Introducción a la Bioinformática

(16.50)

Trabajo Práctico Especial

Julián Palacci - 57157
Agustín Lavarello - 57046
Santiago Swinnen - 57258

Introducción

En el presente informe se presentará y desarrollará el trabajo práctico especial realizado en el marco de la materia Introducción a la Bioinformática del Instituto Tecnológico de Buenos Aires en el primer cuatrimestre de 2019.
El trabajo, desarrollado en el lenguaje Python con el soporte de la librería BioPython, consta de tres partes.
La enfermedad elegida para el presente trabajo fue cáncer de mama, la cual está muchas veces asociado al gen **BRCA2**

Ejercicio 1

Para realizar el ejercicio 1, dada una o varias secuencias de nucleótidos, se buscó para cada una los marcos de lectura abiertos (ORF) tanto en la secuencia original, como en la secuencia complementaria reversa. Para cada marco de lectura se realizó la traducción a la secuencia de aminoácidos correspondiente.
Para determinar cuál era el marco de lectura correcto, se buscó el marco de lectura más largo que comenzara en una M (metionina).
Para este ejercicio se utilizaron expresiones regulares para identificar codones de inicio (ATG).
Se utilizó el mRNA asociado al gen BRCA2.

Ejercicio 2

En este ejercicio, se realizan las consultas blastp de la secuencia de proteínas. Los valores estadísticos que arroja el blast son:

- E-value: Es el número de coincidencias que uno puede esperar encontrar de casualidad cuando busca en una base de datos. Cuanto más cercano a cero sea el E-value, más significativo es el match, aunque las secuencias cortas suelen tener e-values altos debido a que tienen más probabilidad de ocurrir en una base de datos de casualidad
- Score: Se calcula como la suma de los scores de sustitución y gaps. Los scores de sustitución están determinados por la matriz de sustitución (BLOSUM)
- Positives: Es el número de aminoácidos que son iguales entre las secuencias analizadas o que tienen propiedades químicas similares
- Gaps: Es número de espacios que se le agregan a la secuencia para compensar inserciones o borrados en una secuencia relativa a la otra

El resultado del blast muestra que la traducción de la proteína que supusimos fue la correcta, pues se observa que Breast cancer type 2 susceptibility protein matchea en forma completa, es decir su e-value es de 0, no tiene gaps, y las identities y los positives coinciden con el length de la secuencia input. Esto indica que la traducción a la proteína que se hizo en el ejercicio 1, es correcta.
Además, se puede observar que hay otras proteínas similares en otras especies que se asocian a la misma enfermedad como *Felis catus*, *Mus musculus, Rattus norvegicus*

Ejercicio 3

El programa generado por el archivo *ex3.py* espera dos argumentos. En primer lugar un archivo multifasta como input, y en segundo lugar un archivo de output. Para realizar el alineamiento múltiple, se utiliza Clustal Omega.

Una vez obtenidos los alineamientos, se escriben en el archivo de output y el programa finaliza.

El MSA permite alinear tres o más secuencias de proteínas, ADN o ARN. Se asume que las secuencias que se quieren alinear tienen una relación evolutiva y que comparten descendientes de un mismo ancestro. Del resultado del MSA, se pueden inferir secuencias homólogas y realizar un análisis filogenético para poder determinar los ancestros comunes de las especies involucradas. En base a este análisis es posible luego construir un árbol filogenético.

Realizando el alineamiento a través de ClustalW pudimos obtener el árbol filogenético de las especies

Con lo que se observa que *Mus musculus* y *Rattus norvegicus* comparten un ancestro en común y a su vez ese ancestro comparte un ancestro en común con *Felis catus* y *Homo sapiens*

Ejercicio 4

El programa espera dos argumentos. El primero, es el archivo blast.out, resultado de la salida del ejercicio 2. El segundo parámetro es el pattern a buscar en el resultado del blast. El pattern puede ser una palabra a buscar el título o una secuencia de nucleótidos a buscar en las secuencias de query o de match.

El resultado de la búsqueda se guarda en un archivo llamado patternHits.out, y para cada hit encontrado que coincide con la palabra buscada, se realiza una consulta a Uniprot en base a sus accession numbers, para obtener el archivo fasta de la proteína en cuestión.

Ejercicio 5

Este programa espera cuatro argumentos. El primero corresponde a la secuencia de aminoácidos que se obtuvieron en el ejercicio 1. El segundo, corresponde al nombre del archivo que contendrá el análisis de los dominios de la proteína. El tercero, corresponde al archivo gbk con el gen. El último parámetro, es el nombre del archivo que contendrá los opening reading frames.

El programa ejecuta los comandos EMBOSS *prosextract* que convierte los archivos prosite.dat y proside.doc, contenidos en un directorio llamado prosite.
A continuación, analiza los motivos de la proteína, mediante el comando *patmatmotifs*, dejando el resultado en el archivo que se especificó por parámetro.
Finalmente, se realiza mediante el comando *getorf*, un análisis de los ORF del archivo de nucleótidos.

