

PAC1 - Anàlisi bioinformàtic amb el terminal



Universitat
Oberta
de Catalunya



UNIVERSITAT DE
BARCELONA

Roger Massaguer

MU Bioinf. i Bioest.

M0.151 - Eines informàtiques per a
la bioinformàtica - Aula 1 |

Professor col·laborador

Guerau Fernández (gfernandezis@uoc.edu)

Data d'entrega

Març de 2024

Eines informàtiques per a la bioinformàtica

Guerau Fernández Isern i Begoña Hernández-Olasagarre

PAC1.Anàlisi bioinformàtic amb el terminal	
Presentació i objectius	Manipulació de dades Next Generation Sequencing (NGS) amb Unix
Data i format d'entrega	23 al 31 de març del 2024
Criteris de correcció	PAC1 – 15%

PRESENTACIÓ

L'objectiu d'aquest exercici és practicar la manipulació de dades NGS (Next Generation Sequencing) mitjançant comandes UNIX i descobrir el seu potencial. NGS és un grup de tecnologies dissenyades per a seqüenciar gran quantitat de segments d'ADN de manera massiva i en paral·lel, en la menor quantitat de temps i a un menor cost per nucleòtid que utilitzant tecnologies tradicionals de seqüenciació. En aquest exercici us facilitem un arxiu **FASTQ**. El format FASTQ és un format basat en text per a emmagatzemar tant una seqüència de nucleòtids, com les seves puntuacions (**score**) de qualitat corresponents. És necessari que estúdieu la pàgina http://en.wikipedia.org/wiki/fastq_format per a obtenir informació sobre el format FASTQ. El format de les seqüències, **reads** en anglès, que analitzarem consta d'un identificador el qual assigna un "1:" al **paired-end** o un "2:" al **mat-pair** seguit per diferents camps descriptius. Els identificadors procedents dels equips **Illumina** difereixen de la nomenclatura esmentada, assignant "/1" i "/2" al paired-end i al mat-pair, respectivament, a més d'eliminar els camps descriptius que segueixen.

OBJETIUS

Posar en pràctica els coneixements adquirits sobre el maneig del terminal de Linux en un escenari biològic real.

LLISTA DE FIGURES:

Imatge 1 Instruccions apartat 1.....	4
Imatge 2 Descomprimir arxiu testdata.tar	4
Imatge 3 Resultat terminal de la instrucció zcat	5
Imatge 4 Resultat terminal de la instrucció gunzip -c testdata_inter.fastq.gz.....	5
Imatge 5 Resultat de la instrucció gunzip -c testdata_inter.fastq.gz tail	6
Imatge 6 Resultat de la instrucció more testdata_inter.fastq	6
Imatge 7 Resultat de la instrucció head testdata_inter.fastq	7
Imatge 8 Resultat de la instrucció tail per mostrar les ultimes files.....	7
Imatge 9 Resultat de la instrucció cat	7
Imatge 10 Resultat de la instrucció less	8
Imatge 11 Instrucció per obtenir capçalera del tercer read i el resultat corresponent	8
Imatge 12 Instrucció per obtenir capçalera del antepenúltim read i el resultat corresponent	9
Imatge 13 Instruccions realitzades per obtenir les línies del fitxer	10
Imatge 14 Instrucció per obtenir els reads del fitxer	10
Imatge 15 Instrucció i resultat de la quantitat de reads pe i mp, respectivament, que conté el fitxer	11
Imatge 16 Subset del fitxer amb les primeres 30 línies.....	11
Imatge 17 Instrucció per eliminar la línia extra '--'	12
Imatge 18 Instrucció per crear testdata_1.fast i testdata_2.fast.....	12
Imatge 19 Instruccions per esbrinar els reads que contenen la seqüència TGCACTAC i les seqüències que contenen els in-line barcodes amb TGCACTAC.....	12
Imatge 20 Instrucció per la transformació de les capçaleres	13
Imatge 21 Comprobació de la seva eficàcia	13
Imatge 22 Instrucció per extreure el primers 100 reads i posteriorment comprimir el fitxer	14
Imatge 23 Generació mateixes comandes que a l'exercici 10 i 11 utilitzant pipe des del fitxer testdata_2.fastq.....	14
Imatge 24 Instrucció per llistar els reads etiquetats amb el in-line barcode TGCACTAC en el....	14
Imatge 25 Resultat de la instrucció que genera les freqüències dels primers 8 nucleòtids de cada línia	15
Imatge 26 Freqüències generades tenint en compte els diferents conjunts de 8 nucleòtids de cada línia	15

