

Máster en Bioinformática

Secuenciación Genómica y Análisis De Variantes Para Medicina Personalizada y De Precisión

Curso académico 2025-25
Edición Abril



Universidad
Internacional
de Valencia

Dra. Laura Gutiérrez Macías
laura.gutierrez.m@professor.universidadviu.com

14/07/2024

De:
 Planeta Formación y Universidades

Capítulo 4. ¿Cómo analizamos un genoma eucariota? Análisis bioinformático de paneles de captura de genoma humano

4.1.

- Análisis primario. Calidad y filtrado de secuencias.
 - 4.1.1.- ¿Por qué es importante evaluar la calidad de las secuencias y filtrarlas?
 - 4.1.2.- Herramientas de control de calidad
 - 4.1.3.- Análisis de la calidad con FastQC
 - 4.1.4.- Herramientas de filtrado por calidad de las secuencias y eliminación de adaptadores

4.2.

- Mapeo de secuencias. Herramientas de mapeo, visualización y análisis de calidad.
 - 4.2.1.- Proceso de mapeo
 - 4.2.2.- Herramientas para el mapeo
 - 4.2.3.- Archivos de mapeo: formato SAM
 - 4.2.4.- Herramientas para el manejo de archivos SAM. Generación de archivos BAM.
 - 4.2.5.- Visualización del mapeo.
 - 4.2.6.- Análisis de calidad del mapeo. 3.3.

4.3.

- Identificación de variantes
 - 4.3.1.- Preprocesamiento: identificación de secuencias analizadas
 - 4.3.2.- Identificación de variantes
 - 4.3.3.- Post-procesado: Filtrado de variantes y etiquetado

4.4.

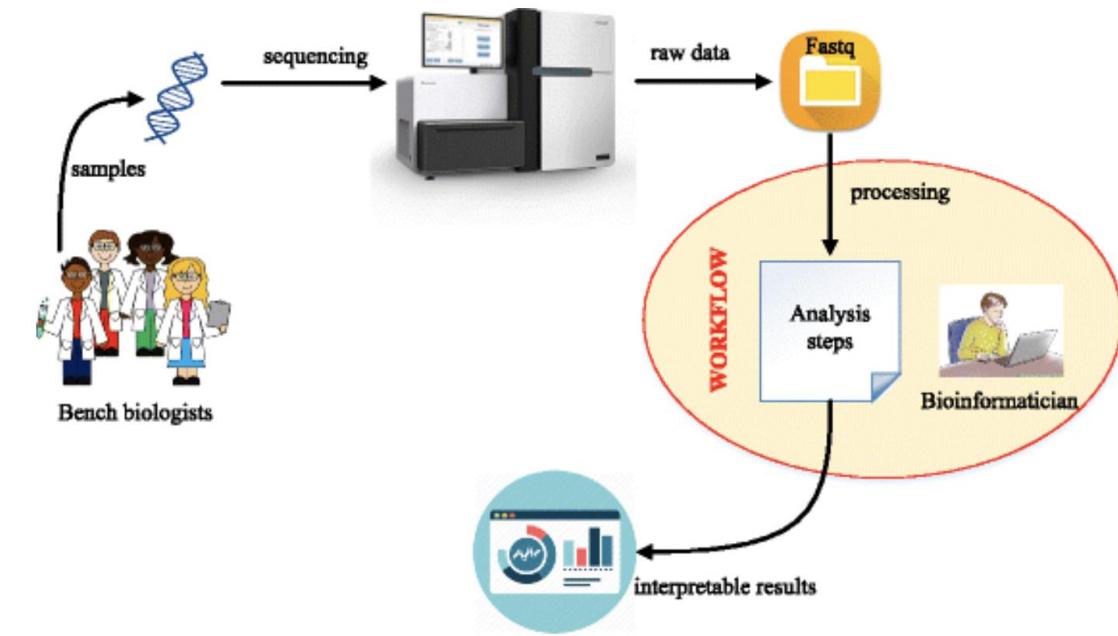
- Anotación de variantes
 - 4.4.1.- Variant Effect Predictor (VEP)
 - 4.4.2.- Annovar

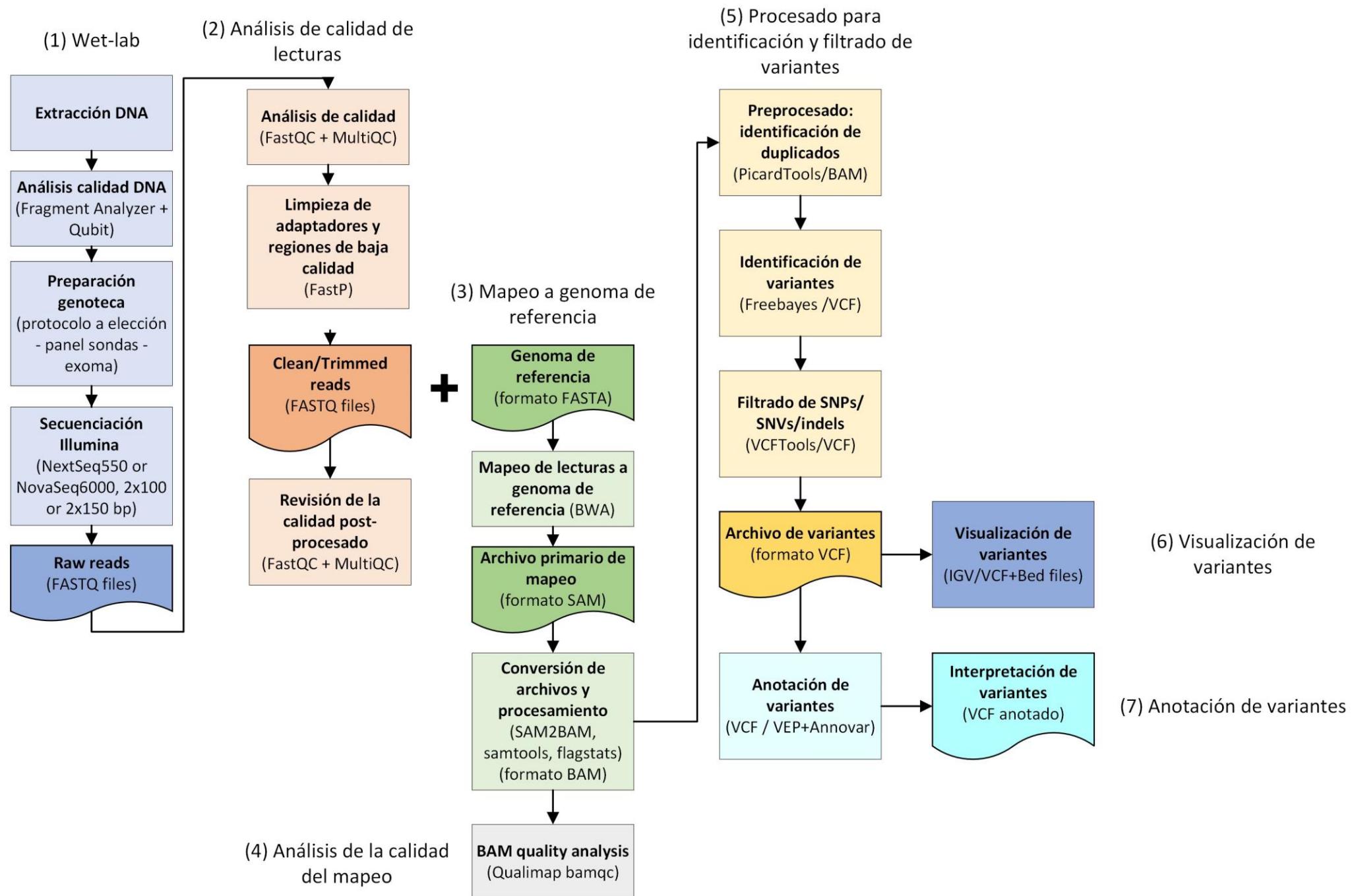
Secuenciación: proceso optimizado, bajo coste por base secuenciada (y bajando). No es un limitante en los proyectos de secuenciación ahora mismo.

Análisis bioinformático: es el principal cuello de botella. Presenta varios retos:

- Es un proceso complejo con muchos pasos
- Automatización
- Optimización: obtener buenos rendimientos de tiempo
- Cada etapa debe diseñarse al detalle
- Usa herramientas y bases de datos que deben mantenerse y actualizarse (reanálisis de muestras)

En este capítulo vamos a ver las etapas generales de un análisis de genoma eucariota
Objetivo: encontrar variantes de interés para diagnóstico clínico





Comandos generales CONDA

Comando	Explicación
conda create -n envX python=3.8	crear un entorno llamado envX basado en python 3.8
conda activate envX	activar el entorno llamado envX
conda env export > envX.yml	exportar el entorno envX y sus dependencias a un archivo YML (texto)
conda env create -n envX --file=envX.yml	crear un entorno llamado envX a partir de un archivo YML
conda remove -n envX --all	eliminar el entorno envX y todas las dependencias
conda clean --all	eliminar dentro de un entorno el caché, archivos bloqueados, no utilizados, archivos log...
conda info --envs	listar los environments disponibles
conda remove --force package	eliminar un paquete (package) de un environment
conda list	listar todos los paquetes instalados en un environment conda
conda install -c chanel package	instalación mediante el canal chanel del paquete llamado package

Environment*	Programa	Comando descarga	Versión	Utilidad	Anotaciones (extras a instalar, referencias...)
04MBIF_humano	IGV	conda install -c bioconda igv	v2.4.9-0	visualización de genomas	
	FastQC	conda install -c bioconda fastqc	v0.12.1	Análisis primario de calidad	
	MultiQC	conda install -c bioconda multiqc	v1.19	Integrar archivos de calidad, mapeo, etc de diferentes análisis	
	FastP	conda install -c bioconda fastp	v0.23.4	Limpieza de adaptadores y trimming	
	BWA	conda install -c bioconda bwa	v0.7.17	mapeo de lecturas sobre genoma de referencia	
	Samtools	conda install -c bioconda samtools	v1.19	manejo de archivos SAM/BAM	
	Qualimap	conda install -c bioconda qualimap	v2.3	Análisis de calidad de archivos de mapeo	
	PicardTools	conda install -c bioconda picard	v3.1.1	Manipulación de secuencias a partir de archivos BAM/SAM/CRAM y VCF	https://broadinstitute.github.io/picard/
	Freebayes	conda install -c bioconda freebayes	v1.3.7	Software para la localización de variantes/polimorfismos	
	RTG-Tools	conda install -c bioconda rtg-tools	v3.12.1	Programa para el manejo y comparación de archivos VCF	https://hpc.nih.gov/apps/rtg-tools.html
	VCFTools	conda install -c bioconda vcftools	v0.1.16	Herramientas para el manejo de archivos VCF	https://vcftools.sourceforge.net/
	conda env export --file=04MBIF_humano.yml				

Archivo YML
disponible en
aula virtual

4.1



Análisis primario. Calidad y filtrado de secuencias.

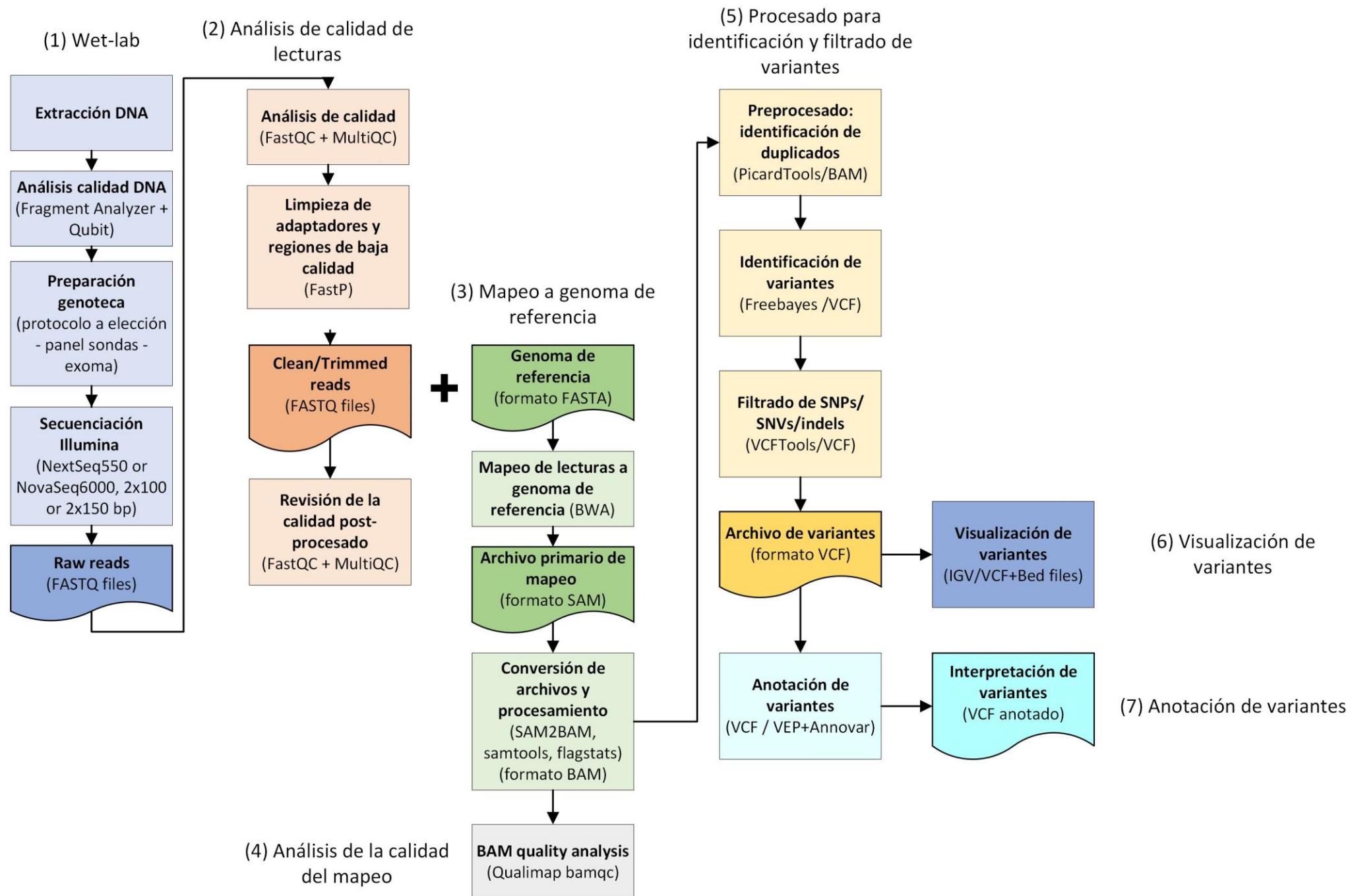
4.1. Análisis primario. Calidad y filtrado de secuencias

4.1.1. ¿Por qué es importante evaluar la calidad de las lecturas y filtrarlas?

4.1.2. Herramientas de control de calidad

4.1.3. Análisis de calidad con FastQC

4.1.4. Herramientas de filtrado por calidad de las secuencias y eliminación de adaptadores.



4.1.1. ¿Por qué es importante evaluar la calidad de las lecturas y filtrarlas?

Como ya hemos dicho, evaluar esta calidad **es el primer paso de todo análisis bioinformático** y debe ser llevado a cabo de manera **obligatoria**, independientemente del objetivo del estudio y del tipo de muestra secuenciado. Los puntos a tener en cuenta son:

- Debemos **evaluar la calidad de la carrera de secuenciación** y preguntarnos si hemos obtenido el **rendimiento** de datos esperado, mayor o menor, así como **evaluar si la concentración de las muestras secuenciadas ha sido adecuada**. Éste es un trabajo conjunto con el departamento de genómica que haya preparado las genotecas y secuenciado las muestras.
- Evaluar **la calidad de las lecturas** nos permitirá valorar cómo de fiables serán nuestros resultados o conclusiones del experimento, incluyendo evaluar la extracción del material genético y la preparación de la librería.
- Determinar si es necesario un **preprocesamiento** o **filtrado** de las secuencias, que de manera habitual realizaremos rutinariamente, con la finalidad de eliminar secuencias de baja calidad, regiones de baja calidad y adaptadores.

4.1.1. ¿Por qué es importante evaluar la calidad de las lecturas y filtrarlas?

- ¿Qué secuenciador hemos utilizado?
 - MiniSeq, MiSeq, HiSeq: 4 canales de color
 - NextSeq, NovaSeq: 2 canales de color. Las **regiones polyG** debe tenerse en cuenta para el procesamiento bioinformático de las muestras.
- Mantener un compromiso entre la **cantidad** de datos secuenciados y la **calidad** de los mismos. **Nuestro nivel de exigencia en el filtrado nos hará perder cantidad de datos, pero aumentar su calidad.**

4.1.2. Herramientas de control de calidad

Existen varias herramientas que realizan el análisis de control de calidad, e incluso la limpieza, en la literatura, como son:

- [Fastx-toolkit](#)
- [NGS QC Toolkit](#) (Patel & Jain, 2012)
- [FastqCleaner](#) (Roser et al., 2019)
- [fastqcr](#) (GitHub - kassambara/fastqcr: fastqcr: Quality Control of Sequencing Data, n.d.)
- [FastQC](#) (Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data, n.d.).

Adicionalmente, la herramienta [MultiQC](#) (Ewels et al., 2016) permite compilar distintos tipos de archivos, como son los archivos de salida de FastQC, pero también archivos de mapeo BAM o de limpieza de lecturas, para proveer informes interactivos a modo de resumen, agrupando todas las muestras en una única visualización.

4.1.3. Análisis de calidad con FastQC

Esta aplicación, bien en su versión de línea de comandos o bien en su versión gráfica, nos ofrece las siguientes métricas para evaluar la calidad de las secuencias crudas FASTQ:

- Estadísticas básicas del conjunto de secuencias
- Calidad de la secuencia por base.
- Valores de calidad por secuencia.
- Contenido de la secuencia por base.
- Contenido en GC por secuencia.
- Contenido de N por base
- Distribución de la longitud de las secuencias
- Secuencias duplicadas
- Secuencias sobrerepresentadas
- Contenido de adaptadores

Summary

<input checked="" type="checkbox"/>	Basic Statistics
<input checked="" type="checkbox"/>	Per base sequence quality
<input checked="" type="checkbox"/>	Per tile sequence quality
<input checked="" type="checkbox"/>	Per sequence quality scores
<input checked="" type="checkbox"/>	Per base sequence content
<input checked="" type="checkbox"/>	Per sequence GC content
<input checked="" type="checkbox"/>	Per base N content
<input type="checkbox"/>	Sequence Length Distribution
<input checked="" type="checkbox"/>	Sequence Duplication Levels
<input type="checkbox"/>	Overrepresented sequences
<input checked="" type="checkbox"/>	Adapter Content

En la web de FastQC podéis encontrar ejemplos explicados sobre la calidad en distintos sets de datos, en el apartado “Example Reports” de este link. : <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>



Tema 4 - Ejemplo 1

En este ejemplo vamos a utilizar el programa FastQC para analizar la calidad de las lecturas secuenciadas.

Éstas son lecturas pareadas y podeis descargarlas de este [link](#). Tened en cuenta que las utilizaremos también en el ejemplo 3.

Proceso del ejemplo:

1) Entrar en WorkSpaces

2) descargar el archivo YML de environment [04MBIF_humano](#)

3) Instalar el entorno de trabajo

```
conda env create -f 04MBIF_humano.yml
```

4) Activar el entorno de trabajo

```
conda activate 04MBIF_humano
```

5) Ejecutar FastQC sobre las lecturas proporcionadas en el link de descarga

```
fastqc *.fastq.gz
```

6) Revisión de resultados, abriendo el archivo HTML

```
firefox *.html
```

7) si quiero compilar los archivos HTML en un único archivo interactivo para la visualización, utilizaremos MultiQC

```
multiqc --interactive .
```

```
firefox multiqc_report.html
```

Tema 4 – Ejemplo 1

Estadísticas básicas del conjunto de lecturas. Basic Statistics.

En este apartado se genera una tabla con información acerca del fichero analizado, que incluye:

- Nombre del fichero analizado
- Tipo de fichero
- Codificación utilizada en los valores de calidad
- Número total de secuencias contenidas
- Secuencias etiquetadas como baja calidad (si se ha indicado que lo haga)
- Longitud de las secuencias
- Porcentaje de contenido GC medio.

Tema 4 – Ejemplo 1



Basic Statistics

Measure	Value
Filename	SRR15170798_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	290645
Sequences flagged as poor quality	0
Sequence length	20-301
%GC	54

Tema 4 – Ejemplo 1

Calidad de las secuencias por base. Per base sequence quality.

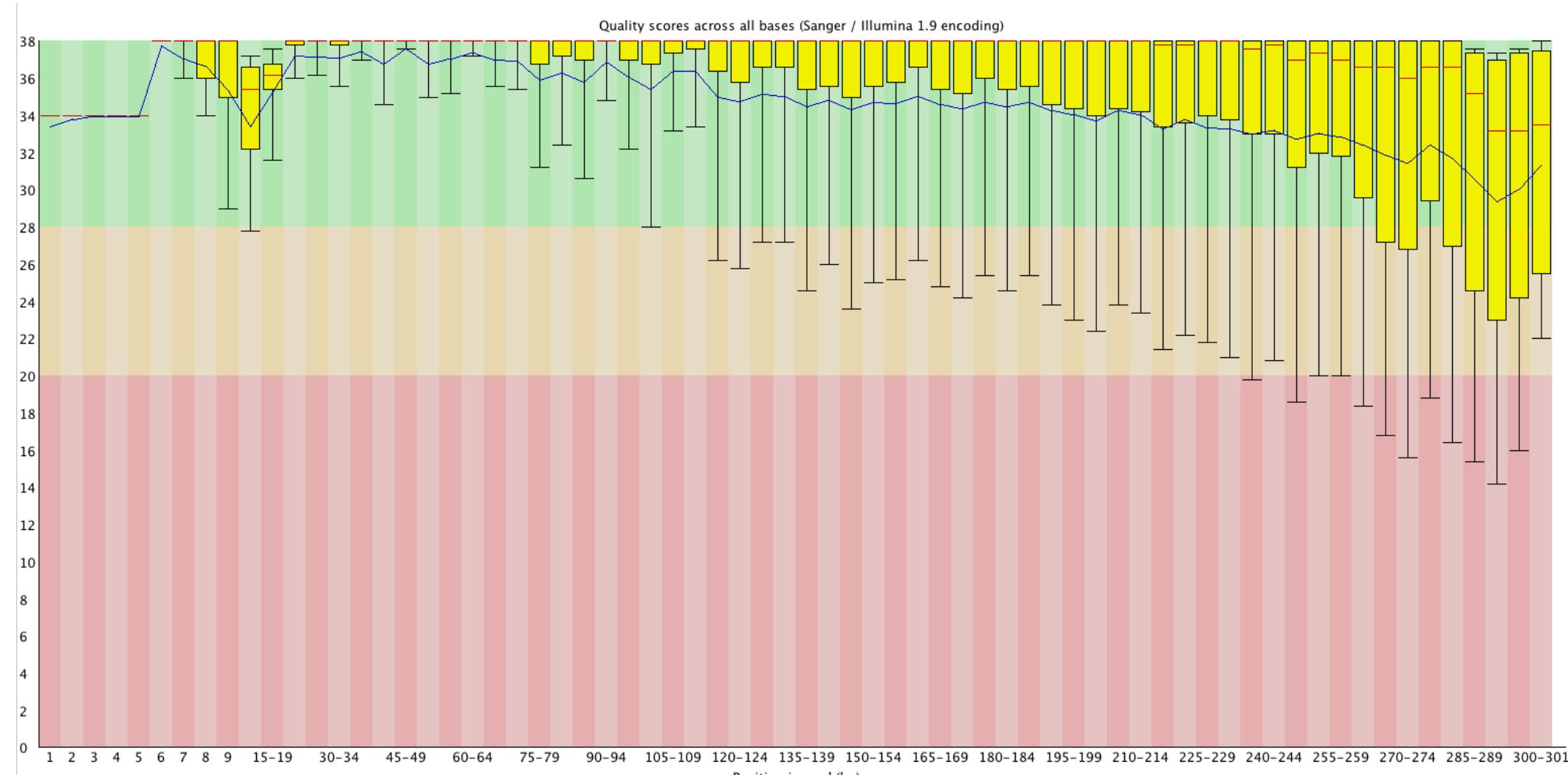
En este apartado se observa el rango de valores de calidad a lo largo de todas las bases secuenciadas teniendo en cuenta todas las lecturas del archivo FASTQ.

En el eje X se representa la posición de la base, mientras que en el eje Y se representan los valores de calidad (**Phred score**). Cuanto **mayor es este valor, mayor precisión en la lectura de esta base, o menor probabilidad de que sea incorrecta**. En este eje se representan tres regiones: **verde** (Phred 28-38, **calidad buena**), **naranja** (Phred 20-28, **calidad razonable**) y **rojo** (Phred 0-20, **mala calidad**).

Lo normal en este tipo de lecturas es que **la calidad se vaya degradando a medida que nos acercamos al final de la secuenciación**. Por tanto, lo normal es que los valores vayan cayendo a regiones naranjas. **Este efecto es más acusado cuanto más largas son las lecturas**. Este hecho se debe a la degradación de la reacción química de fluorescencia a medida que avanzamos en el proceso cíclico de secuenciación.

Las cajas (boxplot) representadas en amarillo tienen en cuenta todas las secuencias en esa posición de la lectura determinada y tienen una línea central roja correspondiente a la mediana de calidad de todos los valores de calidad, la caja amarilla que es el rango intercuartil (25-75%), los bigotes superior e inferior (10 y 90% de calidad, respectivamente) y **la línea azul que es el valor medio del índice de calidad**.

Tema 4 – Ejemplo 1



Tema 4 – Ejemplo 1

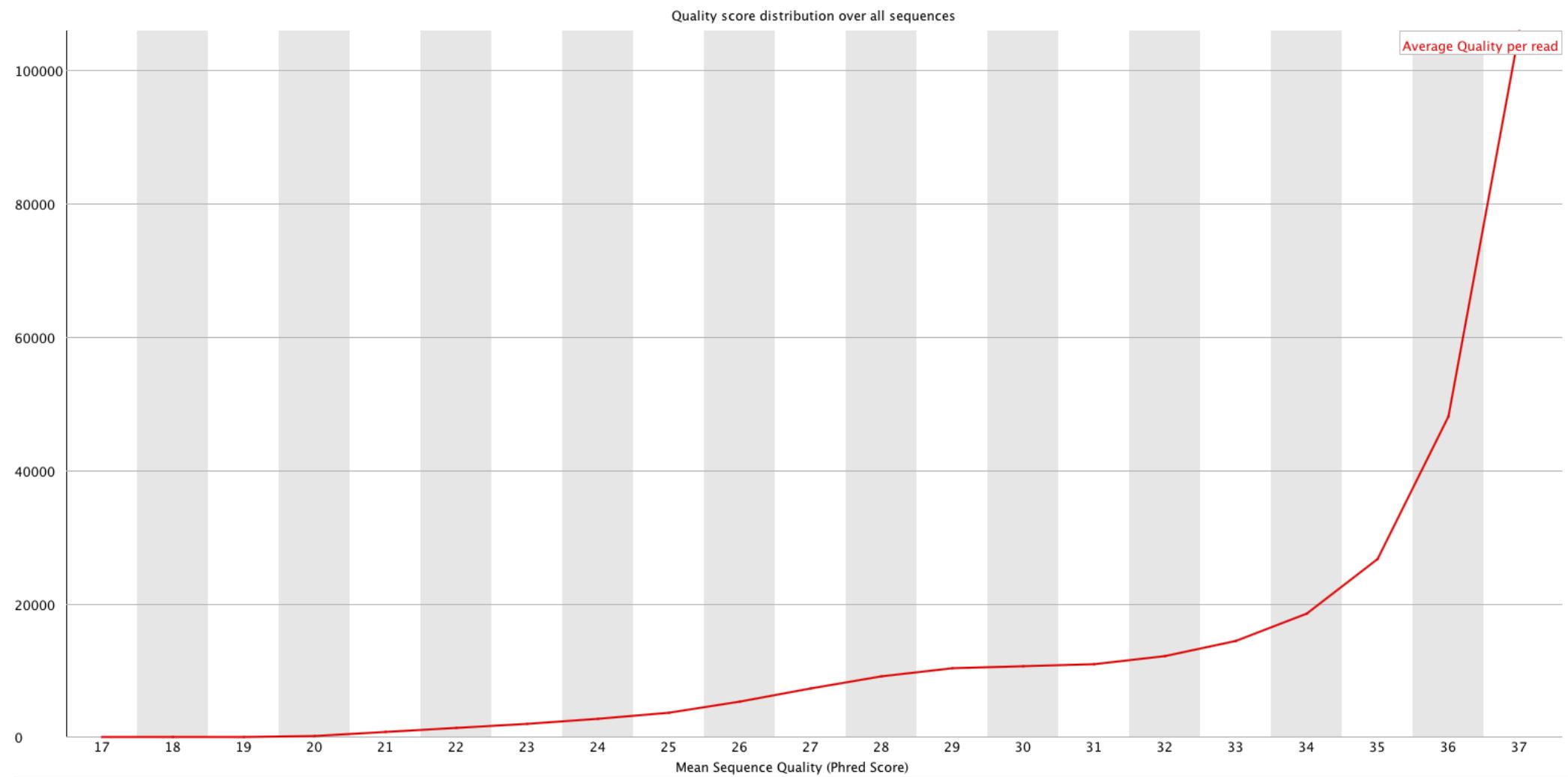
Valores de calidad por secuencia. Per sequence quality score.

En este apartado se calcula el **valor medio de calidad para cada una de las secuencias del archivo FASTQ** y se dibuja los valores en una distribución, de manera que podemos **analizar si un subconjunto de secuencias tiene un valor medio de calidad bajo en comparación con el resto**. En el eje Y se representa la frecuencia de lecturas que tienen un valor de calidad medio dado por el eje X.

En este ejemplo, la mayor parte de nuestras secuencias tienen alto valor de calidad (por encima de 36); sin embargo, hay un subconjunto de secuencias que tienen valores medios de entre 26 a 32. Debemos considerar eliminarlas de nuestro conjunto de datos.

Si una proporción de secuencias tiene un valor de calidad medio-bajo nos puede indicar algún tipo de problema sistemático en una parte de la carrera de secuenciación como, por ejemplo, una celda defectuosa.

Tema 4 – Ejemplo 1



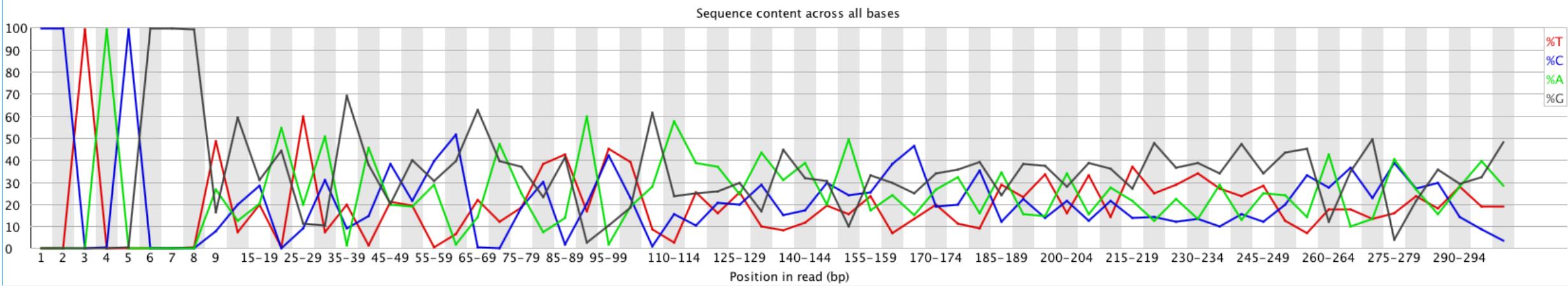
Tema 4 – Ejemplo 1

Contenido de secuencia por base. Per base sequence content.

En esta gráfica vemos la cantidad relativa de cada nucleótido (A, C, G, T) en porcentaje a lo largo de todas las bases de las secuencias del fichero FASTQ.

En una genoteca aleatoria esperaríamos que todas estas bases estuviesen compensadas, de forma que las líneas del gráfico deberían ser paralelas a lo largo de la longitud de la lectura; sin embargo, esto puede verse alterado al inicio de la secuencia si hablamos de experimentos de secuenciación ARN (por el uso de hexámeros en la genoteca); o bien porque sean genotecas de amplicones por ejemplo, de metataxonomía, donde el fragmento es prácticamente idéntico en todas las lecturas.

Tema 4 – Ejemplo 1



Tema 4 – Ejemplo 1

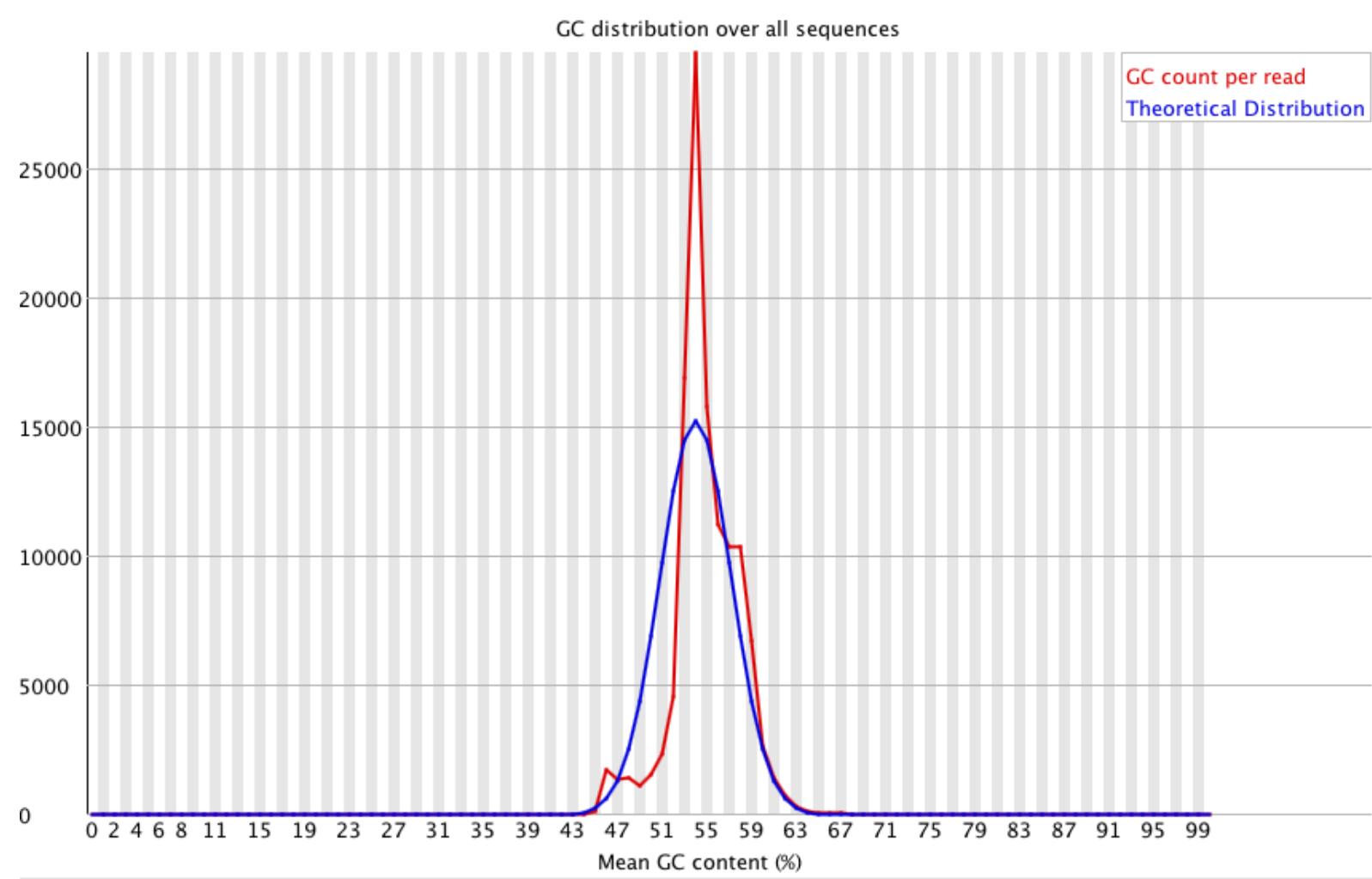
Contenido GC por secuencia. Per sequence GC content.

Este apartado mide el contenido GC medio de cada secuencia contenido en el fichero FASTQ y lo muestra en forma de una distribución, que además se compara con una distribución normal-gaussiana modelada.

En una genoteca aleatoria es esperable una distribución normal de este contenido, donde el pico central es el contenido GC medio del genoma del que se han obtenido las lecturas. Una distribución inusual, como la que vemos en nuestro ejemplo, puede indicarnos algún tipo de sesgo o contaminación.

Si vemos dos picos, podría tratarse de una contaminación de dos genomas diferentes. Los contaminantes específicos, como pueden ser adaptadores de dímeros, que producen picos puntiagudos en la distribución, pueden evaluarse en el apartado de secuencias sobrerrepresentadas.

Tema 4 – Ejemplo 1



Tema 4 – Ejemplo 1

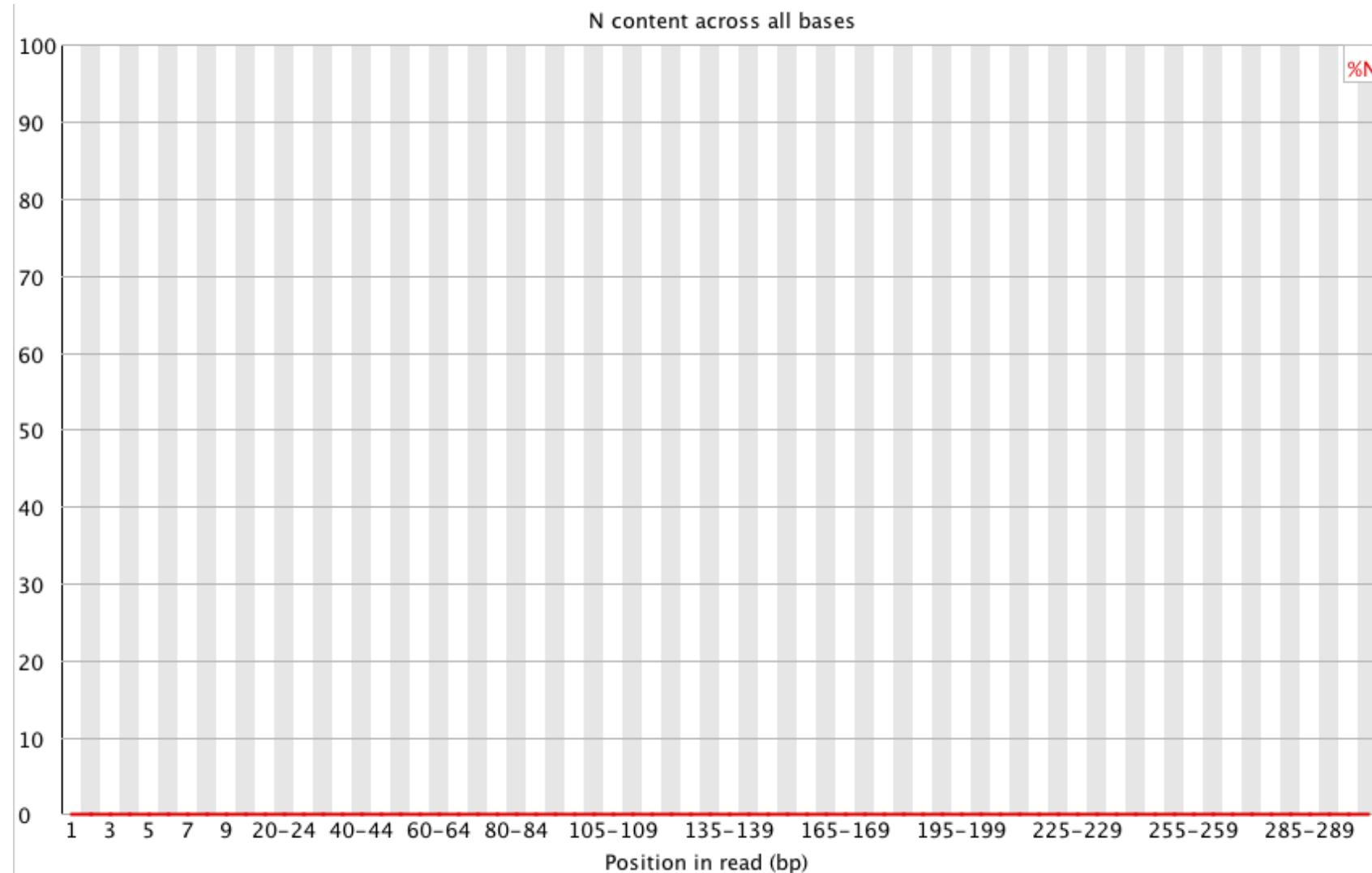
Contenido de 'N' por base. Per base N content.

A lo largo del proceso de secuenciación puede ocurrir que alguna de las bases incorporadas no tenga la resolución suficiente para determinar cuál es exactamente. En este caso, el secuenciador asigna un N en esa posición.

En este gráfico se muestra el porcentaje de N en cada posición de la lectura.

Es habitual detectarlas en baja proporción al final de las lecturas, y pueden ser eliminadas en el proceso de limpieza posterior. En nuestro ejemplo no se observan contenidos en N.

Tema 4 – Ejemplo 1



Tema 4 – Ejemplo 1

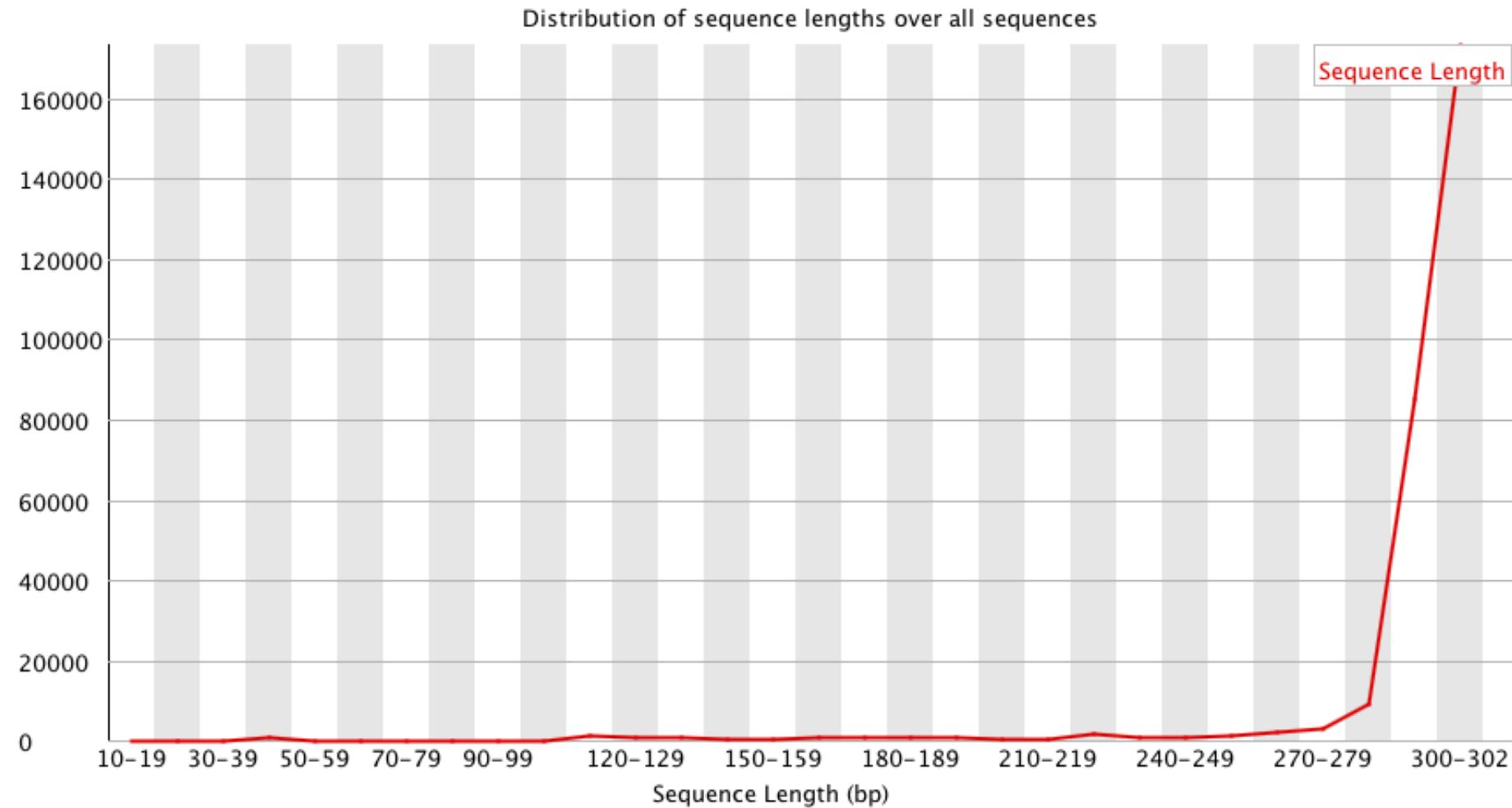
Distribución de la longitud de las secuencias. Sequence length distribution

La gráfica muestra la distribución de la longitud de las secuencias que componen el archivo FASTQ.