Ejercicio 6

1) A partir del gen o proteína de interés para ustedes dar su link a NCBI-Gene como una entrada de Entrez

El link correspondiente a nuestro gen analizado es:
https://www.ncbi.nlm.nih.gov/gene/675

El gen BRCA está localizado en 13q13, es decir, cromosoma 13, en la posición 13 del brazo largo. El gen codifica la proteína breast cancer type 2 susceptibility protein

El gen BRCA2 provee instrucciones para construir proteínas que actúan como supresores de tumores. Las proteínas supresoras de tumores ayudan a prevenir que las células crezcan y se dividan rápidamente o de forma no controlada.

La proteína BRCA2 está involucrada en la reparación del ADN dañado. En el núcleo de muchos tipos de células normales, la proteína BRCA2 interactúa con varias otras proteínas para reparar cortes en el ADN. Estos cortes pueden ser causados por radiación natural o médica u otras exposiciones ambientales, y también ocurren cuando los cromosomas intercambian material genético en preparación para la división celular. Ayudando a reparar el ADN, la proteína BRCA2 juega un papel fundamental en el mantenimiento de la estabilidad de la información genética de la célula.

Los investigadores sospechan que la proteína BRCA2 tiene funciones adicionales dentro de las células. Por ejemplo, la proteína podría ayudar a regular la citocinesis, que es el paso en la división celular en donde el fluido alrededor del núcleo (el citoplasma) se divide para formar dos células separadas. Se están investigando otras actividades potenciales de esta proteína.

https://ghr.nlm.nih.gov/gene/BRCA2

Elegimos esta proteína porque nos pareció interesante que estuviera involucrada en la reparación de ADN dañado, que muchas veces trae como consecuencia la aparición de varios tipos de cáncer, en particular cáncer de ovario y de mama.

2) ¿Cuántos genes / proteínas homólogas se conocen en otros organismos? Utilicen la información que está en la base de datos de HomoloGene y en la 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 realizó primero la siguiente búsqueda en HomoGene:

> brca2[gene name] AND human[orgn]

Se obtuvo la siguiente salida:

HomoloGene:41. Gene conserved in Amniota

Genes

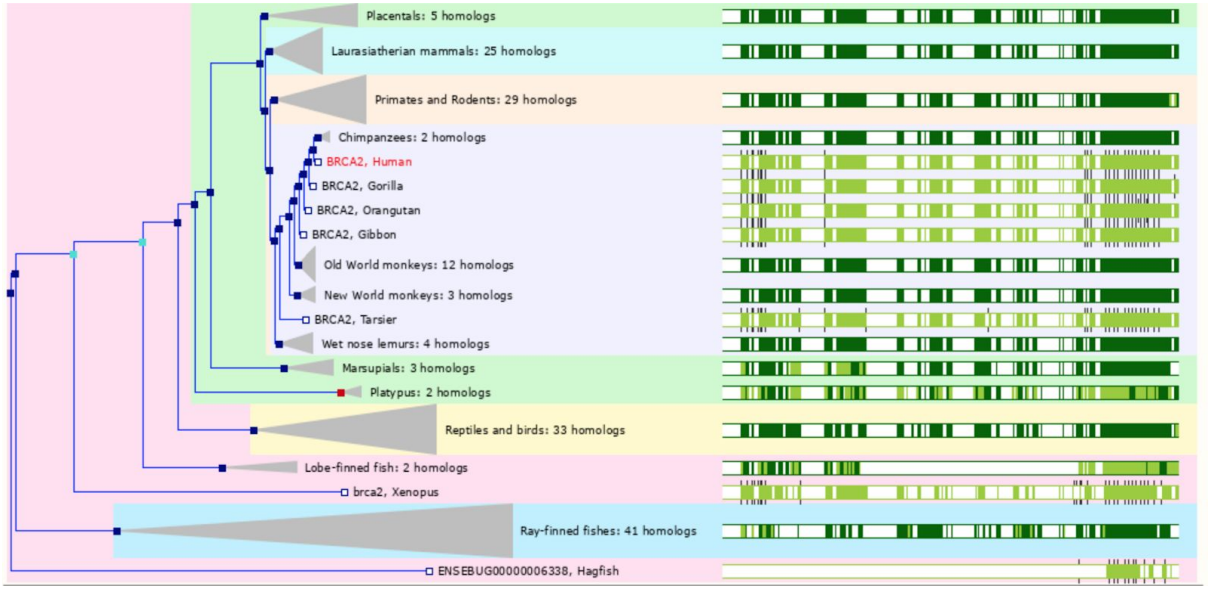
Genes identified as putative homologs of one another during the construction of HomoloGene.

BRCA2, *H.sapiens*
breast cancer 2, early onset
BRCA2, *P.troglodytes*
breast cancer 2, early onset
BRCA2, *M.mulatta*
breast cancer 2, early onset
BRCA2, *C.lupus*
breast cancer 2, early onset
BRCA2, *B.taurus*
breast cancer 2, early onset
Brca2, *M.musculus*
breast cancer 2
Brca2, *R.norvegicus*
breast cancer 2, early onset
BRCA2, *G.gallus*
breast cancer 2, early onset

Aparecen **12 genes conservados en el grupo de los Amniota.**
Los amniota son un clado de vertebrados tetrápodos. Se caracterizan porque el embrión desarrolla cuatro envolturas: el corion, el alantoides , el amnios y el saco vitelino y crea un medio acuoso en el que puede respirar y del que puede alimentarse.La menor cantidad en relación a los obtenidos con Ensembl se debe a que la información en NCBI está curada, por lo que las opciones resultantes son menos.

https://www.ncbi.nlm.nih.gov/guide/howto/find-homolog-gene/

En la base de datos de Ensembl se puede observar el árbol filogenético:



Los árboles de genes en Ensembl se construyen utilizando una proteína representativa por cada gen en cada especie en Ensembl.
El gráfico muestra el *maximum likelihood phylogenetic tree* que representa la historia evolutiva de los genes.

