Trabajo Fin de Grado

BIOTECNOLOGÍA PARA LA MEDICINA DE PRECISIÓN: IDENTIFICACIÓN DE VARIACIONES GENÓMICAS PARA EL DIAGNÓSTICO DE CÁNCER DE MAMA

Universitat Politècnica de València

Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural

Grado en Biotecnología – Curso 2017/2018

Presentado por: **Alba Escalera Balsera**

Tutor: **José Javier Forment Millet** | Cotutor: **Dr. Óscar Pastor López**
Directora experimental: **Ana León Palacio**

Valencia, Julio de 2018

TÍTULO

Biotecnología para la Medicina de Precisión: Identificación de Variaciones Genómicas para el Diagnóstico de Cáncer de Mama.

Biotechnology for Precision Medicine: Identification of Genomic Variations for the Diagnosis of Breast Cancer.

Biotecnologia per a la Medicina de Precisió: Identificació de Variacions Genòmiques per al Diagnòstic de Càncer de Mama.

RESUMEN

El manejo eficiente de datos es un problema esencial para garantizar un diagnóstico correcto en el ámbito de una Medicina de Precisión (MP) efectiva y adecuada. Ciencia de Datos y Biotecnología deben unir sus fuerzas para que ese objetivo sea alcanzable, obteniendo cada vez más datos y convirtiéndolos en información valiosa. En esa dirección, el objetivo de este trabajo es identificar las variaciones genómicas para el diagnóstico de cáncer de mama mediante la metodología SILE: *Search-Identification-Load-Exploitation* (Búsqueda-Identificación-Carga-Explotación). Basándose en dicha metodología, el primer paso es seleccionar un conjunto adecuado de fuentes de datos genómicas. Se iniciará el trabajo explorando cuatro bases de datos (ClinVar, Ensembl, GWAS catalog y SNPedia) y se buscarán en ellas todas las variaciones genómicas relacionadas con el cáncer de mama. Una vez obtenidas las variaciones candidatas, se filtrarán según un conjunto de criterios de calidad que guíen la selección definitiva: entre ellos, que al menos haya un artículo en PubMed (base de datos bibliográfica de referencia en la investigación bioinformática y biogenómica) que sustente la relación entre la variación y el fenotipo de estudio, que esta tenga un determinado significado clínico y que los resultados del estudio cumplan ciertos criterios estadísticos mínimos (lo cual se hace mediante una revisión bibliográfica manual). El objetivo final es identificar aquellas variaciones que tienen mayor riesgo de causar el fenotipo analizado, permitiendo por lo tanto que un diagnóstico fiable sea posible. El proceso de identificación debe de asegurar que las variaciones seleccionadas cumplen los criterios requeridos, y garantizar que los datos proporcionados por las bases de datos son correctos y se corresponden con lo que se busca en este estudio. Finalmente, las variaciones serán cargadas en una plataforma software para el diagnóstico genómico como evidencia experimental de que la MP de la que hablamos es efectivamente viable y puede ser explotada por el personal sanitario en el ámbito del diagnóstico clínico.

Palabras clave: Cáncer de mama, variaciones genómicas, Medicina de Precisión, metodología SILE.

ABSTRACT

The efficient management of data is an essential problem to ensure a correct diagnosis in the field of an effective and appropriate Precision Medicine (PM). Data Science and Biotechnology must join forces to make this goal reachable, obtaining more and more data and converting them into valuable information. In that direction, the objective of this work is identify genomic variations for the diagnosis of breast cancer using the SILE methodology: Search-Identification-Load-Exploitation. Based on this methodology, the first step is to select an appropriate set of genomic data sources. The work will begin by exploring four databases (ClinVar, Ensembl, GWAS catalog and SNPedia) and all the genomic variations related to breast cancer will be explored. Once the candidate variations have been obtained, they will be filtered according to a set of quality criteria that guide the final selection. Such criteria will include that at least one of the articles in which the variations (variation) is studied appears in PubMed (reference bibliographic database in bioinformatics and biogenomics research); that its mutation has a certain clinical significance; and that the article fulfils certain statistic data (which is checked using a literature review by hand). The final objective is to identify those variations that have a higher risk of causing the analysed phenotype, thus, allowing a reliable diagnosis to be possible. The identification process must ensure that the selected variations satisfy the required criteria, and that the data provided by the databases are correct and correspond to what is searched in this study. Finally, the variations will be loaded into a software platform for genomic diagnosis as experimental evidence that the PM we are talking about is indeed viable and can be exploited by health personnel to offer personalized genomic medicine to patients.

Key words: Breast cancer, genomic variations, Precision Medicine, SILE methodology.

Autora del TFG: **Dña. Alba Escalera Balsera.**

Valencia, Julio de 2018.

Tutor Académico: **Prof. D. José Javier Forment Millet.**

Cotutor: **Prof. Dr. Óscar Pastor López.**

Directora experimental: **Dña. Ana León Palacio.**

AGRADECIMIENTOS

Me gustaría agradecer a algunas de las muchas personas que han colaborado y me han ayudado a realizar este Trabajo de Fin de Grado.

Muchas gracias al equipo que forma el centro PROS por permitir que me uniese a ellos y pudiese hacer este trabajo en uno de sus proyectos. Especialmente a Ana, por su ayuda y dedicación en estos meses. También a Óscar por darme la oportunidad de hacer el Trabajo de Fin de Grado en el centro que él dirige.

A Javier Forment darle las gracias por tutorizarme.

Agradecer a mis padres por apoyarme en este y todos los momentos de mi vida, tanto personales como profesionales, ningún logro sería posible sin vosotros. Gracias, también, a mi familia y amigas por todo lo que me aportan día a día.

Finalmente, a los amigos que me han acompañado en estos cuatro años gracias por lo vivido fuera y dentro de las aulas.

ÍNDICES

1. INTRODUCCIÓN 1
1.1. Bases de datos genómicas 3
1.1.1. Tipos de bases de datos. 4
1.1.2. Dificultades que presenta la búsqueda en bases de datos genómicas. 5
1.2. Cáncer de mama 7
2. OBJETIVOS 11
3. METODOLOGÍA 13
3.1. Metodología SILE 13
3.1.1. Búsqueda 13
3.1.2. Identificación 13
3.1.3. Carga 13
3.1.4. Explotación 14
3.2. Modelo conceptual del genoma humano 14
3.3. Criterios de calidad 16
3.4. Base de datos utilizadas 18
4. RESULTADOS Y DISCUSIÓN 20
4.1. Variaciones obtenidas 20
4.2. Cromosoma 22
4.3. Gen 22
4.4. Significado clínico 24
4.5. Fenotipo 25
4.6. Población 26
5. DISCUSIÓN 28
6. CONCLUSIONES 31
7. BIBLIOGRAFÍA 32
8. ANEXOS

ÍNDICE FIGURAS

Figura 1. Esquema de una mama normal…….8
Figura 2. Imagen de la vista de variaciones……...15
Figura 3. Número de variaciones obtenidas de cada una de las bases de datos (o de más de una de ellas)……21
Figura 4. Número de variaciones que hay en cada uno de los cromosomas……..22
Figura 5. Distribución de las variaciones según su fenotipo……26

ÍNDICE TABLAS

Tabla 1. Número de variaciones que hay en cada uno de los genes……...23
Tabla 2. Distribución de las variaciones según su significado clínico…….24
Tabla 3. Distribución de las variaciones según la población sobre la que se han realizado los estudios……27

ABREVIATURAS

CMHG: Modelo Conceptual del Genoma Humano.

GeIS: Sistema de Información Genómica.

GWAS: Estudio de asociación del genoma completo (Genome-wide association study).

HGDB: Base de Datos del Genoma Humano (Human Genome Database).

IC: Intervalo de confianza.

MP: Medicina de Precisión.

NCBI: Centro Nacional de Información Biotecnológica (National Center for Biotechnology Information).

NGS: Tecnologías de secuenciación de nueva generación (Next-generation sequencing).

OR: Odds Ratio.

PMID: PubMed ID.

SILE: Búsqueda-Identificación-Carga-Explotación (Search-Identification-Load-Exploitation).

SNP: Polimorfismo de un solo nucleótido (Single nucleotide polimorfism).

1. INTRODUCCIÓN

En este trabajo se van a buscar variaciones genómicas relacionadas con el cáncer de mama en diferentes bases de datos con el objetivo de identificar cuáles de ellas son realmente relevantes para determinar el riesgo de que el individuo sufra la enfermedad. Esto se hace con la finalidad de poder proporcionar una Medicina de Precisión a los pacientes.

La Medicina de Precisión es un enfoque emergente para el tratamiento y la prevención de enfermedades que tiene en cuenta la variabilidad individual en los genes, en el medio ambiente y en el estilo de vida de cada persona (Shin, 2016). Mediante este enfoque se permite a médicos e investigadores predecir con mayor precisión cuáles son las estrategias de tratamiento y de prevención para una determinada enfermedad que funcionarán correctamente en un grupo característico de personas. Contrasta con el enfoque único para todos, seguido tradicionalmente, en el que las estrategias de tratamiento y de prevención se desarrollan para una persona promedio, con lo que tiene menos consideración por las diferencias entre los individuos ya que pueden ser exitosos para algunas personas pero no para otras (Fradkin, 2016).

El término “Medicina de Precisión” solapa con el término “Medicina Personalizada”, este último más antiguo. Sin embargo, según el Consejo Nacional de Investigación de Estados Unidos la palabra “Personalizada” podría malinterpretarse entendiéndose que los tratamientos y las prevenciones desarrolladas fueran de un modo único para cada paciente; mientras que en Medicina de Precisión el objetivo es identificar los enfoques efectivos para determinados pacientes. Por lo que, según dicho consejo es más correcto el término “Medicina de Precisión”, aunque algunas personas usen ambos indistintamente (National Research Council, 2011).

Uno de los criterios en los que se basa la Medicina de Precisión es la variabilidad del genoma de cada uno de los pacientes, por ello se estudian las variaciones genómicas. Para detectar dichas variaciones se han de secuenciar los genomas de numerosos individuos y compararlos. Se usa una población que presente la enfermedad y otra de control con la finalidad de observar cuáles son las variaciones comunes a todos ellos y cuáles aparecen significativamente en las personas enfermas. En la última década se han realizado numerosos estudios en esta dirección. El hecho de que puedan realizarse tantos es gracias al desarrollo de las tecnologías de secuenciación de nueva generación (conocidas por su acrónimo en inglés NGS*, Next Generation Sequencing*), desde que

comenzara su uso en 2005-2006. Mediante dicha tecnología se pueden secuenciar rápidamente y, cada vez a menor precio, largas secciones del genoma de una persona en particular. Actualmente se puede secuenciar un genoma completo de un individuo en un día por aproximadamente 1.000$ (Van Dijk, 2014), en comparación con los 10 años y los 100 millones de dólares requeridos para secuenciar el primer genoma humano en el denominado Proyecto Genoma Humano (Boeke *et al.,* 2016).

Las variaciones más comunes son los polimorfismos de un solo nucleótido (SNP), cambios en la secuencia del ADN que afectan a una única base nitrogenada. Estas variaciones pueden alterar la expresión o los productos de los genes, por lo que pueden influir en los fenotipos. Algunos de los cambios tienen consecuencias graves, debido a que modifican la estructura de las proteínas alterando o impidiendo su función, lo que conduce a formas de enfermedad mendeliana o compleja. Actualmente, para la búsqueda de estas variaciones se realiza un estudio de asociación del genoma completo (GWAS, *Genome-Wide Association Study*), el cual consiste en la asociación entre SNPs y una determinada enfermedad. Cuando se obtienen alelos más frecuentes en la población que padece la enfermedad, se dice que el SNP está asociado con la enfermedad.

La identificación de dichas variaciones es el enfoque principal del estudio de la genética humana. Estas variaciones genómicas generadas por los estudios de secuenciación actuales, usando -como se ha dicho- tecnología NGS, han de ser anotadas. El objetivo de su anotación es evaluar su impacto clínico para, así, caracterizar y priorizar cada una en función de la gravedad de las consecuencias esperadas (Butkiewicz y Bush, 2016).

En las enfermedades monogénicas, una mutación en un solo gen es responsable de la enfermedad. Las enfermedades de un solo gen pueden ser dominantes o recesivas y autosómicas o relacionadas con el sexo. Cuando las secuencias del genoma comenzaron a estar disponibles públicamente, los investigadores empezaron a cambiar su enfoque de enfermedades monogénicas a poligénicas y complejas, las cuales son más frecuentes en la población general y en las que están involucradas varios genes. La generación de marcadores generalizados de variación genética así como el desarrollo de nuevas tecnologías y bases de datos, permite a los investigadores asociar el fenotipo de la enfermedad a loci genéticos con mayor facilidad (Bianco *et al.*, 2013).

Las variaciones genómicas que han sido encontradas en los estudios de investigación y la información sobre dichos estudios ha de ser almacenada y ser accesible al público.

Con este objetivo se usan las bases de datos, en este caso las bases de datos genómicas. En las bases de datos se guardan todas las variaciones que según los estudios han sido encontradas y se relacionan con el fenotipo que está siendo investigado en el trabajo. Sin embargo, debido al gran número de variaciones que son capaces de encontrarse mediante las NGS, la complejidad de los procesos biológicos y las distintas técnicas de laboratorio utilizadas, no todas tienen un significado clínico verdaderamente fiable para el fenotipo estudiado.

A continuación, se va a explicar lo que son las bases de datos, centrándose en las bases de datos genómicas ya que son las que interesan en el presente trabajo. Posteriormente, se va a dar información sobre la enfermedad del cáncer de mama, ya que en este trabajo la búsqueda e identificación de variaciones se hace para este fenotipo.

1.1. BASES DE DATOS GENÓMICAS

Una base de datos es “una colección de información organizada de tal modo que sea fácilmente accesible, gestionada y actualizada” (Petersson y Breul, 2017).

La velocidad a la que avanza la generación de nueva información y el intercambio de datos es un factor estrechamente relacionado con el desarrollo humano. Y en ello, las bases de datos son esenciales para almacenar información mientras se actualizan y comparan constantemente los datos. La integración de nuevos descubrimientos en las bases de datos existentes las hace, no solo depositarias y guardianas de nuestra ciencia, sino también elementos clave para el progreso y la investigación científica.

Las bases de datos biológicas ofrecen acceso a una amplia variedad de datos biológicos y desempeñan un papel central en la bioinformática. Una base de datos biológicos es un conjunto organizado de información y de datos de estudios desarrollados en laboratorios de investigación (tanto *in vitro* como *in vivo*), a partir de análisis bioinformáticos (*in silico*) y publicaciones científicas (Bianco *et al.*, 2013).

Debido al gran tamaño del genoma humano, a su complejidad y a la variabilidad entre individuos hay una amplia gama de posibles características fenotípicas y de enfermedades, por lo que hay múltiples maneras de capturar, organizar y utilizar los

datos del genotipo y fenotipo humanos. A causa de ello hay muchos y diferentes tipos de bases de datos con información genómica (Brookes y Robinson, 2015). A continuación, se describen los más relevantes.

1.1.1. Tipos de bases de datos.

Las **bases de datos de primer nivel** almacenan nucleótidos (ADN y ARN) o secuencias de aminoácidos (proteínas) y contienen información útil para la identificación de las especies de las que se obtuvieron dichas secuencias, así como de su función.

Las **bases de datos especializadas** reúnen información más específica sobre la taxonomía, las funciones, las publicaciones científicas y las enfermedades relacionadas con las variaciones en las secuencias de nucleótidos.

El primer paso en un proyecto de investigación científica, particularmente si se trata de diseñar un experimento, es la investigación de la literatura, en el caso de experimentos en humanos literatura biomédica. El crecimiento exponencial en el número de artículos biomédicos publicados, junto con el mejor acceso a los recursos de Internet, ha hecho que las redes de datos sean obligatorias. Una red de datos es una infraestructura cuyo diseño posibilita la transmisión de información a través del intercambio de datos.

La **base de datos de artículos científicos** más antigua es PubMed, la cual fue inaugurada en 1997, es accesible en línea, de libre acceso a todos los usuarios y está desarrollada por NCBI (Centro Nacional de Información Biotecnológica). Esta incluye resúmenes médicos y biológicos, además de enlaces a los documentos. PubMed es parte del sistema de recuperación Entrez de NCBI, que proporciona acceso a 38 bases de datos y constituye un puente crucial entre los datos sobre biología molecular, genética y literatura científica. PubMed es el sitio más potente y actualizado para la investigación bibliográfica en biomedicina (Bianco *et al.*, 2013).

El término **“base de datos genotipo-fenotipo”** abarca una amplia gama de implementaciones de sistemas de información, con el objetivo de registrar y poner a disposición de investigadores u otros usuarios conjuntos de datos que incluyen datos genéticos (como las secuencias de ADN, las variaciones o los genotipos), datos fenotípicos (características observables de un individuo) y las correlaciones entre ambos. De este modo proporcionan acceso a datos y conocimientos suficientes para

permitir que la importancia funcional y patogénica de las variaciones genéticas se establezca y documente de manera fiable (Brookes y Robinson, 2015). Sin embargo, esta no es una tarea trivial. A continuación, se describen algunos de los problemas que surgen a la hora de utilizar las bases de datos genómicas.

1.1.2. Dificultades que presenta la búsqueda en bases de datos genómicas.

Es muy importante distinguir entre los alelos que causan enfermedades y las variaciones neutras que ocurren tanto en individuos normales como en afectados por la enfermedad. La asignación incorrecta de patogenicidad a variaciones puede llevar a diagnósticos genéticos equivocados o evaluaciones de riesgo de padecer una enfermedad a personas con estas variaciones. Por ello, lograr dicha distinción es un reto ya que pocos alelos tienen un efecto lo suficientemente grande para diferenciarse del resto, lo cual puede conducir a falsos positivos (Brookes y Robinson, 2015).

Como se ha dicho anteriormente, con las NGS se ha pasado del análisis monogenético (con Sanger) a los análisis mediante un panel de genes o la secuenciación del exoma completo (WES, W*hole-Exome Sequencing*). Sin embargo, la capacidad de estas tecnologías, además de encontrar nuevos genes causantes de enfermedades y variaciones patogénicas, también genera muchas asignaciones de patogenicidad o riesgo engañosas. De todas las variaciones en comparación con el genoma de referencia que se encuentran, no todas poseen consecuencias moleculares como inserciones o deleciones, cambios de aminoácidos que produzcan mutaciones sin sentido, proteínas aberrantes o cambios en el sitio de *splicing.* Por lo que se han de eliminar las variaciones que sean comunes en la población general sana, que se usa como control en los estudios. Es decir, la identificación de variaciones verdaderamente patogénicas o con riesgo de entre las muchas variaciones que se encuentran en un estudio es un gran desafío (Brookes y Robinson, 2015).

La existencia de un número elevado de bases de datos y la gran cantidad de información almacenada en ellas hace que sea difícil la búsqueda y comprensión de sus datos. Esto se debe a que cada una de las bases de datos tiene una estructura y un diseño distinto, con lo que el usuario tiene problemas para encontrar la información necesaria o utilizar las herramientas que la plataforma le ofrece. Además, dado que los grupos de investigación pueden subir, o no, la información resultante de sus estudios a los

diferentes repositorios hay datos que pueden estar repetidos en ellos y, por el contrario, hay otros que puede aparecer únicamente en uno o alguno. También, hay mucha información, en este caso variaciones genómicas, que al ser minuciosamente analizada, está presente en el estudio pero no relacionada con el fenotipo estudiado.

Teniendo en cuenta las dificultades mencionadas anteriormente, en este trabajo se pretende obtener las variaciones genómicas para el diagnóstico del cáncer de mama, lo cual será posteriormente cargado a una plataforma específicamente diseñada para su explotación en un entorno clínico. Para ello se seleccionará un conjunto de bases de datos biológicas, se analizará exhaustiva y sistemáticamente la información que contienen y se seleccionará aquella considerada como relevante según un conjunto de criterios de calidad previamente establecidos. De esta forma podemos asegurar que los resultados son los suficientemente fiables y consistente para proporcionar un diagnóstico genético.

1.2. CÁNCER DE MAMA

La búsqueda de variaciones genómicas se hace generalmente asociada a un fenotipo, en este caso al cáncer de mama.

La importancia de realizar estudios sobre el cáncer de mama se debe a que es la neoplasia maligna que afecta y mata a más mujeres en el mundo y, también, en España (GLOBAL CANCER OBSERVATORY: CANCER TODAY, 2012). Se estima que anualmente hay un millón de casos nuevos y más de 400.000 muertes en todo el mundo (Michailidou *et al.,* 2013). Según la Sociedad Española de Oncología Médica (2018), la incidencia (número de casos nuevos de una enfermedad en una población y en un período de tiempo determinados) estimada de cáncer de mama en España en el año 2017 era de 26.370; siendo el que más incidencia tiene en mujeres y el cuarto tanto en hombres como en mujeres. En el caso de mujeres, el segundo tumor con mayor incidencia es el colorrectal con 13.711 mujeres afectadas en 2017, lo que corresponde con aproximadamente la mitad de las afectadas por cáncer de mama. En el caso de la mortalidad, estimaron que debido al cáncer de mama murieron 6.385 mujeres en España en 2016, siendo también el tumor que mata a más mujeres y el cuarto en ambos sexos. La Asociación Española Contra el Cáncer (2018) dice que una de cada ocho mujeres tendrá cáncer de mama a lo largo de su vida.

El cáncer de mama es una enfermedad en la que se forman células malignas (cancerosas) en los tejidos del seno. La mama está compuesta de lóbulos y ductus (o conductos galactóforos). Cada seno tiene de 15 a 20 secciones llamadas lóbulos y cada lóbulo tiene numerosas secciones, más pequeñas, denominadas lobulillos. Los lobulillos terminan en una gran cantidad de pequeños bulbos que son los que pueden producir la leche. Los lóbulos, lobulillos y bulbos están unidos mediante fino tubos llamados ductus, estos conducen la leche hacia el pezón. Además, en los senos hay vasos sanguíneos y vasos linfáticos. Los primeros tienen como función proporcionar sangre a la glándula y los vasos linfáticos transportan la linfa entre los ganglios linfáticos, los cuales se encuentran en todo el cuerpo y su función es filtrar la linfa con el objetivo de combatir infecciones (NATIONAL CANCER INSTITUTE, 2018).

La mama humana es una glándula exocrina cutánea modificada. Está compuesta de piel, tejido subcutáneo, parénquima mamario (ductos y lóbulos) y estroma de soporte, especialmente el tejido adiposo que proporciona volumen y consistencia (Jesinger, 2014). En la *Figura 1* se observa una representación de la mama.

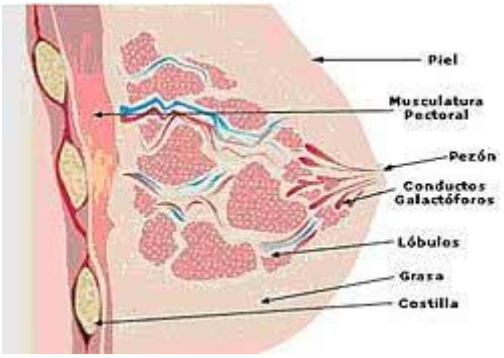

Figura 1. Esquema de una mama normal (ASOCIACIÓN ESPAÑOLA CONTRA EL CÁNCER, 2018).

El tipo más común de cáncer de mama es el carcinoma ductal, este comienza en las células de los conductos (NATIONAL CANCER INSTITUTE, 2018).

El carcinoma ductal *in situ* (DCIS) se define como una proliferación premaligna de células epiteliales neoplásicas contenidas dentro de la luz de los conductos mamarios. Los DCIS están revestidos por una capa de células mioepiteliales semicontinuas y rodeados por una membrana basal intacta. Durante varias décadas se ha aceptado que el DCIS constituye un precursor no obligatorio del carcinoma ductal invasivo; sin embargo, los estudios de observación clínica más recientes han corroborado la hipótesis de que DCIS es un precursor del cáncer de mama invasivo. DCIS se ha convertido en un gran desafío clínico debido a su incidencia creciente, este crecimiento está relacionado con la introducción de la detección mediante mamografías (Cowell *et al.,* 2013). Actualmente, representa alrededor del entre 20% y 25% de los carcinomas de mama (ASOCIACIÓN ESPAÑOLA CONTRA EL CÁNCER, 2018 y Ross *et al.,* 2013).

Debido a la incidencia del cáncer de mama en la población se realizan constantemente numerosos estudios. En ellos se ha determinado que tanto factores genéticos como no genéticos están implicados en la etiología de este. Los factores no genéticos incluyen la historia menstrual y reproductiva, el índice de masa corporal, el consumo de alcohol y la actividad física (Mavaddat *et al.,* 2010). Aunque dichos factores sean muy importantes y han de tenerse en cuenta, este trabajo se centra en los factores genéticos.

Mediante estudios en familias y en gemelos se ha demostrado la contribución de la susceptibilidad de la herencia en este cáncer. Se sabe que muchos loci genéticos contribuyen a este riesgo familiar, incluidos genes con mutaciones de alta penetrancia principalmente en los genes BRCA1 y BRCA2; alelos de riesgo moderado en genes como ATM, CHEK2 y PALB2; y alelos de baja penetrancia comunes, muchos los cuales

han sido identificados principalmente mediante estudios de asociación del genoma completo (GWAS) (Michailidou *et al.,* 2013). Además, TP53 es otro gen de alta penetrancia, conocido como el gen del Síndrome de Li-Fraumeni. Este gen es un supresor de tumores que tiene un papel importante en la regulación del crecimiento celular (Economopoulou *et al.,* 2015).

Por otro lado, los tumores se clasifican en cuanto a los receptores que presentan. Entre el 20% y el 30% de todos los cánceres de mama son negativos al receptor de estrógeno (ER), de ellos una mayor proporción para mujeres más jóvenes y de ascendencia africana. La etiología y el comportamiento clínico de los tumores ER negativo son diferentes al de los tumores que expresan el receptor de estrógeno, denominados ER positivos, incluidas las diferencias en la predisposición genética. Los tumores ER negativos se asocian con un peor pronóstico a corto plazo y están más débilmente asociados a los factores de riesgo reproductivo que los tumores ER positivo. En cuanto a la susceptibilidad genética, las mutaciones en BRCA1 predisponen generalmente la enfermedad a ER negativa, mientras que la mayoría de loci comunes conocidos para el cáncer de mama suele causar que el tumor sea ER positivo. Sin embargo, hay excepciones como la variación rs10069690 en el cromosoma 5 (gen TERT) o rs2284378 en el cromosoma 20 (gen RALY) que confieren mayor riesgo de tumores ER negativos (García-Closas *et al.,* 2014).

En cuanto a su clasificación por los receptores expresados también hay un tipo de cáncer denominado cáncer de mama triple negativo (TNBC). Los TNBC son aquellos que carecen de expresión del receptor de estrógeno (ER) y del receptor de progesterona (PR) y no muestran amplificación del gen del factor de crecimiento epidérmico humano 2 (HER2). Este tipo de cáncer corresponde normalmente al 15% de los cánceres de mama. Los TNBC están formados por un grupo heterogéneo de tumores, cuyas características principales, en comparación con otros subtipos de cáncer de mama, son su naturaleza agresiva, las tasas de recaída más altas y una supervivencia más corta por lo general. También afectan principalmente a mujeres más jóvenes y de ascendencia africana y suelen estar asociados a mutaciones en los genes BRCA1 y BRCA2. La media de supervivencia de las mujeres con este tipo de cáncer es de menos de un año y la mayoría suelen morir.

El problema de estos subtipos de cáncer es que al no presentar bien uno o tres receptores las terapias dirigidas no tienen objetivos moleculares donde poder actuar. En el caso del TNBC la única opción de tratamiento es la quimioterapia citotóxica, aunque

algunas pacientes responden el tratamiento es tóxico y un gran porcentaje recae en la enfermedad. Por lo que es fundamental la búsqueda de objetivos moleculares para mejorar la supervivencia de pacientes con este tipo de cáncer, aunque la heterogeneidad de la enfermedad lo dificulta (Abramson *et al.,* 2015).

También existe el cáncer de mama en hombres. Sin embargo, este no va a ser estudiado en este trabajo. Posteriormente, se realizará otro para esta enfermedad.

2. OBJETIVOS

La finalidad de la Medicina de Precisión es que las pacientes que deseen conocer si tienen mayor riesgo de sufrir cáncer de mama puedan hacerlo. Para ello, en los departamentos de prevención, se secuencia su ADN en busca de las variaciones asociadas con la enfermedad. El listado con variaciones asociadas al cáncer de mama que se pretende crear en este trabajo, facilitaría esta tarea, debido a que únicamente sería necesario secuenciar las regiones del genoma donde se encuentran las variaciones propuestas. Así, una vez obtenida la secuenciación se compararía con los alelos que están asociados al desarrollo de tumores en la mama. De este modo, se puede predecir la probabilidad de una mujer a sufrir cáncer de mama.

El objetivo principal de este trabajo es, desde la biotecnología, identificar la información genética relevante. Dicha identificación es necesaria debido a que existe un gran número de bases de datos con grandes cantidades de información y, en ellas, la información está repetida, es heterogénea y de calidad variable.

Además, en el caso de la búsqueda de variaciones, cuando una publicación hace referencia a una variación, bien porque sea estudiada en él o porque la cite para aportar más información a su trabajo, la variación aparece asociada directamente al fenotipo del que se está hablando. Al revisar la bibliografía, gran parte de las variaciones que aparecen asociadas a un fenotipo en la base de datos realmente no lo están. Debido a ello si se pretende obtener datos sobre un fenotipo es necesaria una correcta revisión previa.

El papel importante de la biotecnología en este contexto es evaluar la validez de las publicaciones que avalan cada una de las variaciones. Es importante estudiar cómo el lugar en el que se encuentra la variación afecta a la expresión génica, la síntesis de proteínas y, finalmente, al fenotipo que causa. También se ha de evaluar la relevancia estadística debido a que la mayoría de las variaciones, aunque inicialmente se crea que pueden estar relacionadas, no son significativas para el fenotipo de estudio.

Una vez identificadas las variaciones significativas para el cáncer de mama, el siguiente objetivo es que éstas sean cargadas en una base de datos, creada por el equipo de trabajo, para su posterior explotación con propósitos clínicos. Se pretende agrupar las variaciones de diferentes bases de datos en una sola, evitando así las duplicaciones de

información y proporcionando un único punto de acceso a la información. También, al introducir los datos en un mismo formato, la información se presentará de un modo homogéneo, facilitando la accesibilidad y la comprensión a los usuarios que precisen obtenerla.

Una vez estén cargadas en ella, se podrán realizar consultas sobre el cáncer de mama generándose un listado con las variaciones confirmadas como relacionadas con la enfermedad. De este modo, los profesionales sanitarios que se dedican a la prevención del cáncer de mama la podrán obtener fácilmente.

Para conseguir los objetivos propuestos es imprescindible seguir una metodología de búsqueda e identificación de variaciones. Esto significa investigar cuáles son las bases de datos que proporcionan mejor información con respecto a lo que se está buscando: variaciones en cáncer de mama. También, se ha de conocer el funcionamiento y la estructura de las fuentes de información, para que se pueda obtener la mayor cantidad de datos posible. Posteriormente y, una vez se tengan las variaciones proporcionadas por las bases de datos, se ha de determinar si son válidas para el diagnóstico clínico. Para ello la información asociada debe ser revisada detenidamente para asegurar que tienen significado estadístico relevante y que las variaciones realmente están asociadas con el fenotipo de interés.

Una vez las variaciones estén cargadas en la base de datos, se podrán realizar consultas sobre el cáncer de mama generándose un listado con las variaciones confirmadas como relacionadas con la enfermedad. De este modo, los profesionales sanitarios que se dedican a la prevención del cáncer de mama la podrán obtener fácilmente.

En conclusión, los objetivos se podrían resumir en:

- Desde la biotecnología, búsqueda e identificación de la información genética relevante.
- Obtención de variaciones genómicas para el cáncer de mama.
- Apoyar el desarrollo de la Medicina de Precisión.

A continuación se describe la metodología utilizada para la identificación de las variaciones.

3. METODOLOGÍA

3.1. METODOLOGÍA SILE

En el presente trabajo, se utiliza una metodología de *Búsqueda-Identificación-Carga-Explotación* (SILE, *Search-Identification-Load-Exploitation*). Como explican Burriel *et al.* (2017), el objetivo de SILE es la búsqueda, identificación, carga y explotación de la información relevante, es decir, las variaciones con significado clínico para el cáncer de mama, en la *Base de Datos del Genoma Humano* (HGDB), propia del grupo de trabajo. Debido a que el cáncer de mama no se basa únicamente en cuestiones biológicas, sino también en características clínicas y genéticas, la metodología SILE es una buena opción para unificar y utilizar de un modo eficiente toda la información disponible.

A continuación, se describe cada una de las fases de SILE:

3.1.1. Búsqueda

Consiste en la búsqueda y posterior selección, de las bases de datos que proporcionan información relevante para el fenotipo en cuestión. De este modo se conoce cuáles son las más adecuadas para dicha tarea. Seguidamente, se buscan los datos relacionados con el fenotipo en los repositorios escogidos.

3.1.2. Identificación

Es el proceso en el cual se analizan los resultados de las búsquedas específicas presentadas en la fase anterior. Se hace con el objetivo de eliminar duplicaciones, así como otros problemas como el no cumplimiento de determinados criterios de calidad, que serán explicados posteriormente. También se verifica que todas las variaciones pertenezcan a cáncer de mama femenino. Además, en este proceso de identificación se detectan problemas de datos tales como inconsistencia, que han de ser resueltos antes del proceso de carga.

3.1.3. Carga

El grupo ha desarrollado una base de datos específica, llamada Human Genome Database (HGDB), con el fin de gestionar de un modo eficiente los datos genómicos. Esta base de datos se basó en el *Modelo Conceptual del Genoma Humano* (CMHG), el cual se va a explicar en el siguiente apartado. Tanto HGDB como CMHG se desarrollaron con el objetivo de definir todas las características que forman el genoma

humano y, así, crear una estructura sólida sobre la que construir un S*istema de Información Genómica* (GeIS).

3.1.4. Explotación

La creación de conocimiento como resultado de la correcta gestión de datos es el objetivo principal de la última etapa de la metodología SILE. La fase de explotación consiste en señalar las variaciones genómicas presentes en una muestra de una paciente y compararlas con las contenidas en el conjunto de datos almacenado en la HGDB. La explotación del conjunto de datos finalmente conduce a una evaluación de riesgo genético basada en evidencias validadas disponibles en fuentes de datos públicos.

En este trabajo únicamente se llevan a cabo los dos primeros procesos (búsqueda e identificación), la carga y la explotación la realizan los compañeros del grupo de trabajo encargados de dichas tareas.

3.2. MODELO CONCEPTUAL DEL GENOMA HUMANO

Según explica Reyes (2018), el ámbito del genoma humano posee una gran cantidad de información biológica, la cual ha de ser capturada, manipulada y controlada de manera eficaz; por lo que son necesarios enfoques avanzados en ingeniería de sistemas de información. Para ello es necesaria la aplicación del modelado conceptual para poder comprender la información relevante en el dominio. De este modo, es más sencillo representarla con claridad, lo que permite desarrollar una estrategia de gestión de datos eficaz.

El objetivo principal del *Modelo Conceptual del Genoma Humano* (MCGH) es comprender los conceptos básicos que explican cómo la estructura genotípica se manifiesta en un fenotipo. Además, las bases de datos que se basan en el MCGH se desarrollan y diseñan desde una perspectiva holística, de modo que posibilitan la integración de diferentes fuentes de datos que representan diferentes perspectivas del conocimiento genómico.

El MCGH se basa en los elementos cromosómicos como bloques de construcción básicos, de este modo se logra una visión centrada en el concepto del cromosoma.

La versión del MCGH utilizada está organizada en cinco vistas principales:

1. Vista estructural: en ella se describe la estructura del genoma, está compuesta por los elementos básicos de la secuencia de ADN.
2. Vista de la transcripción: muestra los componentes y los conceptos relacionados con la síntesis de proteínas, posee los componentes implicados en el paso del ADN a los diferentes ARNs.
3. Vista de variación: en esta vista se caracterizan los cambios en la secuencia de referencia.
4. Vista de rutas metabólicas: con ella se enriquece el modelo, ya que se aporta información sobre las rutas metabólicas.
5. Vista de la bibliografía y el banco de datos: en ella se evalúa la fuente de cualquiera de las informaciones con el objetivo de establecer de donde proviene cada dato.

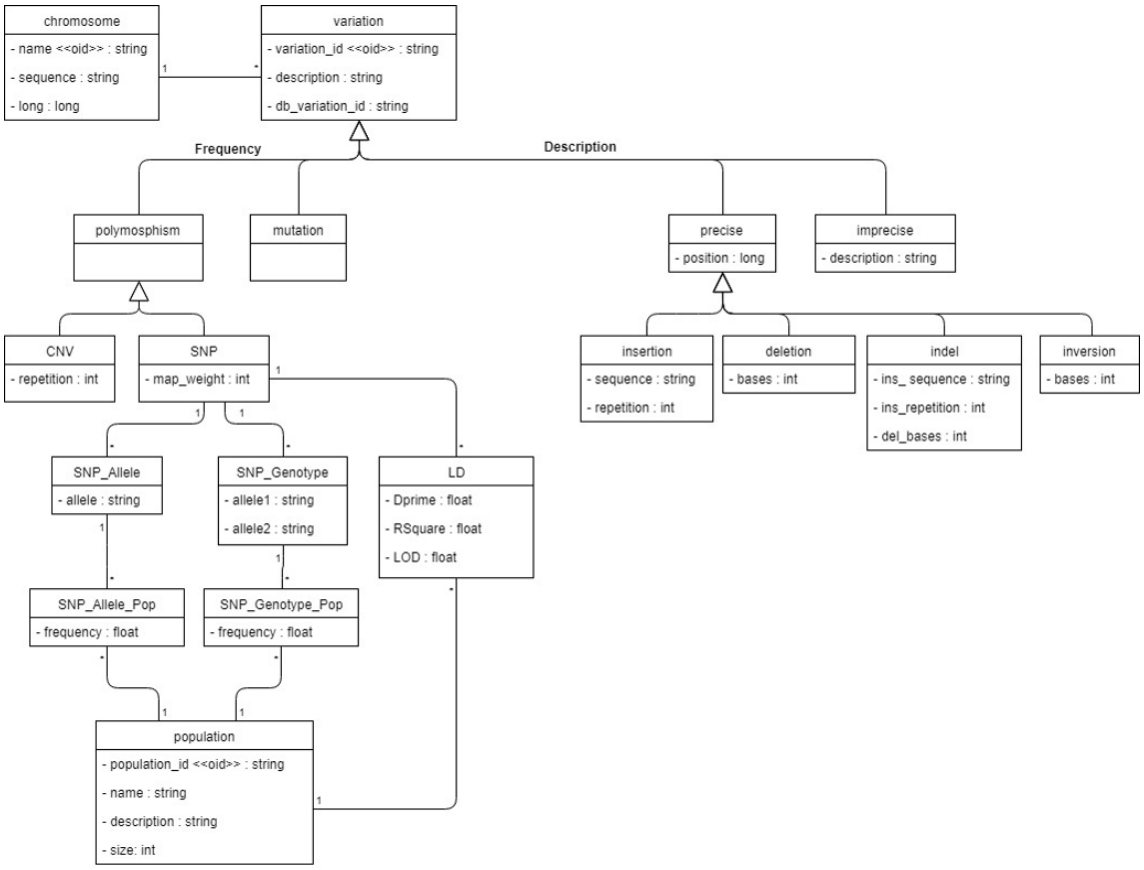

Figura 2. Imagen de la vista de variaciones (Reyes, 2018).

A continuación se detallan los criterios de calidad utilizados para la selección de las variaciones relevantes.

3.3. CRITERIOS DE CALIDAD

Los criterios de calidad se establecen con el objetivo de identificar qué variaciones son significativas para el cáncer de mama. Las variaciones que se descargan de las bases de datos van asociadas a una publicación donde es estudiada. Cada variación puede estar asociada a más de una publicación y, también, en una misma publicación puede ser estudiadas diferentes variaciones. Para comprobar si las variaciones cumplen los criterios de calidad se estudian las publicaciones asociadas en detalle, si los cumplen se puede decir que la variación es relevante para el fenotipo estudiado.

Primero se ha de comprobar que las variaciones tienen el fenotipo que se desea, el cual es cáncer de mama en mujeres. En este trabajo no se estudia el cáncer de mama en hombres ni tampoco otros estudios que no estén directamente relacionados con el cáncer de mama, como son el tamaño de los pechos por causas genéticas o la similitud entre pérdida de cabello debida a quimioterapia con la pérdida de cabello por causas naturales.

A continuación, se van a explicar, los tres criterios de calidad que se siguen para cada una de las variaciones:

1- Cada variación debe tener asociada al menos una publicación, la cual sustente la relación genotipo-fenotipo.
2- El significado clínico de la variación ha de ser distinto de benigno, probablemente benigno, probablemente patogénico, no provisto, incierto o conflicto en la interpretación.
3- Se ha de hacer una revisión manual de lo que dice cada publicación sobre la variación estudiada. En dicha revisión se han de confirmar los siguientes parámetros:
    a. El estudio ha tenido que ser realizado en humanos.
    b. En él han debido de participar más de 500 personas, entre muestra y control.
    c. La evidencia estadística ha de ser la siguiente:
        i. Para las variaciones patogénicas: OR[1] > 1 e IC[2] > 1.
        ii. Para las variaciones protectoras: OR < 0,9 e IC < 1.
        iii. Son deseables estudios que hayan sido reproducidos.

[1] Odds Ratio
[2] Intervalo de Confianza

El primer criterio es necesario porque para poder dar por válida una variación hay que comprobar los criterios en la publicación que tiene asociada. Sin la publicación no hay datos sólidos con los que confirmarla.

La importancia del segundo criterio se debe a que se busca que la certeza sea la mayor posible. El objetivo final del trabajo que se está realizando es encontrar las variaciones que causen riesgo o estén asociadas a padecer cáncer de mama para que en las mujeres que las posean se puedan llevar a cabo técnicas de prevención. Por otro lado, las variaciones que protejan del cáncer de mama son interesantes desde el punto de vista de si una mujer que la posee puede no padecer este cáncer. Además, no se desean las variaciones cuyo significado clínico tiene conflictos en su interpretación, ya que quiere decir que se han hecho diversos estudios sobre ella y no se concluye si está relacionada con el cáncer.

El tercer criterio es el más complicado de aplicar y el que precisa de un mayor conocimiento biológico o biotecnológico. Para llevarlo a cabo se han de estudiar las publicaciones y concluir si en ellas, la variación deseada cumple ciertos parámetros. El estudio ha de haber sido realizado en humanos, no sirven extrapolaciones en modelos animales, ya que para confirmar que la variación verdaderamente está relacionada con el cáncer de mama se ha de asegurar que se ha comprobado con otras mujeres.

En el estudio examinado han tenido que participar más de 500 mujeres, entre muestra y control. Debido al gran número de trabajos que existen sobre el cáncer de mama y la gran cantidad de mujeres que lo sufren en el mundo, es fácil encontrar trabajos con grandes poblaciones.

Por último, en cuanto a la evidencia estadística, la probabilidad de que la variación esté asociada con el fenotipo se ha de medir con el cociente *Odds Ratio* (OD), existente otros cocientes semejantes a él pero en el presente trabajo uno de los criterios es que sea OD. El *Odds Ratio* es el cociente entre la probabilidad que hay de que ocurra un evento y la probabilidad que hay de que no ocurra, por lo que para que haya mayor probabilidad de que ocurra que de que no ocurra ha de ser mayor que 1. Además, se ha de presentar el intervalo de confianza (IC) para dicho OR, el cual será del 95%. En el caso de las variaciones con significado clínico patogénico (factor de riesgo, asociación, susceptibilidad…) el OD tiene ser mayor que 1 y el IC no debe contener el valor 1, es decir, su límite inferior ha de ser mayor que 1. Por el contrario, para las variaciones que

son protectoras del cáncer de mama el OD ha de ser menor que 0,9 y el IC no puede contener el valor 1 o valores menores de 1. Por último, debido a que en los estudios se pueden producir fallos experimentales (errores), interesa que hayan sido replicados, porque eso indica mayor confianza en los datos obtenidos que si únicamente se ha realizado una vez.

3.4. BASE DE DATOS UTILIZADAS

El primer paso en este trabajo fue buscar información sobre las diferentes bases de datos genómicas para escoger las más adecuadas. Una vez contrastada se decidió utilizar ClinVar, Ensembl, GWAS catalog y SNPedia. Un motivo común a todas ellas es que son públicas, por lo que cualquier usuario puede consultar datos. Además, otros de las razones son:

En el caso de ClinVar, esta almacena informes de las relaciones entre variaciones humanas y fenotipos proporcionando como respaldo referencias a las evidencias bibliográficas de ello (PubMed). De este modo, ayuda a conocer la relación que existe entre la variación genómica y el fenotipo observado. La información que almacena esta base de datos son las variaciones, su significado clínico, la información sobre el remitente y otros datos que sean útiles y lo respalden. ClinVar permite representar desde un alelo y su interpretación, hasta múltiples tipos de evidencia experimental sobre el efecto de la variación en el fenotipo (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2016).

Ensembl se diferencia de ClinVar en que es un repositorio de genomas de varias especies, no únicamente del genoma humano. Esta se creó con el objetivo de anotar el genoma humano e integrar dicha anotación con otros datos biológicos. Además, se ha incluido genómica comparativa, variaciones y datos regulatorios, lo cual la hace una base de datos idónea para la búsqueda de variaciones y el efecto que estas tienen en el ser humano (EMBL-EBI, 2018). Posee una herramienta llamada BioMart que permite la descarga de los datos almacenados por Ensembl, facilitando su posterior procesamiento.

GWAS catalog se fundó en respuesta al aumento en el número de estudios publicados de asociaciones genómicas (GWAS). Dichos estudios permiten investigar el impacto de las variaciones comunes en una enfermedad compleja. Mediante esta plataforma se

puede acceder a la información sobre sobre estos estudios con y un resumen de las asociaciones observadas. Una vez obtenida la información de los estudios a partir del texto, GWAS catalog se encarga de evaluarla, extraer el fenotipo, y las asociaciones significativas del SNP al fenotipo y los metadatos de la muestra poblacional (EMBL-EBI, 2017).

Para finalizar, SNPedia es una página de libre acceso que se basa en un modelo wiki, en la que los usuarios pueden editar y añadir su información. Esto se hace con el fin de fomentar la comunicación sobre la variación genética y permitir que los miembros de la comunidad ayuden a que evolucione de modo que se vuelva cada vez más relevante. En ella se almacenan datos de polimorfismos de un único nucleótido (SNP) obtenidos de otras fuentes de información, principalmente de otras bases de datos (SNPEDIA). Los usuarios adicionan datos de forma continua, lo cual se complementa con actualizaciones periódicas de texto extraído de fuentes de datos públicos. Las fuentes citan publicaciones, en particular identificadores PubMed PMID (Cariaso y Lennon, 2012).

Una vez seleccionadas las bases de datos, se descargó de ellas la información de las variaciones genómicas asociadas al fenotipo de interés. Estas variaciones se buscan filtrando por “breast cancer”, pueden ser variaciones bien que lo causen, que protejan de él, etc., es decir, que estén relacionadas con el cáncer de mama. Junto con cada una de las variaciones se obtiene el identificador de la publicación o los identificadores de las publicaciones en las cuales se ha estudiado. En estos artículos se da, además, información sobre las siguientes características: cromosoma, gen, significado clínico y fenotipo. También se explica la población sobre la que se ha realizado el estudio y los parámetros estadísticos resultantes.

4. RESULTADOS Y DISCUSIÓN

Las variaciones genómicas para el cáncer de mama obtenidas, aplicando la metodología explicada, son 125 (*Anexo, Tabla 1*). Todas las variaciones encontradas son SNP, es decir se deben únicamente al cambio de una base nitrogenada en una posición exacta de la secuencia de ADN. Estas han sido extraídas de cuatro bases de datos: ClinVar, Ensembl, GWAS catalog y SNPedia y se han seleccionado según los criterios explicados anteriormente. Algunas de las variaciones tienen asociadas más de una publicación, por lo que se han confirmado en base a más de un estudio, obteniéndose en total 161 publicaciones.

Para cada una de las variaciones genómicas se tiene información del cromosoma en el que se encuentran, el gen del cual modifican su expresión, el significado clínico, el fenotipo que causa dicha variación de la secuencia del genoma, la población en la que se ha realizado el estudio por la que se confirma, el identificador o identificadores de los artículos que apoyan el efecto de la variación y su relación con el fenotipo estudiado y, por último, la base o bases de datos de la que ha sido extraída (*Anexo, Tabla 1*).

4.1. VARIACIONES OBTENIDAS

A continuación se va a explicar las variaciones que se han obtenido en cada una de las bases de datos y el número final de ellas una vez filtradas.

Cuando se descargan las variaciones asociadas a un fenotipo determinado, en este caso “breast cancer”, se obtiene la información por publicaciones. Es decir, las variaciones pueden tener asociadas a ella desde una a numerosas publicaciones y en cada publicación también se puede estudiar una o muchas variaciones. Para cada asociación de una variación con una publicación se aporta información de la posición en la que se produce, cromosoma, gen, etc. y datos del estudio tales como población *Odds Ratio,* intervalo de confianza, etc. En ClinVar, el total de las publicaciones asociadas a una variación fue de 1.224, en Ensembl 42.926, en GWAS catalog 1.569 y en SNPedia 409.

Después de esto se pasan los dos primeros filtros de los explicados anteriormente. Estos son que el PMID pertenezca a los obtenidos en la lista PubMed; que el significado

clínico sea distinto de benigno, probablemente benigno, probablemente patogénico, no proporcionado, incierto y conflicto en la interpretación. Con ello, en ClinVar se obtuvieron 169 publicaciones asociadas a una variación, en total había 43 publicaciones; en Ensembl 8.982 publicaciones y 1.547 variaciones; en GWAS catalog 1.282 publicaciones, 992 variaciones y; por último, en SNPedia 207 publicaciones y 14 variaciones.

Todas ellas fueron filtradas de nuevo, como se ha explicado en la metodología mediante revisión bibliográfica manual y se llegó a la tabla de variaciones final (*Anexo, Tabla 1).* Como se observa en la *Figura 3*, ninguna de las variaciones se encuentra en todas las bases de datos. Solamente 6 de ellas pertenecen a Ensembl, GWAS catalog y SNPedia; y una de ellas a Ensembl, GWAS catalog y ClinVar. La mayor parte de ellas, es decir 136 de las 161, aparecen tanto en Ensembl como en GWAS catalog.

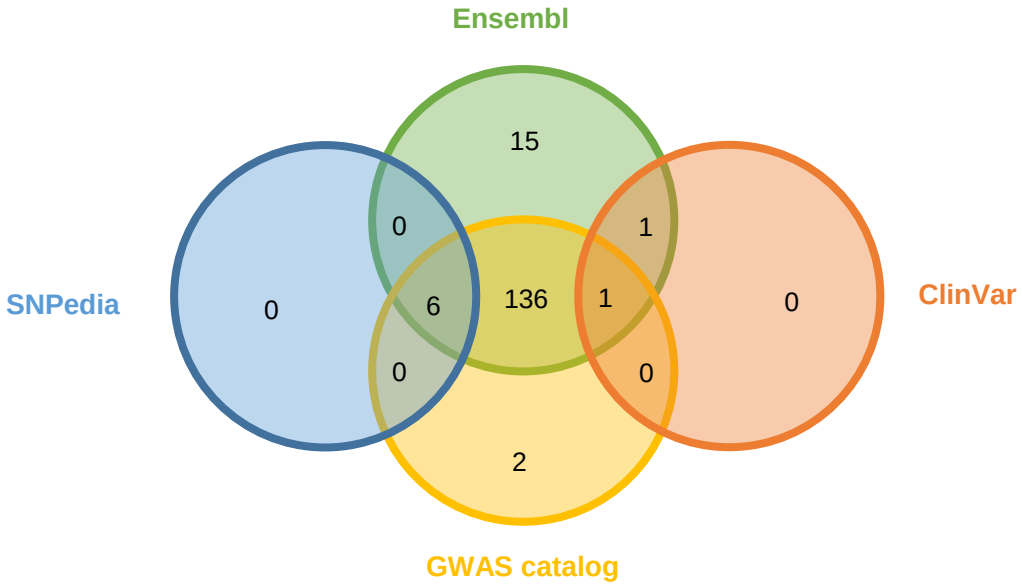

Figura 3. Número de variaciones obtenidas de cada una de las bases de datos (o de más de una de ellas).

A continuación, se va a analizar el conjunto de las variaciones genómicas para cada una de las características que aportan información sobre ella:

4.2. CROMOSOMA

Al estudiar en qué cromosomas aparecen las variaciones se observa que no aparecen en ninguno de los dos cromosomas sexuales (X e Y), sin embargo sí aparecen en todos los demás al menos una vez. Como se presenta en la *Figura 4,* en el que más variaciones hay es en el cromosoma 5, con 22 variaciones; seguido del 10 con 15 y en el resto el número de variaciones es igual o inferior a 10.

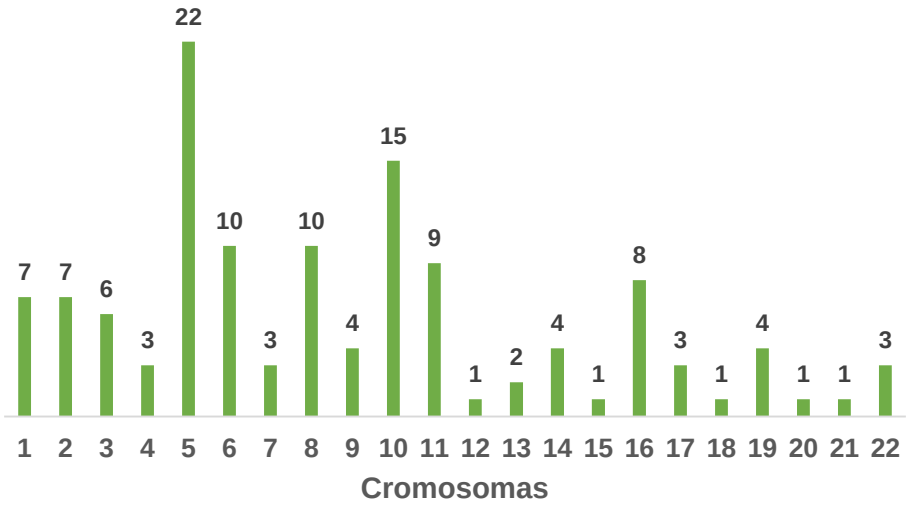

Figura 4. Número de variaciones que hay en cada uno de los cromosomas.

4.3. GEN

En cuanto a los genes que afectan a cada una de las variaciones, bien porque se encuentren en ellos o porque estén cerca y afecten a su regulación, el gen en el que presenta un mayor número de variaciones es MAP3K1, seguido de FGFR2, en el cual se dan 7 variaciones (*Tabla 1)*. Además, hay 37 variaciones que no están relacionadas con ningún gen, debido a que se encuentran en ADN intergénico. Sin embargo, la mayoría de genes únicamente tienen una variación (*Anexo, Tabla 2).* Para saber cuál es el gen que se ve afectado, además de revisar la bibliografía correspondiente, se observa el gen que mapea en la posición en la que se encuentra la variación.

Tabla 1. Número de variaciones que hay en cada uno de los genes.

Gen	Nº variaciones	Gen	Nº variaciones	Gen	Nº variaciones
MAP3K1	9	TERT	3	TOX3	2
FGFR2	7	AS1	2	ZMIZ1	2
ESR1	5	LINC01376	2	ZNF365	2
CASC16	3	LSP1	2	-	37
PCAT1	3	TCF7L2	2	Otros	44

- : Variaciones que no se encuentran asociadas a ningún gen.

A continuación se van a explicar en más profundidad los genes MAP3K1 y FGFR2 ya que son los que más variaciones tienen:

MAP3K1 es el gen que codifica para la *proteína activada por mitógenos quinasa quinasa quinasa 1*, también denominada MAP3K1. Dicha proteína, que es una serina/treonina quinasa, participa en la vía de señalización MAPK y desempeña un papel fundamental en la regulación de genes importantes del cáncer. Las MAPK regulan diversas funciones celulares al modular la actividad del factor de transcripción para afectar la expresión génica. En concreto, el gen MAP3K1 regula el desarrollo y la función del sistema inmune, la reparación de lesiones y la progresión tumoral. Además, recientes estudios han demostrado que este gen es un marcador de susceptibilidad genética para algunos tumores, entre ellos el cáncer de mama. Ciertos alelos relacionados con el riesgo a sufrir cáncer de mama aumentan la expresión del gen MAP3K1 *in vivo* y pueden promover la supervivencia de las células cancerígenas (Xu *et al,* 2016).

En el gen MAP3K1, el cual se halla en el cromosoma 5, se han encontrado las variaciones rs1017226, rs16886034, rs16886113, rs16886165, rs16886181, rs16886364, rs16886448, rs2229882 y rs3822625. Todas ellas, excepto rs16886165, están asociadas a cáncer de mama temprano. Además, la variación rs16886034 se encuentra dentro de la región reguladora del gen, por lo que modifica la expresión de este gen.

Por otra parte, el gen **FGFR2** codifica para la proteína denominada *receptor del factor de crecimiento de fibroblastos 2*. La familia de los FGFR tiene gran importancia en la supervivencia y la proliferación celular ya que pueden actuar como reguladores negativos de la proliferación y como reguladores positivos de la diferenciación (Grose *et al*, 2005). El FGFR2 es uno de los genes comunes de baja penetrancia, codifica un

receptor de tirosina quinasa y ha sido identificado como un gen potencial de susceptibilidad al cáncer de mama, ya que se amplifica y se sobre expresa en este cáncer. La expresión aberrante de isoformas de FGFR2 debidas a errores en el *splicing alternativo* transforma las células en células cancerígenas mediante una transducción de señal sostenida (Zhang *et al,* 2010).

En el gen FGFR2, perteneciente al cromosoma 10, están las variaciones rs10510102, rs11200014, rs1219648, rs2420946, rs2912774, rs2981578 y rs2981579.

4.4. SIGNIFICADO CLÍNICO

Como se puede observar en la *Tabla 2,* de las 125 variaciones hay 122 cuyo significado clínico es asociación, esto corresponde al 97,60% del total. El 1,60% siguiente, es decir 2 variaciones tienen como significado clínico factor de riesgo y, por último, solo una de las variaciones que han cumplido los criterios de calidad, es decir el 0,80% del total, es protectora frente al cáncer de mama.

Tabla 2. Distribución de las variaciones según su significado clínico.

Significado clínico	Nº variaciones	% del total
Asociación	122	97,60
Factor de riesgo	2	1,60
Protectora	1	0,80

Que el significado clínico de una variación sea asociación quiere decir que en los estudios realizados se ha observado que dicha variación es significativa en la población enferma y que no lo es en la control, es decir, que existe una relación entre dicha variación y el fenotipo cáncer de mama. Por otro lado, que una variación sea un factor de riesgo para el cáncer de mama significa que las mujeres que la presenten tienen mayor probabilidad de sufrirlo, debido a que dicha variación puede ser que lo provoque. Sin embargo, que una variación sea protectora se refiere a que tenerla hace que la mujer posea un menor riesgo de padecer dicho cáncer ya que la variación la protege.

4.5. FENOTIPO

En cuanto al estudio del fenotipo (*Anexo, Tabla 3*) y, como se puede observar en la *Figura 5*:

- El 68% de las variaciones, lo que corresponde a 86, causan cáncer de mama en general, no ningún tipo concreto de este cáncer.
- El 12% de ellas, es decir, 15 variaciones, provoca cáncer de mama temprano. Este ocurre en las mujeres más jóvenes.
- El 11% del total, 14 variaciones, causa cáncer de mama en el cual las células no son capaces de reconocer señales de estrógeno, este subtipo se denomina cáncer de mama de receptor con estrógeno negativo (ER-).
- Un poco menos del 3% de variaciones, correspondiente a 3 de ellas, conlleva un cáncer de mama que sí que es capaz de reconocer las señales de estrógeno, llamado cáncer de mama con receptor de estrógeno positivo (ER+). Este fenotipo no significa que únicamente sea positivo para el ER, sino que solamente ha sido confirmada la variación para ese receptor. Podrían hacerse estudios posteriores que demostrasen que ocurre en un tipo general de cáncer de mama.
- Dos de las variaciones, lo que representa menos del 2% del total, están asociadas a cáncer de mama triple negativo (TN). En este tipo de tumores, las células no son capaces de detectar señales de estrógeno, de progesterona ni del factor de crecimiento epidérmico humano.
- Por último, el resto de variaciones causan fenotipos de susceptibilidad al cáncer de mama, de protección frente a este, neoplasma de mama y, además, un cáncer de mama que se produce cuando las mujeres presentan dichas variaciones y se les da una terapia de reemplazo hormonal durante la menopausia.

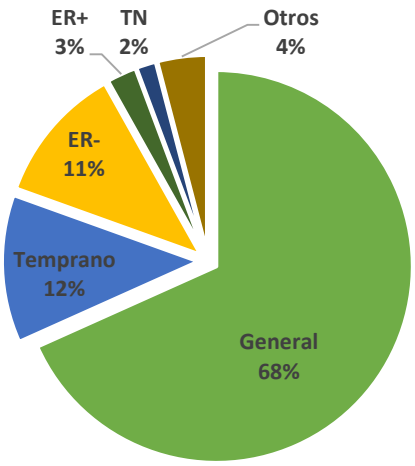

Figura 5. Distribución de las variaciones según su fenotipo.

ER- : Receptor de estrógeno negativo. ER+: Receptor de estrógeno positivo. TN: Triple negativo.

4.6. POBLACIÓN

Al analizar los artículos en los que se confirman las variaciones genómicas es importante conocer la población para la cual se ha realizado el estudio, ya que hay variaciones que son distintas en las diferentes poblaciones. Esto se debe a que en cada región geográfica y en cada raza se conservan unas mutaciones determinadas, las cuales pueden diferir del resto de individuos de otros lugares o etnias.

El cáncer de mama es una enfermedad muy estudiada dada su incidencia en la población, por lo que se realizan muchos estudios sobre esta enfermedad. Además, como afecta a mujeres en todo el mundo se tiene una gran cantidad de datos sobre estas pacientes, lo cual permite realizar grandes estudios con población global. Sin embargo, hay variaciones que se confirman únicamente para una población más pequeña. Esto puede ser debido a que únicamente se presente en dicha población o a que el estudio se haya realizado solamente para ella.

De las variaciones obtenidas en este estudio y, como se observa en la *Tabla 3*, de los 161 estudios realizados que confirman las 125 variaciones, únicamente 7 se han confirmado para la población que representa a la totalidad de las mujeres del planeta. La mayor parte de los estudios, que son 119, se ha realizado sobre población europea. Dentro de la población europea se diferencia la británica, para la que se han realizado 11 estudios. En cuanto a la población asiática, hay 14 publicaciones que confirman

variaciones en mujeres de Asia oriental, 1 en chinas y 2 en chinas de la etnia Han. Por último, hay 3 estudios realizados para población africana, 1 para americana y 3 para afroamericana.

Tabla 3. Distribución de las variaciones según la población sobre la que se han realizado los estudios.

Población	Nº variaciones	% del total
Todas	7	4,34
Europea	119	73,91
Asiática oriental	14	8,69
Británica	11	6,83
Africana	3	1,86
Afroamericana	3	1,86
China Han	2	1,24
Americana	1	0,62
China	1	0,62

Como se ha explicado en el apartado *3.3. Criterios de calidad,* para que un estudio sobre una variación sea aceptado como válido ha de superar los 500 participantes. También se ha explicado que al tratarse de una enfermedad con una gran incidencia existen estudios con un gran número de participantes. De las variaciones confirmadas, el número de participantes más pequeño es de 541 y el mayor de 1.195.919. El promedio de los participantes en todos los estudios es 62.443, lo que confirma que las muestras poblacionales son numerosas.

5. DISCUSIÓN

En el presente trabajo se han encontrado 125 variaciones significativas para el cáncer de mama en función de los criterios aplicados, explicados en el apartado *3.3. Criterios de calidad*. En un principio, contando las cuatro bases de datos utilizadas (ClinVar, Ensembl, GWAS catalog y SNPedia), se encontraron 16.240 variaciones diferentes; muchas de ellas, como se ha explicado en el punto *4.7. Variaciones obtenidas,* aparecían en más de una de las bases de datos. Con ello, se confirma la dificultad supuesta en la búsqueda de las bases de datos.

Michailidou *et al.* (2017), confirmaron 65 nuevos loci relacionados con el cáncer de mama. Según los criterios aplicados en este estudio se encontró una relación significativa en 38 de ellos. Este mismo grupo, anteriormente, confirmó 41 loci también relacionados con el cáncer de mama y en nuestro estudio, de ellos se relacionaron 22 (Michailidou *et al.* ,2013).

Según se indica en la publicación realizada por Joshi *et al.* (2014), asociaron 9 SNPs como factores de riesgo del cáncer de mama y 23 conllevaban susceptibilidad a él. Sin embargo, según los criterios utilizados, en este estudio se llegó a la conclusión de que únicamente 2 de estas variaciones podían relacionarse verdaderamente con el cáncer de mama.

También, Fletcher *et al.* (2011) confirmaron 5 variaciones como relacionadas con el cáncer de mama. De ellas, según el estudio realizado, se confirmó únicamente una.

Por otra parte, existen estudios como el realizado por Cai *et al.* (2011), en el cual se confirmó solamente una variación para la población de Asia oriental. Según los criterios seguidos en el presente trabajo, dicha variación también está relacionada con el cáncer de mama.

De ello se puede extraer que los estudios GWAS proporcionan una oportunidad sin precedentes para investigar el impacto de las variaciones comunes en diferentes individuos en las enfermedades complejas, ya que gracias a ello se han obtenido y, se siguen conociendo, numerosas variaciones que están asociadas a enfermedades o fenotipos de interés. Sin embargo, la identificación de un GWAS puede ser un desafío para los investigadores debido a que muchas de las variaciones que se encuentran en

ellos, como se ha discutido en los últimos párrafos, al ser analizadas no presentan verdadera relación con la enfermedad estudiada, lo cual puede conllevar a falsos positivos en los test de prevención.

En cuanto a los genes en los que se han confirmado las variaciones, una de las diferencias más notable entre la bibliografía consultada y los resultados obtenidos es que únicamente se ha encontrado una variación en el gen BRCA1 y otra en el BRCA2. Sin embargo, en un gran número de artículos sobre el cáncer de mama nombran estos genes como los principales causantes de la enfermedad, en caso de que se produzcan mutaciones en ellos.

En la década de 1990 se descubrió que mutaciones en los genes supresores de tumores BRCA1 y BRCA2 estaban relacionadas con mayor riesgo a padecer cáncer de mama (Mavaddat *et al.,* 2010). Entre el 5% y el 10% de los cánceres de mama pueden deberse a la herencia y, de ellos, más del 30% se deben a mutaciones en dichos genes (Economopoulou *et al.,* 2015). Según indican Tung *et al.* (2015) el 3,9% de las mutaciones encontradas en este estudio fueron en los genes BRCA1 y BRCA2. Según indican Michailidou *et al.* (2013), de las 41 variaciones que encontraron una se encuentra en el gen BRCA2.

No aparecen variaciones para otro de los genes considerados de alta penetrancia, el gen TP53. Según indican Economopoulou *et al.* (2015) está relacionado con aproximadamente el 1% de los casos de cáncer de mama.

Además, de los genes considerados de riesgo moderado, que son ATM, CHEK2 y PALB2; en el presente trabajo únicamente se encuentra una variación en el gen ATM. Sin embargo, según Schutte *et al.* (2003), una deleción en la posición 1100 del gen CHEK2 tiene contribución a la susceptibilidad al cáncer de mama. También, en base a lo estudiado por Fernandes *et al.* (2014), el gen PALB2 tiene una prevalencia del 0,8% en pacientes de familias con cáncer de mama. Sin embargo, los estudios realizados no son relevantes en cuanto a los criterios de calidad establecidos por lo tanto es necesario realizar más estudios que puedan confirmar estos hallazgos.

Según el fenotipo, en este estudio se ha demostrado que el 68% de las variaciones causan cáncer de mama normal. Todos los cánceres de mama considerados normales

expresan receptores para el estrógeno, lo que se confirma con lo que indican Huynh *et al.* (2012): entre el 60% y el 70% de los tumores expresan receptores para el estrógeno. Este es el receptor más estudiado en los tumores en las mamas ya que si se expresa se puede dar a la paciente un tratamiento que tiene como objetivo dicho receptor. En la subsección de fenotipo se diferencia uno que es cáncer de mama positivo para el ER. Esto se hace porque se ha asegurado en los estudios analizados que es positivo para él, pero no en todos se ha estudiado que únicamente sea positivo para el receptor de estrógeno sino que lo puede ser para el resto. En este trabajo se ha obtenido que el cáncer de mama ER+ corresponde al 3%, junto con el 68% del cáncer de mama en general, se puede decir que se encuentra en los valores esperados.

Por otra parte, las variaciones que provocan tumores que son negativos para el receptor del estrógeno según el presente trabajo corresponden al 11% del total. Sin embargo, según Doane *et al.* (2006), el porcentaje que representan estos tumores es de entre el 25% y el 30% de los que causan cáncer de mama. Las cifras pueden variar porque no todos los estudios tienen en cuenta las diferencias entre los receptores que expresan los diferentes tumores, sino que analizan a todas las mujeres asociándolas al fenotipo cáncer de mama.

En cuanto al cáncer de mama triple negativo (TNBC), según Bauer *et al.* (2017), el 15% de tumores de cáncer de mama corresponden con este fenotipo. En nuestro estudio, de las variaciones encontradas solamente el 2% de ellas conllevan TNBC. Es probable que esta diferencia en los resultados se deba a que no todos los estudios que se realizan analizan por separado las diferencias para los tres receptores, como se ha explicado en el caso anterior. Además, debido a que estos tumores son muy agresivos el porcentaje de supervivencia de las mujeres que lo padecen es del 14%, con lo que es más difícil obtener muestras de ellas.

6. CONCLUSIONES

Las conclusiones a las que se ha llegado una vez realizado este trabajo son:

- Ha sido posible la búsqueda e identificación de información genética relevante ya que se ha logrado obtener la información deseada de las diferentes bases de datos e identificar las variaciones cumpliendo los criterios propuestos.
- Por ello, se ha conseguido un listado de variaciones genómicas para el cáncer de mama que están significativamente asociadas a la enfermedad.
- Por último, con dicho listado se puede contribuir a la prevención del cáncer de mama, por lo que se ha podido apoyar, aunque sea con una pequeña aportación, al desarrollo de la Medicina de Precisión.

7. BIBLIOGRAFÍA

ABRAMSON, V. G.; LEHMANN, B. D.; BALLINGER, T. J. y PIETENPOL, J. A. (2015). Subtyping of triple-negative breast cancer: implications for therapy. *Cancer,* 121(1): 8-16.

ASOCIACIÓN ESPAÑOLA CONTRA EL CÁNCER (2018). Madrid, visto el 2 de Julio de 2018
https://www.aecc.es/es/todo-sobre-cancer/tipos-cancer/cancer-mama

BAUER, K. R.; BROWN, M; CRESS, R. D.; PARISE, C. A. y CAGGIANO, V. (2007). Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype. *Cancer,* 109(9): 1721-1728.

BIANCO, A. M.; MARCUZZI, A.; ZANIN, V.; GIRARDELLI, M.; VUCH, J. Y CROVELLA, S. (2013). Database tool in genetic diseases research. *Genomics,* 101(2): 75-85.

BOEKE, J. D. *et al.* (2016). The Genome Project-Write. *Science,* 353(6295): 126-127.

BROOKES, A. J. y ROBINSON, P. N. (2015). Human genotype-phenotype databases: aims, challenges and opportunities. *Nature Reviews Genetics,* 16(12): 702-715.

BURRIEL, V.; REYES, J. F.; CASANOVES, A. H.; INIGUEZ-JARRIN, C. y LEÓN, A. (2017). GeIS based on Conceptual Models for the Risk Assessment of Neuroblastoma. *International Conference on Research Challenges in Information Science.* 451-452 pp. doi:10.1109/RCIS.2017.7956581

BUTKIEWICZ, M. y BUSH, W. S. (2016). In silico functional annotation of genomic variation. *Current protocols in human genetics,* 88(1): 6-15.

CAI, Q. *et al.* (2011). Genome-wide association study identifies breast cancer risk variant 10q21.2: results from the Asia Breast Cancer Consortium. *Human Molecular Genetics,* 20(24): 4991-4999.

CARIASO, M. y LENNON, G. (2012). SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acid Research,* 40(1): 1308-1312.

COWELL, C.F.; WEIGELT, B.; SAKR, R.A.; NG, C. K. Y.; HICKS, J.; KING, T. A. y REIS-FILHO, J. S. (2013). Progression from ductal carcinoma in situ to invasive breast cancer: revisited. *Molecular Oncology,* 7(5): 859-869.

DOANE, A. S.; DANSO, M.; LAL, P.; DONATON, M.; ZHANG, L.; HUDIS, C. y GERALD, W. L. (2006). An estrogen receptor-negative breast cancer subset characterized by a

hormonally regulated transcriptional program and response to androgen. *Nature Oncogene,* 25:3994-4008.

ECONOMOUPOULOU, P; DIMITRIADIS, G. y PSYRRI, A. (2015). Beyond BRCA: New hereditary breast cancer susceptibility genes. *Cancer Treatment Reviews,* 41(1): 1-8.

EMBL-EBI, 2017. *European Molecular Biology Laboratory*, Meyerhofstraße, Heidelberg, Germany, visto el 28 de Junio del 2018
https://www.ebi.ac.uk/gwas/docs/about

EMBL-EBI, 2018. *Wellcome Genome Campus, Hinxton,* Cambridgeshire, UK, visto el 28 de Junio de 2018
https://www.ensembl.org/info/about/index.html

FERNANDES, P. H.; SAAM, J.; PETERSON, J.; HUGHES, E.; KALDATE, R.; CUMMINGS, S.; THEISEN, A.; CHEN, S.; TROST, J. y ROA, B. B. (2014). Comprehensive sequencing of PALB2 in patients with breast cancer suggest PALB2 mutations explain a subset of hereditary breast cancer. *Cancer,* 120(7): 963-967.

FLETCHER, O. *et al.* (2011). Novel Breast Cancer Susceptibility Locus at 9q31.2: Results of a Genome-Wide Association Study. *Journal of the National Cancer Institute,* 103(5): 425-435.

FRADKIN, J. E.; HANLON, M. C. y RODGERS, G. P. (2016). NIH Precision Medicine Initiative: implications for diabetes research. *Diabetes Care,* dc160541.

GARCIA-CLOSAS, M. *et al.* (2014). Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nature Genetics,* 45(4): 392-398.

GLOBAL CANCER OBSERVATORY: CANCER TODAY (2012). *World Health Organization,* Lyon, France, visto el 2 de Julio de 2018
http://gco.iarc.fr/today/online-analysis-multi-bars?mode=cancer&mode_population=continents&population=900&sex=2&cancer=29&type=0&statistic=0&prevalence=0&color_palette=default

GROSE, R. y DICKSON, C. (2005): Fibroblast growth factor signalling in tumorigenesis. *Cytokine & growth factor reviews,* 16(2): 179-186.

HUYNH, K.T.; CHONG, K. K.; GREENBERG, E. S. y HOON, D. S. (2012). Epigenetics of estrogen receptor-negative primary breast cancer. *Expert Review of Molecular Diagnostics,* 12(4): 371-382.

JESINGER, R. A. (2014). Breast anatomy for the interventionalist. *Techniques in Vascular and Interventional Radiology,* 17(1): 3-9.

JOSHI, A. D. *et al.* (2014). Additive Interactions Between Susceptibility Single-Nucleotide Polymorphisms Identified in Genome-Wide Association Studies and Breast Cancer Risk Factors in the Breast and Prostate Cancer Cohort Consortium. *American Journal of Epidemiology,* 180(10): 1018-1027.

MAVADDAT, N.; ANTONIOU, A. C.; EASTON, D. F. y GARCIA-CLOSAS, M. (2010). Genetic susceptibility to breast cancer. *Molecular Oncology,* 4(3): 174-191.

MICHAILIDOU, K. *et al.* (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genetics,* 45(4): 353-361.

MICHAILIDOU, K *et al.* (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature,* 551(7678): 92-94.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2016. *U.S. National Library of Medicine,* Rockville Pike, Bethesda MD, USA, visto el 28 de Junio de 2018 https://www.ncbi.nlm.nih.gov/clinvar/intro/

NATIONAL CANCER INSTITUTE (2018). *Breast Cancer Treatment (PDQ).* National Cancer Institute. Bethesda, USA.

NATIONAL RESEARCH COUNCIL (2011). *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease.* National Academies Press. Washington, D.C., USA. 109 pp.

PETERSSON, G. (Ed.) y BREUL, J. (Ed.) (2017). *Cyber Society, Big Data and Evaluation.* Routledge. New York.

REYES, J. F. (2018). *Diseño y Desarrollo de un Sistema de Información Genómica Basado en un Modelo Conceptual Holístico del Genoma Humano.* Universitat Politècnica de València. 279 pp.

ROSS, D. S.; WEN, Y. H. y BROGI, E. (2013). Ductal carcinoma in situ: morphology-based knowledge and molecular advances. *Advances in Anatomic Pathology,* 20(4): 205-216.

SCHUTTE, M. *et al.* (2003). Variants in CHEK2 Other than 1100delC Do Not Make a Major Contribution to Breast Cancer Susceptibility. *American Journal of Human Genetics,* 72(4): 1023-1028.

SHIN, C.; HAN, C.; PAE, C. U. y PATKAR, A. A. (2016). Precision medicine for psychopharmacology: a general introduction. *Expert review of neurotherapeutics,* 16(7): 831-839.

SNPEDIA. Visto el 28 de Junio de 2018 https://www.snpedia.com/index.php/SNPedia:About

SOCIEDAD ESPAÑOLA DE ONCOLOGÍA MÉDICA (2018). *Las cifras del cáncer en España 2018.* Madrid. 23 pp.

TUNG, N. *et al.* (2015). Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing with a 25-gene panel. *Cancer,* 121(1):25-33.

VAN DIJK, E. L.; AUGER, H.; JASZCZYSZYN, Y. y THERMES, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics,* 30(9): 418-426.

XU, M.; XU, Y.; CHEN, M.; ZHU, J.; ZHANG, M.; CHEN, Z.; ZHANG, X.; LIU, J. y ZHANG, B. (2016): Association study confirms two susceptibility loci for breast cancer in Chinese Han women. *Breast cancer research and treatment,* 159(3): 433-442.

ZHANG, J.; QIU, L.X.; WANG, Z.H.; LEAW, S.J.; WANG, B.Y.; WANG, J.L.; CAO, J.L. y HU, X.C. (2010): Current evidence on the relationship between three polymorphisms in the FGFR2 gene and breast cancer risk: a meta-analysis. *Breast Cancer Research and Treatment,* 124(2): 419-424.

8. ANEXOS

Tabla 1. ID de cada una de las variaciones, asociada a su ID PubMed y características de ellas……37
Tabla 2. Número de variaciones que hay para cada gen……47
Tabla 3. Número de variaciones para cada fenotipo y porcentaje con respecto del total……48

Tabla 1. ID de cada una de las variaciones, asociada a su ID PubMed y características de ellas.

ID dbSNP	Cromosoma	Gen	Significado clínico	Fenotipo	Población	ID PubMed	Bases de datos
rs10069690	5	TERT	Asociación	Cáncer de mama	Afroamericana	22037553	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs1011970	9	CDKN2B-AS1	Asociación	Cáncer de mama	Europea	25255808	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs1017226	5	MAP3K1	Factor de riesgo	Cáncer de mama temprano	China Han	27572337	Ensembl
rs10472076	5	RAB3C	Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
rs10474352	5	ARRDC3-AS1	Asociación	Cáncer de mama	Asiática oriental	25038754	Ensembl, GWAS catalog
rs10510102	10	FGFR2	Asociación	Cáncer de mama	Europea	21263130	Ensembl, GWAS catalog
rs10759243	9		Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs10816625	9		Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog
rs10822013	10	ZNF365	Asociación	Cáncer de mama	Asiática oriental	21908515	Ensembl, GWAS catalog
rs10941679	5		Asociación	Cáncer de mama	Británica	20453838	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs10995190	10	ZNF365	Asociación	Cáncer de mama	Europea	25255808	Ensembl
rs11075995	16	FTO	Asociación	Cáncer de mama (ER-)	Europea	23535733 27117709	Ensembl, GWAS catalog

							Ensembl, GWAS catalog
rs11082321	18		Asociación	Cáncer de mama	Asiática oriental	25038754	Ensembl, GWAS catalog
rs11155804	6	ESR1	Asociación	Cáncer de mama (ER-)	Europea	27117709	Ensembl, GWAS catalog
rs11200014	10	FGFR2	Asociación	Cáncer de mama	Todas	20554749	Ensembl
rs11249433	1	EMBP1	Asociación	Cáncer de mama	Europea	19330030	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs113701136	19	CCNE1	Asociación	Cáncer de mama (ER-)	Europea	29058716	Ensembl, GWAS catalog
rs114962751	2	FAM126B	Asociación	Cáncer de mama	Europea	27117709	Ensembl, GWAS catalog
rs11552449	1	DCLRE1B	Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs115635831	2	PPIL3	Asociación	Cáncer de mama (ER-)	Europea	27117709	Ensembl, GWAS catalog
rs11654964	17		Asociación	Cáncer de mama lobular (Interacción con terapia hormonal para menopausia)	Europea	24080446	Ensembl, GWAS catalog
rs11780156	8		Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs11814448	10		Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog

rs11977670	7		Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog
rs12048493	1	OTUD7B	Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog
rs1219648	10	FGFR2	Asociación	Cáncer de mama	Europea	20872241	Ensembl, GWAS catalog
rs12355688	10	ZMIZ1	Asociación	Cáncer de mama	Africana	22923054	Ensembl
rs12422552	12		Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs12493607	3	TGFBR2	Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs12628403	22		Asociación	Cáncer de mama	Asiática oriental	23411593	Ensembl
rs12655019	5		Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs12662670	6	ESR1	Asociación	Cáncer de mama (ER-)	Europea	27117709	Ensembl, GWAS catalog, ClinVar
rs12710696	2	LINC01376	Asociación	Cáncer de mama (ER-)	Europea	23535733	Ensembl, GWAS catalog
		LINC01376			Europea	29059683	Ensembl, GWAS catalog
rs12870942	13		Asociación	Cáncer de mama (ER-)	Europea	27117709	Ensembl, GWAS catalog
rs12922061	16	CASC16	Asociación	Cáncer de mama	Todas	20332101	Ensembl, GWAS catalog
rs12998806	2		Asociación	Cáncer de mama (ER+)	Afroamericana	21852243	Ensembl, GWAS catalog
rs13074711	3		Asociación	Cáncer de mama	Africana	22923054	Ensembl, GWAS catalog
rs13116936	4	TNIP3	Asociación	Cáncer de mama	Afroamericana	23468962	Ensembl, GWAS catalog

rs132390	22	EMID1	Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
rs13267382	8	LINC00536	Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog
rs13281615	8	PCAT1	Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog, SNPedia
					China	20699374	Ensembl, GWAS catalog, SNPedia
rs13294895	9		Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog
rs13329835	16	CDYL2	Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs13387042	2		Asociación	Cáncer de mama	Europea	17529974	Ensembl, GWAS catalog, SNPedia
					Europea	20453838	Ensembl, GWAS catalog, SNPedia
rs13393577	2	ERBB4	Asociación	Cáncer de mama	Asiática oriental	22452962	Ensembl, GWAS catalog
rs140068132	6	ESR1	Protectora	Protección frente al cáncer de mama	Americana	25327703	Ensembl, GWAS catalog
rs1432679	5	EBF1	Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs144848	13	BRCA2	Asociación	Cáncer de mama	Británica	11062481	Ensembl
					Británica	17341484	Ensembl

rs1562430	8	PCAT1	Asociación	Cáncer de mama	Británica	20453838	Ensembl
rs16857609	2	DIRC3	Asociación	Cáncer de mama	Europea	23535729	Ensembl
					Europea	29059683	GWAS catalog
rs16886034	5	MAP3K1	Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs16886113	5	MAP3K1	Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs16886165	5	MAP3K1	Asociación	Cáncer de mama	Europea	19330030	Ensembl
rs16886181	5	MAP3K1	Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs16886364	5	MAP3K1	Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs16886448	5	MAP3K1	Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs16992204	21		Asociación	Cáncer de mama	Asiática oriental	27354352	Ensembl, GWAS catalog
rs17350191	8		Asociación	Cáncer de mama (ER-)	Europea	29058716	Ensembl, GWAS catalog
rs17530068	6		Asociación	Cáncer de mama	Europea	22976474	Ensembl, GWAS catalog
rs1799950	17	BRCA1	Asociación	Cáncer de mama	Británica	17341484	Ensembl
rs1800054	11	ATM	Factor de riesgo	Susceptibilidad a cáncer de mama	Europea	16652348	Ensembl, ClinVar
rs2012709	5	SUB1	Asociación	Cáncer de mama	Europea	25751625	Ensembl, GWAS catalog
rs204247	6		Asociación	Cáncer de mama	Europea	23535729	Ensembl
					Europea	29059683	GWAS catalog
rs2048672	7		Asociación	Cáncer de mama	Asiática oriental	21908515	Ensembl, GWAS catalog
rs2193094	16	TOX3	Asociación	Cáncer de mama	Europea	25956309	Ensembl, GWAS catalog

rs2229882	5	MAP3K1	Asociación	Cáncer de mama temprano	China Han	27572337	Ensembl
					Europea	24493630	Ensembl, GWAS catalog
rs2242652	5	TERT	Asociación	Cáncer de mama (ER-)	Todas	23535731	Ensembl, GWAS catalog
					Europea	27117709	Ensembl, GWAS catalog
rs2284378	20	RALY	Asociación	Cáncer de mama	Europea	22976474	Ensembl, GWAS catalog
rs2290203	15	PRC1	Asociación	Cáncer de mama	Asiática oriental	25038754	Ensembl, GWAS catalog
rs2392780	8	PCAT1	Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs2420946	10	FGFR2	Asociación	Neoplasma de mama	Todas	21445572	Ensembl
rs2588809	14	RAD51B	Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
rs2912774	10	FGFR2	Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs2943559	8	HNF4G	Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs2981578	10	FGFR2	Asociación	Cáncer de mama	Todas	27966449	Ensembl, GWAS catalog
					Asiática oriental	24143190	Ensembl, GWAS catalog
rs2981579	10	FGFR2	Asociación	Cáncer de mama	Británica	20453838	Ensembl, GWAS catalog
					Europea	19330030	Ensembl, GWAS catalog
rs3757318	6	ESR1	Asociación	Cáncer de mama (TN)	Británica	20453838	Ensembl, GWAS catalog
					Europea	24325915	Ensembl, GWAS catalog

rs3757322	6	CCDC170	Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog
rs3760982	19		Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs3803662	16	CASC16	Asociación	Cáncer de mama	Europea	19330030	Ensembl, GWAS catalog
rs3817198	11	LSP1	Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog, SNPedia
					Todas	21445572	Ensembl, GWAS catalog, SNPedia
rs3822625	5	MAP3K1	Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs4245739	1	MDM4	Asociación	Cáncer de mama (TN)	Europea	24325915	Ensembl, GWAS catalog
					Europea	27117709	Ensembl, GWAS catalog
rs4322600	14		Asociación	Cáncer de mama	Africana	22923054	Ensembl, GWAS catalog
rs4415084	5		Asociación	Cáncer de mama	Europea	20872241	Ensembl, GWAS catalog
rs4784223	16	TOX3	Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs4784227	16	CASC16	Asociación	Cáncer de mama (ER-)	Europea	27117709	Ensembl, GWAS catalog
					Asiática oriental	20585626	Ensembl, GWAS catalog
rs4951011	1	ZC3H11A	Asociación	Cáncer de mama	Asiática oriental	25038754	Ensembl, GWAS catalog
rs4973768	3	SLC4A7	Asociación	Cáncer de mama	Británica	19330027	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog

rs4980383	11	LSP1	Asociación	Cáncer de mama (ER-)	Europea	27117709	Ensembl, GWAS catalog
rs537626	11		Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs554219	11		Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog
rs56069439	19	ANKLE1	Asociación	Cáncer de mama (ER-)	Europea	27117709	Ensembl, GWAS catalog
rs6001930	22		Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
rs614367	11		Asociación	Cáncer de mama	Británica	20453838	Ensembl
rs62355902	5		Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog
rs6472903	8	CASC9	Asociación	Cáncer de mama	Asiática oriental	23535825	Ensembl, GWAS catalog
rs6504950	17	STXBP4	Asociación	Cáncer de mama (ER+)	Europea	22972951	Ensembl, GWAS catalog
rs653465	3	NEK10	Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs6585202	10	TCF7L2	Asociación	Cáncer de mama	Europea	27117709	Ensembl, GWAS catalog
rs6678914	1	LGR6	Asociación	Cáncer de mama	Europea	23535733	Ensembl, GWAS catalog
rs66823261	8	RPL23AP53	Asociación	Cáncer de mama (ER-)	Europea	29058716	Ensembl, GWAS catalog
rs67397200	19		Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog
rs6762644	3	ITPR1	Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs6788895	3	SIAH2	Asociación	Cáncer de mama	Asiática oriental	22951594	Ensembl, GWAS catalog
rs6828523	4	ADAM29	Asociación	Cáncer de mama	Todas	25227710	Ensembl, GWAS catalog

rs6964587	7	AKAP9	Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog
rs704010	10	ZMIZ1	Asociación	Cáncer de mama	Británica	20453838	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs7072776	10		Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs7107217	11		Asociación	Cáncer de mama	Asiática oriental	22383897	Ensembl, GWAS catalog
rs7192724	16	PLCG2	Asociación	Cáncer de mama (Interacción con terapia hormonal para menopausia)	Europea	24080446	Ensembl, GWAS catalog
rs72755295	1	EXO1	Asociación	Cáncer de mama	Europea	25751625	Ensembl, GWAS catalog
rs75915166	11		Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog
rs7716600	5		Asociación	Cáncer de mama	Europea	20872241	Ensembl, GWAS catalog
rs7726159	5	TERT	Asociación	Cáncer de mama (ER-)	Europea	27117709	Ensembl, GWAS catalog
rs7726354	5	MIER3	Asociación	Cáncer de mama temprano	Europea	24493630	Ensembl, GWAS catalog
rs7904519	10	TCF7L2	Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs909116	11	TNNT3	Asociación	Cáncer de mama	Británica	20453838	Ensembl, GWAS catalog
rs9257408	6		Asociación	Cáncer de mama	Europea	25751625	Ensembl, GWAS catalog

rs9383938	6	ESR1	Asociación	Cáncer de mama	Europea	24895409	Ensembl, GWAS catalog
rs9397437	6		Asociación	Cáncer de mama	Europea	29059683	Ensembl, GWAS catalog
rs941764	14	CCDC88C	Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs9693444	8		Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
					Europea	29059683	Ensembl, GWAS catalog
rs9790517	4	TET2	Asociación	Cáncer de mama	Europea	23535729	Ensembl, GWAS catalog
rs999737	14		Asociación	Cáncer de mama (ER+)	Europea	22972951	Ensembl, GWAS catalog

ER-: Receptor de estrógeno negativo. ER+: Receptor de estrógeno positivo. TN: Triple negativo.

Tabla 2. Número de variaciones que hay para cada gen.

Gen	Nº variaciones	Gen	Nº variaciones	Gen	Nº variaciones
MAP3K1	9	CASC9	1	MIER3	1
FGFR2	7	CCDC170	1	NEK10	1
ESR1	5	CCDC88C	1	OTUD7B	1
CASC16	3	CCNE1	1	PLCG2	1
PCAT1	3	CDKN2B	1	PPIL3	1
TERT	3	CDYL2	1	PRC1	1
AS1	2	DCLRE1B	1	RAB3C	1
LINC01376	2	DIRC3	1	RAD51B	1
LSP1	2	EBF1	1	RALY	1
TCF7L2	2	EMBP1	1	RPL23AP53	1
TOX3	2	EMID1	1	SIAH2	1
ZMIZ1	2	ERBB4	1	SLC4A7	1
ZNF365	2	EXO1	1	STXBP4	1
ADAM29	1	FAM126B	1	SUB1	1
AKAP9	1	FTO	1	TET2	1
ANKLE1	1	HNF4G	1	TGFBR2	1
ARRDC3	1	ITPR1	1	TNIP3	1
ATM	1	LGR6	1	TNNT3	1
BRCA1	1	LINC00536	1	ZC3H11A	1
BRCA2	1	MDM4	1	-	37

- : Variaciones que no se encuentran asociadas a ningún gen.

Tabla 3. Número de variaciones para cada fenotipo y porcentaje con respecto del total.

Fenotipo	Nº variaciones	% del total
Cáncer de mama	84	68,29
Cáncer de mama temprano	15	12,20
Cáncer de mama (ER-)	14	11,38
Cáncer de mama (ER+)	3	2,44
Cáncer de mama (TN)	2	1,63
Protección frente al cáncer de mama	1	0,81
Susceptibilidad a cáncer de mama	1	0,81
Neoplasma de mama	1	0,81
Cáncer de mama lobular (Interacción con terapia hormonal para menopausia)	1	0,81
Cáncer de mama (Interacción con terapia hormonal para menopausia)	1	0,81

ER-: Receptor de estrógeno negativo. ER+: Receptor de estrógeno positivo. TN: Triple negativo.