UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Ensamblaje y mapeo de secuencias NGS utilizando Galaxy. Manejo de archivos SAM y BAM.

Apellidos, nombre	Cardona Serrate, Fernando (fcardona@btc.upv.es)
Departamento	Departamento de Biotecnología. Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural.
Centro	Universitat Politècnica de València

1 Resumen de las ideas clave

En este artículo vamos a describir paso a paso cómo analizar los datos de secuenciación de segunda generación usando el servidor Galaxy. En concreto, estudiaremos cómo analizar los archivos de secuencia mapeados con el genoma de referencia (SAM o BAM) que obtenemos en secuenciación de nueva generación (NGS, de *Next Generation Sequencing*).

2 Objetivos

Una vez que el estudiante lea con detenimiento este documento, será capaz de:

- Comprender el contenido de los archivos SAM y BAM.
- Visualizar y analizar los archivos SAM y BAM con los programas IGV y iobio.
- Marcar duplicados en archivos SAM y BAM.
- Analizar la calidad de los archivos SAM y BAM.
- Filtrar los archivos SAM y BAM por calidad.

3 Introducción

La secuenciación de ADN ha sido revolucionaria en los campos de conocimiento de la biología molecular y la genética, ya que permite conocer la secuencia y estructura de los genes, así como las de otros elementos reguladores importantes en la regulación de la expresión génica. En el caso de humanos, es ampliamente conocido el proyecto genoma humano (1990-2001), que permitió conocer casi la totalidad de la secuencia del genoma humano. En este proyecto se utilizó el método de Sanger, conocido también primera generación de secuenciación [1].

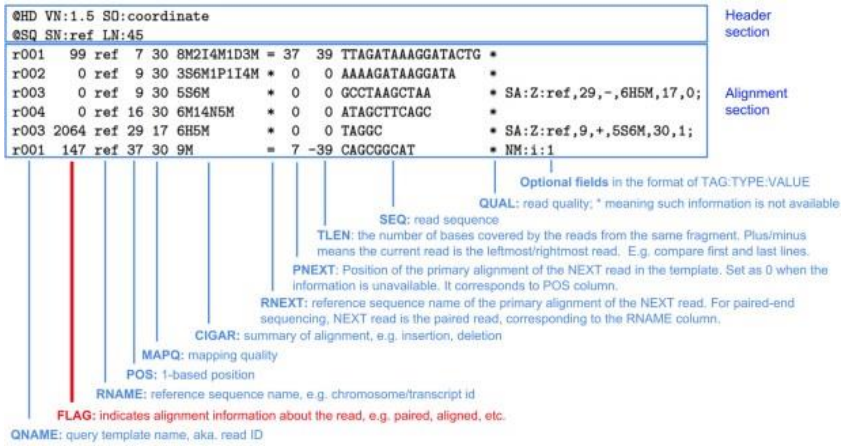
Más recientemente (2008), las técnicas de secuenciación de segunda generación han permitido una secuenciación mucho más rápida de genomas humanos completos, en cuestión de días [2].

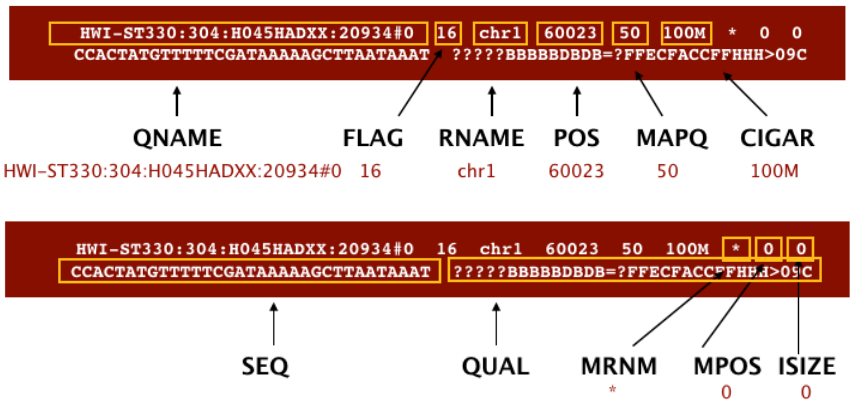
Por último (2010), los secuenciadores de tercera generación permiten obtener secuencias mucho más largas que los de segunda generación, facilitando el ensamblaje y análisis informático, y además no necesitan de un enriquecimiento previo de la muestra [3].