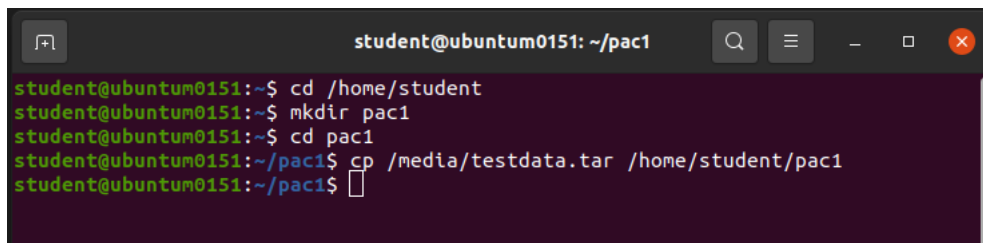
Exercici 1 – Manipulació de dades Next Generation Sequencing (NGS) amb GNU/Linux

1. Desa el fitxer de dades des de l'aula de l'assignatura. Crea un directori de treball en /home/student i còpia en ell, l'arxiu testdata.tar (1 punt).

Comandes necessàries:

```
cd /home/student  
mkdir pac1 #Creació del nou directori: "pac1".  
cd pac1  
cp /media/testdata.tar /home/student/pac1 #Còpia de l'arxiu testdata.tar
```

Imatge resultats:



```
student@ubuntu0151: ~/pac1  
student@ubuntu0151:~$ cd /home/student  
student@ubuntu0151:~$ mkdir pac1  
student@ubuntu0151:~$ cd pac1  
student@ubuntu0151:~/pac1$ cp /media/testdata.tar /home/student/pac1  
student@ubuntu0151:~/pac1$
```

Imatge 1 Instruccions apartat 1

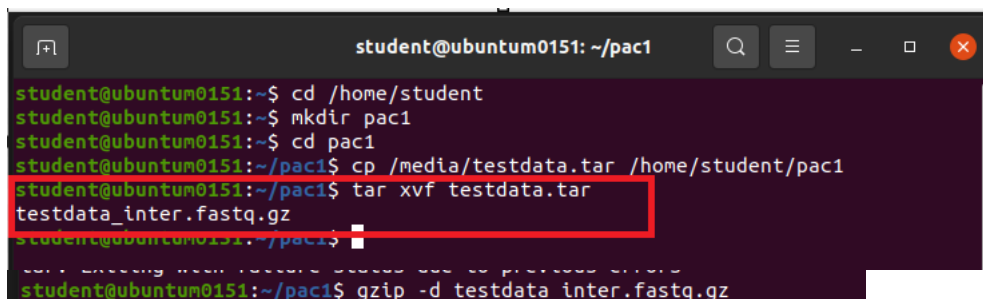
2. Descomprimeix l'arxiu (1 punt)

Comandes necessàries:

`tar xvf testdata.tar` #Amb aquesta instrucció es descomprimeix l'arxiu testdata.tar obtenint l'arxiu: testdata_inter.fastq.gz

`gzip -d testdata_inter.fastq.gz` #Descomprimirà el fitxer gzip i s'obté testdata_inter.fastq al mateix directori.

Imatge resultats:



```
student@ubuntu0151: ~/pac1  
student@ubuntu0151:~$ cd /home/student  
student@ubuntu0151:~$ mkdir pac1  
student@ubuntu0151:~$ cd pac1  
student@ubuntu0151:~/pac1$ cp /media/testdata.tar /home/student/pac1  
student@ubuntu0151:~/pac1$ tar xvf testdata.tar  
testdata_inter.fastq.gz  
student@ubuntu0151:~/pac1$  
student@ubuntu0151:~/pac1$ gzip -d testdata_inter.fastq.gz
```

Imatge 2 Descomprimir arxiu testdata.tar

Imatge 5 Resultat de la instrucció `gunzip -c testdata_inter.fastq.gz | tail`

Imatge 6 Resultat de la instrucció `more testdata inter.fastq`

Imatge 7 Resultat de la instrucció head testdata_inter.fasta

Imatge 8 Resultat de la instrucció tail per mostrar les ultimes files

Imatge 9 Resultat de la instrucció cat

Imatge 10 Resultat de la instrucció less

c. Quina és la capçalera de l'antepenúltim read? Selecciona els comandes perquè únicament es vegi línia que conté la capçalera de l'antepenúltim read És paired-end (pe, forward) o mat-pair (mp, reverse)?

