TRABAJO DE FIN DE GRADO

DISEÑO Y DESARROLLO DE ESTRATEGIAS DE INTEGRACIÓN MULTI-OMICA EN EL CONTEXTO DE LA EVALUACIÓN DE LA RESPUESTA INMUNE HUMORAL EN LEUCEMIA LINFÁTICA CRÓNICA Y SU ESTADIO PREVIO

GRADO EN ESTADÍSTICA

Facultad de Ciencias

Curso 2022/2023

Autora:

Laura Lezaun Pagola

Tutores:

José Manuel Sánchez Santos

Manuel Fuentes García

TRABAJO DE FIN DE GRADO

DISEÑO Y DESARROLLO DE ESTRATEGIAS DE INTEGRACIÓN MULTI-OMICA EN EL CONTEXTO DE LA EVALUACIÓN DE LA RESPUESTA INMUNE HUMORAL EN LEUCEMIA LINFÁTICA CRÓNICA Y SU ESTADIO PREVIO

GRADO EN ESTADÍSTICA

Facultad de Ciencias

Curso 2022/2023

Autora: Laura Lezaun Pagola

Firmado por LEZAUN PAGOLA LAURA - ***7006** el día 03/07/2023 con un certificado emitido por AC FNMT Usuarios

Tutores:

José Manuel Sánchez Santos

Manuel Fuentes García

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

Facvltad D Ciencias
VNiVERSiDAD D SALAMANCA

Certificado del/los tutor/es TFG

D. José Manuel Sánchez Santos, profesor del Departamento de Estadística de la Universidad de Salamanca y D. Manuel Fuentes García del departamento de Medicina e investigador del Centro de Investigación del Cáncer (CiC-IBMCC, USAL/CSIC),

HACE/N CONSTAR:

Que el trabajo titulado "*Diseño y desarrollo de estrategias de integración multi-ómica en el contexto de la evaluación de la respuesta inmune humoral en leucemia linfática crónica y su estadio previo*", que se presenta, ha sido realizado por Laura Lezaun Pagola, con DNI 21070065H, y constituye la memoria del trabajo realizado para la superación de la asignatura Trabajo de Fin de Grado en Estadística en esta Universidad.

Salamanca, 3 de julio de 2023

Firmado digitalmente por SANCHEZ SANTOS JOSE MANUEL - 07870414K
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-07870414K, givenName=JOSE MANUEL, sn=SANCHEZ SANTOS, cn=SANCHEZ SANTOS JOSE MANUEL - 07870414K
Fecha: 2023.07.04 17:39:24 +02'00'

Fdo.: José Manuel Sánchez Santos

FUENTES GARCIA MANUEL - 07974069S
Firmado digitalmente por FUENTES GARCIA MANUEL - 07974069S
Fecha: 2023.07.04 14:39:04 +02'00'

Fdo.: Manuel Fuentes García

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1 INTRODUCCIÓN

La leucemia linfática crónica o leucemia linfocítica crónica (LLC) es el tipo más frecuente de leucemia en adultos en los países occidentales. A día de hoy, se considera una patología incurable, debido a ser una enfermedad muy compleja y muy heterogénea desde el punto de vista genómico y con un pronóstico muy diferente, dado que hay individuos con cursos indolentes y progresión muy lenta; como hay individuos con una progresión extremadamente rápida. Actualmente, existen algunos parámetros que se emplean para el pronóstico; pero, sin embargo, no son del todo eficientes y se tendría que seguir investigando cómo realizar una estratificación correcta de los pacientes tanto de cara a su progresión como en la terapia más eficaz.

Por estos motivos, se propone realizar una caracterización proteómica, con el fin de identificar perfiles diferenciales de expresión proteica entre células LLC y células normales; tanto para identificar mecanismos biológicos de evolución, pronóstico, nuevas dianas terapéuticas y/o desarrollo de nuevos tratamientos.

Desde el punto de vista genómico, existen descritas dos características altamente relevantes a tener en cuenta para el pronóstico de esta enfermedad son el estado mutacional de la región variable de la cadena pesada de inmunoglobulina (IGHV) y las alteraciones cromosómicas, en especial, las deleciones 17p, 13q y 11q, y la trisomía 12 (Delgado et al., 2017). Teniendo en cuenta estudios previos, se ha demostrado que la presencia de la deleción 13q y de mutaciones en el IGHV (M-IGHV) se corresponde con un buen pronóstico, mientras que la presencia de las deleciones 17p y 11q y la ausencia de mutaciones en el IGHV (U-IGHV) está relacionado con un pronóstico desfavorable (Landeira-Viñuela et al., 2022). Sin embargo, la trisomía 12 se considera de pronóstico intermedio ya que puede variar dependiendo de otros factores genéticos. (Bagacean et al., 2017)

También se debe tener presente que existe un estadio previo y asintomático a la LLC denominado Linfocitosis Monoclonal de células B (LMB). Se trata de un tipo de afección caracterizada por la presencia de un número elevado, aunque menor de 5000 células/µl, de linfocitos B en sangre periférica (Mowery & Lanasa, 2012). No todos los pacientes que presentan MBL terminan progresando a LLC, pero se considera un factor de riesgo para su desarrollo y se recomienda hacer un seguimiento de dichos pacientes para detectar cualquier indicio de evolución.

En este contexto, el empleo técnicas estadísticas y herramientas computacionales puede resultar esencial en el estudio de la LLC, ya que permiten explorar y analizar grandes conjuntos de datos. Estas técnicas simplifican la búsqueda de patrones y asociaciones en los datos, lo que puede llevar a la identificación de biomarcadores o de nuevos objetivos terapéuticos para esta enfermedad. De esta manera, el análisis sistemático y masivo de datos puede llegar a mejorar la comprensión biológica de la LLC y, por lo tanto, ayudar a encontrar tanto biomarcadores relacionados con el diagnóstico y pronóstico de la enfermedad como tratamientos más efectivos para los pacientes.

2 FUNDAMENTOS BIOLÓGICOS

2.1 RESPUESTA INMUNE

En el pasado, el término inmunidad hacía referencia a la protección contra enfermedades, especialmente enfermedades infecciosas. El sistema inmunológico está compuesto por células y moléculas que son responsables de la inmunidad, y su respuesta conjunta y coordinada ante la presencia de sustancias extrañas se denomina respuesta inmune (Abbas et al., 2017). Hoy en día, se conoce también como función del sistema inmune, el reconocimiento y protección de lo propio; en general, denominado tolerancia inmune, que juega un papel muy relevante en procesos tumorales, auto-inmunidad y en procesos como trasplantes.

De forma general, existen dos tipos principales de respuesta inmune: A.- Respuesta inmune innata, que es aquella que actúa como primera línea de defensa del organismo frente a señales de peligro de forma rápida y no específica, (Sun et al., 2020). B.- Respuesta inmune adaptativa es un proceso complejo que se desencadena en el organismo en respuesta a la introducción de antígenos extraños, como bacterias, virus o células tumorales. Este tipo de respuesta se caracteriza por su especificidad y memoria inmunológica, lo que significa que una vez que el sistema inmunológico ha encontrado un antígeno, puede recordarlo y producir una respuesta más rápida y efectiva si se expone nuevamente a ese mismo antígeno en el futuro (Abbas et al., 2017).

Los linfocitos B y T son las células características de la respuesta inmune adaptativa. Los linfocitos B producen anticuerpos que pueden unirse a los antígenos y así facilitar su eliminación, mientras que los linfocitos T pueden acabar directamente con las células infectadas o ayudar a los linfocitos B a producir anticuerpos específicos (Punt et al., 2018).

La respuesta inmune adaptativa puede ser dividida en dos tipos dependiendo del linfocito principal involucrado y del tipo de microorganismo que se desea eliminar: la respuesta inmune humoral, llevada a cabo principalmente por los linfocitos B y la respuesta inmune celular, llevada a cabo principalmente por linfocitos T. (Abbas et al., 2017)

2.2 LINFOCITOS B

Los linfocitos B son un tipo de células del sistema inmunitario que se localizan en la sangre periférica y en las regiones foliculares de los ganglios linfáticos, y son los encargados de proporcionar protección contra diferentes microorganismos. Son los responsables de la respuesta humoral. (Abbas et al., 2017)

Estas células B son capaces de reconocer antígenos, que son “moléculas extrañas” que el sistema inmunológico reconoce como no propias, a través de una molécula de inmunoglobulina denominada BCR (“B Cell Receptor”) que se encuentra en su membrana.

Cuando un antígeno se une al BCR de una célula B específica, se desencadena un conjunto de señales intracelulares que provocan que dicha célula se active. Una vez activada, comienza a dividirse y se diferencia en células plasmáticas, las cuales tienen la capacidad de producir y liberar una gran cantidad de anticuerpos específicos que neutralizan o marcan el antígeno causante de la respuesta para que sea eliminado por otras células del sistema inmunológico. (Jr et al., 2001)

Además, los linfocitos B también interactúan en el microambiente tumoral con células inmunológicas (células T y células dendríticas, entre otras) y con células no inmunológicas (células tumorales y células del estroma, entre otras). Estas interacciones pueden ser tanto beneficiosas como perjudiciales en cuanto a la respuesta inmunitaria contra el cáncer. Por una parte, los linfocitos B pueden contribuir a la supresión del crecimiento tumoral mediante la secreción de anticuerpos específicos contra el tumor; pero, por otro lado, también pueden generar anticuerpos que favorezcan al desarrollo de tumores. (Yuen et al., 2016)

2.3 LEUCEMIA LINFÁTICA CRÓNICA (LLC)

La leucemia linfática crónica (LLC) es un tipo de cáncer de la hematológico que se origina debido una proliferación excesiva de linfocitos B que se acumulan en la sangre periférica, médula ósea y ganglios linfáticos. Se caracteriza por tener una gran heterogeneidad durante todo el desarrollo de la enfermedad. (Herbst et al., 2022)

Se trata de la leucemia más común en la población adulta procedente de países occidentales, con una prevalencia de 5 casos por cada 100.000 habitantes. En el momento que se produce el diagnóstico, al menos el 70% de los pacientes superan los 65 años y la media de edad se establece en 70 años. (Briones, 2022)

En algunas ocasiones, los pacientes pueden permanecer asintomáticos, sin requerir ninguna clase de tratamiento y llegar a sobrevivir décadas; mientras que hay otros que fallecen a los pocos años de ser diagnosticados. Esta variabilidad se corresponde con factores intrínsecos y extrínsecos a la célula B leucémica. (Chiorazzi et al., 2021)

Entre los factores intrínsecos podemos encontrar los cambios en la genética y epigenética tanto en genes codificantes (que producen proteínas) como en genes no codificantes (no codifican proteínas). Por otra parte, los factores extrínsecos, que son los que se originan fuera de la célula, incluyen el microambiente tumoral o factores de crecimiento entre otros.

También se considera una enfermedad heterogénea desde el punto de vista biológico. En este punto encontramos las alteraciones cromosómicas (deleciones 13q,17p y 11q y trisomía 12), mutaciones en los genes que codifican la región variable de la cadena pesada de inmunoglobulina (IGHV), variaciones en un solo nucleótido de ciertos genes y alteraciones en los niveles de expresión de algunas proteínas y/o mRNA. (Landeira-Viñuela et al., 2023)

Los biomarcadores son variables biológicas que proporcionan información sobre enfermedades específicas. En el caso de la LLC, los biomarcadores más relevantes son la edad, las alteraciones citogenéticas (del17p, del11q, del13q o trisomía 12), el estado mutacional de la región variable de la cadena pesada de inmunoglobulina (Mutado / No mutado) y las mutaciones genéticas (MYD88, NOTCH1, TP5, SF3B1 y ATM). En la Tabla 1 podemos comprobar si dichos biomarcadores se corresponden con un buen o mal pronóstico. (Landeira-Viñuela et al., 2022)

Tabla 1

Biomarcadores de valor pronóstico

Biomarcador	Buen pronóstico	Mal pronóstico
Edad	< 65 años	> 65 años
Alteraciones cromosómicas	del(13q)	del(17p), del(11q)
Estado mutacional IGHV	IGHV mutado	IGHV no mutado
Mutaciones genéticas	MYD88	NOTCH1, TP5, SF3B1 y ATM

Como se puede comprobar en la Tabla 1, la deleción 17p y la deleción 11q son indicadores de un mal pronóstico, por lo que las vamos a denominar citogenéticas de alto riesgo (cariotipo complejo). Si se tienen en cuenta únicamente el estado mutacional del IGHV y la ausencia o presencia de estas citogenéticas de alto riesgo, los pacientes se pueden separar en tres grupos: Riesgo bajo (IGHV mutado sin deleción 17p o 11q), riesgo intermedio (o bien IGHV no mutado o bien deleción 17p o 11q) y riesgo alto (IGHV no mutado con deleción 11p o 17q) (Tabla 2) (Delgado et al., 2017).

Tabla 2

Clasificación de pacientes según el estado mutacional del IGHV y la presencia de citogenéticas de alto riesgo.

Estado Mutacional del IGHV	Alteraciones citogenéticas de alto riesgo	Grupo de riesgo asociado
Mutado	No	Bajo
Mutado	Sí	Intermedio
No Mutado	No	Intermedio
No Mutado	Sí	Alto

2.3.1 LINFOCITOSIS MONOCLONAL DE CÉLULAS B (LMB)

Es posible hallar linfocitos B con un fenotipo de LLC en la sangre de personas sanas años antes de que se les diagnostique dicha enfermedad. (Georgiadis et al., 2017) Esto se conoce con el nombre de linfocitosis monoclonal de células B o LMB y se considera como el estadio previo a la LLC.

Se estima que alrededor del 5% de la población mayor de 40 años la padece y este porcentaje aumenta hasta el 15% en aquellos que superan los 70. (Chiorazzi et al., 2021) De la misma manera que sucede con la LLC, la prevalencia de la enfermedad es superior en hombres que en mujeres. (Galigalidou et al., 2021)

Se puede considerar LMB de recuento alto cuando se tienen 0.5-4.99× 10^9 células B clonales/L y se calcula que únicamente progresan a LLC entre el 1 y el 4% de individuos con esta característica al año. (Galigalidou et al., 2021)

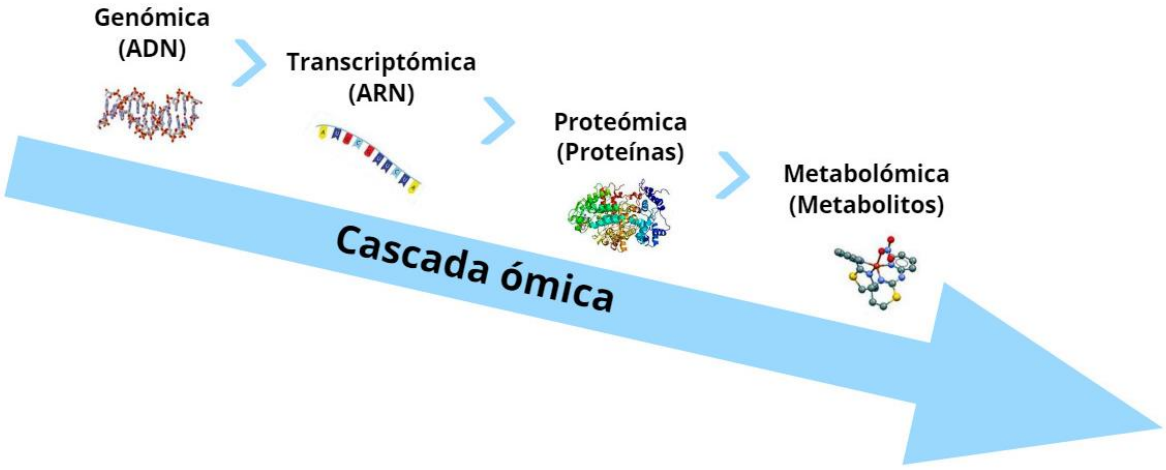
2.4 CIENCIAS ÓMICAS

El concepto de "ómica" se utiliza para hacer referencia a las ciencias que nos permiten investigar un número muy elevado de moléculas que están involucradas en el funcionamiento de un ser vivo. Entre estas moléculas se encuentran el ADN, el ARN, las proteínas y los metabolitos, que son pequeños compuestos químicos. Existen cuatro tipos de ómicas que dependen de la molécula que se esté analizando y de la información biológica que se desee obtener: genómica (ADN), transcriptómica (ARN), proteómica (proteínas) y metabolómica (metabolitos). (Tortosa Viqueira et al., 2017)

El estudio integrado de estos cuatro tipos de datos se denomina cascada ómica (Figura 1) y se basa en la idea de que todas las moléculas biológicas de un organismo están interconectadas y se afectan mutuamente. Por ello, el análisis de múltiples capas de información biológica nos va a permitir obtener una comprensión más completa de los procesos biológicos.

