ReporteT02

Jessica Garcia, Manuel Rivera, Axel Rodriguez

2023-02-10

Contents

1 Para *E.coli*

1.1 Usando bwa para *indexar* el genoma de *E.coli*

Se comienza por crear un entorno de trabajo con los archivos necesarios para la práctica

```
# Se crea un screen de la práctica para poder recurrir a los comandos en distintos momentos
screen -r -D Alignment

# Pasar archivos que se necesitan a carpeta de trabajo
cp -r /mnt/Timina/bioinfoII/data/alignment/ /mnt/Timina/bioinfoII/arodriguez/alignment

# Cargar módulos
module load bwa/0.7.15
module load fastqc/0.11.3
module load samtools/1.9
```

Para este alineamiento, los archivos `Escherichia_coli_K12_MG1655.fasta` y `SRX189773_FNR_ChIP.fastq` se refieren al genoma de referencia y a las lecturas del experimento, respectivamente.

1.2 Usando bwa para *indexar* el genoma de *E.coli*

Los índices permiten a las herramientas de alineación (STAR, BWA, minimap2) recuperar rápida y eficientemente la información de la secuencia de referencia, ahorrando tiempo y memoria.

```
bwa index -p E_index -a bwtsw Escherichia_coli_K12_MG1655.fastax
```

- `-p` indica el prefijo de los archivos generados
- `-a` elige alguno de los algoritmos de indexación de bwa (en este caso bwtsw).

1.3 Alinea las lecturas de un experimento de FNR Chip-seq

```
bwa aln E_index SRX189773_FNR_ChIP.fastq > Escherichia.sai
bwa samse E_index Escherichia.sai SRX189773_FNR_ChIP.fastq > Escherichia.sam
```

- `bwa aln` alinea el archivo de entrada de datos de secuencia con el genoma de referencia y produce un archivo `*.sai` de todos los posibles candidatos.
- `samse` genera los alineamientos en un archivo `.sam` dado un archivo `.sai` y el prefijo dado en `aln` (`E_index` en este caso). Se utiliza en lecturas *single-end*.

1.3.1 Otras alternativas

```
# Opción 1
bwa aln E_index SRX189773_FNR_ChIP.fastq > Escherichia.bwa
bwa samse E_index Escherichia.bwa SRX189773_FNR_ChIP.fastq > Escherichia.sam

# Opcioó 2
bwa mem E_index SRX189773_FNR_ChIP.fastq > Escherichia.sam
```

- `mem` se refiere al algoritmo mem, que es uno de los tres algoritmos proporcionados por BWA. Realiza la alineación local y produce alineaciones para diferentes partes de la secuencia de consulta.

1.4 Análisis de calidad del alineamiento

```
# Opción 1
samtools view -H Escherichia.sam | less -S
# Opción 2
samtools stats Escherichia.sam | grep 'SN'
# Opción 3
samtools flagstat Escherichia.sam
```

2 Para *Mus musculus*

2.1 Control de calidad de las secuencias crudas (*raw sequences*)

FastQC, es una herramienta utilizada para proporcionar una visión general de las métricas básicas de control de calidad para los datos de secuenciación de próxima generación sin procesar. En este caso, los datos crudos de las lecturas se encuentran en el archivo `mus_musculus_CEBPA_liver_ERR005132.fastq.gz`

```
# Analisis fastqc
fastqc mus_musculus_CEBPA_liver_ERR005132.fastq.gz
# Obtener las imagenes a computadora local
rsync -rptuvl arodriguez@dna.lavis.unam.mx:\
/mnt/Timina/bioinfoII/arodriguez/alignment/Musmusculus/fastqc_analisys .
```

La longitud promedio de estas lecturas de de 42 bp, mientras que el porcentaje global de GC en todas las bases de todas las secuencias fue de 45%.

