# Bitacora para el manejo de lecturas crudas que se unificaron con el Flash

**Elaborado por: Dra. Edith Elizondo Reyna**

Como parte de la estancia posdoctoral en el Departamento de Acuicultura bajo la dirección del Dr. Miguel Ángel del Río Portilla, CICESE, Ensenada. Período 2020-2021

## Para el siguiente ejercicio es necesario tener el Blastn instalado en la computadora

https://www.ncbi.nlm.nih.gov/guide/data-software/ (https://www.ncbi.nlm.nih.gov/guide/data-software/)

## Se utilizarán las lecturas crudas de buena calidad

"

In [ ]:
```
import os
from pandas import Series, DataFrame
import pandas as pd
from Bio import SeqIO, AlignIO, SeqRecord
from Bio.SeqRecord import SeqRecord
from Bio.Seq import Seq
import matplotlib.pyplot as plt
```

In [ ]:
```
from Bio import SeqIO, pairwise2, AlignIO, Phylo, Entrez, SeqRecord, Seq, SearchIO
from Bio.Align.Applications import ClustalwCommandline
from Bio.Blast import NCBIWWW, NCBIXML
from Bio.Seq import Seq
from Bio.SeqUtils import GC
from Bio.SeqRecord import SeqRecord

from matplotlib import *
import matplotlib.pyplot as plt
from matplotlib_venn import venn3_unweighted, venn2_unweighted

import os, pylab

from pandas import DataFrame
import pandas as pd

import pylab as pl
from pylab import *
```

In [ ]:
```
cd /home/elizondo/data/microalgas/lecturas_unificadas_flash/lcflash_fastq_fasta/
```

In [ ]:
```
ls
```

# Las lecturas estan en terminacion fastq y se tienen que cambiar a fasta

## Se llama a los programas a utilizar

In [ ]:
```
from Bio import SeqIO
import os
import gzip
```

## Se crea un directorio en donde se guardarán los archivos fasta lcflash_fastq_fasta¶

In [ ]:
```
os.makedirs('lcflash_fastq_fasta',exist_ok=True)
```

## Se asigna a 'archivos' los archivos a procesar

In [ ]:
```
archivos = !ls *extended.fastq
archivos
```

## Procesamiento de los archivos. En este caso los archivos ya estan descomprimidos en formato fastq, solo se deja (open) y se quita (gzip.)

In [ ]:
```
n=1
for archivo in archivos:
    with open(archivo, "rt") as handle:
        f = list(SeqIO.parse(handle, "fastq"))

    archivofasta= "lcflash_fastq_fasta/"+ archivo[:arch
ivo.find(".")]+".fasta"
    print(n, "procesando", archivo, len(f), "secuencias
")
    n+=1
    SeqIO.write(f, archivofasta, "fasta")
```

## Comando que verifica el nodulo donde esta llevandose a cabo el proceso en slurum

In [ ]:
```
!squeue
```

In [ ]:
```
cd /home/elizondo/data/microalgas/lecturas_unificadas_f
lash/lcflash_fastq_fasta
```

In [ ]:
```
ls
```

# C

In [ ]:
```
fout = open("blastn_ccalcitrans_extended1.sh", "w")
linea="""#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi
-blast-2.11.0/lib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/
```

```
NT

#
cd /home/elizondo/data/microalgas/lecturas_unificadas_f
lash/lcflash_fastq_fasta
date > tiempo_ccalcitrans_extended1.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin
/blastn \\
 -query ccalcitrans_extended.fasta \\
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \\
 -out blastn_ccalcitrans_extended1.tsv \\
 -evalue 1E-6 \\
 -max_target_seqs 1 \\
 -num_threads 24 \\
 -outfmt "6 std sskingdoms stitle staxids sscinames sco
mnames sblastnames strand"
date >> tiempo_ccalcitrans_extended1.txt

head blastn_ccalcitrans_extended1.tsv
echo ""
grep -c blastn_ccalcitrans_extended1.tsv

"""
fout.write(linea)
fout.close()
```

