Universidad Europea

Máster en Bioinformática

Estratificación molecular de Enfermedades Inflamatorias Intestinales para la medicina personalizada

Autor: Adrián Miguel Pizarro

Tutor: Daniel Toro Domínguez

Curso 2022-23

ÍNDICE

AGRADECIMIENTOS

Quisiera aprovechar esta ocasión para dar las gracias a las personas que me han acompañado en esta etapa que comenzó hace 8 años, cuando pisé por primera vez la universidad.

En primer lugar, a mis amigos por todos los buenos momentos, los recuerdos y la comprensión cuando no podía salir.

A Rubén, el amigo más fiel que he conocido, y, estoy seguro, que conoceré.

A mi familia, por la ayuda y el amor incondicional que siempre he recibido, y por lo que estaré eternamente agradecido.

A Andrea, por el ánimo, el cariño y el infinito apoyo cuando ni yo lo tenía.

Gracias, de todo corazón, por acompañarme en este largo camino.

Para todos vosotros es este proyecto.

RESUMEN

Objetivos:

- Realizar una búsqueda de múltiples estudios públicos con muestras de sangre de pacientes con la enfermedad de Crohn y colitis ulcerosa, con el objetivo de identificar aquellos patrones moleculares detrás de la heterogeneidad de estas enfermedades. Se definirán subtipos moleculares dentro de las enfermedades.
- Analizar la asociación entre los subtipos moleculares de estas enfermedades y diferentes variables clínicas. Utilizar la información molecular para obtener marcadores predictivos para respuesta a diferentes manifestaciones clínicas y/o fármacos.

Material y métodos: Primero se realizó la obtención de datos de expresión en muestras de sangre de pacientes con Enfermedad Inflamatoria Intestinal a través de los principales repositorios públicos de datos transcriptómicos. Una vez obtenidos, se organizan, depuran y procesan con el objetivo de facilitar su posterior utilización en métodos analíticos. Finalmente se realiza un cálculo de puntuación en la expresión respecto a controles sanos (M-scores) para la búsqueda posterior de patrones moleculares y de subgrupos de la enfermedad mediante técnicas de *clustering* y *machine learning*, con el objetivo de desarrollar modelos predictivos para variables clínicas.

Resultados: Se describieron tres firmas transcripcionales de Enfermedad Inflamatoria Intestinal, que se dividen en inflamación, procesos relacionados con linfocitos, y coagulación. Se detalla una clara co-expresión positiva entre el clúster de inflamación y el clúster de coagulación, y una co-expresión negativa con el grupo de linfocitos. Estos hallazgos se relacionan muy bien con la presencia de enfermedad activa e inactiva, y con el tipo de enfermedad en función de los subgrupos de pacientes, donde se ha demostrado una relación significativa entre estas variables y el subtipo de enfermos. Por otro lado, se ha desarrollado un modelo predictivo con muy alto rendimiento para el diagnóstico de personas enfermas con Enfermedad Inflamatoria Intestinal a través de muestras de sangre. Por último, se han estudiado modelos para predecir diferentes variables clínicas, con resultados prometedores.

Conclusiones: El cálculo de M-scores ha resultado muy eficaz para estratificar los pacientes de Enfermedad Inflamatoria Intestinal e identificar grupos de co-expresión génica en módulos relacionados con patologías autoinmunes. Además, se ha utilizado para desarrollar un robusto modelo para predecir personas enfermas con los datos transcriptómicos obtenidos en una extracción rutinaria de sangre.

Palabras clave: Enfermedad inflamatoria intestinal, IBD, enfermedad de Crohn, colitis ulcerosa, *machine learning*, transcriptómica, inflamación.

ABSTRACT

Objectives:

- Conduct a search for several public studies that use blood samples from patients with Crohn's disease and ulcerative colitis, to recognise molecular patterns underlying the heterogeneity of these diseases. Molecular subtypes within the diseases will be defined.
- Analyse the correlation between the molecular subtypes of these diseases and diverse clinical factors, in order to obtain predictive markers for response to different clinical manifestations and/or drugs using molecular information.

Materials and Methods: First, expression data from blood samples of patients suffering from Inflammatory Bowel Disease were retrieved by utilizing publicly available transcriptomic data repositories. Subsequently, the data were organized, refined, and processed to facilitate their utilization in analytical methods. Finally, M-scores were calculated by comparing expression levels to those of healthy controls. These scores were then utilized to search for molecular patterns and disease subgroups through clustering and machine learning techniques. The ultimate intention is to build predictive models for clinical variables.

Results: Three transcriptional signatures of Inflammatory Bowel Disease were identified, divided into inflammation, lymphocyte-related processes, and coagulation. A clear positive co-expression between the inflammation cluster and the coagulation cluster was identified, along with negative co-expression with the lymphocyte group. The results show a strong correlation between active and inactive forms of the disease and disease type based on patient subgroups, indicating a significant relationship between these variables and the disease subtype.Moreover, a high-performing predictive model was developed to diagnose individuals with Inflammatory Bowel Disease from blood samples. Finally, machine learning models yielded promising results in predicting various clinical variables.

Conclusions: The M-score calculation has proven to be highly effective in stratifying patients with Inflammatory Bowel Disease and identifying gene co-expression groups in modules related to autoimmune diseases. Additionally, a robust model for predicting individuals with the disease has been developed using transcriptomic data obtained from routine blood extraction.

Keywords: Inflammatory Bowel Disease, IBD, Crohn's disease, ulcerative colitis, machine learning, transcriptomics, inflammation.

1. Introducción

Las enfermedades inflamatorias intestinales (IBD por sus siglas en inglés) son patologías crónicas que producen inflamación en una o varias regiones del tracto digestivo, predominantemente en el colon e íleon. Las IBD se pueden dividir principalmente en dos patologías diferenciadas, la enfermedad de Crohn (CD por sus siglas en inglés) y la colitis ulcerosa (UC por sus siglas en inglés). Ambas enfermedades son recurrentes y heterogéneas, tanto en su aparición como en su desarrollo. La UC se caracteriza por afectar solamente al recto y al colon, tanto de manera local como extendiéndose en un continuo por toda la pared del órgano, con una grave inflamación de la mucosa que puede producir dolor abdominal, diarrea, sangrado en las heces, erosiones y úlceras. Por otro lado, la CD ocurre también principalmente en el colón e íleon en aproximadamente 2/3 de los casos, mientras que el tercio restante ocurre en algún punto del intestino delgado. La afectación a la parte superior del tracto digestivo (desde la boca al yeyuno) se estima que ocurre solamente del 0.3-5% de los casos. Los pacientes con CD suelen presentar parches de zonas afectadas intercaladas con regiones de tejido normal. Respecto a la sintomatología, se destaca el dolor abdominal y la diarrea, aunque se puede acabar desarrollando manifestaciones extraintestinales a medida que cursa la enfermedad (Guan, 2019; Medina, 2013; Ordás et al., 2012; Roda et al., 2020).