Analizando la **calidad de secuencia por base**, se denota un severo problema, ya que ni siquiera la mitad de las seceuncias tiene una calidad decente (<30, Fig.1). Por otro lado, parece ser que los **puntajes de calidad de cada carril** (*tile*) fueron constantemente aceptables (Fig.2), sin embargo se pueden observar diferentes zonas con colores un poco mas calidos al resto, lo que significa que se observarán sesgos en las secuencias obtenidas. De igual forma, los **puntajes de calidad por secuencia** arrojan resultados inquietantes, ya que se muchos se encuentran cercanos a 6 (un error cada 4 bp) y el promedio general por secuencia es de ~15 phred (Fig.3).

En conclusión, a pesar de que otros parametros no sean significativamente malos, se aconseja fuertemente no considerar ninguna conclusion obtenida en este informe como significativa.

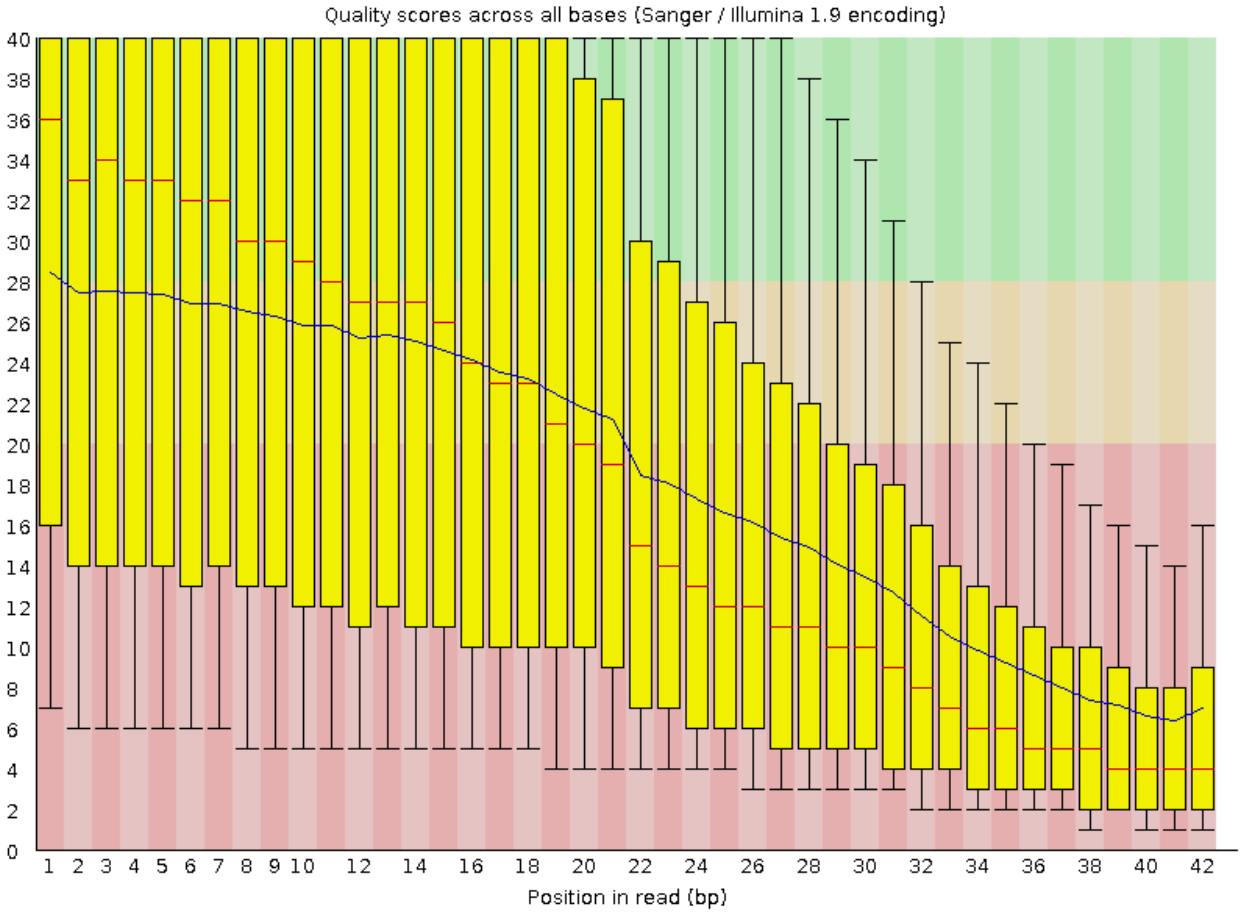

Figure 1: **Calidad de secuencias por base.** La calidad de secuencia por base parece reducirse drasticamente de una puntuacion de calidad media de 28 en la primera base, a 7 en la ultima.

