to replace ‘My’ with ‘Bob’

```
$ sed 's/\bMy\b/Our/' North_of_Boston.txt >New_Boston.txt
```

5. Retrieve the first four lines of the poem ‘Mending Wall’ that is in the file North_of_Boston.txt **HINT**: grep can return a given number of lines before (-B) or after (-A) the match.

```
$ grep -A 5 "^Mending Wall" North_of_Boston.txt
```

GSE49418_series_matrix.txt

6. Create two files: 1. Contains ONLY the comment lines from the GSE49418_series_matrix.txt file and 2. Contains ONLY the non comment lines from GSE49418_series_matrix.txt. **HINT**: All the comment lines start with !

```
$ grep -E '^!' GSE49418_series_matrix.txt >GSE49418_comments.txt

$ grep -vE '^!' GSE49418_series_matrix.txt >GSE49418_matrix.txt
```

7. Find all lines from GSE49418_series_matrix.txt where the gene expression in the first data column is >=10 **HINT**: Use the awk command to check the value of first column ($1) is >=10.

```
$ awk -F "\t" '($2>=10){print $0}' GSE49418_series_matrix.txt
```

SRR6473489.fastq (All retrieved sequences should include (ID, Sequence and quality values) HINT: use the -A and -B options with grep

8. Retrieve all sequences that start with a digit in the quality line

```
$ grep -B 3 '^[0-9]' SRR6473849.fastq
```

9. Retrieve all sequences that contain the sequence GTCGATCTCTCTCGCTCTC

```
$ grep -A2 -B1 GTCGATCTCTCTCGCTCTC SRR6473849.fastq
```

At_Promoters_1KB.fasta contains the 1Kb sequence immediately upstream of every gene in the *Arabidopsis thaliana* genome. The file **bZIP1_TargetIDs.txt** lists the 470 genes up-regulated by the bZIP1 transcription factor Para et. Al., 2014 (PNAS July 15, 2014 111 (28) 10371-10376;). The bZIP1 TF is known to bind a DNA Motif: G[AC]CACGT Using the grep and sed tools identify how many of the 470 genes listed in bZIP1_TargetIDs.txt contain the motif G[AC]CACGT in their promoter.

Here are the steps to achieve this:
1. Use grep to extract the promoter sequences for these 470 genes from the full set of promoters. You can give list a file of IDs to search for using the -f option. Since each promoter sequence is 1000 bases and there are 80 bases per line, each promoter sequence is spread over 1000/80 = 12.5 lines. So, after each gene ID the next 13 lines contain the promoter of that gene. Redirect these lines to a file (Step1.file).

```
$ grep -A 13 -f bZIP1_TargetIDs.txt At_Promoters_1KB.fasta >Step1.file
```

2. When searching for multiple IDs from the file, grep will insert a line containing only '--' between each match. So, use grep again to remove all lines that do not contain an alphabet from the file generated in step 1 and redirect to a new file (Step2.file).

```
$ grep [[:alnum:]] Step1.file >Step2.file
```

3. Step2.file should contain 470 promoter IDs and their sequences. But, the sequence is spread over multiple lines arbitrarily split into 80 characters. If, the motif occurs by chance at the end of the line it will be hard to find. So, we want all of the promoter sequence in one line. We can achieve that as follows:

a. Remove all new line characters from this file using the `tr` command:

```
$ tr -d '\n' < Step2.file > OneLine.file
```

b. Reintroduce new line characters after the string '1000' and before the character '>'. You will have to use two sed commands and use the s///g syntax to achieve this. For example: $ sed 's/A/Z/g' Example.fasta will replace ALL occurrences of A with Z on EVERY line in Example.fasta. Without the 'g' at the end only the first occurrence of A will be replaced on every line.

```
$ sed 's/>/\n>/g' OneLine.file | grep -v '^$' > FusedHeader.file
```

c. Redirect the output of the second sed command to a new file: Step3.file

```
$ sed 's/1000/1000\n/g' FusedHeader.file > Step3.file
```

4. Now use grep -c to find the number of promoters that have the motif G[AC]CACGT in Step3.file.

```
$ grep -c G[AC]CACGT Step3.file
```

OR

```
$ grep -A 13 -f bZIP1_TargetIDs.txt At_Promoters_1KB.fasta |grep
[[:alnum:]] |tr -d '\n' | sed 's/>/\n>/g' | sed 's/1000/1000\n/g'
|grep -c G[AC]CACGT
```