FLUJOS DE TRABAJO SISTEMÁTICOS PARA LA CARACTERIZACIÓN PROTEÓMICA DE LA LEUCEMIA LINFOIDE CRÓNICA Y SU CONTRAPARTIDA NORMAL

GRADO EN ESTADÍSTICA

Trabajo de Fin de Grado

Autora

Laura Díaz Muñoz

Tutores

Dr. José Manuel Sánchez Santos

Dra. Ángela Patricia Hernández García

D. Alberto Berral González

Salamanca, julio de 2022

CIC
CENTRO DE INVESTIGACIÓN DEL CÁNCER

FLUJOS DE TRABAJO SISTEMÁTICOS PARA LA CARACTERIZACIÓN PROTEÓMICA DE LA LEUCEMIA LINFOIDE CRÓNICA Y SU CONTRAPARTIDA NORMAL

GRADO EN ESTADÍSTICA

Trabajo de Fin de Grado

Autora

Laura Díaz Muñoz

Tutores

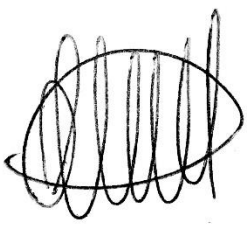
Dr. José Manuel Sánchez Santos

Dra. Ángela Patricia Hernández García

HERNANDEZ GARCIA ANGELA PATRICIA - 70889355C

Firmado digitalmente por HERNANDEZ GARCIA ANGELA PATRICIA - 70889355C
Fecha: 2022.07.06 18:19:25 +02'00'

D. Alberto Berral González



Este trabajo de fin de grado se ha desarrollado en colaboración con el Servicio de Proteómica Funcional del Centro de Investigación del Cáncer de Salamanca, perteneciente a ProteoRed (Red Española de laboratorios de investigación en proteómica) y a ISCIII (Instituto de Salud Carlos III). Parte del contenido está relacionado con el proyecto de investigación FIS PI21/01545.

ÍNDICE

ÍNDICE FIGURAS

ÍNDICE DE TABLAS

1 INTRODUCCIÓN

El sistema inmune es una red de órganos y células compleja, perfectamente integrada y coordinada a nivel corporal. Su desregulación da lugar a enfermedades como las leucemias. En este caso, la prevalencia de la leucemia linfocítica crónica (*LLC)* y la heterogeneidad de la enfermedad hace necesaria la búsqueda de nuevas perspectivas de análisis, como la que plantea la proteómica. Esta ciencia permite obtener una caracterización funcional de las proteínas junto con sus relaciones estructurales, y así conseguir un análisis más preciso de una patología clínica en concreto (Mojica Ph.D et al., 2003).

En este estudio se trabaja con muestras de amígdalas de donde se obtienen los diferentes estadios del linfocito B y muestras de sangre de *LLC* con el objetivo de desarrollar flujos de trabajo sistemáticos para su caracterización proteómica en comparación con su contrapartida de células B normales y sanas.

2 FUNDAMENTOS BIOLÓGICOS

2.1 INMUNOLOGÍA

La inmunología es la ciencia que se ocupa del estudio del sistema inmune, es decir, del conjunto de mecanismos utilizados para la defensa contra agentes infecciosos (Regueiro, 2021)

El sistema inmune está compuesto por un grupo de órganos, tejidos, células y moléculas, que interaccionan entre sí para proporcionar un estado de inmunidad contra una infección (Delves et al., 2014). Un organismo se dice que es inmune cuando tiene la capacidad para mantenerse exento de infecciones por agentes patológicos, enfermedades por alteraciones de células (cáncer) o por trasplantes de órganos o sangre (Regueiro, 2021).

Este sistema está diseñado para detectar y eliminar patógenos. Para ello, se desarrolla una estrategia que incluye en primer lugar, la identificación o el reconocimiento del agente infeccioso, a continuación, la activación de la célula o molécula involucrada y, finalmente, la función efectora asociada a esa célula o molécula crea una respuesta para combatir el patógeno. Es esencial que la etapa de reconocimiento se efectué de forma correcta, de lo contrario, se pueden producir daños irreparables en nuestro organismo, causando enfermedades como por ejemplo las autoinmunes (Regueiro, 2021).

2.1.1 RESPUESTA INNATA Y ADAPTATIVA

La capacidad de un individuo para mantenerse libre de infección depende tanto de su resistencia natural o inmunidad innata como de la resistencia que pueda desarrollar o adquirir durante su vida, inmunidad adquirida o adaptativa (Tabla 1) (Parham et al., 2007).

La inmunidad innata es un mecanismo de defensa poco desarrollado e inespecífico, ya que usa la misma estrategia contra diferentes agentes infecciosos. Actúa a través de receptores inespecíficos que identifican patrones moleculares comunes a un conjunto de patógenos. Además, la exposición repetida de un mismo agente extraño induce respuestas similares y con la misma intensidad; lo que significa que la inmunidad innata carece de memoria inmunológica (Salinas Carmona, 2017).

Aunque sea una respuesta poco desarrollada e inespecífica, es una respuesta rápida, puede combatir la infección de forma inmediata. Además, no precisa de estímulo para estar presente y su actuación es necesaria para el posterior desarrollo de la respuesta inmune adaptativa, a la que ha de exponerse el agente si se ven superadas las primeras líneas de defensa (Vega Robledo, 2014).

La inmunología adaptativa, especifica o adquirida proporciona una defensa con mecanismos muy desarrollados y complejos, que, gracias a la versatilidad de sus diversas interacciones con los patógenos, facilitan una respuesta única y eficaz frente a los patógenos internos y externos. Gracias a la versatilidad anteriormente nombrada, el sistema inmunitario es capaz de diferenciar entre 10^9 y 10^{11} determinantes antigénicos distintos, es decir, la parte específica de una estructura de un antígeno que inducirá una respuesta inmunitaria (Castellanos-Bueno, 2020).

Otra de las características de la respuesta inmune adaptativa es su especificidad. Los efectores fundamentales en la respuesta inmunológica son los linfocitos. En cada linfocito de nuestro organismo se desarrollan receptores únicos para un solo antígeno, pudiendo generar así respuestas específicas en su contra. Las funciones efectoras propias de los linfocitos se activan cuando sus receptores interaccionan con el antígeno (Fainboim & Geffner, 2011).

Además, la respuesta innata adaptativa, tiene memoria inmunológica; tiene la capacidad de generar defensas más rápidas y eficaces, frente a patógenos a lo que haya habido una exposición previa. De esta manera, al exponer al sistema inmune ante un agente extraño por primera vez, se adquiere inmunidad para próximas exposiciones. Esta es la base para el desarrollo de vacunas (Delves et al., 2014). Las vacunas son preparados artificiales de antígenos derivados de patógenos, que carecen de la capacidad infectiva del patógeno frente al que se dirigen. Con las vacunas se producen anticuerpos que impedirán en el futuro que el patógeno pueda infectar y desencadenar la patología correspondiente al ser neutralizado.

Tabla 1. Inmunidad innata y adaptativa

INMUNIDAD INNATA Y ADAPTATIVA		
RESPUESTA	**INNATA**	**ADAPTATIVA**
Especificidad	No	Si
Memoria	No	Si
Tiempo	Rápida (Segundos)	Lenta (Días)
Dirigido a	Patrones moleculares	Antígenos
Ejemplo	Fagocitos	Anticuerpos

2.1.2 DIVISIÓN DEL SISTEMA INMUNE ADAPTATIVO

Hay dos tipos de inmunidad adaptativa, la inmunidad celular y la inmunidad humoral. Los dos trabajan juntos para eliminar microorganismos, sin embargo, gracias a la especialización de los receptores ubicados en los linfocitos, principales efectores de esta respuesta, se producen respuestas óptimas frente a los microorganismos. Por tanto, la estimulación de la inmunidad celular y la inmunidad humoral se genera por diferentes microorganismos (Vega Robledo, 2014).

1. Inmunidad celular

   Los principales efectores de la inmunidad celular son las células o linfocitos T. Se originan en la medula ósea y posteriormente maduran en el timo. Son los encargados de la neutralización de los microorganismos intracelulares.

Dentro del grupo de los linfocitos T existen dos tipos, los linfocitos T citotóxicos (Tc, CD8) y los linfocitos reguladores (Th, CD4) (Vega Robledo, 2014). Los linfocitos reguladores que expresan CD4, son los encargados de coordinar la respuesta inmunitaria, al activar a otras células inmunes como los macrófagos, los linfocitos B y los linfocitos citotóxicos. Del mismo modo regulan cualquier actividad excesiva del sistema e incluso los posibles ataques a otras células. Por otro lado, se encuentran los linfocitos citotóxicos (CD8), encargados de reconocer y adherirse a las células infectadas de microorganismos como bacterias, virus o alteraciones estructurales y destruirlas (Castellanos-Bueno, 2020).

2. Inmunidad humoral

La inmunidad humoral es la principal defensa frente a microorganismos extracelulares, ya que sus elementos efectores suelen encontrarse en circulación en el microambiente celular. La inmunidad humoral se compone de macromoléculas como proteínas y anticuerpos, producidas por los linfocitos B (Regueiro, 2017).

Cuando los receptores específicos de los linfocitos B detectan agentes extraños, los linfocitos se activan y se transforman en células plasmáticas productoras de anticuerpos (Ac), que son glucoproteínas sintetizadas por estos. Al igual que los receptores del linfocito son específicos para el antígeno, los anticuerpos que sintetizan también lo son. Cabe destacar que las células T son también las encargadas de proporcionar citocinas esenciales para el linfocito B, e iniciar la expansión clonal y el cambio de clase en la producción de anticuerpos (Castellanos-Bueno, 2020).

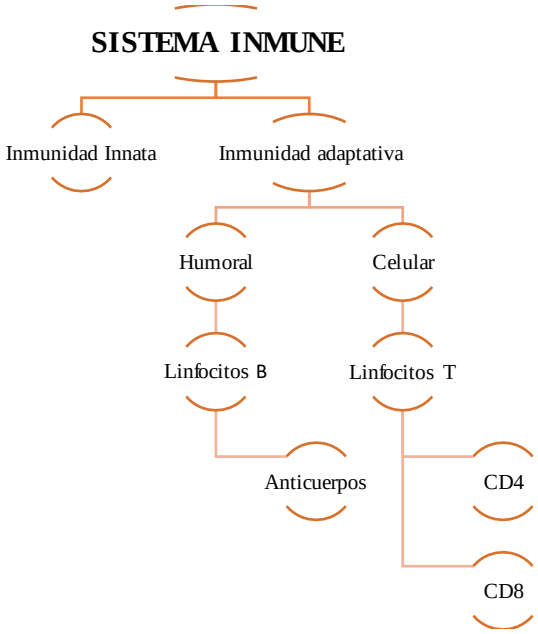


Figura 1. División del sistema inmune.

2.1.3 LA INMUNIDAD ADAPTATIVA ESTÁ MEDIDA POR LOS LINFOCITOS B Y T.

Los receptores de membrana de los linfocitos B y T, se diferencian de los receptores de los demás leucocitos debido a que son específicos para un patógeno en concreto. Sólo aquellos linfocitos activados, proliferan en la respuesta específica y desarrollarán memoria inmunitaria hacia dicho patógeno. Los receptores de membrana capaces de reconocer los patógenos se denominan, *BCR* (del inglés "B cell receptor") y *TCR* (del inglés "T cell receptor"), en los linfocitos B y T respectivamente (Regueiro, 2021).

El proceso de lucha contra los patógenos se lleva a cabo mediante una acción clonal de los linfocitos B y T, que aumentan su número de células por división, garantizando así la lucha frente a los agentes extraños con la especificidad de la respuesta adaptativa. La acción clonal dota además a los linfocitos B y T de memoria inmunitaria desarrollada en dos etapas: la primera etapa es la interacción original con el antígeno, donde en el caso de los linfocitos B, se genera una alta concentración de estos para poder combatir al patógeno, y otros linfocitos de la misma especialidad se referencian como células de memoria, capaces de perdurar en los tejidos desde días hasta años, esperando activarse de nuevo tras la aparición del antígeno característico (Figura 2) (Regueiro, 2021).

La segunda activación de los linfocitos ya asociados a un patógeno en concreto daría paso a la segunda etapa o respuesta secundaria. En este caso se activan los clones de memoria los cuales son más numerosos que en su primera acción y facilitan el proceso siendo este más rápido y eficaz que en la primera etapa (Regueiro, 2021).

Un ejemplo cotidiano donde se puede ver la acción de las dos respuestas, son mediante las vacunas donde recibimos varias dosis, con la finalidad de generar una sucesión de respuestas y así obtener muchos anticuerpos y memoria inmunitaria duradera (Delves et al., 2014).

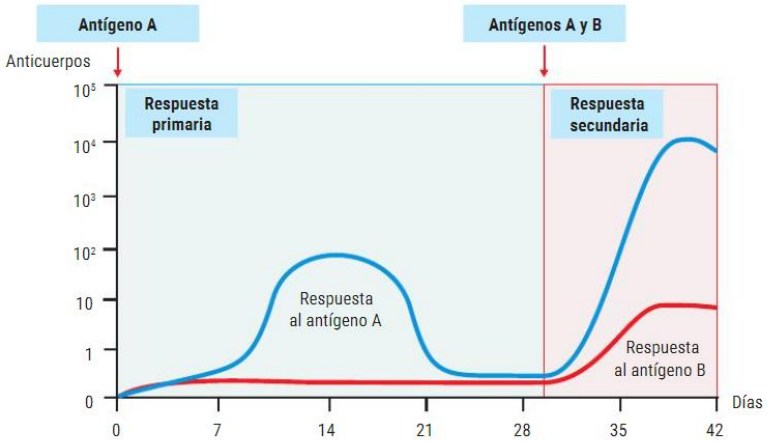


Figura 2. Memoria inmunitaria. Figura adaptada de Regueiro, 2021.

2.1.4 PAPEL DE LOS LINFOCITOS B

Las células B se crean y maduran en medula ósea pasando por varios estadios hasta que se ubican en los ganglios linfáticos, donde se activan en presencia de un agente extraño, con la ayuda de otro tipo celular, los linfocitos T CD4, en la mayoría de los casos. Las células B durante su maduración expresan diferentes moléculas de superficie que son útiles para su identificación y conocimiento de su capacidad funcional.

Los Linfocitos B son células que participan principalmente en la inmunidad humoral por su papel como células productoras de anticuerpos. Estas células se originan a partir de la célula madre hematopoyética pluripotencial, de la cual derivan todas las células de la sangre. Tras diferentes estadios de maduración en la médula ósea, la célula *naïve* migra a los centros germinales donde las células se dividen en el estadio de *centroblastos* y tras el periodo de proliferación, se agrupan como *centrocitos*. Las etapas finales de la diferenciación de las células B humanas son las denominadas "dependientes de antígenos" que conducen a la expansión de las células B hasta su diferenciación terminal en células plasmáticas secretoras de anticuerpos y células B de memoria, que se realiza en tejidos linfoides secundarios (Figura 3).

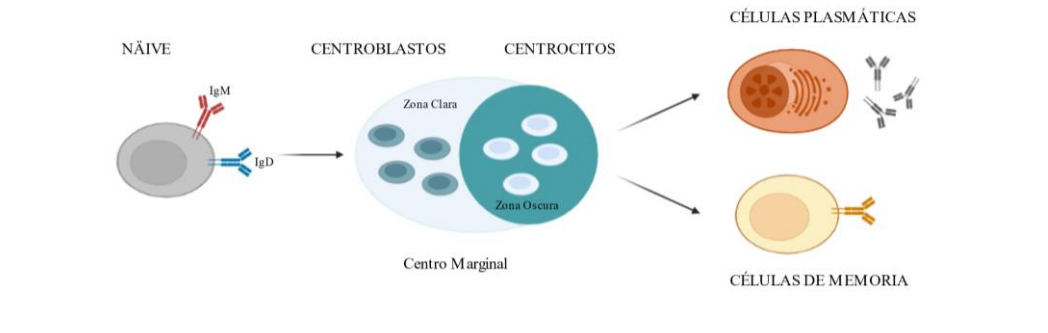


Figura 3. Maduración antígeno dependiente del linfocito. Figura BioRender.

La diferenciación celular del linfocito B se puede realizar por la presencia de antígenos de superficie o cúmulo de diferenciación (CD) presentes en la membrana celular y que permiten la identificación de las diferentes poblaciones, así como su separación fenotípica mediante técnicas de citometría de flujo, a pesar de las grandes similitudes a nivel genómico que presentan estas células (Figura 4) (Díez et al., 2021).

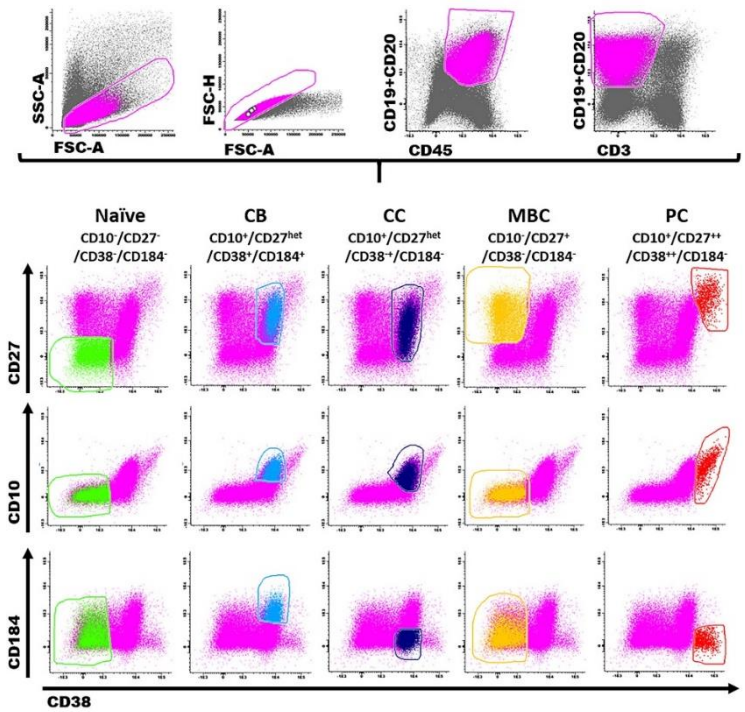


Figura 4. diferencias en los CDs de los estadíos de linfocito B, naïve, centroblasto (CB), centrocito (CC), célula B de memoria (MCB) y célula plasmática (PC). Figura adaptada de Díez P. et al. 2021.

2.2 CÁNCER

El cáncer es una enfermedad genética en la que ciertas células del cuerpo proliferan sin control, pudiendo diseminarse también a otras partes del cuerpo. En circunstancias normales, las células humanas nacen y se multiplican a través de la división celular, creando nuevas células según sea necesario. No obstante, puede existir una variación en el ciclo normal y, células anormales o dañadas se replican sin moderación dando lugar a la aparición de tumores.