Habitualmente, **si no se ha realizado preprocessamiento de las mismas, todas las lecturas de un secuenciador Illumina tienen la misma longitud.**

En nuestro ejemplo, que ha sido obtenido de una base de datos pública, las secuencias tienen una distribución de tamaño de entre 10 pb a 300 pb, siendo este tamaño el mayoritario. Esto es lógico puesto que se trata de secuencias de metataxonomía, secuenciadas en formato pareado de 300 pb. Podemos limpiar las secuencias de bajo tamaño en pasos posteriores.

Tema 4 – Ejemplo 1



Tema 4 – Ejemplo 1

Secuencias duplicadas. Sequence Duplication Levels.

En una genoteca diversa, obtenida a partir de un genoma completo, lo que deberíamos esperar es que la mayor parte de las secuencias aparezcan una sola vez, lo que se observa en un nivel bajo de duplicación.

Por otro lado, **un nivel alto de duplicación puede deberse a un sesgo de enriquecimiento de la genoteca**; sin embargo, puede tener su sentido si hablamos de datos de expresión génica o, como en este caso de un amplicón para análisis metataxonómico.

En la gráfica se observan dos líneas: azul y roja. La **Línea azul** muestra los **niveles de duplicidad totales, del conjunto total de datos**; mientras que la **Línea roja** se calcula **tras eliminar las secuencias duplicadas**. Las proporciones mostradas por la línea roja son del conjunto de datos sin duplicidades.

El porcentaje superior nos indica la proporción del conjunto de datos original que obtendríamos si conserváramos una sola copia de cada secuencia del conjunto de datos. En nuestro caso, tenemos una alta proporción de secuencias duplicadas, y nuestro conjunto de datos quedaría reducido al 4,91% si eliminásemos duplicados.

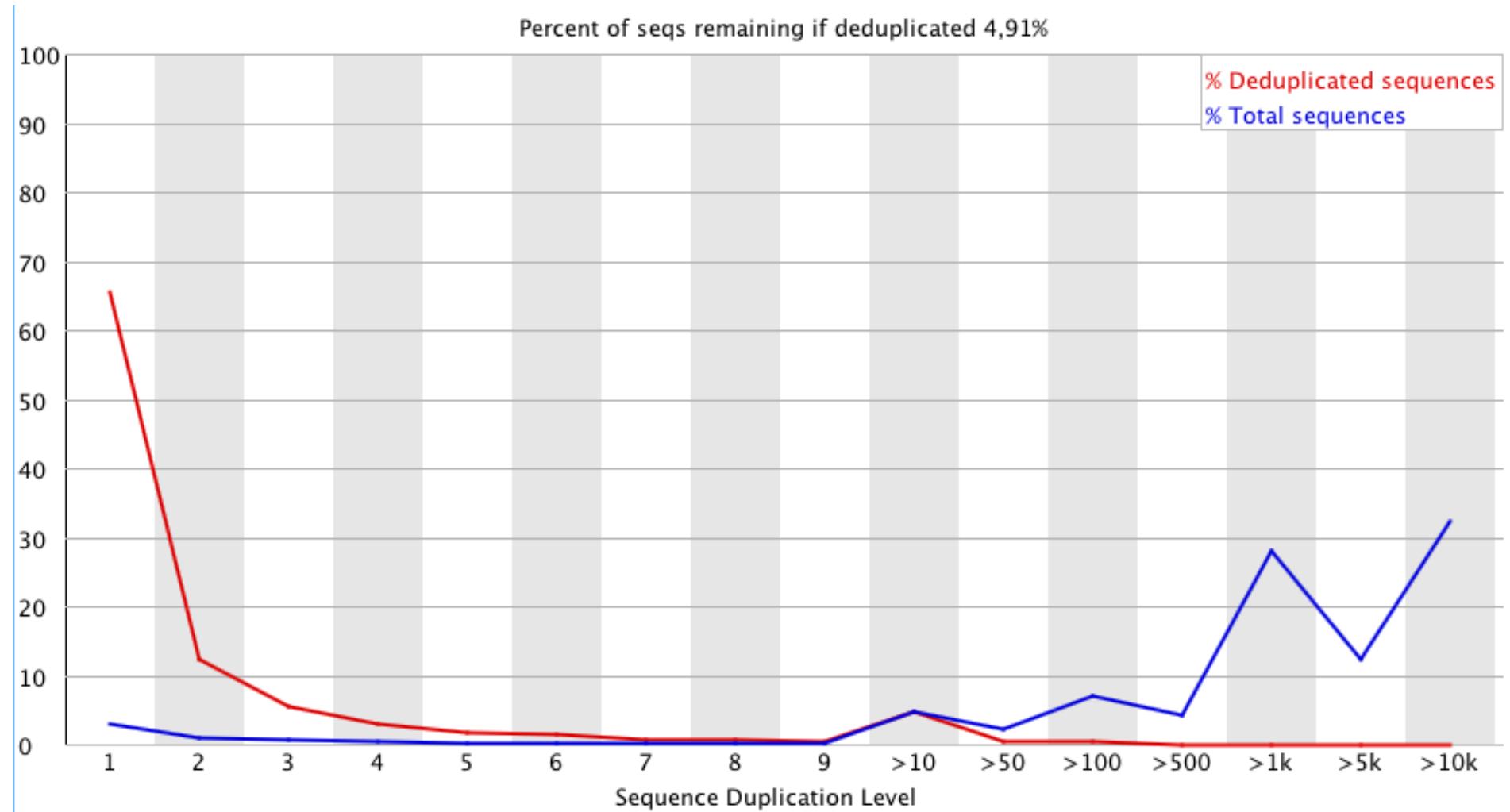
Tema 4 – Ejemplo 1

Secuencias duplicadas. Sequence Duplication Levels.

Las duplicaciones pueden surgir de dos fuentes y es importante diferenciarlas para estimar si es un problema de la genoteca o bien esta duplicidad tiene sentido biológico:

- Duplicaciones técnicas surgidas por problemas de PCR durante el enriquecimiento de la genoteca.
- Duplicación biológica, propia de la naturaleza de nuestra genoteca, por el tipo de experimento que estemos analizando, como puede ser un experimento de transcriptómica o el análisis de un amplicón determinado.

Tema 4 – Ejemplo 1



Tema 4 – Ejemplo 1

Secuencias sobrerepresentadas. Overrepresented sequences.

Una genoteca normal debería ser aleatoria y contener un conjunto diverso de secuencias. Este apartado muestra todas las secuencias que suponen más del 0.1% de las lecturas totales. Asimismo, el programa buscará en su base de datos de contaminantes comunes, entre los que se encuentran los adaptadores de secuenciación, si existen coincidencias.

Si observamos alguna secuencia sobrerepresentada, esto puede significar que esta secuencia es biológicamente significativa (por ejemplo, es un transcripto altamente expresado, si es una genoteca de ARN), que la librería está contaminada (por ejemplo, con adaptadores o cebadores de amplificación); o que la genoteca no sea tan diversa como se esperaba.

En el caso de nuestro ejemplo, se trata de un análisis de un amplicón, ya vemos que hay un tipo del mismo que está presente en el 14.7% de las secuencias generadas.

Tema 4 – Ejemplo 1

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CCTACGGGTGGCAGCGAGAATCATTACAATGGGGAAACCTGAT	42712	14.6955908410604	No Hit
CCTACGGGTGGCTGCAGTCGAGAATCATTACAATGGGGAAACCTGAT	18007	6.195530630150183	No Hit
CCTACGGGAGGCTGCAGTCGAGAATCATTACAATGGGGAAACCTGAT	13635	4.691290061759191	No Hit
CCTACGGGAGGCAGCAGTCGAGAATCATTACAATGGGGAAACCTGAT	10313	3.548314954669786	No Hit
CCTACGGGGGGCAGCAGTCGAGAATCATTACAATGGGGAAACCTGAT	10209	3.512532470883724	No Hit
CCTACGGGTGGCAGCAGTGAGGAATATTGGTCAATGGCGGGAGCCTGAA	9297	3.1987476130674883	No Hit
CCTACGGGGGGCTGCAGTCGAGAATCATTACAATGGGGAAACCTGAT	7449	2.5629204011766933	No Hit
CCTACGGGTGGCAGCAGTGGGAATATTGACAATGGGGAAACCTGAT	6978	2.400867037107124	No Hit
CCTACGGGTGGCAGCAGTGAGGAATATTGGTCAATGGCGCGAGCCTGAA	6452	2.2198902441122335	No Hit
CCTACGGCGGCAGCAGTCGAGAATCATTACAATGGGGAAACCTGAT	6111	2.1025649847752415	No Hit
CCTACGGGAGGCAGCAGTGAGGAATATTGGTCAATGGCGGGAGCCTGAA	4751	1.634640196803661	No Hit
CCTACGGGTGGCAGCAGTGAGGAATATTGGTCAATGGCGGTAGCCTGAA	4457	1.5334858676392162	No Hit
CCTACGGGAGGCTGCAGTGAGGAATATTGGTCAATGGCGGGAGCCTGAA	4249	1.461920900067092	No Hit
CCTACGGGTGGCAGCAGTGAGGAATATTGGTCAATGGCGAGAGCCTGAA	4160	1.4312993514424812	No Hit
CCTACGGGTGGCTGCAGTGAGGAATATTGGTCAATGGCGGGAGCCTGAA	3899	1.3414990796332296	No Hit
CCTACGGGAGGCAGCAGTGGGAATATTGACAATGGGGAAACCTGAT	3677	1.265117239243751	No Hit
CCTACGGGAGGCTGCAGTGGGAATATTGACAATGGGGAAACCTGAT	3476	1.195960708080304	No Hit
CCTACGGGAGGCAGCAGTGAGGAATATTGGTCAATGGCGCGAGCCTGAA	3262	1.1223313664435997	No Hit

Tema 4 – Ejemplo 1

Contenido en adaptadores. Adapter Content.

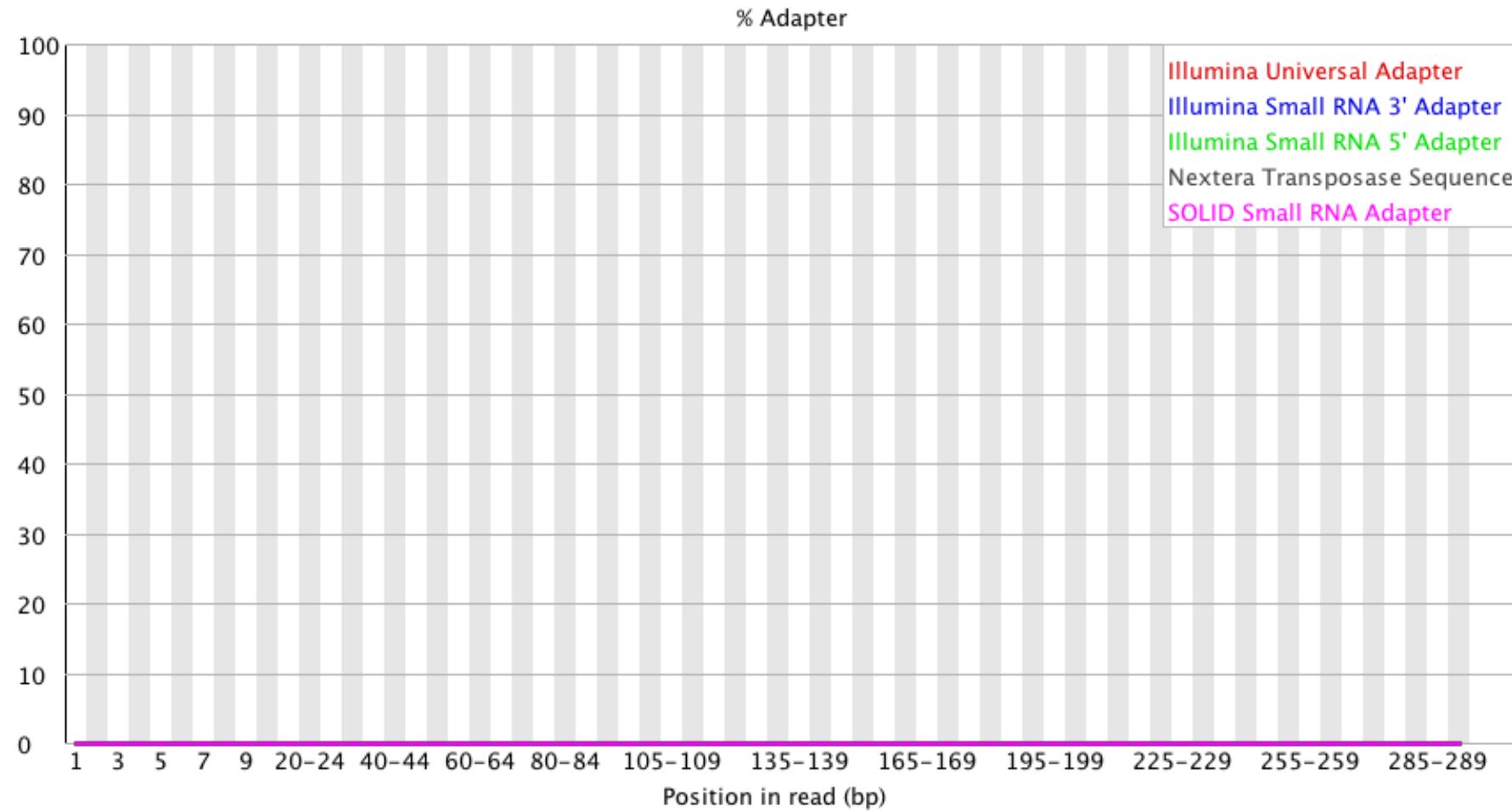
En este apartado únicamente se buscan adaptadores de secuenciación que se encuentren en las lecturas analizadas.

En el gráfico se muestra la **proporción acumulativa de lecturas que tienen adaptadores a lo largo de sus posiciones**. Cuando se detecta un adaptador, se cuenta su presencia en todas las posiciones hasta el final de la lectura, de forma que **los porcentajes siempre se incrementan a medida que avanzamos en la secuencia**.

La presencia de estos adaptadores, habitualmente en una abundancia del 5-10% puede ser normal, y podrán ser eliminados en los pasos posteriores del análisis. En el caso de nuestro ejemplo, no se han detectado adaptadores.

Versiones anteriores de este programa permitían realizar un análisis de contenido de K-meros y un análisis de calidad por celda de secuenciación, pero actualmente se encuentran descatalogados de las versiones más actuales.

Tema 4 – Ejemplo 1



4.1.4. Herramientas de filtrado por calidad de las secuencias y eliminación de adaptadores

Una vez que hemos revisado la calidad de nuestras secuencias debemos tomar la decisión de **limpiarlas y mejorar su calidad, a costa de perder longitud de estas lecturas e incluso lecturas completas.**

Para las siguientes decisiones debemos observar los datos y elegir los parámetros adecuados.

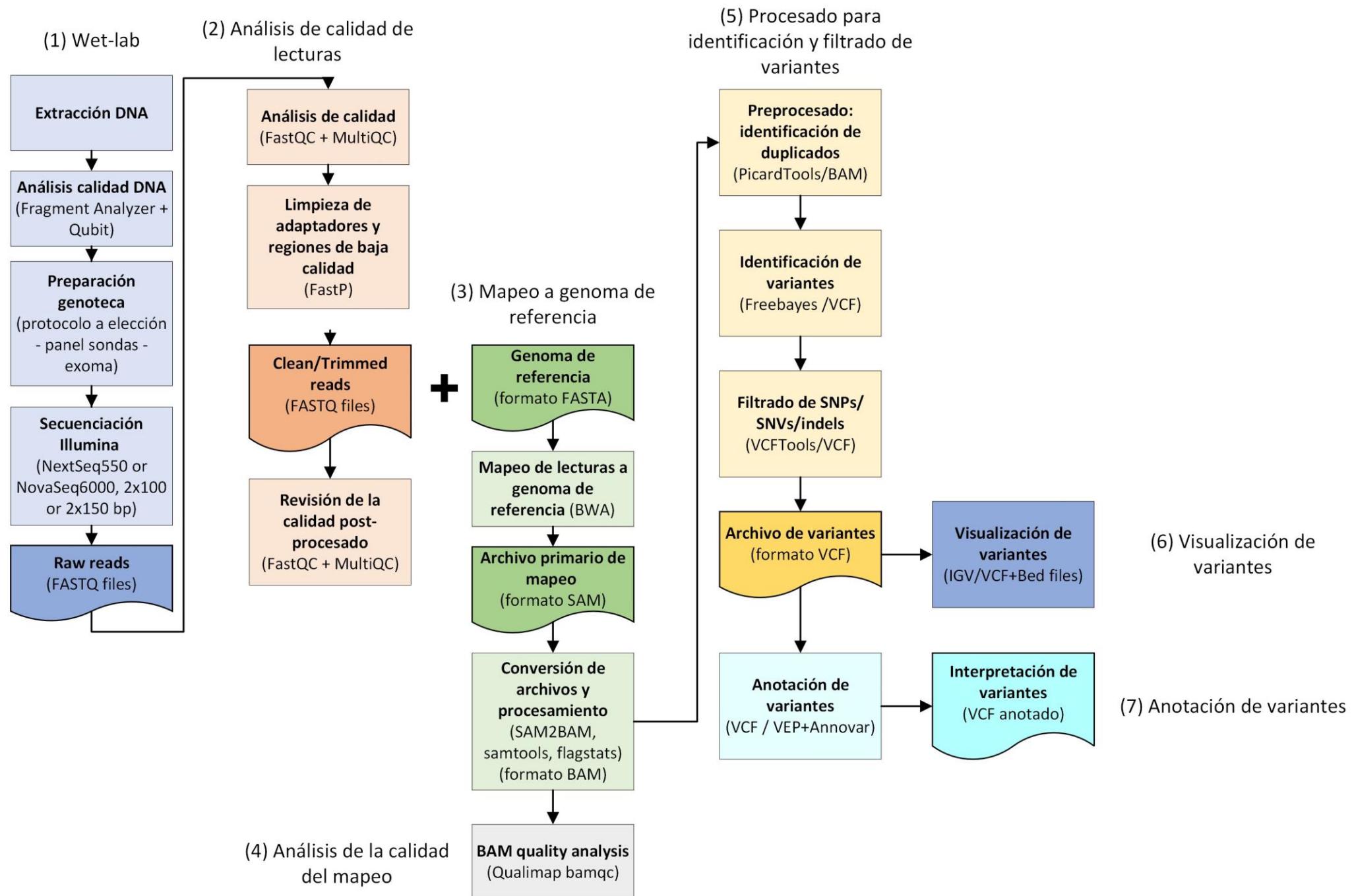
En general, nuestras secuencias serán buenas si todo el proceso previo de preparación de genoteca y secuenciación ha sido adecuado.

Habitualmente lo que haremos será:

- **buscar adaptadores y eliminarlos,**
- **recortar la parte final de las lecturas que decaen en calidad (*trimming de la región final*) y**
- **eliminar aquellas cuya calidad media no supere un umbral.**

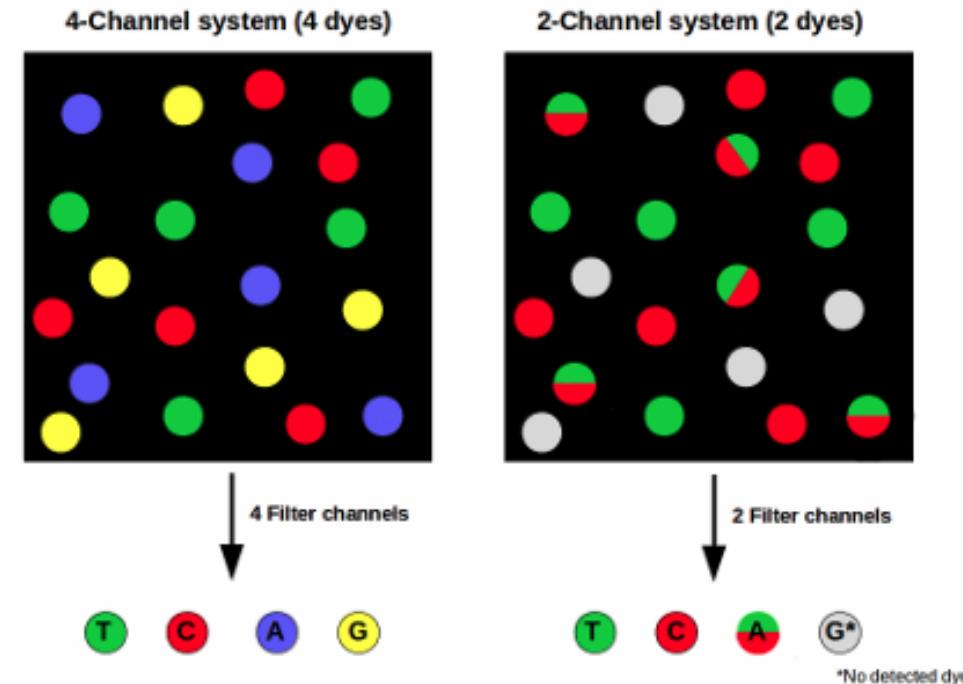
Importante: los archivos *forward* y *reverse* de lecturas pareadas (*paired-end*) se deben limpiar a la vez.

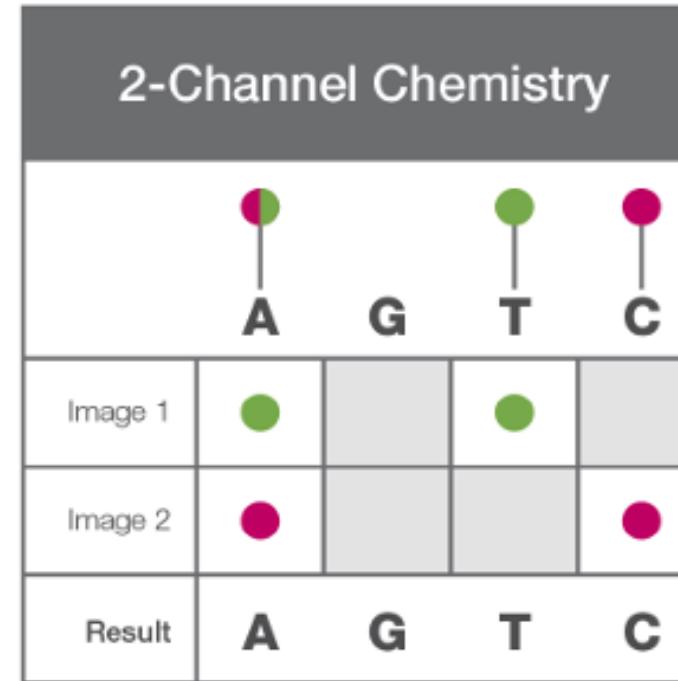
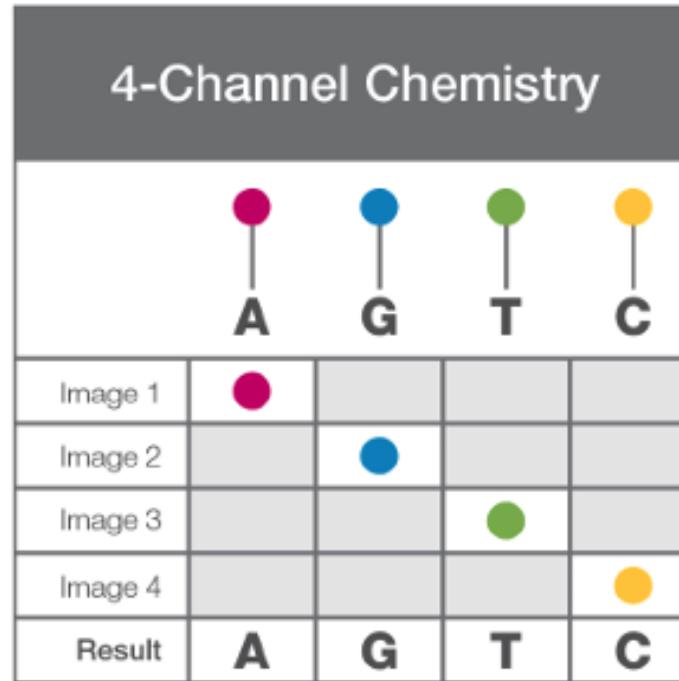
Asimismo, durante el proceso de trimming por calidad puede que algunas secuencias queden muy cortas, y ya no nos interesen, en ese caso, lo idóneo es eliminarlas. Si estamos trabajando con secuencias pareadas hay que recordar que debemos eliminar también su pareja.



4.1.4. Herramientas de filtrado por calidad de las secuencias y eliminación de adaptadores

Adicionalmente, en aquellas secuencias procedentes de secuenciadores de dos canales de color, como son NextSeq y NovaSeq, debemos tener en cuenta la presencia de gran contenido en Guanina (G, **polyG**), debido a que la ausencia de color, debida inclusivamente a un problema del secuenciador, se asocia a la base Guanina. En este caso, debemos seleccionar un programa de limpieza que realice esta tarea adecuadamente.





Four-channel SBS

- Bases are identified using four different fluorescent dyes, one for each base and four images per sequencing cycle

Two-channel SBS

- Simplified nucleotide detection by using two fluorescent dyes and two images to determine all four base calls

Four Channels SBS:

- MiSeq

Two Channels SBS:

- MiniSeq, NextSeq, NovaSeq
- Accelerates sequencing and data processing **times**

4.1.4. Herramientas de filtrado por calidad de las secuencias y eliminación de adaptadores

En este proceso se recomienda realizar el ***trimming***, y posteriormente realizar otro análisis de calidad que nos asegure que la limpieza ha sido adecuada.

Existen varios programas para realizar esta tarea (C. Chen et al., 2014), casi todos ellos basados en Cutadapt (Martin, 2011) para eliminar adaptadores, pero con ampliaciones a trimming y análisis de calidad como Trimmomatic (Bolger et al., 2014), TrimGalore (*Babraham Bioinformatics - Trim Galore!*, n.d.) o FastP (S. Chen et al., 2018).

Por su alta versatilidad, vamos a realizar el ejemplo de eliminación de adaptadores y trimeado con **FastP** (<https://github.com/OpenGene/fastp>).



Tema 4 - Ejemplo 2

El segundo ejemplo tratará de realizar la limpieza de calidad y adaptadores de las lecturas originales mediante el programa FastP.

Para ello utilizaremos las lecturas descargadas en el Ejemplo 1 ([link](#)).

Recordad tener instalado y activo el entorno de trabajo 04MBIF_humano, tal y como vimos en el Ejemplo 1.

1) Ejecutamos FastP para realizar la limpieza de las secuencias

```
fastp -i S8_L001_R1_001.fastq.gz -I S8_L001_R2_001.fastq.gz -o out1.clean fq.gz -O out2.clean fq.gz --trim_poly_g --detect_adapter_for_pe --cut_front 25 --cut_tail 25 --cut_mean_quality 25 -l 100 -h report_fastp.html
```

Nota: revisar las opciones. Si detectáis polyA, polyT... podéis hacer uso de la opción para eliminarlo!

2) Ejecutamos FastQC para analizar si han quedado correctamente, y convergemos el resultado comparativo en MultiQC

```
fastqc *clean.fastq.gz  
multiqc --interactive .
```

3) Revisamos los archivos HTML

Revisemos las siguientes cuestiones:

- ¿Cuál es la calidad media que le hemos pedido que retenga?
- ¿Cuál es el criterio para recortar las lecturas?
- ¿Cuántas lecturas teníamos inicialmente?
- ¿Cuántas han quedado finalmente?
- ¿Cuántas han sido filtradas por baja calidad? ¿Cuántas por contener Ns? ¿Cuántas por ser demasiado cortas?
- ¿Cuál es el tamaño mínimo de lectura?
- ¿Se han detectado adaptadores?

Sobre las lecturas utilizadas en el ejemplo anterior, vamos a realizar la limpieza con el programa FastP. Antes de continuar con el ejemplo, te animo a revisar la documentación del programa, disponible en <https://github.com/OpenGene/fastp>, donde encontrarás detalle de todas las opciones que ofrece.

fastp

```
fastp -i S8_L001_R1_001.fastq.gz -I S8_L001_R2_001.fastq.gz -o out1.clean.fq.gz -O
out2.clean.fq.gz --cut_front 25 --cut_tail 25 --cut_mean_quality 25 --
detect_adapter_for_pe --trim_poly_g --trim_poly_x -l 100 -h report_fastp.html
```

Parámetro	Función
-i -I	Lecturas pareadas crudas: R1 (-i) y R2 (-I).
-o -O	Lecturas pareadas salida: R1 (-o) y R2 (-O)
--cut_front --cut_tail	Calidad recorte extremo 5' (front) y 3' (tail)
--cut_mean_quality	Calidad media para descartar
--detect_adapter_for_pe	Detectar y recortar adaptadores para lecturas pareadas
--trim_poly_g -- trim_poly_x	Recortar colas polyG y polyA
-l	Longitud por debajo de la cual se descartan las lecturas

4.2



Mapeo de secuencias. Herramientas de mapeo, visualización y análisis de calidad

4.2. Mapeo de secuencias. Herramientas de mapeo, visualización y análisis de calidad.

4.2.1. El proceso de mapeo.

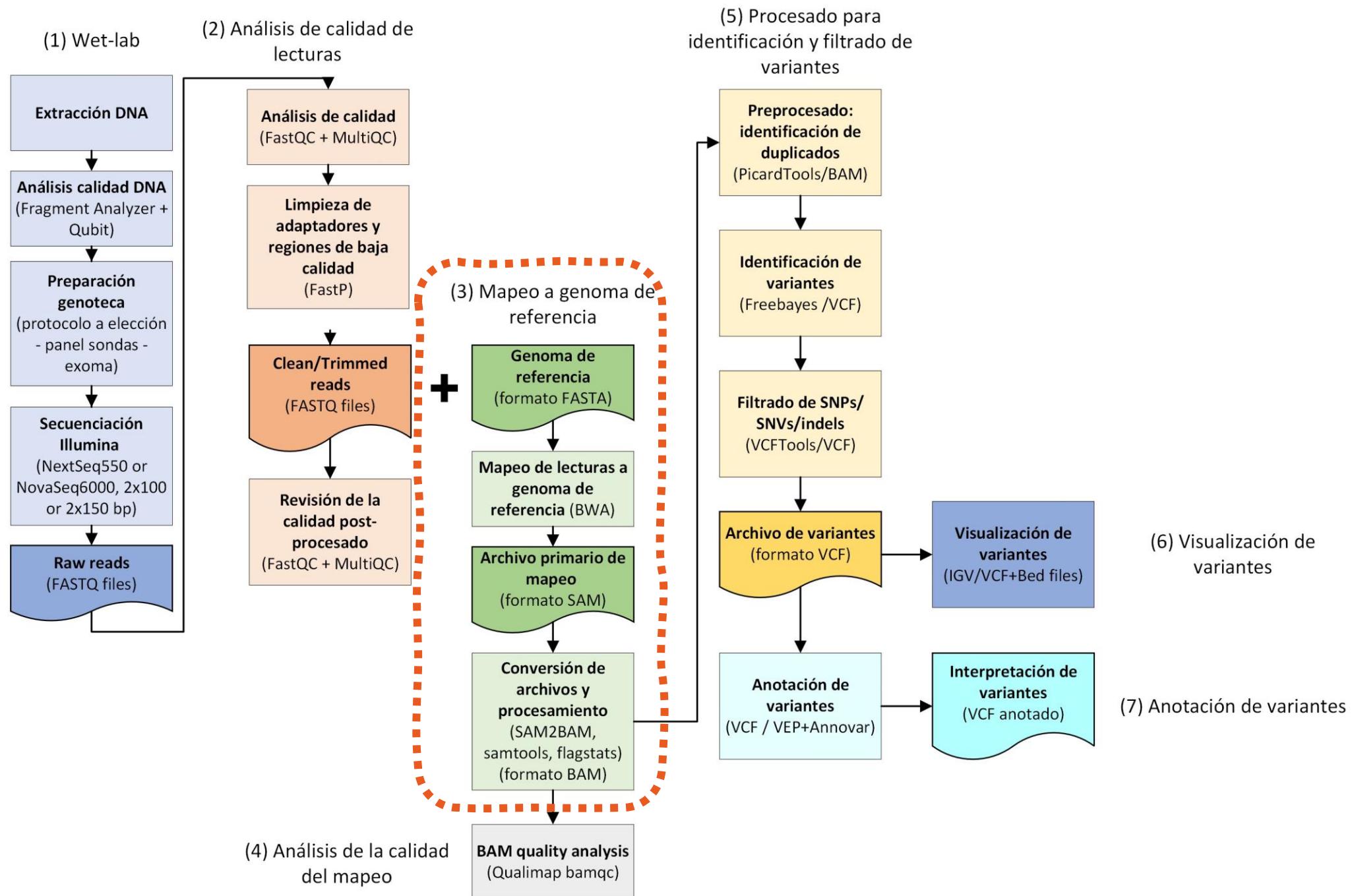
4.2.2. Herramientas para el mapeo.

4.2.3. Archivos de mapeo: el formato SAM

4.2.4. Herramientas para el manejo de los archivos SAM. Generación de archivos BAM.

4.2.5. Visualización del mapeo.

4.2.6. Análisis de la calidad de mapeo.



4.2.1. El proceso de mapeo.

El proceso de alineamiento o mapeo de las lecturas (también denominadas “reads”) es el proceso en el cual un programa llamado mapeador o alineador coloca estas lecturas sobre el genoma de referencia, proporcionando una o más localizaciones probables para cada lectura.

Éste es un proceso fundamental y central en multitud de análisis de datos de secuenciación masiva y, por tanto, debemos cuidar su éxito para lograr el éxito del experimento. Sin embargo, un mapeo mal realizado, con parámetros incorrectos, nos dificultará extraer conclusiones.

Ya habéis visto en otras asignaturas, y hemos recalcado en capítulos anteriores que, durante el proceso de secuenciación masiva, especialmente cuando tratamos de tecnología de lecturas cortas, se realiza una fragmentación del genoma y se secuencian en paralelo millones de veces los mismos fragmentos, obteniendo así una redundancia. Esta es la manera de tener cada base del genoma secuenciada varias veces y calcular la fiabilidad de esa posición. El número de veces que cada base nucleotídica está secuenciada, es decir, el número de lecturas que apoyan dicha base es la **profundidad (depth of coverage)** de secuenciación.

4.2.1. El proceso de mapeo.

Aún así, teniendo cada base secuenciada cientos o miles de veces, **¿por qué el proceso de mapeo es tan complejo?** Más allá de la propia complejidad computacional de manejar millones de datos, habitualmente imaginamos que cada lectura tiene una posición única e inequívoca sobre el genoma de referencia. Pero esto no es cierto, como vimos ya en el capítulo 1, hay algunos factores:

Redundancia genómica. Los genomas, en especial los genomas eucariotas, son muy redundantes. Tenemos regiones iguales a otras, especialmente si pensamos en **pseudogenes y regiones repetitivas**. En las zonas de alta redundancia, como son **telómeros y centrómeros**, ningún mapeador es capaz de encontrar una única posición para cada lectura.

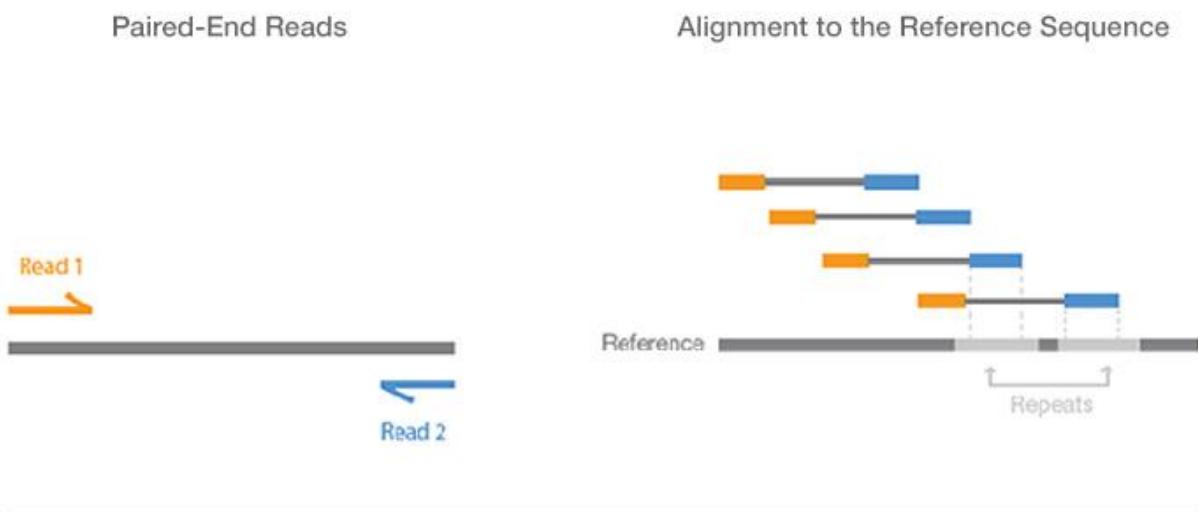
Diversidad genética. En el proceso de mapeo partimos de la base de que todos los genomas de la misma especie son esencialmente iguales. Sin embargo, esto no es estrictamente cierto, ya que sabemos que existen mutaciones que lo hacen peculiar. Esto hará que las **lecturas puedan tener divergencias respecto al genoma de referencia**. Es el objetivo de nuestro estudio y los mapeadores contemplan estas diferencias.

Errores debidos a la técnica. Tanto durante la amplificación de la genoteca/librería, como la amplificación del material y la secuenciación se pueden producir errores en los fragmentos (cambios de un nucleótido por otro, delecciones o inserciones), que se reflejarán en las lecturas del secuenciador. Sin embargo, debemos recordar que la tasa de error asociada a secuenciación en equipos Illumina de lecturas cortas es <0.01%, haciendo que podamos solventar en gran medida este problema con una cobertura del genoma adecuada.

4.2.1. El proceso de mapeo.

Además de estas complejidades, los mapeadores de lecturas cortas deben lidiar con secuencias emparejadas (**paired-end**). Podemos secuenciar cada fragmento preparado en la genoteca de manera sencilla (**single-end**) o emparejada, por los dos extremos (**paired-end**).

Para cada una de las parejas de lecturas secuenciadas de esta forma, se conoce aproximadamente la distancia entre ellas, de manera que el mapeador debe colocarlas sobre el genoma de referencia respetando esta distancia. Aunque es un reto para la computación y supone mayor complejidad para el alineamiento, las lecturas emparejadas nos permiten sortear secuencias repetitivas, siempre y cuando, el tamaño de éstas sea menor que el tamaño que separa las lecturas emparejadas. Ésta es una de las ventajas de este tipo de lecturas, que nos permiten resolver algunas reestructuraciones cromosómicas (traslocaciones e inversiones) y son de ayuda en el ensamblaje de novo de genomas, como veremos en capítulos posteriores.



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

4.2.2. Herramientas para el mapeo.

Desde la aparición de la secuenciación masiva, y por tanto de millones de lecturas obtenidas de distintos genomas, aparecieron diversas herramientas para alinearlas a los genomas de referencia.

Cada una de estas herramientas tiene sus peculiaridades, basadas en el algoritmo base se las dominan y, por tanto, la aplicación a la que van dirigidas.

En este apartado vamos a repasar algunas características generales, aunque en nuestras prácticas seleccionemos solo unos pocos mapeadores en función del objetivo a alcanzar.



4.2.2. Herramientas para el mapeo.

En el siguiente link podéis ver un listado de herramientas de mapeo generales actualizadas y clasificadas en función de su objetivo: búsqueda en base de datos, alineamiento pareado, alineamiento múltiple, análisis genómico, búsqueda de motivos conservados; así como un listado de editores de secuencias y visualizadores.

[Lista actualizada de alineadores](#)

4.2.2. Herramientas para el mapeo.

La clasificación de los programas de mapeo atiende a tres criterios:

- Según tipo y tamaño de las lecturas a alinear.
- Según el tipo de algoritmo base.
- Según la procedencia de la genoteca/librería.

4.2.2. Herramientas para el mapeo.

Según tipo y tamaño de las lecturas a alinear.

Esta es la clasificación más general, ya que atiende a principalmente los dos tipos de lecturas que podemos encontrar:

Alineadores de lecturas largas. Estos algoritmos son más **lentos, pero más específicos**. Son más antiguos, porque son los que tradicionalmente trabajaban con secuencias Sanger o con aquellas que provenían de secuenciadores 454 de Roche. Van a manejar de mejor manera los errores y nos van a permitir mapear de manera más precisa lecturas que vengan de regiones muy variables. Entre ellos están mapeadores como **BLAT, LAST, LASTZ, BLASTZ, SMALT o MUMmer**. Los utilizaremos cuando tengamos **lecturas largas (>300 pb)** y **pocas secuencias (<1M)**. Actualmente, dada la extensión del uso de secuenciadores tales como PacBio y Oxford Nanopore, con lecturas de **>10 Kpb**, existen mapeadores específicos para este tipo de lecturas, como son Minimap2 (H. Li, 2018), Ira (Ren & Chaisson, 2021) o LRA (*GitHub - ChaissonLab/LRA: Long read aligner*, n.d.).

Alineadores de secuencias cortas. Aquí fundamentalmente hacemos referencia a **lecturas procedentes de tecnología Illumina o análoga (MGI)**. Se caracterizan por ser mapeadores **muy rápidos, pero muy sensibles a los errores**. Comprenden el grupo de mapeadores bien conocidos como son **Bowtie/Bowtie2, BWA, SOAP, TopHat/TopHat2, HISAT/HISAT2 o STAR**.

Hay **mapeadores híbridos**, que actualmente nos permiten alinear tanto lecturas cortas como largas. Es el caso de **BWA** en sus últimas versiones, con su algoritmo **swy mem** (opciones -x pacbio o -x ont2d).

4.2.2. Herramientas para el mapeo.

Según tipo de algoritmo base.

Mapeadores basados en hashing. El proceso de hashing (concepto computacional) es, de manera sencilla, crear un **índice a partir de la referencia para encontrar rápidamente la posición de cualquier lectura**. Este es un proceso muy general en el proceso de mapeo: **generar la secuencia índice (index) del genoma de referencia**. Este paso **permite un mapeo posterior con extrema rapidez y muy sensible a errores**. Se utiliza el inicio de cada lectura (semilla, seed) para consultar el índice. Ejemplo de estos mapeadores son STAR, SOAP y BLAT.

Mapeadores basados en el algoritmo de Smith-Waterman. Éste es un algoritmo de programación dinámica que garantiza que el **alineamiento local** es óptimo, basándose en un sistema de puntuación que utiliza una matriz de sustitución. Por tanto, dadas dos secuencias y una tabla de puntuación que actúa como reglas para puntuar el alineamiento, **este algoritmo siempre encontrará el mejor alineamiento**. Este tipo de mapeadores no son tan rápidos como los anteriormente descritos, pero sí son más precisos y menos sensibles a los errores. Ejemplo: BFAST.