Comandes necessàries:

grep -n '^@DHKW5DQ1' testdata_inter.fastq | tail -n 3 | head -n 1 #Mostrarà la línia que conté la capçalera de l'antepenúltim read, juntament amb el número de línia en el qual es troba. Això ens permetrà localitzar la capçalera dins de l'arxiu.

#Resultat:1676249:@DHKW5DQ1:324:C2G0EACXX:4:2316:21119:100793
2:N:0:TGAAGTGG

tail -n 20 testdata_inter.fastq #Per comprovar que la capçalera trobada equival a l'antepenúltim read. Mostrarà les últimes 20 línies de l'arxiu testdata_inter.fastq.

L'antepenúltim read és mat-pair.

Imatge resultats:



```

student@ubuntu0151:~/pac1$ grep -n '^@DHKW5DQ1' testdata_inter.fastq | tail -n 3 | head -n 1
1676249:@DHKW5DQ1:324:C2G0EACXX:4:2316:21119:100793 2:N:0:TGAAGTGG
student@ubuntu0151:~/pac1$ tail -n 20 testdata_inter.fastq
@DHKW5DQ1:324:C2G0EACXX:4:2316:21015:100762 2:N:0:TGAAGTGG
AAGGGTGGCTGGCTCTCCCTTAGAGATAGGTTGAGAGTTCGGCCATCCGGGAGGGGCTCAGAGTAGAGTATCTGCTGCTCCACATCGAAAGGAGCCAG
+
@<;B=?BDDAHGFI@FGEHIEHI>E<CECF@;1)0?B?60BBF@0;@CGIGHHD<>8>?:9ACCA5>;@($>A@DDCC:CCCCC9@CCB<>202228@C#
@DHKW5DQ1:324:C2G0EACXX:4:2316:21119:100793 1:N:0:TGAAGTGG
CCGTCTACTGCAGGGAGGCCACAGACATCCAGCATGAGTTCAAACCTAACGTATTACCTGCCAGCACTTCTGCACACCTGGGACCTCTCATCACAGCG
+
<@CFFFFHCFD<<EGGGIIIGIIHIIJJFGIJJIIJGIIHIIJJIIJJGIIHGHFFFEFFCE@CEDDCA?BDDDDA?CDCCDD<AB
@DHKW5DQ1:324:C2G0EACXX:4:2316:21119:100793 2:N:0:TGAAGTGG
ATTCTTTCTCGTGCGAGTCTTGGGACTCTCGGGCCGAAAAGACTCAGGAGGAGTCGGAGTTTCGTCTCTGCAAAATGCTGGAAGTCTGCTGGCCTCCAG
+
C@CFDDBDAH?FHE>FBHIIIGACHIIIGIIIBHIII;CFGIIIIIGIEE??=BA<=C35,5(9ACC@3>C>C<8ACDCDBC>1: :A>>@(:0<@ABCC
@DHKW5DQ1:324:C2G0EACXX:4:2316:21327:100822 1:N:0:TGAAGTGG
CCACTGACTGCAGGTCTTTGTCTGCAGCAAGCTTTCTGTGTGGCTCCTTCTGCCGCTGCTCCCCGAGACTAACGGCTGTTCCAATTCATTTTCATCTCA
+
@<@FFDDFHFDAAAEH>HHAHFHGGGGBD;FFC?4?CFGHBGGIGGDDF>@FAFBB<@CECEBEEFACCAC('8/8,(,;C>CAA:4@>@4(4@>@
@DHKW5DQ1:324:C2G0EACXX:4:2316:21327:100822 2:N:0:TGAAGTGG
CTTCATCTTTATCTGGAATCGATTTCAACCTTTAATTCTGTGTTTTGAGAGTTTCTCTCTCATATTTCCCATCAAACGTTTCAGTTTCTTCAGGA
+
<<?D=;DDAB><DH><+<EIFDBHIIIGEE@BG@HDA*:CFDG4?DBGGHA?FGFB?FFHGC388CA#####
  
```

Imatge 12 Instrucció per obtenir capçalera del antepenúltim read i el resultat corresponent

5. Quantes línies conté el fitxer? (1 punt)

Comandes necessàries:

wc -l testdata_inter.fastq

cat testdata_inter.fastq|wc -l #Alternativa

#Resultat: 1676260

Imatge resultats:

```
student@ubuntu0151: ~/pac1
student@ubuntu0151:~/pac1$ wc -l testdata_inter.fastq
1676260 testdata_inter.fastq
student@ubuntu0151:~/pac1$ cat testdata_inter.fastq|wc -l
1676260
student@ubuntu0151:~/pac1$
```

