Universitat Oberta de Catalunya

uoc.edu

Estudio de las bases moleculares del cáncer de mama

Gabriel Quintairos Rial
Máster en Ciencia de Datos
Área 2

Carles Barceló
Jordi Casas Roma

08/01/2020

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Estudio de las bases moleculares del cáncer de mama*
Nombre del autor:	*Gabriel Quintairos Rial*
Nombre del consultor/a:	*Carles Barceló*
Nombre del PRA:	*Jordi Casas Roma*
Fecha de entrega (mm/aaaa):	01/2020
Titulación:	*Máster en Ciencia de Datos*
Área del Trabajo Final:	*Trabajo de Fin de Máster – Área 2*
Idioma del trabajo:	*Castellano*
Palabras clave	*DDR, GSEA, cáncer*

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

Existen varios motivos por los cuales se pueden producir daños en el ADN, los cuales pueden provocar enfermedades como el cáncer. Cuando estos daños se producen, algunos mecanismos en nuestras células inician un proceso de reparación del daño en el ADN (en adelante "DDR" por sus siglas en inglés). Conocer y comprender estos mecanismos puede mejorar la predicción de riesgo de cáncer y el tratamiento en las fases tempranas.

El Linfoma de las células del manto (en adelante "MCL por sus siglas en inglés) es una enfermedad con mal pronóstico y difícil de diagnosticar. Dicho tumor se caracteriza por la sobreexpresión de la ciclina D1 y la unión de ésta proteína a determinadas regiones del ADN implicadas en la regulación del DDR.

El cáncer de mama es el más diagnosticado en mujeres por encima de cualquier otro tipo de tumor. El diagnóstico precoz es esencial en el tratamiento, ya que dependiendo de su localización la tasa de supervivencia se puede ver altamente reducida. Por ello, resulta relevante disponer de biomarcadores de la enfermedad que permitan un diagnóstico precoz. El DDR es un proceso que se cree muy importante en el desarrollo de este tumor.

Este proyecto pretende analizar la similitud de la expresión génica regulada por DDR comparando cáncer de mama con leucemia. Se utilizaran los data sets publicados para generar una firma genética en la que se identifiquen los genes significativamente "enriquecidos" para así poder estudiar su posible papel como posible diana terapéutica y como biomarcador. Se utilizará el entorno "R" para alinear los reads, generar quality controls y finalmente generar la firma genética mediante Gene Set Enrichment Analysis (GSEA).

Tras la realización de su desarrollo, se han conseguido identificar varios procesos comunes a los dos tipos de cáncer y a la DDR. Además, también se

ha conseguido identificar aquellos procesos comunes enriquecidos positivamente mediante la utilización de la técnica de Random Forest.

Abstract (in English, 250 words or less):

There are several reasons why DNA damage can occur, which can cause diseases such as cancer. When these damages occur, some mechanisms in our cells initiate a process of DNA damage repair (hereinafter "DDR"). Knowing and understanding these mechanisms can improve cancer risk prediction and treatment in the early stages.

The mantle cell lymphoma (hereinafter “MCL) is a disease with poor prognosis and difficult to diagnose. This tumor is characterized by overexpression of cyclin D1 and the binding of these proteins to certain regions of the DNA involved in the regulation of DDR.

Breast cancer is the most diagnosed in women above any other type of tumor. Early diagnosis is essential in treatment, since it can determine its survival rate can be greatly reduced. Therefore, is crucial to find relevant biomarkers in order to improve early diagnosis. DDR is a process that is considered to be very important in the development of this tumor.

This project aims to analyze the similarity of DDR-regulated gene expression by comparing breast cancer with leukemia. We used the published data sets to generate a genetic signature in which the specifically "enriched" genes are identified in order to study their possible role as a possible therapeutic target and as a biomarker. We used the "R" software to align the readings, generate quality controls and finally generate the genetic signature using Gene Set Enrichment Analysis (GSEA).

After the development, several processes common to both types of cancer and DDR have been identified. In addition, we managed to identify those common processes positively enriched by using the Random Forest technique.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

En las células humanas, tanto las actividades metabólicas como los factores ambientales, como los rayos UV o la radiactividad, pueden causar daños al ADN. Muchas de estas lesiones producen mutaciones potencialmente nocivas en el genoma de la célula, lo que afecta la supervivencia de sus «células hijas» a la hora de la mitosis o induce procesos de malignización que acababan desembocando en un tumor. Se han vinculado varios cánceres humanos a anormalidades de ADN como duplicaciones, translocaciones (se transfiere una parte de un cromosoma a otro), deleciones (se elimina una parte del cromosoma) y mutaciones (se altera la secuencia del cromosoma). La aclaración de los mecanismos que inician el proceso de reparación del daño al ADN (en sus siglas en inglés, DNA-damage response o "DDR") llevará a mejorar la predicción de riesgo de cáncer y el tratamiento en las fases tempranas. Estudios más extensos sobre el daño y las vías de reparación del ADN podrían llevar al desarrollo de nuevas terapias destinadas a reforzar los sistemas de defensa naturales de las células que impiden que se desarrolle un tumor (Jackson & Bartek, 2009).

Dentro de la leucemia, en particular el Linfoma de las células del manto (en adelante, en sus siglas en inglés Mantle Cell Lymphoma "MCL") tiene el peor pronóstico dentro de las Leucemias ya que la supervivencia mediana de los pacientes es de cerca de 3 años. Es una enfermedad difícil de diagnosticar, raramente considerada curada, e identificada en la década de 1990. La investigación para encontrar biomarcadores para diagnosticarlo mejor se persigue activamente por todo el mundo. Dicho tumor se caracteriza por la sobreexpresión de la ciclina D1 y la unión de ésta proteína a determinadas regiones del ADN implicadas en la regulación del DDR.

Por otra parte, el cáncer de mama es el que más se diagnostica en mujeres, por encima de cualquier otro tipo de tumor. Se piensa que la DDR es un proceso muy importante en el desarrollo de este tumor. La probabilidad de padecerlo en algún momento de la vida es de aproximadamente 12%. Aunque la prevalencia no es muy elevada, el diagnóstico precoz es esencial en el tratamiento. Si el cáncer está localizado en los ganglios linfáticos regionales, la tasa de supervivencia a 5 años es del 85% mientras que si el cáncer se ha diseminado a una parte distante del cuerpo, la tasa de supervivencia a 5 años es sólo de 27%. Desgraciadamente, alrededor del 5% de las mujeres tienen cáncer metastásico cuando se les diagnostica cáncer de mama por primera vez. Por ello, resulta relevante disponer de biomarcadores de la enfermedad que permitan un diagóstico precoz (Lee, K.J., Piett, C.G., Andrews, J.F. et al., 2019).

El proyecto que aquí se presenta pretende analizar la similitud de la expresión génica regulada por DDR (respuesta al daño al ADN) comparando cáncer de mama con leucemia. Se utilizaran los data sets publicados para generar una firma genética ("gene signature") en la que se indentifiquen los genes significativamente "enriquecidos" para así poder estudiar su posible papel como posible diana terapéutica y como biomarcador. Se utilizará el

entorno “R” para alinear los reads, generar quality controls y finalmente generar la firma genética mediante Gene Set Enrichment Analysis (GSEA).

Se debe tener en cuenta de que se trata de un proyecto real de investigación biomédica con gran aplicabilidad en el tratamiento de tumores y un retorno social muy elevado. Sin embargo, no exento de incertidumbre en cuanto a la posibilidad de obtener resultados claros y contundentes ya, como sucede en cualquier tipo de investigación científica, nunca antes se ha realizado.

1.1.1 Motivación personal

Actualmente y desde hace más de un año trabajo en el departamento de tecnologías y sistemas informáticos de un hospital. A lo largo de mi estancia he tenido la oportunidad de participar en un proyecto de investigación relacionado con los factores que provocaban un mayor riesgo de reintervención en operaciones de estrabismo. En dicho proyecto utilizamos el programa R y conjuntos de datos del histórico de intervenciones del hospital para generar diversos modelos que fueron presentados en el XXVII Congreso de la Sociedad Española de Estrabología y Oftalmología Pediátrica (Sociedad Española de Estrabología y Oftalmología Pediátrica, 2019).

Estas aplicaciones de la ciencia de datos a la medicina me parecen muy interesantes y pienso que pueden ayudar mucho a solucionar problemas graves que afectan a la humanidad, como son las enfermedades.

Con mi participación en este trabajo espero poder aprender nuevos métodos de minería de datos aplicados a la medicina y a conocer nuevos campos de ésta en los que poder aplicar mi conocimiento.

1.2 Objetivos del Trabajo

A continuación se detallarán los objetivos principales y secundarios del presente proyecto.

1.2.1 Objetivos principales

- Obtener, mediante el uso de técnicas de data mining y herramientas computacionales, una comparación de la firma genética del cáncer de mama y la leucemia en base a los mecanismos de DDR.
- Hallar evidencias que permitan establecer nuevos marcadores terapéuticos para la detección precoz y el desarrollo de nuevos tratamientos adecuados para el cáncer de mama.

1.2.2 Objetivos secundarios

- Investigar acerca de los biomarcadores existentes en el cáncer de mama y en la leucemia y sobre el funcionamiento de la DDR.
- Analizar las aplicaciones actuales del data mining en el tratamiento y prevención del cáncer.
- Crear controles de calidad de los datos obtenidos por open-access sobre los dos tipos de cáncer mencionados, pudiendo aplicar técnicas de limpieza y preprocesamiento de los mismos.
- Analizar utilizando algoritmos de data mining la expresión génica de la DDR del cáncer de mama y de la leucemia y poder correlacionarlas y compararlas.

- Comprender y utilizar el método computacional GSEA para la creación de firmas genéticas.
- Estudiar, evaluar y validar los resultados obtenidos tras la utilización de GSEA.
- Poder exponer conclusiones acerca del estudio que permitan encontrar nuevos tratamientos o marcadores terapéuticos.

1.3 Enfoque y método seguido

La metodología que se utilizará en este trabajo comprenderá las siguientes fases:

- Investigación y estudio del estado del arte: en esta primera fase se investigará acerca de los principales conceptos incluidos en este trabajo (GSEA, DDR, cáncer, biomarcadores, etc.) con el objetivo de ser capaz de comprenderlos y de tener un punto de partida adecuado en nuestra investigación posterior. También se hará un estudio del estado del arte para ver qué trabajos y publicaciones académicas ya han tratado este tema, los resultados que se han obtenido y las líneas de trabajo que hay actualmente abiertas.
- Recolección de información y creación de un conjunto de datos: se deberán obtener datos con los que poder trabajar. Para ello será necesario realizar una búsqueda de fuentes de datos de calidad. Además de su obtención, también será necesario asegurarse de su calidad, veracidad y obtener una descripción de los mismos que nos permita entenderlos.
- Limpieza y preprocesamiento de los datos: los datos obtenidos deberán preprocesarse y limpiarse, eliminando aquellos que no procedan. También se les deberá dar formato y se crearán data frames con los que posteriormente se trabajará.
- Análisis de los datos mediante la aplicación de técnicas de minería de datos: se hará un primer análisis descriptivo de los datos para tener una idea general de la información que manejamos. Posteriormente se hará el análisis por data mining para generar las firmas genéticas utilizando GSEA. Tras esto se crearán, ajustarán y refinarán modelos tantas veces como sea necesario para poder obtener un resultado de calidad.
- Transformación en conocimiento de los resultados obtenidos: se hará un análisis de los modelos creados y se listará el conjunto de resultados obtenidos.
- Evaluación del resultado y extracción de conclusiones: se analizarán los resultados obtenidos y sus posibles aplicaciones clínicas. Para finalizar el trabajo se redactarán y expondrán las conclusiones obtenidas acerca del trabajo realizado.

La metodología escogida suele ser bastante habitual dentro de los proyectos de minería de datos. Esta forma de trabajo permitirá un primer contacto y conocimiento acerca del problema a tratar para tener un punto de partida sólido a la hora de comenzar a modelar los datos. Esta estrategia es la más adecuada para conseguir los objetivos de este trabajo ya que sigue un ciclo que abarca desde la detección del problema hasta su solución. Se ha considerado que es la estrategia mejor a seguir a la hora de afrontar este

proyecto, ya que es necesario desarrollar un análisis nuevo. Esto es debido a que esta investigación nunca antes se ha realizado y no hay ningún punto de partida ni ningún producto existente que se pueda adaptar.

En las fases 3 y 4 de esta metodología se utilizará la tecnología R y todas las librerías necesarias para aplicar los algoritmos que se consideren convenientes. Se ha escogido esta tecnología por el conocimiento y dominio de la misma, la sencillez de su utilización y la buena curva de aprendizaje que tiene. Además, su potencia y su aplicación estadísticas son muy buenas y hacen que sea la herramienta idónea.

Opcionalmente, en las fases de obtención y limpieza de datos también se podrán utilizar scripts de Python para generar los conjuntos de datos.

1.4 Planificación del Trabajo

La planificación del trabajo estará guiada por el calendario de la asignatura y por las distintas PEC que se deberán entregar:

FASE	DESCRIPCIÓN	FECHA INICIAL	FECHA FINAL
1	**Definición y planificación del trabajo final**	**18/09/2019**	**29/09/2019**
1.1	Elección del tema y área del trabajo final	18/09/2019	20/09/2019
1.2	Redacción de la propuesta y planificación del trabajo final	21/09/2019	29/09/2019
2	**Estado del arte o análisis de mercado del proyecto**	**30/09/2019**	**20/10/2019**
	Estudio e investigación acerca de los conceptos abordados en el trabajo	30/09/2019	09/10/2019
	Recopilación de bibliografía, estudios y trabajos académicos relacionados con la temática del trabajo	10/10/2019	20/10/2019
3	**Diseño e implementación del trabajo**	**21/10/2019**	**21/12/2019**
	Búsqueda de fuentes de datos y obtención de la información	21/10/2019	25/10/2019
	Preprocesamiento y limpieza de los datos y creación de los data frames con los que se trabajará	26/10/2019	03/11/2019
	Análisis de los datos mediante técnicas de minería de datos. Creación de modelos	04/11/2019	03/12/2019
	Obtención y refinamiento de resultados	04/12/2019	21/12/2019
4	**Redacción de la memoria**	**22/12/2019**	**08/01/2020**
	Evaluación de los resultados obtenidos y extracción de conclusiones	22/12/2019	28/12/2019
	Redacción de los diferentes apartados de la memoria	29/12/2019	05/01/2020
	Revisión del documento	06/01/2020	08/01/2020

5	**Presentación y defensa del proyecto**	**09/01/2020**	**22/01/2020**
	Realización de una presentación que resuma el trabajo realizado	09/01/2020	14/01/2020
	Defensa pública del proyecto	15/01/2020	22/01/2020

El diagrama de Gantt correspondiente se puede consultar en el Anexo 7.1

1.5 Breve sumario de productos obtenidos

Tras el desarrollo del proyecto se ha conseguido obtener un análisis de los procesos moleculares comunes a la DDR, el cáncer de mama y la MCL, identificando cuáles están enriquecidos positivamente y cuáles negativamente.

1.6 Breve descripción de los otros capítulos de la memoria

La memoria contiene los siguientes capítulos:

- Estado del arte: en este apartado se realizó una investigación sobre estudios y trabajos previos que nos permitiesen conocer las líneas de trabajo actuales. En concreto se ha obtenido información acerca de la DDR, el cáncer de mama, la MCL y la utilización de GSEA.
- Desarrollo e implementación: en este apartado se describe el proceso de obtención y limpieza de los datos, la utilización de GSEA, el análisis de los datos y la validación de los mismos.
- Conclusiones: en este apartado se enumeran las conclusiones obtenidas tras el desarrollo del proyecto y se tratan las líneas de trabajo futuras.
- Glosario: incluye varios términos utilizados a lo largo de la memoria junto a su definición.
- Bibliografía: se incluyen todas las referencias bibliográficas utilizadas durante el desarrollo del proyecto.
- Anexos: se incluye el diagrama de Gantt correspondiente a la planificación del proyecto, el diagrama de flujo de las etapas del desarrollo y el código R correspondiente a la carga de datos y a la aplicación de las técnicas de minería de datos.

2. Estado del arte

El estudio del estado del arte se ha divido en 3 apartados. En un primer apartado se ha investigado acerca de DDR y cáncer para conocer su funcionamiento desde el punto de vista teórico y biológico. Posteriormente se ha buscado información sobre cómo se trabaja actualmente con los mecanismos de DDR en los tratamientos contra el cáncer, así como la aplicación de DDR al cáncer de mama. Para finalizar, se ha incidido en la parte técnica y se han buscado implementaciones que se han hecho con herramientas computacionales y técnicas de data mining para estudiar la DDR en algunos tipos de cánceres.

2.1 DDR y cáncer desde un punto de vista teórico

El ADN de los humanos (y el de cualquier otra especie) recibe decenas de miles de lesiones por día. Si estos daños no se reparan se pueden producir mutaciones del genoma que comprometerían la vida de las células y, en su conjunto, del organismo. El conocimiento de las respuestas al daño del ADN permite mejorar la detección y el manejo de enfermedades. (Jackson & Bartek, 2009).

Una de las características principales del cáncer es la inestabilidad genómica que produce daños en el ADN. Los tratamientos históricos del cáncer, tales como quimioterapia o radioterapia, se basan principalmente en dañar las células cancerígenas afectando a su ADN de tal forma que no se puedan reproducir. El problema de este tipo de tratamientos es que el daño también se produce en las células sanas y se pueden generar efectos secundarios no deseados. Conociendo la DDR específica de los distintos tipos de tumores se pueden desarrollar tratamientos basados en ella que sean más efectivos y menos dañinos (O'Connor, M.J., 2015).

Para eliminar células cancerígenas de manera específica y efectiva mediante terapias que inducen daño al ADN, es importante aprovechar las anormalidades específicas en los mecanismos de DDR que están presentes en las células cancerosas pero no en las células normales. Estas propiedades de las células cancerosas pueden proporcionar biomarcadores para la sensibilización. Por ejemplo, los defectos o la regulación positiva de las vías específicas que reconocen o reparan tipos específicos de daño en el ADN pueden servir como biomarcadores de respuesta favorable o deficiente a las terapias que inducen tales tipos de daño en el ADN. La aplicación más llamativa de esta estrategia es el tratamiento de cánceres deficientes en recombinación homóloga por inhibidores de la poli (ADP-ribosa) polimerasa (Hosoya N. & Miyagawa, K., 2014).

2.2 Utilización de la DDR en la terapia contra el cáncer de mama

Dentro del cáncer de mama, existen tres grupos de receptores que son utilizados a la hora de seleccionar una terapia. Estos receptores son: la hormona femenina estrógeno, la hormona femenina progesterona y una proteína llamada factor de crecimiento epidérmico humano (HER2). Si un cáncer tiene presente alguno de esos receptores, se le llama cáncer de mama con hormonas positivas. En cambio, si no tiene presente ninguno de ellos, se le llama cáncer de mama triple negativo. Este tipo de cáncer es el que tiene peor

pronóstico y es más difícil de tratar (Segaert, P., Lopes, M., Casimiro, S. et al., 2018).

El estudio de la DDR es muy común en los objetivos terapéuticos asociados al cáncer. En concreto, para el cáncer de mama con hormonas positivas se han identificado y propuesto tumores deficientes en XRCC1 como objetivos para terapias combinadas que dañan el ADN e inhiben la DDR. XRCC1 es una proteína de armazón que funciona en la reparación de daños producidos en la base del ADN. En cambio, estos indicadores no están presentes en el cáncer de mama triple negativo, para el cual se necesitan obtener nuevos objetivos terapéuticos y terapias (Lee, K.J., Piett, C.G., Andrews, J.F. et al., 2019).

Como podemos ver, la obtención de biomarcadores es muy importante dentro de los objetivos terapéuticos y de las terapias contra el cáncer. Dentro de ellos, varias investigaciones apuntan a las proteínas ATM, ATR, CHK1 y WEE1. Estas proteínas tienen roles importantes en la respuesta al daño del ADN y como objetivos en la terapia contra el cáncer. En la última década, se han diseñado inhibidores específicos de estas proteínas, y se ha explorado su actividad antineoplásica potencial tanto en estrategias de monoterapia contra tumores con defectos específicos (como el cáncer de mama con hormonas positivas) como en combinación con radioterapia o agentes quimioterapéuticos o dirigidos moleculares (Carrassa, L. & Damia, G., 2017).

2.3 Aplicaciones de la ciencia de datos al estudio de la DDR vinculada al cáncer

La utilización de la ciencia de datos dentro del campo de la medicina es algo que viene siendo común en estos últimos años. Varios de los últimos avances médicos de la década han tenido la ayuda de herramientas computacionales y de algoritmos de data mining. Por ejemplo, en pacientes con cáncer hereditario de mama y ovario, se ha descubierto que la mutación del gen CHK2 (gen que participa en mecanimos de DDR) está relacionada con un alto riesgo de padecer la enfermedad. Gracias a la utilización de herramientas computacionales y de la ciencia de datos se han podido precisar más estas mutaciones e identificar los nsSNP presentes en ellas. Esto ha permitido el desarrollo de medicamentos de precisión para el tratamiento de este tipo de cáncer (Badgujar, N.V., Tarapara, B.V. & Shah, F.D., 2019).

El estudio de las firmas genéticas también es un campo de estudio a la hora de buscar nuevos tratamientos contra el cáncer. La utilización de algoritmos de ciencia de datos para el estudio de las diferentes mutaciones genéticas que se producen en la DDR de algunas células ya ha permitido, en alguna ocasión, identificar genes supresores de tumores o inhibidores de la metástasis. Mediante trabajos de este tipo se han podido proporcionar listas completas de genes candidatos como biomarcadores potenciales para la inestabilidad genómica, nuevos objetivos terapéuticos o predictores de la eficacia de la inmunoterapia (Chae, Y.K., Anker, J.F., Carneiro, B.A et al., 2016).

Las bases de datos de genes y su estudio y análisis son una herramienta más dentro de las investigaciones contra el cáncer y otras enfermedades. En varias investigaciones se ha conseguido correlacionar la sobreexposición de un gen (microRNA-3607) con la baja supervivencia en el

cáncer colorrectal y el cáncer de próstata, consiguiendo de esta forma un objetivo nuevo para la predicción y la terapia de estos tipos de tumores. Estos descubrimientos se han podido hacer gracias al análisis de las firmas genéticas y de las mutaciones en la DDR utilizando técnicas como GSEA, y también el análisis de la ontología genética (GO) o de la Enciclopedia de genes y genomas de Kyoto (KEGG) (Lei, L., Zhao, X., Liu, S. et al., 2019).

La utilización de GSEA para crear perfiles de la DDR ya se ha llevado a cabo en algunos tipos de cánceres, como por ejemplo el de próstata. Un estudio ha permitido crear un perfil de la firma genética de la DDR en este tipo de cáncer, mediante la utilización de GSEA y datos provenientes de 1090 pacientes que se dividieron en un conjunto de test y otro de entrenamiento. El daño en el ADN y el perfil de la vía de reparación revelaron variaciones a nivel del paciente. Los mecanismos de DDR rara vez se ven afectados por la mutación. Una firma de la vía DDR mostró una fuerte relación con los resultados a largo plazo de la supervivencia libre de metástasis. En general, estos resultados pueden ser útiles para la estratificación del riesgo de pacientes con cáncer de próstata y para mejorar su tratamiento (Evans, J.R., Zhao, S.G., Chang, L. et al., 2016).

El análisis de las vías biológicas que están enriquecidas dentro de una lista de genes mediante técnicas de ciencia de datos es otra de las líneas de investigación actuales. Además de la ya mencionada GSEA, existen más herramientas que permiten analizar listas de genes como, por ejemplo, g-Profiler. Esta herramienta accesible como web service o desde su interfaz permite encontrar evidencias estadísticas sobre posibles vías biológicas enriquecidas en una lista de genes. Además de la parte analítica en esta parte de proyectos también es importante la parte visual y la presentación de resultados. Para ello existen herramientas como Cytoscape y su plugin Enrichment Map que permiten visualizar expresiones genéticas, vías biológicas enriquecidas y otros patrones biológicos (Reimand, J., Isserlin, R., Voisin, V. et al., 2019).

3. Desarrollo e implementación

El trabajo se ha desarrollado siguiendo el ciclo de vida de un proyecto de ciencia de datos. Por tanto, ha constado de las siguientes etapas:

3.1. Obtención de los datos:

Para ello se ha utilizado el repositorio Gene Expression Omnibus (GEO). Este repositorio público pertenece al National Center for Biotechnology Information estadounidense. Contiene herramientas para ayudar a los usuarios a consultar y descargar experimentos y perfiles de expresión génica seleccionados (Edgar R., Domrachev M., Lash A.E., 2002).
Se han obtenido tres conjuntos de datos procedentes de este repositorio:

- Conjunto de datos sobre DDR: estos datos se corresponden con la serie GSE25848. La organización encargada de generarlos fue el National Institute of Aging.
  Se recogieron datos de linfocitos B GM02184 ("wild-type", ATM + / +) y GM03332 (AT, ATM - / -) en pacientes con el gen ATM mutado.
  Para cada tipo de linfocito se obtuvieron 6 muestras, 3 de las cuales fueron sometidas a 1 Gy (Gray) de radiación ionizante, dejándose las otras 3 sin tratar. 6h después de recibir la radiación, las muestras tratadas se usaron para inmunoprecipitación de ARN con el anticuerpo HuR. El ARN del material tras la inmunoprecipitación se extrajo y se usó para el análisis de microarrays.
  Este conjunto de datos fue recogido con la plataforma Illumina HumanHT-12 V3.0 expression beadchip.

- Conjunto de datos sobre el cáncer de mama: estos datos se corresponden con la serie GSE48989. La organización encargada de generarlos fue la Universidad Thomas Jefferson.
  Se recogieron datos de células de un tipo de cáncer de mama, MCF-7. Con dichas células se crearon 4 bloques de 10cm que se trataron con siRNA de control y otros 4 bloques del mismo tamaño que se trataron con siRNA de ciclina D1. Además, en cada conjunto de 4 bloques, 2 fueron tratados con estradiol y otros 2 con vehículo.
  El objetivo de esto es ver cómo influye la supresión de la expresión genética por ciclina D1 en el cáncer de mama.
  Este conjunto de datos fue recogido con la plataforma Affymetrix Human Gene 1.0 ST Array.

- Conjunto de datos sobre el linfoma de las células del manto (MCL): estos datos se corresponden con la serie GSE21452. La

organización encargada de generarlos fue el National Cancer Institute de Bethesda (USA).
Se recogieron 64 muestras primarias de MCL de pacientes no tratados previamente. El objetivo es identificar nuevos genes y vías que pueden ser relevantes para la patobiología de MCL.
Este conjunto de datos fue recogido con la plataforma Affymetrix Human Genome U133 Plus 2.0 Array.

3.2. Limpieza y preprocesamiento de datos:

En el proceso de limpieza se cargan los datos en R. Para ello se utiliza la función *fread* que permite leer de ficheros y generar directamente un datatable, siendo esta función y este tipo de dato más rápidos y eficientes que los dataframes y otro tipo de funciones de lectura.

Se crean dos datatables para cada conjunto de datos, uno con los valores obtenidos y otro con sus descripciones.

En cuanto al proceso de limpieza, simplemente fue necesario eliminar varios valores vacíos de los datos correspondientes al DDR. Todo este procedimiento se realizó con el código del fichero cargaDatos.R incluido en el anexo 7.3.

3.3. Creación y preparación de datasets:

Según la documentación de GSEA, necesitaremos 4 tipos de ficheros como entrada para esta herramienta. Estos ficheros deberán estar tabulados y utilizar codificación ASCII (Broad Institute, Inc., Massachusetts Institute of Technology, and Regents of the University of California):

- Datos de expresión: estos datos contienen características, muestras y un valor de expresión para cada característica en cada muestra. Se creará a partir de los datatables procesados en el paso anterior. Deberá estar en formato .res, .gct, .pcl, o .txt.

Para la creación de estos ficheros se optó por utilizar la función fwrite de R y transformar los datatables del apartado anterior a ficheros .txt tabulados según los requisitos de GSEA.

- Etiquetas de fenotipo: estos datos asocian cada muestra a un fenotipo. Se pueden crear manualmente o dejar que GSEA los cree de forma automática. Deberán estar en formato .cls.

En este proyecto se ha optado por utilizar el gen correspondiente a la Ciclina D1 (CCND1) como etiqueta de fenotipo. Esto es debido a que tanto la progresión tumoral del cáncer de mama como de MCL se caracterizan por la sobreexpresión de esta proteína y su unión a determinadas regiones del ADN implicadas en la regulación del DDR. Por tanto, la correlación de la expresión génica con éste gen nos permitirá conocer procesos fundamentales en ambos cánceres y generar biomarcadores o dianas terapéuticas que sirvan para mejorar su terapia.

- Conjuntos de genes: contienen uno o varios conjuntos de genes. Para cada conjunto, incluye su nombre y una lista de las características. Estos datos se pueden exportar del MSigDb (Molecular Signature Database). Esta base de datos recoge 22596 conjuntos de genes que pueden ser utilizados como marcadores para un fenotipo particular. Dichos conjuntos se encuentran divididos en 8 colecciones y varias subcolecciones. (Liberzon, et al., 2015)

Se ha seguido el resumen sobre MSigDb que proporciona GSEA para realizar la selección (Broad Institute, Inc., Massachusetts Institute of Technology, and Regents of the University of California). Finalmente, se ha decidido utilizar las siguientes colecciones de genes por ser considerados relevantes en los procesos tumorales:

O C2 – Curated Genes: se divide en dos subcolecciones: perturbaciones químicas y genéticas (CGP) y vías canónicas (CP). Estos genes están seleccionados de varias fuentes, incluidas las bases de datos de rutas en línea y la literatura biomédica. Muchos conjuntos también son aportados por expertos en dominios individuales.

O C5 – Ontología de Genes: se basan en términos de ontología de genes y sus asociaciones con genes humanos. La colección se divide en tres subcolecciones: función molecular (MF), componente celular (CC) y proceso biológico (BP). Se han omitido los términos GO para categorías muy amplias que producirían conjuntos de genes extremadamente grandes. Los términos GO que produjeron conjuntos de genes con menos de 5 genes también se han omitido.

O C6 – Firmas Oncogénicas: este conjunto contiene firmas de vías celulares que a menudo están desreguladas en el cáncer.

O C7 – Firmas Inmunológicas: se compone de conjuntos de genes que representan tipos de células, estados y perturbaciones dentro del sistema inmunitario. Las firmas se generaron mediante la curación manual de estudios publicados en inmunología humana y de ratón.

- Anotaciones de chip: en estos datos se enumera cada sonda en un chip de ADN y su símbolo del gen HUGO correspondiente. Dependen de la forma en la que se hayan obtenido los genes. Lo habitual es utilizar plataformas de secuenciación, una nueva tecnología que produce un ahorro de tiempo y costos y que permite un rendimiento masivo en la recopilación de datos genómicos. Este nuevo método permite secuenciar en paralelo millones de muestras de ADN y ARN en un gran número de individuos (Stiglic G, Bajgot M, Kokol P., 2010).

Estas anotaciones se pueden obtener descargándolas de GSEA en función de la plataforma de secuenciación utilizada para el conjunto de datos. Para nuestros conjuntos de datos hemos tenido que utilizar los siguientes archivos:

- Illumina HumanHT-12_V3.0.chip para el conjunto de datos sobre DDR.
- HuGene-1_0-st.chip para el conjunto de datos sobre cáncer de mama.
- HG-U133_Plus_2 para el conjunto de datos sobre MCL.

3.4. Creación de perfiles genéticos utilizando GSEA:

Para realizar este proceso se ha utilizado la aplicación de escritorio JavaGSEA Desktop Application, disponible tras un registro gratuito en el apartado de descargas de la web de GSEA.

Los 3 ficheros .txt creados con R en el paso anterior se han cargado en la plataforma utilizando la pestaña “Load Data”.

Posteriormente se ha utilizado la opción “Run GSEA” para comenzar el análisis. Para cada uno de los conjuntos de datos se han seleccionado sus correspondientes etiquetas de fenotipo, conjuntos de genes y anotaciones de chips.

Además, para todos los conjuntos de datos se han seleccionado los siguientes parámetros comunes:

- Número de permutaciones: 1000
- Collapse/Remap dataset to gene symbols: Collapse
- Tipo de permutación: phenotype
- Estadístico de enriquecimiento: weighted
- Métrica de clasificación: Pearson
- Tipo de orden de genes: Real
- Dirección de orden de genes: descendente
- Tamaño máximo de conjunto de genes: 500
- Tamaño mínimo de conjunto de genes: 15

Tras ejecutar el análisis, se han obtenido un conjunto de resultados correspondiente a cada uno de los conjuntos de datos, los cuales están disponibles en el repositorio tfm_gsea de GitHub:

- DDR: https://github.com/gquintairos/tfm_gsea/tree/master/DDR.Gsea.1575483723708
- Cáncer de mama: https://github.com/gquintairos/tfm_gsea/tree/master/CancerDeMama.Gsea.1575389998465
- MCL: https://github.com/gquintairos/tfm_gsea/tree/master/MCL.Gsea.1575495673611

Dentro de los resultados existen 4 valores principales que son los siguientes:

• **ES:** se corresponde con la puntuación de enriquecimiento (Enrichment Score). Este valor refleja el grado en que un conjunto de genes está sobrerrepresentado en la parte superior o inferior de una lista clasificada de genes. Este valor se obtiene recorriendo la lista clasificada de genes, aumentando la suma acumulada cuando un gen está en el conjunto de genes y disminuyéndola cuando no lo está. La magnitud del incremento depende de la correlación del gen con el fenotipo. Su valor final es la desviación máxima de cero encontrada al recorrer la lista. Un ES positivo indica el enriquecimiento del conjunto de genes en la parte superior de la lista clasificada mientras que un ES negativo indica el enriquecimiento del conjunto de genes al final de la lista clasificada.

• **NES:** se corresponde con la puntuación de enriquecimiento normalizada (Normalized Enrichment Score). Es la estadística principal para examinar los resultados de enriquecimiento del conjunto de genes. Al normalizar la puntuación de enriquecimiento, GSEA tiene en cuenta las

diferencias en el tamaño del conjunto de genes y en las correlaciones entre los conjuntos de genes y el conjunto de datos de expresión. Por lo tanto, los valores de NES se pueden usar para comparar los resultados de análisis entre conjuntos de genes. Para calcular este valor, GSEA divide en cada permutación el valor de ES entre la media de ES obtenidos en todas las demás permutaciones del conjunto de datos. Por ello, este valor variará si se modifica el número de permutaciones.

• **FDR:** se corresponde con la tasa de descubrimiento falso (False Discover Rate). Es la probabilidad estimada de que un conjunto de genes con un NES dado represente un hallazgo falso positivo. Por ejemplo, un FDR del 50% indica que es probable que el resultado sea válido la mitad de las veces. El informe de análisis de GSEA destaca los conjuntos de genes de enriquecimiento con un FDR de menos del 25% como los que tienen más probabilidades de generar hipótesis interesantes e impulsar más investigaciones, pero proporciona resultados de análisis para todos los conjuntos de genes analizados. En general, dada la falta de coherencia en la mayoría de los conjuntos de datos de expresión y el número relativamente pequeño de conjuntos de genes que se analizan, es apropiado un límite de FDR del 25%.

• **p-valor nominal:** estima la significancia estadística de la puntuación de enriquecimiento para un solo conjunto de genes. Debido a que el valor p no está ajustado para el tamaño del conjunto de genes y las pruebas de hipótesis múltiples, tiene un valor limitado al comparar conjuntos de genes. En el informe GSEA, un p-valor igual a cero indica un p-valor real de menos de 1 dividido entre el número de permutaciones.

3.5. Aplicación de técnicas de minería de datos:

En primer lugar se han cargado en datatables de R los conjuntos de resultados de enriquecimiento positivo y negativo obtenidos tras el análisis de GSEA. Tras esto se ha realizado un proceso de limpieza en el que nos hemos quedado con las variables que nos interesan según los valores principales explicados en el apartado anterior: nombre del gen, ES, NES, p-valor y FDR.

Tras hacer esto, se han identificado los 10 conjuntos de genes más representativos de cada uno de los conjuntos ordenándolos por su p-valor y FDR, siendo los siguientes:

- <u>Cáncer de mama:</u>
  - O Enriquecimiento positivo:
    1. GO_NUCLEAR_NUCLEOSOME
    2. JACKSON_DNMT1_TARGETS_DN
    3. POMEROY_MEDULLOBLASTOMA_DESMOPLASIC_VS_CLASSIC_UP
    4. GO_CHROMATIN_ORGANIZATION_INVOLVED_IN_REGULATION_OF_TRANSCRIPTION
    5. GSE19888_ADENOSINE_A3R_INH_VS_ACT_WITH_INHIBITOR_PRETREATMENT_IN_MAST_CELL_DN
    6. GSE27859_MACROPHAGE_VS_DC_DN
    7. REACTOME_CHOLESTEROL_BIOSYNTHESIS
    8. SCHMIDT_POR_TARGETS_IN_LIMB_BUD_UP
    9. ATF2_UP.V1_UP

10. GO_PROTEIN_CATABOLIC_PROCESS_IN_THE_VACUOLE

O Enriquecimiento negativo:

1. GO_FILOPODIUM_TIP
2. REACTOME_FGFR1_MUTANT_RECEPTOR_ACTIVATION
3. REACTOME_TP53_REGULATES_TRANSCRIPTION_OF_GENES_INVOLVED_IN_CYTOCHROME_C_RELEASE
4. GSE22342_CD11C_HIGH_VS_LOW_DECIDUAL_MACROPHAGES_UP
5. ACEVEDO_NORMAL_TISSUE_ADJACENT_TO_LIVER_TUMOR_UP
6. REACTOME_REGULATION_OF_TP53_ACTIVITY_THROUGH_METHYLATION
7. GSE40274_FOXP3_VS_FOXP3_AND_HELIOS_TRANSDUCED_ACTIVATED_CD4_TCELL_DN
8. DAZARD_RESPONSE_TO_UV_SCC_UP
9. GSE16266_LPS_VS_HEATSHOCK_AND_LPS_STIM_MEF_DN
10. SCHAEFFER_PROSTATE_DEVELOPMENT_6HR_UP

- MCL:

O Enriquecimiento positivo:

1. GO_ENDOSOME_TO_LYSOSOME_TRANSPORT
2. GO_ARF_GUANYL_NUCLEOTIDE_EXCHANGE_FACTOR_ACTIVITY
3. GO_LYSOSOMAL_TRANSPORT
4. GSE32901_TH17_EMRICHED_VS_TH17_NEG_CD4_TCELL_DN
5. REACTOME_TBC_RABGAPS
6. GO_ACTIVATION_OF_GTPASE_ACTIVITY
7. GO_ARF_PROTEIN_SIGNAL_TRANSDUCTION
8. GO_SNARE_COMPLEX_ASSEMBLY
9. GSE22886_UNSTIM_VS_IL2_STIM_NKCELL_UP
10. GO_SNARE_BINDING

O Enriquecimiento negativo:

1. THILLAINADESAN_ZNF217_TARGETS_UP
2. MARKEY_RB1_CHRONIC_LOF_UP
3. GSE15750_DAY6_VS_DAY10_EFF_CD8_TCELL_UP
4. FLORIO_NEOCORTEX_BASAL_RADIAL_GLIA_DN
5. CHIARADONNA_NEOPLASTIC_TRANSFORMATION_KRAS_UP
6. GSE13547_2H_VS_12_H_ANTI_IGM_STIM_BCELL_UP
7. DANG_MYC_TARGETS_UP
8. VEGF_A_UP.V1_DN
9. DEN_INTERACT_WITH_LCA5
10. GROSS_HYPOXIA_VIA_ELK3_UP

- DDR:

  O Enriquecimiento positivo:

  1. GO_REGULATION_OF_COLLAGEN_BIOSYNTHETIC_PROCESS
  2. REACTOME_PEPTIDE_LIGAND_BINDING_RECEPTORS
  3. GO_ESTABLISHMENT_OR_MAINTENANCE_OF_MONOPOLAR_CELL_POLARITY
  4. GO_CELL_CELL_ADHESION_MEDIATED_BY_CADHERIN
  5. REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES
  6. GO_RESPONSE_TO_MURAMYL_DIPEPTIDE
  7. GO_POSITIVE_REGULATION_OF_FAT_CELL_DIFFERENTIATION
  8. VALK_AML_CLUSTER_9
  9. GO_COLLAGEN_BIOSYNTHETIC_PROCESS
  10. SWEET_KRAS_ONCOGENIC_SIGNATURE

  O Enriquecimiento negativo:

  1. GO_PYRIMIDINE_RIBONUCLEOTIDE_BIOSYNTHETIC_PROCESS
  2. GO_CELLULAR_CARBOHYDRATE_CATABOLIC_PROCESS
  3. ZHAN_MULTIPLE_MYELOMA_HP_DN
  4. GO_POLYSACCHARIDE_CATABOLIC_PROCESS
  5. ONDER_CDH1_SIGNALING_VIA_CTNNB1
  6. GO_REGULATION_OF_T_CELL_RECEPTOR_SIGNALING_PATHWAY
  7. GO_GLUCAN_CATABOLIC_PROCESS
  8. GO_RETROGRADE_VESICLE_MEDIATED_TRANSPORT_GOLGI_TO_ENDOPLASMIC_RETICULUM
  9. GO_ER_TO_GOLGI_TRANSPORT_VESICLE_MEMBRANE
  10. GO_UBIQUITIN_LIKE_PROTEIN_CONJUGATING_ENZYME_BINDING

Tras identificar los genes más representativos, hemos escogido los 500 genes más enriquecidos positiva y negativamente de cada uno de los conjuntos para ver cuáles son los procesos que aparecen en común. Realizando este proceso podemos comprobar que en el enriquecimiento positivo tan solo hay una coincidencia y en el negativo ninguna. Por ello, aumentamos el tamaño a 1000 genes y obtenemos los siguientes procesos comunes, cuya descripción se ha obtenido de MSigDB Collections:

- Enriquecimiento positivo: en este caso aparecen 10 procesos comunes tal y como se puede ver en la figura 1.

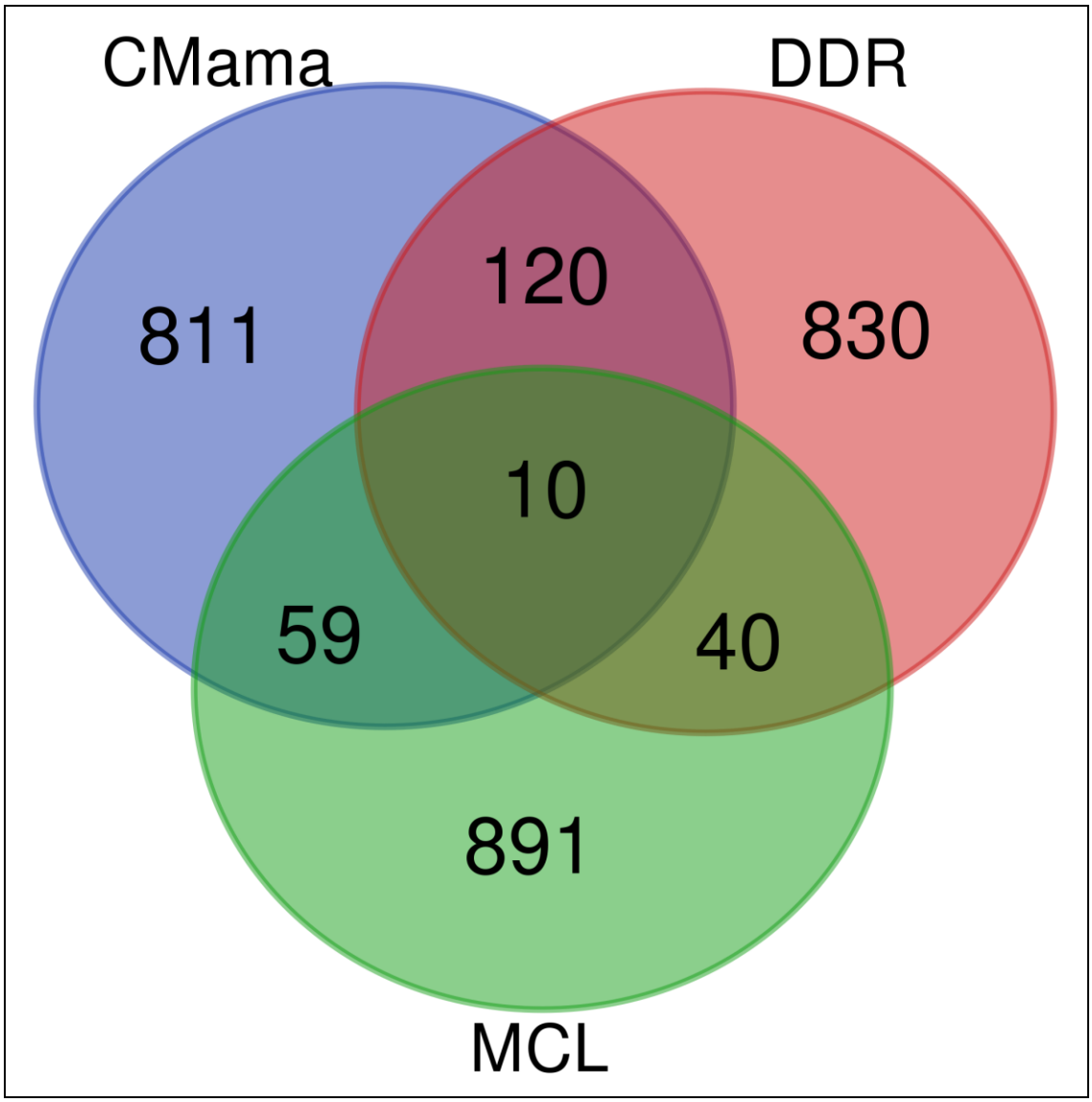


Figura 1: Diagrama de Venn - Enriquecimiento positivo

A continuación mostramos el listado junto con la descripción de dichos procesos:

- **NIKOLSKY_BREAST_CANCER_16Q24_AMPLICON:** este proceso pertenece a la colección C2 de curated gene sets. Está relacionado con muestras genéticas de cáncer de mama. En este estudio se pudo observar que los genes mutados son principalmente reguladores, mientras que los genes enriquecidos están mayormente regulados. Este resultado apoya la hipótesis de que múltiples eventos genéticos, incluyendo ganancias de número de copias y mutaciones somáticas, son necesarios para establecer el fenotipo de células malignas.
- **XU_CREBBP_TARGETS_DN:** pertenece a la colección C2 de curated genes. Son procesos moleculares regulados negativamente en linfocitos pro-B después de la eliminación de CREBBP. Esta proteína está codificada por un gen (CREBBP) cuya traslocación cromosómica está presente en la leucemia mieloide aguda.
- **GSE28726_NAIVE_CD4_TCELL_VS_NAIVE_NKTCELL_DN:** pertenece a la colección C7 de firmas inmunológicas. Son procesos

moleculares regulados negativamente en células T naive. Se obtuvieron tras un análisis de microarrays para determinar los perfiles transcripcionales de las células T NKT, CD1d-aGC+Va24- y CD4.

O GSE22886_NAIVE_CD4_TCELL_VS_NKCELL_UP: pertenece a la colección C7 de firmas inmunológicas. Son procesos moleculares regulados al alza en comparación de las células T CD4 naive con las células NK no estimuladas. Este proceso está presente en la expresión específica de células inmunes. Esto es una indicación de la importancia del papel de un gen en la respuesta inmune.

O HUTTMANN_B_CLL_POOR_SURVIVAL_UP: pertenece a la colección C2 de curated genes. Son procesos moleculares enriquecidos positivamente en pacientes con B-CLL (leucemia crónica de células B) que expresan altos niveles de las proteínas ZAP70 y CD38. Están asociados con una supervivencia deficiente y permite clasificar a los pacientes en buen o mal pronóstico.

O BARRIER_CANCER_RELAPSE_NORMAL_SAMPLE_DN: pertenece a la colección C2 de curated genes. Son procesos moleculares regulados negativamente en muestras de mucosa no neoplásica de pacientes con cáncer de colon que desarrollaron recurrencia de la enfermedad.

O GO_MAMMARY_GLAND_EPITHELIUM_DEVELOPMENT: pertenece a la colección C5 de ontología de genes. Este proceso molecular está presente en el mecanismo cuyo resultado específico es la progresión del epitelio de la glándula mamaria a lo largo del tiempo, desde su formación hasta la estructura madura.

O GO_NEGATIVE_REGULATION_OF_GENE_SILENCING: pertenece a la colección C5 de ontología de genes. Este proceso molecular está presente en aquellos mecanismos que disminuyen la tasa, frecuencia o extensión del silenciamiento génico, el proceso transcripcional o postranscripcional llevado a cabo a nivel celular que resulta en la inactivación génica a largo plazo.

O GO_DENDRITIC_CELL_DIFFERENTIATION: pertenece a la colección C5 de ontología de genes. Aparece en el proceso en el que un tipo de célula precursora adquiere las características especializadas de una célula dendrítica.

O BARRIER_COLON_CANCER_RECURRENCE_DN: pertenece a la colección C2 de curated genes. Se corresponde con un conjunto de procesos moleculares regulados negativamente en el predictor de pronóstico de 70 genes para el cáncer de colon en etapa 2. Están basados en perfiles de expresión génica de la mucosa no neoplásica.

- Enriquecimiento negativo: en este caso aparecen 6 procesos comunes. El resultado gráfico puede verse en la figura 2.

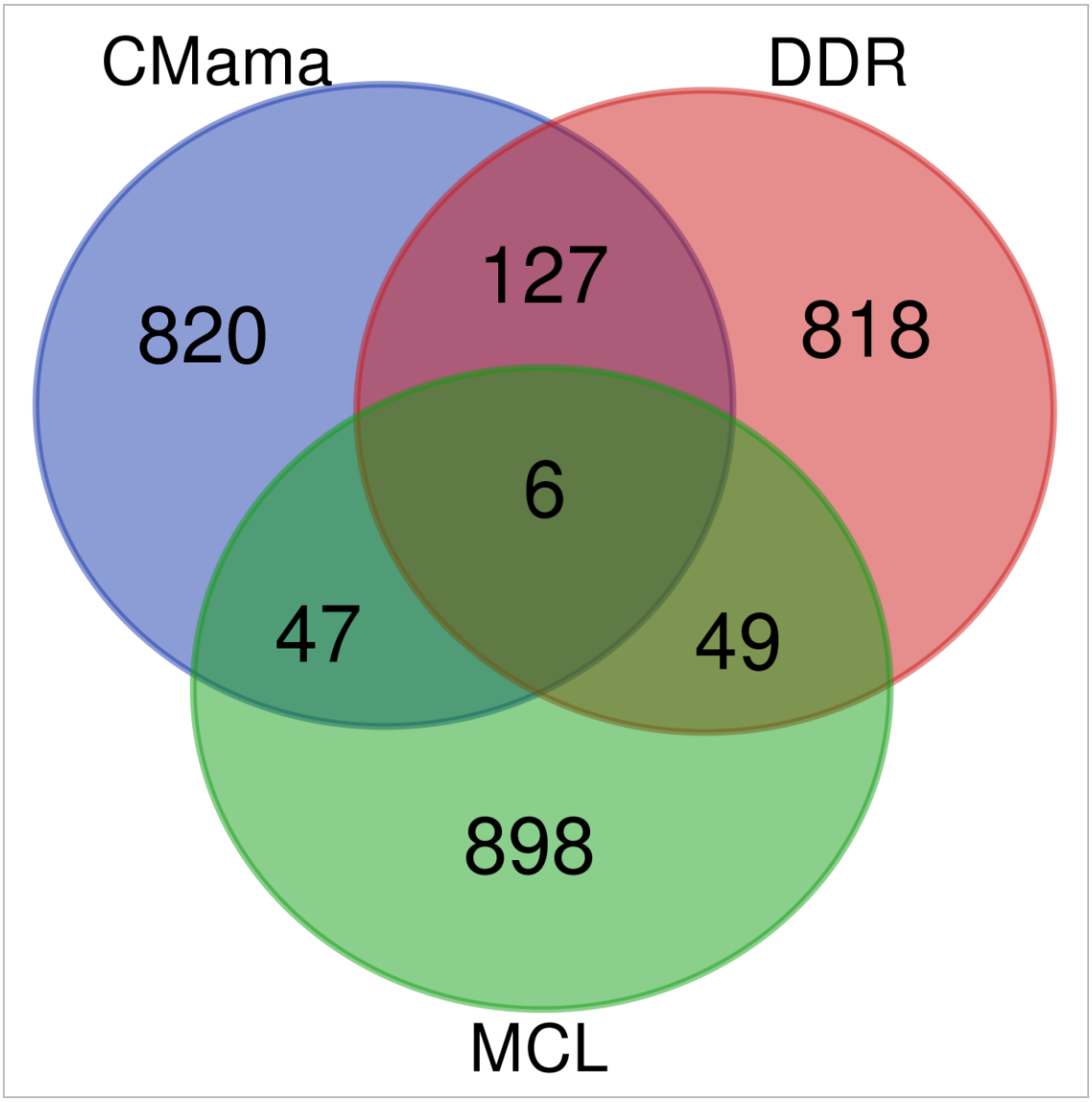


Figura 2: Diagrama de Venn - Enriquecimiento negativo

El listado junto con la descripción de dichos procesos es el siguiente:

- **GSE22886_NAIVE_BCELL_VS_BLOOD_PLASMA_CELL_DN:** pertenece a la colección C7 de firmas inmunológicas. Son procesos moleculares enriquecidos negativamente en comparación de células B naive con células de plasma sanguíneo. La expresión específica de células inmunes es una indicación de la importancia del papel de un gen en la respuesta inmune.
- **MOREAUX_B_LYMPHOCYTE_MATURATION_BY_TACI_DN:** pertenece a la colección C2 de curated genes. Son procesos moleculares enriquecidos negativamente en células plasmáticas de médula ósea (BMPC) normales en comparación con plasmablastos policlonales (PPC) que también distinguieron muestras de mieloma múltiple (MM) por expresión de niveles de la proteína TACI. Se ha detectado que los pacientes con niveles bajos de dicha proteína suelen estar asociados con un mal pronóstico.
- **MOREAUX_MULTIPLE_MYELOMA_BY_TACI_DN**: pertenece a la colección C2 de curated genes. Son procesos moleculares enriquecidos

negativamente que están presentes en muestras de mieloma múltiple (MM) con menor expresión de la proteína TACI. Al igual que se describió en el caso anterior, los pacientes con niveles bajos de dicha proteína suelen estar asociados con un mal pronóstico.

O **GSE14908_RESTING_VS_HDM_STIM_CD4_TCELL_ATOPIC_PATIENT_UP:** pertenece a la colección C7 de firmas inmunológicas. Son procesos moleculares regulados en células T CD4 de pacientes en reposo y en pacientes estimulados con alérgenos (ácaros del polvo doméstico). Se ha detectado que las respuestas de células T CD4 sustentan la enfermedad atópica humana.

O SEIDEN_ONCOGENESIS_BY_MET: pertenece a la colección C2 de curated genes. Son procesos moleculares modificados en tumores de xenoinjerto formados por células DLD-1 o DKO-4 correspondientes al cáncer de colon que sobreexpresan el gen MET. Este gen está presente en este tipo de tumor, así como ocasionalmente en el cáncer de hígado, cabeza y cuello.

O GSE12845_IGD_POS_VS_NEG_BLOOD_BCELL_DN: pertenece a la colección C7 de firmas inmunológicas. Son procesos moleculares enriquecidos negativamente en comparación de células IgD+B con células IgD-B.

Tras este primer análisis descriptivo y la identificación de procesos comunes entre la DDR y ambos tipos de cáncer, se ha decidido utilizar una técnica de minería de datos para intentar clasificar los métodos de la DDR según si son o no comunes a ambos tipos de cáncer y ser capaces de predecir si uno de sus procesos tiene evidencia estadística de poder estar presente en ellos. Para realizar esto hemos utilizado la técnica de Random Forest.

En primer lugar hemos tenido que añadir una variable adicional al resultado de GSEA correspondiente a enriquecimiento positivo y negativo para DDR. Esta variable adicional será un 1 si el proceso es común a los dos tipos de cáncer y un 0 si no lo es.

En segundo lugar, dividimos nuestros conjuntos de datos en datos de entrenamiento (75%) y datos de test (25%). De esta forma podremos entrenar al algoritmo y posteriormente comprobar su eficacia.

Por último, utilizamos la técnica de Random Forest para predecir dicha variable utilizando los dos conjuntos de datos de entrenamiento que acabamos de preparar en el paso anterior.

Una vez hecho esto, tenemos un objeto de R creado para cada uno de los datasets. Analizando estos objetos podemos ver que la tasa de error estimada para el conjunto con enriquecimiento positivo es del 23,75% y para el conjunto con enriquecimiento negativo es del 44,56%. Estas tasas de error son bastante altas, especialmente en el segundo caso.

El código de R utilizado para realizar este apartado se puede observar en el anexo 7.4 de este documento.

3.6. Validación

Tras haber aplicado la técnica de minería de datos, realizamos algunas predicciones con el conjunto de test para ver la eficacia del algoritmo. También calculamos la matriz de confusión para el enriquecimiento positivo y negativo.

Con la matriz de confusión calculada podemos ver que el nivel de eficacia para las predicciones con enriquecimiento positivo es del 75,67% y para el enriquecimiento negativo tan solo del 55,68%. Estos valores se asemejan bastante a la tasa de error que obtuvimos en el paso anterior.

Para intentar mejorarlo realizamos un ajuste de los parámetros del random forest utilizando la función tuneRF. Esta función se encarga de evaluar distintos valores de los parámetros ntree (número de árboles) y mtry (número de variables que se seleccionan aleatoriamente como candidatas en cada iteración). Tras realizar este ajuste y volver a aplicar la técnica de minería de datos obtenemos una pequeña mejora que nos da una eficacia del 78,28% para el enriquecimiento positivo y del 57,1% para el negativo.

Por tanto, hemos obtenido un predictor aceptable para el enriquecimiento positivo pero uno malo para el negativo.

Las posibles causas para que la técnica de minería de datos no haya dado unos resultados muy exactos pueden estar en:

- Tamaño muestral insuficiente.
- Poca cantidad de elementos comunes entre los 3 conjuntos de datos. Este desbalanceo en la variable dependiente puede provocar sobreentrenamiento. Una forma de evitarlo es utilizar muestreo estratificado. Realizamos este ajuste en los parámetros y en el caso del enriquecimiento positivo todo sigue igual, mientras que en el negativo conseguimos mejorar la eficacia al 58,63%, lo cual todavía es insuficiente.

4. Conclusiones

Este proyecto se ha llevado a cabo siguiendo la metodología propia de una investigación de análisis de datos: análisis del problema, investigación del estado del arte, desarrollo de una solución, obtención de los datos, limpieza y preprocesado, aplicación de técnicas de minería de datos, validación y extracción de conclusiones.

Además se han usado conceptos aprendidos en varias asignaturas durante el máster como la utilización de R vista en Estadística Avanzada y Minería de Datos o la planificación y metodología del proceso vista en Fundamentos de la Ciencia de Datos.

Se ha investigado y aprendido a utilizar la herramienta GSEA para realizar los perfiles genéticos de los conjuntos de genes de DDR, MCL y cáncer de mama. Gracias a esto también se han podido conocer procesos genéticos importantes en varias enfermedades. Además, se han obtenido conocimientos nuevos de bioinformática y de genética.

Hemos podido detectar varios procesos y genes que son comunes tanto en la DDR como en los dos tipos de cáncer. Es importante encontrar este tipo de similitudes ya que nos permitirán conocer mejor su funcionamiento y permitirán conocer posibles terapias.

Dentro de los procesos comunes que están enriquecidos positivamente se podría estudiar su inhibición como un posible tratamiento oncológico o la potenciación de sus procesos antagónicos que estarán enriquecidos negativamente. En cambio, los procesos que están enriquecidos negativamente deberían potenciarse.

4.1. Líneas de trabajo futuras

Este proyecto permite líneas futuras de investigación, tales como:

- Repetir el mismo proceso pero con otros dos tipos de cáncer diferentes.
- Utilizar otras técnicas de minería de datos que tengan una mayor eficacia.
- Utilizar los resultados obtenidos para investigar acerca de los procesos comunes en los dos tipos de cáncer y la DDR para evaluar su importancia.
- Investigar acerca de nuevas terapias que inhiban los procesos comunes que están enriquecidos positivamente o potencien los que están enriquecidos negativamente.

5. Glosario

ATM: ataxia-telangiectasia mutada. Gen localizado en el brazo largo del cromosoma 11 humano, entre las posiciones 22 y 23 (11a22-23), codifica la proteína ATM serina/treonina quinasa que está implicada en la regulación de los procesos de control de la división celular y en la reparación de daños sufridos por la molécula de ADN.

Célula dendrítica: es un leucocito de linaje dendrítico especializado en la captación, el procesamiento y el transporte de antígenos a los ganglios linfáticos con el fin de estimular una respuesta inmune a través de la activación de las células.

CREBBP: CREB-Binding Protein. En castellano, proteína de unión a CREB. Es una proteína codificada en humanos por el gen CREBBP. Mutaciones en este gen son las causantes del síndrome de Rubinstein-Taybi (RTS). Se han asociado diversas traslocaciones cromosómicas de este gen con la leucemia mieloide aguda.

DDR: DNA Damage Response (en castellano, respuesta al daño al ADN). Conjunto de procesos por los cuales una célula identifica y corrige daños hechos a las moléculas de ADN que codifican el genoma.

GO: Gene Ontology (en castellano, ontología genética). Sistema para la clasificación jerárquica de genes o genoma en una estructura de grafo llamada ontología. Basándose en esta clasificación, existen técnicas que permiten realizar test estadísticos sobre conjuntos de genes.

Gray: unidad derivada de la dosis de radiación ionizante en el Sistema Internacional de Unidades. Se define como la absorción de un julio de energía de radiación por kilogramo de materia.

GSEA: Gene Set Enrichment Analysis (en castellano, anàlisis de enriquecimiento de conjuntos de genes). Método computacional que determina cuando, en un conjunto de genes definido a priori, existe significación estadística de que hay diferencias concordantes entre dos estados biológicos.

HUGO: Human Genome Organisation. Organización que tiene la misión de aprobar un nombre único y con sentido para cada uno de los genes humanos conocidos basándose en consultas a expertos.

KEGG: Kyoto Encyclopedia of Genes and Genomes (en castellano, Enciclopedia de genes y genomas de Kyoto). Es una colección de bases de datos en línea de genomas, rutas enzimáticas, y químicos biológicos. Esta información puede ser utilizada para la modelización y la simulación, la navegación y extracción de datos.

MCL: Mantle Cell Lymphoma (en castellano, linfoma de las células del manto). Es un tipo de leucemia de muy mal pronóstico. Se caracteriza por la sobreexpresión de la ciclina D1 y la unión de ésta proteína a determinadas regiones del ADN implicadas en la regulación del DDR.

nsSNP: Non-Synonymous Single Nucleotide Polymorphisms (en castellano, polimorfismo de un solo nucleótido no sinónimo). Es una variación en la secuencia de ADN que afecta a una sola base (adenina (A), timina (T), citosina (C) o guanina (G)) de una secuencia del genoma.

Vía Biológica: serie de acciones entre las moléculas en una célula que genera un cierto producto o un cambio. Una vía de tal índole puede activar el ensamblaje de nuevas moléculas, tal como una grasa o una proteína. Las vías también pueden activar y desactivar genes o estimular a una célula para moverse.

6. Bibliografía

[1] Sociedad Española de Estrabología y Oftalmología Pediátrica, XXVII Congreso de la Sociedad Española de Estrabología y Oftalmología Pediátrica. Disponible en: https://www.estrabologia.org/xxvii-congreso-2019/

[2] Jackson, S. & Bartek, J. (2009). "The DNA-damage response in human biology and disease". Nature, 461, páginas 1071–1078 (2009).

[3] O'Connor, M.J. (2015). "Targeting the DNA Damage Response in Cancer". Molecular Cell, Volumen 60, capítulo 4, páginas 547-560, 19 de noviembre de 2015.

[4] Hosoya, N. & Miyagawa, K. (2014). "Targeting DNA damage response in cancer therapy". Cancer Sci, Volumen 105, capítulo 4, páginas 370-388, 21 de marzo de 2014.

[5] Segaert, P., Lopes, M.B., Casimiro, S. et al. (2019). "Robust identification of target genes and outliers in triple-negative breast cancer data". Statistical Methods in Medical Research, Volumen 28, capítulos 10-11, páginas 3042–3056, 27 de agosto de 2018.

[6] Lee, K.J., Piett, C.G., Andrews, J.F. et al. (2019). "Defective base excision repair in the response to DNA damaging agents in triple negative breast cáncer". PLoS One, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0223725, 9 de octubre de 2019.

[7] Carrassa, L. & Damia, G. (2017). "DNA damage response inhibitors: Mechanisms and potential applications in cancer therapy". Cancer Treatment Reviews, Volumen 60, páginas 139-151, 19 de septiembre de 2017.

[8] Badgujar, N.V., Tarapara, B.V. & Shah, F.D. (2019). "Computational analysis of high-risk SNPs in human CHK2 gene responsible for hereditary breast cancer: A functional and structural impact". PLoS One, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0220711, 9 de agosto de 2019.

[9] Chae, Y.K., Anker, J.F., Carneiro, B.A et al. (2016). "Genomic landscape of DNA repair genes in cáncer". Oncotarget, 7(17), 23312–23321. doi:10.18632/oncotarget.8196, 19 de marzo de 2016.

[10] Lei, L., Zhao, X., Liu, S. et al. (2019). "MicroRNA-3607 inhibits the tumorigenesis of colorectal cancer by targeting DDI2 and regulating the DNA damage repair pathway". APOPTOSIS, Volumen 24, Número: 7-8, Páginas: 662-672, agosto de 2019.

[11] Evans, J.R., Zhao, S.G., Chang, L. et al. (2016). "Patient-Level DNA Damage and Repair Pathway Profiles and Prognosis After Prostatectomy for

High-Risk Prostate Cancer". JAMA ONCOLOGY, Volumen: 2, Número: 4, Páginas: 471-480, abril de 2016.

[12] Reimand, J., Isserlin, R., Voisin, V. et al. (2019) "Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap". Nature Protocols, https://doi.org/10.1038/s41596-018-0103-9, publicado online el 21 de enero de 2019.

[13] Edgar R., Domrachev M., Lash A.E. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". Nucleic Acids Res., volumen 30, capítulo 1, páginas 207 a 210. Publicado el 1 de enero de 2002.

[14] Broad Institute, Inc., Massachusetts Institute of Technology, and Regents of the University of California. "GSEA User Guide". (Online) Disponible en: https://software.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html (consultado el 16 de noviembre de 2019)

[15] Liberzon, et al. (2015) "The Molecular Signatures Database Hallmark Gene Set Collection". Cell Systems, Volumen 1, capítulo 6, páginas 417-425, 23 de diciembre de 2015.

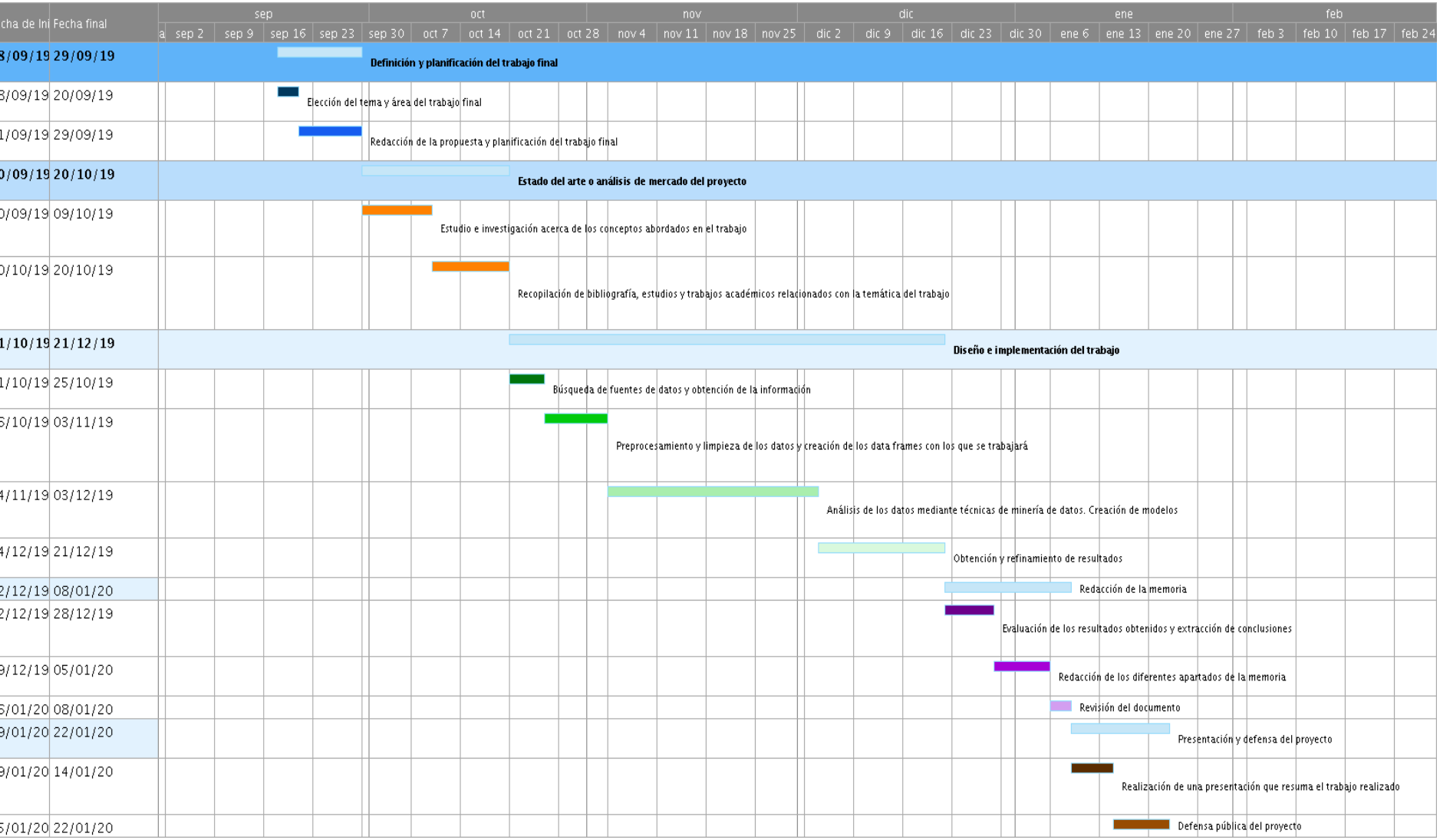
[16] Broad Institute, Inc., Massachusetts Institute of Technology, and Regents of the University of California. "MSigDB Collections: Details and Acknowledgments." Disponible en: http://software.broadinstitute.org/gsea/msigdb (consultado el 4 de diciembre de 2019).

[17] Stiglic G, Bajgot M, Kokol P. (2010) "Gene set enrichment meta-learning analysis: next- generation sequencing versus microarrays". BMC Bioinformatics, volumen 11, 8 de abril de 2010.

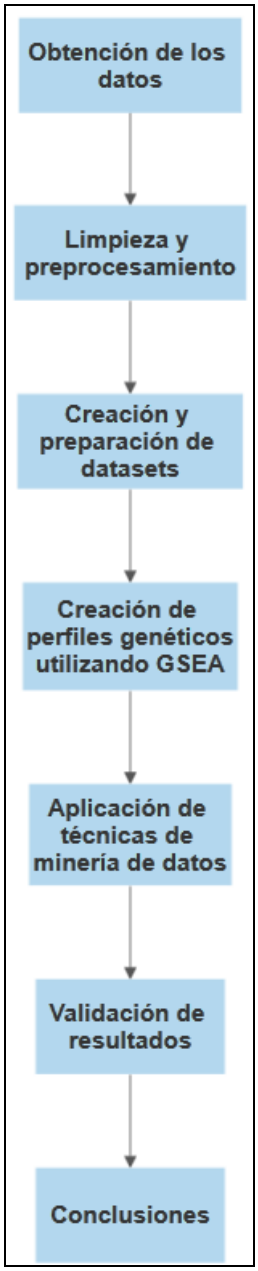
7. Anexos

7.1 Diagrama de Gantt de la planificación

	Nombre de la tarea	Fecha de Ini	Fecha final
1	**Definición y planificación del trabajo final**	**18/09/19**	**29/09/19**
2	Elección del tema y área del trabajo final	18/09/19	20/09/19
3	Redacción de la propuesta y planificación del trabajo final	21/09/19	29/09/19
4	**Estado del arte o análisis de mercado del proyecto**	**30/09/19**	**20/10/19**
5	Estudio e investigación acerca de los conceptos abordados en el trabajo	30/09/19	09/10/19
6	Recopilación de bibliografía, estudios y trabajos académicos relacionados con la temática del trabajo	10/10/19	20/10/19
7	**Diseño e implementación del trabajo**	**21/10/19**	**21/12/19**
8	Búsqueda de fuentes de datos y obtención de la información	21/10/19	25/10/19
9	Preprocesamiento y limpieza de los datos y creación de los data frames con los que se trabajará	26/10/19	03/11/19
10	Análisis de los datos mediante técnicas de minería de datos. Creación de modelos	04/11/19	03/12/19
11	Obtención y refinamiento de resultados	04/12/19	21/12/19
12	**Redacción de la memoria**	22/12/19	08/01/20
13	Evaluación de los resultados obtenidos y extracción de conclusiones	22/12/19	28/12/19
14	Redacción de los diferentes apartados de la memoria	29/12/19	05/01/20
15	Revisión del documento	06/01/20	08/01/20
16	**Presentación y defensa del proyecto**	09/01/20	22/01/20
17	Realización de una presentación que resuma el trabajo realizado	09/01/20	14/01/20
18	Defensa pública del proyecto	15/01/20	22/01/20



7.2. Diagrama de flujo del proyecto



7.3. Carga de datos – Código R

cargaDatos.R – fichero utilizado para cargar los datos y prepararlos para el análisis con GSEA

```
library(data.table) #cargamos la librería para usar data tables
library(tibble) #cargamos la librería para añadir columnas a los data tables
dataDescription_DDR <- fread("data/GPL6947_HumanHT-12_V3_0_R1_11283641_A.bgx", nrows=48803)
data_DDR <- fread("data/GSE25848_series_matrix.txt")
data_DDR <- data_DDR[!is.na(data_DDR$GSM634846)]
data_DDR <- add_column(data_DDR, DESCRIPTION = NA, .after = 1) #añadimos la columna DESCRIPTION con valor NA tal y como indica el manual de uso de GSEA
fwrite(data_DDR, "data/DDR_GSEA.txt", quote=FALSE, sep="\t")

dataDescription_cancerMama <- fread("data/GPL6244-17930.txt", nrows=28869)
data_cancerMama <- fread("data/GSE48989_series_matrix.txt") #cargamos los datos sobre cáncer de mama
data_cancerMama <- data_cancerMama[!is.na(data_cancerMama$GSM1191286)] #eliminamos las filas que tengan valores nulos
data_cancerMama <- add_column(data_cancerMama, DESCRIPTION = NA, .after = 1) #añadimos la columna DESCRIPTION con valor NA tal y como indica el manual de uso de GSEA
fwrite(data_cancerMama, "data/cancerMama_GSEA.txt", quote=FALSE, sep="\t")

dataDescription_MCL <- fread("data/GPL570-55999.txt", nrows=54675)
data_MCL <- fread("data/GSE21452_series_matrix.txt")
data_MCL <- data_MCL[!is.na(data_MCL$GSM536113)]
data_MCL <- add_column(data_MCL, DESCRIPTION = NA, .after = 1) #añadimos la columna DESCRIPTION con valor NA tal y como indica el manual de uso de GSEA
fwrite(data_MCL, "data/MCL_GSEA.txt", quote=FALSE, sep="\t")
```

7.4. Clasificación con Random Forest – Código R:
clasificacion.R – código utilizado para aplicar las técnicas de minería de datos

```
library(data.table)
library(dplyr)
library(caTools)
library(randomForest)
library(textshape)

#carga de los resultados de GSEA
ddr_gseaResult_neg <- fread("data/CCND1_neg_DDR.xls")
ddr_gseaResult_neg <- ddr_gseaResult_neg[,-c(2,3,4,9,10,11,12)]
ddr_gseaResult_pos <- fread("data/CCND1_pos_DDR.xls")
ddr_gseaResult_pos <- ddr_gseaResult_pos[,-c(2,3,4,9,10,11,12)]

mcl_gseaResult_neg <- fread("data/CCND1_neg_MCL.xls")
mcl_gseaResult_neg <- mcl_gseaResult_neg[,-c(2,3,4,9,10,11,12)]
mcl_gseaResult_pos <- fread("data/CCND1_pos_MCL.xls")
mcl_gseaResult_pos <- mcl_gseaResult_pos[,-c(2,3,4,9,10,11,12)]

cmama_gseaResult_neg <- fread("data/CCND1_neg_CancerDeMama.xls")
cmama_gseaResult_neg <- cmama_gseaResult_neg[,-c(2,3,4,9,10,11,12)]
cmama_gseaResult_pos <- fread("data/CCND1_pos_CancerDeMama.xls")
cmama_gseaResult_pos <- cmama_gseaResult_pos[,-c(2,3,4,9,10,11,12)]

#selección del top 500 de NES
ddr_neg_top500NES <- top_n(ddr_gseaResult_neg, 1000, "NES")[1:1000,]
cmama_neg_top500NES <- top_n(cmama_gseaResult_neg, 1000,
"NES")[1:1000,]
mcl_neg_top500NES <- top_n(mcl_gseaResult_neg, 1000, "NES")[1:1000,]
ddr_pos_top500NES <- top_n(ddr_gseaResult_pos, 1000, "NES")[1:1000,]
cmama_pos_top500NES <- top_n(cmama_gseaResult_pos, 1000,
"NES")[1:1000,]
mcl_pos_top500NES <- top_n(mcl_gseaResult_pos, 1000, "NES")[1:1000,]

#preparación de ficheros para detectar elementos comunes
fwrite(list(ddr_neg_top500NES$NAME), file="ddr_neg.txt")
fwrite(list(cmama_neg_top500NES$NAME), file="cmama_neg.txt")
fwrite(list(mcl_neg_top500NES$NAME), file="mcl_neg.txt")
fwrite(list(ddr_pos_top500NES$NAME), file="ddr_pos.txt")
fwrite(list(cmama_pos_top500NES$NAME), file="cmama_pos.txt")
fwrite(list(mcl_pos_top500NES$NAME), file="mcl_pos.txt")

#preparacion para random forest
ddr_gseaResult_pos$COMUN <- 0
for (row in 1:nrow(ddr_gseaResult_pos)) {
  if(ddr_gseaResult_pos[row]$NAME %in% cmama_gseaResult_pos$NAME &&
     ddr_gseaResult_pos[row]$NAME %in% mcl_gseaResult_pos$NAME){
    ddr_gseaResult_pos[row]$COMUN <- 1
  }
```

```
}
ddr_gseaResult_pos <- column_to_rownames(ddr_gseaResult_pos, "NAME")
names(ddr_gseaResult_pos) <- c("ES", "NES", "p", "FDR", "COMUN")
ddr_gseaResult_pos$COMUN <- as.factor(ddr_gseaResult_pos$COMUN)

ddr_gseaResult_neg$COMUN <- 0
for (row in 1:nrow(ddr_gseaResult_neg)) {
  if(ddr_gseaResult_neg[row]$NAME    %in%    cmama_gseaResult_neg$NAME
&&
     ddr_gseaResult_neg[row]$NAME %in% mcl_gseaResult_neg$NAME){
    ddr_gseaResult_neg[row]$COMUN <- 1
  }
}
ddr_gseaResult_neg <- column_to_rownames(ddr_gseaResult_neg, "NAME")
names(ddr_gseaResult_neg) <- c("ES", "NES", "p", "FDR", "COMUN")
ddr_gseaResult_neg$COMUN <- as.factor(ddr_gseaResult_neg$COMUN)

#aplicación de la técnica
sample_pos = sample.split(ddr_gseaResult_pos$COMUN, SplitRatio = .75)
train_pos = subset(ddr_gseaResult_pos, sample_pos == TRUE)
test_pos  = subset(ddr_gseaResult_pos, sample_pos == FALSE)
rf_pos <- randomForest(COMUN ~ ., data=train_pos)
predicciones_pos <- predict(rf_pos, test_pos)
(mc_pos <- with(test_pos,table(predicciones_pos, COMUN)))
100 * sum(diag(mc_pos)) / sum(mc_pos)

sample_neg = sample.split(ddr_gseaResult_neg$COMUN, SplitRatio = .75)
train_neg = subset(ddr_gseaResult_neg, sample_neg == TRUE)
test_neg  = subset(ddr_gseaResult_neg, sample_neg == FALSE)
rf_neg <- randomForest(COMUN ~ ., data=train_neg)
predicciones_neg <- predict(rf_neg, test_neg)
(mc_neg <- with(test_neg,table(predicciones_neg, COMUN)))
100 * sum(diag(mc_neg)) / sum(mc_neg)



#validación
mtry_pos <- tuneRF(ddr_gseaResult_pos[,-5], ddr_gseaResult_pos$COMUN)
sample_pos = sample.split(ddr_gseaResult_pos$COMUN, SplitRatio = .75)
train_pos = subset(ddr_gseaResult_pos, sample_pos == TRUE)
test_pos  = subset(ddr_gseaResult_pos, sample_pos == FALSE)
rf_pos <- randomForest(COMUN ~ ., data=train_pos, mtry=mtry_pos)
predicciones_pos <- predict(rf_pos, test_pos)
(mc_pos <- with(test_pos,table(predicciones_pos, COMUN)))
100 * sum(diag(mc_pos)) / sum(mc_pos)

mtry_neg <- tuneRF(ddr_gseaResult_neg[,-5], ddr_gseaResult_neg$COMUN)
sample_neg = sample.split(ddr_gseaResult_neg$COMUN, SplitRatio = .75)
train_neg = subset(ddr_gseaResult_neg, sample_neg == TRUE)
test_neg  = subset(ddr_gseaResult_neg, sample_neg == FALSE)
rf_neg <- randomForest(COMUN ~ ., data=train_neg, mtry=mtry_neg)
```

```
predicciones_neg <- predict(rf_neg, test_neg)
(mc_neg <- with(test_neg,table(predicciones_neg, COMUN)))
100 * sum(diag(mc_neg)) / sum(mc_neg)

#estratificación del muestreo
mtry_pos <- tuneRF(ddr_gseaResult_pos[,-5], ddr_gseaResult_pos$COMUN)
sample_pos = sample.split(ddr_gseaResult_pos$COMUN, SplitRatio = .75)
train_pos = subset(ddr_gseaResult_pos, sample_pos == TRUE)
test_pos  = subset(ddr_gseaResult_pos, sample_pos == FALSE)
rf_pos <- randomForest(COMUN ~ ., data=train_pos, mtry=mtry_pos, strata =
ddr_gseaResult_pos$COMUN, sampsize=c('0'=25, '1'=75))
predicciones_pos <- predict(rf_pos, test_pos)
(mc_pos <- with(test_pos,table(predicciones_pos, COMUN)))
100 * sum(diag(mc_pos)) / sum(mc_pos)

mtry_neg <- tuneRF(ddr_gseaResult_neg[,-5], ddr_gseaResult_neg$COMUN)
sample_neg = sample.split(ddr_gseaResult_neg$COMUN, SplitRatio = .75)
train_neg = subset(ddr_gseaResult_neg, sample_neg == TRUE)
test_neg  = subset(ddr_gseaResult_neg, sample_neg == FALSE)
rf_neg <- randomForest(COMUN ~ ., data=train_neg, mtry=mtry_neg, strata =
ddr_gseaResult_neg$COMUN, sampsize=c('0'=25, '1'=75))
predicciones_neg <- predict(rf_neg, test_neg)
(mc_neg <- with(test_neg,table(predicciones_neg, COMUN)))
100 * sum(diag(mc_neg)) / sum(mc_neg)
```