Universitat Oberta de Catalunya

uoc.edu

Búsqueda de genes involucrados en el grado histológico de tumores de mama

Sandra Garrido Romero
Máster Universitario en Ciencia De Datos
Área Medicina

Ana Belén Nieto Librero
Jordi Casas Roma

06/06/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Búsqueda de genes involucrados en el grado histológico de tumores de mama*
Nombre del autor:	*Sandra Garrido Romero*
Nombre del consultor/a:	*Ana Belén Nieto Librero*
Nombre del PRA:	*Jordi Casas Roma*
Fecha de entrega (mm/aaaa):	06/2021
Titulación:	*Máster Universitario en Ciencia De Datos*
Área del Trabajo Final:	*Trabajo Final De Máster*
Idioma del trabajo:	*Español*
Palabras clave	*Ciencia de datos, cáncer de mama, genética*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

En las últimas décadas, las líneas de investigación siguen una tendencia hacia la obtención de terapias personalizadas para el tratamiento de cada paciente realizando un análisis genético. La finalidad de este trabajo es analizar un conjunto de datos génicos clasificados por grado histológico que permita obtener los genes responsables de cada grado. Para ello, se utilizarán bases de datos públicas que contienen la información genética de diversos pacientes. En este trabajo, en concreto, se analiza un conjunto abierto de datos génicos de diversas mujeres que han padecido cáncer de mama.

La metodología utilizada consiste en el entrenamiento de diversos modelos de distintos algoritmos de clasificación para poder obtener los genes involucrados en el desarrollo de tumores de mama. El principal objetivo es seleccionar el mejor modelo para poder ofrecer un tratamiento totalmente personalizado a cada paciente acorde a su genética. Además, obteniendo los genes involucrados en el desarrollo del tumor, se podrían llegar a predecir tumores de mama en descendientes de personas que ya lo hayan desarrollado, antes de que apareciesen, simplemente por la presencia de los genes responsables del cáncer obtenido en los algoritmos estudiados.

Abstract (in English, 250 words or less):

In recent years, the lines of research have followed a trend towards obtaining personalized therapies for the treatment of each patient with a genetic analysis. The purpose of this work is to analyse a set of genetic data classified by histological grade. Public databases containing the genetic information of various women who have suffered from breast cancer will be used in this work.

The methodology used consists of training various models of different classification algorithms in order to obtain the genes involved in the growth of breast tumour. The main objective is to select the model with the best results. This can offer a totally personalized treatment to each patient according to their genetics. Furthermore, if I ii can achieve the genes involved in tumour development, breast tumours could be predicted in descendants of people who have already developed it, before the cancer appears, simply by the presence of the genes responsible for cancer obtained in the algorithms studied.

Índice

Lista de figuras

Lista de tablas

Lista de ecuaciones

1. Introducción

En este apartado se plantean los objetivos que se pretenden conseguir en la realización de este proyecto junto con la metodología y la planificación que se pretende seguir.

1.1 Contexto y justificación del Trabajo

El tema escogido es bastante relevante a día de hoy ya que el cáncer de mama es el que más aparece en mujeres y es uno de los más mortales a nivel global. Este tipo de tumores surge por diversas razones y una de ellas es por herencia genética. Además, en los últimos años se ha intentado crear terapias personalizadas para el tratamiento de cada persona, lo que hace que sea muy relevante el estudio del genoma de pacientes que han padecido la enfermedad.

De momento, no se encuentra muy extendida la utilización de metodologías de la Ciencia de Datos para el estudio de los genes. Esa es la necesidad que se quiere cubrir con el desarrollo de este trabajo. El principal objetivo es determinar que genes se ven involucrados en el desarrollo de los tumores de mama y su posible relación con el grado histológico que presenta el tumor.

Mi motivación personal para la elección de esta línea de investigación ha sido por un caso de cáncer de mama que me ha tocado bastante cerca a nivel familiar. Mi madre sufrió esta enfermedad hace unos años y actualmente se encuentra recuperada, lo que no quiere decir que no pueda volver a padecer la enfermedad. Como ya he mencionado anteriormente, el cáncer de mamá es hereditario por lo que puede que en algún momento del futuro yo también deba soportarlo. De ahí mi gran interés en poder aportar un granito de arena con este trabajo a la comunidad científica.

1.2 Objetivos del Trabajo

En este se pretende realizar un estudio de los genes involucrados en el grado histológico del cáncer de mama, utilizando herramientas estadísticas y algoritmos de machine learning dentro de la metodología de la Ciencia de Datos.

La idea principal es poder encontrar los factores genéticos responsables del grado desarrollado por el tumor y, que este estudio aporte valor a las investigaciones llevadas a cabo por diversos profesionales de la medicina, que se encuentran actualmente en la literatura publicada.

A continuación, se detallan los objetivos principales y secundarios de este estudio.

1.2.1 Objetivos principales

Los objetivos principales de este trabajo son los siguientes:

- Análisis de bases de datos disponibles con información genética e histológica de pacientes de cáncer de mama.

- Análisis comparativo de diversos modelos basados en distintas técnicas sobre la clasificación de los genes.

- Seleccionar los genes responsables de las diferencias en el grado histológico de los tumores de mama gracias a los modelos implementados.

1.2.2 Objetivos secundarios

Los siguientes objetivos se derivan de los resultados obtenidos en el apartado anterior:

- Mostrar la utilidad que puede tener la Ciencia de Datos en estudios relacionados con medicina para poder ampliar el conocimiento sobre distintas enfermedades.

- Aprovechar los resultados obtenidos para predecir posibles casos de tumores en personas que cuenten con los genes responsables del desarrollo del cáncer.

- Utilizar distintas herramientas de ciencia de datos que puedan enriquecer el análisis realizado.

- Conseguir que la identificación de estos genes ayude a la creación de tratamientos personalizados para pacientes.

1.3 Enfoque y método seguido

Existen dos estrategias distintas, desarrollar una investigación totalmente nueva desde cero o, intentar partir de algo ya creado.

La idea de este trabajo es la creación de modelos que sirvan para clasificar los datos existentes y decidir cuál es el modelo que mejor resultados obtiene. Esta estrategia es la que mejor se adapta a los objetivos expuestos porque permite la total libertad de análisis con los datos, y poder centrarse con los datos que se estimen oportunos. Si se partiese de una investigación que ya haya sido desarrollada, habría que adaptarse a ella.

Aun así, las dos estrategias no son excluyentes ya que cuando se realice el estudio del estado del arte, es probable que se encuentre algún algoritmo ya implementado que pueda servir como base a este proyecto. O, algún estudio que ayude a complementar el desarrollo de este trabajo.

El método a seguir será la búsqueda de información sobre el tumor de mamá y los posibles genes relacionados dentro de repositorios especializados. También, se buscará posible código abierto o investigaciones ya realizadas sobre el tema.

El método de trabajo seguirá una metodología en cascada con posibles modificaciones. Se ha decidido este modelo ya que el diseño de la solución no comenzará hasta que no se haya recabado información, y la implementación no comenzará hasta que no se haya realizado el diseño. Pero, siempre se van a poder modificar decisiones tomadas en fases anteriores por lo que no será un modelo en cascada totalmente rígido.

1.4 Planificación del Trabajo

El principal recurso necesario para la realización del trabajo es la base de datos que contienen la información genética con la que se va a trabajar. En este

caso, se contará con datos de tumores provenientes de 129 pacientes. Además, se necesitarán los recursos informáticos básicos para realizar el análisis de los datos.

Las tareas que realizar son:

- Definición del trabajo final → Concretar la línea de investigación que se va a lleva a cabo para la realización del trabajo.

- Planificación del trabajo final → Listar las taras a realizar durante la investigación y hacer una estimación del tiempo llevado por cada una.

- Redacción PEC1 → Redactar la primera entrega del TFM.

- Búsqueda de bibliografía → Buscar información relacionada con posibles estudios anteriores tales como publicaciones, artículos o libros.

- Análisis de las tecnologías y el código disponible → Analizar las tecnologías disponibles para el desarrollo de los algoritmos. Buscar posibles librerías o código ya existente relacionado con el tema.

- Lectura de bibliografía → Leer la bibliografía y documentación encontrada.

- Revisar la definición del trabajo final → Redefinir el trabajo final si fuera necesario tras la búsqueda de información.

- Redacción PEC2 → Redactar la segunda entrega del TFM.

- Estudio de la base de datos → Estudiar la base de datos que va a ser utilizada para saber con qué tipo de datos se cuenta en la investigación.

- Preparación de los datos → Tratar los datos para el estudio posterior.

- Diseño del algoritmo → Diseñar los modelos y algoritmos que se utilizarán en el trabajo.

- Implementación del algoritmo → Implementar los modelos diseñados anteriormente.

- Realización de pruebas → Realizar pruebas a los modelos implementados para encontrar posibles errores.

- Documentación (PEC3) → Documentar el diseño y la implementación realizada.

- Realización de experimentos → Poner a prueba el código desarrollado y realizar múltiples experimentos para obtener métricas de cada modelo.

- Conclusiones → Pensar las posibles conclusiones a las que ha llevado la realización del trabajo.

- Redacción de la memoria (PEC4) → Redactar la memoria final del TFM.

- PEC5 - Presentación y defensa del proyecto → Preparar la presentación y defensa del trabajo final.

- Defensa pública → Preparar la defensa pública que se llevará a cabo tras la defensa del proyecto.

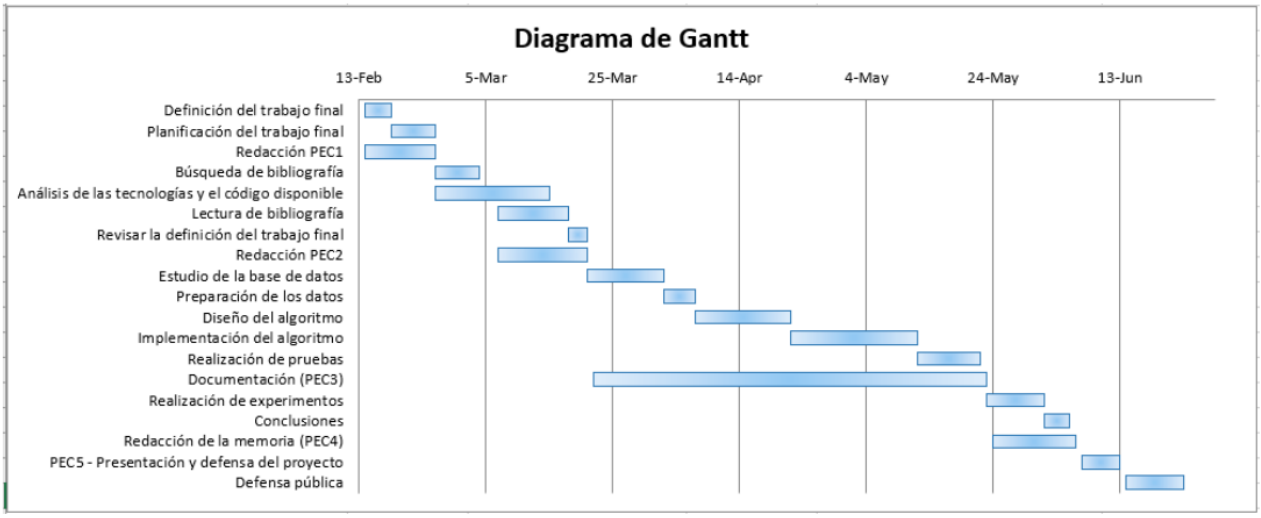


Ilustración 1: Diagrama de Gantt

1.5 Breve sumario de productos obtenidos

El principal producto obtenido es el análisis de los genes de diversas muestras de pacientes con tumores de mama. Con estos análisis se han obtenido conjuntos de genes implicados en el desarrollo de ciertos aspectos fenotípicos dentro de un tumor.

