

Bitácora para juntar lecturas (pair end) de secuenciación masiva

Se usa el programa **FLASH**
(<https://doi.org/10.1093/bioinformatics/btr507>)

Los datos están en el directorio

La instalación del programa se localiza en la página de distribución de **FLASH**
(<http://ccb.jhu.edu/software/FLASH/>) ¶

Se descarga el programa desde la página **FLASH** (<http://ccb.jhu.edu/software/FLASH/>) al directorio Downloads

Visualizamos la descarga

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

```
In [ ]: ls ~/Downloads/FLASH*
```

se procede a cambiarlo de lugar a ~/Documents/programas_genetica/

```
In [ ]: mv ~/Downloads/FLASH* ~/Documents/programas_genetica/
```

```
In [ ]: ls ~/Documents/programas_genetica/FLASH*
```

Procedemos a descomprimirlo

```
In [ ]: cd ~/Documents/programas_genetica/
```

Se procede como lo indica las instrucciones en el archivo README

```
In [ ]: !head -100 README
```

```
In [ ]: !tar -xzvf FLASH-1.2.11.tar.gz
```

```
In [ ]: cd FLASH-1.2.11/
```

```
In [ ]: !grep -A 15 "INSTALLATION" README
```

El comando make se usa para generar el archivo.

```
In [ ]: !make
```

Verificamos la creación del archivo ejecutable, debido a que el programa no se encuentra en el directorio de las aplicaciones, es necesario, señalar toda la ruta para su ejecución.

```
In [ ]: !~/Documents/programas_genetica/FLASH-1.2.11/flash -h
```

Alternativamente, se puede comprobar que está funcionando con:

```
In [ ]: !./flash -h
```

Se procede a continuar con el análisis

```
In [ ]: cd ~/Desktop/data/sec_masiva/
```

```
In [ ]: ls
```

El programa flash está en:

```
In [ ]: ls ~/Documents/programas_genetica/FLASH-1.2.11/
```

```
In [ ]: ### Por lo que se ejecuta de acuerdo con el contenido en la ayuda  
!~/Documents/programas_genetica/FLASH-1.2.11/flash -h
```

```
In [ ]: ### Por lo que se ejecuta de acuerdo con el contenido en la ayuda  
!~/Documents/programas_genetica/FLASH-1.2.11/flash -h |  
grep "$ flash"
```

utilizando los archivos que se encuentran en el directorio actual

```
In [ ]: ls
```

```
In [ ]: !~/Documents/programas_genetica/FLASH-1.2.11/flash 8_S3  
56_L001_R1_001.fastq.gz 8_S356_L001_R2_001.fastq.gz 2>  
&1 | tee flash.log
```

Se observa una advertencia, por lo que se modificará el traslape máximo

```
flash -M 100
```

```
In [ ]: !~/Documents/programas_genetica/FLASH-1.2.11/flash -M 1
00 8_S356_L001_R1_001.fastq.gz 8_S356_L001_R2_001.fast
q.gz 2>&1 | tee flash.log
```

Se sigue observando la advertencia, por lo que se modificará el traslape máximo

```
flash -M 250
```

```
In [ ]: !~/Documents/programas_genetica/FLASH-1.2.11/flash -M 2
50 8_S356_L001_R1_001.fastq.gz 8_S356_L001_R2_001.fast
q.gz 2>&1 | tee flash.log
```

Ya no hay advertencia. Se sigue el análisis.

```
In [ ]: ls out*
```

Visualizamos los archivos de salida

```
In [ ]: !head out.extendedFragments.fastq
```

```
In [ ]: !zgrep -c "^@" 8_S356_L001_R1_001.fastq.gz
```

```
In [ ]: !zgrep -c "^@" 8_S356_L001_R2_001.fastq.gz
```

```
In [ ]: !zgrep -c "^@" out.extendedFragments.fastq
```

Los comandos de las tres celdas anteriores se pueden visualizar de una manera más fácil con estos comandos:

```
In [ ]: %%bash
        for f in *.gz
        do
            echo $f
            gzcat $f | head -5
            echo
        done
```

Visualizando las primeras líneas

```
In [ ]: %%bash
        for f in *.gz
        do
            echo $f
            zgrep "^@" $f | head -5
            echo
        done
```

```
In [ ]: !grep "^@" out.extendedFrag.fastq | head -5
```

```
In [ ]: ls out*.fastq
```

```
In [ ]: %%bash
        for f in out*.fastq
        do
            echo $f
            grep "^@" $f | head -5
            echo "número de secuencias"
            grep -c "^@" $f
            echo
        done
```

visualización de los archivos de histograma

```
In [ ]: ls out*
```

```
In [ ]: !wc -l out.hist
```

```
In [ ]: !head out.hist
```

```
In [ ]: !head out.histogram
```

Se trabajará con el archivo `out.hist`

```
In [ ]: import pandas as pd
        from pandas import DataFrame
```

```
In [ ]: import pylab
        import matplotlib.pyplot as plt
```

```
In [ ]: histograma = pd.read_csv("out.hist", sep= '\t', header
        = None, names = ["readsize", "number"])
        histograma.set_index("readsize", inplace =True)
        histograma.head()
```

```
In [ ]: histograma1.plot( kind= "line")
        plt.xlabel("Read size")
        plt.ylabel("Frecuency")
        plt.title("Read frequency distribution after flash run"
        )
        plt.legend().set_visible(False)

        plt.show()
```