Tarea 1 25/08/2025

Roberto Naranjo Partarrieu

Bioinformática e investigación reproducible para análisis genómicos

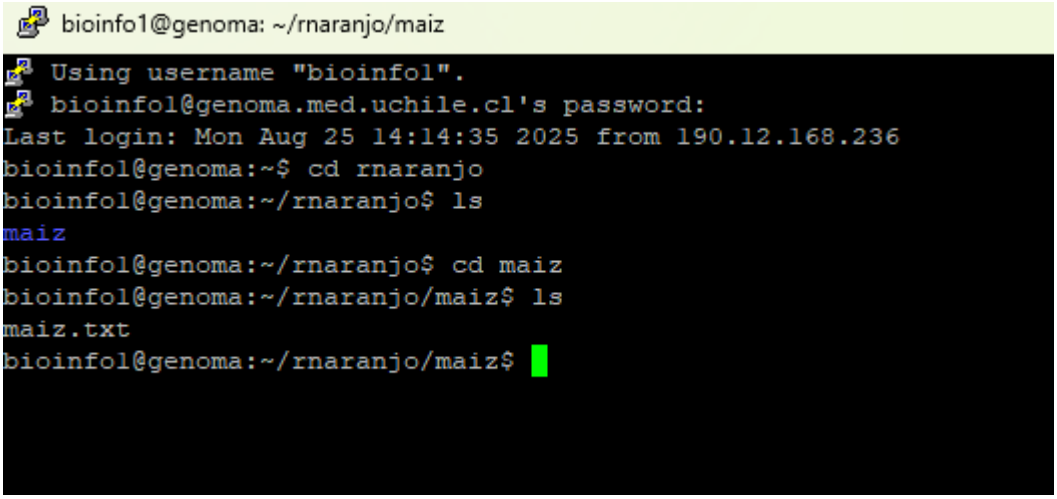
La tarea se encuentra en el directorio “rnaranjo” en el cual se generaron 3 directorios

Directorio	Ejercicio
Maiz	Ejercicio 1
Pob	Ejercicio 2
Sec1	Ejercicio 3

```
bioinfo1@genoma:~/rnaranjo$ ls
maiz  Pob  sec1
bioinfo1@genoma:~/rnaranjo$
```

Ejercicio 1

Escribe una línea de código que cree un archivo con los nombres de las muestras de maíz enlistadas en /Unidad1/Sesion1/Prac_Uni1/Maiz/nuevos_final.fam.



Archivo maíz.txt con la información del archivo nuevos_final.fam

```
bioinfo1@genoma:~/rnaranjo/maiz$ less maiz.txt
1 maiz_3 0 0 0 -9
2 maiz_68 0 0 0 -9
3 maiz_91 0 0 0 -9
4 maiz_39 0 0 0 -9
5 maiz_12 0 0 0 -9
6 maiz_41 0 0 0 -9
7 maiz_35 0 0 0 -9
8 maiz_58 0 0 0 -9
9 maiz_51 0 0 0 -9
10 maiz_82 0 0 0 -9
```

Ejercicio 2

Escribe un script que cree 4 directorios llamados PobA, PobB, PobC, PobD y dentro de cada uno de ellos un archivo de texto que diga "Este es un individuo de la población x" donde x debe corresponder al nombre del directorio.

```
bioinfo1@genoma:~/rnaranjo/Pob$ ls
PobA  PobB  PobC  PobD
bioinfo1@genoma:~/rnaranjo/Pob$
```

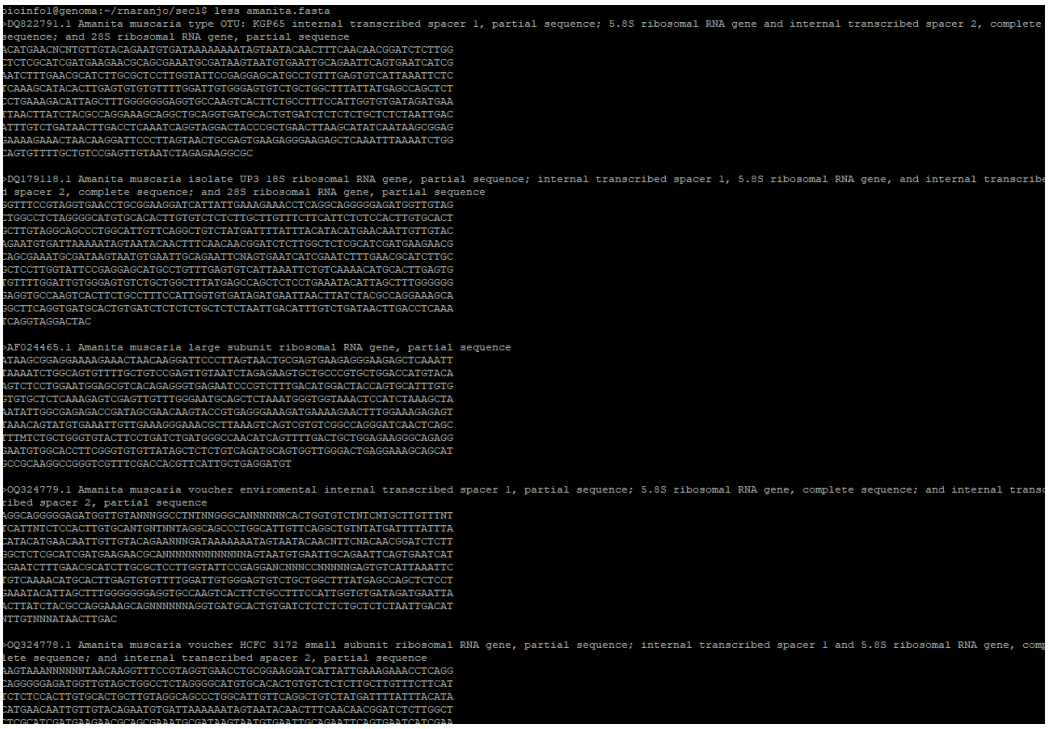
```
bioinfo1@genoma:~/rnaranjo/Pob/PobA$ ls
pobA
bioinfo1@genoma:~/rnaranjo/Pob/PobA$
```

```
Este es un individuo de la población A
pobA (END)
```

