

Universidad Regional Amazónica IKIAM

Nombre: Emily Benavides

Tarea TC2 FOR LOOP

1.10 Exercises

1.10.1 Next Generation Sequencing Data

In this exercise we work with next generation sequencing (NGS) data. Unix is excellent at manipulating the huge FASTA files that are generated in NGS experiments. FASTA files contain sequence data in text format. Each sequence segment is preceded by a single-line description. The first character of the description line is a “greater than” sign (>). The NGS data set we will be working with was published by Marra and DeWoody (2014), who investigated the immunogenetic repertoire of rodents. You will find the sequence file Marra2014_data.fasta in the directory CSB/unix/data. The file contains sequence segments (contigs) of variable size. The description of each contig provides its length, the number of reads that contributed to the contig, its isogroup (representing the collection of alternative splice products of a possible gene), and the isotig status.

1. Change directory to CSB/unix/sandbox.
2. What is the size of the file Marra2014_data.fasta?
3. Create a copy of Marra2014_data.fasta in the sandbox and name it my_file.fasta.
4. How many contigs are classified as isogroup00036?
5. Replace the original “two-spaces” delimiter with a comma.
6. How many unique isogroups are in the file?
7. Which contig has the highest number of reads (numreads)? How many reads does it have?

```
MINGW64:/c/Users/Emily Benavides/Documents/CSB-master/unix/sandbox
Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~ (EmilyBenavides)
$ cd Documents/CSB-master/unix/sandbox/
```

```
Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi Benavides)
$ echo "1. Cual es el tamaño del archivo?"
1. Cual es el tamaño del archivo?

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi Benavides)
$ ls -lh ../data/Marra2014_data.fasta | cut -d " " -f 6
553K
```

```
Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi Benavides)
$ echo "2. Hacer una copia del archivo en la carpeta sandbox"
2. Hacer una copia del archivo en la carpeta sandbox

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi Benavides)
$ cp ../data/Marra2014_data.fasta my_file.fasta

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi Benavides)
$ du -h ../data/Marra2014_data.fasta
556K    ../data/Marra2014_data.fasta

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi Benavides)
$ echo "3. Cuantos contigs se clasifican como isogruppo00036?"
3. Cuantos contigs se clasifican como isogruppo00036?

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi Benavides)
$ grep -c isogruppo00036 my_file.fasta
0

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi Benavides)
$ echo "4. Reemplace el delimitador"
4. Reemplace el delimitador

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi Benavides)
$ head -n 4 my_file.fasta
>contig00001 length=527 numreads=2 gene=isogroup00001 status=it_thresh
ATCCTAGCTACTCTGGAGACTGAGGATTGAAGTTCAAAGTCAGCTCAAGCAAGAGATTTG
TTTACAATTAACCCACAAAAGGCTGTTACTGAAGGTGTGGCTTAAGTGTGAGAGCAACAG
CTATGAGTGGAGGAATTTTCTATTACAATATAATTTTCATCTCTGGTAAATTGACCAATTA

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi Benavides)
$ cat my_file.fasta | tr -s " " "," > my_file.fasta

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi Benavides)
$ mv my_file.fasta my_fasta_emi.fasta
```

```

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi
Benavides)
$ cat my_fasta_emi.fasta | tr -s " " "," | head -n 6

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi
Benavides)
$ cat my_fasta_emi.fasta | tr -s " " "," | head -n 5

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi
Benavides)
$ echo "isogrupos unicos presentes en la carpeta"
isogrupos unicos presentes en la carpeta

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi
Benavides)
$ grep ">" my_fasta_emi.fasta | cut -d "," -f 4 | uniq -c | wc -l
0

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi
Benavides)
$ echo "el contig con el mayor numero de lecturas"
el contig con el mayor numero de lecturas

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/sandbox (Emi
Benavides)
$ grep ">" my_fasta_emi.fasta | cut -d "," -f 1,3 | sort -n -r -t "=" -k 2 | hea
d -n 1

```

1.10.2 Hormone Levels in Baboons

Gesquiere et al. (2011) studied hormone levels in the blood of baboons. Every individual was sampled several times.

1. How many times were the levels of individuals 3 and 27 recorded?
2. Write a script taking as input the file name and the ID of the individual, and returning the number of records for that ID.
3. Write a script that returns the number of times each individual was sampled.

MINGW64:/c/Users/Emily Benavides/Documents/CSB-master/unix/data

```

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master (EmiBenavides)
$ cd unix/data/

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data (master)
$ head -n 3 Gesquiere2011_data.csv
maleID  GC      T
1       66.9   64.57
1       51.09  35.57

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data (master)
$ cut -f 1 Gesquiere2011_data.csv | head -n 4
maleID
1
1
1

```

```

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data (master
)
$ echo "individuo 3"
individuo 3

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data (master
)
$ bash ../solutions/count_baboons.sh Gesquiere2011_data.csv 3
61

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data (master
)
$ echo "individuo 27"
individuo 27

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data (master
)
$ bash ../solutions/count_baboons.sh Gesquiere2011_data.csv 27
5

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data (master
)
$ echo "escriba un script que devuelva el numero de veces que se muestro a cada
individuo"
escriba un script que devuelva el numero de veces que se muestro a cada individu
o

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data (master
)
$ tail -n +2 Gesquiere2011_data.csv | cut -f 1 | sort -n | uniq
1
2
3
4
5

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data (master
)
$ echo "numero de veces que se muestra"
numero de veces que se muestra

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data (master
)
$ myIDS= tail -n +2 Gesquiere2011_data.csv | cut -f 1 | sort -n | uniq
1
2
3
4
5
6
7
8
9
10
11
12

```

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 separated by spaces and that there is a space at the end of each line. Your script should return
2. Write a script that prints the numbers of rows and columns for each network:
3. Which file has the largest number of rows? Which has the largest number of columns?

```
MINGW64:/c/Users/Emily Benavides/Documents/CSB-master/unix/data/Saavedra2013
Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data (master)
$ ls
Buzzard2015_about.txt      Marra2014_about.txt      Saavedra2013/            mirRNA/
Buzzard2015_data.csv      Marra2014_data.fasta    Saavedra2013_about.txt
Gesquiere2011_about.txt  Pacifici2013_about.txt  clase4_for.sh*
Gesquiere2011_data.csv   Pacifici2013_data.csv   filtrado1.txt

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data (master)
$ cd Saavedra2013

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data/Saavedra2013 (master)
$ nano ForLoop3.sh

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data/Saavedra2013 (master)
$ cat ForLoop3.sh
echo "parte1"
echo "El numero de columna 7"
  head -n 1 n7.txt | grep -o " " | wc -l
echo "el numero de filas"
  wc -l n7.txt

echo "parte 2"
for i in n*.txt
do
  echo "nombre de columna"
  head -n 1 $i | grep -o " " | wc -l
  echo "nombre de filas"
  wc -l $i
echo "final"
done

Emily Benavides@LAPTOP-I58SGI7Q MINGW64 ~/Documents/CSB-master/unix/data/Saavedra2013 (master)
$ bash ForLoop3.sh
parte1
El numero de columna 7
25
el numero de filas
16 n7.txt
parte 2
nombre de columna
80
nombre de filas
97 n1.txt
final
```