IDENTIFICACIÓN DE GENES CON PAPEL DRIVER EN CÁNCER DE MAMA Y SU RED GÉNICA

Patricia Asensio Calavia
Máster en bioinformática y bioestadística
Estudio genómico del cáncer

Nombre Consultor/a: Laia Bassaganyas
Nombre PRA: José Antonio Morán Moreno
05/06/2018


Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

B) GNU Free Documentation License (GNU FDL)

Copyright © 2018 Patricia Asensio Calavia.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.
A copy of the license is included in the section entitled "GNU Free Documentation License".

C) Copyright

© (Patricia Asensio Calavia)
Reservados todos los derechos. Está prohibido la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la impresión, la reprografía, el microfilme, el tratamiento informático o cualquier otro sistema, así como la distribución de ejemplares mediante alquiler y préstamo, sin la autorización escrita del autor o de los límites que autorice la Ley de Propiedad Intelectual.

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Descripción del trabajo*
Nombre del autor:	*Patricia Asensio Calavia*
Nombre del consultor/a:	*Laia Bassaganyas*
Nombre del PRA:	*José Antonio Morán Moreno*
Fecha de entrega (mm/aaaa):	06/2018
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Estudio genómico del cáncer*
Idioma del trabajo:	*castellano*
Palabras clave	*Breast cancer, Driver genes, Bioinformatics tools.*

Resumen del Trabajo (máximo 250 palabras):

La investigación del cáncer de mama es importante hoy en día porque es el más frecuente en mujeres, se estima que el 12% de ellas se verán afectadas alguna vez en sus vidas. El objetivo del presente estudio es identificar posibles genes drivers, la interacción entre ellos y sus funciones, con la finalidad de entender mejor cómo pueden influir en la enfermedad, ayudar en el diagnóstico y diseño de nuevos fármacos. A partir de 990 datos mutagénicos de pacientes con cáncer de mama invasivo se identifican 239 genes drivers putativos con la herramienta DriverDBv2. A partir de estos genes se crea una red de interacción con el software FunRich, donde 9 genes aparecen como nodos con un mayor número de interacciones: EGFR, TP53, BRCA1, FLNA, EP300, ERBB3, AKT1, ERBB2 y PIK3R1. Después, Se realiza un análisis funcional con PANTHER dando como resultado un mayor porcentaje de genes con función molecular de unión y catalítica. Las vías de acción con un mayor número de posibles genes drivers son: la vía de señalización Wnt (6.7%), señalización de receptor EGF (6.7%) y la vía del receptor hormonal de gonadotropina (6.7%). Hay varios estudios sobre genes drivers del cáncer de mama pero aún no se ha consensuado un método de identificación. Los resultados muestran algunos posibles nuevos genes driver identificados, además de su clasificación mayoritaria como genes con función receptor. Finalmente, estos resultados podrían servir como base para futuras investigaciones en busca de dianas farmacológicas, mejora de diagnóstico y medicina personalizada.

Abstract (in English, 250 words or less):

Breast cancer research is important nowadays because it is the most common cancer in women, and it is estimated that 12% of women will be affected along their lives. The aim of this study is to identify possible driver genes, the interaction between them and their functions, in order to understand how they can influence the disease, help in the diagnosis and design of new drugs. 239 putative driver genes are identified with the DriverDBv2 tool, from the initial 990 mutagenic data from patients with invasive breast cancer. An interaction network is created with the FunRich software from identified genes, 9 of them appear as nodes with a greater number of interactions: EGFR, TP53, BRCA1, FLNA, EP300, ERBB3, AKT1, ERBB2 and PIK3R1. A functional analysis is then performed with PANTHER, presenting a higher percentage of genes with binding and catalytic molecular function. The pathways with the greatest number of possible driver genes involve are: Wnt signalling pathway (6.7%), EGF receptor signalling pathway (6.7%) and Gonadotropin-releasing hormone receptor pathway (6.7%). There are several studies on breast cancer driver genes but no method of identification has been reached agreement yet. The results show some possible new driver genes identified, in addition, receptor function classification genes. Finally, these results could serve as a basis for future research of pharmacological targets, to improve diagnosis and personalized medicine.

Índice

Lista de figuras y tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

El cáncer de mama es el tumor más frecuente en las mujeres occidentales. Solo en Estados Unidos se estimó que más de 180.000 mujeres y 2.000 hombres serían diagnosticados con cáncer de mama y 40.000 personas morirían por la enfermedad en 2009. Según varios estudios cerca de 12% de las mujeres de la población padecerán cáncer de seno alguna vez en sus vidas. Aunque también las muertes por cáncer de mama han disminuido constantemente desde la década de 1990 debido a la detección temprana a través del uso de mamografías y a las mejoras en el tratamiento.[1]

La tasa de incidencia más alta en el mundo está en Norte América y Europa y la menor en África y Asia, aunque la incidencia en China y Japón está creciendo en los últimos tiempos. A pesar de que la incidencia en África sea menor, las mujeres aproximadamente son 10 años más jóvenes a la hora de su diagnóstico con estados más avanzados de la enfermedad y mayor tasa de mortalidad comparada con las mujeres occidentales. También, se encuentran diferencias entre mujeres de distintas regiones del mismo país, pero no se ha llegado a poder demostrar ninguna hipótesis de variabilidad que lo explique. Esta disparidad de datos sugiere una compleja interacción entre los genes y el ambiente.[1]

En este estudio se va a investigar el cáncer de mama invasivo. Hay varios tipos de cáncer de mama invasivo que incluyen aquellos que se han diseminado al tejido mamario circundante:

- Carcinoma ductal infiltrante (80%)
- Carcinoma lobulillar infiltrante o invasor (10%)
- Carcinoma medular (5%)
- Carcinoma mucinoso o coloide (2%)
- Carcinoma papilar infiltrante (2%)
- Carcinoma tubular (2%)

El programa Cancer Genome Atlas (TCGA) se centra principalmente en dos tipos de cáncer de mama invasivo: carcinoma ductal y carcinoma lobular. El carcinoma ductal invasivo es el tipo más común de cáncer de mama y se desarrolla en los conductos lácteos de la mama. Por otro lado, alrededor del 10 por ciento de todos los casos de cáncer de mama avanzado son carcinoma de mama lobular invasivo. Este cáncer se desarrolla en los lóbulos o glándulas productoras de leche materna.[2]

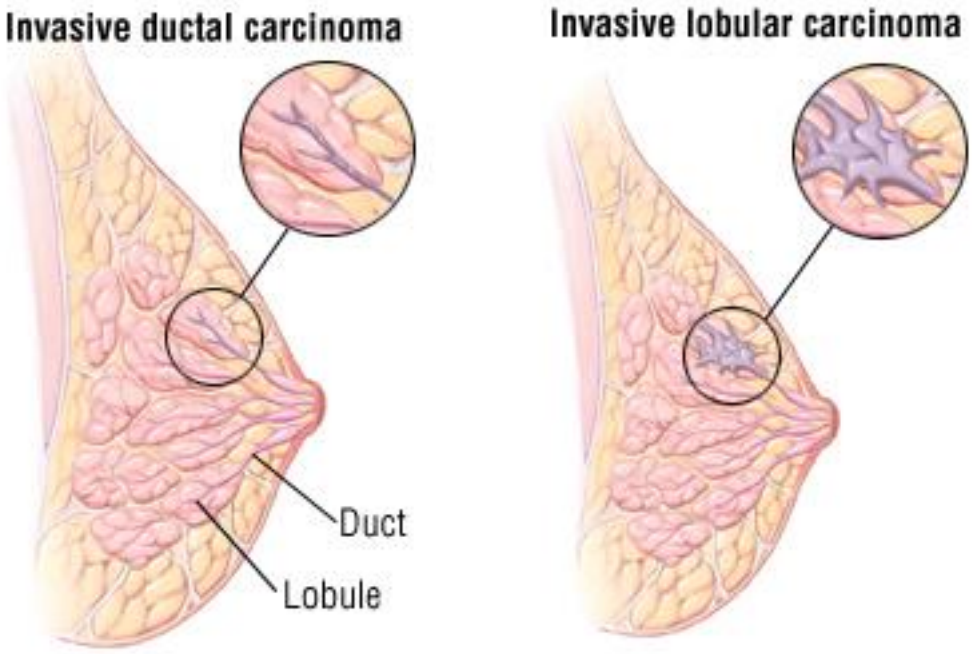


Figura 1: Tipos de tumores invasivos en el cáncer de mama.[3]

Carcinoma ductal invasivo

El carcinoma ductal invasivo (CDI), también denominado carcinoma ductal infiltrante, es el tipo más común de cáncer de mama.
Este tipo de cáncer comienza en los conductos lácteos, que son las vías que transportan la leche desde los lobulillos hasta el pezón. En la Figura 1 se puede ver la distribución de estos conductos donde se puede formar el tumor. La palabra invasivo significa que el cáncer atraviesa la pared del conducto lácteo y comienza a invadir los tejidos de la mama. Con el paso del tiempo este tipo de tumor invasivo puede llegar a propagarse hacia los ganglios linfáticos y otras zonas del cuerpo.

Según estudios realizados por la Sociedad Americana del Cáncer, anualmente en los Estados Unidos diagnostican cáncer de mama invasivo a más de 180.000 mujeres, la mayoría carcinoma ductal invasivo.
Además, los análisis muestran que la prevalencia de este tipo de carcinoma en mujeres es mayor a partir de los 55 años. Aunque puede afectar a mujeres de cualquier edad y también a hombres.[4]

Carcinoma lobular invasivo

El carcinoma lobular invasivo (CLI), también denominado como carcinoma lobular infiltrante, es el segundo tipo de cáncer de mama más común después del carcinoma ductal invasivo. Cerca de un 10 % de los casos de cáncer de mama invasivo diagnosticados en Estados Unidos son carcinomas lobulares invasivos.
Este tipo de cáncer tiene su comienzo en los lobulillos productores de leche, los cuales a su vez vacían su contenido en los conductos lácteos que llevan la leche hasta el pezón. Al igual que el carcinoma ductal

invasivo el tumor atraviesa, esta vez, la pared del lobulillo e invade los tejidos colindantes de la mama hasta incluso poder llegar a los ganglios linfáticos u otras zonas.

Los CLI tienden a aparecer en edades más avanzadas que los carcinomas ductales invasivos: alrededor de los 60 años.
Algunas investigaciones sugieren que el uso de terapias de reemplazo hormonal durante y después de la menopausia pueden aumentar el riesgo de desarrollar un CLI.[5]

Genética del Cáncer de Mama

La mayoría de las mutaciones drivers ocurren a nivel somático, mientras que un pequeño número de mutaciones se transmiten a la descendencia. Se cree que entre el 5 % y el 10 % de los cánceres de mama son hereditarios, causados por genes anormales que se transmiten de padres a hijos. La mayoría de los casos hereditarios están relacionados con dos genes presentes en el genoma humano que presentan anomalías: BRCA1 y BRCA2. La función de los genes BRCA es reparar el daño celular y mantener el crecimiento regular de las células mamarias, ováricas y de otros tipos. Sin embargo, cuando estos genes contienen anomalías o mutaciones que se transmiten de una generación a otra, no funcionan normalmente, y el riesgo de cáncer de mama, de ovario y de otros tipos aumenta.
La mujer promedio en los Estados Unidos tiene un riesgo de 12% de desarrollar cáncer de mama en su vida. Mientras que las mujeres que tienen una anomalía de los genes BRCA1 y/o BRCA2 pueden tener hasta un 80 % de riesgo de ser diagnosticadas con cáncer de mama.

