Imagen que contiene nombre de la empresa

Descripción generada automáticamente

**Trabajo práctico**

**Bioinformática**

**Profesor**: Patricio Yankilevich

**Curso**: K5551

**Año**: 2022

**Cuatrimestre**: 2do

|  |  |
| --- | --- |
| **Integrantes** | |
| Jacquemin, Solana | 156.758-5 |
| Luquet, María Azul | 163.415-0 |
| Walsh, Alejandro | 153.054-9 |

Ejercicio 6

Trabajo con Bases de Datos Biológicas (útil para la presentación del Ejercicio 7).

1. *A partir del gen o proteína de interés para ustedes dar su link a NCBI-Gene como una entrada de Entrez, por ej.:* [*http://www.ncbi.nlm.nih.gov/gene/3630*](http://www.ncbi.nlm.nih.gov/gene/3630) *Expliquen brevemente lo que hace la proteína y por qué la eligieron.*

El link de NCBI-Gene de nuestro gen de interés es:  
<https://www.ncbi.nlm.nih.gov/gene/170680>

Este gen es uno de varios genes que se cree que confieren susceptibilidad a la psoriasis y la esclerosis sistémica, ubicados en el cromosoma 6 cerca de la región de clase I del complejo mayor de histocompatibilidad (MHC).

La elegimos principalmente debido a que nos pareció una enfermedad de la cual no conocemos mucho y también nos pareció interesante por la variedad de presentaciones que puede provocar, es decir psoriasis en placas, en gotas, entre otras.

1. *¿Cuántos genes / proteínas homólogas se conocen en otros organismos? Utilicen la información que está en la base de datos de HomoloGene y en* *las bases de datos Ensembl. Describan los resultados en ambas bases de datos, y en qué se diferencian. Mencionen sobre qué tan común creen son estos genes o proteínas y a qué grupos taxonómicos pertenecen (sólo en las bacterias, en los vertebrados, etc.)*

Se buscó en la base de datos de HomoloGene, el código asociado al gen de nuestro interés (PSORS1C2) (<https://www.ncbi.nlm.nih.gov/homologene/?term=PSORS1C2>)

Se encontraron dos resultados:

1. PSORS1C1 - conservado en Homininae, a continuación, se muestran los genes homólogos entre sí:
   1. [PSORS1C1, *H.sapiens*](https://www.ncbi.nlm.nih.gov/gene/170679)
   2. [PSORS1C1, *P.troglodytes*](https://www.ncbi.nlm.nih.gov/gene/100169686)
2. PSORS1C2 – conservado en Boreoeutheria y estos son los genes homólogos entre sí:
   1. [PSORS1C2, *H.sapiens*](https://www.ncbi.nlm.nih.gov/gene/170680)
   2. [PSORS1C2, *P.troglodytes*](https://www.ncbi.nlm.nih.gov/gene/462554)
   3. [PSORS1C2, *M.mulatta*](https://www.ncbi.nlm.nih.gov/gene/714607)
   4. [PSORS1C2, *C.lupus*](https://www.ncbi.nlm.nih.gov/gene/100682937)
   5. [PSORS1C2, *M.musculus*](https://www.ncbi.nlm.nih.gov/gene/57390)
   6. [PSORS1C2, *R.norvegicus*](https://www.ncbi.nlm.nih.gov/gene/686254)

Luego, realizamos la misma búsqueda en la base de datos Ensembl. Se encontraron más coincidencias que con HomoloGen. (<https://www.ensembl.org/Multi/Search/Results?q=PSORS1C2;site=ensembl;page=1;style=table;perpage=10;fall_species=1>)

Analizando un poco estos resultados, vemos que el gen aparece relacionado a la mayoría de las especies, tales como: Homo sapiens, Mus musculus, Rhinopithecus bieti, Ursus americanus Pan paniscus, Pan troglodytes, Mandrillus leucophaeus, Aotus nancymaaen, Nomascus leucogenys, Pongo abelii, Gorilla gorilla, entre otros.

Con esto podemos entender que el gen está presente en la gran mayoría de las especies, y no es exclusivo de ninguna en particular. Indicando que es un gen relativamente común.

1. *¿Cuántos transcriptos y cuántas formas alternativas de splicing son conocidos para este gen / proteína? ¿Cuáles de estos splicing alternativos se expresan? ¿Tienen funciones alternativas? Buscar evidencia de esto en* *las bases de datos de NCBI y en los transcriptos de Ensembl ¿Cómo el número de splicings alternativos diferente entre las dos bases de datos y cuál piensan que es más precisa y por qué?*

En NCBI podemos ver que tiene solo un único transcripto para el gen y una proteína asociada. Por lo tanto, podemos decir que solo hay un splicing expresado del gen.

Por otra parte, en Ensembl vemos que hay 118 transcriptos, algunos de ellos son:

* 6 transcriptos del Homo Sapiens
* 16 de Mus musculus
* 9 de Canis lupus familiaris
* 9 de Sus scrofa (distribuido en varias razas de cerdos)

Vemos que en Ensembl hay una mayor cantidad de información que en NCBI, y asumimos que es porque se han analizado más variantes del gen y se ha buscado en múltiples especies, mientras que en NCBI se ha priorizado específicamente al Homo sapiens.