Figura 1

Cascada ómica



La cascada ómica comienza analizando el genoma de un organismo, es decir su ADN. A partir de este punto se pueden realizar análisis transcriptómicos para medir la expresión de dichos genes, proteómicos para estudiar las proteínas producidas por los genes y metabolómicos para medir los metabolitos.

Todos estos datos se integran y se analizan utilizando técnicas bioinformáticas para identificar patrones y relaciones entre los distintos niveles de información. A su vez, la biología de sistemas utiliza esta información para comprender cómo están regulados los procesos biológicos y cómo se relacionan con el fenotipo.

De todas estas técnicas, se considera que la proteómica puede llegar a ser una herramienta muy prometedora en la identificación de biomarcadores, ya que las proteínas son más susceptibles a sufrir cambios de manera generalizada tanto en la enfermedad como en la respuesta del organismo a ella. (Horgan & Kenny, 2011)

En los pacientes con LLC, hay múltiples eventos moleculares que dificultan la selección de la terapia. Para poder tratar a dichos pacientes de manera más efectiva, es necesario comprender cómo los eventos moleculares internos y externos afectan a la biología celular de cada individuo en particular. Se plantea la hipótesis de que realizar un análisis proteómico puede aportar esa información faltante, ya que el resultado final de todas las influencias epigenéticas, genéticas y ambientales en la célula ocurre a nivel de proteína. (Griffen et al., 2022)

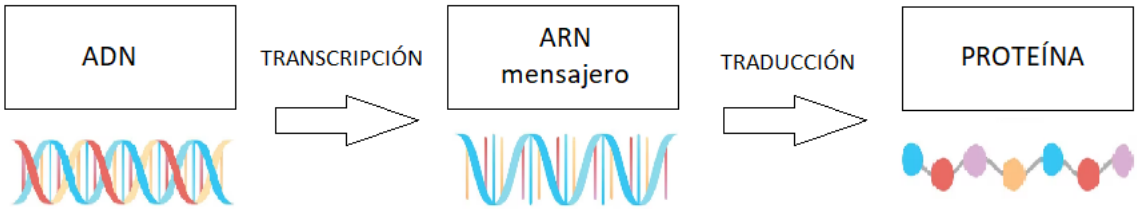
2.4.1 PROTEÓMICA

Para comprender mejor qué es la proteómica, es importante abordar previamente el concepto de genómica. La genómica es la rama de la biología que se dedica al estudio de la totalidad del ADN de un ser vivo, es decir, su genoma. (National Human Genome Research Institute, 2023)

Esta disciplina se basa en el dogma central de la biología molecular (Figura 2), que establece que la información genética que se encuentra en el ADN se transcribe en forma de ARN para posteriormente traducirse en proteínas. (Lesher Gordillo & Tovilla Zárate, 2013)

Figura 2

Dogma central de la biología molecular



La proteómica se considera una continuación natural de la genómica, ya que las proteínas son fundamentales para llevar a cabo la mayoría de los procesos biológicos. (López Arias, 2013) Podemos definir la proteómica como la disciplina científica que se centra en el estudio y análisis de los proteomas, es decir, el conjunto de proteínas que son expresadas por un genoma, célula o tejido. (Centro de Apoyo a la Investigación (CAI), 2018)

Entre las técnicas más frecuentes para estudiar la proteómica se encuentran la cromatografía liquida acoplada a espectrometría de masas en tándem y los microarrays de proteínas (Figura 3):

- Cromatografía líquida acoplada a espectrometría de masas en tándem (LC-MS/MS): Es una técnica utilizada para cuantificar proteínas en una muestra biológica. Se basa en la separación de las proteínas, mediante cromatografía liquida (LC) y su posterior análisis

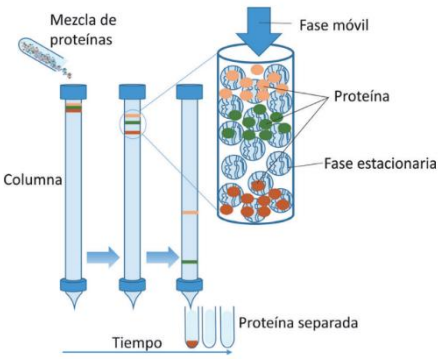
mediante espectrometría de masas (MS). Los objetivos de este método son la identificación de las proteínas presentes en la muestra y la cuantificación de sus niveles de abundancia. (Centro de Apoyo a la Investigación (CAI), s. f.)

  - La cromatografía líquida o LC por sus siglas en inglés, es una estrategia de fraccionamiento utilizada para separar proteínas o péptidos dependiendo de sus propiedades físicas o químicas. (De la Torre Gómez, 2012)
  - La espectrometría de masas (MS) es una técnica analítica que puede proporcionar información acerca de la masa molecular y estructura de un compuesto, o bien detectar su presencia y/o medir su concentración. Se lleva a cabo en un instrumento denominado espectrómetro de masas, el cual está compuesto por tres elementos: fuente de ionización, analizador y detector. (Carrera Aguado, s. f.)

- Microarrays de proteínas: Son una técnica que consiste en la inmovilización de una gran cantidad de proteínas, dispuestas de forma ordenada y específica, sobre una superficie sólida. La creación de los microarrays de proteínas supone un avance significativo en la automatización y miniaturización del estudio del proteoma, al igual que en la disminución de los costes debido a la reducción del consumo de reactivos. (San Miguel-Hernández et al., 2009)

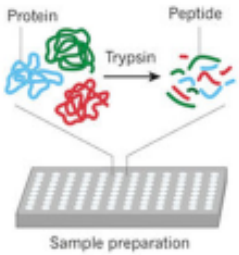
Figura 3

Técnicas para estudiar la proteómica



Cromatografía líquida

Espectrómetro de masas



Microarray de proteínas

En el estudio que vamos a realizar se busca determinar y estudiar una gran cantidad de proteínas de cada una de las muestras, por lo que se va a requerir la aplicación de un análisis multidimensional. De esta manera, se pretende integrar los datos cuantitativos referentes a las proteínas junto con las variables clínico-biológicas para obtener una visión completa de la información.

3 OBJETIVOS

En el actual estudio se trabaja con cuatro tipos de datos dentro de la proteómica: autoanticuerpos (aAbs), antígenos microbianos (mAgs), puntos de control inmunitario (immune checkpoints) y proteoma, obtenidos de individuos con leucemia linfática crónica (LLC) y Linfocitosis Monoclonal de células B (LMB). En el caso de autoanticuerpos y antígenos microbianos se han tomado datos tanto de IgG como de IgM.

IgG (inmunoglobulina G) e IgM (inmunoglobulina M) son proteínas que genera el organismo con la finalidad de protegerlo contra patógenos y sustancias tóxicas que éstos producen. En primer lugar, se produce la IgM como respuesta inicial ante la presencia de un agente infeccioso, considerándose un marcador de fase aguda. Ésta activa el sistema complementario, un conjunto de proteínas que detectan los patógenos y ayudan a eliminarlos. Posteriormente, se genera la IgG proporcionando una respuesta inmunitaria más específica y de mayor duración a largo plazo contra el patógeno.

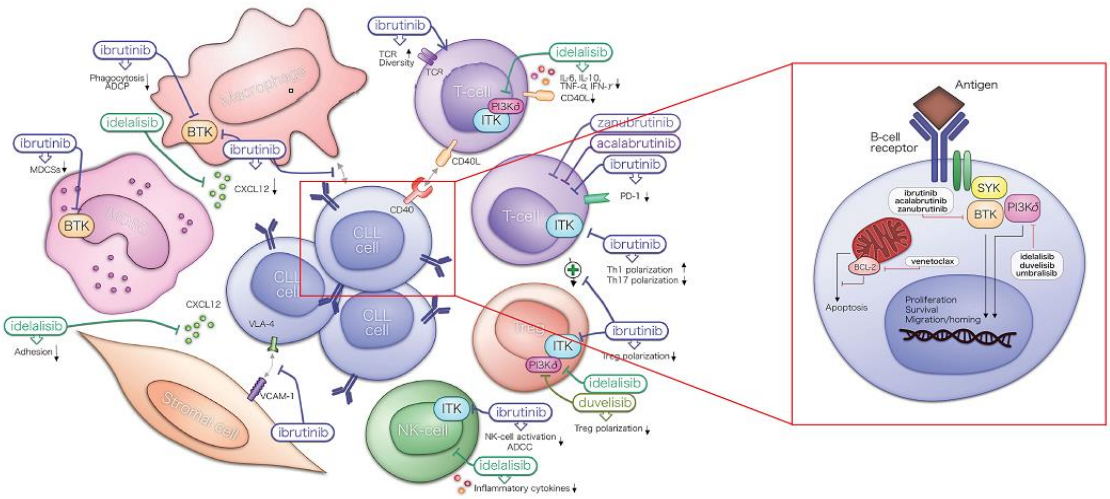
En la Figura 4 se observa como las células de LLC interaccionan y se comunican con otras células en el microambiente tumoral mediante una serie de proteínas denominadas receptores de membrana (CD40, VLA-4) o mediante quimiocinas (CXCL12, IL-6, IL-10, TNF-α, IFN-γ). Cuando se produce una respuesta inmunológica, los immuno checkpoints (PD-1, BCL-2) son los encargados de regular esa respuesta mediante inhibidores como el ibrutinib o venetoclax, respectivamente.

Por otro lado, en la ampliación del linfocito B, se representa como a través del BCR se reconocen los antígenos y esto provoca que se activen proteínas como BTK o PI3Kδ, las cuales desempeñan un papel fundamental en la proliferación y supervivencia de los linfocitos B. En algunas ocasiones, la activación del BCR también conlleva a la producción de autoanticuerpos.

De este modo, se pretende visualizar cómo por medio de los diferentes conjuntos de datos empleados en este estudio se están estudiando diversas partes de los linfocitos B en el ámbito de Leucemia Linfocítica Crónica.

Figura 4

Comunicación de los linfocitos B con otras células



Por lo tanto, el objetivo general de este trabajo es integrar todos los datos de los que se dispone en un único análisis y, de esta manera, poder llegar a formar nuevos grupos de pacientes que no se habían contemplado anteriormente. Por otro lado, también se busca identificar biomarcadores (autoanticuerpos, antígenos microbianos, puntos de control inmunitario, proteínas) que puedan estar relacionados con la Leucemia Linfocítica Crónica para comprender mejor dicha enfermedad y tenerlos en cuenta a la hora de desarrollar nuevos tratamientos.

4 MATERIAL Y MÉTODOS

4.1 MATERIAL

Todos los datos que se han utilizado para realizar este estudio han sido proporcionados y anonimizados por el laboratorio 11 del CIC, pertenecientes al proyecto de investigación activo FIS PI21/01545 financiado por el ISCIII siendo investigador Principal Manuel Fuentes; realizado en el Centro de Investigación del Cáncer y pacientes reclutados en el Hospital Universitario de Salamanca (consentimiento informado y aprobación del CEIm).

4.1.1 DESCRIPCIÓN DE LA BASE DE DATOS

Se dispone de una cohorte de 67 pacientes de LLC de los cuales tenemos información sobre 9 variables clínico-biológicas que son: *Gender, Age, Disease stage, Disease evolution, Dx according to therapy response, Treatment, IGHV mutational status, Karyotype* y *Chromosomal aberrations*. Todas ellas son cualitativas, excepto la variable edad (*Age*). En la siguiente tabla (Tabla 3) se describen dichas variables con sus correspondientes categorías.

Tabla 3

Descripción de las variables cualitativas

Variable	Descripción	Categorías
Gender	Género del paciente	Female: Mujer Male: Hombre
Disease stage	Estadio de la enfermedad	CLL: Leucemia Linfática Crónica MBL: Linfocitosis Monoclonal de células B
Disease evolution	Evolución de la enfermedad	c-CLL: Enfermedad constante p-CLL: Progresión de la enfermedad MBL: Linfocitosis Monoclonal de células B
Dx according to therapy response	Diagnóstico según la respuesta a la terapia	c-CLL: Enfermedad constante CLL-PFT: Antes de la 1ª línea de tratamiento CLL-TFT: Después de la 1ª línea de tratamiento MBL: Linfocitosis Monoclonal de células B
Treatment	Tratamiento	Yes: Recibe tratamiento No: No recibe tratamiento
IGHV mutational status	Estado mutacional del IGHV	Mutated: IGHV mutado Unmutated: IGHV no mutado
Karyotype	Cariotipo	Aberration: Hay alteraciones Normal: No hay alteraciones

Chromosomal aberrations	Alteraciones cromosómicas	13q: Deleción 13q 17p: Deleción 17p Normal: No hay alteraciones Trisomy 12: Trisomía 12

Además, para cada uno de los individuos incluidos en el estudio, disponemos de seis tipos de datos multiómicos diferentes que son: "Autoanticuerpos IgG", "Autoanticuerpos IgM", "ESPA IgG", "ESPA IgM", "Immunocheckpoints", y "Proteoma". Todos estos datos proceden de análisis proteómicos masivos y proporcionan información cuantitativa. Han sido obtenidos de los siguientes artículos: "Systematic Evaluation of Antigenic Stimulation in Chronic Lymphocytic Leukemia: Humoral Immunity as Biomarkers for Disease Evolution" (Landeira-Viñuela et al., 2023) y "Unravelling soluble immune checkpoints in chronic lymphocytic leukemia: Physiological immunomodulators or immune dysfunction" (Landeira-Viñuela et al., 2022)

En cuanto a las bases de datos Autoanticuerpos, ambas están formadas por las mismas 122 proteínas, pero una contiene información de IgG y otra de IgM. Lo mismo sucede con los ESPA y cada una de ellas está compuesta por 37 patógenos medidos en 5 concentraciones (0.00001, 0.0001, 0.001, 0.01, 0.08 µg/µL). Las bases de datos "Immunockeckpoints" y "Proteoma" están formadas por 103 y 2946 proteínas respectivamente.

Se observó que la base de datos "Autoanticuerpos IgM" presentaba valores perdidos para 12 proteínas en las 67 muestras: *CA125, BRCA2/BRCA1, CXCL10, Catalase, Glutamate decarboxylase-65/GAD2, Carbonic Anhydrase 6 (CA6), Parotid Secretory Protein (PSP), Sm/RNP, HPV-16 E6, HPV-11 E6 (aa 1-150), HPV-6 E2*. Por lo tanto, éstas no se tuvieron en cuenta en los posteriores análisis.

Basándonos en el artículo "Systematic Evaluation of Antigenic Stimulation in Chronic Lymphocytic Leukemia: Humoral Immunity as Biomarkers for Disease Evolution", se obtuvieron mejores resultados para los patógenos que estaban medidos en una concentración de 0.0001 µg/µL. En consecuencia, únicamente utilizaremos estos en nuestro estudio.

Teniendo en cuenta la información mencionada en los párrafos anteriores, el tamaño final de cada base de datos se recoge en la Tabla 4.

Tabla 4

Tamaño de las bases de datos para los 67 individuos de LLC en el estudio

Base de datos	Número de variables
Autoanticuerpos IgG	122
Autoanticuerpos IgM	110
ESPA IgG	37
ESPA IgM	37
Immunocheckpoints	103
Proteoma	2946

Si juntamos todos estos datos, nos queda una matriz final de 67 individuos y 3364 variables, 9 clínico-biológicas y 3355 correspondientes a datos multiómicos.