En esta tabla se pueden observar los genes ortólogos. Aparecen 167 genes ortólogos.
Dentro de la homología de secuencia se distinguen dos tipos de homología: la ortología y la paralogía. Se llaman genes ortólogos a los que son semejantes por pertenecer a dos especies que tienen un antepasado común. Existen además genes parálogos, que son aquellos que se encuentran en el mismo organismo, y cuya semejanza revela que uno procede de la duplicación del otro. La ortología requiere que se haya producido especiación, mientras que esta no es necesaria en el caso de la paralogía, que puede producirse sólo en los individuos de una misma especie.

Summary of orthologues of this gene Hide

Click on 'Show details' to display the orthologues for one or more groups of species. Alternatively, click on 'Configure this page' to choose a custom list of species.

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
Primates (*26 species*) Humans and other primates	☐	25	0	0	1
Rodents and related species (*31 species*) Rodents, lagomorphs and tree shrews	☐	26	2	0	3
Laurasiatheria (*25 species*) Carnivores, ungulates and insectivores	☐	25	0	0	0
Placental Mammals (*87 species*) All placental mammals	☐	81	2	0	4
Sauropsida (*34 species*) Birds and Reptiles	☐	33	0	0	1
Fish (*48 species*) Ray-finned fishes	☐	39	1	0	8
All (*183 species*) All species, including invertebrates	☑	159	4	0	20

En esta sección, las especies se agrupan en conjuntos, como Primates, Roedores, Laurasiatheria, Mamíferos placentarios, Saudaropsida y Peces.

Los tipos de ortologia se asignan comparando dos especies de la siguiente forma:

- Ortólogos 1-1: Solo una copia se encuentra en cada especie
- Ortólogos 1 a varios: Un gen en una especia es ortólogo a varios otros genes en otras especies
- Ortólogos varios a varios: Múltiples ortólogos se encuentran en ambas especies

Se puede observar que se encuentran más genes ortólogos en los mamíferos (87). Además, la proteína pertenece a la familia PTHR11289_SF0

c) ¿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 base 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?

Este gen presenta 7 transcriptos (o splicings alternativos).

Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq Match	Flags
BRCA2-206	ENST00000544455.5	10984	3418aa	Protein coding	CCDS9344	P51587	-	TSL:1 GENCODE basic APPRIS P1
BRCA2-201	ENST00000380152.7	11986	3418aa	Protein coding	CCDS9344	P51587	-	TSL:5 GENCODE basic APPRIS P1
BRCA2-202	ENST00000470094.1	842	186aa	Nonsense mediated decay	-	H0YE37	-	CDS 5' incomplete TSL:5
BRCA2-203	ENST00000528762.1	495	64aa	Nonsense mediated decay	-	H0YD86	-	CDS 5' incomplete TSL:4
BRCA2-207	ENST00000614259.1	7950	No protein	Processed transcript	-	-	-	TSL:2
BRCA2-205	ENST00000533776.1	523	No protein	Retained intron	-	-	-	TSL:3
BRCA2-204	ENST00000530893.6	2011	No protein	Processed transcript	-	-	-	TSL:1

Si se analiza el transcript table, se puede observar que de los 7 transcriptos, 2 forman protein coding mientras que los tres últimos no son expresados. Esto se puede observar en la columna “Biotype”, donde “Protein Coding” significa que la secuencia tiene un marco abierto de lectura y “Processed Transcript” es un transcrito que no contiene un ORF y por ende no puede ser codificado, “Retained intrón” que tiene un splicing alternativo que se cree que tiene intrones relativos uno a otro. Cabe destacar que el color presente en la columna de “biotype” determina el grado de confiabilidad de la información. Si está dorado quiere decir que es muy confiable ya que la información fue curada tanto por Ensembl automated annotation como VEGA/Havana manual curation y la información es idéntica. Si está en rojo, quiere decir que el transcrito viene de uno de los dos sistemas antes mencionados pero no presenta gran confiabilidad porque no fue revisada aún por el otro sistema.

d) ¿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?

El producto génico del gen BRCA tiene 315 interacciones con un total de 120 interactores. Esta información fue obtenida de BioGrid, sitio accedido desde UniProt. Investigando en los documentos de PubMed se puede encontrar que dichas interacciones están principalmente relacionadas con dos funciones: la prevención de la reproducción descontrolada y la reparación de ADN dañado. La interacción más significativa (por orden de evidencias) se da con el el interactor RAD51, cuya función está directamente vinculada a la reparación de ADN.

La base de datos de NCBI, por otro lado, presenta una tabla de 144 interacciones. Esta tabla recopila información de BIND, HPRD y BioGRID. Se muestra a continuación la red de interacciones con un mínimo de evidencia 1 del gen BRCA obtenido BioGRID. Allí se puede observar su estrecha vinculación con BRCA1 y con RAD51.

