Actividad 1 – 1a convocatoria.

Edición abril 2024

Introducción a la actividad

Esta actividad corresponde al ejercicio de evaluación del Tema 4. Se proporcionan los datos crudos de secuenciación Illumina pareados (*paired*-end) de un panel de genes de captura sobre genoma humano Hg19. El ejercicio debe realizarse dentro de lo establecido en clase. Para ello debéis hacer uso del entorno de trabajo proporcionado 04MBIF_humano, utilizado en clase, y que contiene todos los programas necesarios para la ejecución.

Nota: recordad que los entornos están disponibles en el apartado Recursos y materiales / 01. Materiales docentes / Environments Conda de la web de la asignatura.

Link de descarga de datos

El link de descarga para las lecturas y el archivo de sondas BED, se encuentra en el link siguiente:

https://alumnosviu-my.sharepoint.com/:f:/g/personal/laura_gutierrez_m_professor_universidadviu_com/EtVR89Mx7bFPif7wKDDhQIYBHLq9M7e1cjunzMiOADQohA?e=c20MWe

Los archivos proporcionados son:

- **Lecturas crudas directas del secuenciador, en formato FASTQ correspondientes a la muestra llamada S15**
- **Archivo en formato BED de las sondas utilizadas para la captura de este panel de genes.**

Objetivo

Realizar el análisis completo de esta muestra, incluyendo los pasos siguientes:

1. **Realiza el análisis de calidad de las lecturas iniciales.**
2. **Realiza la limpieza de calidad de las lecturas, junto con el análisis de calidad de las lecturas resultantes**
3. **Descargar el genoma de referencia, prepararlo y mapear las lecturas limpias al genoma de referencia**
4. **Procesamiento de los archivos de mapeo y estadísticas básicas de mapeo.**
5. **Post-procesamiento del mapeo. Eliminación de duplicados y extracción de variantes en formato VCF.**
6. **Análisis VEP del archivo VCF (recordad que el genoma de referencia corresponde a Hg19/GRCh37 y usad todos los parámetros tal y como se especifica en este ejercicio, excluyendo las variantes comunes al filtrar por frecuencia)**
7. **Visualización en IGV del análisis realizado.**

Formato de entrega

Importante: Leed atentamente y contestad a lo que se pregunta

1. **Toma las lecturas iniciales y realiza un análisis de calidad de las mismas. Contesta a las siguientes preguntas (1 punto):**
    a. **¿Cuántas lecturas se han secuenciado?**
    Se han secuenciado 1382239 lecturas.

    b. **¿Qué longitud de secuencia tienes?**
    La longitud de las secuencias está entre 35 y 151 pb.

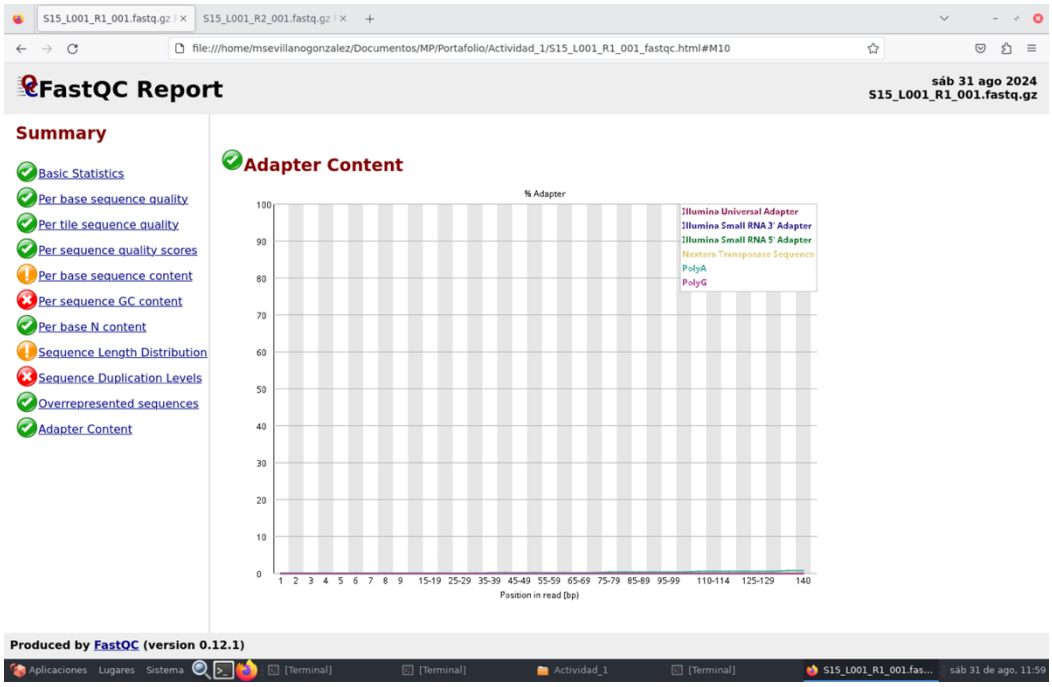
    c. **Incluye el/los comandos utilizados para obtener la información**
    Comando utilizado para obtener las lecturas de cada archivo:

        zgrep -c '@M02899:19' *fastq.gz

    Para realizar el análisis de calidad de ambos archivos .fastq.gz a la vez:

        fastqc *fastq.gz

    d. **Incluye el pantallazo de la gráfica "Adapter Content" obtenida para las lecturas R1 e indica si consideras necesario realizar la limpieza de adaptadores.**



Dado que parece que hay una pequeña cantidad de polyA, y que Fastq no muestra todos los adaptadores posibles, por lo que, que no aparezcan no significa que no haya, es recomendable siempre hacer una limpieza de adaptadores, aunque reamente la cantidad que aparece es ínfima y no sería tan necesario.

2. **Realiza la limpieza de calidad de las lecturas iniciales, junto con el análisis de calidad de las lecturas resultantes (1 punto). Parámetros para considerar: longitud mínima 100 bp, calidad media, inicial y final de lectura = Q20. Entrega:**
    a. **Comando/s necesarios para realizar la tarea.**
    El comando utilizado para el filtrado de calidad de las secuencias y la eliminación de adaptadores usando la herramienta fastp es:

        fastp -i S15_L001_R1_001.fastq.gz -I S15_L001_R2_001.fastq.gz -o S15_R1clean.fq.gz -O S15_R2clean.fq.gz --cut_front 20 --cut_tail 20 --cut_mean_quality 20 --detect_adapter_for_pe --trim_poly_g --trim_poly_x -l 100 -h report_fastp.html

    Comando utilizado para hacer el análisis de calidad de las lecturas resultantes del filtrado (de

ambos archivos)

fastqc *clean.fq.gz

b. ¿Cuántas lecturas finales han pasado filtros? ¿Cuántas lecturas se han eliminado por ser demasiado cortas? ¿Y por baja calidad?

Lecturas que han pasado filtros: 1954540

Lecturas eliminadas por ser demasiado cortas: 793198

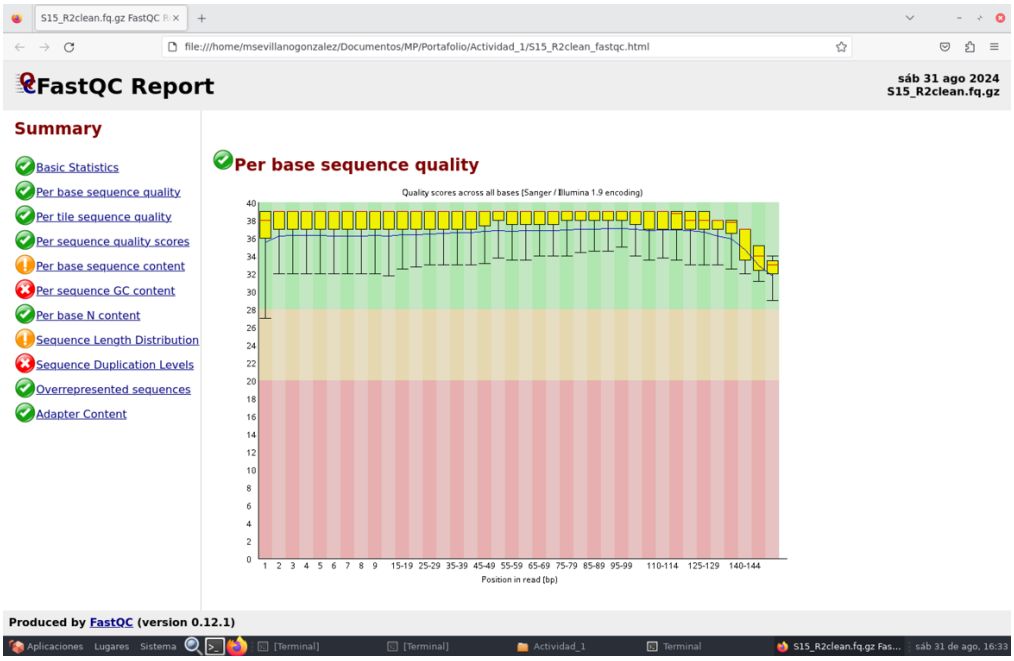
Lecturas eliminadas por baja calidad: 16740

c. Indica cuál es el tamaño medio de las lecturas resultantes R1 y R2, así como el porcentaje de duplicación.