4.1.2 OBTENCIÓN DE LOS DATOS

Se recolectaron un total de 67 muestras de sangre periférica de adultos con diagnóstico de Leucemia Linfática Crónica o Linfocitosis Monoclonal de células B. El diagnóstico se llevó a cabo de acuerdo con las directrices nacionales del Grupo Español de Leucemia Linfocítica Crónica (GELLC). En cuanto a la evaluación del estadio de la enfermedad, se utilizaron los criterios de Binet y Rai. (Landeira-Viñuela et al., 2023)

Para el estudio de los autoanticuerpos, se utilizó un microarray comercial de proteínas de GeneCopoeia™ que contenía 122 proteínas. Las muestras de suero se analizaron y evaluaron siguiendo las recomendaciones del fabricante. (Landeira-Viñuela et al., 2023)

En cuanto a los patógenos, se elaboró una recopilación de 37 antígenos específicos (array ESPA). Además, se incluyeron en el array 10 controles tanto positivos como negativos únicamente para validar los resultados obtenidos, es decir, no se tuvieron en cuenta para el análisis estadístico. Cada subarray estaba compuesto por 5 diluciones (0.00001, 0.0001, 0.001, 0.01, 0.08 µg/µL) de cada antígeno e inmunoglobulina, aunque al final solo se tuvo en cuenta la dilución 0.0001 µg/µL. (Landeira-Viñuela et al., 2023)

Después de realizar las convenientes incubaciones y lavados de los arrays, éstos se secaron y se escanearon utilizando el escáner GenePix® 4000B Microarray Scanner. Se ajustaron los parámetros para poder cuantificar los valores de intensidad de señal correspondientes a la detección de anticuerpos IgG e IgM. Para ello, se utilizó el marcador Cy3 (λ = 532 nm) para la señal de IgG y Alexa Fluor 647 (λ = 635 nm) para IgM. Mediante este proceso, se obtuvieron las imágenes TIFF (Tagged Image File Format) que se analizaron con el software GenePix Pro 6.0. (Landeira-Viñuela et al., 2023)

Respecto al estudio de immunocheckpoints, se utilizaron cuatro kits diferentes de Luminex siguiendo las instrucciones del fabricante. Las muestras fueron analizadas con tecnología Luminex, haciendo uso del instrumento MAGPIX® y la versión 4.2 del software xMAP® (Landeira-Viñuela et al., 2022).

En cuanto al estudio del proteoma, cada una de las muestras de células B, extraídas de la sangre periférica, fue procesada siguiendo el protocolo óptimo correspondiente a cada ensayo específico de proteómica (caracterización mediante LC-MS/MS o caracterización mediante proteómica de afinidad).

4.1.2.1 ANÁLISIS CUANTITATIVO DE AUTOANTICUERPOS Y ANTÍGENOS MICROBIANOS

Para obtener la señal de fluorescencia correspondiente a cada uno de los puntos, se siguieron los siguientes pasos (Landeira-Viñuela et al., 2023):

1. Normalización de la señal (S):

$$S = \frac{\widetilde{F_\lambda} - \widetilde{B_\lambda}}{B\sigma_\lambda}$$

siendo $\widetilde{F_\lambda}$ la mediana de la intensidad de señal de las características, $\widetilde{B_\lambda}$ la mediana de la intensidad de señal del background y $B\sigma_\lambda$ la desviación típica del background, teniendo en cuenta la longitud de onda.

2. Intensidad de la señal de control negativa (S_{neg}):

$$S_{neg} = \frac{\sum S_p}{N}$$

siendo S_p la señal estandarizada para un determinado punto que se ha incubado con un control negativo y N el número de subarrays que pertenecen a una misma incubación.

3. Intensidad de señal de los spots (I):

$$I = S_y - S_{neg}$$

siendo S_y la señal normalizada para un punto y S_{neg} la señal normalizada para su respectivo punto de control negativo.

4. Promedio de la intensidad de señal del antígeno ($\bar{I}_m$):

$$\bar{I}_m = \frac{\sum I_y}{Z}$$

siendo I_y la señal de intensidad de un punto específico y Z la cantidad de repeticiones.

5. Selección de los spots con respecto a la referencia (NIST, muestra de suero referencia de población general, proporcionado por el National Institute of Standards del NIH (EE.UU.).

   Únicamente se tuvieron en cuenta los antígenos con una señal de intensidad positiva y se escogieron las señales positivas de antígeno que superaban a la referencia:

$$\bar{I}_m^S > \bar{I}_m^r + I_m^r\sigma$$

siendo $\bar{I}_m^S$ es la intensidad de señal media de los antígenos, $\bar{I}_m^r$ la intensidad de señal media de la referencia y $I_m^r\sigma$ la desviación típica de la intensidad de señal del antígeno para la referencia.

6. Z-Score de los antígenos microbianos: Solamente se realiza para aquellos que tengan valores positivos:

$$Z\ score = \frac{\bar{I}_m^S}{\bar{I}_m^r + I_m^r\sigma}$$

4.1.2.2 ANÁLISIS CUANTITATIVO DE LAS PROTEÍNAS DE CONTROL INMUNITARIO

Para cada una de las proteínas, se determinó la concentración utilizando un algoritmo de ajuste de curva logística de 5 parámetros (Landeira-Viñuela et al., 2022). Este se empleó para los datos de concentración y su fórmula es:

$$y = a + \frac{b - a}{\left(1 + \left(\frac{x}{c}\right)^{d}\right)^{f}}$$

siendo x la concentración medida en pg/ml, a, b, c, d y f constantes y, por último, y la intensidad mediana neta de fluorescencia.

4.1.2.3 ANÁLISIS CUANTITATIVO DEL PROTEOMA

La cuantificación de la concentración de las proteínas caracterizadas por LC-MS/MS se llevó a cabo mediante el ensayo de Bradford a partir del reactivo *Coomassie Plus Protein Assay Reagent*. Para aquellas proteínas que se caracterizaron mediante proteómica de afinidad, se determinó su concentración mediante el kit de ensayo de proteínas *Pierce™ BCA Protein Assay Kit.*

En la base de datos resultante, los datos se encontraban estandarizados (media 0 y desviación típica 1) por filas, es decir, por los pacientes.

4.1.2.4 ESTANDARIZACIÓN DE LOS DATOS

No todos los conjuntos de datos se encuentran en la misma escala y para poder trabajar con todos ellos a la vez, se ha realizado una estandarización por columnas (variables). Para ello se ha aplicado la fórmula del *Z-Score*:

$$ZScore = \frac{x - \bar{x}}{S}$$

donde x es el valor que se está evaluando, $\bar{x}$ es la media de la proteína correspondiente y S la desviación típica. De esta manera, se consigue que cada una de las variables tenga media 0 y desviación típica 1 y así poder comparar las unas con las otras sin problemas con las unidades de medida.

4.1.3 SOFTWARE

Para la realización de los análisis de este estudio se han empleado los programas estadísticos RStudio versión 4.3.0 (R Core Team, 2023) y Microsoft Office Excel 2016 (Microsoft, 2016).

Excel se utilizó en un primer momento para realizar el preprocesamiento y limpieza de las bases de datos y el programa RStudio se empleó para la realización de gráficos y análisis estadísticos.

4.2 METODOLOGÍA

4.2.1 ANÁLISIS FACTORIAL

El Análisis Factorial (AF) es una técnica estadística multivariante cuyo objetivo principal es reducir la dimensionalidad de un conjunto de datos, intentando mantener siempre la mayor cantidad de información posible (de la Fuente Fernández, 2011).

Esta técnica tiene sus raíces en los análisis de regresión lineal llevados a cabo por Galton. Sin embargo, fue Karl Pearson la primera persona en presentar una primera versión el método de componentes principales, un método utilizado para reducir la dimensionalidad de un conjunto de datos, en el año 1901. Tres años más tarde, Charles Spearman propuso la existencia de un factor general en la inteligencia humana que se encuentra presente en todas las diferentes habilidades cognitivas. Este hallazgo sirvió para asentar las bases del análisis factorial. Con el paso de los años, otros investigadores han ido realizando aportaciones y complementando idea de Spearman.

Existen dos tipos de Análisis Factorial: Análisis Factorial Exploratorio (AFE) y Análisis Factorial Confirmatorio (AFC). Podemos comparar ambas técnicas en la Tabla 5.

Tabla 5

Diferencias entre AFE y AFC

ANÁLISIS FACTORIAL EXPLORATORIO (AFE)	ANÁLISIS FACTORIAL CONFIRMATORIO (AFC)
El objetivo es explorar la estructura oculta de un conjunto de datos	El objetivo es confirmar una estructura previamente definida de un conjunto de datos
No se establecen hipótesis a priori sobre la estructura factorial	Se establecen hipótesis a priori sobre la estructura factorial
A partir de este análisis se determina el número de factores	El número de factores se especifica antes de realizar el análisis
No se puede replicar	Se puede replicar

De manera general, este método trata de explicar un conjunto de *p* variables observadas mediante un número más pequeño, *k<p* de factores o variables latentes. Estos factores son variables que no se pueden observar a simple vista y se infieren utilizando las variables originales. Podemos distinguir entre dos tipos de factores: los factores comunes y los factores únicos o específicos, uno por cada variable (*Reducción de la dimensión: Análisis factorial*, 2016).

Las variables observadas y los factores se relacionan de la siguiente manera:

$$X_1 = a_{11}F_1 + \cdots + a_{1k}F_k + d_1U_1$$
$$X_2 = a_{21}F_1 + \cdots + a_{2k}F_k + d_2U_2$$

…

$$X_p = a_{p1}F_1 + \cdots + a_{pk}F_k + d_pU_p$$

O de forma matricial:

$$X = AF + DU$$

Donde:

- *X* es un vector de tipo columna constituido por las *p* variables observables,
- *A* es una matriz de pesos factoriales o cargas de dimensiones *p×k* que representa la relación entre las variables observadas y los factores latentes,
- *F* es un vector columna formado por los *k* factores latentes comunes,
- *D* es una matriz diagonal de tamaño *p×p* cuyos elementos son las saturaciones entre las variables observables y los factores únicos,
- *U* es un vector columna compuesto por los coeficientes de los factores únicos.

El objetivo del Análisis Factorial es hallar una estimación de la matriz de pesos factoriales *A* e interpretarla (Cuadras, 2014). Hay que tener en cuenta dos aspectos a la hora de interpretar los pesos. La primera es el tamaño, es decir, un peso grande nos muestra que existe una fuerte relación entre la variable y el factor, lo que significa que dicha variable ha sido importante a la hora de definir el factor, mientras que un peso pequeño o cercano a 0 indica una relación débil. El segundo aspecto a tener en cuenta es el signo: un signo positivo quiere decir que la relación entre variable y factor es directa, y un signo negativo indica que la relación es inversa.

El análisis factorial se basa en tres supuestos que son:

- El número de factores latentes ha de ser menor que el número de variables observadas: $k < p$.
- Tanto variables como factores tienen media 0 y varianza 1, es decir, están centrados y estandarizados:

$$E(X_p) = E(F_k) = E(U_p) = 0$$

$$S^2_{X_p} = S^2_{F_k} = S^2_{U_p} = 1 \text{ donde } p = 1 \dots P, k = 1 \dots K$$

- Los K + P factores comunes y específicos están incorrelacionados para todo i≠j

$$cor(F_i, F_j) = 0$$

$$cor(U_i, U_j) = 0$$

$$cor(F_i, U_j) = 0$$

Considerando el esquema del modelo factorial y el segundo supuesto que dice que la varianza de cada una de las variables X_i es igual a 1, dicha varianza se puede descomponer de la siguiente manera:

$$S_i^2 = a_{i1}^2 S_{F_1}^2 + \cdots + a_{ik}^2 S_{F_k}^2 + \sum_{j=1}^{k} \sum_{t \neq j} a_j a_t S_{F_j F_t} + d_i^2 s_{U_i}^2 + d_i \sum_{j=1}^{k} a_{ik} S_{F_j U_i} = 1$$

Si tenemos en cuenta el segundo supuesto, las varianzas de los factores serán igual a 1 y las covarianzas serán iguales que los coeficientes de correlación:

$$S_i^2 = a_{i1}^2 + \cdots + a_{ik}^2 + \sum_{j=1}^{k} \sum_{t \neq j} a_j a_t r_{F_j F_t} + d_i^2 + d_i \sum_{j=1}^{k} a_{ik} r_{F_j U_i} = 1$$

Como consecuencia del tercer supuesto, la expresión se reduce a:

$$S_i^2 = a_{i1}^2 + \cdots + a_{ik}^2 + d_i^2 = 1$$

La suma de los k primeros términos ($a_{i1}^2 + \cdots + a_{ik}^2$) se denomina comunalidad de la i-ésima variable observable, se simboliza con h_i^2 y se corresponde con la varianza explicada por los factores comunes. Por otra parte, d_i^2 se conoce por unicidad y hace referencia a la varianza explicada por el factor único. (*Reducción de la dimensión: Análisis factorial*, 2016)

Las técnicas más utilizadas para calcular la matriz factorial son el método del factor principal y el método de la máxima verosimilitud (Cuadras, 2014).

4.2.1.1 MÉTODO DEL FACTOR PRINCIPAL

Este método busca obtener una matriz factorial donde los factores latentes se encuentren incorrelacionados unos con otros y expliquen la mayor cantidad de varianza posible.

La suma de todas las varianzas explicadas por el factor F_j viene dada por la expresión:

$$V_j = a_{1j}^2 + \cdots + a_{pj}^2$$

Donde a_{ij}^2 se corresponde con la variabilidad que explica el factor F_j de la variable X_i.

Se define el primer factor principal F_1 como aquel que hace que V_1 sea máximo. Por lo tanto, se considera el problema de maximizar la suma de variabilidades explicadas por el factor 1 (V_1) sujeto a la restricción $R^* = AA'$. Haciendo uso del método de los multiplicadores de Lagrange, se tiene en cuenta la función:

$$V_1 + \sum_{j,j'=1}^{p} q_{jj'} \left(r_{jj'} - \sum_{k=1}^{m} a_{jk} a_{j'k} \right),$$

siendo $q_{jj'} = q_{j'j}$ los multiplicadores. Si realizamos las derivadas y las igualamos a 0, obtenemos como resultado que las saturaciones del primer factor principal $a_1 = (a_{11}, \dots, a_{p1})'$ cumplen:

$$R^* a_1 = \lambda_1 a_1,$$

lo que quiere decir que a_1 se corresponde con el primer vector propio de la matriz R^* y λ_1 con el primer valor propio y también con el máximo valor de V_1.

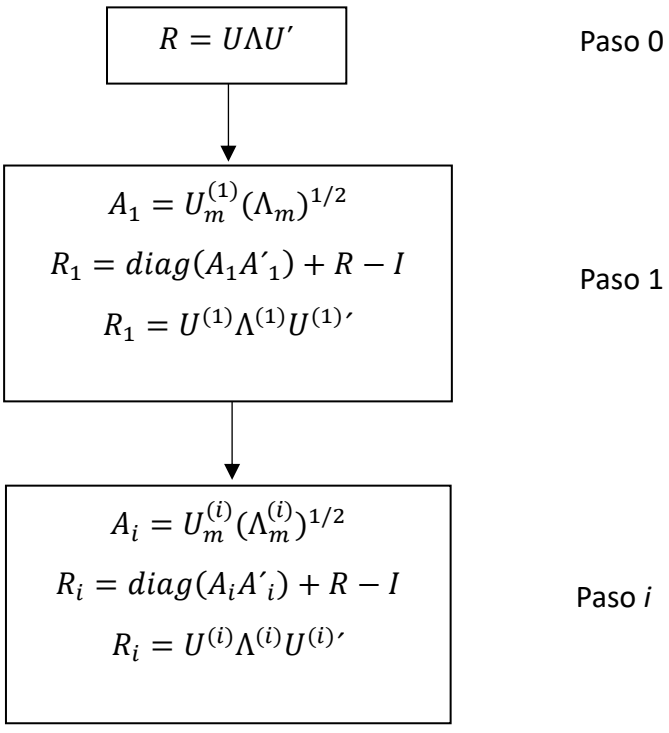
Si volvemos a aplicar el criterio del factor principal al modelo que nos queda restando el primer factor, volvemos a obtener que $a_2 = (a_{12}, \dots, a_{p2})'$ se corresponde con el segundo vector propio de la matriz R^* y λ_2 con el segundo valor propio y también con el máximo valor de V_2. Por lo tanto, si R^* se descompone espectralmente como $R^* = U\Lambda U'$, la solución del factor principal será:

$$A = U\Lambda^{1/2}$$

En la Figura 5 podemos encontrar de manera iterativa y resumida el algoritmo para obtener la matriz factorial mediante el método del factor principal. Podemos decir que A_i convergerá a A cuando en dos iteraciones consecutivas los valores de $diag(A_i A'_i)$ no varíen significativamente.

Figura 5

Algoritmo del método del factor principal



4.2.1.2 MÉTODO DE LA MÁXIMA VEROSIMILITUD

Se parte de la base de que la matriz de covarianzas Σ se puede descomponer de manera que:

$$\Sigma = AA' + V$$

donde V es una matriz diagonal que se corresponde con el cuadrado de la matriz de unicidades (D^2).