Todos los secuenciadores, a veces tras el paso inicial por un sistema de archivos propio, acaban dando las secuencias en formato FASTQ. Este tipo de archivos está basado en texto, y contienen, además de la secuencia de nucleótidos como texto (formato FASTA), las correspondientes puntuaciones de calidad de secuencia para cada nucleótido codificada en formato ASCII.

El procesado de estas secuencias en formato FASTQ implica un paso de alineado o mapeo con el genoma de referencia del organismo correspondiente, de manera que pueda añadirse la información de coordenadas genómicas. Es decir, el análisis de la secuencia implica el paso a archivos SAM (*Sequence Alignment Map*) o BAM (*Binary Alignment Map*, SAM comprimido).

El formato SAM (*.sam) es un tipo de archivo que se encuentra delimitado por tabulaciones (TAB-delimited text). El esquema de este formato se divide en las líneas de cabecera que son opcionales y están precedidas por el símbolo "@", y las líneas de alineamiento. Además, existen otros campos opcionales que pueden aparecen en un archivo SAM que pueden aparecer en cualquier orden (imagen 1).





Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	*|[:rname:^*=][:rname:]*	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[:rname:^*=][:rname:]*	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Imagen 1. Ejemplo y estructura de un archivo SAM/BAM.
Fuentes: Bioinformáticamente [4]; SAM tools GitHub [5]; Learning the BAM format [6]

BIT		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented the first segment in the template
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

OP	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference
=	7	sequence match
X	8	sequence mismatch

Imagen 2. Interpretación de un archivo SAM/BAM.
Fuentes: SAM tools GitHub [5]; Learning the BAM format [6]; File Formats Tutorial [7].

Un archivo BAM (*.bam) es la versión binaria comprimida de un archivo SAM que contiene secuencias alineadas de hasta 128 Mb. Los archivos BAM emplean el formato de nomenclatura de archivos SampleName_S#.bam, donde # es el número de la muestra determinado por el orden en que aparecen las muestras en el experimento. Los archivos BAM contienen una sección de encabezado y una sección de alineaciones:

Encabezado: contiene información sobre todo el archivo como, por ejemplo, el nombre de la muestra, su longitud y el método de alineación. Las alineaciones de la sección de alineaciones están asociadas a la información específica de la sección del encabezado.

Alineaciones: contiene el nombre, la secuencia y la calidad de la lectura e información sobre la alineación, así como etiquetas personalizadas. El nombre de la lectura incluye el cromosoma, la coordenada de inicio, la calidad de la alineación y la secuencia del descriptor de la coincidencia. La sección de alineaciones contiene la siguiente información por cada lectura o par de lectura (se describen únicamente aquellas cuyo significado no es evidente):

FLAG: Indica un código numérico que nos dice cómo la lectura observada en la presente línea fue alineada sobre el genoma de referencia. Esta columna es esencial para obtener posteriormente estadísticas relativas a la calidad del alineamiento. Por ejemplo, un valor

FLAG igual a 4. ¿Qué significa? Pues bien, a través de la tabla de valores (imagen 2) sabemos que este valor indica que dicha lectura no ha sido mapeada ya que no se ha encontrado ningún punto en el genoma con el que alinearse.

POS: Indica la posición inicial de alineamiento sobre la referencia con un número que indica la posición del primer nucleótido del alineamiento. Si el valor es cero significa que la lectura no ha sido mapeada para confirmar el valor 4 colocado en la columna FLAG.