El tamaño medio de las lecturas es entre 100 y 151 pb.

Porcentaje de duplicación: 18.8941%

d. Incluye el pantallazo de la gráfica "Per base sequence quality" de las lecturas R2 de este apartado.



3. **Descargar el genoma de referencia, prepararlo y mapear las lecturas limpias al genoma de referencia. Descarga el cromosoma 7 del genoma de referencia GRCh37/Hg19 de la base de datos ENSEMBL (https://grch37.ensembl.org/Homo_sapiens/Info/Index). Prepara esta secuencia para su utilización. Mapea las lecturas apropiadas obtenidas en apartados anteriores a este cromosoma de referencia (1 punto).**

a. Comando/s necesarios para realizar este apartado.

Para realizar el mapeo de las lecturas obtenidas frente al genoma de referencia primero hay que descomprimir el archivo fasta del cromosoma 7 del genoma de referencia a usar, con el comando:

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

Después se indexa el genoma de referencia con el comando:

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

Por último, el comando utilizado para mapear las lecturas, ya filtradas y limpiadas, al cromosoma 7 del genoma de referencia es:

bwa mem -t 2 -a Homo_sapiens.GRCh37.dna.chromosome.7.fa trimming/S15_R1clean.fq.gz trimming/S15_R2clean.fq.gz -o chr7.sam 2> chr7.out

4. **Procesamiento de los archivos de mapeo y estadísticas básicas de mapeo. Prepara los archivos para los pasos posteriores de análisis y procesamiento (1 punto).**
    a. **Indica el/los comandos que utilizas para conversión del archivo a formato BAM, así como para adecuarlo a los análisis posteriores de uso y visualización.**
    Primero hay que convertir el archivo sam a archivo bam mediante la herramienta samtools utilizando el comando:

        samtools view -bS chr7.sam > chr7.bam

    Después se ordena por coordenadas el archivo bam con el comando:

        samtools sort chr7.bam > chr7.sorted.bam

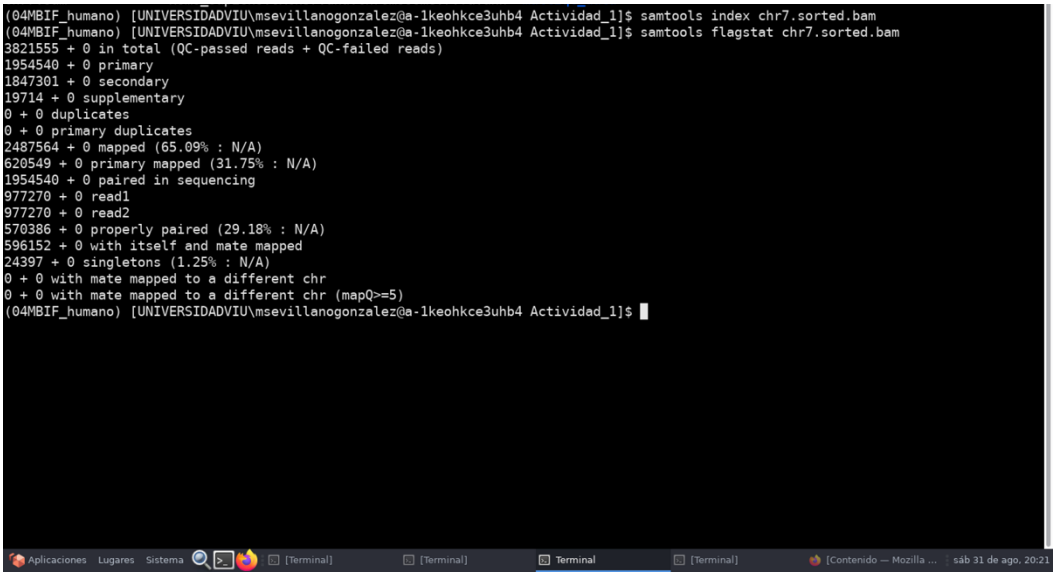
    Y, por último, se indexa el archivo bam ordenado antes de realizar el análisis de calidad del mapeo, utilizando el comando:

        samtools index chr7.sorted.bam

    b. **Realiza un análisis de calidad del mapeo de manera sencilla utilizando "samtools flagstat".**
    Indica el comando necesario. Muestra la salida que obtienes y contesta a estas preguntas: ¿Cuál es el porcentaje de mapeo global? ¿Cuál el de mapeo primario? ¿Y el de singletones?

    Comando utilizado para realizar el análisis de calidad del mapeo:

        samtools flagstat chr7.sorted.bam



    Porcentaje del mapeo global: 65,09%
    Porcentaje de mapeo primario: 31,75%
    Porcentaje de singletones: 1,25%

5. **Post-procesamiento del mapeo. Eliminación de duplicados y extracción de variantes en formato VCF. Nota: para la llamada de variantes utiliza el parámetro C=50. Contesta a las siguientes preguntas (1,5 puntos).**
    a. **Indica el/los comandos necesarios para realizar el marcaje de duplicados y la adición de Read Groups. Explica qué es el RGID y cómo podemos saber su valor.**

    Comando para marcar los duplicados:

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

    Comando para añadir los Read Groups:

        picard AddOrReplaceReadGroups -I chr7.dedup.bam -O chr7.dedup.RG.bam -RGID M02899 -RGLB run19 -RGPL ILLUMINA -RGPU unit1 -RGSM S15

Comando para indexer el archivo bam de nuevo:

samtools index chr7.dedup.RG.bam

El RGID, o Read Group ID, es un identificador único que se asocia a cada grupo de lecturas y nos indica el nombre del secuenciador del que provienen. Se puede obtener de la cabecera del archivo fastq.

b. Indica el/los comandos necesarios para realizar la llamada de variantes.

Comando para realizar la llamada de variantes:
freebayes -C 50 -f Homo_sapiens.GRCh37.dna.chromosome.7.fa chr7.dedup.RG.bam > chr7_C50.vcf

c. Analiza las estadísticas básicas sobre la llamada de variantes e indica: ¿cuántas variantes se han detectado en total? ¿Cuántas son SNPs? ¿Cuál es el ratio de SNPs Het/Hom?

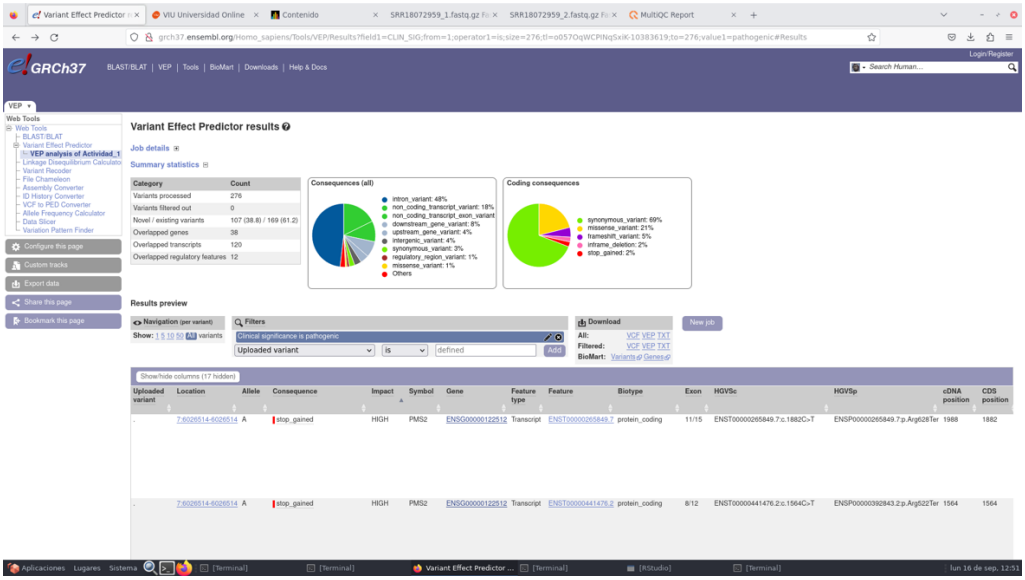
Comando para obtener las estadísticas básicas de la llamada de variantes:
rtg vcfstats chr7_C50.vcf > chr7_C50.vcfstats