Imatge 13 Instruccions realitzades per obtenir les línies del fitxer

6. Utilitzant només comandes *bash*, quants *reads* conté el fitxer?(1 punt)

Comandes necessàries:

#Per determinar el nombre de lectures que conté el fitxer, s'analitzarà específicament l'identificador de cada seqüència, utilitzant el patró de cerca representatiu en aquest arxiu: @DHKW5DQ1.

```
grep "@DHKW5DQ1" testdata_inter.fastq|wc -l
```

#Resultat:419065. Per comprovar-ho, podem dividir el nombre total de línies de l'arxiu entre 4, ja que el format FastQ conté 4 línies per cada lectura. Per tant: $1676260/4 = 419065$.

Imatge resultats:

```
student@ubuntu0151: ~/pac1
student@ubuntu0151:~/pac1$ grep "@DHKW5DQ1" testdata_inter.fastq|wc -l
419065
student@ubuntu0151:~/pac1$
```

Imatge 14 Instrucció per obtenir els reads del fitxer

7. Quants reads tipus paired-end (pe) i mate-pair (mp) conté el fitxer? (1 punt)

Comandes necessàries:

#paired-end(pe) s'identifica amb 1 i mate-pair (mp) amb 2

```
grep -c "1:" testdata_inter.fastq
grep "2:" testdata_inter.fastq|wc -l
```

#Resultat: 443046

Imatge resultats:

```

student@ubuntu0151: ~/pac1
student@ubuntu0151:~/pac1$ grep -c "1:" testdata_inter.fastq
443046
student@ubuntu0151:~/pac1$ grep "2:" testdata_inter.fastq|wc -l
232876
student@ubuntu0151:~/pac1$
  
```

Imatge 15 Instrucció i resultat de la quantitat de reads pe i mp, respectivament, que conté el fitxer

8. Extreure els pe/mp reads continguts en el fitxer de dades i escriure'ls de manera separada en els fitxers testdata_1.fastq i testdata_2.fastq (1 punt)

Comandes necessàries:

#Per realitzar aquest apartat, s'iniciarà executant un subet representatiu del fitxer, comprovada la seva funcionalitat, es procedirà a realitzar l'execució del fitxer complet

head -n 30 testdata_inter.fastq| grep ' 1:' -A 3 #Mostra de les primeres 30 línies

#Comprobant la sortida de resultats, s'observa que el flag -A afegeix un línia extra '--', s'haurà d'eliminar

head -n 30 testdata_inter.fastq| grep ' 1:' -A 3|grep -v '^--\$'

#Línies pel fitxer complet:

cat testdata_inter.fastq| grep ' 1:' -A 3|grep -v '^--\$'> testdata_1.fast

cat testdata_inter.fastq| grep ' 2:' -A 3|grep -v '^--\$'> testdata_2.fast

Imatge resultats:

```

student@ubuntu0151: ~/pac1
+
@DHKW5DQ1:324:C2G0EACXX:4:2314:10729:77825 1:N:0:TGAAGTGG
ATCAGTGTTCAGGCCCGAGATGTGTGTTTACATGTTAGAGGAGGCAAAGTTACCGTAAGCTAGAAGAATGCACC
GGGTTTTTAAAAAGTTCTTGACAA
+
+11ADDDFHCFAA6CGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHGIIIIIIIIIIIIFHHHHGFFFF
DDDBDDDDDDDDCEEDDDDDDD
+
@DHKW5DQ1:324:C2G0EACXX:4:2314:10623:77858 1:N:0:TGAAGTGG
CCGTCTACTGCAGGTTAATACTGGCAGTGTACCATGCCAGCTGACTGTATTCATCATCAGCCAGTGTGTCTT
TTATACATGAACATTACCAAAGTTT
+
@@@FFDDFHCFFHGGIJJJJIGIHGIBGIIGEGIGIJJIIIGIIGFHIJJJIIGCGGIGIIGIFGIJJJJG
GHHFHHHFEEBDFFCACCCDCCDD
+
@DHKW5DQ1:324:C2G0EACXX:4:2314:10582:77926 1:N:0:TGAAGTGG
CCACGACATGCAACAGTCGAGTGGATAGCCCTGCCTGAGTGTGGTTATACTGGAGATTTCTTTTAAAAAGAG
AGTTTTTCTTCCCACTGTCACCAA
  