MAPQ: Indica el valor de calidad del alineamiento. Equivale a -10 log10 Pr {posición de mapeo incorrecta}, redondeado al entero más próximo. Un valor 255 indica que la calidad de mapeo no está disponible. Así, si supiera que la probabilidad de mapear correctamente una lectura aleatoria es de 0,99, la puntuación MAPQ sería de 20 (es decir, log10 de 0,01 * -10). Si la probabilidad de una coincidencia correcta aumentara a 0,999, la puntuación MAPQ aumentaría a 30. Por tanto, el límite superior de una puntuación MAPQ depende del nivel de precisión de su probabilidad (aunque suele definirse un límite superior de 255). A la inversa, a medida que la probabilidad de una coincidencia correcta tiende a cero, también lo hace la puntuación MAPQ.

CIGAR: En esta columna encontramos una cadena formada por un número entero y una letra, que hace referencia a una operación, que en conjunto resumen la información relativa al alineamiento. Esto es muy útil ya que permite visualizar gráficamente el alineamiento de las lecturas sobre la referencia. Existe una tabla (imagen 2) que relaciona cada letra que podemos encontrar en la cadena. Pongamos un ejemplo, supongamos que tenemos una cadena 76H130M; esto significa que 130 bases de la lectura considerada han sido alineadas con la referencia mientras que 76 bases restantes no han sido alineadas.

RNEXT: Indica el nombre de la lectura que está en *paired-end* con la lectura considerada. El símbolo «*» indica que no hay información disponible mientras que el símbolo «=» indica que la lectura en emparejamiento tiene el mismo ID (nombre) que la lectura de esa línea.

PNEXT: Indica la posición inicial de la lectura que está emparejada con la lectura considerada.

TLEN: Representa la longitud del segmento de referencia mapeado por las dos lecturas *paired-end*.

QUAL o PHRED: Expresa el valor de calidad relacionado con la secuenciación de la lectura, es decir, expresa la probabilidad de tener un error al asignar las bases durante la secuenciación.

Los archivos de índice BAM (*.bam.bai) facilitan un índice del archivo BAM correspondiente.

El servidor Galaxy [8] es un sistema gratuito y de código abierto para el análisis de datos, la creación de flujos de trabajo, la formación y la educación, la publicación de herramientas, la gestión de infraestructuras, entre otros, que facilita el análisis de secuencias NGS sin necesidad de tener conocimientos avanzados de bioinformática. Esta herramienta aglutina distintas herramientas bioinformáticas, de análisis NGS y de otros tipos (por ejemplo, estructura de proteínas), que procesan los datos en su propia nube y están siempre accesibles. De esta forma, es una herramienta muy adecuada para este tipo de procesos en usuarios no avanzados con recursos de hardware limitados y/o pocos conocimientos informáticos. Es posible darse de alta en el sistema de forma gratuita con una cuenta académica en https://usegalaxy.org/.

4 Desarrollo

Con el objetivo de obtener los archivos BAM a partir de los archivos FASTQ, es necesario alinear estos con el genoma de referencia, en el caso de este ejemplo el genoma humano, en este caso utilizaremos la herramienta BWA-MEM en Galaxy. Para ello, cargar los archivos FASTQ de calidad suficiente en Galaxy en la herramienta BWA-MEM como *paired-end* utilizando las secuencias F y R de cada archivo (pueden hacerse todos a la vez), y posteriormente realizar el alineamiento según se indica en la **imagen 3**.