Los genes con papel driver son aquellos genes mutados que presentan variaciones de un solo nucleótido (SNPs), variaciones en el número de copias, etc y proporcionan una ventaja a las células para la supervivencia del tumor. [6]

Muchos son los estudios alrededor de la identificación de genes drivers del cáncer de mama. Por ejemplo, encontramos The genetics home reference, la cual sugirió una lista de genes drivers con alta asociación con los tumores de mama: BARD1, BRCA1, BRCA2, CASP8, CHEK2, CTLA4, CYP19A1, FGFR2, H19, LSP1, MAP3K1, MRE11A, RAD51C, STK11, TERT, TOX3, XRCC2, GATA3, PIK3CA, AKT1, CDH1, RB1, TP53, PTEN y XRCC3.[7]
Por otro lado, Stephens et al. (2012) indicaron los siguientes genes envueltos en la carcinogénesis e la mama: *TP53, ERBB2, GATA3, FGFR1, CCND1, ARID5B, CDH1, CTCF, HDAC9, KDM5B, NCOR2, SETD1A y SXL2*.[8]
Como podemos observar, cada uno de los datos publicados sobre los impulsores del cáncer publica un conjunto distinto de genes drivers con alguna coincidencia. Hasta el momento, no se ha desarrollado un procedimiento estándar para identificar y validar los genes impulsores del cáncer de mama.

También podemos ver como la mutación de unos genes puede afectar a otros y provocar así la proliferación celular incontrolada. Por ejemplo, el gen supresor tumoral mutado, PALB2, afecta en gran medida al BRCA2 y aumenta el riesgo de cáncer de mama.
Además, varios estudios han asociado también al cáncer de mama los genes: CDH1, STK11, PALB2, CHEK2, BRIP1, CDKN2A, CTNNB1, MLH1, MSH2, MSH6, NBN, RAD50, RAD51, TP53, entre otros.[9]

Para combatir el cáncer se están estudiando posibles dianas genéticas para nuevos fármacos. Estas dianas genéticas se basan en los estudios actuales que hay sobre vías de actuación (pathways) que afectan a los genes con papel driver.[10]

Una de las vías más importantes hasta el momento ha sido la de **PI3K/AKT/mTOR** (Figura 2). La activación oncogénica de la vía PI3K puede ocurrir a través de una variedad de mecanismos como mutaciones, amplificación de genes que codifican RTK, subunidades de PI3K, AKT o isoformas activadoras de RAS, pérdida de la expresión de PTEN, etc. que involucran otros genes drivers identificados.[11]

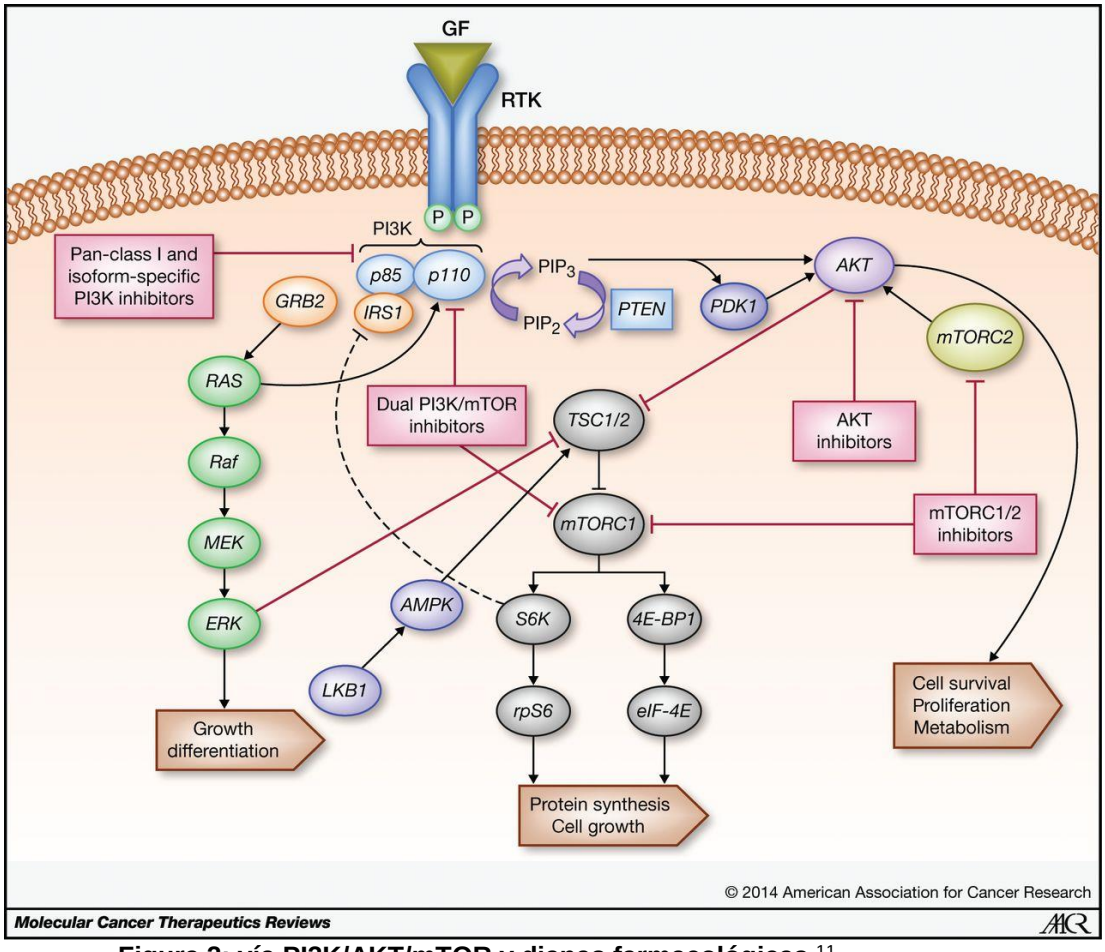


Figura 2: vía PI3K/AKT/mTOR y dianas farmacológicas.[11]

Otra de las vías relacionada, es la de señalización de receptores de estrógeno. Los receptores de estrógeno **ERs** dependen de ligandos que regulan los genes responsables de la diferenciación, apoptosis y proliferación celular. Por tanto, las mutaciones en esta vía se asocian a la invasión del tumor de mama. Los genes implicados con mayor relevancia son: ESR1, PGR y AR. En la Figura 3 podemos ver como esta vía a su vez interacciona con otras implicadas en esta enfermedad como la PI3K/AKT/mTOR o la de MAPK.[12]

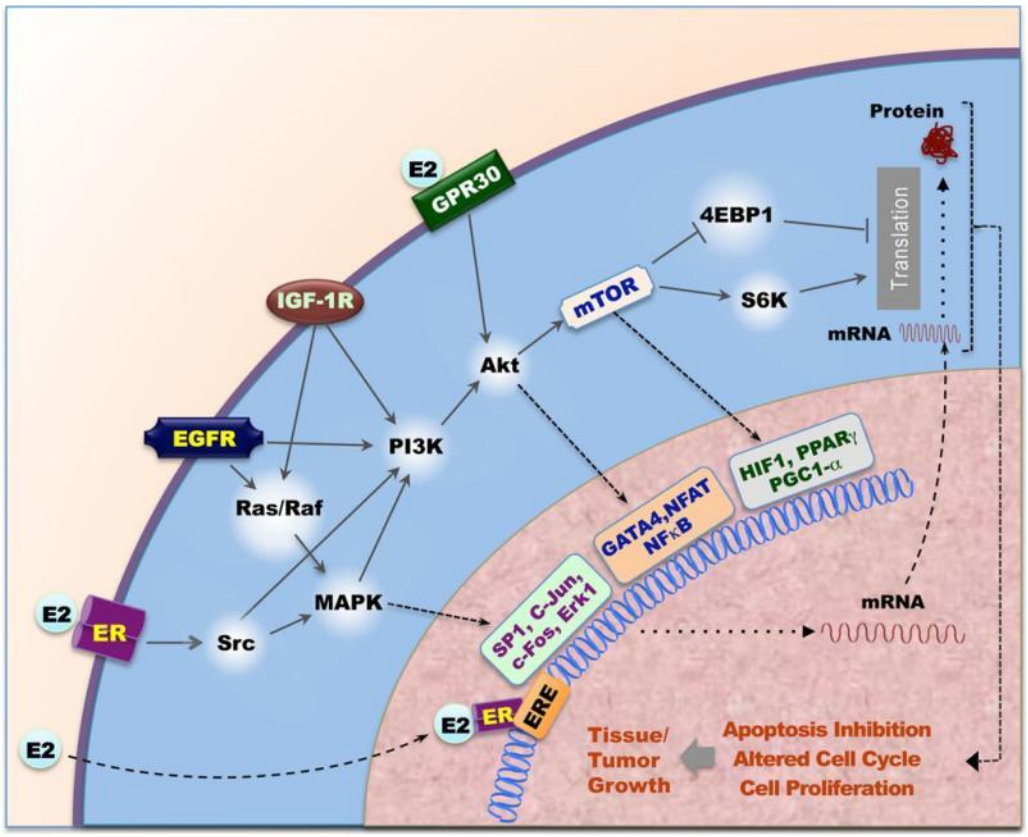


Figura 3: Vía de señalización del receptor de estrógeno.[12]

La siguiente vía de la que vamos a hablar es la de los **receptores tirosina-quinasa** que regulan los factores de crecimiento. Algunos de los genes implicados son: ERBB2, FGFR1 y FGFR2. Su activación está relacionada con las vías de señalización MAP quinasa, JAK/STAT y PI3K/AKT1/MTOR. Esta vía es importante ya que regula procesos como el crecimiento celular, proliferación, diferenciación, transcripción, regulación metabólica o supervivencia celular. En la Figura 4 podemos ver un esquema de su funcionamiento.[13]

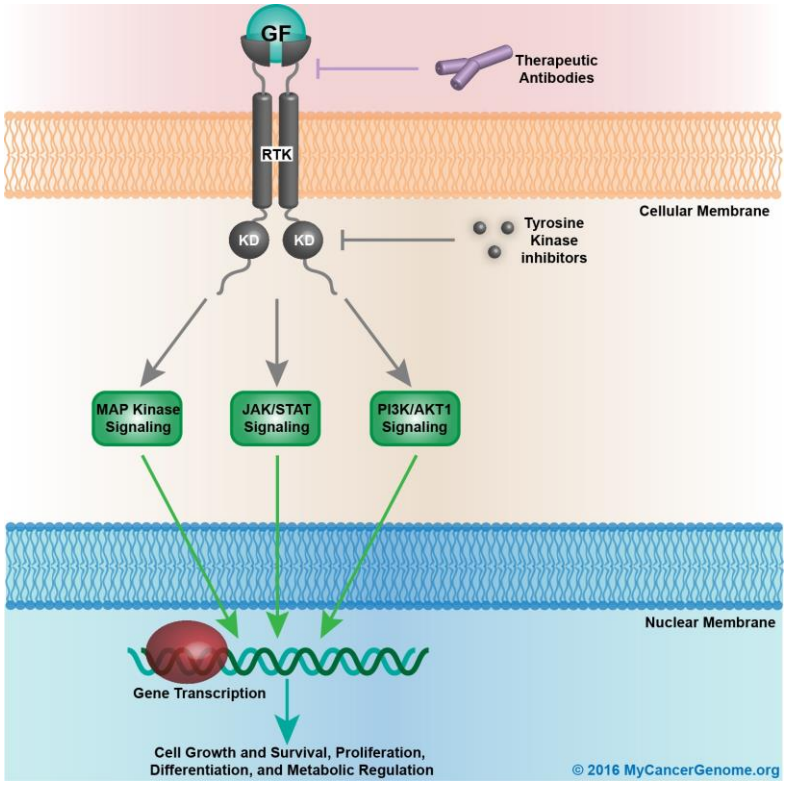


Figura 4: vía de actuación receptor tirosina-quinasa.