Estos tumores se pueden clasifican en malignos o benignos, la diferencia principal entre estos dos grupos viene dada por la capacidad de diseminación. Los tumores malignos o cancerosos son capaces de propagarse e invadir otras partes del cuerpo creando nuevos tumores, proceso definido como metástasis; mientras que los benignos no se pueden expandir y una vez extirpados no suelen volver a salir.

En este caso, el interés del trabajo serán las células tumorales. A continuación, se expone una comparación entre las células tumorales y las células normales (Tabla 2).

Tabla 2: Características de las células normales frente a las células tumorales

CÉLULAS NORMALES	**CÉLULAS CANCEROSAS**
Reciben señales para formarse	No reciben señal de que se deben formar
Atienden a la muerte celular programada	No atienden a la muerte celular programada
No se mueven ni se multiplican sin control	Invaden áreas cercanas y se propagan a otras partes del cuerpo
Los vasos sanguíneos no se modifican	Hacen que los vasos sanguíneos crezcan (llevan más oxígeno y nutrientes al tumor)
Son reconocidas por el sistema inmune	Se camuflan del sistema inmune
No alteran los cromosomas	Incorporan modificaciones en los cromosomas
Se multiplican de manera normal	Se multiplican más rápido de lo normal

Dentro de los tumores, hay dos grandes divisiones a la hora de clasificar los tipos de cáncer, por un lado, están los denominados tumores sólidos:

- Carcinomas, se forman en las células epiteliales son las células que cubren las zonas internas y externas del cuerpo.
- Sarcomas, se desarrollan en los huesos y tejidos blandos como músculos, grasa, vasos sanguíneos…
- Tumores de encéfalo y médula espinal, aparecen en el sistema nervioso central.
- Tumores carcinoides, normalmente crecen en el aparato digestivo (recto, intestino delgado…).
- Tumores de células germinativas, provienen de las células que crean los espermas y los óvulos.

Por lado otro, están los tumores de las células sanguíneas de la serie blanca (linfocitos, monocitos y granulocitos) denominados tumores hematológicos. Estas células se forman en la médula ósea, maduran en los órganos linfoides y circulan por la sangre donde ejercen su papel como células del sistema inmune. Estas patologías se pueden clasificar por diferentes criterios:

- Lugar donde se produce la proliferación:

- Leucemia: cuando la patología parece en la médula ósea.
- Linfoma: la proliferación ocurre en el sistema linfoide o en órganos asociados como, por ejemplo, en los ganglios.

- Tasa de crecimiento celular:
  - Aguda: de crecimiento rápido.
  - Crónica: de crecimiento lento.

- Tipo de célula afectada:
  - Linfocítica: cuando son los linfocitos lo que presentan una proliferación anormal.
  - Mieloide: cuando afecta a otras células sanguíneas (plaquetas, glóbulos rojos u otros glóbulos blancos que no son linfocitos.

Una característica que muestra el impacto que tiene el cáncer como enfermedad, son los estudios de incidencia-prevalencia, y los de incidencia-mortalidad. En la siguiente figura se muestran datos comparativos de incidencia-mortalidad, de los 15 tipos de cánceres con mayor importancia en España (Figura 5).

Este trabajo se basa en la leucemia linfocítica crónica, que es un cáncer con una incidencia media-baja, 5.800 casos aproximadamente en 2020; en comparación con cánceres más frecuentes, como el de próstata que supera los 30.000. Además, la leucemia presenta una tasa media-alta de mortalidad por incidencia de pacientes mayor que la del cáncer de próstata.

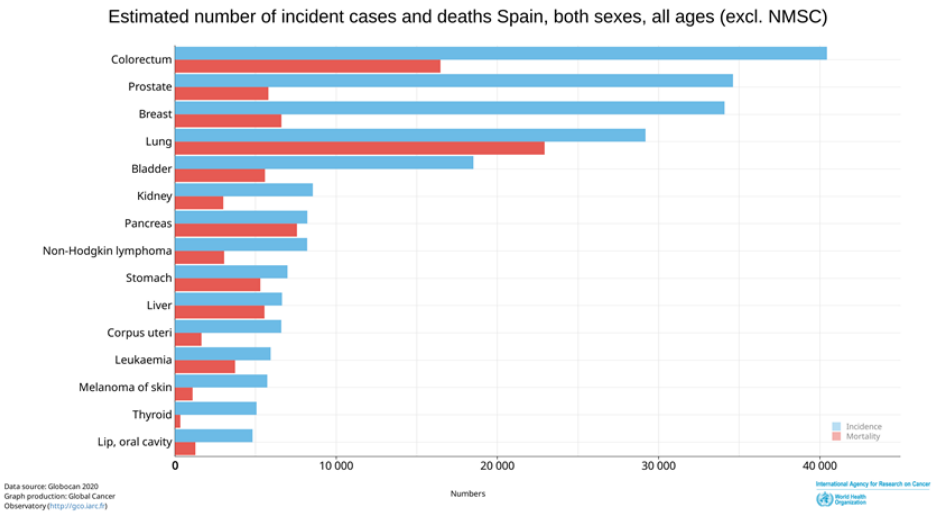


Figura 5. Gráfico comparativo de incidencia-mortalidad de la población española, año 2020.

Con el mismo patrón de estudio, se muestra datos comparativos de incidencia-prevalencia a lo largo de 5 años para los distintos tipos de cáncer en España (Figura 6). En el estudio se ve que la leucemia tiene una prevalencia alta, con un aumento aproximadamente del 300% de los casos. Ahora bien, en el cáncer de próstata donde se veía en la anterior gráfica que su impacto en mortalidad no era tan significativo como en la leucemia; en prevalencia si se puede afirmar que tiene una tasa también alta, incluso mayor a la de la leucemia, con un incremento del 400% de los casos a los 5 años.

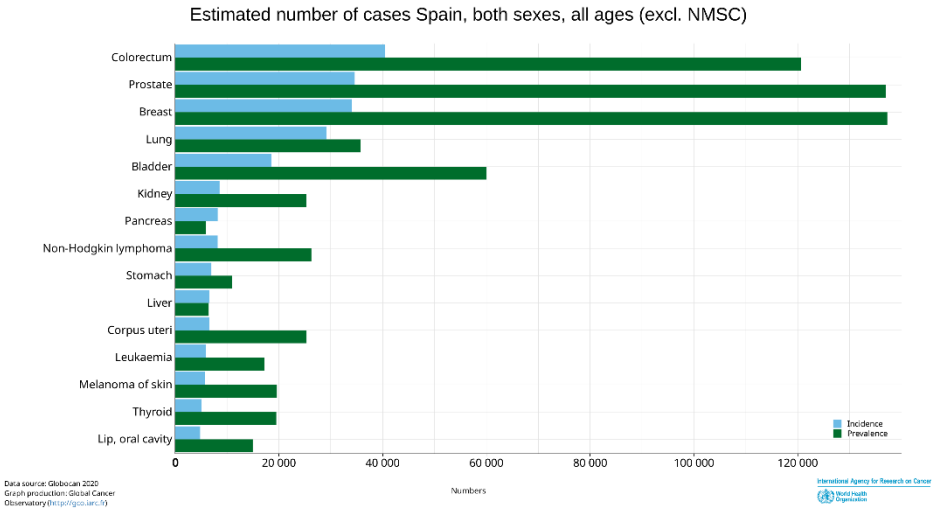


Figura 6. Gráfico comparativo de incidencia-prevalencia de la población española, año 2020.

La complejidad del cáncer como enfermedad es abrumadora por la amplitud de factores que abarca como, por ejemplo, la genética, las células tisulares, patologías y tipos de respuesta a la terapia. Las herramientas experimentales y computacionales cada vez son más poderosas y proporcionan una avalancha de "grandes datos" sobre las innumerables manifestaciones de las enfermedades que engloba el cáncer. Esto está ayudando a destilar esta complejidad en una ciencia cada vez más lógica, que comprende mejor los mecanismos y aplica un conocimiento más profundo a la medicina del cáncer (Hanahan, 2022).

2.3 LEUCEMIA LINFOCÍTICA CRÓNICA Y PROTEÓMICA

La leucemia linfocítica crónica (*LLC*) es uno de los tipos más frecuentes de leucemia. Habitualmente ocurre en pacientes de edad avanzada y tiene un diagnóstico y curso clínico muy variable. Esta enfermedad es iniciada por alteraciones genéticas específicas, que interfieren en la proliferación de células B clonales y en su proceso de apoptosis. Estas células B tumorales se acumulan, en sangre, médula ósea, ganglios linfáticos o el bazo, provocando disfunciones del sistema inmune y problemas en la homeostasis de los pacientes (Díez et al., 2017).

La incidencia ajustada por edad según el SEER *"The Surveillance, Epidemiology, and End Results"* es de 4,9 por 100.000 habitantes al año, con una media de edad de diagnóstico de 70 años y con un número mayor de hombres que de mujeres. La supervivencia relativa a 5 años de los pacientes con *LLC* fue del 65,1 % en 1975 y ha aumentado constantemente durante las últimas décadas; se estimó un 87,2 % en 2021(Hallek & Al-Sawaf, 2021).

Es necesario caracterizar funcionalmente la diferenciación de las células B para el diagnóstico, pronóstico y tratamiento de la enfermedad. Como ya se ha mencionado anteriormente, la *LLC* se caracteriza por su alta heterogeneidad a lo largo de la evolución de la enfermedad. Esta heterogeneidad a nivel genético se refleja en intra- e inter- variaciones tumorales. Las deleciones del cromosoma 13q, 17p y 11q juntos con la trisomía 12 son consideradas como las lesiones genéticas más abundante en casos de *LLC*; y, pueden están presentes simultáneamente más de una alteración genética.

Para la caracterización y selección de un tratamiento optimo en la enfermedad de *LLC*, todavía existen ciertas limitaciones que requieren nuevos enfoques. En este sentido, la proteómica aparece como una tecnología prometedora para descifrar la alteración vías de señalización celular implicadas en la enfermedad (Díez et al., 2021).

Las funcionalidades de las células de *LLC* siguen siendo una cuestión abierta que, una vez superada, podría aumentar notablemente la supervivencia de los pacientes, proporcionando una mejora en el diagnóstico, pronostico y tratamiento. Es necesaria una caracterización cualitativa, pero también cuantitativa en las que diferentes proteínas expresadas en estas células descifren alteraciones en las rutas de señalización.

Esto ayudaría a la selección de los tratamientos adecuados según el paciente (medicina personalizada), y a identificar nuevas dianas para futuros tratamientos, crucial para encontrar una cura y, en este punto, él se requieren perfiles de modificaciones postraduccionales (del inglés *"Post-translational modification"*, PTM) ya que las fosforilaciones (entre otras PTMs) aparecen como alteraciones importantes que desencadenan cascadas de señalización de manera concreta.

Antes de definir del concepto proteómica es conveniente introducir el término de genómica que se basa en el estudio del genoma. El genoma constituido principalmente por ADN (ácido desoxirribonucleico) se ordena en el núcleo celular en 23 pares de fragmentos denominados cromosomas. El ADN lo forman cuatro nucleótidos que son la timina, la citosina, la guanina y la adenina, que se interrelacionan entre sí por combinaciones dos a dos (Mojica Ph.D et al., 2003).

Para utilizar la información del genoma, se recurre al dogma central de la biología molecular. Este proceso se basa en la replicación llevada a cabo en las dos etapas, la transcripción y la traducción. Tiene como finalidad sintetizar proteínas, que serán las encargadas de realizar todas las tareas de las células.

Por tanto, nace una nueva necesidad de estudiar el comportamiento de las proteínas en los organismos y células. Dicha ciencia se denomina proteómica, basada en el estudio de las proteínas y proteomas de los organismos y células y sus modificaciones en sus diferentes interacciones; además define niveles celulares, funciones metabólicas e interrelaciones. Se podría decir que la proteómica realiza una caracterización funcional y estructural de las proteínas (Benito Jiménez, 2019).

Las técnicas más comunes utilizadas en proteómica son: electroforesis en geles 2D, espectrometría de masas y microarrays de proteínas para ensayos masivos (Figura 7) (Mojica Ph.D et al., 2003).

- Electroforesis en geles 2D: es una técnica de separación de moléculas según su movilidad en un campo eléctrico. El método utilizado es en 2D, ya que en la primera dimensión las proteínas se separan por su carga eléctrica y en la segunda por su masa, a continuación, las manchas del gel se extraen y se tratan con tripsina para obtener patrones de péptidos que se analizarán en una espectrometría de masas.

- Microarrays: es una técnica que permite especificar la información necesaria de las proteínas que son traducidas del ARNm a través de un soporte sólido con forma de matriz, en el cual se imprime en un orden determinado una serie de moléculas. Esta técnica supone un gran avance en la proteómica al poder identificar las proteínas modificadas por los diversos procesos, posibilitando aplicaciones para enzima-sustrato, ADN-proteína y otras interacciones entre proteína-proteína.

- Espectrometría de masas: es una técnica de identificación de péptidos y proteínas, que se realiza a través de una fuente de iones y un aparato medidor.

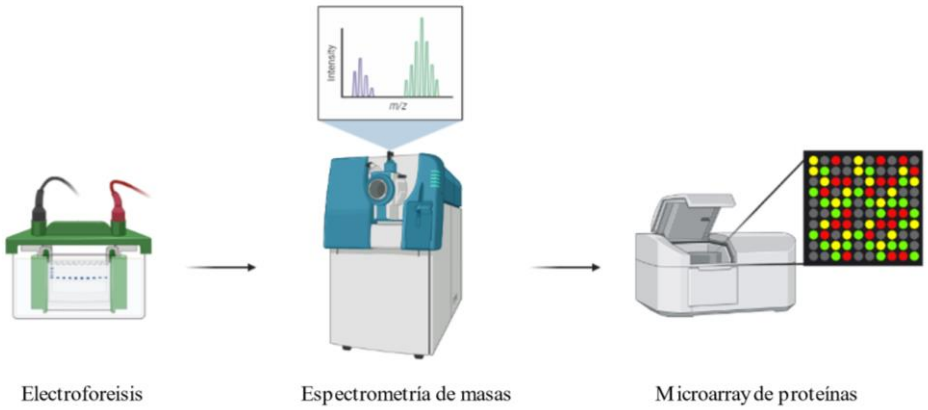


Figura 7. Técnicas en proteómica. Figura BioRender.

3 OBJETIVO

El sistema inmune es una red de órganos y células compleja, perfectamente integrada y coordinada a nivel corporal. Su desregulación da lugar a enfermedades como las leucemias. En este caso, la prevalencia de la *LLC* y la heterogeneidad de la enfermedad hace necesaria la búsqueda de nuevas perspectivas de análisis de esta patología con el fin de obtener una información más precisa y adecuada que pueda servir para la aplicación directa en los pacientes. En el contexto de la bioquímica y la biología celular, la proteómica presenta un gran auge como metodología traslacional para determinación de perfiles que de proteínas que puedan ayudar a conocer más en profundidad las patologías clínicas. La gran cantidad de datos que se generan gracias al avance en tecnologías como la proteómica y la complejidad de la *LLC* a nivel fenotípico ha planteado el objetivo general de este trabajo basado en el diseño y desarrollo de flujos de trabajo sistemáticos para la caracterización proteómica de leucemia linfoide crónica en comparación con su contrapartida de células B normales y sanas.

Este objetivo general se puede concretar en los siguientes objetivos específicos:

1. Plantear una estrategia en R que permita un análisis completo de datos proteómicos.
    - Estudio descriptivo cuantitativo y cualitativo de los seis tipos celulares.
    - Comprobar la calidad de las muestras, revisión e imputación de valores perdidos.
    - Uso de técnicas de reducción de dimensiones.
    - Comparación de las muestras de *LLC* con cada uno de los Estadios del linfocito B.

2. Conocer el propósito, la metodología y aplicación práctica en R, de la herramienta bioinformática SAM.

3. Aplicar técnicas de enriquecimiento funcional con el objetivo de relacionar las proteínas significativas entre grupos, con fenotipos de enfermedades y desarrollar sus respectivas rutas de señalización.

4 OBTENCIÓN DE LA BASE DE DATOS

En este trabajo se han empleado datos de proteómicos de células de LLC y su contrapartida normal subdividida en los diferentes estadios de diferenciación (*Naïve*, *Centroblastos*, *Centrocitos*, *Memoria* y *Células plasmáticas*). El origen de la muestra difiere por las particularidades de la localización de las células: sangre periférica en LLC y estadio de células B en tejidos linfiodes (amígdalas). Una vez obtenidas las células, el procesamiento de las muestras y la obtención de los resultados se realizó mediante un procedimiento común como se representa en la figura 8

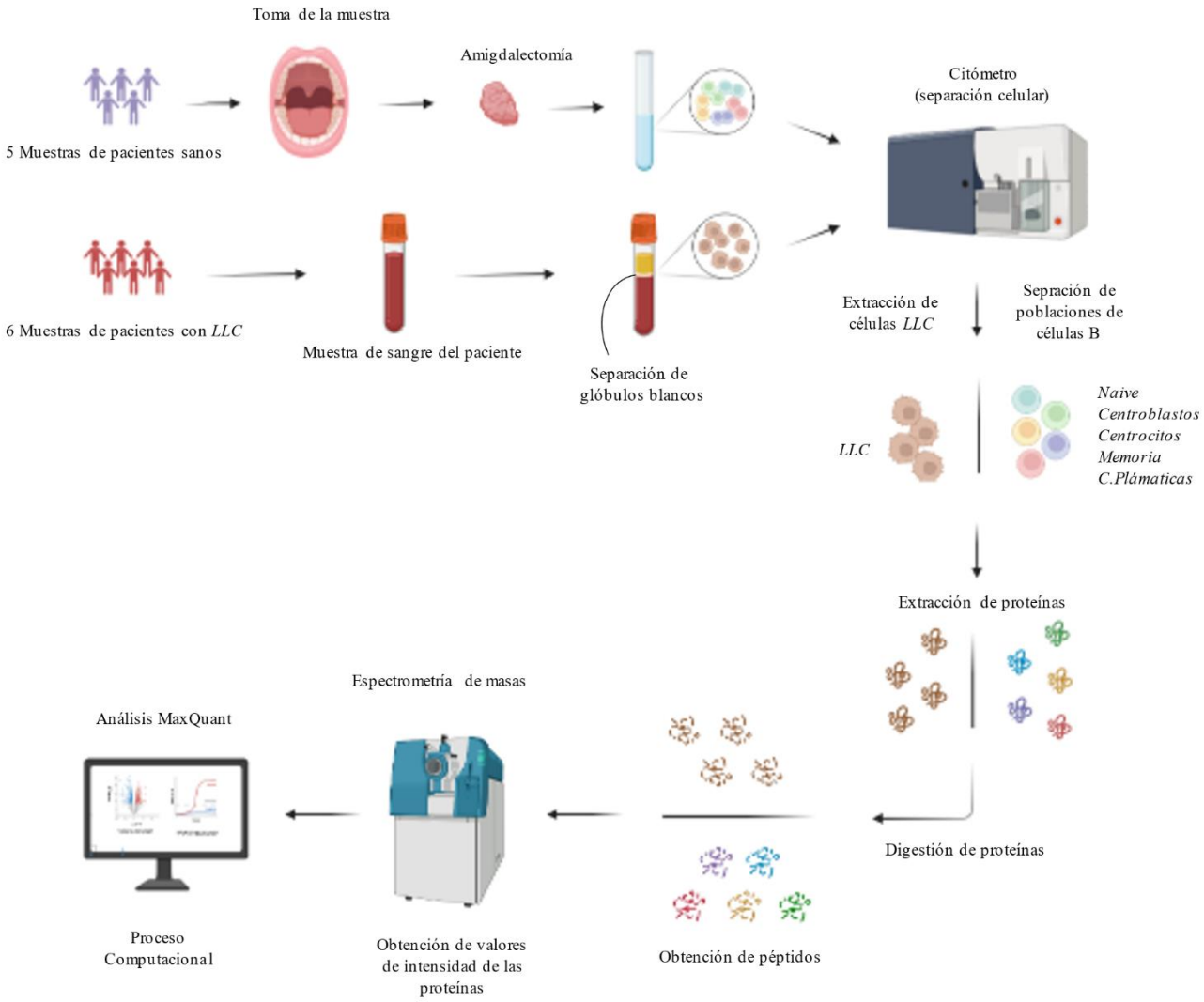


Figura 8. Esquema obtención base de datos. Figura BioRender.