Ejercicio 3

Escribe un script que baje 5 secuencias (algún loci corto, no un genoma) de una especie que te interese y señala cuántas veces existe la secuencia "TGCA" en cada una de ellas.

Debo mencionar que es este ejercicio me tuve que apoyar en el uso de IA para elaborar y generar el código, con la consideración de estudiar porque la construcción de este y los componentes que presentan



Resultado usando opción A (awk)

```
bioinfol@genoma:~/rnaranjo/sec1$ less amanita.fasta
bioinfol@genoma:~/rnaranjo/sec1$ awk -v motif="TGCA" '    #dandole el motivo a buscar"TGCA"
> BEGIN{OFS="\t"}
> /^>/{
>   if (seq!="") {
>     up=seq; gsub(/[a-z]/, toupper("&"), up); gsub(/[^ACGT]/,"", up)       #Normalizamos la secuencia
>     c=0; for(i=1;i<=length(up)-length(motif)+1;i++) if(substr(up,i,length(motif))==motif) c++
>     print id, length(up), c
>   }
>   id=substr($0,2); split(id,a,/[ \t]/); id=a[1]          #Conservando el id utilizando el simbolo ">"
>   seq=""; next
> }
> { gsub(/\r/,""); seq=seq $0 }                           #concatena las lineas en una solacadena
> END{
>   if (seq!="") {
>     up=seq; gsub(/[a-z]/, toupper("&"), up); gsub(/[^ACGT]/,"", up)
>     c=0; for(i=1;i<=length(up)-length(motif)+1;i++) if(substr(up,i,length(motif))==motif) c++
>     print id, length(up), c
>   }
> }' amanita.fasta | awk 'BEGIN{print "secuencia\tlongitud_bp\tTGCA_count"}1' | column -t
secuencia   longitud_bp  TGCA_count
DQ822791.1  597          3
DQ179118.1  643          5
AF024465.1  604          3
OQ324779.1  520          4
OQ324778.1  635          5
bioinfol@genoma:~/rnaranjo/sec1$
```

Resultado usando opción B (grep)

```
bioinfol@genoma:~/rnaranjo/sec1$ csplit -z amanita.muscaria '/>/' '{*}'
  # El nombre de la secuencia es la primecsplit: ra linecannot open ‘amanita.muscaria’ for readinga
  nombre=$(head -n 1 $archivo)

  # Juntar todas las : No such file or directorylinea
s de la secuencia (quitar los >)
  secuencia=$(grep -v ">" $archivo | tr -d '\n')

  # Contar cuantas veces aparece TGCA
  cantidad=$(echo $secuencia | grep -o "TGCA" | wc -l)

  # Mostrar resultado
  echo "$nombre -> TGCA aparece $cantidad veces"
donebioinfol@genoma:~/rnaranjo/sec1$
bioinfol@genoma:~/rnaranjo/sec1$ # Revisar cada archivo creado
bioinfol@genoma:~/rnaranjo/sec1$ for archivo in xx*; do
>   # El nombre de la secuencia es la primera linea
>   nombre=$(head -n 1 $archivo)
>
>   # Juntar todas las lineas de la secuencia (quitar los >)
>   secuencia=$(grep -v ">" $archivo | tr -d '\n')
>
>   # Contar cuantas veces aparece TGCA
>   cantidad=$(echo $secuencia | grep -o "TGCA" | wc -l)
>
>   # Mostrar resultado
>   echo "$nombre -> TGCA aparece $cantidad veces"
> done
>DQ822791.1 Amanita muscaria type OTU: KGP65 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete
sequence; and 28S ribosomal RNA gene, partial sequence -> TGCA aparece 3 veces
>DQ179118.1 Amanita muscaria isolate UP3 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribe
d spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence -> TGCA aparece 5 veces
>AF024465.1 Amanita muscaria large subunit ribosomal RNA gene, partial sequence -> TGCA aparece 3 veces
>OQ324779.1 Amanita muscaria voucher enviromental internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transc
ribed spacer 2, partial sequence -> TGCA aparece 4 veces
>OQ324778.1 Amanita muscaria voucher HCFC 3172 small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1 and 5.8S ribosomal RNA gene, comp
lete sequence; and internal transcribed spacer 2, partial sequence -> TGCA aparece 5 veces
bioinfol@genoma:~/rnaranjo/sec1$
```

secuencia	longitud_bp	TGCA_count
DQ822791.1	597	3
DQ179118.1	643	5
AF024465.1	604	3
OQ324779.1	520	4
OQ324778.1	635	5

*En el script se detallan los comentarios y metodología para lograr los resultados