Mapeadores basados en la transformación de Burrows-Wheeler (BWT). Estos mapeadores utilizan la transformada de Burrows-Wheeler para optimizar el uso de la memoria. Se suelen utilizar junto con mapeadores basados en hashes. Son **los preferidos para lecturas cortas**, ya que ofrecen un buen balance entre eficacia, sensibilidad y especificidad. Ejemplo: BWA, Bowtie o HISAT2.

4.2.2. Herramientas para el mapeo.

Los mapeadores más actuales, e incluso las nuevas versiones de mapeadores clásicos como **BWA**, combinan más de una de estas estrategias para sacar el mejor partido de ellas. Ejemplo es HISAT2, que combina BWT junto con índices de tipo GFM. Asimismo, el alineador BWA con sus funciones SW y MEM realiza una combinación, por lo que hoy en día sigue siendo uno de los mapeadores más ampliamente utilizados (<http://bio-bwa.sourceforge.net/>).

En este video podéis profundizar en el algoritmo de **Smith-Waterman**.

<https://www.youtube.com/watch?v=lu9ScxSejSE>.

En este video podéis profundizar en la transformación de **Burrows-Wheeler**.

<https://www.youtube.com/watch?v=4n7NPk5lwbl>

4.2.2. Herramientas para el mapeo.

Según la procedencia de la librería.

Por último, los mapeadores pueden clasificarse según su optimización para distintos tipos de librerías genómicas. Detallamos algunos de ellos:

Mapeadores para secuencias cromosómicas (DNAseq). Incluye aquellos capaces de mapear lecturas obtenidas de librerías genómicas de ADN. La mayor parte de los mapeadores pueden realizarlo. Los ejemplos más conocidos: **BWA**, **Bowtie/Bowtie2**.

Mapeadores para secuencias de ADN complementario (RNAseq). Cuando la librería se ha obtenido a partir de ADN complementario de una secuencia de RNA (bien total o mensajero), **lo normal es que no tenga intrones**, al proceder de transcriptos maduros. Bien es cierto que debido a la profundidad de la técnica encontraremos algunos transcriptos sin procesar totalmente. **La eliminación de los intrones hace que el mapeo sea más dificultoso, ya que los alineadores deben saltar estos intrones** que están en el genoma de referencia y pueden incluso medir varios Kb. Si la lectura solapa con dos exones adyacentes del ARNm, el alineador debe fragmentar la lectura para colocarla, sin que esto suponga una penalización en el puntaje del mapeo. En este grupo están los alineadores **TopHat/TopHat2** y **HISAT/HISAT2**.

4.2.2. Herramientas para el mapeo.

Las características que le pedimos a un mapeador son las siguientes:

Capacidad de manejar errores de secuenciación y diferencias genéticas (polimorfismos). Debido a los errores de secuenciación y a que cada individuo posee mutaciones propias respecto al genoma de referencia, nuestro mapeador debe poder alinear estas lecturas pese a no ser exactamente idénticas a la referencia.

Especificidad. Como hemos dicho, existe redundancia genética, y una lectura puede alinearse con más de una región del genoma. El alineador debe encontrar la posición más probable.

Eficacia. El mapeo masivo de millones de lecturas debe realizarse en un tiempo razonable y consumiendo unos recursos computacionales razonables.

Tema 4 – Ejemplo 3



Tema 4 - Ejemplo 3



El siguiente paso de nuestro protocolo es mapear las lecturas secuenciadas, analizadas y filtradas (en los pasos anteriores de los ejemplos 1 y 2) sobre el genoma de referencia. Para ello utilizaremos BWA como alineador.

En este ejemplo vamos a necesitar los siguientes archivos:

- lecturas a mapear: lecturas limpias, obtenidas en el ejemplo 2 ([link](#))

- genoma de referencia. Utilizaremos el genoma de referencia humano en su versión Hg19/GRCh37 (ya que las sondas fueron diseñadas sobre él). Para agilizar el mapeo utilizaremos sólo el cromosoma 7 de referencia. El genoma de referencia puede descargarse de la web:

https://ftp.ensembl.org/pub/grch37/current/fasta/homo_sapiens/dna/, de donde tomaremos el cromosoma 7 (https://ftp.ensembl.org/pub/grch37/current/fasta/homo_sapiens/dna/Homo_sapiens.GRCh37.dna.chromosome.7.fa.gz).

Pasos a realizar:

1) Descarga del cromosoma 7 de referencia y descomprimir en la carpeta donde vayamos a trabajar.

```
gzip -d Homo_sapiens.GRCh37.dna.chromosome.7.fa.gz
```

2) indexar el genoma de referencia (178 sg)

```
bwa index Homo_sapiens.GRCh37.dna.chromosome.7.fa
```

3) mapeo al genoma de referencia (5-10 min)

```
bwa mem -a Homo_sapiens.GRCh37.dna.chromosome.7.fa out1.clean.fq.gz out2.clean.fq.gz -o chr7.sam 2> chr7.out
```

Ahora ya tenemos nuestras lecturas mapeadas al chr7 del genoma de referencia!

Resultados: [Tema4_Ejemplo3_Mapeo](#)

4.2.3. Archivos de mapeo. Formato SAM.

Los archivos resultantes del proceso de mapeo son archivos SAM (Sequence Alignment Map). En este formato se representan cada una de las lecturas sobre el genoma de referencia, indicando su secuencia, posición exacta de inicio, de final, hebra sobre la que mapea y parámetros de calidad del mapeo.

En este enlace se encuentra la documentación oficial sobre el archivo estándar de tipo SAM.
<https://samtools.github.io/hts-specs/SAMv1.pdf>

El formato SAM es un archivo de texto plano delimitado por tabuladores, con dos secciones definidas, la sección cabecera y los alineamientos.

```
@SQ SN:7 LN:159138663
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem -a Homo_sapiens.GRCh37.dna.chromosome.7.fa out_1.clean.fq.gz out_2.clean.fq.gz -o chr7.sam
M02899:19:000000000-JFB5J:1:1101:17405:2045 83 7 100912013 3 5S79M49S = 100912013 -79 CGTTTAAAAATTAGCTGGGTGTGGTGGCGTGTAGTGCCAGTACCCAGGAGGCTGAGGTGGAGGATCACCTGAGCCCCAGAGGCCAAGGC
TGCAGTGACCATGACCGCACCCTGACTCTAGCC AAAABFFFFFGGGGGGAEEEDHGCGLF2E2ECGGHB5DCEADBCBCHBCGG3FE1F?EG11FAA1101101FGCHGB3F3????//EFEEC/FFEEF3B334?B?FFFFHCFCDGHFFGFFFFHDGBG NM:i:6 MD:Z:12A9T1G1C9C8T33 MC:Z:5S79M49S AS:i:49 XS:i:44 SA:Z:7,73730619,+56M775,3,5;
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 110472164 0 SH59M69H = 100912013 -9560210 * * NM:i:3 MD:Z:28C7C8T13 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 44591270 0 34H99M = 100912013 56320646 * * NM:i:11 MD:Z:7C6T31A2T0A0G3A0T14G6T15C4 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 116784570 0 42H63M28H = 100912013 -15872620 * * NM:i:4 MD:Z:41T1G0T3T14 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 151266200 0 20H58M55H = 100912013 -50354188 * * NM:i:3 MD:Z:7C14A4C30 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 5142163 0 99M34H = 100912013 95769851 * * NM:i:12 MD:Z:2G15T0A0A21A1T2C0T3T3C6T27G7 MC:Z:5H79M49H AS:i:41
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 283971 0 26H65M42H = 100912013 100628043 * * NM:i:5 MD:Z:21C0T20T10C1C8 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 148973073 0 48H80M5H = 100912013 -48061061 * * NM:i:8 MD:Z:29C4A8A7G1G0T1T16C6 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 132493804 0 49H50M34H = 100912013 -31581792 * * NM:i:2 MD:Z:33A867 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 17046803 0 69H59M5H = 100912013 83865211 * * NM:i:4 MD:Z:11C1A8G7G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 41146108 0 69H59M5H = 100912013 59765906 * * NM:i:4 MD:Z:13A8G1T5G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 81351976 0 5H59M69H = 100912013 19559980 * * NM:i:4 MD:Z:28C5A1C8T13 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 43949171 0 5H59M69H = 100912013 56962785 * * NM:i:4 MD:Z:28C7C8T0A12 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 138980542 0 34H44M55H = 100912013 -38068573 * * NM:i:1 MD:Z:T36 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 97540837 0 42H63M28H = 100912013 3371115 * * NM:i:5 MD:Z:8T33A0G0A3T14 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 128693225 0 5H58M70H = 100912013 -27781270 * * NM:i:4 MD:Z:28C5A1C8T12 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 98072245 0 49H42M42H = 100912013 2839769 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 28419599 0 49H42M42H = 100912013 72492415 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 111786632 0 4H77M52H = 100912013 -10874696 * * NM:i:8 MD:Z:27C0A0C2A4C8T13C0A15 MC:Z:5H79M49H AS:i:37
```

```

@SQ SN:7 LN:159138663
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem -a Homo_sapiens.GRCh37.dna.chromosome.7.fa out_1.clean.fq.gz out_2.clean.fq.gz -o chr7.sam
M02899:19:000000000-JFB5J:1:1101:17405:2045 83 7 100912013 3 5S79M49S = 100912013 -79 CGTTTAAAAATTAGCTGGGTGTTGGCGTGTGGCTTAGTGCCTACCCAGGAGGCTGAGGTGGAGGATCACCTGAGCCGAGAGGCCAAGGC
TGCAGTGACCCATGACCGACCACTGACTCTAGCC AAAAABFFFFFVGGGGGGAEEEDHGCFF2E2ECGGHB5DCEADBCHBCGG3FE1F?EG11FAA1101101FGCHGB3F3?//EFEEC/FFEEF3B334?B?FFFFHCFCDGHFFGFFFFHDGBG NM:i:6 MD:Z:12A9T16C9C8T33 MC:Z:5S79M49S AS:i:49 XS:i:44 SA:Z:7,73730619,+56M77S,3,5;
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 110472164 0 5H59M69H = 100912013 -9560210 * * NM:i:3 MD:Z:28C7C8T13 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 44591270 0 34H99M = 100912013 56320646 * * NM:i:11 MD:Z:7C6T31A2T0A0G3A0T14G6T15C4 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 116784570 0 42H63M28H = 100912013 -15872620 * * NM:i:4 MD:Z:41T1G0T3T14 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 151266200 0 20H58M55H = 100912013 -50354188 * * NM:i:3 MD:Z:7C14A4C30 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 5142163 0 99M34H = 100912013 95769851 * * NM:i:12 MD:Z:2G15T0A0A21A1T2C0T3T3C6T27G7 MC:Z:5H79M49H AS:i:41
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 283971 0 26H65M42H = 100912013 100628043 * * NM:i:5 MD:Z:21C0T20T10C1C8 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 148973073 0 48H80M5H = 100912013 -48061061 * * NM:i:8 MD:Z:29C4A8A7G1G0T16C6 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 132493804 0 49H50M34H = 100912013 -31581792 * * NM:i:2 MD:Z:3A8G7 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 17046803 0 69H59M5H = 100912013 83865211 * * NM:i:4 MD:Z:11C1A8G7G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 41146108 0 69H59M5H = 100912013 59765906 * * NM:i:4 MD:Z:13A8G1T5G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 81351976 0 5H59M69H = 100912013 19559980 * * NM:i:4 MD:Z:28C5A1C8T13 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 43949171 0 5H59M69H = 100912013 56962785 * * NM:i:4 MD:Z:28C7C8T0A12 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 138980542 0 34H44M55H = 100912013 -38068573 * * NM:i:1 MD:Z:7C36 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 97540837 0 42H63M28H = 100912013 3371115 * * NM:i:5 MD:Z:8T33A0G0A3T14 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 128693225 0 5H58M70H = 100912013 -27781270 * * NM:i:4 MD:Z:28C5A1C8T12 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 98072245 0 49H42M42H = 100912013 2839769 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 28419599 0 49H42M42H = 100912013 72492415 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 111786632 0 4H77M52H = 100912013 -10874696 * * NM:i:8 MD:Z:27C0A0C2A4C8T13C0A15 MC:Z:5H79M49H AS:i:37

```

Sección de cabecera

La **sección de cabecera** comienza con un carácter @, seguido de dos letras mayúsculas que codifican el tipo de registro, tras las que vienen las etiquetas (TAG) separadas por tabuladores, en formato TAG:VALOR: Cada TAG son dos letras mayúsculas que definen el formato y contenido del VALOR. Existen pocas TAG y solo son válidas las definidas en las directrices del formato SAM. En la figura se encuentran dos líneas correspondientes a la cabecera:

- **@SQ:** este registro indica la secuencia de referencia.
 - SN:12, nombre de la secuencia de referencia, en este caso del cromosoma 12.
 - LN:133275309, longitud de la secuencia de referencia.
- **@PG:** este registro hace referencia al programa utilizado para mapear/alinear las lecturas a la secuencia de referencia.
 - ID:bwa, identificador del programa. BWA en este caso
 - PN:bwa, nombre del programa para alinear.
 - VN:0.7.17-r1188, versión del programa utilizado.
 - CL:..., es la línea que indica la línea de comandos utilizada para ejecutar este programa

```

@SQ SN:7 LN:159138663
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem -a Homo_sapiens.GRCh37.dna.chromosome.7.fa out_1.clean.fq.gz out_2.clean.fq.gz -o chr7.sam
M02899:19:000000000-JFB5J:1:1101:17405:2045 83 7 100912013 3 5S79M49S = 100912013 -79 CGTTTAAAAATTAGCTGGGTGTTGGCGCTGTGGCTGTAGTGCAGCTACCCAGGAGGCTGAGGTGGAGGATCACCTGAGGCCGAGAGGCCAAGGC
:TGCA GTACC ATGACCGCACCACTGCACTCTAGCC AAAAABFFFFFFGGGGGGAAEEDHCGFF2E2ECGGHB5DCEADBCHBCGG3FE1F?EG11FAA1101101FGCHCGB3F3????//EFFEEC/FFEEF3B334?B?FFFFHCFCDGHFFFFFFHDDBG NM:i:6 MD:Z:12A9T1G1C9C8T33 MC:Z:5S79M49S AS:i:49 XS
:i:44 SA:Z:7,73730619,+,56M77S,3,5;
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 110472164 0 5H59M69H = 100912013 -9560210 * * NM:i:3 MD:Z:28C7C8T13 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 44591270 0 34H99M = 100912013 56320646 * * NM:i:11 MD:Z:7C6T31A2T0A0G3A0T14G6T15C4 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 116784570 0 42H63M28H = 100912013 -15872620 * * NM:i:4 MD:Z:41T1G0T3T14 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 151266200 0 20H58M55H = 100912013 -50354188 * * NM:i:3 MD:Z:7C14A4C30 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 5142163 0 99M34H = 100912013 95769851 * * NM:i:12 MD:Z:2G15T0A021A1T2C0T3C6T27G7 MC:Z:5H79M49H AS:i:41
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 283971 0 26H65M42H = 100912013 100628043 * * NM:i:5 MD:Z:2I0C0T20T10C1C8 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 148973073 0 48H80M5H = 100912013 -48061061 * * NM:i:8 MD:Z:29C4A8A7G1G0T1T16C6 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 132493804 0 49H50M34H = 100912013 -31581792 * * NM:i:2 MD:Z:33A8G7 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 17046803 0 69H59M5H = 100912013 83865211 * * NM:i:4 MD:Z:11C1A8G7G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 41146108 0 69H59M5H = 100912013 59765906 * * NM:i:4 MD:Z:13A8G1T5G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 81351976 0 5H59M69H = 100912013 195559980 * * NM:i:4 MD:Z:28C5A1C8T13 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 43949171 0 5H59M69H = 100912013 56962785 * * NM:i:4 MD:Z:28C7C8T0A12 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 138980542 0 34H44M55H = 100912013 -38068573 * * NM:i:1 MD:Z:7C36 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 97540837 0 42H63M28H = 100912013 3371115 * * NM:i:5 MD:Z:8T33A0G0A3T14 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 128693225 0 5H58M70H = 100912013 -27781270 * * NM:i:4 MD:Z:28C5A1C8T12 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 98072245 0 49H42M42H = 100912013 2839769 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 28419599 0 49H42M42H = 100912013 72492415 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 111786632 0 4H77M52H = 100912013 -10874696 * * NM:i:8 MD:Z:27C0A0C2A4C8T13C0A15 MC:Z:5H79M49H AS:i:37

```

Sección de alineamientos

La **sección de alineamientos** va desde la primera línea que no comienza por @ hasta la última del fichero.

Contiene la información del alineamiento o mapeo de las lecturas con el genoma de referencia.

En esta sección cada línea corresponde al alineamiento de un segmento. Deben aparecer todos los campos obligatorios, siempre en el mismo orden, aunque su valor sea '0' o '*' cuando la información no está disponible.

Tras estos valores vienen los campos opcionales. Las columnas separadas por tabulador son las siguientes:

QNAME (Query name): nombre de la lectura en el alineamiento

FLAG: describe las propiedades del alineamiento. Es un campo fundamental, en el que ahondaremos más adelante.

RNAME: nombre de la secuencia de referencia.

POS: posición del inicio del alineamiento.

```

@SQ SN:7 LN:159138663
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem -a Homo_sapiens.GRCh37.dna.chromosome.7.fa out_1.clean.fq.gz out_2.clean.fq.gz -o chr7.sam
M02899:19:000000000-JFB5J:1:1101:17405:2045 83 7 100912013 3 5S79M49S = 100912013 -79 CGTTTAAAAATTAGCTGGGTGTGGTGGCGTGTTGAGGTGAGGTGGAGGATCACCTGAGCCGAGAGGCCAAGGC
TGAGTGACCATGACCGCACCACTGCACTCTAGCC AAAAABFFFFFFGGGGGGGAEEDHCGGFF2E2ECGGHB5DCEADBCHBCGG3FE1F?EG11FAA1101101FGCHCGB3F3????//EFEEC/FFEEF3B334?B?FFFFHCFCDGHFFFHDGBG NM:i:6 MD:Z:12A9T1G1C9C8T33 MC:Z:5S79M49S AS:i:49 XS
:i:44 SA:Z:7,73730619,+,56M77S,3,5;
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 110472164 0 5H59M69H = 100912013 -9560210 * * NM:i:3 MD:Z:28C7C8T13 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 44591270 0 34H9M = 100912013 56320646 * * NM:i:11 MD:Z:7C6T3IA2TOA0G3A0T14G6T15C4 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 116784570 0 42H63M28H = 100912013 -15872620 * * NM:i:4 MD:Z:41T1G0T3T14 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 151266200 0 20H58M55H = 100912013 -50354188 * * NM:i:3 MD:Z:7C14A4C30 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 5142163 0 99M34H = 100912013 95769851 * * NM:i:12 MD:Z:2G15T0A0A21A1T2C0T3C6T27G7 MC:Z:5H79M49H AS:i:41
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 283971 0 26H65M42H = 100912013 100628043 * * NM:i:5 MD:Z:21C0T20T10C1C8 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 148973073 0 48H80M5H = 100912013 -48061061 * * NM:i:8 MD:Z:29C4A8A7G1G0T1T16C6 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 132493804 0 49H50M34H = 100912013 -31581792 * * NM:i:2 MD:Z:33A8G7 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 17046803 0 69H59M5H = 100912013 83865211 * * NM:i:4 MD:Z:11C1A8G7G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 41146108 0 69H59M5H = 100912013 59765906 * * NM:i:4 MD:Z:13A8G1T5G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 81351976 0 5H59M69H = 100912013 19559980 * * NM:i:4 MD:Z:28C5A1C8T13 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 43949171 0 5H59M69H = 100912013 56962785 * * NM:i:4 MD:Z:28C7C8T0A12 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 138980542 0 34H44M55H = 100912013 -38068573 * * NM:i:1 MD:Z:7C36 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 97540837 0 42H63M28H = 100912013 3371115 * * NM:i:5 MD:Z:8T33A0G0A3T14 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 128693225 0 5H58M70H = 100912013 -27781270 * * NM:i:4 MD:Z:28C5A1C8T12 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 98072245 0 49H42M42H = 100912013 2839769 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 28419599 0 49H42M42H = 100912013 72492415 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 111786632 0 4H77M52H = 100912013 -10874696 * * NM:i:8 MD:Z:27C0A0C2A4C8T13C0A15 MC:Z:5H79M49H AS:i:37

```

MAPQ: calidad del mapeo. Este valor lo otorga el programa de mapeo en función de la calidad del alineamiento, y su valor es más alto cuanto más perfecto sea. Valora también la presencia de mapeos secundarios de esa lectura.

CIGAR: este campo muestra las operaciones realizadas para alinear las dos secuencias (lectura y referencia).

RNEXT: nombre de la secuencia en al que ha alineado su pareja, cuando las lecturas son pareadas. El símbolo '=' indica que la pareja está mapeada sobre la misma referencia.

PNEXT: posición de la pareja en el alineamiento.

TLEN: longitud del alineamiento, calculada según las coordenadas de la secuencia de referencia.

SEQ: secuencia de la lectura que participa en el alineamiento.

QUAL: calidad de la lectura, extraída del archivo FASTQ original.

Los campos opcionales aparecen a continuación, separados por tabuladores, en el formato TAG:TIPO:VALOR.

```

@SQ SN:7 LN:159138663
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem -a Homo_sapiens.GRCh37.dna.chromosome.7.fa out_1.clean.fq.gz out_2.clean.fq.gz -o chr7.sam
M02899:19:000000000-JFB5J:1:1101:17405:2045 83 7 100912013 3 5S79M49S = 100912013 -79 CGTTAAAAATTAGCTGGGTGTTGGCGTGTGGCTGTAGTGCCAGCTACCCAGGAGGCTGAGGTGGAGGATCACCTGAGCCGAGAGGCCAAGGC
TGCA GTGACCCATGACCGACCCTGACTCTAGCC AAAAABFFFFFFGGGGGGGAEEEDHCGFF2E2ECGGHB5DCEADBCHBCGG3FE1F?EG11FAA1101101FGCHGB3F3????//EFEEC/FFEEF3B334?B?FFFFHCFCDGHFFGFFFFHDGBG NM:i:6 MD:Z:12A9T1G1C9C8T33 MC:Z:5S79M49S AS:i:49 XS
:i:44 SA:Z:7,73730619,+,.56M77S,3,5;
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 110472164 0 5H59M69H = 100912013 -9560210 * * NM:i:3 MD:Z:28C7C8T13 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 44591270 0 34H99M = 100912013 56320646 * * NM:i:11 MD:Z:7C6T31A2T0A0G3A0T14G6T15C4 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 116784570 0 42H63M28H = 100912013 -15872620 * * NM:i:4 MD:Z:41T1G0T3T14 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 151266200 0 20H58M55H = 100912013 -50354188 * * NM:i:3 MD:Z:7C14A4C30 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 5142163 0 99M34H = 100912013 95769851 * * NM:i:12 MD:Z:2G15T0A0A21ALT2C0T3T3C6T27G7 MC:Z:5H79M49H AS:i:41
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 283971 0 26H65M42H = 100912013 100628043 * * NM:i:5 MD:Z:21C0T20T10C1C8 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 148973073 0 48H80M5H = 100912013 -48061061 * * NM:i:8 MD:Z:29C4A8A7G1G0T1T16C6 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 132493804 0 49H50M34H = 100912013 -31581792 * * NM:i:2 MD:Z:33A8G7 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 17046803 0 69H59M5H = 100912013 83865211 * * NM:i:4 MD:Z:11C1A8G7G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 41146108 0 69H59MSH = 100912013 59765906 * * NM:i:4 MD:Z:13A8G1T5G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 81351976 0 5H59M69H = 100912013 19559980 * * NM:i:4 MD:Z:28C5A1C8T13 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 43949171 0 5H59M69H = 100912013 56962785 * * NM:i:4 MD:Z:28C7C8T0A12 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 138980542 0 34H44M55H = 100912013 -38068573 * * NM:i:1 MD:Z:7C36 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 97540837 0 42H63M28H = 100912013 3371115 * * NM:i:5 MD:Z:8T33A0G0A3T14 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 128693225 0 5H58M70H = 100912013 -27781270 * * NM:i:4 MD:Z:28C5A1C8T12 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 98072245 0 49H42M42H = 100912013 2839769 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 28419599 0 49H42M42H = 100912013 72492415 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 111786632 0 4H77M52H = 100912013 -10874696 * * NM:i:8 MD:Z:27C0A0C2A4C8T13C0A15 MC:Z:5H79M49H AS:i:37

```

En este documento se pueden encontrar las TAG que hay para los campos opcionales.

<https://samtools.github.io/hts-specs/SAMtags.pdf>

```

@SQ SN:7 LN:159138663
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem -a Homo_sapiens.GRCh37.dna.chromosome.7.fa out_1.clean.fq.gz out_2.clean.fq.gz -o chr7.sam
M02899:19:000000000-JFB5J:1:1101:17405:2045 83 7 100912013 3 5579M49S = 100912013 -79 CGTTTAAAAATTAGCTGGGTGTTGGCGTGTGGCTGTAGTGCAGTACCCAGGAGGCTGAGGTGGGAGGATCACCTGAGCCCCGAGAGGCCAAGGC
TGCAGTGACCATGACCGCACCCTGACTCTAGCC AAAAABFFFFFGGGGGGGAAEEDHGCFF2E2ECGGHB5DCEADBCHBCGG3FE1?EG11FAA1101101FGCHGB3F????//EFEEC/FFEEF3B34?B?FFFFHFCFDGHFFGFFFFHDGBG NM:i:6 MD:Z:12A9T1G1C9C8T33 MC:Z:5579M49S AS:i:49 XS
:i:44 SA:Z:7,73730619,+,56M77S,3,5;

M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 110472164 0 5H59M69H = 100912013 -9560210 * * NM:i:3 MD:Z:28C7C8T13 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 44591270 0 34H99M = 100912013 56320646 * * NM:i:11 MD:Z:7C6T31A2T0A0G3A0T14G6T15C4 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 116784570 0 42H63M28H = 100912013 -15872620 * * NM:i:4 MD:Z:41T1G0T3T14 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 151266200 0 20H58M55H = 100912013 -50354188 * * NM:i:3 MD:Z:7C14A4C30 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 5142163 0 99M34H = 100912013 95769851 * * NM:i:12 MD:Z:2G15T0A0A21A1T2C0T3T3C6T27G7 MC:Z:5H79M49H AS:i:41
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 283971 0 26H65M42H = 100912013 100628043 * * NM:i:5 MD:Z:21C0T20T10C1C8 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 148973073 0 48H80M5H = 100912013 -48061061 * * NM:i:8 MD:Z:29C4A8A7G1G0T1T16C6 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 132493804 0 49H50M34H = 100912013 -31581792 * * NM:i:2 MD:Z:33A8G7 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 17046803 0 69H59M5H = 100912013 83865211 * * NM:i:4 MD:Z:11C1A8G7G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 41146108 0 69H59M5H = 100912013 59765906 * * NM:i:4 MD:Z:13A8G1T5G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 81351976 0 5H59M69H = 100912013 19559980 * * NM:i:4 MD:Z:28C5A1C8T13 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 43949171 0 5H59M69H = 100912013 56962785 * * NM:i:4 MD:Z:28C7C8T0A12 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 138980542 0 34H44M55H = 100912013 -38068573 * * NM:i:1 MD:Z:7C36 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 97540837 0 42H63M28H = 100912013 3371115 * * NM:i:5 MD:Z:8T33A0G0A3T14 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 128693225 0 5H58M70H = 100912013 -27781270 * * NM:i:4 MD:Z:28C5A1C8T12 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 98072245 0 49H42M42H = 100912013 2839769 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 28419599 0 49H42M42H = 100912013 72492415 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 111786632 0 4H77M52H = 100912013 -10874696 * * NM:i:8 MD:Z:27C0A0C2A4C8T13C0A15 MC:Z:5H79M49H AS:i:37

```

Anteriormente se ha comentado que el campo FLAG es obligatorio y que se utiliza para describir con precisión las propiedades del alineamiento. Cada una de las flag está definida, pero en el valor que se observa en el archivo SAM es la suma total de los valores en base decimal asignados para cada una de las flag.

Detalle de las 'flag' utilizadas en el formato SAM.

Nota. Adaptado de <https://www.samformat.info/sam-format-flag>

Número	Decimal	Hexadecimal	Descripción
1	1	0x1	La lectura forma parte de un par
2	2	0x2	La lectura forma parte de un emparejamiento correcto
3	4	0x4	Lectura no mapeada
4	8	0x8	Pareja no mapeada
5	16	0x10	Lectura mapeada en la cadena complementaria (negativa) del genoma de referencia
6	32	0x20	Pareja mapeada en la cadena complementaria (negativa) del genoma de referencia
7	64	0x40	Primera lectura del par
8	128	0x80	Segunda lectura del par
9	256	0x100	Alineamiento secundario
10	512	0x200	La lectura no ha pasado los filtros de calidad
11	1024	0x400	La lectura es un duplicado óptico o de PCR
12	2048	0x800	Alineamiento suplementario

```

@SQ SN:7 LN:159138663
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem -a Homo_sapiens.GRCh37.dna.chromosome.7.fa out_1.clean.fq.gz out_2.clean.fq.gz -o chr7.sam
M02899:19:000000000-JFB5J:1:1101:17405:2045 83 7 100912013 3 5S79M49S = 100912013 -79 CGTTTAAAAATTAGCTGGGTGTTGGCGTGTGGCTGTAGTGCAGTACCCAGGAGGCTGAGGTGGGAGGATCACCTGAGCCCCGAGAGGCCAAGGC
TGCA GTGACCCATGACCGCACCCTGCACCTAGCC AAAAABFFFFFGGGGGGGAAEEDHGCFF2E2ECGGHB5DCEADBCHBCGG3FE1F?EG11FAA1101101FGCHGB3F3??????//EFEEC/FFEEF3B34?B?FFFFHFCFDGHFFGFFFFHDGBG NM:i:6 MD:Z:12A9T1G1C9C8T33 MC:Z:5S79M49S AS:i:49 XS
:i:44 SA:Z:7,73730619,+,56M77S,3,5;

M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 110472164 0 5H59M69H = 100912013 -9560210 * * NM:i:3 MD:Z:28C7C8T13 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 44591270 0 34H99M = 100912013 56320646 * * NM:i:11 MD:Z:7C6T31A2T0A0G3A0T14G6T15C4 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 116784570 0 42H63M28H = 100912013 -15872620 * * NM:i:4 MD:Z:41T1G0T3T14 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 151266200 0 20H58M55H = 100912013 -50354188 * * NM:i:3 MD:Z:7C14A4C30 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 5142163 0 99M34H = 100912013 95769851 * * NM:i:12 MD:Z:2G15T0A0A21A1T2C0T3T3C6T27G7 MC:Z:5H79M49H AS:i:41
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 283971 0 26H65M42H = 100912013 100628043 * * NM:i:5 MD:Z:21C0T20T10C1C8 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 148973073 0 48H80M5H = 100912013 -48061061 * * NM:i:8 MD:Z:29C4A8A7G1G0T1T16C6 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 132493804 0 49H50M34H = 100912013 -31581792 * * NM:i:2 MD:Z:33A8G7 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 17046803 0 69H59M5H = 100912013 83865211 * * NM:i:4 MD:Z:11C1A8G7G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 41146108 0 69H59M5H = 100912013 59765906 * * NM:i:4 MD:Z:13A8G1T5G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 81351976 0 5H59M69H = 100912013 19559980 * * NM:i:4 MD:Z:28C5A1C8T13 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 43949171 0 5H59M69H = 100912013 56962785 * * NM:i:4 MD:Z:28C7C8T0A12 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 138980542 0 34H44M55H = 100912013 -38068573 * * NM:i:1 MD:Z:7C36 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 97540837 0 42H63M28H = 100912013 3371115 * * NM:i:5 MD:Z:8T33A0G0A3T14 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 128693225 0 5H58M70H = 100912013 -27781270 * * NM:i:4 MD:Z:28C5A1C8T12 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 98072245 0 49H42M42H = 100912013 2839769 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 28419599 0 49H42M42H = 100912013 72492415 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 111786632 0 4H77M52H = 100912013 -10874696 * * NM:i:8 MD:Z:27C0A0C2A4C8T13C0A15 MC:Z:5H79M49H AS:i:37

```

Ejemplos:

Flag 83 (1+2+16+64) : la lectura forma parte de un par (1) correctamente alineado (2), es la primera lectura del par (64) y se ha mapeado en la cadena complementaria (antisentido) del genoma de referencia (16).

Flag 99 (1+2+32+64) : La lectura forma parte de un par (1) correctamente alineado (2), es la primera lectura del par (64) y se ha mapeado en la cadena positiva (sentido) del genoma de referencia. Este último punto lo sabemos porque el código 32 indica que su pareja se ha mapeado en la cadena antisentido del genoma de referencia.

Flag 73 (1+8+64) : La lectura forma parte de un par (1), es la primera lectura (64) del par, pero su pareja no ha sido mapeada.

```

@SQ SN:7 LN:159138663
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem -a Homo_sapiens.GRCh37.dna.chromosome.7.fa out_1.clean.fq.gz out_2.clean.fq.gz -o chr7.sam
M02899:19:000000000-JFB5J:1:1101:17405:2045 83 7 100912013 3 5579M49S = 100912013 -79 CGTTTAAAAATTAGCTGGGTGTTGGCGTGTGGCTGTAGTGCAGTACCCAGGAGGCTGAGGTGGGAGGATCACCTGAGCCCCGAGAGGCCAAGGC
TGCA GTGACCCATGACCGCACCCTGACTCTAGCC AAAAABFFFFFGGGGGGGAAEEDHGCGFF2E2ECGGHB5DCEADBCHBCGG3FE1?EG11FAA1101101FGCHGB3F3??????//EFEEC/FFEEF3B334?B?FFFFHFCFDGHFFGFFFFHDGBG NM:i:6 MD:Z:12A9T1G1C9C8T33 MC:Z:5579M49S AS:i:49 XS
:i:44 SA:Z:7,73730619,+,56M77S,3,5;

M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 110472164 0 5H59M69H = 100912013 -9560210 * * NM:i:3 MD:Z:28C7C8T13 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 44591270 0 34H99M = 100912013 56320646 * * NM:i:11 MD:Z:7C6T31A2T0A0G3A0T14G6T15C4 MC:Z:5H79M49H AS:i:44
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 116784570 0 42H63M28H = 100912013 -15872620 * * NM:i:4 MD:Z:41T1G0T3T14 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 151266200 0 20H58M55H = 100912013 -50354188 * * NM:i:3 MD:Z:7C14A4C30 MC:Z:5H79M49H AS:i:43
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 5142163 0 99M34H = 100912013 95769851 * * NM:i:12 MD:Z:2G15T0A0A21A1T2C0T3T3C6T27G7 MC:Z:5H79M49H AS:i:41
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 283971 0 26H65M42H = 100912013 100628043 * * NM:i:5 MD:Z:21C0T20T10C1C8 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 148973073 0 48H80M5H = 100912013 -48061061 * * NM:i:8 MD:Z:29C4A8A7G1G0T1T16C6 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 132493804 0 49H50M34H = 100912013 -31581792 * * NM:i:2 MD:Z:33A8G7 MC:Z:5H79M49H AS:i:40
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 17046803 0 69H59M5H = 100912013 83865211 * * NM:i:4 MD:Z:11C1A8G7G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 41146108 0 69H59M5H = 100912013 59765906 * * NM:i:4 MD:Z:13A8G1T5G28 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 81351976 0 5H59M69H = 100912013 19559980 * * NM:i:4 MD:Z:28C5A1C8T13 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 43949171 0 5H59M69H = 100912013 56962785 * * NM:i:4 MD:Z:28C7C8T0A12 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 138980542 0 34H44M55H = 100912013 -38068573 * * NM:i:1 MD:Z:7C36 MC:Z:5H79M49H AS:i:39
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 97540837 0 42H63M28H = 100912013 3371115 * * NM:i:5 MD:Z:8T33A0G0A3T14 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 128693225 0 5H58M70H = 100912013 -27781270 * * NM:i:4 MD:Z:28C5A1C8T12 MC:Z:5H79M49H AS:i:38
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 98072245 0 49H42M42H = 100912013 2839769 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 323 7 28419599 0 49H42M42H = 100912013 72492415 * * NM:i:1 MD:Z:33A8 MC:Z:5H79M49H AS:i:37
M02899:19:000000000-JFB5J:1:1101:17405:2045 339 7 111786632 0 4H77M52H = 100912013 -10874696 * * NM:i:8 MD:Z:27C0A0C2A4C8T13C0A15 MC:Z:5H79M49H AS:i:37

```

Si deseáis ahondar en el formato SAM, especialmente en las flag, cabeceras y etiquetas de alineamiento, aquí se encuentra mucha información : <https://www.samformat.info/sam-format-flag>

Para poder realizar el cálculo automático de las flag podéis utilizar el recurso siguiente:
<https://broadinstitute.github.io/picard/explain-flags.html>

This utility makes it easy to identify what are the properties of a read based on its SAM flag value, or conversely, to find what the SAM Flag value would be for a given combination of properties.

To decode a given SAM flag value, just enter the number in the field below. The encoded properties will be listed under Summary below, to the right.

SAM Flag:

[Explain](#)

[Switch to mate](#) Toggle first in pair / second in pair

Find SAM flag by property:

To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- read paired
- read mapped in proper pair
- read unmapped
- mate unmapped
- read reverse strand
- mate reverse strand
- first in pair
- second in pair
- not primary alignment
- read fails platform/vendor quality checks
- read is PCR or optical duplicate
- supplementary alignment

Summary:

- read paired (0x1)
- read mapped in proper pair (0x2)
- read reverse strand (0x10)
- first in pair (0x40)
- not primary alignment (0x100)

This utility makes it easy to identify what are the properties of a read based on its SAM flag value, or conversely, to find what the SAM Flag value would be for a given combination of properties.

To decode a given SAM flag value, just enter the number in the field below. The encoded properties will be listed under Summary below, to the right.

SAM Flag: [Explain](#)

[Switch to mate](#) Toggle first in pair / second in pair

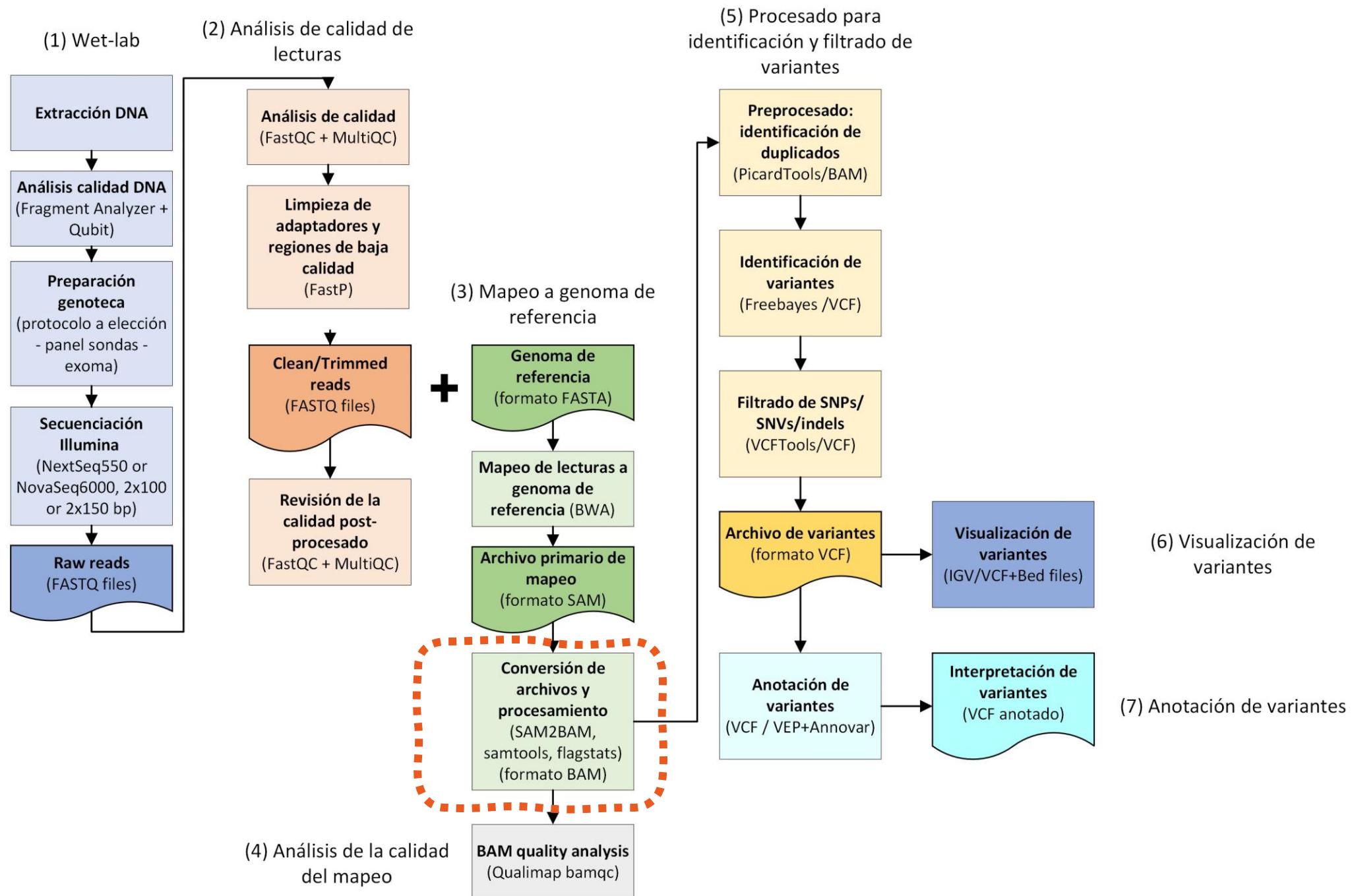
Find SAM flag by property:

To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- read paired
- read mapped in proper pair
- read unmapped
- mate unmapped
- read reverse strand
- mate reverse strand
- first in pair
- second in pair
- not primary alignment
- read fails platform/vendor quality checks
- read is PCR or optical duplicate
- supplementary alignment

Summary:

- read paired (0x1)
- read mapped in proper pair (0x2)
- first in pair (0x40)
- not primary alignment (0x100)



4.2.4. Herramientas para el manejo de los archivos SAM. Generación de archivos BAM.

Hasta ahora hemos visto archivos de mapeo SAM, de texto plano, que habitualmente tienen un “peso” alto, computacionalmente hablando. Además, manejar archivos de texto no es computacionalmente rentable. Para ahorrar espacio y agilizar los cálculos posteriores, los archivos SAM se convierten en **archivos binarios**, denominados **BAM**. Ambos archivos contienen la misma información.

Para realizar esta conversión se utiliza un paquete de herramientas llamadas “**samtools**”, que nos permiten leer, filtrar, transformar y extraer información de los ficheros BAM y SAM. <http://www.htslib.org/>

Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

Samtools Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format

BCFtools Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants

HTSlib A C library for reading/writing high-throughput sequencing data

Samtools and BCFtools both use HTSlib internally, but these source packages contain their own copies of htslib so they can be built independently.