5 ANÁLISIS ESTADÍSTICO

5.1 DESCRIPCIÓN DE LA BASE DE DATOS

Para este trabajo, se parte de dos bases datos compuestas por un gran número de proteínas con sus respectivos valores de intensidad. Para realizar los análisis, se necesita unificar ambas bases de datos con las proteínas pertenecientes a los 5 estadios del linfocito B y a las muestras de *LLC*.

El primer conjunto de datos está formado por 25 muestras, 5 para cada estadio del linfocito B (*Naïve*, *Centroblastos*, *Centrocitos*, *Memoria* y *Células plasmáticas*). Cada uno de estos grupos viene definido por el nombre de las proteínas identificadas y cuantificadas por el proceso de MS/MS, y sus respectivos valores de intensidad para cada una de las cinco muestras. Cabe destacar que en cada estadio hay un número diferente de proteínas (Tabla 3).

Tabla 3. Nº de proteínas para cada estadio del Linfocito B

Muestras	N.º de proteínas
Naive	2007
Centroblastos	2493
Centrocitos	2572
Memoria	2593
Células Plasmáticas	393

Además, se revisa la presencia de valores perdidos y se realiza con el paquete de R "ggplot2" (Wickham, 2016) un gráfico de barras con el sumatorio de los valores de intensidad de las proteínas, para cada una de las 25 muestras (Figura 9). Se tiene que no hay ningún valor perdido y que la carga de las proteínas se ha realizado correctamente, ya que, las muestras en cada uno de los estadios se distribuyen de manera alterna, a excepción del grupo de *Células Plasmáticas*. Era de esperar que su total fuese significativamente menor, porque el número de proteínas que se tienen en este grupo es de 393 frente al resto de grupos que tienen una media de 2200 proteínas.

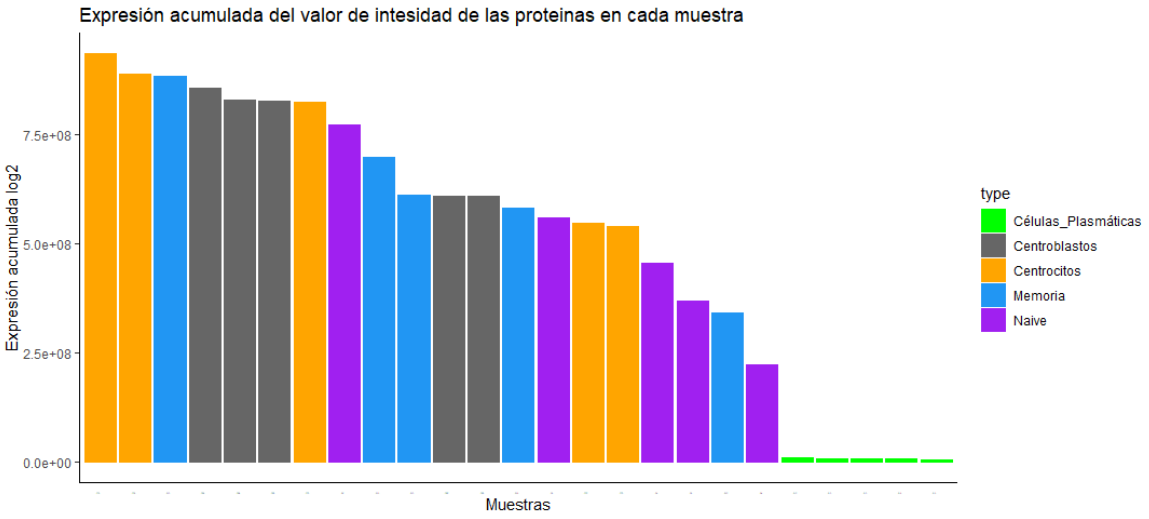


Figura 9. Expresión acumulada de las proteínas de las muestras de los estadios del Linfocito B.

El segundo conjunto de datos consta de 6 muestras de *LLC* con diferente número de proteínas para cada individuo (Tabla 4). Las proteínas también vienen dadas en valores de intensidad reconocidas y cuantificadas mediante el proceso de MS/MS.

Tabla 4. Nº de proteínas para cada muestra de LLC

Muestras	N.º de proteínas
LLC 1	3904
LLC 2	2156
LLC 3	2885
LLC 4	2969
LLC 5	2969
LLC 6	3397

Como se quiere analizar conjuntamente estas 6 muestras de *LLC*, se crea una base de datos auxiliar donde aparezcan las proteínas con valores de intensidad en al menos una de las muestras. (ANEXO). Se utilizan, con ayuda del software R, las funciones "excel_sheet"," read_excel" (Wickham et al., 2022), "unlist", "unique" y "match" (Huber et al., 2015).

La base de datos que se obtiene está compuesta por 4626 proteínas identificas. Se procede a examinar los valores perdidos denotados NA, que es la notación habitual en el software R (Figura 10). Se obtienen 2192 proteínas (47%) que tienen más de 4 muestras con valores perdidos, por tanto, se procede a eliminar estas proteínas utilizando únicamente 2434 en el estudio.

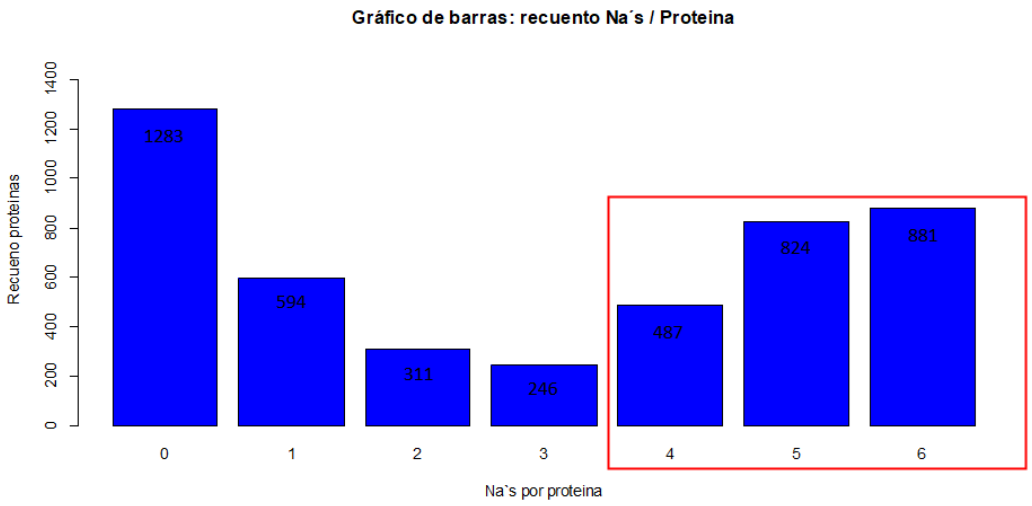


Figura 10. Recuento valores perdidos por proteína, el recuadro rojo indica las proteínas a eliminar.

5.2 OBTENCIÓN DE LA BASE DE DATOS

Una vez filtradas, se procede a unir la base de datos de Linfocitos B y *LLC*. En primer lugar, se calcula el número de proteínas en común entre los estadios del linfocito B y las muestras de *LLC*. Para ello calculamos la intersección entre ambos conjuntos y es habitual representarlo mediante un diagrama de Venn. Este tipo de gráfico está asociado a la teoría de conjuntos y muestra cómo se relacionan los elementos entre dos o más espacios muestrales (Lorenzo, 2021).

Pero como queremos mostrar el número de proteínas comunes en cada intersección entre los tipos celulares dos a dos, tres a tres, etc. el diagrama de Venn resulta muy complejo de visualizar y se ha decidido utilizar el gráfico UpSet que se muestra a continuación (Figura 5). Se ha realizado con el paquete de R "UpSetR" (Conway et al., 2017) incorporando los seis tipos celulares del estudio (N*aive, Centroblastos, Centrocitos, Memoria, Células plasmáticas y LLC)*. Se observan 302 proteínas que pertenecen a estos seis grupos y hay 609 solo identificadas en las muestras de LLC.

Cuando se quieren ver las relaciones entre espacios muestrales, el mayor problema para tener una buena representación, son las múltiples combinaciones de intersecciones si el número de conjuntos es elevado. El grafico UpSet es el más eficaz para el análisis de datos cuantitativo teniendo, en este caso, seis conjuntos.

Para poder representar este gráfico, es necesario transformar los datos a un formato aceptado por el paquete de R "UpSetR". Uno de estos formatos se corresponde con una matriz de ceros y unos donde, las filas representan proteínas y las columnas los seis tipos celulares que hacen las veces de los conjuntos a interseccionan. Dada una proteína i y un conjunto j, el valor x_{ij} vendrá definido por un 0 si la proteína i no pertenece al conjunto j y por un 1 si pertenece. A partir de aquí, se van clasificando las proteínas en las distintas intersecciones, según la cantidad y la posición de ceros y unos asignados para cada una de ellas.

Por tanto, el gráfico UpSet representa las intersecciones en una matriz, cada fila viene definida por un conjunto, y el grafico de barras en la parte izquierda muestra el tamaño de este. A su vez, cada columna determina una intersección, si el circulo (celda) se muestra relleno, quiere decir que ese conjunto pertenece a esa intersección, si por el contrario ese círculo está vacío (hueco en blanco), ese tipo celular no pertenece a esa intersección. Si solamente hay un círculo, ese conjunto es único de ese tipo celular. Además, se puede incorporar un gráfico de barras, como el de la parte superior, que hace un recuento de las proteínas que pertenecen a cada intersección dada ordenando las intersecciones de mayor a menor tamaño.

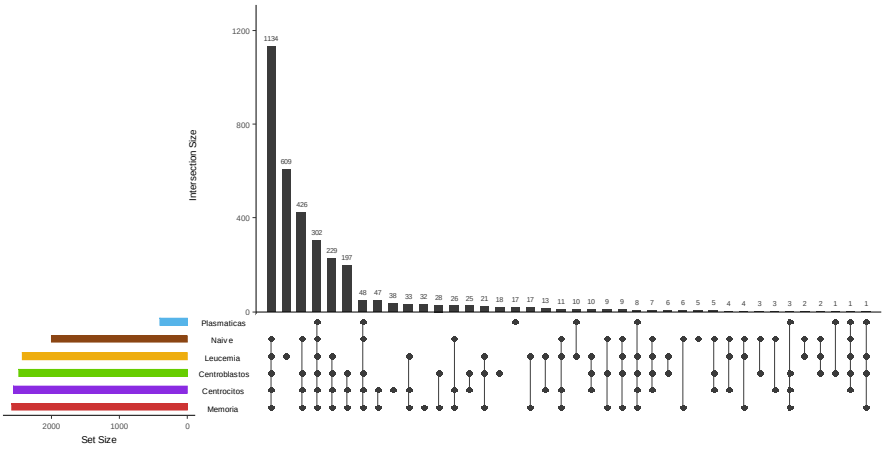


Figura 11. Gráfico Upset para los 6 tipos celulares.

Una vez realizado el diagrama, se ha creado una función en R con el comando "merge" (R Core Team, 2022) y "Reduce" (Wilke, 2021), para construir una hoja de datos con la expresión de las proteínas comunes en los seis tipos celulares. La base de datos estará compuesta por los valores de intensidad de las 302 proteínas en las 25 muestras pertenecientes a cada estadio del linfocito B y en las 6 muestras de leucemia.

```
>MyMerge_venn <- function (x, y) {
         df <- data.frame()
         df <- merge (x, y, by= "row.names")
         rownames(df) <- df$Row.names
         df$Row.names <- NULL
         return(df)
       }
>all_celltype <- Reduce (MyMerge_venn, list (df_Naive, df_Cenblastos,
                    df_Centrocitos, df_Memoria,df_Plasmaticas,df_LLC))
```

Antes de comenzar con el análisis estadístico, se realiza una normalización de los datos para que las distribuciones de la intensidad de las proteínas puedan ser comparables. El propósito de la normalización es reconocer y eliminar cambios sistemáticos manteniendo la señal biológica. Pretende garantizar que la diferencia de intensidad refleje verdaderamente la expresión diferencial de las proteínas y que no haya un sesgo debido a factores técnicos.

El método empleado para la normalización es el método de estandarización de los datos escalados con el logaritmo (que nos permite reducir la escala para entender mejor los valores de expresión). Según los autores García, Ramos y Ruiz (2016) "Dada una variable estadística X, con media $\bar{x}$ y desviación típica Sx, la tipificación consiste en la transformación:"

$$z = \frac{x - \bar{x}}{S_x}$$

Fórmula 1. Tipificación de una variable

Donde se obtiene una variable tipificada con media $\bar{z} = 0$ y desviación típica $S_z = 1$. Previamente se han escalado, por tanto, X es el log2 de los datos originales. El proceso de tipificación no modifica la distribución de la intensidad de las proteínas y el log2 simplemente cambia la escala.

Para ver cómo se distribuyen las proteínas en cada muestra, se crea un boxplot con los datos crudos (Figura 12) y se observa que los datos no han sido previamente normalizados.

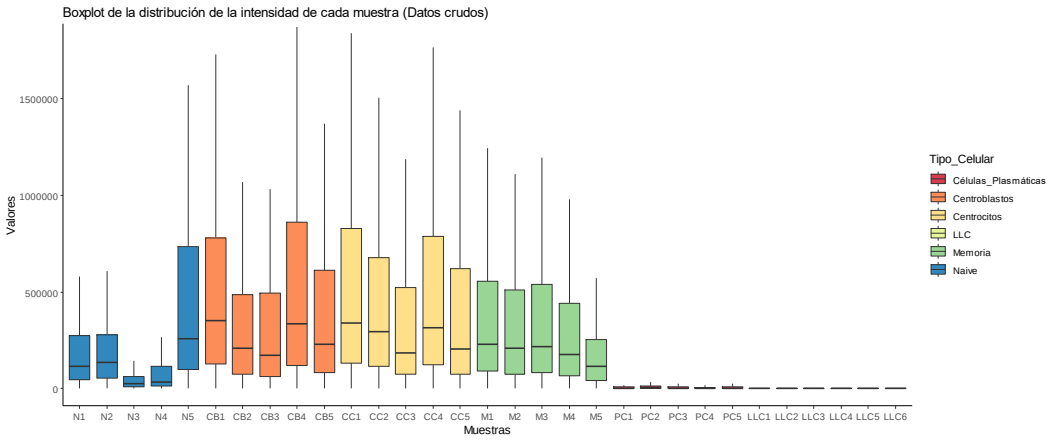


Figura 12. Boxplots de la distribución de la intensidad de cada muestra (Datos sin normalizar).

A continuación, se aplica a la base de datos el logaritmo en base 2 con el fin de reducir la escala de los datos y la tipificación previamente mencionada. Las funciones de R que se aplican son “log2” y “scale” (R Core Team, 2022). Se comprueba a través del boxplot (Figura 13) que las muestras tienen media cero, las medianas se parecen y el recorrido intercuartílico, los máximos y los mínimos son similares. Por tanto, se podría decir que las distribuciones son parecidas y las muestras pueden ser comparables.

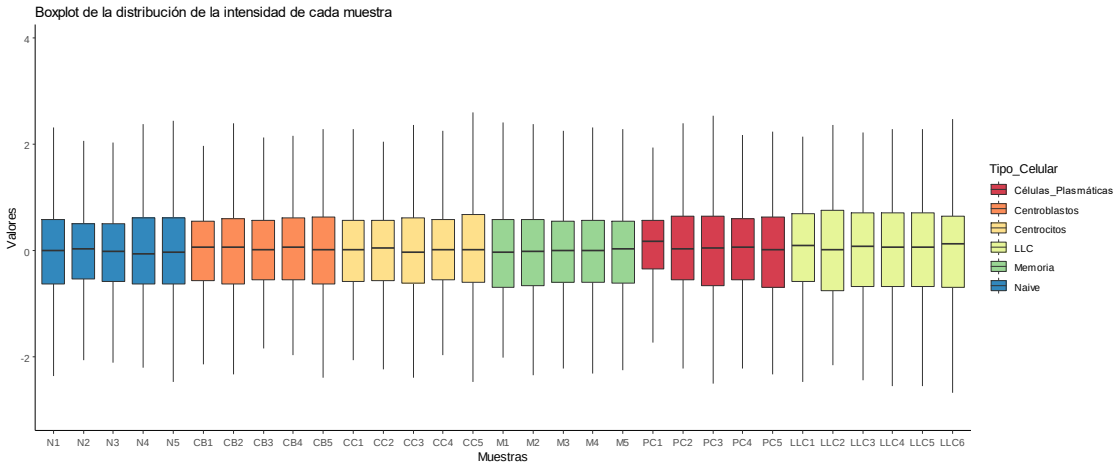


Figura 13. Boxplots de la distribución de la intensidad normalizada de cada muestra.

Para visualizar con mayor precisión el rango en el que se encuentra la distribución de las intensidades y comprobar la posible presencia de valores atípicos, se ha construido un histograma (Figura 14). Se considera como valor atípico a toda observación tan diferente de cualquier otra que se sospeche que fue producida por un mecanismo diferente (Atkinson & Hawkins, 1981).

Al observar el histograma, la distribución se encuentra en un intervalo de (-5,5) y la mayoría de los valores se concentran en torno al cero. Además, a excepción de un valor que se encuentra alrededor del -8, se podría decir que no hay presencia de outliers.