```

Imatge 16 Subet del fitxer amb les primeres 30 línies

```

student@ubuntu0151: ~/pac1
+
CCCCFFFFHCHHCCGHIIJIIIIIDHHHIGIIGIIJIIIIJJJIHJJJEIIJJJJJIGHEHFFDFFFEDEBDDDE
EDCCDEDDDDDCDEDDDDDDDD@CC
@DHKWS1:324:C2G0EACXX:4:2314:10729:77825 1:N:0:TGAACGG
ATCAGTGTTCAGGCCCCAGATGTGTGTTACATGTTAGAGGAGGCAAAGTTACCGTAAGCTAGAAGATGCACC
GGGTTTTTAAAGTTCTTGACAA
+
+11ADDDFHCFAA6CGIIIIIIHIIIIIIIIIIIIIIIIIIIIIIHGIIIIIIIIIIFHHHGGFFFF
DDDBDDDDDDDDDCDEDDDDDDDD
@DHKWS1:324:C2G0EACXX:4:2314:10623:77858 1:N:0:TGAACGG
CCGTCTACTGCAGGTTAATACTGGCAGTGTCCATGCCAGTGACTGTATTCATCATCAGCCAGTGTGTCT
TTATACATGAACATTACCAAAGTTT
+
@@@FDDDFHCFHFFHGGHIIJIIJIGIHGIBGIIGEGIGIJJIIIGIIFHIIJJJJIGCGGIGIIGIFGIIJJJIG
GHFHFFHFEEDFFCACCCDCCDD
@DHKWS1:324:C2G0EACXX:4:2314:10582:77926 1:N:0:TGAACGG
CCACGACATGCAACCACTCGCAGTGGATAGCCCTGCCTGAGTGTGTTACTGGAGATTTCTTTTAAAGAG
AGTTTTTCCTCCCACTGTACCAA
  
```

Imatge 17 Instrucció per eliminar la línia extra '--'.

```

student@ubuntu0151: ~/pac1
student@ubuntu0151:~/pac1$ cat testdata_inter.fastq | grep ' 1:' -A 3 | grep
-v '^--$'> testdata_1.fast
student@ubuntu0151:~/pac1$
student@ubuntu0151:~/pac1$ cat testdata_inter.fastq | grep ' 2:' -A 3 | grep
-v '^--$'> testdata_2.fast
student@ubuntu0151:~/pac1$
  
```

Imatge 18 Instrucció per crear testdata_1.fast i testdata_2.fast

9.Respon:

a. Esbrinar el número de reads que contenen la seqüència TGCACTAC en testdata_1.fastq.

Comandes necessàries:

```
grep -c "TGCACTAC" testdata_1.fast
```

#Resultat: 2442 número de reads que contenen la seqüència TGCACTAC

b. Es denomina in-line barcode als 8 primers nucleòtids de cada seqüència. Esbrinar el número de reads que comencen amb la seqüència TGCACTAC en testdata_1.fastq. (1 punt)

Comandes necessàries:

```
grep -c "^TGCACTAC" testdata_1.fast
```

#Resultat: 1965 número de reads que comencen amb la seqüència TGCACTAC

Imatge resultats:

```

student@ubuntu0151:~/pac1
student@ubuntu0151:~/pac1$ grep -c "TGCACTAC" testdata_1.fast
2442
student@ubuntu0151:~/pac1$ grep -c "^TGCACTAC" testdata_1.fast
1965
  
```

Imatge 19 Instruccions per esbrinar els reads que contenen la seqüència TGCACTAC i les seqüències que contenen els in-line barcodes amb TGCACTAC

En els exercicis 10, 12 i 14 és obligatori utilitzar el comando `sed` per respondre

10. Transformar totes les capçaleres dels reads de l'arxiu "testdata_1.fastq" a format Illumina. Això implica reemplaçar l'espai de la capçalera que identifica el read com a `pe` i els diversos camps descriptius amb un `/1`. Guarda el resultat en un nou arxiu anomenat "testdata_1_nova_capçalera.fastq". Finalment, mostra algunes de les línies del resultat (1 punt).

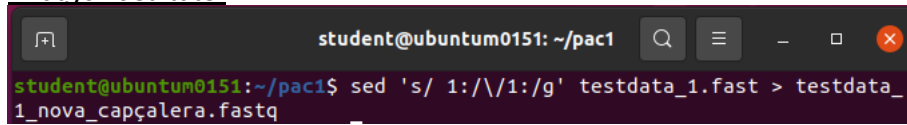
Comandes necessàries:

`sed 's/ 1:/\1:/g' testdata_1.fast > testdata_1_nova_capçalera.fastq` #Aquesta ordre buscarà cada aparició de ' 1:' a l'arxiu 'testdata_1.fast' i la substituirà per '/1:', i després guardarà el resultat en un nou arxiu anomenat 'testdata_1_nova_capçalera.fastq'.