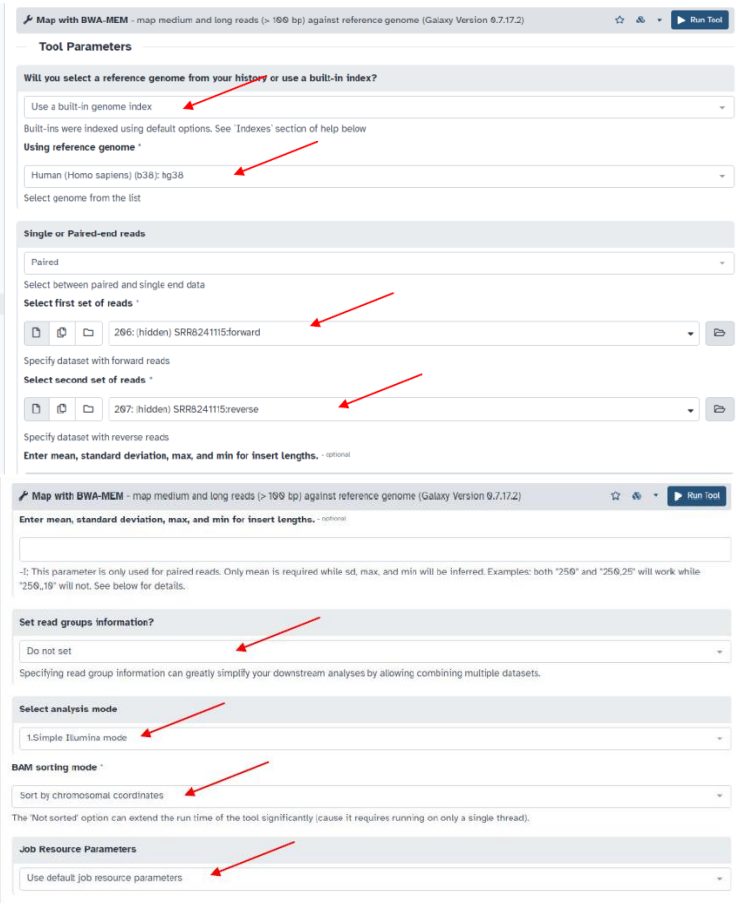


Imagen 3. Carga de archivos de secuencias FASTQ en Galaxy para su alineamiento con BWA-MEM.

Posteriormente se puede visualizar utilizando IGV, preferiblemente la versión "Desktop" (imagen 4). También puede usarse la versión "web" o la integrada en Galaxy. También es útil (y rápido) verlo directamente en la web de la base de datos de UCSC (https://genome-euro.ucsc.edu/cgi-bin/hgGateway?redirect=manual&source=genome.ucsc.edu) accediendo desde el mismo Galaxy, desde los botones habilitados para ello en Galaxy (imagen 5).

4.1 Marcaje de duplicados en los archivos BAM

Para marcar los duplicados en los archivos BAM podemos utilizar la herramienta *MarkDuplicates* de Galaxy, como se muestra en la imagen 6. Esta herramienta localiza y etiqueta lecturas duplicadas, es decir originadas a partir de un único fragmento de ADN, en

un archivo BAM o SAM. Los duplicados pueden surgir durante la preparación de librerías mediante PCR, o bien ser el resultado de un único clúster de amplificación, que se detecta como múltiples clústeres por el sensor óptico (duplicados ópticos).