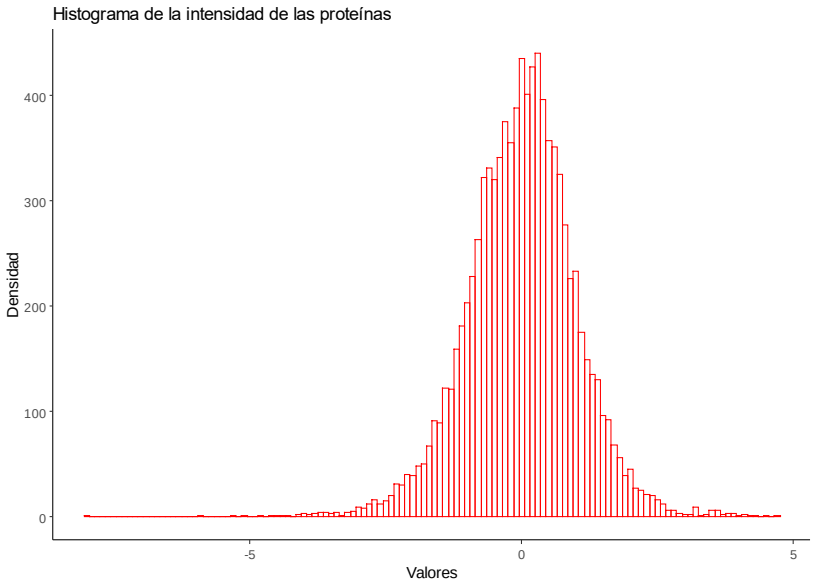


Figura 14. Histograma de la intensidad de las proteínas.

5.3 IMPUTACIÓN VALORES PERDIDOS

Antes de comenzar con la aplicación de técnicas multivariantes se debe realizar una imputación de valores perdidos. Esto se debe, a que estás técnicas no admiten muestras que tengan ni un solo valor ausente y, por tanto, si se eliminasen, se perdería una gran cantidad de información. Se comprobó que las muestras de los diferentes estadios del linfocito B no presentaban ningún valor perdido. Al contrario que las muestras de *LLC*, donde en promedio se tienen 7 valores perdidos por muestra. Por tanto, omitiendo los valores perdidos, todas las muestras de *LLC* se verían fuera del estudio.

El método utilizado para imputar los datos faltantes es el denominado imputación múltiple (MI). Esta imputación trata de predecir los valores ausentes a partir de otros datos presentes en una misma muestra, reconociendo la incertidumbre asociada con los valores imputados (Little & Rubin, 2014). El procedimiento que sigue este método es el siguiente (Figura 15):

- Imputa los valores ausentes n veces generando n grupos de datos completos con los valores presentes. Cada uno de estos grupos proporciona una estimación única de los valores perdidos.

- Calcula las estimaciones con sus respectivos errores estándar para cada uno de los n conjuntos de imputación.

- Agrupa las n estimaciones para obtener una estimación general y su error estándar correspondiente.

El error estándar agrupado obtenido de la estimación general es superior al error estándar generado en un único método de imputación (p. ej., sustitución de medias), ya que no se tiene en cuenta la incertidumbre entre imputaciones (Dong & Peng, 2013). Por ello, el método de imputación múltiple reduce el sesgo en el error estándar, siendo este método uno de los más efectivos. En R se utiliza la función “impute.knn”, proveniente del paquete impute (Hastie et al., 2022).

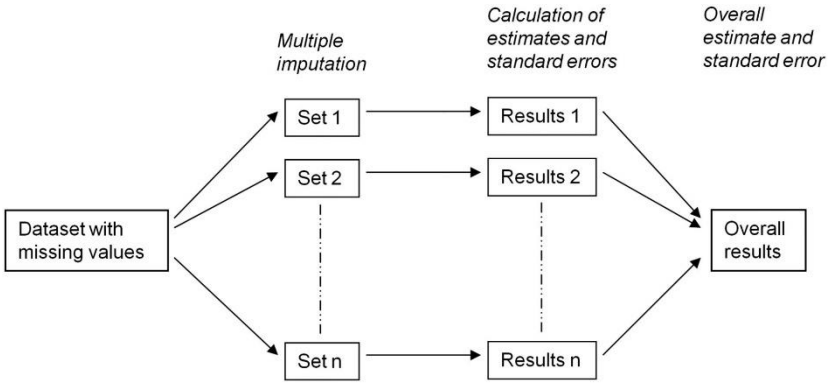


Figura 15. Representación gráfica del procedimiento de imputación múltiple.

5.4 TÉCNICAS MULTIVARIANTES

5.4.1 ANÁLISIS DE COMPONENTES PRINCIPALES

En este punto se va a aplicar el Análisis de Componentes Principales (PCA). Es un análisis de aprendizaje no supervisado de reducción de la dimensionalidad. El objetivo de esta técnica es encontrar transformaciones ortogonales de las variables originales para conseguir un nuevo conjunto de variables incorrelacionadas llamadas Componentes Principales. Estas componentes se calculan

como una suma ponderada de las variables originales, por tanto, las variables iniciales contribuyen, en mayor o menor medida, de acuerdo con su ponderación, en cada una de las variables latentes (Amat, 2017).

Los individuos tienen un valor en cada variable y esto se puede pensar como que están representados en el espacio de dichas variables. La intención del PCA es representar los individuos en un espacio de con menos dimensiones (variables) que el espacio inicial con la menor pérdida de información posible. Y normalmente este espacio se suele elegir de dimensión 2 por motivos de representación y visualización. Es decir, cada una de las 2 componentes que se elegirán, retendrán la mayor variabilidad posible de los datos. Esto se debe a que las componentes principales, se van construyendo según el orden de importancia en cuanto a la información total que recogen de la muestra.

Para que el uso de esta técnica sea eficaz, se necesita que tanto los individuos como las variables estén correlacionadas, por ello, antes de realizar el PCA se va a comprobar el grado de correlación entre ellas y su significación.

5.4.2 CORRELACIONES

El término correlación se define como una medida de relación lineal entre dos variables cuantitativas continuas (Vinuesa, 2016). Dos variables se dicen están correlacionadas si cambian de manera conjunta y a una tasa constante. Es importante destacar que la correlación no involucra obligatoriamente causalidad, es decir, la causalidad representa una relación causa-efecto mientras que la correlación una relación de similitud(Maxwell et al., 2022).

En función de las características que presenten los datos, se utilizan distintos coeficientes para medir la correlación entre dos variables cuantitativas. Los más comunes son el coeficiente de correlación de Pearson, el de Spearman o el de Kendall. En este estudio se aplica el coeficiente de correlación de Pearson debido a que es la técnica más robusta para datos normalizados y se tiene un número elevado de variables para cada muestra.

Dadas dos variables aleatorias continuas X e Y, se define el coeficiente de correlación muestral de Pearson ($\rho_{X,Y}$,) como:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Fórmula 2. Coeficiente de correlación de Pearson

siendo σ_{XY} la covarianza de X e Y y σ_X, σ_Y las desviaciones típicas de las variables X e Y respectivamente.

Este coeficiente toma valores entre -1 y +1. Los valores próximos a 1; indican una correlación fuerte y positiva, mientras que los valores próximos a –1 indican una correlación fuerte y negativa. Los valores próximos a cero indican que no hay correlación lineal. Aparte del valor del coeficiente de correlación, se necesita comprobar su significación estadística. Aun teniendo un coeficiente muy elevado, si la prueba resulta no significativa, no se tienen evidencias para afirmar que la correlación es real y no viene dada por simple aleatoriedad. A partir de la función de R "cor.test" proveniente del paquete "stats" (R Core Team, 2022), se hallan los p-valores de cada una de las correlaciones, bajo las siguientes hipótesis:

H0: No existe correlación entre las variables
H1: Existe correlación entre las variables

Este contraste se basa en el coeficiente de correlación de Pearson $\rho_{X,Y}$ y sigue una distribución t con 2 grados de libertad si las muestras siguen distribuciones normales independientes.

Para representar de manera rápida y visual las correlaciones entre los individuos, se va a realizar un gráfico con la función de R "corrplot", que permite acceder a amplia gama de formatos en función de los gustos del investigador (Taiyun Wei et al., 2021). En este caso, se trata de una visualización de puntos, que varían en tamaño y color. Cuanto mayor tamaño presentan, más fuerte es la correlación entre esas dos variables. Además, si el punto es azul, determina que esa relación es positiva, las variables son directamente correlacionadas y, por el contrario, si el color es rojo, se dice que las variables tienen una correlación negativa, son inversamente correlacionadas. Cuando aparece una cruz significa que, para ese par de variables su coeficiente de correlación no es significativo (p-valor > 0,05) y, por tanto, se considera que es nula la correlación.

Antes de comenzar con la interpretación del gráfico, se deben especificar las nomenclaturas usadas: *Naive* "N", *Centroblastos* "CB", *Centrocitos* "CC", *Memoria* "M", *Células Plasmáticas* "PC" y *Leucemia* "LLC". Se observa que las correlaciones de la muestra 1 de *Células Plasmáticas* son no significativas con casi todas las demás muestras. Por lo general, los 31 individuos a excepción de los pertenecientes a *Células Plasmáticas* presentan correlaciones moderadas o altas (Figura 16).

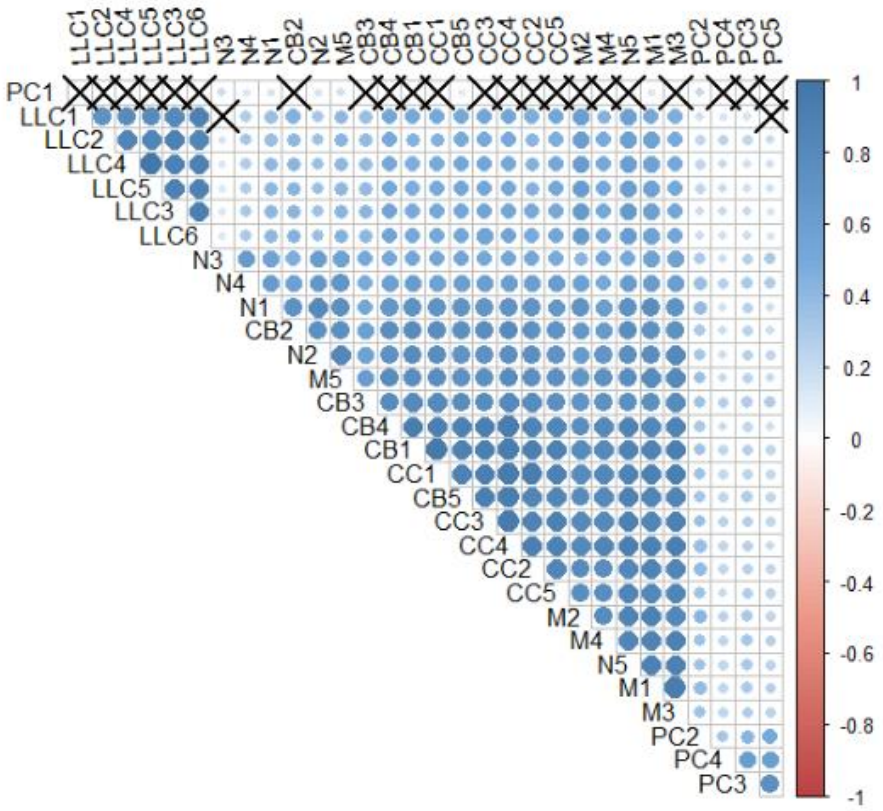


Figura 16.Gráfico de correlaciones ente todas las muestras del estudio.

Para poder visualizar correctamente las correlaciones entre las proteínas, se utiliza un heatmap ya que el gráfico corrplot no es adecuado cuando queremos representar un alto número de correlaciones. Un heatmap es una representación gráfica de una matriz numérica en el que se colorea cada celda con una escala continua de color, proporcional al valor de la proteína en cada individuo (Amat, 2017).

En este caso, se representa la matriz de correlaciones (Figura 17) entre proteínas, que es simétrica y en su diagonal está la correlación de una proteína consigo misma (valor 1). Se identifican grupos de proteínas con correlaciones directas altas (en rojo), grupos de proteínas con correlaciones inversas altas (en azul) y muy pocos espacios blancos que serían grupos de proteínas sin correlación.

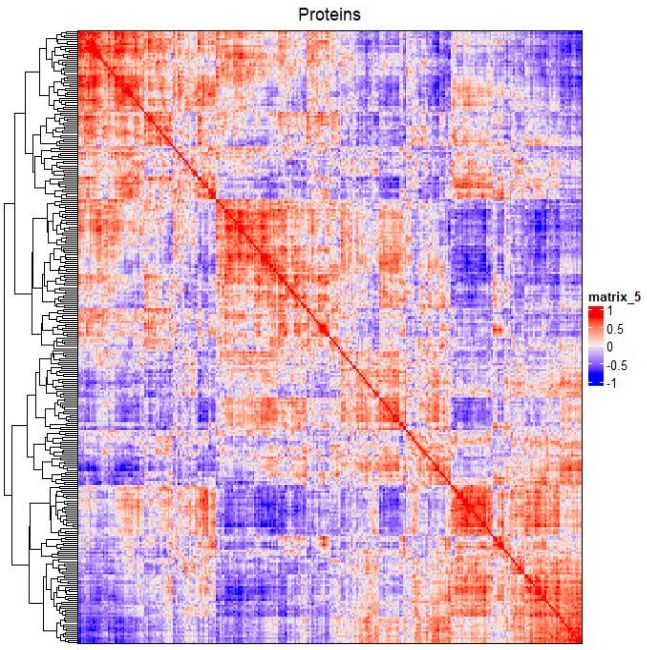


Figura 17. Gráfico de correlaciones ente las proteínas.

Una vez comprobado que las correlaciones entre las proteínas parecen prácticamente todas significativas y mayoritariamente altas, se puede decir que el PCA es viable. Con ayuda de R y los paquetes "stats" (R Core Team, 2022) y "factoextra" (Kassambara & Mundt, 2020) se han efectuado las siguientes salidas sobre el análisis de componentes principales.

En primer lugar, se crea un gráfico de barras que dará una idea clara del porcentaje de variabilidad que explican las 10 primeras componentes principales de manera no acumulativa.

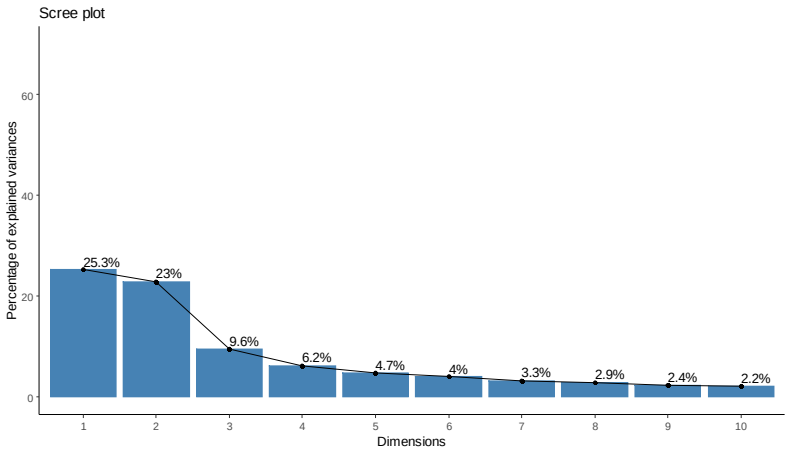


Figura 18. Gráfico de sedimentación.

Se debe tener en cuenta el número de dimensiones de las que se partían (302 proteínas), para realmente comprender como de bien o mal se están representando los individuos en las componentes

que elijamos, que en este caso serán 2. Si el porcentaje de variabilidad explicado por las 2 primeras componentes fuera muy pequeño está claro que la representación no sería muy fiable. En nuestro caso, la variabilidad total de los datos era explicada por 302 variables, y las dos primeras componentes principales explican prácticamente la mitad (48.3%) de la variabilidad total, luego es síntoma de una representación fiable el elegir sólo 2 componentes (Figura 18).

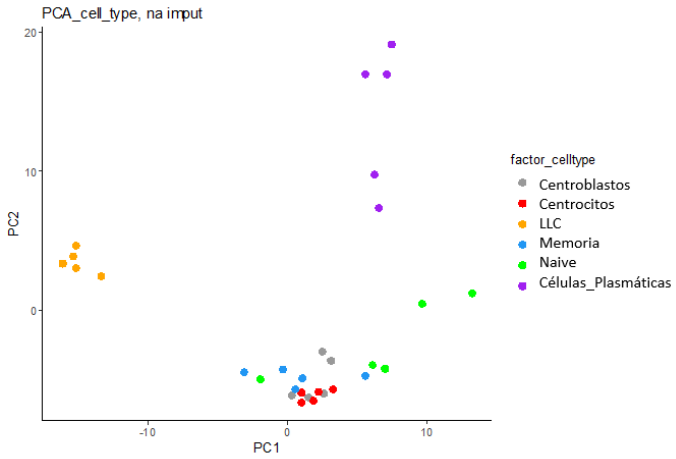


Figura 19. PCA de los 6 tipos celulares.

En el PCA se observan tres agrupaciones o clústeres (Figura 19). El primero está compuesto por las muestras de *LLC*, que están bastante agrupadas entre ellas y presentan valores muy negativos en la primera dimensión y valores entorno al 0 en la segunda. El siguiente clúster contiene las muestras de células plasmáticas que están dispersas entre ellas y tienen valores altos en ambas dimensiones. Por último, el tercer grupo está formado por las muestras de *Naive, Centroblastos, Centrocitos y Memoria*. Estos individuos son bastante homogéneos entre ellos y se sitúan en torno al 0 en la primera componente y con valores bajos en la segunda. Cabe destacar la presencia de dos individuos pertenecientes a las muestras de *Naive*, que sí que presentan diferencias con el resto.

5.4.3 HEATMAP

Otra manera de visualizar esta estructura de agrupamiento que se observa en el PCA es construir un heatmap de la matriz de valores de intensidad de las proteínas en las muestras y añadir sendos dendrogramas correspondientes a agrupamientos jerárquicos (Amat, 2017) realizados sobre las muestras (filas) y sobre las proteínas (columnas).

Un dendrograma es una técnica de clúster jerárquico, técnica no supervisada, que tiene como objetivo determinar patrones o agrupaciones dentro de un conjunto de datos (Amat, 2017). Estos patrones se establecen utilizando medidas de distancia o criterios de enlace, que permiten cuantificar la semejanza o diferencia entre las observaciones. Así, una agrupación está formada por un conjunto de observaciones parecidas entre ellas y diferentes al resto. Al ser una técnica de clustering jerárquico, no es necesario determinar a priori el número de clústeres que se van a formar y las agrupaciones siguen una estructura de árbol donde, las agrupaciones formadas en los niveles superiores contienen a los grupos de los niveles inferiores.