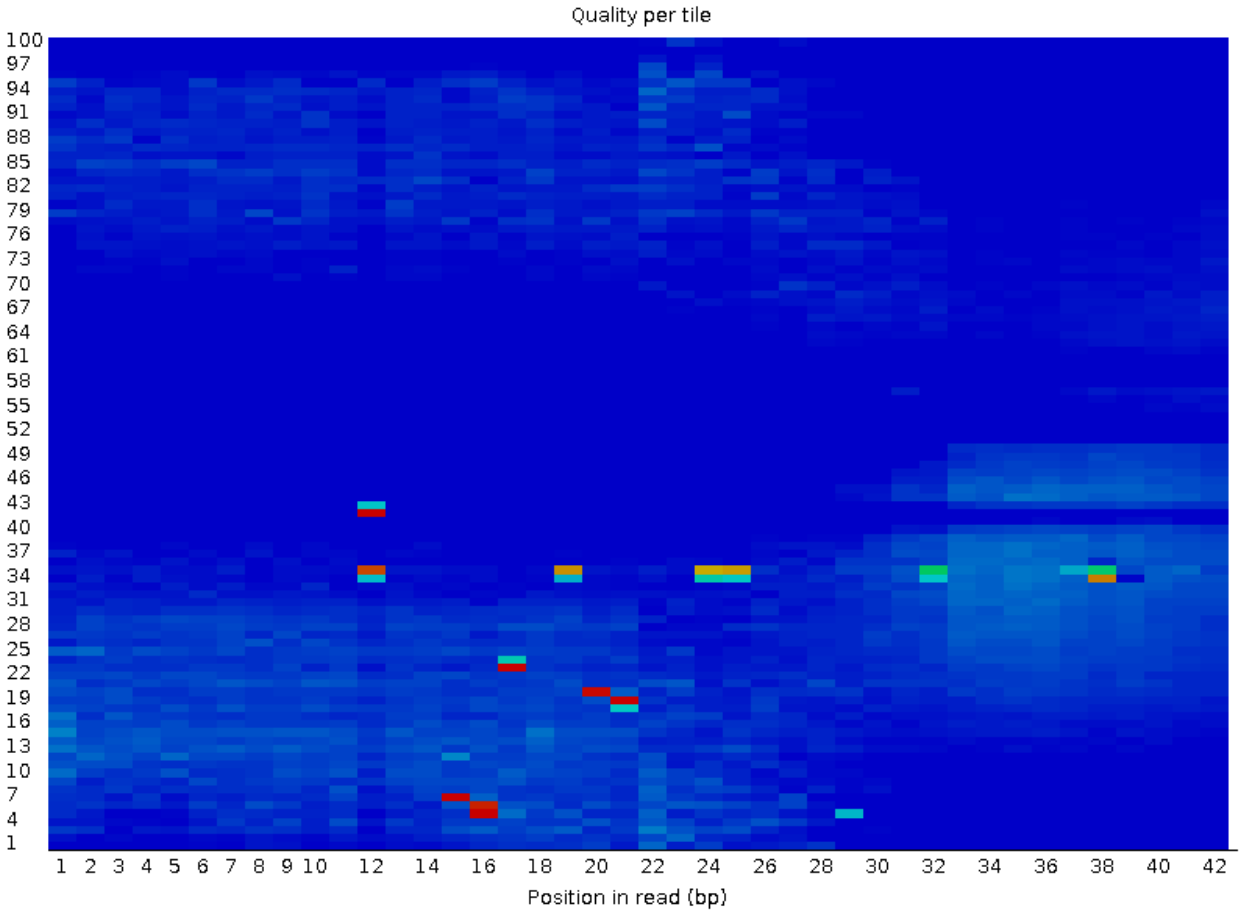

Figure 2: **Calidad por carril.** Las razones para ver advertencias o errores en este gráfico podrían ser problemas transitorios, como burbujas que atraviesan la celda de flujo, o podrían ser problemas más permanentes, como manchas en la celda de flujo o desechos dentro del carril de la celda de flujo.