`head -n 10 testdata_1.fast` #Mostra l'abans de la transformació

`head -n 10 testdata_1_nova_capçalera.fastq` #Mostra el després de la transformació

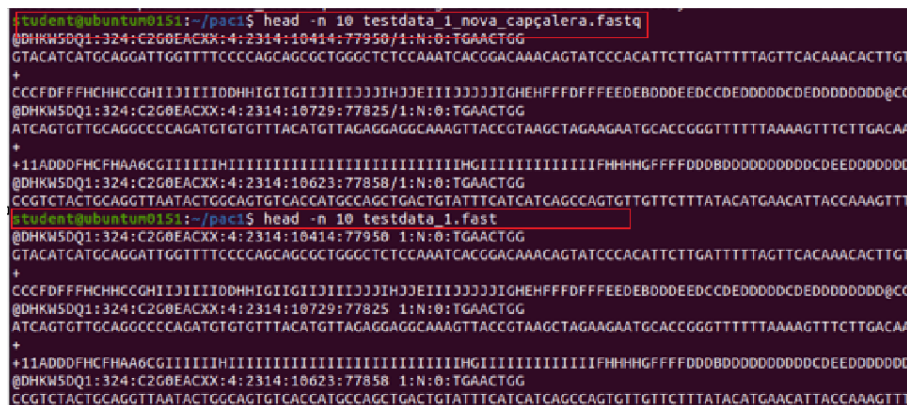
Imatge resultats:



```

student@ubuntu0151: ~/pac1
student@ubuntu0151:~/pac1$ sed 's/ 1:/\1:/g' testdata_1.fast > testdata_1_nova_capçalera.fastq
  
```

Imatge 20 Instrucció per la transformació de les capçaleres



```

student@ubuntu0151:~/pac1$ head -n 10 testdata_1_nova_capçalera.fastq
@DHKMSDQ1:324:C2G0EACXX:4:2314:10414:77950/1:N:0:TGAAGTGG
GTACATCATGCAGGATTGGTTTCCCCAGCAGCGCTGGGCTCTCCAAATCAGCGACAACAGTATCCACATTCTTGATTTTAGTTCACAAACACTTGT
+
CCCFDFFFHCHHCCGHIJIIIIIDHHHIGIIGIJJIIJJJJHJJJEIIJJJJJIGHEHFFDFFFEEDBDDDEDCDEDDDDDCDEDDDDDDDD@CC
@DHKMSDQ1:324:C2G0EACXX:4:2314:10729:77825/1:N:0:TGAAGTGG
ATCAGTGTTCAGGCCCCAGATGTGTGTACATGTAGAGGAGGCAAGTTACCGTAAGCTAGAAGAATGCACCGGTTTTTAAAGTTTCTTGACAA
+
11ADD0FHCFAA6CGIIIIIIHIIIIIIIIIIIIIIIIIIIIHGIIIIIIIIIIFHHHGGFFFDDBDDDDDDDDDCDEDDDDDDDD
@DHKMSDQ1:324:C2G0EACXX:4:2314:10623:77858/1:N:0:TGAAGTGG
CCGTCTACTGCAGGTTAATCTGCGAGTGTACCATGCCAGCTGACTGTATTTTCATCATCAGCCAGTGTGTCTTTATACATGAACATTACCAAGTTT
student@ubuntu0151:~/pac1$ head -n 10 testdata_1.fast
@DHKMSDQ1:324:C2G0EACXX:4:2314:10414:77950 1:N:0:TGAAGTGG
GTACATCATGCAGGATTGGTTTCCCCAGCAGCGCTGGGCTCTCCAAATCAGCGACAACAGTATCCACATTCTTGATTTTAGTTCACAAACACTTGT
+
CCCFDFFFHCHHCCGHIJIIIIIDHHHIGIIGIJJIIJJJJHJJJEIIJJJJJIGHEHFFDFFFEEDBDDDEDCDEDDDDDCDEDDDDDDDD@CC
@DHKMSDQ1:324:C2G0EACXX:4:2314:10729:77825 1:N:0:TGAAGTGG
ATCAGTGTTCAGGCCCCAGATGTGTGTACATGTAGAGGAGGCAAGTTACCGTAAGCTAGAAGAATGCACCGGTTTTTAAAGTTTCTTGACAA
+
11ADD0FHCFAA6CGIIIIIIHIIIIIIIIIIIIIIIIIIIIHGIIIIIIIIIIFHHHGGFFFDDBDDDDDDDDDCDEDDDDDDDD
@DHKMSDQ1:324:C2G0EACXX:4:2314:10623:77858 1:N:0:TGAAGTGG
CCGTCTACTGCAGGTTAATCTGCGAGTGTACCATGCCAGCTGACTGTATTTTCATCATCAGCCAGTGTGTCTTTATACATGAACATTACCAAGTTT
  