Bajo el supuesto de que los datos siguen una distribución normal con media 0 ($\mu = 0$), se establece la distancia F, entre la matriz de covarianzas observada y los valores estimados de dicha matriz mediante el AF como:

$$F = \log|\Sigma| + \mathrm{tr}(\Sigma^{-1}S) - \log|S| - p$$

Para poder conocer las estimaciones de las matrices A y V, derivamos la función F respecto de ellas e igualamos a 0, obteniendo las ecuaciones:

$$(AA' + V)^{-1}(AA' + V - S)(AA' + V)^{-1}A = 0$$
$$diag((AA' + V)^{-1}(AA' + V - S)(AA' + V)^{-1}) = 0$$

Resolviendo las ecuaciones anteriores y aplicando algún tipo de rotación a la solución, se obtiene finalmente la estimación de la matriz factorial, A.

4.2.2 CORRELACIÓN

La correlación es una medida estadística que nos permite conocer el grado de intensidad con el que se relacionan dos variables numéricas. Para poder determinar su valor, se establecen los coeficientes de correlación (Vargas Sabadías, 1995). Los dos coeficientes de correlación más conocidos son el de Pearson y el de Spearman.

Si las dos variables presentan una distribución normal, se suele utilizar el coeficiente de correlación lineal de Pearson (r) que mide el grado de asociación lineal entre ambas variables:

$$r = \frac{s_{XY}}{s_Y s_Y}$$

donde s_{xy} se corresponde con la covarianza de X e Y, s_X es la desviación típica de X y s_Y la desviación típica de Y.

Si alguna de las dos variables no cumple con el supuesto de normalidad o queremos medir el grado de asociación no lineal entre ambas, se recomienda utilizar el coeficiente de correlación de Spearman (r_s). Éste sirve para evaluar la asociación monótona entre las variables y tiene la siguiente fórmula (Caridad y Ocerin, 2016):

$$r_s = 1 - \frac{6\sum_{i=1}^{n}\left(r_{X_i} - r_{Y_i}\right)^2}{n(n^2 - 1)}$$

donde r_{X_i} y r_{Y_i} son los rangos de los valores de X e Y respectivamente y n es el número de parejas de datos de los que se dispone.

Ambos coeficientes están comprendidos entre -1 y 1 y se interpretan de la misma manera:

- Valores cercanos a -1 indican fuerte relación inversa.
- Valores cercanos a 0 indican ausencia de relación.
- Valores cercanos a 1 indican fuerte relación directa.

Únicamente podremos realizar estas interpretaciones si la relación entre las variables es significativa. Para ello se aplica una prueba de hipótesis donde la hipótesis nula (H_0) establece que no hay correlación significativa entre X e Y, y la hipótesis alternativa (H_1) expresa que sí existe correlación significativa entre estas variables (Illowsky et al., 2022).

4.2.3 ANÁLISIS DE CLÚSTER O DE CONGLOMERADOS

El clustering, también conocido como análisis de agrupamiento, es una técnica estadística de aprendizaje no supervisado que se utiliza para agrupar elementos similares en grupos distintos. De manera más específica, el objetivo de esta técnica es dividir un conjunto de datos compuesto por *n* unidades en *k* subconjuntos (*k*<<*n*) de modo que todos los grupos estén compuestos por al menos un elemento, cada elemento sea asignado exclusivamente a un grupo y los grupos no puedan compartir elementos entre ellos. (Giordani et al., 2020)

Al tratarse de un problema de aprendizaje no supervisado, se parte de datos no etiquetados para intentar descubrir estructuras subyacentes en dichos datos. Por ello, se define un clúster como un conjunto de elementos que presentan similitudes entre sí y que, al mismo tiempo, se diferencian significativamente de los elementos pertenecientes a otros clústeres. (Madhulatha, 2012)

Existen dos tipos principales de algoritmos de clustering: jerárquicos y particionales. Los algoritmos jerárquicos consisten en ir realizando grupos de manera sucesiva, utilizando para ello grupos que ya se habían establecido anteriormente. Dentro de los algoritmos jerárquicos podemos encontrar los aglomerativos, que como indica la palabra, comienzan por un elemento y conforme avanzan se van fusionando en estructuras más grandes; y los divisivos, que realizan el proceso contrario, se parte de la totalidad de los datos y se va dividiendo en grupos más pequeños. En el caso de los algoritmos particionales, todos los clústeres se determinan de manera simultánea. (Madhulatha, 2012)

4.2.3.1 MEDIDAS DE SIMILARIDAD Y DISTANCIA ENTRE OBJETOS

Para agrupar individuos en un análisis de clúster, se requiere de alguna medida numérica que permita describir de alguna manera la relación existente entre ellos. Estas medidas muestran distintos tipos de asociación y es muy importante seleccionar una medida adecuada dependiendo del problema específico que se esté abordando.

Existen dos tipos de medidas de asociación: distancias, que miden la separación o diferencia entre los individuos y similitudes, que indican qué tan parecidos son dichos individuos. Entre

las medidas de similitud más comunes se encuentran los coeficientes de correlación de Pearson, Spearman y Kendal. En cuanto a las medidas de distancia destacan: la distancia euclídea, la distancia de Manhattan, la distancia de Minkowski y la distancia de Canberra.

Para dos observaciones i, j donde x_{ir} y x_{jr} $(r = 1, \dots, m)$ son los valores que toman dichas observaciones en la variable r, se definen las siguientes distancias $d_{(i,j)}$:

- **Distancia euclídea**: Se basa en el Teorema de Pitágoras y mide la longitud del segmento que une dos puntos en un espacio multidimensional.

$$d_{(i,j)} = \sqrt{\sum_{r=1}^{m} (x_{ir} - x_{jr})^2}$$

- **Distancia de Manhattan**: Mide la distancia entre dos puntos teniendo en cuenta únicamente los desplazamientos verticales y horizontales. Es preferible utilizar la distancia de Manhattan cuando las variables están medidas en diferentes unidades.

$$d_{(i,j)} = \sum_{r=1}^{m} |x_{ir} - x_{jr}|$$

- **Distancia de Minkowski**: Esta distancia depende de un parámetro p, que puede ser cualquier número real, de modo que permite adaptarse a distintas situaciones.
  - $p = 1 \rightarrow$ Distancia de Manhattan
  - $p = 2 \rightarrow$ Distancia euclídea

$$d_{(i,j)} = \sqrt[p]{\sum_{r=1}^{m} |x_{ir} - x_{jr}|^p}$$

- **Distancia de Canberra**: Mide la distancia entre dos puntos teniendo en cuenta tanto los valores relativos como los valores absolutos de las variables.

$$d_{(i,j)} = \sum_{r=1}^{m} \frac{|x_{ir} - x_{jr}|}{(x_{ir} + x_{jr})}$$

4.2.3.2 FORMAS DE MEDIR LA DISTANCIA ENTRE CLÚSTERES

Se conocen varias formas de calcular la distancia entre dos clústeres y cada una de ellas lleva a la construcción de grupos distintos. Teniendo en cuenta que δ_i son constantes de

ponderación, la distancia que separa un objeto R de un grupo (P,Q) se calcula a partir de la siguiente función:

$$d(R, P+Q) = \delta_1 d(R,P) + \delta_2 d(R,Q) + \delta_3 d(P,Q) + \delta_4 |d(R,P) - d(R,Q)|$$

Según los valores que se asignen a las diferentes δ_i, se obtendrán los siguientes métodos jerárquicos:

- **MÉTODO DE LA MEDIA (AVERAGE LINKAGE)**: Es una técnica que genera clústers compactos, con una varianza similar y de tamaño óptimo, evitando agrupaciones muy grandes o muy pequeñas. Hay que tener en cuenta que los resultados son susceptibles a los cambios en la escala de las medidas. Además, este método ofrece una buena representación gráfica de los resultados, lo que facilita la interpretación visual de los grupos obtenidos. La distancia entre los clústeres se calcula como:

$$d(R, P+Q) = \frac{1}{2} d(R,P) + \frac{1}{2} d(R,Q)$$

- **MÉTODO DE WARD**: También se llama método de la varianza mínima. Se trata de una técnica altamente eficiente que tiende a generar clústeres pequeños, es decir, con pocas observaciones y puede ser influenciado por la presencia de valores atípicos (outliers). La distancia entre los clústeres se calcula sustituyendo en la fórmula inicial:

$$\delta_1 = \frac{n_R + n_P}{n_R + n_P + n_Q},\ \delta_2 = \frac{n_R + n_Q}{n_R + n_P + n_Q},\ \delta_3 = -\frac{n_R}{n_R + n_P + n_Q} \text{ y } \delta_4 = 0$$

  donde n_R, n_Q, n_P se corresponden con el número de objetos que hay en cada grupo.

- **MÉTODO DEL VECINO MÁS CERCANO**: Es un método que tiene tendencia a generar clústeres de gran tamaño y poco significativos. No es una técnica buena para resumir datos, aunque es de gran utilidad para identificar outliers ya que estos serán los últimos en ser agrupados.
  La distancia que separa dos clústeres se define como el valor mínimo de las distancias entre dos objetos, uno correspondiente a cada clúster, es decir:

$$d(R, P+Q) = min\ (d(R,P), d(R,Q))$$

- **MÉTODO DEL VECINO MÁS LEJANO (COMPLETE LINKAGE)**: Al contrario que en el caso anterior, esta técnica proporciona clústeres de tamaño pequeño pero compactos, aunque también es útil para la identificación de outliers.
  La distancia que separa dos clústeres se define como el valor máximo de las distancias entre dos objetos, uno correspondiente a cada clúster, es decir:

$$d(R, P+Q) = max\ (d(R,P), d(R,Q))$$

- **ALGORITMO DIANA (Divisive ANAlysis):** Se parte de un único clúster formado por todos los individuos y a medida que va avanzando el algoritmo, cada clúster se va dividiendo en dos clústeres más pequeños hasta que en cada clúster haya únicamente un individuo.

  Se describe a continuación el procedimiento a seguir para realizar el algoritmo DIANA (Kaufman & Rousseeuw, 1990):

  En cada etapa, el algoritmo divide un clúster, al que llamaremos R, en dos clústeres más pequeños A y B. En la etapa inicial, A se corresponde con R y en B no hay ningún elemento, por lo tanto, para cada elemento i en A se calcula la disimilitud promedio con respecto a los demás objetos en A de la siguiente manera:

  $$d_{(i,A\setminus\{i\})} = \frac{1}{|A|-1}\sum_{\substack{j\in A\\ j\neq i}} d_{(i,j)}$$

  Se selecciona el elemento i que maximice el valor de la ecuación anterior y se cambia dicho elemento del clúster A al clúster B. Siempre y cuando en A haya más de un elemento se explora la posibilidad de mover otros puntos de A a B y se calcula:

  $$d_{(i,A\setminus\{i\})} - d_{(i,B)} = \frac{1}{|A|-1}\sum_{\substack{j\in A\\ j\neq i}} d_{(i,j)} - \frac{1}{|B|}\sum_{h\in B} d_{(i,h)}$$

  Al igual que en el paso anterior, se van seleccionan aquellos i que hagan máxima esta función y den como resultado un valor positivo. Es decir, el criterio de parada se dará cuando el valor que maximice $d_{(i,A\setminus\{i\})} - d_{(i,B)}$ sea menor o igual a 0.

  A continuación, se evalúa el diámetro de cada clúster Q de la siguiente manera:

  $$diam(Q) = \max_{\substack{j\in Q\\ h\in Q}} d_{(j,h)}$$

  Se seleccionará aquel clúster que tenga un diámetro mayor para ser dividido. El valor obtenido se usa también como punto de referencia para representar la división de Q en el gráfico. Se puede realizar esta acción debido a que el valor de la ecuación anterior es monótono, es decir:

  $$A \subset R \Rightarrow diam(A) \leq diam(R)$$

  Esta propiedad denota que los niveles de los siguientes pasos constituyen una secuencia que no incrementa.

En cuanto a los algoritmos de clustering particionales, el método más común es el de k-medias. Este método busca dividir un conjunto de n observaciones en un número menor predefinido k de clústeres minimizando la suma de los cuadrados de las distancias entre cada una de las observaciones con el centroide de su clúster correspondiente. Es decir, se busca minimizar la dispersión dentro de cada grupo.

Este algoritmo se divide en 4 pasos:

1. Se escogen de manera aleatoria un número de puntos igual a los clústeres que se deseen formar. Estos puntos se utilizarán como los centroides iniciales.

2. Se crean los grupos asignando cada observación al centroide que se encuentra a una menor distancia (teniendo en cuenta el tipo de distancia elegida).

3. Después de que todas las observaciones hayan sido asignadas a un clúster, se procede a recalcular los centroides a partir de las medias de los clústeres creados en el punto 2.

4. Se repiten los pasos 2 y 3 iterativamente hasta lograr que el algoritmo converja, en otras palabras, hasta que de una iteración a otra no varíen ni los centroides ni las asignaciones.

4.2.3.3 CONSENSUS CLUSTERING (Agrupamiento por Consenso)

Al aplicar diferentes métodos de agrupamiento a un mismo conjunto de datos, es posible obtener resultados diferentes debido a la variabilidad en las fuentes de los datos o debido a que los algoritmos de clustering utilizados tienen una naturaleza no determinista. Como solución a este problema, surge el concepto de *Consensus Clustering* o clustering de consenso, ya que permite combinar y sintetizar varios algoritmos de agrupamiento para obtener una única solución más fiable que refleje la estructura subyacente de un conjunto de datos (Goder & Filkov, 2008).

Los algoritmos de clustering de consenso ofrecen múltiples ventajas en el análisis de datos: son capaces de formar agrupamientos combinados de mayor calidad que no pueden ser obtenidos por algoritmos de clustering individuales, son más robustos frente a los outliers y al ruido, es decir, son menos sensibles frente a ellos y permiten combinar soluciones procedentes de diversas fuentes de datos (Nguyen & Caruana, 2007).

De forma matemática, se podría definir el problema de la siguiente manera (Nguyen & Caruana, 2007):

Partimos de un conjunto de N elementos, $X = \{x_1, x_2, \dots, x_N\}$ agrupados de C diferentes maneras, $\pi = \{\pi_1, \pi_2, \dots, \pi_C\}$. Cada uno de los agrupamientos π_i se corresponde con una asignación de los objetos de X al clúster $\{1, \dots, n_{\pi_i}\}$, donde n_{π_i} representa el número de clústeres de π_i. Por lo tanto, el objetivo será hallar un nuevo agrupamiento π^* que sintetice de manera óptima el conjunto π del que se disponía.

Una vez se haya seleccionado un algoritmo de agrupamiento y un método de remuestreo, se necesita encontrar una forma de representar y medir el grado de concordancia entre los resultados de los diferentes agrupamientos realizados. Para abordar este problema se crea una matriz de dimensión $N \times N$ denominada matriz de consenso $(\mathcal{M})$, la cual se obtiene de la siguiente manera (Monti et al., 2003):

$$\mathcal{M}(i,j) = \frac{\sum_{h=1}^{H} M^h(i,j)}{\sum_{h=1}^{H} I^h(i,j)}$$

Para poder desarrollar la ecuación anterior, se parte de una lista de conjuntos de datos $(X^{(1)}, \ldots, X^{(H)})$ que se obtienen al aplicar el método de remuestreo al conjunto de datos original X. Cada uno de los conjuntos $X^{(h)}$ de datos se somete al algoritmo de agrupamiento elegido dando lugar a la matriz de conectividad M^h. Esta matriz tiene una dimensión $N \times N$ y sus elementos se definen de manera que:

$$M^{(h)}(i,j) = \begin{cases} 1\ si\ i\ y\ j\ pertenecen\ al\ mismo\ cl\acute{u}ster, \\ 0\ si\ i\ y\ j\ no\ pertenecen\ al\ mismo\ cl\acute{u}ster \end{cases}$$

Por otro lado, hay que tener en cuenta que la mayoría de los métodos de remuestreo proporcionan conjuntos de datos que no presentan todos los elementos del conjunto de datos original. Por ello, se necesita de una matriz de indicadores I^h, también de dimensión $N \times N$, tal que sus elementos se definan de la siguiente manera:

$$I^{(h)}(i,j) = \begin{cases} 1\ si\ i\ y\ j\ están\ presentes\ en\ el\ conjunto\ de\ datos, \\ 0\ en\ caso\ contrario \end{cases}$$

Tal y como se han definido la matriz $\mathcal{M}$, cada elemento $\mathcal{M}(i,j)$ toma valores entre 0 y 1, de manera que valores cercanos a 0 se corresponden con un bajo consenso y valores cercanos a 1 con un alto consenso.

Por último, se representa visualmente la matriz de consenso mediante un mapa de calor o heatmap. Se suele utilizar una escala de colores donde el blanco representa los valores cercanos a 0 y colores más oscuros como el rojo o el azul para valores cercanos a 1.

4.2.4 HEATMAP (MAPA de Calor)

Un Heatmap o mapa de calor es una técnica estadística que se basa en una representación gráfica de una matriz de datos bidimensional numérica por medio de colores (Gehlenborg & Wong, 2012). En la rama de la Biología se utilizan frecuentemente para representar datos multivariados, de manera que la matriz de la que se parte contenga los datos de cada uno de los individuos en las columnas y las características a estudiar (en nuestro caso proteínas) en filas (Key, 2012).

Un Heatmap lleva a cabo dos operaciones fundamentales en la matriz de datos. En primer lugar, reestructura las filas y las columnas, de manera que se agrupan aquellas que tengan unos perfiles similares, para que estas similitudes sean más evidentes a simple vista. Posteriormente, se asigna un color a cada entrada de la matriz, de tal manera que permite visualizar de manera gráfica los patrones o tendencias que siguen los datos (Key, 2012).

La elección de una escala de colores adecuada es muy importante, ya que puede llegar a afectar a la interpretación de los datos. Por ello, la escala seleccionada debe resaltar las similitudes y diferencias que existen en los datos, de modo que no puedan llevar a confusiones. Habitualmente, se utiliza una gama de colores que va desde tonos claros como el blanco para representar los valores más pequeños, hasta tonos más oscuros como el azul o el rojo para valores más altos.

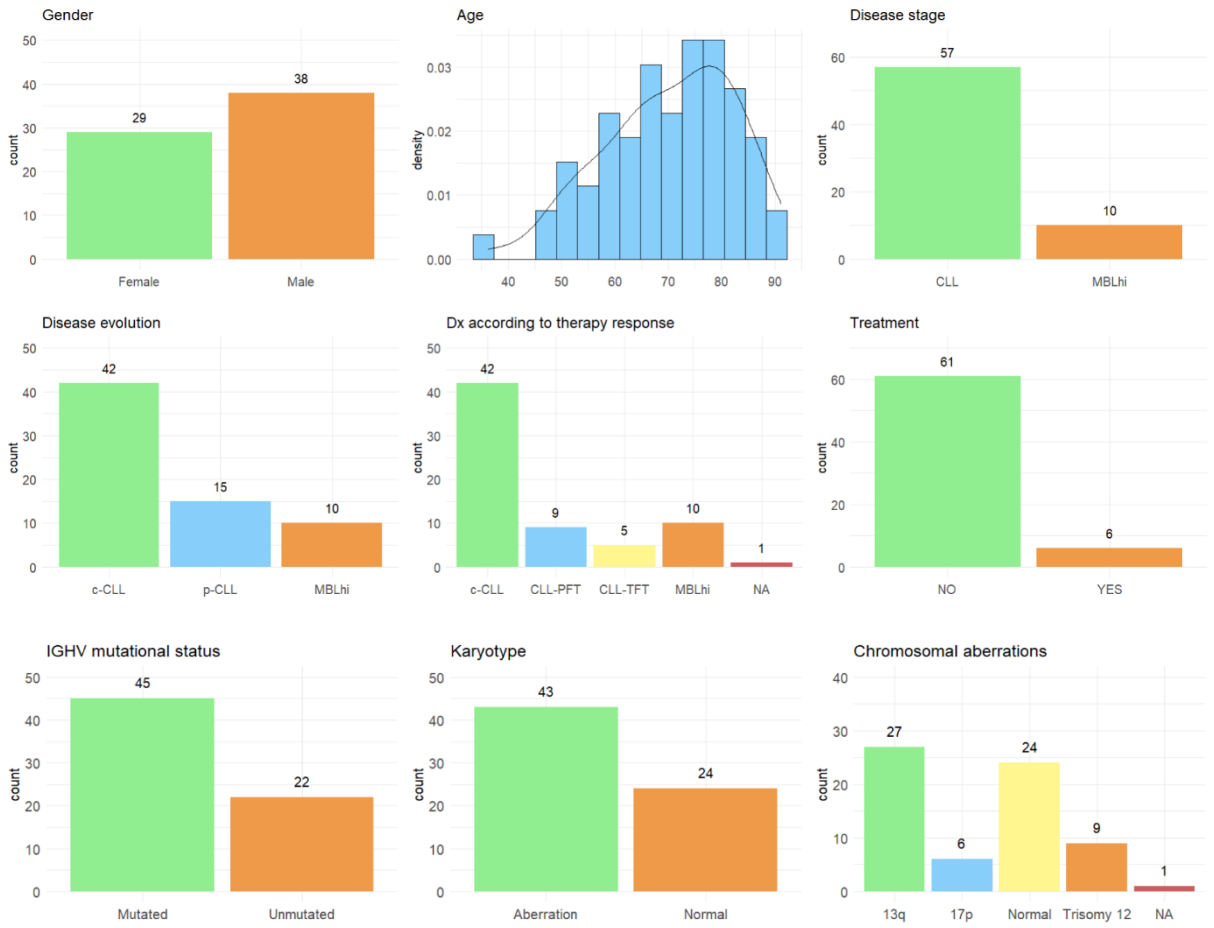
5 RESULTADOS

5.1 ANÁLISIS DESCRIPTIVO

En primer lugar, se realizó un análisis descriptivo (Figura 6) de las variables clínico-biológicas mencionadas en el apartado 0. Las variables cualitativas se representaron mediante diagramas de barras y, la variable correspondiente edad (*Age*), se representó combinando un histograma y la curva de densidad. Para ello, se empleó el paquete de RStudio "*ggplot2*" (Wickham, 2016).

Figura 6

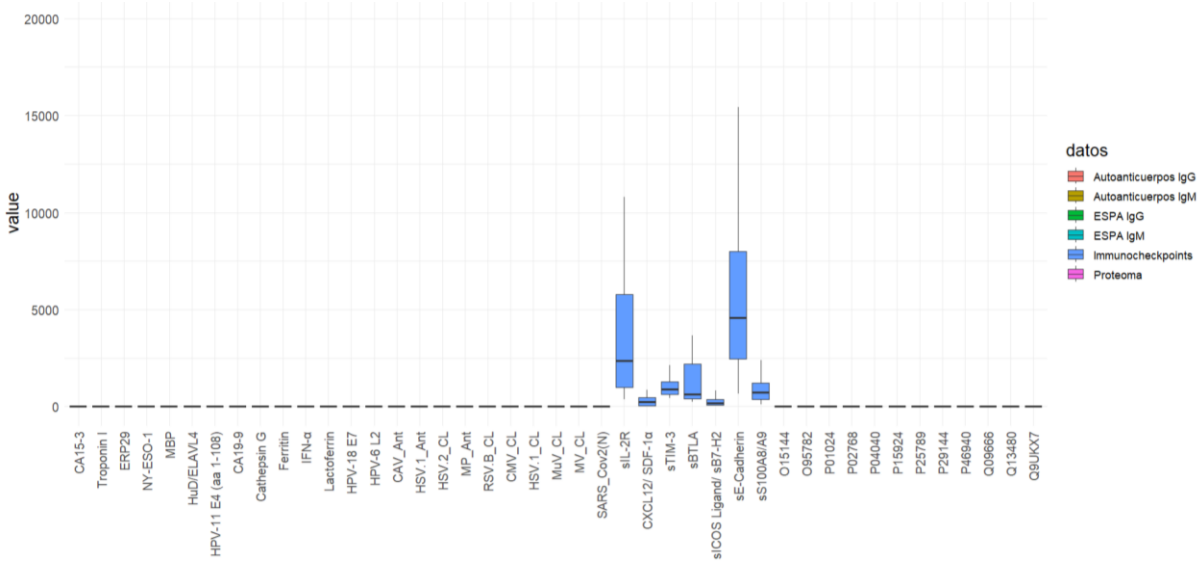
Análisis descriptivo de las variables clínico-biológicas



En cuanto a las variables correspondientes a los datos ómicos, se realizó en un primer momento un boxplot a partir de los datos originales (Figura 7). Se seleccionó para cada base de datos una muestra de proteínas para poder visualizarlas de manera conjunta. Podemos observar cómo los datos no se encuentran en la misma escala, especialmente los correspondientes a "Immunocheckpoints" presentan valores muy superiores al resto de datos.

Figura 7

Boxplots de los datos originales

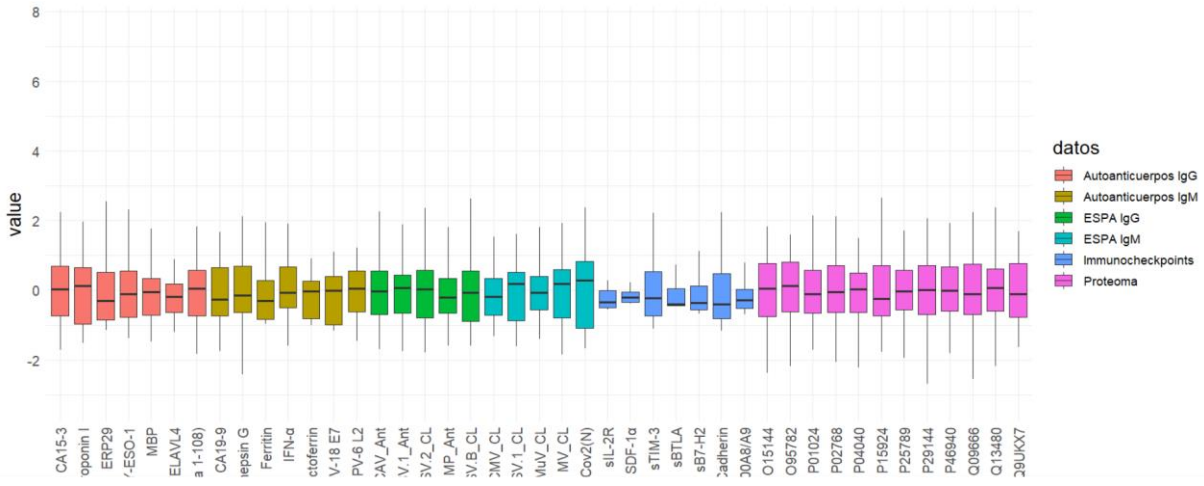


Para poder comparar las diferentes proteínas e integrarlas en un mismo análisis de manera que todas aporten los mismos pesos, es decir, que no tengan más peso aquellas que posean unos valores mayores, se realizó un Z-Score para cada una de ellas. En la Figura 8 se observa la misma muestra de proteínas que en la Figura 7, pero habiendo estandarizado los datos. De esta manera, se puede ver como todas las variables tienen la mediana en torno al 0 y su dispersión es similar.

En los anexos se encuentra el código utilizado para los gráficos sobre las variables clínico-biológicas y los boxplots tanto de los datos originales como de los datos estandarizados para cada una de las bases de datos por separado.

Figura 8

Boxplots de los datos estandarizados



5.2 INTEGRACIÓN DE DATOS MULTIÓMICOS

Para poder analizar y trabajar de manera conjunta con las distintas modalidades de datos, se empleó una técnica estadística denominada MOFA (*Multi-Omics Factor Analysis*). Se trata de un método de aprendizaje no supervisado basado en el enfoque estadístico del análisis factorial, es decir, extrae factores latentes interpretables que recogen las fuentes de variabilidad más importantes en múltiples tipos de datos ómicos. (Yao et al., 2022)

De forma general, se parte de M matrices de datos numéricos $Y^1, \dots, Y^M$ de tamaños $N \times D_m$ donde N se corresponde con el número de individuos y D_m con la cantidad de características de cada matriz de datos m. En este estudio, $M = 6$, $N = 67$ y D_m depende del tipo de datos proteómicos. MOFA permite descomponer las matrices mencionadas de la siguiente manera (Argelaguet et al., 2018):

$$Y^m = ZW^{mT} + \varepsilon^m \ \ (m = 1, \dots, M)$$

donde Z es independiente de m y se corresponde con la matriz factorial. Por otro lado, W^m y ε^m varían dependiendo de la matriz de datos m y se corresponden con la matriz de pesos y la matriz de residuos, respectivamente.

El enfoque utilizado para desarrollar este modelo se basa en un marco bayesiano probabilístico, por lo que se asignan distribuciones iniciales a todas las variables del modelo que no se observan directamente (Z, W^m y ε^m) (Argelaguet et al., 2018).

Para realizar este análisis se ha utilizado el paquete "MOFA2" de RStudio (Argelaguet et al., 2018). En primer lugar, se modificó la base de datos original para poder aplicar MOFA, de tal manera que tuviese únicamente 4 columnas: Sample (nombre de las muestras), View (nombre de la base de datos), Feature (nombre de las características) y Value (valor en Z-Score). Una vez que se tiene la base de datos preparada, se crea un objeto MOFA sin entrenar con la función "*create_mofa()*" para comprobar que hayamos introducido bien los datos. Si visualizamos este objeto, R proporciona una salida con los siguientes datos:

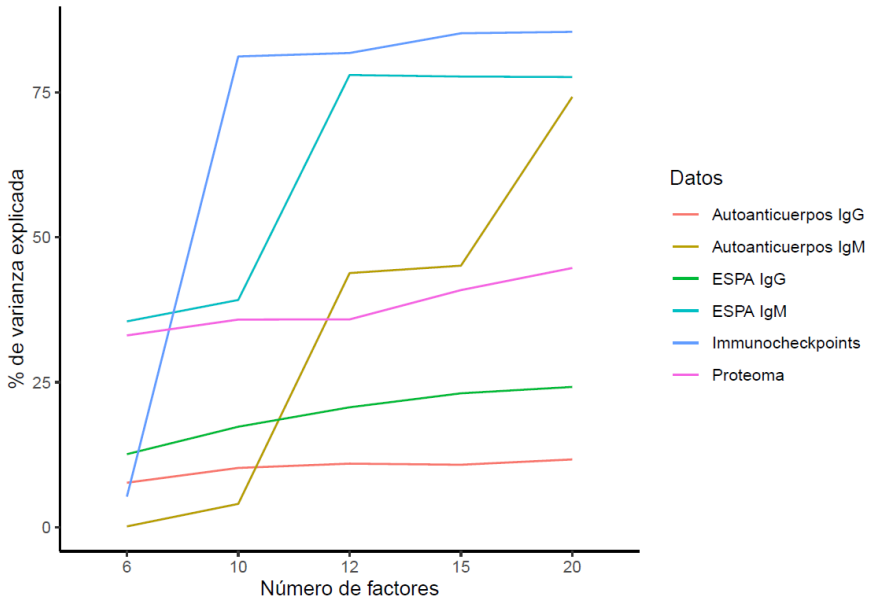
- Número de vistas: 6
- Nombres de las vistas: Autoanticuerpos IgG, Autoanticuerpos IgM, ESPA IgG, ESPA IgM, Immunocheckpoints, Proteoma
- Número de características por vista: 122, 110, 37, 37, 103, 2946
- Número de grupos: 1
- Nombre de los grupos: group1
- Número de muestras por grupo: 67

Antes de entrenar el modelo, se procede a definir una serie de opciones acerca de los datos, del modelo y del entrenamiento para poder ajustarlo de la manera que queramos. En este punto, se decide el número de factores que va a tener el modelo. Para ello, se fue probando con diferentes valores y observando el porcentaje de varianza explicada por todos ellos en cada base de datos (Figura 9). Se puede observar como a partir de 12 factores, el porcentaje de varianza explicada se mantiene más o menos estable para todas las matrices de datos excepto para Autoanticuerpos IgM que continúa creciendo. Aun así, para este estudio se

decidió utilizar 12 factores ya que es un número moderado de factores con el que conseguimos explicar una gran cantidad de información de todos los tipos de datos.

Figura 9

Varianza total explicada con respecto al número de factores



Teniendo en cuenta el paso anterior, establecemos las opciones para los datos. En este caso, vamos a dejar las opciones que proporciona R por defecto, ya que no nos interesa escalar las vistas debido a que todas ellas están en Z-Score. Por lo tanto, se utiliza el siguiente código:

```
> data_opts <- get_default_data_options(mofaobj)
```

En cuanto a las opciones del modelo únicamente se va a cambiar el número de factores que vienen por defecto por 12:

```
> model_opts <- get_default_model_options(mofaobj)
> model_opts$num_factors <- 12
```

Finalmente, para las opciones referentes al entrenamiento, establecemos un modo de convergencia intermedio, es decir que no sea muy rápido ni muy lento y dejamos las demás opciones como están:

```
> train_opts <- get_default_training_options(mofaobj)
> train_opts$convergence_mode <- "medium"
```

A continuación, se prepara el objeto MOFA para el entrenamiento mediante la función "*prepare_mofa()*", a la que introducimos como datos de entrada el objeto MOFA sin entrenar, las opciones de los datos, las opciones del modelo y las opciones de entrenamiento.