Tema 4 - Ejemplo 4 - Indexado

Vamos a manipular ahora el archivo SAM obtenido en el ejemplo 3. Para ello vamos a utilizar diferentes herramientas del software Samtools, que tenéis descritas aquí:

“samtools view” para leer un fichero SAM/BAM. Esta herramienta nos permite leer un fichero SAM y convertirlo en un fichero BAM. `samtools view -bS mapping.sam > mapping.bam`

“samtools view” para visualizar un archivo BAM `<samtools view -h mapping.bam>`

“samtools view” para seleccionar los alineamientos con una calidad superior a la especificada. `<samtools view -q 30 mapping.bam>`

“samtools sort” permite ordenar un fichero por las coordenadas genómicas. Este paso es imprescindible para visualización en programas específicos. `<samtools sort mapping.bam -o mapping.sorted.bam>`

“samtools index” genera un archivo índice (fichero con extensión BAI) para un fichero BAM ordenado. `<samtools index mapping.sorted.bam>`

“samtools [flagstat](#)” nos da unas estadísticas básicas pero útiles para tener una visión general sobre los alineamientos en el archivo BAM. `<samtools flagstat mapping.sorted.bam>`

Ejemplo (como mucho cada proceso debe durar 5 min):

1) Conversión del archivo SAM a BAM

```
 samtools view -bS chr7.sam > chr7.bam
```

2) Visualizar el archivo BAM

```
 samtools view -h chr7.bam  
 samtools view -h chr7.bam | more
```

3) Seleccionar alineamientos por encima de una calidad (-q [integer])

```
 samtools view -q 30 chr7.bam  
 samtools view -q 30 chr7.bam | more
```

4) Ordenar por coordenadas un bam

```
 samtools sort chr7.bam > chr7.sorted.bam
```

5) Indexar un archivo bam

```
 samtools index chr7.sorted.bam
```

6) Obtener unas estadísticas básicas del mapeo

```
 samtools flagstat chr7.sorted.bam
```

EXTRA!

Samtools flagstat

<http://www.htslib.org/doc/samtools-flagstat.html>

DESCRIPTION

Does a full pass through the input file to calculate and print statistics to stdout.

Provides counts for each of 13 categories based primarily on bit flags in the FLAG field. Information on the meaning of the flags is given in the SAM specification document <<https://samtools.github.io/hts-specs/SAMv1.pdf>>.

Each category in the output is broken down into QC pass and QC fail. In the default output format, these are presented as "#PASS + #FAIL" followed by a description of the category.

The first row of output gives the total number of reads that are QC pass and fail (according to flag bit 0x200). For example:

122 + 28 in total (QC-passed reads + QC-failed reads)

Which would indicate that there are a total of 150 reads in the input file, 122 of which are marked as QC pass and 28 of which are marked as "not passing quality controls"

Singlets: solo una de las lecturas de la pareja alinea. Puede ser también que de la pareja, uno alinee múltiples veces y el otro solo una, en ese caso uno tendrá un registro individual y el otro múltiple.

"properly paired": ambas lecturas de la pareja alinean sobre el mismo cromosoma

"with itself and mate mapped": ambas lecturas de la pareja alinean. Pueden ser "properly paired" (mismo chr) o no. De todos modos, si ambos se alinean en algún lugar al menos una vez, entonces cuentan para esto.

```
(04MBIF_humano) [UNIVERSIDADVIU\laura.gutierrez.m@a-3uau7bcnl6t98 mapping]$ samtools flagstat chr7.sorted.bam  
3209165 + 0 in total (QC-passed reads + QC-failed reads)  
1638116 + 0 primary  
1552388 + 0 secondary  
18661 + 0 supplementary  
0 + 0 duplicates  
0 + 0 primary duplicates  
2127870 + 0 mapped (66.31% : N/A)  
556821 + 0 primary mapped (33.99% : N/A)  
1638116 + 0 paired in sequencing  
819058 + 0 read1  
819058 + 0 read2  
515896 + 0 properly paired (31.49% : N/A)  
540348 + 0 with itself and mate mapped  
16473 + 0 singletons (1.01% : N/A)  
0 + 0 with mate mapped to a different chr  
0 + 0 with mate mapped to a different chr (mapQ<=5)
```

```
04@4MBTF_humano [JUNTVFSTDADVTU]\laura.gutierrez.m@alumni.bcnl6t98 mapping]$ samtools flagstat chr7.sorted.bam
3209165 + 0 in total (QC-passed reads + QC-failed reads)
0 1638116 + 0 primary
1552388 + 0 secondary
18661 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
2127870 + 0 mapped (66.31% : N/A)
556821 + 0 primary mapped (33.99% : N/A)
1638116 + 0 secondary mapped (55.25% : N/A)
18661 + 0 supplementary mapped (0.65% : N/A)
```

3209165 reads totales que pasan filtro, de ellos:

- 1638116 son alineamientos primarios
- 1552388 son alineamientos secundarios (cuando un read mapea sobre lugares múltiples debido a la ambigüedad de localizaciones múltiples, i.e. debido a repeticiones, sólo uno de los múltiples alineamientos se considera primario – decisión arbitraria-, mientras que todos los demás son considerados secundarios)
- 18661 son alineamiento suplementarios (reads quiméricos, alinean en distintas porciones del genoma con poco solapamiento. Suelen representar variaciones estructurales)

```
0 + 0 duplicates  
0 + 0 primary duplicates
```

En este punto del ejercicio no hemos marcado duplicados sobre el BAM, por ello no aparecen duplicados marcados en esos campos.

Marcaremos duplicados en una etapa posterior de nuestro pipeline

```
o + o primary duplicates  
n 2127870 + 0 mapped (66.31% : N/A)  
n 556821 + 0 primary mapped (33.99% : N/A)  
4 1638116 + 0 paired in sequencing  
U 819058 + 0 read1  
U 819058 + 0 read2  
U 515896 + 0 properly paired (31.49% : N/A)  
U 540348 + 0 with itself and mate mapped  
U 16473 + 0 singletons (1.01% : N/A)  
U 0 + 0 with mate mapped to a different chr  
d 0 + 0 with mate mapped to a different chr (mapQ>=5)
```

- Un total de 2127870 lecturas mapeadas a referencia (chr7) --> 66,31%
- De ellos, el 33,99% son considerados mapeo primario
- Del total de lecturas mapeadas, 1638116 son pareadas, lo que corresponde a 819058 reads1 (forward) y 819058 reads2 (reverse)
- De éstos, un total de 515896 "properly paired" --> mismo cromosoma en orientación opuesta y con poca desviación del tamaño de inserto esperado
- Un total de 540348 lecturas mapean de manera "with itself and mate mapped", lo que indica que mapean ambos reads de la pareja en algún lugar del genoma, pero no tienen por qué estar en el mismo chr o en orientación opuesta. Nota: más abajo vemos que no existen parejas mapeadas en chr diferentes (normal porque sólo hemos mapeado un cromosoma)
- 16473 reads son singletones (no tienen pareja)

4.2.5. Visualización del mapeo

Antes de continuar en el análisis del mapeo, se puede realizar una inspección visual del alineamiento de las lecturas contra el genoma de referencia, siendo un paso complementario a cualquier proceso de análisis de variantes genómicas.

Para este fin se utiliza comúnmente el programa **Integrative Genomics Viewer (IGV)**.

En el siguiente link se encuentra la información de este software, así como su descarga.

<https://software.broadinstitute.org/software/igv/>

Este visualizador funciona en distintos sistemas operativos, como Windows, Linux o Mac.

Se encuentra instalado en la máquina virtual de trabajo, y puede abrirse desde la terminal, tecleando 'igv' en el prompt.

Para ahondar en la visualización de archivos BAM en el explorador IGV, se incluye el siguiente tutorial.

https://www.youtube.com/watch?v=E_G8z_2gTYM

4.2.5. Visualización del mapeo

Los puntos más importantes para tener en cuenta cuando revisemos la visualización de un genoma son:

- Cargar en el programa el genoma de referencia adecuado a nuestro análisis. IGV incluye algunos genomas de referencia estándar, pero pueden descargarse un gran número de ellos e incluso cargar un genoma de referencia propio, a partir de un archivo FASTA.
- Tener un archivo BAM del mapeo previamente ordenador por coordenadas (tal y como se ha visto en el apartado 4.2.4 de este documento) y con su índice BAI, localizado en la misma carpeta del archivo.
- Es opcional cargar en el programa el archivo de sondas de captura, que se trata de un archivo en formato BED donde se especifica para cada cromosoma, qué región cubre cada sonda.



Tema 4 - Ejemplo 5 - IGV

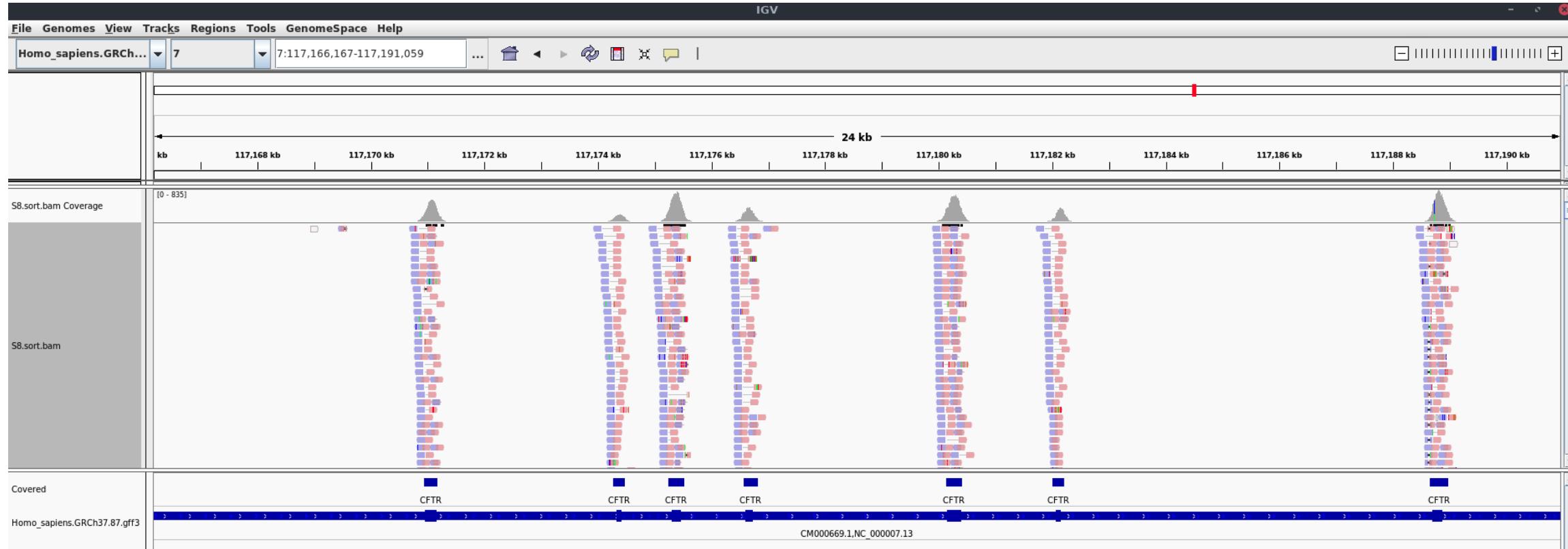


En este paso vamos a visualizar el archivo BAM obtenido en el ejemplo anterior en IGV. Para ello necesitamos:

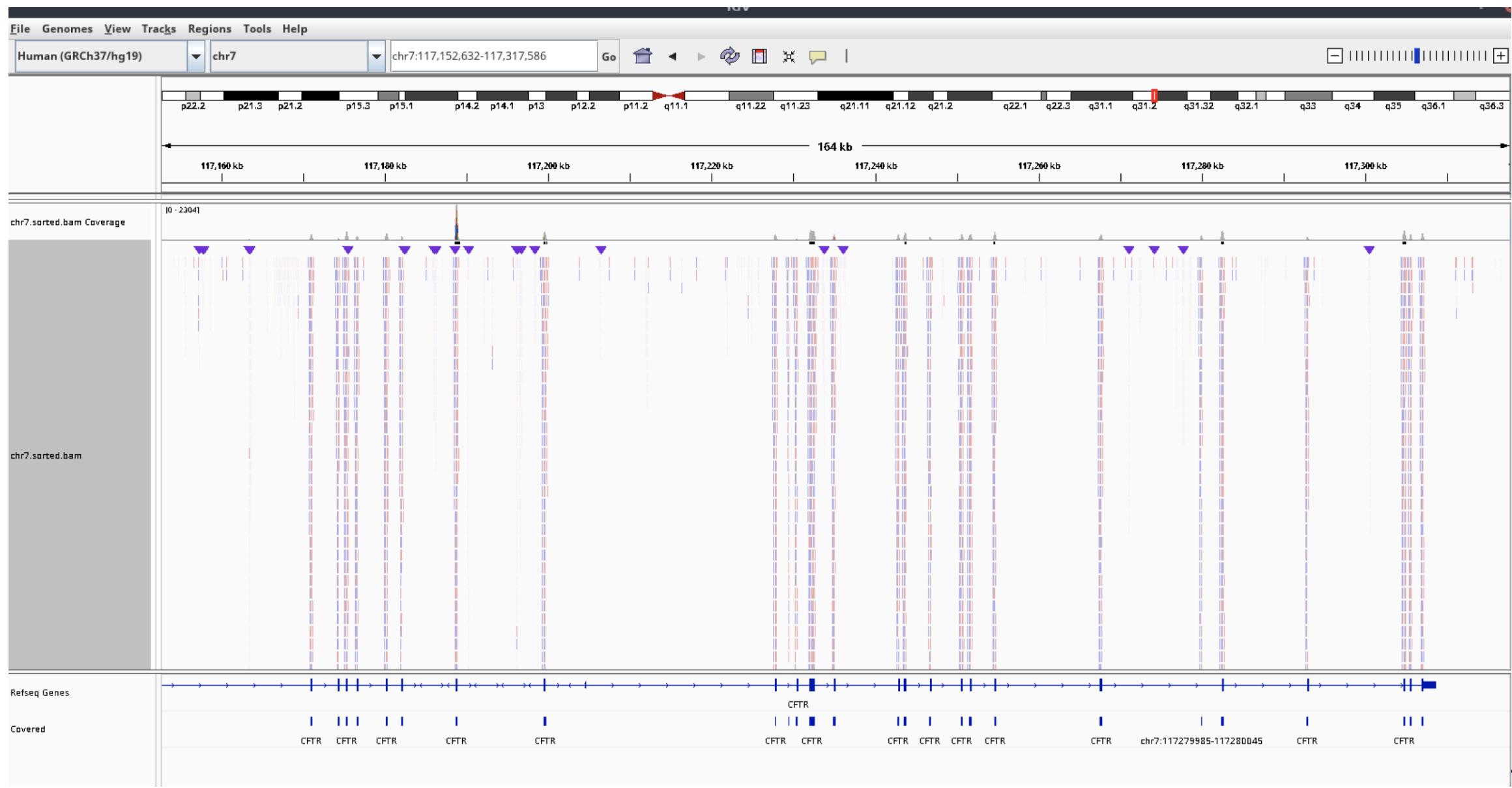
- archivo BAM (sorted, ordenado)
- archivo BAM indexado (.bam.bai). Es obligatorio siempre tener el archivo indexado.
- sondas utilizadas en este panel (archivo BED)

Tenéis los archivos necesarios aquí: [Tema4_Ejemplo5_Visualización](#)

Tema 4 – Ejemplo 5

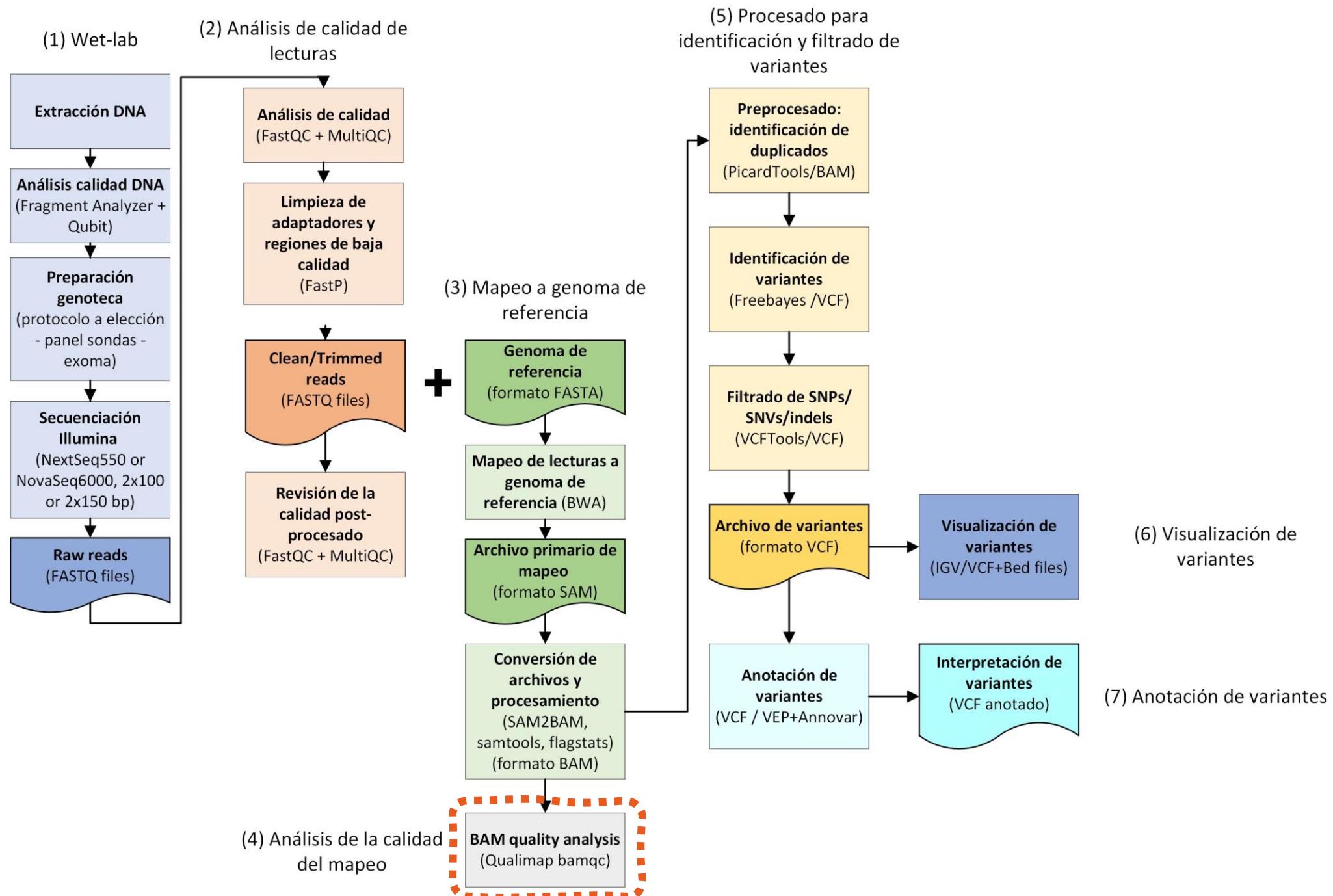


Tema 4 – Ejemplo 5



Tema 4 – Ejemplo 5





4.2.6. Análisis de la calidad del mapeo

El proceso de análisis de la calidad del mapeo es fácil si disponemos de unas pocas secuencias, pero en este caso, donde se han manejado millones de ellas, puede resultar muy complejo.

Una persona altamente entrenada en visualizar los mapeos tiene una idea bastante aproximada de la calidad de este, simplemente realizando una inspección visual del mismo.

En este apartado analizaremos la calidad del mapeo utilizando **Qualimap2** (Okonechnikov et al., 2015), una herramienta con interfaz gráfica y por línea de comandos que nos permiten la evaluación de experimentos de secuenciación de genoma o exoma o incluso experimentos de secuenciación de ARN.

La última versión del programa, así como una completa guía de uso está disponible en los siguientes enlaces.

<http://qualimap.conesalab.org/>

http://qualimap.conesalab.org/doc_html/index.html

Los siguientes programas son herramientas muy útiles en determinar la calidad de un mapeo sobre un archivo BAM, siendo programas más sencillos en cuanto a resultados ofrecidos.

SAMstat: <http://samstat.sourceforge.net/>

ngSCAT (especializado en evaluar capturas dirigidas): http://ngscat.clinbioinfosspa.es/media/ngscat/documentation/two_examples/captureqc.html

4.2.6. Análisis de la calidad del mapeo

El módulo **BamQC** de **Qualimap** genera un informe en el que se evalúa la calidad de los datos de alineamiento del archivo BAM. Nos genera la siguiente información

Ejemplos:

https://public-docs.crg.es/biocore/lcozzuto/course-PhD-Oct-2018/QC/QUALIMAP_B7_H3K4me1/qualimapReport.html

https://kokonech.github.io/qualimap/HG00096.chr20_bamqc/qualimapReport.html#genome_coverage_histogram.png

Input data and parameters

QualiMap command line

```
qualimap bamqc -bam B7_H3K4me1_d_s.bam -gff genome.gtf -nw 400 -hm 3
```

Alignment

Command line:	"/usr/local/bin/bowtie2-align-s --wrapper basic-0 --non-deterministic -x bowtie2genome -p 8 -U B7_H3K4me1.fastq.gz"
Draw chromosome limits:	no
Analyze overlapping paired-end reads:	no
Program:	bowtie2 (2.3.2)
Analysis date:	Fri Aug 10 15:55:44 UTC 2018
Size of a homopolymer:	3
Skip duplicate alignments:	no
Number of windows:	400
BAM file:	B7_H3K4me1_d_s.bam

GFF region

Library protocol:	non-strand-specific
Outside statistics:	no
GFF file:	genome.gtf

4.2.6. Análisis de la calidad del mapeo

Resumen. Información y estadísticas básicas de alineamiento, como son datos globales sobre el número de lecturas, número de lecturas mapeadas, eficacia del mapeo emparejado, distribución de la longitud de las lecturas, número de lecturas recortadas, tasa de duplicación, contenido en nucleótidos y porcentaje GC, cobertura media y desviación típica de la profundidad, calidad media de los mapeos, tamaño medio de inserto, visión general de la tasa de error y estadísticas sobre el cromosoma (número de bases mapeadas, media y desviación estándar para cada uno de los cromosomas).

Summary

Globals

Reference size	2,725,537,669
Number of reads	36,547,722
Mapped reads	35,531,328 / 97.22%
Unmapped reads	1,016,394 / 2.78%
Mapped paired reads	0 / 0%
Read min/max/mean length	35 / 76 / 75.07
Clipped reads	0 / 0%

Globals (inside of regions)

Regions size/percentage of reference	1,168,286,534 / 42.86%
Mapped reads	19,674,747 / 53.83%
Duplicated reads (estimated)	383,413 / 1.95%
Duplication rate	1.91%

4.2.6. Análisis de la calidad del mapeo

Resumen. Información y estadísticas básicas de alineamiento, como son datos globales sobre el número de lecturas, número de lecturas mapeadas, eficacia del mapeo emparejado, distribución de la longitud de las lecturas, número de lecturas recortadas, tasa de duplicación, contenido en nucleótidos y porcentaje GC, cobertura media y desviación típica de la profundidad, calidad media de los mapeos, tamaño medio de inserto, visión general de la tasa de error y estadísticas sobre el cromosoma (número de bases mapeadas, media y desviación estándar para cada uno de los cromosomas).

ACGT Content (inside of regions)

Number/percentage of A's	400,128,493 / 27.13%
Number/percentage of C's	337,158,553 / 22.86%
Number/percentage of T's	400,846,538 / 27.18%
Number/percentage of G's	336,873,983 / 22.84%
Number/percentage of N's	18,359 / 0%
GC Percentage	45.7%

Coverage (inside of regions)

Mean	1.2627
Standard Deviation	1.921

Mapping Quality (inside of regions)

Mean Mapping Quality	38.78
----------------------	-------

4.2.6. Análisis de la calidad del mapeo

Resumen. Información y estadísticas básicas de alineamiento, como son datos globales sobre el número de lecturas, número de lecturas mapeadas, eficacia del mapeo emparejado, distribución de la longitud de las lecturas, número de lecturas recortadas, tasa de duplicación, contenido en nucleótidos y porcentaje GC, cobertura media y desviación típica de la profundidad, calidad media de los mapeos, tamaño medio de inserto, visión general de la tasa de error y estadísticas sobre el cromosoma (número de bases mapeadas, media y desviación estándar para cada uno de los cromosomas).

Mismatches and indels (inside of regions)

General error rate	0.36%
Mismatches	4,960,265
Insertions	139,218
Mapped reads with at least one insertion	0.36%
Deletions	112,116
Mapped reads with at least one deletion	0.31%
Homopolymer indels	47.41%

4.2. Mapeo de secuencias. Herramientas de mapeo, visualización y análisis de calidad.

4.2.6. Análisis de la calidad del mapeo

Resumen. Información y estadísticas básicas de alineamiento, como son datos globales sobre el número de lecturas, número de lecturas mapeadas, eficacia del mapeo emparejado, distribución de la longitud de las lecturas, número de lecturas recortadas, tasa de duplicación, contenido en nucleótidos y porcentaje GC, cobertura media y desviación típica de la profundidad, calidad media de los mapeos, tamaño medio de inserto, visión general de la tasa de error y estadísticas sobre el cromosoma (número de bases mapeadas, media y desviación estándar para cada uno de los cromosomas).

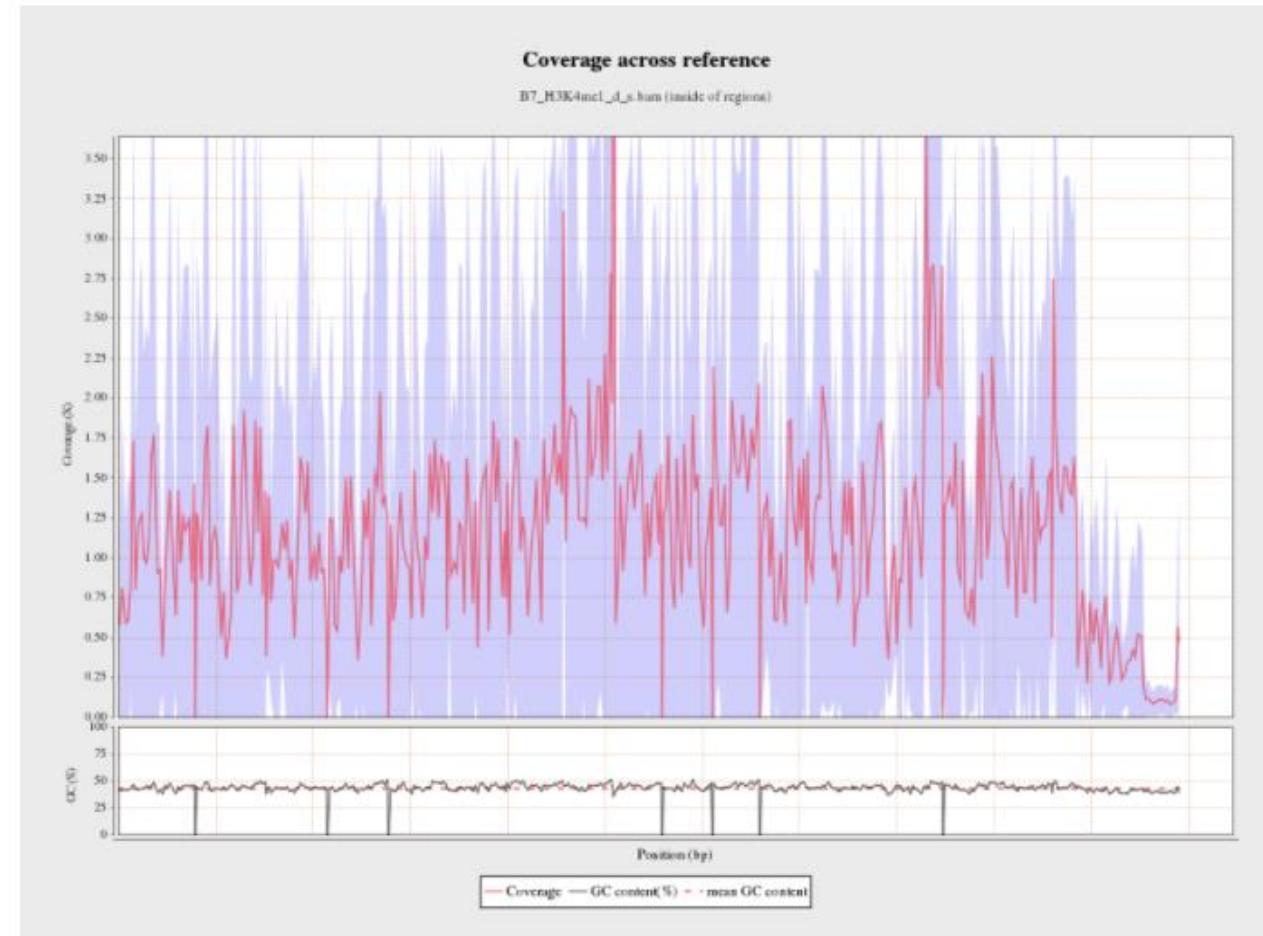
Chromosome stats (inside of regions)

Name	Length	Mapped bases	Mean coverage	Standard deviation
1	87833224	98754330	1.1243	1.7019
2	91729544	107134369	1.1679	1.8148
3	63799692	68389698	1.0719	1.6198
4	70909143	82853719	1.1684	1.7229
5	74314364	90547793	1.2184	1.7837
6	70481271	84576352	1.2	1.7784
7	72394548	94934128	1.3113	1.8448
8	62305007	112509915	1.8058	2.4018
9	65205555	87619480	1.3437	2.5807
10	64472795	79582346	1.2344	1.7795
11	63004071	101016969	1.6033	2.1472
12	45140598	55943130	1.2393	1.7842
13	44652192	58058169	1.3002	1.7919
14	48751186	57181195	1.1729	1.8765
15	40767811	79485248	1.9497	2.6005
16	38505094	46992001	1.2204	1.7624
17	41899931	58486618	1.3959	1.9805
18	33695330	40316418	1.1965	1.7335
19	29501366	45733440	1.5502	2.0198
X	51516852	24029265	0.4664	0.9102
Y	7391595	1047127	0.1417	0.6023
MT	15365	7289	0.4744	0.7732

4.2.6. Análisis de la calidad del mapeo

Profundidad (o cobertura) a lo largo de la referencia. La gráfica muestra una distribución de la profundidad, así como la desviación sobre la referencia. También muestra el porcentaje GC a lo largo de la referencia.

Coverage across reference



4.2.6. Análisis de la calidad del mapeo

Profundidad (o cobertura) a lo largo de la referencia. La gráfica muestra una distribución de la profundidad, así como la desviación sobre la referencia. También muestra el porcentaje GC a lo largo de la referencia.

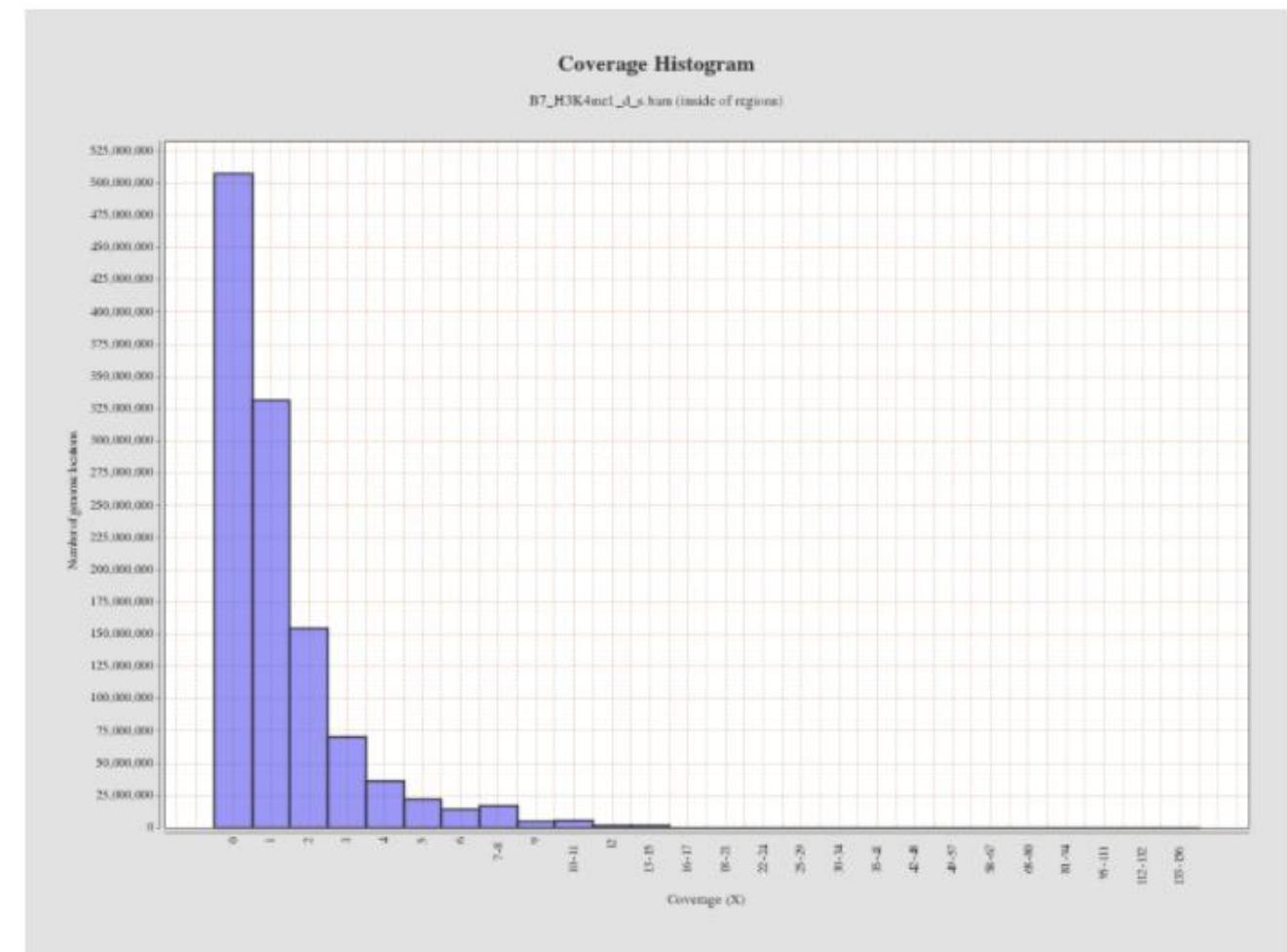
Coverage across reference



4.2.6. Análisis de la calidad del mapeo

Histograma de profundidad. Representa el número de regiones genómicas que tienen un determinado intervalo de profundidad.

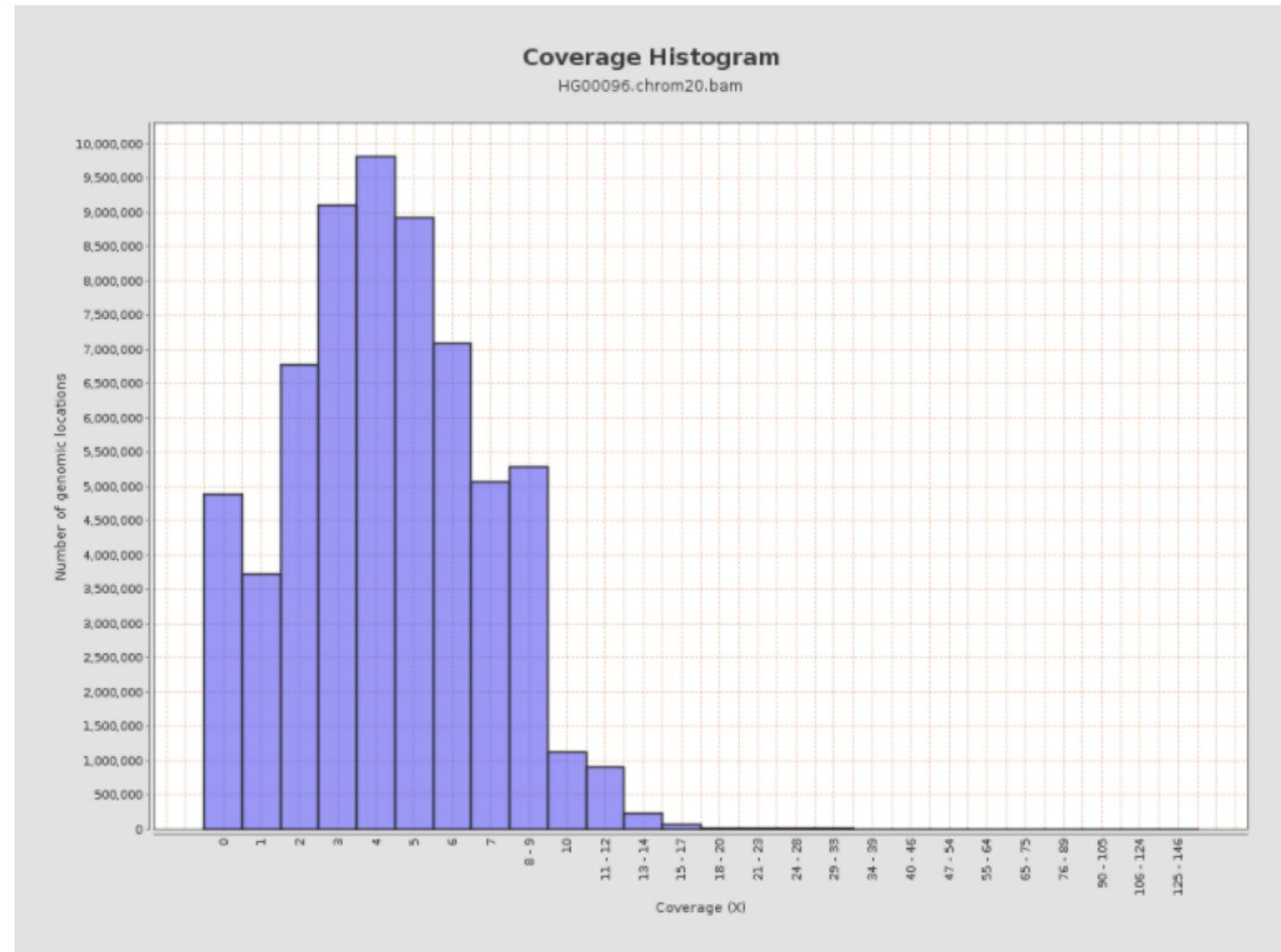
Coverage Histogram



4.2.6. Análisis de la calidad del mapeo

Coverage Histogram

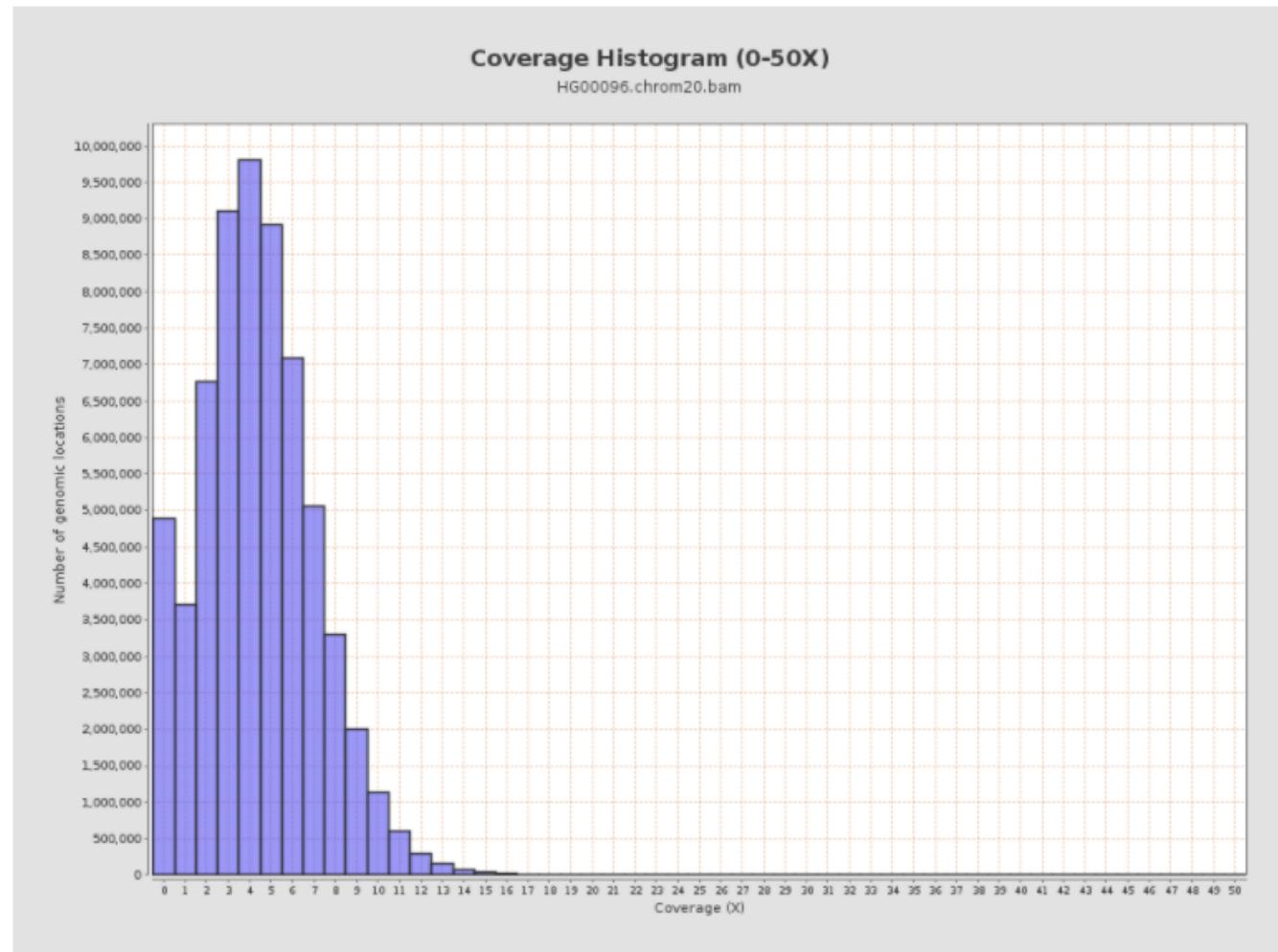
Histograma de profundidad. Representa el número de regiones genómicas que tienen un determinado intervalo de profundidad.



4.2.6. Análisis de la calidad del mapeo

Coverage Histogram (0-50X)

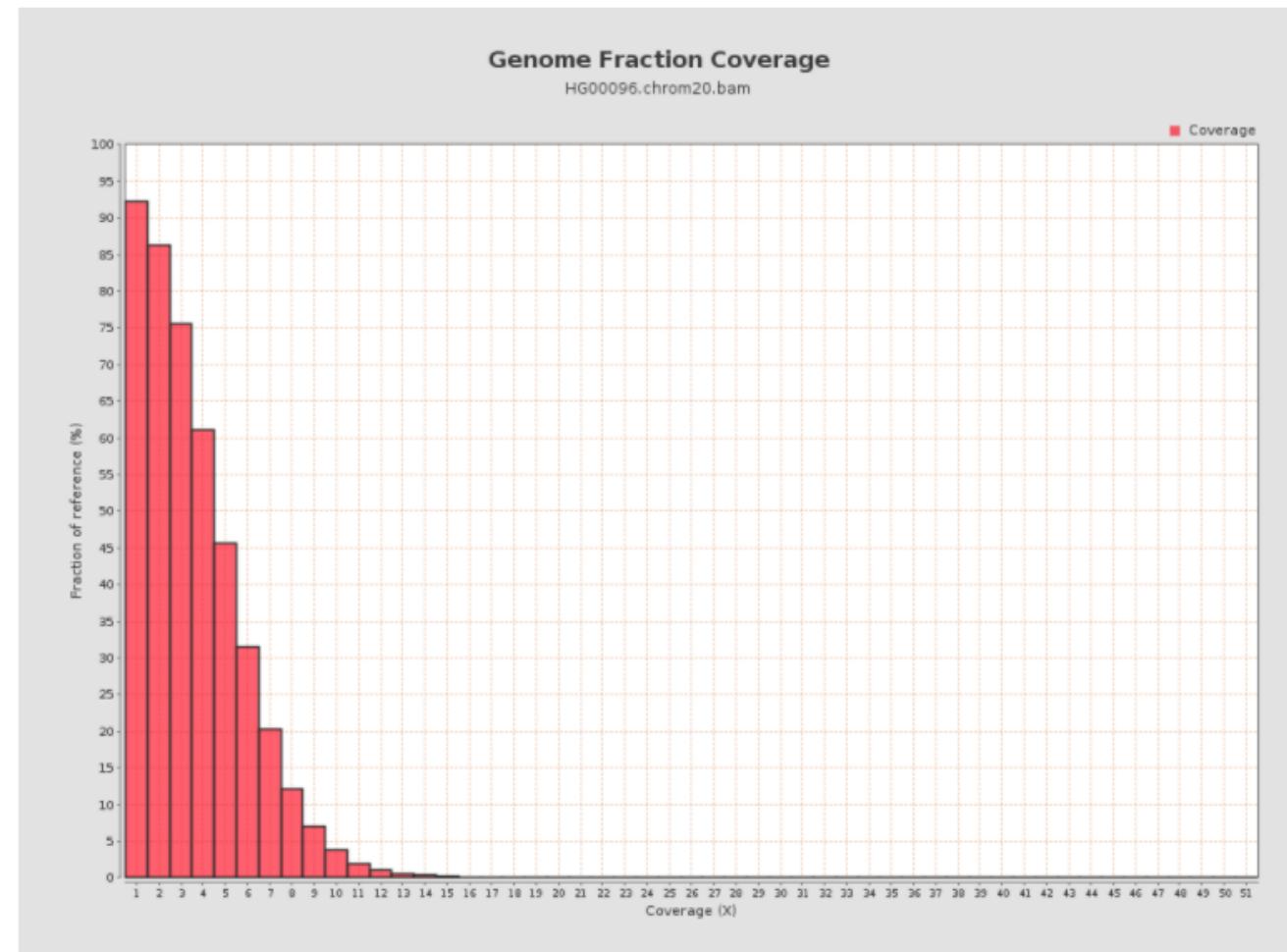
Histograma de profundidad 0-50x. Es un histograma similar al anterior, pero solo se representa la profundidad hasta 50x.