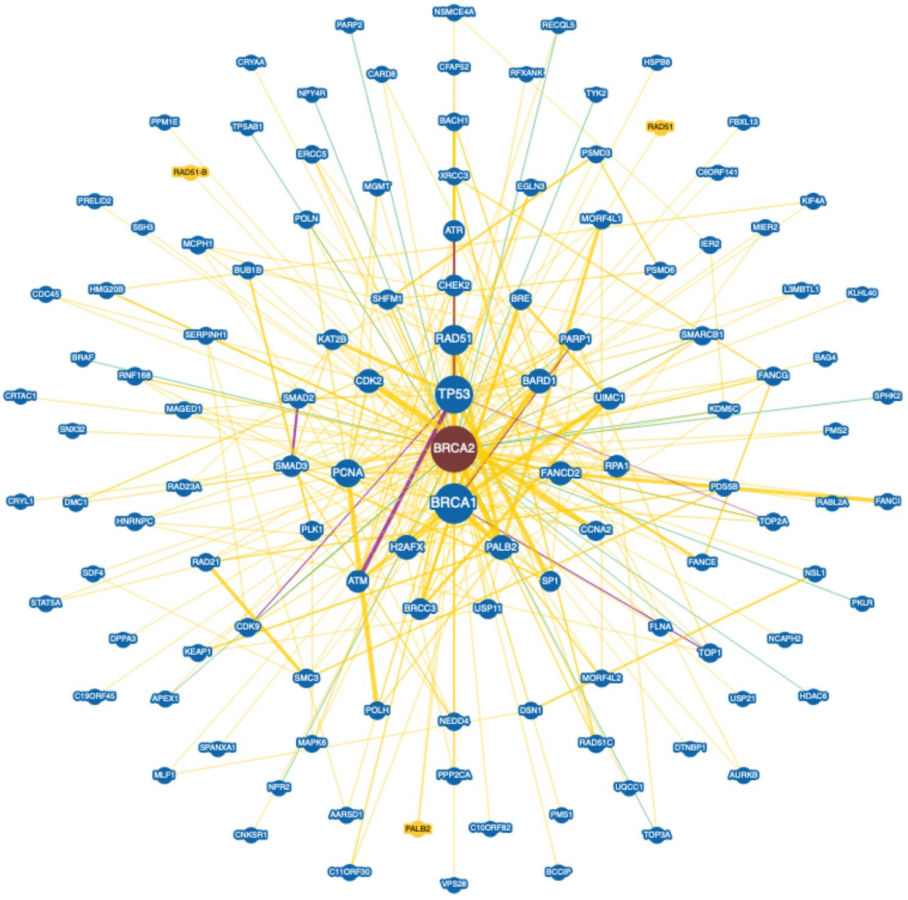


Figura : Red de interacciones con nivel mínimo de evidencia 1.

Es interesante observar cómo a medida que se filtra el mínimo de evidencia se obtiene como resultado una red mucho más compacta, como la que se observa a continuación, con un mínimo de 15, en donde se pueden observar las interacciones más importantes. Del análisis de las publicaciones de PubMed de estos interactores más importantes se puede determinar lo comentado anteriormente, es decir que las interacciones siguen un patrón en torno a la prevención del cáncer.