Por último, se entrena el objeto MOFA de la salida anterior con la función "*run_mofa()*". En este paso, R se comunica con Python (Rossum & Drake, 1995) mediante el paquete "*reticulate*" (Kalinowski et al., 2023) y se entrena el modelo con el paquete "*mofapy2"* (Argelaguet et al., 2018) de Python. La salida de esta función es un objeto de tipo MOFA que está entrenado y tiene los siguientes elementos:

- data: lista con los datos de entrada que se han utilizado para entrenar el modelo.
- Intercepts: lista con los interceptos del modelo
- imputed_data: lista con los datos que han sido imputados
- samples_metadata: data frame con los metadatos para las muestras
- features_metadata: data frame con los metadatos para las características
- expectations: valores esperados para las muestras y características en cada factor
- training_stats: lista con las estadísticas del modelo de entrenamiento
- training_options: lista con las opciones de entrenamiento del modelo
- stochastic_options: lista con las opciones de inferencia variacional estocástica
- data_options: lista con las opciones de los datos
- model_options: lista con las opciones del modelo
- dimensions: lista con las dimensiones del modelo. Contiene información sobre:
    - M: número de vistas o bases de datos
    - G: número de grupos de muestras
    - N: número de muestras
    - D: lista con la cantidad de características de lada vista
    - K: número de factores
- on_disk: indicador lógico que determina si los datos se cargan del disco
- dim_red: variedades en la reducción de la dimensionalidad no lineales
- cache: caché
- status: variable de tipo carácter que indica si el modelo está o no entrenado

En este punto, el modelo únicamente contiene los datos numéricos. Para añadir también las variables clínico-biológicas asociadas a los pacientes (metadatos), debemos crear un *data frame* (hoja de datos) que contenga, además de las variables, una columna "*sample*" con los nombres de las muestras. Incluimos estos datos de la siguiente manera:

```
> samples_metadata(mofaobj) <- metadata
```

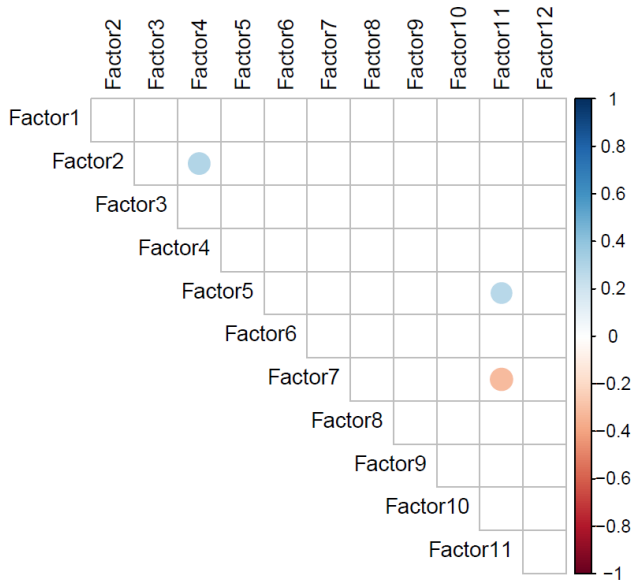
donde "mofaobj" es el objeto MOFA entrenado y "metadata" es el *data frame* con las variables clínico-biologicas.

Al ser una técnica que se basa en el análisis factorial, los factores deberían estar incorrelacionados. De este modo, si existe una alta correlación entre factores podríamos suponer que el modelo no es el adecuado y esto se podría relacionar con una normalización de los datos inadecuada o con un número excesivo de factores.

Para ello realizamos una prueba de correlaciones para cada pareja de factores y calculamos el coeficiente de correlación mediante el coeficiente de correlación lineal de Pearson. Para representar los resultados realizamos un gráfico "*corrplot*" (Figura 10) de tal manera que en los contrastes que hayan salido significativos se dibuja un círculo de tamaño y color correspondientes con el coeficiente de correlación. Las zonas en blanco representan que el contraste entre esos dos factores no es significativo.

Figura 10

Gráfico de correlaciones entre los factores



Una vez entrenado el modelo, el primer paso a realizar es calcular la cantidad de varianza explicada (R^2) por cada uno de los factores k en cada vista m (Figura 11) (Argelaguet et al., 2018):

$$R^2_{m,k} = 1 - \frac{\left(\sum_{n,d} y^m_{nd} - z_{nk} w^m_{kd} - \mu^m_d\right)^2}{\left(\sum_{n,d} y^m_{nd} - \mu^m_d\right)^2}$$

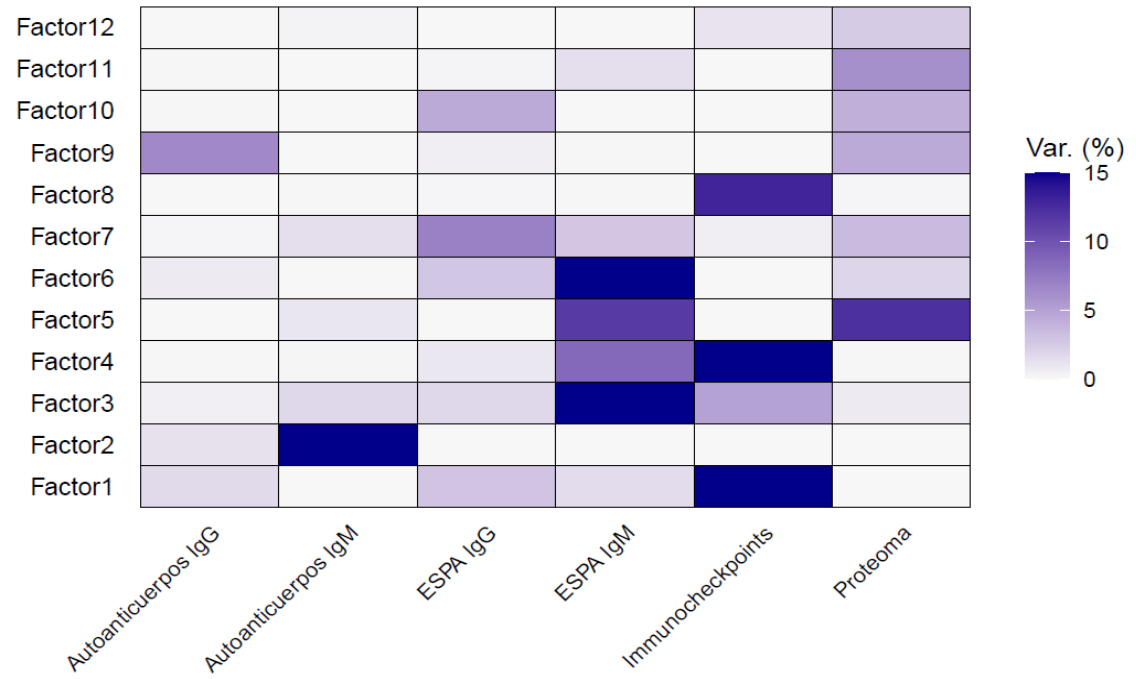
y la cantidad de varianza explicada por todos los factores en cada vista (Figura 12):

$$R^2_m = 1 - \frac{\left(\sum_{n,d} y^m_{nd} - \sum_k z_{nk} w^m_{kd} - \mu^m_d\right)^2}{\left(\sum_{n,d} y^m_{nd} - \mu^m_d\right)^2}$$

donde μ^m_d es la media de la característica d en cada matriz de datos m.

Figura 11

Varianza explicada por cada uno de los factores en cada vista



Tomando como referencia el artículo "*Proteogenomics refines the molecular classification of chronic lymphocytic leukemia*" (Herbst et al., 2022), para cada una de las bases de datos vamos a seleccionar e interpretar únicamente aquellos factores que expliquen como mínimo el 1.5% de la variabilidad (ANEXO).

En cuanto a los datos sobre Autoanticuerpos IgG, los factores que cumplen esa condición, ordenados de mayor a menor varianza explicada son el Factor 9 y el Factor 1. Siguiendo el mismo criterio para todos los tipos de datos, los Autoanticuerpos IgM se explican mediante el Factor 2 y el Factor 3. Los Factores 7, 10, 1, 6 y 3 explican la mayor parte de la variabilidad sobre los antígenos microbianos IgG (ESPA IgG) y los Factores 3, 6, 5, 4, 7 y 1 sobre los antígenos microbianos IgM (ESPA IgM). Con respecto a los puntos de control inmunitario, se seleccionan el Factor 1, Factor 4, Factor 8 y Factor 3. Finalmente, la base de datos Proteoma se explica mayoritariamente por los Factores 5, 11, 9, 10, 7, 12 y 6.

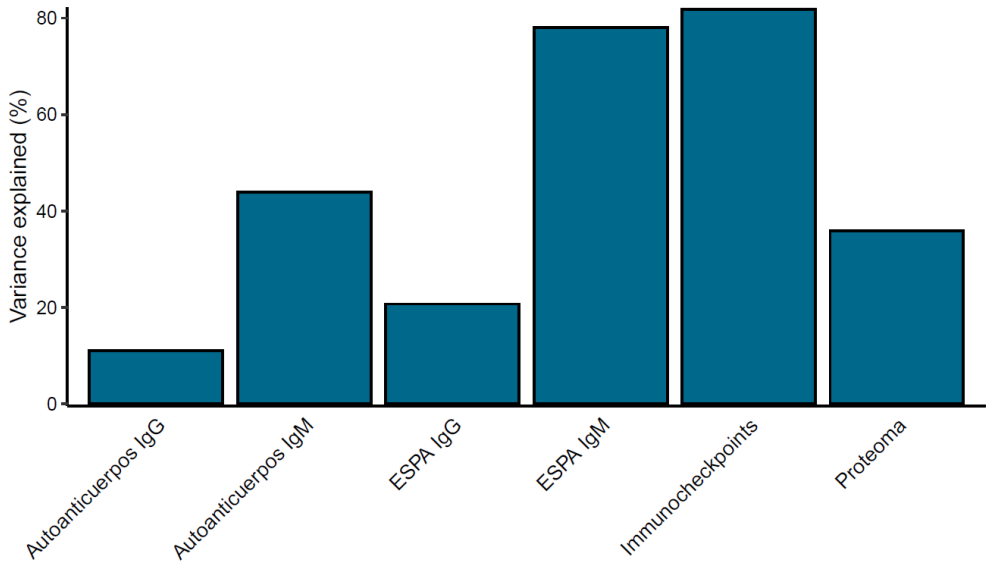
Si lo enfocamos desde el punto de vista biológico, el orden de aparición de las diferentes bases de datos puede proporcionar información importante sobre su relevancia en el contexto de la LLC. En primer lugar, aparecen los "Immunocheckpoints", esto puede sugerir que los puntos de control inmunitario se activan desde las fases más tempranas de la LLC para mantener el equilibrio del sistema inmunológico. A continuación, se expresan los "Autoanticuerpos IgM" seguidos de "ESPA IgM", lo que indica una respuesta autoinmune temprana en la enfermedad y hace sentido que se manifiesten antes que los IgG, ya que primero se produce la inmunoglobulina M. Posteriormente se comienza a expresar el proteoma, que va a indicar los cambios que se producen en la expresión de las proteínas. Por último, aparecen "ESPA IgG" y "Autoanticuerpos IgG".

En la Figura 12 encontramos un gráfico de barras que representa el porcentaje total de varianza explicada de cada una de las bases de datos por todos los factores. Observamos que

Autoanticuerpos IgG es la vista que menos se consigue explicar, con un 10.99%. Seguidamente se encuentra ESPA IgG con un 20.7%, Proteoma con un 35.85%, Autoanticuerpos IgM con un 43.84%, ESPA IgM con un 77.98% y, finalmente, la base de datos que más se consigue explicar mediante este modelo es Immunocheckpoints con un 81.77%.

Figura 12

Varianza explicada por todos los factores en cada vista



5.3 IDENTIFICACIÓN DE LAS CARACTERÍSTICAS MÁS IMPORTANTES

Una vez que se ha estudiado la descomposición de la varianza de todas las bases de datos, es de gran interés identificar cuáles son las características de cada vista que más influyen a la hora de la formación de los factores, es decir, cuáles son las características con mayor peso. Para poder visualizarlo de una forma más sencilla, se propone construir un mapa de calor (heatmap) a partir de las 10 características más influyentes de cada base de datos y factor.

En primer lugar, se crea un objeto de tipo lista "mofa_weights" aplicando la función "*get_weights*" al objeto MOFA "mofaobj". Los argumentos de esta función son:

- object: un objeto del tipo MOFA entrenado.
- views: un vector que puede ser numérico o de caracteres para indicar qué vistas se quieren extraer. Si se quieren extraer todas, se utiliza "all".
- factors: un vector que puede ser numérico o de caracteres para indicar qué factores se quieren extraer. Si se quieren extraer todos, se utiliza "all".
- abs: argumento lógico para seleccionar si los pesos se deben tomar en valor absoluto o no.

- scale: argumento lógico para determinar si se deben escalar los pesos de tal manera que el valor mínimo sea -1 y el máximo 1. En el caso de que se tomen valores absolutos, la escala sería de 0 a 1.
- as.data.frame: Argumento lógico para indicar si el objeto creado debe transformarse en un data frame o se mantiene como una lista.

Para este estudio se decidió emplear todos los factores y vistas, con los valores escalados y en formato lista, de tal manera que:

```
> mofa_weights <- get_weights(mofaobj, views = "all",
factors = "all", abs = FALSE, scale = TRUE, as.data.frame
= FALSE)
```

De la forma que hemos definido nuestros datos, los pesos se interpretarán de la siguiente manera:

- Valores próximos a 1: Fuerte asociación positiva
- Valores próximos a 0: Falta de asociación
- Valores próximos a -1: Fuerte asociación negativa

A continuación, se seleccionan para cada base de datos aquellos factores con los que se consiga explicar al menos el 1.5% de su varianza, tal y como se ha explicado en el apartado 5.2. Para cada una de las vistas en cada factor seleccionado, se extrae el *top 10* de características, es decir, las 10 características con los pesos en valor absoluto más altos (más próximos a 1). Por ejemplo, para la base de datos "Autoanticuerpos IgG" en el factor 1:

```
> top10_F1_AutoanticuerposIgG <-
names(sort(abs(mofa_weights$`Autoanticuerpos
IgG`[,"Factor1"]), decreasing = TRUE )[1:10])
```

Una vez que se haya implementado la función anterior para todas las bases de datos y factores, se procede a juntar toda la información en un único marco de datos con las siguientes columnas: en la primera se almacena el nombre de las características, en la segunda se determina a que base de datos pertenece la característica, en la tercera se especifica de qué factor se trata y en la última se proporciona el valor escalado de los pesos. El código de R utilizado para realizar este proceso y la tabla resultante se incluyen en los anexos.

Posteriormente, las columnas referentes a las características y a los factores latentes se convierten en factores mediante la función "*factor*" del paquete "*base*" (R Core Team, 2023) para poderlos ordenar del modo que queramos que aparezcan en el gráfico. En este paso también se hace uso del paquete "*dplyr*" (Wickham et al., 2023).

