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 practica

```
# Se crea un screen de la practica 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 modulos
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 indices permiten a las herramientas de alineacion (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 indexacion 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

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

# Opcion 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 Analisis de calidad del alineamiento

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

1.4.1 Radio de Mapeo (*Mapping rate*)

1.4.2 DRadio de duplicacion (*Duplication rate*)

1.4.3 Contenido de GC vs Profundidad de Lectura

1.4.4 Indeles por ciclo (*Indels by cycle*)

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`

```
fastqc mus_musculus_CEBPA_liver_ERR005132.fastq.gz
```

2.1.1 Per Base Sequence Quality

2.1.2 Per Sequence Quality Scores

2.1.3 Per Base Sequence Content

2.1.4 Per Base GC Content

2.1.5 Per Sequence GC Content

2.1.6 Per Base N Content

2.1.7 Sequence Length Distribution

2.1.8 Duplicate Sequences

2.1.9 Overrepresented Kmers

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 llamadao `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_CEBP
```

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

2.4 Analisis decalidad del alineamiento

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

Nota: los archivos de alineamiento se enceuntran en la carpeta `/mnt/Timina/bioinfoII/arodriguez/alignment/Musmusculu`