Variantes en total: 276
SNPs: 214
Ratio SNPs Het/Hom: 2.72 (155/59)

6. **Análisis de variantes detectadas en el archivo VCF utilizando VEP (recordad que el genoma de referencia corresponde a Hg19/GRCh37 y usad todos los parámetros tal y como los vimos en clase, excluyendo las variantes comunes al filtrar por frecuencia e incluyendo frecuencia alélica global de los 1000 genomas y las continentales, así como los cálculos SIFT y Polyphen, y la indicación de fenotipos) y visualización en IGV de este archivo. En este último paso son varias las cuestiones a contestar (2,5 puntos):**
    a. **De acuerdo con la salida en pantalla del análisis VEP: ¿cuántas variantes de las analizadas son nuevas? ¿qué porcentaje de variantes son intrónicas? ¿Qué porcentaje de variantes son sinónimas del total de variantes en regiones codificantes? ¿Qué porcentaje de variantes suponen un cambio de marco de lectura del total de variantes en regiones codificantes? (0.5 puntos)**

Variantes nuevas analizadas: 107
Porcentaje de variantes intrónicas: 48%
Porcentaje de variantes sinónimas del total de variantes en regiones codificantes: 69%
Porcentaje de variantes que suponen un cambio del marco de lectura del total de variantes en regiones codificantes: 5%

b. Teniendo en cuenta todos los genes del estudio, ¿Se detectan variantes patogénicas en este análisis? Si es así, indica qué gen/genes están involucrados. Muestra en una captura de pantalla cómo las has filtrado y el resultado. (0,5 puntos)
Sí existen variantes patogénicas. El gen involucrado es el PMS2.



c. **Busca el gen/genes (symbol) en el que se han encontrado variantes patogénicas y revisa las mutaciones en posiciones diferentes (1,5 puntos).**

   i. **Consulta la base de datos OMIM e indica: localización citogenética del gen, fenotipo/s asociados y tipo de herencia.**
   La localización citogenética del gen PMS2 es: 7p.22.1
   Fenotipos asociados con los diferentes tipos de mutaciones: Síndrome de Lynch, cáncer colorrectal y otros tipos de cánceres.
   El tipo de herencia para las variantes patogénicas del Síndrome de Lynch es autosómica recesiva.

   ii. **Haciendo uso de VEP, ¿Cuántas variantes tipo *missense* <u>en localizaciones diferentes</u> encuentras en ese gen? ¿Y cuántas con una ganancia de un codón de stop (*stop gained*)?**
   Aparecen 2 variantes tipo missense en el gen PMS2, y con ganancia de un codón de stop (stop gained) solo 1 variante.