Aunque las IBD fueron señaladas por su sintomatología en la Antigua Grecia, no fue hasta 1875, que gracias al avance en la medicina se pudo identificar y definir la enfermedad de la colitis ulcerosa. Posteriormente, en el año 1932, se identificó y reconoció a la enfermedad de Crohn como una patología diferente a UC. A pesar de que estas enfermedades son conocidas desde hace mucho tiempo, su etiología es heterogénea y desconocida, y su incidencia se encuentra en progresivo aumento (Kaplan, 2015; Mulder et al., 2014).

La inflamación es la seña de identidad de las IBD, y, de hecho, los principales biomarcadores usados en la práctica clínica son indicadores que muestran algún proceso inflamatorio en el cuerpo. La respuesta inflamatoria es un proceso autolimitante en personas sanas, que se inicia cuando las células inmunes detectan patrones moleculares asociados a patógenos (PAMPs) o asociados a daños (DAMPs) a través de sus receptores

de reconocimientos de patrones (PRR). Una vez reconocidos, se liberan sustancias proinflamatorias que tienen efecto paracrino y autocrino, dando lugar a la activación de más células inmunitarias y a una cascada de respuesta inmune, que afecta a múltiples tipos celulares, tejidos y órganos. El proceso inflamatorio no termina solamente con la eliminación del agente causante o con la reparación del tejido, sino con la reprogramación funcional de las células implicadas a través de múltiples mecanismos para restaurar la homeostasis. La aparición de una inflamación crónica en el sistema gastrointestinal proviene por la incapacidad de los mecanismos inmunes de finalizar el proceso de inflamación, y su causa es incierta, pero parece estar relacionada con personas genéticamente susceptibles a una respuesta inmune anormal frente a los microorganismos comensales (Netea et al., 2017; Saez et al., 2023; Schett & Neurath, 2018).

Aunque su causa no sea comprendida, recientemente se han descubierto múltiples y diversas asociaciones en la patogenia de las IBD además de la susceptibilidad genética, tales como variaciones en el microbiota intestinal, trastornos emocionales, baja ingesta de vitamina D, uso de fármacos como antibióticos, AINEs y pastillas anticonceptivas, o el tabaquismo (Alatab et al., 2020; Roda et al., 2020). Por tanto, se considera que las IBD tienen una patogénesis muy compleja, con interacciones entre factores genéticos, ambientales, microbiológicos e inmunes. Además, el incremento en la prevalencia y en los nuevos casos puede considerarse alarmante. Un estudio de *Global Burden of Disease* (GDB) realizado por la Universidad de Washington señala que la estimación de personas que padecen IBD ha aumentado de 3.32 millones de personas en 1990 a 6.8 millones en 2017, siendo Norteamérica, Norte Europa y Asia oriental las regiones con mayor prevalencia (Alatab et al., 2020; Guan, 2019; Roda et al., 2020).

Una de las principales limitaciones en cuanto al diagnóstico de la enfermedad se encuentra en las pruebas metodológicas, en la que la ileocolonoscopia es el "*gold standard*". Es un procedimiento invasivo y hospitalario, y normalmente transcurre un largo tiempo desde el inicio de los síntomas, la sospecha de la enfermedad y la prueba diagnóstica. Además, para el caso de CD, puede ser que la patología curse en el intestino delgado, por lo que sería imposible de detectar con la colonoscopia y se requeriría una

entero-resonancia magnética o una endoscopia mediante video-cápsula para la identificación, lo que retrasaría aún más el diagnóstico (Alghoul et al., 2022). Un diagnóstico temprano es crucial para evitar un posible daño intestinal irreversible que reduzca aún más la calidad de vida de los pacientes, ya que cada vez hay más evidencia de que si la IBD es detectada y tratada de manera precoz mediante inmunosupresión, se reduce significativamente las tasas de recidivas y la necesidad de tratamiento quirúrgico (Roda et al., 2020; Ungaro et al., 2021).

Se han establecido diferentes tipos de biomarcadores para ayudar a identificar la gravedad de la patología en pacientes con IBD. Estas técnicas son más accesibles que la ileocolonoscopia al tener menor coste y, sobre todo, menos invasivas, optimizando la elección de pacientes que requerirán diagnóstico por imagen y extracción de biopsias. En la práctica clínica se utilizan principalmente 3 de estos biomarcadores: Proteína C-Reactiva (CRP por sus siglas en inglés), Calprotectina Fecal (FC por sus siglas en inglés), y Lactoferrina Fecal (FL por sus siglas en inglés). Todos ellos son útiles para acelerar el diagnóstico de IBD, sin embargo, tienen sus limitaciones. La CRP solamente ayuda a identificar inflamación aguda en pacientes con CD, no mostrando cambio significativo en pacientes con UC. La FC es el biomarcador más usado, pero la inflamación detectada no es exclusiva de IBD, su alteración puede provenir de otra patología digestiva, como alergias, infecciones o parasitosis. Asimismo, todos ellos muestran inconvenientes respecto a la precisión de su detección. Se ha estudiado sus coeficientes de correlación entre los valores obtenidos por el biomarcador y los datos obtenidos tras la colonoscopia, obteniendo que la FC sufre una variación de 0.48 a 0.83 y la FL de 0.19 a 0.87 (Alghoul et al., 2022).

El uso de biomarcadores es altamente útil como comienzo en el proceso de diagnóstico de IBD, pero es necesario encontrar otras formas de identificar más eficazmente, tanto la presencia de enfermedad, como el subtipo de ésta y su perfil molecular, de manera que se pueda iniciar un tratamiento óptimo y personalizado de la forma más temprana posible. Se han realizado algunos estudios con análisis transcriptómicos en muestras de sangre de pacientes con IBD, mostrando resultados muy prometedores, incluso comparando con la tasa de respuesta de algunos fármacos como anti-TNF. Aunque los

resultados son halagüeños, una limitación de estos estudios es el tamaño muestral y la necesidad de validar los resultados con una población de mayor tamaño (Biasci et al., 2019; Nowak, Szymańska, et al., 2022; Park et al., 2022; Ungaro et al., 2021).

Con la motivación de profundizar más en esta esperanzadora área, en este estudio se propone realizar una agrupación de todas las bases de datos de expresión génica disponibles en los principales repositorios que provengan de muestras de sangre de pacientes con IBD. De esta manera, se tratará de identificar los patrones moleculares que se desarrollan bajo estas patologías con un tamaño muestral mayor, describiendo subgrupos en base a datos de expresión e intentando predecir su correlación con otras variables clínicas como la respuesta a fármacos. Todo esto con el objetivo de identificar el tipo de enfermedad para tratar de manera más eficiente y temprana a los pacientes con IBD.

2. Hipótesis y Objetivos

Las enfermedades autoinmunes inflamatorias del sistema digestivo muestran a menudo diferencias entre pacientes, tanto en manifestaciones clínicas como en respuesta a fármacos. Estas diferencias son un reflejo de diferencias a nivel molecular, por lo que profundizar en las bases de la heterogeneidad molecular en estas patologías puede ayudar en el tratamiento de las mismas.

Objetivos:

- Realizar una búsqueda de múltiples estudios públicos con muestras de sangre de pacientes con la enfermedad de Crohn y colitis ulcerosa, con el objetivo de identificar aquellos patrones moleculares detrás de la heterogeneidad de estas enfermedades. Se definirán subtipos moleculares dentro de las enfermedades.
- Analizar la asociación entre los subtipos moleculares de estas enfermedades y diferentes variables clínicas. Utilizar la información molecular para obtener marcadores predictivos para respuesta a diferentes manifestaciones clínicas y/o fármacos.

3. Metodología

Obtención de datos de expresión.

Los datos de expresión han sido obtenidos de repositorios públicos, principalmente de la base de datos NCBI Gene Expression Omnibus (GEO) y de ArrayExpress. Como se ha mencionado, el requisito esencial es que las muestras procedan de sangre total (*whole blood*) o periférica (*peripheral blood*) humana, y que no haya ninguna línea celular aislada en la secuenciación. Además, los estudios deben contener datos transcriptómicos de pacientes sanos, que actuarán como controles. Los perfiles de expresión de las muestras deben de haber sido obtenidos mediante RNA-seq o microarrays, teniendo correctamente identificada la plataforma usada para poder realizar la anotación de genes.

Procesamiento de los datos.

Se han tratado y depurado las bases de datos mencionadas mediante el uso de Rstudio (versión 2023.06.0+421), el cual es un entorno de desarrollo integrado para el lenguaje de programación R. Para la descarga de los datos clínicos y de expresión se ha utilizado el paquete "***Geoquery***" (versión 2.68) (R Core Team, 2023; Sean & Meltzer, 2007). La normalización se ha realizado mediante el escalado en logaritmos de base 2 para el caso de los perfiles de expresión obtenidos por microarray. Si los datos de expresión fueron obtenidos usando RNA-seq, la normalización ha sido empleando el método TMM (The Trimmed Mean of the M-values), gracias al paquete "***NOISeq***" (versión 2.44) (Tarazona et al., 2015). Aunque sean dos procesos de normalización diferentes, no afecta al cálculo de la puntuación de transcripción de cada gen ya que ésta se realiza a través de los controles sanos de cada estudio, relativizándose la distribución en la expresión. Basándose en el paquete "***caret***" (versión 6.0-94), se eliminaron aquellas medidas cuya expresión entre muestras tienen una varianza próxima a 0 (Kuhn, 2008). Se realizó la identificación y conversión de las sondas a un identificador estandarizado y universal de genes (Gene Symbol) gracias al paquete "***biomaRt***" (versión 2.56.1), señalando la plataforma de secuenciación utilizada en cada base de datos y traduciéndolo a *Gene Symbol* (Durinck et al., 2009). Por último, se filtraron las matrices clínicas de cada estudio para mantener solamente las muestras y las variables de interés.

Cálculo y categorización de expresión génica

Con el uso del paquete "***pathMED***" (versión 0.1.19), desarrollado por la Unidad de Bioinformática de GENYO, se calculan y cuantifican los genes y rutas moleculares que se encuentran alteradas en las muestras de los pacientes respecto a las muestras control. Gracias a ello, se podrá distinguir diferentes patrones de expresión y definir los posibles subgrupos moleculares más homogéneos dentro de la diversidad de este tipo de enfermedades. En este paso, se sigue el procedimiento empleado en el estudio de Toro-Domínguez et al., 2022.

Para llevarlo a cabo, es necesario agrupar la expresión de miles de genes en una red de módulos génicos los cuales representan funciones moleculares concretas. Los genes no actúan de forma aislada en el organismo, sino que existe una co-regulación entre ellos dentro de las funciones biológicas. Por lo que con este primer paso se resume la expresión de los genes en las funciones que desempeñan, reduciendo enormemente la dimensionalidad de los datos. Estos módulos de funciones biológicas se obtuvieron a través del paquete "***tmod***" ("*transcriptional modules related with immune funciones*") (versión 0.50.13), el cual contiene bibliotecas de módulos relacionadas con el sistema inmune e intervienen en una función biológica. Este paquete divide los genes en 606 módulos de expresión, que fueron a su vez subdivididos en múltiples subgrupos con el paquete "***pathMED***" hasta obtener 1085 submódulos de co-expresión, con los que se obtiene una mayor resolución al dividir una determinada función biológica en base a grupos de genes que se co-expresan dentro de ella (Chaussabel et al., 2008; Li et al., 2014; Toro-Domínguez et al., 2022).

La forma de medir la alteración de una ruta biológica en un paciente respecto a los controles es mediante el cálculo de M-scores de cada módulo, utilizando la metodología empleada por de Toro-Domínguez et al., 2022. De esta forma, se cuantifican las variaciones en la expresión de cada función biológica en las muestras de pacientes enfermos respecto a los controles sanos. La puntuación proviene del cálculo de z-score de la expresión de cada gen respecto a la expresión del mismo gen en el grupo de controles sanos. Con las z-score de cada gen calculada, se proporciona una puntuación

para cada módulo (M-score), compuesta por la media aritmética de todos los genes que componen dicho módulo.