Con el paquete de R "ComplexHeatmap" procedente de Bioconductor (Gu et al., 2016) se realiza el heatmap mostrado a continuación (Figura 20). Se observan los tres clústeres de las muestras que comentamos anteriormente y dos clústeres de las proteínas correspondientes a las dos componentes principales.

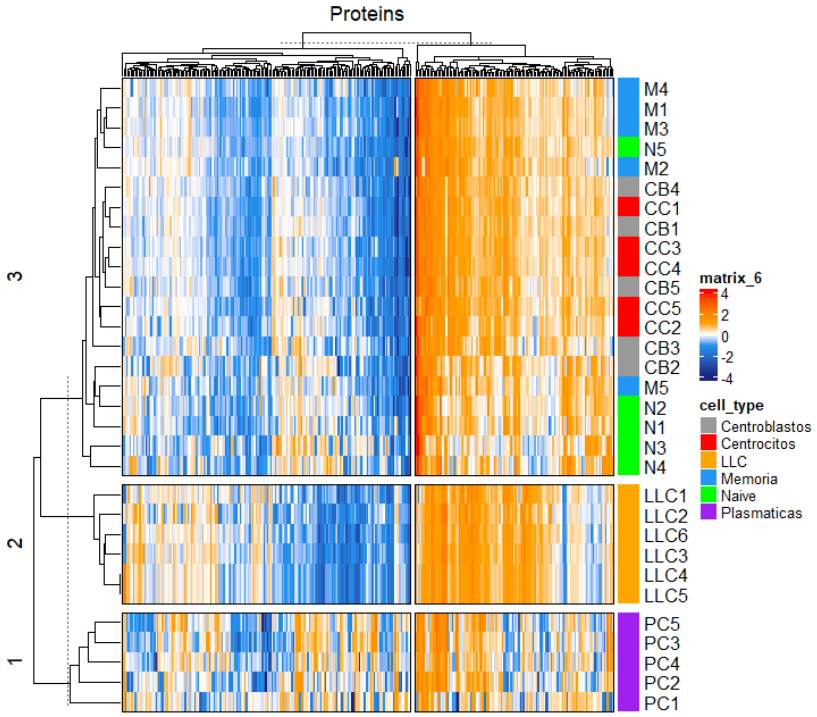


Figura 20. Heatmap matriz de expresiones

El estudio se continúa realizando comparaciones dos a dos entre las muestras de *LLC* versus las muestras de cada uno de los 5 estadios del linfocito B por separado, es decir, un total de 5 comparaciones. Debido a las limitaciones en el número de páginas, se van a exponer los resultados de un único estadio, el considerado más interesante desde el punto de vista biológico son las muestras de *Centroblastos*. El resto de las comparaciones están recogidas en los anexos.

5.5 DIAGRAMAS DE VENN *LLC* vs Estadios Linfocito B

En este apartado, se van a exponer un gráfico Upset (Conway et al., 2017)y un diagrama de Venn, ambos definidos anteriormente, para representa el número de proteínas comunes entre las muestras de *LLC* y *Centroblastos* . El resto de los gráficos Upset y diagramas de Venn comparando las muestras de *LLC* con cada uno de los estadios del Linfocito B se reflejan en los anexos.

Se obtiene que de las 2434 proteínas con valores de intesidad en las muestras de leucemia (*LLC*), 1730 (54%) de ellas tambien pertenecen a las muestras de *Centroblastos*. Por tanto, hay 704 (22%) proteinas perteneceintes unicamente a las muestras de *LLC* y 763 (24%) proteinas presenten solamente en *Centroblastos* (Figura 21).

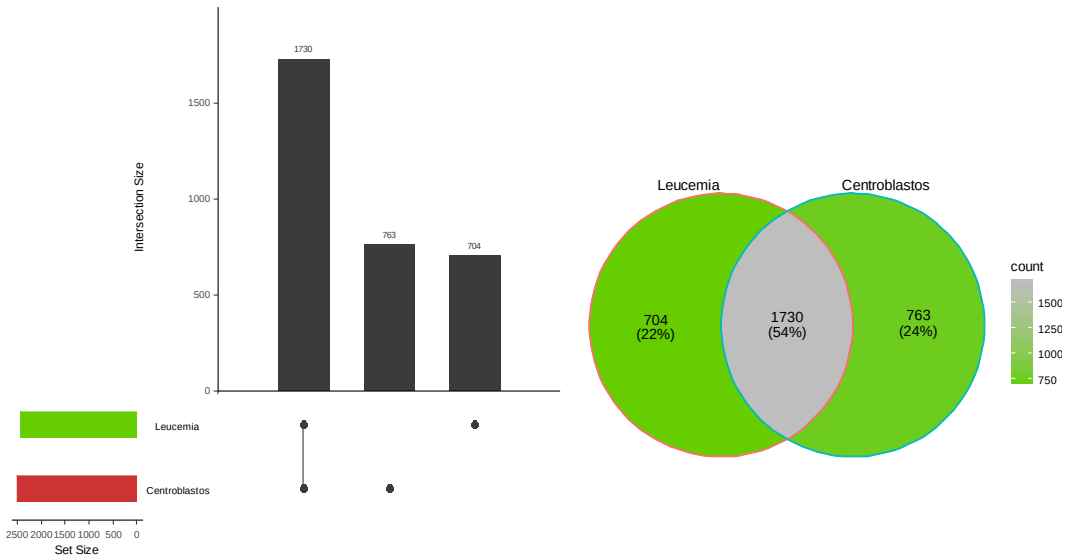


Figura 21. Gráfico Upset y Diagrama de Venn para muestras de LLC vs Centroblastos

5.6 ALGORITMO SAM

5.6.1 VENTAJAS DEL USO DE TÉCNICAS DE BIOINFORMÁTICA MODERNAS

En este estudio, para encontrar expresión diferencial entre un alto grupo de proteínas, se hace uso de la técnica bioinformática SAM, "Significance analysis of microarrays" (Ayala, 2020). Este método permite detectar, qué cambios de valores de intensidad de las proteínas son significativos entre individuos de dos o más grupos (Figura 22).

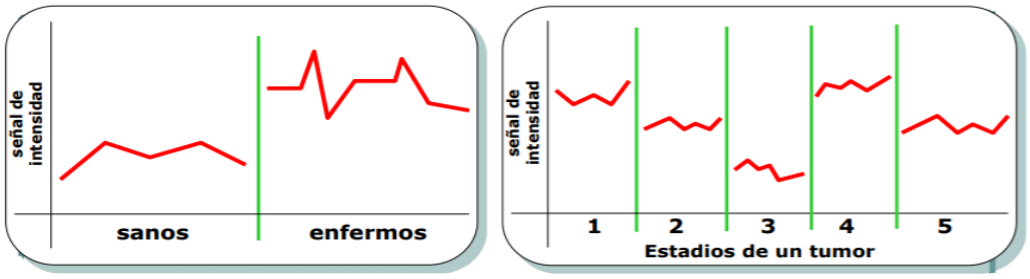


Figura 22. Diferencias en la señal de intensidad de un gen entre grupos

Esta técnica sustituye a los métodos tradicionales, agregando una serie de ventajas. Por un lado, permite trabajar con grandes cantidades de proteínas, y analizarlas a la vez. Habitualmente, se tiene un numero grande de variables (proteínas), frente a una cantidad mucho menor de muestras. Esto para el algoritmo SAM no es un problema, pero sí para técnicas clásicas de modelización de variables como la regresión, donde es imposible estimar el modelo a partir de mínimos cuadrados, ya que, no se puede invertir la matriz de dispersión (matrices no cuadradas).

A su vez, SAM no necesita verificar que los datos cumplan con suposiciones previas como la presencia de normalidad o la igualdad de varianzas, al contrario que para muchos de los test estadísticos estándar para los que el incumplimiento de estas características hace que estos test no sean

robustos y eficientes. Se sabe que aplicando el logaritmo en base dos a los datos, se consigue mejorar la asimetría, pero las varianzas seguirían siendo diferentes y, por tanto, muchos de los test como el ANOVA, no serían útiles al tratarse de datos heterocedásticos.

Como se ha mencionado anteriormente, el interés se centra en examinar la variación significativa de las proteínas entre 2 o más grupos, también llamada expresión diferencial entre grupos Para ello, se podrían usar técnicas de contraste para k muestras independientes como t-test, ANOVA o Kruskal-Wallis. Dado un conjunto de proteínas con i =1, …, N la hipótesis para la i-ésima proteína es la siguiente:

Ho : La proteína i no tiene expresión diferencial entre los grupos considerados.

H_1: La proteína i tiene expresión diferencial entre los grupos considerados.

Si se rechaza la hipótesis nula, se podría decir que la proteína tiene distinto valor de intensidad entre al menos dos grupos y, se consideraría significativa para el estudio. Si, por el contrario, se acepta la hipótesis nula, la proteína tendría la misma intensidad para todos los grupos y no sería significativa.

El uso de estas técnicas requiere el control de los errores del contraste, puesto que, al realizarlas reiteradamente, cada vez que se considera una proteína significativa, se irá acumulando un error tipo I (según define Ayala (2022) no hay expresión diferencial, pero de forma errónea se determina que sí hay) (Tabla 5).

Tabla 5. Errores tipo I y II en contraste de múltiples hipótesis

	DECISIÓN ESTADÍSTICA	
REALIDAD	**Proteína no cambia (Aceptar H_0)**	**Proteína cambia (Rechazar H_0)**
Proteína no cambia (H_0 Verdadera)	Acierto No significativa	Error tipo I (α) Falso positivo
Proteína cambia (H_0 Falsa)	Error tipo II (β) Falso negativo	Acierto Significativa

La probabilidad del error de tipo I se denota como α y se denomina nivel de significación de la hipótesis nula, y la probabilidad del error de tipo II se denota con α. Por ejemplo, si se fija un nivel de significación de α=0,01 y ejecutamos un t-test para 10.000 proteínas individualmente, se esperan 100 falsos positivos, se están considerando 100 proteínas significativas que realmente no lo son. Debido a la gran acumulación de errores tipo I, estos contrastes no resultan del todo eficaces. Por ello, es recomendable el uso de técnicas más modernas, como el algoritmo SAM, que logra suprimir esta problemática con el cálculo de un error global de tipo I, denominado" False Discovery Rate" (FDR).

Dada la variable aleatoria $Q = \frac{Nº\ de\ pruebas\ rechazadas\ siendo\ verdaderas\ (F)}{Nº\ de\ pruebas\ rechazadas\ (R)}$, se define la "tasa de falsamente rechazados" (FDR) como la esperanza de Q.

$$FDR = E(Q) = E \cdot P(R > 0)$$

Fórmula 2. Tasa de falsamente rechazados

Determina la proporción esperada de hipótesis erróneamente rechazadas entre aquellas hipótesis que hemos rechazado (Ayala, 2020). Dado un FDR de 0,05 se dice que el 5% de las pruebas nulas que rechazamos en realidad son ciertas.

Es necesario mencionar la diferencia entre el FDR y la tasa de falsos positivos para evitar posibles confusiones. Se define la tasa de falsos positivos como la proporción esperada de pruebas rechazadas siendo ciertas entre todas las hipótesis que son verdaderas. Por tanto, si se tiene una tasa de falsos positivos de 0,05, en promedio el 5% de las hipótesis nulas ciertas serán rechazadas por error.

5.6.2 METODOLOGÍA

SAM está diseñado para detectar un grupo de proteínas expresadas diferencialmente, es decir, un conjunto que represente señales de intensidad diferentes en los grupos de la variable respuesta, determinando previamente el FDR.

Este algoritmo guarda en una matriz las señales de intensidad de las proteínas en las muestras del estudio. Dicha tabla representa en las filas a las variables (proteínas) y en las columnas a los individuos. De esta forma, se dice que el valor que ocupa la i-ésima fila (i=1…n) y la j-ésima columna (j=1…N), representa la señal de intensidad de la proteína i en la muestra j (Tabla 6). Además, se tiene una variable respuesta Y=(y_j) (j=1…n) que agrupa a los n individuos en función de una característica. En nuestro trabajo, esta variable vendrá definida por las muestras de *LLC* vs cada uno de los estadios del Linfocito B, es decir, tomará valor 1 si el individuo es *LLC* y 2 si el individuo es de otro estadio.

Tabla 6. Matriz de señal de intensidad SAM

	Individuos					
Proteínas	**1**	**2**	**…**	**j**	**…**	**n**
1	x_{11}	x_{12}	…	x_{1j}	…	x_{1n}
2	x_{21}	x_{22}	…	x_{2j}	…	x_{2n}
…	…	…	…	…	…	…
i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{in}
…	…	…	…	…	…	…
N	x_{N1}	x_{N2}	…	x_{Nj}	…	x_{Nn}
Y_j	Y_1	Y_2	…	Y_j	…	Y_n

Se expone la metodología para el caso de una variable respuesta formada por 2 grupos independientes: las muestras de *LLC*, consideradas como pacientes o enfermos y uno de los estadios del linfocito B por separado, definidos como controles o sanos.

Dado un conjunto de muestras $\{1, \dots, n\}$ y un vector de agrupación $Y = (y_1, \dots, y_n)$, la muestra j-ésima pertenecerá al grupo 1 si y_j toma el valor 1 y por el contrario, formará parte del grupo 2 si y_j es igual a 2. Se considera J_g al conjunto de muestras pertenecientes al grupo g donde $g = 1,2$ y n_1, n_2 serán los cardinales de J_1 y J_2 respectivamente. A partir de aquí, se puede denotar:

$$\bar{x}_{i,J_g} = \frac{\sum_{j \epsilon J_g} x_{ij}}{n_g} \text{ con g=1,2}$$

Fórmula 3. Valor medio de las proteínas para cada grupo, enfermos vs sanos

Así pues, para la i-ésima proteína se calcula la diferencia relativa definida como:

$$d_i = \frac{\bar{x}_{i1} - \bar{x}_{i2}}{s_i + s_0}$$

Fórmula 4. Estadístico de contraste

donde

$$s_i = \sqrt{a \sum_{g=1}^{2} \sum \left(x_{ij} - \bar{x}_{i,g}\right)^2} \text{ siendo } a = \frac{\frac{1}{n_1} + \frac{1}{n_2}}{n_1 + n_2}$$

Fórmula 5. Desviación típica de la diferencia de medias de expresión de la proteína.

Como se puede observar en la fórmula 5, el parámetro s_i, es una ponderación de las desviaciones s de la expresión de cada proteína a la media de expresión de la proteína i en cada grupo.

El parámetro s_0 es una constante que corrige la influencia de las proteínas con baja intensidad, en la varianza de d_i. Esto se debe a que el estadístico de contraste d_i, va a tener un valor muy grande si la desviación típica de la diferencia de medias s_i es muy pequeña. Por esta razón es necesario sumarle a esta desviación típica (s_i) un valor constante s_0 que estabilice la varianza y así hacer comparables los valores d_i entre sí. Este valor se calcula a partir del propio conjunto de datos.

A continuación, se va a definir el método SAM secuencialmente (Ayala, 2020):

1. Se obtiene a partir de la ecuación anterior (Fórmula 4) la diferencia relativa d_i para cada una de las proteínas del estudio.

2. Se ordenan de forma ascendente los $d_i's$ obtenidos: $d_1 \leq \cdots \leq d_N$.

3. Se realizan P permutaciones aleatorias entre las n muestras divididas en los dos grupos dados por el vector de agrupación (Y). Como Y viene definido por dos grupos, con n_1 sanos y n_2 enfermos, el número total de permutaciones se obtiene a partir del coeficiente multinomial:

$$P = \frac{n!}{n_1! \cdot n_2!}$$

Fórmula 6. Número de permutaciones posibles

Las P permutaciones son agrupaciones de los n individuos en los grupos de sanos y pacientes, no variando el tamaño original de los grupos n_1, n_2.

4. Para cada permutación P y cada proteína i se calculan todos los valores d_i de la fórmula 4, lo que nos proporciona las distribuciones empíricas de cada d_i, que denotamos como $d_i(p)$. Todos estos valores nos proporcionan una distribución de probabilidad con la que calcular los momentos, en concreto la media. Este método de permutaciones (Berry et al., 2011) nos proporciona la distribución esperada debida al azar del estadístico d_i.

5. Se halla el promedio de las P distribuciones empíricas $d_i(p)$

$$\bar{d}_i(p) = \sum \frac{d_i(p)}{P}$$

Fórmula 6. Valor medio de las diferencias relativas empíricas

Para las proteinas donde la diferencia relativa observada d_i sea muy similiar a la diferencia relativa esperada $\bar{d}_i(p)$, se tiene una expresion diferencial debida al azar y por tanto no debemos aceptar que la proteina cambia entre los grupos.

6. Para considerar hasta que punto es similar $d_i \approx \bar{d}_i(p)$, se fija un valor positivo Δ como umbral (Figura 23), de tal modo que si:

$$|d_i - \bar{d}_i(p)| \leq \Delta \rightarrow Se\ acepta\ H_0$$
$$|d_i - \bar{d}_i(p)| > \Delta \rightarrow Se\ rechaza\ H_0$$

siendo,

H_0: La proteína i-ésima no tiene expresion diferencial, no cambia entre grupos.
H_1: La proteína i-ésima tiene expresion diferencial, cambia significativamente entre grupos.

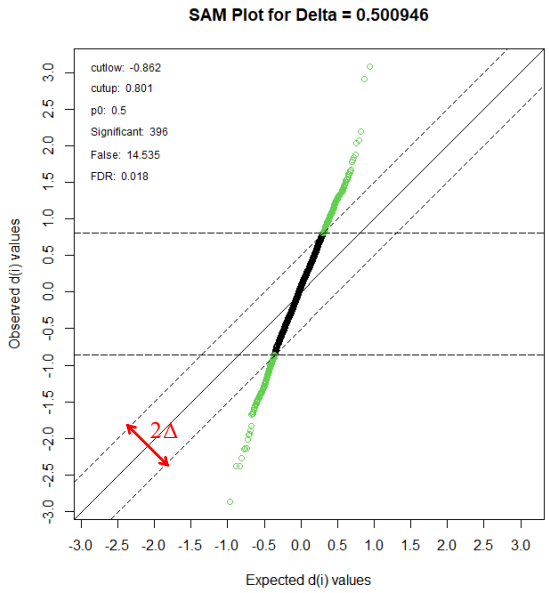


Figura 23. Gráfico SAM para un Delta de 0,500946

Para un umbral Δ, se puede determinar el número de proteinas significativas y estimar la tasa de falsamente rechazados (FDR). Cabe destacar que, estos términos estan inversamente relacionados, cuanto más aumenta uno, más disminuirá el otro, ya que si aumentamos el umbral axceptamos más proteinas no significativas y habrá menos falsas positivas (Figura 24).

