Identificación de variantes en una muestra de tumor

José de la Fuente R.

Para el desarrollo de esta actividad, se usarán datos de una muestra de carcinoma acínico masivo, cuya secuencia fue obtenida a través de la plataforma Illumina 2000 y reportada en 2013 por A.C Nichols et al. en *Case reports in Oncological Medicine*. La data de la muestra se obtuvo desde el siguiente enlace.

El carcinoma acínico es un tipo de cáncer que se desarrolla en las glándulas salivales, especialmente en la glándula parótida, la cual es la glándula salival más grande.

Identificación de variantes

Para la identificación de varientes, el primer paso es el mapeo de las lecturas de secuenciación de la muestra, que en este caso corresponderá cromosoma 5, acompañado de la detección de indels los cuales son una guía de las variantes para el paso de realinamiento local, que es usado para mejorar el mapeo y facilitar una mejor detección de las variantes.

Siguiendo las instrucciones del tutorial se realizó la identificación de variantes, cuyo resultado se guardó en la carpeta Analyzed data, como se muestra en la figura 1.

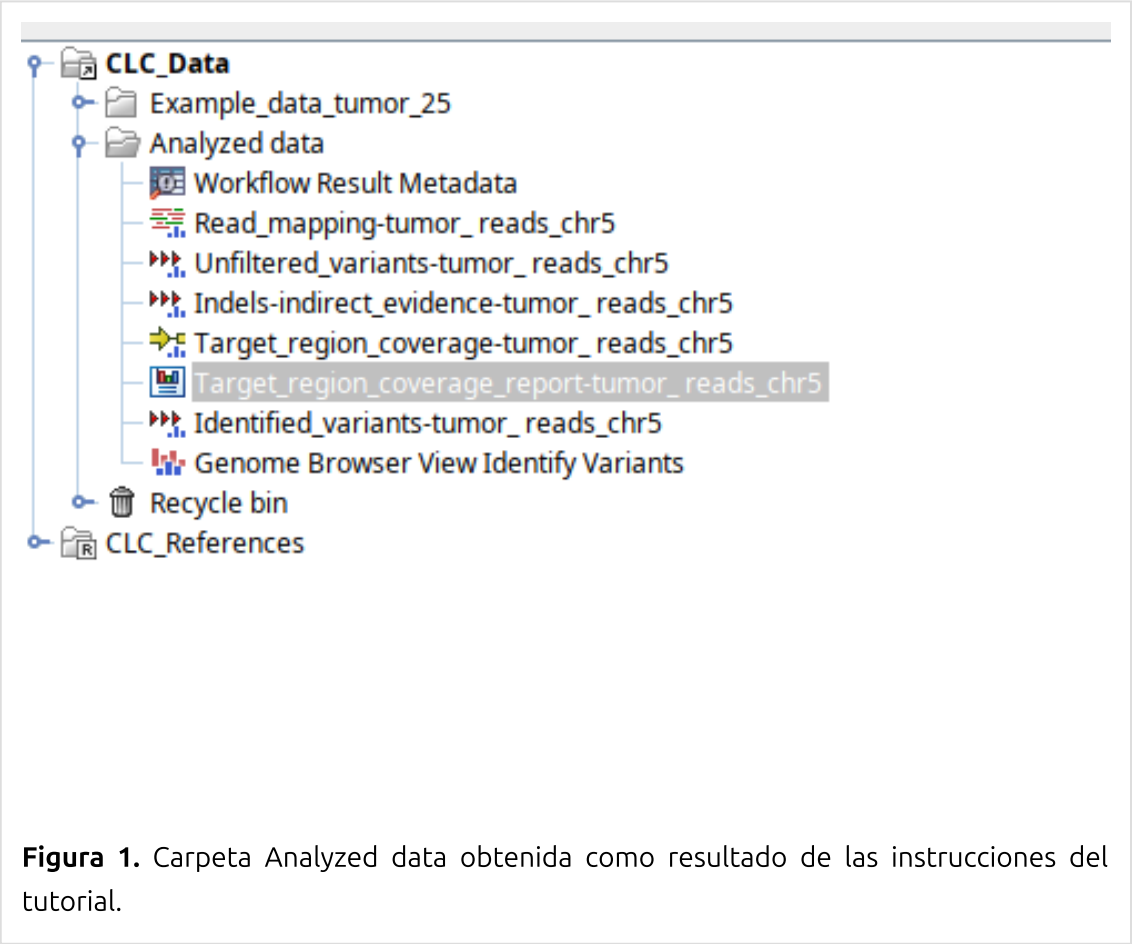


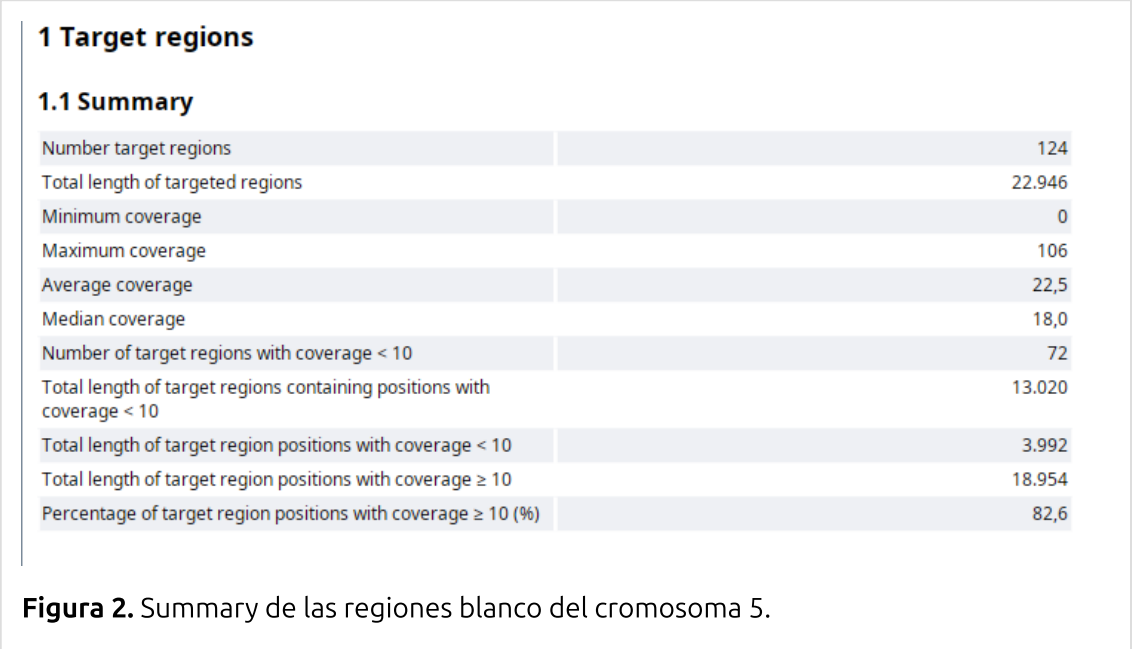
Figura 1. Carpeta Analyzed data obtenida como resultado de las instrucciones del tutorial.

Chequeo del reporte de calidad (QC) de las regiones objetivo.

Para comprobarar si el enriquecimiento de las zonas objetivo fue correcto, se debe revisar *Target_region_coverage_report-tumor_reads_chr5* que aparece marcado en la figura 1. Con ello podemos responder las siguientes preguntas:

¿La cobertura promedio en las regiones objetivo es suficiente?

Para responder esta pregunta, se puede analizar la figura 2. En ella se puede observar que la región objetivo tiene una cobertura de un 82,6% de las bases con el parametro de minima cobertura de 10X que se fijó según el tutorial.



1 Target regions

1.1 Summary

Number target regions	124
Total length of targeted regions	22.946
Minimum coverage	0
Maximum coverage	106
Average coverage	22,5
Median coverage	18,0
Number of target regions with coverage < 10	72
Total length of target regions containing positions with coverage < 10	13.020
Total length of target region positions with coverage < 10	3.992
Total length of target region positions with coverage ≥ 10	18.954
Percentage of target region positions with coverage ≥ 10 (%)	82,6

Figura 2. Summary de las regiones blanco del cromosoma 5.

¿La especificidad de las lecturas que se alinean a las regiones objetivo está dentro del rango esperado (por ejemplo, superior al 50% para secuenciación de exoma y superior al 90% para secuenciación de amplicones dirigidos)?

La especificidad de las lecturas fue de un 36,48%, como se muestra en la figura 3. Si bien en el tutorial se nos indica que se está considerando una pequeña fracción del total de las regiones objetivo, lo que hace que el valor esté subestimado, un resultado como este nos podría sugerir que quizás la captura del cromosoma 5 fue poco eficiente, que hay mucho ruido fuera del objetivo, que el enriquecimiento fue limitado y/o incluso una mala calidad en el diseño de la sonda.