1. *¿Con cuántas otras proteínas interactúa el producto génico de su gen? ¿Existe un patrón o relación entre las interacciones? Mencione las interacciones interesantes o inusuales. Usted encontrará las interacciones de su gene/proteína tanto en la base de datos NCBI Gene como en la base de datos UniProt . Compare las dos tablas entre sí. ¿Hay proteínas que interactúan únicas para cada tabla?*

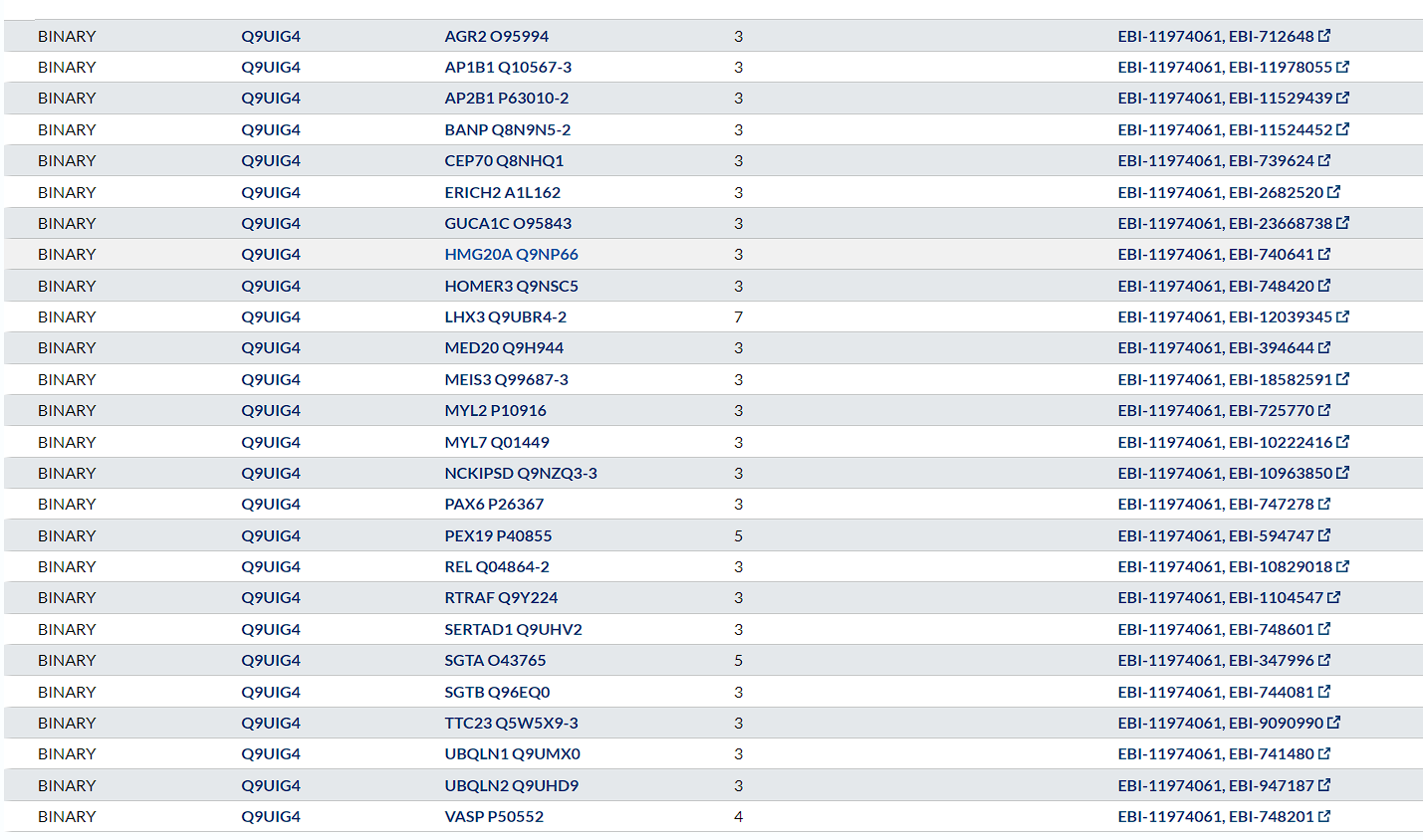
En NCBI vemos 28 interacciones con este gen, adjuntamos una imagen con algunas de referencia:

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

Por ejemplo, la primera que aparece es con el gen EVPL   
(<https://www.ncbi.nlm.nih.gov/gene/2125>) que, según vimos, es un gen que codifica un miembro de la familia de proteínas plaquinas y los pacientes que posean alguna enfermedad de desorden autoinmune pueden generar anticuerpos que ataquen a la proteína codificada.