4.2.6. Análisis de la calidad del mapeo

Profundidad de la fracción del genoma. Muestra la fracción del genoma cubierto al menos a una determinada profundidad. Resulta útil para saber si la secuenciación ha cubierto nuestras expectativas.

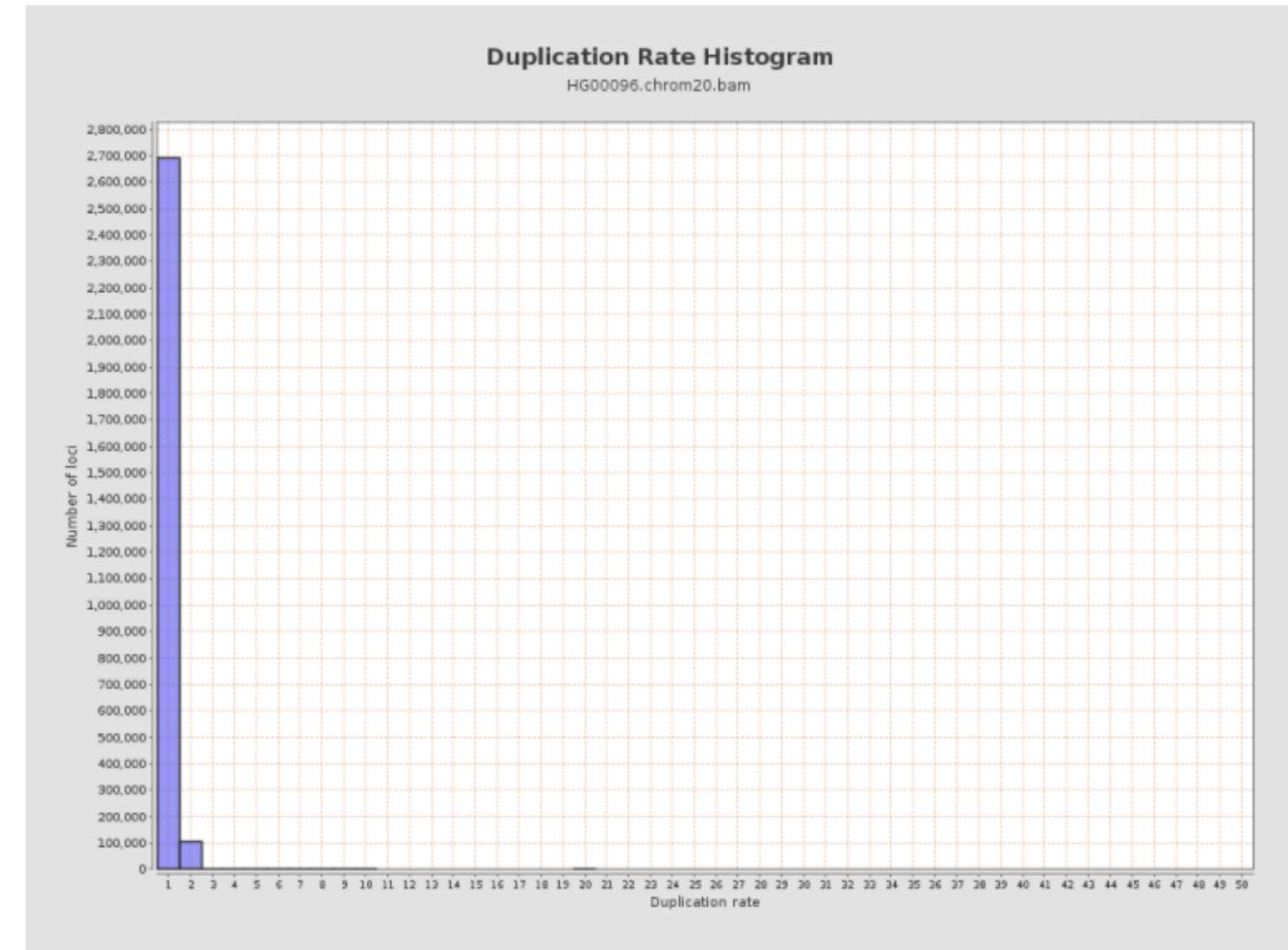
Genome Fraction Coverage



4.2.6. Análisis de la calidad del mapeo

Duplication Rate Histogram

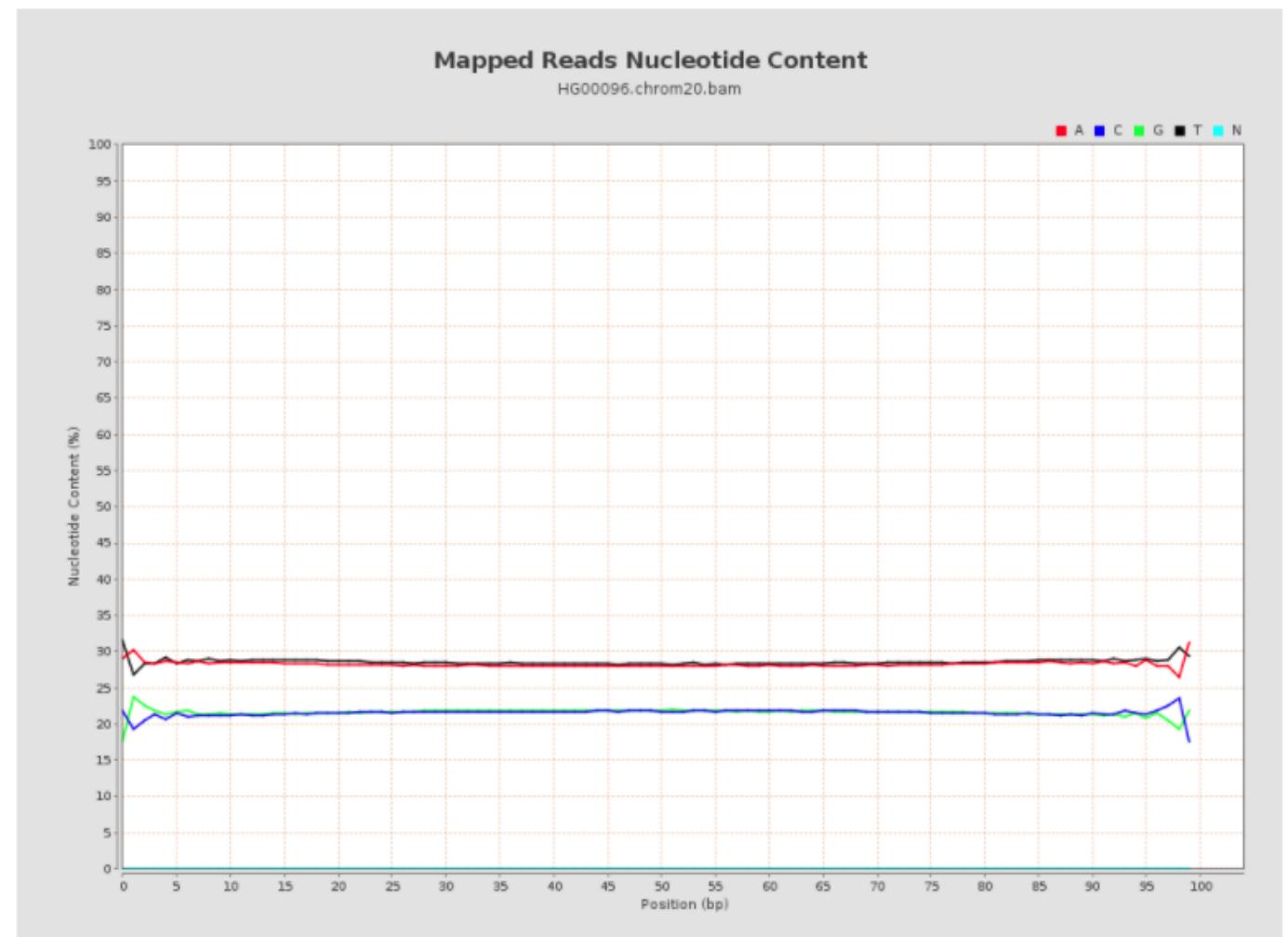
Histograma de la tasa de duplicación. Nos muestra la tasa de duplicación.



4.2.6. Análisis de la calidad del mapeo

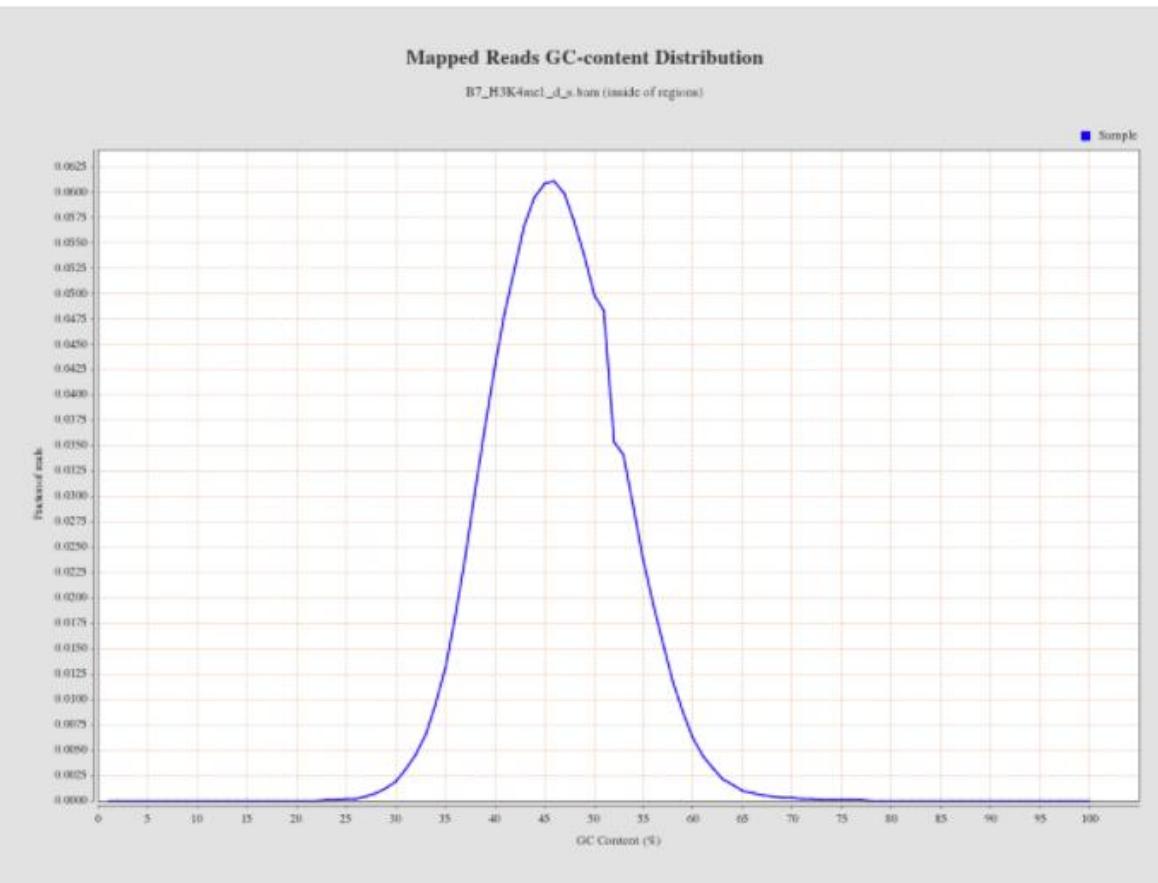
Mapped Reads Nucleotide Content

Contenido nucleotídico de las lecturas mapeadas.



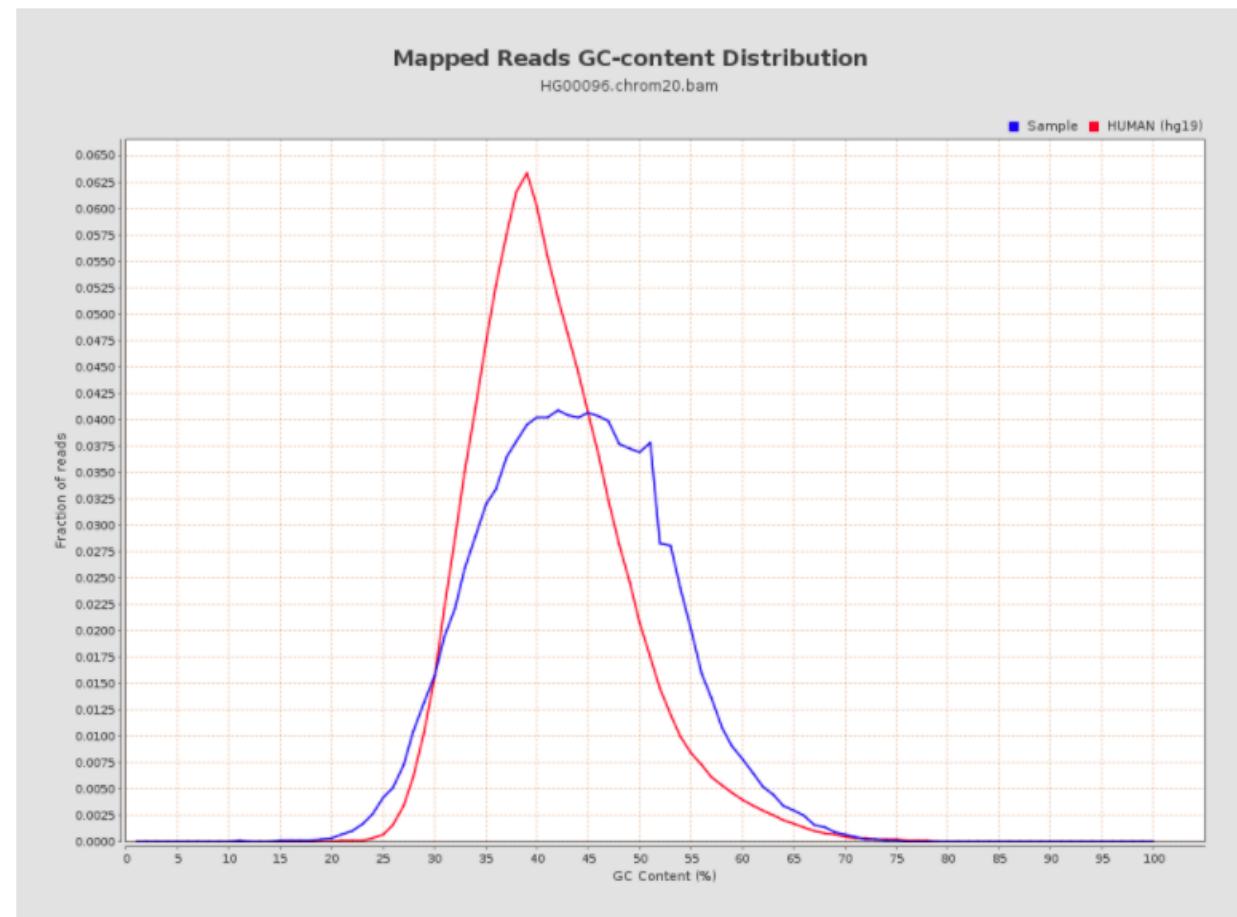
4.2.6. Análisis de la calidad del mapeo

Mapped Reads GC-content Distribution



Distribución GC en las lecturas mapeadas. Nos muestra el contenido GC solo de las lecturas mapeadas. Si se provee de un genoma de referencia, realizará la comparativa con el mismo.

Mapped Reads GC-content Distribution

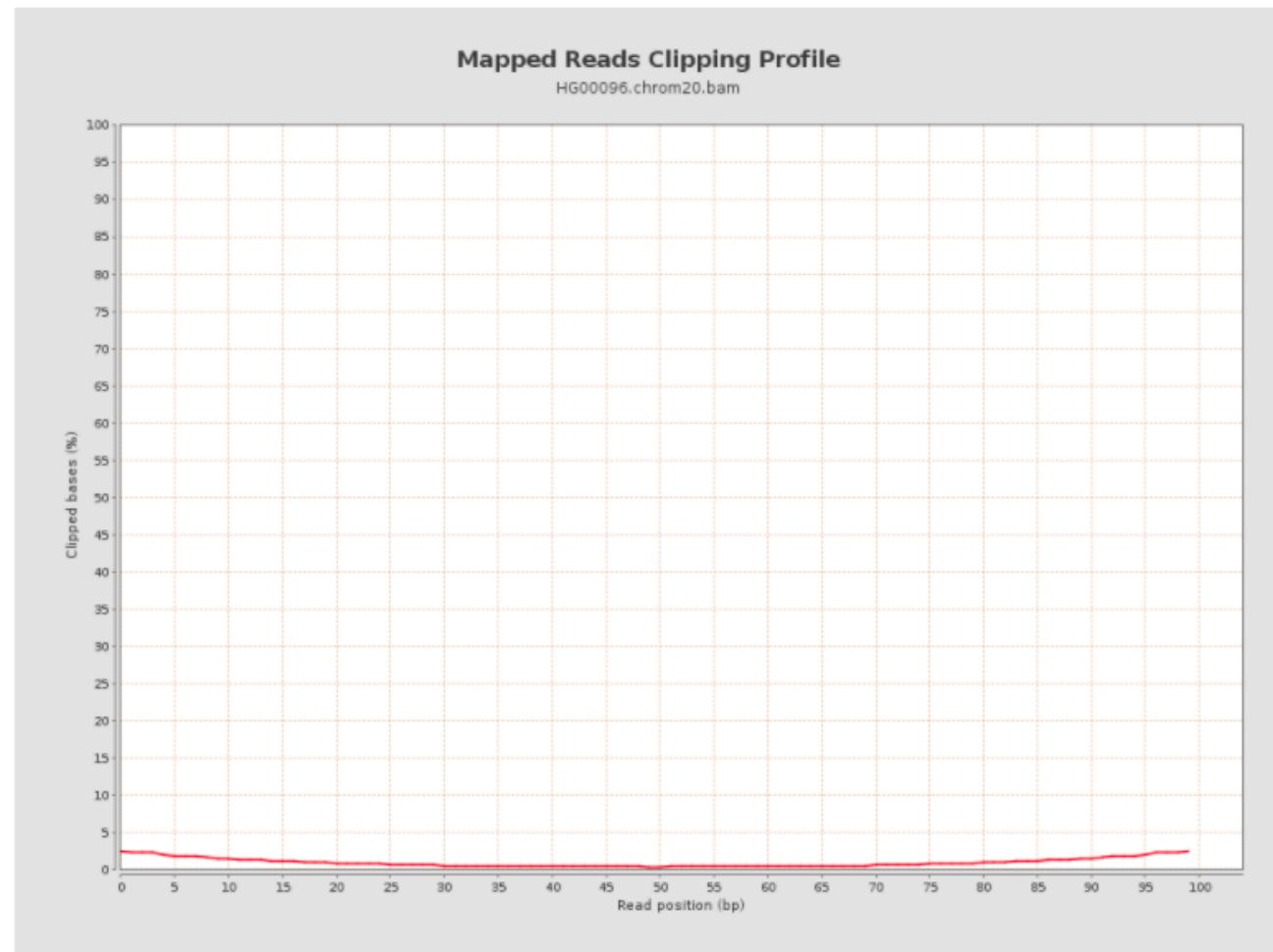


4.2.6. Análisis de la calidad del mapeo

Perfil de recorte de las lecturas mapeadas.

Nos muestra el porcentaje de bases recortadas a lo largo de las lecturas en el proceso de mapeo. El número total de lecturas recortadas aparece en el apartado de resumen.

Mapped Reads Clipping Profile

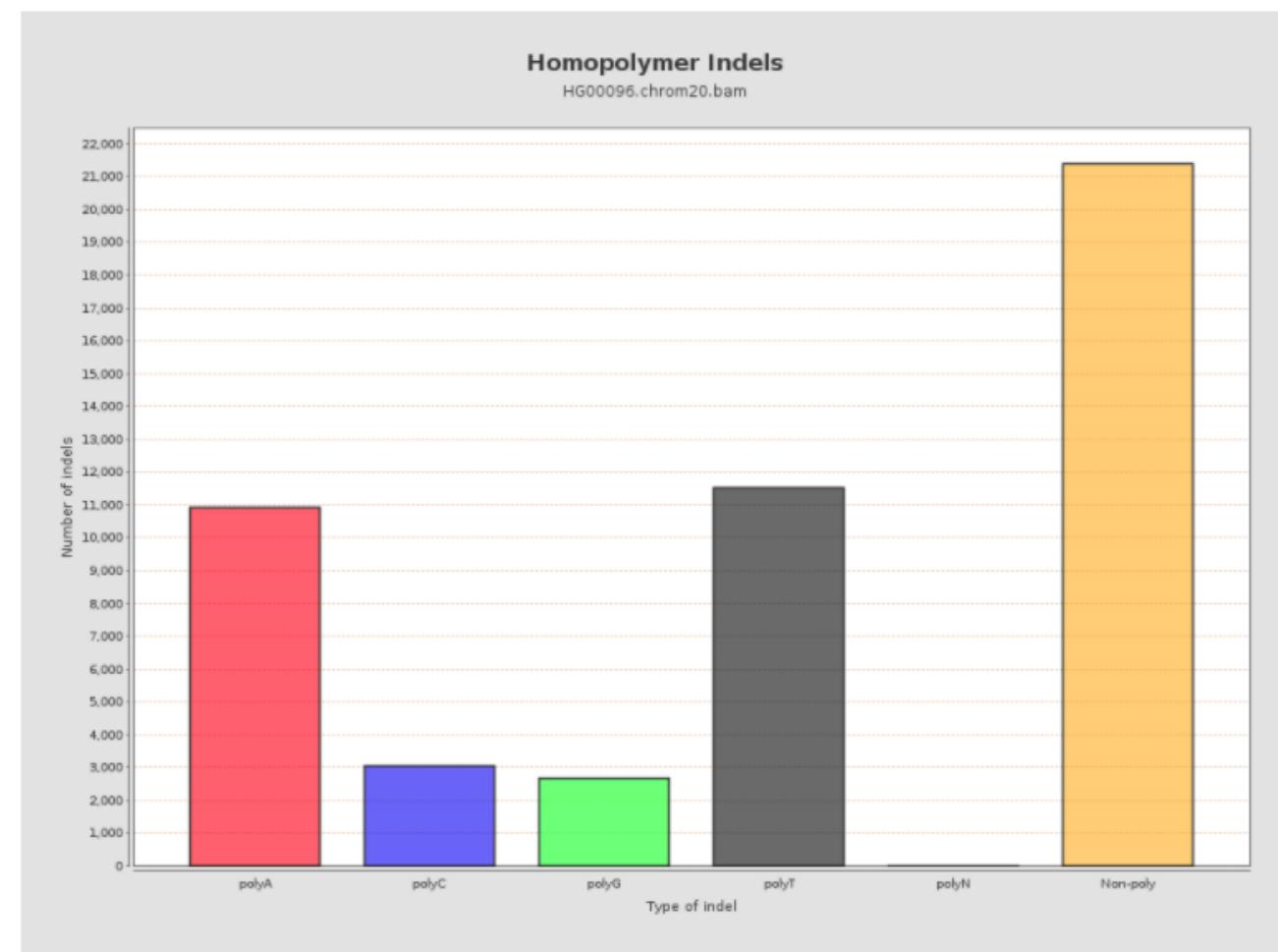


4.2.6. Análisis de la calidad del mapeo

Indels en homopolímeros.

Nos muestra el número de indels que están dentro de homopolímeros de cualquiera de los nucleótidos. Un número elevado de homopolímeros indica un problema de secuenciación.

Homopolymer Indels

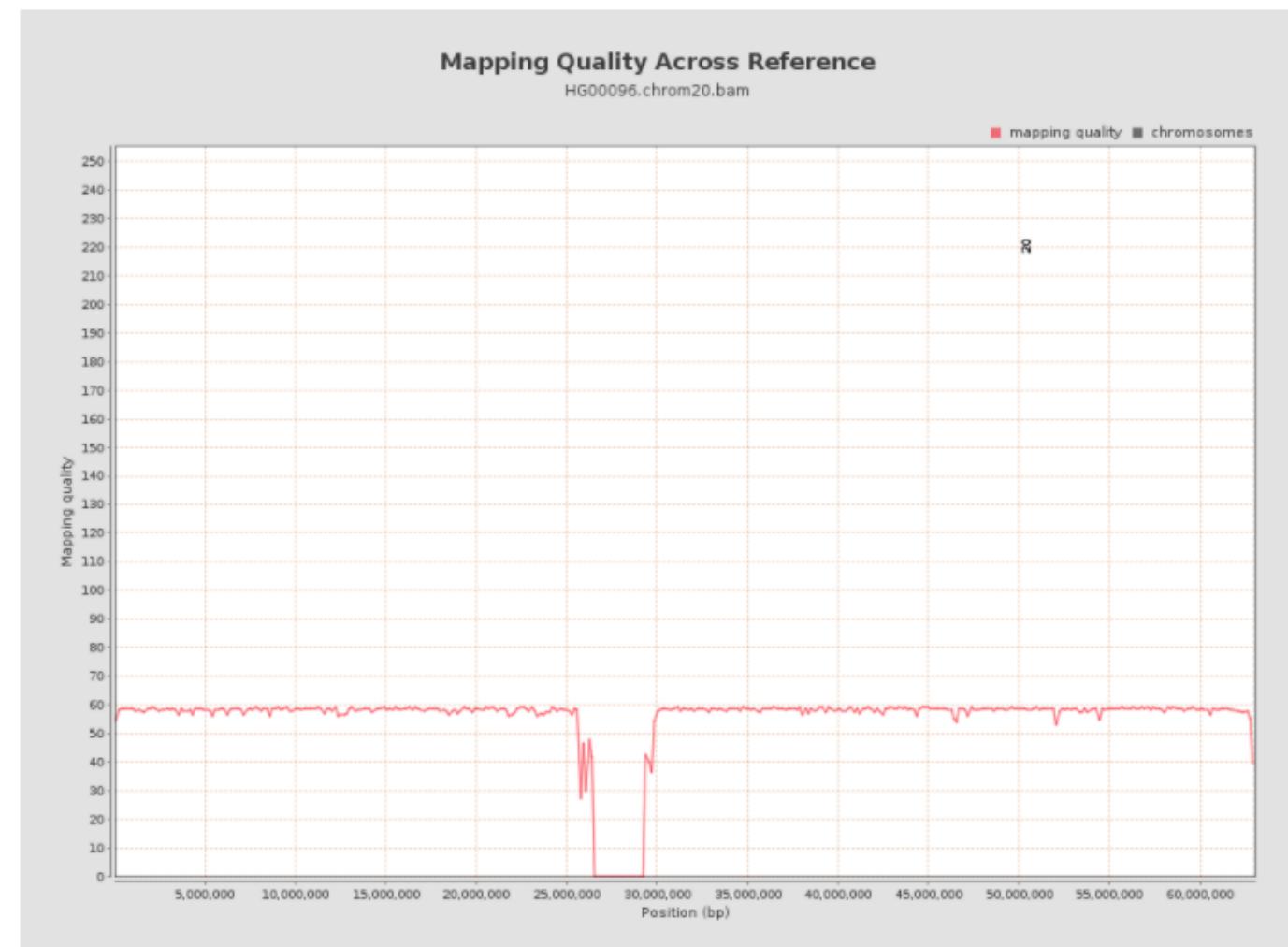


4.2.6. Análisis de la calidad del mapeo

Mapping Quality Across Reference

Calidad del mapeo a lo largo de la referencia.

Distribución de la calidad a lo largo de la referencia analizada.

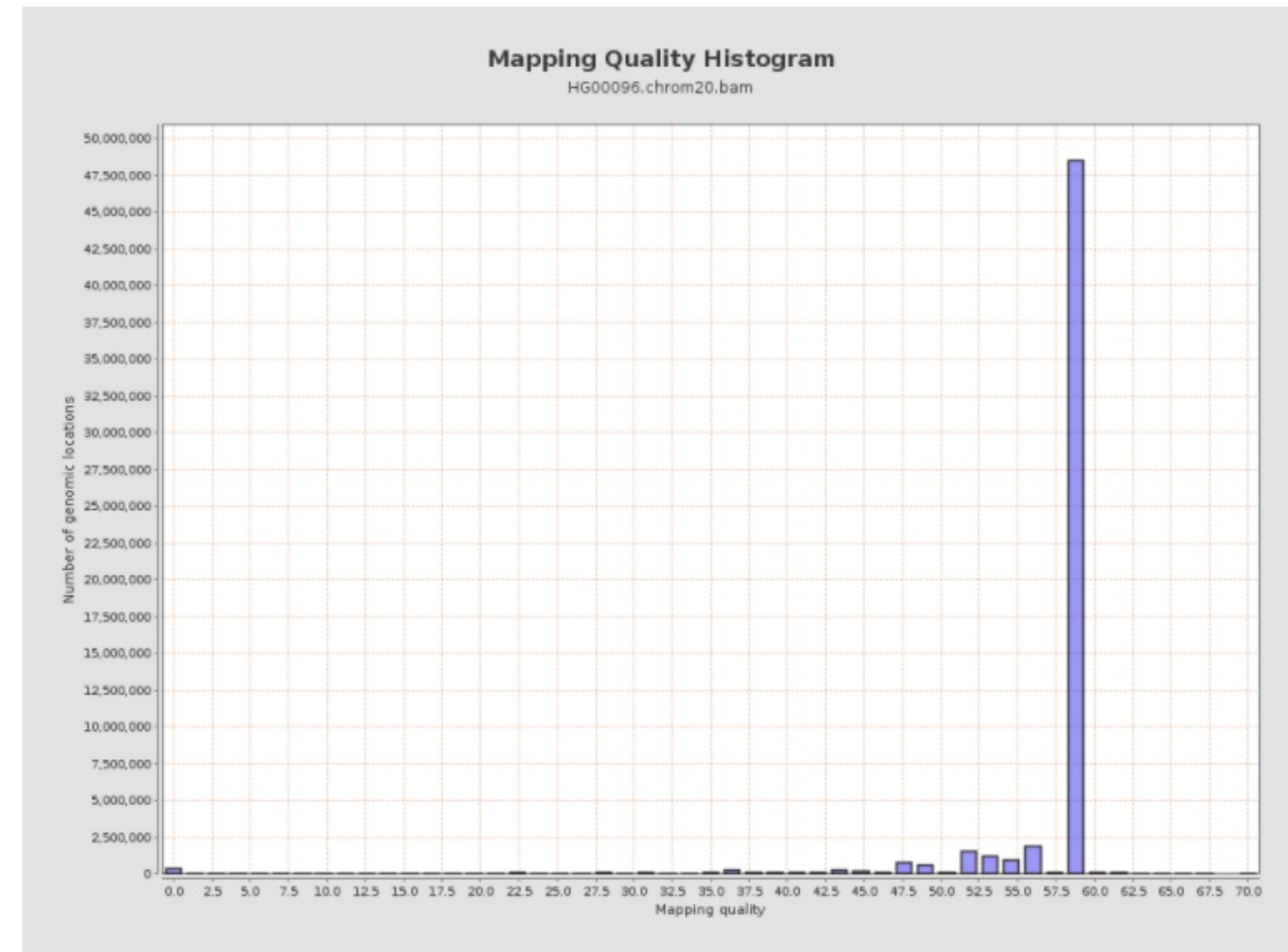


4.2.6. Análisis de la calidad del mapeo

Mapping Quality Histogram

Histograma de la calidad del mapeo.

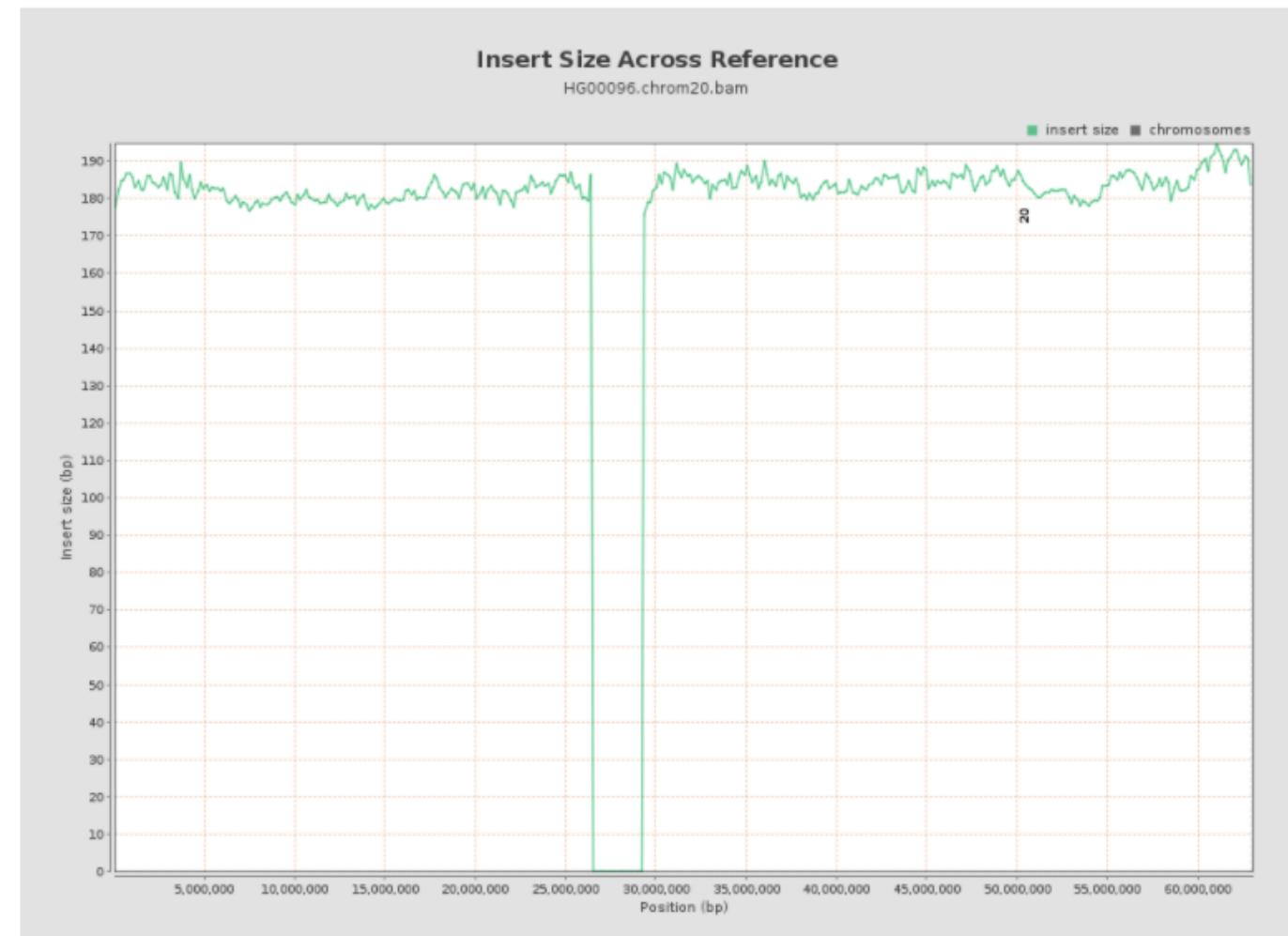
Histograma del número de localizaciones genómica con una calidad de mapeo dada.



4.2.6. Análisis de la calidad del mapeo

Tamaño de los insertos a lo largo de la referencia.

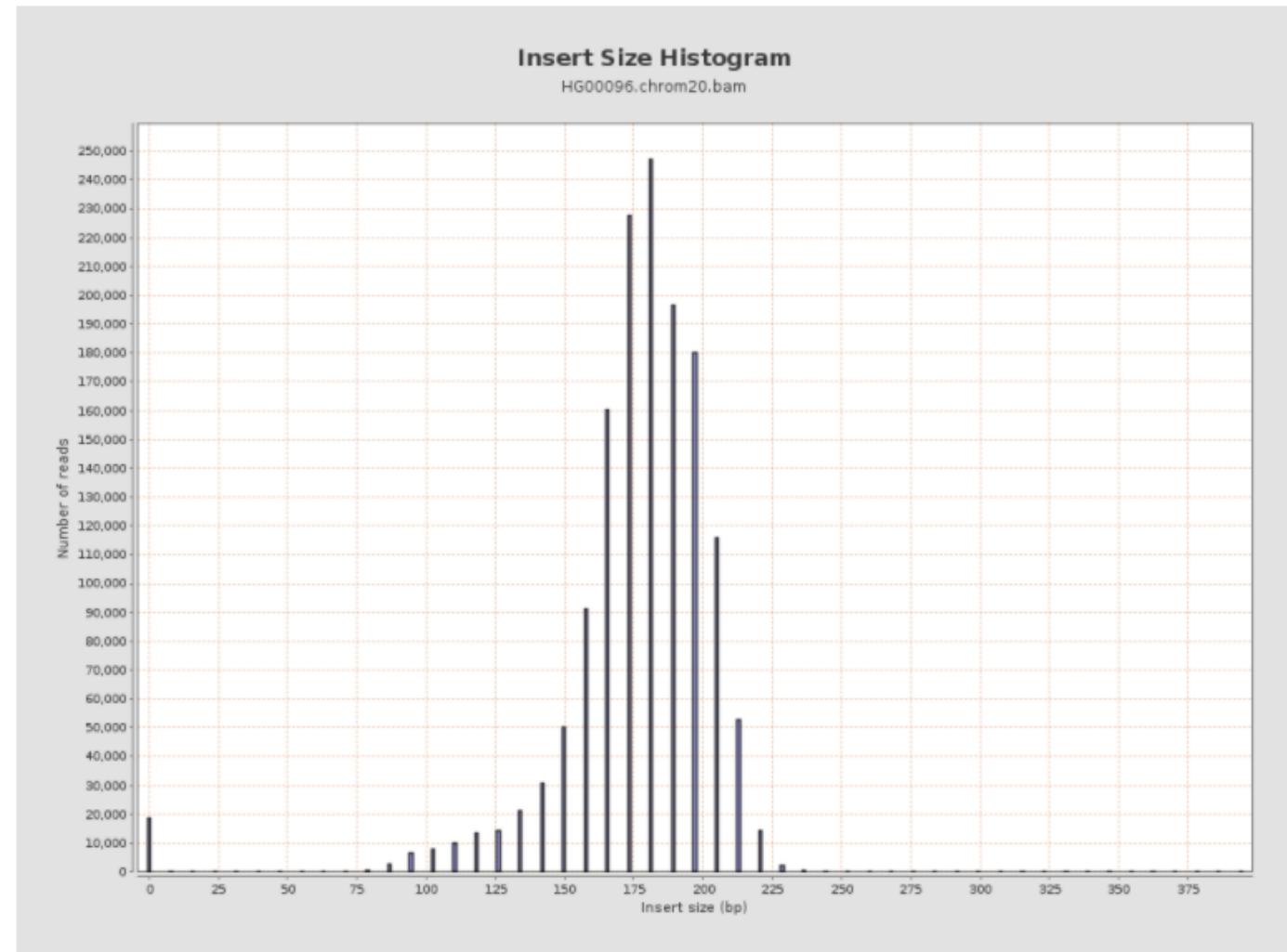
Insert Size Across Reference



4.2.6. Análisis de la calidad del mapeo

Insert Size Histogram

Histograma del tamaño de insertos.



Tema 4 – Ejemplo 6



Tema 4 - Ejemplo 6 - Calidad del Mapeo

Además de la visualización, llevada a cabo en el Ejemplo 5, podemos realizar el análisis de la calidad del mapeo con el software Qualimap.

Qualimap puede utilizarse en modo interactivo, al que se accede simplemente invocando al programa desde la terminal, o bien, por vía de comandos, que es la mejor forma si queremos automatizar y escalar el proceso.

El comando general a utilizar sería:

```
qualimap bamqc --bam mapping.bam --c --nt 2 --gd HUMAN --gff genome.gff --java-mem-size=32G
```

En nuestro caso, al haber mapeado sobre un cromosoma únicamente, prescindiremos del archivo de anotación GFF, y además nuestra máquina es más humilde, así que nuestro comando sería así:

```
qualimap bamqc --bam chr7.sorted.bam --c --nt 2 --gd HUMAN --java-mem-size=4G
```

Resultados: [Tema4_Ejemplo6_Qualimap](#)

Qualimap Report: BAM QC

QualiMap

Input data and parameters

QualiMap command line

```
qualimap bamqc -bam chr7.sorted.bam -c -nw 400 -hm 3
```

Alignment

Command line:	bwa mem -a Homo_sapiens.GRCh37.dna.chromosome.7.fa FastP_result/out1.clean.fq.gz FastP_result/out2.clean.fq.gz -o chr7.sam
Draw chromosome limits:	yes
Analyze overlapping paired-end reads:	no
Program:	bwa (0.7.17-r1188)
Analysis date:	Wed Dec 28 20:55:28 CET 2022
Size of a homopolymer:	3
Skip duplicate alignments:	no
Number of windows:	400
BAM file:	chr7.sorted.bam

CONTENTS

- Input data & parameters
- Summary
- Coverage across reference
- Coverage Histogram
- Coverage Histogram (0-50X)
- Genome Fraction Coverage
- Duplication Rate Histogram
- Mapped Reads Nucleotide Content
- Mapped Reads GC-content Distribution
- Mapped Reads Clipping Profile
- Homopolymer Indels
- Mapping Quality Across Reference
- Mapping Quality Histogram
- Insert Size Across Reference
- Insert Size Histogram

Summary

Globals

Reference size	159,138,663
Number of reads	1,639,750
Mapped reads	558,642 / 34.07%
Unmapped reads	1,081,108 / 65.93%
Mapped paired reads	558,642 / 34.07%
Mapped reads, first in pair	279,410 / 17.04%
Mapped reads, second in pair	279,232 / 17.03%
Mapped reads, both in pair	542,134 / 33.06%
Mapped reads, singletons	16,508 / 1.01%
Secondary alignments	1,560,932
Supplementary alignments	18,790 / 1.15%
Read min/max/mean length	0 / 151 / 134.99
Duplicated reads (estimated)	428,180 / 26.11%
Duplication rate	57.71%
Clipped reads	428,323 / 26.12%

ACGT Content

Number/percentage of A's	9,722,027 / 28.04%
Number/percentage of C's	7,632,250 / 22.01%
Number/percentage of T's	9,704,002 / 27.99%
Number/percentage of G's	7,610,433 / 21.95%
Number/percentage of N's	0 / 0%
GC Percentage	43.97%

Coverage

Mean	0.2182
Standard Deviation	7.253

Mapping Quality

Mean Mapping Quality	25.42
----------------------	-------

Insert size

Mean	3,259,137.8
Standard Deviation	15,636,530.66
P25/Median/P75	19 / 21 / 111

Mismatches and indels

General error rate	2.27%
Mismatches	750,859
Insertions	19,123
Mapped reads with at least one insertion	3.18%
Deletions	21,944
Mapped reads with at least one deletion	3.55%
Homopolymer indels	34.28%

Chromosome stats

Name	Length	Mapped bases	Mean coverage	Standard deviation
7	159138663	34720568	0.2182	7.253

4.3

Identificación de variantes

4.3. Identificación de variantes.

4.3.1. Preprocesamiento: Identificación de secuencias duplicadas.

4.3.2. Identificación de variantes.

4.3.3. Post-procesado: Filtrado de variantes y etiquetado

4.3. Identificación de variantes

Como ya se explicó en el Capítulo 1 de este documento, la variación dentro del genoma humano es un proceso habitual y normal. Estas variaciones incluyen la alteración en el número de cromosomas de la célula o la alteración éstos o de los genes.

