

Stage 1

Project 1: BASH Basic

You are to achieve this short story with the command line alone.


Create your copy of the file and enter your command in the terminal space (\$) below each action.

Your Team Name: fredrick-sanger

Participants who contributed significantly (slack handle alone):

N/B: The story here is fictional and the files are just hypothetical. Please don't use it for any serious research work.

Please copy exactly what worked. Do not paraphrase. A single mismatch makes you lose your point.

To submit this project, make this document open using the  share icon at the top right corner. Copy the link and submit it on HackBio platform (See submission Guide).

1. Login to your coding workspace → I USE UBUNTU (VM)
2. Create a folder titled your name

→ mkdir: make directory

```
$ mkdir mariam
```

3. Create another new directory titled biocomputing and change to that directory with one line of command

→ it will make a new directory called biocomputing AND will change the working directory using cd (change directory) command to the final argument of previous command executed using "\$_"

Ref:

<https://stackoverflow.com/questions/14136635/one-command-to-create-and-change-directory>

```
$ mkdir biocomputing && cd "$_"
```

4. Download these 3 files:
 - a. <https://raw.githubusercontent.com/josoga2/dataset-repos/main/wildtype.fasta>
 - b. <https://raw.githubusercontent.com/josoga2/dataset-repos/main/wildtype.gb>
 - c. <https://raw.githubusercontent.com/josoga2/dataset-repos/main/wildtype.gbk>

→ using wget command followed by the link

```
$ wget
https://raw.githubusercontent.com/josoga2/dataset-repos/main/wildtype.fna
$ wget
https://raw.githubusercontent.com/josoga2/dataset-repos/main/wildtype.gbk
$ wget
https://raw.githubusercontent.com/josoga2/dataset-repos/main/wildtype.gbk
```

5. OH! You made a mistake. You have to **move the .fna file** to the folder titled your name directly. (Do this with one command. Hint: [See our cheatsheet](#))

→ `mv file1 file2`

File1 is the .fna file and file2 is the folder entitled my name

```
$ mv wildtype.fna /home/mariam/mariam
```

6. OH No! The `gbk` file is a duplicate, they are actually the same thing. Please delete it.

→ check that the file was removed using `ls` command as `ls` will list all the files in the working directory

```
$ rm wildtype.gbk.1
```

7. The **.fna file** is actually from a bacteria, and it should definitely have a TATA (`tata`) box for initiating gene transcription. The molecular biologist is trying to understand the implication of dual TATA sequences. The files got mixed up and we are not sure which is wildtype and which is mutant. The mutant should have "tatatata" while the normal should have just "`tata`". Can you confirm if the file is **mutant or wild type**

→ `grep` function will catch the `tatata` sequence in `wildtype.fna` file

ANOTHER SOLUTION (CHECK IF IT IS CORRECT)

1. Create a bash file

```
vim wildtype_or_mutant.sh
#!/bin/bash
if grep -q "tatata" wildtype.fna; then
    echo "mutant"
else
    echo "wildtype"
fi
```

2. In terminal: `./wildtype_or_mutant.sh wildtype.fna`

```
$ grep 'tatata' wildtype.fna
```

8. If it is mutant, print all the lines that show it is a mutant into a new file

```
$ grep 'tatata' wildtype.fna > mutant.fna
```

9. What is your favorite gene? (In any organism). Each team member should pick a unique gene different from every other person

CYP2C9

```
$
```

10. Download the fasta format of the gene from NCBI Nucleotide
(check last badge in document)

```
$ sh -c "$(wget -q
https://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/install-edirect
.sh -O -)"
export PATH=${HOME}/edirect:${PATH}
esearch -db nucleotide -query "NG_008385.2" | efetch -format
fasta > CYP2C9.fna

ORRR
$wget
"https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&id=NM_001256799 &rettype=fasta&retmode=text" -O
gabdh.fasta
```

11. How many lines are in the FASTA file (with the exception of the header)

Output: 842

wc -l → counts the number of lines

awk → will filter the file

\$1-1 → will exclude the 1st line in the file

Check the no. of lines of the whole file first → wc -l < CYP2C9.fna

```
$ wc -l < CYP2C9.fna | awk '{print($1-1)}'
```

12. How many times does A occur

Remove the header and save the fasta file without header in another file called CYP2C9_noheader.fna

```
$ vim CYP2C9_noheader.fna
$ sed 's/>NG_008385.2 Homo sapiens cytochrome P450 family 2
subfamily C member 9 (CYP2C9), RefSeqGene (LRG_1195) on
chromosome 10/ /g' CYP2C9.fna > CYP2C9_noheader.fna
$ grep -o "A" CYP2C9_noheader.fna | wc -l
```