2. Targeted region overview

Reference	Total mapped reads	Mapped reads in targeted region	Specificity (%)	Total mapped reads excl ignored
5	21.952	8.008	36,48	21.952

Mapped reads in targeted region excl ignored	Specificity excl ignored (%)
8.008	36,48

Reference	Total mapped bases	Mapped bases in targeted region	Specificity (%)	Total mapped bases excl ignored
5	1.769.422	515.539	29,14	1.769.422

Mapped bases in targeted region excl ignored	Specificity excl ignored (%)
515.539	29,14

Figura 3. Summary de las regiones blanco del cromosoma 5.

¿Todas las regiones específicas están suficientemente cubiertas?

De las 124 regiones objetivo totales, 88 regiones (71%) tienen al menos un 80% de sus bases con una cobertura mayor o igual a 10X, como se puede ver en la figura 4, lo se considerará como aceptable para continuar con el estudio de variantes.

1.2 Fractions of targets with coverage at least 10

Number of targeted regions for which	Count	Percentage
≥100% of the targeted region has coverage at least 10	52	41,94
≥90% of the targeted region has coverage at least 10	75	60,48
≥80% of the targeted region has coverage at least 10	88	70,97
≥70% of the targeted region has coverage at least 10	96	77,42
≥60% of the targeted region has coverage at least 10	100	80,65
≥50% of the targeted region has coverage at least 10	104	83,87
≥40% of the targeted region has coverage at least 10	107	86,29
≥30% of the targeted region has coverage at least 10	107	86,29
≥20% of the targeted region has coverage at least 10	107	86,29
≥10% of the targeted region has coverage at least 10	107	86,29
≥0% of the targeted region has coverage at least 10	124	100,00

3

Figura 4. Análisis por región con cobertura de al menos 10X.

¿Está bien cubierta una región objetivo en particular con lecturas?

Para responder esta pregunta, se seleccionará el *Genome Browser Identify Variants*, que se muestra en la figura 1, la cual mostrará un panel como el de la figura 5.

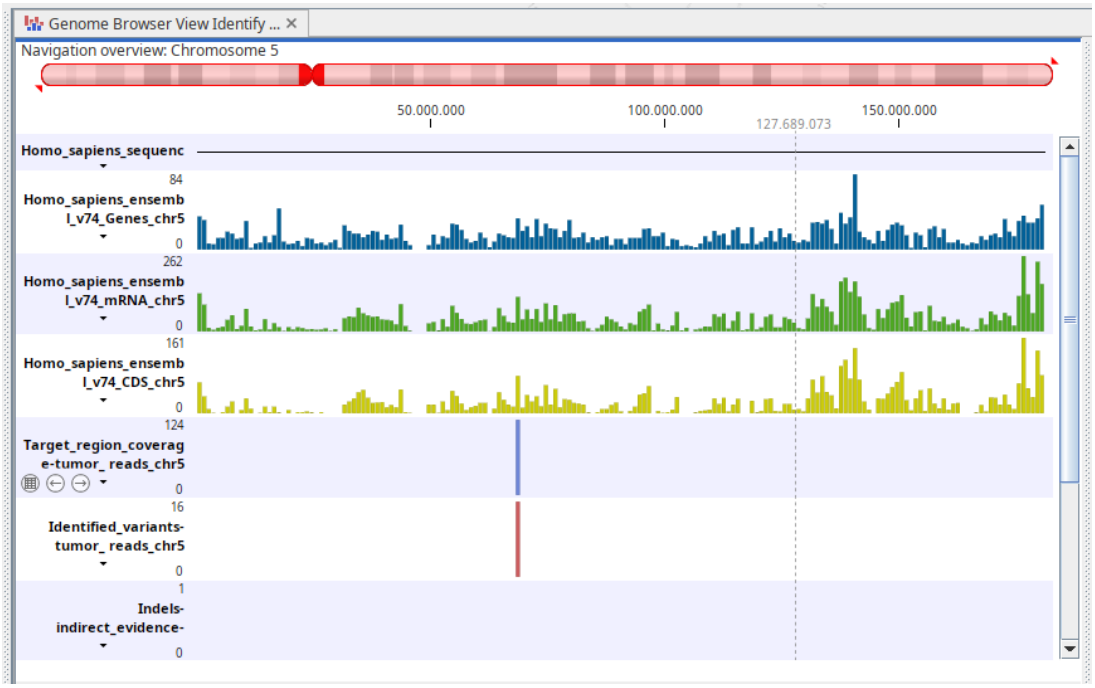
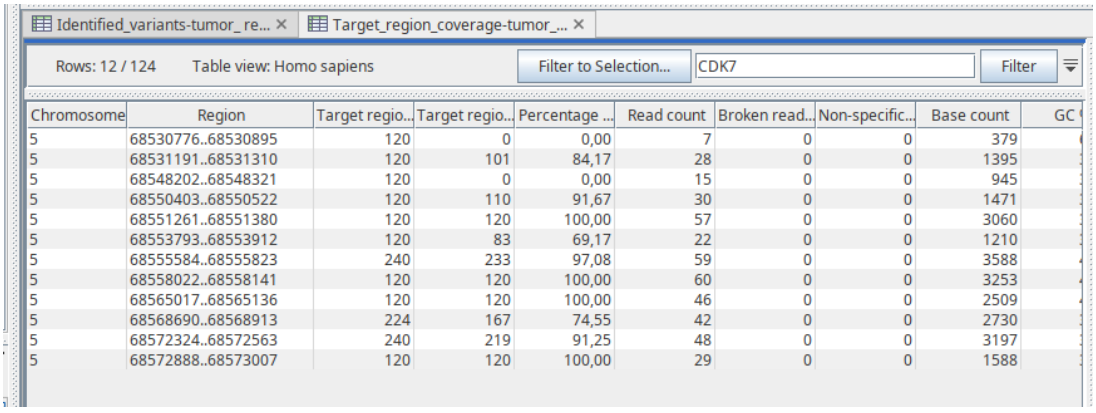