La puntuación de M-score se obtiene a través de la siguiente ecuación:

$$Mscore_i = \frac{\sum_{j=1}^{n_i} \left(\frac{x_j - \mu_{jH}}{\sigma_{jH}} \right)}{n_i}$$

Ecuación 1. Fórmula empleada para el cálculo de M-scores (Toro-Domínguez et al., 2022)

En la ecuación, x_j hace referencia a la expresión de un gen (j) en la muestra de un paciente. μ_{jH} es la media de la expresión del gen y σ_{jH} es la desviación estándar, ambas de los controles sanos. Por último, n_i es el número de genes del módulo (i). La distribución centrada que se obtiene de esta ecuación marca un intervalo de -1.65 y 1.65, dentro del cual contiene el 90% de los datos, ya que cada extremo del intervalo corresponde a un P-valor de 0.05. Es por esto por lo que los módulos que muestran una desregulación significativa son aquellos cuyas M-scores están fuera de dicho intervalo.

Con las M-scores calculadas, se seleccionan aquellos módulos que interfieren significativamente en la presencia de la enfermedad mediante el paquete "***pathMED***".

Análisis de patrones moleculares y subgrupos.

Para poder realizar la anotación génica de los módulos, se utilizó de nuevo el paquete "***tmod***" con el objetivo de relacionar los grupos de genes con sus respectivas funciones biológicas y su relevancia en la respuesta inmune. Con el objetivo de encontrar el número de subtipos de enfermedad, o clústeres, se emplearon múltiples paquetes que permiten agrupar y valorar los posibles subtipos de enfermedad que se pueden generar. Se utilizó el método de *clustering* por k-means debido a su efectividad en datos de expresión génica (D'haeseleer, 2005). Se aplicó el paquete "***NbClust***" (versión 3.0.1) para calcular el número óptimo de clústeres (Charrad et al., 2014). Posteriormente se empleó "***bios2mds***" (versión 1.2.3) para calcular la "*silhouette score*", la cual sirvió de ayuda para validar el número óptimo de clústeres calculado por "***NbClust***" (Pelé et al., 2012). La medida de *Silhouette Score* permite valorar, en clusterización por *k-means*, como de similares son los puntos dentro de un clúster comparado con otros clústeres. El valor

puede variar en un intervalo de -1 a 1, siendo 1 una perfecta compactación (Rousseeuw, 1987). Para asignar cada muestra a uno de esos clústeres, se hizo uso del paquete "***ConsensusClusterPlus***" (versión 1.64), aplicándose el método de descrito de *Consensus Clustering* para *k-means*. (Wilkerson & Hayes, 2010).

Asimismo, se utilizaron los paquetes "***FactoMiner***" (versión 2.9) y "***factoextra***" (versión 1.0.7) para el Análisis de Componentes Principales del conjunto de datos (Alboukadel Kassambara & Fabian Mundt., 2016; Lê et al., 2008). Para representar gráficamente las variaciones de expresión de las muestras y el *clustering*, se usaron varios paquetes de R, como "***gplots***" (versión 3.1.3), "***pheatmap***" (versión 1.0.12) y "***RColorBrewer***" (versión 1.1-3) (Gregory R. Warnes et al., 2016; Neuwirth, 2022; Raivo Kolde, 2018).

Por último, gracias al paquete "***stats***" (versión 4.3.0) se realizó un análisis de chi cuadrado (χ^2) de independencia para evaluar si existe una correlación significativa entre el clúster asignado a una muestra y diversas variables clínicas, como el tipo de enfermedad, su estado de actividad y la severidad en las heridas observadas en la colonoscopia (R Core Team, 2023).

Desarrollo de predictores mediante *machine learning*

Los módulos de genes calculados previamente se utilizan para buscar asociaciones con diversas variables clínicas. Para ello, se utilizan los paquetes "***pathMED***" y "***caret***" para la creación de modelos predictivos basados en *machine learning* con varios objetivos. Primero, ser capaz de diferenciar a través de M-scores calculados a partir de sus datos transcriptómicos, si una persona es sana o si padece IBD. Seguidamente, y usando también M-scores, intentar predecir variables clínicas en nuevas muestras, ayudando a clasificar el estado y la tipología de la enfermedad (CD o UC). Debido a la falta de datos públicos de estas patologías autoinmunes inflamatorias, la variedad de características clínicas a predecir es limitada. Sin embargo, se usarán estos métodos para intentar identificar si una persona padece IBD, clasificar el tipo de IBD, actividad, gravedad, y la respuesta al tratamiento de corticoesteroides intravenosos (*IVCS therapy*).

Para mejorar la robustez en las estimaciones, se aplicó el método de validación cruzada repetida (*repeated-CV*), la cual es una estrategia que se utiliza para evaluar la capacidad de generalización de un modelo predictivo. Esta técnica se basa en dividir el conjunto

de datos en múltiples subconjuntos de entrenamiento y prueba de manera repetida (también llamados "*folds*"). Cada repetición implica dividir los datos en particiones distintas, alternando el conjunto de prueba y el de entrenamiento en cada iteración. De esta manera se reduce el impacto de la aleatoriedad en la partición de los datos y proporciona una estimación más precisa del rendimiento del modelo en datos no vistos. Según las muestras disponibles para cada modelo, se empleó un número diferente de *folds* para evitar la escasez de datos de entrenamiento, pero siempre con 30 repeticiones por cada *fold* (Yang et al., 2011)

Como medida para comprobar la fiabilidad del modelo se empleó el Coeficiente de Correlación de Matthews (MCC), el cual es un índice estadístico más fiable que produce una puntuación en un intervalo de -1 a 1 según los resultados obtenidos en las cuatro categorías de la matriz de confusión y proporcionalmente a los tamaños de las clases positiva y negativa del conjunto de datos (Chicco & Jurman, 2020). Se empleó el paquete "***mccr***" (versión 0.4.4) para usar este índice en los modelos predictivos desarrollados (Iuchi H, 2017).

4. Resultados

Obtención de las bases de datos

Aplicando los criterios descritos en la metodología, se han encontrado 10 estudios con información transcriptómica de pacientes con CD o con UC. Se han obtenido 1982 muestras de expresión de pacientes con IBD y 791 muestras de controles sanos, dando una población total de 2773 muestras para el estudio. Esta población de estudio es muy superior a la realizada en estudios previos de análisis de expresión transcripcional en muestras de sangre en pacientes con IBD, sin emplear muestras de biopsias procedentes de colonoscopia.

Tabla 1

Bases de datos utilizadas en el estudio

Nº Dataset	Accesion ID	CD	UC	Sanos	Muestras totales
1	GSE94648 (Planell et al., 2017)	50	25	22	97
2	GSE86434 (Ventham et al., 2016)	24	21	13	58
3	GSE119600 (Ostrowski et al., 2019)	95	93	47	235
4	GSE169568 (Juzenas et al., 2022)	52	58	95	205
5	GSE3365 (Burczynski et al., 2006)	59	26	42	127
6	GSE186507 (Argmann et al., 2023)	430	386	207	1023
7	GSE112057 (Mo et al., 2018)	60	15	12	87
8	GSE126124 (Palmer et al., 2019)	43	18	37	98
9	GSE100833 (Peters et al., 2017)	204	0	49	253
10	E-MTAB-11349 (*) (Nowak et al., 2022)	156	167	267	590
Muestras totales		1173	809	791	2773

Nota. Todas las muestras fueron obtenidas del repositorio NCBI GEO, salvo (*), que fueron adquiridas del portal ArrayExpress.

Identificación de módulos con desregulación significativa

Con las M-scores calculadas, se seleccionan aquellos módulos que interfieren significativamente en la presencia de la enfermedad mediante el paquete "***pathMED***". Para identificar los módulos que participan notablemente en el curso de la enfermedad es necesario indicar el porcentaje de muestras en los que los módulos deben de ser significativos en cada dataset, y el número mínimo de dataset en los que el módulo debe ser significativo. Con el objetivo de elegir los valores óptimos, se testaron diferentes configuraciones para esos dos parámetros, obteniendo los resultados mostrados en el Gráfico 1.

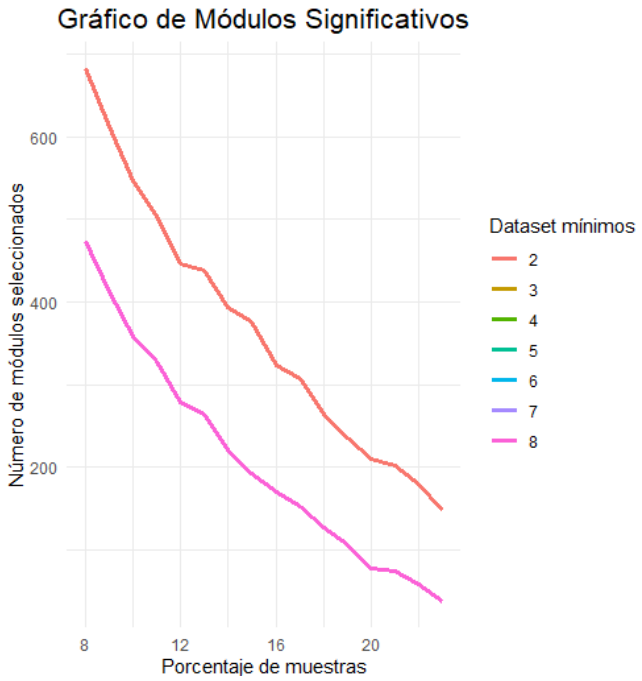

Gráfico 1. Número de módulos seleccionados según el porcentaje de muestras y el número mínimo de dataset en el que es significativo.

Solamente aparecen dos curvas debido a que a partir de establecer el mínimo en 3 dataset, el número de módulos significativos se mantienen igualados hasta llegar a 8 dataset. Debido a esto, se eligió como criterio de restricción que un módulo debe ser significativo en al menos 3 de los dataset. Además, el porcentaje de muestras de cada módulo donde debe ser significativo decrece hasta llegar al 23%, observándose una estabilización en la pendiente entre los valores del 13-20%. Tal como se comprobó en el estudio de Toro-Domínguez et al., 2022, un conjunto cercano a 200 módulos ha demostrado ser eficaz a la hora de identificar asociaciones y subgrupos de pacientes. Por todo ello, se estableció el valor del 15% como porcentaje de muestras en las que el módulo sea significativo, obteniendo un total de 192 módulos útiles para el posterior desarrollo del estudio.

***Clustering* de módulos génicos**

Los resultados del cálculo de las M-score de las muestras de pacientes respecto a los niveles de expresión de los controles sanos se obtuvieron satisfactoriamente, calculándose para los 192 módulos significativos para cada muestra. Para certificar que no existiese ningún sesgo por parte de alguno de los datasets empleados en el estudio, se realizó un Análisis de Componentes Principales de los M-score de las 1982 muestras

de pacientes enfermos (Figura 1). En la Figura 1 se puede comprobar como todas las bases de datos se agrupan de una manera similar hacia el mismo punto, sin segregación evidente de ningún dataset en particular. Existen muestras con valores más atípicos pertenecientes a los dataset 6 y 10, pero es esperable que aparezcan valores extremos precisamente en ellos, dado que son las dos bases de datos con mayor número de pacientes. También es notable la poca flexibilidad en la varianza que exhibe el dataset 9, a pesar de contar con un elevado número de muestras. Pero, debido a que se agrupa dentro del rango desplegado por parte del resto de bases de datos, se toma la decisión de mantenerlo en el estudio. Cabe señalar que la varianza explicada por los dos componentes principales es del 55%, bastante alta tratándose de datos biológicos y suficiente para valorar que no existe sesgo particular por ninguna de las bases de datos.

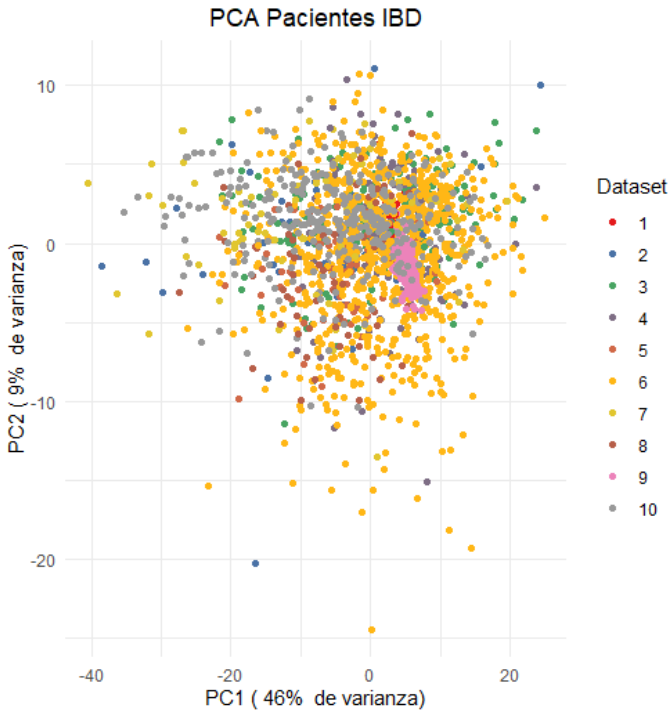

Figura 1. Análisis de Componentes Principales con los M-score de cada paciente. Los colores varían según a qué dataset pertenece la muestra.

Los resultados del *heatmap* permiten ver de manera preliminar como se distribuye la expresión génica en la muestra, representado en la Figura 2. Previo a realizar cualquier procedimiento de agrupación, se puede apreciar a simple vista como hay evidencia de dos patrones claros de co-expresión en las muestras. Otro punto que destacar, es que se observan grupos de muestras donde la diferencia de su expresión respecto a los controles sanos es más leve, teniendo M-scores menores (en valor absoluto) a lo largo

de todos los módulos génicos, mientras que hay muestras que su desregulación es mucho más acentuada.

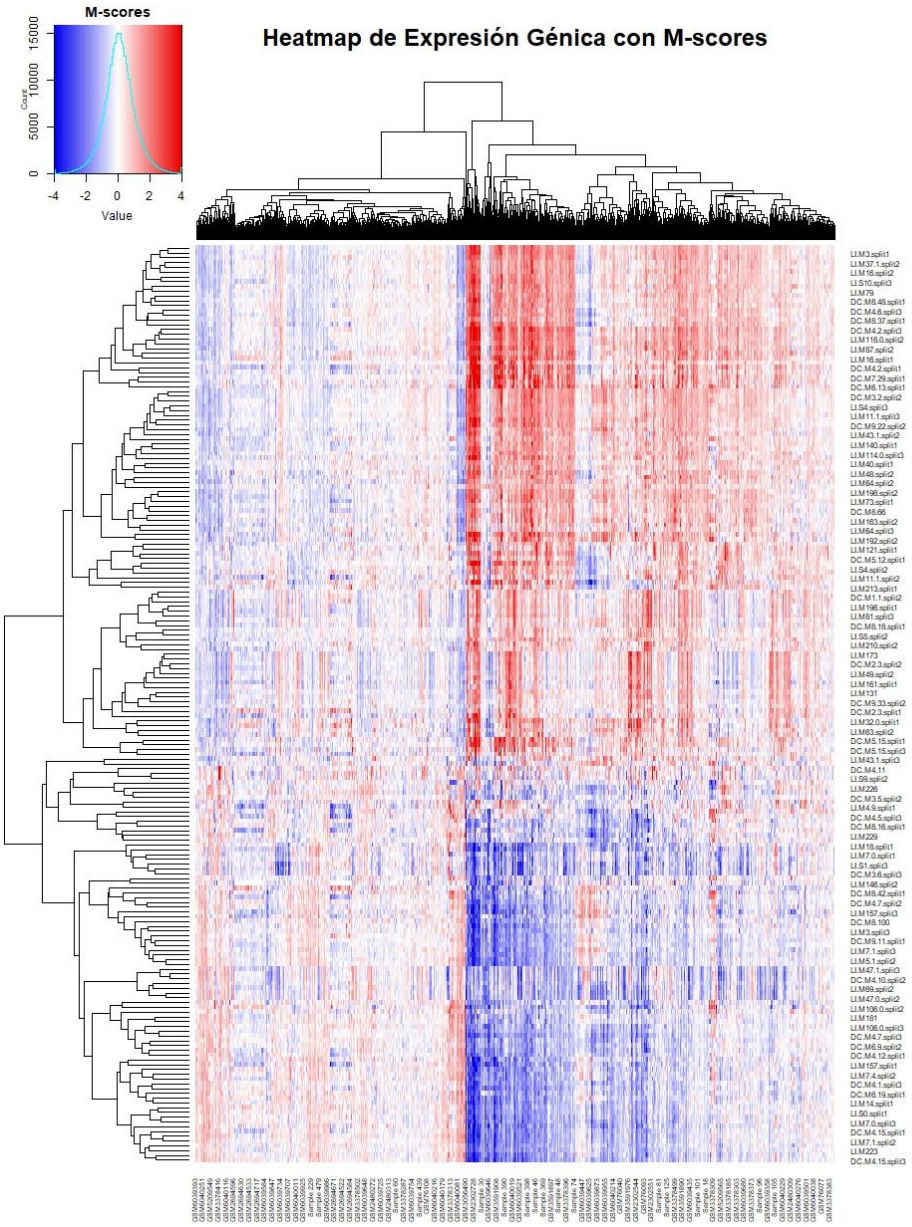

Figura 2. Heatmap de M-scores. Las filas hacen referencia a los módulos genéticos y las columnas a las muestras de pacientes. La variación de color depende del valor de M-scores, siendo rojo los módulos con una puntuación mayor (upregulated) y azul los valores más bajos (downregulated).

Los resultados del estudio de *clustering* por *k-means* de los módulos génicos (Figura 3) resaltan los dos evidentes patrones de co-expresión. Agrupar los módulos en dos conjuntos sería lo más adecuado según el Índice Callinski-Harabasz (CH), pero dotaría de poco sentido biológico al análisis. El gráfico muestra una estabilización del Índice CH a partir de los cuatro clústeres, sugiriendo que 4-5 grupos serían los ideales. Sin embargo,

al pasarlo a la práctica se observó que a partir de cuatro clústeres se formaban conjuntos con muy pocos módulos, incluso con solamente uno, resultando en una distribución irregular. Teniendo en cuenta estos factores, se optó por elegir tres clústeres de agrupación para los módulos, que poseen más estabilidad y logran una agrupación más precisa a la hora de asignar funciones biológicas. Además, la formación de tres conjuntos obtiene un *Silhouette Score* de 0.32, lo que sustenta la elección tomada.

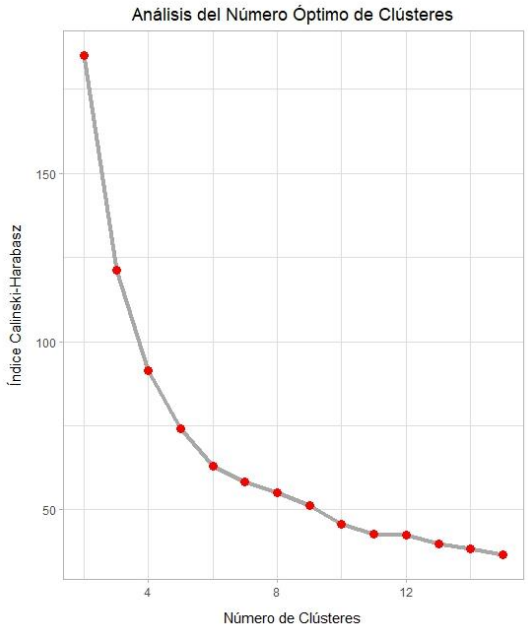

Figura 3. Distribución del número de clústeres de módulos génicos según el Índice Callinski-Harabasz.

Los resultados de la anotación tras los procesos de *clustering* se pueden ver en la Tabla 2. El primer clúster formado es el que abarca más módulos y está claramente enriquecido en rutas que intervienen el proceso de inflamación, habiendo además enriquecimiento de módulos de activación de células dendríticas y monocitos, módulos que participan en la secreción de interferón y rutas que participan en procesos de apoptosis. El segundo clúster está claramente agrupado en módulos que regulan la diferenciación de linfocitos B y T. Esté grupo además está dotado de módulos que regulan el ciclo celular y el metabolismo de las células NK. Por último, el tercer clúster está muy enfocado en procesos de coagulación y de diferenciación de plaquetas y eritrocitos. Todos ellos presentan características desreguladas que han sido descritas previamente en numerosos estudios, tanto en IBD, como en otras enfermedades autoinmunes.

Tabla 2

Clasificación de clústeres transcripcionales

Clústeres	Firma transcripcional	Rutas enriquecidas
1	Inflamación	Rutas de inflamación. Diferenciación y activación de monocitos, neutrófilos, fagocitos y células dendríticas. Interferón.
2	Metabolismo de linfocitos	Activación y diferenciación de linfocitos B, T y NK. Ciclo celular.
3	Coagulación	Rutas de coagulación. Activación y diferenciación plaquetaria. Mitosis.

Estos clústeres están claramente diferenciados como se puede ver en la Figura 4, donde en el gráfico del Análisis de Componentes Principales se observa lo bien definido que está cada grupo, en especial el segundo clúster, enfocado en la diferenciación de linfocitos. El tercer clúster está más cercano al primero, y posteriormente se podrá observar como las rutas de inflamación y de coagulación se co-expresan en la misma dirección en todos los grupos.

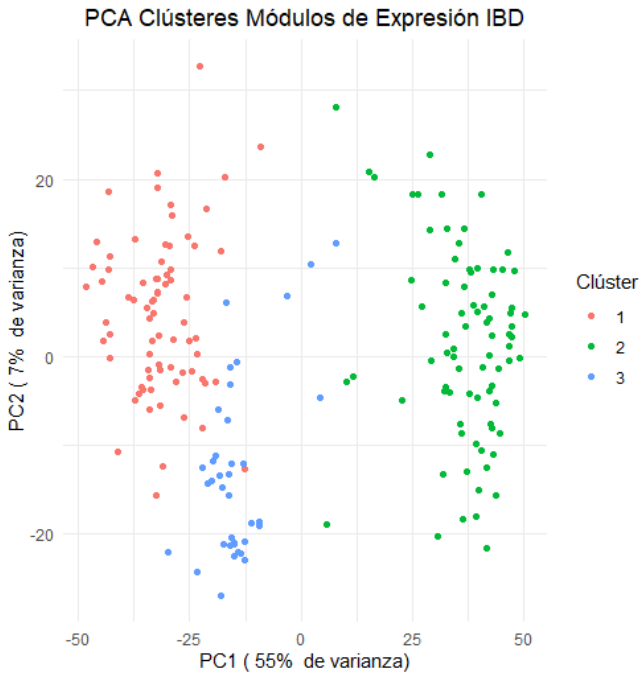

Figura 4. Análisis de Componentes Principales para los clústeres de co-expresión de módulos en pacientes de IBD.

Identificación de subgrupos de pacientes

Para lograr hallar subtipos de pacientes en IBD, se siguió el mismo protocolo empleado en el *clustering* de módulos génicos. El estudio de agrupación indicó un punto de estabilidad cuando el número de clústeres era igual a 6 como se puede ver en la Figura 5, por lo que se eligió ese valor como referencia.

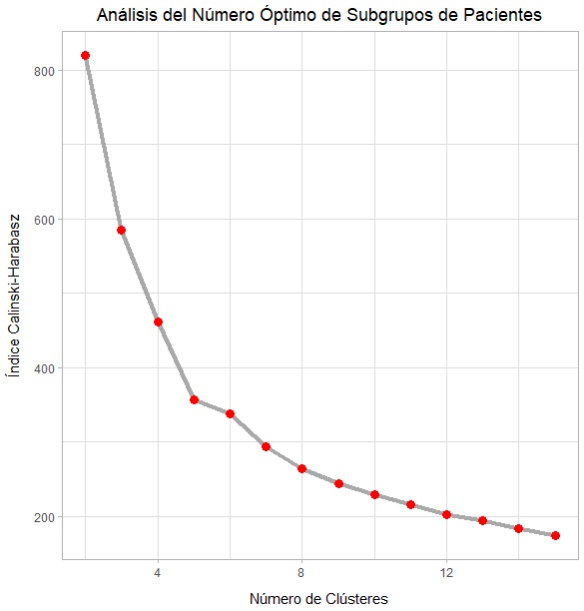

Figura 5. Distribución del número de subgrupos de pacientes según el Índice Callinski-Harabasz.