1.6 Breve descripción de los otros capítulos de la memoria

Esta memoria está dividida en cinco capítulos:

- **Capítulo 1**: Se expone una introducción al proyecto junto con los principales objetivos y la planificación creada para el desarrollo del proyecto**.**

- **Capítulo 2**: Se resumen las últimas investigaciones y el estado actual de los temas que se van a tratar. Además, se determinan las preguntas que se pretenden responder con el presente trabajo.

- **Capítulo 3**: Se especifica paso por paso el proceso llevado a cabo para la implementación del proyecto. Además de definir y explicar los modelos de clasificación de genes utilizados.

- **Capítulo 4**: Se presentan los resultados obtenidos del proyecto. También se analiza si se han respondido a las preguntas planteadas en el segundo capítulo.

- **Capítulo 5**: Se analizan las conclusiones obtenidas y se plantean las posibles líneas de futuro que puede seguir el proyecto.

- **Capítulo 6**: Se detallan los recursos bibliográficos utilizados como apoyo para el desarrollo del trabajo.

2. Estado del Arte

En el presente documento, se abordará el contexto en el que se desarrollará el proyecto. Primeramente, se tratará el cáncer de mamá en términos numéricos, ofreciendo datos sobre esta enfermedad alrededor del mundo y los avances médicos que se han realizado al respecto. También se mencionarán los principales factores hereditarios de riesgo para esta enfermedad.

A continuación, se presentarán el concepto de microarray genético y la información que contiene para poder contextualizar las herramientas y métodos utilizados durante el proyecto.

Para terminar, se resumirán las principales aportaciones que se han realizado durante los últimos años respecto a los genes implicados en el desarrollo de los distintos tipos de cáncer de mama. Se comenzará con aportaciones efectuadas a finales de la década de los 90 hasta la actualidad dado que se ha encontrado información relevante en cada uno de los estudios.

2.1 Cáncer de mama

El cáncer de mama es uno de los más frecuentes en todo el mundo junto con el cáncer de pulmón, colon o piel.

Este tipo de tumores cada vez es más frecuente entre la población femenina de los países desarrollados, en 2018 se diagnosticaron 2.088.849 nuevos casos de cáncer de mama en todo el mundo [1]. Sólo en España, hubo 33.307 nuevos casos durante el año 2019, representando un 30% de los tumores padecidos por mujeres en el país.

Aunque este tipo de tumores son la causa más frecuente de muerte por cáncer en algunos países, la tasa de mortalidad es relativamente baja y continúa disminuyendo gracias a la mejora en los tratamientos terapéuticos.

Las principales herramientas de diagnóstico existen hoy en día son exámenes clínicos, mamografías o ecografías y biopsia con biomarcadores [2]. Tras la realización de algunas de estas pruebas, se lleva a cabo la toma de decisión sobre el tratamiento a seguir con cada paciente dependiendo de diversos factores como edad, subtipo del tumor o factores de riesgo existentes. Algunos de los posibles tratamientos son cirugía, radioterapia o quimioterapia, sin ser excluyentes entre ellos. La supervivencia es mayor en pacientes con tratamientos más dirigidos o particularizados a su caso y subtipo de cáncer.

Uno de los mayores problemas, o beneficios, de este tipo de tumores es que el factor hereditario es clave. Una de cada siete mujeres diagnosticadas con cáncer de mama cuenta con un familiar cercano (madre, hermana o hija) que ha sido diagnosticada también [2]. Esto es debido a que la presencia de ciertos genes incrementa el riesgo de padecer la enfermedad y se podrían utilizar para predecir casos de tumores antes de aparecer, y llevar un seguimiento de los posibles futuros pacientes para descubrir la enfermedad en un estado inicial.

La presencia de unos genes u otros también es importante a la hora de decidir el tratamiento que se va a administrar a un paciente. Algunos de los genes implicados afectan a la mortalidad o supervivencia de la persona y el tratamiento que se vaya a seleccionar debe de tener en cuenta este aspecto. Los últimos

estudios realizados sobre este tema pretenden personalizar los tratamientos de cáncer a partir de la realización de un estudio genómico del paciente y de las células cancerosas.

2.2 Microarrays genéticos

"La tecnología de microarrays es una tecnología en desarrollo para estudiar la expresión de muchos genes a la vez" [3]. Explicado de otro modo, consiste en obtener multitud de secuencias de genes y situarlos en un portaobjetos denominado chip [3]. Por otro lado, se obtienen muestras de ADN o ARN que se ponen en contacto con el chip anterior. Al unirse las dos muestras, las áreas del chip que contienen genes que aparecen en la muestra de ADN o ARN se iluminan.

En la siguiente imagen se puede ver el proceso descrito:

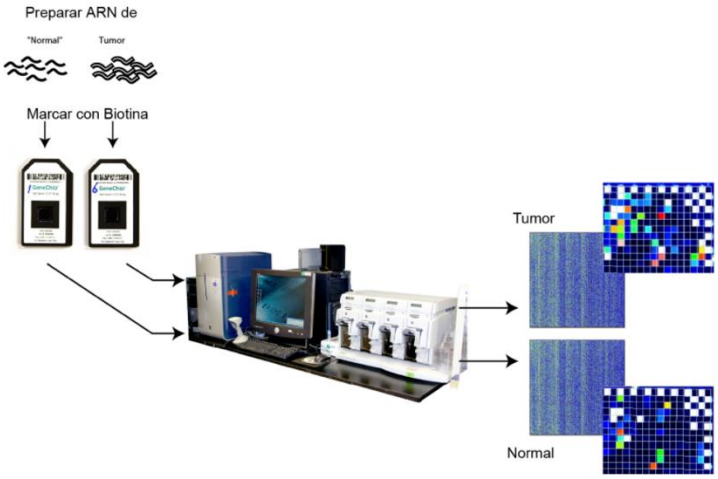


Ilustración 2:Tecnología de microarrays. [3]

Tanto en el presente trabajo como en las investigaciones realizadas previamente y analizadas para contextualizar este desarrollo, se utilizan bases de datos de microarrays que contienen genes de pacientes con tumores de mamas.

2.3 Investigaciones previas

Las principales líneas de investigación respecto a los genes implicados en el desarrollo de tumores de mama empezaron a aparecer a finales de los años 90 y siguen desarrollándose en la actualidad.

Todas estas líneas parten del empleo de microarrays con información genética. Esta información varía en función de cada artículo que ha sido consultado. El nexo de los datos común a todos ellos son los datos del genoma de pacientes con tumores de mama y pacientes sanos que actúan como registros de control. Los datos particulares a cada investigación varían bastante. Algunos de ellos añaden datos de los mismos pacientes, una vez que se encuentran libres del cáncer. Otros, utilizan datos que de las propiedades intrínsecas de los tumores para poder obtener una correlación. Incluso, en algún caso, la información utilizada corresponde a personas que presentan algún factor de riesgo.

Pese a las diferencias en los datos de partida, la mayoría comparte el método de clasificación utilizado: algoritmos de *clustering*. Este tipo de algoritmos,

denominados de agrupación, consiste en dividir el conjunto de datos en diversos subgrupos en función de la similitud que presentan entre, en este caso, los genes utilizados. Es importante determinar qué tipo de métrica se va a utilizar para definir si dos datos se encuentran dentro de un mismo grupo o no. En los diversos estudios que se han consultado, se parte de cinco centroides (dato que caracteriza a un grupo) que representan los cinco subtipos de cáncer de mama existentes. Estos centroides no son escogidos al azar, se obtienen de los datos pertenecientes a "*Stanford Genomics Breast Cancer Consortium*" [4] y sus estudios realizados [5].

Los tumores de mama se pueden clasificar en subtipos distintos según las diferencias en los patrones de los genes de cada individuo. En uno de los artículos más antiguos consultados [6] se menciona el concepto "*Molecular portraits*" que refleja las similitudes y diferencias que aparecen en los distintos tumores y su posible interpretación biológica como la tasa de crecimiento o la composición de la célula cancerosa. Además, este estudio obtiene 8 *clusters* o agrupaciones en función de la variación del tipo de células cancerosas como células adiposas o endoteliales.

En un artículo del mismo año [7], se encuentra que la proteína ERB2 se encuentra en niveles relativamente altos en el 20-30% de los tumores de mama. Concretamente, existe una correlación directa entre esta proteína y el cáncer ductal in situ. Este tipo de cáncer se da cuando las células cancerosas cubren los conductos por donde circula la leche, pero no se propagan al resto del tejido mamario. Algunas de las proteínas de la misma familia, como la ERB1, también están relacionadas con este tipo de tumores. Es decir, hace veinte años se empezaron a encontrar posibles genes implicados en los tumores de mama.

En uno de los estudios consultados [8], se buscaba la correlación de la expresión de los genes y los parámetros clínicos relevantes. Se parte de que los tumores se pueden clasificar en tres grandes grupos: epitelial basal, tumores con ERBB2 sobre expresado y los tumores normales. Y los principales parámetros clínicos relacionados con el pronóstico son: la metástasis de los nodos linfáticos, el grado histológico, los receptores del factor de crecimiento, los protooncogenes y las mutaciones en el gen TP53. Después de realizar la clasificación, se realiza un análisis estadístico mediante el método analítico SAM que asigna una puntuación para cada gen que mide la correlación de ese gen con la supervivencia del paciente. Los resultados obtenidos muestran la implicación de diversos genes en función de cada subtipo de tumor. Algunos de los genes implicados son ERBB2, GRB7, genes epiteliales basales o el TP53.

A partir de este momento, el interés en los genes implicados en los tumores de mama va en aumento y la aparición de nuevas herramientas para la investigación ayuda a que se encuentren más evidencias de ello. En uno de los estudios [9] se estudia la mutación de la proteína GATA3 responsable de regular la respuesta de las células inmunes. La presencia de GATA3 está correlacionada con la presencia de algunos genes importantes en el desarrollo de las células cancerosas epiteliales en los tumores de mama.

Además, el nivel de presencia de GATA3 influye en el pronóstico de supervivencia de los pacientes.

Por último, los estudios más recientes estudian los genes de personas que cuentan con algún tipo de factor de riesgo hereditario y la probabilidad de desarrollar un nuevo tumor [10]. En este caso se encuentra que los portadores de genes BRCA con mutaciones obtienen unos resultados peores y los tumores desarrollados

suelen afectar a los ganglios linfáticos. Por otro lado, asegura que la supervivencia aumenta en los pacientes que no presentan este tipo de genes.

El enfoque de este trabajo se realizará a partir de los estudios analizados. Se ha podido afirmar que la presencia de algunos genes favorece el desarrollo de ciertos tipos de tumores de mama. Y en algunos casos, incluso pueden mejorar o empeorar el pronóstico de supervivencia. Es interesante el estudio génico para realizar una personalización de los tratamientos en función de la información genética del paciente y de las células cancerosas presentes en él.

2.4 Preguntas a responder

Tras el análisis realizado se pretende, mediante técnicas de análisis y clasificación de datos, responder a interrogantes que aparecen respecto a los genes implicados y posibles mejoras a realizar en los algoritmos utilizados.

Algunas de estas preguntas son:

¿Se pueden segmentar los tipos de tumores en función de la presencia de ciertos genes?

Las personas que presentan mutaciones en los genes relacionados con el cáncer, ¿tienen una alta probabilidad de desarrollar la enfermedad?

¿Se puede crear un tratamiento único para cada persona, que mejore su probabilidad de supervivencia?

¿Es posible predecir el desarrollo de un futuro tumor realizando un estudio genético?

¿Se puede determinar un listado de genes implicados en el desarrollo de tumores de mama?

3. Proceso de implementación

El proceso de implementación que se ha seguido ha sido definido por las fases de cualquier proyecto de ciencia de datos. Según la literatura tradicional la gestión de datos consta de las siguientes fases que se pueden solapar entre ellas [11]:

- Captura. Su objetivo es realizar la recopilación de los datos mediante diversas técnicas.
- Almacenamiento. Se determina el formato y la tipología en la que se van a guardar los datos.
- Preprocesado. Por norma general, los datos necesitan un proceso de preparación antes de iniciar el análisis.
- Análisis. Se crean varios modelos que intentan responder a las preguntas planteadas al inicio del proyecto.
- Visualización. Se componen recursos de transmisión visual para trasmitir los resultados del análisis.
- Publicación. El proceso puede finalizar con la publicación de los resultados de manera que puedan ser utilizados por otras personas ajenas al proyecto.

El objetivo principal de la aplicación del ciclo de vidas de los datos en el proyecto es la extracción de conocimiento a partir de una base de datos genética de diversos pacientes.

En la siguiente figura se puede ver la pirámide DIKW donde se vislumbra la extracción de conocimiento mencionada. Partiendo de los datos se extrae cierta información de ellos. A partir de esta información se puede extraer conocimiento y este conocimiento al transmitirse se entiende como sabiduría.

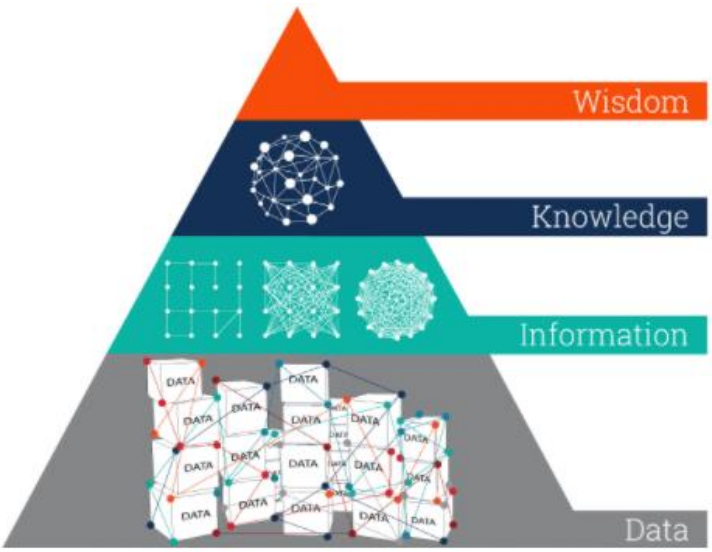


Ilustración 3: Pirámide DIKW [12]

Este proceso se ha llevado a cabo utilizando el lenguaje de programación *R* integrado en el *software RStudio.* La principal librería que se ha utilizado ha sido *Bioconductor* que esta creada especialmente para el tratamiento de datos genéticos.

3.1 Recogida de datos y almacenamiento

Los datos han sido obtenidos del sitio web del *National Center for Biotechnology Information* de Estados Unidos, que contiene múltiples conjuntos de datos genómicos.

El *dataset* escogido es GSE5460 denominado *Predicting Features of Breast Cancer with Gene Expression Patterns*. Está formado por los datos a nivel de sonda de 54675 genes de 129 muestras de pacientes que presentan distintos tipos de cáncer de mama. [13]

La recogida de datos se ha realizado a través del software *R* y la librería *GEOquery* almacenando el resultado en una variable de tipo *ExpressionSet*. Este tipo de variable contiene información de los datos en múltiples formatos. Para este trabajo, resulta interesante centrarse en la matriz de expresión y en los metadatos o variables fenotípicas.

La matriz de expresión contiene la abundancia de cada gen observado en cada una de las muestras. Cada una de las columnas corresponde a una muestra mientras que cada una de las filas corresponde a un gen. Cada una de las filas es denominada *Expression profile* al contener los valores observados de un mismo gen.

Por otro lado, los metadatos son la información de las variables que describen a las muestras. En este caso, los metadatos están compuestos por las siguientes 48 variables:

- *Title*
- *Geo_accession*
- *Status*
- *Submission_date*
- *Last_update_date*
- *Type*
- *Channel_count*
- *Source_name_ch1*
- *Organism_ch1*
- *Characteristics_ch1*
- *Characteristics_ch1.1*
- *Characteristics_ch1.2*
- *Characteristics_ch1.3*
- *Characteristics_ch1.4*
- *Characteristics_ch1.5*
- *Characteristics_ch1.6*
- *Treatment_protocol_ch1*
- *Molecule_ch1*
- *Extract_protocol_ch1*
- *Label_ch1*
- *Label_protocol_ch1*

- *Taxid_ch1*
- *Hyb_protocol*
- *Scan_protocol*
- *Description*
- *Data_processing*
- *Platform_id*
- *Contact_name*
- *Contact_email*
- *Contact_laboratory*
- *Contact_department*
- *Contact_institute*
- *Contact_address*
- *Contact_city*
- *Contact_state*
- *Contact_zip/postal_code*
- *Contact_country*
- *Supplementary_file*
- *Data_row_count*
- *Relation*
- *Relation.1*
- *B-R grade:ch1*
- *ER:ch1*
- *HER2:ch1*
- *LVI:ch1*
- *Node status:ch1*
- *Tumor size:ch1*
- *Tumor type:ch1*

3.2 Limpieza, preprocesado y preparación de datos

Las muestras tomadas cuentan con un gran número de variables fenotípicas que no son relevantes para este estudio. El primer paso va a ser seleccionar qué variables tienen importancia o se pretenden estudiar.

En segundo lugar, es necesario homogeneizar los datos para poder realizar un estudio que los compare. Esta homogeneización se lleva a cabo mediante un proceso de corrección de fondo, normalización y asociación de todos los valores de un gen en un solo valor.

Por último, es necesario llevar a cabo alguna técnica que permita evaluar los microarrays de cada muestra para determinar si hay alguna muestra que no sea relevante para el experimento y pueda ser eliminada del conjunto de datos.

3.2.1 Selección de variables

En primer lugar, se han seleccionado las variables fenotípicas interesantes para el estudio [14]:

- *B-R grade*: El grado es asignado tras la obtención de una muestra y su posterior análisis en laboratorio. El grado depende del nivel de división que tengan las células cancerosas y de su similitud con el resto de las células. Existen tres grados distintos numerados del uno al tres.

- *ER*: Proteína receptora de estrógenos que facilita el crecimiento de las células cancerosas.
- *HER2*: Proteína que promueve el crecimiento de las células cancerosas, también llamada biomarcador en algunos casos. Tiene dos valores posibles en función de la presencia de esta proteína en las células cancerosas: positivo o negativo.
- *LVI*: Esta variable hace referencia la invasión linfovascular que se define como la presencia de células tumorales dentro de los vasos linfáticos o los vasos sanguíneos. Está asociada a un mayor riesgo y peor pronóstico de la enfermedad. Cuenta con dos valores posibles: positivo o negativo.
- *Node status*: Esta variable determina si el tumor se ha propagado a los nodos linfáticos o no. Tiene dos valores posibles: positivo o negativo.
- *Tumor size*: Determina el tamaño del tumor y está relacionado con el pronóstico de la enfermedad. Un mayor tamaño suele estar relacionado con una menor oportunidad de supervivencia.
- *Tumor type*: Hay distintos tipos de tumores de mama en función de si son invasivos o no-invasivos, la localización o como se ven en el microscopio. En este conjunto de datos se manejan tumores de tipo ductal, lobular y una mezcla de ambos.

El resto de las variables no se van a utilizar en el estudio porque no hace referencia a características del cáncer de mama si no, a variables relacionadas con la recogida de datos como los datos del laboratorio responsable.

3.2.2 Método MAS5

Las muestras con las que se trabaja han sido tomadas mediante el escáner *Affymetrix GeneChip* por lo que se utiliza el método MAS5, proporcionado por este *software*, para poder corregir cada chip. Las muestras tomadas se encuentran separadas en ficheros CEL donde se guarda el valor de expresión de cada gen en forma de microarray [15].

El método MAS5 se divide en dos pasos. En primer lugar, realiza una corrección de fondo, donde se deciden unos valores para restarlos a cada uno de los datos originales de manera que el ruido del conjunto de datos disminuye notablemente. Después, se pasa a calcular el valor de expresión de cada conjunto de genes. Todo este proceso se encuentra incluido dentro del software facilitado por *Affymetrix* dentro de su librería para *Rstudio*.

3.2.3 Filtrado de genes

El filtrado de genes es un paso muy importante en este tipo de estudios porque encontrar una posible relación entre la abundancia observada de un gen y un valor concreto de una variable fenotípica aumenta de dificultad al aumentar el número de genes.

En este trabajo, se ha utilizado el denominado filtrado independiente o selección no especifica. La principal peculiaridad de este filtrado es que no utiliza la información de los valores de las muestras para un valor concreto de una variable fenotípica.

La premisa para que un gen sea eliminado es que su nivel de expresión sea alto y que la variabilidad con la que se expresa en las distintas muestras es alta. Para conseguir esta selección, se utiliza la función *nsFilter* que filtra las muestras midiendo la variabilidad de ellas. En este caso, se realizan dos filtrados. En el primero se realiza el filtrado midiendo la variabilidad con el rango intercuartílico del conjunto de datos. En el segundo se utiliza la desviación típica para medir esta variabilidad. Por último, se combinan los dos filtrados generando un nuevo conjunto de datos. Este nuevo conjunto de datos es el utilizado a partir de este momento en todo el desarrollo del trabajo [15].

3.2.4 Análisis de componentes principales

Otro posible método de reducción de dimensionalidad es el análisis de componentes principales o PCA [16]. Este método permite obtener un conjunto de datos con una dimensionalidad menor al original minimizando el error cuadrático sin perder la información representada. La manera de reducir la dimensionalidad que se utiliza es escoger un nuevo sistema de coordenadas en función de las varianzas de los datos. Estas varianzas se van recogiendo en distintos ejes denominados componentes principales. A la hora de reducir la dimensionalidad, se seleccionan los primeros componentes que expliquen la mayor parte de la varianza de los datos de manera que se determinen los atributos más importantes.

El software R cuenta un método que realiza este tipo de análisis centrando y tipificando los datos [17]. Se efectúa el método utilizando las muestras como observaciones y las expresiones de los genes como variables puesto que existe mucha más dimensionalidad en los datos de las expresiones de los genes que en los datos de las muestras. Tras el PCA, se obtienen como resultado 129 componentes principales de las cuales las 7 primeras explican el 90% de la variación total. En la imagen se muestra la información de las primeras componentes obtenidas.

```
## Importance of components:
##                             PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     10.3999 1.71346 1.21635 1.18928 1.01394 0.98792 0.93344
## Proportion of Variance  0.8384 0.02276 0.01147 0.01096 0.00797 0.00757 0.00675
## Cumulative Proportion   0.8384 0.86120 0.87267 0.88363 0.89160 0.89917 0.90592
##                             PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.85799 0.73500 0.67422 0.61056 0.60603 0.59513 0.56853
## Proportion of Variance 0.00571 0.00419 0.00352 0.00289 0.00285 0.00275 0.00251
## Cumulative Proportion  0.91163 0.91582 0.91934 0.92223 0.92508 0.92782 0.93033
##                            PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.53411 0.53027 0.51546 0.49741 0.47841 0.47072 0.46466
## Proportion of Variance 0.00221 0.00218 0.00206 0.00192 0.00177 0.00172 0.00167
## Cumulative Proportion  0.93254 0.93472 0.93678 0.93870 0.94047 0.94219 0.94386
##                            PC22    PC23   PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.45323 0.43326 0.4245 0.42230 0.41937 0.41226 0.39639
## Proportion of Variance 0.00159 0.00146 0.0014 0.00138 0.00136 0.00132 0.00122
## Cumulative Proportion  0.94546 0.94691 0.9483 0.94969 0.95105 0.95237 0.95359
##                            PC29    PC30    PC31    PC32    PC33    PC34    PC35
## Standard deviation     0.3928 0.38261 0.37828 0.37206 0.37071 0.36316 0.35639
## Proportion of Variance 0.0012 0.00113 0.00111 0.00107 0.00107 0.00102 0.00098
## Cumulative Proportion  0.9548 0.95592 0.95703 0.95810 0.95917 0.96019 0.96117
##                            PC36    PC37   PC38    PC39    PC40    PC41    PC42
## Standard deviation     0.35401 0.34416 0.3413 0.33644 0.33466 0.33113 0.32693
## Proportion of Variance 0.00097 0.00092 0.0009 0.00088 0.00087 0.00085 0.00083
## Cumulative Proportion  0.96215 0.96306 0.9640 0.96484 0.96571 0.96656 0.96739
##                            PC43   PC44    PC45    PC46    PC47    PC48    PC49
## Standard deviation     0.32402 0.3212 0.31530 0.31071 0.30979 0.30722 0.30184
## Proportion of Variance 0.00081 0.0008 0.00077 0.00075 0.00074 0.00073 0.00071
## Cumulative Proportion  0.96821 0.9690 0.96978 0.97052 0.97127 0.97200 0.97271
##                            PC50    PC51    PC52    PC53    PC54    PC55    PC56
## Standard deviation     0.29701 0.29507 0.29349 0.28981 0.28856 0.28380 0.28165
## Proportion of Variance 0.00068 0.00067 0.00067 0.00065 0.00065 0.00062 0.00061
## Cumulative Proportion  0.97339 0.97406 0.97473 0.97538 0.97603 0.97665 0.97727
##                            PC57    PC58    PC59    PC60    PC61    PC62    PC63
## Standard deviation     0.27973 0.27338 0.27283 0.26925 0.26474 0.26455 0.26326
## Proportion of Variance 0.00061 0.00058 0.00058 0.00056 0.00054 0.00054 0.00054
## Cumulative Proportion  0.97787 0.97845 0.97903 0.97959 0.98014 0.98068 0.98122
##                            PC64    PC65    PC66    PC67    PC68    PC69    PC70
## Standard deviation     0.25890 0.25557 0.25202 0.25131 0.24664 0.24452 0.24280
## Proportion of Variance 0.00052 0.00051 0.00049 0.00049 0.00047 0.00046 0.00046
## Cumulative Proportion  0.98174 0.98224 0.98273 0.98322 0.98370 0.98416 0.98462
##                            PC71    PC72    PC73    PC74    PC75    PC76    PC77
## Standard deviation     0.24075 0.23908 0.23625 0.23374 0.23224 0.23106 0.22951
## Proportion of Variance 0.00045 0.00044 0.00043 0.00042 0.00042 0.00041 0.00041
## Cumulative Proportion  0.98507 0.98551 0.98594 0.98636 0.98678 0.98720 0.98761
```

Ilustración 4: Resultado de PCA

Los resultados obtenidos implican que con los siete primeros genes es suficiente para realizar un análisis de expresión diferencial reduciendo notablemente el número de genes a utilizar. Ya que con estas siete primeras componentes se explica un 90% de la variación total [18].

En R se pueden representar gráficamente las dos primeras componentes separando los datos por colores en función del valor de una de las variables fenotípicas. Para empezar, se va a seleccionar la variable *node_status* para comprobar si el PCA es capaz de separar las muestras. En la siguiente imagen se puede ver la representación.

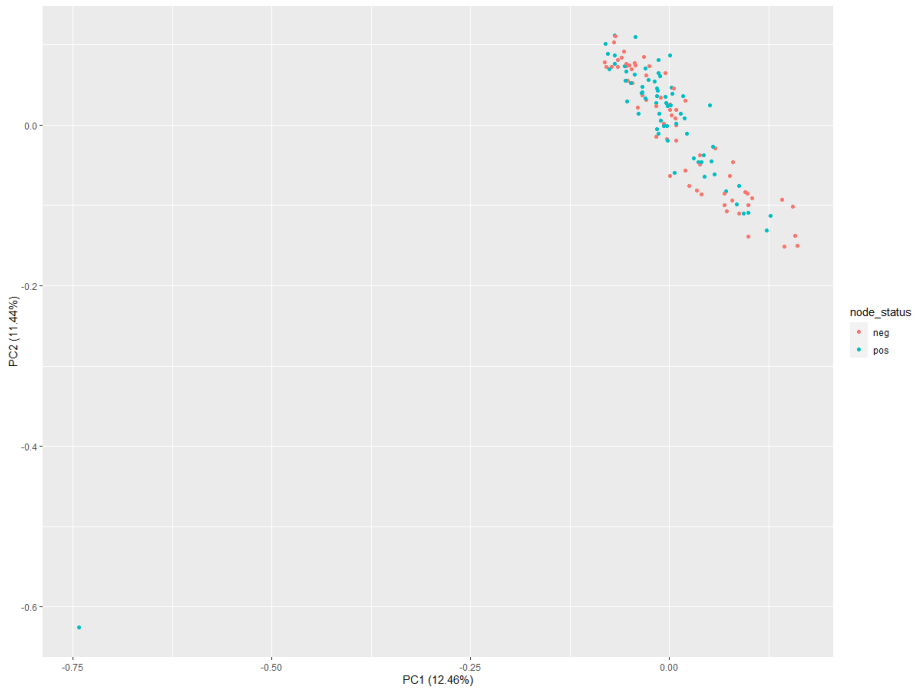


Ilustración 5: Muestras en función de la variable node_status con outlier

La imagen deja claro que las componentes no son capaces de diferenciar bien las muestras en dos grupos en función de si los nodos presentan células cancerosas o no. Además, se puede comprobar que hay un dato con unos valores muy alejados del resto de las muestras, que se puede ver en la parte inferior izquierda de la gráfica. Este dato se ha delimitado para su posterior eliminación de manera que no repercuta en el análisis.

Se ha vuelto a realizar la representación de las muestras junto con las dos primeras componentes del PCA para comprobar, como ya se había mencionado, que con la variable *node_status* no se pueden diferenciar las muestras.

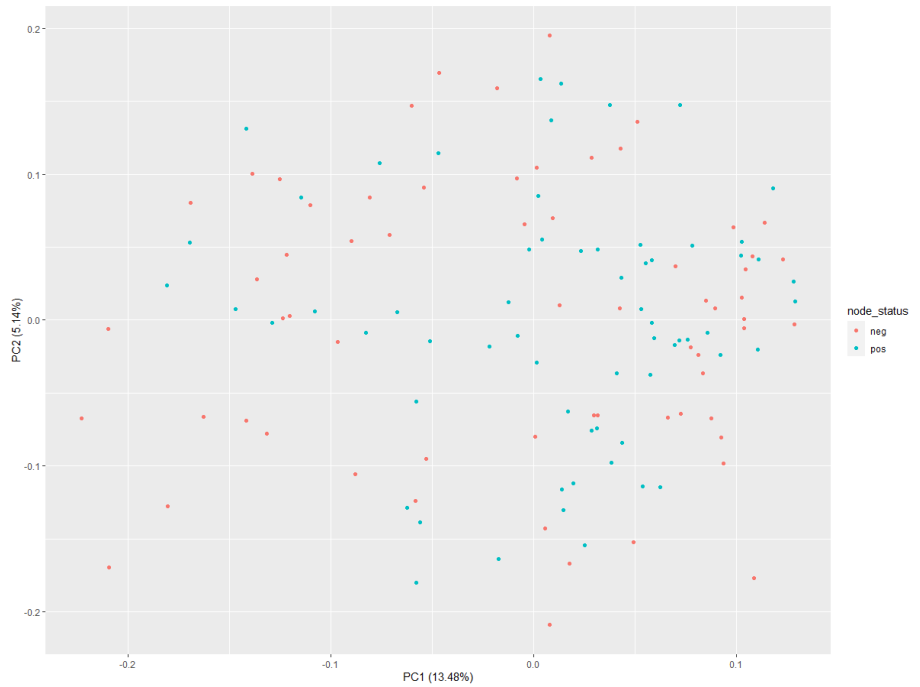


Ilustración 6: Muestras en función de la variable node_status

A continuación, se realiza el mismo procedimiento, pero con la variable *ER* y sin el dato *outlier* con el que se contaba al principio.

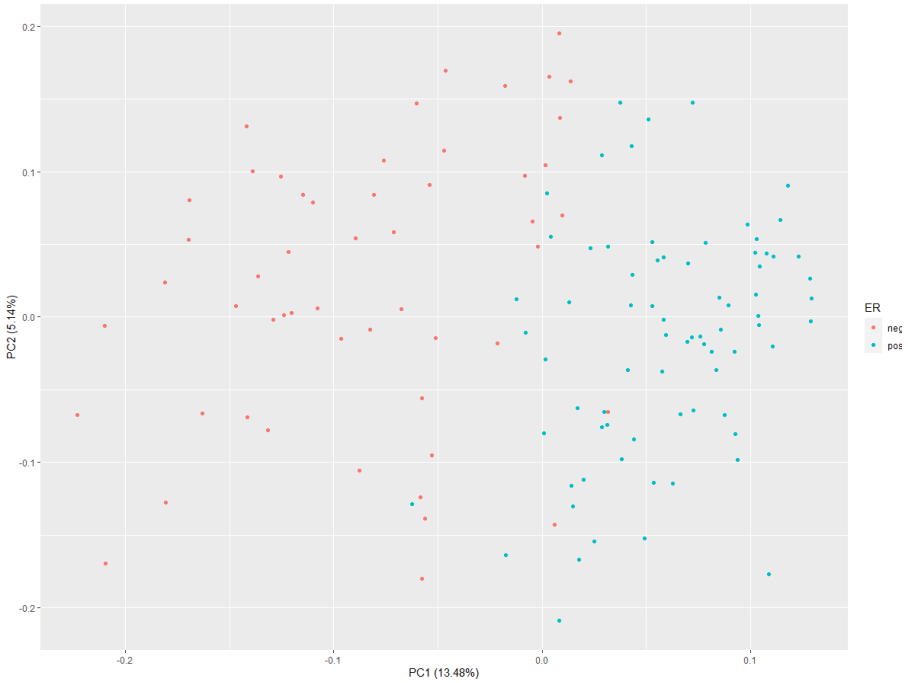


Ilustración 7: Muestras en función de la variable ER

En este caso, tan solo las dos primeras componentes son capaces de generar dos grupos bien diferenciados de pacientes con niveles positivos de la proteína o negativos.

Por último, se realiza el mismo esquema con la variable *HER2*.

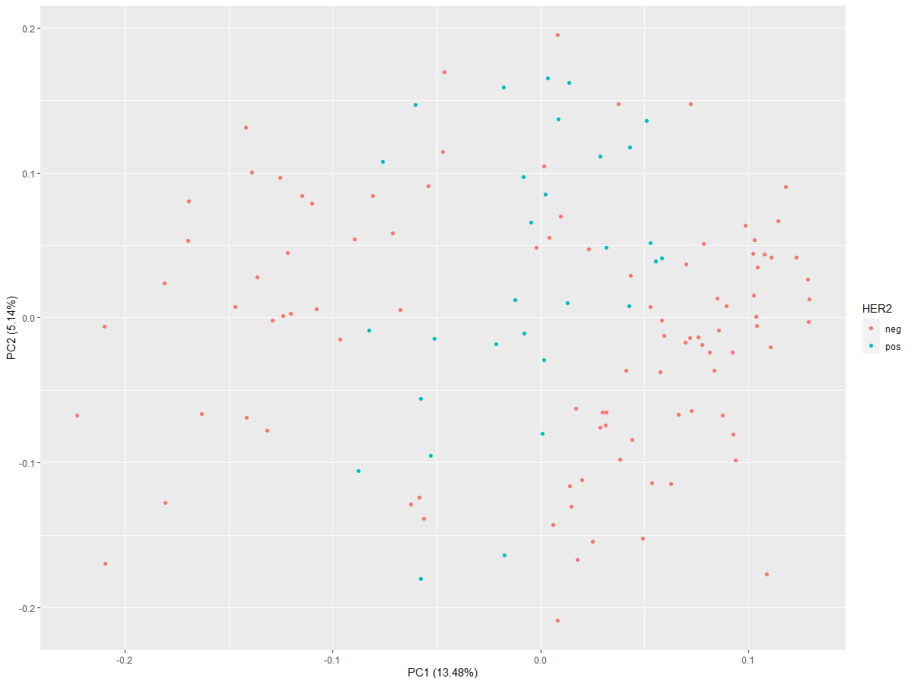


Ilustración 8: Muestras en función de la variable HER2

En estas condiciones, las dos componentes son capaces de separar las muestras ya que en la imagen se puede ver que las pacientes con niveles positivos de esta proteína se encuentran agrupadas entre las que presentan niveles negativos.

Este análisis revela bastante información del conjunto de datos porque se empieza a vislumbrar que las variables *HER2* y *ER* pueden dividir los datos en grupos según las dos primeras componentes principales que se han determinado. Además, es probable que no haya una expresión diferencial entre la variable *node_status* y los distintos genes de las muestras.

3.3 Análisis de grupos de genes

Los análisis realizados en los apartados anteriores se han realizado a nivel de gen, pero dado que se busca encontrar una posible relación entre gen y fenotipo, tiene sentido que existan relaciones entre fenotipo y grupos de genes. A partir de este momento, los análisis se realizan formando grupos de genes obteniendo la información de estos grupos de la base de datos de *Gene Ontology* [19].

Gene Ontology nace como organización en 1998 cuando varios investigadores comenzaron a estudiar el modelo genómico de los siguientes organismos: *Drosophila melanogaster* (mosca de la fruta), *Mus musculus* (ratón) y *Saccharomyces cerevisiae* (levadura de cerveza). Desde entonces el número de organismos de los que se ha añadido información ha aumentado considerablemente hasta llegar a miles. El principal recurso que ofrecen es una relación ontológica de la información biológica de distintas especies. Esto permite tener predefinida una agrupación de genes que tienen cierta relación entre ellos y que puede ayudar a clasificarlos.

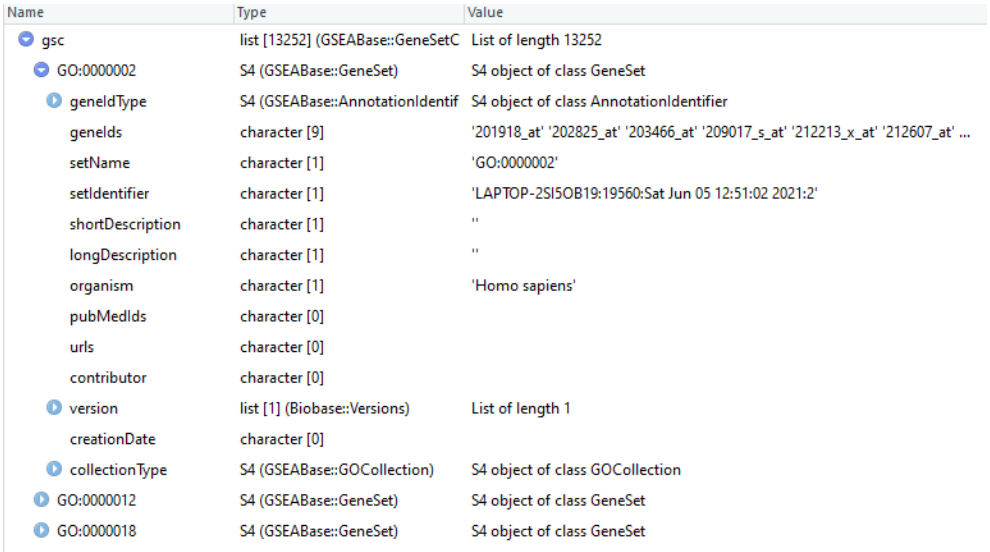
Así, se reduce considerablemente el espacio muestral ya que no se va a contar con los miles de genes de los que se partía al principio, si no con los grupos formados por estos genes. Y de estos grupos, se trabajará con los que sean suficientemente grandes para resultar interesantes en el análisis.

La librería *GSEABase* contiene funciones que permiten crear conjuntos a partir de la anotación con la que han sido tomadas las muestras y de la información disponible en *Gene Ontology*. En la siguiente imagen se puede ver la creación de los grupos a partir de los datos filtrados:

```
gse5460.gsc = GeneSetCollection(GSE5460_mas5_filt,setType=GOCollection())
names(gse5460.gsc) = unlist(lapply(gse5460.gsc,setName))
```

Ilustración 9: Código necesario para agrupar genes

Si se accede a la nueva variable creada con la agrupación de los genes se puede comprobar el número de grupos que se han creado y la información asociada a cada uno de ellos. En la siguiente imagen se muestra un ejemplo de estos datos:



Name	Type	Value
gsc	list [13252] (GSEABase::GeneSetC	List of length 13252
GO:0000002	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
geneIdType	S4 (GSEABase::AnnotationIdentif	S4 object of class AnnotationIdentifier
geneIds	character [9]	'201918_at' '202825_at' '203466_at' '209017_s_at' '212213_x_at' '212607_at' ...
setName	character [1]	'GO:0000002'
setIdentifier	character [1]	'LAPTOP-2SI5OB19:19560:Sat Jun 05 12:51:02 2021:2'
shortDescription	character [1]	''
longDescription	character [1]	''
organism	character [1]	'Homo sapiens'
pubMedIds	character [0]	
urls	character [0]	
contributor	character [0]	
version	list [1] (Biobase::Versions)	List of length 1
creationDate	character [0]	
collectionType	S4 (GSEABase::GOCollection)	S4 object of class GOCollection
GO:0000012	S4 (GSEABase::GeneSet)	S4 object of class GeneSet
GO:0000018	S4 (GSEABase::GeneSet)	S4 object of class GeneSet

Ilustración 10: Agrupación de genes

Uno de los aspectos más importantes de la realización de esta agrupación es que los genes no se agrupan según sus datos de expresión si no, a partir de ontologías ya creadas en base a otras especificaciones como ya se han mencionado.

El principal resultado que se pretende conseguir en esta fase es una ordenación de los grupos de genes. De manera que los grupos que aparezcan más arriba en el listado tengan una mayor asociación con la variable fenotípica que se esté estudiando en cada caso. Respecto a esto, los especialistas en la materia determinan dos puntos de vista en los que se plantean las siguientes hipótesis nulas [20]:

- **Q1**: Los genes pertenecientes a un mismo conjunto muestran el mismo tipo de asociación con el fenotipo comparado con la asociación mostrada por el resto de los genes. O, dicho de otro modo, los genes de un conjunto se encuentran tan diferencialmente expresados como los del resto de conjuntos. Los modelos que utilizan esta hipótesis se denominan test competitivo.

- **Q2**: El conjunto de genes no contiene ningún gen cuyo nivel de expresión esté asociado con el fenotipo del que se está realizando el estudio. Es decir, ningún gen del conjunto se encuentra diferencialmente expresado. Los modelos que utilizan esta hipótesis se denominan test autocontenido.

Estas hipótesis son con las que trabajan los modelos utilizados a continuación dependiendo de si buscan la asociación entre conjuntos de genes o lo contrario.

3.3.1 Test de Fisher

El test de Fisher permite comprobar si dos muestras de una misma población se encuentran relacionadas o son independientes entre sí. Este test utiliza como estadístico de contraste la distribución hipergeométrica. [21]

La distribución hipergeométrica se utiliza en experimentos en los que hay que seleccionar elementos, no independientes entre sí, de una muestra sin reemplazamiento. Este tipo de distribución se utiliza en muestras pequeñas y puede

ser confundida con la distribución binomial. La principal diferencia que tiene con la distribución binomial es que la probabilidad va variando al no haber reemplazamiento de un elemento seleccionado [22].

Si se define la variable aleatoria X = "número de éxitos obtenidos", la función de probabilidad es la siguiente:

$$P(X = x) = \frac{\binom{k}{x} \cdot \binom{N-k}{n-x}}{\binom{N}{n}}$$

Ecuación 1: Expresión matemática para el cálculo de la probabilidad

Donde N es el tamaño de la población, k es el número de individuos que cumple cierta característica, n es el tamaño de la muestra y x es el valor que toma la variable. Esta probabilidad se puede utilizar cuando se tienen los datos en forma de tabla de los que se conoce sus frecuencias marginales. Como se verá a continuación, se pueden agrupar los genes y formar una tabla de contingencia en la que se puede aplicar la distribución hipergeométrica.

El primer paso es definir los grupos de genes que se van a comparar para realizar el test. Primero, se necesita definir un grupo de genes significativo que puede venir determinado por un análisis previo. Después, se define el grupo con el que se quiere comparar a partir de la información sustraída de la base de datos *Gene Ontology* en función de su proximidad dentro de un mismo cromosoma o sus funciones biológicas.

En la siguiente imagen se puede ver la tabla de contingencias de los distintos grupos que se han formado de genes. Donde G identifica al grupo total de genes, S_0 es el grupo de genes significativo y S_1 es el grupo de genes con el que se quiere comparar. Los valores de la tabla indican el número de genes que se encuentran en solo un conjunto, o en ambos. Y N identifica el número total de genes. [15]

	S_1	**$G \setminus S_1$**	
S_0	n_{11}	n_{12}	n_{1-}
$G \setminus S_0$	n_{21}	n_{22}	n_{2-}
	n_{-1}	n_{-2}	N

Tabla 1: Tabla de contingencia

Prosiguiendo con los análisis anteriores, se realiza el Test de Fisher en las variables *node_status*, *HER2* y *ER*. Este test se va a realizar con el paquete *GOstats* que permite utilizar la distribución hipergeométrica para determinar cuáles son los grupos de genes significativos en la relación genotipo-fenotipo con cada una de las variables seleccionadas.

Primero, se empieza con el estudio de la variable *node_status*. Se determinan los grupos con los que se va a realizar el test y se obtienen dos grupos de genes importantes en la expresión diferencial de la variable seleccionada.

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0015728	0,001	Inf	0	1	1	*Mevalonate transport*
GO:0051780	0,001	Inf	0	1	1	*Behavioral response to nutrient*

Tabla 2: Resultado Test de Fisher para node_status

Se han obtenido dos grupos de genes con una expresión diferencial mayor al resto de grupos respecto a la variable *node_status*. Se puede ver que el valor de odds es infinito, a mayor valor de odds, mayor sobre expresión. Los grupos obtenidos son los siguientes:

- *Mevalonate transport*. Este grupo de genes son genes que se dedican a dirigir el movimiento del mevalonato entre las células, dentro o fuera de ellas. El mevalonato es un intermediario metabólico que participa en la ruta de síntesis del colesterol [23].

- *Behavioral response to nutrient*. Este grupo de genes se encuentran involucrados en procesos en los que aparece un cambio de comportamiento frente a un estímulo proveniente de un nutriente.

Se realiza el mismo procedimiento con la variable *HER2* y se obtienen los siguientes resultados:

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006487	0,001	4,445	2	9	42	*Protein N-linked glycosylation*
GO:0051055	0,001	6,950	1	6	20	*Negative regulation of lipid biosynthetic process*
GO:0042180	0,001	2,684	7	16	114	*Cellular ketone metabolic process*

Tabla 3: Resultado Test de Fisher para HER2

En este caso se han obtenido tres grupos de genes con expresión diferencial respecto a la variable *HER2*. Los niveles de odds son mayores a uno lo que demuestra una sobre expresión como ocurría en el caso anterior. Los grupos obtenidos son:

- *Protein N-linked glycosylation*. Estos genes participan en procesos en los que se agrega un carbohidrato, o derivado de este, a una proteína a través de átomos concretos.

- *Negative regulation of lipid biosynthetic process*. Este grupo contiene a los genes que intervienen en procesos que detienen, evitan o reducen la frecuencia, velocidad o extensión de las reacciones que dan lugar a la formación de lípidos.

- *Cellular ketone metabolic process*. Estos genes participan en reacciones químicas en las que se trata con compuestos orgánicos que contienen al grupo carbonilo (CO).

Por último, se realiza el análisis con la última variable elegida, *ER* obteniendo la siguiente tabla de resultados:

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0051983	0,000	3,114	35	48	58	*Regulation of chromosome segregation*

Tabla 4: Resultado Test de Fisher para ER

En este último caso solo se obtiene un grupo significativo frente a la expresión diferencial de la variable fenotípica *ER*. El grupo obtenido ha sido:

- *Regulation of chromosome segregation*. Este grupo contiene a los genes que intervienen en procesos que detienen, evitan o reducen la frecuencia, velocidad o extensión del procedimiento de separación del material genético en forma de cromosomas.

No hay ningún grupo en común en los tres análisis efectuados así que no se puede determinar que exista un determinado grupo de genes relacionados con los tumores de mama. Lo que sí se puede determinar es que la expresión de distintos grupos está relacionada con distintas variables fenotípicas, es decir, existe una expresión diferencial entre ellos. Esto puede llevar a pensar que los grupos de genes obtenidos en el análisis pueden determinar el posible desarrollo de la enfermedad.

3.3.2 GSA

En este análisis se busca un valor marginal de cada gen que represente el nivel de asociación entre un gen y una variable fenotípica concreta. Al estudiar grupos de genes, este valor puede ser asociado a cada uno de los grupos realizando la media aritmética del valor marginal de cada uno de los genes que forman el grupo. Esto tiene sentido porque genes que están relacionados o sean cercanos tienen patrones de expresión similares. Estos valores que se les asigna a cada grupo de genes generan el estadístico de contraste que se utiliza en el método GSA. Este estadístico es conocido como *maxmean* debido a que se calcula la media aritmética de los valores positivos y negativos que se les ha asignado a cada gen y se elige el mayor en valor absoluto [24].

Este método obtiene dos listados de grupos de genes. Por un lado, aquellos que tienen una diferenciación positiva y por otro, los que tienen una diferenciación negativa.

RStudio cuenta con un paquete llamado *GSA* que implementa el método explicado. Utilizando este paquete se obtienen los grupos significativos para cada variable fenotípica. Este estudio se va a realizar con las variables utilizadas durante todo el proyecto: *nodes_status*, *HER2* y *ER*.

Primero, se empieza por la variable *node_status* de la que se obtiene la siguiente representación del p-valor en función de la tasa de falsos positivos. En verde aparecen representados los grupos con una asociación negativa y en azul aquellos que tienen una asociación positiva.

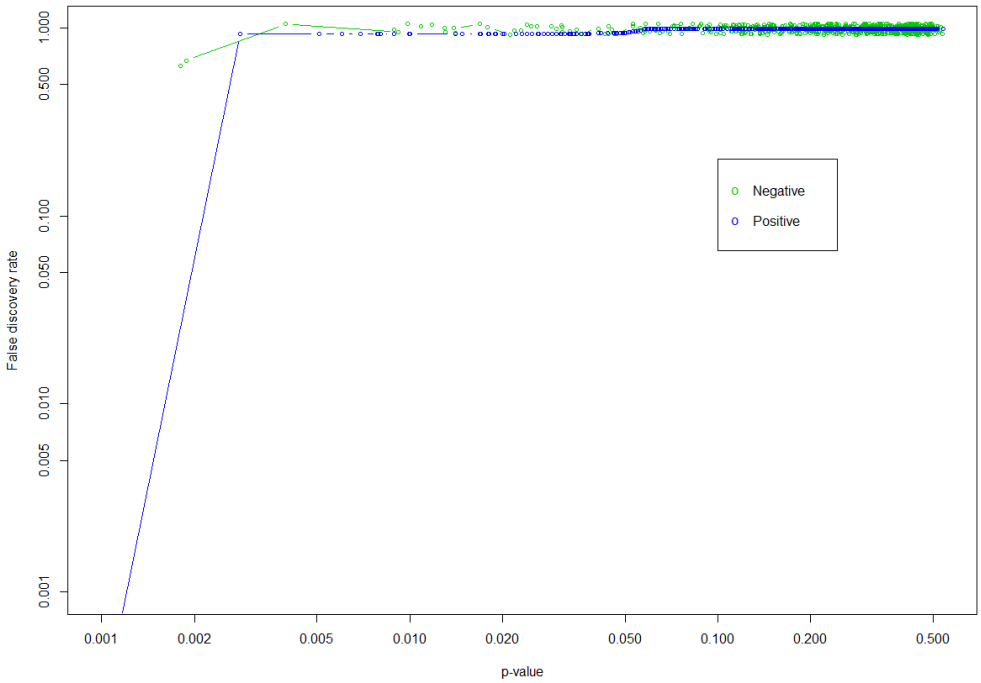


Ilustración 11: Resultado GSA para node_status

En la gráfica no se aprecia una gran diferencia entre los grupos positivos o negativos. La información de los conjuntos de genes, tanto positivos como negativos, es accesible mediante el modelo. A continuación, se numeran los tres grupos de genes más significativos con asociación positiva para esta variable:

- GO:0006294. *Nucleotide-excision repair, preincision complex assembly*. Se encarga de formar complejos multiproteicos con las proteínas del ADN.

- GO:0006298. *Mismatch repair*. Repara errores surgidos en la replicación y recombinación del ADN.

- GO:0006351. *Transcription, DNA-templated*. Participan en la síntesis celular de ARN.

Y ahora, los tres grupos de genes más significativos con asociación negativa:

- GO:0001933. *Negative regulation of protein phosphorylation*. Procesos que detienen, evitan o reducen la unión de grupos fosfato a aminoácidos dentro de una proteína.

- GO:0006486. *Protein glycosylation*. Proceso que tiene como resultado la adición de un carbohidrato a una proteína.

- GO:0006936. *Muscle contraction*. Proceso que modifica la geometría de los músculos.

A continuación, se realiza el mismo análisis sobre la variable *HER2*.

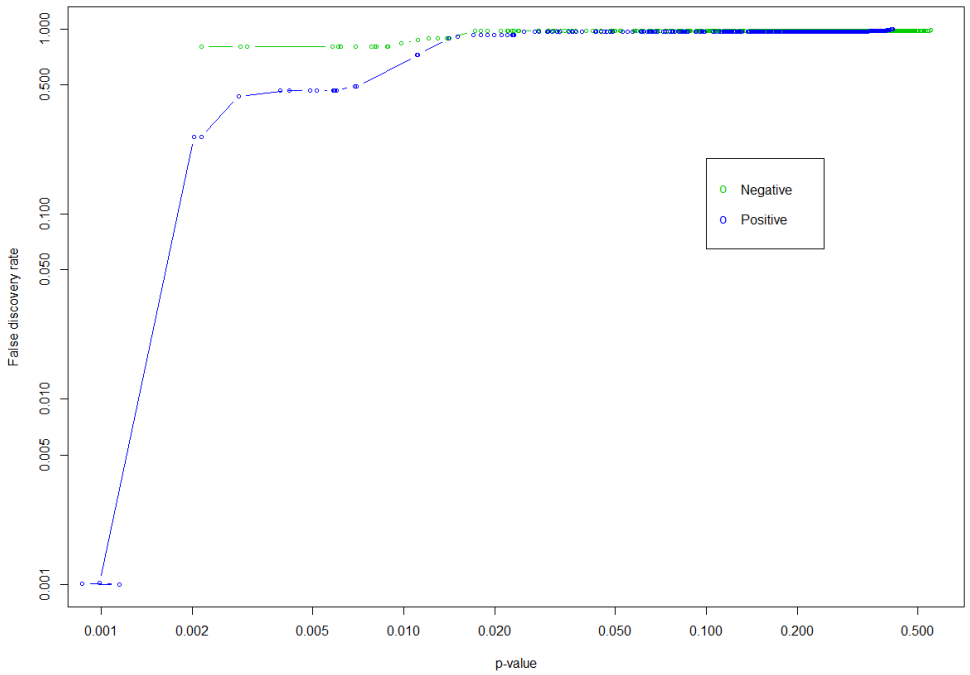


Ilustración 12: Resultado GSA para HER2

En este caso, se puede aprecia una ligera diferencia entre los grupos de genes cuando el p-valor de ellos es relativamente pequeño. Se procede a listar los grupos de genes más significativos con asociación positiva:

- GO:0000188. *Inactivation of MAPK activity*. Procesos relacionados con la activación de una enzima.
- GO:0006487. *Protein N-linked glycosylation*. Proceso que tiene como resultado la adición de un carbohidrato a una proteína.
- GO:0006898. *Receptor-mediated endocytosis*. Se involucran en procesos de transporte de macromoléculas dentro de la célula.

Los conjuntos de genes más significativos con asociación negativa son:

- GO:0001658. *Branching involved in ureteric bud morphogenesis*. Proceso que genera y organiza la estructura ramificada de un tubo epitelial entre los riñones y el uréter.
- GO:0001843. *Neural tube closure*. Este proceso es el último paso en la formación de los tubos neuronales.
- GO:0001942. *Hair follicle development*. Como su nombre indica participan en el proceso de crecimiento del cabello.

Por último, se realiza el análisis GSA sobre la variable *ER*.

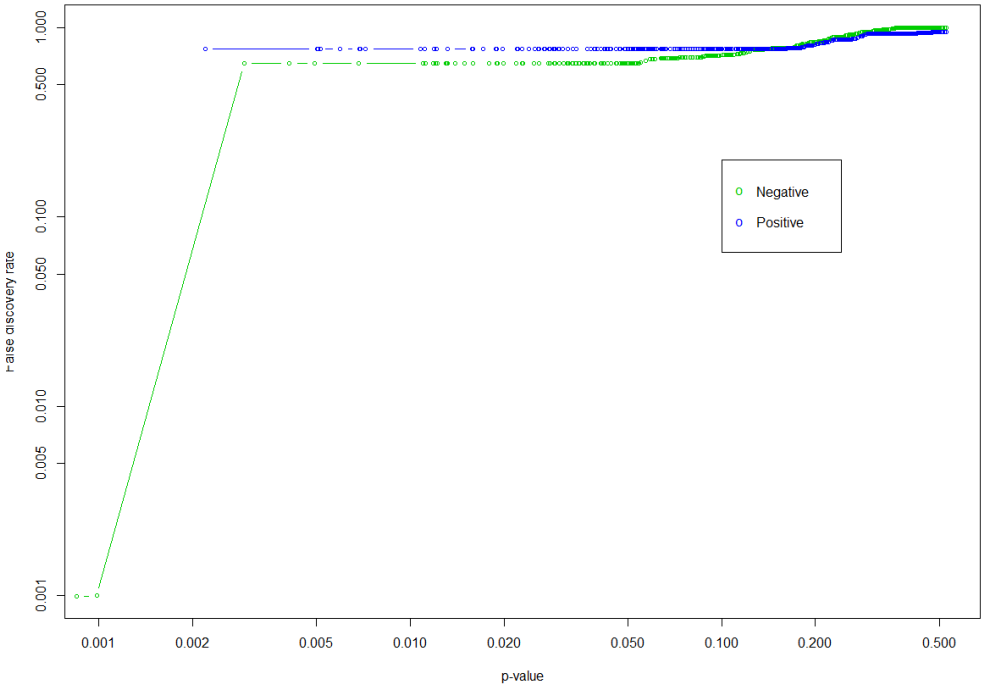


Ilustración 13: Resultado GSA para ER

Los conjuntos de genes de la variable fenotípica *ER* sí que presentan una diferencia entre ambos grupos, aunque esta es mínima. Los grupos de genes positiva con mayor significación son:

- GO:0000381. *Regulation of alternative mRNA splicing, via spliceosome*. Procesos que modulan la frecuencia y velocidad en las uniones de cadenas de ARN.

- GO:0001701. *In utero embryonic development*. Procesos que se encargan del crecimiento del cigoto hasta el nacimiento.

- GO:0001755. *Neural crest cell migration*. Procesos de movimiento de células en un embrión.

Y los grupos de genes negativos con mayor significación son:

- GO:0000070. *Mitotic sister chromatid segregation*. Se encargan de procesos de mitosis celular.

- GO:0000082. *G1/S transition of mitotic cell cycle*. Intervienen en una de las fases de la mitosis celular.

- GO:0000278. *Mitotic cell cycle*. Aparecen en las transiciones de las fases de la mitosis celular.

Ninguno de los grupos de genes más significativos, tanto positivos como negativos, han coincido en el análisis de las tres variables. Esto puede implicar que no exista un grupo concreto que esté implicado en el desarrollo de cáncer de mama, aunque niveles de expresión altos de algunos grupos sí que pueden estar relacionados con la aparición de ciertos fenotipos.

4. Análisis de los resultados

En este capítulo se aborda el análisis de los resultados obtenidos tras la realización de los análisis descritos anteriormente para comprobar si se ha podido dar respuesta a las preguntas planteadas al inicio del estudio.

El trabajo se ha realizado centrándose en tres variables fenotípicas distintas para comprobar si existen genes, o grupos de genes, cuya abundancia genética determine si un tumor presentará alguna de estas tres características.

El análisis de componentes principales concluía que no existían en las muestras componentes principales para esta variable fenotípica. Sin embargo, el test de Fisher y el GSA obtienen grupos de genes significativos para esta variable. Aunque no concluyen con los mismos conjuntos, el resultado del modelo GSA es interesante porque los grupos de genes forman parte de procesos relacionados con replicación de ADN y ARN.

En cambio, la presencia de la proteína *HER2* sí que puede estar relacionada con la presencia de ciertos genes en un paciente por los resultados aportados por el análisis de componentes principales. Además, tanto el test de Fisher como el GSA han obtenido dos grupos de genes muy similares que participan en procesos de agregación de carbohidratos. Sería interesante comprobar mediante otros análisis si se vuelve a obtener el mismo conjunto de genes.

Al igual que con la proteína anterior, la proteína *ER* sí que ha obtenido componentes principales que permitían dividir a las muestras del *dataset*. El test de Fisher y el GSA han obtenido conjuntos de genes involucrados en procesos separación de material genético como es durante el desarrollo embrionario. La principal función de esta proteína es favorecer el crecimiento de las células cancerosas y dado que los genes obtenidos se encuentran involucrados en procesos de crecimiento, es posible que un mayor nivel de expresión de los genes significativos obtenidos tenga relación con los niveles de esta proteína.

Una de las preguntas que se plantearon al inicio estaba relacionada con mutaciones en ciertos genes que podrían implicar una alta probabilidad de desarrollar tumor de mama, esta pregunta no ha sido respondida debido a que no se contaba con información del estado de los genes. Es decir, no se sabía si un gen presentaba mutaciones o no. Por eso, no se ha podido determinar la relación de mutaciones en genes relacionados con el cáncer.

Otra pregunta planteada inicialmente estaba relacionada con la creación de un tratamiento único para cada persona a raíz de su estudio genético para aumentar su probabilidad de supervivencia. No se ha podido llegar a una conclusión certera sobre esta pregunta. Sí que se puede determinar el tipo de cáncer, sus características y su posible evolución en el futuro, y esto puede influir en el tipo de tratamiento que un especialista pueda designar a un paciente. Además, a través del estudio genético se puede prever si una persona es más propensa a padecer esta enfermedad y tomar medidas preventivas antes de que aparezca.

En este estudio se han determinado algunos genes y grupos de genes cuya expresión determina el estado de algunas variables fenotípicas. Es muy probable que ampliando el estudio al resto de variables y contando con un mayor número de muestras se pueda determinar un listado de genes implicados en el desarrollo de tumores de mama.

5. Conclusiones y líneas de futuro

5.1 Problemas encontrados

Durante estos meses de trabajo han ido apareciendo abundantes problemas que han afectado al desarrollo del análisis expuesto como aparecen en cualquier proyecto de Ciencia de Datos. Pero en este apartado no se van a detallar todos, sólo se van a plantear los más relevantes y que más han afectado al desarrollo. Ha sido necesario introducir algunas modificaciones en la planificación inicial del proyecto debido a los problemas que se describen a continuación.

El principal problema que se ha encontrado a lo largo del proceso ha sido el tratamiento de los datos por el formato en el que se encontraban inicialmente. Este formato es específico para las expresiones de los genes tomados con una tecnología concreta. Además, el archivo obtenido inicialmente contaba con muchísima información no relevante para el análisis, que se ha tenido que limpiar, lo que ha hecho que el proceso de preprocesado y análisis fuese más extenso de lo esperado.

También ha sido complicada la selección de los modelos sobre los que realizar el análisis. Actualmente existen varios modelos que se pueden utilizar con datos génicos, pero hay poca información práctica relativa a ellos lo que ha provocado retrasos en el desarrollo del proyecto. Debido a ello, el tiempo de elección de modelo se ha alargado junto con el tiempo empleado para la comprensión de los resultados obtenidos.

5.2 Conclusiones

La parte más importante y eje central del trabajo ha sido la comprensión y el tratamiento de los datos obtenidos. Tal y como se ha mencionado en el apartado anterior, el formato de los datos ha ocasionado problemas. Esto ha hecho que se pueda comprender como funciona realmente un proyecto de Ciencia de Datos con datos reales con todos los problemas que puedan surgir. Aparte de utilizar conceptos aprendidos en algunas asignaturas del máster como Estadística Avanzada o Minería Avanzada de Datos, se han aprendido nuevos algoritmos de clasificación aplicados a biomedicina que se desconocían hasta el momento de desarrollo de este proyecto.

En relación con los objetivos planteados inicialmente, los objetivos principales se han cumplido. Se ha trabajado con una base de datos de información genética e histológica con 129 muestras de pacientes con tumores de mama. Estos datos se han procesado, filtrado y analizado con diversos modelos que utilizan distintas técnicas para la clasificación de genes o grupos de genes. El trabajo se ha centrado en el estudio de tres variables de las que se han obtenido algunos genes que pueden estar implicados en la presencia de ciertos fenotipos, pero no se puede asegurar que estos genes impliquen diferencias en el resto de las variables fenotípicas.

Al principio del proyecto se marcaron una serie de objetivos secundarios derivados de los objetivos principales. Uno de ellos hacía referencia a la utilidad de la Ciencia de Datos en estudios relacionados con la medicina y este objetivo se ha cumplido. A partir de la utilización de modelos específicos de la Ciencia de Datos con datos de tipo genético, se ha podido extraer información relevante para al ámbito de medicina. La importancia de la Ciencia de Datos en los estudios de

bioinformática queda patente al poder determinar cómo pueden influir ciertos genes en el desarrollo de la enfermedad a partir de la utilización de algoritmos de clasificación. Estos resultados pueden ser aprovechados para predecir posibles casos de tumores antes de que el paciente comience a mostrar síntomas físicos y pueda ser demasiado tarde. O, llevar un seguimiento preventivo de personas que por los niveles de expresión de sus genes puedan ser propensos a padecer esta enfermedad.

5.3 Líneas de trabajo futuras

Este proyecto abre varias posibles líneas de trabajo futuras en relación con los resultados obtenidos.

Primeramente, existen múltiples modelos de clasificación de datos que no ha sido posible utilizar durante el desarrollo de este proyecto. Sería interesante como posible línea de trabajo comparar los resultados obtenidos en el presente trabajo con los obtenidos con otros métodos de análisis. Y comprobar si los genes o grupos de genes obtenidos con distintos métodos tienen algún tipo de relación entre ellos. Especialmente los obtenidos para la variable *ER*, que han sido los más concluyentes. Incluso, se podrían replicar los modelos expuestos en este trabajo con el resto de las variables fenotípicas que no se han utilizado.

Asimismo, sería interesante poder contar con una línea de investigación en la que se llevasen a cabo ensayos de laboratorio, para comprobar si los resultados obtenidos mediante estos ensayos coinciden con los obtenidos mediante técnicas de Ciencia de Datos. Y corroborar que ciertos niveles de expresión de genes dan lugar a un fenotipo concreto.

Una posible manera de enriquecer este proyecto mediante técnicas de Ciencia de Datos sería investigar algún método de visualización que pudiese abarcar los resultados obtenidos en laboratorio junto con los resultados obtenidos mediante algoritmos de clasificación. De este modo, los productos y conclusiones obtenidas se podrían publicar y llegar a personas del ámbito de la medicina y del ámbito de la Ciencia de Datos.

En última instancia, se podría ampliar el estudio genético a la creación de tratamientos personalizados en función del nivel de expresión de los genes de cada persona. Ya que algunas características fenotípicas que presentan los tumores determinan el tipo de tratamiento a seguir por el paciente.

6. Glosario

Affymetrix: Empresa especializada en el diseño de microarrays de ADN.

GEO: Base de datos para perfiles de expresión génica gestionada por el Centro Nacional de Información Biotecnológica (NCBI).

R: Lenguaje de programación enfocado al análisis estadístico.

RStudio: Entorno de desarrollo para el lenguaje de programación R.

Bioconductor: Proyecto software de código abierto para el análisis de datos génicos.

7. Bibliografía

[1] Asociación Española Contra el Cáncer. (2021). Recuperado de https://www.aecc.es/es

[2] Rossing, M., Storgaard Sørensen, C., Ejlertsen, B., & Cilius Nielsen, F. (2019, 28 enero).
Whole genome sequencing of breast cancer. Recuperado de https://onlinelibrary.wiley.com/doi/full/10.1111/apm.12920

[3] Tecnología de microarrays (chips de ADN o ARN) | NHGRI. (2021). Recuperado de https://www.genome.gov/es/genetics-glossary/Tecnologia-de-microarrays

[4] Stanford Genomics Breast Cancer Consortium Portal. (2021). Recuperado de http://genome-www.stanford.edu/breast_cancer/

[5] Sorlie, T., Tibshirani, R., S. Parker, J., & Hastie, T. (2003, enero). Repeated Observation of
breast tumor subtypes in independent gene expression data sets. Recuperado de https://www.researchgate.net/publication/222110628_Repeated_Observation_of_breast_tumor_subtypes_in_independent_gene_expression_data_sets

[6] M. Perou, C., Sorlie, T., B. Eisen, M., & van de Rijn, M. (2000, septiembre). Molecular portraits of human breast tumours. Nature. Recuperado de https://www.researchgate.net

[7] Harari, D., & Yarden, Y. (2000, 9 diciembre). Molecular mechanisms underlying ErbB2/HER2 action in breast cancer. Oncogene. Recuperado de https://www.nature.com

[8] Sorlie, T., M. Perou, C., Tibshirani, R., & Aas, T. (2001, octubre). Gene expression patterns
of breast carcinomas distinguish tumor subclasses with clinical implication. Proceedings of
the National Academy of Sciences. Recuperado de https://www.researchgate.net

[9] Usary, J. (2004, 13 septiembre). Mutation of GATA3 in human breast tumors. Oncogene.
Recuperado de https://www.nature.com

[10] Wang, Y. A. (2018, 22 marzo). Germline breast cancer susceptibility gene mutations and
breast cancer outcomes. PubMed. Recuperado de https://pubmed.ncbi.nlm.nih.gov

[11] Julià Minguillón. Fundamentos de data science. UOC. PID_00235534

[12] What is the Data Information Knowledge Wisdom Pyramid? (2021). Ontotext. https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/

[13] Lu X, Lu X, Wang ZC, Iglehart JD et al. Predicting features of breast cancer with gene expression patterns. Breast Cancer Res Treat 2008 Mar;108(2):191-201. PMID: 18297396

[14] Susan G. Komen®. (2021, 12 febrero). Factors that Affect Treatment and Prognosis. https://www.komen.org/breast-cancer/diagnosis/factors-that-affect-prognosis/

[15] Ayala, G. (2019). Bioinformática Estadística. Análisis estadístico de datos ómicos. https://www.uv.es/ayala/docencia/tami/tami13.pdf

[16] L. (2021, 28 febrero). Lindseynicer/How-to-analyze-GEO-microarray-data. GitHub. https://github.com/Lindseynicer/How-to-analyze-GEO-microarray-data/blob/main/GSE_analysis_microarray.Rmd

[17] Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE by Joaquín Amat Rodrigo, available under a Attribution 4.0 International (CC BY 4.0) at https://www.cienciadedatos.net/documentos/35_principal_component_analysis

[18] Gironés Roig, J., Casas Roma, J., Minguillón Alfonso, J., & Caihuelas Quiles, R. (2017). Minería de datos. Modelos y algoritmos. Editorial UOC.

[19] GeneOntology. (1999). GeneOntology. http://geneontology.org/

[20] Lu Tian y col. «Discovering statistically significant pathways in expression profiling studies». En: Proceedings of the National Academy of Sciences of the United States of America 102.38 (2005), págs. 13544-13549. doi: 10.1073/pnas.0506577102.
eprint: http : / / www . pnas . org / content / 102 / 38 / 13544 .full.pdf+html. url: http://www.pnas.org/content/102/38/13544.abstract.

[21] RPubs - Test exacto de Fisher, chi-cuadrado de Pearson, McNemar y Q-Cochran. (2016, 21 octubre). Test estadísticos para variables cualitativas: test exacto de Fisher, chi-cuadrado de Pearson, McNemar y Q-Cochran. https://rpubs.com/Joaquin_AR/220579#:%7E:text=El%20test%20exacto%20de%20Fisher,eventos%20dentro%20de%20una%20tabla.&text=Si%20las%20frecuencias%20marginales%20son,el%20valor%20de%20las%20dem%C3%A1s.

[22] Modelo Estadística. (s. f.). José R. Galo Sánchez. https://proyectodescartes.org/iCartesiLibri/materiales_didacticos/EstadisticaProbabilidadInferencia/VAdiscreta/4_1DistribucionHipergeometrica/index.html

[23] Clínica Universidad de Navarra. (s. f.). Mevalonato. Diccionario médico. Clínica Universidad de Navarra. https://www.cun.es/diccionario-medico/terminos/mevalonato

[24] Bradley Efron y Robert Tibshirani. «On testing the significance of sets». En: Annals of Applied Statistics 1.1 (2007). Gene set analysis, págs. 107-129. doi: 10.1214/07-AOAS101.

8. Anexos

8.1 Anexo 1. Código proyecto.

https://github.com/sgarridoromero/TFM