13. How many times does G occur

```
$ grep -o "G" CYP2C9_noheader.fna | wc -l
```

14. How many times does C occur

```
$ grep -o "C" CYP2C9_noheader.fna | wc -l
```

15. How many times does T occur

```
$ grep -o "T" CYP2C9_noheader.fna | wc -l
```

16. Calculate the %GC content of your gene

```
$vim GCcontent.sh
$chmod u+x GCcontent.sh

#!/bin/bash

G=$(grep -o "G" CYP2C9_noheader.fna | wc -l)

echo "G count is: $G"

C=$(grep -o "C" CYP2C9_noheader.fna | wc -l)

echo "C count is: $C"

A=$(grep -o "A" CYP2C9_noheader.fna | wc -l)

echo "A count is: $A"

T=$(grep -o "T" CYP2C9_noheader.fna | wc -l)

echo "T count is: $T"

GC=$(( $G + $C ))

echo "summation of GC is: $GC"

total=$(( $G + $C + $A + $T ))

echo "total is: $total"
```

```
GCcontent= echo "scale=4; $GC / $total" | bc
```

17. Create a nucleotide file title your name

```
$ vim mariam.fasta
```

18. "echo" the following into the file using >>

```
$ vim mariam.fasta
$ A=$(grep -o "A" CYP2C9_noheader.fna | wc -l)

C=$(grep -o "C" CYP2C9_noheader.fna | wc -l)

T=$(grep -o "T" CYP2C9_noheader.fna | wc -l)

G=$(grep -o "G" CYP2C9_noheader.fna | wc -l)

echo "C count is:$C" >> mariam.fasta && echo "A count is:$A" >>
mariam.fasta && echo "G count is:$G" >> mariam.fasta && echo "T
count is:$T" >> mariam.fasta

vim mariam.fasta
```

19. Upload the file to your team's github repo in a folder called /output

```
$ link here
```

20. Save all the codes you have used in this project in a file named yourname.sh Upload all the codes you have used to your team's github repo in a folder called /script

```
$ link here
```

21. Clear your terminal space and print all the commands you have used today.

```
$ history
$ clear
```

22. List the files in the two folders and share a screenshot of your terminal below

```
$ ls
```

23. Take a screenshot of your terminal screen currently and paste it below

Project 2: Installing Bioinformatics Softwares on the terminal

N/B: You need to install and setup your conda environment with either anaconda or miniconda.

Please copy exactly what worked. Do not paraphrase. A single mismatch makes you loose your point.

