Tarea 3 de bioinformática

1.10.1

1. Change directory to CSB/unix/sandbox.

```
DELL@Christian-Lescano MINGW32 ~ (master)
$ cd gbi6g02/

DELL@Christian-Lescano MINGW32 ~/gbi6g02 (master)
$ ls
CSB/

DELL@Christian-Lescano MINGW32 ~/gbi6g02 (master)
$ cd CSB

DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB (master)
$ ls
LICENSE    data_wrangling/  good_code/  python/  regex/       sql/
README.md  git/             latex/      r/       scientific/  unix/

DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB (master)
$ cd unix/

DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix (master)
$ ls
data/  installation/  sandbox/  solutions/

DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix (master)
$ cd sandbox/
```

2. What is the size of the file Marra2014_data.fasta?

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ ls -lh ../data/Marra2014_data.fasta
-rw-r--r-- 1 DELL 197121 563K Apr 19 16:55 ../data/Marra2014_data.fasta
```

3. Create a copy of Marra2014_data.fasta in the sandbox and name it

my_file.fasta.

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ cp ../data/Marra2014_data.fasta Tarea3.fasta

DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ ls
'Papers and reviews'/   buzzard.sh    contador2.sh   osito/      tarea1/
 Tarea1.csv             buzzard2.sh   g02.sh         results/
 Tarea3.fasta           contador.sh   hola.sh*       rutas.txt
```

4. How many contigs are classified as isogroup00036?

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ grep -c isogroup00036 Tarea3.fasta
16

DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ grep isogroup00036 Tarea3.fasta | wc -l
16
```

5. Replace the original “two-spaces” delimiter with a comma.

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ cat Tarea3.fasta | tr -s "  " "," | head -n 3
>contig00001,length=527,numreads=2,gene=isogroup00001,status=it_thresh
ATCCTAGCTACTCTGGAGACTGAGGATTGAAGTTCAAAGTCAGCTCAAGCAAGAGATTTG
TTTACAATTAACCCACAAAAGGCTGTTACTGAAGGTGTGGCTTAAGTGTCAGAGCAACAG
```

6. How many unique isogroups are in the file?

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ grep ">" Tarea3.fasta | cut -d "," -f 4 | head -n 2
>contig00001  length=527  numreads=2  gene=isogroup00001  status=it_thresh
>contig00002  length=551  numreads=8  gene=isogroup00001  status=it_thresh
```

7. Which contig has the highest number of reads (numreads)? How many

reads does it have?

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ grep ">" Tarea3.fasta | cut -d "," -f 1,3 | head -n 3
>contig00001  length=527  numreads=2  gene=isogroup00001  status=it_thresh
>contig00002  length=551  numreads=8  gene=isogroup00001  status=it_thresh
>contig00003  length=541  numreads=2  gene=isogroup00001  status=it_thresh
```

1.10.2

1. Howmany times were the levels of individuals 3 and 27 recorded?

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ cut -f 1 Tarea3.fasta | grep -c -w 3
29
```

2. Write a script taking as input the filename and the ID of the individual,

and returning the number of records for that ID.

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ cat Ejercicio2_1.10.2.sh
#!/bin/bash

read -p "Ingrese el nombre del archivo: " archivo

read -p "Ingrese el número de ID a buscar: " id

registro=$(grep -n "$id" "$archivo" | cut -d: -f1)

if [ -n "$registro" ]; then
  echo "El número de registro del individuo con el ID $id es: $registro"
else
  echo "No se encontró ningún registro con el ID $id en el archivo $archivo"
fi

DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ bash Ejercicio2_1.10.2.sh
Ingrese el nombre del archivo: Tarea3.fasta
Ingrese el número de ID a buscar: 1
El número de registro del individuo con el ID 1 es: 1
11
22
33
39
50
106
```

3. [Advanced]17 Write a script that returns the number of times each

individual was sampled.

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ cat Ejercicio3_1.10.2.sh
#!/bin/bash

echo "Ingrese el nombre del archivo:"
read archivo
echo "Ingrese el número de ID:"
read id

registros=$(tail -n +2 "$archivo" | cut -d ',' -f 1 | sort -n | uniq)

for registro in $registros
do

  num_registros=$(grep -c "^$registro," "$archivo")

