VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

GRADO DE ESTADÍSTICA

Trabajo de fin de grado

FLUJOS DE TRABAJO SISTEMÁTICOS PARA EL ANÁLISIS COMPUTACIONAL DE DATOS PROTEÓMICOS

Autor: Héctor Lorenzo Gil

Tutores: Dr. José Manuel Sánchez Santos

Dr. Manuel Fuentes García

Marina Luque García-Vaquero

Salamanca, julio de 2021

GRADO DE ESTADÍSTICA

Trabajo de fin de grado

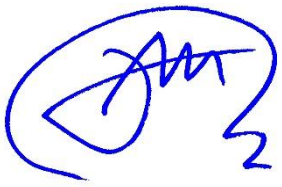
FLUJOS DE TRABAJO SISTEMÁTICOS PARA EL ANÁLISIS COMPUTACIONAL DE DATOS PROTEÓMICOS

Autor:

Héctor Lorenzo Gil

Tutores:

Dr. José Manuel Sánchez Santos



Dr. Manuel Fuentes García

Marina Luque García-Vaquero

Salamanca, julio de 2021

Este trabajo de fin de grado se ha desarrollado en colaboración con ProteoRed (Red Española de laboratorios de investigación en proteómica) perteneciente al ISCIII (Instituto de Salud Carlos III) y coordinado por Manuel Fuentes.

A su vez, parte del contenido de este trabajo está relacionado con el Proyecto de Investigación: FIS17/01930; siendo Investigador Principal Manuel Fuentes.

Índice

CAPÍTULO 1: Introducción

La proteómica es el estudio de los proteomas, donde se separan, identifican y caracterizan proteínas a gran escala. También define niveles de proteínas celularmente, explica sus funciones metabólicas e interrelaciones. En definitiva, la proteómica realiza una caracterización funcional de las proteínas y sus relaciones estructurales, además de su previo análisis (Mojica et al., 2003).

En el presente trabajo se han obtenido muestras de amígdalas los estadios del linfocitos B y muestras de Leucemia Linfocítica Crónica (*LLC*) y de Linfocitosis Monoclonal de las células B (*LBM*) con el objetivo de comparar los datos obtenidos a través de técnicas estadísticas y bioinformáticas. Se realizan dos estudios paralelos utilizando las mismas técnicas para ambos. El primero estudio incluye las muestras de los diferentes estadios del linfocito B y, el segundo incluye las muestras de todos los tipos celulares (las de los estadios del linfocito B y las de *LLC* y *LBM*).

Con el software estadístico *R* se aplican las siguientes técnicas:

- → Descripción de la base de datos con técnicas estadísticas descriptivas.

- → Comparaciones cuantitativas a través de diagramas de correlación y dispersión entre los estadios del linfocito B, *LLC* y *LBM.*

- → Comparaciones cualitativas a través de diagramas de Venn para observar cuántas proteínas coinciden entre las muestras del linfocito B y las de *LLC* y *LBM*.

- → Reducción de dimensiones a través de Análisis de Componentes Principales para una explicación de la variabilidad de los datos.

- → Creación de Self Organizing Maps para la clasificación de proteínas en función de sus características, con sus correspondientes gráficos para explicar la calidad de la clasificación, la distribución de las proteínas y su interpretación biológica.

- → Técnicas de enriquecimiento funcional con paquetes como GO para explicar posibles funciones de las proteínas más significativas.

Además, se utilizan STRING y *Reactome* para conocer las interacciones entre proteínas significativas que se han obtenido del enriquecimiento funcional en *R*, sus rutas de señalización y su posición dentro de la célula.

Los pasos que se siguen en el presente trabajo permitirán una buena clasificación de proteínas en función de sus características, buscar sus funciones dentro del organismo y sus posiciones dentro de la célula.

CAPÍTULO 2: Introducción biológica

2.1. Concepto de genoma y proteoma:

Se define como *genoma* , según Hans Winkler, botánico alemán, en 1920 "el juego completo de cromosomas y sus genes en una especie biológica determinada"(Cazzulo, 2014). En 1911 se introdujeron por parte de Wilhelm Johanssen los conceptos de "genotipo" y de "fenotipo" (Cazzulo, 2014).

El genotipo es la información hereditaria en un individuo de forma completa, se haya expresado o no; sin embargo, el fenotipo es la manifestación del genotipo, por ejemplo, la morfología o el desarrollo de una célula. La genómica se dedica al estudio de los genomas de los organismos mediante la secuenciación completa del ADN y la ubicación de los genes en los cromosomas (Cazzulo, 2014).

El genoma se constituye principalmente por ADN (ácido desoxirribonucleico), que son moléculas poliméricas formadas por subunidades monoméricas denominadas nucleótidos. Los nucleótidos son de cuatro tipos llamados: Timina, Guanina, Citosina y Adenina que, enlazados unos con otros, completan una cadena. Uniéndose la Adenina con la Timina y la Guanina con la Citosina se forman los pares de bases nitrogenadas que confeccionan todo el genoma humano, el cual está dividido en 23 fragmentos llamados cromosomas. La denominada "expresión del genoma" consiste en una serie de reacciones bioquímicas realizadas por enzimas y proteínas para utilizar la información del genoma (Brown, 2008).

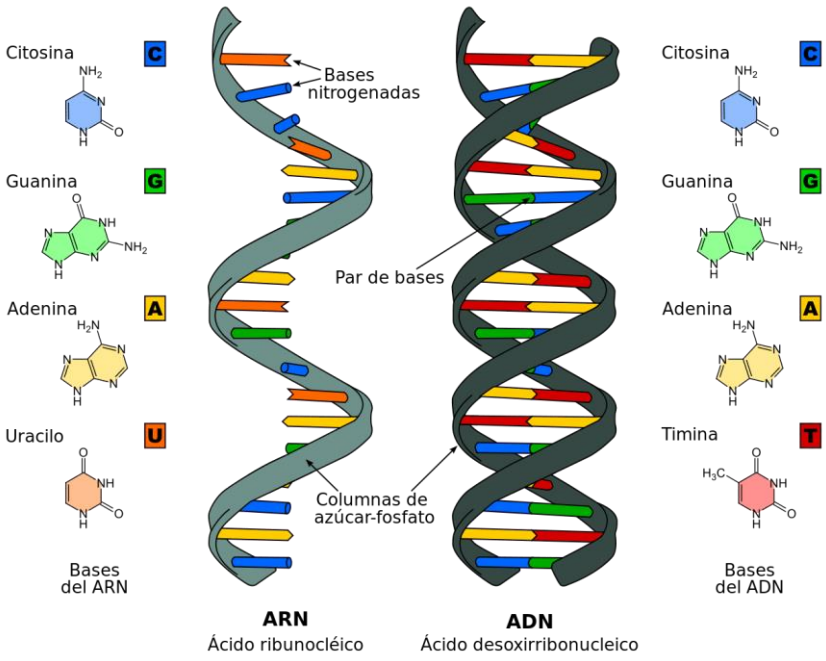


Figura A. Bases de ARN, ARN, ADN y Bases de ADN. Figura adaptada de («Ácidos Nucleicos Qué son, Funciones y Estructura - ADN y ARN», 2017)

Para comprender la función del genoma, es importante adentrarnos en el Dogma central de la biología molecular. La información genética que se conserva en el ADN se transmite a las posteriores generaciones por medio de la replicación, llevada a cabo en dos etapas: transcripción y traducción. En la transcripción, la expresión del genoma origina el transcriptoma, que son las moléculas de ARN (ácido ribonucleico) que se encuentran en una

célula en un instante determinado. Hay dos tipos de ARN: codificante y no codificante (Brown, 2008).

→ ARN codificante: está compuesto por el ARNm, es decir, el ARN mensajero. Este tipo de ARN sintetiza las proteínas y forma parte del transcriptoma.

→ ARN no codificante: no está presente en el transcriptoma y dentro de este tipo se engloban otros subtipos: el ARN ribosómico (ARNr), donde se realiza la síntesis proteica; el ARN de transferencia (ARNt), que se encarga de enviar aminoácidos al ribosoma y el ARN nuclear pequeño.

Posteriormente, en la etapa de traducción, se confeccionan las proteínas a través del uso de la información almacenada en el ARNm. El ARNm se traslada al citoplasma y se asocia al ribosoma donde se lee la información genética. Mientras tanto, los ARNt llevan los aminoácidos al ribosoma para formar las proteínas al juntarse con el ARNm.

Los ARNm se traducirán a proteínas y realizarán entonces todas las funciones de la célula. La información no está contenida directamente en el genoma, los sistemas de enzimas se encuentran codificados allí y son los encargados de ella. Por lo tanto, la expresión de los genes llevará al fenotipo a través de la acción de las proteínas. Es por esto que, en 1994, Marc Wilkins introdujo el concepto proteoma que se define como "el conjunto completo de las proteínas que una célula u organismo puede expresar"(Cazzulo, 2014).

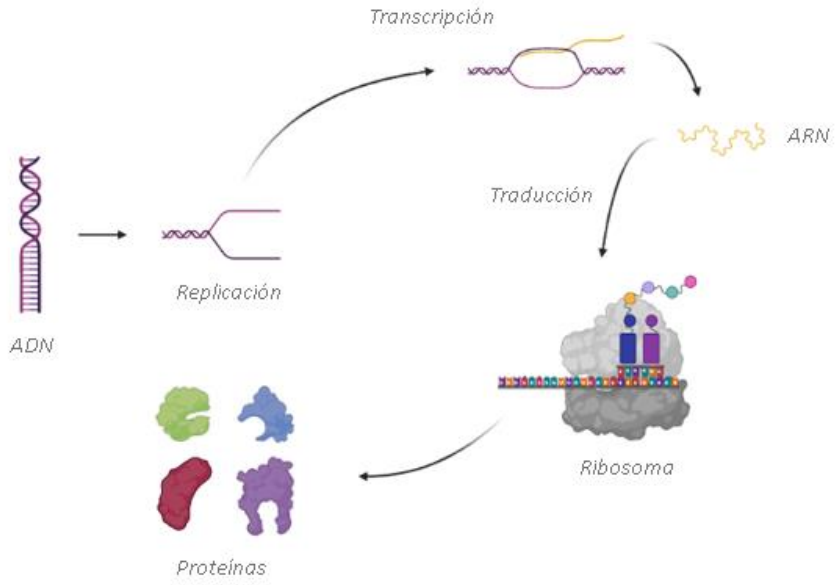


Figura B. Dogma central de la biología molecular. Figura creada con BioRender.

Una propiedad muy importante del proteoma es que es dinámico, es decir, cambia de manera constante para adaptarse a las necesidades que requiere la célula en cada momento fisiológico. En los organismos pluricelulares el genoma es igual para todas las células, pero el proteoma es distinto en cada tipo celular, hay proteínas que se expresan en casi todos los tipos celulares, pero hay otras proteínas que solo se expresan en algún tipo celular. El organismo humano tiene más de 200 tipos celulares y la diferencia es originada por la variación en la expresión de sus proteínas. Además, es necesario destacar el alto dinamismo

que presenta el proteoma porque cambia constantemente según las condiciones en las que se encuentre la célula: su entorno, tiempo de vida, etc.

2.2. Concepto de proteómica

La proteómica es el estudio de los proteomas, donde se separan, identifican y caracterizan proteínas a gran escala. También define niveles de proteínas celularmente, explica sus funciones metabólicas e interrelaciones. En definitiva, la proteómica realiza una caracterización funcional de las proteínas y sus relaciones estructurales, además de su previo análisis (Mojica et al., 2003).

El proceso de replicación hasta la obtención de las proteínas de la *Figura B*, nos ofrece como resultado la estructura primaria de la proteína y las proteínas codificadas en un genoma. A partir de la estructura primaria surgen estructuras secundarias y terciarias en las proteínas, de las cuales algunas tienen modificaciones posteriores a la etapa de traducción e incluso puede adquirir estructura cuaternaria. Ni la estructura primaria del gen ni la del ARNm son indicativos del producto proteico ni de las funciones que cumple, es por eso por lo que la proteómica comienza por identificar las proteínas creadas en un genoma (Mojica et al., 2003).

La proteómica se divide en 3 tipos de acuerdo con su campo de acción (Castellanos et al., 2004):

- → Proteómica estructural: se encarga de determinar las proteínas expresadas en un momento dado y sus modificaciones tras la etapa de traducción.

- → Proteómica comparativa: identifica cambios en el nivel de expresión de las proteínas.

- → Proteómica funcional: identifica grupos funcionales de proteínas.

Existen múltiples técnicas en proteómica, siendo las más conocidas: electroforesis en geles en 2D y tecnología para la identificación de proteínas, como la espectrometría de masas, y microarrays de proteínas para ensayos masivos. A continuación, se describen brevemente las técnicas anteriormente mencionadas:

1. La electroforesis, técnica para la separación de moléculas según la movilidad de estas en un campo eléctrico, fue inventada hace más de 50 años y en 1975 Patrick O'Farell consiguió realizar el proceso de separación en 2D, que es de lo que trata la técnica que se usa hoy en día. En la primera dimensión las proteínas se separan por su carga y en la segunda dimensión por su masa. Posteriormente, las manchas de gel se extraen y se trata con tripsina para obtener patrones de péptidos, los cuales se identifican a través de espectrometría de masas, técnica que mide la masa de las moléculas .

2. La espectrometría de masas es una técnica de identificación de péptidos y proteínas. Los espectrómetros de masas, dispositivo que permite analizar con gran precisión la composición de diferentes elementos químicos, están formados por: una fuente de iones y un aparato medidor (cuadrupolos, confinamiento de iones y TOF). Las composiciones de esta técnica que más se utilizan son (Mojica et al., 2003):

→ MALDI-TOF-MS: se trata de espectrometría de masas en la que se realiza un proceso de desorción ionización asistida por una matriz sólida. Este tipo de identificación tiene lugar cuando están inmovilizadas en una membrana, pero es de baja eficiencia.

→ MALDI o ESI-MS: es un método de espectrometría de masas por ionización en electroatomizado. Analiza proteínas inmovilizadas en membranas y también digeridas en geles.

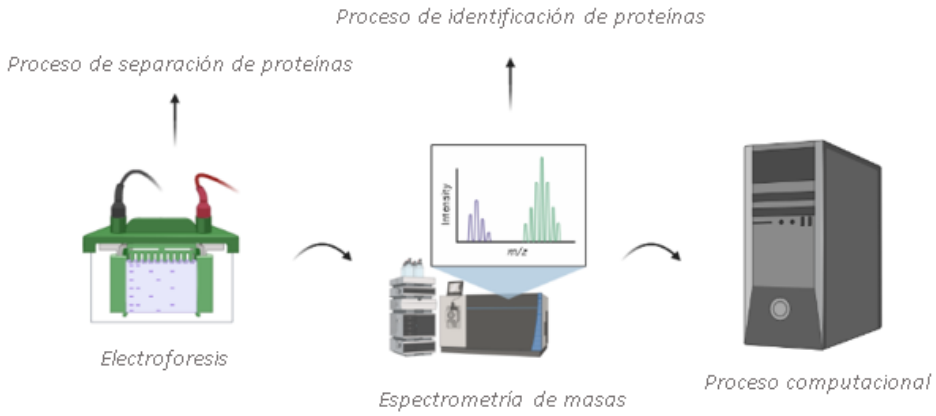


Figura C. Proceso de la proteómica para separación, identificación y análisis de proteínas. Figura creada con BioRender.

3. Un microarray es un soporte sólido con forma de matriz en el cual se imprime en un orden determinado una colección de moléculas *(Figura D)* (Schrenzel et al., 2009).
   Este tipo de microarrays permiten especificar la información necesaria de las proteínas que son traducidas del ARNm, ya que después de la etapa de traducción sufren cambios o alteraciones. Suponen un gran avance en la proteómica ya que muestran aplicaciones para enzima-sustrato, ADN-proteína y otras interacciones entre proteína-proteína. (Templin et al., 2002). A parte, algunas otras ventajas que tienen los microarrays de proteínas son: las matrices permiten monitorear varias proteínas en un único ensayo, posibilitan el cribado de suero, descubrimiento de biomarcadores y realizar análisis proteómicos funcionales, se posee un control sencillo de las condiciones experimentales, tiene un bajo consumo de muestra y es muy sensible en comparación con otros tipos de microarray (microarray de ADN).

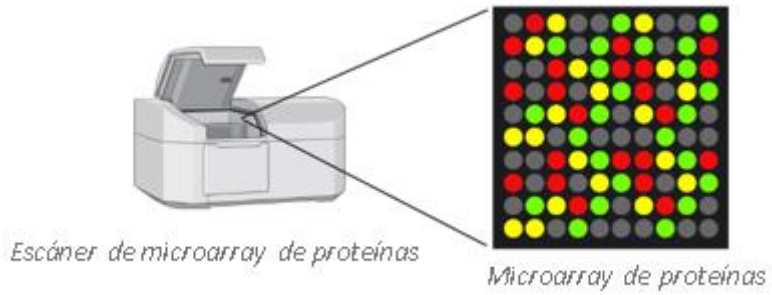


Figura D. Escáner y microarray de proteínas. Figura creada con BioRender.

Finalmente, tras haber obtenido la información de la masa y la carga de cada péptido; se lleva a cabo un proceso computacional en el que se compara con la base de datos y se determina la proteína más probable (Mojica et al., 2003).

Para llevar a cabo el proceso o análisis computacional hay varios programas informáticos en los que se apoya la proteómica, pero una de las plataformas más usadas es el MaxQuant. Dicho programa informático permite visualizar datos sin procesar de espectrometría de masas a la vez que se puede observar los resultados de la tubería de identificación y cuantificación. Además, MaxQuant permite la navegación entre conjuntos de datos masivos a través de la indexación de estructuras de datos subyacentes. Este programa nos ofrece como resultado final valores de intensidad LFQ (*label-free quantitation*) para las proteínas que se encuentren presentes en la previa identificación de las mismas a partir de la espectrometría de masas (Tyanova et al., 2015).

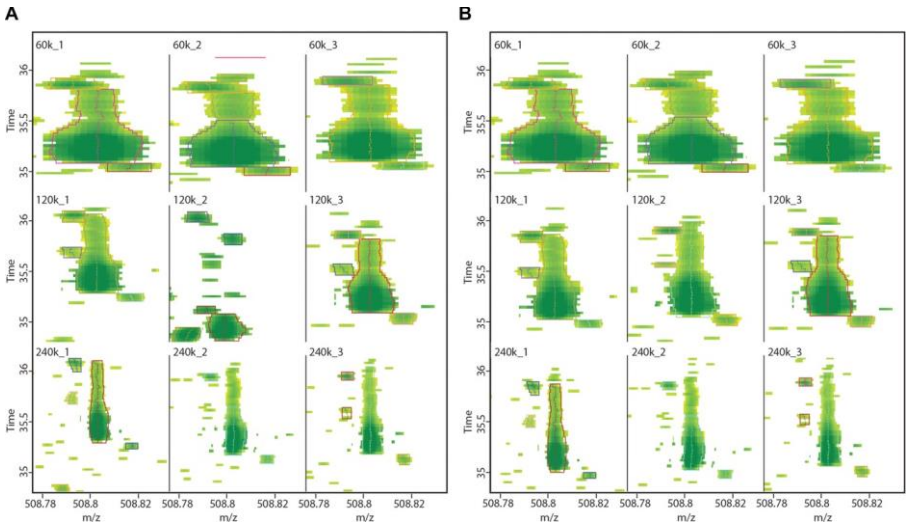


Figura E. Visualización de múltiples mapas con m/z en MaxQuant. Figura adaptada de (Tyanova et al., 2015).

Tras el uso de programas como MaxQuant, se lleva a cabo el análisis de los resultados obtenidos (los valores de intensidades LFQ). Para ello se utilizan softwares estadísticos, que pueden estar realizados con MaxQuant (Perseus o Skyline) u otros softwares como R con su interfaz RStudio. Todo ello permite crear mapas de interacción entre proteínas, es decir, las interacciones que se producen entre los elementos de un proteoma. Cada red se constituye alrededor de un grupo de proteínas que evidencian interacciones entre ellas. Todo esto permite la búsqueda de *biomarcadores* que se definen como las proteínas que podrían ayudar a diagnosticar una enfermedad o estudiar la evolución de la misma según (Cazzulo, 2014).

2.2. 1. Técnicas de espectrometría de masas en proteómica

La espectrometría de masas se introdujo a finales de los años 70 con técnicas de ionización suave, por ejemplo, la desorción por campo eléctrico, la desorción por plasma y la ionización por bombardeo de átomos rápidos. Estas técnicas permitían la ionización de moléculas termolábiles de gran tamaño sin producir demasiada degradación. En los 90, aparecen métodos como el electrospray (ESI) o la desorción por láser asistida por matriz (MALDI), que se establecieron como la base fundamental de la espectrometría de masas en la proteómica contemporánea (Abián et al., 2008).

Posteriormente, comenzó el desarrollo de analizadores de masa de creciente velocidad de análisis y resolución que, con su incremento de las prestaciones el estudio del proteoma por espectrometría de masas ha conseguido avanzar grandes pasos en los últimos años.

Los analizadores de masas pueden dividirse en 4 grupos: analizadores de sectores (eléctricos o magnéticos), de cuadrupolo (Q), de tiempo de vuelo (TOF) y de confinamiento de iones (Ion Traps, Orbitraps, FT-ICR). Hoy en día, todos estos tipos de analizadores de masas son utilizados ampliamente en proteómica excepto los analizadores de sectores. La ventaja que ofrecían los analizadores de sectores en referencia a la resolución actualmente es proporcionada por el Orbitrap o los reTOF (analizadores de tiempo de vuelos provistos de reflectrón) (Abián et al., 2008).

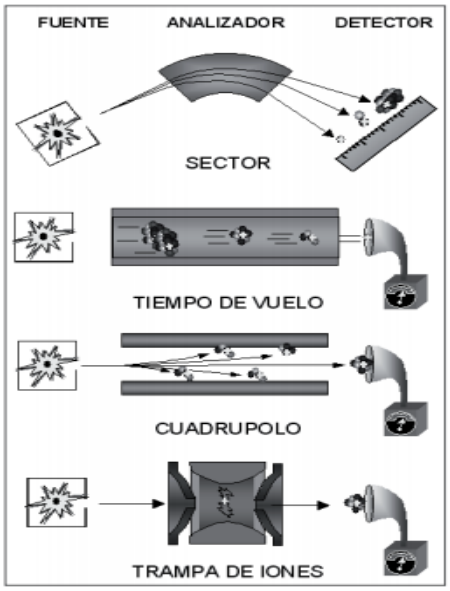


Figura F. Tipos de analizadores de masa. Figura adaptada de (Abián et al., 2008)

Los principales analizadores de masas que son utilizados hoy en día son:

→ Cuadrupolo: consiste en cuatro barras paralelas en las que se aplica un potencial de corriente continua y de radiofrecuencia que forman un campo llamado cuadrupolar. Los iones generados en la fuente traspasan longitudinalmente el espacio entre las barras para llegar al detector *(Figura F)*. Solamente algunos iones con determinada m/z consiguen incidir en el detector ya que, al entrar al analizador, son sometidos al efecto del campo cuadrupolar que hace que muchos de ellos se desvíen. Por lo tanto, solo se puede monitorizar una fracción de los iones mientras que los demás se descartan (Abián et al., 2008).

→ TOF: en este analizador de masas los iones acelerados por un campo eléctrico obtienen velocidades en función del valor de su m/z, por tanto, lo que mide este analizado es el tiempo de vuelo. Es importante mencionar que la carga y la energía cinética de los iones lanzados desde la fuente es constante. Este analizador no desecha ningún ion a diferencia de los cuadrupolos, ya que separa y detecta una escala de tiempo de todos los iones que salen de la fuente *(Figura F)*. El TOF es un analizador adecuado para técnicas de ionización en régimen discontinuo como el MALDI. Si se quiere acoplar a técnicas de producción continua como el ESI, es necesario realizar sistemas de confinamiento de iones intermedios que acumulan e

inyectan paquetes de iones en el analizador de ciclos. En resumen, es un analizador barato, simple, con un ciclo de barrido rápido y un rango de masas en teoría ilimitado (Abián et al., 2008).

→ Confinamiento de iones:

~ Trampas de Iones: estos analizadores posibilitan el confinamiento de iones en una cámara utilizando campos eléctricos o magnéticos y proporcionan la posibilidad de analizar iones formados en la misma trampa o de otras fuentes. Como los iones pueden estar en el interior de la trampa en rangos de tiempo duraderos, se pueden formar fragmentos de iones por la colisión con moléculas de gas. Estos fragmentos se pueden fragmentar otra vez en la misma trampa por lo que se correspondería con un sistema de espectrometría de masas en tándem múltiple. El espacio de la trampa está formado por dos tapas laterales, con agujeros de entrada y salida para los iones, y un anillo central. Las tapas se conectan a un campo de corriente continua y al electrodo anular se le fija un voltaje de radiofrecuencia *(Figura F)*. Finalmente, los iones se extraen de la trampa en base a su m/z a través de técnicas como ESI o MALDI (Abián et al., 2008).

~ Orbitrap: este analizador está formado por una barra y un cilindro con paredes en forma de huso. Los iones adquieren un movimiento radial alrededor de la barra y un movimiento axial en el que su frecuencia es una función de su valor m/z. La señal que se registra es una combinación provocada por las diferentes frecuencias (m/z) y la abundancia de cada uno de los iones. A esta señal generada se le aplica la *transformada de Fourier*, que proporciona la frecuencia de cada ion y, como resultado final, permite obtener el espectro de masas. Este analizador es el más reciente en espectrometría de masas (Abián et al., 2008).

~ FT-ICR: la resonancia ciclotrónica de iones con transformada de Fourier ha sido recientemente introducida en el ámbito de la proteómica, siendo uno de los instrumentos más potentes. Consiste en un campo eléctrico cuadrupolar compuesto con un campo magnético que posibilita el confinamiento de iones, de forma axial por el campo cuadrupolar y de forma radial por el magnético. Para obtener la m/z del ion se calcula la frecuencia de giro para un valor de campo constante (Abián et al., 2008).

