ReporteT01

Jessica Garcia, Manuel Rivera, Axel Rodriguez

2023-02-10

Contents

1 Practica archivo BAM

```
# Entrar a un qlogin
qlogin
# Entrar a la carpeta donde se encuentra el archivp
cd /mnt/Timina/bioinfoII/format_qc
# Copiar el archivo a la carpeta donde se estará trabajando
cp NA20538.bam /mnt/Timina/bioinfoII/arodriguez
# Dirigirse a la carpeta donde se trabaja
cd /mnt/Timina/bioinfoII/arodriguez
# Cargar Samtools versión 1.9
module load samtools/1.9
# Visualizar información del BAM
samtools view -H NA20538.bam
```

1.1 ¿Qué significan las siglas RG?

RG: *read group*, lecturas de grupo que se generan a partir de una única ejecución de un secuenciador.

1.2 ¿Qué es "lane ID"?

Identificador del grupo de lectura, esta etiqueta identifica a qué grupo de lectura pertenece cada lectura, por lo que el ID de cada grupo de lectura debe ser único.

1.3 ¿Cuál es la plataforma de secuenciación?

Para conocer la plataforma utilizada, existe la etiqueta PL (*platform)*, que se puede encontrar en el encabezado del archivo BAM.

```
# Buscamos la etiqueta PL dentro del header
samtools view -H NA20538.bam | grep 'PL'
```

Para estas muestras se utiizo **Illumina**.

1.4 ¿Qué versión de ensamblaje humano se usó para realizar las alineaciones?

Las entradas @SQ del encabezado contienen información acerca de las secuencias de referencia. La etiqueta AS (*genome assembly identifier*) nos habla acerca del genoma de referencia utilizado para realizar el alineamiento.

```
samtools view -H NA20538.bam | grep 'AS'
```

el genoma de ensamblaje utilizado fue el **NCBI37: GRCh37 (*Genome Reference Consortium Human Build 37*)**, especificamente el descargado del Proyecto de los 1000 genomas.

1.5 ¿Qué programadas fueron utlizados para crear este archivo BAM?

La entrada @PG (*program*) del encabezado contiene la informacion de los programas utilizados.

```
samtools view -H NA20538.bam | grep 'PG'| less -S
```

Los programas utilizados fueron:

- Genome Analysis Toolkit (GATK) IndelRealigner (version 1.0.4487)
- Genome Analysis Toolkit (GATK) TableRecalibration (version 1.0.4487)
- Burrows-Wheeler Aligner (bwa) (version 0.5.5)

1.6 ¿Qué versión de bwa fue usada para alinear las lecturas?

VN *program version*, versión del programa: 0.5.5

```
#Observar información particular del BAM para responder las siguientes preguntas
samtools view NA20538.bam | head
```

1.7 ¿Cuál es el nombre de la primera lectura?

ERR003814.1408899

1.8 ¿En qué posición inicia el alineamiento de la lectura?

19999970

1.9 ¿Cuál es la calidad de la primera lectura?

23

2 Practica archivo bfc

2.1 ¿Que es un BCF?

Un BCF es el archivo binario del VCF.

2.2 ¿Como convertir un BCF–>VCF?

Si es posible convertir un bcf a vcf utilizando el comando *view*, el argumento $-o$ permite indicar el nombre del archivo de salida.

```
#Convertir un BCG --> VCF
bcftools view 1kg.bcf -o 1kg.vcf
```

2.3 ¿Cuantas muestras hay en el BCF?

Hay 50 muestras en total.

```
# Muestra los nombres de las muestras
bcftools query -l 1kg.vcf

# Muestra el numero de muestras con un pípeline
bcftools query -l 1kg.bcf | wc -l
```

2.4 ¿Cual es el genotipo de la muestra HG00107 en la posicion 20:24019472?

El genotipo es "A/T".

```
#Comprimir VCF --> GZIP
bcftools view 1kg.vcf -Oz -o 1kg.vcf.gz

#Filtrar por nombre de la muestra
#-s es para filtrar por el nombre de la muestra
bcftools view -s HG00107 1kg.vcf.gz

#Indexar el archivo GZIP, esto es necesario para filtrar por posiciones
bcftools index 1kg.vcf.gz

#Filtrar por posicion de la muestra y nombre
bcftools view -r 20:24019472 -s HG00107 1kg.vcf.gz

#Mostrar el genotipo
bcftools view -r 20:24019472 -s HG00107 1kg.vcf.gz | bcftools query -f '[ %TGT]\n'
```

2.5 ¿Cuantas posiciones tienen mas de 10 alelos diferentes?

Hay 4822 posiciones que tienen mas de 10 alelos diferentes.

```
# Saber posiciones
bcftools view -i 'AC>10' 1kg.vcf| wc -l #DUDA arroja mas

# Comando duda
bcftools query -f '%INFO/AC\n' 1kg.bcf -i'AC>10' | wc -l
```

2.6 Lista todas las posiciones en las que HG00107 no tiene un genotipo de referencia y la cobertura es mayor a 10.

Hay 451 posiciones en las que no hay un genotipo de referencia y tienen cobretura mayor a 10.

```
# Mostrar la cobertura
bcftools query -f '[ % DP]\n'

# Filtrar por nombre
```

```
bcftools view -s HG00107 1kg.vcf

# Filtrar por no tener genoma de referencia
bcftools view -i 'GT="alt"' 1kg.vcf

# Ambos filtros
bcftools query -s HG00107 -f '[%POS]\n' -i 'GT ="alt" & DP>10' 1kg.vcf.gz

#Contar las posiciones
bcftools query -s HG00107 -f '[%POS]\n' -i 'GT ="alt" & DP>10' 1kg.vcf.gz | wc -l

#DUDA EN COMANDO
bcftools view -s HG00107 | query -i ' REF="." && DP>10' 1kg.vcf | wc -l  #arroja menos
```

3 Practicas estadisticas

3.1 ¿Cual es numero total de lecturas?