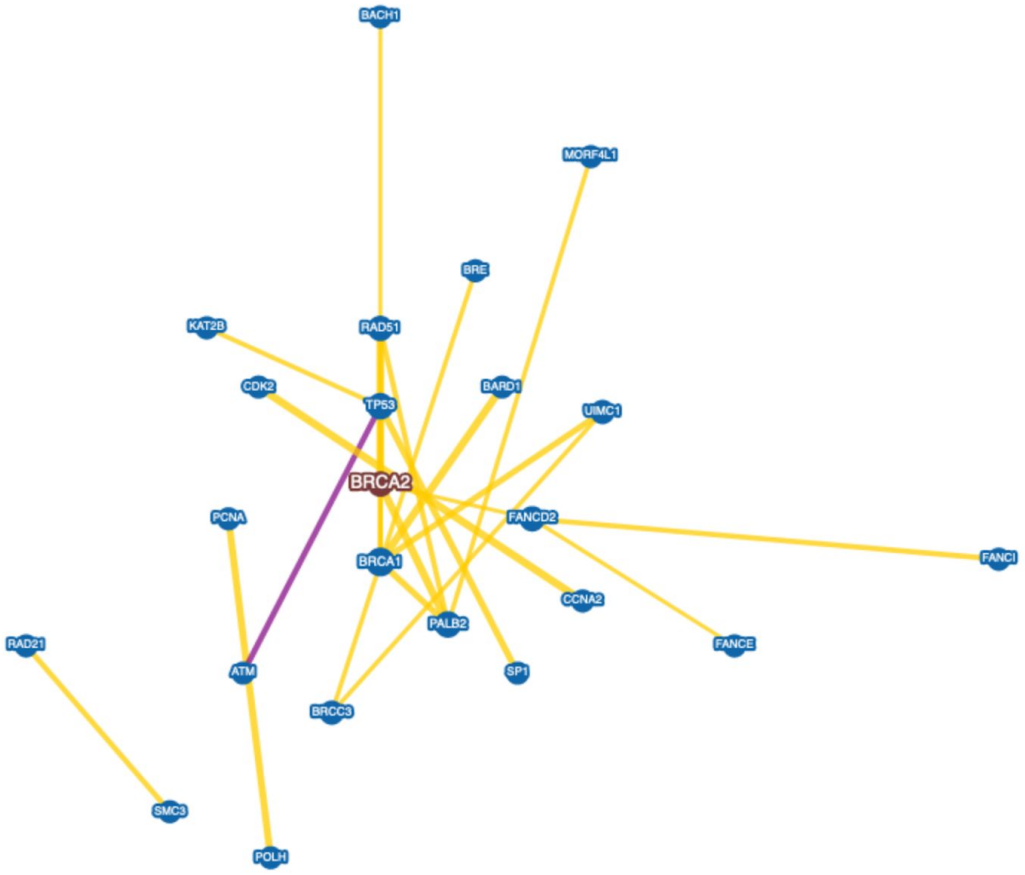


Figura : Red de interacciones con nivel mínimo de evidencia 15.

e) 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.

El componente celular del que forma parte BRCA2 es **BRCA2-BRAF35 complex**. Este componente celular suele estar asociado con la cromatina condensada durante la mitosis y está formado por las proteínas BRCA2 and BRAF35.

También forma parte de **BRCA2-MAGE-D1 complex**, que según se explica Gene Ontology puede mediar las actividades sinérgicas de las dos proteínas para regular el crecimiento celular y está formado por las proteínas BRCA2 and MAGE-D1.
A continuación se listan las funciones moleculares y los procesos biológicos en los que participa (UniProt).

GO - Molecular function [i]

- gamma-tubulin binding — Source: UniProtKB
- H3 histone acetyltransferase activity — Source: UniProtKB
- H4 histone acetyltransferase activity — Source: UniProtKB
- identical protein binding — Source: IntAct
- protease binding — Source: UniProtKB
- protein C-terminus binding — Source: MGI
- single-stranded DNA binding — Source: UniProtKB

Complete GO annotation on QuickGO ...

GO - Biological process [i]

- brain development — Source: Ensembl
- cell aging — Source: Ensembl
- centrosome duplication — Source: UniProtKB
- DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator — Source: Ensembl
- double-strand break repair — Source: UniProtKB
- double-strand break repair via homologous recombination — Source: UniProtKB
- establishment of protein localization to telomere — Source: BHF-UCL
- female gonad development — Source: Ensembl
- hemopoiesis — Source: Ensembl
- histone H3 acetylation — Source: UniProtKB
- histone H4 acetylation — Source: UniProtKB
- inner cell mass cell proliferation — Source: Ensembl
- intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator — Source: Ensembl
- male meiosis I — Source: Ensembl
- mitotic cytokinesis — Source: UniProtKB
- mitotic recombination-dependent replication fork processing — Source: BHF-UCL
- negative regulation of mammary gland epithelial cell proliferation — Source: UniProtKB
- nucleotide-excision repair — Source: UniProtKB
- oocyte maturation — Source: Ensembl
- positive regulation of mitotic cell cycle — Source: Ensembl
- positive regulation of transcription, DNA-templated — Source: UniProtKB
- regulation of cytokinesis — Source: Ensembl
- replication fork protection — Source: Ensembl
- response to gamma radiation — Source: Ensembl
- response to UV-C — Source: Ensembl
- response to X-ray — Source: Ensembl
- spermatogenesis — Source: Ensembl
- telomere maintenance via recombination — Source: Ensembl

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

Una vía metabólica es una serie de reacciones químicas que se producen dentro una célula que puede ser de tipo anabólica o catabólica. El gen participa en 4 pathways (resultado de la búsqueda en la base de datos de KEGG.

Se muestran los diagramas correspondientes a continuación:

Cáncer de páncreas

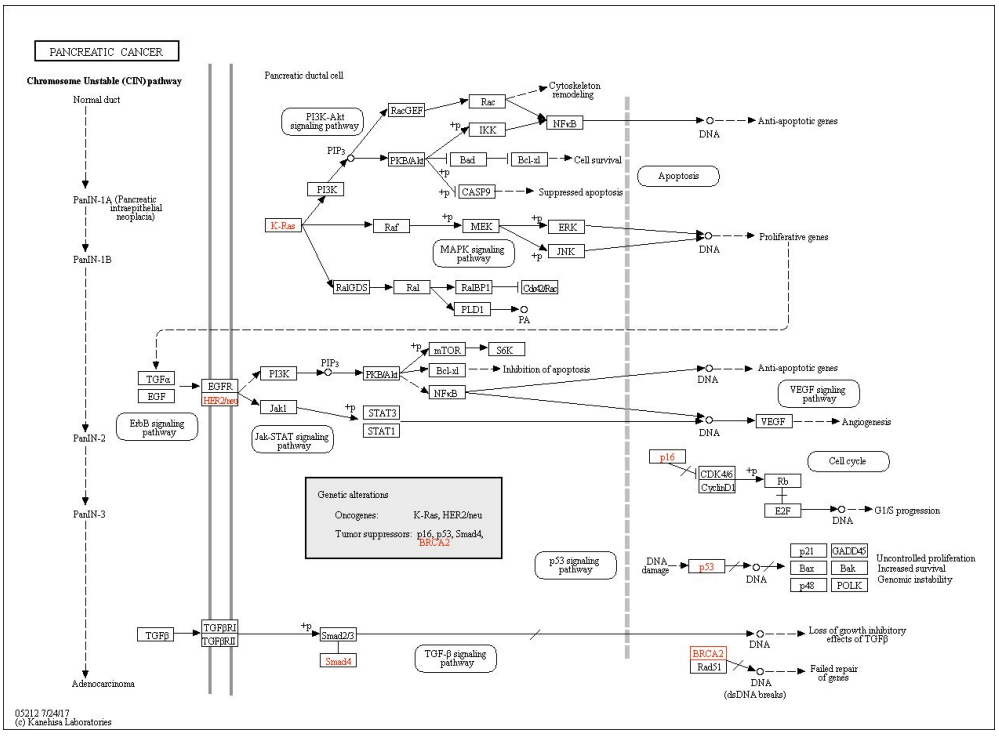


Figura : Mapa de vías metabólicas en el cáncer de páncreas

Recombinación homóloga

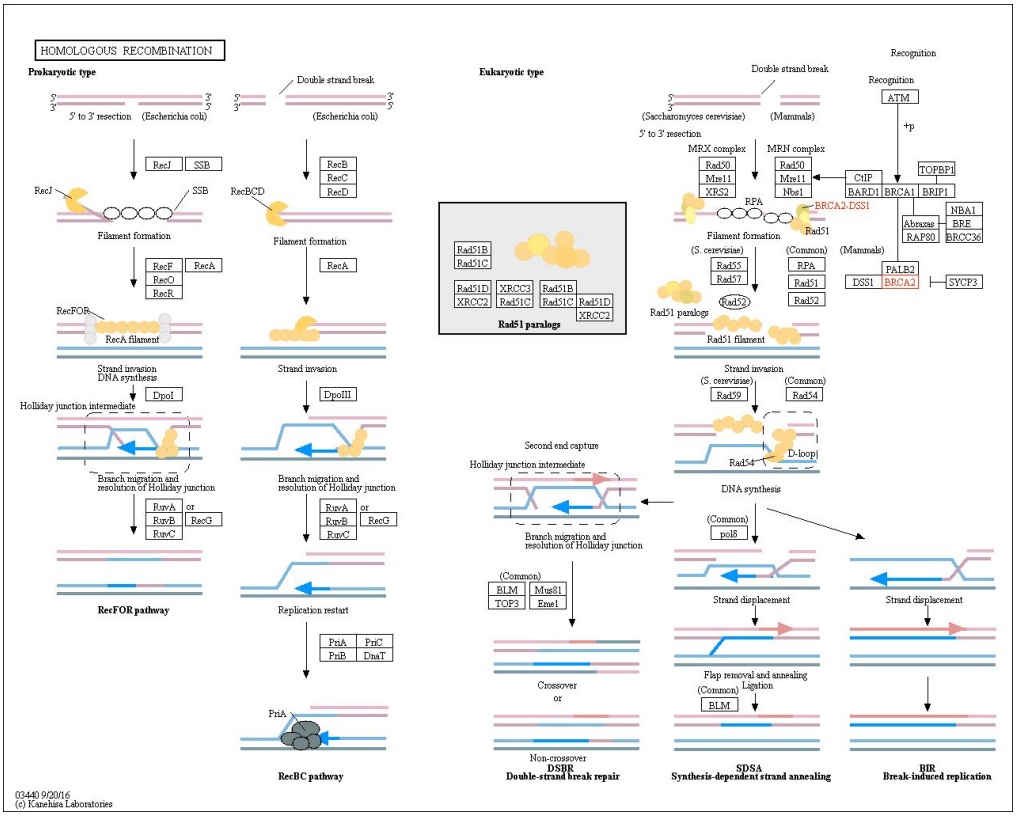


Figura : Mapa de vías metabólicas de recombinación homóloga

Cáncer de mama

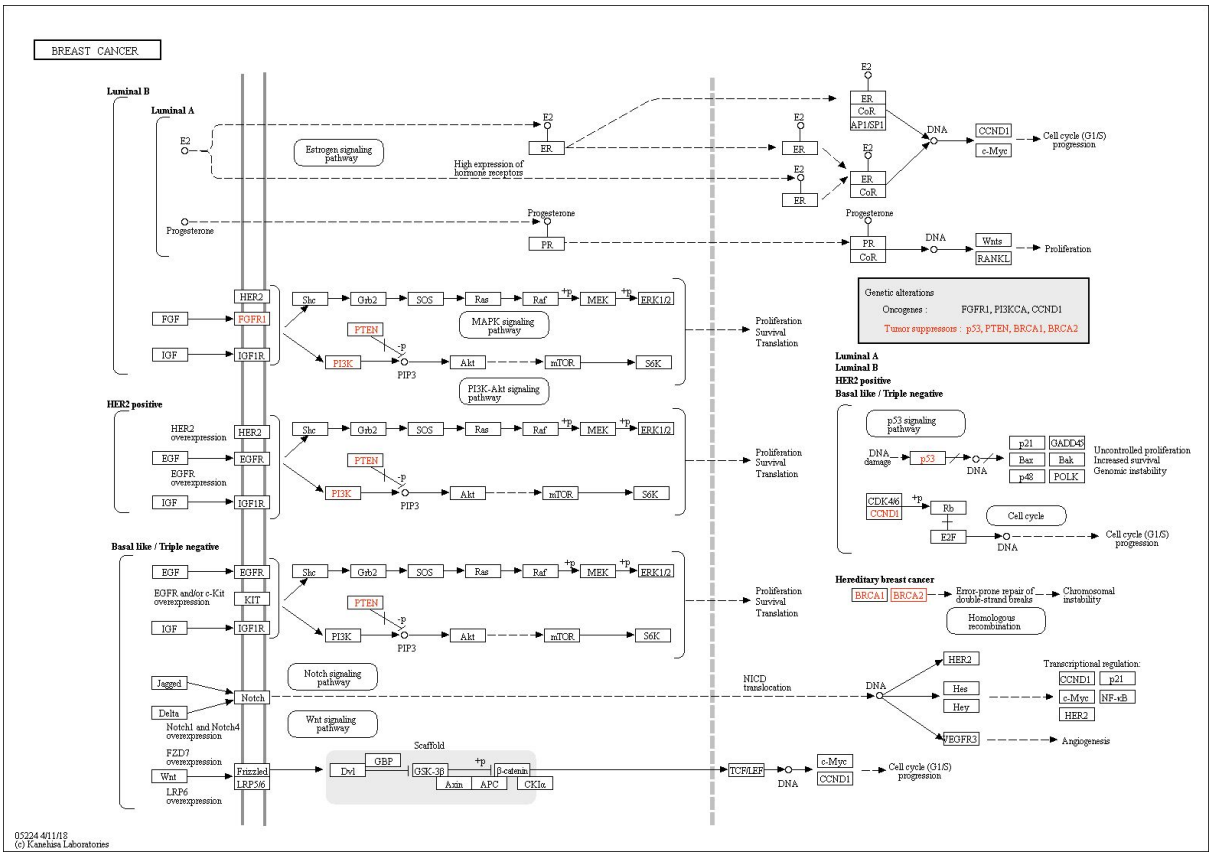


Figura : Mapa de vías metabólicas de cáncer de mama

Anemia de Fanconi

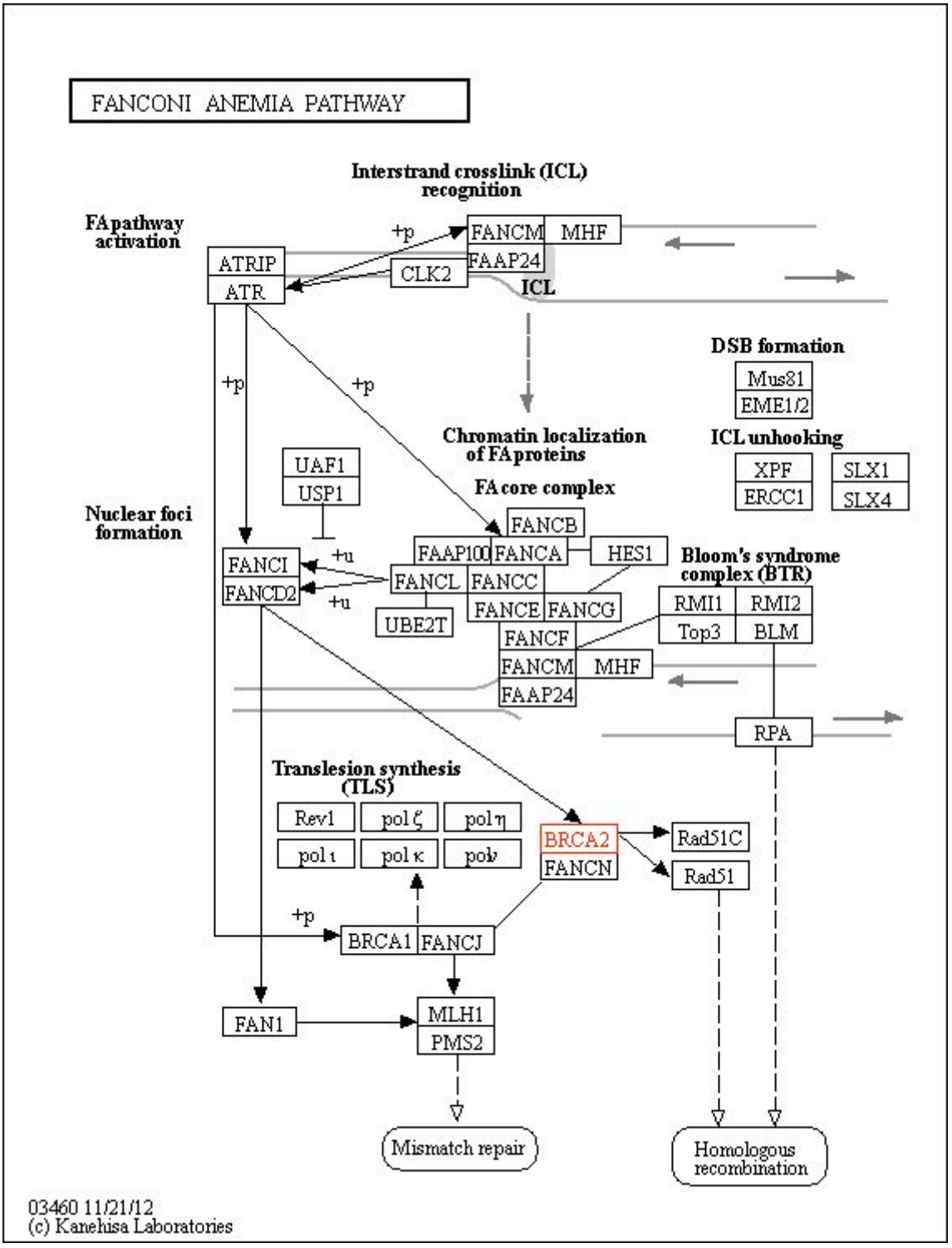


Figura : Mapa de vías metabólicas de la Anemia de Fanconi

Vías metabólicas en Cáncer

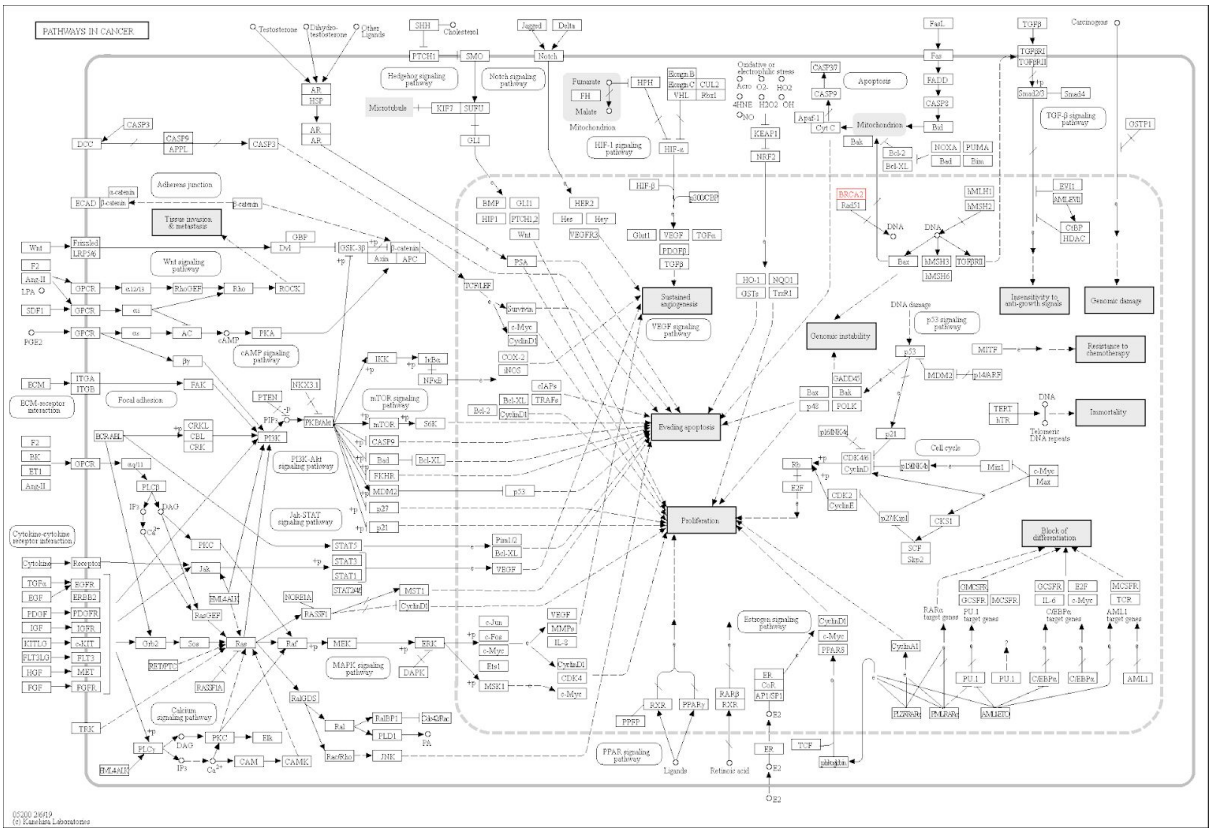


Figura : Mapa de vías metabólicas en el cáncer