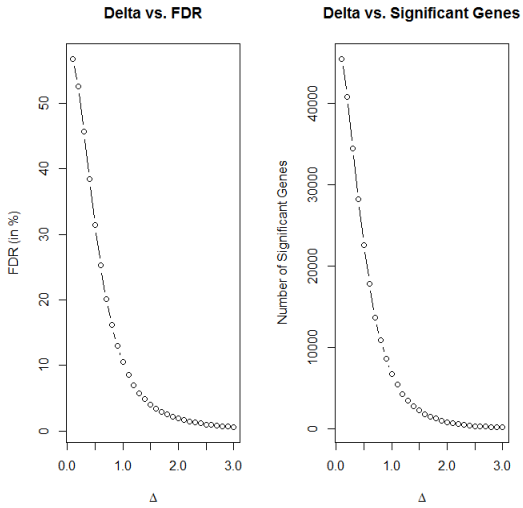


Figura 24. Gráficos Delta

7. La estimación del FDR viene dada por el cociente entre el número medio de falsos positivos en las permutaciones y el número de proteinas consideradas significativas (positivas).

   Se denota :
   - t_1 como el menor valor i tal que $d_i - \overline{d}_i(p) \geq \Delta$
   - t_2 como el mayor valor i tal que $d_i - \overline{d}_i(p) \leq \Delta$

Una vez calculados estos índices, se calcula el número medio de falsos positivos en las permutaciones:

$$\frac{1}{\#P}\sum_{p=1}^{P} \#\{i | d_i(p) \geq d_{t1} \cup d_i(p) \leq d_{t2}\}$$

Y el número de positivos o proteinas significativas será:

$$\#\{i | d_i \geq d_{t1} \cup d_i \leq d_{t2}\}$$

Por tanto, la estimacion de la tasa de falsamente rechazados (FDR) es:

$$FDR \approx \frac{\frac{1}{\#P}\sum_{p=1}^{P} \#\{i | d_i(p) \geq d_{t1} \cup d_i(p) \leq d_{t2}\}}{\#\{i | d_i \geq d_{t1} \cup d_i \leq d_{t2}\}}$$

En resumen, el algoritmo SAM obtiene un valor d_i para cada proteína i, que mide el grado de dependencia entre la expresión de la proteína y la variable respuesta Y. Se basa en permutaciones aleatorias de los datos para ver si los valores de intensidad de las proteínas se asocian significativamente con la variable de agrupación. El investigador determina el parámetro delta (umbral), basado en el FDR, es decir, en función del porcentaje de error tipo I que se esté dispuesto a asumir se determina un valor delta u otro. Cuanto menor sea el valor delta mayor cantidad de genes significativos se tiene y por tanto mayor cantidad de falsos positivos.

5.6.3 SAM EN R

En este trabajo se utiliza el algoritmo SAM con R y RStudio (interfaz del lenguaje de programación R). Estos programas son software libre y se mantienen en constante desarrollo, ya que R está formado por una gran variedad de paquetes almacenados en repositorios de acesso libre como CRAN y Bioconductor. CRAN es el repositorio oficial formado por los paquetes más generales, mientras que Bioconductor es específico para la bioinformática; de este último se extraerán todos los paquetes necesarios para ejecutar este algoritmo.

Para la aplicación de esta técnica, es necesario detallar los argumentos que utiliza la función SAM en R (Maintainer & Schwender, 2022):

- data: data frame donde cada fila corresponde a una variable (proteína) y cada columna a una muestra.

- cl: vector de agrupación de longitud igual al número de columnas. Contiene la etiqueta de clase de cada individuo, en este caso es un vector que contiene 0 para los enfermos y 1 para los sanos.

- method: especifica el método por el cual se calculan las puntuaciones (d_i). Hay tres opciones:
    - "d.stat": se usa para matrices de expresión numérica, si son dos las categorías de la variable respuesta, se utiliza un estadístico t asumiendo que las varianzas entre los grupos son iguales, o el estadístico t de Welch si las varianzas son diferentes. Si, por el contrario, la variable tiene más de dos categorías se aplica un estadístico F.
    - "wilc.stat" : se emplea si se tiene el rango de la expresión numérica, utiliza la prueba de los rangos con signo de Wilcoxon
    - "chisq.stat": es la opción para un análisis de datos categóricos, se calcula con la Chi-cuadrado de Pearson para cada fila.

- var.equal: valor lógico, TRUE presencia de homocedasticidad en las variables (proteínas) y FALSE ausencia de homocedasticidad.

- B: número de permutaciones para calcular la distribución empírica, permuta la variable grupos para las n muestras y hace la distribución del estadístico d.stat para todas esas permutaciones.

- rand: fija un numero aleatorio para realizar siempre las mismas permutaciones y que los resultados no varíen de un usuario a otro.

- control: con la función "samcontrol" se determina el delta, viene definida por los siguientes parámetros:
    - delta: vector que define los valores para el umbral delta (Δ). Es una secuencia, ya que a simple vista es imposible saber qué valor delta se necesita para obtener el FDR indicado
    - p0: es la probabilidad de que una proteína sea significativa o no. Se pone por defecto el peor de los casos p0 = 0.5 (50%), probabilidad de que la proteína cambie.

```
>salida_sam <- sam (LLC_CENTROBLASTOS, factor_celltype_5,
          method = d.stat, var.equal = T, B = 600, rand = 9999,
          control = samControl (delta = seq (0.1,6.0,0.2), p0 = 0.5))
```

R devuelve un objeto propio de la función SAM compuesto por los siguientes elementos:

- d: conjunto de puntuaciones observadas d_i para cada proteína.
- d.bar: conjunto de puntuaciones esperadas $\overline{d_i}(p)$ para cada proteína
- vec.false: número esperado de falsos positivos de cada proteína en un único umbral
- p.value: p-valores de las proteínas, obtenidos al realizar el contraste de hipótesis (punto 6)
- s: ponderación de las desviaciones de la diferencia de las medias de expresión de cada proteína entre los grupos
- s0: valor de la constante que corrige la influencia de las proteínas con baja intensidad
- mat.samp: matriz que representa en cada fila una permutación y en cada columna un individuo, contiene para cada permutación el grupo de la variable respuesta asignado a cada individuo.
- p0: probabilidad de que una proteína sea significativa (establecido con anterioridad)
- mat.fdr: matriz que contiene la información general sobre el conjunto de proteínas significativas, viene definida por:
    - Una secuencia de deltas
    - El p-valor establecido para considerar una proteína significativa o no, (p0 = 0,5)
    - Número medio de proteínas falsamente rechazadas
    - FDR: tasa de falsamente rechazados
    - cutlow /cutup: puntos de corte para determinar el umbral delta
    - j1/j2:
- q.value: valor que asocia a cada test un valor numérico que diga lo extremo que es su p-valor considerando el resto de p-valores.
- fold: r.fold de cada proteína, cuanto varia en proporción una proteína en función del grupo al que pertenezca
- msg: contiene información general del análisis de forma resumida
- chip: nombre el microarray utilizado, si no hay información aparece como “”

El propio paquete SAM proporciona una serie de funciones, que permiten completar los análisis. Las funciones que se han utilizado son (Maintainer & Schwender, 2022):

- findDelta: es una función que extrae una tabla los FDR más próximos al impuesto por el investigador, con sus respectivos valores de delta y el número de proteínas significativamente expresadas.

```
>findDelta (salida_sam, FDR)
```

- summary: resume los resultados de un análisis SAM para un delta determinado

```
>sam.out <- summary (salida_sam, DELTA)
```

- plot: genera un gráfico SAM si se especifica el delta (Figura 16) o los gráficos delta si no se especifica (Figura 17)

```
>plot (salida_sam, DELTA)

>plot (salida_sam)
```

APLICACIÓN PRÁCTICA SAM

En este estudio se ha ejecutado el algoritmo SAM para 5 conjuntos de muestras:

1) LLC vs Naive
2) LLC vs Centroblastos
3) LLC vs Centrocitos
4) LLC vs Células de Memoria
5) LLC vs Células Plasmáticas

Se han realizado estas comparaciones, ya que, el interés se centra en encontrar conjuntos de proteínas con niveles de expresión significativamente diferentes entre dos grupos de individuos, sanos (cada estadio del linfocito B) frente enfermos (muestras de *LLC*). Se van a exponer los resultados obtenidos para el grupo 2, ya que, es el grupo que biológicamente más interesa, el resto de los análisis se encuentran en el anexo.

En primer lugar, se crea una base de datos con los valores de intensidad de las proteínas comunes en las muestras de *LLC* y *Centroblastos*, se encuentra un total de 1730 proteínas conjuntas para al menos una muestra de cada tipo celular. Además, se aplica el log2 para reducir la escala y se tipifica (Fórmula 1) para que las muestras del conjunto de datos puedan ser comparables.

Antes de aplicar la función SAM, se requiere crear un factor que asocie a cada muestra, el tipo celular al que pertenece, en este caso se tienen 5 muestras de *Centroblastos* y 6 muestras de *LLC*. Hay que tener cuidado con el nombre de las categorías del factor, porque R va a asignar al 0 el nombre que alfabéticamente este antes y, por tanto, puede cambiar la interpretación para cada conjunto de muestras. En este estudio, se va a forzar que las muestras de *LLC* siempre estén asociadas al primer nivel del factor, es decir, al valor 0.

Al tener una matriz de expresión numérica, conviene utilizar el método "d.stats" para calcular las puntuaciones (d_i). Este método emplea pruebas paramétricas, son pruebas útiles en este análisis, ya que, la base de datos ha sido previamente normalizada. Si solo se tienen dos categorías para la variable respuesta, este método utiliza el estadístico t si hay homocedasticidad en la mayoría de las proteínas o el estadístico t de Welch si hay heterocedasticidad.

Por ello, se requiere realizar un contraste para determinar si las proteínas presentan igualdad de varianzas entre las muestras, al tener presencia de normalidad se aplica el test de Bartlett, teniendo como hipótesis (Hossein Arsham, Lovric, 2011):

Ho: las muestras presentan varianzas iguales

H_1: al menos una muestra presenta varianzas distintas

El estadístico de contraste para la prueba de Bartlett se calcula a partir de la siguiente fórmula:

$$X^2 = \frac{(N-k)\ln\left(S_p^2\right) - \sum_{i=1}^{k}(n_i - 1)\ln\left(S_p^2\right)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i - 1} - \frac{1}{N-k}\right)\right.}$$

Siendo:

- k el nº de muestras
- n_i el tamaño de cada muestra
- $N = \sum_{i=1}^{k} n_i$
- $S_p^2 = \frac{1}{N-k}\sum_i (n_i - 1)S_i^2$ donde S_i^2 es la varianza de la muestra i

Sigue aproximadamente una distribución Chi-cuadrado con $k - 1$ grados de libertad.

Se tiene un total de 1435 proteínas con un p-valor mayor a 0,05 frente a 295 proteínas con un p-valor menor a 0,05. Por tanto, se considera que, globalmente, las proteínas presentan varianzas iguales, hay presencia de homocedasticidad.

Una vez comprobado este requisito, se ejecuta el algoritmo SAM con 600 permutaciones, una semilla aleatoria 24 y con un rango de deltas de 0,1 a 6,0 (de 0,2 en 0,2). Se obtiene el siguiente resultado (Tabla 7):

Tabla 7. Tabla salida SAM

Delta	p0	False (nº estimado de proteínas falsamente rechazadas)	Called (nº proteínas significativas)	FDR
0,1	0,5	938,158	1422	0,3298
0,3	0,5	126,197	790	0,0798
0,5	0,5	14,535	396	0,0183
0,7	0,5	1,714	159	0,0053
0,9	0,5	0,223	44	0,0025
…	…	…	…	…
5,9	0,5	0	0	0

El FDR elegido ha sido 0,0053, con este valor se dice que el 0,053% de las hipótesis nulas que se rechazan son ciertas. Habría 159 proteínas significativas y una media de 1,714 falsos positivos, un 0,053% (FDR). Se ha elegido este FDR con el objetivo de encontrar un número considerable de proteínas y pocos falsos positivos. Con la función "FindDelta" se encuentra para el FDR más a aproximado a 0,0053 el valor delta correspondiente (Tabla 8).

Tabla 8. Tabla de la función FindDelta

Delta	Called (nº proteínas significativas)	FDR
0,703954	156	0,005363
0,703955	155	0,005300

En este caso se ha encontrado el FDR igual a 0,0053 con un Delta de 0,703955. Se tendrán 155 proteínas que tienen expresión diferencial entre las muestras de *LLC* y *centroblastos*.

Por último, se obtiene el resultado de SAM para el valor delta 0,703955. Debido a las dimensiones de la tabla, se van a exponer únicamente 5 de las 155 proteínas con las puntuaciones d_i más altas en valor absoluto, la tabla completa aparece en los anexos (Tabla 9).

Tabla 9. Tabla resumen resultado SAM para un delta concreto

UNIPROT	Row	d.value (d_i)	Stdev (s_i)	Rawp (p-valor)	q.value	R.fold
P68871	923	3,081706	0,172477	1.251157e-06	0.001082	9.928132
P69905	924	2,907580	0,268403	2.502315e-06	0.001082	10.58041
Q09666	1003	-2,864283	0,246133	3,753472e-06	0.001082	0.102318
Q15149	1121	-2,381087	0,116505	5,004629e-06	0,001082	0,171825
P84243	943	-2,380101	0.19606465	6.255787e-06	0.001082	0.163375

Las proteínas vienen ordenadas por el valor absoluto del "d-value", cuanto mayor sea este valor absoluto, mayor es la diferencia de expresión de la proteína entre sanos y enfermos. Cabe destacar, que las proteínas están ordenadas por dicho valor, no por cuánto más o menos se expresa la proteína en las muestras de *LLC* sobre las muestras de *centroblastos*, que es el R.fold.

Esta diferencia ("d.value") puede ser positiva o negativa. Si la diferencia es positiva, como en el caso de "P68871" y "P69905", se dice que los enfermos presentan valores de intensidad superiores que los sanos. Del mismo modo, si la diferencia es negativa, como muestran las proteínas "Q09666", "Q15149" y "P84243", significa que la señal en los enfermos es más pequeña que en los sanos, la leucemia, en este caso, hace que esa proteína se apague, ha reprimido su expresión.

Para cuantificar ese cambio en la señal de intensidad se observa el "R.fold", que es una ratio (cociente) entre las señales de intensidad de los pacientes y los sanos (*LLC* / *centroblastos*). Para la primera proteína "P68871", donde la diferencia 'd-value' es positiva el "R.fold" es de 9.928132, es decir, la expresión ha aumentado en *LLC* un 993% con respecto a las muestras de *centroblastos*, los valores de intensidad en *LLC* son 9.93 veces más grandes que los valores en *centroblastos*. Ahora bien, para el tercer caso "Q09666", donde el "d.value" es negativo se tiene un "R.fold" de 0.102318, lo que quiere decir que la expresión ha disminuido en los pacientes aproximadamente un 10% con respecto a los sanos, los valores de intensidad en *LLC* son 0,1 más pequeños que en *centroblastos*.

A continuación, se muestra el gráfico SAM, donde las proteínas que están en negro no cambian, no son significativas, mientras que las verdes sí son consideradas significativas para un FDR de 0,0053; las que se sitúan por arriba se sobre expresa y si están por debajo se bajo-expresan (Figura 25).

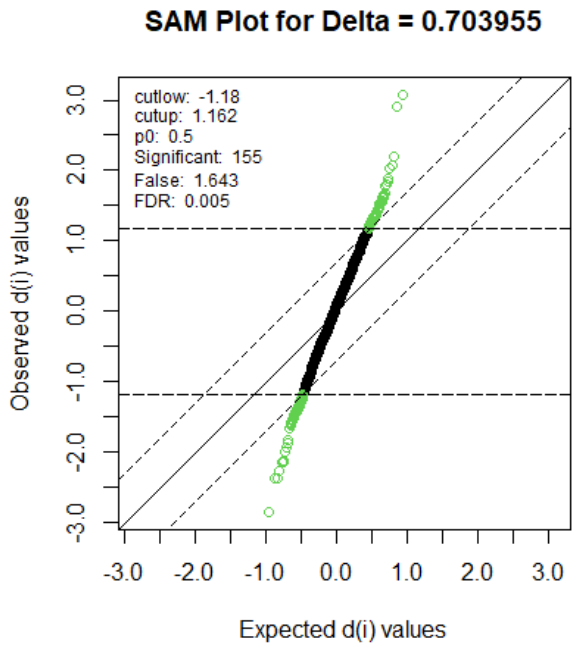


Figura 25. Gráfico SAM para delta = 0,703955

Por último y a modo resumen se presenta un gráfico con el recuento de proteínas que se sobre-expresan y bajo-expresan para las muestras de *LLC* vs cada uno de los estadios del linfocito B (Figura 26). Para el conjunto de proteínas significativas entre las muestras de *LLC* y *centroblastos* se tiene un número muy similar de proteínas sobre- y bajo- expresadas.

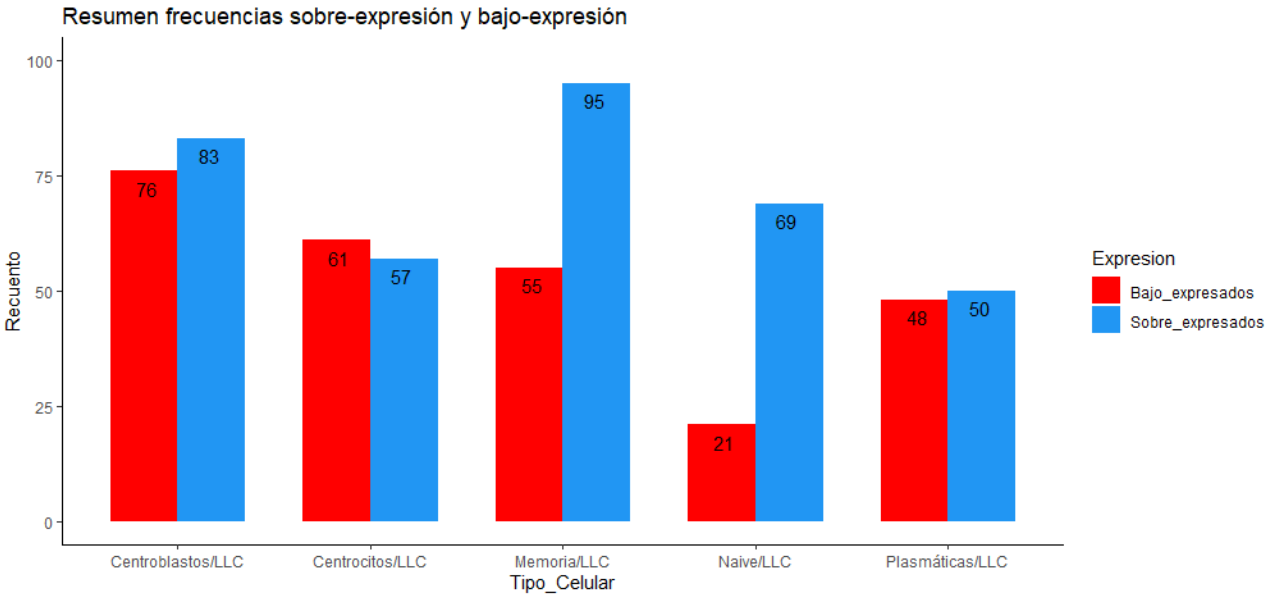


Figura 26. Número de proteínas sobre- y bajo- expresada para cada comparación.