Los métodos más distinguidos de ionización son el ESI y MALDI, técnicas que cambiaron el rumbo del estudio en los sistemas biológicos y que, hoy en día, siguen estando presentes en la proteómica. El ESI es una técnica de ionización a presión atmosférica mediante la aplicación de un campo eléctrico, donde se generan iones cuya carga depende del pH de la solución y el número de grupos básicos de la molécula. En principio se demostró su capacidad de análisis de péptidos, pero más tarde se manifestó su competencia para el análisis de proteínas de gran tamaño. Los analizadores a los cuales puede ir este método acoplado son a trampas de iones, orbitraps y FT-ICR. El MALDI es una técnica de desorción por láser que se compone de una matriz química para la deposición de la muestra. Esta matriz absorbe la radiación en función de la frecuencia del láser y es preferentemente acoplada a analizadores TOF (Abián et al., 2008).

2.3. Proteómica e Inmunología

La inmunología es el estudio de los mecanismos fisiológicos que los seres humanos utilizan para la defensa frente a la invasión de otros organismos. Históricamente se empezó a observar que las personas que habían sobrevivido a una enfermedad infecciosa eran inmunes a la infección y, más tarde, se comprobó que los microorganismos externos eran capaces de enviar poblaciones muy grandes contra un solo Homo sapiens. En consecuencia, el Homo sapiens se resiste a través de células dedicadas a la defensa, las cuales forman el sistema inmunitario (Parham, 2006).

El sistema inmunitario es esencial para la supervivencia humana, sin la presencia de este hasta las infecciones menores podrían resultar mortales. Los seres humanos deben enfrentarse a un microorganismo para proporcionar inmunidad protectora, esto expone al sistema inmunitario a un riesgo durante su primera infección por parte de ese microorganismo (Parham, 2006).

2.3.1. Proyecto del proteoma humano

La Organización del Proteoma Humano (HUPO) es un organismo internacional que promueve actividades científicas dirigidas al progreso del estudio del proteoma humano. Este organismo ha desarrollado el Proyecto de Proteoma Humano (HPP), diseñado para mapear todo el conjunto de proteínas humanas. El espacio proteómico incluye un millón de isoformas proteicas distintas y numerosas alteraciones postraduccionales que varían en función del tiempo, la ubicación, la fisiología, etc. (Legrain et al., 2011).

Figura G. HUPO. Figura adaptada de "Human Proteome Organization"

Los grupos de experimentación de HPP utilizan como estrategia general tres bases de trabajo: la espectrometría de masas, la captura de anticuerpos y herramientas de conocimiento bioinformáticas. Estas estrategias están centradas en el mapeo de proteínas con fijación en los cromosomas (C-HPP) y en el Proyecto del proteoma humano de biología y enfermedad (B/D-HPP), basada en incrementar la compresión del proteoma humano a través del estudio de las enfermedades. Esto permitirá tener una base más sólida para la medicina personalizada (Legrain et al., 2011).

Los principales objetivos de la HUPO son:

- → Fomentar la colaboración mundial para el estudio y la profundización en proteómica.
- → Apoyo a proyectos proteómicos orientados al estudio directo de enfermedades humanas.
- → Coordinar la relación de las agencias de financiación con la comunidad proteómica.
- → Producción de una política ética para la reserva y uso de muestras de tejido humano en proyectos proteómicos.

2.3.2. Estudio del sistema inmune mediante proteómica

El sistema inmunológico tiene una heterogeneidad y plasticidad celular única en el organismo humano. Para lograr la inmunidad, la interacción de múltiples tipos de células es muy importante. Debido a que una desregulación de la función inmunológica se vincula con patologías, es necesario un análisis metódico del sistema inmunológico para conseguir avances consistentes en la investigación médica (Meissner & Mann, 2014).

Un planteamiento productivo para estudiar la funcionalidad de las redes celulares elevadamente interconectadas es deconvolucionar su complejidad y determinar los niveles celulares o incluso subcelulares de un sistema modelo concreto. La producción celular es determinada por la integración constante de señales intracelulares y extracelulares y su interacción funcional con distintos procesos biológicos. El desarrollo de técnicas capaces de capturar la composición dinámica de los elementos moleculares en espacio y tiempo permite caracterizar estos procesos y verificar sus mecanismos moleculares subyacentes. Para los ácidos nucleicos, hay asentadas tecnologías que posibilitan la cuantificación completa de la expresión génica, que son capaces de indicar la abundancia de proteínas. Sin embargo, a veces el número de ARN mensajero y proteínas no se correlacionan con los procesos postraduccionales que dirigen la síntesis, degradación y localización de proteínas. Debido a los procesos inmunológicos y biológicos, ocurren modificaciones en las proteínas que alteran las funciones en estas regulando la traducción de señales e interacción de proteínas. Esos mecanismos reguladores se sincronizan de manera dinámica para dirigir la función inmunológica (Meissner & Mann, 2014).

Para el estudio de proteomas en condiciones cambiantes, la proteómica ofrece herramientas que permiten la caracterización de los procesos celulares. Estas herramientas dan lugar a un número infinito de aplicaciones, ya que cualquier muestra biológica que contenga proteínas puede ser analizada. Algunos ejemplos pioneros de aplicación de proteómica en inmunología se pueden visualizar en la *Figura H* (Meissner & Mann, 2014).

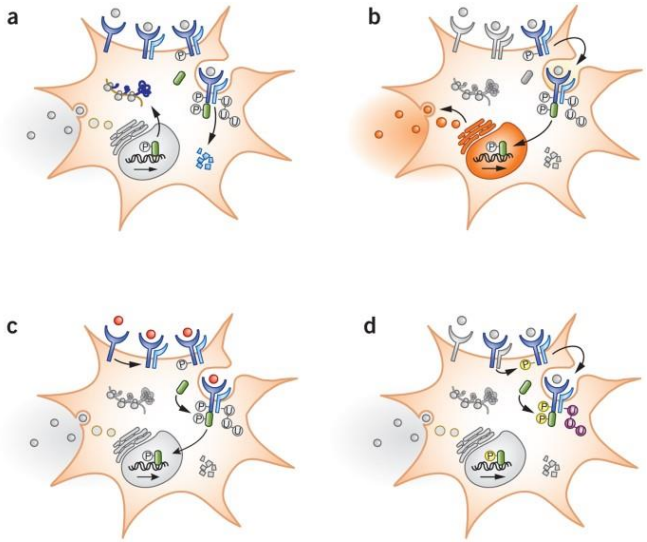


*Figura H. (a) Proteomas totales: composición dinámica de los proteomas totales, incluida la síntesis y degradación de proteínas. (**b**) Proteomas subcelulares: dinámica de proteomas de orgánulos y estructuras subcelulares, tráfico intracelular o secreción de proteínas. (**c**) Interacción proteína-proteína e interacción proteína-ácido nucleico: interacción de proteínas con ADN. (**d**) Modificaciones postraduccionales: dinámica de modificaciones de proteínas covalentes. Imagen adaptada de (Meissner & Mann, 2014)*

2.3.3. Diferenciación de células B humanas antígeno dependiente

Las respuestas inmunes adaptativas son protagonizadas por células que reconocen y responden al antígeno con receptores antígeno-específicos codificados por genes que han sufrido un proceso de recombinación somática (Prieto Martín et al., 2017).

La diferenciación de células B humanas se ha estudiado de forma amplia a niveles genómicos y transcriptómicos, pero a nivel proteómico este estudio, desarrollado por el Servicio de Medicina y Citometría -Núcleo y la Unidad de Proteómica del Centro de Investigación del Cáncer, ha realizado un análisis de las poblaciones de células B humanas que se asocian a la maduración dependiente de antígenos desde un punto de vista proteómico.

El proceso de diferenciación de células B humanas antígeno dependiente comienza cuando los linfocitos B, aún siendo inmaduros, abandonan la médula ósea y finaliza cuando completan su diferenciación en células plasmáticas secretoras de anticuerpos (*Células Plasmáticas*) y *Memoria B*, que son células en órganos linfoides secundarios. Un ejemplo de este tipo de órganos son las amígdalas (Díez et al., 2012).

Este proceso es dinámico y estrictamente regulado y, como ya se ha comentado, comienza tras el abandono de la médula ósea de los linfocitos B y posterior migración por medio de la sangre periférica (PB) al bazo. Allí, maduran en células B vírgenes (*naïve*), que son células B que no han sido expuestas a un antígeno, y se desplazan a los centros marginales de los órganos linfoides secundarios (Díez et al., 2012).

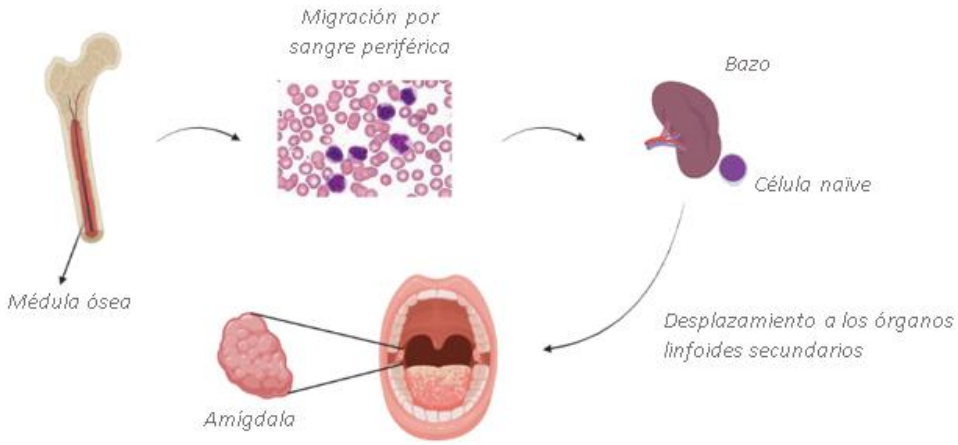


Figura I. Dinámica fisiológica de un linfocito B inmaduro desde la médula ósea hasta el órgano linfoide secundario. Figura creada con BioRender.

Una vez llega a los órganos linfoides secundarios, entra en la zona oscura del centro germinal, donde tiene lugar la proliferación y la hipermutación somática. Esta zona aloja a los *centroblastos* que según (*Tesis Hto Gómez.pdf*, s. f.) son "células de tamaño grande (2-3 veces mayor que un linfocito), con un núcleo redondo, frecuentemente multilobulado y cromatina vesicular, 1 o 3 nucléolos periféricos, un citoplasma escaso y estrecho".

Cuando los *centroblastos* pasan a la zona clara del centro germinal tras la hipermutación somática, se denominan *centrocitos* y son definidos como "células de tamaño pequeño a mediano, con un núcleo alargado, angulado, un nucléolo pequeño y escaso citoplasma" (*Tesis Hto Gómez.pdf*, s. f.). Pese a que *centroblastos* y *centrocitos* tienen aspectos en cuanto

forma muy diferentes, en relación a sus perfiles de expresión genética son muy parecidos (Díez et al., 2012).

La maduración dependiente comienza por la manifestación de antígenos a las células T colaboradoras y la producción de citocinas. Por tanto, esto permite, a través de la recombinación somática, un número grande de anticuerpos en células B en desarrollo. Con la finalidad de incrementar la afinidad de los anticuerpos, se consiguen obtener receptores de células de unión fuerte por medio de un proceso de maduración como resultado de la hipermutación somática de los genes de inmunoglobulina en las células B (Díez et al., 2021).

Posteriormente, las células B abandonan los centros marginales de dos formas posibles:

→ *Células plasmáticas*: este tipo de células secretan anticuerpos específicos contra su antígeno. Pueden residir de forma indefinida en la médula hasta encontrarse con el antígeno y responder de manera muy rápida y eficaz (Turner & Gil-Pulido, s. f.).

→ *Memoria B*: estas células circulan por el organismo en busca de organismos que tengan afinidad con su receptor de célula B. Responden rápida y eficazmente también. Este tipo de células tienen gran importancia en el desarrollo de vacunas ya que cuando el sistema inmunitario sea expuesto a un antígeno previo, las células de *memoria B* serán capaces de responder rápidamente la próxima vez que el organismo se exponga a la infección, impidiendo que se enferme (Turner & Gil-Pulido, s. f.).

En la *Figura J* se puede observar el proceso por el cual una célula *naïve* entra en el centro marginal, pasa ambas zonas y, finalmente, se divide en *células plasmáticas* y de *memoria B*.

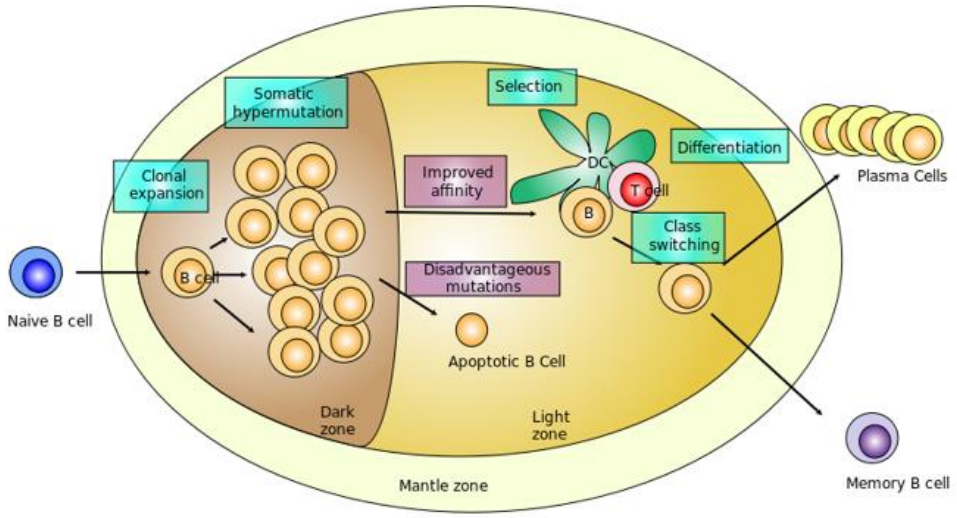


Figura J. Diferenciación de células B antígeno dependiente en el centro germinal. Figura adaptada de (G, 2012)

Se desconocen ciertos rasgos de la vía molecular por la cual se generan células de *memoria B* tras el centro germinal. Según estudios recientes, la afinidad de los receptores de la célula B están implicados en el proceso. Según (Kulis et al., 2015), a través del análisis del metiloma del ADN en 5 subpoblaciones de células B (células pre-BII, células B *naïve*, células B del centro marginal, células de *memoria B* y *células plasmáticas*) se descubrieron niveles más bajos de metilación para las células *memoria B* y *células plasmáticas* de la médula ósea.

Se han realizado varios estudios a través de la estimulación in vitro y para la búsqueda de similitudes entre el proteoma de las células del centro marginal y el proteoma de las células del linfoma de células del manto. Sin embargo, no existen estudios del proteoma cuantitativo de células B primarias humanas no tumorales que derivan de los órganos linfoides secundario durante la maduración de células B dependientes de antígeno (Díez et al., 2021).

Identificar perfiles proteicos ligados a condiciones fisiológicas o patológicas tiene gran relevancia en la comprensión de los mecanismos implicados. Por lo tanto, el estudio del proteoma de células B tumorales se transforma en un paso crítico. El proteoma general cuantitativo de poblaciones de células B no tumorales desde células B *naïve* hasta las *células plasmáticas* y de *memoria B* revela el papel de las diferentes vías metabólicas en las etapas de maduración dependiente del antígeno (Díez et al., 2021).

La leucemia linfocítica crónica (*LLC*) es una neoplasia que se diferencia por una expansión continua de linfocitos B en la sangre periférica, la médula ósea, los ganglios linfáticos y el bazo. Su característica primordial es la acumulación de linfocitos B en la médula ósea y posterior infiltración a los ganglios linfáticos, el bazo y el hígado, desplazando las células normales (Valdespino-Gómez, 2014).

Las células B se activan de forma continua a través de la adquisición de mutaciones que dirigen a la linfocitosis monoclonal de las células B (*LBM*). La acumulación añadida de anormalidades genéticas y la posterior evolución oncogénica de las células B monoclonales produce *LLC*. Es el tipo de leucemia más común en el mundo occidental (Emadi & York Law, 2020).

En la siguiente figura se observa un frotis de sangre periférica de una leucemia linfocítica crónica con linfocitos pequeños, escaso citoplasma y núcleo con forma circular, al lado de linfocitos atípicos, como una célula de gran tamaño y apariencia estimulada (flecha azul) y otra célula con dos núcleos (flecha negra).

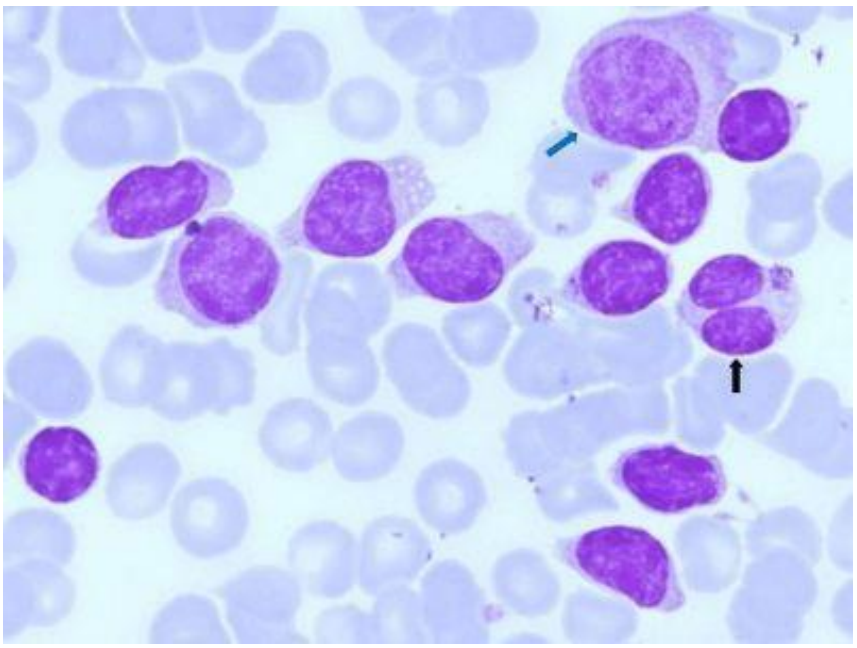


Figura K. Sangre periférica con linfocitos normales y células LLC. Figura adaptada de (Villamor et al., s. f.)

Algunos de los síntomas que poseen los pacientes con *LLC* son fatiga, fiebre, sudores nocturnos y pérdida de peso involuntaria. A medida que avanza la *LLC*, puede causar (Emadi & York Law, 2020):

→ Anemia: es una cantidad disminuida de glóbulos rojos o un valor de hematocrito inferior de lo normal

→ Neutropenia: es un decrecimiento de neutrófilos sanguíneos. Si es una neutropenia avanzada, incrementa el riesgo de infecciones bacterianas.

→ Trombocitopenia: es un trastorno en el que hay una cantidad muy baja de plaquetas en el organismo. Se relaciona a menudo con un sangrado anormal ya que las plaquetas son células que actúan en la coagulación de la sangre.

→ Hipogammaglobulemia: es un trastorno provocado por una deficiencia en las células B, lo que aumenta las gammaglobulinas en el sistema circulatorio. Ocurre en aproximadamente dos tercios de los pacientes y aumenta el riesgo de padecer complicaciones infecciosas.

Los principales métodos de diagnóstico para *LLC* son el hemograma completo y frotis periférico, la citometría de flujo de la sangre periférica y la inmunofenotipificación. A través de la citometría de flujo de la sangre periférica se puede confirmar la clonalidad de las células B, entonces, para los pacientes con un recuento absoluto de linfocitos $< 5 \times 10^9$ / L pero con certeza de clonalidad se diagnostica la linfocitosis de células B monoclonales (*LBM*). El diagnóstico de *LLC* se prevé cuando se encuentra una linfocitosis periférica absoluta, es decir, necesario un recuento de linfocitos $> 5 \times 10^9$ / L (Emadi & York Law, 2020).

CAPÍTULO 3: Objetivos

El objetivo principal de este trabajo es desarrollar flujos de trabajo y estrategias computacionales para el análisis de datos proteómicos. Con el propósito de lograr este objetivo se proponen los siguientes subobjetivos:

1. Subobjetivo 1: Comparación cuantitativa y cualitativa de los tipos celulares.

2. Subobjetivo 2: Hacer uso de técnicas de reducción de dimensiones para una buena visualización.

3. Subobjetivo 3: Creación de mapas de redes neuronales para la clasificación de los datos.

4. Subobjetivo 4: Aplicación de técnicas bioinformáticas con el propósito de relacionar los datos con bases de datos como GO y hallar redes de interacción con sus respectivas rutas de señalización.

CAPÍTULO 4: Material y métodos

4.1. Obtención de las bases de datos

Para este trabajo, se obtuvieron dos bases de datos con valores de intensidad para un gran número de proteínas. Una base de datos contiene valores de intensidad de proteínas para muestras de linfocitos B en sus diferentes estadios (*naïve*, *centroblasto, centrocito, memoria* y *célula plasmática)* y la otra base contiene valores de intensidad de proteínas para muestras de *LLC* y *LBM*. Para la base de datos de linfocitos B se obtuvieron muestras de 5 pacientes y para la base de *LLC* y *LBM* se disponía de 80 muestras de *LLC* y 11 muestras de *LBM*.

Posteriormente, se utilizó el mismo proceso para llegar a ambas bases de datos por lo que se explicará el procedimiento para la obtención de la base de datos con linfocitos B en sus diversos estadios, siendo pasos paralelos para la obtención de la otra base de datos. El camino que se siguió hasta obtener las bases de datos es el siguiente:

→ Muestras y procesamiento:
Se obtuvieron amígdalas humanas recogidas de 5 donantes tras amigdalectomías de rutina. Todos los donantes otorgaron el consentimiento informado de acuerdo con las directrices de ética local del Hospital Universitario de Salamanca, en base a la Declaración de Helsinki. Se obtuvieron suspensiones de células de una única amígdala por desagregación mecánica en PBS. Se tiñeron 150 x 10^6 células de amígdalas en paralelo en varios tubos con un panel de combinación de 8 colores de anticuerpos monoclonales. Posteriormente, las poblaciones de células B se clasificaron sistemáticamente en base a los siguientes fenotipos para clasificar sus valores de pureza (Díez et al., 2021):

- *Naïve*: CD45 $^+$, CD184 $^-$, CD38 $^-$, CD10 $^-$, CD19 / CD20 $^+$, CD3 $^-$, CD27 $^-$
- *Centroblastos*: CD45 $^+$, CD184 $^+$, CD38 $^+$, CD10 $^+$, CD19 / CD20 $^+$, CD3 $^-$, CD27^{het}
- *Centrocitos*: CD45 $^+$, CD184 $^-$, CD38 $^+$, CD10 $^+$, CD19 / CD20 $^+$, CD3 $^-$, CD27^{het}
- Células B de *memoria*: CD45 $^+$, CD184 $^-$, CD38 $^-$, CD10 $^-$, CD19 / CD20 $^+$, CD3 $^-$, CD27 $^+$
- *Células plasmáticas*: CD45 $^+$, CD184 $^-$, CD38 $^{++}$, CD10 $^+$, CD19 / CD20 $^+$, CD3 $^-$, CD27 $^{++}$

→ Extracción de proteínas:
Tras obtener las células purificadas se procede a la extracción de proteínas. Cada población se lavó tres veces con PBS y, tras secar el volumen total de PBS sin modificar el sedimento celular, se agregó un tampón de lisis a la célula. Posteriormente, se centrifugaron a velocidad máxima las muestras y el sobrenadante que contenía las proteínas se acumuló a -80 ^{o}C hasta su procesamiento (Díez et al., 2021).

→ Cuantificación de proteínas:
La cuantificación de proteínas se llevó a cabo utilizando el kit de ensayo de proteínas DC^{TM}II. Se separaron 15 microgramos de proteínas de células B *naïve, centroblasto, centrocitos* y *memoria* en un gel SDS-PAGE, que es el proceso de electroforesis. Tras

esto, los geles se tiñeron y se guardaron en una solución acuosa con ácido acético hasta el análisis. Las *células plasmáticas* se procesaron en solución (Díez et al., 2021).

→ Digestión de proteínas:

La enorme dificultad para aislar *células plasmáticas* de las amígdalas se debe a la presencia de un bajo número de éstas. Esto provoca que la muestra no sea suficiente como para ser procesadas en gel como el resto de los tipos celulares. Por lo tanto, se realizan dos planteamientos diferentes: en gel y en solución. Para la digestión en gel, cada línea de gel se cortó en cinco fragmentos y se digirió con tripsina, después se secaron las muestras parcialmente, se almacenaron a -20 ºC hasta su análisis por LC-MS/MS. En cuanto a la digestión en solución, se redujeron 4 microgramos de proteína con DDT (ditiotreitol, reactivo usado comúnmente como agente reductor), la proteínas se digirieron con tripsina y, finalmente, se utilizó el mismo proceso de digestión en gel (Díez et al., 2021).

→ Análisis LC-MS/MS:

El análisis por espectrometría de masas se realizó en un sistema acoplado a un espectrómetro de masas LTQ Orbitrap, que funciona por confinamiento de iones. La fuente de iones es de nanoelectrospray para el análisis de LC-MS/MS de fase inversa. Los péptidos se cargaron en una columna de captura (columna de trampa Symmetry300 C18 UPLC) y se separaron en una columna BEH130C18 durante 140 minutos para las proteínas digeridas en gel y 170 minutos para las digeridas en solución. El Orbitrap se utilizó en el modo de iones positivos y las exploraciones se obtuvieron en un rango de m/z de 400 a 2000 con una resolución de 30000 en m/z. Los 6 iones más intensos fueron seleccionados en la trampa de iones para su fragmentación inducida por colisión (Díez et al., 2021).