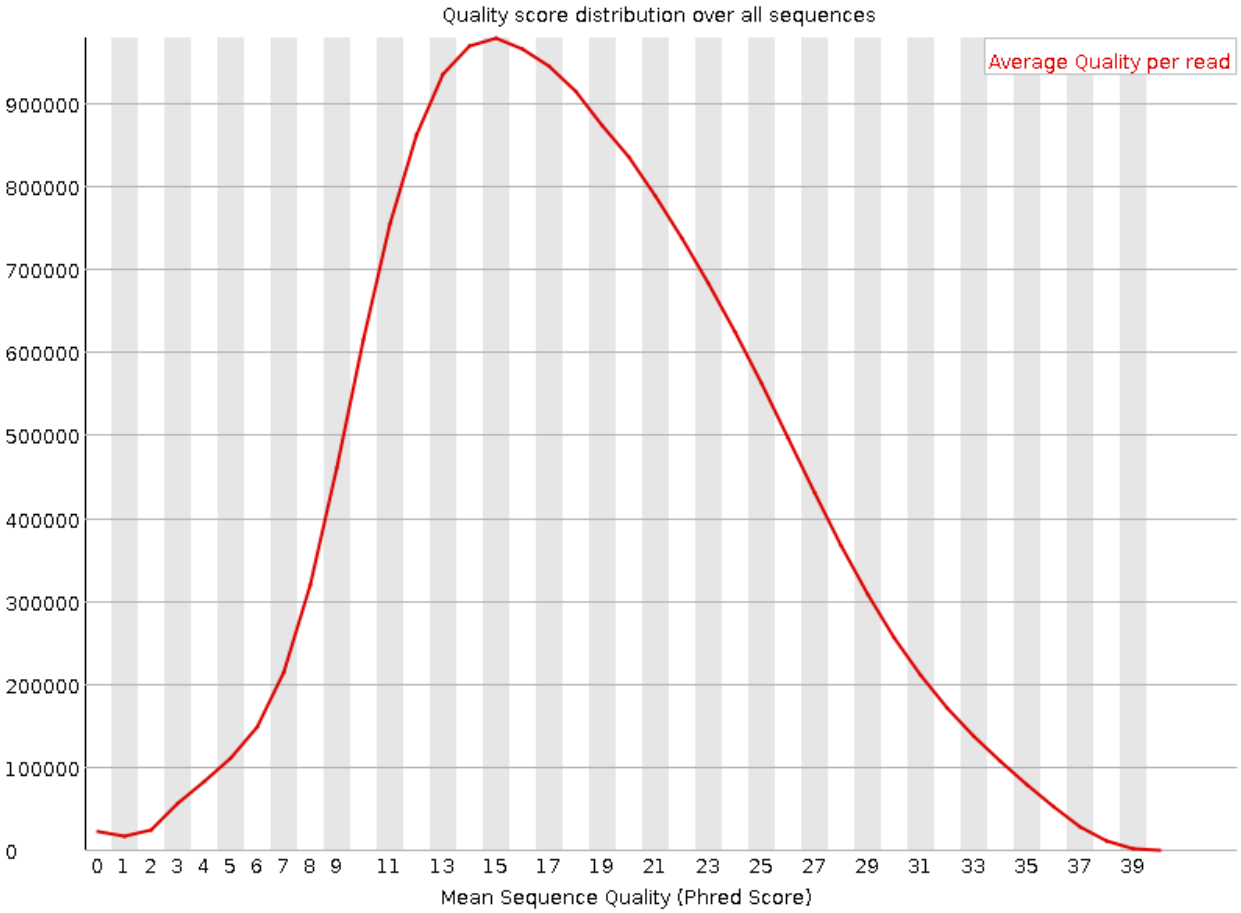

Figure 3: **Puntajes de calidad por secuencia.** Esta grafica permite ver si un subconjunto de sus secuencias tiene valores de calidad universalmente bajos. Se emite una advertencia si la calidad media observada con mayor frecuencia es inferior a 27, lo que equivale a una tasa de error del 0,2%.

2.2 Index para el genoma de *Mus musculus*

Los archivos de *indexación* y el genoma de referencia se encuentran en el siguiente directorio del cluster:

```
/mnt/Archives/genome/mouse/mm10/UCSC/bwa-0.7.15-index/index
```

Donde `mm10.fa` se refiere al genoma de referencia (Genome Reference Consortium Mouse Build 38 (GCA_000001635.2)); mientras que `mm10.fa.amb`, `mm10.fa.ann`, `mm10.fa.bwt`, `mm10.fa.pac`, y `mm10.fa.sa` son los archivos del *index*.

2.3 Alineacion del genoma de raton de un experimento ChiP-seq utilizando bwa

Antes de realizar el alineamiento, verificamos el tamaño de los archivos

```
cd /mnt/Archives/genome/mouse/mm10/UCSC/bwa-0.7.15-index/index
ls -lh
```

Teniendo en cuenta el tamaño de los archivos tanto del genoma de referencia *Mus musculus* como de los datos de ChIP-seq de CEBPA, la memoria disponible para las sesiones de `qlogin` podría no ser suficiente, y probablemente haría que el proceso se suspendiera indefinidamente. Por lo tanto se generó un script sge llamado `MusculusAlignment.sge`, el cual contiene lo siguiente:

```
#!/bin/bash
#
# Use Current working directory
#$ -cwd
#
# Join stdout and stderr
#$ -j n
#
# Run job through bash shell
#$ -S /bin/bash