In [ ]:
```
!sbatch blastn_ccalcitrans_extended1.sh
```

In [ ]:
```
fout = open("blastn_ccalcitransNA_extended1.sh", "w")
linea="""#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
```

```
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi
-blast-2.11.0/lib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/
NT

#
cd /home/elizondo/data/microalgas/lecturas_unificadas_f
lash/lcflash_fastq_fasta
date > tiempo_ccalcitransNA_extended1.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin
/blastn \\
 -query ccalcitransNA_extended.fasta  \\
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \\
 -out blastn_ccalcitransNA_extended1.tsv \\
 -evalue 1E-6 \\
 -max_target_seqs 1 \\
 -num_threads 24 \\
 -outfmt "6 std sskingdoms stitle staxids sscinames sco
mnames sblastnames strand"
date >> tiempo_ccalcitransNA_extended1.txt

head blastn_ccalcitransNA_extended1.tsv
echo ""
grep -c blastn_ccalcitransNA_extended1.tsv

"""
fout.write(linea)
fout.close()
```

In [ ]:
```
!sbatch blastn_ccalcitransNA_extended1.sh
```

# M

In [ ]:
```
fout = open("blastn_cmurelli_extended12.sh", "w")
linea="""#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi
-blast-2.11.0/lib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/
NT

#
cd /home/elizondo/data/microalgas/lecturas_unificadas_f
```

```
lash/lcflash_fastq_fasta
date > tiempo_cmurelli_extended12.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin
/blastn \\
 -query cmurelli_extended.fasta \\
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \\
 -out blastn_cmurelli_extended12.tsv \\
 -evalue 1E-6 \\
 -max_target_seqs 1 \\
 -num_threads 24 \\
 -outfmt "6 std sskingdoms stitle staxids sscinames sco
mnames sblastnames strand"
date >> tiempo_cmurelli_extended12.txt

head blastn_cmurelli_extended12.tsv
echo ""
grep -c blastn_cmurelli_extended12.tsv

"""
fout.write(linea)
fout.close()
```

In [ ]:
```
!sbatch blastn_cmurelli_extended12.sh
```

In [ ]:
```
fout = open("blastn_cmurelliNA_extended12.sh", "w")
linea="""#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#
```

```
shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi
-blast-2.11.0/lib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/
NT

#
cd /home/elizondo/data/microalgas/lecturas_unificadas_f
lash/lcflash_fastq_fasta
date > tiempo_cmurelliNA_extended12.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin
/blastn \\
 -query cmurelliNA_extended.fasta \\
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \\
 -out blastn_cmurelliNA_extended12.tsv \\
 -evalue 1E-6 \\
 -max_target_seqs 1 \\
 -num_threads 24 \\
 -outfmt "6 std sskingdoms stitle staxids sscinames sco
mnames sblastnames strand"
date >> tiempo_cmurelliNA_extended12.txt

head blastn_cmurelliNA_extended12.tsv
echo ""
grep -c blastn_cmurelliNA_extended12.tsv

"""
fout.write(linea)
fout.close()
```

In [ ]:
```
!sbatch blastn_cmurelliNA_extended12.sh
```

# X

In [ ]:
```
fout = open("blastn_cx_extended12.sh", "w")
linea="""#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi
-blast-2.11.0/lib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/
NT

#
cd /home/elizondo/data/microalgas/lecturas_unificadas_f
lash/lcflash_fastq_fasta
date > tiempo_cx_extended12.txt
```

```
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin
/blastn \\
 -query cx_extended.fasta \\
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \\
 -out blastn_cx_extended12.tsv \\
 -evalue 1E-6 \\
 -max_target_seqs 1 \\
 -num_threads 24 \\
 -outfmt "6 std sskingdoms stitle staxids sscinames sco
mnames sblastnames strand"
date >> tiempo_cx_extended12.txt

head blastn_cx_extended12.tsv
echo ""
grep -c blastn_cx_extended12.tsv

"""
fout.write(linea)
fout.close()
```