Figura 5. Genome Browser Identify Variants.

En el se amplía la zona en la cual coincidan los tracks de referencia, junto al *Target_region_coverage-tumor_reads_chr5*. En esa zona se pueden notar varios genes, como CCDC125, SLC30A5, RAD17, CCNB1 y CDK7. Este último ha sido detectado en varios tipos de cáncer y ha sido asociado con resultados clínicos como blanco terapéutico (Sava G.P. et al.,2020). Por lo que el análisis de variantes en esa zona puede ser valiosa. Para ello, se filtra con el nombre de CDK7, que la mayoría de las regiones poseen una cobertura por sobre el filtro aplicado (10X), con 7 regiones exónicas con valores sobre el 90% de cobertura. Como se ve en la fiugra 6, es posible también notar que hay 2 regiones con cobertura nula. Con todo lo anterior, se puede afirmar que el gen CDK7 está bien cubierto.



Chromosome	Region	Target regio...	Target regio...	Percentage ...	Read count	Broken read...	Non-specific...	Base count	GC
5	68530776..68530895	120	0	0,00	7	0	0	379	
5	68531191..68531310	120	101	84,17	28	0	0	1395	
5	68548202..68548321	120	0	0,00	15	0	0	945	
5	68550403..68550522	120	110	91,67	30	0	0	1471	
5	68551261..68551380	120	120	100,00	57	0	0	3060	
5	68553793..68553912	120	83	69,17	22	0	0	1210	
5	68555584..68555823	240	233	97,08	59	0	0	3588	
5	68558022..68558141	120	120	100,00	60	0	0	3253	
5	68565017..68565136	120	120	100,00	46	0	0	2509	
5	68568690..68568913	224	167	74,55	42	0	0	2730	
5	68572324..68572563	240	219	91,25	48	0	0	3197	
5	68572888..68573007	120	120	100,00	29	0	0	1588	

Figura 6. Tabla filtrada para gen CDK7.

Chequeo de falsos positivos de la variante escogida

Para ello, se filtró en el *Genome Browser Identify Variants* siguiendo las instrucciones del tutorial de filtrado y además del chequeo de falsos positivos, cuyos criterios son:

- La calidad promedio de las bases de la variante : Una calidad promedio baja (por debajo de 20) podría sugerir que se trata de un error de secuenciación.

- El número de lecturas únicas que apoyan la variante :revisar los valores en las columnas: unique start positions y unique end positions. Estos valores deben ser mayores que uno. Si no lo son, la variante podría deberse a un error de PCR durante el enriquecimiento.

- Las regiones que rodean la variante : en la lista de tracks, observa las regiones del genoma alrededor de la variante en la secuencia de referencia. ¿Está en una región de homopolímeros (por ejemplo, una secuencia continua de As)? ¿Es una deleción o una inserción? Si es así, la variante podría ser un error de secuenciación.

- El número de lecturas que respaldan la variante: este valor debe ser al menos 1, pero preferiblemente 5 o más.