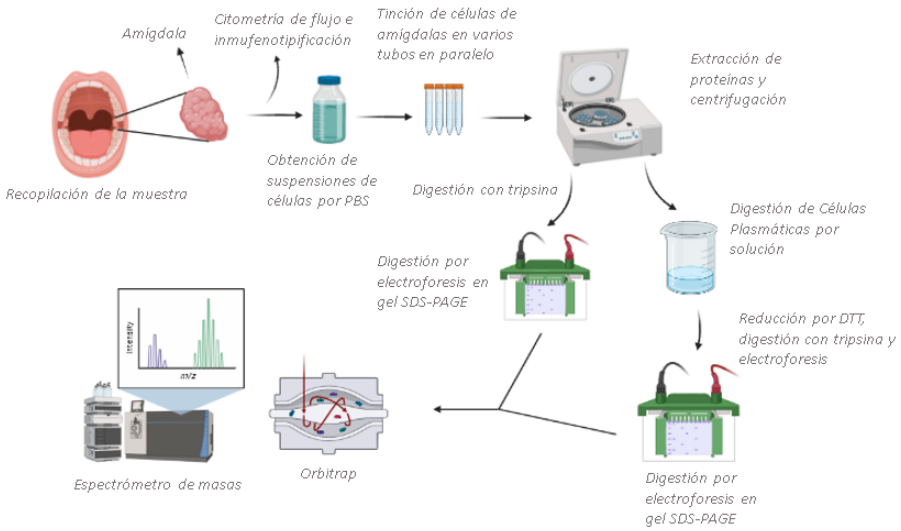


Figura L. Proceso de obtención de la base de datos. Figura creada con BioRender.

→ Análisis de Max-Quant:

Finalmente, a través del software Max-Quant, se analizaron los datos obtenidos por espectrometría de masas, de donde se obtuvieron las intensidades que aparecen en la base de datos final.

4.2. Descripción de las bases de datos

4.2.1 Descripción de las muestras

En primer lugar, la base de datos de linfocitos B se compone de 3098 proteínas identificadas y cuantificadas del proceso anterior de MS/MS con valores de intensidad para 25 muestras en sus diferentes estadios: *naïve*, *centroblastos*, *centrocitos*, *memoria* y *células plasmáticas*. Cada estadio posee 5 muestras. En el siguiente gráfico de barras agrupadas se puede observar el número de proteínas en función de su presencia en, al menos una muestra, y en todas las muestras para cada uno de los estadios del linfocito B.

```
>library(ggplot2)
>ggplot() +
geom_bar(data=barras_agrup,aes(x=Estadio_LinfocitoB,  y=Número_proteinas,fill=Número
_muestras), stat='identity', position='dodge') +
theme_minimal() +
scale_fill_manual(values=c("#132f49","#56b86f"))
```

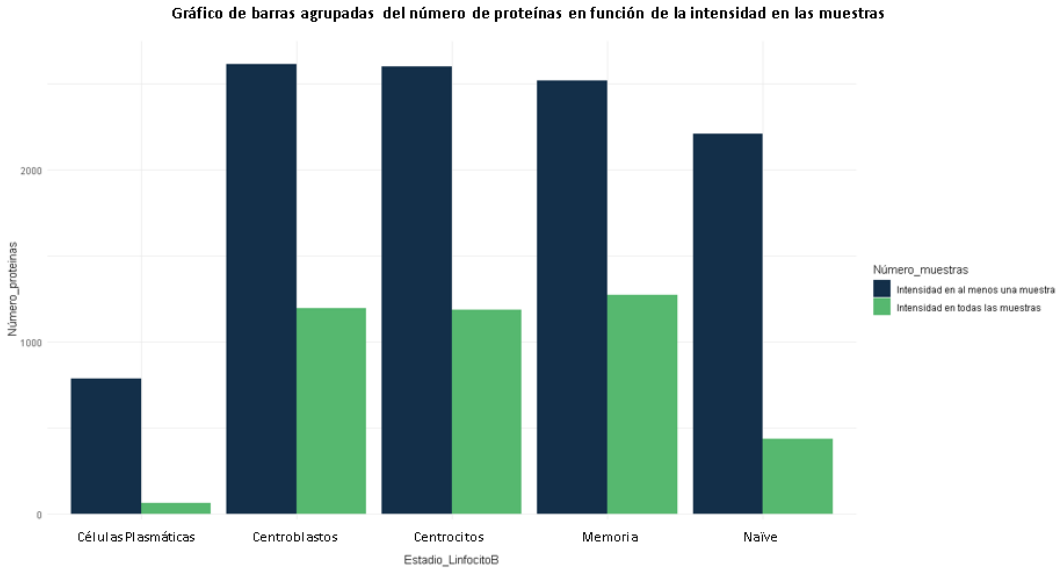


Figura 1. Gráfico de barras agrupadas del número de proteínas en función de la intensidad en las muestras

Se observa que para *células plasmáticas* el número de proteínas identificadas es notablemente inferior en comparación con el resto de los estadios del linfocito B. La explicación biológica es que se dispone de un número muy reducido de este tipo de células. Eso no quiere decir que las proteínas no estén presentes, sino que al haber tan pocas células, no es posible identificar sus proteínas. Por otro lado, se observa que en *centroblastos*, *centrocitos* y *memoria* hay un número de proteínas identificadas muy similar, tanto para todas las muestras como para cuando aparece solamente en una de ellas.

Por otro lado, la base de datos que contiene las muestras de *LLC* y *LBM* contiene 4626 proteínas identificadas y cuantificadas a través del proceso de MS/MS para 80 muestras de pacientes con células *LLC* y 11 muestras para *LBM*. En este caso, apenas hay muestras en las que no haya presencia de intensidad para las proteínas por lo que no es necesario realizar un gráfico de barras para su comparación por muestras.

Además, para que las distribuciones de la intensidad de las proteínas sean comparables, la base de datos debe ser normalizada a través del método de tipificación por el cual a cada

proteína se le resta su intensidad media y se divide por su desviación típica, obteniendo una variable tipificada con media igual a 0 y varianza igual a 1 (MARIA, 2007).

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

Fórmula 1. Tipificación de una variable

Para poder visualizar en qué rango de valores se encuentran los datos originalmente se realiza un boxplot sobre las bases de datos originales (tanto para las muestras de Linfocitos B como para las muestras de *LLC* y *LBM*).

```
>data_peptides$Proteins <- factor(data_peptides$Proteins)
>data_peptides_long <- gather(data_peptides, celulas, intensidad, LLC_1:LBM_65,
factor_key = T)
>ggplot(data_peptides_long, aes(x=celulas, y=intensidad, fill=celulas)) +
geom_boxplot(outlier.shape =NA) +
coord_cartesian(ylim = quantile(data_peptides_long$intensidad, c(0.001,0.95))) +
theme_classic()+stat_summary(fun=mean, geom="point", size=2, color="red")+
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

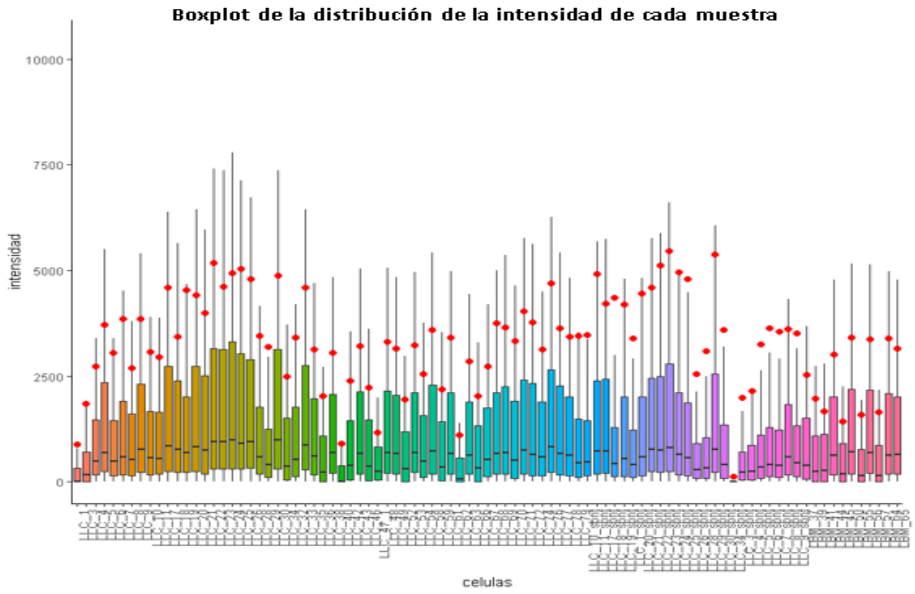


Figura 2. Boxplot de la distribución de intensidad de cada muestra de LLC y LBM

Sobre las intensidades de estas bases de datos se aplica el logaritmo en base 2 para reducir la escala y, posteriormente, se aplica la normalización. En la *Figura 3* se puede observar la distribución de las muestras cuando están normalizadas.

```
>data_peptides_norm <-  mutate_if(data_peptides, is.numeric, log2)
>data_peptides_norm <-  mutate_if(data_peptides_norm, is.numeric, scale)
>data_peptides_norm$Proteins <- factor(data_peptides_norm$Proteins)
>data_peptides_long <- gather(data_peptides_norm, celulas, intensidad, LLC_1:LBM_65,
factor_key = T)
>ggplot(data_peptides_long, aes(x=celulas, y=intensidad, fill=celulas)) +
geom_boxplot(outlier.shape =NA) +
coord_cartesian(ylim = quantile(data_peptides_long$intensidad, c(0.001, 0.9999))) +
theme_classic()+stat_summary(fun=mean, geom="point", size=2, color="red")+
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

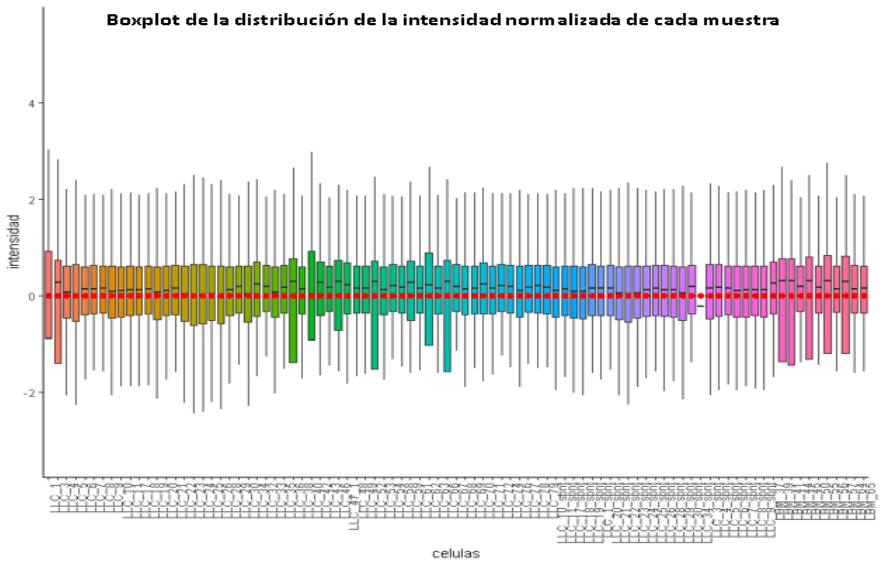


Figura 3. Boxplot de la distribución de la intensidad normalizada de cada muestra de LLC y LBM

Tras ser normalizadas, la distribución de todas las muestras de *LLC* y *LBM* tiene media igual a 0 y están en disposición de ser comparadas entre sí. Las figuras para la base de datos del linfocito B se encuentran en los anexos. Para una visualización más específica del intervalo en el que se encuentran la distribución de intensidades y para la comprobación de una posible presencia de outliers se realiza el histograma de la *Figura 4*. Según (Hawkins, 1980), un *outlier* se define como " valor atípico: es una observación que se desvía tanto de las otras observaciones como para despertar sospechas de que fue generado por un mecanismo diferente".

```
>hist(as.matrix(data_peptides_norm[,-1]), main = "Histograma de la intensidad de las proteínas", freq=F, xlab = "Distribución de la intensidad de las muestras", ylab = "Densidad", col="darkolivegreen1")
```

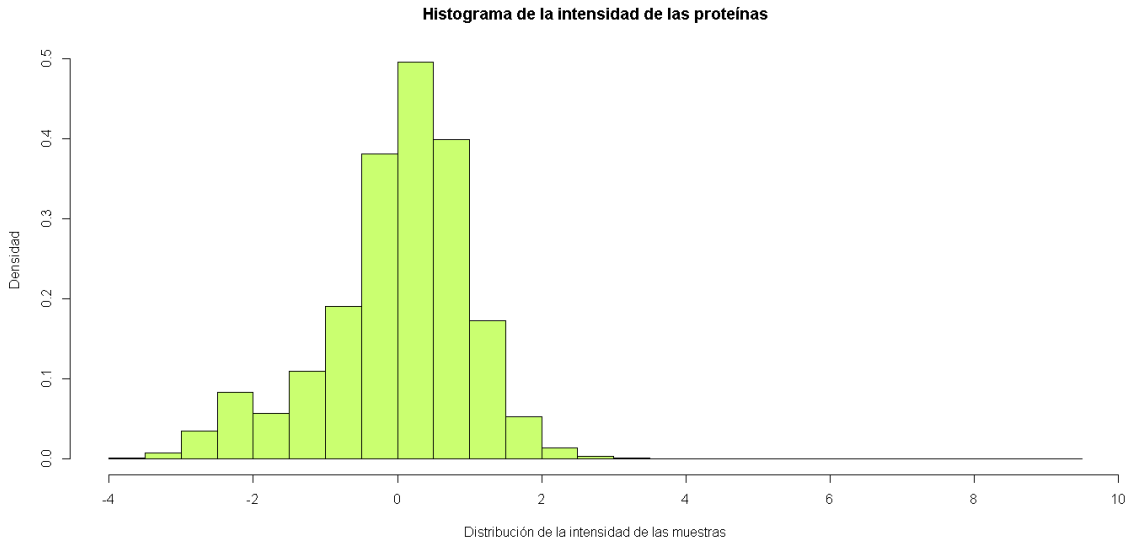


Figura 4. Histograma de la intensidad de las proteínas de LLC y LBM

Según la *Figura 4*, el intervalo de la distribución es (-4,4) y la mayoría de los valores están distribuidos alrededor del 0. Además, se ha comprobado que no hay presencia de outliers por lo que las bases de datos están preparadas para trabajar sobre ellas.

4.2.2. Descripción de los valores objeto

Las dos direcciones para seguir en este trabajo son el estudio de los linfocitos B en sus diferentes estadios y el estudio de los estadios comparándolas con las muestras de *LLC* y *LBM*. Para ello, se calcula el valor medio de cada proteína para los diferentes estadios de la célula B, para *LLC* y *LBM*, formando así una base de datos conjunta donde aparecen todas las proteínas que tengan intensidad para al menos una muestra de uno de los estadios y que tengan intensidad para al menos una muestra de *LLC* y *LBM*. Se decide calcular el valor medio porque para aquellos estadios del linfocito B que tienen 3 o más muestras sin intensidad, el cálculo de la mediana supondría obtener que esa proteína no tiene intensidad para ese estadio. Sin embargo, lo que interesa en el presente trabajo es que, si hay al menos una muestra con intensidad para una determinada proteína, es necesario tener en cuenta la proteína para su posterior estudio. Ambas bases tienen 2512 proteínas en común. En la *Figura 5* se puede observar la distribución del valor medio de las intensidades para cada tipo de célula de las 2512 proteínas que coinciden entre ambas bases de datos.

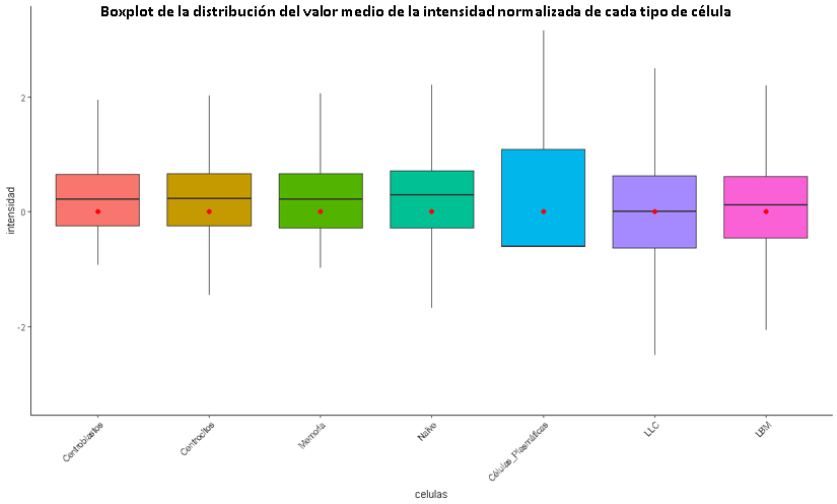


Figura 5. Boxplots de la distribución del valor medio de la intensidad normalizada de cada tipo de célula

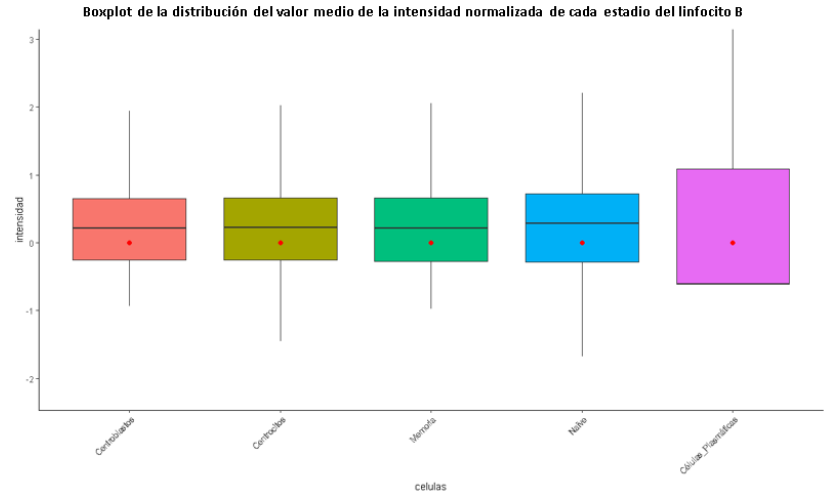


Figura 6. Boxplots de la distribución del valor medio de la intensidad normalizada de cada estadio del linfocito B

En la *Figura 6* se muestra la distribución del valor medio de la intensidad normalizada para cada estadio del linfocito B, teniendo en cuenta las 3098 proteínas presentes inicialmente.

4.3. Correlación de los tipos celulares y diagrama de Venn

Se realizan matrices de correlación para ver la fuerza y dirección de la relación lineal y la proporcionalidad entre dos variables estadísticas. Dos variables cuantitativas se correlacionan cuando los valores de una variable cambian sistemáticamente en relación con sus valores homónimos de la otra variable. Es necesario especificar que la correlación no es lo mismo que causalidad, es decir, la correlación representa una posible relación de similitud, pero no una relación causa-efecto (J.F. Kenney, 1962).

Para medir el grado de correlación que hay entre dos variables cuantitativas, existen diferentes coeficientes estadísticos, como el coeficiente de Pearson o el coeficiente de Spearman. En este trabajo se utiliza el coeficiente de correlación de Pearson puesto que se dispone de unas muestras de grandes dimensiones y es más robusto que el de Spearman que funciona mejor con muestras más pequeñas. Además, las muestras han sido normalizadas con anterioridad, por lo que hacer uso de una técnica paramétrica es eficiente.

Siendo *X* e *Y* dos variables aleatorias, el coeficiente de correlación de Pearson, expresado como $\rho_{X,Y}$, se define:

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

Fórmula 2. Coeficiente de correlación de Pearson

Donde:

- σ_{XY} es la covarianza de las variables *X* e *Y*.
- σ_X es la desviación típica de la variable *X*.
- σ_Y es la desviación típica de la variable *Y*.

Su interpretación geométrica es la siguiente:

Dadas dos variables aleatorias $X(x_1,\ldots,x_n)$ e $Y(y_1,\ldots,y_n)$, se pueden crear los siguientes vectores centrados (J.F. Kenney, 1962):

$X(x_1 - \bar{x},\ldots,x_n - \bar{x})$ e $Y(y_1 - \bar{y},\ldots,y_n - \bar{y})$

El coseno del ángulo alfa es dado por la siguiente fórmula:

$$\rho = \cos(\alpha) = \frac{\sum_{i-1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i-1}^{N}(x_1 - \bar{x})^2} \cdot \sqrt{\sum_{i-1}^{N}(y_1 - \bar{y})^2}}$$

Fórmula 3. Interpretación geométrica del coeficiente de correlación de Pearson

El coeficiente de correlación muestral de Pearson ρ es igual al cos(α), es decir, el coeficiente de Pearson es el coseno del ángulo formado entre los vectores centrados (J.F. Kenney, 1962):

- → Si ρ = 1, α = 0 °, los vectores son colineales y existe una correlación positiva perfecta.
- → Si ρ = 0, α = 90 °, los vectores son ortogonales y no existe una relación lineal entre las variables.
- → Si ρ = -1, α = 180 °, los vectores son colineales en la dirección opuesta y existe una correlación negativa perfecta.

Para representar la correlación entre todos los tipos celulares y sus diagramas de dispersión se hace uso de la función *pairs.panels* sobre los valores normalizados.

```
>pairs.panels(data_peptides_norm, pch=20, stars=T, main="Matriz de correlación de los estadios del linfocito B")
```

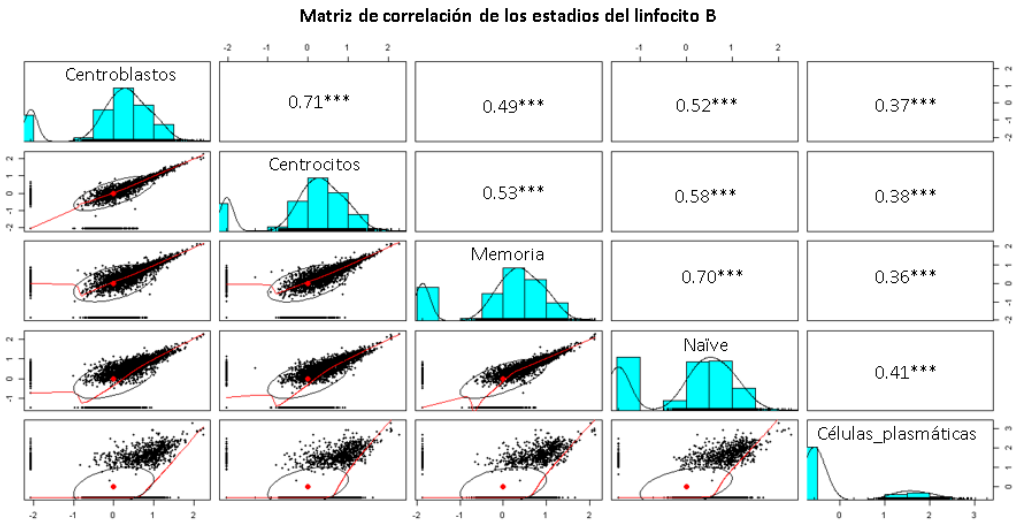


Figura 7. Matriz de correlación de los estadios del linfocito B

Para los valores de intensidad de las 3098 proteínas con presencia en los estadios del linfocito B, se observa que hay una correlación moderada-alta entre todos los estadios, excepto el de *células plasmáticas*, la cual tiene una correlación un poco inferior con los demás estadios. Destaca la correlación alta de 0.71 entre *centroblastos* y *centrocitos*, la cual tiene sentido ya que, como se explica en (Díez et al., 2021), aunque no tengan un aspecto similar, su expresión genética es semejante. Destaca también la correlación de 0.70 entre *naïve* y *memoria*, siendo lo más lógico biológicamente que las células *naïve* se parecieran más a los *centroblastos* o *centrocitos* por su cercanía en los centros marginales de los órganos linfoides secundarios, resulta peculiar una correlación más alta entre esos dos estadios.

Además, a partir del parámetro *stars* en la función *pairs.panels*, se puede obtener la significación de las correlaciones obtenidas, siendo la hipótesis de contraste la siguiente:

- → H_0: No existe relación entre las variables
- → H_1: Existe relación entre las variables

Los tres asteriscos junto a la correlación muestran una muy alta significatividad, por lo que se confirma que no es debida al azar, sino que existe relación entre las variables. Tras realizar un test de significación de correlaciones 2 a 2 con la función *cor.test()* se comprueba que todos los p-valores para las correlaciones entre variables son <2.2 e-16, es decir, se corrobora que existe una relación lineal significativa entre los estadios del linfocito B.

Para visualizar las correlaciones anteriormente obtenidas, se realiza un gráfico de las correlaciones:

```
>correlac <- cor(data_peptides_norm[,-1], method = "pearson")
>corrplot(correlac, method = "shade", shade.col ="NA", tl.col= "black", addCoef.col = "black", main="Gráfico de correlación de los estadios del linfocito B"
```

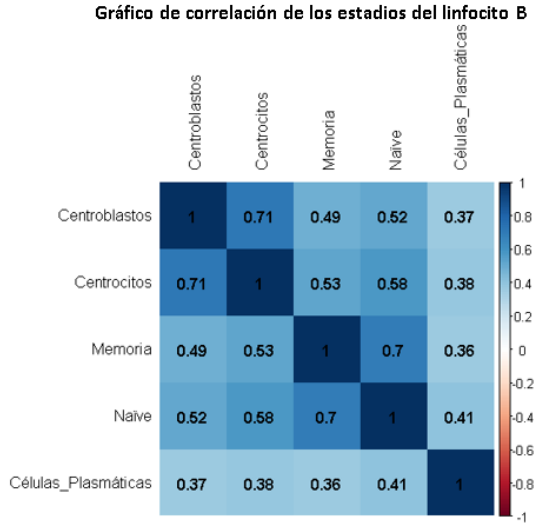


Figura 8. Gráfico de correlación de los estadios del linfocito B

En cuanto a la correlación de todos los tipos celulares, incluidos *LLC* y *LBM*, para las 2512 proteínas comunes, se observa en la *Figura 27* en los anexos que, aunque el número de proteínas se reduzca, la correlación entre *centroblastos* - *centrocitos* y *memoria – naïve* sigue siendo alta: 0.75 y 0.68 respectivamente. Se observa una correlación muy alta, 0.88, entre *LLC* y *LBM*, en cambio, la correlación entre esos dos tipos celulares y los estadios del linfocito B es más baja. La correlación entre *LLC* con el resto de los estadios varía entre 0.3 y 0.4, y la correlación entre *LBM* y el resto de los estadios varía entre 0.2 y 0.3.

Además de la correlación, se realizan diagramas de Venn con el objetivo de comparar el número de proteínas que coinciden entre cada estadio del linfocito B con las células *LLC* y *LBM*. Un diagrama de Venn es un gráfico vinculado con la teoría de conjuntos que muestra las relaciones lógicas entre dos o más conjuntos (Gomero Mancesidor & Gomero Mancesidor, 2017). Se utilizan para representar cómo se relacionan los elementos entre sí en un universo.