In [ ]:
```
!sbatch blastn_cx_extended12.sh
```

In [ ]:
```
fout = open("blastn_cxNA_extended12.sh", "w")
linea="""#!/bin/sh

#
#SBATCH -p cicese
#SBATCH --job-name=blastn
#SBATCH -e blastn.%N.%j.err

#SBATCH -o blastn.%N.%j.log
#SBATCH -t 6-00:00:00
#
#SBATCH -N 1
#SBATCH -n 24
#
#SBATCH --exclusive

cd $SLURM_SUBMIT_DIR
#

shell=`/bin/basename \`/bin/ps -p $$ -ocomm=\``
```

```
if [ -f /usr/share/Modules/init/$shell ]
then
  . /usr/share/Modules/init/$shell
else
  . /usr/share/Modules/init/sh
fi

module load gcc-7.2
export LD_LIBRARY_PATH=/LUSTRE/apps/bioinformatica/ncbi
-blast-2.11.0/lib:$LD_LIBRARY_PATH
export BLASTDB=/LUSTRE/bioinformatica_data/BD/blast/db/
NT

#
cd /home/elizondo/data/microalgas/lecturas_unificadas_f
lash/lcflash_fastq_fasta
date > tiempo_cxNA_extended12.txt
time  /LUSTRE/apps/bioinformatica/ncbi-blast-2.11.0/bin
/blastn \\
 -query cxNA_extended.fasta \\
 -db /LUSTRE/bioinformatica_data/BD/blast/db/NT/nt \\
 -out blastn_cxNA_extended12.tsv \\
 -evalue 1E-6 \\
 -max_target_seqs 1 \\
 -num_threads 24 \\
 -outfmt "6 std sskingdoms stitle staxids sscinames sco
mnames sblastnames strand"
date >> tiempo_cxNA_extended12.txt

head blastn_cxNA_extended12.tsv
echo ""
grep -c blastn_cxNA_extended12.tsv

"""
fout.write(linea)
fout.close()
```

In [ ]:
```
!sbatch blastn_cxNA_extended12.sh
```

## Comando para verificar el contenido de los archivos *.err que se generan como resultado de las corridas

In [ ]:
```
!for f in blastn.*.err; do echo $f; ls -lh $f; head $f;
echo "--------------"; done
```

## Visualizar el archivo blastn

In [ ]:
```
!head blastn_ccalcitrans_extended1.tsv
```

In [ ]:
```
!head blastn_ccalcitransNA_extended1.tsv
```

In [ ]:
```
!head blastn_cmurelli_extended12.tsv
```

In [ ]:
```
!head blastn_cmurelliNA_extended12.tsv
```

In [ ]:
```
!head blastn_cx_extended12.tsv
```

In [ ]:
```
!head blastn_cxNA_extended12.tsv
```

## se visualiza el contenido del archivo de salida de blastn de lecturas crudas .tsv

In [ ]:
```
%%bash
head blastn_ccalcitrans_extended1.tsv
echo "numero de resultados es:"
wc -l blastn_ccalcitrans_extended1.tsv
```

In [ ]:
```
%%bash
head blastn_ccalcitransNA_extended1.tsv
echo "numero de resultados es:"
wc -l blastn_ccalcitransNA_extended1.tsv
```

In [ ]:
```
%%bash
head blastn_cmurelli_extended12.tsv
echo "numero de resultados es:"
wc -l blastn_cmurelli_extended12.tsv
```

In [ ]:
```
%%bash
head blastn_cmurelliNA_extended12.tsv
echo "numero de resultados es:"
wc -l blastn_cmurelliNA_extended12.tsv
```

In [ ]:
```
%%bash
head blastn_cx_extended12.tsv
echo "numero de resultados es:"
wc -l blastn_cx_extended12.tsv
```

In [ ]:
```
%%bash
head blastn_cxNA_extended12.tsv
echo "numero de resultados es:"
wc -l blastn_cxNA_extended12.tsv
```

In [ ]:
```
cd /home/elizondo/data/microalgas/lecturas_unificadas_f
lash/lcflash_fastq_fasta
```

In [ ]:
```
ls /home/elizondo/data/microalgas/lecturas_unificadas_f
lash/lcflash_fastq_fasta
```

## se visualizan los archivos .tsv que son los que tienen la informacion del blastn

In [ ]:
```
ls *.tsv
```

# FALTAN TERMINAR OTROS EL PROCESO

## se copian los archivos .tsv desde Lustre hasta mi caprteta tsv en mi direccion de omica

In [ ]:
```
%%bash
for f in ls *.tsv
do
echo $f
cp $f ~/data/microalgas/tsv/tsv_lc/
done
```

# CA

In [ ]:
```
!head -2 blastn_ccalcitrans_extended1.tsv
```

In [ ]:
```
encabezado =("qseqid", "sseqid", "pident", "length", "mismatch", "gapopen", "qstart",
             "qend", "sstart", "send", "evalue", "bitscore", "sskingdoms", "stitle",
             "staxids", "sscinames", "scomnames", "sblastnames")
```

In [ ]:
```
ftsv=pd.read_csv("blastn_ccalcitrans_extended1.tsv", sep = "\t", header=None , names= encabezado, engine="c")
ftsv.head()
```

In [ ]:
```
ftsv.to_csv("blastn_ccalcitrans_extended1.csv", header=True, index= None)
```

In [ ]:
```
ftab1= ftsv.groupby("sskingdoms")["qseqid"].count()
ftab1 = DataFrame(ftab1)
ftab1
```