Como se puede ver en la figura 7, la variante de CDK7, marcada en gris en la figura,cumple con todos los criterios para no ser considerada un falso positivo.

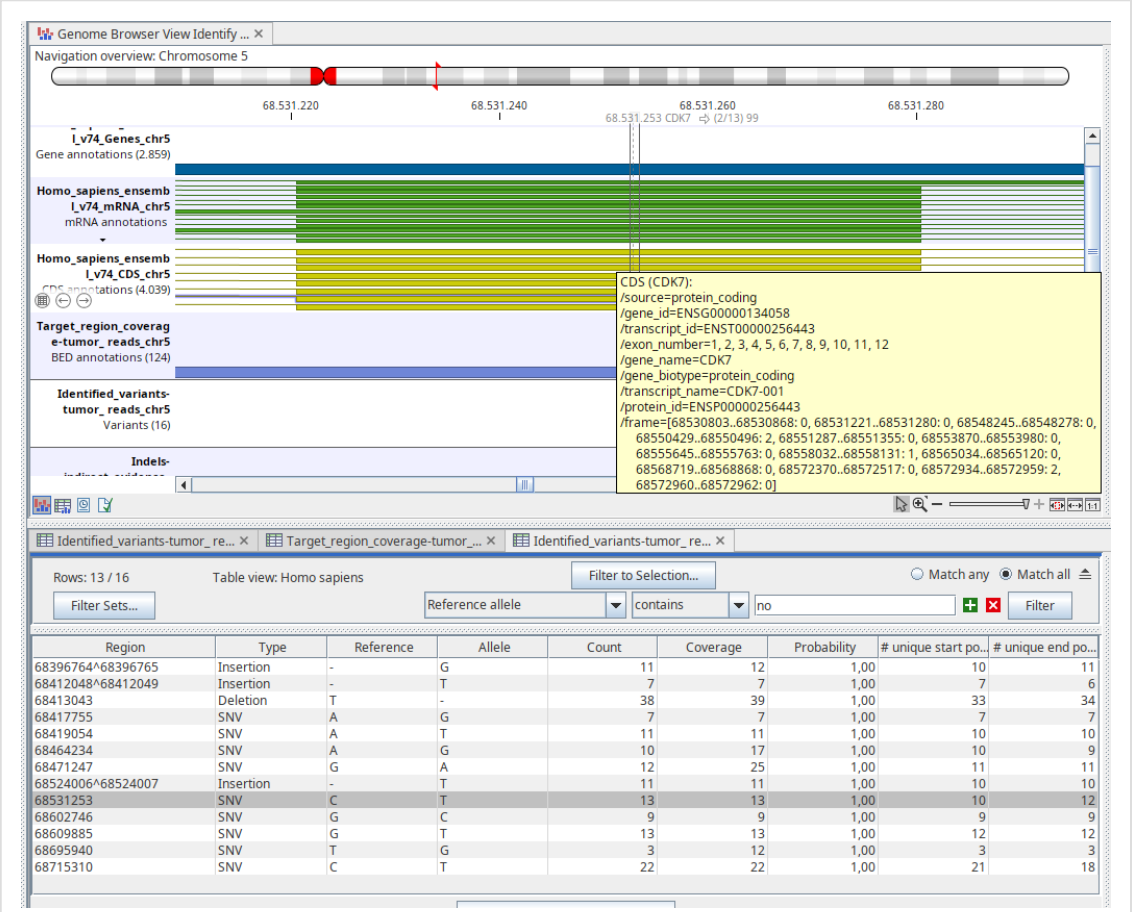


Figura 7. Chequeo de falso positivo para CDK7.

Análisis e interpretación clínica de la variante CDK7

Para el análisis e interpretación, se usará VEP, y ClinVar. De ellas el uso combinado de ellas da una idea del efecto de la variante. De la figura 7, se ve marcada la variante a analizar. Los resultados de los llamados en ambas plataformas se muestra a continuación:

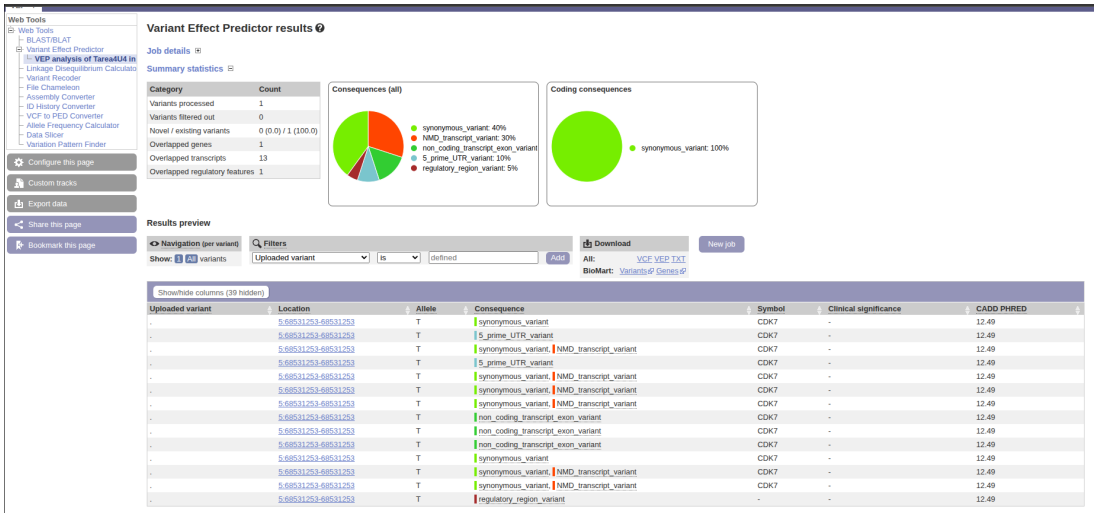


Figura 8. Resultado de plataforma VEP.

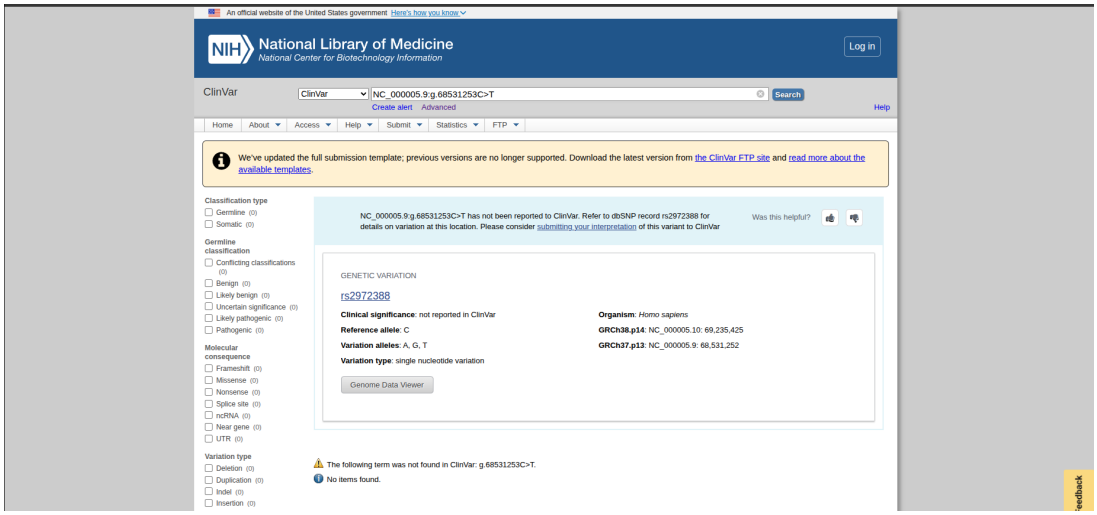


Figura 9. Resultado de plataforma ClinVar.

Como se puede observar en ninguna de las dos plataformas se muestran resutlados que indiquen su significancia clínica. Además, como se muestra en la figura 8, por el valor CADD, es posible que exista algún impacto, pero que esté no es fuerte.

Conclusiones

La variante 5:68531253 en la figura 9, se muestra que la variente está asociada a rsID: rs2972388, lo que indica que la variante es conocida en la población humana . Además no hay evidencia o reportes en cuanto a su patogenicidad, ni en nivel somático o germinal. Tampoco parece estar asociada a alguna enfermedad distinta al cáncer. En resumen, no parece existir evidencia clínica al respecto de esta variante. Los resultados de VEP, en la figura 8, muestran que es una variante sinónima, lo que implica que no hay alteraciones ni en el aminoácido, y por consiguiente en la proteína, tampoco es missense o nonsense ni afecta el sitio de splicing. Con lo anterior, podemos afirmar que hasta el momento esta variante se encuentra en el tier IV, ya que nos encuentra reportada en ClinVar y su CADD es de 12.49.