  echo "ID: $registro - Registros: $num_registros"
done

DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/sandbox (master)
$ bash Ejercicio3_1.10.2.sh
Ingrese el nombre del archivo:
Tarea3.fasta
Ingrese el número de ID:
2
ID: >contig00002 - Registros: 0
ID: length=551 - Registros: 0
ID: numreads=8 - Registros: 0
```

1.10.3 Plant–Pollinator Networks

Saavedra and Stouffer (2013) studied several plant–pollinator networks. These can be represented as rectangular matrices where the rows are polli-nators, the columns plants, a 0 indicates the

absence and 1 the presence of an interaction between the plant and the pollinator. The data of Saavedra and Stouffer (2013) can be found in the directory CSB/unix/data/Saavedra2013.

1. Write a script that takes one of these files and determines the number of rows (pollinators) and columns (plants). Note that columns are sep-arated by spaces and that there is a space at the end of each line. Your script should return.

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/data/Saavedra2013 (master)
$ cat Ejercicio1_1.10.3.sh
echo "Ingrese el nombre del archivo:"
read archivo

num_filas=$(wc -l < "$archivo")

echo "El número de filas en el archivo es: $num_filas"

echo "Ingrese el nombre del archivo:"
read archivo

primera_fila=$(head -n 1 "$archivo")

sin_espacios=$(echo "$primera_fila" | tr -d '[:space:]')

num_columnas=$(echo -n "$sin_espacios" | wc -c)

echo "El número de columnas en la primera fila del archivo es: $num_columnas"

DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/data/Saavedra2013 (master)
$ bash Ejercicio1_1.10.3.sh
Ingrese el nombre del archivo:
n1.txt
El número de filas en el archivo es: 97
Ingrese el nombre del archivo:
n1.txt
El número de columnas en la primera fila del archivo es: 80
```

2. [Advanced] 18 Write a script that prints the numbers of rows and columns for each network:

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/data/Saavedra2013 (master)
$ cat Ejercicio2_1.10.3.sh
echo "Ingrese el nombre del archivo:"
read archivo

num_filas=$(wc -l < "$archivo")

echo "El número de filas en el archivo es: $num_filas"

for ((i=1; i<=num_filas; i++))
do

  fila=$(head -n $i "$archivo" | tail -n 1)

  sin_espacios=$(echo "$fila" | tr -d '[:space:]')

  num_columnas=$(echo -n "$sin_espacios" | wc -c)

  echo "Número de columnas en la fila $i: $num_columnas"
done
```

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/data/Saavedra2013 (master)
$ bash Ejercicio2_1.10.3.sh
Ingrese el nombre del archivo:
n1.txt
El número de filas en el archivo es: 97
Número de columnas en la fila 1: 80
Número de columnas en la fila 2: 80
Número de columnas en la fila 3: 80
Número de columnas en la fila 4: 80
Número de columnas en la fila 5: 80
Número de columnas en la fila 6: 80
Número de columnas en la fila 7: 80
Número de columnas en la fila 8: 80
Número de columnas en la fila 9: 80
Número de columnas en la fila 10: 80
Número de columnas en la fila 11: 80
Número de columnas en la fila 12: 80
Número de columnas en la fila 13: 80
Número de columnas en la fila 14: 80
Número de columnas en la fila 15: 80
Número de columnas en la fila 16: 80
Número de columnas en la fila 17: 80
Número de columnas en la fila 18: 80
Número de columnas en la fila 19: 80
```

3. Which file has the largest number of rows? Which has the largest number of columns?

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/data/Saavedra2013 (master)
$ bash Ejercicio3_1.10.3.sh
Ingrese el nombre del archivo:
n2.txt
El número de filas en el archivo es: 62
El número máximo de filas es: 62
El número máximo de columnas es: 41

DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/data/Saavedra2013 (master)
$ bash Ejercicio3_1.10.3.sh
Ingrese el nombre del archivo:
n7.txt
El número de filas en el archivo es: 16
El número máximo de filas es: 16
El número máximo de columnas es: 25
```

1.10.4 Data Explorer

1. Write a script that, for a given CSV file and column number, prints

- the corresponding column name;
- the number of distinct values in the column;
- the minimum value;
- the maximum value

```
DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/data (master)
$ nano Ejercicio1_1.10.4.sh

DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/data (master)
$ cat Ejercicio1_1.10.4.sh
# Nombre de la columna
cut -d ',' -f 7 Buzzard2015_data.csv | head -n 1
# Valores distintos
cut -d ',' -f 7 Buzzard2015_data.csv | tail -n +2 | sort | uniq | wc -l
# Mínimo
cut -d ',' -f 7 Buzzard2015_data.csv | tail -n +2 | sort -n | head -n 1
# Máximo
cut -d ',' -f 7 ../data/Buzzard2015_data.csv | tail -n +2 | sort -n | tail -n 1

DELL@Christian-Lescano MINGW32 ~/gbi6g02/CSB/unix/data (master)
$ bash Ejercicio1_1.10.4.sh
biomass
285
1.048466198
14897.29471
```