5.7 GSEA

En este apartado, se va a describir el método de Análisis de Enriquecimiento de Conjuntos de Genes (GSEA) para la interpretación de datos de expresión génica. El método se centra en grupos de genes que comparten función biológica, ubicación cromosómica, o procesos regulatorios. Concretamente, GSEA es incorporado en un paquete de software gratuito, junto con un base de datos inicial de firmas moleculares "MSigDB", que cuenta con 1.325 conjuntos de genes biológicamente definidos (Subramanian et al., 2005).

El interés de GSEA es analizar si los genes de una colección perteneciente a "MSigDB", tienden a situarse hacia la parte superior o inferior, de una lista de genes ordenados según la expresión diferencial entre dos grupos. Esta lista es la extraída con la herramienta SAM para las muestras de *LLC* y *centroblastos*. El objetivo es comprobar si es significativo el número de genes encontrados de los extremos, ya que, en este caso podrá haber diferencias fenotípicas.

5.7.1 METODOLOGÍA

Se expone de manera resumida la metodología de GSEA dividida en tres fases (Subramanian et al., 2005):

Dado un conjunto de genes S definido a priori, y una lista de genes de expresión diferencial L, se quiere ver si los miembros de S se distribuyen aleatoriamente en L o se sitúan en los extremos, posible asociación fenotípica.

1. *Cálculo de una puntuación de enriquecimiento (ES):* refleja el grado en que un conjunto S es sobrerrepresentado en los extremos (superior o inferior) de toda la lista L. La puntuación se calcula recorriendo la lista L, cuando se encuentra un gen en S la suma acumulada aumenta, y disminuye cuando el gen no está presente en S. La magnitud del incremento o disminución depende de la correlación del gen con el fenotipo. Por tanto, la puntuación de enriquecimiento es la igual al valor absoluto más grande que se obtiene en la suma acumulada. Este procedimiento utiliza un estadístico de contraste similar a Kolmogorov-Smirnov.

2. *Estimación del nivel de significación de ES:* Se realiza mediante el método de permutaciones, específicamente, se permutan las etiquetas de fenotipo y se vuelve a calcular el ES del conjunto de genes para los datos permutados, generando así una distribución nula para cada ES. El p-valor se obtiene a partir de la comparación entre la distribución nula original y la obtenida con en cada permutación, si la prueba es significativa se podría decir que hay cierta asociación entre el fenotipo y las categorías estudiadas en la lista de genes de expresión diferencial L.

3. *Corrección del p-valor para pruebas de hipótesis múltiples:* Cuando se evalúa una base de datos con numerosos conjuntos de genes, ajustamos el nivel de significancia estimado para tener en cuenta las múltiples pruebas de hipótesis. Se normaliza el ES para cada conjunto de genes para tener en cuenta el tamaño del conjunto; y se calcula la tasa de falsamente rechazados (FDR) correspondiente a cada NES.

5.7.2 FGSEA EN R

Para realizar este enriquecimiento funcional, en primer lugar, se ha creado una base de datos con los uniprot de las proteínas significativas entre las muestras de *LLC* y *centroblastos* con sus respectivos "R.fold" extraídos del algoritmo SAM.

A continuación, se han mapeado los identificadores de las proteínas de formato *uniprot* a formato *entrez id* con la función "mapIds" (Carlson & Id, 2022). Esto es necesario ya que el conjunto de genes que proviene de la base de datos de firmas moleculares "MSigDB", concretamente de la colección *'hallmark'*, vienen definidos por *entrez id*.

Se carga la base de datos "MSigDB" para la categoría "*Hallmark*" y la especie humana (Subramanian et al., 2005) ; con la función "subset" (Huber et al., 2015)se crea un subconjunto de datos a partir de esta base de datos, donde solo aparezcan los entrez id que sean comunes con los *entrez* obtenidos a partir del SAM. Además, con la función "split" (R Core Team, 2022) se dividen los genes del SAM en los diferentes conjuntos de genes definidos en "*Hallmark*".

A partir de aquí, se aplica el análisis de enriquecimiento funcional en R, con la función "fgsea" (Korotkevich et al., n.d.). A continuación, se van a especificar únicamente los argumentos que se han utilizado en esta aplicación práctica:

- pathways: es un objeto de tipo lista, donde las componentes vienen definidas por grupos de genes asociados a las distintas funciones Hallmarks.
- stats: vector de nombres entrez con su respectivo "R.fold", determina el nivel de cada proteína dentro del conjunto. Estos nombres tienen que coincidir con los introducidos en el "pathways"
- nperm: número de permutaciones a realizar
- minSize: tamaño mínimo del conjunto de proteínas a probar. Se excluyen todas las rutas de señalización por debajo del umbral, por defecto este valor es 1.
- maxSize: tamaño máximo del conjunto de proteínas a probar. Se excluyen todas las rutas de señalización por encima de este umbral, por defecto el valor que se pone es infinito "Inf".

Se obtiene una tabla con los resultados de GSEA, cada fila corresponde al análisis de una función "Hallamark" donde en cada columna se tienen las siguientes características del análisis:

- pathway: el nombre la función Hallmark
- pval: el p-valor del análisis de enriquecimiento
- padj: el p-valor ajustado del análisis de enriquecimiento
- ES: valor del enriquecimiento
- NES: valor del enriquecimiento normalizado
- nMoreExtreme: el número de veces que un conjunto de genes aleatorio (no significativo) tiene el valor del enriquecimiento más grande.
- size: número de genes incluidos en la ruta de señalización
- leadingEdge: código entrez de los genes asociados a la ruta de señalización

Por último, se expone una tabla resumen del resultado de GSEA para cada uno de los conjuntos de datos obtenido en el SAM entre las muestras de *LLC* y cada uno de los 5 estadios del Linfocito B (Tabla 10). Se observa que para *LLC-Memoria* el p-valor obtenido no es significativo, se ha expuesto por ser la función con un nivel de significación más próximo a 0,05.

Tabla 10. Tabla resumen enriquecimiento funcional

Hallmarks	p-valor	Gene Ratio	Ranking Hallmarks	Entrez	Gen symbol
LLC-NAÏVE					
HALLMARK_IL2_STAT5_SIGNALING	0.00304	0,02	Down (1º)	79026 5339	AHNAK PLEC
HALLMARK_MYOGENESIS	0.03212	0,03	Down (2º)	1974 6709 4627	EIF4A2 SPTAN1 MYH9
LLC-CENTROBLASTOS					
HALLMARK_IL2_STAT5_SIGNALING	0.00709	0,01	Down (1º)	79026 5339	AHNAK PLEC
HALLMARK_MYC_TARGETS_V1	0.02814	0,08	Up (1º)	9377 6428 7332 57819 3336 4673 5902 5111 6432 6434	COX5A SRSF3 UBE2L3 LSM2 HSPE1 NAP1L1 RANBP1 PCNA SRSF7 TRA2B
HALLMARK_KRAS_SIGNALING_DN	0.04576	0,01	Up (2º)	8115 3848	TCL1A KRT1
LLC-CENTROCITOS					
HALLMARK_MYC_TARGETS_V2	0.00751	0,01	Down (1º)	6652 26354	SORD GNL3
HALLMARK_HEME_METABOLISM	0.01084	0,01	Up (1º)	3043 831	HBB CAST
LLC-MEMORIA					
HALLMARK_PI3K_AKT_MTOR_ SIGNALING	0.09737	0,02	Up (1º)	7334 396 5216	UBE2N ARHGDI A PFN1
LLC-PLAMÁTICAS					
HALLMARK_MITOTIC_SPINDLE	0.00952	0,05	Down (1º)	7430 8243 4926 9126 4627	EZR SMC1A NUMA1 SMC3 MYH9

6 DISCUSIÓN Y CONCLUSIONES

Hoy en día, es necesario continuar con las investigaciones para el desarrollo de tratamientos más eficaces para contrarrestar la Leucemia Linfoide Crónica, ya que, tras los resultados obtenidos en este estudio se observa heterogeneidad entre la enfermedad y su contrapartida normal. Esto dificulta la selección de un tratamiento optimo y la elección de un diagnóstico temprano.

Considerando los objetivos marcados, se ha logrado establecer un flujo de trabajo a nivel computacional en R, que ha permitido extraer un listado de proteínas con expresión diferencial significativa entre las muestras de *LLC* y cada uno de los Estadios del Linfocito B, además de un conjunto más reducido de proteínas que pueden tener cierta relación con distintos fenotipos.

Para ello, se extrajeron por análisis de sangre 6 muestras de *LLC* y se obtuvieron otras 25, repartidas en 5 grupos de 5 para cada estadio del linfocito B, a través de una amigdalectomía. Se parte de este conjunto de datos para realizar el análisis bioestadístico.

En la primera fase del análisis se realizó una estadística descriptiva para revisar la calidad de los datos, conocer las distribuciones de las proteínas, normalizar las muestras para que fueran comparables y comprobar e imputar valores perdidos. Como los datos venían de fuentes distintas se creó una base de datos auxiliar que contenía las proteínas comunes para al menos una muestra de cada uno de los 6 tipos celulares; y se llevó a cabo un análisis cuantitativo con técnicas de reducción de dimensiones para ver posibles patrones de agrupación y un análisis cualitativo con diagramas de Venn y gráficos "Upset" para comprobar las intersecciones entre los distintos tipos celulares.

Con los resultados obtenidos se observa que las células de *LLC* son muy diferentes de las 5 poblaciones celulares de linfocitos B. Dentro de estas poblaciones, existe mucha homogeneidad entre *naïve, centroblastos, centrocitos y célula de memoria,* así como una diferencia de estas con las *células plasmáticas.*

En la segunda etapa, se aplicó la herramienta bioestadística SAM para extraer las proteínas con expresión diferencial significativa para cada comparación a estudiar (*LLC vs näive, LLC vs centroblastos, LLC vs centrocitos, LLC vs células de memoria* y *LLC vs células plasmáticas*).

Se consiguieron entre 90 y 160 proteínas significativas en cada comparación, con un FDR menor a 0,01, a excepción de las *células plasmáticas*, donde el FDR es de 0,019 debido a la poca cantidad de proteínas con las que se trabaja.

En la última etapa, con los conjuntos obtenidos a partir del algoritmo SAM, se realizó un análisis de enriquecimiento funcional (GSEA). Atendiendo al número de proteínas y las funciones "*hallmarks*" obtenidas, se puede observar una tendencia en el proceso de diferenciación celular y la relación con las células de *LLC* (Figura 27).

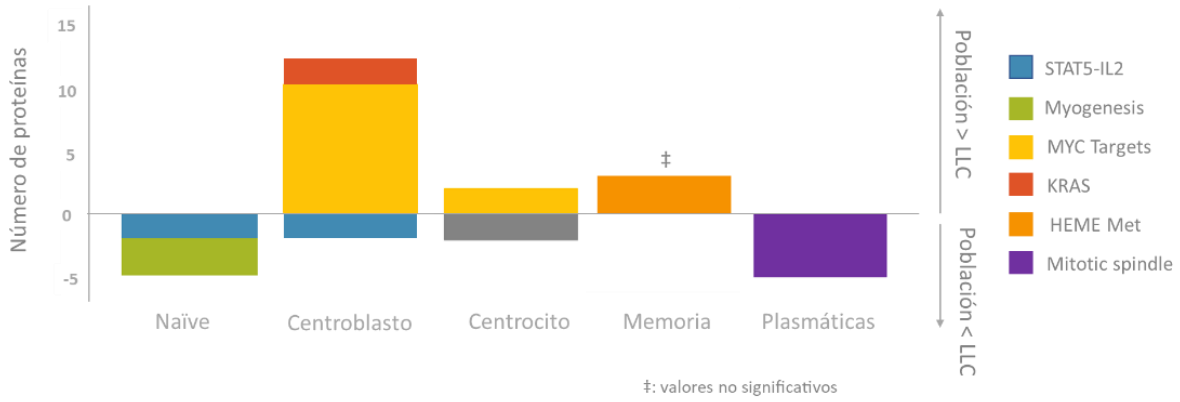


Figura 27. Representación gráfica de los resultados de enriquecimiento funcional

En los primeros estadios de diferenciación, se encuentran resultados muy llamativos. Si bien cierto que es esperable un cambio en la expresión de las proteínas entre los diferentes estadios, parecería más razonable encontrar los mayores cambios en los estadios finales cuando las células están más especializadas. En este caso, con respecto a la *LLC* los mayores cambios se reflejan en la población de *centroblastos*. Esta población, además, presenta unas proteínas relacionadas con la función "*IL2-STAT5 siganling"* que también coinciden con las que aparecen en esa función en las células *naïve*.

Así mismo, la función más destacable "*MYC targets*" también aparece en la población *centrocitos*. En ambos tipos celulares, se observa desregulación en ambos sentidos ("*upregulation*" y "*downregulation*"). Por lo general, las proteínas que varían significativamente en estos tres grupos son proteínas relacionadas con el citoesqueleto, la migración celular y el metabolismo del ARN. Estas funciones son generalmente celulares y ponen de manifiesto los cambios heterogéneos y poco específicos que se pueden observar entre estas poblaciones y las células de *LLC*.

Siguiendo con el análisis de los dos últimos estudios de diferenciación, en las *células de memoria*, no se han encontrado diferencias significativas con respecto a las células de *LLC*. Este resultado poco interesante desde el punto de vista estadístico resulta de gran importancia a nivel biológico, ya que, pone de manifiesto una cierta similitud funcional de las células de *LLC* con *las células de memoria*.

Por último, entre las *células plasmáticas* y las células de *LLC* se ha obtenido diferencia en proteínas relacionadas con el ciclo celular ("*mitotic spindle*"), una desregulación de esta función es clave en el crecimiento de las células tumorales.

Este listado de proteínas asociadas a funciones fenotípicas específicas puede abrir una nueva visión para estudios posteriores, si se analiza más exhaustivamente el comportamiento, función y regulación de las proteínas obtenidas.

7 BIBLIOGRAFÍA

Amat, R. J. (2017). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. *RStudio Pubs*, 1–38. https://www.cienciadedatos.net/documentos/35_principal_component_analysis%0Ahttps://www.cienciadedatos.net/documentos/35_principal_component_analysis%0Ahttps://www.cienciadedatos.net/documentos/35_principal_component_analysis#ejemplo_pca_aplicado_a_genómi

Atkinson, A. C., & Hawkins, D. M. (1981). Identification of Outliers. *Biometrics*, *37*(4), 860. https://doi.org/10.2307/2530182

Ayala, G. (2020). *Bioinformática Estadística Análisis estadístico de datos ómicos*. 369.

Benito Jiménez, C. (n.d.). *Proyecto Editorial MANUALES DE GENÉTICA Coordinador*. Retrieved July 4, 2022, from www.sintesis.com

Berry, K. J., Johnston, J. E., & Mielke, P. W. (2011). Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(6), 527–542. https://doi.org/10.1002/WICS.177

Carlson, M., & Id, G. (2022). *AnnotationDbi: Introduction To Bioconduc-tor Annotation Packages PLATFORM PKGS*.

Castellanos-Bueno, R. (2020). La respuesta inmunitaria. *Revista Colombiana de Endocrinología, Diabetes & Metabolismo*, *7*(2S), 55–61. https://doi.org/10.53853/encr.7.2s.584

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*, *33*(18), 2938–2940. https://doi.org/10.1093/bioinformatics/btx364

Díez, P., Góngora, R., Orfao, A., & Fuentes, M. (2017). Functional proteomic insights in B-cell chronic lymphocytic leukemia. *Expert Review of Proteomics*, *14*(2), 137–146. https://doi.org/10.1080/14789450.2017.1275967

Díez, P., Pérez-Andrés, M., Bøgsted, M., Azkargorta, M., García-Valiente, R., Dégano, R. M., Blanco, E., Mateos-Gomez, S., Bárcena, P., Santa Cruz, S., Góngora, R., Elortza, F., Landeira-Viñuela, A., Juanes-Velasco, P., Segura, V., Manzano-Román, R., Almeida, J., Dybkaer, K., Orfao, A., & Fuentes, M. (2021). Dynamic Intracellular Metabolic Cell Signaling Profiles During Ag-Dependent B-Cell Differentiation. *Frontiers in Immunology*, *12*, 637832. https://doi.org/10.3389/FIMMU.2021.637832

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 1–17. https://doi.org/10.1186/2193-1801-2-222

Fainboim, L., & Geffner, J. (2011). *Introducción a la Inmunología Humana*. Editorial Médica Panamericana.

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/BIOINFORMATICS/BTW313

Hallek, M., & Al-Sawaf, O. (2021). Chronic lymphocytic leukemia: 2022 update on diagnostic and therapeutic procedures. *American Journal of Hematology*, *96*(12), 1679–1705. https://doi.org/10.1002/ajh.26367

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discovery*, *12*(1), 31–46. https://doi.org/10.1158/2159-8290.CD-21-1059

Hossein Arsham, Lovric, M. (2011). Bartlett ' S Test Bartlett ' S Test. *International Journal of Ecological Economics & Statistics*, *10*.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M.,

Love, M. I., MaCdonald, J., Obenchain, V., Oleš, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods 2015 12:2*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252

Kassambara, A., & Mundt, F. (2020). Package 'factoextra.' *CRAN- R Package*, 84. https://cran.r-project.org/package=factoextra

Korotkevich, G., Sukhov, V., & Sergushichev, A. (n.d.). *Fast gene set enrichment analysis*. https://doi.org/10.1101/060012

Little, R. J. A., & Rubin, D. B. (2014). Statistical analysis with missing data. *Statistical Analysis with Missing Data*, 1–381. https://doi.org/10.1002/9781119013563

Maintainer, S., & Schwender, H. (2022). *Package "siggenes" Title Multiple Testing using SAM and Efron's Empirical Bayes Approaches*. https://git.bioconductor.org/packages/siggenes

Maxwell, O., Onyedikachi, I. P., Aidi, K., Akpa, C. I., Seddik-Ameur, N., Maxwell, O., Onyedikachi, I. P., Aidi, K., Akpa, C. I., & Seddik-Ameur, N. (2022). Generalized Kumaraswamy Generalized Power Gompertz Distribution: Statistical Properties, Application, and Validation Using a Modified Chi-Squared Goodness of Fit Test. *Applied Mathematics*, *13*(3), 243–262. https://doi.org/10.4236/AM.2022.133019

Mojica Ph.D, T., Sánchez, O., & Bobadilla, L. (2003). La Proteómica, otra cara de la genómica. *Nova*, *1*(1), 13. https://doi.org/10.22490/24629448.1060

Parham, P., Rojas-espinosa, Ó., & Lozano, F. (2007). *Inmunología (2ª ed) Inmunología (de memoria) (3ª ed)*. *26*, 62–63. https://books.google.com/books/about/Inmunología_de_memoria.html?hl=es&id=CtWACreo-BkC

Regueiro, J. R. (n.d.). *Inmunología : biología y patología del sistema inmunitario*.