A lo largo de esta sección se desgranará un protocolo bioinformático base para analizar las mutaciones germinales. El análisis de mutaciones somáticas sería similar, variando alguno de los pasos de la llamada de variantes, para tener en cuenta el mosaicismo genético.

La identificación y anotación de variantes es el proceso que comprende la extracción de los sitios variables (inserciones, delecciones, cambios nucleotídicos) o diferencias respecto al genoma de referencia a partir del alineamiento de las secuencias obtenido en pasos anteriores.

4.3. Identificación de variantes

El proceso de identificación de variantes no es sencillo, ya que existen numerosos factores que complican la identificación de éstas, y que pueden hacernos ver variantes (SNVs, indels, CNVs) cuando realmente no las hay.

Entre los factores que nos conducen a error están:

- **Sesgos en el proceso de amplificación** cuando se preparan **las genotecas** en el laboratorio.
- **Errores de secuenciación durante el proceso de lectura** que realiza la máquina. Para paliar este efecto, cada base nucleotídica leída lleva asociado un índice de calidad que nos da la fiabilidad determinada de que ese nucleótido sea real.
- **Cuestiones relacionadas con el software de alineamiento o mapeo**, especialmente en aquellas regiones altamente repetidas. Cada software calcula una probabilidad de que esa lectura esté asignada a esa posición.

Por ello, los programas de identificación de variantes deben tratar de distinguir entre variaciones reales y artefactos debidos a estos errores generados durante el proceso de la generar la genoteca, secuenciación y análisis.

4.3. Identificación de variantes

Por otra parte, **Freebayes** es un detector de variantes bayesiano desarrollado para la detección de pequeños polimorfismos, como SNPs, indels, polimorfismos múltiples de nucleótido y eventos complejos (como inserciones y sustituciones complejas) (Garrison & Marth, 2012).

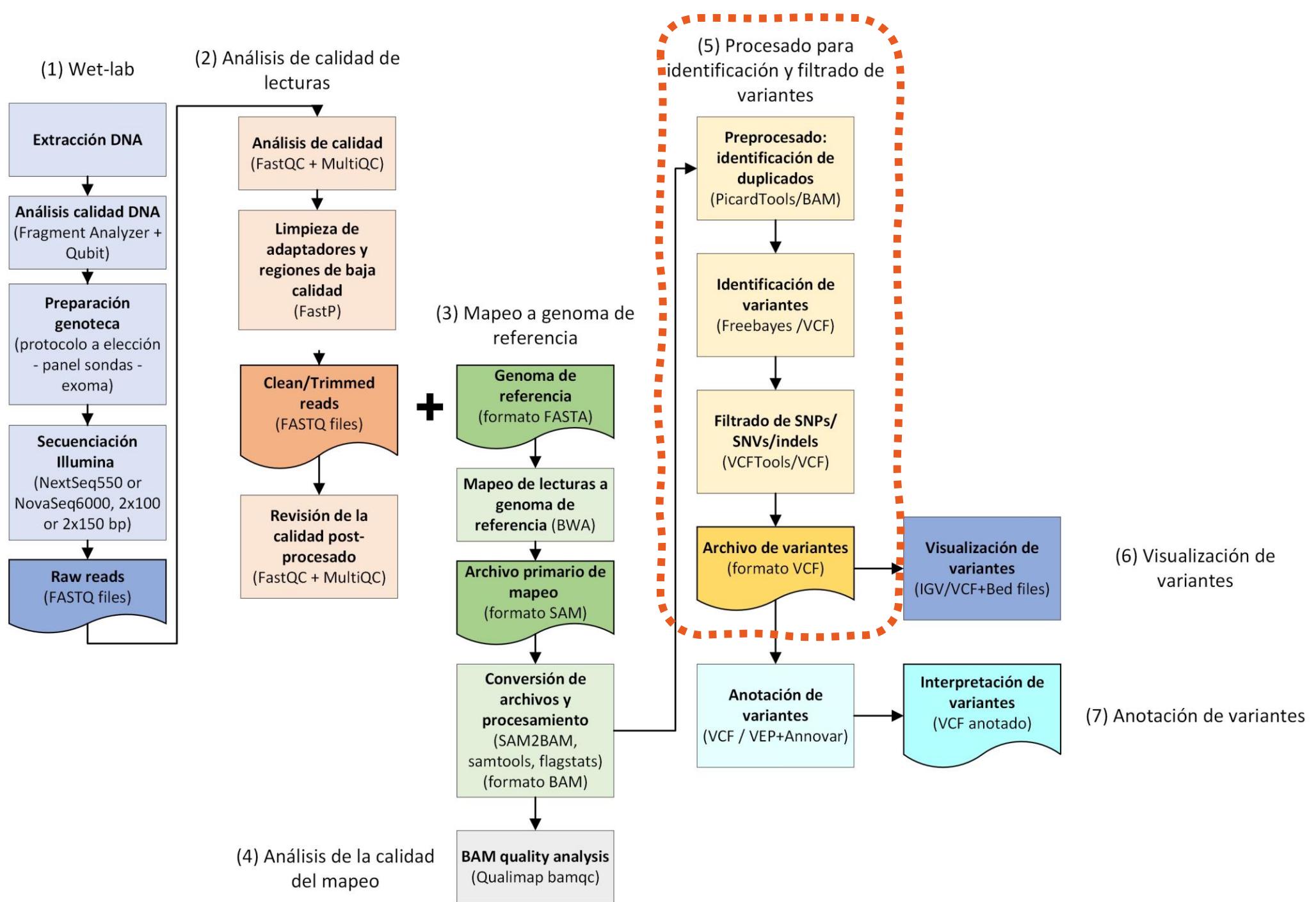
Se basa en haplotipos, ya que llama variantes basadas en las lecturas literales alineadas, no en su alineamiento exacto sobre el genoma de referencia. Su funcionamiento es más sencillo que otros programas como **GATK**. Este software presenta una mayor complejidad en su manejo, así que, para introducirnos, utilizaremos este en las prácticas.

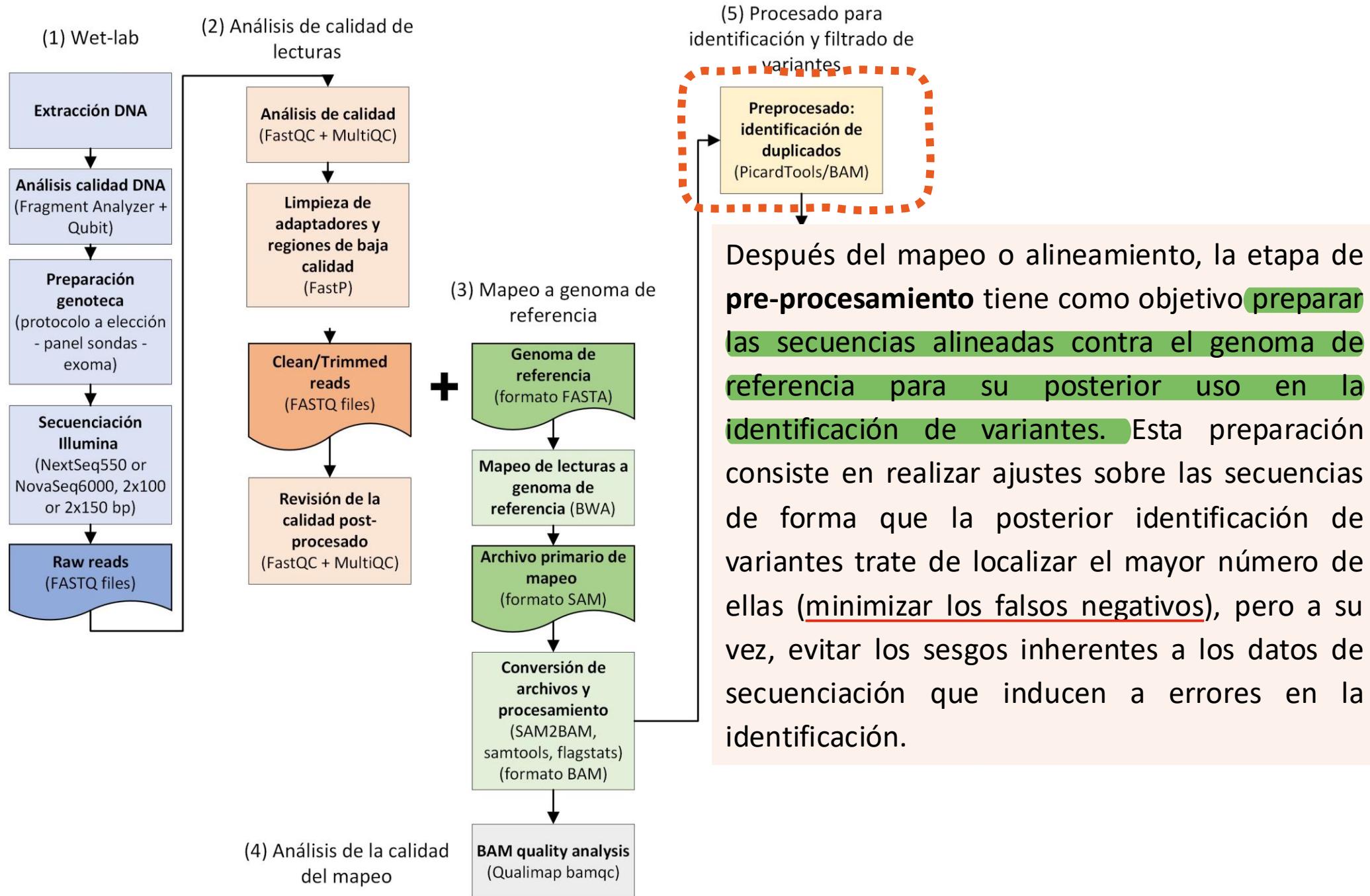
4.3. Identificación de variantes

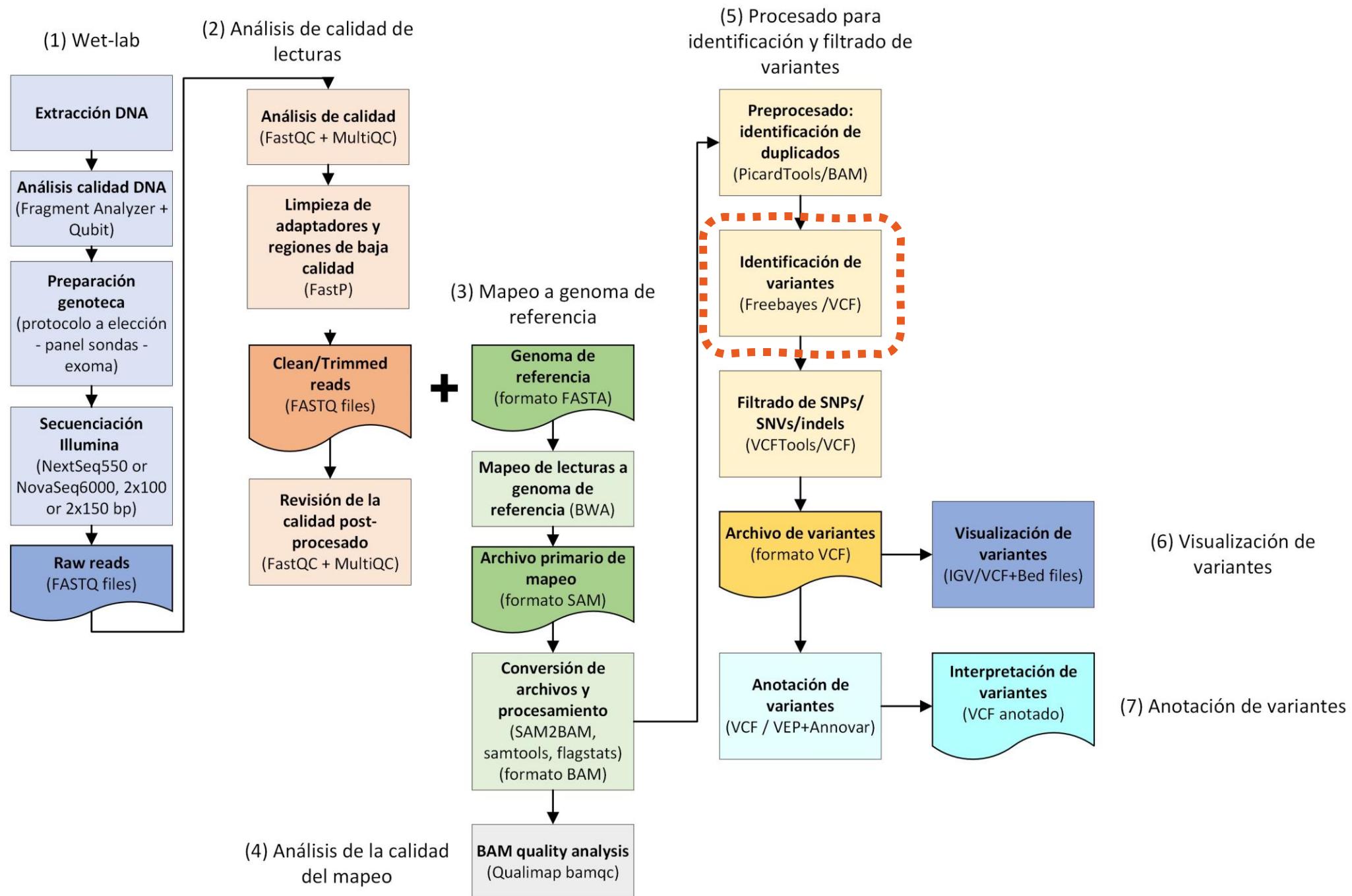
En los siguientes artículos se encuentran varias revisiones de software utilizados para identificación de variantes, donde se indica para qué tipo de variantes están indicados y la tecnología de secuenciación más apropiada.

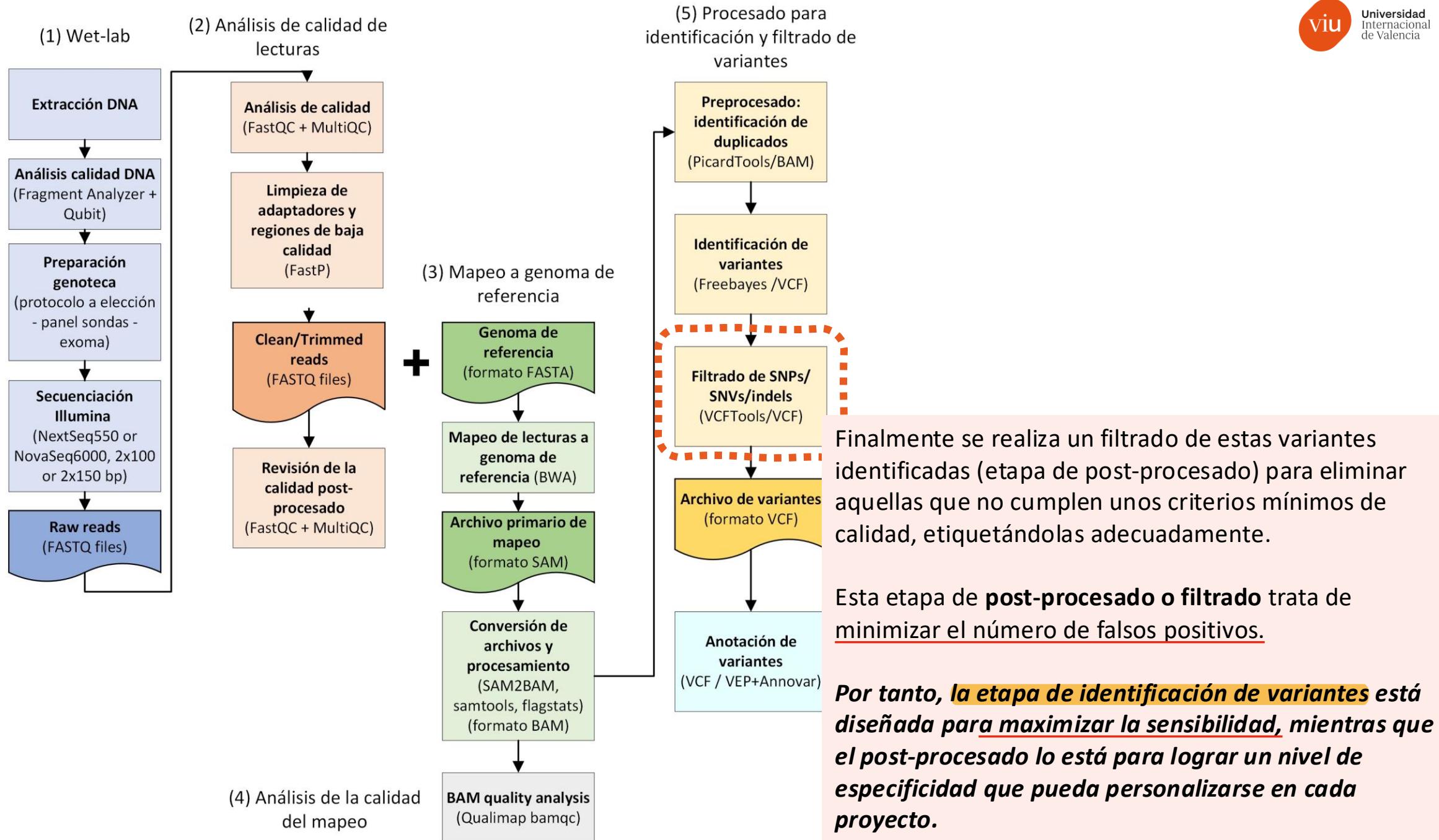
Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). **A survey of tools for variant analysis of next-generation genome sequencing data**. *Briefings in bioinformatics*, 15(2), 256–278. <https://doi.org/10.1093/bib/bbs086> [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3956068/]

Zhao, S., Agafonov, O., Azab, A. et al. **Accuracy and efficiency of germline variant calling pipelines for human genome data**. *Sci Rep* 10, 20222 (2020). <https://doi.org/10.1038/s41598-020-77218-4> [https://www.nature.com/articles/s41598-020-77218-4]









4.3.1. Preprocesamiento: Identificación de secuencias duplicadas

Tras el alineamiento podemos encontrarnos **secuencias duplicadas**, es decir, cuya posición de inicio y final sobre el genoma de referencia es el mismo lugar.

Estas lecturas duplicadas **pueden deberse a artefactos en el paso de amplificación de PCR de la librería/genoteca, o a la lectura por parte del secuenciador de ese fragmento de ADN más de una vez.**

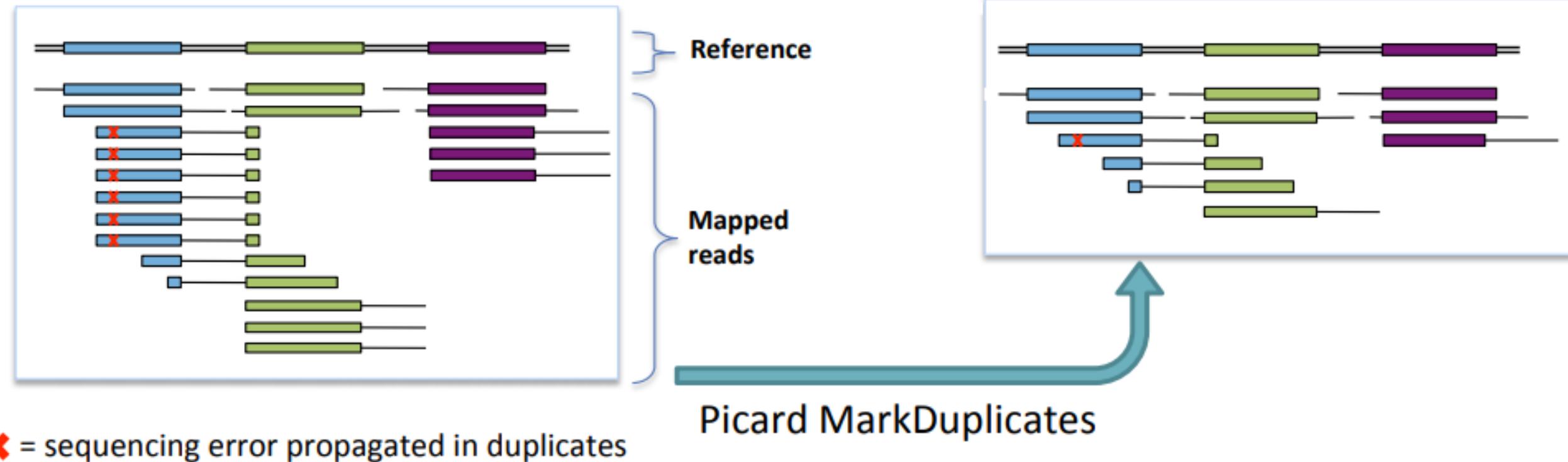
Esto puede suponer un problema cuando intentamos detectar variantes. Imaginemos que un error de secuenciación o de amplificación es propagado en distintas secuencias duplicadas, dando lugar a variantes falsas.

Por tanto, el primer paso que realizamos es identificar las lecturas cuyas coordenadas externas mapeen en la misma posición del genoma de referencia, y nos quedamos con una de ellas.

El proceso de eliminación de lecturas duplicadas se realiza con la herramienta **PicardTools** (<https://broadinstitute.github.io/picard/>), que identifica todas las lecturas que tienen la misma posición de inicio, final y orientación. **Para ello se fija en el valor CIGAR del archivo SAM o BAM. Marca todas las lecturas duplicadas excepto una, la que posee mayor calidad, y marca como duplicadas las demás.**

En el siguiente enlace se encuentra una descripción más detallada del valor CIGAR, presente en los archivos de alineamiento SAM o BAM.

<https://sites.google.com/site/bioinformaticsremarks/bioinfo/sam-bam-format/what-is-a-cigar>



Tema 4 – Ejemplo 7



Tema 4 - Ejemplo 7 - Marcaje duplicados



El siguiente paso a realizar es marcar las secuencias duplicadas que han mapeado sobre la referencia. De esta forma intentaremos "limpiar" el mapeo de posibles artefactos que den lugar a falsos positivos (falsas variantes). En este paso se localizan las lecturas cuyas coordenadas externas mapeen en la misma posición del genoma de referencia, y nos quedaremos tan sólo con una de ellas. Utilizaremos la herramienta PicardTools, que localizará, identificará y marcará las secuencias duplicadas, dejando como lectura válida la de mayor calidad.

Para este paso utilizaremos el archivo BAM ordenado (chr7.sorted.bam) obtenido en pasos anteriores:

1) Marcamos duplicados (1.5 min)

```
picard MarkDuplicates --INPUT chr7.sorted.bam --OUTPUT chr7.dedup.bam --METRICS_FILE markDuplicatesMetrics.txt --ASSUME_SORTED True
```

2) Añadir Read Groups (addReadGroups, dado que no los incluimos en el mapeo inicial, debemos hacerlo ahora, ya que si no nos da problemas en el siguiente paso (1 - 1.5 min aprox)

```
#RGID = Read Group ID : proviene del nombre de la lectura, nos indica el nombre del secuenciador (se revisa en los archivos FASTQ)
```

```
#GPL = Read group platform, equipo de secuenciación (ILLUMINA o SOLID)
```

```
#GPU = Read Group platform unit (run barcode); por defecto le ponemos unit1
```

```
#GSM = Read Group sample name, es el nombre de la muestra. Importante!!
```

```
picard AddOrReplaceReadGroups -I chr7.dedup.bam -O chr7.dedup.RG.bam -RGID M02899 -RGLB lib1 -GPL ILLUMINA -GPU unit1 -GSM S8
```

3) Indexamos bam nuevo (10-30 sg)

```
samtools index chr7.dedup.RG.bam
```

Resultados: [Tema4_Ejemplo7_MarcarDuplicados](#)

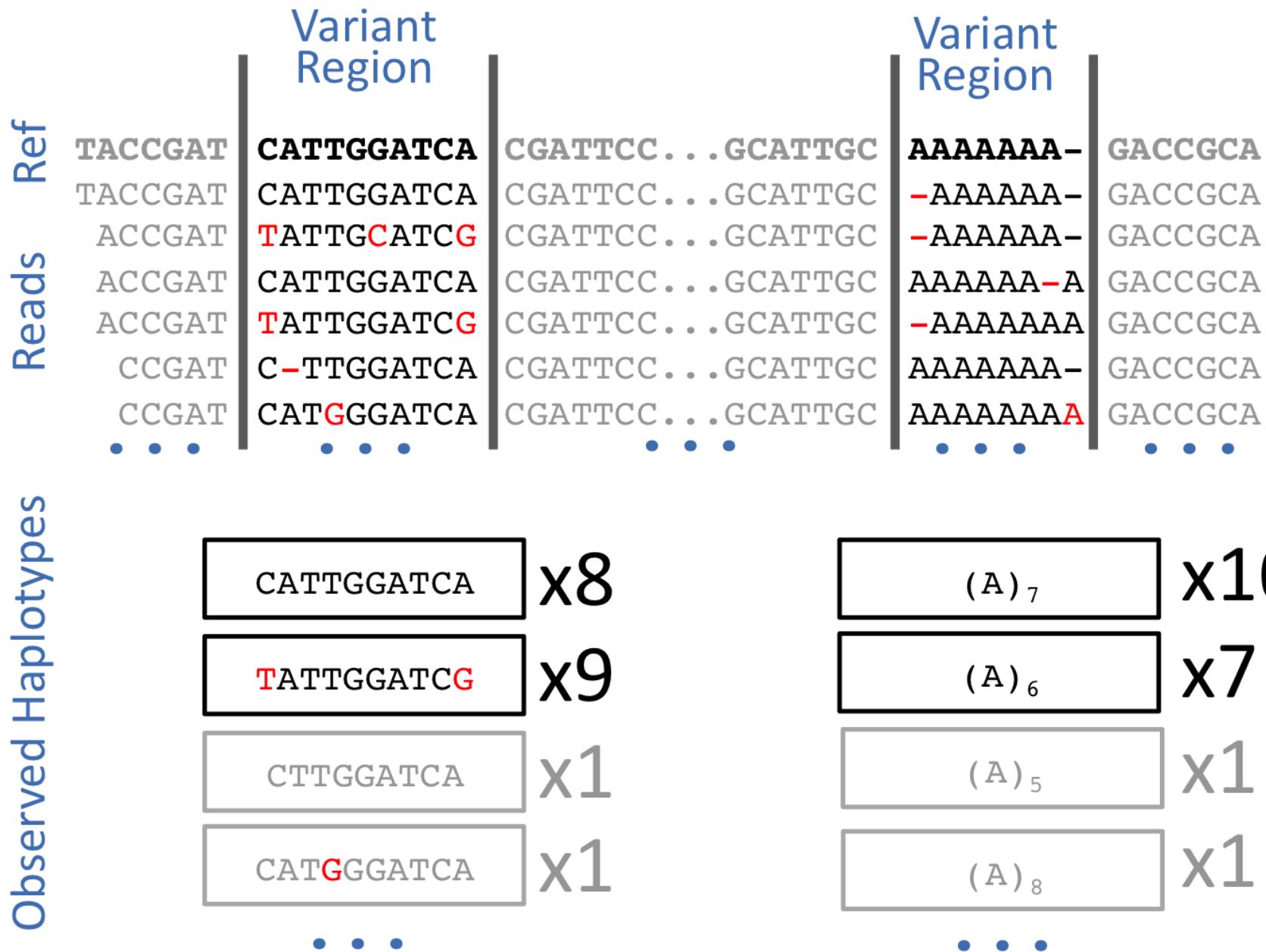
4.3.2. Identificación de variantes.

En este punto tenemos un archivo BAM de alineamiento donde hemos introducido una serie de mejoras para evitar sesgos producidos durante pasos anteriores, y con el objetivo de identificar todas las variantes potenciales de la manera más precisa posible, minimizando el número de falsos positivos.

En este punto identificaremos las **variantes**, determinando en qué posición al menos una base difiere con la referencia, e identificaremos el **genotipo** para el individuo en esas posiciones modificadas.

Recontaremos para cada posición modificada el número de ocurrencias de cada nucleótido teniendo en cuenta todas las lecturas alineadas en esa posición.

Para cada gen podemos encontrar versiones alternativas que se diferencian en la secuencia, es lo que denominamos alelo. Por otra parte, la combinación particular de alelos para un locus es el genotipo. Se denomina genotipo homocigoto si los alelos que lo componen son exactamente iguales o heterocigoto si los alelos difieren.



4.3.2. Identificación de variantes.

Para la detección de las distintas variantes o alelos que podemos encontrar en un locus, deben tenerse en cuenta los sesgos intrínsecos a los datos de secuenciación y análisis bioinformático. Por ello, se siguen métodos probabilísticos que incorporan estas incertidumbres en la detección de variantes, como es el software GATK o FreeBayes, donde se tienen en cuenta:

- La calidad de todas las bases (valor Phred) en la posición objeto de estudio, teniendo en cuenta todas las lecturas que apoyan esa posición.
- Proximidad a una región de inserciones, delecciones o regiones homopolímeras.
- Calidad del mapeo de las lecturas que apoyan la variante. Valores bajos de calidad pueden indicar región repetida.
- Longitud de las secuencias que soportan la variante. Secuencias cortas tienen más posibilidad de alinear en múltiples localizaciones del genoma.
- Profundidad de la secuenciación. Cuando una variante está soportada por un número muy elevado de secuencias, esto apoya que sea real.

Tema 4 – Ejemplo 8



Tema 4 - Ejemplo 8 - Variant Calling



El siguiente paso del protocolo es la llamada de variantes (Variant Calling) que realizaremos con el software FreeBayes.

Para realizar este apartado utilizaremos el archivo BAM deduplicado con los read groups del ejemplo anterior (chr7.dedup.RG.bam).

1) Llamada de variantes (aprox 5 min)

opciones:

-C = mínimo de lecturas que deben mapear en una posición para ser considerado un alelo alternativo

-f = genoma de referencia

```
freebayes -C 25 -f Homo_sapiens.GRCh37.dna.chromosome.7.fa chr7.dedup.RG.bam > chr7.vcf
```

El -C 25: requiere al menos 25 observaciones que soportan la evidencia del alelo alternativo, es decir, al menos 25 lecturas que apoyen que ahí hay una variante

2) Estadísticas sobre las variantes detectadas (10 sg)

```
rtg vcfstats chr7.vcf > chr7.vcfstats
```

3) Repite la llamada de variantes y las estadísticas buscando un número mínimo de lecturas de 150 para la llamada de variantes. ¿Cómo varía la detección?

Resultados: [Tema4_Ejemplo8_VariantCalling](#)

```

Location : chr7.vcf
Failed Filters : 0
Passed Filters : 515
SNPs : 396
MNP : 59
Insertions : 7
Deletions : 13
Indels : 6
Same as reference : 34
SNP Transitions/Transversions: 1.53 (343/224)
Total Het/Hom ratio : 1.32 (274/207)
SNP Het/Hom ratio : 1.34 (227/169)
MNP Het/Hom ratio : 1.03 (30/29)
Insertion Het/Hom ratio : 1.33 (4/3)
Deletion Het/Hom ratio : 3.33 (10/3)
Indel Het/Hom ratio : 1.00 (3/3)
Insertion/Deletion ratio : 0.54 (7/13)
Indel/SNP+MNP ratio : 0.06 (26/455)
  
```

```

Location : chr7_C150.vcf
Failed Filters : 0
Passed Filters : 69
SNPs : 59
MNP : 6
Insertions : 2
Deletions : 1
Indels : 0
Same as reference : 1
SNP Transitions/Transversions: 2.00 (44/22)
Total Het/Hom ratio : 8.71 (61/7)
SNP Het/Hom ratio : 8.83 (53/6)
MNP Het/Hom ratio : - (6/0)
Insertion Het/Hom ratio : 1.00 (1/1)
Deletion Het/Hom ratio : - (1/0)
Indel Het/Hom ratio : - (0/0)
Insertion/Deletion ratio : 2.00 (2/1)
Indel/SNP+MNP ratio : 0.05 (3/65)
  
```

4.3.2. Identificación de variantes. El formato VCF.

Una vez realizado el procesamiento bioinformático para la detección de variantes, donde FreeBayes toma el archivo BAM pre-procesado, el archivo resultante es un archivo en formato **Variant Call Format (VCF)**.

Éste tiene un formato de texto plano, tabular, donde están presentes todas las variantes que han sido encontradas en el genoma, con numerosa información sobre cada una de ellas. Éste es un archivo estándar que puede ser visualizado en programas como IGV, que ya hemos visto.

El formato VCF es bastante complejo, así que vamos a describir las características principales. Son ficheros de texto plano, pero pueden llegar a ser muy pesados. Es importante no abrir un archivo de este tipo en un procesador Word, puesto que elimina el formato del archivo, sino en editores de texto plano.

En el siguiente enlace podéis encontrar una descripción detallada del formato VCF en sus dos últimas versiones.

<https://samtools.github.io/hts-specs/VCFv4.2.pdf>

<https://samtools.github.io/hts-specs/VCFv4.3.pdf>

4.3.2. Identificación de variantes. El formato VCF.

Este fichero está formado por dos partes principales:

Cabecera: contiene información acerca del conjunto de datos y otro tipo de información, como organismo y versión del genoma de referencia, como las definiciones o descripciones de todas las anotaciones utilizadas.

Versión del fichero VCF: ##fileformat=VCFv4.2

Líneas FILTER, que nos indican el tipo de filtros que se han utilizado para filtrar las variantes detectadas que no cumplen los parámetros de calidad.
##FILTER=<ID=LowQual,Description="Low quality">

Líneas FORMAT e INFO. Definen las anotaciones contenidas en las columnas FORMAT e INFO de la sección de variantes identificadas.

Líneas de contigs y referencia. En estas líneas están los nombres de los contigs o cromosomas, sus longitudes y qué versión de referencia se utilizó en el fichero BAM de entrada.

4.3.2. Identificación de variantes. El formato VCF.

Apartado “Records” tras la cabecera, aparecen las variantes identificadas, donde tenemos una línea de cabecera y a continuación una línea por cada variante identificada.

Línea de cabecera, define las columnas de las siguientes líneas, que tienen un formato separado por tabuladores. Las primeras ocho columnas, hasta el campo INFO, inclusive, representan las propiedades observadas a nivel de variante identificada. La información específica de la muestra (genotipo) se muestra en la columna FORMAT (novena columna). Además, hay una columna adicional de información por cada muestra representada en el archivo VCF.

```
[#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 ...]
```

4.3.2. Identificación de variantes. El formato VCF.

Existen dos tipos de anotaciones en el apartado “Records”:

Anotaciones a nivel de variante. Se refiere a las siete primeras columnas, que son obligatorias en todo archivo VCF. Si el valor para ese dato no se conoce, se indica con un punto ('.'). De esta forma:

- CHROM: define el contig o cromosoma donde está la variante
- POS: define la posición sobre el genoma de referencia. Si esta variación es una delección, la posición corresponde a la base anterior a la delección.
- ID: identificador opcional de la variante, si hemos utilizado una base de datos de variantes como dbSNP.
- REF: alelo de referencia.
- ALT: alelo alternativo. (Nota: tanto REF como ALT se proporcionan respecto a la hebra sentido 5'-3' -forward-; adicionalmente, para las inserciones, el alelo ALT contiene la base insertada y su precedente; mientras que para delecciones, el alelo ALT es la base anterior a la delección.)

4.3.2. Identificación de variantes. El formato VCF.

- **QUAL:** probabilidad en escala Phred de que el polimorfismo REF/ALT exista.
- **FILTER:** este campo tiene el nombre o nombres de los filtros en los que la variante no ha pasado. Si ha cumplido con todos los filtros, indicará ‘PASS’. Si el valor FILTER es ‘.’, significa que no se ha aplicado ningún filtro a las variantes.
- **INFO:** este campo no es obligatorio y corresponde a diferentes anotaciones realizadas sobre la variante. Estas anotaciones nos proporcionan información variada de las muestras. La información descrita en este campo siempre está definida en la cabecera del archivo VCF (líneas ##INFO), lo que facilita su comprensión. Las anotaciones más frecuentes son: AC (número de alelos para el genotipo, para cada alelo ALT), AF (frecuencia alélica para cada alelo ALT), AN (Número total de alelos identificados), DP (profundidad combinada de todos los alelos). Para otras anotaciones adicionales, podéis consultar la guía de formato VCF indicada anteriormente en enlaces de interés.

Anotaciones a nivel de muestra. El resto de las columnas situadas a la derecha de la columna INFO nos muestran información a nivel de muestra. Como mínimo esta información indica el genotipo inferido en la muestra.

4.3.2. Identificación de variantes. El formato VCF.

VCF header	<pre>##fileformat=VCFv4.0 ##fileDate=20100707 ##source=VCFtools ##reference=NCBI36 ##INFO=<ID=AA,Number=1,Type=String>Description="Ancestral Allele"> ##INFO=<ID=H2,Number=0,Type=Flag>Description="HapMap2 membership"> ##FORMAT=<ID=GT,Number=1,Type=String>Description="Genotype"> ##FORMAT=<ID=GQ,Number=1,Type=Integer>Description="Genotype Quality (phred score)"> ##FORMAT=<ID=GL,Number=3,Type=Float>Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)"> ##FORMAT=<ID=DP,Number=1,Type=Integer>Description="Read Depth"> ##ALT=<ID=DEL,Description="Deletion"> ##INFO=<ID=SVTYPE,Number=1,Type=String>Description="Type of structural variant"> ##INFO=<ID=END,Number=1,Type=Integer>Description="End position of the variant"></pre>	Mandatory header lines Optional header lines (meta-data about the annotations in the VCF body)																																																							
	<pre>#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2</pre>																																																								
Body	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;">#CHROM</th> <th style="text-align: left;">POS</th> <th style="text-align: left;">ID</th> <th style="text-align: left;">REF</th> <th style="text-align: left;">ALT</th> <th style="text-align: left;">QUAL</th> <th style="text-align: left;">FILTER</th> <th style="text-align: left;">INFO</th> <th style="text-align: left;">FORMAT</th> <th style="text-align: left;">SAMPLE1</th> <th style="text-align: left;">SAMPLE2</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>1</td> <td>.</td> <td>ACG</td> <td>A,AT</td> <td>.</td> <td>PASS</td> <td>.</td> <td>GT:DP</td> <td>1/2:13</td> <td>0/0:29</td> </tr> <tr> <td>1</td> <td>2</td> <td>rs1</td> <td>C</td> <td>T,CT</td> <td>.</td> <td>PASS</td> <td>H2;AA=T</td> <td>GT:GQ</td> <td>0 1:100</td> <td>2/2:70</td> </tr> <tr> <td>1</td> <td>5</td> <td>.</td> <td>A</td> <td>G</td> <td>.</td> <td>PASS</td> <td>.</td> <td>GT:GQ</td> <td>1 0:77</td> <td>1/1:95</td> </tr> <tr> <td>1</td> <td>100</td> <td>.</td> <td>T</td> <td></td> <td>.</td> <td>PASS</td> <td>SVTYPE=DEL;END=300</td> <td>GT:GQ:DP</td> <td>1/1:12:3</td> <td>0/0:20</td> </tr> </tbody> </table>	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2	1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29	1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70	1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95	1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20	Reference alleles (GT=0) Alternate alleles (GT>0 is an index to the ALT column) Phased data (G and C above are on the same chromosome)
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2																																															
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29																																															
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70																																															
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95																																															
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20																																															
<p>Deletion</p> <p>SNP</p> <p>Large SV</p> <p>Insertion</p> <p>Other event</p>																																																									

Si tenemos 0 | 1 quiere decir que esa variante es heterocigota, si tenemos 1 | 1 es homocigota. Cuando tenemos la barra / significa que las variantes no están en fase, y cuando tenemos la | significa que están en fase.

4.3.3. Post-procesado. Filtrado de variantes.

Una vez identificadas todas las potenciales variantes, que en nuestro caso son SNVs, inserciones y delecciones, es recomendable realizar un filtrado de todas las variantes.

En la etapa anterior de identificación de variantes hemos tratado de maximizar la sensibilidad (minimizar falsos negativos), pero en la etapa de filtrado vamos a maximizar la especificidad (minimizar falsos positivos). Esta etapa debe adaptarse a cada tipo de proyecto.

Tema 4 – Ejemplo 9



Tema 4 - Ejemplo 9 - Filtrado de Variantes

Una vez que tenemos el archivo VCF, podemos realizar este paso OPCIONAL, que consiste en filtrar las variantes por diferentes criterios. Utilizaremos uno de los archivos VCF obtenidos en el ejemplo anterior:

- 1) filtrar para quedarnos sólo con los indels (< 1 sg). ¿Cuántos indels se retienen?

```
vcftools --vcf chr7.vcf --keep-only-indels --recode --recode-INFO-all --out chr7_indels.vcf
```

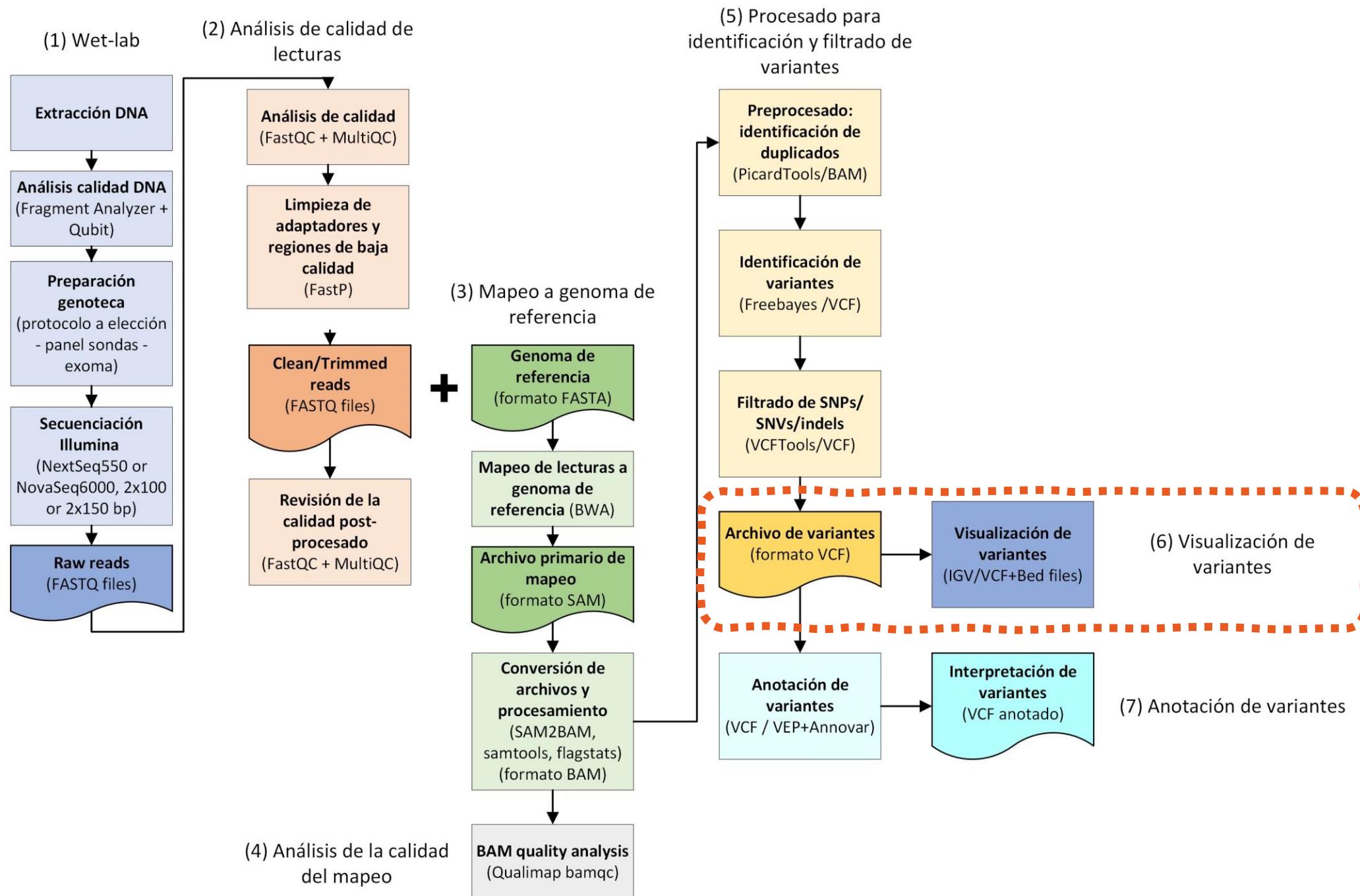
- 2) filtrar para eliminar los indels. Nos quedamos con los SNVs (< 1 sg). ¿Cuántos SNVs retiene?

```
vcftools --vcf chr7.vcf --remove-indels --recode --recode-INFO-all --out chr7_snvs.vcf
```

- 3) filtrar para quedarnos con las variantes con un mínimo de calidad de 15 y una profundidad de 10 (sample.GQ > 15 y sample.DP > 10)

```
vcftools --vcf chr7.vcf --minGQ 15 --minDP 10 --recode --recode-INFO-all --out chr7_GQ15DP10.vcf
```

Resultado: [Tema4_Ejemplo9_FiltradoVariantes](#) **No es el de `--minDP`, es `min-meanDP`**



Tema 4 – Ejemplo 10: Visualización de variantes en IGV



Tema 4 - Ejemplo 10 - Visualización de Variantes



Como paso opcional, pero recomendable, vamos a aprender a visualizar las variantes en el programa IGV. De esta manera podremos analizar:

- comparar la asignación de variantes a profundidades diferentes C=25 y C=150
- analizar las características de cada variante (posición, alelo referencia, alelo alternativo, calidad, tipo, frecuencia alélica, profundidad)

Para ello necesitamos los archivos anteriormente calculados y suministrados:

- archivo BAM (y su índice): chr7.dedup.RG.bam y chr7.dedup.RG.bam.bai
- archivo BED con las sondas (3313701_Coverered.bed)
- VCF con variantes (cargamos C25 y C150).