```

Imatge 21 Comprobació de la seva eficàcia

11. Extreure els primers 1000 reads de testdata_1_nova capçalera.fastq i salva el fitxer com testdata_1_sub1000.fastq. Comprimeix el fitxer (1 punt)

Comandes necessàries:

`head -n 4000 testdata_1_nova_capçalera.fastq > testdata_1_sub1000.fastq` #head per seleccionar les primeres 4000 línies de l'arxiu "testdata_1_nova_capçalera.fastq", ja que cada "read" ocupa 4 línies en el format FastQ.

`gzip testdata_1_sub1000.fastq` #Compressió de l'arxiu "testdata_1_sub1000.fastq"

Imatge resultats:

```
student@ubuntu0151: ~/pac1
student@ubuntu0151:~/pac1$ head -n 4000 testdata_1_nova_capçalera.fastq >
testdata_1_sub1000.fastq
student@ubuntu0151:~/pac1$ gzip testdata_1_sub1000.fastq
student@ubuntu0151:~/pac1$
```

Imatge 22 Instrucció per extreure el primers 100 reads i posteriorment comprimir el fitxer

12. Repeteix l'exercici 10 & 11 intercanviant "/" per "/" en una única línia de comandos utilitzant pipe "|" des del fitxer testdata_2.fastq fins al fitxer comprimit testdata_2_sub1000.fastq.gz (1 punt)

Comandes necessàries:

```
sed 's/ 2:/\2:/g' testdata_2.fast | head -n 4000 | gzip >testdata_2_sub1000.fastq
```

Imatge resultats:

```
student@ubuntu0151: ~/pac1
student@ubuntu0151:~/pac1$ sed 's/ 2:/\2:/g' testdata_2.fast | head -n 4000 | g
zip >testdata_2_sub1000.fastq.gz
```

Imatge 23 Generació mateixes comandes que a l'exercici 10 i 11 utilitzant pipe | des del fitxer testdata_2.fastq

13. Llista tots els reads etiquetats amb el in-line barcode TGCACTAC en el fitxer testdata_1_sub1000.fastq.gz i salva'ls en el fitxer sample_TGCACTAC_sub.1.fastq . Mostra el resultat. (1 punt)

Comandes necessàries:

```
grep "^TGCACTAC" -A 2 -B 1 testdata_1_sub1000.fastq.gz | grep -v "^--" >
sample_TGCACTAC_sub.1.fastq
```

#Resultat: no conté seqüències.

Imatge resultats:

```
student@ubuntu0151:~/pac1$ grep "TGCACTAC" -A 2 -B 1 testdata_1_sub1000.fastq.gz
| grep -v "^--" > sample_TGCACTAC_sub.1.fastq
```

Imatge 24 Instrucció per llistar els reads etiquetats amb el in-line barcode TGCACTAC en el

14. Quines són les 12 seqüències de codis de in-line barcodes més freqüents (de 8 nucleòtids de longitud) en l'arxiu "testdata_1.fastq", i quina és la seva freqüència d'aparició?

a. Es consideraran només els primers 8 nucleòtids de cada línia (1 punt)