# CN

In [ ]:
```
!head -2 blastn_ccalcitransNA_extended1.tsv
```

In [ ]:
```
encabezado =("qseqid", "sseqid", "pident", "length", "m
ismatch", "gapopen", "qstart",
             "qend", "sstart", "send", "evalue", "bitsc
ore", "sskingdoms", "stitle",
             "staxids", "sscinames", "scomnames", "sbla
stnames")
```

In [ ]:
```
ftsv=pd.read_csv("blastn_ccalcitransNA_extended1.tsv",
sep = "\t", header=None , names= encabezado, engine="c"
)
ftsv.head()
```

In [ ]:
```
ftsv.to_csv("blastn_ccalcitransNA_extended1.csv", heade
r=True, index= None)
```

In [ ]:
```
ftab1= ftsv.groupby("sskingdoms")["qseqid"].count()
ftab1 = DataFrame(ftab1)
ftab1
```

# MA

In [ ]:
```
!head -2 blastn_cmurelli_extended12.tsv
```

In [ ]:
```
encabezado =("qseqid", "sseqid", "pident", "length", "m
ismatch", "gapopen", "qstart",
             "qend", "sstart", "send", "evalue", "bitsc
ore", "sskingdoms", "stitle",
             "staxids", "sscinames", "scomnames", "sbla
stnames")
```

In [ ]:
```
ftsv=pd.read_csv("blastn_cmurelli_extended12.tsv", sep
= "\t", header=None , names= encabezado, engine="c")
ftsv.head()
```

In [ ]:
```
ftsv.to_csv("blastn_cmurelli_extended12.csv", header=Tr
ue, index= None)
```

In [ ]:
```
ftab1= ftsv.groupby("sskingdoms")["qseqid"].count()
ftab1 = DataFrame(ftab1)
ftab1
```

# MN

In [ ]:
```
!head -2 blastn_cmurelliNA_extended12.tsv
```

In [ ]:
```
encabezado =("qseqid", "sseqid", "pident", "length", "m
ismatch", "gapopen", "qstart",
             "qend", "sstart", "send", "evalue", "bitsc
ore", "sskingdoms", "stitle",
             "staxids", "sscinames", "scomnames", "sbla
stnames")
```

In [ ]:
```
ftsv=pd.read_csv("blastn_cmurelliNA_extended12.tsv", se
p = "\t", header=None , names= encabezado, engine="c")
ftsv.head()
```

In [ ]:
```
ftsv.to_csv("blastn_cmurelliNA_extended12.csv", header=
True, index= None)
```

In [ ]:
```
ftab1= ftsv.groupby("sskingdoms")["qseqid"].count()
ftab1 = DataFrame(ftab1)
ftab1
```

# XA

In [ ]:
```
!head -2 blastn_cx_extended12.tsv
```

In [ ]:
```
encabezado =("qseqid", "sseqid", "pident", "length", "m
ismatch", "gapopen", "qstart",
             "qend", "sstart", "send", "evalue", "bitsc
ore", "sskingdoms", "stitle",
             "staxids", "sscinames", "scomnames", "sbla
stnames")
```

In [ ]:
```
ftsv=pd.read_csv("blastn_cx_extended12.tsv", sep = "\t"
, header=None , names= encabezado, engine="c")
ftsv.head()
```

In [ ]:
```
ftsv.to_csv("blastn_cx_extended12.csv", header=True, in
dex= None)
```

In [ ]:
```
ftab1= ftsv.groupby("sskingdoms")["qseqid"].count()
ftab1 = DataFrame(ftab1)
ftab1
```

# XN

In [ ]:
```
!head -2 blastn_cxNA_extended12.tsv
```

In [ ]:
```
encabezado =("qseqid", "sseqid", "pident", "length", "m
ismatch", "gapopen", "qstart",
             "qend", "sstart", "send", "evalue", "bitsc
ore", "sskingdoms", "stitle",
             "staxids", "sscinames", "scomnames", "sbla
stnames")
```

In [ ]:
```
ftsv=pd.read_csv("blastn_cxNA_extended12.tsv", sep = "\
t", header=None , names= encabezado, engine="c")
ftsv.head()
```

In [ ]:
```
ftsv.to_csv("blastn_cxNA_extended12.csv", header=True,
index= None)
```

In [ ]:
```
ftab1= ftsv.groupby("sskingdoms")["qseqid"].count()
ftab1 = DataFrame(ftab1)
ftab1
```