Archivos requeridos: [Tema4_Ejemplo10_VisualizaciónVariantes](#)

Revisa las cuestiones siguientes y contesta:

- 1) Visualiza el chr7 al completo. ¿Ves diferencias a simple vista?
- 2) Revisa el gen PMS2 (ayuda en la protección contra el cáncer de endometrio y ovario, se estudia como parte del panel para el síndrome de Lynch) y revisa a simple vista ¿cuántas variantes hay en común? Analiza esas variantes en común (posición, alelo referencia y alternativo, calidad, tipo, frecuencia alélica, profundidad y exón).
- 3) Realiza el mismo ejercicio para el gen MET (proto-oncogen, receptor de la tirosin-kinasa)
- 4) Realiza el mismo ejercicio para el gen CFTR (involucrado en fibrosis quística)

4.3. Identificación de variantes.

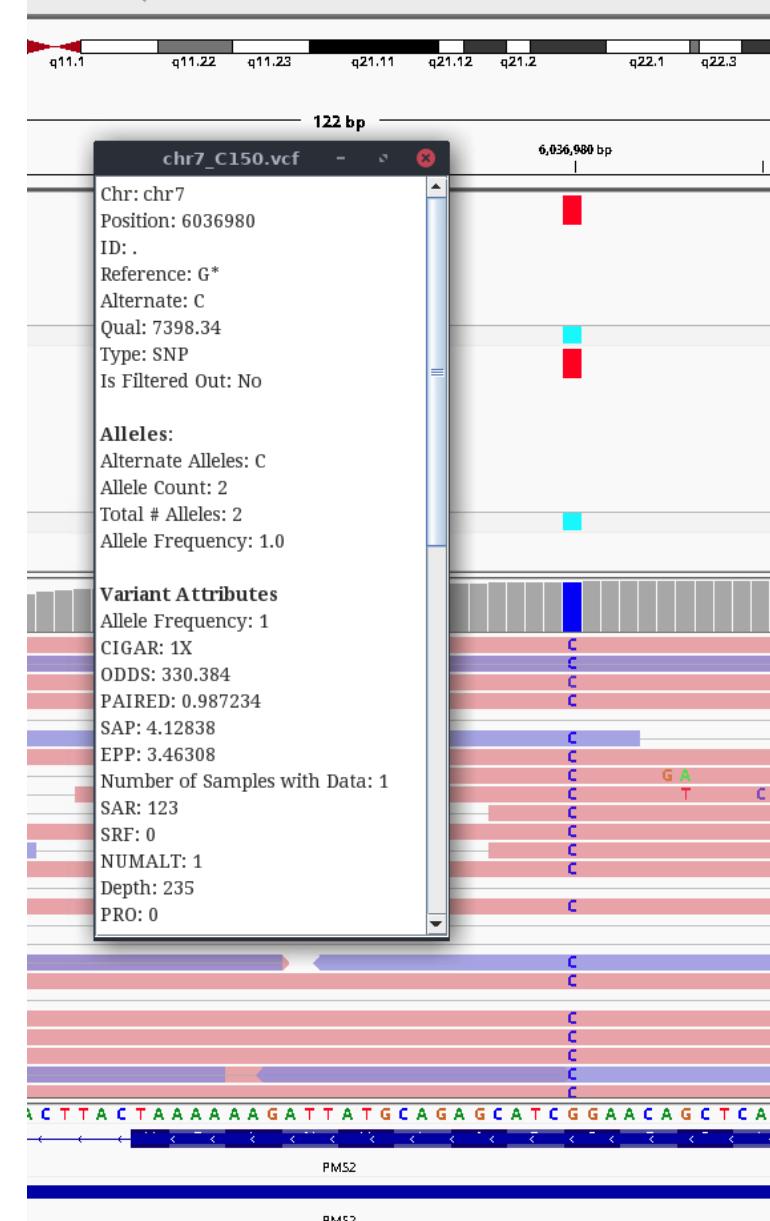
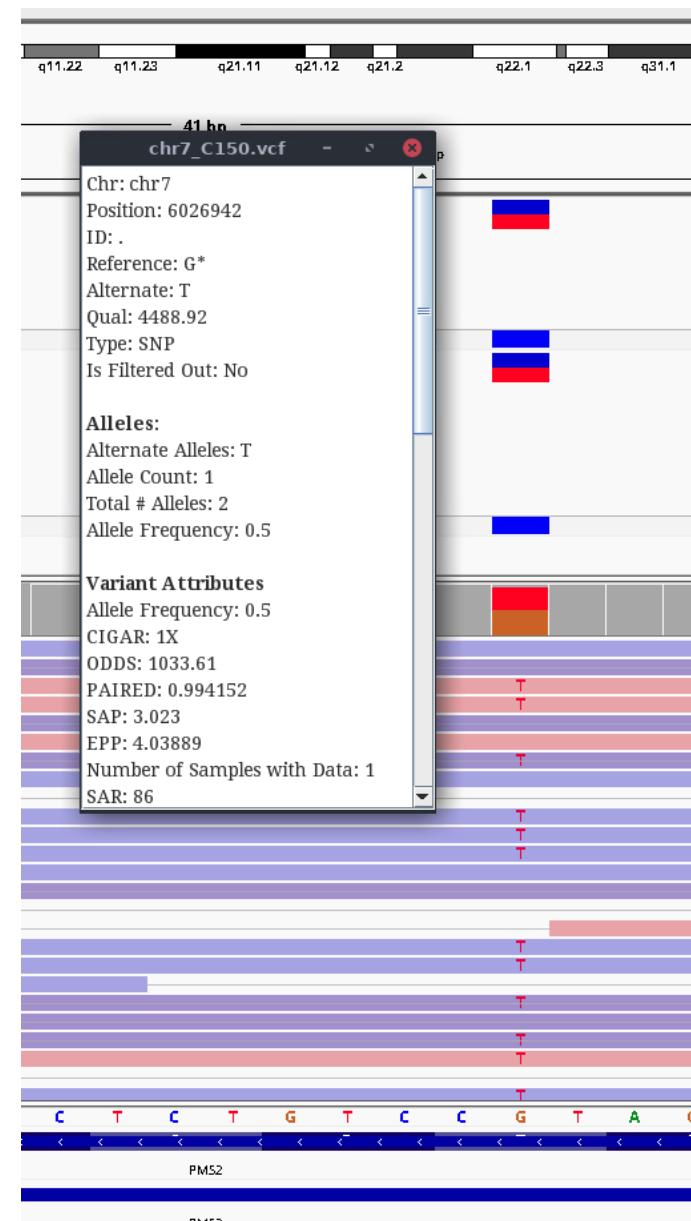
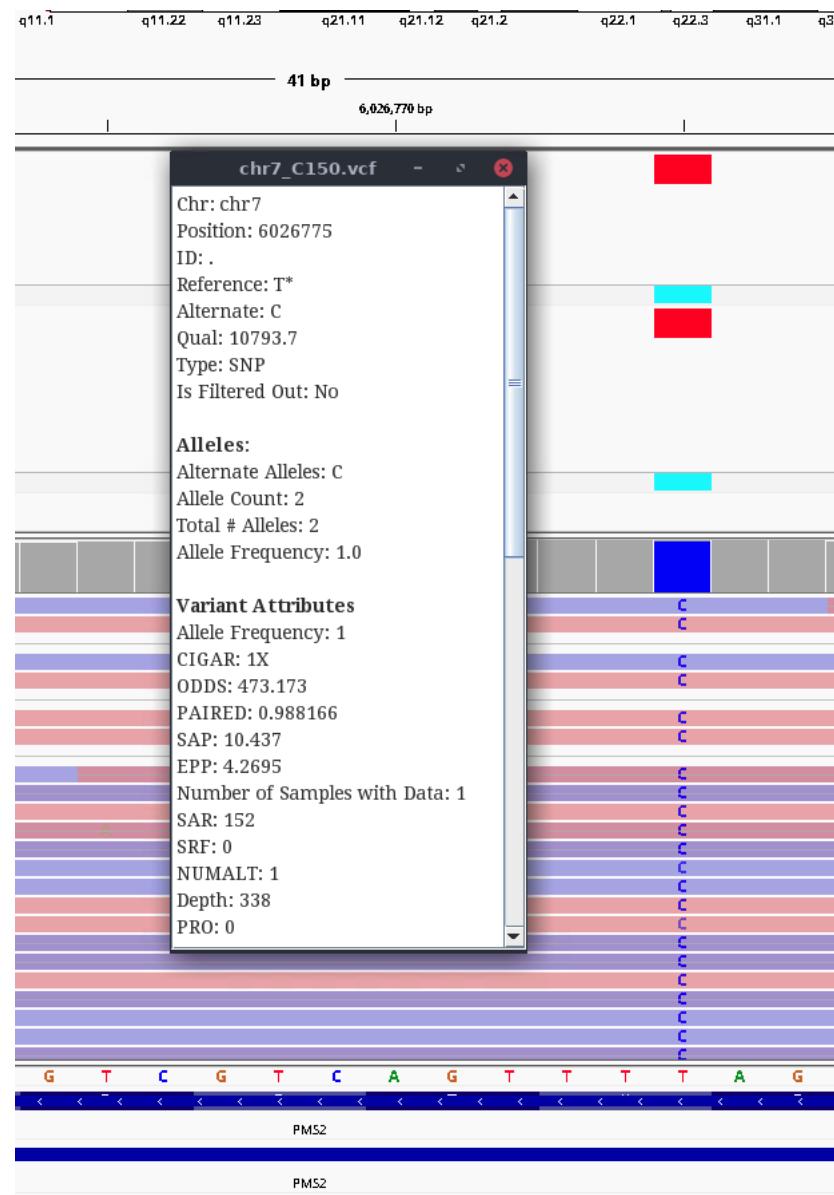


4.3. Identificación de variantes.

PMS2

4.3. Identificación de variantes.

PMS2



Ejemplo de registro manual:

Chr	Posición	Gen	Ref	Alt	Exón	Tipo	HOM/HET	Qual	Depth
7	6026775	PMS2	T	C	11	SNP	HOM	10793,7	338
7	6026942	PMS2	G	T	11	SNP	HET	4488,92	366
7	6036980	PMS2	G	C	7	SNP	HOM	7398,34	235
7	55229255	EGFR	G	A	13	SNP	HET	4531,82	326
7	55249063	EGFR	G	A	19	SNP	HET	3957,84	321

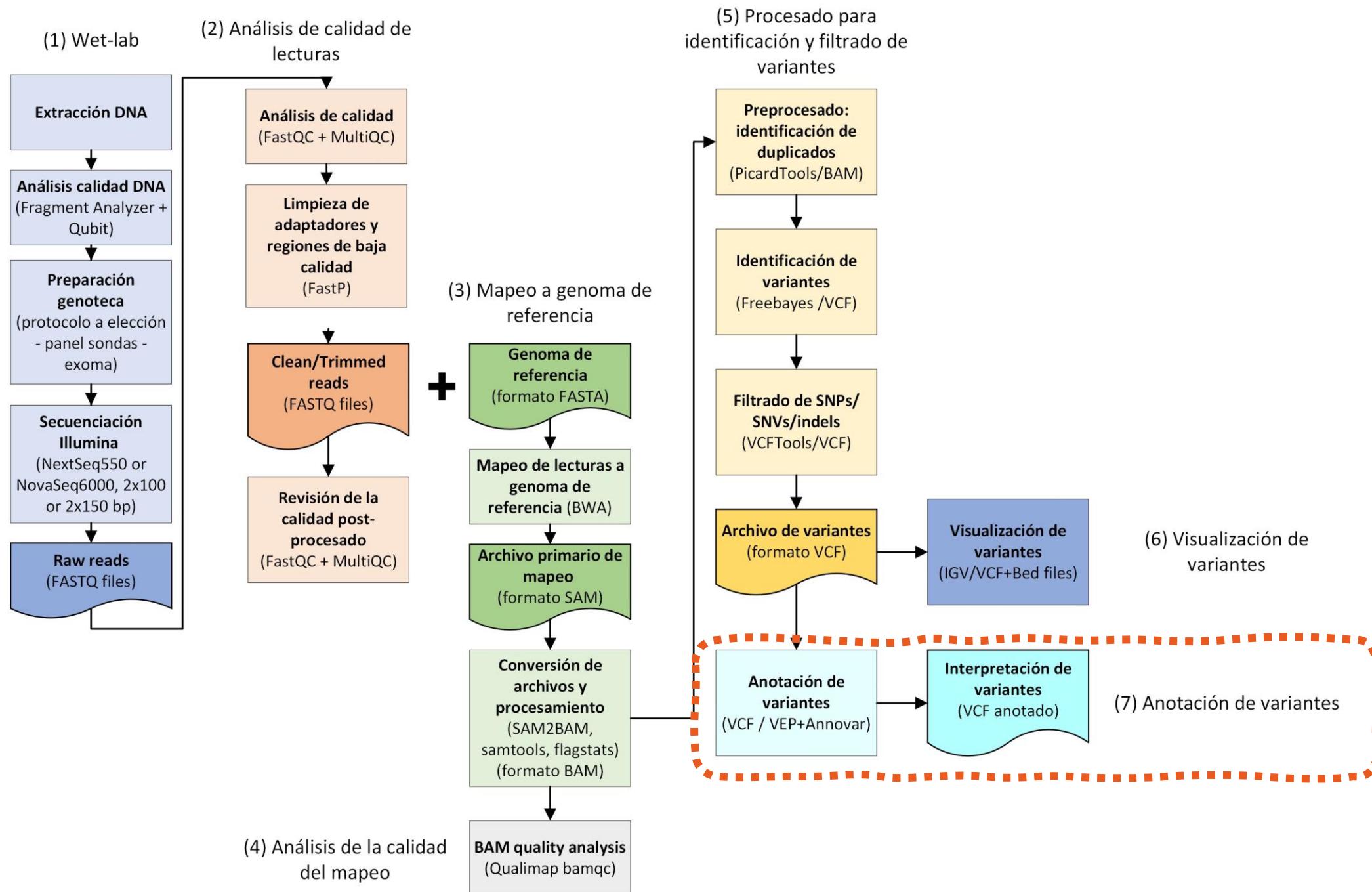
4.4

Anotación de variantes

4.4. Anotación de variantes.

4.4.1. Variant Effect Predictor (VEP).

4.4.2. Annotar



4.4. Anotación de variantes

El último paso, bioinformáticamente hablando, en el estudio de variantes es la anotación de las variantes pre-procesadas y filtradas.

Debemos tener en cuenta algunos datos, que nos dan idea de que en las etapas previas tenemos un listado de miles de variantes, cuando analizamos un exoma, o millones si estamos analizando un genoma completo.

Pero no todas ellas, sino un número muy reducido, serán las responsables de la enfermedad que estamos analizando.

Una persona sana, en promedio, tiene los siguientes números de variantes en su exoma:

>10.000 variantes no sinónimas, donde la variación de un solo nucleótido hace que el codón correspondiente codifique para otro aminoácido.

>10.000 variantes sinónimas, donde la variación del nucleótido no implica una variación en el aminoácido codificado.

>250 variantes con pérdida de función en el gen anotado.

>50 variantes que poseen relación con enfermedades previamente descritas.

4.4. Anotación de variantes

En esta etapa identificaremos un subconjunto pequeño de variantes potencialmente interesantes, a partir del total de variantes candidatas.

Esta anotación nos ayudará a reducir el número de variantes y focalizar nuestra atención y esfuerzos en las más interesantes. **Este proceso de anotación es el proceso de asignar información biológica relevante a las variantes detectadas**, haciendo uso de diversas bases de datos.

Este proceso puede ser iterativo, y ser necesarias varias bases de datos para recabar información suficiente para determinar las variantes sospechosas.

4.4. Anotación de variantes

Lo primero que nos va a interesar de una variante es su consecuencia, ¿qué consecuencia tiene esa variante? Puede tener las siguientes consecuencias:

Variante sinónima / synonymous variant: Variante en una región codificante, cuyo cambio nucleotídico **no supone un cambio en la proteína.**

Variante no-sinónima / missense variant: La variante cambia una o más bases, **produciendo una secuencia de aminoácidos diferente.**

Parada prematura / Stop gained: La variante provoca la aparición de un codón de parada prematuro, **producido un transcripto de menor longitud.**

Cambio marco de lectura / frameshift variant: La variante provoca el desplazamiento del marco de lectura, **habitual cuando la variante es una inserción o delección.**

Adicionalmente, podremos preguntarnos cuestiones como si la variante ha sido reportada previamente, qué frecuencia alélica tiene en población general, qué impacto tiene en la función de la proteína y en su estructura, si existen enfermedades asociadas a dicha variante o si se encuentra en una región alta o bajamente conservada. En este tema veremos la anotación con las bases de datos VEP y Annovar.

4.4. Anotación de variantes

En los siguientes enlaces encontrarás bases de datos que proporcionan información sobre distintos aspectos de las variantes encontradas.

Descripción de todos los tipos de consecuencia, descritos por ENSEMBL, donde se categorizan por orden de importancia de la consecuencia que provocan.

https://m.ensembl.org/info/genome/variation/prediction/predicted_data.html

Estudio de la relación del gen con distintas enfermedades según la base de datos **OMIM (NCBI)**. Esta base de datos incluye enfermedades de base genética y ofrece información sobre su manifestación genotípica.

<https://www.omim.org/>

La base de datos **ClinVar (NCBI)** proporciona un repositorio de relación entre las variaciones en humano y los fenotipos observados. <https://www.ncbi.nlm.nih.gov/clinvar/>

4.4.1. Variant Effect Predictor (VEP)

Esta es una herramienta de software gratuita que se puede ejecutar *on line* o bien por vía de comandos (<https://www.ensembl.org/info/docs/tools/vep/index.html>), así como utilizarlo en el programa R (<https://bioconductor.org/packages/release/bioc/html/ensemblVEP.html>).

VEP soporta varias especies y determina el efecto de una variante de tipo SNV, indel, CNV o variación estructural en genes, transcritos y secuencias de proteínas. De esta forma nos proporciona la siguiente información:

Genes y transcritos afectados

Localización de la variante (en secuencia codificante, región reguladora, etc)

Consecuencia de la variante en la secuencia de la proteína

Información sobre si esta variante se ha descrito anteriormente o su frecuencia alélica en el proyecto 1000 genomas

Predicción del cambio de aminoácido, mediante la utilización de las herramientas SIFT (<https://sift.bii.a-star.edu.sg/>) y Polyphen (<http://genetics.bwh.harvard.edu/pph2/>). Ambas se basan en métodos de aprendizaje supervisado para predecir el nivel de malignidad de una variante. SIFT nos dirá si el cambio de aminoácido afecta a la función de la proteína, basándose en homología de secuencias; mientras que Polyphen realizará una predicción sobre el impacto de una variación de aminoácido en la estructura y función proteica.

4.4.1. Variant Effect Predictor (VEP)

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Login/Register

Using this website Annotation and prediction Data access API & software About us

In this section

- VEP web interface
 - Input form
 - Results
- VEP command line
 - Tutorial
 - Download and install
 - Running VEP
 - Annotation sources
 - Filtering results
 - Custom annotations
 - Plugins
 - Examples and use cases
 - Other information
- Data formats
- Variant Recoder
- HaploSaurus
- VEP FAQ

On this page

- VEP interfaces
- Publication
- VEP related tools

Search documentation Go

Ensembl Variant Effect Predictor (VEP)

Ve!P

VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.

Simply input the coordinates of your variants and the nucleotide changes to find out the:

- Genes and Transcripts affected by the variants
- Location of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- Consequence of your variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift), see [variant consequences](#)
- Known variants that match yours, and associated minor allele frequencies from the [1000 Genomes Project](#)
- SIFT and PolyPhen-2 scores for changes to protein sequence
- ... And more! See [data types](#), [versions](#).

★ [What's new in release 105?](#)

VEP interfaces

Web interface



- Point-and-click interface
- Suits smaller volumes of data

[Documentation](#)

Launch Ve!P

Command line tool



- More options and flexibility
- For large volumes of data

[Documentation](#)

[Clone from GitHub](#)

[Download \(.zip\)](#)

[Pull Docker image from DockerHub](#)

REST API



- Language-independent API
- Simple URL-based queries

[Documentation](#)

[VEP REST API](#)



Tema 4 - Ejemplo 11 - Variant Effect Predictor (VEP)

En este ejemplo vamos a realizar la anotación de variantes en VEP con los datos obtenidos en ejemplos anteriores ([Tema4_Ejemplo11_VEP](#))
Accedemos vía web a la herramienta VEP (http://grch37.ensembl.org/Homo_sapiens/Tools/VEP), siendo cuidadosos de seleccionar el genoma de referencia correspondiente a nuestro mapeo (en nuestro caso GRCh37/Hg19).

Importante: vamos a utilizar el archivo VCF completo (chr7.vcf).

Indicamos en las opciones:

- Species: por defecto humano, cuidando que sea la versión de genoma adecuado (GRCh37 o GRCh38).
- Name for the job: Tema4_Ejemplo11
- Subimos el archivo VCF generado en la identificación de variants
- Bases de datos a utilizar Ensembl/Gencode
- Identifiers and frequency data: podemos añadir identificadores adicionales para genes, transcritos y variantes, así como información relacionada con la frecuencia de las variantes en bases de datos como los 1000 genomas o el de los 6000 exomas. Vamos a seleccionar:
 - Identificador de proteína Ensembl
 - Frecuencia alélica global de los 1000 genomas
 - El resto lo dejamos por defecto
- Opciones extra.
 - Información miscelánea de la variante: biotipo del transcripto (miRNA, secuencia codificante, etc...). Dejamos por defecto
 - Predicciones acerca de la patogenicidad de la variante. Elegimos:
 - SIFT y Polyphen: son dos algoritmos que predicen cómo un cambio de aminoácido afecta a la función de la proteína. Los dejamos activados (predicción y score).
- Opciones de filtrado: podemos añadir opciones para filtrar las variantes. Por ejemplo, por frecuencia de aparición en los 1000 genomas: podríamos eliminar aquellas variantes que son frecuentes en la población y por tanto, no tienen significado relevante en patogenicidad.

Nota: puedes repetir el ejercicio utilizando el archivo VCF con opción C150

Variant Effect Predictor ?

New job

Species:



Homo_sapiens

X

Assembly: GRCh37.p13

VEP for non-human species is now only available on this site for Human (GRCh37). For other species, please visit our [main site](#).

Name for this job (optional):

Tema4_Ejemplo10

Input data:

Either paste data:



Examples: [Ensembl default](#), [VCF](#), Variant identifiers, HGVS notations

Or upload file:

Seleccionar archivo chr7.vcf

Or provide file URL:

Transcript database to use:

Ensembl/GENCODE transcripts

Ensembl/GENCODE basic transcripts

RefSeq transcripts

Ensembl/GENCODE and RefSeq transcripts

Identifiers □ Additional identifiers for genes, transcripts and variants**Identifiers****Gene symbol:****Transcript version:****CCDS:****Protein:****UniProt:****HGVS:****Variants and frequency data** □ Co-located variants and frequency data**Variants and frequency data****Find co-located known variants:****Variant synonyms:****Frequency data for co-located variants:**

- 1000 Genomes global minor allele frequency
- 1000 Genomes continental allele frequencies
- ESP allele frequencies
- gnomAD (exomes) allele frequencies

PubMed IDs for citations of co-located variants:**Include flagged variants:**

Transcript annotation

Transcript biotype:

Exon and intron numbers:

Transcript support level:

APPRIS:

MANE:

Identify canonical transcripts:

Upstream/Downstream distance (bp):

5000

miRNA structure:

Protein annotation

Protein domains:

Regulatory data

Get regulatory region consequences:

Yes

Phenotype data and citations

Phenotypes:

Mastermind:

Pathogenicity predictions

SIFT:**PolyPhen:****dbNSFP:** Disabled Enabled**CADD:****LoFtool:****MPC:**

Splicing predictions

dbScSNV:**MaxEntScan:****SpliceAI:** Disabled Enabled

Conservation

BLOSUM62:**Ancestral allele:**

Filtering options □ Pre-filter results by frequency or consequence type

Filters

Filter by frequency:

- No filtering
- Exclude common variants
- Advanced filtering

Return results for variants in coding regions

only:

Restrict results:

Show all results ▾

NB: Restricting results may exclude biologically important data!

Advanced options □ Additional enhancements

Advanced options

Buffer size:

5000 ▾

NB: When the **Regulatory data** option is selected then due to the large amount of regulatory data available, the **maximum buffer size** is automatically reduced from the default value of **5000** to **500**. This reduces the memory requirement but might increase the run time. If you find that your jobs are still failing due to memory limitations then you can select a value **lower than 500**.

Right align variants prior to consequence calculation:

No ▾

Variant Effect Predictor

New job

Recent jobs

1 Refresh

Show/hide columns (1 hidden)

Filter

Analysis	Jobs	Submitted at	
Variant Effect Predictor	 VEP analysis of Tema4_Ejemplo10 in Homo_sapiens Queued	09/01/2022, 21:04 (GMT)	  

Variant Effect Predictor

New job

Recent jobs

Refresh

Show/hide columns (1 hidden)

Filter

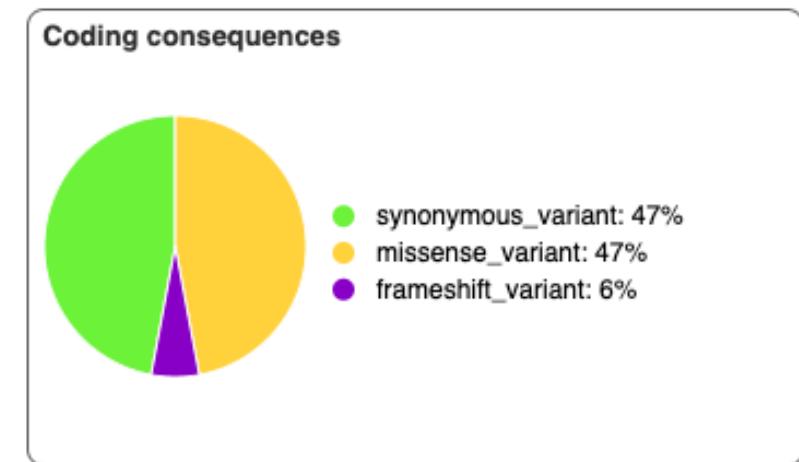
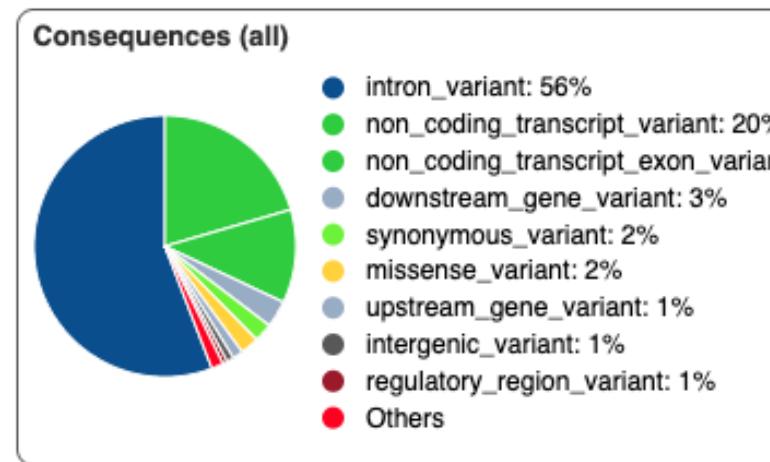
Analysis	Jobs	Submitted at	
Variant Effect Predictor	 VEP analysis of Tema4_Ejemplo10 in Homo_sapiens Done View results	09/01/2022, 21:04 (GMT)	  

Variant Effect Predictor results

Job details 

Summary statistics 

Category	Count
Variants processed	69
Variants filtered out	0
Novel / existing variants	29 (42.0) / 40 (58.0)
Overlapped genes	12
Overlapped transcripts	44
Overlapped regulatory features	4



Results preview

Navigation (per variant)				Filters				Download				New job			
				Uploaded variant is defined				All: VCF VEP TXT				BioMart: Variants Genes			
Show/hide columns (16 hidden)															
Uploaded variant	Location	Allele	Consequence	Symbol	Gene	Feature type	Feature	Biotype	Exon	cDNA position	CDS pos	Scroll to see more columns »		Co	Co
	7:6026775-6026775	C	missense_variant	PMS2	ENSG00000122512	Transcript	ENST00000265849.7	protein_coding	11/15	1727	1621	541	K/E	AA/	
	7:6026775-6026775	C	intron_variant	PMS2	ENSG00000122512	Transcript	ENST00000382321.4	protein_coding	-	-	-	-	-	-	
	7:6026775-6026775	C	missense_variant	PMS2	ENSG00000122512	Transcript	ENST00000406569.3	protein_coding	11/12	1621	1621	541	K/E	AA/	
	7:6026775-6026775	C	missense_variant	PMS2	ENSG00000122512	Transcript	ENST00000441476.2	protein_coding	8/12	1303	1303	435	K/E	AA/	
	7:6026775-6026775	C	intron_variant, non_coding_transcript_variant	PMS2	ENSG00000122512	Transcript	ENST00000469652.1	processed_transcript	-	-	-	-	-	-	
	7:6026942-6026942	T	missense_variant	PMS2	ENSG00000122512	Transcript	ENST00000265849.7	protein_coding	11/15	1560	1454	485	T/K	ACG	
	7:6026942-6026942	T	intron_variant	PMS2	ENSG00000122512	Transcript	ENST00000382321.4	protein_coding	-	-	-	-	-	-	
	7:6026942-6026942	T	missense_variant	PMS2	ENSG00000122512	Transcript	ENST00000406569.3	protein_coding	11/12	1454	1454	485	T/K	ACG	
	7:6026942-6026942	T	missense_variant	PMS2	ENSG00000122512	Transcript	ENST00000441476.2	protein_coding	8/12	1136	1136	379	T/K	ACG	
	7:6026942-6026942	T	intron_variant, non_coding_transcript_variant	PMS2	ENSG00000122512	Transcript	ENST00000469652.1	processed_transcript	-	-	-	-	-	-	
	7:6036980-6036980	C	synonymous_variant	PMS2	ENSG00000122512	Transcript	ENST00000265849.7	protein_coding	7/15	886	780	260	S	TCC	
	7:6036980-6036980	C	downstream_gene_variant	PMS2	ENSG00000122512	Transcript	ENST00000380416.5	retained_intron	-	-	-	-	-	-	
	7:6036980-6036980	C	synonymous_variant	PMS2	ENSG00000122512	Transcript	ENST00000382321.4	protein_coding	7/11	780	780	260	S	TCC	
	7:6036980-6036980	C	synonymous_variant	PMS2	ENSG00000122512	Transcript	ENST00000406569.3	protein_coding	7/12	780	780	260	S	TCC	
	7:6036980-6036980	C	synonymous_variant	PMS2	ENSG00000122512	Transcript	ENST00000441476.2	protein_coding	4/12	462	462	154	S	TCC	

Navigation (per variant)
Show: 1 5 10 50 All variants

Filters
Symbol is PMS2
Uploaded variant defined

Download
All: VCF VEP TXT
Filtered: VCF VEP TXT
BioMart: Variants Genes

New job

PMS2

Show/hide columns (12 hidden)																		
Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon	Intron	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	F
7:6026775-6026775	C		missense_variant	MODERATE	PMS2	ENSG00000122512	Transcript	ENST00000265849.7	protein_coding	11/15	-	1727	1621	541	K/E	AAA/GAA	rs2228006, COSV56223170	-
7:6026775-6026775	C		intron_variant	MODIFIER	PMS2	ENSG00000122512	Transcript	ENST00000382321.4	protein_coding	-	7/10	-	-	-	-	rs2228006, COSV56223170	-	
7:6026775-6026775	C		missense_variant	MODERATE	PMS2	ENSG00000122512	Transcript	ENST00000406569.3	protein_coding	11/12	-	1621	1621	541	K/E	AAA/GAA	rs2228006, COSV56223170	-
7:6026775-6026775	C		missense_variant	MODERATE	PMS2	ENSG00000122512	Transcript	ENST00000441476.2	protein_coding	8/12	-	1303	1303	435	K/E	AAA/GAA	rs2228006, COSV56223170	-
7:6026775-6026775	C		intron_variant, non_coding_transcript_variant	MODIFIER	PMS2	ENSG00000122512	Transcript	ENST00000469652.1	processed_transcript	-	2/2	-	-	-	-	rs2228006, COSV56223170	-	
7:6026942-6026942	T		missense_variant	MODERATE	PMS2	ENSG00000122512	Transcript	ENST00000265849.7	protein_coding	11/15	-	1560	1454	485	T/K	ACG/AAG	rs1805323, COSV56219377	-
7:6026942-6026942	T		intron_variant	MODIFIER	PMS2	ENSG00000122512	Transcript	ENST00000382321.4	protein_coding	-	7/10	-	-	-	-	rs1805323, COSV56219377	-	
7:6026942-6026942	T		missense_variant	MODERATE	PMS2	ENSG00000122512	Transcript	ENST00000406569.3	protein_coding	11/12	-	1454	1454	485	T/K	ACG/AAG	rs1805323, COSV56219377	-
7:6026942-6026942	T		missense_variant	MODERATE	PMS2	ENSG00000122512	Transcript	ENST00000441476.2	protein_coding	8/12	-	1136	1136	379	T/K	ACG/AAG	rs1805323, COSV56219377	-
7:6026942-6026942	T		intron_variant, non_coding_transcript_variant	MODIFIER	PMS2	ENSG00000122512	Transcript	ENST00000469652.1	processed_transcript	-	2/2	-	-	-	-	rs1805323, COSV56219377	-	
7:6036980-6036980	C		synonymous_variant	LOW	PMS2	ENSG00000122512	Transcript	ENST00000265849.7	protein_coding	7/15	-	886	780	260	S	TCC/TCG	rs1805319, CD136761, COSV56223185	-
7:6036980-6036980	C		downstream_gene_variant	MODIFIER	PMS2	ENSG00000122512	Transcript	ENST00000380416.5	retained_intron	-	-	-	-	-	-	rs1805319, CD136761, COSV56223185	-	
7:6036980-6036980	C		synonymous_variant	LOW	PMS2	ENSG00000122512	Transcript	ENST00000382321.4	protein_coding	7/11	-	780	780	260	S	TCC/TCG	rs1805319, CD136761, COSV56223185	-
7:6036980-6036980	C		synonymous_variant	LOW	PMS2	ENSG00000122512	Transcript	ENST00000406569.3	protein_coding	7/12	-	780	780	260	S	TCC/TCG	rs1805319, CD136761, COSV56223185	-
7:6036980-6036980	C		synonymous_variant	LOW	PMS2	ENSG00000122512	Transcript	ENST00000441476.2	protein_coding	4/12	-	462	462	154	S	TCC/TCG	rs1805319, CD136761, COSV56223185	-
7:6036980-6036980	C		intron_variant, non_coding_transcript_variant	MODIFIER	PMS2	ENSG00000122512	Transcript	ENST00000469652.1	processed_transcript	-	2/2	-	-	-	-	rs1805319, CD136761, COSV56223185	-	

PMS2

Navigation (per variant) **Filters** **Download** **New job**

Show: [1](#) [5](#) [10](#) [50](#) [All](#) variants

Symbol is PMS2 All: [VCF](#) [VEP](#) [TXT](#)
 Filtered: [VCF](#) [VEP](#) [TXT](#)
 BioMart: [Variants](#) [Genes](#)

Show/hide columns (12 hidden)

variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Protein	Exon	Intron	cDNA position	CDSS position	Protein position	Amino acids	Codon	Transcript	Ref	variant
7:6026775-6026775		C	missense_variant	MODERATE	PMS2	ENSG00000122512	Transcript	ENST00000265849.7	protein_coding	11/15	-	1727	1621	541	K/E	AAA/GAA	rs2228006, COSV56223170	-	
7:6026775-6026775		C	intron_variant	MODIFIER	PMS2	ENSG00000122512	Transcript	ENST00000382321.4	protein_coding	-	7/10	-	-	-	-	-	rs2228006, COSV56223170	-	
7:6026775-6026775		C	missense_variant	MODERATE	PMS2	ENSG00000122512	Transcript	ENST00000406569.3	protein_coding	11/12	-	1621	1621	541	K/E	AAA/GAA	rs2228006, COSV56223170	-	
7:6026775-6026775		C	missense_variant	MODERATE	PMS2	ENSG00000122512	Transcript	ENST00000441476.2	protein_coding	8/12	-	1303	1303	435	K/E	AAA/GAA	rs2228006, COSV56223170	-	
7:6026775-6026775		C	intron_variant, non_coding_transcript_variant	MODIFIER	PMS2	ENSG00000122512	Transcript	ENST00000469652.1	processed_transcript	-	2/2	-	-	-	-	-	rs2228006, COSV56223170	-	
7:6026942-6026942		C	missense_variant	MODERATE	PMS2	ENSG00000122512	Transcript	ENST00000265849.7	protein_coding	7/15	-	1500	1454	485	T/K	ACG/AAG	rs1805323, COSV56219377	-	
7:6026942-6026942		T	intron_variant	MODIFIER	PMS2	ENSG00000122512	Transcript	ENST00000382321.4	protein_coding	-	7/10	-	-	-	-	-	rs1805323, COSV56219377	-	
7:6026942-6026942		T	missense_variant	MODERATE	PMS2	ENSG00000122512	Transcript	ENST00000406569.3	protein_coding	11/12	-	1454	1454	485	T/K	ACG/AAG	rs1805323, COSV56219377	-	
7:6026942-6026942		T	missense_variant	MODERATE	PMS2	ENSG00000122512	Transcript	ENST00000441476.2	protein_coding	8/12	-	1136	1136	379	T/K	ACG/AAG	rs1805323, COSV56219377	-	
7:6026942-6026942		T	intron_variant, non_coding_transcript_variant	MODIFIER	PMS2	ENSG00000122512	Transcript	ENST00000469652.1	processed_transcript	-	2/2	-	-	-	-	-	rs1805323, COSV56219377	-	
7:6036980-6036980		C	synonymous_variant	LOW	PMS2	ENSG00000122512	Transcript	ENST00000265849.7	protein_coding	7/15	-	886	780	260	S	TCC/TCG	rs1805319, CD136761, COSV56223185	-	
7:6036980-6036980		C	downstream_gene_variant	MODIFIER	PMS2	ENSG00000122512	Transcript	ENST00000380416.5	retained_intron	-	-	-	-	-	-	-	rs1805319, CD136761, COSV56223185	-	
7:6036980-6036980		C	synonymous_variant	LOW	PMS2	ENSG00000122512	Transcript	ENST00000382321.4	protein_coding	7/11	-	780	780	260	S	TCC/TCG	rs1805319, CD136761, COSV56223185	-	
7:6036980-6036980		C	synonymous_variant	LOW	PMS2	ENSG00000122512	Transcript	ENST00000406569.3	protein_coding	7/12	-	780	780	260	S	TCC/TCG	rs1805319, CD136761, COSV56223185	-	
7:6036980-6036980		C	synonymous_variant	LOW	PMS2	ENSG00000122512	Transcript	ENST00000441476.2	protein_coding	4/12	-	462	462	154	S	TCC/TCG	rs1805319, CD136761, COSV56223185	-	
7:6036980-6036980		C	intron_variant, non_coding_transcript_variant	MODIFIER	PMS2	ENSG00000122512	Transcript	ENST00000469652.1	processed_transcript	-	2/2	-	-	-	-	-	rs1805319, CD136761, COSV56223185	-	

PMS2

AF, frecuencia alélica poblacional: va de 0 - 1, si tiene una frecuencia muy alta se descarta. Se descarta si es una FA que esté en más de un 5%, es decir, mayor que un 0,05

Navigation (per variant) Show: 1 5 10 50 All variants

Filters Symbol is PMS2 Impact is defined Add

Download All: VCF VEP TXT Filtered: VCF VEP TXT BioMart: Variants Genes New job

Show/hide columns (12 hidden)

cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand	Symbol source	ENSP	SIFT	PolyPhen	AF	AFR AF	AMR AF	EAS AF	EUR AF	SAS AF	Clinical significance	Somatic status	Phenotype or disease	Pubmed	Associated phenotypes
1727	1621	541	K/E	AAA/GAA	rs2228006, COSV56223170	-1	HGNC	ENSP00000265849	1	0	0.8832	0.9418	0.7738	0.878	0.8767	0.8937	benign	0, 1	1, 1	23 PubMed IDs	8 Phenotype associations
-	-	-	-	-	rs2228006, COSV56223170	-1	HGNC	ENSP00000371758	-	-	0.8832	0.9418	0.7738	0.878	0.8767	0.8937	benign	0, 1	1, 1	23 PubMed IDs	8 Phenotype associations
1621	1621	541	K/E	AAA/GAA	rs2228006, COSV56223170	-1	HGNC	ENSP00000384308	1	0	0.8832	0.9418	0.7738	0.878	0.8767	0.8937	benign	0, 1	1, 1	23 PubMed IDs	8 Phenotype associations
1303	1303	435	K/E	AAA/GAA	rs2228006, COSV56223170	-1	HGNC	ENSP00000392843	1	0	0.8832	0.9418	0.7738	0.878	0.8767	0.8937	benign	0, 1	1, 1	23 PubMed IDs	8 Phenotype associations
-	-	-	-	-	rs2228006, COSV56223170	-1	HGNC	-	-	-	0.8832	0.9418	0.7738	0.878	0.8767	0.8937	benign	0, 1	1, 1	23 PubMed IDs	8 Phenotype associations
1560	1454	485	T/K	ACG/AAG	rs1805323, COSV56219377	-1	HGNC	ENSP00000265849	0.92	0	0.1120	0.0098	0.0173	0.3373	0.0457	0.1534	benign	0, 1	1, 1	17 PubMed IDs	5 Phenotype associations

Protein position	Amino acids	Codons	Existing variant	Feature strand	Symbol source	ENSP	SIFT	PolyPhen	AF	AFR AF
541	K/E	AAA/GAA	rs2228006, COSV56223170	-1	HGNC	ENSP00000265849	1	0	0.8832	0.9418
-	-	-	rs2228006, COSV56223170	-1	Variant: rs2228006 more about rs2228006	ENSP00000371758	-	-	0.8832	0.9418
541	K/E	AAA/GAA	rs2228006, COSV56223170	-1	Class SNP Location 7:6026775 Alleles T/A/C/G Ambiguity code N	ENSP00000384308	1	0	0.8832	0.9418
435	K/E	AAA/GAA	rs2228006, COSV56223170	-1	Global MAF 0.1168 (T) Consequence stop gained Evidences Frequency, 1000Genomes, Cited, ESP, Phenotype or Disease, ExAC, TOPMed, gnomAD	ENSP00000392843	1	0	0.8832	0.9418
-	-	-	rs2228006, COSV56223170	-1	Sources UniProt, PhenCode, dbSNP, dbSNP HGVS, ClinVar, Archive dbSNP	ENSP00000265849	0.92	0	0.1120	0.0098
485	T/K	ACG/AAG	rs1805323, COSV56219377	-1	Population genetics Phenotype data	ENSP00000371758	-	-	0.1120	0.0098
-	-	-	rs1805323, COSV56219377	-1	Sources UniProt, PhenCode, dbSNP, dbSNP HGVS, ClinVar, Archive dbSNP	ENSP00000384308	0.96	0	0.1120	0.0098
485	T/K	ACG/AAG	rs1805323, COSV56219377	-1	Population genetics Phenotype data	ENSP00000392843	0.92	0	0.1120	0.0098
379	T/K	ACG/AAG	rs1805323, COSV56219377	-1	HGNC	ENSP00000392843	0.92	0	0.1120	0.0098

Protein position	Amino acids	Codons	Existing variant	Feature strand	Symbol source	ENSP	SIFT	PolyPhen	AF
541	K/E	AAA/GAA	rs2228006, COSV56223170	-1	HGNC	ENSP00000265849	1	0	0.8832
-	-	-	rs2228006, COSV56223170	-1	more about rs2228006	ENSP00000371758	-	-	0.8832
541	K/E	AAA/GAA	rs2228006, COSV56223170	-1	Class SNP Location 7:6026775 Alleles T/A/C/G Ambiguity code N	ENSP00000384308	1	0	0.8832
435	K/E	AAA/GAA	rs2228006, COSV56223170	-1	Global MAF 0.1168 (T) Consequence stop gained Evidences Frequency, 1000Genomes, Cited, ESP, Phenotype or Disease, ExAC, TOPMed, gnomAD	ENSP00000392843	1	0	0.8832
-	-	-	rs2228006, COSV56223170	-1	Sources UniProt, PhenCode, dbSNP, dbSNP HGVS, ClinVar, Archive dbSNP	ENSP00000265849	0.92	0	0.1120
485	T/K	ACG/AAG	rs1805323, COSV56219377	-1	Population genetics Phenotype data	ENSP00000371758	-	-	0.1120
-	-	-	rs1805323, COSV56219377	-1	Sources UniProt, PhenCode, dbSNP, dbSNP HGVS, ClinVar, Archive dbSNP	ENSP00000384308	0.96	0	0.1120
485	T/K	ACG/AAG	rs1805323, COSV56219377	-1	Population genetics Phenotype data	ENSP00000392843	0.92	0	0.1120

Hay que quedarse con las variantes que no tienen valor AF o que tienen muy bajo valor

Create alert Advanced

[Home](#)[About](#) ▾[Access](#) ▾[Help](#) ▾[Submit](#) ▾[Statistics](#) ▾[FTP](#) ▾**Clinical significance**

Conflicting interpretations (0)

Benign (2)

Likely benign (0)

Uncertain significance (0)

Likely pathogenic (0)

Pathogenic (0)

Molecular consequence

Frameshift (0)

Missense (1)

Nonsense (0)

Splice site (0)

ncRNA (2)

Near gene (0)

UTR (0)

Variation type

Deletion (0)

Duplication (0)

Indel (0)

Insertion (0)

Single nucleotide (2)

Variation size

Short variant (< 50 bps) (1)

Structural variant (>= 50 bps) (0)

Variant length

< 1kb, single gene (1)

> 1kb, single gene (0)

> 1kb, multiple genes (0)

Review status

Practice guideline (0)

Expert panel (1)

Multiple submitters (1)

Single submitter (0)

At least one star (2)

Conflicting interpretations (0)

[Clear all](#)[Show additional filters](#)[Display options](#) ▾ [Sort by Variation](#) ▾ [Download](#) ▾rs2228006 has not been reported to ClinVar. Refer to dbSNP record rs2228006 for details on variation at this location. Please consider [submitting your interpretation](#) of this variant to ClinVar

Was this helpful?