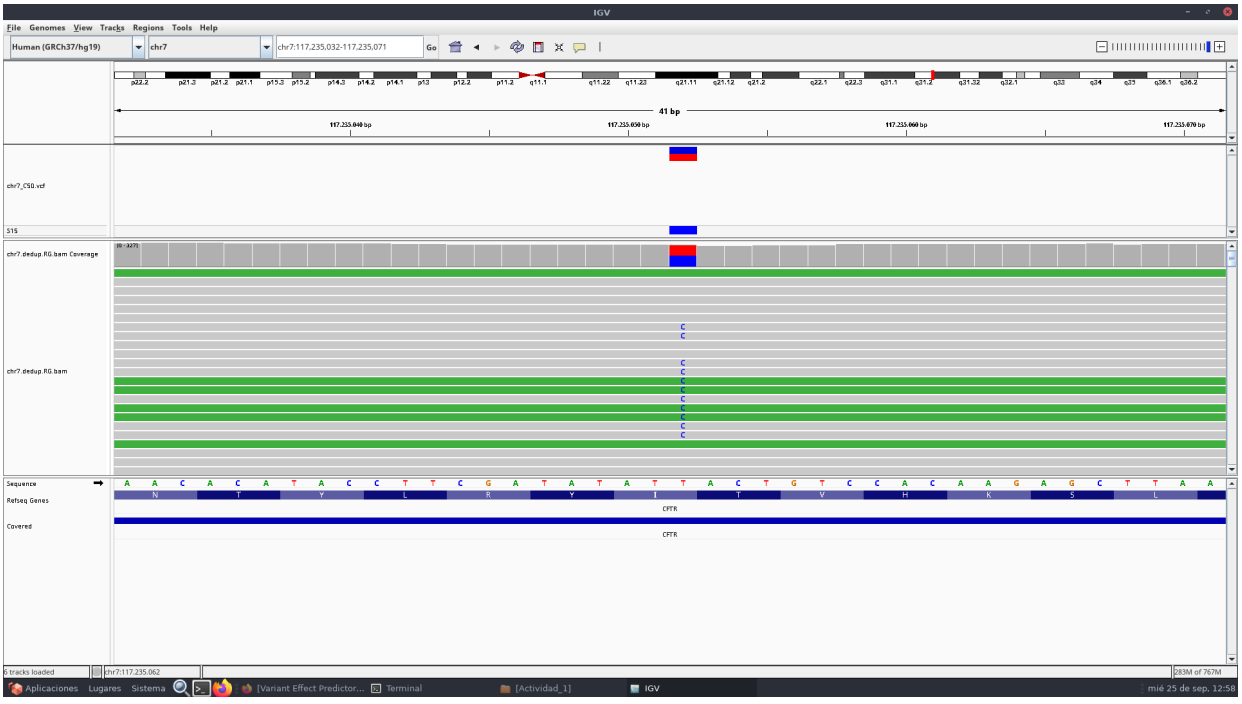
iii. **Rellena una tabla como la que se indica con la información de las variantes detectadas en el anterior apartado, donde se indique: localización, alelo de referencia, alelo alternativo, impacto esperado de la variante, frecuencia alélica en población europea y significado clínico asociado. Rellena tantas filas como variantes en localizaciones diferentes tengas. ¿Hay alguna variante patogénica? ¿Qué patología está relacionada con esta variante?**

Localización	Alelo referencia	Alelo alternativo	Impacto	FA EUR	Significado clínico
7:6026514-6026514	G	A	Alto	-	Patogénica
7:6026775-6026775	T	C	Moderado	0.8767	Benigno
7:6026942-6026942	G	T	Moderado	0.0457	Benigno

Si, hay una variante patogénica que está relacionada con el Síndrome del neurodesarrollo asociado a EIF2AK1.

7. **Visualiza en IGV los archivos de mapeo y análisis de variantes. Vete a la región 7:117235052-117235052 e indica (2 puntos):**
   a. **¿Qué archivos has utilizado en IGV?**
   He utilizado el archivo chr7.dedup.RG.bam y el chr7.dedup.RG.bam.bai indexado, el archivo bed con las sondas, 3313701_Covered.bed y el archivo vcf con las variantes, chr7_C50.vcf

   b. **Muestra en una captura de pantalla esta posición e indica ¿a qué gen corresponde la región seleccionada? ¿Esta región está cubierta por alguna sonda? ¿Qué exón del gen está representado?**



La región seleccionada corresponde al gen CFTR y está cubierta por una sonda.
Está representando el exón número 15.

   c. **Relativa a la variante seleccionada en ese punto: ¿cuál es la variante visualizada? ¿cuál es el alelo de referencia? ¿Cuál es el alelo alternativo? ¿Cuál es la frecuencia alélica?**
   Se trata de un polimorfismo de un solo nucleótido, SNP. El alelo de referencia sería la T, y el alelo alternativo la C.
   La frecuencia alélica es de 0.5

d. **Consulta en VEP esta variante e indica: ¿qué tipo de variante es? ¿cuál es su impacto? ¿cuál es su referencia 'rs'? ¿cuál es su frecuencia en la población europea (EUR AF)?**
Es una variante sinónima, por lo que no hay ningún cambio en el aminoácido codificado.
El impacto de la variante es bajo. Su referencia es: rs1800104
La frecuencia en la población europea es 0.

Instrucciones entrega

Cada captura debe contener la hora y fecha del ordenador (captura de pantalla completa). Cualquier captura que no cumpla este requisito no será válida para la entrega del ejercicio.

La entrega se debe realizar en un único documento en formato PDF.

Límite de entrega:

18/10/2024 a las 23:59

Criterios de evaluación.

La actividad tiene 7 apartados, con valor variable (indicado en el apartado FORMATO DE ENTREGA).

Debe numerarse apropiadamente cada uno de los apartados para su entrega. La entrega en un apartado que no corresponde, no se puntuará.

Se valorará el correcto seguimiento de las instrucciones de la actividad.