Por otra parte, en UniProt encontramos 26 interacciones:



En conclusión, los genes que aparecen únicamente en NCBI y no en UniProt son EVPL (<https://www.ncbi.nlm.nih.gov/gene/2125>) y MICAL1 (<https://www.ncbi.nlm.nih.gov/gene/64780>)

1. *Expliquen brevemente de qué componente celular forma parte su proteína (pista: se puede estudiar la información de Gene Ontology - GO), ¿A qué procesos biológicos pertenece (pista idem)? y ¿En qué función molecular trabaja esta proteína? Los términos ontológicos de genes los pueden encontrar tanto en NCBI Gene y en la base de datos UniProt como haciendo una búsqueda en AmiGO.*

En Gene Ontology – GO no encontramos información sobre el gen PSORS1C1, solo encontramos de PSORS1C2.

Según la información obtenida en NCBI Gene ([https://www.ncbi.nlm.nih.gov/gene/170680#gene-ontology](https://www.ncbi.nlm.nih.gov/gene/170680%23gene-ontology)), vemos lo siguiente:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Entendemos que esto implica que a nivel función es un gen que permite la vinculación de proteínas y a nivel componente está localizado en la región extracelular. Los demás ítems al tener la leyenda ND indica que no hay información biológica para definir una función molecular, ni para confirmar que es parte de un proceso biológico y tampoco que está activo en un componente celular.

En AmiGO (<http://amigo.geneontology.org/amigo/gene_product/UniProtKB:Q9UIG4>) también podemos ver información muy similar a esto indicando que la mayoría de la información es desconocida, con excepción de lo ya mencionado para NCBI, a nivel componente está localizado en la región extracelular y a nivel función permite la vinculación de proteínas.

Y, por último, UniProt ([https://www.uniprot.org/uniprotkb/Q9UIG4/entry#subcellular\_location](https://www.uniprot.org/uniprotkb/Q9UIG4/entry%23subcellular_location)) confirma esta información de que está localizado en la región extracelular incluso indicando que pertenece a los datos curados por lo cual ya es algo confirmado y verificado.

1. *Discutan brevemente en qué estructura o vías metabólicas específicas (pathways) estaría participando su gen / proteína? (Reactome, KEGG son algunas bases de datos de pathways).*

En Reactome (<https://reactome.org/content/detail/interactor/Q9UIG4>) encontramos 4 interacciones con la proteína correspondiente a nuestro gen:

Imagen que contiene Tabla

Descripción generada automáticamente

Y, por otra parte, en KEGG (<https://www.genome.jp/entry/hsa:170680>) se indica que no está incluido en sus pathways.

1. *Entrar en la base de datos de variantes genéticas dbSNP e intentar interpretar o encontrar información sobre alguna variante (reference SNP - rsXXXX) asociada con la patología investigada en su gen de interés. ¿Qué variante es? ¿Hay información sobre la frecuencia que tiene esta variante en la población? ¿Qué grupo étnico parece ser el más afectado?*   
   *NOTA: Para hacer este ejercicio les pueden servir algunas otras bases de datos como estas (entre otras):*

* [*http://www.genecards.org*](http://www.genecards.org)
* [*https://www.ncbi.nlm.nih.gov/snp/*](https://www.ncbi.nlm.nih.gov/snp/) *(para obtener información de la variante en la población)*
* [*http://www.ncbi.nlm.nih.gov/clinvar/*](http://www.ncbi.nlm.nih.gov/clinvar/) *(para obtener información clínica del gen y sus variantes)*
* [*https://ghr.nlm.nih.gov*](https://ghr.nlm.nih.gov)

Hicimos la primera búsqueda en dbSNP de NCBI (<https://www.ncbi.nlm.nih.gov/snp/?term=PSORS1C2>), arrojando 1763 resultados, pero todos de variantes benignas de la patología. Sin embargo, si volvemos al gen con el que habíamos iniciado (<https://www.ncbi.nlm.nih.gov/snp/?term=PSORS1C1>), encontramos 6 variantes patológicas:

* [rs121917819](https://www.ncbi.nlm.nih.gov/snp/rs121917819): es una variación de nucleótido único siendo G>A / G>C con un impacto del 0,03% en la población. No hay información concluyente sobre cuál es el grupo étnico más afectado.
* [rs121917820](https://www.ncbi.nlm.nih.gov/snp/rs121917820): es una variación de nucleótido único siendo G>A y no hay información de la frecuencia de esta. No hay información concluyente sobre cuál es el grupo étnico más afectado.
* [rs387906841](https://www.ncbi.nlm.nih.gov/snp/rs387906841): es una variación de nucleótido único siendo T>A con una frecuencia del 0,0034%. El grupo étnico más afectado es el europeo.
* [rs606231274](https://www.ncbi.nlm.nih.gov/snp/rs606231274): es una variación de tipo Indel siendo una cadena duplicada de CAGG, la frecuencia es de 0,0031% y el grupo étnico más afectado son los judíos asquenazí.
* [rs606231275](https://www.ncbi.nlm.nih.gov/snp/rs606231275): es una variación de nucleótido único siendo C>A sin información de la frecuencia ni de grupos étnicos.
* [rs672601343](https://www.ncbi.nlm.nih.gov/snp/rs672601343): es una variación de tipo Indel siendo una cadena eliminada de C, no hay información de la frecuencia ni de grupos étnicos.