Usando el comando `samtools flagstat` se puede observar las estadisticas generales del archivo BAM. Si lo que se requiere saber es el **numero total de lecturas**, entonces se utiliza el comando:

```
samtools flagstat NA20538.bam | head -n 1
```

En donde el comando `head -n 1` funciona para imprimir solo la primera linea del output, donde se encuentran el numero de lecturas totales.

```
347367 + 0 in total (QC-passed reads + QC-failed reads)
```

Como se observa, el numero de lecturas totales fue de **347,367**.

3.1.1 Otras alternativas

```
# Estadisticas generales
samtools flagstat NA20538.bam
# Lecturas totales ejemplo 2
samtools view -c NA20538.bam # 347367
# Lecturas totales ejemplo 3
samtools stats NA20538.bam | grep 'SN' | cut -f 2- # 347367
# Lecturas totales ejemplo 4
samtools stats NA20538.bam | grep 'raw total sequences' | cut -f 2- # 347367
```

- `view -c` cuenta los alineamientos e imprime el total.
- `samtools stats` muestra estadísticas relevantes sobre cada lane y grupo de lectura, así como información sobre las secuencias.
- `grep` busca e imprime las lineas coincidentes con un patron (en el ej.3 `'SN'`)
- La seccin `SN` del comando `samtools stats` brinda un resumen con conteos, porcentajes y promedios, en un estilo similar al de `samtools flagstat`, pero más completo.
- `cut` se utiliza para seleccionar una columna del output. Con el parametro `-f 2-` le indicamos que solo queremos las ultimas dos columnas del output.
- `grep 'raw total sequences'` se utiliza para especificar que solo se requieren el total de lecturas, lo que nos da un output mas ordenado.

3.2 ¿Que proporcion de las lecturas fueron mapeadas?

Con el mismo comando de `samtools flagstat NA20538.bam`, se observa que el porcentaje de *reads* mapeados fue de **93.26%** (323,966 lecturas).

```
# Filtramos los resultados que coincidan con 'mapped' e imprimimos
# todos los argumentos de la primera fila
samtools flagstat NA20538.bam | grep 'mapped'| awk '{print $0}'| head -n 1
323966 + 0 mapped (93.26% : N/A)
```

El comando `awk '{print $0}'` imprimir la primera fila del output del comando `samtools flagstat NA20538.bam | grep 'mapped'`.

Generalmente `awk '{print $n}'` es utilizado para imprimir la n-esima columna de un output. `awk 'NR==m {print $n}'` se utiliza para imprimir la m-esima fila y la n-esima columna.

3.2.1 Otras alternativas

```
# Lecturas mapeadas (Mapped alignments)

## Ejemplo 2
samtools view -F 0x904 -c NA20538.bam # 323966
## Ejemplo 3
samtools view -c -F 260 NA20538.bam # # 323966
## Ejemplo 4
samtools view -F 0x04 -c NA20538.bam # 323966
## Ejemplo 5
samtools stats NA20538.bam | grep 'SN' | grep 'reads mapped' | cut -f 2- #323966
```

3.3 ¿Cuantas lecturas fueron mapeadas a un cromosoma diferente?

```
samtools stats NA20538.bam | grep 'pairs on different chromosomes:' | cut -f 2-
```

De acuerdo a la seccion `SN` del archivo, **4,055** lecturas fueron mapeadas a un cromosoma diferente.

3.4 ¿Cual es el tamaño de inserción promedio y su desviación estándar?

```
samtools stats -F SECONDARY NA20538.bam | grep "insert size" | cut -f 2-
```

Finalmente, en la seccion de *Sumary Numbers* (`SN`) generada por `samtools stats` se encuentra la información requerida.

- tamaño promedio de inserción: **190.3**
- desviación estándar del tamaño de inserción: **136.4**

4 Referencias

1. Copiar en Linux: con CP es muy sencillo. (s/f). IONOS Digital Guide. Recuperado el 20 de febrero de 2023, de https://www.ionos.mx/digitalguide/servidores/configuracion/comando-cp-de-linux/

2. Markdown: introducción al lenguaje de marcado. (s/f). IONOS Digital Guide. Recuperado el 20 de febrero de 2023, de https://www.ionos.es/digitalguide/paginas-web/desarrollo-web/tutorial-de-markdown/

3. The SAM/BAM Format Specification Working Group. (s/f). Sequence alignment/map format specification. Github.io. Recuperado el 20 de febrero de 2023, de https://samtools.github.io/hts-specs/SAMv1.pdf

4. Anónimo. (s/f). VCF AND BCF. Evolution and Genomics. Recuperado el 20 de febrero de 2023, de https://evomics.org/vcf-and-bcf/#:~:text=BCF%2C%20or%20the%20binary%20variant,that%20between%20BAM%2