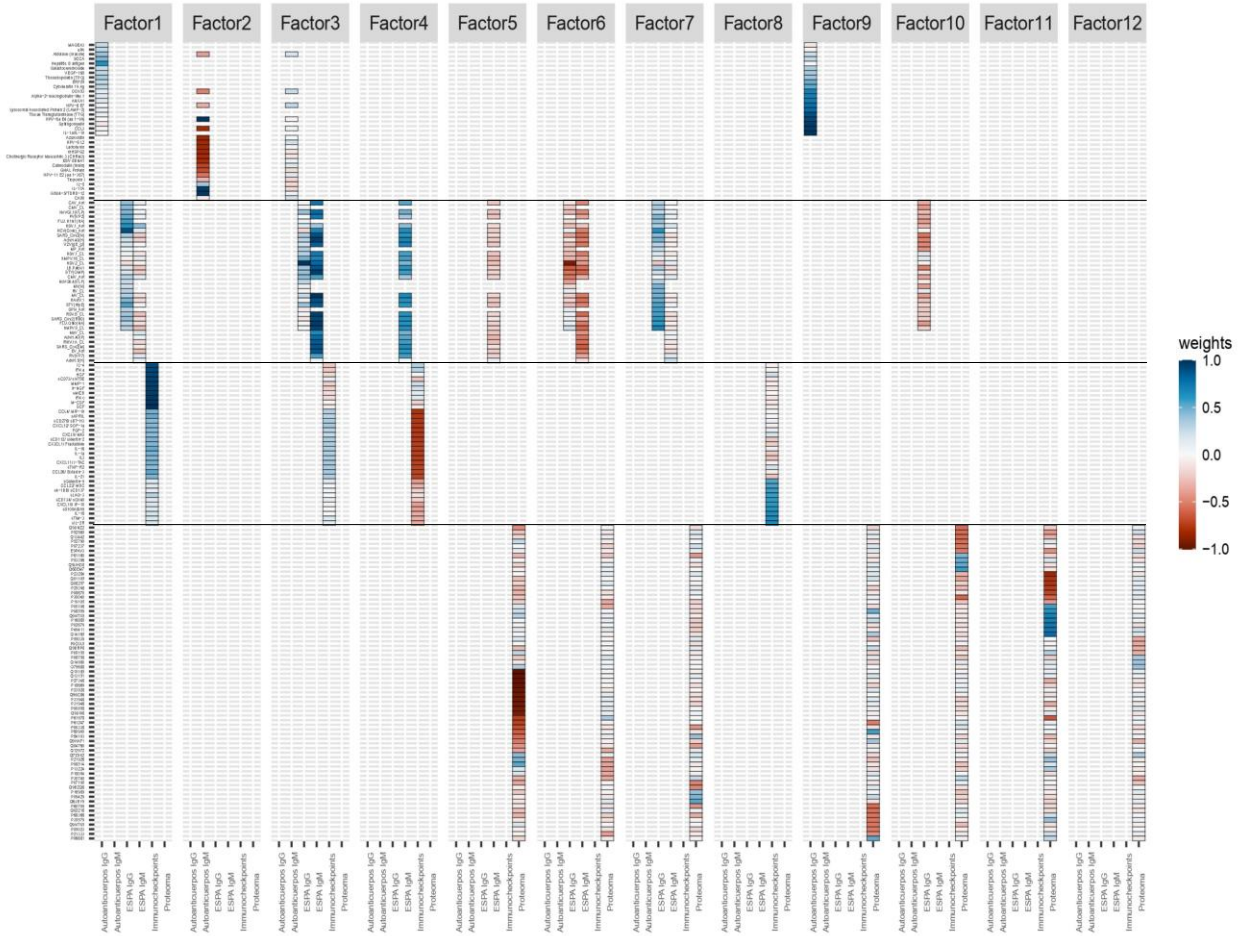
Por último, se procede a realizar el gráfico mediante la función "*ggplot*" y se utiliza el paquete "*colorspace*" (Stauffer et al., 2015) para definir la paleta de colores utilizada. El código para realizar el gráfico se encuentra en los anexos.

En la Figura 13 podemos observar el gráfico resultante que se interpreta de la siguiente manera:

- A la izquierda del gráfico encontramos las características extraídas de los respectivos *top 10* ordenadas por base de datos y separadas mediante las líneas horizontales: Arriba del todo tenemos las referentes a autoanticuerpos, en segundo lugar, se encuentran los antígenos microbianos, más abajo las proteínas de control inmunitario y, en último lugar, las proteínas del proteoma.
- Para cada uno de los factores, encontramos las bases de datos ordenadas de izquierda a derecha: “Autoanticuerpos IgG”, “Autoanticuerpos IgM”, “ESPA IgG”, “ESPA IgM”, “Immunocheckpoints”, “Proteoma”.
- Cada uno de los rectángulos representa, por tanto, una característica en cada factor y base de datos correspondiente. Estos rectángulos se colorean siguiendo el patrón de la leyenda

Figura 13

Heatmap de los pesos para los 12 factores



A modo de resumen, se recoge en la Tabla 6 el factor que mayor cantidad de varianza explica de cada vista y las 10 características que mayor peso han tenido a la hora de formarse dicho factor.

Tabla 6

Top 10 características de cada base de datos

Base de datos	Factor que explica más varianza	Top 10 características
Autoanticuerpos IgG	Factor 9	IL-1a/IL-1b, CCL3, Sphingomyelin, HPV-6a E4 (aa 1-99), Tissue Transglutaminase (TTG), Lysosomal Associated Protein 2 (LAMP-2), HPV-6 E7, ANXA1, Alpha-2-macroglobulin-like 1, DDX53
Autoanticuerpos IgM	Factor 2	HPV-6a E4 (aa 1-99), GBU4-5/TDRD-12, IL-17ª, Azurocidin, HPV-6 L2, Lactoferrin, rhHSPG2, CCL3, Cholinergic Receptor Muscarinic 3 (CHRM3), EBV EBNA1
ESPA IgG	Factor 7	hMPV.9_CL, HCV(Core), FLU.Ohio(NA), SARS_Cov2(RBD), RSV.B_CL, SPN_Ant, STY(HlyE), HAstV.1, MV_CL, RV_CL
ESPA IgM	Factor 3	MV_CL, FLU.Ohio(NA), STY(OMP), RSV.B_CL, AdvH.40(H), hMPV.9_CL, SARS_Cov2(RBD), SARS_Cov2(N), HAstV.1, EV_Ant
Immunocheckpoints	Factor 1	SCF, M-CSF, IFN γ, sMICB, b-NGF, MMP-1, sCD73/ sNT5E, HGF, IFN α, IL-4
Proteoma	Factor 5	Q15185, Q13151, P27348, P18669, P23528, Q96C86, P31948, P31946, P00558, Q9UI08

5.4 BÚSQUEDA DE NUEVOS GRUPOS DE PACIENTES

A partir de los datos extraídos del análisis MOFA, utilizando la función "*get_data*", se procede a realizar una búsqueda de nuevos grupos de pacientes. Para ello, se lleva a cabo un Agrupamiento por Consenso o *Consensus Clustering*. De esta manera se va a obtener una solución más fiable y representativa de los datos que aplicando un análisis de clúster tradicional.

Empleamos el paquete de RStudio "*ConsensusClusterPlus*" (Wilkerson & Hayes, 2010) para efectuar las agrupaciones de individuos y su posterior representación. La función utilizada también se llama "*ConsensusClusterPlus*" y sus argumentos más importantes son:

- d: Matriz de datos que se quieren agrupar, donde las filas representan las características y las columnas los individuos. El conjunto de datos empleado puede ser una matriz de observaciones por variables o una matriz de disimilaridades.
- maxK: número máximo de clústeres para ser evaluados
- reps: número de repeticiones
- clusterAlg: tipo de algoritmo de clustering a utilizar.
- distance: medida de similaridad o de distancia que se desea utilizar para realizar el agrupamiento
- seed: número aleatorio que sirve como semilla para que el análisis pueda ser reproducible.

Esta función no admite valores faltantes (NA's) en la matriz de datos, por lo que antes de ejecutarla, se sustituyeron estos valores por la media de la característica correspondiente. Las dos únicas bases de datos con este problema son Autoanticuerpos IgG Y Autoanticuerpos IgM. Teniendo en cuenta de que las características se encuentran en las filas y los pacientes en las columnas, por ejemplo, para Autoanticuerpos IgG, se ha utilizado el siguiente código:

```
> medias <- rowMeans(Autoanticuerpos_IgG, na.rm = TRUE)
> for (x in 1:122) {
  Autoanticuerpos_IgG [x, is.na(Autoanticuerpos_IgG [x,])]
  <-
  medias[x]
  }
```

Para cada una de las bases de datos utilizadas en este trabajo se empleó la función "*ConsensusClusterPlus*" utilizando un número máximo de clústeres igual a 7, 50 repeticiones, el algoritmo "DIANA", distancia de Pearson y una semilla con un valor aleatorio de 856814. Si continuamos utilizando la base de datos de Autoanticuerpos IgG como ejemplo, el código queda de la siguiente manera:

```
> ConsensusClusterPlus(Autoanticuerpos_IgG, maxK = 7,
reps = 50, clusterAlg = "diana_alg", distance =
"pearson",seed = 856814)
```

Para este estudio se decidió utilizar el algoritmo DIANA (DIvisive ANAlysis) ya que es muy recomendable cuando se dispone de una gran cantidad de datos y, al tratarse de un método jerárquico divisivo, permite realizar una exploración más exhaustiva sobre el modo de agrupamiento de los datos. Además, este proceso de ir dividiendo cada clúster en clústeres más pequeños ayuda a capturar mejor los patrones existentes en los datos. Para poder utilizarlo en el Agrupamiento por Consenso, se creó en RStudio una función "*diana_alg*":

```
> diana_alg <- function(mat, n){
  x <- diana(mat,diss=TRUE)
  a <- cutree(x, n)
  return(a) }
```

Esta función tiene dos argumentos de entrada: "mat" que se refiere a la matriz de entrada y "n" que indica la cantidad de clústeres deseados. En primer lugar, se calcula la estructura jerárquica con la función "*diana*" del paquete "*clúster*", donde se especifica que la matriz de entrada se corresponde con una matriz de disimilaridades, ya que esta función se va a realizar dentro de "*ConsensusClusterPlus*" y recibe la matriz con las distancias calculadas por el método seleccionado. Después con la función "*cutree*" se realiza el corte del dendrograma en "n" clústeres. Por último, la función devuelve la asignación de cada individuo al clúster correspondiente.

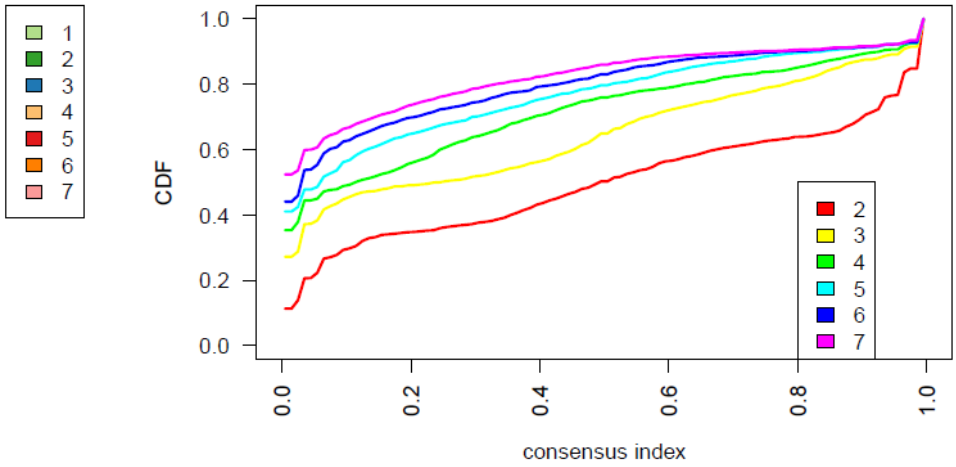
Aplicando lo anterior, se obtienen una serie de gráficos sobre cómo quedaría la división de individuos seleccionando de 2 a 7 grupos y otros para ayudar a tomar una decisión del número óptimo de clústeres a seleccionar (ANEXO). En este trabajo únicamente se muestra el gráfico que se ha tomado de referencia para elegir el número óptimo de grupos (*k*).

En la Figura 14 se puede observar un gráfico denominado "Consensus CDF" ("Cumulative Distribution Function") que muestra la función de distribución de consenso acumulada. El eje Y representa la probabilidad acumulada de que una observación sea asignada al mismo clúster en repetidas ocasiones. El eje X muestra los índices de consenso, que se utilizan para medir la calidad de los clústeres obtenidos. Si el índice de consenso toma valores cercanos a 0, quiere decir que existe mucha variabilidad en las asignaciones de los clústeres, mientras que valores cercanos a 1 denotan una consistencia alta en las asignaciones.

A partir de este gráfico y observando la trayectoria de las diferentes curvas, se tomó la decisión de realizar 5 clústeres. Es decir, las curvas correspondientes a 2, 3 y 4 grupos presentan trayectorias que son muy diferentes entre sí, por lo que la calidad del agrupamiento no es óptima. Sin embargo, al seleccionar 5 grupos se puede observar como la trayectoria de la curva correspondiente pasa a ser más estable y conforme se va aumentando el número de grupos no se observan mejoras significativas en el agrupamiento de los datos.

Figura 14

Función de Distribución Acumulativa de Consenso para Autoanticuerpos IgG



El proceso de selección del número de clústeres se realizó de manera análoga para el resto de las bases de datos obteniendo como resultado final el número óptimo de grupos indicado en la Tabla 7

Tabla 7

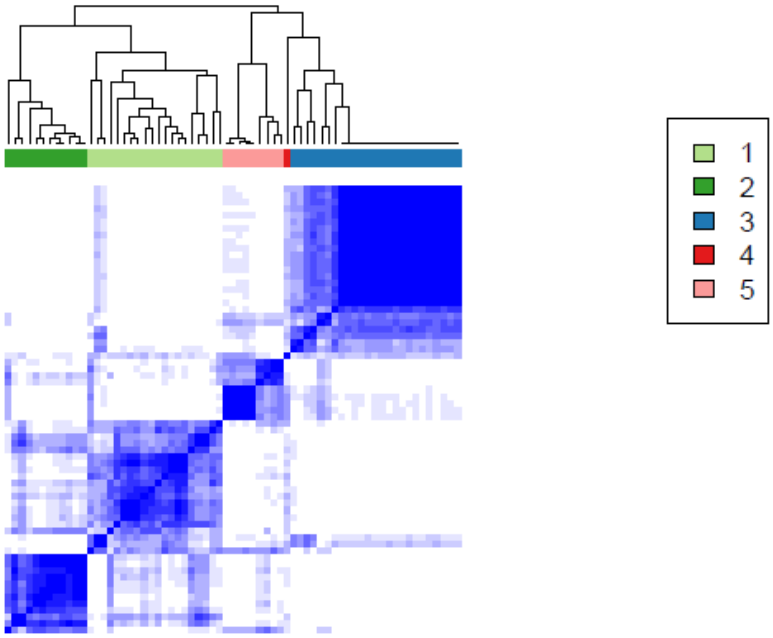
Número óptimo de grupos para cada base de datos

Base de datos	Número de grupos
Autoanticuerpos IgG	5
Autoanticuerpos IgM	6
ESPA IgG	6
ESPA IgM	5
Immunocheckpoints	5
Proteoma	6

En la Figura 15 se muestra la matriz de consenso para un número de grupos igual a 5 con su respectivo dendrograma. Como podemos observar, todos los grupos poseen varios individuos, excepto el grupo 4 que solo contiene uno. Sería interesante estudiar la naturaleza de este individuo para poder averiguar por qué se encuentra en un clúster aparte.

Figura 15

Heatmap de la matriz de consenso sobre Autoanticuerpos IgG para k = 5



En la siguiente tabla (Tabla 8) se muestra cada uno de los individuos con el clúster que se le ha asignado. Como podemos comprobar, el paciente que se encuentra solo en un clúster es LLC-23 y se corresponde con un hombre 60 años con LLC que ha recibido la primera línea de tratamiento y posee la deleción 13q. Las tablas para las demás bases de datos se incluyen en los anexos.

Tabla 8

Individuos y clúster asignado para Autoanticuerpos IgG

ID del paciente	Clúster	ID del paciente	Clúster	ID del paciente	Clúster
LLC-1	1	LLC-38	2	LLC-6	1
LLC-10	1	LLC-39	2	LLC-61	3
LLC-11	1	LLC-4	2	LLC-62	3
LLC-17	2	LLC-40	2	LLC-63	3
LLC-18	1	LLC-41	2	LLC-64	3
LLC-19	1	LLC-42	2	LLC-65	3
LLC-20	3	LLC-43	2	LLC-66	3
LLC-21	2	LLC-44	2	LLC-67	3
LLC-22	2	LLC-45	2	LLC-68	3
LLC-23	4	LLC-46	5	LLC-69	5

LLC-24	2	LLC-47-1	2	LLC-7	1
LLC-25	2	LLC-48	2	LLC-70	5
LLC-26	2	LLC-49	2	LLC-71	5
LLC-28	3	LLC-5	1	LLC-72	5
LLC-29	2	LLC-50	2	LLC-74	5
LLC-3	2	LLC-52	2	LLC-76	5
LLC-30	3	LLC-53	2	LLC-77	5
LLC-32	2	LLC-54	3	LLC-78	1
LLC-33	5	LLC-55	3	LLC-79	5
LLC-34	3	LLC-56	3	LLC-8	1
LLC-35	2	LLC-57	3	LLC-9	1
LLC-36	3	LLC-58	3		
LLC-37	5	LLC-59	3		

6 DISCUSIÓN Y CONCLUSIONES

La Leucemia Linfática Crónica (LLC) es un tipo de cáncer de la sangre que presenta un curso clínico muy heterogéneo y afecta a una gran parte de la población adulta occidental. Debido a la falta de información que se tiene sobre este tipo de leucemia, se propuso realizar un estudio desde el punto de vista proteómico sobre la respuesta inmune humoral en 57 pacientes de LLC y 10 pacientes de su estadio anterior (LMB). Para ello se tuvieron en cuenta datos relacionados con autoanticuerpos IgG e IgM, antígenos IgG e IgM, puntos de control inmunitario y proteínas en general.