#
# You can edit the script since this line
#
# Your job name
#$ -N MusculusAlignment
#
# Send an email after the job has finished
#$ -m e
#$ -M axelrdz5205@gmail.com
#
#
# If modules are needed, source modules environment (Do not delete the next line):
. /etc/profile.d/modules.sh
#
# Add any modules you might require:
module load bwa/0.7.15
#
# Write your commands in the next line
```

```
# Generar alineamiento
# trabajando en /mnt/Timina/bioinfoII/arodriguez/alignment/Musmusculus/
bwa mem -M -t 8 /mnt/Archives/genome/mouse/mm10/UCSC/bwa-0.7.15-index/index/mm10.fa \
./mus_musculus_CEBPA_liver_ERR005132.fastq.gz > Mus_alg.sam

# -M deja la mejor (más larga) alineación para una lectura como alineación primaria
#    y alineaciones adicionales para la lectura como secundaria
# -t indica el número de núcleos de procesador utilizados para hacer la alineación
```

2.4 Analisis de calidad del alineamiento

```
# Opción 1
samtools view -H Mus_alg.sam | less -S
# Opción 2
samtools stats Mus_alg.sam | grep 'SN'
# Opción 3
samtools flagstat Mus_alg.sam
```

Primeramente, dado los datos arrojados podemos observar que seguramente existe un error de secuenciación ya que tenemos un promedio de calidad de 18.7%, además, una taza de error mayor a un 1% (2.094056e-02 por cada 1 000 bases mapeadas), de tal forma que la calidad de secuenciación es baja.

Ahora bien, las lecturas mapeadas fueron 6,530,881 y las no mapeadas cuentan con un número mayor, siendo este de 10, 640, 249. Finalmente, la calidad de mapeo fue de MQ = 0, lo cual una vez más nos hace pensar que hubo un error durante la secuenciación.