Finalmente, hablaremos de la vía que regula el **ciclo celular y el daño al ADN**. La regulación del ciclo celular depende de unos puntos de control que previenen el daño de ADN o un ciclo anormal. Los genes implicados en esta vía son: CCND1, CDK4, CDK6, RB1, TP53. En el cáncer de mama suelen encontrarse mutaciones en estos puntos de control, de tal manera que no se repara correctamente el ADN. La Figura 5 muestra un esquema de la vía del ciclo celular.[14]

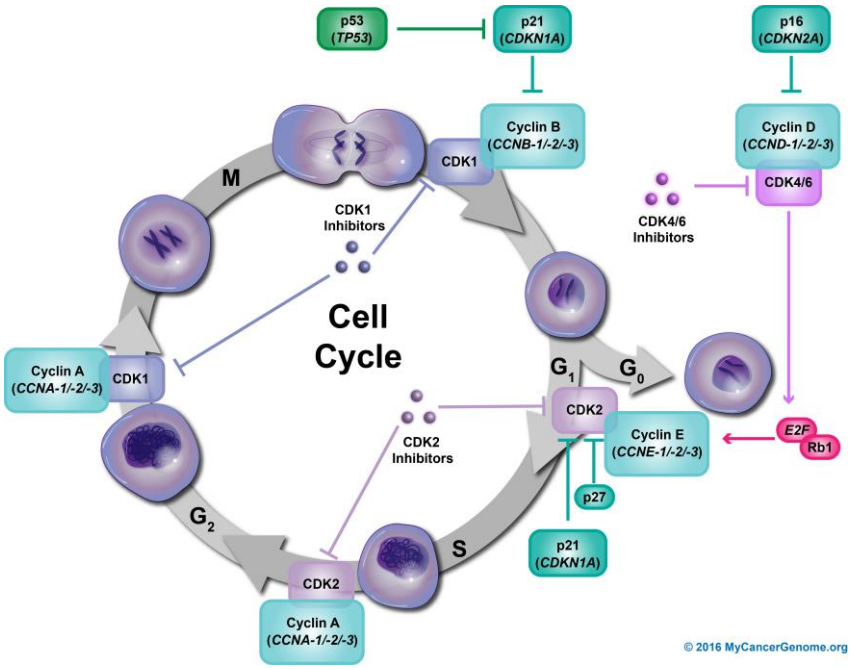


Figura 5: esquema ciclo celular con los genes implicados en los puntos control.[14]

Bioinformática y herramientas computacionales en la investigación

Con la creciente evidencia de que las interacciones, la red génica y de proteínas juegan un papel importante en la investigación de los mecanismos moleculares del cáncer, es necesario e importante introducir la biología de sistemas, la tecnología basada en la ómica, la bioinformática y la ciencia computacional para mejorar el diagnóstico, las terapias y el pronóstico de las enfermedades.

Actualmente hay un gran número de datos y es necesario su análisis. En el caso de las muestras de pacientes con cáncer, el avance en la next generation sequencing y sus técnicas computacionales aliadas, han allanado el camino para identificar un gran número de mutaciones genéticas y sus impactos. De esta forma, con ayuda de la bioinformática, es posible analizar rápidamente una gran cantidad de datos mutagénicos e interpretarlos con ayuda de las bases de datos ya existentes.

No cabe duda de que en los últimos años el avance de la investigación del cáncer de mama, así como de otras enfermedades, está avanzando a pasos agigantados gracias a las herramientas bioinformáticas y software de dominio público. Compartir datos clínicos como hace TCGA agiliza la investigación y el acceso a todos los datos sin necesidad de haberlos obtenido previamente.

La aplicabilidad de métodos, software, herramientas computacionales y bases de datos que pueden utilizarse para explorar los mecanismos moleculares del cáncer e identificar y validar nuevos biomarcadores, y medicina individualizada en el cáncer deben considerarse seriamente.[15]

1.2 Objetivos del Trabajo

El objetivo de este trabajo es estudiar los factores genéticos que influyen en el desarrollo del tumor mamario para que en un futuro se pueda usar como prevención de la enfermedad, un diagnóstico precoz, o para un seguimiento o medicación personalizada a cada paciente. La investigación de vías biológicas afectadas por mutaciones de estos genes nos ayudará a entender los determinantes del inicio, progresión y otras funciones biológicas del cáncer. Incluso también podría ser útil para la creación de nuevos fármacos contra dianas específicas.
Para ello se estudian las SNVs específicas de pacientes, variación en el número de copias, etc. con este tipo de cáncer, qué genes y proteínas afectan, su interacción, funciones y vías de acción afectadas por las mutaciones.

El objetivo principal de este Trabajo Final de Máster es el estudio de los genes mutados en el Cáncer de Mama. En concreto, se investiga cuáles son los posibles genes con papel driver de la enfermedad, sus funciones y la interacción entre ellos, con la finalidad de entender mejor cómo pueden influir estos genes en el Cáncer de Mama.

Para esta investigación se plantean una serie de tareas u objetivos a cumplir:

- Recopilación de datos mutagénicos de Cáncer de Mama y contextualizar el estudio de la enfermedad en la actualidad.
- Identificación de posibles genes con papel driver a partir de datos mutagénicos de pacientes con Cáncer de Mama.
- Creación de un modelo de red de interacción génica con los posibles genes drivers identificados.
- Definición de la función molecular, biológica y vía de actuación de los genes identificados.

1.3 Enfoque y método seguido

Los datos mutagénicos de partida se han obtenido de la base de datos de The Cancer Genome Atlas (TCGA). A partir de ellos, se ha utilizado principalmente la herramienta DriverDBv2, ya que permite hacer un análisis con múltiples herramientas a la vez y poder así comparar los resultados. Estas herramientas son: ActiveDriver, Dendrix, MDPFinder, Simon, NetBox, OncodriveFM, MutSigCV, MEMo, CoMDP, DanRank, DriverNet, e-Driver, iPAC, MSEA y OncodriveCLUST.

Con el objetivo de encontrar la posible red génica con papel driver en el cáncer de mama se hicieron varios análisis con el servidor String. Esta herramienta identifica los genes relacionados a partir de fuentes fiables de experimentación, bases de datos, coexpresión, minería de textos, cercanía, coocurrencia y fusión génica. La información que se puede obtener con esta herramienta es muy amplia, desde el nombre de los genes relacionados hasta la estructura 3D de la proteína, pasando por la predicción de la acción de dos genes relacionados. Esto ayuda a relacionar algunos de los genes entré sí.

Por otro lado, se pensó que a resultar un análisis muy lento y difícil para encontrar una red consenso, se buscó una alternativa, otras herramientas y encontramos el software FunRich. Esta herramienta es capaz de crear una red génica a partir de los IDs de los genes que interesan y los datos de fuentes fiables como Uniprot. Crea una network muy visual donde además se pueden ver los genes que más interaccionan con claridad.

Una vez que tenemos la red génica se investiga la función de estos genes. Para ello se utiliza una herramienta asociada a GeneOntology lo cual asegura que los datos obtenidos del análisis son totalmente fiables. Esta herramienta se llama Protein ANalysis THrough Evolutionary Relationships (PANTHER) Classification System. Tras un análisis bioinformático de los genes muestra una clasificación según su función molecular, biológica, tipo de proteína que sintetizan, localización celular y vía de acción (pathway).

1.4 Planificación del Trabajo

En primer lugar, se necesitan los datos mutagénicos de pacientes con cáncer de mama por lo que la primera tarea es su recopilación. A continuación, se procede con el análisis inicial de los datos, que consiste en identificar los posibles genes con papel driver dentro de los datos mutagénicos que se han obtenido anteriormente. Esto forma parte de la fase 1 del desarrollo de la investigación, al terminar, se obtienen los genes identificados como posibles drivers de este tipo de cáncer y se puede continuar con la siguiente fase. La fase 2 del desarrollo de la investigación parte de los genes identificados y continúa con la creación de un modelo de red génica con los posibles drivers. Más adelante, se clasifican estos genes según su función molecular, biológica y su vía de acción.

Por último, se redacta la memoria y escriben las conclusiones del trabajo. También se realiza una presentación y exposición pública.
Las fechas programadas y la duración para cada tarea se muestran en la Tabla 6. Y en la Figura 7 se puede visualizar las distintas fases por colores y su duración.

Nombre de la tarea	Fecha de inicio	Fecha final	Duración (dias)
1- Propuesta de TFM	21/02/2018	05/03/2018	12
2- Plan de trabajo	06/03/2018	19/03/2018	13
3- Recuperar datos mutagénicos	20/03/2018	06/04/2018	17
4- Identificación posibles genes drivers	07/04/2018	19/04/2018	12
5- Red génica	20/04/2018	06/05/2018	16
6- Función génica	07/05/2018	21/05/2018	14
7- Conclusiones y redacción	22/05/2018	05/06/2018	14
8- Elaboración de la presentación	06/06/2018	13/06/2018	7
9- Presentación pública	14/06/2018	25/06/2018	11

Tabla 6: Fechas programadas para cada tarea del Trabajo Final de Máster.

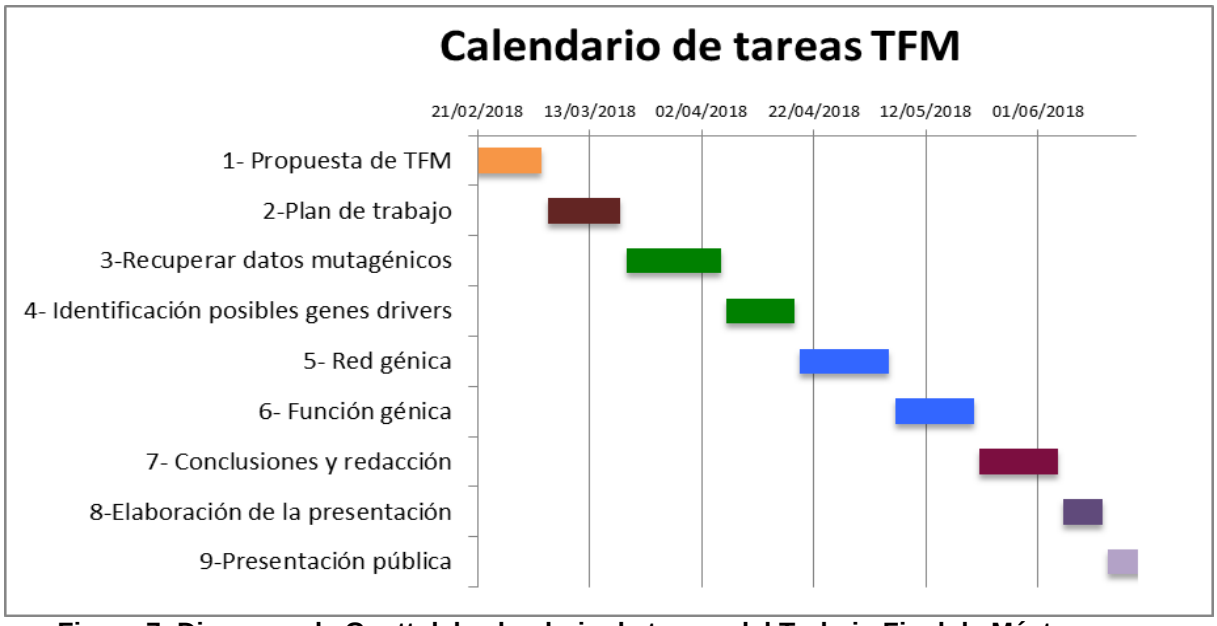


Figura 7: Diagrama de Gantt del calendario de tareas del Trabajo Final de Máster.

Los colores de las barras de la Figura 7 representan cada una de las fases de entrega de las PEC o pruebas de evaluación. La barra naranja representa la PEC0–Propuesta de TFM; la granate representa la PEC1-Plan de trabajo; verde es la PEC2- Desarrollo del trabajo fase 1; azul PEC3-Desarrollo del trabajo fase 2; fucsia PEC4-Redacción de la memoria; y por último la barra morada y lila representan la PEC5 a y b que son la elaboración de la presentación y exposición pública.

1.5 Breve sumario de productos obtenidos

Productos obtenidos del Trabajo Final de Máster del área “Estudio genómico del cáncer”:

-Propuesta de TFM
-Plan del trabajo PEC1
-Informes de seguimiento PEC2 y PEC3
-Memoria final del trabajo
-Presentación del TFM

Como resultados adjuntos a la investigación se han obtenido: la lista de posibles genes drivers en cáncer de mama, esquema de red de interacción génica y clasificación funcional de genes identificados.

1.6 Breve descripción de los otros capítulos de la memoria

En primer lugar, se encuentra el capítulo de Introducción, donde se ponen de maniefiesto los avances que hay hasta la fecha del tema en cuestión, investigación sobre la genética del cáncer de mama, vías de actuación de los genes y la bioinformática. En este capítulo se explica la problemática del tema y la importancia de su estudio. Además, contiene los objetivos del Trabajo y la planificación del mismo.
En el segundo capítulo, se habla sobre los materiales y métodos utilizados en la investigación. Tanto los datos de partida como software y herramientas que han servido de ayuda en el presente trabajo.

En tercer lugar, tenemos los resultados y discusión donde se trata de recopilar los datos obtenidos tras el análisis y relacionarlos con otros trabajos o información recopilada de artículos o revistas. Los resultados se dividen a su vez en tres subapartados: identificación de genes con posible papel driver, diseño de red génica, y análisis funcional de genes drivers.
En este punto se da una opinión objetiva del estado de la investigación y de la posible vía futura para continuar.

Por último, se muestran las conclusiones del trabajo, un glosario y bibliografía.

2. Material y métodos

2.1 Datos de partida y metodología

Utilizamos los datos mutagénicos de 990 muestras de pacientes de cáncer de mama invasivo de la base de datos The Cancer Genome Atlas (TCGA). El conjunto de datos de TCGA está disponible públicamente y es utilizado ampliamente por la comunidad científica. Este proyecto forma parte de una colaboración entre el Instituto Nacional de Investigación del Genoma Humano (NHGRI) y el Instituto Nacional del Cáncer (NCI) que cuenta con datos de 33 tipos de cáncer distintos.

Con estos datos se realizó un análisis de identificación de posibles genes con papel driver en el cáncer de mama. Esta identificación consiste en utilizar varias herramientas bioinformáticas y bioestadísticas capaces de distinguir dónde se encuentran las mutaciones y clasificar los genes humanos según la frecuencia de estas mutaciones. Utilizamos el software DriverDBv2 [16] que es la nueva versión actualizada de DriverDB en la cual se han incorporado más de 9500 RNA-seq datasets de cáncer y más de 7000 exome-seq datasets procedentes de TCGA, ICGC y publicaciones.

A continuación, se constituyó una red génica con los posibles genes drivers ya identificados, con ayuda de la herramienta FunRich y su fuente de datos. Esta network nos sirvió además para identificar aquellos posibles genes con papel driver que tenían mayor número de interacciones entre ellos. Es decir, aquellos genes mutados con una posible mayor influencia que otros en el cáncer de mama.

Por último, para completar nuestro estudio se clasificaron los genes drivers según su función y su vía de acción para poder entender mejor el papel de los genes mutados en el desarrollo del cáncer invasivo de mama.

2.2 Software utilizado

DriverDB

DriverDB es un software multiherramienta que utiliza 15 métodos computacionales diferentes para identificar genes drivers del cáncer. Analiza los datos provenientes de un dataset dado o también de fuentes como TCGA o IGCG entre otras.
Concretamente, tiene disponibles tres datasets de cáncer de mama que son: Breast invasive carcinoma (TCGA, US); Breast triple negative/lobular cáncer (ICGC); y Primary triple negative breast cáncer (British Columbia Cancer Research Center). Como hemos dicho anteriormente en este trabajo estudiamos los datos provenientes del cáncer invasivo de mama utilizando los datos obtenidos a partir de TCGA.
Las herramientas en las que se basa este multianálisis son: MDPFinder, Simon, NetBox, OncodriveFM, MutSigCV, MEMo, CoMDP, DawnRank, DriverNet, e-Driver, iPAC, MSEA y OncodriveCLUST.
En resumen, con este software y con la base de datos de TCGA obtenemos la identificación de genes driver al menos por tres de las sub-herramientas. Cada una de estas sub-herramientas utiliza un método o algoritmo para la identificación de genes con papel driver. Con ello, conseguimos ampliar nuestro espectro de posibles genes driver para estudiar.

El criterio de identificación de genes con papel driver de cada herramienta se basa en lo siguiente:

- **ActiveDriver** → Se centra en los loci principales de un gen, como los sitios de fosforilación y el dominio de la kinasa, para predecir los genes drivers. Los genes con un valor de P ajustado por FDR < 0.05 son estadísticamente mutados de forma inesperada en sitios de fosforilación de proteínas o dominios de proteína kinasa.[17]

- **Dendrix** → Encuentra genes reportados en al menos el 10% de los módulos en cualquier K, que es el número de genes en un módulo, pero no puede soportar datos de expresión génica o red de referencia.[18]

- **MDPFinder** → Utiliza el mismo mecanismo de “K” que Dendrix, a diferencia de que también aporta datos de expresión génica.[19]

- **Simon** → Esta herramienta tiene un modelo background de mutaciones que se puede utilizar para todo tipo de cáncer. Además, considera múltiples codones que codifican para el mismo aminoácido.[20] Los genes con un valor de P ajustado por FDR de < 0.05 son mutados estadísticamente.

- **NetBox** → Consulta la red de referencia para encontrar módulos de genes drivers. Los genes están incluidos en todos los módulos. Esta herramienta además proporciona resultados de red que se pueden ver en Cytoscape.[21]

- **OncodriveFM** → Se mide cualquier sesgo a favor de la acumulación de variantes de alto impacto funcional observadas en un gen o grupo de genes para detectar genes conductores candidatos o módulos génicos.[22] Los genes con un valor de P ajustado por FDR < 0.05 son estadísticamente acumulados de variantes con alto impacto funcional.

- **MutSigCV** → Utiliza la frecuencia y el espectro de mutaciones específicas del paciente, así como las tasas de mutaciones background específicas del gen.[23] Los genes con un valor de P ajustado por FDR < 0.05 están estadísticamente mutados.

- **MEMo** → Considera como genes drivers aquellos que contienen mutaciones somáticas y variaciones en el número de copias y proporciona la red de relaciones entre genes drivers.[24] Los genes están incluidos en todos los módulos con un valor de P ajustado por FDR < 0,05.

- **CoMDP** → Descubre las vías de genes co-ocurrentes en la generación del cáncer y su desarrollo. Las vías están formadas por un conjunto de genes mutados con alta cobertura y alta exclusividad. Las mutaciones entre genes de la vía exhiben una co-ocurrencia estadísticamente significativa en muestras de cáncer. CoMDP Obtiene el conjunto óptimo de rutas utilizando un algoritmo eficiente a partir de los datos de los perfiles de mutación. [25]

- **DawnRank** → Clasifica los genes mutados de cada paciente según su potencial para tener papel driver. Esta herramienta requiere el conocimiento de una red de interacción génica, alteraciones genómicas somáticas del paciente, y el perfil de expresión génica. Un gen poseerá una puntuación de impacto más alta si está altamente conectado a genes expresados diferencialmente downstream. [26]

- **DriverNet** → relaciona las aberraciones genómicas con patrones transcripcionales interrumpidos, informados por asociaciones o interacciones conocidas entre genes. Formula asociaciones entre mutaciones y niveles de expresión utilizando un gráfico bipartito donde los nodos representan el estado de la mutación y el conjunto de genes que representan el estado de expresión periférico. Los genes que explican el mayor número de eventos de expresión periféricos son nominados como genes driver putativos, después se aplica la significación estadística. [27]

- **E-Driver** → se basa en la hipótesis de que no todas las regiones funcionales de una proteína dada pueden ser igualmente relevantes para la carcinogénesis. Si este es el caso, debería reflejarse en la distribución de las mutaciones missense a lo largo de la proteína, con regiones bajo selección que muestran enriquecimiento o agotamiento de tales mutaciones comparadas con regiones con mutaciones aleatorias.
  Asumimos que cada mutación es un evento independiente, y que todos los residuos de la proteína tienen la misma probabilidad de ser mutados. Entonces, dado el número total de mutaciones en la proteína y las longitudes de la región podemos calcular la probabilidad de observar mutaciones en la región bajo la hipótesis nula de que las mutaciones se distribuyen aleatoriamente a través de la proteína.[28]

- **iPAC** → Identifica mutaciones somáticas no aleatorias en proteínas utilizando información de la estructura proteica terciaria.

- **MSEA** → Predice los genes conductores del cáncer en base a los patrones del punto caliente de la mutación.[29]
- **OncodriveCLUST** → Identifican los genes drivers por la existencia de mutaciones que provocan ganancia de función en regiones de proteínas específicas, una señal de que esas mutaciones proporcionan una ventaja adaptativa a las células cancerosas y, en consecuencia, se seleccionan positivamente durante la evolución clonal de los tumores.[30]

String

STRING es una herramienta que evalúa las interacciones proteína-proteína dándoles una puntuación indicadora de confianza. Es decir, la herramienta juzga la probabilidad de que una interacción sea verdadera con las evidencias disponibles (puntuación de 0 a 1 siendo 1 la de mayor confianza y 0 un falso positivo).[31]

La puntuación total final es una evaluación de sub-puntuaciones de las distintas evidencias de la interacción proteína-proteína. Se puntúan los experimentos, bases de datos, textmeaning, coexpresión, cercanía, fusión y coocurrencia.

La primera evidencia proviene de experimentos reales en el laboratorio a partir de las principales bases de datos de interacción organizadas en el consorcio IMEx y BioGRID. La segunda evidencia es importada de bases de datos de pathays. En la tercera STRING busca menciones de nombres de proteínas en todos los resúmenes de PubMed, en una colección interna de más de tres millones de artículos de texto completo, y en otras colecciones de texto. A los pares de proteínas se les da una puntuación de asociación cuando se mencionan frecuentemente juntos

en el mismo trabajo, resumen o incluso oración (en relación con la frecuencia con la que se mencionan por separado). La evidencia de coexpresión se basa en los datos de expresión génica provenientes de una variedad de experimentos de expresión, son normalizados, podados y luego correlacionados (similares patrones de expresión tendrán mayor puntuación), además se procesan datos de RNAseq. En cuanto a la localización o cercanía a los genes se les da una puntuación de asociación cuando se observan consistentemente en la cercanía del genoma del otro. En cuanto a la fusión se les da una puntuación a las parejas de proteínas cuando hay al menos un organismo en el que sus respectivos ortólogos se han fusionado en un único gen que codifica las proteínas. Finalmente, STRING evalúa la distribución filogenética de ortólogos (coocurrencia) de todas las proteínas de un organismo dado. Si dos proteínas muestran una alta similitud en esta distribución, es decir, si sus ortólogos tienden a ser observados como "presentes" o "ausentes" en los mismos subconjuntos de organismos, entonces se asigna una puntuación de asociación.
Otra contribución importante de las interacciones en STRING proviene de la transferencia de evidencia de un organismo a otro. Esta transferencia se llama 'interolog' y se basa en la observación de que los ortólogos de proteínas interactivas en un organismo a menudo también están interactuando en otro organismo.

FunRich

La herramienta FunRich se encarga de crear una red de interacción entre genes utilizando varias opciones de bases de datos como Uniprot, la propia de FunRich o incluso customizarlo. Actualmente tiene muchos tipos de diseño de redes como planetarias, circulares, empaquetadas y cuadradas.
Planetario: Por defecto, en función de la distribución de grados de los nodos, aquella cuya distribución de grados superior siempre está situada en el centro y el resto de los nodos de la periferia.
Circular: Por defecto, independientemente de la distribución de grados, todos los nodos y sus socios interactivos son circularizados aleatoriamente.
Empaquetado: Por defecto, los nodos con mayor distribución de grados siempre se representan con mayor radio y siempre se enfatizan de forma circular, así como se colocan cerca de los otros grupos que tienen una distribución de grados similar.
Cuadrado: Este diseño es muy similar al diseño de red empaquetado. Sin embargo, los nodos con mayor grado de conectividad y los nodos con menor grado de conectividad contienen el mismo radio.

Se ha elegido la distribución empaquetada porque su diseño nos parece más visual. Se distribuyen los genes en función del número de interacciones en la red, los genes con mayor número de interacciones forman los nodos de la red.[32]

Esta herramienta también es capaz de hacer un análisis funcional encontrando los GO correspondientes de los genes.
Para la estadística usan test hipergromético, BH y Bonferroni.

PANTHER

Protein ANalysis THrough Evolutionary Relationships (PANTHER) es una herramienta que forma parte del Proyecto Gene Ontology Phylogenetic Annotation. La versión utilizada en la 13.1 que cuenta con 112 genomas, 15524 familias de proteínas y utiliza la versión 1.2 de Gene Ontology como fuente de datos.
La función de esta herramienta es hacer un análisis funcional de los genes y clasificarlos según su función molecular, biológica, el tipo de componente celular de la proteína codificante, tipo de proteína y la vía de actuación (pathway).
Las anotaciones funcionales son incluidas manualmente y también electrónicamente por algoritmos informáticos basados en la similitud de secuencias.
Los resultados que obtenemos son unos gráficos con los resultados y también la lista de funciones con el número de genes pertenecientes a ese grupo y el porcentaje de genes respecto al total que representa dicho grupo.[33]

3. Resultados y Discusión

3.1 Resultados

El Proyecto Genoma Humano estima que el ser humano tiene entre 20000 y 25000 genes. En este estudio se quiere identificar cuáles de ellos están mayormente mutados en los pacientes de cáncer de mama y estudiar su función y red génica de interacción en la especie humana para poder entender un poco más a fondo esta enfermedad.
Como se ha dicho ya anteriormente, se ha trabajado con datos mutagénicos de pacientes con cáncer invasivo de mama. Datos que hemos obtenido de TCGA y a partir de los cuales comienza el estudio.

Identificación de genes con papel driver

El primer paso en la investigación es la identificación de los putativos genes con papel driver en cáncer invasivo de mama. Para ello, hemos utilizado la herramienta DriverDBv2 que a partir de los datos mutagénicos de 990 muestras de pacientes con dicho tipo de cáncer y mediante 15 sub-herramientas distintas, es capaz de identificar 239 posibles genes con papel driver. Los genes identificados son resultados coincidentes en al menos tres de las sub-herramientas, hecho que aporta una mayor validez a los resultados del análisis bioinformático (Figura 8).
Si hubiéramos elevado el nivel de exigencia a resultados coincidentes en más de 4 herramientas obtendríamos 66 posibles genes drivers más fiables. Pero lo que nos interesa también, es hallar aquellos genes que no se hayan identificado anteriormente en otros estudios para poder abrir un nuevo campo de investigación y dianas terapéuticas. Por otro lado, también se podría haber disminuido la exigencia a dos sub-herramientas pero sería difícil trabajar con una cantidad de datos tan grande (953 genes).

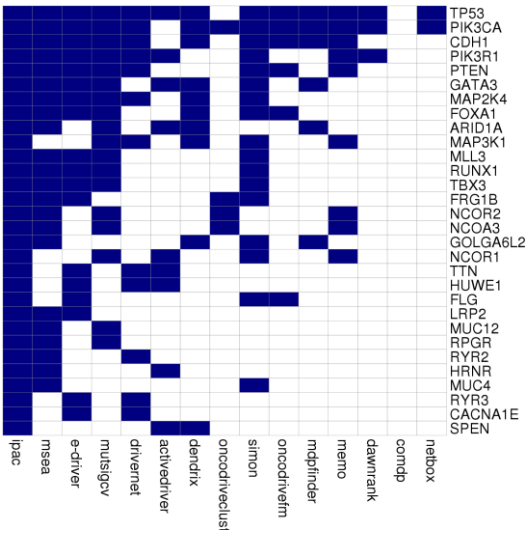


Figura 8: Esquema de 30 genes identificados como posibles drivers en cáncer de mama frente a las sub-herramientas encargadas de indentificarlos. Las casillas marcadas son las coincidentes en la identificación.

A continuación, en la Tabla 9 se muestra una lista de todos los genes identificados con la herramienta DriverDBv2 (coincidentes por 3 sub-herramientas). Observamos que algunos de los genes coinciden con los ya estudiados anteriormente, pero otros tantos no, se hablará de ello más adelante.

TP53	PIK3CA	CDH1	AKT1	PIK3R1	PTEN	GATA3	MAP2K4	CTCF
FOXA1	ERBB2	ITPR1	ARID1A	TPRX1	MAP3K1	OR5P2	NR1H2	SH3PXD2A
AOAH	USP36	MLL3	RUNX1	ZFP36L1	TBX3	EGFR	FRG1B	C9orf43
MEF2A	SF3B1	NCOR2	NCOA3	CDC27	ANK2	CRIPAK	GOLGA6L2	NCOR1
FRG1	VILL	ZFP36L2	POLE	TBL1XR1	SLC38A10	RBM5	RBMX	ZNF384
VEZF1	ASB10	BCL6B	GPRIN2	CBFB	FGFR2	CLTC	MAP3K4	LAMA1
HRAS	DSPP	HS6ST1	ASH1L	TTN	HUWE1	FLG	CASP8	BRCA1
RB1	HLA-DRB1	GPR32	ABCB10	ABR	ADAMTS7	ADAMTSL1	AXL	BAI1
BAZ2B	BZRAP1	CDC42BPG	CDH24	CHD3	CNGA3	COL1A2	CSPG4	CTU2
DSCAM	DYSF	EPPK1	FBXW7	GLYR1	GRB7	HECTD4	HLA-DRB5	KCNH1
LIFR	LRP2	LTBP2	MAP1B	MOCOS	NBPF12	NFE2L3	NOTCH2	NRCAM
OLFML2B	OR2T2	OR6C76	PABPC3	PARP4	PCDHB14	PGR	PTHLH	PXDNL
RP11-419C5.2	SCN9A	SLC4A5	TBX22	TFE3	TYRP1	VWF	AQP7	ART5
CCDC144NL	CD300LG	CELA1	CYP11B1	EPDR1	FAM104A	FAM98C	GIPC3	GPS2
HDAC2	KRT38	KRTAP4-5	KRTAP9-9	LGALS9B	MAPRE3	MST1	MUC12	NUPL2
PCBP2	PFKP	PHLDA1	PTH2	RALY	RBM15B	SAAL1	SLC25A5	TERF2IP
THEM5	TICAM1	U2AF2	WNT9A	RPGR	ACTL6B	FNDC4	BAX	SELPLG
ATN1	TAF1B	KRTAP5-1	RYR2	MTOR	CACNA1F	RYR1	MSH3	FRMD4A
BCL9L	HRNR	SMARCC2	AKAP11	USP8	HERC1	SRRM2	SUPT6H	LSR
RANBP2	IDH1	CASR	PHC3	CRMP1	KCNN2	RREB1	UBR5	ECM2
DENND4B	DCHS1	SHROOM4	APOBR	RP1L1	CDK12	GIGYF2	CCDC66	KCNN3
ASCC3	ARSB	HIST1H3B	MUC4	SOS2	RYR3	CACNA1A	ROCK2	CACNA1E
ARHGAP35	MYLK	BIRC6	CACNA1B	DDX11	SRCAP	SPTAN1	FLNA	SETD2
SDK2	PCDH11X	SVEP1	FBN1	GRIA3	PDE3A	ITGB4	BRCA2	ERBB3
PRKCB	AKT3	EP300	PLCG2	RAF1	SPEN	VPS13B	TP53BP1	SOS1
FLNB	RELN	MAST1	NRK	OTOF	ATM	SEPT10	OGFR	MGAM
BTNL8	PCDHGA4	ZNF302	KRAS	PIK3CB				

Tabla 9: Lista de genes identificados como drivers por la herramienta DriverDBv2.

Red génica de los genes con papel driver en cáncer de mama

Una vez identificados los genes el siguiente paso es relacionarlos entre sí, tal y como se plantea en los objetivos del proyecto. Es una gran cantidad de genes por lo que la red de interacción resultante es interesante para la mejor comprensión del cáncer de mama. Lo más lógico sería pensar que los genes con un número mayor de interacciones con otros genes drivers podrían tener mayor influencia en el desarrollo de la enfermedad.

Con la finalidad de crear una red génica de interacción entre los genes drivers identificados se ha utilizado la herramienta FunRich. Con la introducción de los IDs de los genes y mediante un análisis en la propia base de datos de FunRich, es capaz de crear un network en forma de imagen mostrando todas esas interacciones encontradas.
Al tener 239 genes identificados la cantidad de interacciones es inmensa, pero con la red creada los resultados son más visuales. Esta herramienta ha sido capaz de situar en la red 235 genes, es decir un 98.33% de los genes identificados. De los cuales, 97 genes no presentan ninguna interacción con otros genes identificados según la base de datos de FunRich.

Los genes se distribuyen de forma empaquetada, esto quiere decir que los nodos con mayor distribución de grados siempre se representan con mayor radio y siempre se enfatizan de forma circular, así como se colocan cerca de los otros grupos que tienen una distribución de grados similar. Por tanto, los nodos con un mayor número de interacciones tendrán un radio también mayor.

De esta forma podemos ver en la Figura 10 que los genes con mayor número de interacciones son los que se representan en los nodos: EGFR, TP53, BRCA1, FLNA, EP300, ERBB3, AKT1, ERBB2 y PIK3R1.
Y por otro lado los genes drivers con menor número de interacciones se encuentran en los extremos. Se representan con líneas naranjas las interacciones entre los genes.

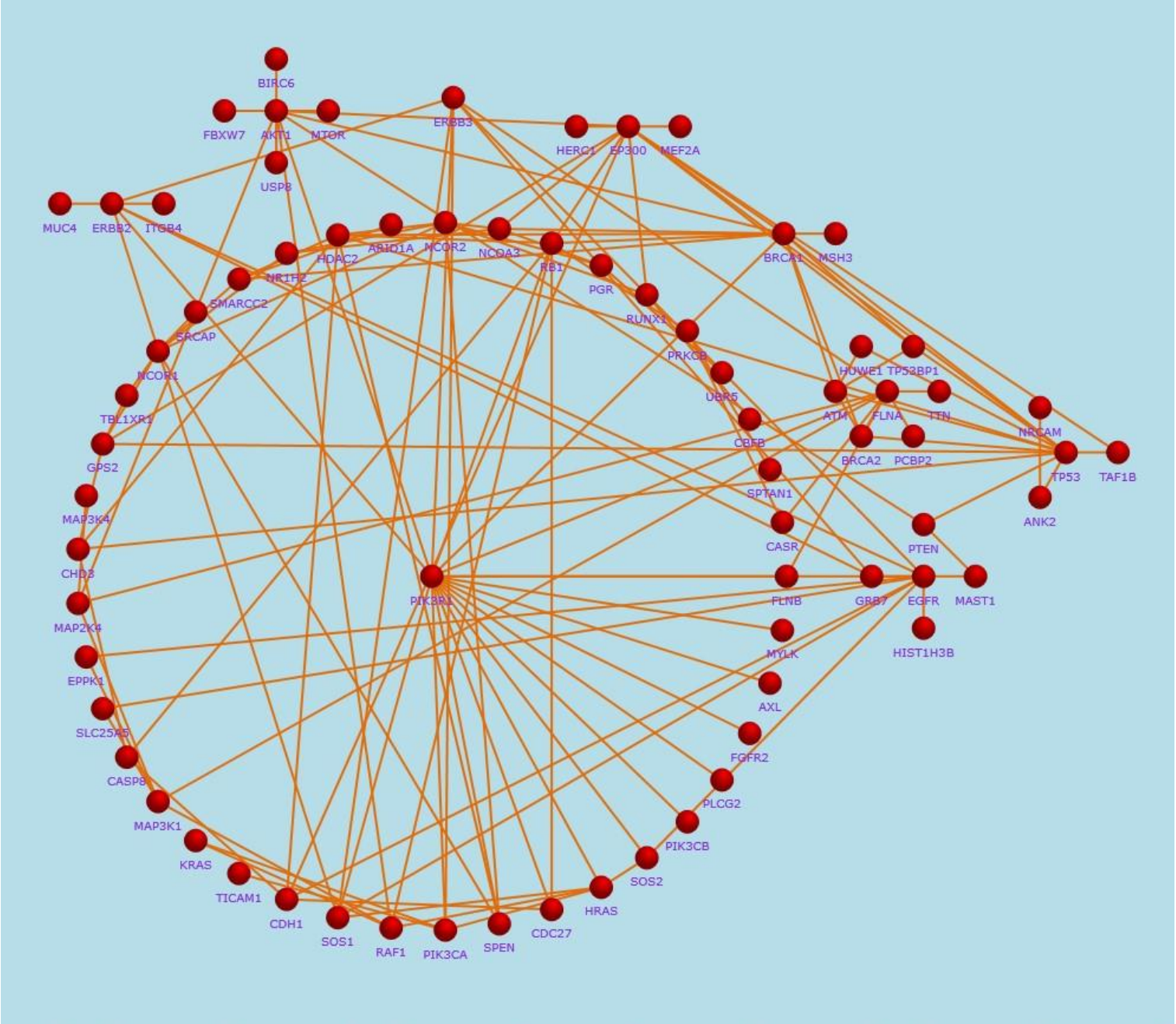


Figura 10: Red de interacción de posibles genes drivers en cáncer de mama creada con FunRich.

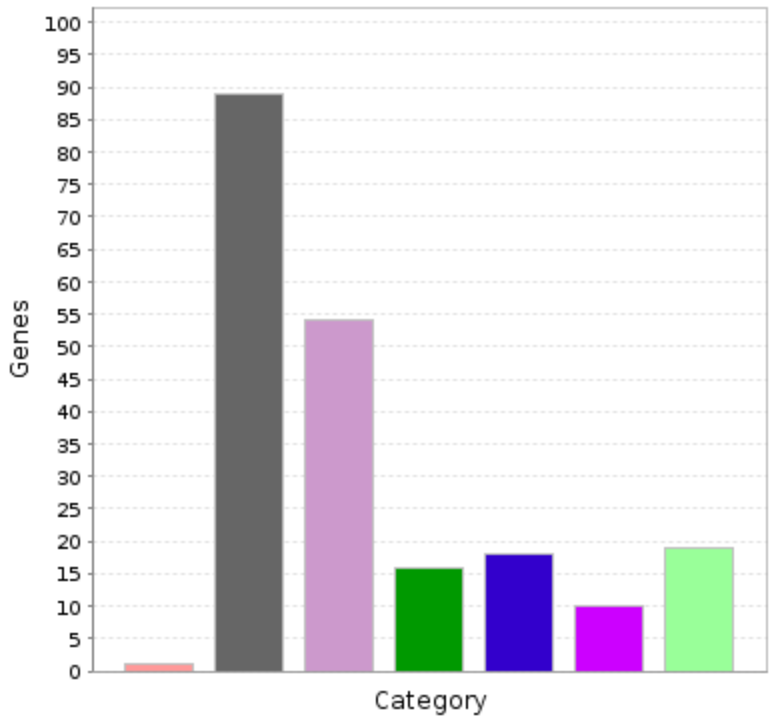
Análisis funcional

Una vez que se tiene los genes drivers identificados y también una red génica establecida pasamos al análisis funcional de los genes. Se han estudiado cinco características de los genes: función molecular, función biológica, vía de actuación, tipo de proteína sintetizada y tipo de componente celular. A continuación, se habla de cada una de ellas en particular, dando mayor importancia a la función molecular, biológica y a la vía de actuación.

El objetivo de esta parte de la investigación es intentar comprender lo que pueden llegar a implicar in vivo las mutaciones en estos genes. A qué vías o estructuras afectan y si podrían desembocar, por si solas o en conjunto, en proliferación celular incontrolada.

Función molecular

Utilizamos la herramienta PANTHER para el análisis funcional que como se ha comentado en el apartado de material y métodos en un software asociado a Gene Ontology. Introducimos los IDs de los genes identificados como drivers, seleccionamos el organismo *Homo sapiens* y la opción de la clasificación funcional mostrada en un gráfico de barras.
Como todas las herramientas bioinformáticas tienen un margen de error la clasificación no es perfecta pero el error es bajo. Si comparásemos los resultados de más herramientas funcionales este error se podría ir reduciendo.



Función	Nº genes	%
antioxidant activity (GO:0016209)	1	0.4%
binding (GO:0005488)	89	37.2%
catalytic activity (GO:0003824)	54	22.6%
receptor activity (GO:0004872)	16	6.7%
signal transducer activity (GO:0004871)	18	7.5%
structural molecule activity (GO:0005198)	10	4.2%
transporter activity (GO:0005215)	19	7.9%

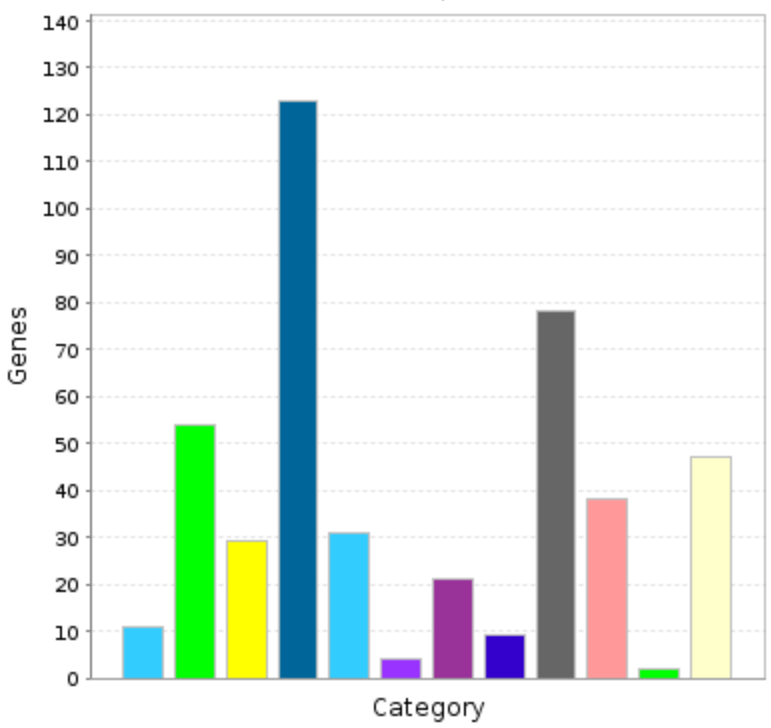
Figura 11: Gráfico de barras de la distribución de genes drivers según su función molecular y tabla leyenda.

Como podemos observar en los resultados de la Figura 11, la función molecular que caracteriza a un mayor porcentaje de genes (37.2%) es la de **unión**. En la base de datos de Gene Ontology define GO:0005488 como “la interacción selectiva, no covalente, a menudo estequiométrica, de una molécula con uno o más sitios específicos en otra molécula”[34]. En segundo lugar, con un 22.6% de los genes estaría la **actividad catalítica**, que se refiere básicamente a la reacción bioquímica mediada por enzimas. Después con un 7.9, un 6.7 y un 7.5% estaría la función **transportadora, receptora, y transducción de señales** respectivamente. Por último, función **estructural y antioxidante** (4.2 y 0.4%).

Función biológica

En cuanto a la función biológica lo genes drivers identificados se dividen en 12 grupos. Un 51.5% del total identificado, casi la mitad de ellos, cumple con un papel en el **proceso celular** (no tiene por qué implicar una sola célula). El segundo gran grupo es el de los genes con función en el **proceso metabólico**, un 32.6% del total. Como podemos observar estas funciones biológicas son de las más importantes tanto a nivel celular como a nivel del organismo entero. Las mutaciones causadas en estos genes pueden acarrear graves cambios en vías muy importantes y variar su funcionamiento.
Después, con un menor número de genes tenemos la función de **regulación biológica** (22.6%) y la **respuesta a estímulos** (19.7%).

En la Figura 12 podemos ver representados estos genes por su función biológica y una tabla con el porcentaje del número de genes en cada grupo.



Función	Nº genes	%
biological adhesion (GO:0022610)	11	4.6%
biological regulation (GO:0065007)	54	22.6%
cellular component organization or biogenesis (GO:0071840)	29	12.1%
cellular process (GO:0009987)	123	51.5%
developmental process (GO:0032502)	31	13.0%
immune system process (GO:0002376)	4	1.7%
localization (GO:0051179)	21	8.8%
locomotion (GO:0040011)	9	3.8%
metabolic process (GO:0008152)	78	32.6%
multicellular organismal process (GO:0032501)	38	15.9%
reproduction (GO:0000003)	2	0.8%
response to stimulus (GO:0050896)	47	19.7%

Figura 12: Gráfico de barras de la distribución de genes drivers según su función biológica y tabla leyenda.

Tipo de proteína y tipo de componente celular

Un 31% de los genes identificados producen proteínas constituyentes de la célula (como la membrana plasmática o citoplasma), un 19.2% orgánulos y en menos medida complejos macromoleculares y membrana (15.9 y 13.4% respectivamente).

El tipo de proteína sintetizada más representada por estos genes es la de unión al ácido nucléico con un 12.6 %, a continuación, se encuentran los factores transcripcionales (8.4%), modulador enzimático (7.5%) y la hidrolasa (7.1%). También se destacan las proteínas de tipo transferasa (6.7%), transportador (5%), receptor (3.8%) y proteínas del citoesqueleto (3.8%).

Pathway

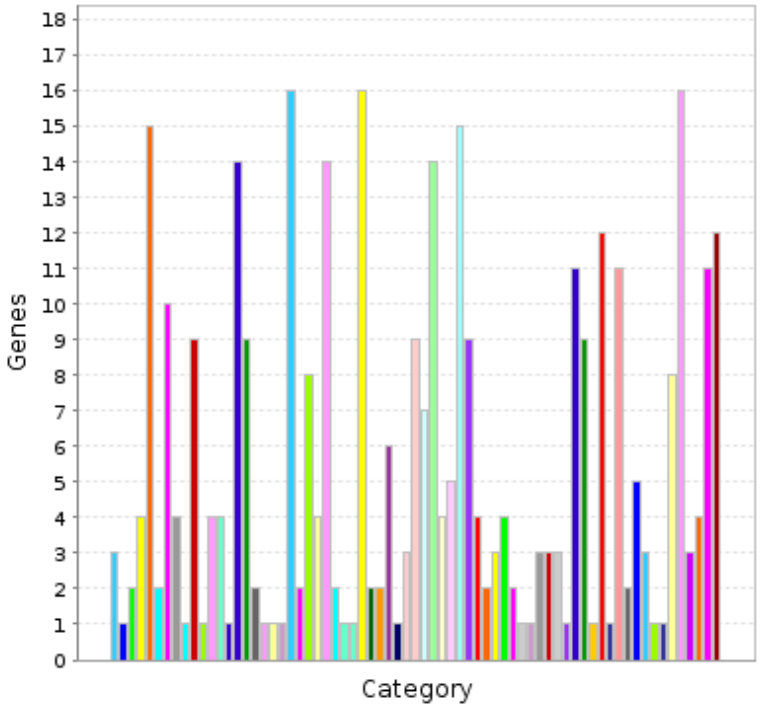
En la Figura 13 se pueden distinguir tres grupos muy marcados, son los de la vía de señalización **Wnt** (6.7%), vía de señalización de **receptor EGF** (6.7%) y la vía del receptor hormonal relacionado con la **gonadotropina** (6.7%). A continuación, vienen los genes relacionados con la **angiogénesis** (6.3%), la vía de señalización de la **integrina** (6.3%), vía de señalización de la **inflamación** mediada por citocina y quimiocina (5.9%), vía de señalización **FGF** (5.9%) y mapa de señalización **CCKR** (5.9%).