Dadas dos variables aleatorias A y B, la intersección de esos dos conjuntos viene dada por:

$$A \cap B = \{x \mid x_i \in A \,\&\, x_i \in B\}$$

Fórmula 4. Intersección de dos conjuntos

Se representa el diagrama de Venn del número de proteínas presentes en al menos una muestra para *centroblastos* y *LLC-LBM* en la *Figura 9*. El resto de los diagramas de Venn se reflejan en los anexos.

```
>Centroblastos<-as.vector(Combined_tonsils_peptide_measurements$Centroblastos)
>LLC_LBM <- as.vector(Combined_tonsils_peptide_measurements$LLC_LBM)
>lista_venn <- list(Centroblastos, LLC_LBM)
>names(lista_venn) <- c("Centroblastos","LLC_LBM")
>vennset=overLapper(lista_venn, type="vennsets")
>vennPlot(vennset,lines=c("#2196f3","#1b5e20","#d50000"),
lcol=c("#2196f3","#1b5e20","#d50000"), ccex=1,lcex=0, mylwd=2)
>upset(fromList(lista_venn), order.by = "degree",point.size = 3.2,cutoff =0,
sets.bar.color =c("#d50000", "#2196f3"),set_size.show = F, text.scale=c(1,1.3,1,1,1.5,1.5) )
```

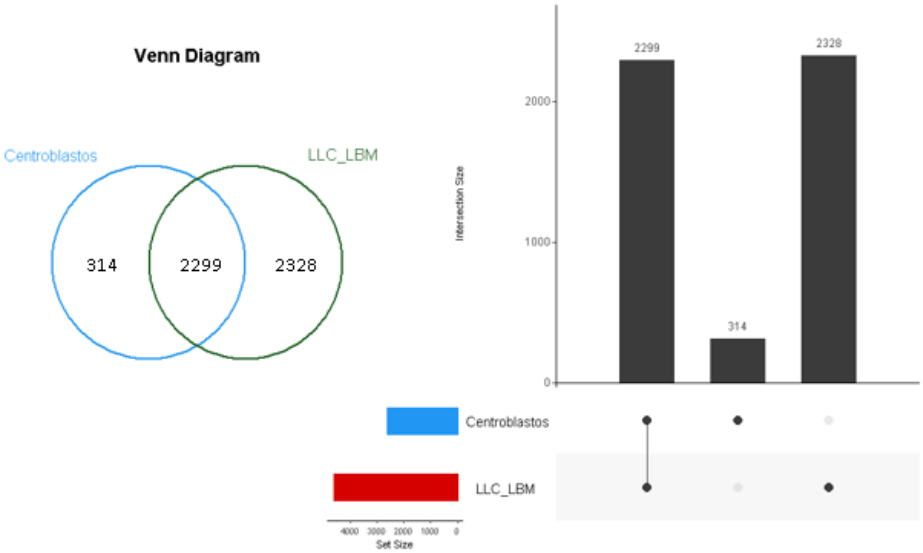


Figura 9. Diagrama de Venn de Centroblastos y LLC-LBM

Se observa que de las 2613 proteínas con valores de intensidad en *centroblastos*, 2299 (87.98%) proteínas coinciden con las presentes en *LLC-LBM*. Para *centrocitos*, de las 2593 proteínas con valores de intensidad, 2279 (87.89%) también se encuentran en las células *LLC-LBM*. Para las células de *memoria*, de las 2512 proteínas con valores de intensidad, 2266 (90.21%) se encuentran en las células *LLC-LBM*. En cuanto a las células *naïve*, de las 2202, 1985 (90.14%) se encuentran presentes en las *LLC-LBM*. Finalmente, de las 786 proteínas con valores de intensidad para *células plasmáticas*, 731 (93%) coinciden con las presentes en las células *LLC-LBM*.

4.4. Análisis de componentes principales

El Análisis de Componentes Principales, también conocido como Principal Component Analysis (PCA), es un método de aprendizaje no supervisado que consiste en encontrar transformaciones ortogonales de las variables originales para conseguir unas nuevas variables no correlacionadas que se ordenan en función de la cantidad de varianza que explican (Aguilar Gutierrez & Vasquez Valdivia, 2017).

La finalidad de esta técnica es conseguir una reducción de dimensiones y explicar la variabilidad inicial de los datos. Para obtener una reducción de dimensiones efectiva es necesario que las variables estén correlacionadas. En el presente trabajo, las correlaciones obtenidas entre las variables son moderadas-altas, además de significativas, por lo que el uso de esta técnica puede ser eficaz.

En el *Gráfico 10* se representan las dos primeras componentes de un PCA de las 5 muestras de cada estadio del linfocito B en función de los valores de intensidad de las 3098 proteínas presentes.

```
>data_peptides_norm_pca <- select(data_peptides_norm, -1)
>pca_normales <- prcomp(data_peptides_norm_pca)
>fviz_pca_ind(X=pca_normales, geom = "point", title="Gráfico PCA de los tipos celulares",
col.ind = data_peptides$Tipo_Celula)
```

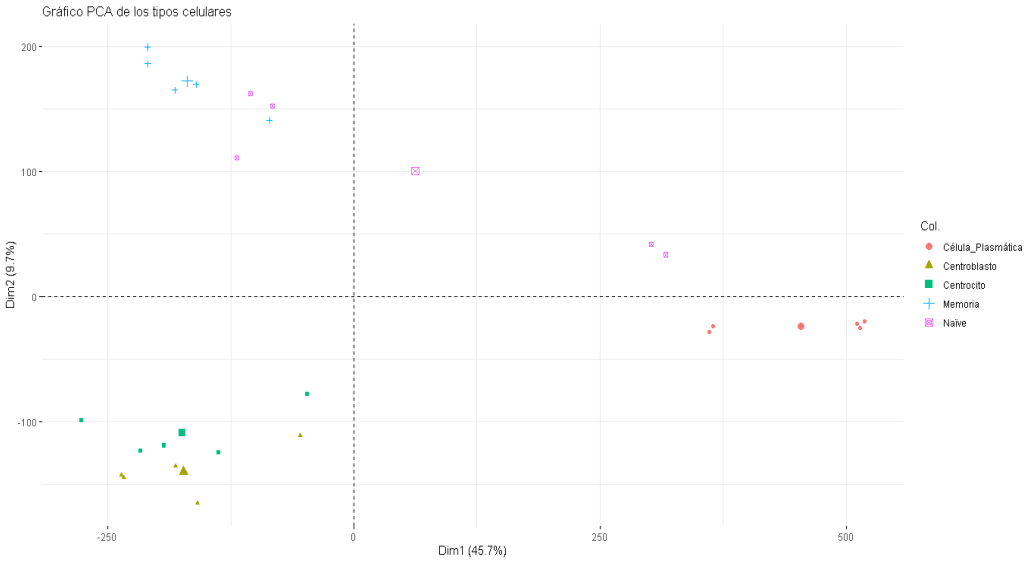


Figura 10. Gráfico PCA de los tipos celulares del linfocito B

Se observa que las *células plasmáticas* se diferencian notablemente del resto de estadios del linfocito B (debido al reducido número de células que se pueden obtener, como se explica en el apartado de *Descripción de las muestras*). Se observan semejanzas entre las muestras de *naïve* y *memoria* y entre las muestras de *centroblastos* y *centrocitos*, por lo que confirman los resultados obtenidos al realizar las correlaciones.

En cuanto al gráfico PCA realizado para comparar las muestras de *LCM* y *LBM* con los estadios del linfocito B, se puede observar en la *Figura 33* en los anexos tres grupos de muestras diferenciados: *LLC-LBM*, *células plasmáticas* y *centroblastos-centrocitos-naïve-memoria*.

4.5. Self-organizing maps (SOM)

4.5.1. Redes neuronales

Los modelos de redes neuronales se utilizan en problemas de reconocimiento de patrones. Estos modelos de redes neuronales artificiales nacen de células neuronales en el cerebro que responden selectivamente a ciertos estímulos sensoriales. Esas células se agrupan en conjuntos locales, cuya ubicación topográfica pertenece a algún valor de característica de un estímulo específico de manera ordenada. Estos conjuntos de células se denominan *mapas cerebrales* (Kohonen, 2013).

En un primer momento, se pensó que estos mapas están establecidos genéticamente, sin embargo, más tarde se descubrió que dependían de experiencias y que se encontraban alterados por estas (Merzenich et al., 1983).

A partir de ahí surgió la idea de crear mapas sensoriales artificiales por aprendizaje, entre ellos los *modelos de redes neuronales de aprendizaje competitivo*. En un subconjunto de células, la adaptación de las células activadas más fuertes a las señales de entrada similares hizo que se conectaran con características de entrada determinadas o sus combinaciones. Además, es necesario tener en cuenta otros tipos de factores adicionales a parte de las conexiones neuronales que tenga la capacidad de mediar la información sin mediar las actividades (Kohonen, 2013).

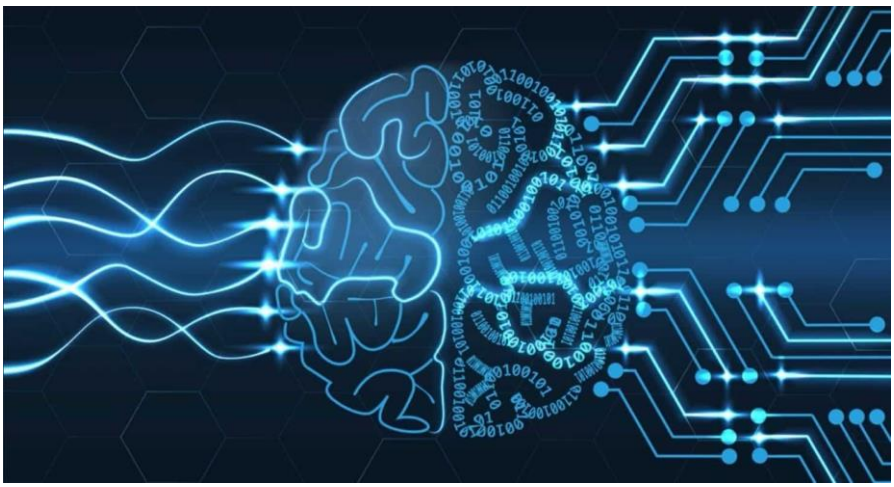


Figura M. Mapas cerebrales formados por células neuronales. Figura adaptada de (Villanueva, s. f.).

En los modelos artificiales de redes neuronales existen dos tipos de aprendizaje: supervisado y no supervisado. La diferencia esencial es que, en el aprendizaje supervisado, el modelo conoce para cada dato de entrada($x_i = i = \{1, \dots, N\}$) su resultado y_i, y produce una predicción $\hat{y}_i$ con fundamento de las $(N-1)$ medidas de respuesta que ya ha entrenado (siendo N el total de resultados que se conoce), sin embargo, en el aprendizaje no supervisado, el modelo únicamente conoce las características de los datos de entrenamiento x_i pero desconoce las medidas de resultado (Pérez, s. f.).

Como en el presente trabajo se utiliza SOM, que es una técnica de aprendizaje no supervisado, se desarrolla éste. Al conocer solamente los datos de entrenamiento, se tiene como entrada un vector $X(x_1, \dots, x_N)$ y una densidad conjunta $\Pr(X)$. El vector $X(x_1, \dots, x_N)$ simboliza las características (variables del conjunto de datos), en consecuencia, la densidad $\Pr(X)$no está condicionada por las inferencias entre las variables.

Para la resolución de problemas con dimensiones altas (gran número de variables), es frecuente el uso de la estadística descriptiva, como gráficas y parámetros estadísticos para describir los datos. Este tipo de técnicas representan conjuntos de valores de X y logran visualizaciones de alta densidad y bajas dimensiones. Otro tipo de técnicas asociadas al aprendizaje no supervisado son los Análisis de Componentes Principales, los Análisis de Clusters o los Self Organizing Maps (Pérez, s. f.)

4.5.2. SOM

(Kohonen, 1982) introdujo un modelo de redes neuronales de aprendizaje no supervisado competitivo denominado *mapas autoorganizados (SOM)*. Las redes neuronales son competitivas cuando las neuronas compiten por activarse, es decir, cuando se introducen los datos de entrada al modelo, se activan una o más neuronas de salida. A estas neuronas se les denomina neuronas vencedoras. Los SOM se fundamentan en aprendizaje no supervisado, por tanto, tras introducir los datos en la red, los datos que pertenezcan a las mismas categorías activan la misma neurona de salida. Solo se activa una neurona, la adecuada para la categoría correspondiente. Esas categorías las crea la propia red, de ahí que se denomine aprendizaje no supervisado (Pérez, s. f.).

El SOM representa conjuntos de datos multidimensionales en una red con menos dimensiones, normalmente es bidimensional, de forma que aquellos datos que sean

similares o adyacentes en el espacio multidimensional también lo sean en el espacio bidimensional (Kohonen & Honkela, 2007).

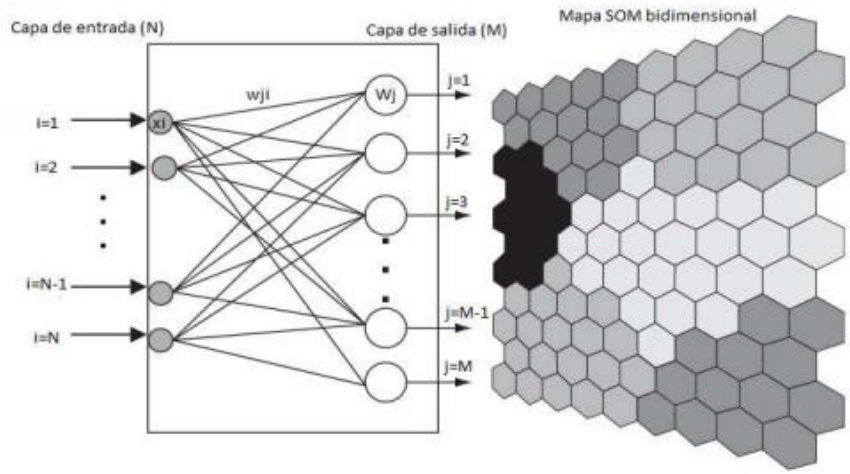


Figura N. Mapa bidimensional del algoritmo del SOM. Figura adaptada de (Pérez, s. f.)

Un modelo de SOM se compone de dos capas de neuronas. La capa de entrada está formada por *n* neuronas, una neurona por cada dato de entrada, que recibe y transmite la información a la capa de salida. La capa de salida es la encargada de procesar la información, crear patrones e identificar las categorías. Las M neuronas de esta capa se estructuran en función de las categorías y se obtiene un mapa bidimensional.

La transmisión de información es *forward*, es decir, se traspasa desde la capa de entrada hasta la capa de salida. Cada neurona de la capa de entrada *i* se encuentra conectada a todas las neuronas de la capa de salida *j* con un peso w_{ij}. De este modo, a las neuronas de la capa de salida (nodos) se les asigna un vector de pesos W_j, que es el vector promedio de la categoría que representa la neurona *j*.

El algoritmo de SOM sigue los siguientes pasos (Kohonen, 1982):

1. El primer paso es asignar a cada nodo un vector de pesos w_j.

2. En el segundo paso se selecciona, para cada dato de entrada x_i, el nodo *j* que sea más cercano en términos de similitud. Se calcula la distancia euclídea del dato x_i a los vectores de pesos W_j y se escoge la neurona a la cual esta distancia sea mínima. A esa neurona *j* se le califica como neurona vencedora. Según (Kohonen, 2013), si los datos están normalizados, la distancia euclídea es aplicable en este tipo de redes neuronales debido a que el SOM es capaz de mostrar complejas interdependencias de las variables en su visualización.

$$j = \underbrace{argmin\left\|x_i - w_j\right\|^2}_{1 \leq j \leq M}$$

Fórmula 5. Selección del nodo j por distancia euclídea

3. En este paso se calculan las tasas de vecindad y aprendizaje. Los nodos vecinos son aquellos nodos con peso w_k, cuya su distancia al peso w_j de la neurona vencedora es

pequeña. La tasa de vecindad es la función que relaciona esa distancia y asigna más peso cuanto más próximo se encuentre del nodo vencedor.

$$h(w_j, w_k) = \exp\left(\frac{-\|w_k - w_j\|^2}{2\sigma(t)^2}\right) \qquad \text{donde} \quad \sigma(t) = \textit{función monótonamente decreciente}$$

Fórmula 6. Tasa de vecindad

La tasa de aprendizaje α es una función que indica cuánto afecta la actualización del peso y es decreciente con el tiempo. Varía entre 0 y 1, aunque normalmente su valor más bajo es aproximadamente 0.01.

$$\alpha(t) = \alpha_0 \left(\frac{\alpha_f}{\alpha_0}\right)^{t/t_\alpha}$$

Fórmula 7. Tasa de aprendizaje

5. En este paso se actualizan los pesos de todos los nodos, siendo el nodo vencedor y sus nodos vecinos los más afectados. En la *Fórmula 8* se especifica el peso de cada nodo tras la actualización (t+1).

$$w_k(t+1) = w_k(t) + \alpha(t) \cdot h(w_j, w_k) \cdot (x_i - w_k)$$

Fórmula 8. Actualización de peso de los nodos

6. Se repiten los pasos 2,3 y 4 hasta que se verifique alguno de los dos criterios de pausa: se alcanza el número máximo de iteraciones o tras varias iteraciones el cambio de vectores de peso no es significativo.

4.5.3. Algoritmo SOM en R

El algoritmo SOM en R permite utilizar las librerías *som* y *kohonen*. La librería *kohonen* está principalmente creada para los SOM de aprendizaje no supervisado por lo que es la que se utiliza en el presente trabajo.

Al aplicar la función *som*, los argumentos necesarios son la base de datos, las dimensiones del mapa que se desea obtener, la topología del mapa y el tamaño de *rlen*. El argumento *rlen* es importante porque, como el SOM es un procedimiento iterativo, marca cuántas veces debe ejecutar el programa. Por tanto, indica el número máximo de iteraciones o, si se programa un número de iteraciones muy alto, permite ver a través de un gráfico el número mínimo de iteraciones para que la media de la distancia al nodo más próximo sea lo más pequeña posible.

Un factor a tener en cuenta al aplicar la función *som* es que hay una componente aleatoria, entonces, cada vez que se produce un entrenamiento, es muy probable que las proteínas no se distribuyan en los mismos nodos.

Cuando aplicas esta función, R devuelve una lista con los siguientes elementos:

- → Los datos originales que se introdujeron como argumento.
- → *Unit.classif*: un objeto numérico que especifica en qué nodo de la capa de salida se encuentra cada individuo (en el presente trabajo se clasifican proteínas).
- → *Grid*: un objeto lista que contiene los parámetros introducidos, es decir, las dimensiones, topología del mapa y un objeto carácter que establece el tipo de función de vecindad.
- → *Codes*: un objeto lista que indica el vector de los pesos de cada nodo de la capa de salida en función de las características de las variables.
- → *Changes*: un objeto numérico que devuelve la media de la distancia al nodo más próximo en cada iteración.
- → *Alpha*: un objeto vector de dos componentes, la tasa de aprendizaje en su inicio y en su final.
- → *Distances*: un objeto numérico que devuelve la distancia de cada proteínas al centro de su nodo

A partir de la función *som*, no solamente se distribuyen las proteínas en categorías en función de sus características, sino que permite realizar gráficos de diferentes estilos:

- → *Counts*: representa el número de proteínas que se asigna a cada nodo. Cuanto más proteínas haya en el nodo, este tomará un color más claro. Si hay algún nodo al cual no se le asigna ninguna proteína, tomará el color gris.
- → *Codes*: muestra los vectores de peso de cada nodo en función de las variables de la base de datos original.
- → *Mapping*: representa donde están situadas las proteínas dentro del nodo. Cuanto más céntricas estén, significa que más se identifican con las características de ese nodo. En este tipo de gráfico se busca observar que haya poca dispersión dentro de cada nodo para obtener una clasificación buena.
- → *DIst.neighbours*: muestra la suma de las distancias a los nodos vecinos, es decir, a todos los nodos contiguos. Es necesario diferenciar por qué la suma de distancias de un nodo respecto a sus vecinos es mayor o menor. Una de las causas principales de tener un nodo con una suma de distancias alta es que los nodos contiguos tengan características muy diferentes.
- → *Quality*: indica la distancia media de las proteínas a su nodo. Cuanto menor sea esa distancia media significa que las proteínas se verán mejor representadas en ese nodo. Este gráfico está relacionado con el de *Mapping* ya que la distancia media se calcula a partir de la distancia de todos los puntos de este gráfico al centro del nodo.
- → *Changes*: muestra la media de la suma de las distancias de un nodo a los nodos contiguos en el eje vertical y el número de iteraciones en el eje horizontal. Este gráfico muestra saltos bruscos hasta que finalmente se obtiene la mínima distancia media en un número determinado de iteraciones.
- → *Property*: representa en qué grado se ve representada cada variable en cada nodo del mapa, permitiendo visualizar nodos con características similares. Para ello, se llama a la función *getCodes*, que proporciona la variable que se quiera observar.

Uno de los argumentos más importantes al aplicar la función *som* es la dimensión del mapa. No hay un proceso específico para la elección de las dimensiones del mapa ni unas dimensiones óptimas. Las dimensiones cambian en función de la información, el tipo y el

detalle de análisis que se quiera realizar por lo que este argumento es una decisión del analista. Escoger unas dimensiones demasiado pequeñas provocaría que en cada nodo hubiera demasiadas proteínas, lo que dificultaría el análisis posterior. Sin embargo, escoger unas dimensiones demasiado grandes, produciría la aparición de varios nodos vacíos y de información no muy concreta.

Para especificar unas dimensiones correctas para el análisis es necesario implementar la función *som* con diferentes combinaciones y elegir la combinación que más favorezca al estudio. En el código de R aparecen las dimensiones 7x8 porque son las que se escogen finalmente para el SOM de las células de los estadios del linfocito B y las células *LLC* y *LBM*, pero *xdim* y *ydim* varían desde las dimensiones 2x2 hasta 10x10 para encontrar las dimensiones finales. También se representa el gráfico *counts* para visualizar la distribución de las proteínas en los nodos.

```
>SOM_proteinas <- as.matrix(SOM_TFG)
>tfg_grid <- somgrid(xdim=7, ydim=8, topo = "rectangular")
>tfg_SOM_model <- som(X = SOM_proteinas, grid = tfg_grid, neigh = "gaussian",
keep.data=T,    alpha=c(0.05,0.01), rlen=230)
>counts <- plot(tfg_SOM_model, type = "counts")
```

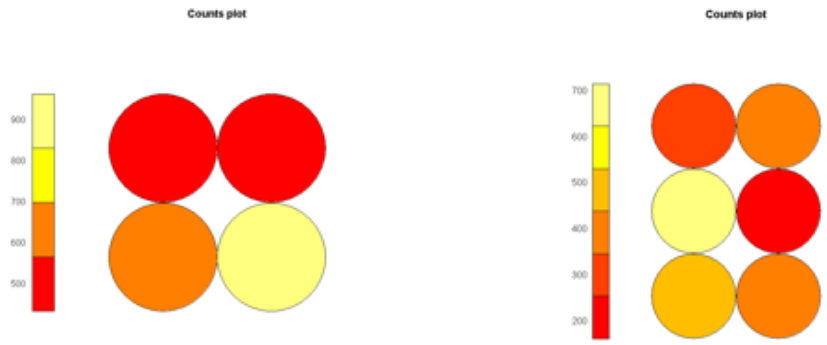


Figura 11. Gráfico counts para dimensiones 2x2 y 2x3

En la *Figura 11* se observan los SOM para unas dimensiones reducidas, lo que provoca que en cada nodo haya demasiadas proteínas. Por ejemplo, en el de dimensiones 2x2 algún nodo contiene 900 proteínas aproximadamente, lo que hace imposible el análisis posterior ya que es muy difícil extraer información de un número tan grande de proteínas.

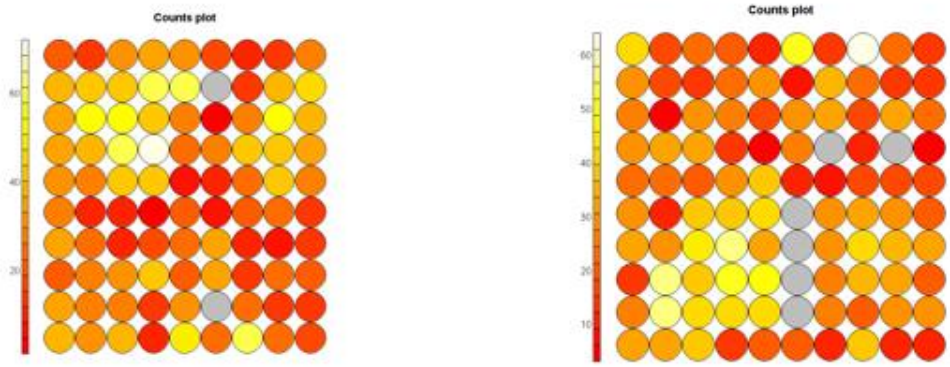


Figura 12. Gráfico counts para dimensiones 9x10 y 10x10

En la *Figura 12* se representan los SOM con dimensiones demasiado grandes. Por un lado, en cada nodo hay un número pequeño de proteínas, pero hay varios nodos vacíos, lo cual no es conveniente.

Por tanto, las dimensiones apropiadas para el SOM son las 6x7, 7x7 y 7x8, ya que no contienen nodos vacíos (la de 7x8 si contiene un nodo vacío, pero solamente uno es insignificante) pero, a su vez, hay una buena distribución de proteínas entre los nodos permitiendo el posterior análisis. Para la elección del mejor mapa entre las tres posibilidades, se aplica la siguiente línea de código en R:

```
>sum(tfg_SOM_model$distances)
```

Este código devuelve la suma de las distancias de cada proteína al centro del nodo que la contiene. Por ende, tener la menor suma de distancias quiere decir que las proteínas se encuentran más cercanas al centro del nodo, es decir, que esas dimensiones para el SOM clasifican mejor. Se obtiene como resultado que las dimensiones 7x8 son las que menor suma de distancias tienen, por tanto, se escogen como las dimensiones finales del SOM para los estadios del linfocito B y las células *LLC* y *LBM*. Con el gráfico *counts* se observa el mapa del SOM para las dimensiones escogidas en la *Figura 13*.

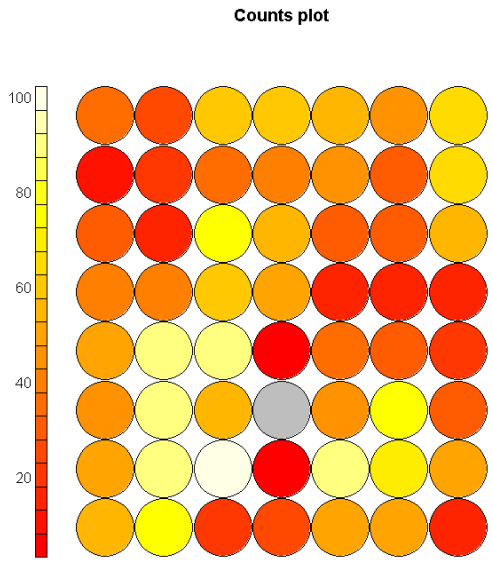


Figura 13. Gráfico counts para el mapa SOM final de dimensiones 7x8 para todos los tipos celulares

Además, es importante tener en cuenta el parámetro *rlen* en la función *som* puesto que va a marcar el número de iteraciones necesarias para el SOM. Para ello, se pone un *rlen* aleatorio con todos los parámetros fijados, incluidas las dimensiones 7x8.

Un *rlen* igual a 300 es suficiente para visualizar el número de iteraciones. Con el gráfico *changes* se puede observar el número de iteraciones mínimas en función de la media de la suma de las distancias de los nodos a sus nodos contiguos. Con un *rlen* igual a 230 sería suficiente, puesto que ya es la media más pequeña, por tanto, 230 son el número mínimo de iteraciones para este SOM.

```
>changes <- plot(tfg_SOM_model, type= "changes")
```

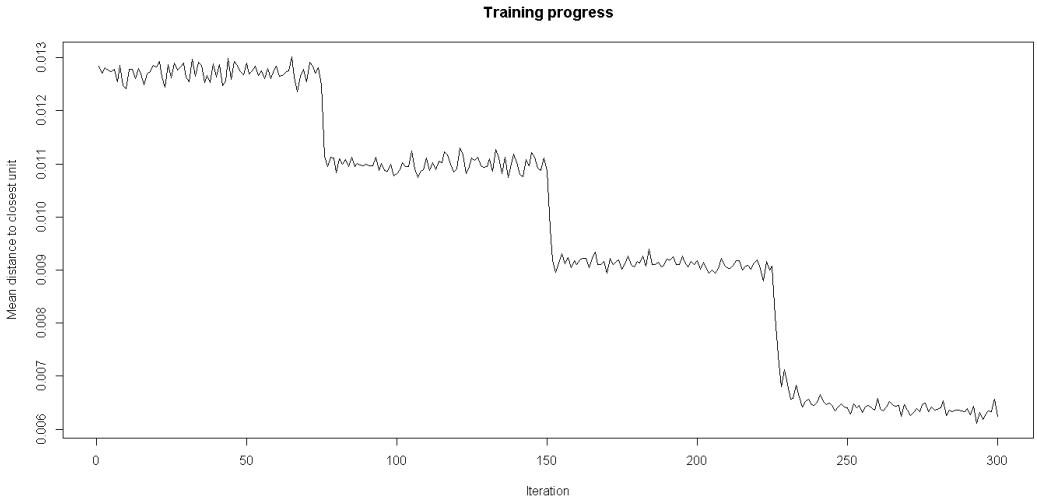


Figura 14. Gráfico changes para visualizar el número de iteraciones

En cuanto al SOM que contiene únicamente los estadios del linfocito B, se llevó exactamente el mismo proceso. Se comparó el gráfico *counts* desde las dimensiones 2x2 hasta las dimensiones 8x8, obteniendo que las dimensiones más adecuadas eran 5x5, 5x6 y 6x6. Posteriormente, se observó la suma de las distancias de cada proteínas al centro de su nodo, consiguiendo como dimensiones finales 6x6 porque tenían la suma de distancias menor. En cuanto al parámetro *rlen*, se escoge 300 como número inicial sobre las dimensiones 6x6. Con el gráfico *changes* se observa que el número mínimo de iteraciones para el SOM que contiene a los linfocitos B únicamente es de 210. Los gráficos de *counts* y de *changes* relacionados con el SOM para linfocitos B se encuentran en los anexos.

Posteriormente, para visualizar la calidad del SOM, se utiliza el gráfico *Quality*, a través del cual se puede observar la media de las distancias de las proteínas al nodo que las contiene. Cuanto menor sea la distancia, más representadas están las proteínas dentro de ese nodo. La *Figura 14* muestra en una escala de color del calor blanco al rojo en función de su calidad, siendo más rojo cuanto mayor calidad.

```
>quality <- plot(tfg_SOM_model, type="quality")
```

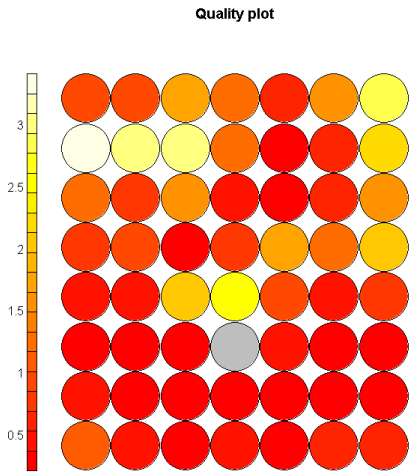


Figura 14. Gráfico quality para visualizar la calidad del SOM

Se observa en la *Figura 14* que la mayoría de los nodos tienen una buena clasificación de las proteínas que contienen. De los 56 nodos, únicamente 5 superan la distancia media al centro del nodo igual a 2.5, con los que habrá que tener precaución si se quieren utilizar para un análisis posterior. Para observar cómo se distribuyen las proteínas dentro de los nodos, se representa el gráfico *Mapping*. La *Figura 18* representa la dispersión de los puntos en los nodos.

```
>mapping <- plot(tfg_SOM_model, type = "mapping", pchs = 20, main = "Mapping Type SOM")
```

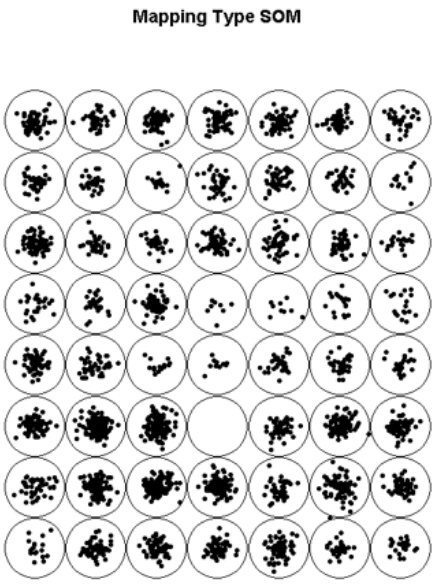


Figura 15. Gráfico mapping para visualizar la dispersión de las proteínas en los nodos

Se observa que los nodos que tienen una media de distancias altas contienen un número bajo de proteínas por lo que, si algunas proteínas se alejan del centro del nodo, la media se verá afectada ya que es una medida de tendencia central sensible a los valores extremos. Por tanto, no porque la media de distancias salga más alta, quiere decir que todas las proteínas de ese nodo estén peor clasificadas, sino que alguna proteína no está clasificada de la forma óptima. Sin embargo, el resto de las proteínas que se encuentran más centradas sí tienen una buena clasificación. En los nodos con grandes números de proteínas, cuya clasificación es buena según la *Figura 15*, los valores extremos no afectan a la media de las distancias de forma tan agresiva ya que hay muchas más proteínas cercanas al centro del nodo. Por tanto, las 2512 proteínas están bien clasificadas en su mayoría.

En cuanto al SOM en el que aparecen solamente los estadios del linfocito B, el gráfico *quality* muestra una buena calidad en la mayoría de los nodos con una media de distancia por debajo de 1.5 (solamente existen 3 nodos de los 36 que superan esa media), por lo que la clasificación de las 3098 proteínas parece correcta casi en su totalidad. En cuanto al gráfico *mapping*, tiene lugar el mismo fenómeno que en el SOM con *LLC* y *LBM*, la calidad de clasificación de los nodos que contienen pocas proteínas es peor. Esto ocurre porque dos o tres proteínas que se alejan bastante del centro del nodo condicionan la media de las distancias. Sin embargo, en los nodos que contienen un gran número de proteínas, las que se alejan del centro del nodo no hacen grandes cambios en la media. Los gráficos de *quality* y de *mapping* para el SOM de únicamente linfocitos B se encuentran en los anexos.

El gráfico *codes* permite representar las características topológicas en forma bidimensional, y agrupar los nodos en función de las características de las proteínas permitiendo una visualización sencilla. Este gráfico muestra a través de un código de colores los vectores de peso de cada nodo. Cuanto mayor es el peso de una variable, mayor porción es representada.

```
codes <- plot(tfg_SOM_model, type = "codes")
```

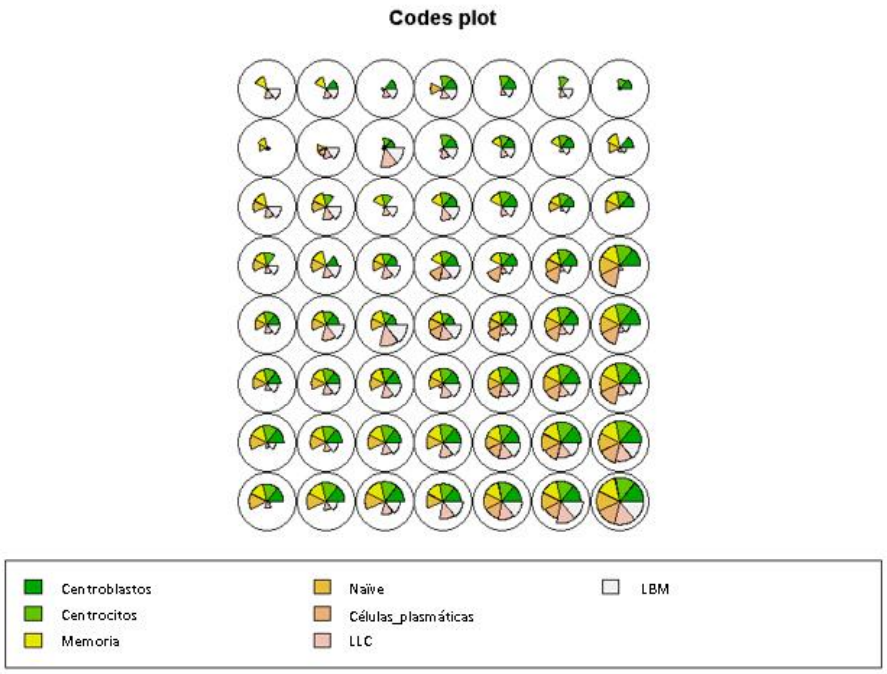


Figura 17. Gráfico codes para visualizar las características topológicas

En la *Figura 17* se puede observar cómo se han distribuido las características en los 56 nodos. Algunos nodos son semejantes a otros contiguos, pero no por estar dos nodos juntos quiere decir que se parezcan. Por ejemplo, las proteínas que se encuentran en el nodo que se encuentra más a la derecha y abajo tienen valores de intensidad altos para todos los tipos celulares. Sin embargo, las proteínas que se encuentran en el nodo de arriba a la derecha poseen valores de intensidad bajos para *centroblastos* y *centrocitos*, y no tienen presencia en los demás tipos celulares.

Si se quiere observar el peso de las variables por separado se puede utilizar el gráfico *property* con la función *getCodes*. Este gráfico permite visualizar nodos con características semejantes. Con la siguiente función se establece un abanico de colores para que los nodos con valores de intensidad altos para un tipo celular se visualicen en rojo y los nodos con valores de intensidad bajos se visualicen en azul.

```
>colores_rainbow <- function(n, alpha = 1) {
rainbow(n, end=4/6, alpha=alpha)[n:1]
}
>property1 <- plot(tfg_SOM_model,
type = "property",property=getCodes(tfg_SOM_model)[,1],
main=colnames(getCodes(tfg_SOM_model))[1], palette.name=colores_rainbow)
```

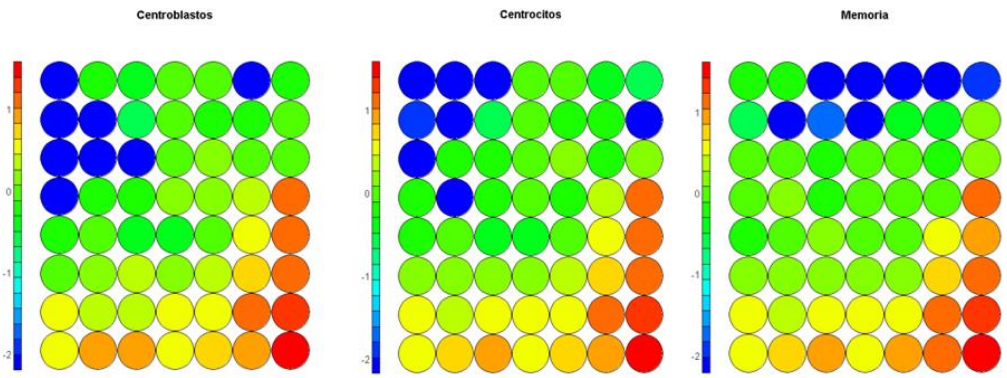


Figura 18. Gráfico property para visualizar el peso de Centroblastos, Centrocitos y Memoria en cada nodo

En la *Figura 18* se observan los nodos con los pesos de las variables *Centroblastos*, *Centrocito* y *Memoria*. El resto de los gráficos *property* tanto del SOM de todos los tipos celulares, como del SOM de los estadios del linfocito B se encuentran en los anexos.

Finalmente, se realiza un dendrograma de los nodos para una posterior agrupación entre nodos con características similares. Para ello se escoge la distancia euclídea y, a través del siguiente código, el método de clustering óptimo para este conjunto de datos.

La función *object.distances* devuelve la distancia euclídea y la función creada *cophe* devuelve la correlación entre la distancia euclídea y el coeficiente cofenético de cada uno de los métodos de clustering. El *coeficiente cofenético* es una medida de bondad de ajuste entre los datos de partida y la estructura del dendrograma, y se calcula como el coeficiente de correlación de Pearson entre la matriz de distancias euclídeas de las observaciones y la matriz de distancias dentro del dendrograma de las observaciones (Santana, 1991). Un valor muy próximo a 1 indica que el dendrograma refleja muy bien la estructura de distancias euclídeas entre las observaciones. Posteriormente, a través de un bucle *for* se pueden visualizar las correlaciones que se obtienen entre la distancia euclídea y el coeficiente cofenético de cada método.

```
>object.dist <- object.distances(tfg_SOM_model, "codes")
>metodos <-c("single", "complete", "average", "ward.D", "ward.D2", "mcquitty",
"median", "centroid")
>cophe <- function(y) {
        distance <- object.dist
        hc <- hclust(d = distance, method = y)
        round(cor(distance, cophenetic(hc)), 4)
}
>for(j in metodos) {
  print(paste(j, cophe(j), sep = ' --> '))
}
   [1] "single --> 0.6352"
   [1] "complete --> 0.7036"
   [1] "average --> 0.6869"
   [1] "ward.D --> 0.6954"
   [1] "ward.D2 --> 0.7103"
   [1] "mcquitty --> 0.6748"
   [1] "median --> 0.5572"
   [1] "centroid --> 0.6719”
```

Se observa que la correlación más alta entre la distancia euclídea y la "distancia dendrográmica" es con el método Ward.D2, por lo que ese método clasificará mejor los nodos. El método Ward.D no aplica directamente el método de clustering de Ward, puesto que se necesitan las distancias al cuadrado. Es el método Ward.D2 el que aplica el criterio correcto (Murtagh & Legendre, 2014).

La distancia euclídea en un espacio n-dimensional se calcula:

$$d(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Fórmula 9. Distancia euclídea en espacio n-dimensional

El método de Ward consiste en ir clasificando observaciones en uno u otro clúster según la distancia entre las observaciones del clúster y el centroide del clúster. El centroide es el punto en el que se minimiza la suma de distancias euclidianas al cuadrado entre ese punto y cada punto del grupo (Strauss & Maltitz, 2017).

A partir del siguiente código, se representa en la *Figura 21* el dendrograma de los nodos pertenecientes al SOM con todos los tipos celulares.

```
>dendrograma<-hclust(object.distances(tfg_SOM_model, "codes"),method="ward.D2")
>plot(dendrograma)
```

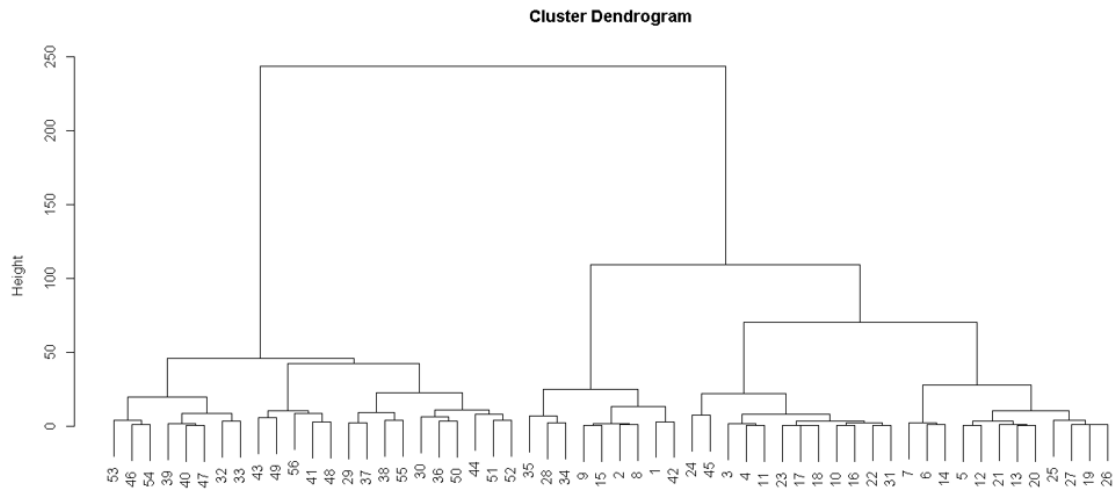


Figura 19. Dendrograma de los nodos pertenecientes al SOM

A partir del dendrograma de la *Figura 19* y del gráfico *codes* de la *Figura 17* se escogen las agrupaciones de nodos que se consideran de interés biológico y estadístico, por ejemplo: buscando agrupaciones de nodos que contengan proteínas con valores de intensidad altos para todos los tipos celulares, que solo tengan valores altos para algún tipo celular, etc.

Para el SOM de los linfocitos B, el gráfico *codes* y el dendrograma se encuentran en los anexos. A partir de esas figuras se obtienen los grupos de nodos para el análisis bioinformático final.

4.6. Enriquecimiento funcional

El análisis de enriquecimiento funcional es un método que identifica clases de genes o proteínas que están sobrerrepresentadas en un gran conjunto de genes o proteínas. Para estudiar la relación entre proteínas que se agrupan en nodos similares se emplearán técnicas bioinformáticas para crear redes de interacción, vías de señalización, etc. Se utilizarán técnicas como KEGG o Gene Ontology (GO). El GO se divide en varios tipos: proceso biológico, función molecular y componente celular. Para cada término GO (expresión de genes, regulación de procesos metabólicos, etc.) se calcula la frecuencia *k* de genes en el conjunto *n* asociados al término, y la frecuencia *K* de genes en el conjunto de población *N* asociados al mismo término. Posteriormente, se prueba cuán probable sería obtener al menos *k* genes asociados a ese término si *n* genes se muestrearan aleatoriamente de la población, dada la frecuencia *K* y el tamaño de *N* de la población. Se aplica la variante de una cola de la prueba exacta de Fisher, la prueba hipergeométrica de sobrerrepresentación. La distribución hipergeométrica mide la probabilidad así (Doyle & Batut, 2021):

$$\mathrm{P}(\mathrm{X} = k) = \frac{\binom{K}{k}\binom{N-n}{K-k}}{\binom{N}{n}}$$

Fórmula 10. Función de probabilidad de una distribución hipergeométrica

A partir de la función *getBM* se crea el universo de proteínas, con la nomenclatura específica para las proteínas, que se utiliza en el paquete *clusterProfiler*. Posteriormente, se crea un objeto de las proteínas escogidas de los nodos con características similares y con la función *match* se llaman con la nomenclatura necesaria. Finalmente se aplica el enriquecimiento funcional a través de *enrichGO*.

```
>library(biomaRt)
>library(clusterProfiler)
>allentrezzz=getBM(attributes = c("uniprotswissprot", "entrezgene_id"), filters =
"uniprotswissprot", values=Proteinas_SOM_LLC_y_LBM_$Proteins, mart =ensembl)
>match_proteins5671214 <- allentrezzz$entrezgene_id[match(nodos5671214$Proteins,
allentrezzz$uniprotswissprot, nomatch=0)]
>match_proteins5671214data <- as.data.frame(match_proteins5671214)
>nodo5671214 <- enrichGO(match_proteins5671214data[,1], pvalueCutoff = 0.05, OrgDb =
org.Hs.eg.db, universe = as.character(allentrezzz$entrezgene_id), readable=T)
>nodo5671214result <- nodo5671214@result
```

Description	*Gene Ratio*	*BgRatio*	*p.adjust*	*geneID*	*Count*
cell adhesion molecule binding	44/200	178/2493	3,3212E-10	ACTN4/P4HB/ATIC/CAPG/CCT8/LASP1/SND1 /DBNL/EZR/MSN/CALR/PTPN6/ MYH9/EHD1/TLN1/ALDOA/ENO1/ PFN1/HMGB1/HSPA5/HSPA8/LCP1/PPIA/ YWHAZ/PDLIM1/PTPN1/PCMT1/TMPO/ SYK/SEPTIN9/CLIC1/HSP90AB1/RPSA/EEF2/ PRDX6/TAGLN2/CAPZB/CAPZA1/HNRNPK/ YWHAE/RAN/RACK1/PRDX1/PARK7	44
protein-containing complex binding	63/200	339/2493	9,7049E-10	ACTN4/WDR1/ATP5PD/GNAI2/P4HB/XRCC5/H CLS1/UQCRC2/EIF4A3/CAPG/C1QBP/LASP1/SF 3B3/SND1/DBNL/ANXA6/EZR/CALR/MYH9/NA PA/TLN1/VIM/HMGB1/HSPD1/HSPA5/HSPA8/ LCP1/ACTB/PPIA/MTA2/TPR/PTPN1/SYK/NSF/ TPM4/SSRP1/SMC1A/NUMA1/CPSF6/SMC3/N	63

				PM1/TUBB/HSP90AA1/HNRNPC/HSP90AB1/RPSA/XRCC6/EEF2/H1-5/VDAC1/PPIB/CFL1/CORO1A/CAPZB/CAPZA1/VCP/ACTR3/ACTR2/YWHAE/EIF5A/RACK1/HNRNPU/PARK7	
actin binding	26/200	100/2493	2,4829E-06	ACTN4/WDR1/P4HB/HCLS1/LSP1/CAPG/LASP1/DBNL/ANXA6/EZR/MSN/MYH9/TLN1/ALDOA/PFN1/LCP1/PDLIM1/TPM4/EEF2/CFL1/CORO1A/CAPZB/CAPZA1/ACTR3/ACTR2/HNRNPU	26
cadherin binding	34/200	158/2493	2,4829E-06	ATIC/CAPG/CCT8/LASP1/SND1/DBNL/EZR/MYH9/EHD1/TLN1/ALDOA/ENO1/PFN1/HSPA5/HSPA8/YWHAZ/PDLIM1/PTPN1/PCMT1/TMPO/SEPTIN9/CLIC1/HSP90AB1/EEF2/PRDX6/TAGLN2/CAPZB/CAPZA1/HNRNPK/YWHAE/RAN/RACK1/PRDX1/PARK7	34
actin filament binding	19/200	60/2493	5,1023E-06	ACTN4/WDR1/HCLS1/CAPG/LASP1/DBNL/ANXA6/EZR/MYH9/TLN1/LCP1/TPM4/EEF2/CFL1/CORO1A/CAPZB/CAPZA1/ACTR3/ACTR2	19

Tabla 1. Enriquecimiento funcional de las proteínas pertenecientes a los nodos 5, 6, 7, 11 y 12

La *Tabla 1* contiene los 5 resultados más significativos de la función *enrichGO* para las proteínas de los nodos 5,6,7,11 y 12. El *p.adjust* es el p-valor ajustado por el método *Benjamini-Hochberg*, que es un test de hipótesis múltiples. El método de *Fisher* es un test global cuando se pretende contrastar la misma hipótesis en estudios independientes (acepta o rechaza H_0), sin embargo, si se pretende especificar qué $H_{0,i}$ se quiere rechazar, se hará uso de test de hipótesis múltiples.

El método de *Benjamini-Hochberg* reduce la tasa de descubrimiento falso, es decir, ayuda a prevenir falsos positivos, errores de tipo I. La probabilidad de rechazar un H_0 siendo verdadero será menor cuanto mayor sea el número de hipótesis que se prueben simultáneamente.

Posteriormente, se utiliza la herramienta STRING, una base de datos que contiene información de diversas fuentes como datos experimentales y métodos de predicción computacional. Por medio de las redes de interacción proteína-proteína se pueden estudiar los procesos celulares, es decir, esas redes pueden servir para evaluar datos genómicos-funcionales. Además, proporciona una plataforma para obtener información sobre las propiedades estructurales, funcionales y evolutivas de las proteínas (Schwartz et al., 2009).

Los nodos de la red representan todas las proteínas producidas por un único locus génico que codifica proteínas. El locus génico es una posición física fija dentro de un cromosoma para localizar un gen. Los nodos coloreados representan una interacción directa entre proteínas y los nodos blancos una relación indirecta. Las aristas representan asociaciones significativas y específicas entre proteínas.

En función del color de la arista, las relaciones serán:

→ Interacciones conocidas: en azul se representan las interacciones de bases de datos seleccionadas y en morado se representan las interacciones que han sido determinadas experimentalmente.

→ Interacciones previstas: en verde se representan las interacciones entre genes vecinos, en rojo fusiones de genes y en azul oscuro coocurrencia de genes.

→ Otras interacciones: las interacciones amarillas representan interacción entre proteínas basadas en publicaciones, las negras coexpresión y las azules oscuras homología de proteínas.

Se han introducido las proteínas de la *Tabla 1* en STRING y se ha obtenido el gráfico de interacciones de la *Figura 20*.

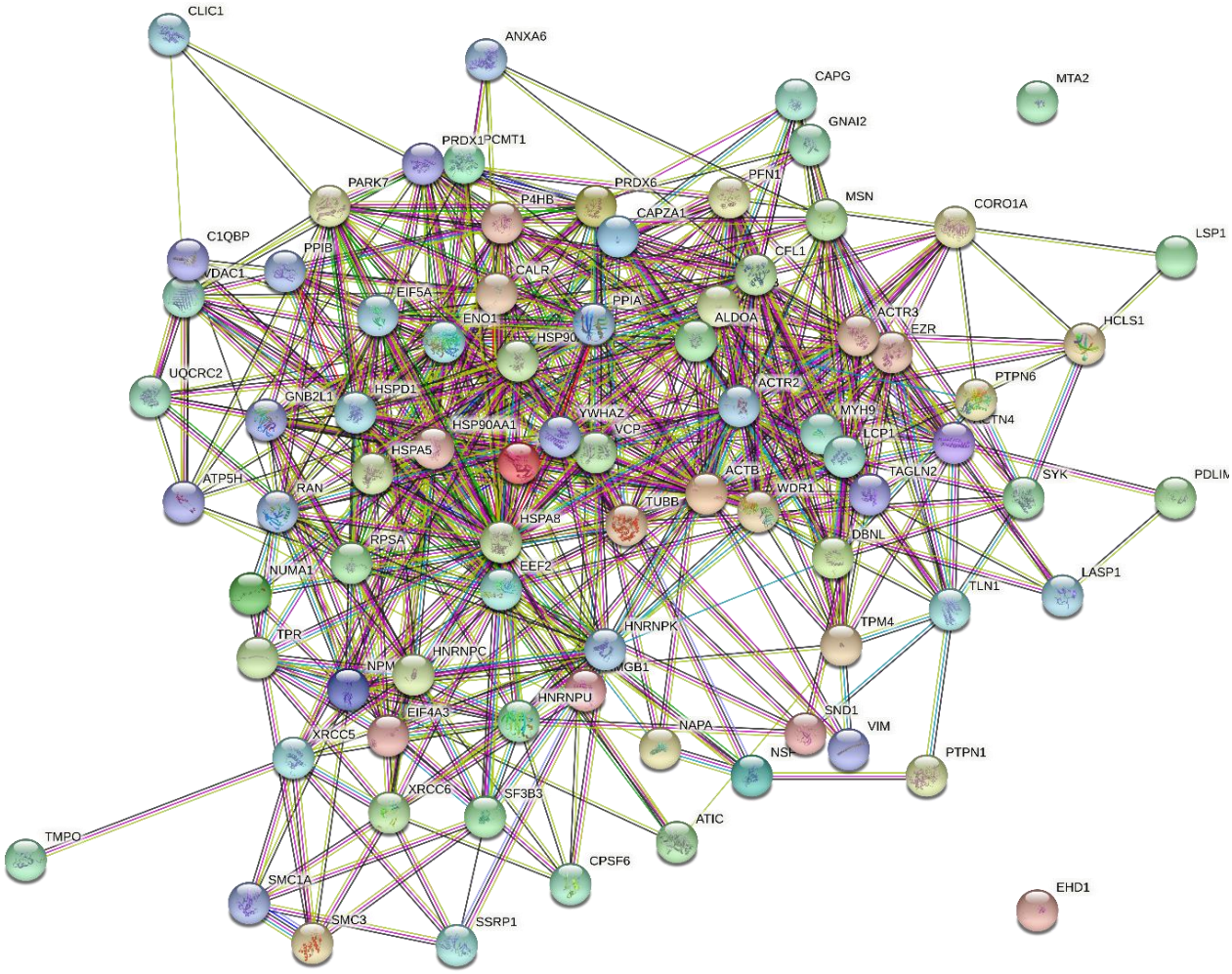


Figura 20. Red de interacción entre proteínas creada en STRING, nodos 5, 6, 7, 11 y 12

Por ejemplo, *PPIA* es una proteína que se encuentra en el centro de la red, cataliza la isomerización cis-trans de enlaces peptídicos imídicos de prolina en oligopéptidos. Esta proteína interacciona con bastantes otras proteínas cercanas a ella, por ejemplo, *CFL1*, *ENO1*, *TAGLN2*, etc. Según el enriquecimiento funcional realizado por la función *enrichGO* en R, *PPIA* realiza la función de unión compleja que contiene proteínas.

Se escoge *CLF1* observando que tiene un *combined-score* de 0.914 sobre 1 de interacción. Sus interacciones proceden de co-expresión, datos experimentales y basadas en publicaciones.

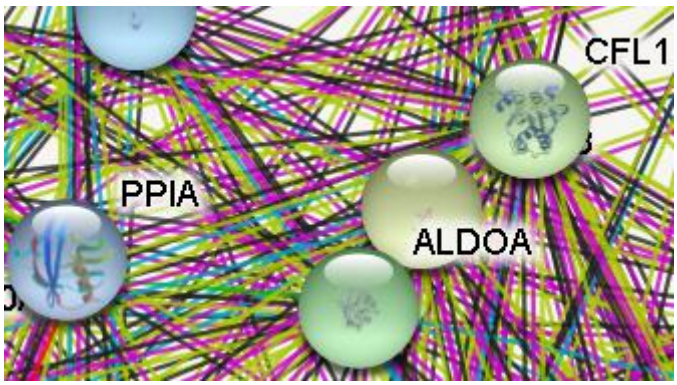


Figura 21. Red de interacción entre proteínas creada en STRING, proteínas (PPIA-CFL1), nodos 5, 6, 7, 11 y 12

Para ampliar la información que se obtiene de STRING, se examinarán las rutas de señalización a través de *Reactome*.

Reactome_Pathway es una biblioteca gratuita, de código abierto y curada, que suministra herramientas bioinformáticas de visualización, interpretación y análisis del conocimiento en vías de señalización (Fabregat et al., 2017).

En la *Figura 22* se observan las rutas de señalización de las proteínas correspondientes a los nodos 5,6,7,11 y 12.

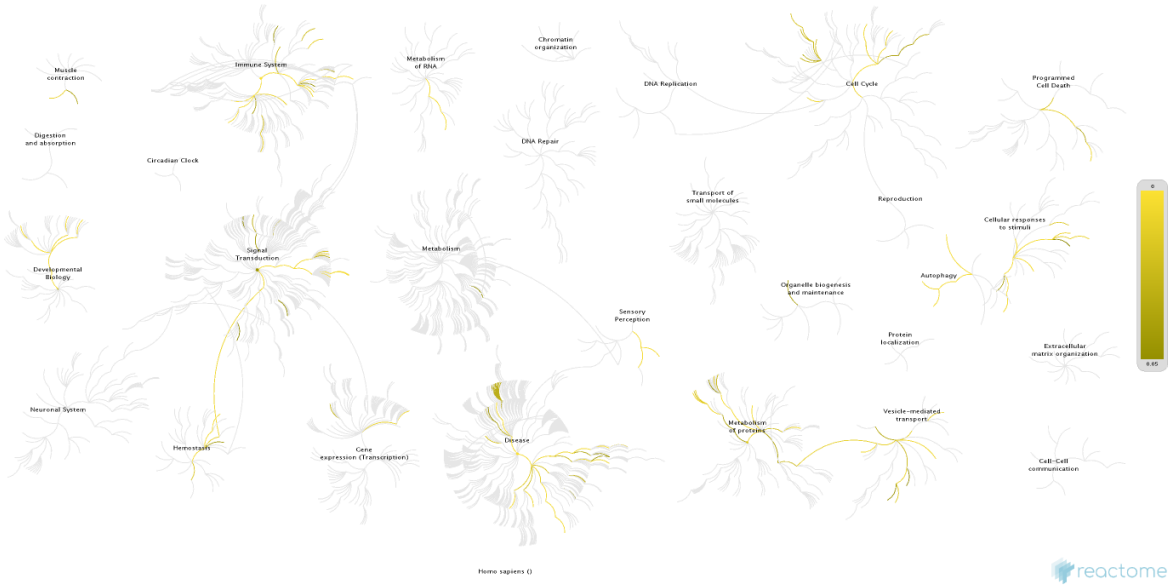


Figura 22. Gráficos con las vías de señalización de los nodos 5, 6, 7, 11 y 12. Figura adaptada de Reactome.org

En la *Figura 22* se observan ciertas rutas marcadas en amarillo. Cuanto más amarillo más se acercará el FDR al 0, por lo que esa ruta de señalización será más significativa. La leyenda solo comprende FDR menores de 0.05.

Pathway name	***Count***	***FDR***
Interleukin-12 signaling	11/84	2.19e-08
Gene and protein expression by JAK- STAT signaling after Interleukin-12 stimulation	11/96	3.87e-08
Interleukin-12 family signaling	10/73	3.87e-08
Signaling by Interleukins	22/643	1.51e-07
Immune System	44/2681	2.13e-06

Pathway name	Count	FDR
Platelet activation, signaling andaggregation	14/291	2.71e-06
Cytokine Signaling in Immune system	25/1092	1.53e-05
Neutrophil degranulation	16/480	2.80e-05
Signaling by Rho GTPases	19/709	4.78e-05
Chaperone Mediated Autophagy	5/23	4.79e-05
Platelet degranulation	9/139	4.79e-05
Signaling by Rho GTPases, Miro GTPases and RHOBTB3	19/725	4.79e-05
Hemostasis	20/801	4.79e-05
Response to elevated plateletcytosolic Ca2+	9/146	5.79e-05
Infectious disease	26/1348	9.62e-05
Regulation of actin dynamics for phagocytic cup formation	9/158	9.62e-05
Selective autophagy	7/89	1.59e-04
ATF6 (ATF6-alpha) activates chaperone genes	4/15	1.77e-04
RHOBTB GTPase Cycle	5/36	2.44e-04
ATF6 (ATF6-alpha) activates chaperones	4/17	2.58e-04
Signaling by ALK fusions and activated point mutants	6/66	2.72e-04
Signaling by ALK in cancer	6/66	2.72e-04
Fcgamma receptor (FCGR) dependent phagocytosis	9/193	3.27e-04
HSP90 chaperone cycle for steroidhormone receptors (SHR) in the presence of ligand	6/72	4.13e-04
Axon guidance	15/584	4.63e-04

Tabla 2. Jerarquía de eventos de las proteínas de la Tabla 1, resultado obtenido de Reactome_Pathway

La proteína escogida *PPIA* se encuentra presente en numerosos *pathways*. Esos pathways se representan en la *Figura 25*.

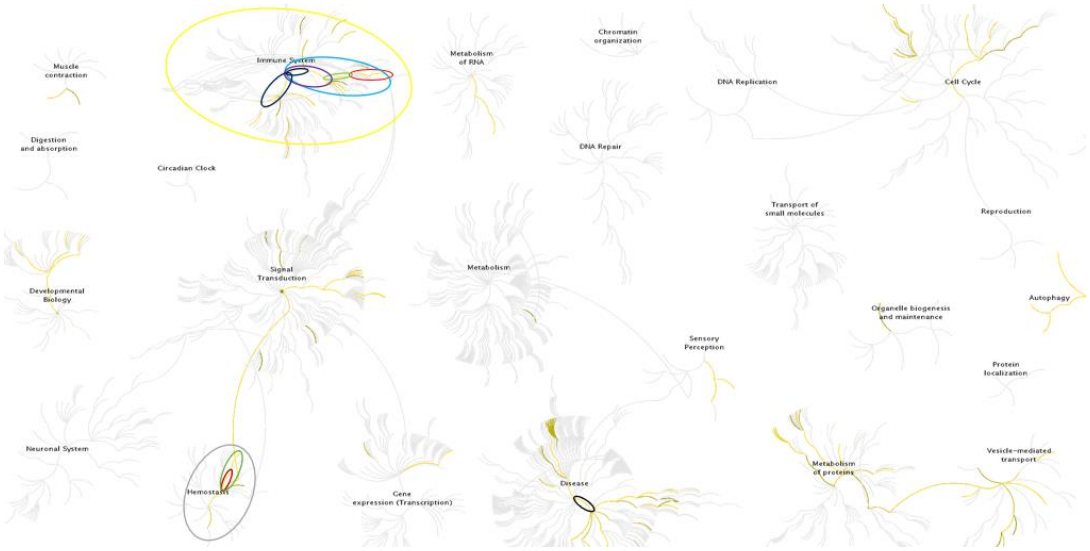


Figura 23. Gráficos con las vías de señalización en las que aparece PPIA. Figura adaptada de Reactome.org

Los colores de la *Figura 23* representan los siguientes *pathways* donde *PPIA* tiene presencia en sus funciones:

→ Interleukin-12 signaling
→ Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation
→ Interleukin-12 family signaling
→ Signaling by Interleukins
→ Immune System

→ Platelet activation, signaling and aggregation
→ Cytokine Signaling in Immune system
→ Neutrophil degranulation
→ Platelet degranulation
→ Hemostasis
→ Infectious disease

Según la *Tabla 2* la rutas de señalización con menor FDR son *interleukin-12 (IL-12) signaling* y *Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation* pertenecientes al *Immune System* por lo que se estudia su comportamiento en esa ruta. Para llegar a esa ruta, se parte del *Cytokine Signaling in Immune System*, posteriormente la ruta continúa por *Signaling by Interleukins*, hacia *Interleukin-12 family signaling* y, finalmente, se llega a *IL-12 signaling* y *Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation.*

Interleukin-12 family signaling se divide en *Interleukin-12, Interleukin-23, Interleukin-27* y *Interleukin-35* por lo que están muy relacionados. En la *Figura 24* se representa la ruta de señalización de *Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation* en la que se señalan las proteínas pertenecientes a los nodos 5, 6, 7, 11 y 12 del presente trabajo y forman parte de la ruta de señalización.

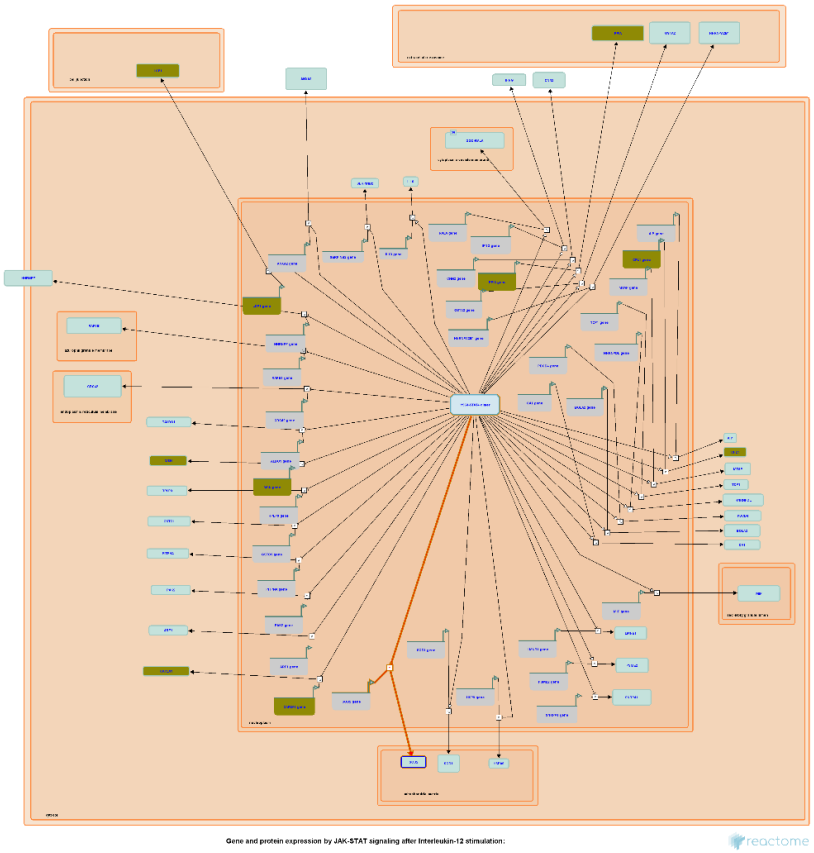


Figura 24. Ruta de señalización de Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation. Figura adaptada de Reactome.org

Las proteínas señaladas en verde en la *Figura 24* son las pertenecientes a la base de datos de este trabajo. El gen de todas las proteínas se encuentra en el nucleoplasma, pero solo nos

centramos en las señaladas. *CAPZA1*, *MSN* y *CFL1* se encuentran en el citosol, que es el componente acuoso del citoplasma de una célula, *LCP1* se encuentra en las uniones celulares y, por último, *PPIA* se encuentra en el exosoma extracelular teniendo 17 interacciones con otras proteínas.

IL-12 es una citocina heterodimérica que induce la producción de interferón-g por asesinos naturales y linfocitos T. Además, *IL-12* está relacionada con la activación de las células B humanas a través del complejo del receptor de *IL-12 (IL-12R)*. Los componentes de *IL-12R*, es decir, las cadenas *b1* y *b2*, se expresan en células B de amígdala de *memoria*, *centro marginal* y *naïve* humanas. Las transcripciones de *IL-12 p35* y p40 se detectaron en todos los subconjuntos, pero solamente las células B de amígdalas *memoria* y *naïve* produjeron *IL-12. IL-12R* se expresa en los principales subconjuntos de células B humanas, pero es funcional en células B *naïve* (Airoldi et al., 2002).

CAPÍTULO 5: Conclusiones

Tomando en consideración los objetivos que se describieron para el presente trabajo, se ha desarrollado una estrategia computacional para llevar a cabo el análisis sistemático de datos proteómica.

En primer lugar, se obtuvieron muestras a partir de las amígdalas de donantes. Se sacaron suspensiones de célula de una amígdala y se tiñeron para realizar un proceso de inmufenotipificación en función de la población de célula B. Posteriormente, se extrajeron las proteínas y se centrifugaron, para realizar la digestión por electroforesis en gel. Finalmente, se realizó el análisis por LC-MS/MS con un espectrómetro Orbitrap y, a través del software Max-Quant, se obtuvieron las señales de intensidad que aparecen en la base de datos final.

Después, se realizó un análisis descriptivo de los datos con el propósito de conocer la distribución de estos. Con el objetivo de poder realizar comparaciones y análisis, se tipificaron las muestras de los estadios del linfocito B para 3098 proteínas y las muestras de *LLC* y *LBM* para las 2512 proteínas comunes entre todos los tipos celulares.

Tras tipificar todas las muestras se procedió al análisis cuantitativo y cualitativo de las muestras. En cuanto al análisis cualitativo, se realizaron diagramas de Venn donde se observó que aproximadamente el 90% de las proteínas que se encontraban en al menos una muestra de cada estadio del linfocito B, también se encontraban en muestras de *LLC* y *LBM*. Para ser más concretos, el estadio con mayor porcentaje de coincidencia fue *células plasmáticas* con un 93%, y el estadio con mejor porcentaje fue *centrocitos* con un 87.89%. Sin embargo, todos son porcentajes de coincidencia altos por lo que se observó a nivel cuantitativo también.

A través de los diagramas de correlación, se observó que *centroblastos* y *centrocitos* tenían una correlación alta, al igual que *memoria* y *naïve*, resultado que se percibe como extraño ya que, a priori, tendría sentido biológico que las células de *memoria* se parecieran más los estadios del *centro germinal*. Todos los tipos celulares tuvieron una correlación baja con *células plasmáticas*, al igual que con *LLC* y *LBM*, cuyas correlaciones con los estadios del linfocito B variaban entre 0.2 y 0.4 para las 2512 proteínas coincidentes.

Posteriormente, a través del uso de Análisis de Componentes Principales (PCA), se hizo una reducción de las dimensiones. Con el gráfico PCA de los estadios del linfocito B, se pretende observar cómo se distribuyen las muestras y explicar la variabilidad de estas. Corroborando los resultados obtenidos en los diagramas de correlación, se observó que había 3 grupos diferenciados en el PCA de los estadios: *centroblastos-centrocitos*, *naïve-memoria* y *células plasmáticas* para las 3098 proteínas. En cuanto al PCA de todos los tipos celulares, se visualizaron 3 grupos igualmente: *LLC-LBM, células plasmáticas y centroblastos-centrocitos-naïve-memoria.*

Con el objetivo de hacer una clasificación de las proteínas en función de sus características se utilizaron Self Organizing Maps, que es un modelo de redes neuronales de aprendizaje no supervisado competitivo. Con esta técnica se crearon mapas de dimensiones 6x6 para las 3098 proteínas presentes en linfocitos B y 7x8 para las 2512 proteínas presentes en todos los tipos celulares. A través de distintos tipos de gráficos, se pudo analizar la calidad de la clasificación, la distribución de las proteínas en los nodos y las características de estos. Después, se buscó que método de clustering correlacionaba mejor con la distancia euclídea, obteniendo el método Ward.D2 y se obtuvieron los dendrogramas donde se visualizaban los nodos.

Finalmente, se seleccionaron nodos con características comunes que tuvieran distancia cercana para realizar el enriquecimiento funcional. Para llevar a cabo el enriquecimiento funcional, se aplicaron técnicas bioinformáticas con el propósito de relacionar los datos con bases de datos como GO y hallar redes de interacción con sus respectivas rutas de señalización.

En primer lugar, a traves de la función *enrichGO* se obtuvieron las funciones más significativas y las proteínas que formaban parte de ellas de los nodos 5, 6, 7, 11 y 12. Para visualizar las conexiones entre las proteínas se utilizó STRING obteniendo un mapa en el que señalaban si las interacciones eran conocidas, previstas u otro tipo de interacciones. Se escogió como ejemplo la proteína *PPIA*, que interactuaba con *CLF1* con un *combined-score* de 0.914 sobre 1 de interacción y sus interacciones procedían de co-expresión, datos experimentales y basadas en publicaciones.

Para aportar más información de las proteínas que aparecían en la cadena de STRING y de *PPIA*, se usó *Reactome*, en el cual, a parte de obtener las conexiones entre proteínas, se obtienen rutas de señalización. Según la *Tabla 2* las funciones con menor FDR estaban relacionadas con el sistema inmune, específicamente en *Interleukin-12 signaling*, obteniendo finalmente la ruta de señalización en la *Figura 26* de *Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation*. Se pudo observar que de las proteínas que aparecen, 5 están contenidas en las proteínas escogidas para estos nodos. Esas proteínas son: *CAPZA1*, *MSN*, *CFL1*, *LCP1* y *PPIA*.

Además de la ruta de señalización, *Reactome* permitió visualizar que *CAPZA1*, *MSN* y *CFL1* se encuentran en el citoplasma, *LCP1* en las uniones celulares y, por último, *PPIA* se encuentra en el exosoma extracelular teniendo 17 interacciones con otras proteínas.

A partir del presente trabajo, se pueden escoger otros grupos de nodos en función de las características biológicas que se requieran y estudiar la expresión de las proteínas que contienen. Además, la clasificación realizada permite obtener nuevas muestras que seguirán los parámetros ya creados.

CAPÍTULO 6: Bibliografía

Abián, J., Carrascal, M., & Gay, M. (2008). *Introducción a la Espectrometría de Masas para la caracterización de péptidos y proteínas en Proteómica*. 20.

Ácidos Nucleicos Qué son, Funciones y Estructura—ADN y ARN. (2017, septiembre 21). *Muy Educativo*. https://muyeducativo.com/biologia/acidos-nucleicos/

Aguilar Gutierrez, L. A., & Vasquez Valdivia, Y. O. (2017). Principal Component Analysis (PCA) para mejorar la performance de aprendizaje de los algoritmos Support Vector Machine (SVM) y Red Neuronal Multicapa (MLNN). *Universidad Privada Antenor Orrego*. https://repositorio.upao.edu.pe/handle/20.500.12759/3398

Airoldi, I., Guglielmino, R., Carra, G., Corcione, A., Gerosa, F., Taborelli, G., Trinchieri, G., & Pistoia, V. (2002). The interleukin-12 and interleukin-12 receptor system in normal and transformed human B lymphocytes. *Haematologica*, *87*(4), 434-442. https://doi.org/10.3324/%x

Brown. (2008). *Genomas/ Genome*. https://books.google.com/books/about/Genomas_Genome.html?hl=es&id=4tYIcMOdsBwC

Castellanos, L. G., Gonzalez, J. L., & PADRÓN, G. (2004). Proteómica. *Combinatoria Molecular. Elfos Scientiae. La Habana*, 367-403.

Cazzulo, J. J. (2014). De la Genómica a la Proteómica. *Manual de Proteómica, Sociedad Española de Proteómica, INTECH, UNSAM-CONICET, Argentina*, 13-20.

Díez, P., Dasilva, N., González-González, M., Matarraz, S., Casado-Vela, J., Orfao, A., & Fuentes, M. (2012). Data Analysis Strategies for Protein Microarrays. *Microarrays*, *1*(2), 64-83. https://doi.org/10.3390/microarrays1020064

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*. https://doi.org/10.3389/fimmu.2021.637832

Doyle, M., & Batut. (2021). *Galaxy Training: GO Enrichment Analysis*. https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/goenrichment/tutorial.html#functional-enrichment-analysis

Emadi, A., & York Law, J. (2020). *Leucemia linfocítica crónica (LLC)—Hematología y oncología*. Manual MSD versión para profesionales. https://www.msdmanuals.com/es-mx/professional/hematolog%C3%ADa-y-oncolog%C3%ADa/leucemias/leucemia-linfoc%C3%ADtica-cr%C3%B3nica-llc

Fabregat, A., Sidiropoulos, K., Viteri, G., Forner, O., Marin-Garcia, P., Arnau, V., D'Eustachio, P., Stein, L., & Hermjakob, H. (2017). Reactome pathway analysis: A high-performance in-memory approach. *BMC Bioinformatics*, *18*(1), 142. https://doi.org/10.1186/s12859-017-1559-2

G, P. (2012, septiembre 15). *Research: EBV persistence in B cells*. The MS-Blog. https://multiple-sclerosis-research.org/2012/09/research-ebv-persistence-in-b-cells/

Gomero Mancesidor, J. M., & Gomero Mancesidor, F. R. (2017). Teoría de Conjuntos y Lógica Matemática. *Universidad Nacional de Barranca*. http://repositorio.unab.edu.pe/handle/UNAB/19

Hawkins, D. M. (1980). A single outlier in normal samples. En D. M. Hawkins (Ed.), *Identification of Outliers* (pp. 27-41). Springer Netherlands. https://doi.org/10.1007/978-94-015-3994-4_3

J.F. Kenney. (1962). *Mathematics Of Statistics Part Two*. http://archive.org/details/in.ernet.dli.2015.223161

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59-69. https://doi.org/10.1007/BF00337288

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, *37*, 52-65. https://doi.org/10.1016/j.neunet.2012.09.018

Kohonen, T., & Honkela, T. (2007). Kohonen network. *Scholarpedia*, *2*(1), 1568. https://doi.org/10.4249/scholarpedia.1568

Kulis, M., Merkel, A., Heath, S., Queirós, A. C., Schuyler, R. P., Castellano, G., Beekman, R., Raineri, E., Esteve, A., Clot, G., Verdaguer-Dot, N., Duran-Ferrer, M., Russiñol, N., Vilarrasa-Blasi, R., Ecker, S., Pancaldi, V., Rico, D., Agueda, L., Blanc, J., … Martín-Subero, J. I. (2015). Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nature Genetics*, *47*(7), 746-756. https://doi.org/10.1038/ng.3291

Legrain, P., Aebersold, R., Archakov, A., Bairoch, A., Bala, K., Beretta, L., Bergeron, J., Borchers, C. H., Corthals, G. L., Costello, C. E., Deutsch, E. W., Domon, B., Hancock, W., He, F., Hochstrasser, D., Marko-Varga, G., Salekdeh, G. H., Sechi, S., Snyder, M., … Omenn, G. S. (2011). The Human Proteome Project: Current State and Future Direction. *Molecular & Cellular Proteomics*, *10*(7), M111.009993. https://doi.org/10.1074/mcp.M111.009993

MARIA, M. L., JOSE. (2007). *Estadística descriptiva*. Editorial Paraninfo.

Meissner, F., & Mann, M. (2014). Quantitative shotgun proteomics: Considerations for a high-quality workflow in immunology. *Nature Immunology*, *15*(2), 112-117. https://doi.org/10.1038/ni.2781

Merzenich, M. M., Kaas, J. H., Wall, J., Nelson, R. J., Sur, M., & Felleman, D. (1983). Topographic reorganization of somatosensory cortical areas 3b and 1 in adult

monkeys following restricted deafferentation. *Neuroscience*, *8*(1), 33-55. https://doi.org/10.1016/0306-4522(83)90024-6

Mojica, T., Sánchez, O., & Bobadilla, L. (2003). La Proteómica, otra cara de la genómica. *nova*, *1*(1), 13-16.

Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, *31*(3), 274-295. https://doi.org/10.1007/s00357-014-9161-z

Parham, P. (2006). *Inmunología.* Ed. Médica Panamericana.

Pérez, M. M. (s. f.). *APLICACIÓN DE K-MEANS Y SOM (SELF-ORGANIZING MAPS) AL ANÁLISIS MICRO DE ACCIDENTES DE TRÁFICO*. 116.

Prieto Martín, A., Barbarroja Escudero, J., Haro Girón, S., & Monserrat Sanz, J. (2017). Respuesta inmune adaptativa y sus implicaciones fisiopatológicas. *Medicine - Programa de Formación Médica Continuada Acreditado*, *12*(24), 1398-1407. https://doi.org/10.1016/j.med.2016.12.008

Santana, O. F. (1991). El análisis de cluster: Aplicación, interpretación y validación. *Papers: revista de sociologia*, 65-76.

Schrenzel, J., Kostic, T., Bodrossy, L., & Francois, P. (2009). Introduction to Microarray-Based Detection Methods. En *Detection of Highly Dangerous Pathogens* (pp. 1-34). John Wiley & Sons, Ltd. https://doi.org/10.1002/9783527626687.ch1

Schwartz, A. S., Yu, J., Gardenour, K. R., Finley Jr, R. L., & Ideker, T. (2009). Cost-effective strategies for completing the interactome. *Nature Methods*, *6*(1), 55-61. https://doi.org/10.1038/nmeth.1283

Strauss, T., & Maltitz, M. J. von. (2017). Generalising Ward's Method for Use with Manhattan Distances. *PLOS ONE*, *12*(1), e0168288. https://doi.org/10.1371/journal.pone.0168288

Templin, M. F., Stoll, D., Schrenk, M., Traub, P. C., Vöhringer, C. F., & Joos, T. O. (2002). Protein microarray technology. *Drug Discovery Today*, *7*(15), 815-822. https://doi.org/10.1016/S1359-6446(00)01910-2

Tesis Hto Gómez.pdf. (s. f.). Recuperado 7 de junio de 2021, de http://incan-mexico.org/incan/docs/tesis/2014/subespecialidad/Tesis%20Hto%20G%C3%B3mez.pdf

Turner, V., & Gil-Pulido, J. (s. f.). *Activación de las células B y formación de centros germinales*.

Tyanova, S., Temu, T., Carlson, A., Sinitcyn, P., Mann, M., & Cox, J. (2015). Visualization of LC-MS/MS proteomics data in MaxQuant. *PROTEOMICS*, *15*(8), 1453-1456. https://doi.org/10.1002/pmic.201400449

Valdespino-Gómez, V. M. (2014). *Leucemia linfocítica crónica de linfocitos B: un modelo personalizado de valoración clínica y molecular*. 19.

Villamor, Rozman, & Calvo. (s. f.). *Leucemia linfática crónica*. Recuperado 17 de junio de 2021, de http://atlas.gechem.org/es/component/k2/item/598-leucemia-linfatica-cronica

Villanueva, J. D. (s. f.). *Redes neuronales desde cero (I)—Introducción—IArtificial.net*. Recuperado 24 de junio de 2021, de https://www.iartificial.net/redes-neuronales-desde-cero-i-introduccion/

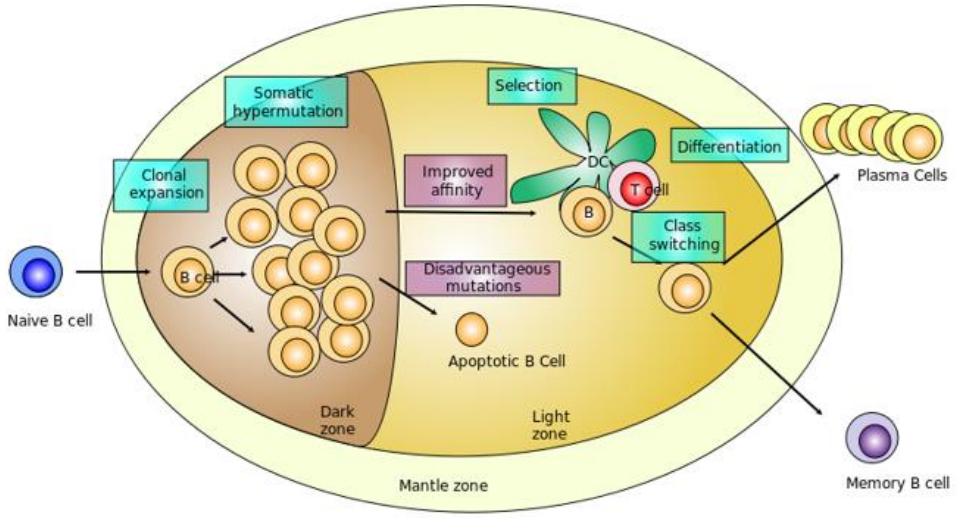
CAPÍTULO 7: Summary

Proteomics is the study of proteomes, where large-scale proteins are separated, identified, and characterized. It also defines cell protein levels, explains their metabolic functions and interrelations. In short, proteomics performs a functional characterization of proteins and their structural relationships, in addition to their previous analysis.

There are multiple ways to do proteomics studies, but the main ones are electrophoresis and mass spectrometry. In the present study, these two techniques are performed for the study of B-cell dependent differentiation.

The stages of B-cell from when it leaves the bone marrow until it finally undergoes a splitting process are the following: *naïve, centroblasts, centrocytes, memory and plasma cells*. This process is dynamic and strictly regulated and, as already mentioned, begins after the abandonment of the bone marrow of B lymphocytes and subsequent migration through the peripheral blood (PB) to the spleen. There, they mature into *naïve* which are B cells that have not been exposed to an antigen and move to the marginal centers of secondary lymphoid organs. Once it reaches the secondary lymphoid organs, it enters the dark zone of the marginal centre, where proliferation and somatic hypermutation take place. This area accommodates the *centroblasts*. When the *centroblasts* pass into the clear zone of the germinal center after somatic hypermutation, they are called *centrocytes*.

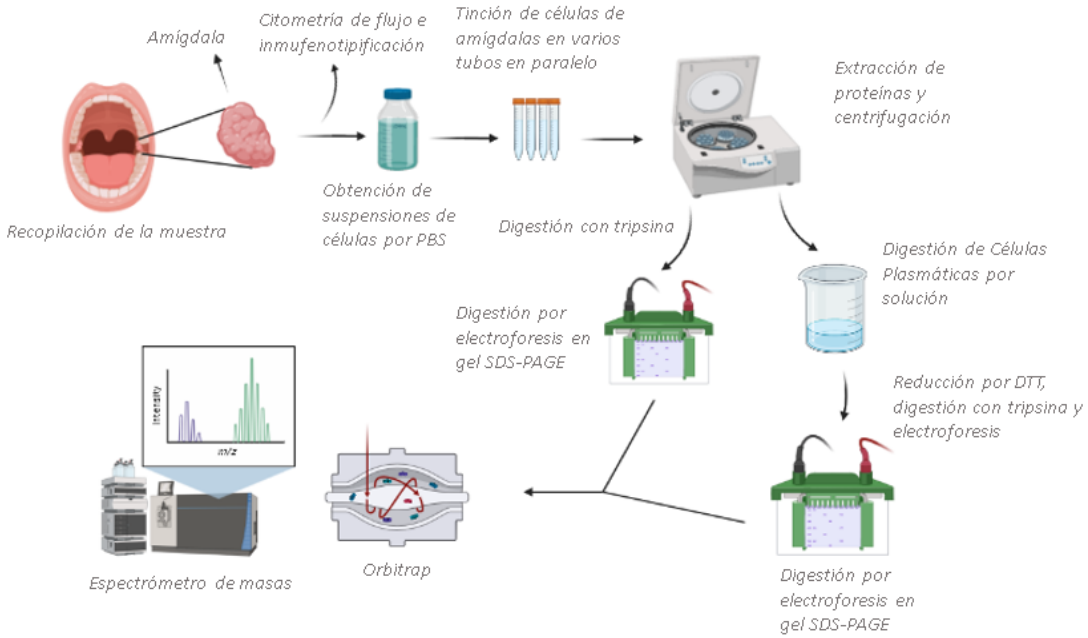
Dependent maturation begins with the manifestation of antigens to collaborating T cells and the production of cytokines. Therefore, this allows, through somatic recombination, a large number of antibodies in developing B cells. In order to increase the affinity of antibodies, strong-binding cell receptors are obtained through a maturation process as a result of somatic hypermutation of immunoglobulin genes in B cells. Subsequently, these cells leave the marginal center in the form of *plasma* or *memory cells*. *Plasma cells* secrete specific antibodies against their antigen. They can reside indefinitely in the bone marrow until they meet the antigen and respond very quickly and effectively. *Memory* cells circulate throughout the body in search of organisms that have affinity with their B cell receptor. They respond quickly and effectively as well.



Chronic lymphocytic leukemia (*CLL*) is a neoplasm that is differentiated by a continuous expansion of B lymphocytes in the peripheral blood, bone marrow, lymph nodes, and spleen. Its primary characteristic is the accumulation of B lymphocytes in the bone marrow and subsequent infiltration to the lymph nodes, spleen and liver, displacing normal cells. B cells are continuously activated through the acquisition of mutations that target monoclonal B cell lymphocytosis (*LBM*). The added accumulation of genetic abnormalities and the subsequent oncogenic evolution of monoclonal B cells produces *LLC*. It is the most common type of leukemia in the Western world.

For this study, two databases were obtained with intensity values for a large number of proteins. One database contains protein intensity values for B lymphocyte samples in their different stages (*naïve*, *centroblast*, *centrocyte*, *memory* and *plasma cell*) and the other database contains protein intensity values for *LLC* and *LBM* samples. the same process was used to reach both databases.

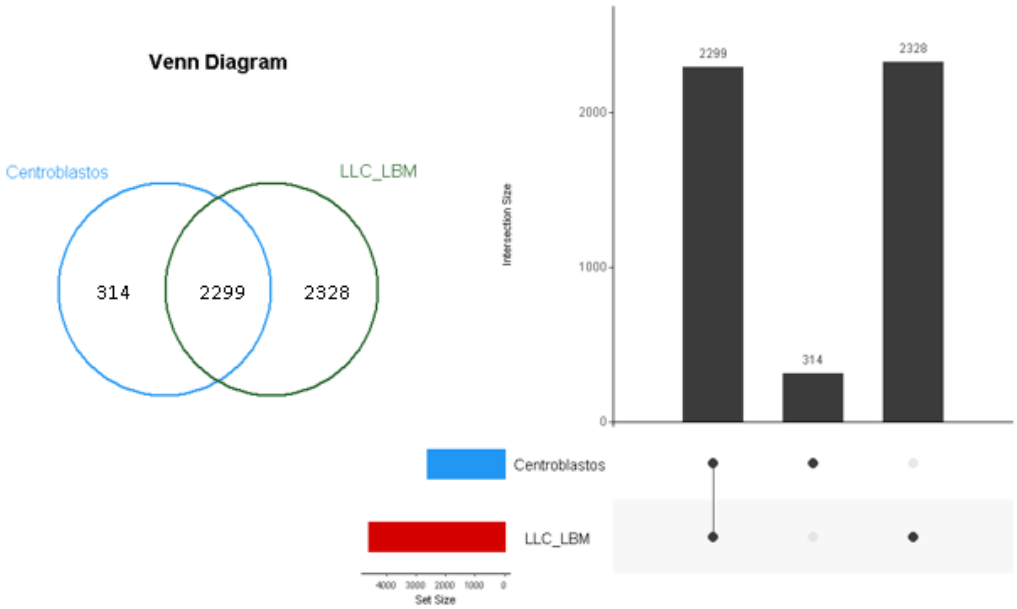
Human tonsils were collected from 5 donors after routine tonsillectomies. Cell suspensions of a single tonsil were obtained by mechanical disaggregation in PBS. After the purified cells are obtained, proteins are extracted. 15 micrograms of *naïve*, *centroblasts*, *centrocytes* and *memory* B-cell proteins were separated by electrophoresis. Mass spectrometry analysis was performed in a system coupled to an LTQ Orbitrap mass spectrometer, which operates by ion confinement. The ion source is nanoelectrospray for reverse phase LC-MS/MS analysis. Finally, through the Max-Quant software, the data obtained by mass spectrometry were analyzed, from which the intensities that appear in the final database were obtained.



Correlation diagrams and Venn diagrams are used for qualitative and quantitative data comparison. To measure the degree of correlation between two quantitative variables, there are different statistical coefficients, such as the Pearson coefficient or the Spearman coefficient. In this study, the Pearson correlation coefficient is used since large samples are available and it is more robust than that of Spearman which works better with smaller samples. The high correlation of 0.71 between *centroblasts* and *centrocytes* stands out,

which makes sense because, as explained in (Díez et al., 2021), although they do not have a similar appearance, their genetic expression is similar. Also highlights the correlation of 0.70 between *naïve* and *memory*, being the most logical biologically that the *naïve* cells are more similar to the *centroblasts* or *centrocytes* by their proximity in the marginal centers of the secondary lymphoid organs, A higher correlation between these two stages is peculiar. Regarding the correlation of all cell types, including *LLC* and *LBM*, for the 2512 common proteins, it is observed that, although the number of proteins is reduced, the correlation between *centroblasts-centrocytes* and *memory-naïve* remains high: 0.75 and 0.68 respectively. A very high correlation is observed, 0.88, between *LLC* and *LBM*, however, the correlation between these two cell types and the B lymphocyte stages is lower. The correlation between *LLC* and the rest of the stages varies between 0.3 and 0.4, and the correlation between *LBM* and the rest of the stages varies between 0.2 and 0.3.

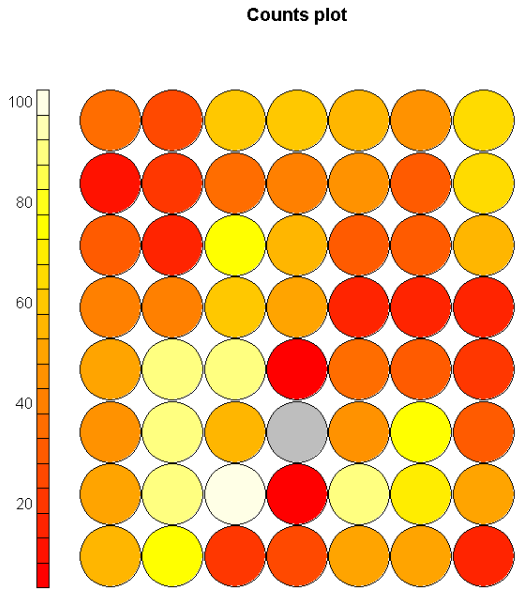
As for the Venn diagrams, it is observed that of the 2613 proteins with intensity values in *centroblasts*, 2299 (87.98%) proteins coincide with those present in *LLC-LBM*. For *centrocytes*, of the 2593 proteins with intensity values, 2279 (87.89%) are also found in *LLC-LBM* cells. For *memory* cells, of the 2512 proteins with intensity values, 2266 (90.21%) are found in *LLC-LBM* cells. As for the *naïve* cells, of the 2202, 1985 (90.14%) they are present in the *LLC-LBM*. Finally, of the 786 proteins with intensity values for *plasma cells*, 731 (93%) coincide with those present in *LLC-LBM* cells.



With regard to PCA, it is observed that *plasma cells* differ significantly from other stages of B lymphocyte (due to the small number of cells that can be obtained). There are similarities between the *naïve* and *memory* samples and between the *centroblasts* and *centrocytes* samples, so they confirm the results obtained when making the correlations. As for the PCA graph made to compare the samples of *LCM* and *LBM* with the stages of lymphocyte B, it can be observed three groups of differentiated samples: *LLC-LBM*, *plasma cells* and *centroblasts-centrocytes-naïve-memory*.

The SOM are based on unsupervised learning, therefore, after entering the data in the network, the data belonging to the same categories activate the same output neuron. Only one neuron is activated, the one suitable for the corresponding category. These categories are created by the network itself. SOM represents multidimensional data sets in a network with fewer dimensions, usually two-dimensional, so that those data that are similar or adjacent in multidimensional space are also similar in two-dimensional space.

Based on the *som* function, not only are proteins distributed in categories according to their characteristics, but it allows to make graphs of different styles: *counts*, *mapping*, *quality*… As a result, the dimensions 7x8 are the smallest sum of distances, therefore, they are chosen as the final dimensions of the SOM for the stages of lymphocyte B and the cells *LLC* and *LBM*. The *counts* chart shows the SOM map for the chosen dimensions.



A *rlen* equal to 300 is enough to visualize the number of iterations. With the *changes* graph you can see the number of minimum iterations as a function of the mean of the sum of the distances of the nodes to their contiguous nodes. With a *rlen* equal to 230 would be enough, since it is already the smallest average, therefore, 230 are the minimum number of iterations for this SOM.

As for the SOM containing only the stages of B lymphocyte, exactly the same process was carried out. The sum of the distances of each protein to the center of its node was observed, achieving as final dimensions 6x6 because they had the sum of smaller distances. For the *rlen* parameter, choose 300 as the starting number over the dimensions 6x6. The *changes* graph shows that the minimum number of iterations for SOM containing only B lymphocytes is 210. A dendrogram of the nodes is performed for a later grouping between nodes with similar characteristics. To do this, choose the *Euclidean distance* and the optimal clustering method for this data set. Clusters of nodes that are considered of biological and statistical interest are selected.

Functional enrichment analysis is a method that identifies classes of genes or proteins that are overrepresented in a large set of genes or proteins.

Description	*Gene Ratio*	*BgRatio*	*p.adjust*	*geneID*	*Count*
cell adhesion molecule binding	44/200	178/2493	3,3212E-10	ACTN4/P4HB/ATIC/CAPG/CCT8/LASP1/SND1 /DBNL/EZR/MSN/CALR/PTPN6/ MYH9/EHD1/TLN1/ALDOA/ENO1/ PFN1/HMGB1/HSPA5/HSPA8/LCP1/PPIA/ YWHAZ/PDLIM1/PTPN1/PCMT1/TMPO/ SYK/SEPTIN9/CLIC1/HSP90AB1/RPSA/EEF2/ PRDX6/TAGLN2/CAPZB/CAPZA1/HNRNPK/ YWHAE/RAN/RACK1/PRDX1/PARK7	44
protein-containing complex binding	63/200	339/2493	9,7049E-10	ACTN4/WDR1/ATP5PD/GNAI2/P4HB/XRCC5/H CLS1/UQCRC2/EIF4A3/CAPG/C1QBP/LASP1/SF 3B3/SND1/DBNL/ANXA6/EZR/CALR/MYH9/NA PA/TLN1/VIM/HMGB1/HSPD1/HSPA5/HSPA8/ LCP1/ACTB/PPIA/MTA2/TPR/PTPN1/SYK/NSF/ TPM4/SSRP1/SMC1A/NUMA1/CPSF6/SMC3/N PM1/TUBB/HSP90AA1/HNRNPC/HSP90AB1/R PSA/XRCC6/EEF2/H1-5/VDAC1/PPIB/CFL1/CORO1A/CAPZB/CAPZA1 /VCP/ACTR3/ACTR2/YWHAE/EIF5A/RACK1/HN RNPU/PARK7	63
actin binding	26/200	100/2493	2,4829E-06	ACTN4/WDR1/P4HB/HCLS1/LSP1/CAPG/LASP1 /DBNL/ANXA6/EZR/MSN/MYH9/TLN1/ALDOA/ PFN1/LCP1/PDLIM1/TPM4/EEF2/CFL1/CORO1 A/CAPZB/CAPZA1/ACTR3/ACTR2/HNRNPU	26
cadherin binding	34/200	158/2493	2,4829E-06	ATIC/CAPG/CCT8/LASP1/SND1/DBNL/EZR/MY H9/EHD1/TLN1/ALDOA/ENO1/PFN1/HSPA5/H SPA8/YWHAZ/PDLIM1/PTPN1/PCMT1/TMPO/ SEPTIN9/CLIC1/HSP90AB1/EEF2/PRDX6/TAGL N2/CAPZB/CAPZA1/HNRNPK/YWHAE/RAN/RA CK1/PRDX1/PARK7	34
actin filament binding	19/200	60/2493	5,1023E-06	ACTN4/WDR1/HCLS1/CAPG/LASP1/DBNL/ANX A6/EZR/MYH9/TLN1/LCP1/TPM4/EEF2/CFL1/C ORO1A/CAPZB/CAPZA1/ACTR3/ACTR2	19

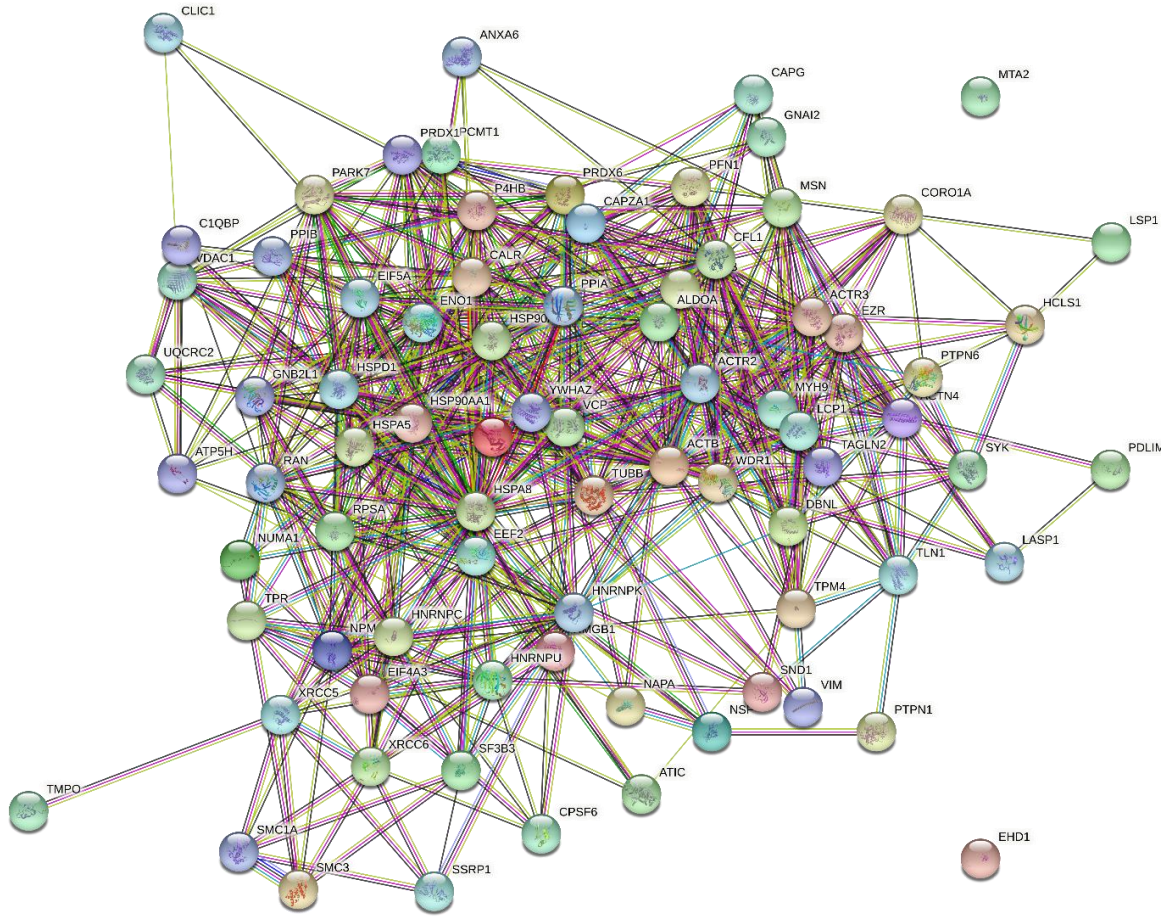
The p.adjust is the p-value adjusted by the Benjamini-Hochberg method, which is a multiple hypothesis test. The Benjamini-Hochberg method reduces the rate of false discovery, i.e., it helps prevent false positives, type I errors. The probability of rejecting an H_0 being true will be lower the greater the number of hypotheses being tested simultaneously.