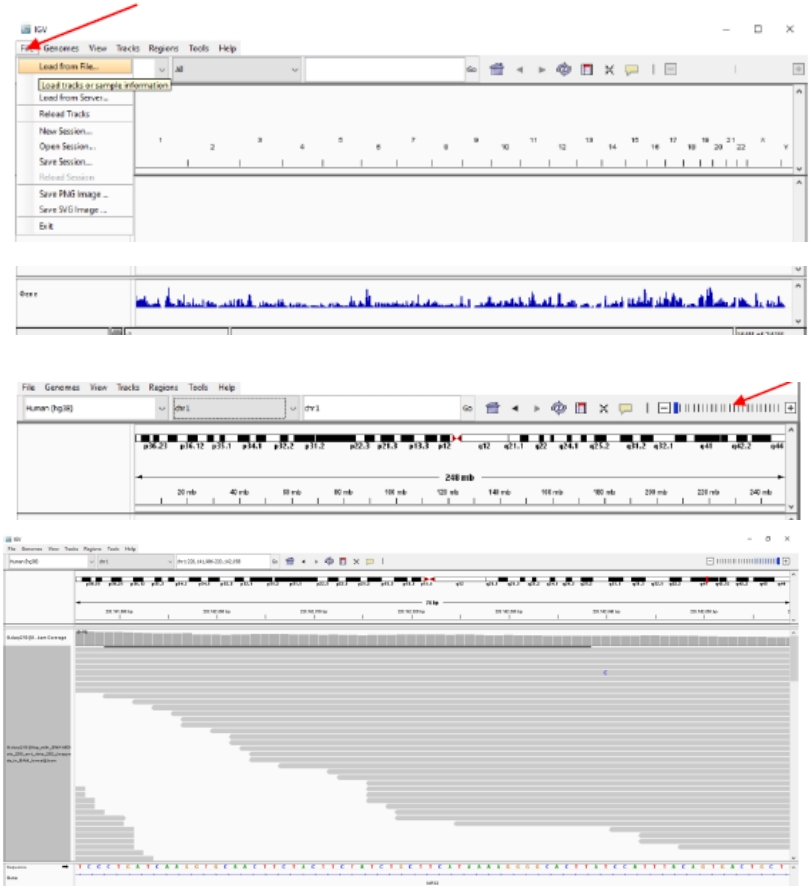


Imagen 4. Visualización de archivos BAM utilizando la versión de escritorio de IGV.

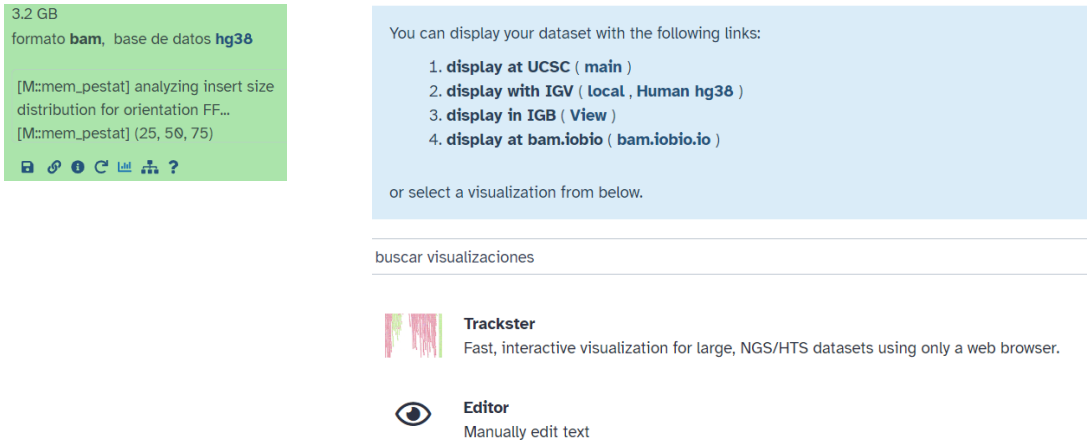


Imagen 5. Opciones de visualización de archivos BAM en Galaxy.