1. Activate your base conda environment

```
$ conda activate base
```

2. Create a conda environment names funtools

```
$ conda create -n funtools
```

3. Activate the funtools environment

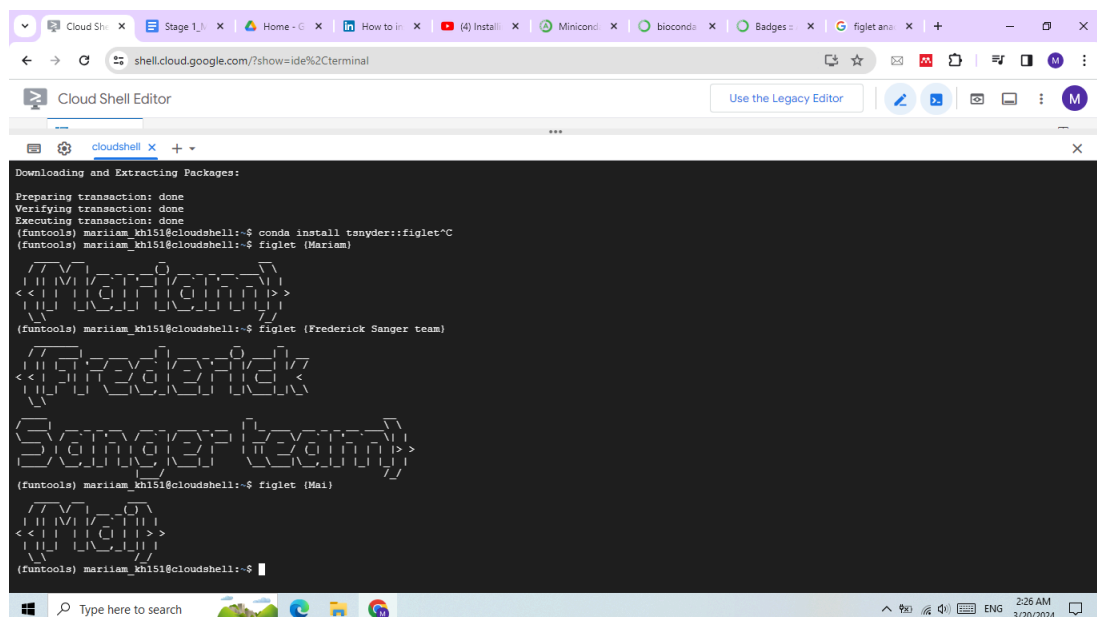
```
$ conda activate funtools
```

4. Install Figlet using conda

<https://anaconda.org/tsnyder/figlet>

```
$ conda install tsnyder::figlet
```

5. Run the following command figlet {your name}. Put a screenshot of what you see below 😊



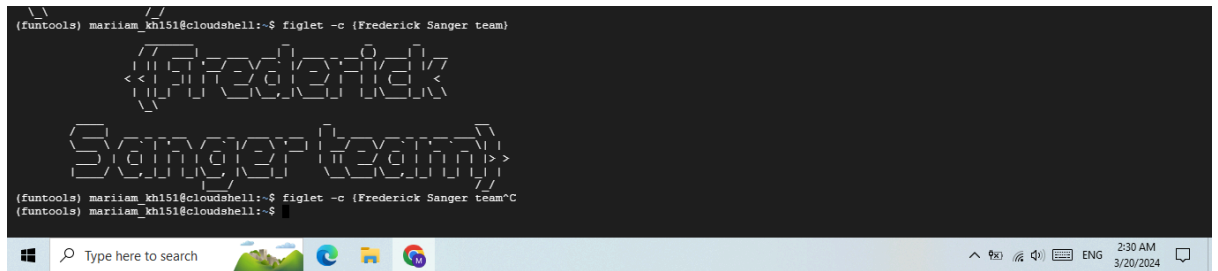
The screenshot shows a web browser window with the Cloud Shell Editor interface. The terminal window displays the following commands and output:

```
Downloading and Extracting Packages:
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
(funtools) mariam_kh151@cloudshell:~$ conda install tsnyder::figlet^C
(funtools) mariam_kh151@cloudshell:~$ figlet (Mariam)
<M>
<I>
<A>
<R>
<I>
<A>
<M>
(funtools) mariam_kh151@cloudshell:~$ figlet (Frederick Sanger team)
<F>
<R>
<E>
<D>
<E>
<R>
<I>
<C>
<K>
<T>
<E>
<A>
<M>
(funtools) mariam_kh151@cloudshell:~$ figlet (Mai)
<M>
<A>
<I>
(funtools) mariam_kh151@cloudshell:~$
```

NOTE:

`figlet -c {Frederick Sanger team}`

Will display the input {Frederick Sanger team} at the center



```
(funtools) mariam_khl51@cloudshell:~$ figlet -c {Frederick Sanger team}
      Frederick
    Sanger team
(funtools) mariam_khl51@cloudshell:~$ figlet -c {Frederick Sanger team}C
(funtools) mariam_khl51@cloudshell:~$
```

More arguments for figlet command: <https://linuxhint.com/figlet-command-linux/>

6. Install bwa through the bioconda channel

```
$ conda install -c bioconda bwa
```

7. Install blast through the bioconda channel

```
$ conda install -c bioconda blast
```

8. Install samtools through the bioconda channel

```
$ conda install -c bioconda samtools
```

9. Install bedtools through the bioconda channel

```
$ conda install -c bioconda bedtools
```

10. Install spades.py through the bioconda channel

```
$ conda install bioconda::spades
```

11. Install bcftools through the bioconda channel


```
$ conda install -c bioconda bcftools
```

12. Install fastp through the bioconda channel

```
$ conda install -c bioconda fastp
```


13. Install multiqc through the bioconda channel

```
$ pip install multiqc
```

To submit this project, make this document open using the  **share icon** at the top right corner. Copy the link and submit it on HackBio platform.

Finally, everyone in your team should be ready to discuss your code submission with everyone.

Learning Resources:

The Official learning resource for this internship is [HackBio's Genomics Course](#). Sign up to enjoy uninterrupted and synchronized flow of bioinformatics knowledge. If you have access to the course already, everything you need for the internship is already provided in the course.

However, we have plans for you if you are unable to purchase the course. We have gathered some resources for you to help you learn and navigate the internship better.

Stage 1

- How to [Access the terminal](#) for the purpose of stage 1 and 2
- [Introduction to BASH](#)
- [How to setup and use Conda](#)

- [An article on bioconda usage from HackBio](#)
- The rest is practice! practice!! practice!!!

How to download fasta file from ncbi:

Using Entrez Direct E-utilities: <https://www.ncbi.nlm.nih.gov/books/NBK179288/>

1. Install E-Direct software

```
sh -c "$(wget -q https://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/install-edirect.sh -O -)"
```

2. During installation a message will appear: write y

In order to complete the configuration process, please execute the following:

```
echo "export PATH=/home/mariam/edirect:${PATH}" >> ${HOME}/.bashrc
```

or manually edit the PATH variable assignment in your .bashrc file.

Would you like to do that automatically now? [y/N]

y

3. Edit the path: run the following

```
export PATH=${HOME}/edirect:${PATH}
```

4. Use the following syntax. Check the website of E-utilities to download proteins

```
esearch -db nucleotide -query "NC_001552" | efetch -format fasta > output.fasta
```

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi>