Universitat Oberta de Catalunya

uoc.edu

Comparación de la expresión génica regulada por ciclina D1 en Cáncer de mama y Linfoma de las células del manto mediante GSEA

Irene Calvo Cuesta
Máster Universitario en Ciencia de Datos
Minería de datos y machine learning

Director del TFM/Profesor colaborador:
Carles Barceló, PhD

PRA
Jordi Casas Roma

Junio de 2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Comparación de la expresión génica regulada por ciclina D1 en Cáncer de mama y Linfoma de las células del manto mediante GSEA*
Nombre del autor:	*Irene Calvo Cuesta*
Nombre del consultor/a:	*Carles Barceló*
Nombre del PRA:	*Jordi Casas Roma*
Fecha de entrega (mm/aaaa):	06/2019
Titulación:	*Máster Universitario en Ciencia de Datos*
Área del Trabajo Final:	*Minería de datos y machine learning*
Idioma del trabajo:	*Español*
Palabras clave	*Linfoma, Cáncer de Mama, GSEA*

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

Tanto el Cáncer de mama y Linfoma de las células del manto se caracterizan por tener una elevada sobreexpresión de ciclina D1. Por consiguiente, conocer los mecanismos básicos de la función transcripcional de la ciclina D1 es vital para entender por qué sucede su desregulación. Recientes estudios de análisis transcripcional global han generado una gran cantidad de datos. El presente trabajo pretende analizar los data set publicados para Cáncer de mama y Linfoma de las células del manto y buscar así una similitud en la expresión génica de estas neoplasias, para generar una firma genética (*gene signature*) y estudiar su papel como biomarcador que contribuya a la detección precoz de estos tumores.

Esta investigación biomédica va a utilizar la aplicación de escritorio de *Gene Set Enrichment Analysis* (GSEA) para llevar a cabo el análisis y generar la firma genética. Inicialmente el análisis será individual para cada tipo de tumor con el fin de conocer sus particularidades respecto a la sobreexpresión de ciclina D1. A partir de aquí se podrá conocer si dos tumores tan diferentes pueden tener mecanismos comunes de carcinogénesis.

Se han detectado procesos comunes diferencialmente enriquecidos, como la quimiotaxis, que podría ser tratada mediante inmunoterapia. Otros posibles

tratamientos serían la inhibición de los procesos enriquecidos positivamente, tales como la inhibición de la quinasa IRAK4 o el potenciamiento de los procesos significativamente contrarios a la sobreexpresión del gen tales como la activación del EGFR.

Con la validación experimental en laboratorio de los resultados obtenidos se espera un gran retorno social que contribuya de manera positiva en el tratamiento de tumores.

Abstract (in English, 250 words or less):

Both Breast Cancer and Mantle Cell lymphoma are characterized by an overexpression of Cyclin D1. Therefore, knowing the basic mechanisms of the transcriptional function of cyclin D1 is vital to understand why its dysregulation happens. Recent studies of global transcriptional analysis have generated a large amount of data. The present work aims to analyze the published data sets for breast cancer and Mantle Cell Lymphoma and thus seek a similarity in the gene expression of these neoplasms, to generate a genetic signature and study its role as a biomarker that contributes to the early detection of these tumors.

This biomedical research will use the desktop application of Gene Set Enrichment Analysis (GSEA) to carry out the analysis and generate the genetic signature. Initially the analysis will be individual for each type of tumor in order to know its particularities regarding the overexpression of cyclin D1. Then, it will be possible to study whether those two different tumors can have common mechanisms of carcinogenesis.

Differentially enriched common processes have been detected, such as chemotaxis, which could be treated by immunotherapy. Alternative treatments would be the inhibition of other positively enriched processes, such as inhibition of IRAK4 kinase activity or the potentiation of the processes significantly contrary to the overexpression of the gene as the activation of EGFR.

With the experimental validation in the laboratory of the results obtained, a great social return is expected which would positively contribute to the treatment of tumors.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN

1.1. CONTEXTO Y JUSTIFICACIÓN DEL PROYECTO

La ciclina D1 es una proteína encargada de controlar la multiplicación celular, que frecuentemente se encuentra desregulada en procesos tumorales (Qie y Diehl, 2016). Desempeña un papel importante en el control de la proliferación celular, pero se desconoce su función como regulador de la transcripción.

El cáncer de mama es, según la Organización Mundial de la Salud, el más diagnosticado en mujeres y supone la segunda causa de muerte por cáncer en el mundo desarrollado (Li et al., 2017). Su detección precoz es esencial para su tratamiento, y por ello resulta relevante disponer de biomarcadores que permitan su diagnóstico temprano.

Por su parte el Linfoma de las células del manto (*Mantle Cell Lymphoma*, "MCL" en adelante) es una neoplasia linfoide agresiva que, dentro de los linfomas no Hodgkin, cuenta con uno de los peores pronósticos (Klapper, 2011). Se considera una enfermedad difícil de diagnosticar y con una supervivencia mediana de 3 a 5 años (Royo et al., 2012). Por ello su investigación para encontrar biomarcadores que permitan detectarlo se persigue activamente por todo el mundo.

Concretamente los tumores de mama y los linfomas de las células del manto tienen en común la sobreexpresión de la ciclina D1 (Roy y Thompson, 2006; Royo et al., 2012). No se conoce la similitud de la expresión génica regulada por ciclina D1 en MCL respecto a la del Cáncer de mama. Se estudiará su función transcripcional en estos dos procesos tumorales.

Comparar la expresión génica permitiría descubrir nuevos biomarcadores que ayudaran a un diagnóstico precoz, ligado frecuentemente a una mejor tasa de supervivencia. Para interpretar los datos de expresión génica se va a utilizar un poderoso método computacional llamado *Gene Set Enrichment Analysis* (GSEA, en adelante). Este método se basa en analizar conjuntos de genes significativamente enriquecidos para descubrir su comportamiento colectivo en enfermedades (Subramanian et al., 2005).

Se dispone de una enorme cantidad de datos de expresión génica tanto de cáncer de mama como de Linfoma de las células del manto susceptibles de ser analizados mediante data mining. La interpretación de estos datos es fundamental para mejorar el

conocimiento sobre los mecanismos comunes básicos de la célula en estas neoplasias. El conocimiento de estos procesos podrá ayudar a la generación de nuevos fármacos para tratar estos tumores.

1.2. EXPLICACIÓN DE LA MOTIVACIÓN PERSONAL

Contribuir a la investigación en biomedicina es la principal motivación para el desarrollo de este proyecto. Me interesa poder trabajar este tema a través de la inteligencia artificial y la minería de datos, ya que son herramientas básicas en los avances de numerosos sectores. Concretamente en el ámbito de la medicina el uso del machine learning está contribuyendo en gran medida en el campo de la investigación, generando un impacto positivo en la sociedad.

Mis estudios en economía me permitieron conocer que las ramas que me atraían eran las matemáticas y la estadística. Siempre me ha interesado poder dedicarme a la investigación, sobre todo a la de una enfermedad como el cáncer, por el interés que han despertado en mí las ciencias de la salud. Por tanto, después de descubrir la ciencia de datos vi una ventana que me permitía dedicarme a todo ello utilizando mis conocimientos en estadística.

El Máster Universitario de Ciencia de Datos me ha dotado de la base necesaria para realizar este proyecto, al cursar asignaturas de minería de datos, estadística y programación. Personalmente, el proyecto puede aportarme una base en la investigación científica, campo donde quiero proyectar mis conocimientos, y adquirir práctica en el análisis de los datos generados para investigaciones. Concretamente abordar el reto que supone conocer los mecanismos básicos de la célula que pueden conducir a la aparición de tumores me parece muy positivo para mi carrera, complementando mis conocimientos en biología.

1.3. OBJETIVOS DEL PROYECTO

Los objetivos principales fijados para este trabajo son:

- Generar una firma genética a partir de los data sets publicados para cáncer de mama y MCL, en la que se identifiquen los genes significativamente enriquecidos.

- Encontrar similitudes en la expresión génica regulada por ciclina D1 para MCL respecto a cáncer de mama.

- Determinar la función de la firma generada como biomarcador para la detección de estas neoplasias.

1.4. ENFOQUE Y MÉTODO SEGUIDO EN EL DESARROLLO DEL PROYECTO

La inteligencia artificial y la minería de datos son herramientas muy potentes, capaces de analizar enormes cantidades de datos que permiten importantes avances. Concretamente en análisis biológico, y unidas a métodos de enriquecimiento funcional, se han observado muy buenos resultados (Fabris et al., 2019).

En el desarrollo del proyecto se van a analizar dos data sets publicados para cáncer de mama (Casimiro et al., 2013) y MCL (Hartmann et al., 2010), GSE48989 y GSE21452, respectivamente. Para llevar a cabo el análisis y generar la firma genética se va a utilizar la implementación de software para Java del método de Análisis de Enriquecimiento de Conjuntos de Genes (GSEA).

GSEA se presenta como un enfoque prometedor a nivel de microarrays de expresión génica (Subramanian et al., 2005), y presenta una serie de ventajas comparado con los análisis de gen individual (Tamayo et al., 2012), ya que genes que pueden ser descartados en un análisis de enriquecimiento singular aquí pueden ser útiles contribuyendo al enriquecimiento de otros términos. Su principal ventaja es que la selección de genes significativos para el análisis de enriquecimiento funcional no se basa en ningún umbral predefinido y, sobre todo, en la investigación de enfermedades humanas los individuos están sujetos a una variación biológica mucho mayor que la que puede darse en organismos modelo.

Este método persigue determinar si los elementos del conjunto de genes *S*, definido a priori, se distribuyen al azar a lo largo de la lista *L*, o si aparecen principalmente en la parte superior o inferior de *L*. Cada gen recibe una puntuación que permite decidir si estará en la lista final de genes significativos, como se puede ver en la Figura 1 (Subramanian et al., 2005). El método original utiliza como puntaje de enriquecimiento una estadística ponderada similar a Kolmogorov-Smirnov (Fabris et al., 2019), más

adecuada para conjuntos de datos menos coherentes porque necesita menos elementos significativos para obtener una buena puntuación.

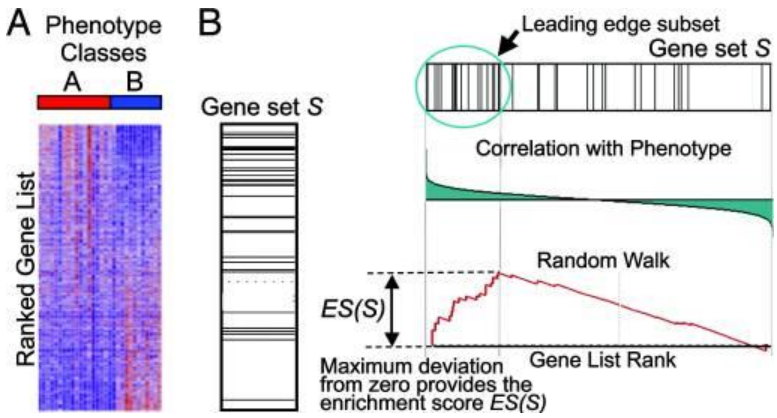

Figura 1. Descripción general de GSEA

En definitiva, GSEA permite descubrir conjuntos de genes que no se conocía que estaban relacionados, y describir la compleja relación entre los cambios de la expresión génica en diferentes condiciones experimentales.

1.5. PLANIFICACIÓN DEL PROYECTO

El plan del proyecto consta de 5 fases con sus correspondientes tareas:

- Primera fase: Definición y planificación

  Desarrollo de la introducción y primeros apartados del trabajo.

- Segunda fase: Estado del arte

  Revisión bibliográfica extensa.

  Desarrollo de éxitos conseguidos y problemas actuales.

- Tercera fase: Diseño e implementación del proyecto

  Selección y preparación de los conjuntos de datos.

  Preparación de los archivos necesarios para la implementación del método GSEA.

Establecimiento de parámetros y desarrollo del proceso seguido.

Ejecución de los análisis de enriquecimiento.

- Cuarta fase: Redacción de la memoria

  Interpretación de los resultados obtenidos.

  Revisión final de la memoria.

- Quinta fase: Presentación y defensa del proyecto

  Elaboración de la presentación.

1.6. BREVE DESCRIPCIÓN DE LOS OTROS CAPÍTULOS DE LA MEMORIA

La memoria está compuesta por los siguientes capítulos:

- Estado del arte: recopilación de estudios previos que permite analizar la evolución que ha existido en la línea de investigación propuesta, y conocer la situación actual de cáncer de mama y MCL, y de la tecnología GSEA.

- Diseño e implementación. Análisis de enriquecimiento de conjuntos de genes: capítulo en el que se aplica el método GSEA a los conjuntos de datos para cáncer de mama (GSE48989) y MCL (GSE21452), y se interpretan sus resultados.

- Conclusiones: capitulo que expone las conclusiones extraídas durante el desarrollo del proyecto.

- Glosario: recopilación de los términos más relevantes utilizados en la memoria con sus correspondientes definiciones.

- Bibliografía: catálogo de las referencias bibliográficas consultadas para la realización del proyecto.

- Anexos: ampliación de los resultados obtenidos durante los análisis de enriquecimiento.

2. ESTADO DEL ARTE

El cáncer es uno de los principales problemas de salud mundial, por ello su investigación se persigue de manera activa. Durante años, las investigaciones en cáncer se basaron en el enfoque de gen individual. Desde hace un tiempo, el uso de microarrays de ADN para identificar cambios en la expresión génica característicos de enfermedades humanas ha permitido obtener resultados exitosos, ofreciendo nuevas formas de mejorar el diagnóstico (Davalos et al., 2017).

En los últimos 20 años el progreso internacional en la investigación de la genómica del cáncer ha sido exponencial (Low et al., 2018). Recientes estudios han demostrado que la inteligencia artificial y el aprendizaje automático juegan un papel muy importante en el análisis de los datos de expresión génica, debido a la creciente complejidad y tamaño de los conjuntos de datos (Bashiri et al., 2017). Su fin es descubrir nuevos biomarcadores basados en los componentes genéticos de los tumores, que contribuyan en el diagnóstico y tratamiento de estas enfermedades. La identificación de patrones genéticos es clave para el análisis molecular de tumores y su relación gen-proteína, gen-enfermedad (Xu et al., 2019)

La inteligencia artificial ha superado a expertos en patología en la detección de varias neoplasias como el cáncer de mama metastásico, y la clasificación de tejidos de los principales subtipos de linfoma no-Hodgkin, a través del procesamiento de imágenes médicas con Deep Learning o algoritmos de clasificación (Ehteshami Bejnordi et al., 2017; do Nascimento et al., 2018).

El auge de las ciencias ómicas y el análisis de la expresión génica han revolucionado la investigación en biomedicina, generando enormes cantidades de datos susceptibles de ser analizados con nuevos métodos estadísticos. El análisis de enriquecimiento funcional de grandes listas de genes o proteínas de interés es esencial para comprender los procesos biológicos en los que se encuentran involucrados. Este tipo de análisis consiste en extraer información biológica significativa de una lista de genes de interés procedente de un experimento, para realizar un análisis estadístico donde se observe que características son relevantes a la hora de explicar dicha lista (Tabas Madrid, 2017).

Para llevar a cabo este análisis existen multitud de metodologías diferentes, y su elección la determinará el tipo de investigación a desarrollar. Normalmente, los

métodos basados en el análisis de enriquecimiento funcional se pueden clasificar en tres categorías: Análisis de enriquecimiento singular (SEA), Análisis de enriquecimiento de conjuntos de genes (GSEA), y Análisis de enriquecimiento modular (MEA) (Fabris et al., 2019). Estos métodos utilizan pruebas estadísticas para descubrir términos biológicos que estén significativamente enriquecidos en una lista de genes (Tabas Madrid, 2017).

El Análisis de Enriquecimiento de Conjuntos de Genes (GSEA) es uno de los métodos estadísticos más utilizados en bioinformática para explicar los procesos biológicos cuya expresión genética está regulada de manera similar (Subramanian et al., 2005). Este método se presenta por primera vez en un análisis de datos de biopsias musculares de pacientes diabéticos frente a controles de personas sanas (Mootha et al., 2003) con el fin de detectar cambios coordinados en la expresión de grupos de genes funcionalmente relacionados. Posteriormente este enfoque se ha aplicado en diversos estudios oncológicos para comparar los modelos de cáncer en ratón con tumores humanos, concretamente en cáncer de pulmón, y comparar en qué medida estos modelos imitan fielmente la enfermedad en humanos (Sweet-Cordero et al., 2005); descubrir procesos genéticos involucrados en el Linfoma difuso de células B grandes, obtener firmas moleculares con más información sobre las células B malignas y compararlo con otras firmas para comprobar si comparten características similares (Monti et al., 2005); identificar cambios biológicos complejos que ocurren durante el desarrollo del carcinoma de células renales metastásico (Khan et al., 2016); encontrar genes clave que regulan y controlan el desarrollo y pronóstico del carcinoma de células escamosas de esófago (He et al., 2018); buscar nuevos biomarcadores en cáncer de mama comparando tejidos cruzados (Li et al., 2017). En este último estudio se identificaron ocho factores de transcripción correlacionados con el cáncer de mama, cuatro de ellos nuevos hasta el momento.

El uso de GSEA también ha permitido avances en otros ámbitos contribuyendo, entre otros, al análisis de organismos vegetales (Yi et al., 2013), al conocimiento de la compleja biología del *Toxoplasma gondii*, parásito causante de la toxoplasmosis (Croken et al., 2014), o al conocimiento del genoma completo del hombre moderno (Akkuratov et al., 2018).

3. DISEÑO E IMPLEMENTACIÓN. ANÁLISIS DE ENRIQUECIMIENTO DE CONJUNTOS DE GENES

Gene Set Enrichment Analysis es compatible con Java, R, y Gene Pattern. En este proyecto, los análisis de enriquecimiento de conjuntos de genes se van a realizar utilizando la implementación *JavaGSEA Desktop Application*, en su versión 3.0 (Figura 2), que utiliza la tecnología Java Web Start, y a parte de implementar el método GSEA proporciona herramientas de procesamiento y métodos de análisis y visualización adicionales. Se encuentra disponible en la página de descargas del sitio web oficial de GSEA (Broad Institute, Inc., Massachusetts Institute of Technology, and Regents of the University of California, 2004-2017).

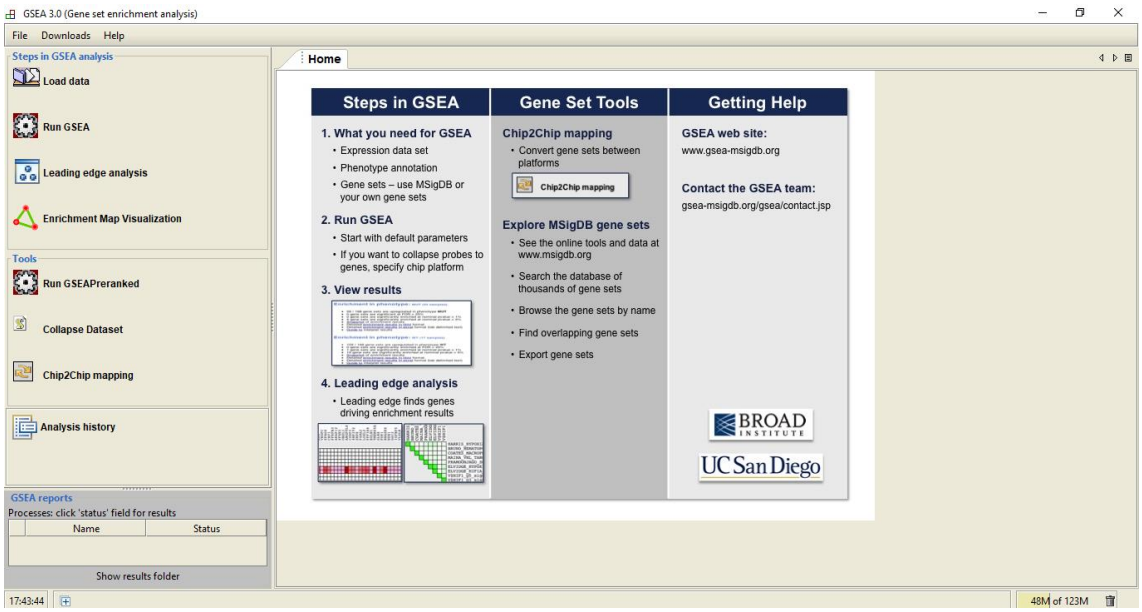

Figura 2. Pantalla principal *JavaGSEA Desktop Application*

Los pasos principales para la implementación del método GSEA podrían recogerse en cuatro fases. En primer lugar, la preparación de los archivos de datos; a continuación, la carga de los datos en la aplicación de escritorio de GSEA; en tercer lugar, el establecimiento de los parámetros para el análisis y su ejecución; por último, la visualización e interpretación de los resultados obtenidos.

GSEA requiere como entrada cuatro archivos para su implementación, todos ellos archivos de texto ASCII delimitados por tabulaciones con un formato específico:

- Conjunto de datos de expresión: contiene los datos de expresión génica de cada tipo de tumor, características, muestras, y un valor de expresión para cada característica en cada muestra. Este archivo debe crearse a partir de los datasets originales. Su formato puede ser .res, .gct, .pcl, o .txt.

- Etiquetas de fenotipo: asocia cada muestra de su conjunto de datos con un fenotipo. Su formato es de tipo .cls, y puede crearse de forma manual u obtenerse de la propia aplicación de GSEA.

- Conjunto de genes: define los conjuntos de genes que se van a analizar, proporcionando su nombre y la lista de características (genes o sondas) de ese conjunto de genes. Se puede generar manualmente o exportarse desde la base de datos de firmas moleculares MSigDB (Liberzon et al., 2011).

- Anotaciones de chip: contiene anotaciones sobre un microarray y asigna identificadores de sonda a los símbolos de los genes. Este archivo también puede generarse de forma manual, o descargarse desde el sitio web de GSEA en función de la plataforma de secuenciación utilizada.

3.1. CONJUNTOS DE DATOS GSE48989 y GSE21452

La creación de los archivos de datos para GSEA parte de los conjuntos de datos originales GSE48989 y GSE21452 seleccionados para este proyecto. Estos conjuntos de datos de expresión génica se obtuvieron del repositorio Gene Expression Omnibus (GEO) perteneciente al Centro Nacional para la Información Biotecnológica (NCBI) (Clough y Barrett, 2016).

GSE48989 recoge datos de ocho placas de células MCF-7, cuatro de ellas tratadas con siRNA de control, y las otras cuatro con siRNA de ciclina D1 (Casimiro et al., 2013). Brevemente, el tratamiento con el siRNA específico permite suprimir la expresión génica de ciclina D1.

GSE21452 comprende datos de sesenta y cuatro muestras primarias de MCL de pacientes que no fueron tratados previamente (Hartmann et al., 2010).

Para recoger los datos se utilizaron las plataformas de secuenciación masiva Affymetrix Human Gene 1.0 ST Array y Affymetrix Human Genome U133 Plus 2.0

Array. Estas plataformas permiten secuenciar en paralelo, de forma rápida, millones de fragmentos de ADN y ARN en un gran número de individuos. El uso de estas técnicas revolucionó la genómica y la biología molecular al ser más rentable que los métodos tradicionales.

En la Tabla 1 se pueden ver algunas características principales de los conjuntos de datos.

Acceso GEO	Organismo	Tipo de Tumor	Tipo Experimento	Centro de Investigación	Plataforma
GSE48989	*Homo Sapiens*	Cáncer de Mama	Perfil de expresión por array	Thomas Jefferson University	Affymetrix Human Gene 1.0 ST Array
GSE21452	*Homo Sapiens*	Linfoma de las células del manto	Perfil de expresión por array	National Cancer Institute	Affymetrix Human Genome U133 Plus 2.0 Array

Tabla 1. Características principales GSE48989 y GSE21452

3.2. PREPARACIÓN Y CARGA DE ARCHIVOS

La creación y elección de los archivos que van a utilizarse durante el análisis es una de las fases más importantes, ya que deben seguir un patrón establecido.

Para la creación de los archivos de conjunto de datos de expresión, en formato .gct, se utiliza el editor Excel, siguiendo la guía oficial disponible en (Broad Institute, Inc., Massachusetts Institute of Technology, and Regents of the University of California, 2004-2017). La primera columna del archivo contiene identificadores de características, la segunda columna una descripción de la característica, y todas las columnas siguientes contienen los valores de expresión para cada característica. También es posible crear estos archivos mediante módulos que convierten los datos de expresión automáticamente en archivos .gct, como *ExpressionFileCreator* o *GEOImporter*.

Los archivos de etiquetas de fenotipo pueden generarse a través del módulo de *ClsFileCreator* de Gene Pattern, disponible en Jupyter Notebook, o se pueden obtener desde la propia aplicación de GSEA en el momento de establecer los parámetros de análisis. En este caso se seleccionan utilizando un gen de los conjuntos de datos como etiqueta de fenotipo. El gen que interesa en ambos casos es CCND1, encargado de codificar la ciclina D1 que se encuentra sobreexpresada tanto en MCL como en cáncer de mama.

Los archivos de conjuntos de genes se exportan directamente de la base de firmas moleculares MSigDB (Liberzon et al., 2011). Esta base de datos recoge 17810 conjuntos de genes, en formato .gmt, divididos en ocho colecciones principales y varias subcolecciones (Broad Institute, Inc., Massachusetts Institute of Technology, and Regents of the University of California, 2004-2017). Se van a utilizar cuatro de estas colecciones conjuntamente, al ser las más relevantes para el estudio por reunir procesos implicados en la progresión tumoral.

- La colección C2 de *curated genes*, que se divide en dos subcolecciones, perturbaciones químicas y genéticas (CGP) y vías canónicas (CP).

- La colección C5 de conjuntos de genes de GO (Ontología de genes) dividida en tres subcolecciones, proceso biológico (BP), componente celular (CC), y función molecular (MF).

- La colección C6 de firmas oncogénicas.

- La colección C7 de firmas inmunológicas (Godec et al., 2016), por la implicación de la respuesta inmune en el cáncer.

Por último, los archivos de anotaciones de chip se descargan desde el sitio web de GSEA, en función de la plataforma de secuenciación utilizada. *HuGene_1_0_st.chip* para el conjunto GSE48989 y *HG_U133_Plus_2.chip* para el conjunto GSE21452.

Una vez preparados todos los data sets necesarios se procede a cargar los archivos en la aplicación de GSEA. Se cargan los dos archivos de conjuntos de datos de expresión (GSE48989_series_matrix.gct y GSE21452_series_matrix.gct) desde la pestaña *Load Data*, y los demás conjuntos se seleccionan directamente al establecer los parámetros del análisis.

3.3. ESTABLECIMIENTO DE PARÁMETROS Y EJECUCIÓN DEL ANÁLISIS

La pestaña Run Gsea de la aplicación de escritorio permite establecer los parámetros con los que se llevará a cabo el análisis de enriquecimiento de conjuntos de genes.

El establecimiento de parámetros para llevar a cabo el análisis referente a cáncer de mama se realiza como muestra la figura 3. Se selecciona el archivo cargado GSE48989_series_matrix.gct como conjunto de datos de expresión; como conjunto de genes se seleccionan conjuntamente las cuatro colecciones c2.all.v6.2.symbols.gmt, c5.all.v6.2.symbols.gmt, c6.all.v6.2.symbols.gmt, c7.all.v6.2.symbols.gmt que contienen los conjuntos de genes, exportados la base de firmas moleculares MSigDB, de las colecciones *curated gene set*, *GO gene set*, *oncogenic signatures*, e *inmunologic signatures*, respectivamente, en su versión 6.2; para las etiquetas de fenotipo se utiliza el gen CCND1, que aparece con el nombre 7942123 al haber utilizado una plataforma de secuenciacion *Affymetrix Human Gene 1.0 ST Array*; como plataforma de chip se selecciona la plataforma de secuenciación HuGene_1_0_st.chip; el último cambio sería modificar la métrica de clasificación (*Metric for ranking genes*) a la opción “Pearson” para correlacionar el resto de genes con la expresión del gen CCND1 y poder obtener un GSEA continuo.

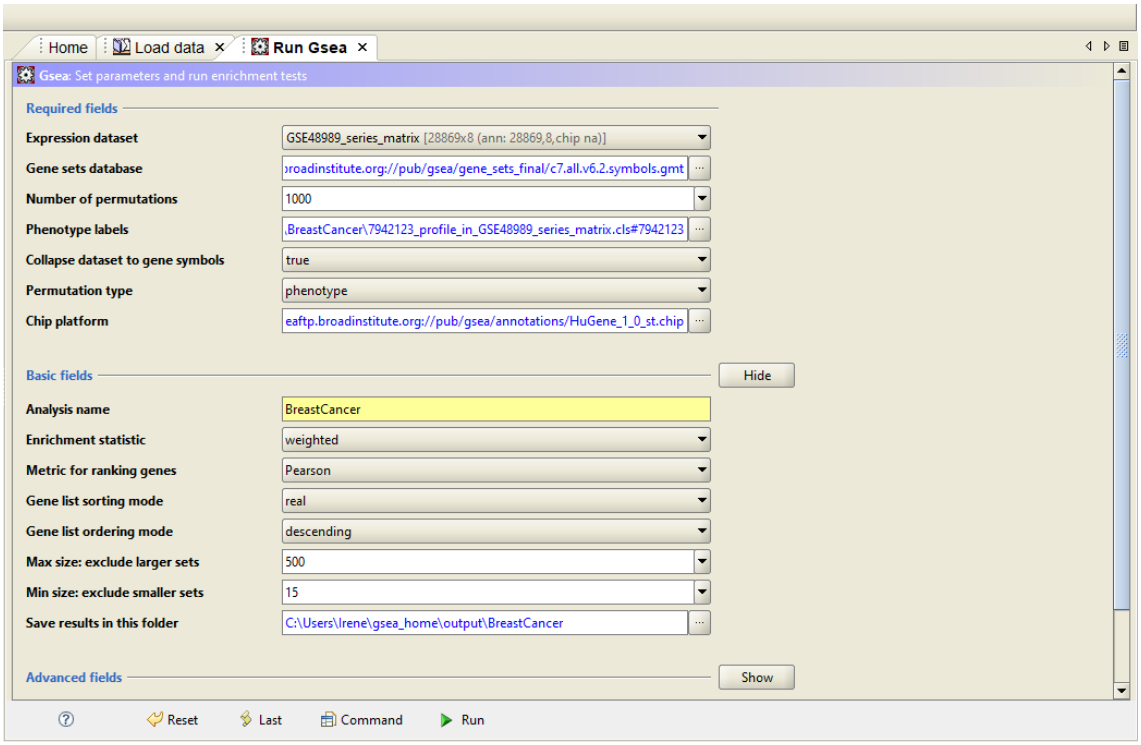

Figura 3. Establecimiento de parámetros GSE48989

Para implementar el análisis de enriquecimiento de conjuntos de genes en MCL se establecen los parámetros como muestra la figura 4. Se selecciona el conjunto cargado de datos de expresión (GSE21452_series_matrix.gct); las colecciones c2.all.v6.2.symbols.gmt, c5.all.v6.2.symbols.gmt, c6.all.v6.2.symbols.gmt, c7.all.v6.2.symbols.gmt con los conjuntos de genes como en el análisis anterior; como etiqueta de fenotipo se utiliza también el gen CCND1 que en este caso, al haber utilizado una plataforma de secuenciación *Affymetrix Human Genome U133 Plus 2.0 Array*, corresponde al 208712_at; en cuando a la plataforma de chip se selecciona HG_U133_Plus_2.chip; por último, también se modifica la métrica de clasificación a Pearson.

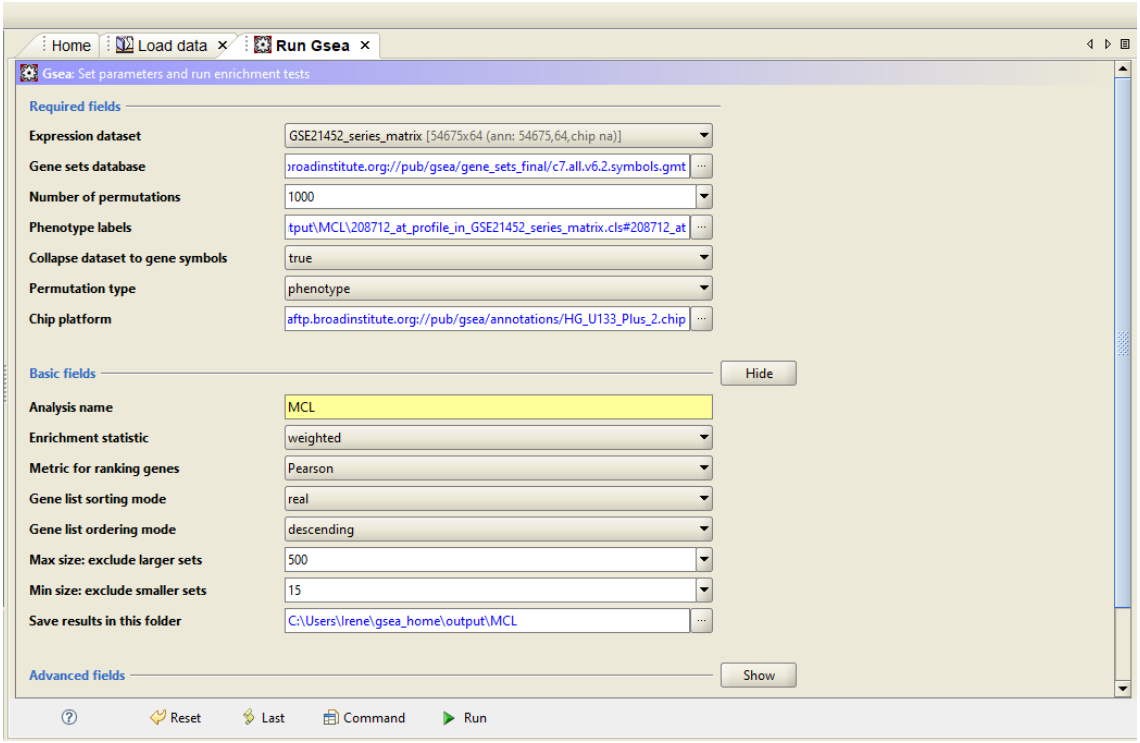

Figura 4. Establecimiento de parámetros GSE21452

Después del establecimiento de parámetros se lleva a cabo la ejecución de ambos análisis de enriquecimiento. Los resultados de esta implementación del método GSEA se pueden consultar y descargar desde el repositorio GSEA de GitHub disponible en la siguiente dirección: https://github.com/icalvocu/GSEA.

3.4. INTERPRETACIÓN DE RESULTADOS

GSEA calcula cuatro estadísticas clave para la interpretación de los resultados.

- La puntuación de enriquecimiento (ES) refleja el grado en que un conjunto de genes está sobrerrepresentado en la parte superior o inferior de la lista de genes clasificada. La puntuación de enriquecimiento se calcula recorriendo la lista de genes clasificada, cuando un gen se encuentra en el conjunto de genes la suma acumulada de la puntuación aumenta, y disminuye cuando se encuentra un gen que no está en el conjunto de genes. La magnitud de incremento depende de la correlación de dicho gen con el fenotipo (Subramanian et al., 2005). Esta puntuación es la desviación máxima de cero encontrada en el recorrido aleatorio de la lista.

- La puntuación de enriquecimiento normalizada (NES) es la estadística principal en el análisis de enriquecimiento. GSEA normaliza automáticamente las puntuaciones de enriquecimiento para tener en cuenta las diferencias en el tamaño de los conjuntos de genes y en las correlaciones entre los conjuntos de genes y el conjunto de datos de expresión. Esta normalización coloca las puntuaciones de enriquecimiento del conjunto de genes en la misma escala, y así permite comparar los resultados del análisis entre conjuntos de genes.

- El valor P nominal es la estimación del nivel de significación de la puntuación de enriquecimiento (ES) para un único conjunto de genes, es decir, la probabilidad de rechazar la hipótesis nula cuando esta es verdadera, conocido como error tipo I o falso positivo.

- La tasa de descubrimiento falso (FDR) es la probabilidad estimada de que un conjunto de genes con una puntuación determinada de enriquecimiento normalizada (NES) represente un resultado falso positivo. GSEA destaca conjuntos de genes con una tasa de descubrimiento falso inferior a 25%, ya que pueden ser los más propensos a generar hipótesis interesantes y por los que conducir la investigación. Es importante evitar que aparezcan conjuntos de genes duplicados durante la implementación del método GSEA, ya que esto puede sesgar la estadística de falsos descubrimientos al basarse esta en todos los conjuntos de genes. Puede ocurrir que conjuntos de genes con distinto nombre contengan los mismos genes identificados, y esto puede ser crítico en los resultados análisis.

Estas estadísticas pueden consultarse en los informes de análisis obtenidos tras la ejecución. El informe para cáncer de mama se puede observar de forma más detallada en el anexo 1, y puede ser descargado mediante el archivo *index.htlm* que se encuentra dentro de la carpeta GSE48989 del repositorio GSEA de GitHub, https://github.com/icalvocu/GSEA/tree/master/GSE48989. El archivo que contiene el informe con los resultados del análisis para MCL, también denominado *index.htlm*, se encuentra dentro de la carpeta GSE21452 del repositorio GSEA de GitHub, https://github.com/icalvocu/GSEA/tree/master/GSE21452, y se puede consultar detalladamente en el anexo 2.

3.4.1. COMPARACIÓN ENTRE CÁNCER DE MAMA Y MCL

La tabla 2 muestra una comparativa de los resultados obtenidos en los GSEA para cáncer de mama contra MCL, indicando cuantos conjuntos de genes están significativamente enriquecidos por cada estadístico (FDR, p-value 1%, y p-value 5%).

Las cuatro colecciones de firmas genéticas, *curated gene set*, *GO gene set*, *oncogenic signatures*, e *inmunologic signatures*, recogen un total de 15740 conjuntos de genes. Se ha comprobado que para cáncer de mama hay 6869 conjuntos de genes con una puntuación de enriquecimiento positiva, y 6292 que muestran un enriquecimiento en la parte inferior de la lista clasificada, de un total de 13161 conjuntos de genes filtrados para el conjunto GSE48989.

De los 13135 conjuntos de genes que han pasado el filtro para MCL, 5263 se encuentran enriquecidos en la parte superior de la lista y 7872 están enriquecidos de manera negativa (tabla 2).

		Cáncer de mama	***MCL***
Enriquecimiento positivo	Genes enriquecidos	6869/13161	5263/13135
	FDR	0	0

	p-valor 1%	18	127
	p-valor 5%	148	435
Enriquecimiento negativo	Genes enriquecidos	6292/13161	7872/13135
	FDR	0	14
	p-valor 1%	14	283
	p-valor 5%	143	1015

Tabla 2. Cáncer de mama Vs. MCL

A continuación, se estudian más en detalle los conjuntos de genes enriquecidos que figuran en la tabla 2, a través de los archivos de enriquecimiento positivo y negativo, en formato Excel, disponibles en el repositorio GSEA de GitHub. *gsea_report_for_7942123_pos_1559558652078.xls* contiene los conjuntos de genes enriquecidos positivamente para cáncer de mama, *gsea_report_for_7942123_neg_1559558652078.xls* recoge los procesos negativos más significativos para cáncer de mama, *gsea_report_for_226623_at_pos_1559557211336.xls* almacena los conjuntos de genes sobrerrepresentados de forma positiva para MCL, y *gsea_report_for_226623_at_neg_1559557211336.xls* contiene los conjuntos de genes enriquecidos de manera negativa en MCL.

Las firmas genéticas de estos cuatro archivos se ordenan de mayor a menor significancia estadística, en función de los estadísticos FDR y p-valor, para conocer cuáles son procesos más significativos de cada conjunto en particular. A modo de ejemplo se seleccionan las 10 primeras firmas genéticas de cada listado, y se muestran en las tablas 3 y 4.

La tabla 3 compara los procesos positivos más significativos en cáncer de mama y MCL.

Cáncer de mama	MCL
GSE360_T_GONDII_VS_B_MALAYI_HIGH_DOSE_DC_UP	GSE17186_MEMORY_VS_CD21HIGH_TRANSITIONAL_BCELL_UP
GO_NEGATIVE_REGULATION_OF_LYMPHOCYTE_DIFFERENTIATION	GSE13306_RA_VS_UNTREATED_TREG_DN
GO_LEUKOTRIENE_METABOLIC_PROCESS	GSE39916_B_CELL_SPLEEN_VS_PLASMA_CELL_BONE_MARROW_UP
GSE40274_CTRL_VS_FOXP3_AND_GATA1_TRANSDUCED_ACTIVATED_CD4_TCELL_UP	GSE19888_CTRL_VS_A3R_ACT_TREATED_MAST_CELL_PRETREATED_WITH_A3R_INH_DN
WIELAND_UP_BY_HBV_INFECTION	GO_REGULATION_OF_IMMUNE_EFFECTOR_PROCESS
GO_RESPONSE_TO_FUNGUS	GO_REGULATION_OF_DEPHOSPHORYLATION
GSE21546_UNSTIM_VS_ANTI_CD3_STIM_ELK1_KO_DP_THYMOCYTES_UP	BONOME_OVARIAN_CANCER_SURVIVAL_SUBOPTIMAL_DEBULKING
GO_MYOBLAST_FUSION	GSE26343_WT_VS_NFAT5_KO_MACROPHAGE_UP
GO_REGULATION_OF_LYMPHOCYTE_MEDIATED_IMMUNITY	GSE10211_UV_INACT_SENDAI_VS_LIVE_SENDAI_VIRUS_TRACHEAL_EPITHELIAL_CELLS_DN
GO_NEGATIVE_REGULATION_OF_IMMUNE_EFFECTOR_PROCESS	GO_PROTEIN_HETEROOLIGOMERIZATION

Tabla 3. Procesos enriquecidos positivamente

La tabla 4 compara, también a modo de ejemplo, las 10 primeras firmas genéticas de cáncer de mama y MCL que están enriquecidas negativamente.

Cáncer de mama	**MCL**
GO_T_CELL_DIFFERENTIATION_INVOLVED_IN_IMMUNE_RESPONSE	GSE10325_CD4_TCELL_VS_BCELL_UP
GO_REGULATION_OF_LYMPHOCYTE_MIGRATION	GSE11057_CD4_EFF_MEM_VS_PBMC_UP
GAUSSMANN_MLL_AF4_FUSION_TARGETS_D_UP	GO_NEGATIVE_REGULATION_OF_MITOTIC_NUCLEAR_DIVISION
POOLA_INVASIVE_BREAST_CANCER_UP	GO_LYMPHOCYTE_CHEMOTAXIS
SEITZ_NEOPLASTIC_TRANSFORMATION_BY_8P_DELETION_UP	GO_REGULATION_OF_ALPHA_BETA_T_CELL_DIFFERENTIATION
GO_CELL_ACTIVATION_INVOLVED_IN_IMMUNE_RESPONSE	GO_ALCOHOL_DEHYDROGENASE_NADP_ACTIVITY
KANG_DOXORUBICIN_RESISTANCE_UP	LEE_LIVER_CANCER_DENA_UP
BIOCARTA_IL12_PATHWAY	PID_IL27_PATHWAY
KONG_E2F3_TARGETS	MASRI_RESISTANCE_TO_TAMOXIFEN_AND_AROMATASE_INHIBITORS_DN
GOLDRATH_EFF_VS_MEMORY_CD8_TCELL_UP	SANA_RESPONSE_TO_IFNG_UP

Tabla 4. Procesos enriquecidos negativamente

3.4.2. PROCESOS COMUNES

Por último, después de conocer cuáles son los procesos positivos y negativos más significativos de cada enfermedad por individual, se van a identificar aquellos procesos con mayor significancia estadística que son comunes en ambos tumores.

Se van a reconocer mediante diagramas de Venn a través del sitio web http://bioinformatics.psb.ugent.be/webtools/Venn/. Este tipo de diagramas permite calcular la intersección entre dos listas de elementos. Para ello se seleccionan los 500 procesos más significativos según el estadístico FDR, que define la tasa de descubrimiento falso, de las listas ordenadas anteriores y se comparan para obtener los procesos comunes tanto positivos como negativos.

Para obtener los procesos comunes de enriquecimiento positivo se introducen en la herramienta del diagrama de Venn las 500 firmas genéticas más enriquecidas positivamente de la lista ordenada para cáncer de mama y las 500 de la lista para MCL. El resultado obtenido es el que muestra la figura 5, donde se comprueba que 28 procesos son comunes en ambos tumores.

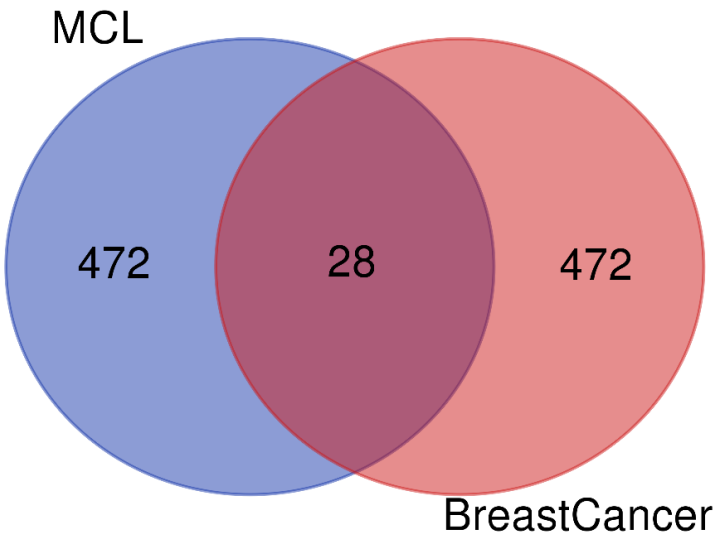

Figura 5. Diagrama de Venn - Enriquecimiento positivo

Además del resultado gráfico, se obtiene una lista que indica que firmas genéticas se encuentran en la intersección. De estas 28 firmas enriquecidas positivamente se seleccionan las 6 más significativas para hacer una breve descripción.

- REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES: se trata de receptores de quimioquinas, también conocidas como citocinas quimiotácticas,

que se encuentran en la superficie de ciertas células. Estos receptores desencadenan un flujo de calcio intracelular que conduce a la quimiotaxis: movimiento de las células a través de un gradiente de concentración de determinadas sustancias Se encuentran altamente expresados en tumores malignos, y actualmente están en el objetivo de la inmunoterapia (Mollica Poeta et al., 2019).

- GO_POSITIVE_REGULATION_OF_OXIDOREDUCTASE_ACTIVITY: es una enzima encargada de catalizar la transferencia de electrones de una molécula a otra. Esta enzima es secretada por células cancerosas e impulsa la progresión del cáncer.

- GSE9037_WT_VS_IRAK4_KO_BMDM_DN: la quinasa 4 asociada al receptor de la interleucina 1 (IRAK4) es una diana a diferentes tipos de cáncer, por ello se ha fomentado el desarrollo de inhibidores de la quinasa IRAK y actualmente fármacos contra el cáncer se centran en ello.

- CAHOY_OLIGODENDROCYTIC: conocidos como oligodendrocitos, son un tipo de célula de la microglía, que forman la capa de mielina del cerebro y la médula espinal. Alterados en tumores que comienzan en el encéfalo o la médula espinal, y mayoritariamente malignos.

- GO_POSITIVE_REGULATION_OF_HEMOPOIESIS: se trata de un proceso de formación, desarrollo y maduración de componentes celulares en la sangre. En cáncer este proceso se desregula produciendo células de forma anormal.

- GO_REGULATION_OF_CHEMOTAXIS: es el fenómeno por el cual el movimiento celular se dirige en respuesta a un gradiente químico extracelular. Los factores que median la quimiotaxis son frecuentemente mutados en cáncer (Roussos et al., 2011). Guarda una estrecha relación con el primer proceso descrito (REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES).

El mismo proceso se repite para las firmas genéticas enriquecidas negativamente. Se insertan las 500 firmas genéticas con mayor enriquecimiento negativo de la lista ordenada para cáncer de mama y las 500 firmas de la lista para MCL. La figura 6 muestra en la intersección los 15 procesos comunes en ambas neoplasias.

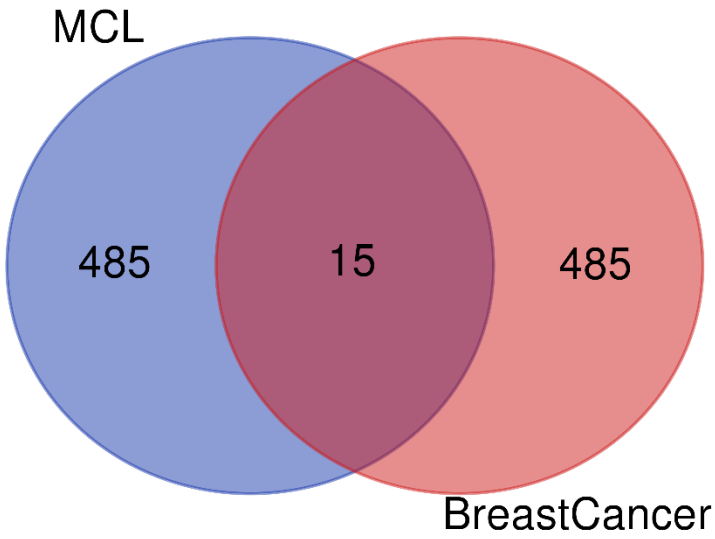

Figura 6. Diagrama de Venn - Enriquecimiento negativo

De los 15 procesos comunes significativamente contrarios a la sobreexpresión del gen CCND1 también se van a seleccionar los 6 más relevantes para llevar a cabo una breve explicación sobre por qué interesa potenciarlos como clave de un mecanismo anti-tumorigénico.

- GO_HIPPO_SIGNALING: controla el tamaño de los órganos mediante la regulación de la proliferación celular y la renovación de las células madre. El hecho de que no ocurra este proceso lleva a la progresión del tumor. Una posible acción terapéutica sería la activación del *Hippo pathway*.

- GINESTIER_BREAST_CANCER_20Q13_AMPLIFICATION_UP: se trata de la banda del cromosoma 20q13, una región de ADN que no se encuentra amplificada en estos procesos tumorales.

- ZWANG_CLASS_2_TRANSIENTLY_INDUCED_BY_EGF: se trata de un factor de crecimiento epidérmico. Las mutaciones que llevan a su infraexpresión se han asociado con un elevado número de cánceres. Su identificación como un oncogén ha llevado al desarrollo de terapias contra el cáncer dirigidas contra el EGFR (receptor del factor de crecimiento epidérmico). Una posible acción terapéutica sería la activación del receptor EGFR con su sustrato EGF.

- BURTON_ADIPOGENESIS_11: proceso múltiple de diferenciación celular mediante el cual los preadipocitos se convierten en adipocitos que requiere la activación secuencial de numerosos factores de transcripción. La influencia de

este proceso en las células tumorales se atribuye a diversos efectos hormonales.

- GO_SEX_CHROMOSOME: son cromosomas particulares que participan en la determinación del sexo de un organismo. En las células de los humanos consisten en un par de cromosomas llamados X e Y. Investigar las diferencias sexuales en cáncer podrá mejorar la terapia para ambos sexos (Arnold y Disteche, 2018).

- COLLIS_PRKDC_SUBSTRATES: subunidad catalítica de la proteína quinasa dependiente de ADN que se encuentra diferencialmente regulada en diferentes tipos de cáncer. En este caso no interesa inhibirlos como terapia, sino dar los sustratos para activar la PRKDC.

4. CONCLUSIONES

Durante el desarrollo de este proyecto se han conocido las etapas imprescindibles para llevar a cabo una investigación preliminar real con resultados inesperados. Asignaturas del Máster de Ciencia de datos como Estadística avanzada, Minería de datos, y Tipología y ciclo de vida de los datos, entre otras, me han permitido trabajar con el entorno GitHub para alojar el proyecto y desarrollar conceptos con más destreza durante el análisis. También se han conocidos nuevos métodos muy utilizados en bioinformática que han permitido alcanzar el nivel de esta investigación, cumpliendo los objetivos establecidos.

El método GSEA se utilizó en cada conjunto de datos (GSE48989 y GSE21452) de manera individual para identificar genes significativamente alterados respecto a la sobreexpresión de ciclina D1, y conocer qué particularidades tiene cada uno de ellos. Se han podido identificar cambios en la expresión génica propios de diferentes enfermedades.

Después del análisis comparativo entre los resultados de cáncer de mama y MCL se ha podido comprobar como dos tumores que parecen tan diferentes, y teniendo en común la sobreexpresión un gen, pueden generar procesos comunes de carcinogénesis. Descubrimiento interesante ya que, si dos tumores diferentes, ambos muy agresivos, tienen mecanismos comunes, se está llegando a la causa que hace que, tanto cáncer de mama como MCL sean tan agresivos.

La inhibición de algunos procesos enriquecidos positivamente podría estudiarse como posible terapia. Uno de los procesos comunes donde centrarse es la quimiotaxis, que consiste en movimiento de las células a través de un gradiente de concentración de determinadas sustancias. Se ha encontrado dos veces de manera independiente y, en relación con la metástasis, se encuentra en el punto de mira de la inmunoterapia (Kakimi et al., 2016; Mollica Poeta et al., 2019). Por su parte, potenciar los procesos comunes enriquecidos negativamente podría ser la clave de un mecanismos anti-tumorigénico.

A partir de aquí sería muy interesante continuar el análisis del resto de procesos tumorales comunes que permita buscar herramientas de detección temprana y promover el desarrollo de nuevos tratamientos, como la inmunoterapia o el silenciamiento génico.

Como líneas futuras de investigación se propone realizar el mismo análisis a través de la aplicación web GEO2R, basada en el lenguaje de programación R (Barrett et al., 2012), para comparar los resultados obtenidos.

También se sugiere la realización de ensayos de laboratorio que permitan comprobar algunos procesos comunes como la quimiotaxis. Para ello pueden utilizarse técnicas cuantitativas y cualitativas que evalúen la actividad quimiotáctica de las células como experimentos en placa-agar, la estimulación temporal de células, ensayos de dos cámaras, o el conteo automatizado de células al microscopio (Liao et al., 2016).

A través de este estudio se ha querido contribuir a una comprensión más profunda de la biología del cáncer, a partir del cual expertos en biomedicina puedan abordar los desafíos pendientes que todavía quedan en torno a la investigación del cáncer.

5. GLOSARIO

Affymetrix: compañía estadounidense especializada en el diseño de micromatrices de ADN. Fue fundada en 19992.

Ciclina D1: proteína, codificada en humanos por el gen CCND1, encargada de la regulación del ciclo celular.

Diagrama de Venn: diagrama que muestra visualmente todas las posibles relaciones lógicas entre una colección de conjuntos.

GEO: Gene Expression Omnibus es un repositorio público de datos genómicos funcionales.

GitHub: plataforma online de desarrollo colaborativo para alojar proyectos utilizando el sistema de control de versiones Git.

GSEA: *Gene Set Enrichment Analysis*, método computacional que determina si un conjunto de genes, definidos a priori, muestra diferencias estadísticamente significativas y concordantes entre dos estados biológicos.

Kolmogorov-Smirnov: prueba estadística, no paramétrica, que determina la bondad de ajuste entre dos distribuciones de probabilidad.

Leucemia: conjunto de procesos tumorales que provocan un aumento descontrolado de leucocitos en la sangre u órganos linfáticos.

MCF-7: línea celular de cáncer de mama que fue aislada en 1970 de una mujer 69 años.

NCBI: Centro Nacional para la Información Biotecnológica.

Neoplasia: Formación anormal, en alguna parte del cuerpo, de un tejido nuevo de carácter tumoral, benigno o maligno.

SiRNA: ARN pequeño de interferencia, es un tipo de ARN interferente.

Transcripción: Proceso por el que se generan las proteínas que controlan todos los procesos celulares.

6. BIBLIOGRAFÍA

Aibar, S., Fontanillo, C., Droste, C., Roson-Burgo, B., Campos-Laborie, F., Hernandez-Rivas, J. and De Las Rivas, J. (2015). Analyse multiple disease subtypes and build associated gene networks using genome-wide expression profiles. *BMC Genomics*, 16(Suppl 5), p.S3.

Akkuratov, E., Gelfand, M. and Khrameeva, E. (2018). Neanderthal and Denisovan ancestry in Papuans: A functional study. *Journal of Bioinformatics and Computational Biology*, 16(02), p.1840011.

Arnold, A. and Disteche, C. (2018). Sexual Inequality in the Cancer Cell. *Cancer Research*, 78(19), pp.5504-5505.

Barrett, T., Wilhite, S., Ledoux, P., Evangelista, C., Kim, I., Tomashevsky, M., Marshall, K. and Phillippy, K. et al. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1), pp.D991-D995.

Bashiri, A., Ghazisaeedi, M., Safdari, R., Shahmoradi, L. and Ehtesham, H. (2017). Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review. *Iranian journal of public health*, 46(2), pp.165-172.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G. and Geessink, O. et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, *318*(22), pp.2199-2210.

Broad Institute, Inc., Massachusetts Institute of Technology, and Regents of the University of California (2004-2017). *GSEA*. [Online]. Available at: http://software.broadinstitute.org/gsea/index.jsp

Casimiro, M., Wang, C., Li, Z., Di Sante, G., Willmart, N., Addya, S., Chen, L., Liu, Y., Lisanti, M. and Pestell, R. (2013). Cyclin D1 Determines Estrogen Signaling in the Mammary Gland In Vivo. *Molecular Endocrinology*, 27(9), pp.1415-1428.

Clot, G., Jares, P., Giné, E., Navarro, A., Royo, C., Pinyol, M., Martín-Garcia, D. and Demajo, S. et al. (2018). A gene signature that distinguishes conventional and leukemic nonnodal mantle cell lymphoma helps predict outcome. *Blood*, 132(4), pp.413-422.

Clough, E. and Barrett, T. (2016). The Gene Expression Omnibus database. *Methods in Molecular Biology, 1418*, pp.93-110.

Croken, M., Qiu, W., White, M. and Kim, K. (2014). Gene Set Enrichment Analysis (GSEA) of Toxoplasma gondii expression datasets links cell cycle progression and the bradyzoite developmental program. *BMC Genomics*, 15(1), p.515.

Davalos, V., Martinez-Cardus, A. and Esteller, M. (2017). The Epigenomic Revolution in Breast Cancer: From Single-Gene to Genome-Wide Next-Generation Approaches. *The American Journal of Pathology*, 187(10), pp.2163-2174.

do Nascimento, M., Martins, A., Azevedo Tosta, T. and Neves, L. (2018). Lymphoma images analysis using morphological and non-morphological descriptors for classification. *Computer Methods and Programs in Biomedicine*, 163, pp.65-77.

Fabris, F., Palmer, D., de Magalhães, J. and Freitas, A. (2019). Comparing enrichment analysis and machine learning for identifying gene properties that discriminate between gene classes. *Briefings in Bioinformatics*, 0(0), pp.1-12.

Godec, J., Tan, Y., Liberzon, A., Tamayo, P., Bhattacharya, S., Butte, A., Mesirov, J. and Haining, W. (2016). Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. *Immunity*, 44(1), pp.194-206.

Guo, L., Liu, S., Jakulin, A., Yilamu, D., Wang, B. and Yan, J. (2015). Positive expression of cyclin D1 is an indicator for the evaluation of the prognosis of breast cancer. *International Journal of Clinical and Experimental Medicine, 8*(10), pp.18656-18664.

Hartmann, E., Campo, E., Wright, G., Lenz, G., Salaverria, I., Jares, P., Xiao, W. and Braziel, R. et al. (2010). Pathway discovery in mantle cell lymphoma by integrated analysis of high-resolution gene expression and copy number profiling. *Blood*, 116(6), pp.953-961.

He, W., Chen, L., Yuan, K., Zhou, Q., Peng, L. and Han, Y. (2018). Gene set enrichment analysis and meta-analysis to identify six key genes regulating and controlling the prognosis of esophageal squamous cell carcinoma. *Journal of Thoracic Disease*, 10(10), pp.5714-5726.

Kakimi, K., Karasaki, T., Matsushita, H. and Sugie, T. (2016). Advances in personalized cancer immunotherapy. *Breast Cancer*, 24(1), pp.16-24.

Khan, M., Dębski, K., Dabrowski, M., Czarnecka, A. and Szczylik, C. (2016). Gene set enrichment analysis and ingenuity pathway analysis of metastatic clear cell renal cell

carcinoma cell line. *American Journal of Physiology-Renal Physiology*, 311(2), pp.F424-F436.

Klapper, W. (2011). Histopathology of Mantle Cell Lymphoma. *Seminars in Hematology*, 48(3), pp.148-154.

Li, W., He, K., Tang, L., Dai, S., Li, G., Lv, W., Guo, Y., An, S., Wu, G., Liu, D. and Huang, J. (2016). Comprehensive tissue-specific gene set enrichment analysis and transcription factor analysis of breast cancer by integrating 14 gene expression datasets. *Oncotarget*, 8(4), pp.6775-6786.

Liao, X., Meena, N., Southall, N., Liu, L., Swaroop, M., Zhang, A., Xiang, J., Parent, C., Zheng, W. and Kimmel, A. (2016). A High-Throughput, Multi-Cell Phenotype Assay for the Identification of Novel Inhibitors of Chemotaxis/Migration. *Scientific Reports*, 6(1).

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P. and Mesirov, J. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12), pp.1739-1740.

Liu, J., Wei, H., Zhu, K., Lai, L., Han, X. and Yang, Y. (2017). Male breast cancer and mantle cell lymphoma in a single patient. *Medicine*, 96(48), p.e8911.

Low, S., Zembutsu, H. and Nakamura, Y. (2017). Breast cancer: The translation of big genomic data to cancer precision medicine. *Cancer Science*, 109(3), pp.497-506.

Mollica Poeta, V., Massara, M., Capucetti, A. and Bonecchi, R. (2019). Chemokines and Chemokine Receptors: New Targets for Cancer Immunotherapy. *Frontiers in Immunology*, 10(379).

Monti, S., Savage, K. and Kutok, J. (2005). Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood*, 105(5), pp.1851-1861.

Mootha, V., Lindgren, C., Eriksson, K., Subramanian, A., Sihag, S., Lehar, J. and Puigserver, P. et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3), pp.267-273.

Ortiz, A., Garcia, D., Vicente, Y., Palka, M., Bellas, C. and Martin, P. (2017). Prognostic significance of cyclin D1 protein expression and gene amplification in invasive breast carcinoma. *PLOS ONE*, 12(11), p.e0188068.

Qie, S. and Diehl, J. (2016). Cyclin D1, cancer progression, and opportunities in cancer treatment. *Journal of Molecular Medicine*, 94(12), pp.1313-1326.

Roussos, E., Condeelis, J. and Patsialou, A. (2011). Chemotaxis in cancer. *Nature Reviews Cancer*, 11(8), pp.573-587.

Roy, P. and Thompson, A. (2006). Cyclin D1 and breast cancer. *The Breast*, 15(6), pp.718-727.

Royo, C., Navarro, A., Clot, G., Salaverria, I., Giné, E., Jares, P., Colomer, D. and Wiestner, A. et al. (2012). Non-nodal type of mantle cell lymphoma is a specific biological and clinical subgroup of the disease. *Leukemia*, 26(8), pp.1895-1898.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. and Mesirov, J. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), pp.15545-15550.

Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J., Ladd-Acosta, C., Mesirov, J., Golub, T. and Jacks, T. (2004). An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nature Genetics*, 37(1), pp.48-55.

Tabas Madrid, D. (2017). *Herramientas eficientes para el análisis masivo de datos ómicos*. Doctorado. Universidad Complutense de Madrid.

Tamayo, P., Steinhardt, G., Liberzon, A. and Mesirov, J. (2012). The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical Methods in Medical Research*, 25(1), pp.472-487.

Xu, J., Yang, P., Xue, S., Sharma, B., Sanchez-Martin, M., Wang, F., Beaty, K., Dehan, E. and Parikh, B. (2019). Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. *Human Genetics*, 138(2), pp.109-124.

Yi, X., Du, Z. and Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Research, 41(W1), pp.W98-W103.

7. ANEXOS

Anexo 1 – Informe GSEA para cáncer de mama

GSEA Report for Dataset GSE48989_series_matrix

Enrichment in phenotype: positive correlation with profile

- 6869 / 13161 gene sets are upregulated in phenotype **7942123_pos**
- 0 gene sets are significant at FDR < 25%
- 18 gene sets are significantly enriched at nominal pvalue < 1%
- 148 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: negative correlation with profile

- 6292 / 13161 gene sets are upregulated in phenotype **7942123_neg**
- 0 gene sets are significantly enriched at FDR < 25%
- 14 gene sets are significantly enriched at nominal pvalue < 1%
- 143 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Dataset details

- The dataset has 28869 native features
- After collapsing features into gene symbols, there are: 20693 genes

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 2579 / 15740 gene sets
- The remaining 13161 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

Gene markers for the neighbors of 7942123_pos

- The dataset has 20693 features (genes)
- Detailed rank ordered gene list for all features in the dataset
- Heat map and gene list correlation profile for all features in the dataset
- Buttefly plot of significant genes

Global statistics and plots

- Plot of p-values *vs.* NES
- Global ES histogram

Other

- Parameters used for this analysis

Comments

- Timestamp used as random seed: 1559558822136

Anexo 2 – Informe GSEA para MCL

GSEA Report for Dataset GSE21452_series_matrix

Enrichment in phenotype: positive correlation with profile

- 5263 / 13135 gene sets are upregulated in phenotype **226623_at_pos**
- 0 gene sets are significant at FDR < 25%
- 127 gene sets are significantly enriched at nominal pvalue < 1%
- 435 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: negative correlation with profile

- 7872 / 13135 gene sets are upregulated in phenotype **226623_at_neg**
- 14 gene sets are significantly enriched at FDR < 25%
- 283 gene sets are significantly enriched at nominal pvalue < 1%
- 1015 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Dataset details

- The dataset has 54675 native features
- After collapsing features into gene symbols, there are: 20606 genes

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 2605 / 15740 gene sets
- The remaining 13135 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

Gene markers for the neighbors of 226623_at_pos

- The dataset has 20606 features (genes)
- Detailed rank ordered gene list for all features in the dataset
- Heat map and gene list correlation profile for all features in the dataset
- Buttefly plot of significant genes

Global statistics and plots

- Plot of p-values *vs.* NES
- Global ES histogram

Other

- Parameters used for this analysis

Comments

- Timestamp used as random seed: 1559557375984