Si tomamos como referencia los objetivos definidos para este estudio en el apartado 3, se pueden extraer las siguientes conclusiones:

- El objetivo principal consistía en conseguir integrar todos estos datos en un único análisis de tal manera que se redujese la dimensión para poder interpretar mejor los resultados, pero sin perder una gran cantidad de información. Mediante la técnica MOFA (Multi-Omics Factor Analysis) se ha logrado crear un modelo de 12 factores que explica la siguiente cantidad de información de cada base de datos: un 10.99% de Autoanticuerpos IgG, un 43.84% de Autoanticuerpos IgM, un 20.7% de ESPA IgG, un 77.98% de ESPA IgM, un 81.77% de Immunocheckpoints y un 35.85% de Proteoma. Si tenemos en cuenta la dimensión de cada base de datos, de manera global se consigue explicar aproximadamente un 37% de la información total con tan solo 12 factores.
- Para poder identificar posibles biomarcadores, se han seleccionado para cada una de las vistas aquellos factores con los que se lograba explicar al menos el 1.5% de su varianza. Mediante un mapa de calor (*heatmap*) se ha conseguido representar de manera conjunta en un único gráfico aquellas características que habían tenido una mayor influencia para la creación de dichos factores. De este modo, se representan mediante tonos azules las características que tienen una relación positiva con el factor y mediante tonos rojos las que tienen una relación negativa con el factor.
- Por último, con estos nuevos factores o variables latentes se ha podido obtener una nueva clasificación de los individuos con referencia a la respuesta inmune. Para ello se ha empleado el Agrupamiento por Consenso, ya que la solución obtenida es más robusta y fiable que si utilizamos un *Clustering* tradicional. Utilizando el algoritmo jerárquico divisivo DIANA, la distancia de Pearson y estudiando la Función de Distribución Acumulativa de Consenso mediante el gráfico “Consensus CDF”, se ha establecido el siguiente número óptimo de grupos para cada base de datos: 5 para Autoanticuerpos IgG, 6 para Autoanticuerpos IgM, 6 para ESPA IgG, 5 para ESPA IgM, 5 para Immunocheckpoints y 6 para Proteoma.

El actual trabajo se considera un ensayo piloto con un tamaño reducido de pacientes. Debido a las limitaciones y a la variabilidad relacionadas con el tamaño de la muestra, se debe tener en cuenta que los resultados obtenidos pueden estar sesgados.

Por ello, se propone como perspectiva de futuro validar este estudio mediante una muestra de pacientes más grande y, de este modo, poder evaluar la reproducibilidad de los resultados obtenidos y comprobar si dichos resultados se pueden extrapolar a una cohorte más amplia de pacientes.

Además, al realizar este estudio utilizando un número superior de pacientes, se conseguiría obtener una visión más completa y representativa sobre las características de los individuos con LLC o LMB. De esta manera se lograría aumentar la relevancia tanto estadística como médica de los resultados y proporcionar una base más sólida para investigaciones futuras y el desarrollo de nuevas terapias.

7 BIBLIOGRAFÍA

Abbas, A. K., Lichtman, A. H., & Pillai, S. (2017). *Cellular and Molecular Immunology* (9.ª ed.). https://evolve.elsevier.com/cs/product/9780323479783?role=student

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-Omics Factor Analysis—A framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, *14*(6), e8124. https://doi.org/10.15252/msb.20178124

Bagacean, C., Le Dantec, C., Berthou, C., Tempescul, A., Saad, H., Bordron, A., Zdrenghea, M., Cristea, V., Douet-Guilbert, N., & Renaudineau, Y. (2017). Combining cytogenetic and epigenetic approaches in chronic lymphocytic leukemia improves prognosis prediction for patients with isolated 13q deletion. *Clinical Epigenetics*, *9*, 122. https://doi.org/10.1186/s13148-017-0422-7

Briones, J. (2022). Precision medicine in chronic lymphatic leukemia: Cost-effectiveness analysis of the new targeted therapies. *Farmacia Hospitalaria*, *46*(3), 103-104. https://doi.org/10.7399/fh.13280

Caridad y Ocerin, J. M. (2016). *Econometría: Modelos econométricos y series temporales. Tomo 1*. https://www.google.es/books/edition/Econometr%C3%ADa_modelos_econom%C3%A9tricos_y_se/zpHmDwAAQBAJ?hl=es&gbpv=0&bshm=nce/1

Carrera Aguado, N. (s. f.). *Espectrometría de masas*. Laboratorio de Técnicas Instrumentales UVa. Recuperado 5 de mayo de 2023, de https://laboratoriotecnicasinstrumentales.es/analisis-qumicos/espectrometra-de-masas

Centro de Apoyo a la Investigación (CAI). (s. f.). *Técnica LC-MS/MS*. Recuperado 5 de mayo de 2023, de https://cai.ucm.es/tecnicas-biologicas/proteomica/tecnicas/lc-msms/123/

Centro de Apoyo a la Investigación (CAI). (2018, octubre 4). *Unidad de Proteómica*. Universidad Complutense de Madrid. https://www.google.com/maps/d/viewer?mid=1QzZG5nS5OqvYJDGkWroxAI-VAz6drOhy

Chiorazzi, N., Chen, S.-S., & Rai, K. R. (2021). Chronic Lymphocytic Leukemia. *Cold Spring Harbor Perspectives in Medicine*, *11*(2), a035220. https://doi.org/10.1101/cshperspect.a035220

Cuadras, C. M. (2014). *Nuevos métodos de analisis multivariante* [Recurso electrónico]. El autor.

de la Fuente Fernández, S. (2011). *ANÁLISIS FACTORIAL*.

De la Torre Gómez, C. (2012). *Aplicación de técnicas de proteómica para el estudio de enfermedades neuromusculares*. Universitat de Barcelona.

Delgado, J., Doubek, M., Baumann, T., Kotaskova, J., Molica, S., Mozas, P., Rivas-Delgado, A., Morabito, F., Pospisilova, S., & Montserrat, E. (2017). Chronic lymphocytic leukemia: A prognostic model comprising only two biomarkers (IGHV mutational status and FISH cytogenetics) separates patients with different outcome and simplifies the CLL-IPI. *American Journal of Hematology*, *92*(4), 375-380. https://doi.org/10.1002/ajh.24660

Galigalidou, C., Zaragoza-Infante, L., Iatrou, A., Chatzidimitriou, A., Stamatopoulos, K., & Agathangelidis, A. (2021). Understanding Monoclonal B Cell Lymphocytosis: An Interplay of Genetic and Microenvironmental Factors. *Frontiers in Oncology*, *11*. https://www.frontiersin.org/articles/10.3389/fonc.2021.769612

Gehlenborg, N., & Wong, B. (2012). Heat maps. *Nature Methods*, *9*(3), Article 3. https://doi.org/10.1038/nmeth.1902

Georgiadis, P., Liampa, I., Hebels, D. G., Krauskopf, J., Chatziioannou, A., Valavanis, I., de Kok, T. M. C. M., Kleinjans, J. C. S., Bergdahl, I. A., Melin, B., Spaeth, F., Palli, D., Vermeulen, R. C. H., Vlaanderen, J., Chadeau-Hyam, M., Vineis, P., Kyrtopoulos, S. A., Gottschalk, R., van Leeuwen, D., … on behalf of the EnviroGenomarkers consortium. (2017). Evolving DNA methylation and gene expression markers of B-cell chronic lymphocytic leukemia are present in pre-diagnostic blood samples more than 10 years prior to diagnosis. *BMC Genomics*, *18*(1), 728. https://doi.org/10.1186/s12864-017-4117-4

Giordani, P., Ferraro, M. B., & Martella, F. (2020). Introduction to Clustering. En P. Giordani, M. B. Ferraro, & F. Martella (Eds.), *An Introduction to Clustering with R* (pp. 3-5). Springer. https://doi.org/10.1007/978-981-13-0553-5_1

Goder, A., & Filkov, V. (2008). Consensus Clustering Algorithms: Comparison and Refinement. En *2008 Proceedings of the Workshop on Algorithm Engineering and Experiments (ALENEX)* (pp. 109-117). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972887.11

Griffen, T. L., Hoff, F. W., Qiu, Y., Lillard, J. W., Ferrajoli, A., Thompson, P., Toro, E., Ruiz, K., Burger, J., Wierda, W., & Kornblau, S. M. (2022). Proteomic profiling based classification of CLL provides prognostication for modern therapy and identifies novel therapeutic targets. *Blood Cancer Journal*, *12*(3), Article 3. https://doi.org/10.1038/s41408-022-00623-7

Herbst, S. A., Vesterlund, M., Helmboldt, A. J., Jafari, R., Siavelis, I., Stahl, M., Schitter, E. C., Liebers, N., Brinkmann, B. J., Czernilofsky, F., Roider, T., Bruch, P.-M., Iskar, M., Kittai, A., Huang, Y., Lu, J., Richter, S., Mermelekas, G., Umer, H. M., … Dietrich, S. (2022). Proteogenomics refines the molecular classification of chronic lymphocytic leukemia. *Nature Communications*, *13*(1), 6226. https://doi.org/10.1038/s41467-022-33385-8

Horgan, R. P., & Kenny, L. C. (2011). 'Omic' technologies: Genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, *13*(3), 189-195. https://doi.org/10.1576/toag.13.3.189.27672

Illowsky, B., College, D., Dean, S., & College, D. (2022). *Introducción a la estadística*.

Jr, C. A. J., Travers, P., Walport, M., Shlomchik, M. J., Jr, C. A. J., Travers, P., Walport, M., & Shlomchik, M. J. (2001). *Immunobiology* (5th ed.). Garland Science.

Kalinowski, T., Ushey, K., Allaire, J. J., RStudio, Tang [aut, Y., cph, Eddelbuettel, D., Lewis, B., Keydana, S., Hafen, R., library, M. G. (TinyThread, & http://tinythreadpp.bitsnbites.eu/). (2023). *reticulate: Interface to «Python»* (1.30). https://cran.r-project.org/web/packages/reticulate/index.html

Kaufman, L., & Rousseeuw, P. J. (1990). Divisive Analysis (Program DIANA). En *Finding Groups in Data* (pp. 253-279). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316801.ch6

Key, M. (2012). A tutorial in displaying mass spectrometry-based proteomic data using heat maps. *BMC Bioinformatics*, *13*(16), S10. https://doi.org/10.1186/1471-2105-13-S16-S10

Landeira-Viñuela, A., Alcoceba-Sanchez, M., Navarro-Bailón, A., Arias-Hidalgo, C., Juanes-Velasco, P., Sánchez-Santos, J. M., Lecrevisse, Q., Pedreira, C. E., García-Vaquero, M. L., Hernández, Á.-P., Montalvillo, E., Góngora, R., De las Rivas, J., González-Díaz, M., Orfao, A., & Fuentes, M. (2023). Systematic Evaluation of Antigenic Stimulation in Chronic Lymphocytic Leukemia: Humoral Immunity as Biomarkers for Disease Evolution. *Cancers*, *15*(3), Article 3. https://doi.org/10.3390/cancers15030891

Landeira-Viñuela, A., Arias-Hidalgo, C., Juanes-Velasco, P., Alcoceba, M., Navarro-Bailón, A., Pedreira, C. E., Lecrevisse, Q., Díaz-Muñoz, L., Sánchez-Santos, J. M., Hernández, Á.-P., García-Vaquero, M. L., Góngora, R., De Las Rivas, J., González, M., Orfao, A., & Fuentes, M. (2022). Unravelling soluble immune checkpoints in chronic lymphocytic leukemia: Physiological immunomodulators or immune dysfunction. *Frontiers in Immunology*, *13*. https://www.frontiersin.org/articles/10.3389/fimmu.2022.965905

Lesher Gordillo, J. M., & Tovilla Zárate, C. A. (2013). *Introducción a la genómica*.

López Arias, E. A. (2013). *BIOMARCADORES PARA EL DIAGNÓSTICO TEMPRANO DE CÁNCER DE MAMA; BÚSQUEDA E IDENTIFICACIÓN*.

Madhulatha, T. S. (2012). *An Overview on Clustering Methods* (arXiv:1205.1117). arXiv. https://doi.org/10.48550/arXiv.1205.1117

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, *52*, 91-118. https://doi.org/10.1023/A:1023949509487

Mowery, Y. M., & Lanasa, M. C. (2012). Clinical Aspects of Monoclonal B-Cell Lymphocytosis. *Cancer Control: Journal of the Moffitt Cancer Center*, *19*(1), 8-17. https://doi.org/10.1177/107327481201900102

National Human Genome Research Institute. (2023, mayo 2). *Genómica*. Genome.gov. https://www.genome.gov/es/genetics-glossary/Genomica

Nguyen, N., & Caruana, R. (2007). *Consensus Clusterings*. 607-612. https://doi.org/10.1109/ICDM.2007.73

Punt, J., Stanford, S. A., Jones, P. P., & Owen, J. A. (2018). *Kuby Immunology* (8.ª ed.).

R Core Team. (2023). *R: The R Project for Statistical Computing*. https://www.r-project.org/

Reducción de la dimensión: Análisis factorial. (2016). https://www.uv.es/mlejarza/actuariales/tam/afact.pdf

Rossum, G. V., & Drake, F. L. (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

San Miguel-Hernández, Á., Martín-Gil, F. J., & Armentia-Medina, A. (2009). Metodología y aplicaciones en proteómica clínica. *Diálisis y Trasplante*, *30*(4), 139-143. https://doi.org/10.1016/S1886-2845(09)72698-0

Stauffer, R., Mayr, G. J., Dabernig, M., & Zeileis, A. (2015). Somewhere Over the Rainbow: How to Make Effective Use of Colors in Meteorological Visualizations. *Bulletin of the American Meteorological Society*, *96*(2), 203-216. https://doi.org/10.1175/BAMS-D-13-00155.1

Sun, L., Wang, X., Saredy, J., Yuan, Z., Yang, X., & Wang, H. (2020). Innate-adaptive immunity interplay and redox regulation in immune response. *Redox Biology*, *37*, 101759. https://doi.org/10.1016/j.redox.2020.101759

Tortosa Viqueira, M., Cartea González, M. E., Abilleira Ambroa, R., & Velasco Pazos, P. (2017). *Técnicas de análisis masivo para el estudio de factores de resistencia a enfermedades en cultivos de brásicas*. https://digital.csic.es/handle/10261/157943

Vargas Sabadías, A. (1995). *Estadística descriptiva e inferencial*.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer International Publishing.

Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D., Software, P., & PBC. (2023). *dplyr: A Grammar of Data Manipulation* (1.1.2). https://cran.r-project.org/web/packages/dplyr/index.html

Wilkerson, M. D., & Hayes, D. N. (2010). ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics*, *26*(12), 1572-1573. https://doi.org/10.1093/bioinformatics/btq170

Yao, H., Xu, H., Qiu, S., Chen, J., Lin, Z., Zhu, J., Sun, X., Gao, Q., Chen, X., Xi, C., Huang, D., Zhang, F., Gao, S., Wang, Z., Zhang, J., Liu, X., Ren, G., Tao, X., Li, M., & Chen, W. (2022). Choline deficiency-related multi-omics characteristics are susceptible factors for chemotherapy-induced thrombocytopenia. *Pharmacological Research*, *178*, 106155. https://doi.org/10.1016/j.phrs.2022.106155

Yuen, G. J., Demissie, E., & Pillai, S. (2016). B lymphocytes and cancer: A love-hate relationship. *Trends in cancer*, *2*(12), 747-757. https://doi.org/10.1016/j.trecan.2016.10.010