Las vías de señalización Wnt agrupa todas aquellas vías formadas por proteínas que transfieren las señales del exterior al interior de la célula a través de la superficie receptora de señales.

La angiogénesis tumoral es el crecimiento de vasos sanguíneos nuevos que los tumores necesitan para crecer. Por tanto, es lógico pensar que la mutación de algún gen de esta vía pueda beneficiar el desarrollo del cáncer.

El nivel de hormona gonadotropina coriónica humana se ha utilizado como marcador tumoral en diversos trastornos neoplásicos.

La integrina participa en la unión célula-célula o célula matriz extracelular y además tiene un papel de señalización. Las señales que recibe la célula a través de las integrinas pueden dar lugar a diversos procesos celulares como crecimiento, división, supervivencia, diferenciación celular y apoptosis.



Vía	Nº genes	%
DPP signaling pathway (P06213)	1	0.4%
DPP-SCW signaling pathway (P06212)	1	0.4%
BMP/activin signaling pathway-drosophila (P06211)	1	0.4%
Axon guidance mediated by netrin (P00009)	4	1.7%
Metabotropic glutamate receptor group III pathway (P00039)	4	1.7%
Beta2 adrenergic receptor signaling pathway (P04378)	4	1.7%
Axon guidance mediated by semaphorins (P00007)	1	0.4%
Beta1 adrenergic receptor signaling pathway (P04377)	4	1.7%
Apoptosis signaling pathway (P00006)	10	4.2%
Ionotropic glutamate receptor pathway (P00037)	4	1.7%
Angiogenesis (P00005)	15	6.3%
Interleukin signaling pathway (P00036)	9	3.8%
Alzheimer disease-presenilin pathway (P00004)	4	1.7%
5HT2 type receptor mediated signaling pathway (P04374)	3	1.3%
Alzheimer disease-amyloid secretase pathway (P00003)	2	0.8%
Integrin signalling pathway (P00034)	15	6.3%
Alpha adrenergic receptor signaling pathway (P00002)	1	0.4%
Insulin/IGF pathway-protein kinase B signaling cascade (P00033)	5	2.1%
Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade (P00032)	4	1.7%
Inflammation mediated by chemokine and cytokine signaling pathway (P00031)	14	5.9%
Hypoxia response via HIF activation (P00030)	7	2.9%
Ubiquitin proteasome pathway (P00060)	1	0.4%
Nicotine pharmacodynamics pathway (P06587)	1	0.4%
GABA-B receptor II signaling (P05731)	2	0.8%
Huntington disease (P00029)	9	3.8%
Endogenous cannabinoid signaling (P05730)	2	0.8%
Heterotrimeric G-protein signaling pathway-rod outer segment phototransduction (P00028)	1	0.4%
p53 pathway (P00059)	12	5.0%
p53 pathway feedback loops 2 (P04398)	11	4.6%
Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway (P00027)	6	2.5%
p53 pathway by glucose deprivation (P04397)	4	1.7%
Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway (P00026)	2	0.8%
Wnt signaling pathway (P00057)	16	6.7%
Hedgehog signaling pathway (P00025)	2	0.8%
VEGF signaling pathway (P00056)	8	3.3%
Transcription regulation by bZIP transcription factor (P00055)	1	0.4%
Thyrotropin-releasing hormone receptor signaling pathway (P04394)	5	2.1%
Toll receptor signaling pathway (P00054)	3	1.3%
Ras Pathway (P04393)	12	5.0%
General transcription by RNA polymerase I (P00022)	1	0.4%
T cell activation (P00053)	11	4.6%
P53 pathway feedback loops 1 (P04392)	1	0.4%
FGF signaling pathway (P00021)	14	5.9%
TGF-beta signaling pathway (P00052)	2	0.8%
Oxytocin receptor mediated signaling pathway (P04391)	3	1.3%
FAS signaling pathway (P00020)	4	1.7%
Endothelin signaling pathway (P00019)	8	3.3%
EGF receptor signaling pathway (P00018)	16	6.7%
p38 MAPK pathway (P05918)	3	1.3%
Parkinson disease (P00049)	1	0.4%
PI3 kinase pathway (P00048)	9	3.8%
Cytoskeletal regulation by Rho GTPase (P00016)	2	0.8%
PDGF signaling pathway (P00047)	11	4.6%
Oxidative stress response (P00046)	3	1.3%
Histamine H1 receptor mediated signaling pathway (P04385)	3	1.3%
Notch signaling pathway (P00045)	3	1.3%
Nicotinic acetylcholine receptor signaling pathway (P00044)	1	0.4%
Cadherin signaling pathway (P00012)	9	3.8%
Blood coagulation (P00011)	1	0.4%
Dopamine receptor mediated signaling pathway (P05912)	1	0.4%
Muscarinic acetylcholine receptor 1 and 3 signaling pathway (P00042)	2	0.8%
B cell activation (P00010)	9	3.8%
Angiotensin II-stimulated signaling through G proteins and beta-arrestin (P05911)	2	0.8%
Metabotropic glutamate receptor group I pathway (P00041)	2	0.8%
Metabotropic glutamate receptor group II pathway (P00040)	3	1.3%
CCKR signaling map (P06959)	14	5.9%
Gonadotropin-releasing hormone receptor pathway (P06664)	16	6.7%
SCW signaling pathway (P06216)	1	0.4%
GBB signaling pathway (P06214)	1	0.4%

Figura 13: Gráfico de barras de la distribución de genes drivers según su vía de actuación y tabla leyenda.

Dentro de las vías descritas con un mayor porcentaje de genes driver identificados, están los genes asociados a la cadherina (5 genes drivers), con Phosphatidylinositol 3-kinase (PI3K, 3 genes), conexión de la vía no fermentativa de la sacarosa (3 genes), receptor factor de crecimiento epidermal (3 genes) y receptor rianodina RYR 1/2/3 (3 genes).
En los resultados se ha visto sobre todo que la vía PI3K es muy representativa dentro de las categorías marcadas con anterioridad, es decir, que interacciona con vías importantes.

3.2 Discusión

El Cáncer de Mama está siendo actualmente muy estudiado ya que afecta a gran parte de la población, sobre todo mujeres. Los cánceres diagnosticados en una etapa temprana pueden ser tratados para su eliminación con diferentes métodos como fármacos o cirugía. El estudio génico del cáncer de mama abre nuevas puertas al desarrollo de nuevos fármacos y medicina personalizada. Gracias a la investigación de posibles dianas terapéuticas, se pretende mejorar el tratamiento de esta enfermedad y con ello reducir el número de casos de mujeres y hombres afectados.

En anteriores estudios se ha visto una asociación entre las mutaciones de genes concretos a los que se llama genes drivers con el cáncer de mama. Estos genes aparecen recurrentemente mutados en una gran proporción de datos mutagénicos de pacientes con dicha enfermedad. Por ello es importante centrarse en la investigación de estos genes mutados e interpretar su funcionamiento e interacción con otros genes.

En el presente trabajo, los datos mutagénicos de 990 muestras de pacientes con cáncer de mama invasivo fueron analizados para la identificación de posibles genes con papel driver. Además, los genes resultantes fueron expuestos a un análisis funcional y de interacciones.
239 posibles genes drivers fueron identificados por al menos tres sub-herramientas distintas del software DriverDBv2 proporcionando así una alta fiabilidad de los resultados. Dichos genes se sometieron a un análisis en busca de redes de interacción, dando como resultado una red génica de 235 posibles genes drivers que interaccionan entre sí en menor o mayor medida. Concretamente, 9 genes fueron clasificados como nodos de la red de interacción: EGFR, TP53, BRCA1, FLNA, EP300, ERBB3, AKT1, ERBB3 y PIK3R1. Además, también los genes identificados como posibles drivers fueron clasificados según su función biológica, molecular y su vía de actuación. Los resultados de este análisis muestran una distribución molecular variada de los genes con un 37.2% de ellos con función de unión, un 22.6% de actividad catalítica y

en menor medida función transportadora (7.9%), receptora (6.7%) y transductora de señales (7.5%). En cuanto a la función biológica, se distribuyen el 51.5% de los genes con proceso celular, el 32.6% en proceso metabólico, el 22.6% en regulación biológica y el 19.7% respuesta a estímulos. Por otro lado, donde más variabilidad se encuentra es en las vías de actuación ya que un solo gen puede actuar en varias vías a la vez. Las vías más destacadas son: señalización Wnt (6.7%), señalización de receptor EGF (6.7%), la vía del receptor hormonal relacionado con la gonadotropina (6.7%), genes relacionados con la angiogénesis (6.3%), señalización de la integrina (6.3%), señalización de la inflamación mediada por citocina y quimiocina (5.9%), señalización FGF (5.9%) y mapa de señalización CCKR (5.9%).

Si nos centramos en los 9 genes clasificados como nodos en la red de interacción se puede observar que muchos de ellos ya han sido identificados anteriormente como posibles genes drivers del cáncer de mama (TP53, BRCA1, ERBB2, AKT1 y PIK3R1). Pero otros como EP300, FLNA, ERBB3 o EGFR no son tan conocidos.

EP300 produce la proteína p300 encargada de la regulación del crecimiento, división y diferenciación celular. Se suele denominar como coactivador transcripcional.[35]

FLNA ayuda a producir la proteína filamina A, que se encarga de la formación citoesqueleto y la organización de la matriz extracelular. De esta forma se puede unir a las integrinas y permiten intercambiar señales entre la célula y la matriz extracelular.[36] [37]

ERBB3 codifica para un receptor del factor de crecimiento epidérmico, la tirosina quinasa. Por sí sola no puede transmitir la señal y por ello forma dímeros con otros receptores EGF. Estos dímeros conducen a la activación de vías de proliferación o diferenciación celular. La amplificación de este gen y/o la sobreexpresión de su proteína se han reportado en numerosos cánceres, incluyendo tumores de próstata, vejiga y mama.[38]

EGFR sintetiza la proteína llamada receptor del factor de crecimiento epidérmico, que se extiende a lo largo de la membrana celular. Como resultado de la unión del receptor a un ligando se desencadenan vías de señalización dentro de la célula que promueven el crecimiento y la proliferación y la supervivencia celular. [39]
Anteriormente se había hablado de la vía de señalización que implicaba los factores de crecimiento, pero concretamente, los resultados de nuestro estudio implican directamente al gen EGFR como posible gen driver del cáncer de mama.

Los resultados sugieren que el gen PIK3R1 es muy importante en el cáncer de mama ya que es el gen identificado con un mayor número de interacciones en la red génica que se ha creado y además, está relacionado con varias vías de acción diferentes (Figura 14).