Roitt inmunología [Recurs electrònic] : fundamentos / Peter J. Delves ... [et al.]. (2014). Médica Panamericana,. https://discovery.udl.cat/iii/encore/record/C__Rb1329191?lang=cat

Salinas Carmona, M. C. (n.d.). *La Inmunología en la Salud y la Enfermedad*. Editorial Médica Panamericana.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*. www.pnas.orgcgidoi10.1073pnas.0506580102

Taiyun Wei, M., Taiyun Wei cre, A., Simko aut, V., Levy ctb, M., Xie ctb, Y., Jin ctb, Y., Zemla ctb, J., Freidank ctb, M., Cai ctb, J., & Protivinsky ctb, T. (2021). *Title Visualization of a Correlation Matrix NeedsCompilation no*.

Vega Robledo, G. B. (n.d.). *Inmunología Básica y su Correlación Clínica*. Editorial Médica Panamericana.

Vinuesa, P. (2016). *Tema 8-Correlación: teoría y práctica*. http://www.ccg.unam.mx/~vinuesa/

Wilke, C. O. (2021). *Package ' ggridges . '* 1–8.

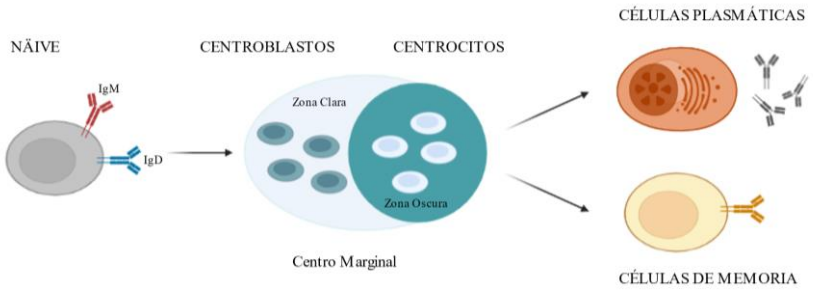
8 ABSTRACT

The immune system is a complex network of organs and cells perfectly integrated and coordinated at t he body level. In this case, the prevalence of chronic lymphocytic leukemia (CLL) and the heterogeneity of the disease search for new perspectives of analysis, such as that proposed by proteomics. This study enables to obtain a functional characterization of the proteins with their structural relationships, an accurate analysis of a specific clinical pathology.

Two of the most common techniques used in proteomics, and also in this study, are electrophoresis and mass spectrometry. These two techniques allow for the extraction and quantification of the proteins defined in CLL cells and in those of the 5 stages of the B-cells.

B lymphocytes are cells that are primarily involved in humoral immunity due to their role as antibody-producing cells. These cells originate from the pluripotent hematopoietic stem cell, from which all blood cells derive.

After different stages of maturation in the bone marrow, the naïve cell migrates to the germinal centers where the cells divide at the centroblast stage and, after the proliferation period, as centrocytes. The final stages of human B cell differentiation are the "antigen-dependent" stages leading to the expansion of B cells to their terminal differentiation into antibody-secreting plasma cells and memory B cells, which occurs in secondary lymphoid tissues.

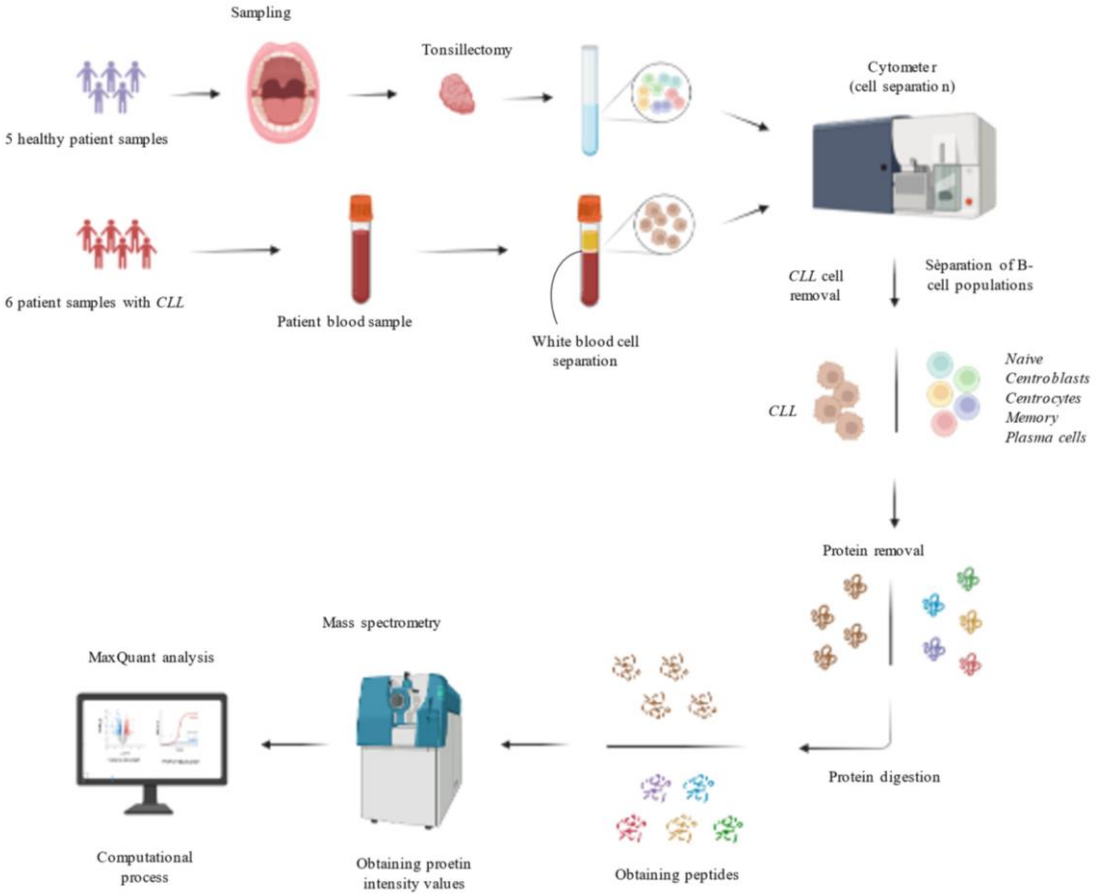


The B-cell dependent differentiation can be performed by the presence of surface antigens or cluster of differentiation (CD) present in the cell membrane and allows for the identification of the different populations, as well as their phenotypic separation by flow cytometry techniques, despite the great similarities at the genomic level that these cells present.

Chronic lymphocytic leukemia (CLL) is one of the most common types of leukemia. It usually occurs in elderly patients and has a highly variable diagnosis and clinical course. This disease is initiated by specific genetic alterations that interfere with the appearance of clonal B cells and their apoptosis process. These tumor B cells accumulate in the blood, bone marrow, lymph nodes or the spleen, causing immune system dysfunction and homeostasis problems in patients.

The large amount of data generated by advances in these technologies such as proteomics and the complexity of CLL at the phenotypic level has raised the overall objective of this work based on the design and development of systematic workflows for the proteomic characterization of chronic lymphoid leukemia in comparison with its normal B-cell counterpart.

The two datasets were obtained by the procedure below:



The following strategy is established in the R statistical language, in order to achieve a good biostatistical analysis of the proteins in the study:

1. Descriptive analysis

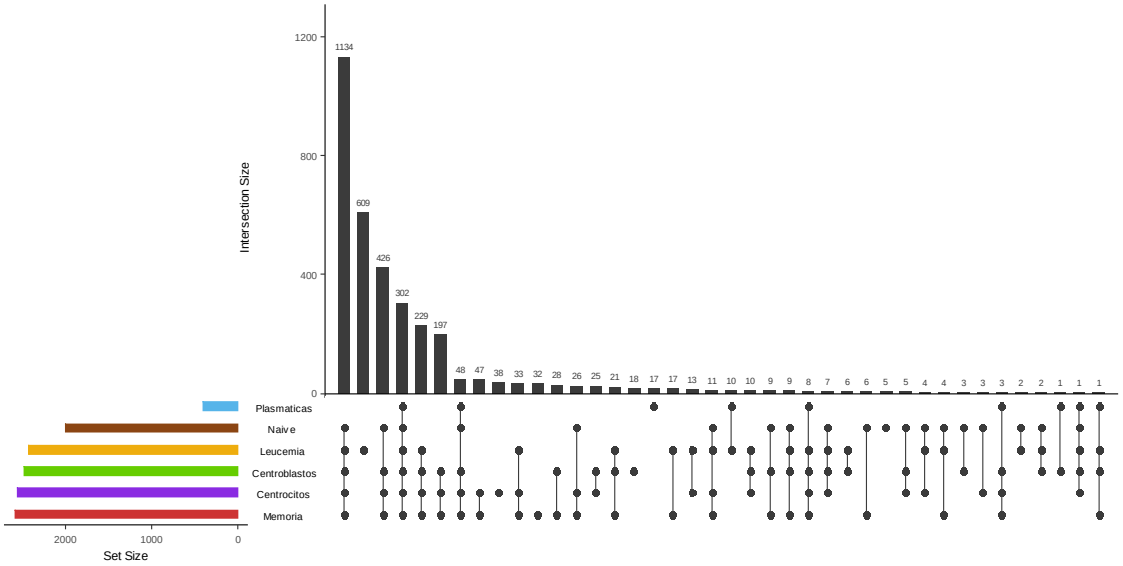
A descriptive analysis is performed to check the quality of the data, to know the distribution of the proteins, to normalize the sample and to impute missing values.

Both datasets were normalized using the typing/standardization method. The 5 B-cell populations did not present any missing values, whereas in the CLL samples there were 47% of proteins with missing values in more than 4 samples, therefore, they were eliminated from the study with a total of 2434 proteins. Of this remaining set, there were 52% of proteins with at least one missing value, these were imputed by the multiple imputation method.

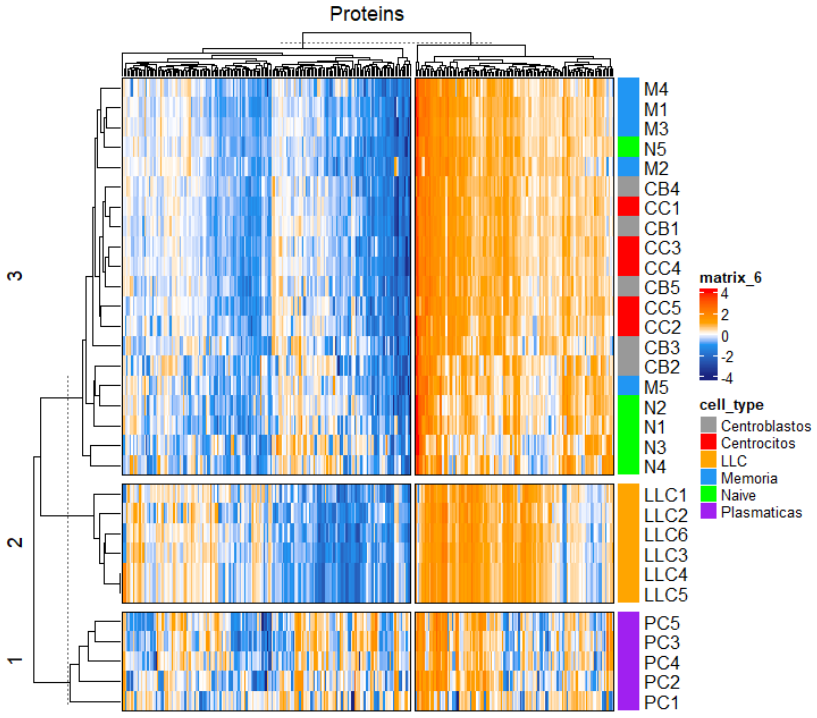
2. Union of the dataset

These two are matched with common proteins for at least one sample from each of the 6 cell types; and a quantitative analysis was carried out with dimension reduction techniques to see possible clustering patterns and a qualitative analysis with Venn diagrams and 'Upset' graphs to verify the intersections between the different cell types.

This analysis shows that CLL cells are very different from the 5 B lymphocyte cell populations. There are 609 proteins uniquely identified in CLL samples.



Furthermore, it can be observed that within the 5 populations of B cells and LLC cells, there homogeneity between *naïve, centroblasts, centrocytes and memory cells*, as well as a difference between these and plasma cells and LLC, both classified individually.

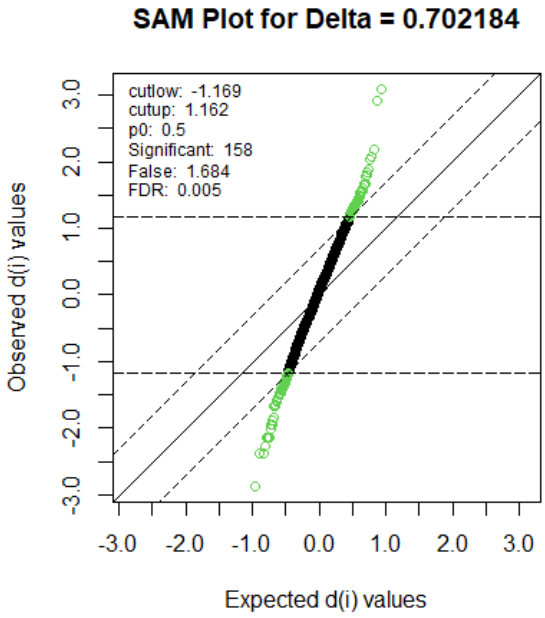


3. Significance analysis of microarrays (SAM)

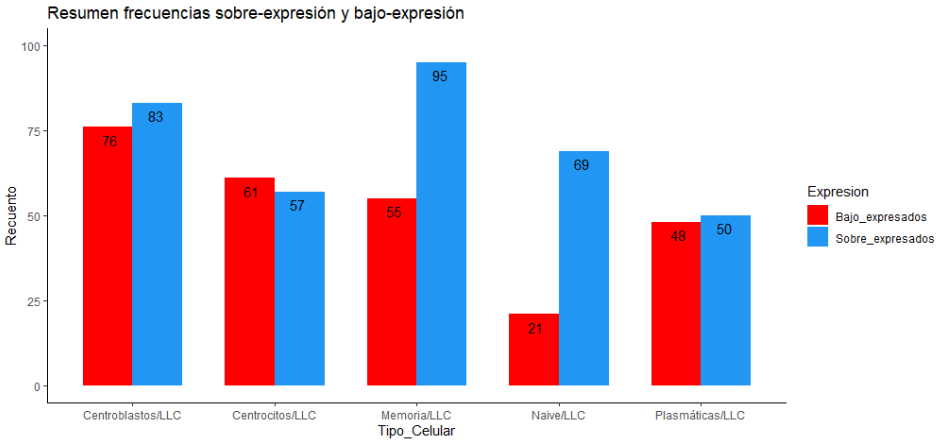
In this third stage, we want to extract the proteins with significant differential expression for each comparison to be studied (LLC vs näive, LLC vs centroblasts, LLC vs centrocytes, LLC vs memory cells and LLC vs plasma cells).

Between 90 and 160 significant proteins were obtained in each comparison, with an FDR of less than 0.01, except in the case of plasma cells, where the FDR is 0.019 due to the small number of proteins used.

With an FDR of 0.00539 associated with a delta of 0.702184, 158 significant proteins were obtained with an average of 1.684 false positives in CLL vs centroblast comparison.



The significant over-expressed and under-expressed proteins for each comparison are shown in the following graph



1. GSEA functional enrichment

   In the last stage, with the sets obtained from the SAM algorithm, an enrichment analysis (GSEA) was performed. Based on the number of proteins and the hallmark functions obtained, we can observe a relationship between the process of cell differentiation and CLL cells.

Hallmarks	p-valor	Gene Ratio	Ranking Hallmarks	Entrez	Gen symbol
CLL-NAÏVE					
HALLMARK_IL2_STAT5_SIGNALING	0.00304	0,02	Down (1º)	79026 5339	AHNAK PLEC
HALLMARK_MYOGENESIS	0.03212	0,03	Down (2º)	1974 6709 4627	EIF4A2 SPTAN1 MYH9
CLL-CENTROBLAST					
HALLMARK_IL2_STAT5_SIGNALING	0.00709	0,01	Down (1º)	79026 5339	AHNAK PLEC
HALLMARK_MYC_TARGETS_V1	0.02814	0,08	Up (1º)	9377 6428 7332 57819 3336 4673 5902 5111 6432 6434	COX5A SRSF3 UBE2L3 LSM2 HSPE1 NAP1L1 RANBP1 PCNA SRSF7 TRA2B
HALLMARK_KRAS_SIGNALING_DN	0.04576	0,01	Up (2º)	8115 3848	TCL1A KRT1
CLL-CENTROCYTES					
HALLMARK_MYC_TARGETS_V2	0.00751	0,01	Down (1º)	6652 26354	SORD GNL3
HALLMARK_HEME_METABOLISM	0.01084	0,01	Up (1º)	3043 831	HBB CAST
CLL-MEMORY					
HALLMARK_PI3K_AKT_MTOR_ SIGNALING	0.09737	0,02	Up (1º)	7334 396 5216	UBE2N ARHGDIA PFN1
CLL-PLASMA CELLS					
HALLMARK_MITOTIC_SPINDLE	0.00952	0,05	Down (1º)	7430 8243 4926 9126 4627	EZR SMC1A NUMA1 SMC3 MYH9

Very striking results are found in the early stages of differentiation. While it is true that change in protein expression between the different stages is expected, it seems more reasonable to find the greatest changes in the later stages, when the cells are more specialized. In this case, with respect to CLL, the greatest changes are reflected in the population of centroblasts. This population has proteins related to the "IL2-STAT5-follow-on" function that also coincide with those appearing in naïve cells.

Likewise, the most remarkable function “MYC targets” appears not only in centroblast but also in centrocytes with dysregulation, in both directions (“upregulation” for and “downregulation”). Typically, the proteins that vary significantly in these three groups are proteins related to the cytoskeleton, cell migration, and RNA metabolism. These functions are generally cellular and lay bare the heterogeneous and non-specific changes that can be observed between these populations and CLL cells.

In memory cells, no significant differences were found with respect to CLL cells. This result, which is not very interesting from a statistical point of view, is of great importance at the biological level since it reveals a certain functional similarity of CLL cells with memory cells.

Finally, between plasma cells and CLL cells, a difference in proteins related to the cell cycle ("mitotic spindle") has been obtained, a dysregulation of this function is key in the growth of tumor cells.

This list of proteins associated with specific phenotypic functions may open a new vision for further studies if the behavior, function, and regulation of proteins obtained are more thoroughly analyzed.