BRCA1 y BRCA 2 participan principalmente, como se especificó anteriormente, en la reparación de ADN dañado. Dentro de esta función BRCA1 se desempeña en variadas cuestiones a diferencia de BRCA2, cuya actividad fundamental es la mediar el reclutamiento de la recombinasa RAD51 a DSB. El reclutamiento de RAD51 no solo es esencial para la recombinación homóloga, sino que también es responsable de la función supresora de tumores de este proceso de reparación.

g) Entrar en la base de datos de variantes genéticas dbSNP e intentar interpretar o encontrar info 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?

La búsqueda arroja 25750 resultados. Examinaremos en detalle la variante más común, rs80359550, que es aquella asociada al cáncer de mama. La mayoría de las variantes presentan una probabilidad de ocurrencia muy baja, menor al 3%.

Se estima que el cáncer de mama se da en 1 de cada 8 mujeres, con mayor frecuencia en algunos grupos étnicos que en otros.

Dicha variante también está asociada al cáncer de ovario y también se han conocido diversos desórdenes relacionados con esta variante (reportados en la base de datos de NCBI). Esta variante tiene como longitud un par de bases y su localización exacta es 13q13.1 y corresponde a una deleción.

Particularmente se ha encontrado que la variante se halla en las personas de raza Judía Asquenazí con mayor frecuencia que lo común.