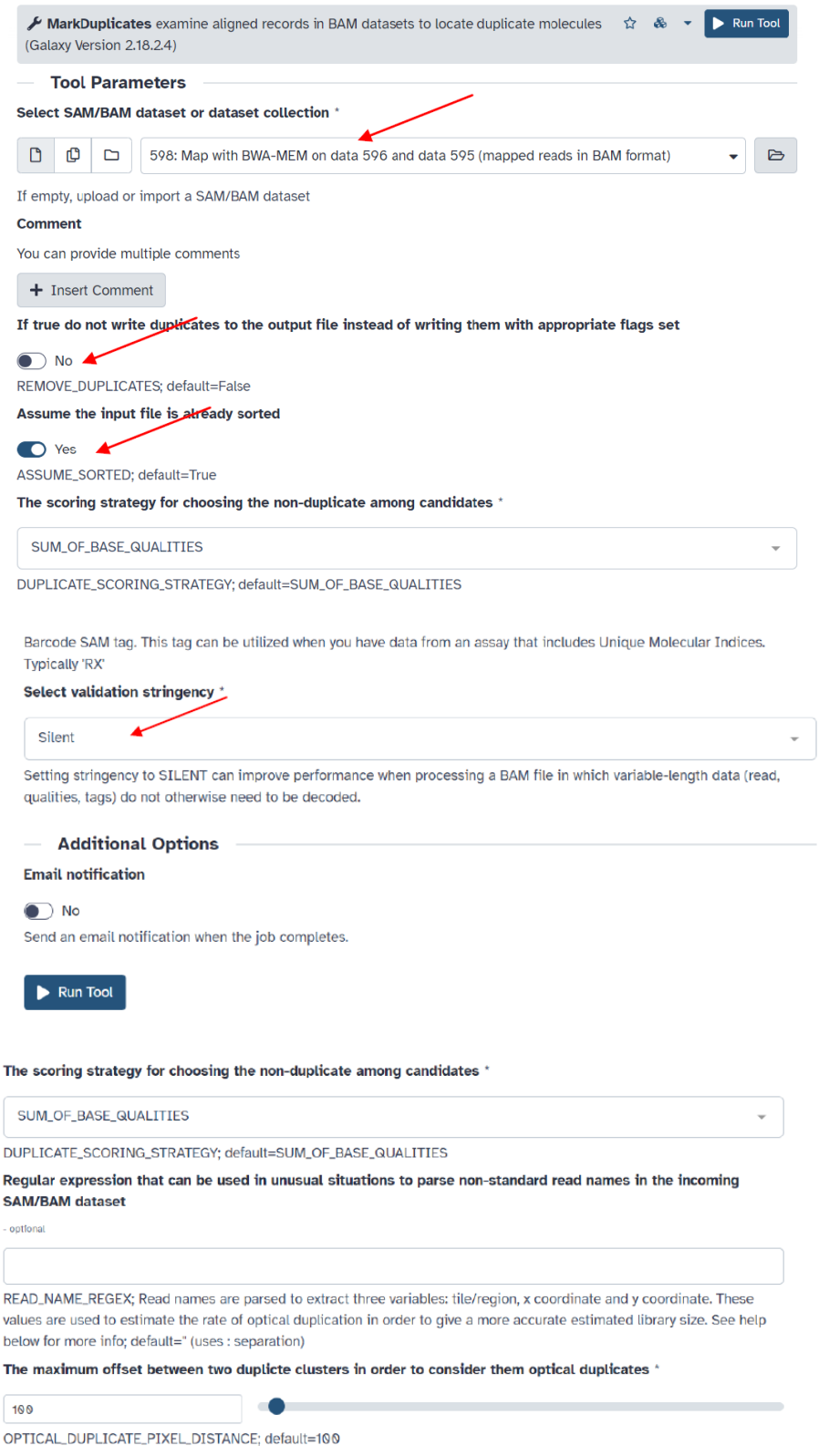


Imagen 6. Marcaje de duplicados utilizando MarkDuplicates en Galaxy.

Imagen 7. Visualización del mapeo de los archivos BAM en la herramienta Iobio.