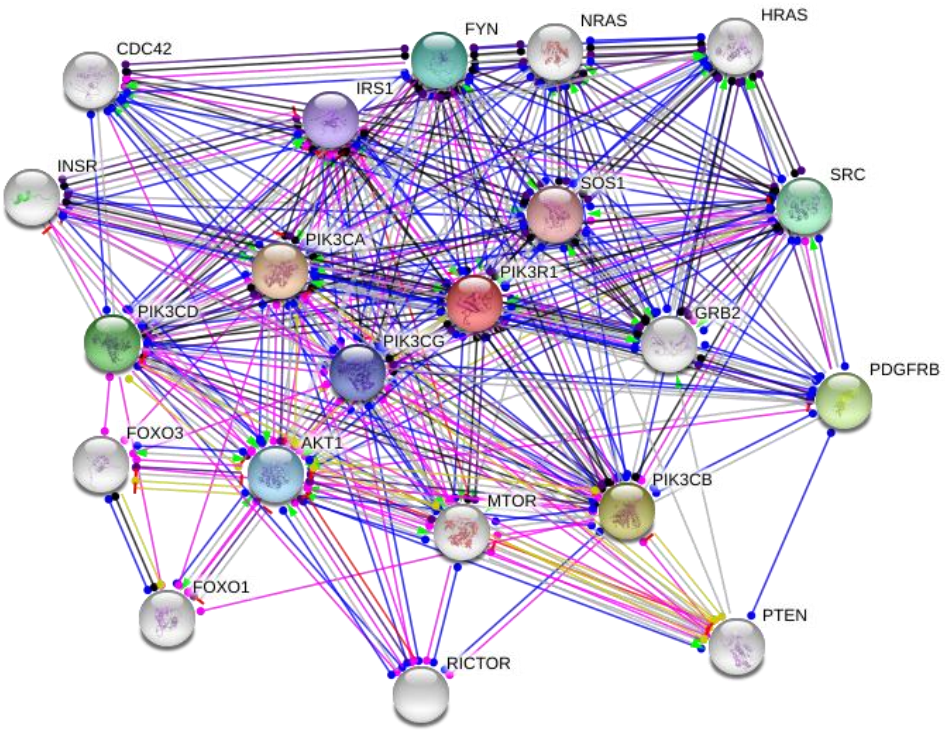


Figura 14: Network de PIK3R1 creada con el software String. Resultados con más de un 90% de confianza, muchos de ellos son posibles genes con papel driver.

En cuanto a las vías de actuación, la mayoría de genes identificados como posibles drivers sugieren tener relación con vías de señalización, receptores y unión a la membrana. Tienen que ver con los procesos de regulación del ciclo celular y proceso de carcinogénesis. Por tanto, es lógico pensar que las mutaciones que afectan dichos genes pueden provocar una proliferación incontrolada de células, daño en el ADN, o desregulación de las vías importantes para el buen funcionamiento de la célula como los factores de crecimiento.

La finalidad de este trabajo es intentar clarificar el tema de los genes con papel driver en el cáncer de mama. Los resultados en general de los estudios son muy dispersos por la falta de consenso, aun así, la identificación de genes drivers está bastante estudiada. La parte más novedosa y funcional del estudio es el análisis de la red génica de interacción entre los genes y el análisis funcional de los mismos. El objetivo principal es poder abrir puertas para la medicina personalizada y la creación de nuevos fármacos contra los genes drivers o las vías de acción implicadas en el cáncer de mama.

4. Conclusiones

Basándonos en los resultados presentados proponemos las siguientes conclusiones:

- Las muestras de cáncer de mama invasivo presentan mutaciones en diversos genes. Los genes que prevalecen con mutaciones en la mayoría de muestras de los pacientes son clasificados para su posible papel driver. Con herramientas bioinformáticas se han identificado 239 posibles genes drivers. Algunos de ellos en línea con otros estudios realizados y otros nuevos.
- Los posibles genes drivers identificados crean una red de interacción, donde podemos ver genes-nodo que participan en varias vías de acción e interaccionan con otros genes identificados.
- Los resultados del análisis funcional se muestran de acuerdo con otros estudios, proponiendo las posibles vías de acción más importantes en el cáncer de mama invasivo.
- Los datos obtenidos de este análisis pueden servir para continuar con la investigación de este tipo de cáncer que afecta a muchas personas en el mundo. En un futuro podría ser el punto de partida para la investigación de nuevas posibles dianas terapéuticas.

Con el presente trabajo, se ha investigado acerca del cáncer de mama. Partiendo de un punto básico hasta el estudio más profundo de los posibles genes drivers y sus funciones dentro de la célula humana.
Se han conseguido los objetivos propuestos para la investigación obteniendo resultados interesantes con la posibilidad de continuar en un futuro. Los métodos planteados al principio se han ido cambiando a lo largo del estudio según las necesidades que iban apareciendo y se han reflejado dichos cambios en los informes de seguimiento. Se ha utilizado, por ejemplo, algún software que al inicio no se planteó, pero nos ha sido útil en la investigación. Por otro lado, también hemos ido adecuando los datos y resultados a las necesidades y manejabilidad del estudio. Es decir, se han determinado ciertos límites o factores de cribado para que los resultados fueran más fiables (como en el caso de identificación de genes por al menos 3 herramientas distintas).
Una propuesta futura es poder mejorar las herramientas utilizadas para los análisis. Los resultados han sido buenos pero las clasificaciones, interacciones, o identificación no son 100% efectivas.

En cuanto a las líneas de trabajo futuro queda pendiente el análisis profundo de los genes identificados como posiblemente más influyentes, sobre todo su estructura y forma específica de interacción para poder diseñar fármacos nuevos que puedan ayudar a la cura de cáncer de mama. Y también es importante consensuar un listado de genes drivers para un posible diagnóstico prematuro de la enfermedad.

Aún hay muchas cosas por entender de este proceso tumoral cómo por qué se producen estas mutaciones espontáneas o cómo se traspasan de forma hereditaria.

5. Glosario

Gen driver: gen mutado que presenta variaciones de un solo nucleótido (SNPs), variaciones en el número de copias, etc y proporcionan una ventaja a las células para la supervivencia del tumor.

Pathway (vía de acción): es una serie de interacciones entre moléculas en la célula que llevan a un determinado producto o cambio en la célula.

Génico: perteneciente o relativo a los genes.

Mutación: es el cambio en la secuencia de un nucleótido o en la organización del ADN de un ser vivo, que produce una variación en las características de este y que no necesariamente se transmite a la descendencia. Se presenta de manera espontánea o por la acción de mutágenos. [40]

SNV: es una variación o mutación en la secuencia de ADN que afecta a una sola base (adenina (A), timina (T), citosina (C) o guanina (G) de una secuencia del genoma.[41]

TCGA: The Cancer Genome Atlas, cuya finalidad es catalogar cambios moleculares de importancia biológica responsables de la aparición de cáncer haciendo uso de la secuenciación genómica y la bioinformática. [2]

Citoesqueleto: se compone principalmente de proteínas y da soporte interno a las células, organiza las estructuras internas e interviene en los fenómenos de transporte, tráfico y división celular.

Dímero: complejo formado por dos proteínas, generalmente la unión se produce con un enlace no covalente.

6. Bibliografía

1. DeVita, V. T., Lawrence, T. S. & Rosenberg, S. A. *Cancer : principles & practice of oncology. Volume 2 annual advances in oncology*.
2. Home - The Cancer Genome Atlas - Cancer Genome - TCGA. Available at: https://cancergenome.nih.gov/. (Accessed: 4th June 2018)
3. Breast Cancer - Harvard Health. Available at: https://www.health.harvard.edu/womens-health/breast-cancer. (Accessed: 4th June 2018)
4. CDI: carcinoma ductal invasivo. Available at: http://www.breastcancer.org/es/sintomas/tipos/cdi. (Accessed: 4th June 2018)
5. CLI: carcinoma lobular invasivo. Available at: http://www.breastcancer.org/es/sintomas/tipos/cli. (Accessed: 4th June 2018)
6. Ashworth, A., Lord, C. J. & Reis-Filho, J. S. Genetic Interactions in Cancer Progression and Treatment. *Cell* **145,** 30–38 (2011).
7. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
8. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486,** 400–404 (2012).
9. Kumar Rajendran, B. & Deng, C.-X. Characterization of potential driver mutations involved in human breast cancer by computational approaches. *Oncotarget* **8,** 50252–50272 (2017).
10. Chalhoub, N. & Baker, S. J. PTEN and the PI3-kinase pathway in cancer. *Annu. Rev. Pathol.* **4,** 127–50 (2009).
11. Dienstmann, R., Rodon, J., Serra, V. & Tabernero, J. Picking the point of inhibition: a comparative review of PI3K/AKT/mTOR pathway inhibitors. *Mol. Cancer Ther.* **13,** 1021–31 (2014).
12. Zhou, Z. *et al.* Regulation of estrogen receptor signaling in breast carcinogenesis and breast cancer therapy. *Cell. Mol. Life Sci.* **71,** 1549–1549 (2014).
13. RTK-growth-factor-signaling - My Cancer Genome. Available at: https://www.mycancergenome.org/content/molecular-medicine/pathways/RTK-growth-factor-signaling. (Accessed: 4th June 2018)
14. cell-cycle-control-DNA-damage - My Cancer Genome. Available at: https://www.mycancergenome.org/content/molecular-medicine/pathways/cell-cycle-control-DNA-damage. (Accessed: 4th June 2018)
15. Wu, D., Rice, C. M. & Wang, X. Cancer bioinformatics: A new approach to systems clinical medicine. *BMC Bioinformatics* **13,** 71 (2012).
16. Chung, I.-F. *et al.* DriverDBv2: a database for human cancer driver gene research. *Nucleic Acids Res.* **44,** D975–D979 (2016).
17. Reimand, J. & Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **9,** 637–637 (2014).
18. Vandin, F., Upfal, E. & Raphael, B. J. De novo discovery of mutated driver

pathways in cancer. *Genome Res.* **22,** 375–385 (2012).
19. Zhao, J., Zhang, S., Wu, L.-Y. & Zhang, X.-S. Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* **28,** 2940–2947 (2012).
20. Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27,** 175–181 (2011).
21. Cerami, E., Demir, E., Schultz, N., Taylor, B. S. & Sander, C. Automated Network Analysis Identifies Core Pathways in Glioblastoma. *PLoS One* **5,** e8918 (2010).
22. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40,** e169–e169 (2012).
23. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
24. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22,** 398–406 (2012).
25. Zhang, J., Wu, L.-Y., Zhang, X.-S. & Zhang, S. Discovery of co-occurring driver pathways in cancer. *BMC Bioinformatics* **15,** 271 (2014).
26. Hou, J. P. & Ma, J. DawnRank: discovering personalized driver genes in cancer. *Genome Med.* **6,** 56 (2014).
27. Bashashati, A. *et al.* DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.* **13,** R124 (2012).
28. Porta-Pardo, E. & Godzik, A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics* **30,** 3109–14 (2014).
29. Xia, J. & Wishart, D. S. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* **38,** W71-7 (2010).
30. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29,** 2238–2244 (2013).
31. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45,** D362–D368 (2017).
32. Keerthikumar, S., Pathan, M., Agbinya, J. & Mathivanan, S. FunRich Tool Documentation.
33. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41,** D377–D386 (2012).
34. AmiGO 2: Search. Available at: http://amigo.geneontology.org/amigo/search/ontology. (Accessed: 31st May 2018)
35. EP300 gene - Genetics Home Reference. Available at: https://ghr.nlm.nih.gov/gene/EP300. (Accessed: 5th June 2018)
36. Jenkins, Z. A. *et al.* Differential regulation of two *FLNA* transcripts explains some of the phenotypic heterogeneity in the loss-of-function filaminopathies. *Hum. Mutat.* **39,** 103–113 (2018).
37. Duval, D. *et al.* Valvular dystrophy associated filamin A mutations reveal a new role of its first repeats in small-GTPase regulation. *Biochim. Biophys. Acta - Mol. Cell Res.* **1843,** 234–244 (2014).

38. ERBB3 gene - Genetics Home Reference. Available at: https://ghr.nlm.nih.gov/gene/ERBB3. (Accessed: 5th June 2018)
39. Lemmon, M. A., Schlessinger, J. & Ferguson, K. M. The EGFR Family: Not So Prototypical Receptor Tyrosine Kinases. *Cold Spring Harb. Perspect. Biol.* **6,** a020768–a020768 (2014).
40. Jorde, L. B., Carey, J. C. & Bamshad, M. J. *Genética médica*.
41. Boyle, J. A primer of genome science (2nd ed.): Gibson, G., and Muse, S. *Biochem. Mol. Biol. Educ.* **33,** 313–313 (2005).