Subsequently, the STRING tool is used, a database containing information from various sources such as experimental data and computational prediction methods. The edges represent significant and specific associations between proteins. Depending on the colour of the edge, the ratios are:

→ Known interactions: in blue are represented the interactions of selected databases and in purple are represented the interactions that have been experimentally determined.

→ Intended interactions: in green are represented interactions between neighboring genes, in red gene fusions and in dark blue gene co-occurrence.

→ Other interactions: Yellow interactions represent interaction between publications-based proteins, black coexpression, and dark blue protein homology.

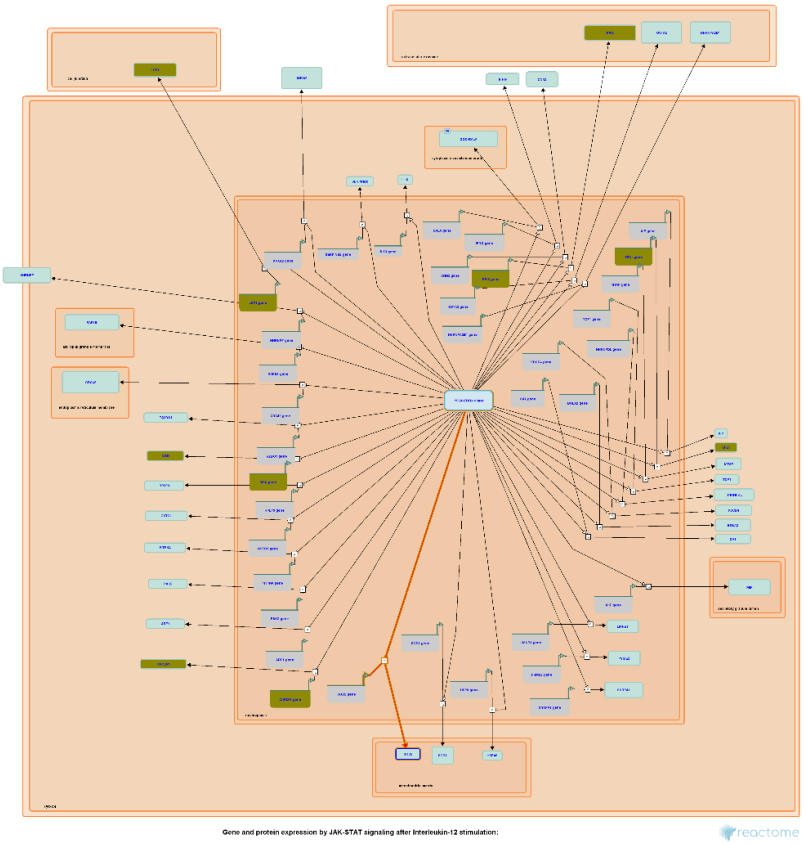


For example, *PPIA* is a protein that is found in the center of the network, catalyzing the cis-trans isomerization of lipid peptide bonds of proline in oligopeptides. This protein interacts with several other proteins close to it, for example *CFL1, ENO1, TAGLN2*, etc. According to the functional enrichment performed by the R-*enrichGO* function, *PPIA* performs the complex binding function containing proteins. *CLF1* is chosen observing that it has a combined-score of 0.914 on 1 of interaction. Their interactions come from co-expression, experimental data and publications. To expand the information obtained from STRING, signal paths through *Reactome* will be examined. *Reactome_Pathway* is a free, open source, curated library that provides bioinformatics tools for visualization, interpretation, and knowledge analysis in signaling pathways.

Pathway name	***Count***	***FDR***
Interleukin-12 signaling	11/84	2.19e-08
Gene and protein expression by JAK- STAT signaling after Interleukin-12 stimulation	11/96	3.87e-08
Interleukin-12 family signaling	10/73	3.87e-08
Signaling by Interleukins	22/643	1.51e-07
Immune System	44/2681	2.13e-06

Pathway name	Count	FDR
Platelet activation, signaling andaggregation	14/291	2.71e-06
Cytokine Signaling in Immune system	25/1092	1.53e-05
Neutrophil degranulation	16/480	2.80e-05
Signaling by Rho GTPases	19/709	4.78e-05
Chaperone Mediated Autophagy	5/23	4.79e-05
Platelet degranulation	9/139	4.79e-05
Signaling by Rho GTPases, Miro GTPases and RHOBTB3	19/725	4.79e-05
Hemostasis	20/801	4.79e-05
Response to elevated plateletcytosolic Ca2+	9/146	5.79e-05
Infectious disease	26/1348	9.62e-05
Regulation of actin dynamics for phagocytic cup formation	9/158	9.62e-05
Selective autophagy	7/89	1.59e-04
ATF6 (ATF6-alpha) activates chaperone genes	4/15	1.77e-04
RHOBTB GTPase Cycle	5/36	2.44e-04
ATF6 (ATF6-alpha) activates chaperones	4/17	2.58e-04
Signaling by ALK fusions and activated point mutants	6/66	2.72e-04
Signaling by ALK in cancer	6/66	2.72e-04
Fcgamma receptor (FCGR) dependent phagocytosis	9/193	3.27e-04
HSP90 chaperone cycle for steroidhormone receptors (SHR) in the presence of ligand	6/72	4.13e-04
Axon guidance	15/584	4.63e-04

According to Table 2, the signaling pathways with lower FDR are *interleukin-12 (IL-12) signaling* and *Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation* belonging to the *Immune System*, so its behavior in that route is studied.



The proteins indicated in green are those belonging to the database of this work. The gene for all proteins is found in nucleoplasm, but we only focus on those indicated. *CAPZA1, MSN* and *CFL1* are found in the cytosol, which is the aqueous component of a cell's cytoplasm, *LCP1* is found in cell junctions and, finally, *PPIA* is found in the extracellular exosome having 17 interactions with other proteins.

IL-12 is a heterodimeric cytokine that induces the production of interferon-g by natural killers and T lymphocytes. In addition, *IL-12* is related to the activation of human B cells through the *IL-12* receptor complex (*IL-12R*). The components of *IL-12R*, i.e., chains b1 and b2, are expressed in B-cells of *memory* tonsil, marginal center and human *naïve*. Transcripts of *IL-12* p35 and p40 were detected in all subsets, but only *memory* and *naïve* tonsil B cells produced *IL-12. IL-12R* is expressed in the main subsets of human B cells but is functional in *naïve* B cells.

CAPÍTULO 7: Anexos

Figura 25. Boxplot de la distribución de intensidad de cada muestra del linfocito B

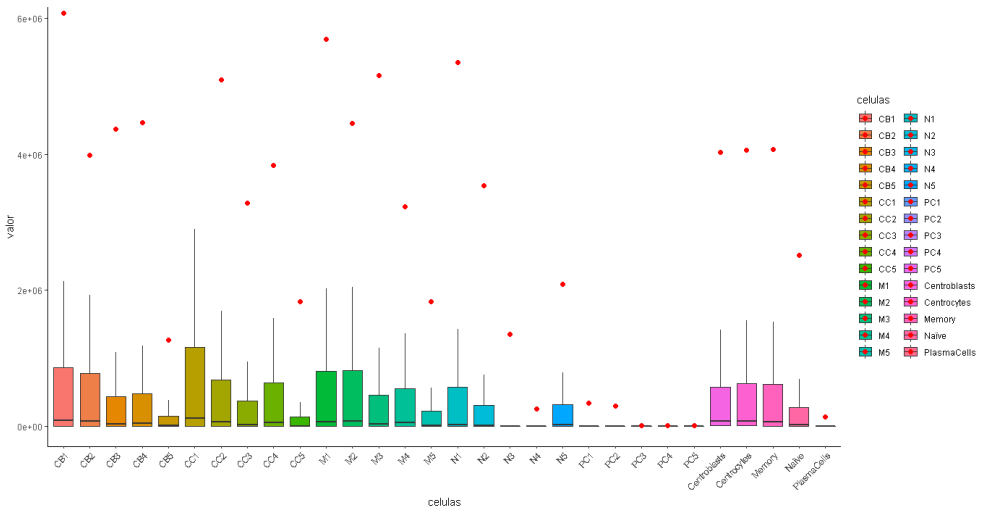


Figura 26. Boxplot de la distribución de la intensidad normalizada de cada muestra de linfocito B

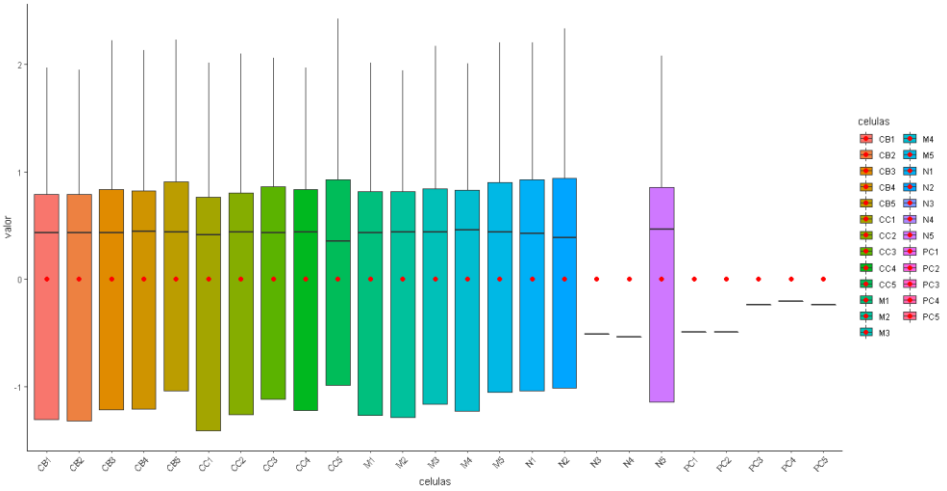


Figura 27. Matriz de correlación de los estadios del linfocito B, LLC y LBM

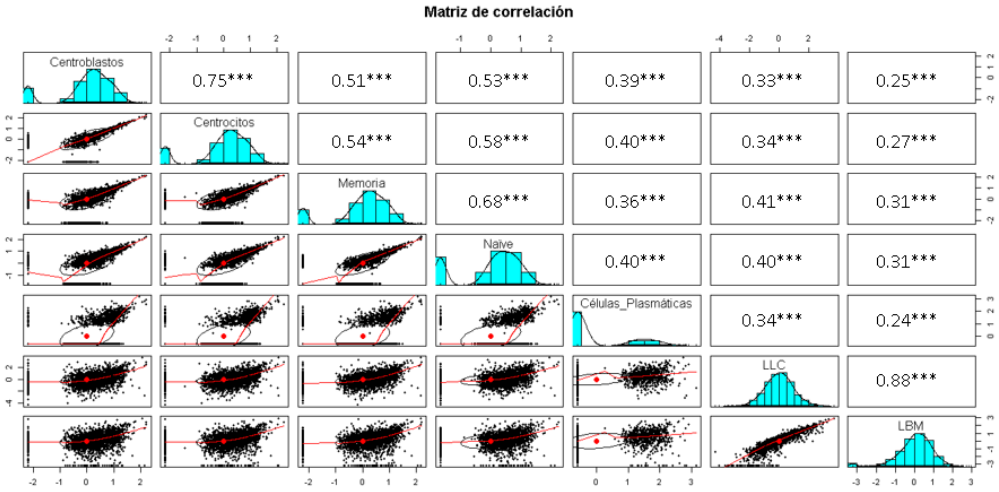


Figura 28. Gráfico de correlación de los estadios del linfocito B, LLC y LBM

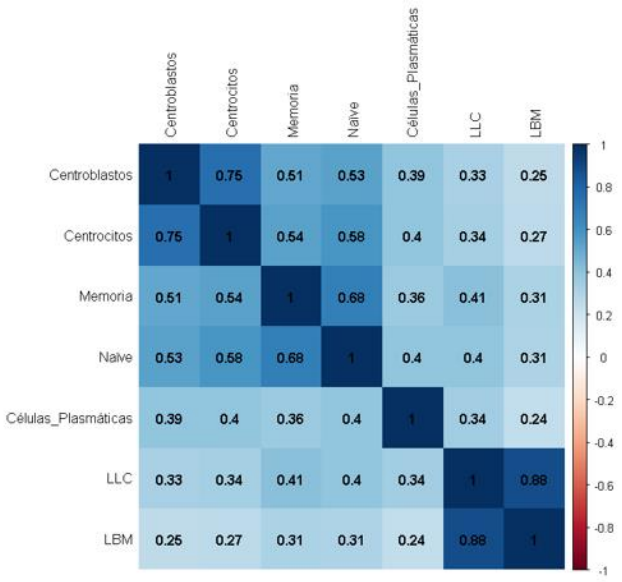


Figura 29. Diagrama de Venn de centrocitos y LLC-LBM

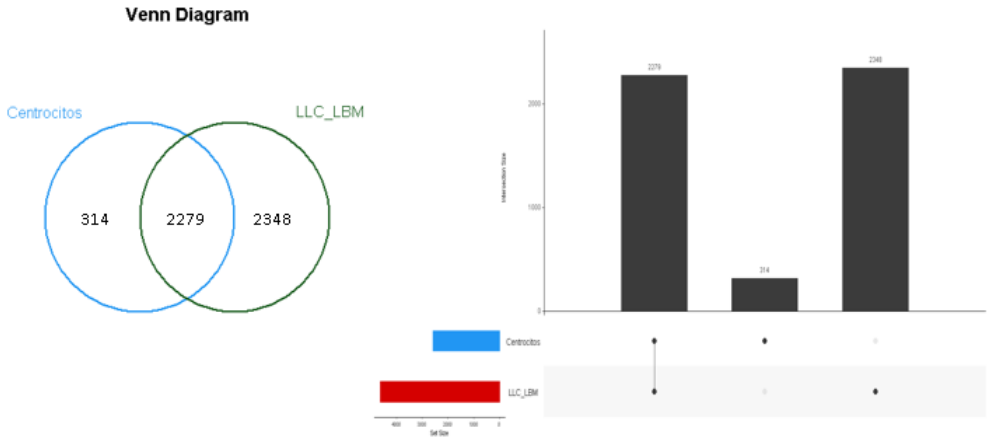


Figura 30. Diagrama de Venn de memoria y LLC-LBM

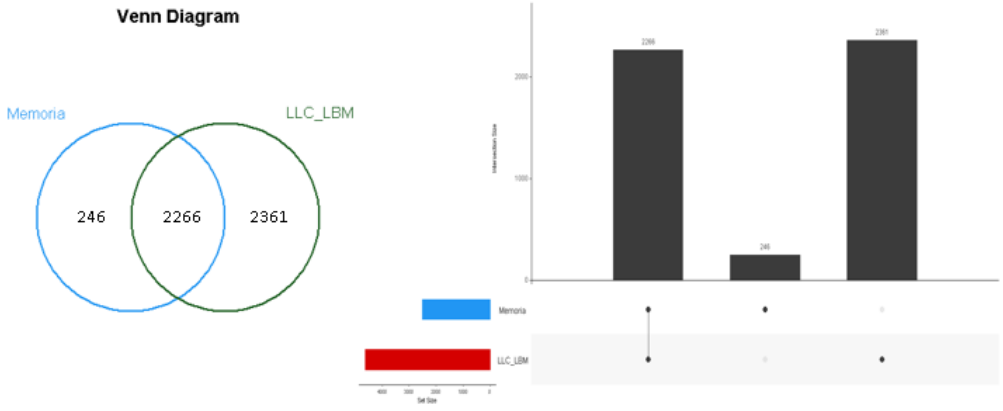


Figura 31. Diagrama de Venn de naïve y LLC-LBM

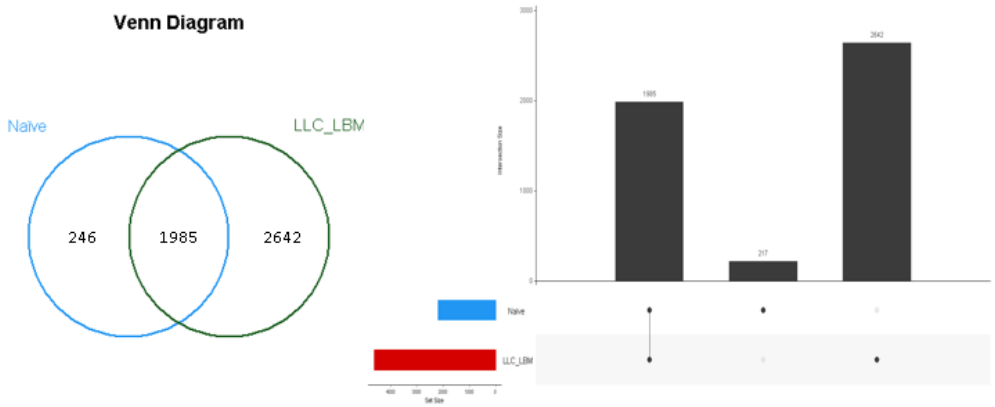


Figura 32. Diagrama de Venn de células plasmáticas y LLC-LBM

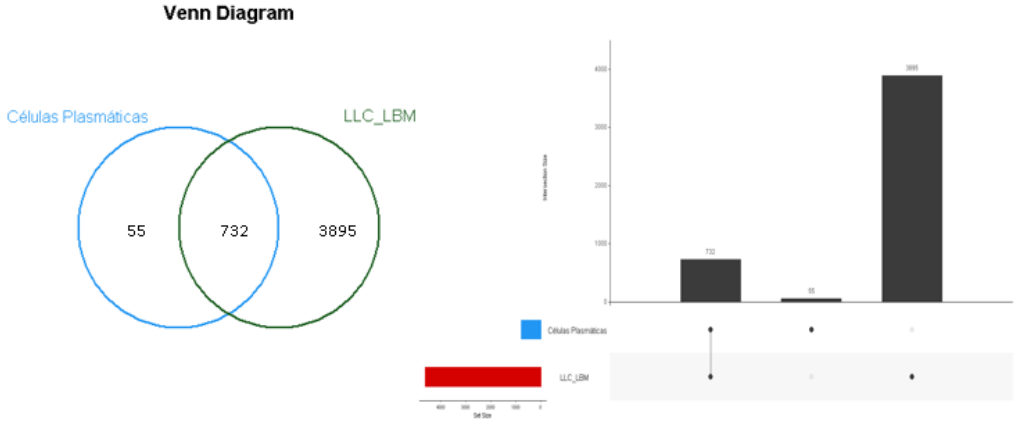


Figura 33. Gráfico PCA de todos los tipos celulares

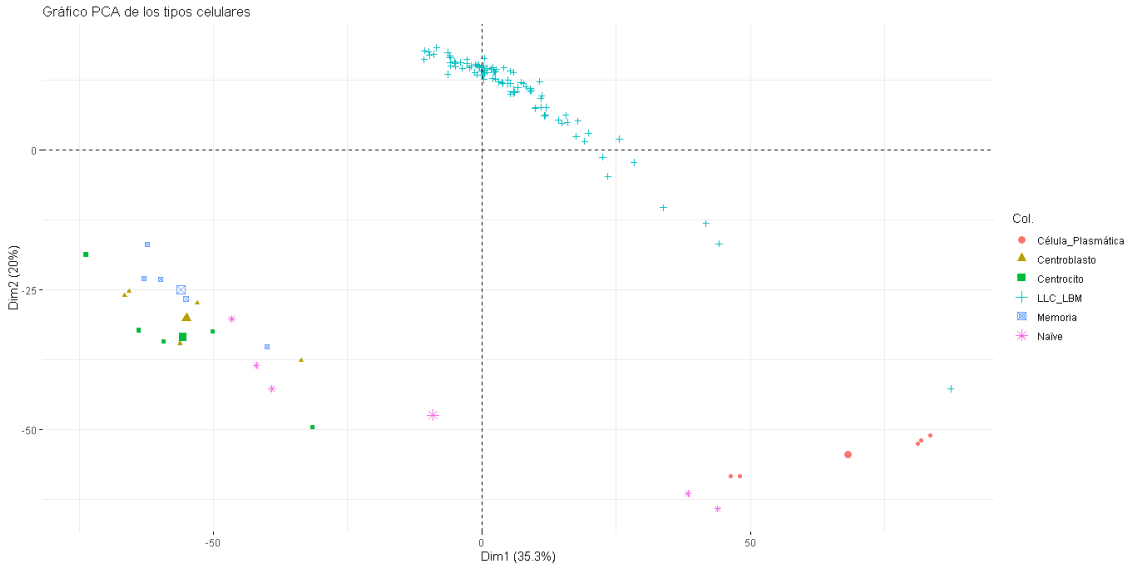


Figura 34. Gráfico counts para el mapa SOM final de dimensiones 6x6 para los estadios del linfocito B

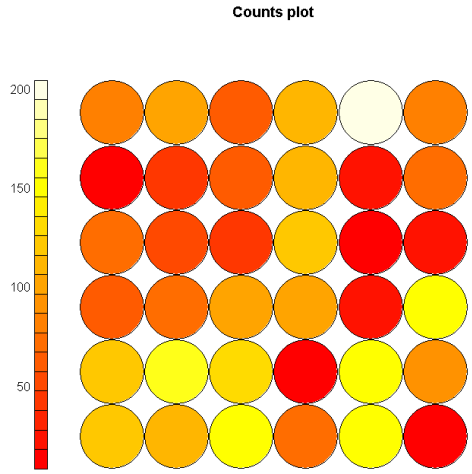


Figura 35. Gráfico changes para visualizar el número de iteraciones

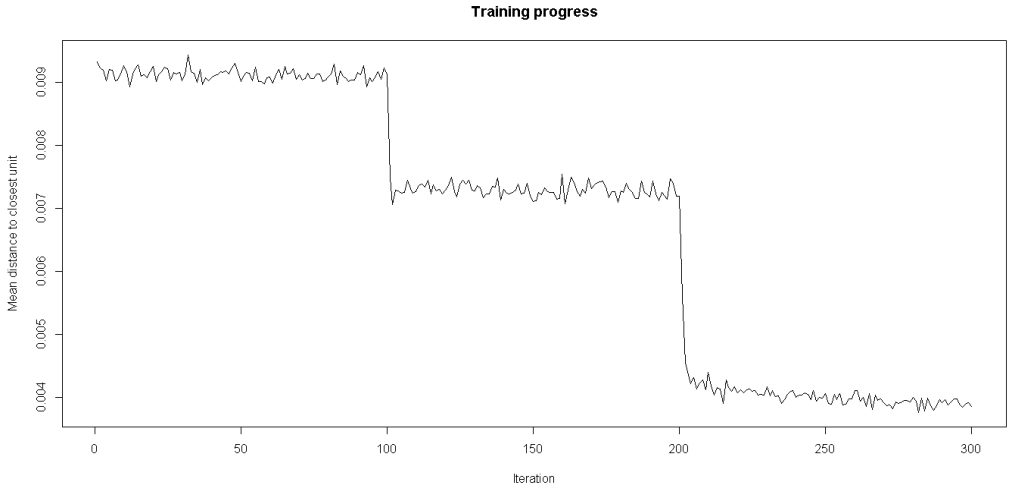


Figura 36. Gráfico quality para visualizar la calidad del SOM

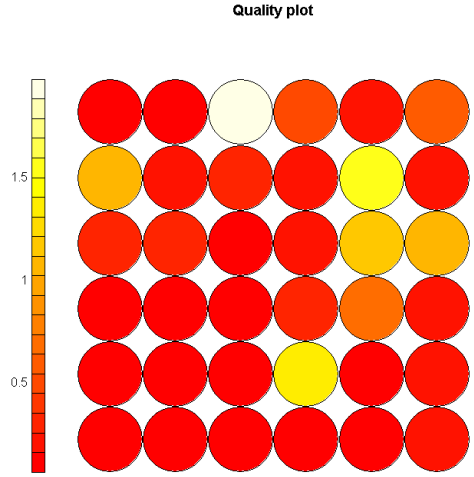


Figura 37. Gráfico mapping para visualizar la dispersión de las proteínas en los nodos

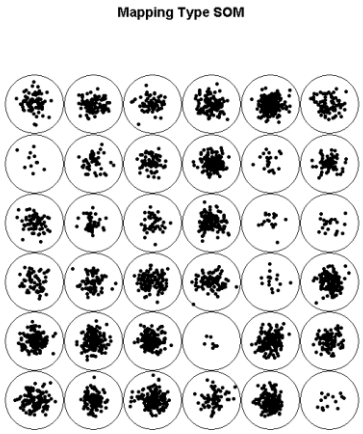


Figura 38. Gráfico codes para visualizar las características topológicas

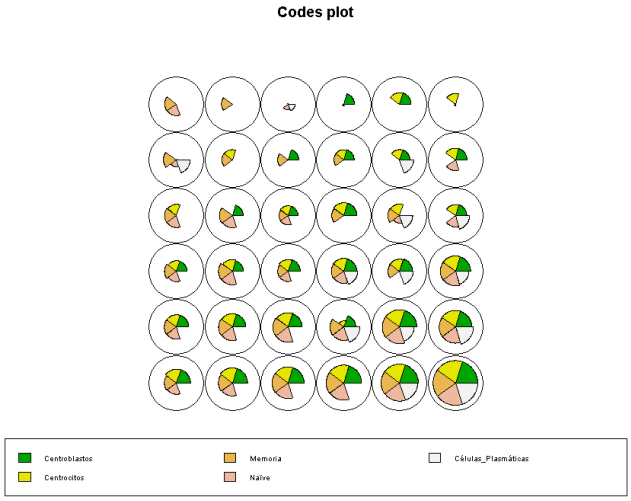


Figura 39. Dendrograma de los nodos pertenecientes al SOM