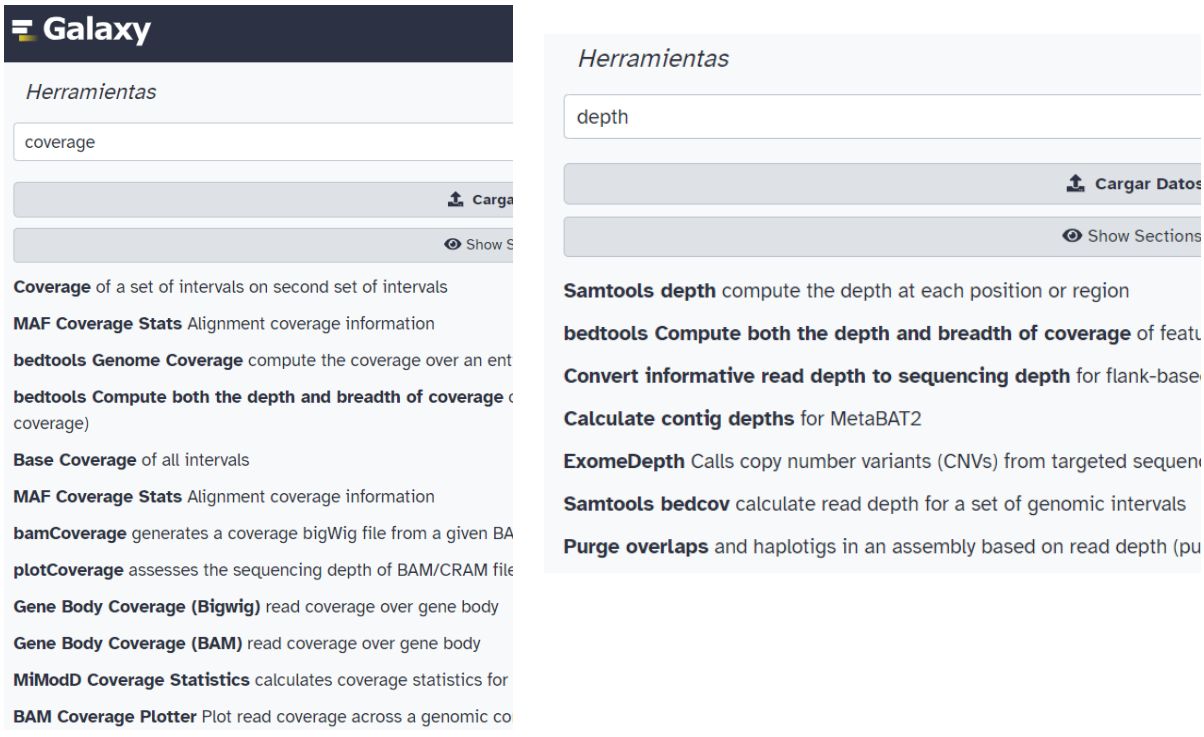


Imagen 8. Herramientas disponibles en Galaxy para analizar la cobertura y la profundidad de lectura de los mapeos.

4.2 Análisis de la calidad del mapeo en los archivos BAM

Normalmente conviene analizar la calidad de los mapeos de los archivos BAM, lo que también puede hacerse utilizando las herramientas disponibles en Galaxy. Puede verse la calidad aproximada del mapeo utilizando la herramienta integrada en Galaxy (bam.iobio) como se indica en la imagen 7. Las mimas visualizaciones pueden realizarse en IGV o UCSC, como ya se ha indicado anteriormente.

De esta manera puede observarse la cobertura del mapeo, tanto visualmente como por los valores obtenidos. Para continuar con los análisis el mapeo debe cubrir la zona a secuenciar con una profundidad de lectura suficiente, lo que se considera mínimo 20X (20 secuencias de la zona). Existen también una serie de herramientas en Galaxy que nos sirven para analizar la cobertura de nuestro mapeo y la profundidad de lectura (imagen 8). Utilizaremos como ejemplos *BAM coverage plotter* y *Samtools stats*. Bam coverage (imagen 9) nos genera un gráfico vectorial en formato .svg. Samtools statas (imagen 10) nos genera un archivo .tsv (texto compatible con excel) con los parámetros del mapeo.

5 Cierre

A lo largo de este objeto de aprendizaje hemos visto qué formato tienen los archivos de mapeo de secuencia SAM y BAM, además de cómo se pueden visualizar y analizar su calidad, principalmente analizando la cobertura y la profundidad de lectura. Todo ello podemos hacerlo utilizando la herramienta Galaxy y sin necesidad de conocimientos bioinformáticos avanzados.

6 Bibliografía

1- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature.* **409**, 860–921 (2001). https://doi.org/10.1038/35057062

2- Wadman, M. James Watson's genome sequenced at high speed. *Nature.* 452, 788 (2008). https://doi.org/10.1038/452788b

3- Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*. **7(6)**, 461-465 (2010). https://doi.org/10.1038%2Fnmeth.1459

4- Bioinformáticamente. https://bioinformaticamente.com/2021/03/03/sam-bam/

5- Sequence Alignment/Map Format Specification. The SAM/BAM Format Specification Working Group. 2023. SAM tools GitHub https://samtools.github.io/hts-specs/SAMv1.pdf

6- Learning the BAM format https://bookdown.org/content/24942ad6-9ed7-44e9-b214-1ea8ba9f0224/learning-the-bam-format.html

7- File Formats Tutorial https://bioinformatics.uconn.edu/resources-and-events/tutorials-2/file-formats-tutorial/#fileformats_bam

8- The Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Research*. gkae410 (2024). https://doi.org/10.1093/nar/gkae410