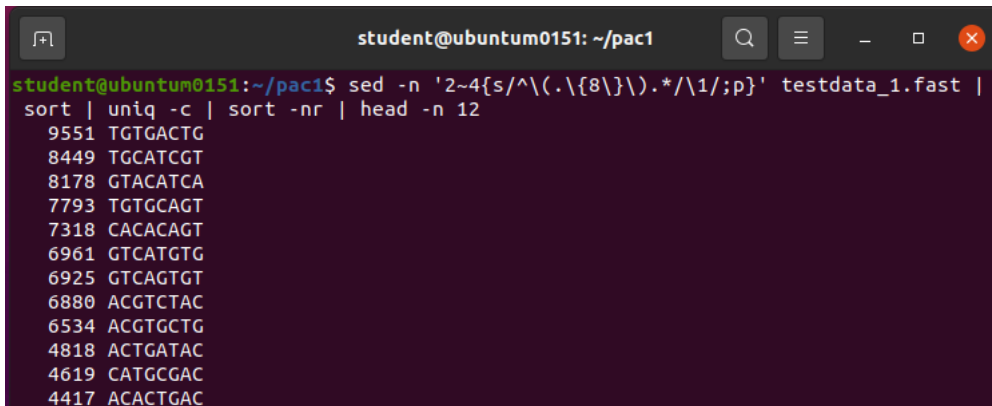
Comandes necessàries:

```
sed -n '2~4{s/^\(.{8}\).*\1/;p}' testdata_1.fast | sort | uniq -c | sort -nr | head -n 12
```

#sed per seleccionar només els primers 8 caràcters de cada segona línia (seqüències) de cada grup de 4 línies. Després, fem servir sort per ordenar les seqüències, uniq -c per comptar les freqüències de cada seqüència única, sort -nr per ordenar-les per

freqüència de més a menys, i finalment, head -n 12 per mostrar les 12 seqüències més freqüents.

Imatge resultats:



```

student@ubuntu0151: ~/pac1
student@ubuntu0151:~/pac1$ sed -n '2~4{s/^(\{8\}).*/\1;p}' testdata_1.fast |
sort | uniq -c | sort -nr | head -n 12
 9551 TGTGACTG
 8449 TGCATCGT
 8178 GTACATCA
 7793 TGTGCAGT
 7318 CACACAGT
 6961 GTCATGTG
 6925 GTCAGTGT
 6880 ACGTCTAC
 6534 ACGTGCTG
 4818 ACTGATAC
 4619 CATGCGAC
 4417 ACACTGAC
  
```

Imatge 25 Resultat de la instrucció que genera les freqüències dels primers 8 nucleòtids de cada línia

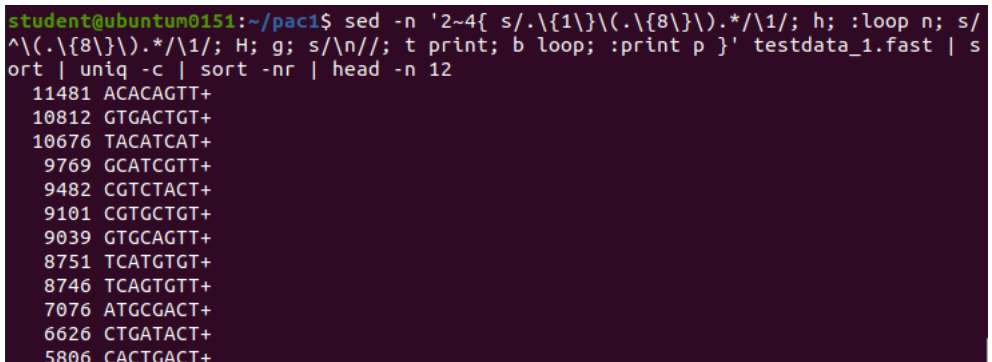
b. Si es tenen en compte els diferents conjunts de 8 nucleòtids de cada línia (1 punt)

Comandes necessàries:

```
sed -n '2~4{ s/\{1\}(\{8\}).*/\1/; h; :loop n; s/^\{8\}/H; g; s/\n//; t print; b loop; :print p }' testdata_1.fast | sort | uniq -c | sort -nr | head -n 12
```

sed per recórrer tots els conjunts de 8 nucleòtids de cada segona línia de cada grup de 4 línies. Després, fa el mateix procés de classificació i comptatge com a l'opció a.

Imatge resultats:



```

student@ubuntu0151:~/pac1$ sed -n '2~4{ s/\{1\}(\{8\}).*/\1/; h; :loop n; s/
^\{8\}/H; g; s/\n//; t print; b loop; :print p }' testdata_1.fast | s
ort | uniq -c | sort -nr | head -n 12
 11481 ACACAGTT+
 10812 GTGACTGT+
 10676 TACATCAT+
 9769 GCATCGTT+
 9482 CGTCTACT+
 9101 CGTGCTGT+
 9039 GTGCAGTT+
 8751 TCATGTGT+
 8746 TCAGTGTT+
 7076 ATGCGACT+
 6626 CTGATACT+
 5806 CACTGACT+
  
```

Imatge 26 Freqüències generades tenint en compte els diferents conjunts de 8 nucleòtids de cada línia