**GENETIC VARIATION**[rs2228006](#)**Clinical significance:** not reported in ClinVar**Organism:** *Homo sapiens***Reference allele:** T**GRCh38.p13:** NC_000007.14: 5,987,143**Variation alleles:** A, C, G**GRCh37.p13:** NC_000007.13: 6,026,774**Variation type:** single nucleotide variation[Genome Data Viewer](#)**Search results****Items: 2**

	Variation Location	Gene(s)	Protein change	Condition(s)	Clinical significance (Last reviewed)	Review status
<input type="checkbox"/>	NM_000535.7(PMS2):c.1621= (p.Lys541=) 1. GRCh37: Chr7:6026775 GRCh38: Chr7:5987144	PMS2		Lynch syndrome	Benign (Sep 5, 2013)	reviewed by expert panel
<input type="checkbox"/>	NM_000535.7(PMS2):c.1621A>G (p.Lys541 Glu) 2. GRCh37: Chr7:6026775 GRCh38: Chr7:5987144	PMS2	K541E, K406E, K354E, K438E, K489E, K230E, K350E, K435E	Hereditary nonpolyposis colorectal neoplasms, Mismatch repair cancer syndrome 4, not specified, Lynch syndrome 4	Benign (Dec 18, 2021)	criteria provided, multiple submitters, no conflicts



Clinvar y
dbSNP del
NCBI son
muy
importantes,
pero ahora
mismo la
base de
datos más
relevante es
gnomAD a
nivel
mundial

dbSNP SNP rs2228006 Create alert Advanced

Clinical Significance: benign Display Settings: Summary, Sorted by SNP_ID Send to:

Validation Status: by-ALFA, by-cluster, by-frequency

Publication: LitVar Annotated, PubMed Cited, PubMed Linked

Function Class: missense, non coding transcript variant

Annotation: somatic

Global MAF: Custom range...

[Clear all](#) [Show additional filters](#)

Search results
Items: 5

rs2228006 [*Homo sapiens*]
1. Variant type: SNV
Alleles: T>A,C,G [Show Flanks]
Chromosome: 7:5987144 (GRCh38)
7:6026775 (GRCh37)
Canonical SPDI: NC_000007.14:5987143:T:A,NC_000007.14:5987143:T:C,NC_000007.14:5987143:T:G
Gene: PMS2 ([Varview](#))
Functional Consequence:
non_coding_transcript_variant,stop_gained,missense_variant,coding_sequence_variant
Clinical significance: benign
Validated: by frequency,by alfa,by cluster
MAF: T=0.147268/27256 ([ALFA](#))
G=0./0 (KOREAN)
T=0.061941/1038 (TOMMO)
[...more](#)

HGVS: NC_000007.14:g.5987144T>A, NC_000007.14:g.5987144T>C,
NC_000007.14:g.5987144T>G, NC_000007.13:g.6026775T>A,
NC_000007.13:g.6026775T>C, NC_000007.13:g.6026775T>G, NG_008466.1:g.26963G>A,
NG_008466.1:g.26963G>C, NG_008466.1:g.26963G>T, NM_000525.7:l.4001A>T
[...more](#)

[PubMed](#) [LitVar](#)

Franklin es otra herramienta también gratuita parecida a varsome

VarSome logo | rs2228006 | hg19 | Search | Editions | About | Community | News | Demo | Sign in | Join

Your query results in several genomic alleles, click on each to see its data

chr7-6026775-T-A (PMS2:p.K541*) **chr7-6026775-T-C** (PMS2:p.K541E) chr7-6026775-T-G (PMS2:p.K541Q)

1 user classified this variant as Benign.

[Submit to ClinVar](#) [Link publication](#) [Classify](#) [Share](#) [API Link](#) [Favorites](#)

General Information SNV PMS2(NM_000535.7):c.1621A>G (p.Lys541Glu)	PharmGKB No data available	ACMG Classification Benign -21 points = 0 P - 21 B	Frequencies exomes: f = 0.841 (cov:93.7) genomes: f = 0.872 (cov:31.8)	Conservation Scores phyloP100: 1.443	Region Browser New	Structural Variants New
Genes PMS2	Publications View Variant: 5 Gene: 1422	ClinVar Benign ★☆☆☆ Submissions: 14	MitoMap No data available	Pathogenicity Scores 	Expression Data Top: Cells - EBV-transformed lymphocytes Tissues: 54	Beacon Network
Community Contributions Classifications: 1 Comments: 0	Transcripts NM_000535.7 - missense MANE Select	Uniprot Variants Benign	OMIM ® No data available	ClinGen No data available	GWAS No data available	Protein Viewer View

Variant [Explain](#)

Chromosome chr7	Position 6026775	REF Sequence View T	ALT Sequence View C	Variant type View SNV	Cytoband View 7p22.1-7p22.1	HGVS PMS2(NM_000535.7):c.1621A>G (p.Lys541Glu)	RS ID View rs2228006 dbSNP	Gene symbol View PMS2
--------------------	---------------------	--	--	--	--	--	---	---

[UCSC genome browser](#) [Mastermind](#) [TraP Score](#)

This variant has been viewed **1802** times on VarSome.

[Connect with past and future viewers of this variant...](#)

⚠ [VarSome.com](#) is for research use only. Find out about our clinically certified platform: [VarSome Clinical](#).



Location: 7:6,026,442-6,027,442

Variant: rs1805323

Variant displays

Explore this variant

- Genomic context
 - Genes and regulation
 - Flanking sequence
 - Population genetics
 - Phenotype data
 - Sample genotypes
 - Linkage disequilibrium
 - Phylogenetic context
 - Citations
 - 3D Protein model

[Configure this page](#)[Custom tracks](#)[Export data](#)[Share this page](#)[Bookmark this page](#)

rs1805323 SNP

Most severe consequence

missense variant | [See all predicted consequences](#)

Alleles

G/A/T | Ancestral: G | MAF: 0.11 (T) | Highest population MAF: 0.37

Change tolerance

CADD: A:0.015, T:0.001

Location

Chromosome 7:6026942 (forward strand) | VCF: 7 6026942 rs1805323 G A,T

Co-located variant

COSMIC COSV56219377

This variant has 108 HGVS names - [Show](#)This variant has 7 synonyms - [Show](#)

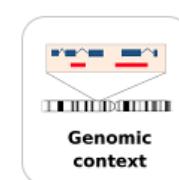
This variant has assays on: Illumina_HumanOmni5, Illumina_ExomeChip

Variants (including SNPs and indels) imported from dbSNP (release 154) | [View in dbSNP](#)

This variant overlaps 6 transcripts, has 2504 sample genotypes, is associated with 9 phenotypes and is mentioned in 17 citations.

modifies the age of onset of poly-glutamine (aka Poly-Q) diseases, such as Huntington disease and multiple spinocerebellar atrophy, according to

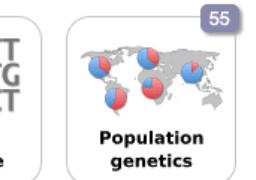
Explore this variant



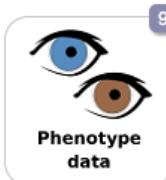
Genomic context



Genes and regulation



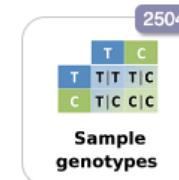
Flanking sequence



Population genetics



Phenotype data



Sample genotypes



Linkage disequilibrium



Phylogenetic context



Citations



3D Protein model

mirar sesión 7 1.18h

Using the website

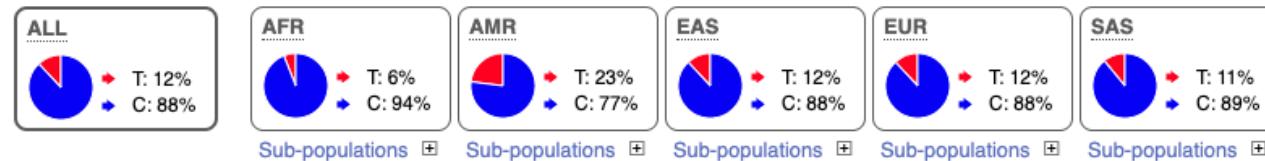
- Video: [Browsing SNPs and CNVs in Ensembl](#)
- Video: [Clip: Genome Variation](#)
- Video: [BioMart: Variation IDs to HGNC Symbols](#)

Programmatic access

- Tutorial: [Accessing variation data with the Variation API](#)

Reference materials

1000 Genomes Project Phase 3 allele frequencies



Jump to: [1000 Genomes Project Phase 3 \(32\)](#) | [gnomAD exomes \(9\)](#) | [gnomAD genomes \(8\)](#) | [GEM-J \(1\)](#) | [TOPMed \(1\)](#) | [UK10K \(2\)](#) | [NHLBI Exome Sequencing Project \(2\)](#)

1000 Genomes Project Phase 3 (32) ☰

Population	Allele: frequency (count)	Genotype: frequency (count)	Genotypes
ALL	T: 0.117 (585) C: 0.883 (4423)	TIT: 0.020 (51) CIC: 0.787 (1970) CIT: 0.193 (483)	Show
AFR	T: 0.058 (77) C: 0.942 (1245)	TIT: 0.002 (1) CIC: 0.885 (585) CIT: 0.113 (75)	Show
ACB	T: 0.047 (9) C: 0.953 (183)	TIT: 0.010 (1) CIC: 0.917 (88) CIT: 0.073 (7)	Show
ASW	T: 0.082 (10) C: 0.918 (112)	CIC: 0.836 (51) CIT: 0.164 (10)	Show
ESN	T: 0.040 (8) C: 0.960 (190)	CIC: 0.919 (91) CIT: 0.081 (8)	Show
GWD	T: 0.075 (17) C: 0.925 (209)	CIC: 0.850 (96) CIT: 0.150 (17)	Show
LWK	T: 0.086 (17) C: 0.914 (181)	CIC: 0.828 (82) CIT: 0.172 (17)	Show
MSL	T: 0.035 (6) C: 0.965 (164)	CIC: 0.929 (79) CIT: 0.071 (6)	Show
YRI	T: 0.046 (10) C: 0.954 (206)	CIC: 0.907 (98) CIT: 0.093 (10)	Show
AMR	T: 0.226 (157) C: 0.774 (537)	TIT: 0.061 (21) CIC: 0.608 (211) CIT: 0.331 (115)	Show
CLM	T: 0.191 (36) C: 0.809 (152)	TIT: 0.043 (4) CIC: 0.660 (62) CIT: 0.298 (28)	Show
MXL	T: 0.328 (42) C: 0.672 (86)	TIT: 0.141 (9) CIC: 0.484 (31) CIT: 0.375 (24)	Show
PEL	T: 0.265 (45) C: 0.735 (125)	TIT: 0.071 (6) CIC: 0.541 (46) CIT: 0.388 (33)	Show
PUR	T: 0.163 (34) C: 0.837 (174)	TIT: 0.019 (2) CIC: 0.692 (72) CIT: 0.288 (30)	Show
EAS	T: 0.122 (123) C: 0.878 (885)	TIT: 0.024 (12) CIC: 0.780 (393) CIT: 0.196 (99)	Show
CDX	T: 0.172 (32) C: 0.828 (154)	TIT: 0.054 (5) CIC: 0.710 (66) CIT: 0.237 (22)	Show
CHB	T: 0.073 (15) C: 0.927 (191)	TIT: 0.019 (2) CIC: 0.874 (90) CIT: 0.107 (11)	Show
CHS	T: 0.124 (26) C: 0.876 (184)	CIC: 0.752 (79) CIT: 0.248 (26)	Show
JPT	T: 0.053 (11) C: 0.947 (197)	CIC: 0.894 (93) CIT: 0.106 (11)	Show
KHV	T: 0.197 (39) C: 0.803 (159)	TIT: 0.051 (5) CIC: 0.657 (65) CIT: 0.293 (29)	Show
FIN	T: 0.123 (124) C: 0.877 (882)	TIT: 0.018 (9) CIC: 0.771 (388) CIT: 0.211 (106)	Show

Repite el ejercicio para las siguientes variantes:

Chr	Posición	Gen	Ref	Alt	Exón	Tipo	HOM/HET	Qual	Depth	dbSNP	PolyPhen	EUR AF	ClinVar	Varsome	ClinVar Condition
7	6026775	PMS2	T	C	11	SNP	HOM	10793,7	338	rs2228006	0	0,8767	bening	bening	Lynch syndrome
7	6026942	PMS2	G	T	11	SNP	HET	4488,92	366						
7	6036980	PMS2	G	C	7	SNP	HOM	7398,34	235						
7	55229255	EGFR	G	A	13	SNP	HET	4531,82	326						
7	55249063	EGFR	G	A	19	SNP	HET	3957,84	321						
											0 = bening				

Chr	Posición	Gen	Ref	Alt	Exón	Tipo	HOM/HET	Qual	Depth	dbSNP	PolyPhen	EUR AF	ClinVar	Varsome	ClinVar Condition
7	6026775	PMS2	T	C	11	SNP	HOM	10793,7	338	rs2228006	0	0,8767	benign	benign	Lynch syndrome
7	6026942	PMS2	G	T	11	SNP	HET	4488,92	366	rs1805323	0	0,0457	benign	benign	Lynch syndrome
7	6036980	PMS2	G	C	7	SNP	HOM	7398,34	235	rs1805319	-	0,836	benign	benign	Lynch syndrome
7	55229255	EGFR	G	A	13	SNP	HET	4531,82	326	rs2227983	0	0,2763	benign	benign	Lung cancer
7	55249063	EGFR	G	A	19	SNP	HET	3957,84	321	rs1050171	-	0,6074	benign	benign	Lung cancer
											0 = benign				

Del resultado de VEP, ¿Qué variante/s es/son patogénicas?

Navigation (per variant)

Show: [1](#) [5](#) [10](#) [50](#) [All](#) variants

Filters

Clinical significance is pathogenic

Uploaded variant defined

Download

All: [VCF](#) [VEP](#) [TXT](#)

Filtered: [VCF](#) [VEP](#) [TXT](#)

BioMart: [Variants](#) [Genes](#)

Show/hide columns (12 hidden)

Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon	Intron	cDNA position	CDS position	Protein position	Amino acids	Code	Score
7:117188877- T 117188877		T	missense_variant, splice_region_variant	MODERATE	CFTR	ENSG00000001626	Transcript	ENST00000003084.6	protein_coding	10/27	-	1524	1392	464	K/N	AAG/AAT	rs397508198 , CM034784 , COSV50041004
7:117188877- T 117188877		T	missense_variant, splice_region_variant	MODERATE	CFTR	ENSG00000001626	Transcript	ENST00000426809.1	protein_coding	9/26	-	1302	1302	434	K/N	AAG/AAT	rs397508198 , CM034784 , COSV50041004
7:117188877- T 117188877		T	intron_variant	MODIFIER	CFTR	ENSG00000001626	Transcript	ENST00000454343.1	protein_coding	-	9/25	-	-	-	-	-	rs397508198 , CM034784 , COSV50041004

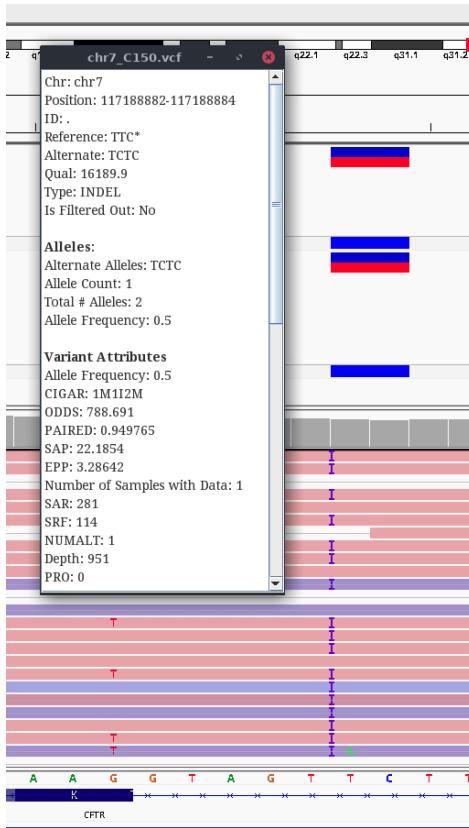
(12 hidden)

Amino acids	Codons	Existing variant	Feature strand	Symbol source	ENSP	SIFT	PolyPhen	AF	AFR	AMR	EAS	EUR	SAS	Clinical significance	Somatic status	Phenotype or disease	Pubmed	Associated phenotypes
K/N	AAG/AAT	rs397508198 , CM034784 , COSV50041004	1	HGNC	ENSP00000003084	0	0.985	-	-	-	-	-	-	pathogenic	0, 0, 1	1, 1, 1	-	Cystic Fibrosis (rs397508198 , ClinVar) Stage 5 chronic kidney disease (rs397508198 , ClinVar)
K/N	AAG/AAT	rs397508198 , CM034784 , COSV50041004	1	HGNC	ENSP00000389119	0	0.985	-	-	-	-	-	-	pathogenic	0, 0, 1	1, 1, 1	-	Cystic Fibrosis (rs397508198 , ClinVar) Stage 5 chronic kidney disease (rs397508198 , ClinVar)
-	-	rs397508198 , CM034784 , COSV50041004	1	HGNC	ENSP00000403677	-	-	-	-	-	-	-	-	pathogenic	0, 0, 1	1, 1, 1	-	Cystic Fibrosis (rs397508198 , ClinVar) Stage 5 chronic kidney disease (rs397508198 , ClinVar)



Chr	Posición	Gen	Ref	Alt	Exón	Tipo	HOM/HET	Qual	Depth	dbSNP	PolyPhen	EUR AF	ClinVar	Varsome	ClinVar Condition
7	117188877	CFTR	G	T	10	SNP	HET	1563,85	1065	rs397508198	0,985 probably damaging	(ExAC = 0,00001/1; 1000Genomes = 0,0002/1)	pathogenic	uncertain significance (ACMG)	Cystic fibrosis, Hereditary pancreatitis

Chr	Posición	Gen	Ref	Alt	Exón	Tipo	HOM/HET	Qual	Depth
7	117188882-117188884	CFTR	TCC	TCTC	-	INDEL	HET	16189,9	951



Navigation (per variant)		Filters		Download		New job										
Show: 1 5 10 50 All variants		<input type="text"/> Location is 7:11718882-11718884 Filter		VCF VEP TXT												
		<input type="text"/> Symbol is CFTR Filter		VCF VEP TXT												
Clear filters				BioMart: Variants Genes												
		Uploaded variant Select is defined Add														
Show/hide columns (12 hidden) Expand																
Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon	Intron	cDNA position	CDS position	Protein position	Amino acids	Scalable column headers
7:11718882-11718884	CTC	splice_donor_region_variant intron_variant	LOW		CFTR	ENSG00000001626	Transcript	ENST00000003084.6	protein_coding	-	10/26	-	-	-	-	rs1474892453 1
7:11718882-11718884	CTC	splice_donor_region_variant intron_variant	LOW		CFTR	ENSG00000001626	Transcript	ENST00000426809.1	protein_coding	-	9/25	-	-	-	-	rs1474892453 1
7:11718882-11718884	CTC	intron_variant		MODIFIER	CFTR	ENSG00000001626	Transcript	ENST00000454343.1	protein_coding	-	9/25	-	-	-	-	rs1474892453 1

Chr	Posición	Gen	Ref	Alt	Exón	Tipo	HOM/HET	Qual	Depth	dbSNP	PolyPhen	EUR AF	ClinVar	Varsome	ClinVar Condition
7	117188882-117188884	CFTR	TCC	TCTC	-	INDEL	HET	16189,9	951	rs1474892453	-	-	not reported	uncertain significance (ACMG)	-

4.4.2. Annovar

Annovar (<https://annovar.openbioinformatics.org/en/latest/>) es una herramienta software de línea de comandos gratuita para la anotación funcional de varios genomas estándar. De esta forma, suministrada una lista de variantes (cromosoma, posición, referencia y nucleótido alternativo), puede darnos la información a tres niveles principales:

Anotaciones basadas en gen: identifica si una variante de tipo SNP, indel o CNV causa cambios en la proteína codificada y los aminoácidos que están afectados. En este paso puede elegirse utilizar las bases de datos de genes como RefSeq, UCSC, ENSEMBL o GENCODE, entre otras.

Anotaciones basadas en región: identifica variantes en regiones genómicas específicas, por ejemplo, regiones conservadas entre especies, lugares de unión de factores de transcripción, etc.

Anotaciones basadas en filtros: identifica variantes que están previamente descritas en bases de datos como 1000 genomas, 6500 exomas, Exome Aggregation Consortium (ExAC), etc.

Otras funcionalidades: identificar una lista de genes candidatos para enfermedades mendelianas a partir de datos de exoma.

4.4.2. Annotar en versión web: wANNOVAR



Tema 4 - Ejemplo 12 - Anotación wANNOVAR

De manera alternativa al ejemplo anterior, podemos utilizar Annotar en su versión web (wANNOVAR) para la anotación funcional a nivel de gen, región o basadas en varias bases de datos como 1000 genomes, 6500 genomes, ExAC...

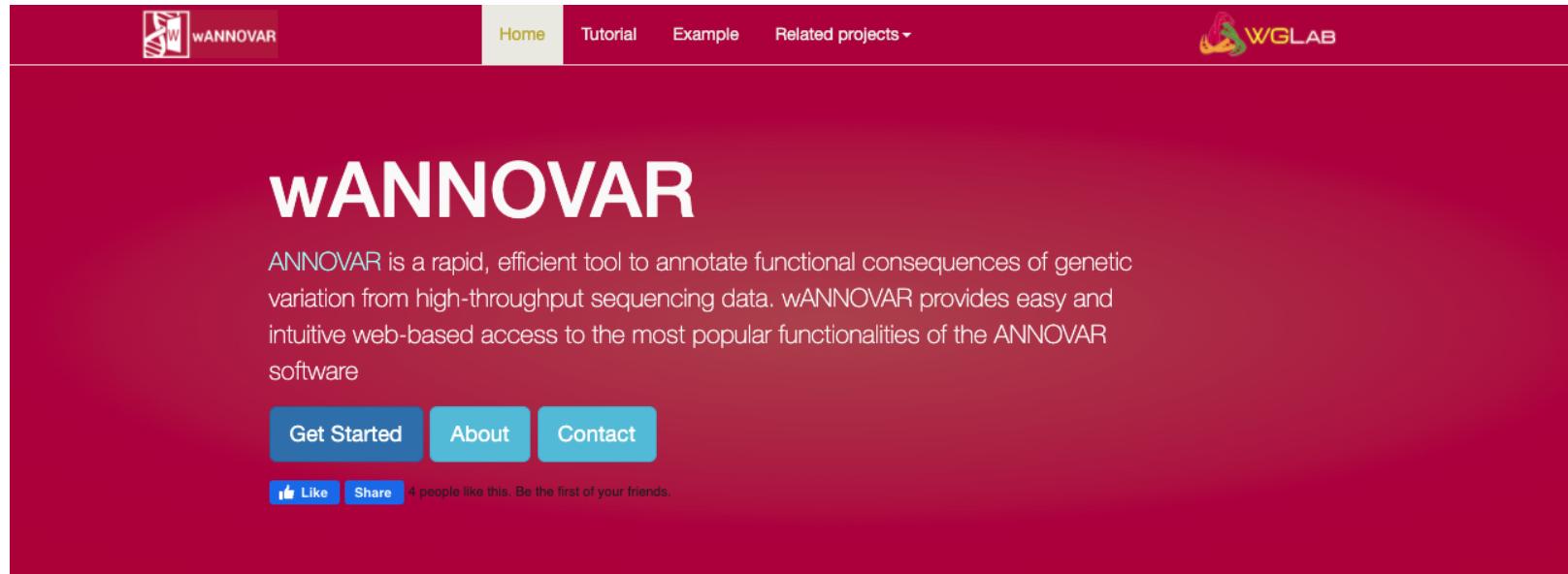
Web: <https://wannovar.wglab.org/>

Archivos: [Tema4_Ejemplo11_VEP](#)

Debemos llenar:

- email al que queremos que nos llegue el link con el resultado
- identificador de la muestra
- archivo input, que es el archivo VCF a analizar (chr7.vcf)
- opcionalmente podemos indicar la enfermedad o término de fenotipo sobre el que queremos centrar el estudio
- parámetros, entre los que destaca el genoma de referencia (hg19 o hg38), formato input (VCF, en nuestro caso), definición del gen (RefSeq, UCSC o ENSEMBL, nosotros elegiremos RefSeq), modelo de enfermedad (si lo sabemos o queremos afinar sobre ello).

4.4.2. Annotar en versión web: wANNOVAR



By default, wANNOVAR performs "individual analysis" on the first sample in your VCF file to help find disease genes (you may need to split your multi-sample VCF file to individual files for annotation separately to find disease genes). If you only want to annotate all variant sites in a multi-sample VCF file, select "All Annotations" option below.

Basic Information

Email	<input type="text" value="Email"/>
Sample Identifier	<input type="text" value="Sample Identifier"/>
Input File	<input type="button" value="+ Input File"/>

Recent Updates

[5/9/2019] The wANNOVAR server is migrated to a new host. All submissions will be deleted **within 1 day (new rule in June 2019)** due to lack of storage space. Please download your results promptly.

[10/19/2017] The detailed amino acid changes for indels are now included in the output (through -polish argument in table_annovar). The server also handles duplicated

By default, wANNOVAR performs "individual analysis" on the first sample in your VCF file to help find disease genes (you may need to split your multi-sample VCF file to individual files for annotation separately to find disease genes). If you only want to annotate all variant sites in a multi-sample VCF file, select "All Annotations" option below.

If you see an error message "cannot create submission directory for submission: no space left", it means our server's storage space is full. You can wait 1-2 days to submit again, or you can email Kai to resolve the issue quickly.

Basic Information

Email

maria.detoro@campusviu.es

Sample Identifier

S8_chr7

Input File

+ Input File

chr7.vcf

or Paste Variant Calls

paste your variant call here

Submit

Reset

Monitor Progress

I agree to the [Terms of Use](#). Please note that commercial users would need to obtain a license.

Disease/Phenotype

Enter Disease or Phenotype Terms

please enter your focused disease/phenotype terms

Please use semicolon or enter as separators. Like "alzheimer;brain".
Try to use multiple terms instead of a super long term
OMIM IDs are also accepted, like 114480 for 'Breast cancer'
Better Combined with wANNOVAR's disease model.

Parameter Settings

Result duration

1 day



Reference Genome

hg19



Input Format

VCF



Gene Definition

RefSeq Gene



Individual analysis

All annotations



Disease Model

none



Submission 394757

Your submission has been received by the ANNOVAR server at Thu Dec 29 12:41:26 2022.

The results will be generated at </done/394757/JvizdjwtjtYpt3d/index.html> after the computation is done.

WARNING WARNING WARNING: Many email servers nowadays block suspicious emails with URL inside, so you may NOT receive an email, and all results will be deleted every 24 hours. Therefore, it is best that you record the URL for results (such as bookmarking it), or keep the browser open until results are shown.

A summary of input file

User input contains 577 lines

All Rights Reserved @Wang Genomics Lab 2010-2019

Submission 2023/01/19 (chr7.150.vcf)

: https://wannovar.wglab.org/done/397761/E2_3mNAqtGRm7b7i/index.html

Submission ID: 397761

Sample identifier = S8_chr7_C150
File_name=chr7.150.vcf
File_format=vcf4
Reference_genome=hg19
Disease_model=no filtering
Processed variants=73

Basic Information

exome summary results	view	CSV file	TXT file
genome summary results	view	CSV file	TXT file

Submission ID: 394757

Sample identifier = S8_chr7
File_name=chr7.vcf
File_format=vcf4
Reference_genome=hg19
Disease_model=no filtering
Processed variants=539

Basic Information

exome summary results	view	CSV file	TXT file
genome summary results	view	CSV file	TXT file

100 results per page

[back to HOME](#)

Page: 1

Chr	Start	End	Ref	Alt	Func	Gene	GeneDetail	ExonicFunc	AAChange
7	6026775	6026775	T	C	exonic	PMS2		nonsynonymous SNV	PMS2:NM_001322008:exon9:c.A1303G:p.K435E,PMS2:NM_001322010:exon9:c.A1060G:p.K354E,PMS2:NM_001322004:exon10:c.A1216G:p.
7	6026942	6026942	G	T	exonic	PMS2		nonsynonymous SNV	PMS2:NM_001322008:exon9:c.C1136A:p.T379K,PMS2:NM_001322010:exon9:c.C893A:p.T298K,PMS2:NM_001322004:exon10:c.C1049A:p.
7	6036980	6036980	G	C	exonic	PMS2		synonymous SNV	PMS2:NM_001322008:exon5:c.C462G:p.S154S,PMS2:NM_001322004:exon6:c.C375G:p.S125S,PMS2:NM_001322007:exon6:c.C462G:p.S154S
7	55229255	55229255	G	A	exonic	EGFR		nonsynonymous SNV	EGFR:NM_001346941:exon7:c.G761A:p.R254K,EGFR:NM_001346897:exon12:c.G1427A:p.R476K,EGFR:NM_001346899:exon12:c.G1427A:p.R476K
7	55233089	55233089	C	T	exonic	EGFR		synonymous SNV	EGFR:NM_001346941:exon9:c.C1038T:p.A346A,EGFR:NM_001346897:exon14:c.C1704T:p.A568A,EGFR:NM_001346899:exon14:c.C1704T:p.A568A
7	55238087	55238087	C	T	exonic	EGFR		synonymous SNV	EGFR:NM_201284:exon16:c.C1968T:p.H656H
7	55238874	55238874	T	A	exonic	EGFR		synonymous SNV	EGFR:NM_001346941:exon10:c.T1086A:p.T362T,EGFR:NM_001346897:exon15:c.T1752A:p.T584T,EGFR:NM_001346899:exon15:c.T1752A:p.T584T
7	55249063	55249063	G	A	exonic	EGFR		synonymous SNV	EGFR:NM_001346941:exon14:c.G1560A:p.Q520Q,EGFR:NM_001346897:exon19:c.G2226A:p.Q742Q,EGFR:NM_001346899:exon19:c.G2226A:p.Q742Q
7	55266417	55266417	T	C	exonic	EGFR		synonymous SNV	EGFR:NM_001346941:exon17:c.T1908C:p.T636T,EGFR:NM_001346897:exon22:c.T2574C:p.T858T,EGFR:NM_001346899:exon22:c.T2574C:p.T858T
7	55268916	55268916	C	T	exonic	EGFR		synonymous SNV	EGFR:NM_001346941:exon19:c.C2181T:p.D727D,EGFR:NM_001346897:exon24:c.C2847T:p.D949D,EGFR:NM_001346899:exon24:c.C2847T:p.D949D
7	116435768	116435768	C	T	exonic	MET		synonymous SNV	MET:NM_001324402:exon19:c.C2568T:p.D856D,MET:NM_000245:exon20:c.C3858T:p.D1286D,MET:NM_001127500:exon20:c.C3912T:p.D1286D
7	116436022	116436022	G	A	exonic	MET		synonymous SNV	MET:NM_001324402:exon20:c.G2727A:p.A909A,MET:NM_000245:exon21:c.G4017A:p.A1339A,MET:NM_001127500:exon21:c.G4071A:p.A1339A
7	116436097	116436097	G	A	exonic	MET		synonymous SNV	MET:NM_001324402:exon20:c.G2802A:p.P934P,MET:NM_000245:exon21:c.G4092A:p.P1364P,MET:NM_001127500:exon21:c.G4146A:p.P1364P
7	117188717	117188721	AAGCA	-	exonic	CFTR		frameshift deletion	CFTR:NM_000492:exon10:c.1232_1236del:p.A412Tfs*4
7	117188726	117188726	A	C	exonic	CFTR		nonsynonymous SNV	CFTR:NM_000492:exon10:c.A1241C:p.Q414P
7	117188736	117188736	C	A	exonic	CFTR		nonsynonymous SNV	CFTR:NM_000492:exon10:c.C1251A:p.N417K
7	117188739	117188739	T	C	exonic	CFTR		synonymous SNV	CFTR:NM_000492:exon10:c.T1254C:p.N418N
7	117188750	117188750	C	T	exonic	CFTR		nonsynonymous SNV	CFTR:NM_000492:exon10:c.C1265T:p.S422F
7	117188763	117188763	C	T	exonic	CFTR		synonymous SNV	CFTR:NM_000492:exon10:c.C1278T:p.D426D
7	117188797	117188797	A	G	exonic	CFTR		nonsynonymous SNV	CFTR:NM_000492:exon10:c.A1312G:p.T438A
7	117188816	117188819	TTAA	-	exonic	CFTR		frameshift deletion	CFTR:NM_000492:exon10:c.1331_1334del:p.N445Sfs*3
7	117188842	117188842	T	C	exonic	CFTR		synonymous SNV	CFTR:NM_000492:exon10:c.T1357C:p.L453L
7	117188849	117188849	C	T	exonic	CFTR		nonsynonymous SNV	CFTR:NM_000492:exon10:c.C1364T:p.A455V
7	117188850	117188850	G	T	exonic	CFTR		synonymous SNV	CFTR:NM_000492:exon10:c.G1365T:p.A455A
7	117188877	117188877	G	T	exonic	CFTR		nonsynonymous SNV	CFTR:NM_000492:exon10:c.G1392T:p.K464N
7	117199533	117199533	G	A	exonic	CFTR		nonsynonymous SNV	CFTR:NM_000492:exon11:c.G1408A:p.V470M
7	117234999	117234999	G	T	exonic	CFTR		nonsynonymous SNV	CFTR:NM_000492:exon15:c.G2506T:p.D836Y
7	148504762	148504762	G	A	exonic	EZH2		synonymous SNV	EZH2:NM_001203249:exon19:c.C2064T:p.I688I,EZH2:NM_152998:exon19:c.C2100T:p.I700I,EZH2:NM_001203247:exon20:c.C2217T:p.I700I
7	148504768	148504768	G	A	exonic	EZH2		synonymous SNV	EZH2:NM_001203249:exon19:c.C2058T:p.V686V,EZH2:NM_152998:exon19:c.C2094T:p.V698V,EZH2:NM_001203247:exon20:c.C2211T:p.I700I
7	148506462	148506462	G	A	exonic	EZH2		nonsynonymous SNV	EZH2:NM_001203249:exon17:c.C1882T:p.R628C,EZH2:NM_152998:exon17:c.C1918T:p.R640C,EZH2:NM_001203247:exon18:c.C2035T:p.V698V

Page: 1

100 results per page

[back to HOME](#)

Page: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#)

Chr	Start	End	Ref	Alt	Func	Gene	GeneDetail
7	10081	10083	AAC	-	intergenic	NONE;LOC102723672	dist=NONE;dist=134362
7	10116	10116	A	C	intergenic	NONE;LOC102723672	dist=NONE;dist=134329
7	3838128	3838130	TGT	CGC	intronic	SDK1	
7	6022629	6022629	G	A	intronic	PMS2	
7	6026384	6026384	C	T	intronic	PMS2	
7	6026775	6026775	T	C	exonic	PMS2	
7	6026942	6026942	G	T	exonic	PMS2	
7	6036980	6036980	G	C	exonic	PMS2	
7	6037058	6037058	A	-	intronic	PMS2	
7	6038621	6038621	A	-	intronic	PMS2	
7	6043319	6043319	A	G	splicing	PMS2	NM_001322010:exon3:UTR5;NM_001322009:exon4:UTR5;NM_001322006:exon4:c.353+2T>C;NM_001322005:exon4:UTR5;NM_001322004:exon3:UTR
7	6048538	6048538	G	T	UTR5	PMS2	NM_001322009:c.-6323C>A;NM_001322005:c.-6323C>A
7	6776956	6776956	A	G	ncRNA_exonic	PMS2CL	
7	6777369	6777369	T	C	ncRNA_exonic	PMS2CL	
7	8327283	8327287	CTTAC	TTTAT	ncRNA_intronic	LOC100505938	
7	9934234	9934237	ACTA	TCT	intergenic	PER4;MGC4859	dist=258787;dist=555210
7	9934250	9934250	T	C	intergenic	PER4;MGC4859	dist=258803;dist=555197
7	9989632	9989632	A	C	intergenic	PER4;MGC4859	dist=314185;dist=499815
7	14799217	14799217	C	T	intronic	DGKB	
7	16862736	16862736	C	A	intergenic	AGR2;AGR3	dist=17998;dist=36294
7	16862741	16862741	G	T	intergenic	AGR2;AGR3	dist=18003;dist=36289
7	20042353	20042353	C	T	ncRNA_intronic	LOC101927668	
7	20042358	20042358	G	A	ncRNA_intronic	LOC101927668	
7	20042363	20042363	C	T	ncRNA_intronic	LOC101927668	
7	20042364	20042364	G	A	ncRNA_intronic	LOC101927668	
7	20042369	20042369	C	T	ncRNA_intronic	LOC101927668	
7	20042372	20042372	C	T	ncRNA_intronic	LOC101927668	
7	20042373	20042373	C	T	ncRNA_intronic	LOC101927668	
7	20042381	20042381	C	T	ncRNA_intronic	LOC101927668	
7	20042382	20042382	G	A	ncRNA_intronic	LOC101927668	
7	20042392	20042392	C	T	ncRNA_intronic	LOC101927668	
7	20042393	20042393	G	A	ncRNA_intronic	LOC101927668	
7	20042421	20042421	C	T	ncRNA_intronic	LOC101927668	
7	20042422	20042422	G	A	ncRNA_intronic	LOC101927668	
7	20042424	20042424	T	C	ncRNA_intronic	LOC101927668	

El proceso de anotación de variantes es muy amplio, por lo que existen varios artículos en literatura con revisiones profundas sobre herramientas disponibles, que nos ayudan a guiarnos en el proceso de anotación y correcta interpretación de las variantes. A continuación, se proporcionan los enlaces a algunos de estos artículos:

(Cordero & Ashley, 2012): <https://pubmed.ncbi.nlm.nih.gov/22549284/>

(Pabinger et al., 2014): <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3956068/>

(Richards et al., 2015): <https://pubmed.ncbi.nlm.nih.gov/25741868/>

(Seaby et al., 2016): <https://academic.oup.com/bfg/article/15/5/374/2240049>

(M. M. Li et al., 2017): <https://pubmed.ncbi.nlm.nih.gov/27993330/>

<https://www.acgs.uk.com/media/11631/uk-practice-guidelines-for-variant-classification-v4-01-2020.pdf>

Base de datos	Web
VEP	https://www.ensembl.org/info/docs/tools/vep/index.html
wAnnotar	https://annovar.openbioinformatics.org/en/latest/
ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/ ; https://www.youtube.com/watch?v=A8G3ej83ZgU&feature=youtu.be
dbSNP	https://www.ncbi.nlm.nih.gov/snp/
Varsome	https://varsome.com/
OMIM	https://www.omim.org/
GeneCards	https://www.genecards.org/ [suite de herramientas]
GeneReviews	https://www.ncbi.nlm.nih.gov/books/NBK1116/
ClinGen	https://clinicalgenome.org/
HGMD	https://www.hgmd.cf.ac.uk/ac/index.php

¡Gracias!

The logo consists of the lowercase letters "viu" in white, centered within a dark orange, rounded rectangular shape.

viu

Universidad
Internacional
de Valencia

universidadviu.com

De:
 Planeta Formación y Universidades