La Figura 6 representa un *heatmap* con los 6 subgrupos de pacientes en las filas y los 3 clúster de módulos génicos en las columnas. El primer clúster de pacientes es el que abarca más casos, y muestra una desregulación más leve, con una expresión más similar a los controles sanos, por lo que cabe pensar que pertenecen a casos más leves de IBD, sin un patrón claro de expresión, o de procesos biológicos fuertemente alterados. En el segundo subgrupo de pacientes se observa un ligero aumento en la desregulación de expresión de rutas enfocadas a procesos inflamatorios, mientras que hay una menor activación de los módulos típicos de linfocitos B y T, y células NK. El tercer, cuarto y quinto clúster presentan un patrón de co-expresión similar, donde se encuentran una muy alta activación de rutas de inflamación y coagulación, mientras que el conjunto de rutas de linfocitos se encuentra muy inhibido. Es de destacable mención la intensidad en la que este patrón se hace visible en el tercer clúster, donde la desregulación es máxima. En el cuarto grupo la desregulación es menor, mientras que en el quinto se presenta algo más de activación en procesos de coagulación respecto a las rutas de

inflamación. Por último, el sexto grupo de pacientes presenta un patrón de expresión totalmente opuesto a los anteriores, con los clústeres de inflamación y de coagulación regulados a la baja y una alta activación del clúster de los linfocitos. Por tanto, a grandes rasgos, se pueden ver tres subgrupos de pacientes: un subgrupo similar a los controles sanos, otro subgrupo donde se activan rutas inflamatorias y de coagulación a expensas de una baja expresión de módulos génicos de linfocitos, y un tercer conjunto de pacientes que presentan una alta expresión de funciones biológicas de linfocitos junto con una baja expresión de rutas inflamatorias y de coagulación. Estos dos últimos grupos, a su vez, se pueden dividir en base a la magnitud de los M-scores que presenten. La Figura 6 representa un *heatmap* con los seis subgrupos de pacientes en las filas y los tres clústeres de módulos génicos en las columnas.

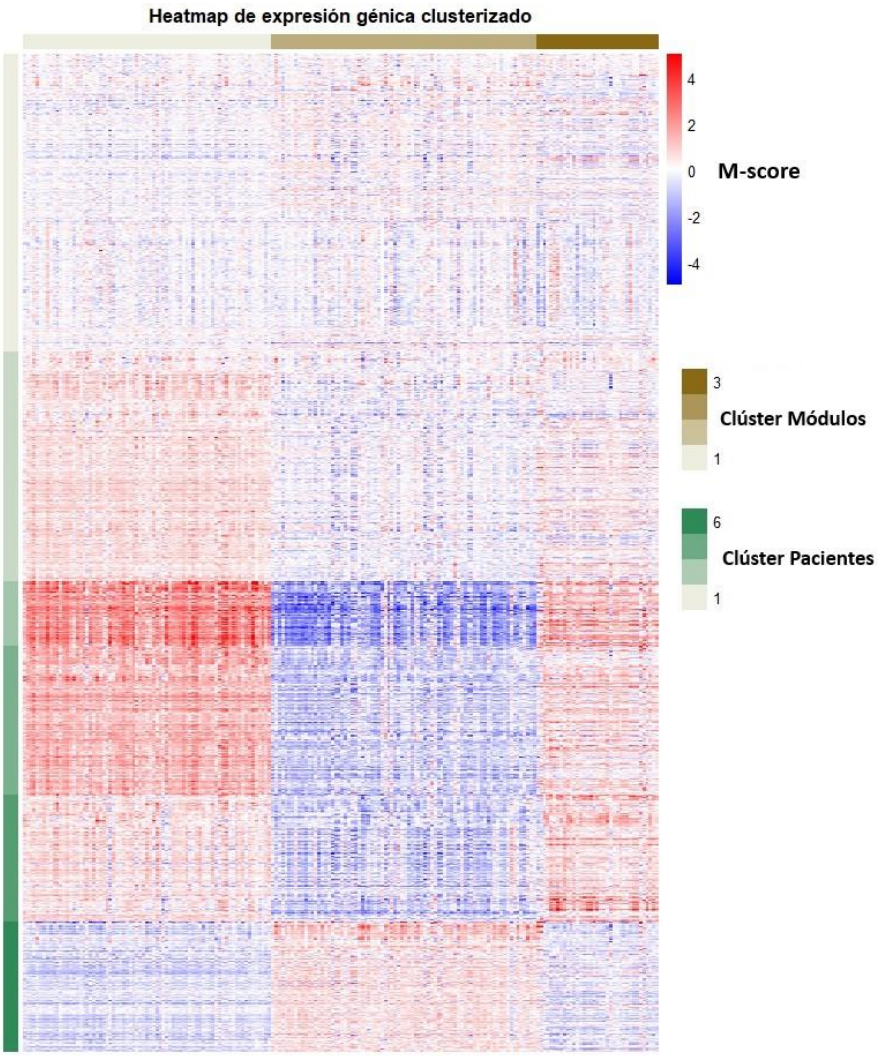

Figura 6. Heatmap de M-score con los subgrupos de pacientes en las filas y los módulos de genes en las columnas. Se ha dispuesto de esta manera para facilitar la detección de las diferencias entre los subgrupos de enfermedad. Los cambios de intensidad de color marcan cada clúster, tal como marca la leyenda.

En la Tabla 2 se ve el balance de pertenencia de ambas patologías de IBD a cada subtipo de pacientes. Es destacable que no hay una clara clasificación de enfermedades por subgrupo, lo que enfatiza lo similares que son en la expresión génica ambas enfermedades autoinmunes. Sin embargo, existe un desbalanceo en los subgrupos con mayor diferencia de desregulación. En el subgrupo más parecido a los controles sanos, existe una mayor proporción de casos con CD que con UC, con un 68% de las muestras. En cambio, el grupo donde hay una mayor desregulación respecto a las muestras sanas, es el único donde existen más casos de UC que de CD, lo cual se puede deber a que suele ser una enfermedad más agresiva en la sintomatología y que requiere una mayor tasa de cirugía. No obstante, la alta variabilidad en las manifestaciones clínicas dificulta clasificar a cada enfermedad en un fenotipo más grave.

Tabla 2

Clasificación de enfermedad por subgrupo de pacientes

Subgrupo	Colitis ulcerosa	Crohn	Pacientes totales
1	190 (32%)	401 (68%)	591
2	215 (47%)	240 (53%)	455
3	65 (50%)	64 (50%)	129
4	119 (40%)	178 (60%)	297
5	106 (42%)	145 (58%)	251
6	114 (44%)	145 (56%)	259
Muestras totales	809	1173	1982

Esta distribución también se puede observar de manera gráfica en la Figura 7, donde está el anterior *heatmap* junto con la clasificación de la enfermedad, y un Análisis de Componentes principales mostrando cómo ambas patologías ocupan un espacio muy similar, resaltando su similitud en la expresión.

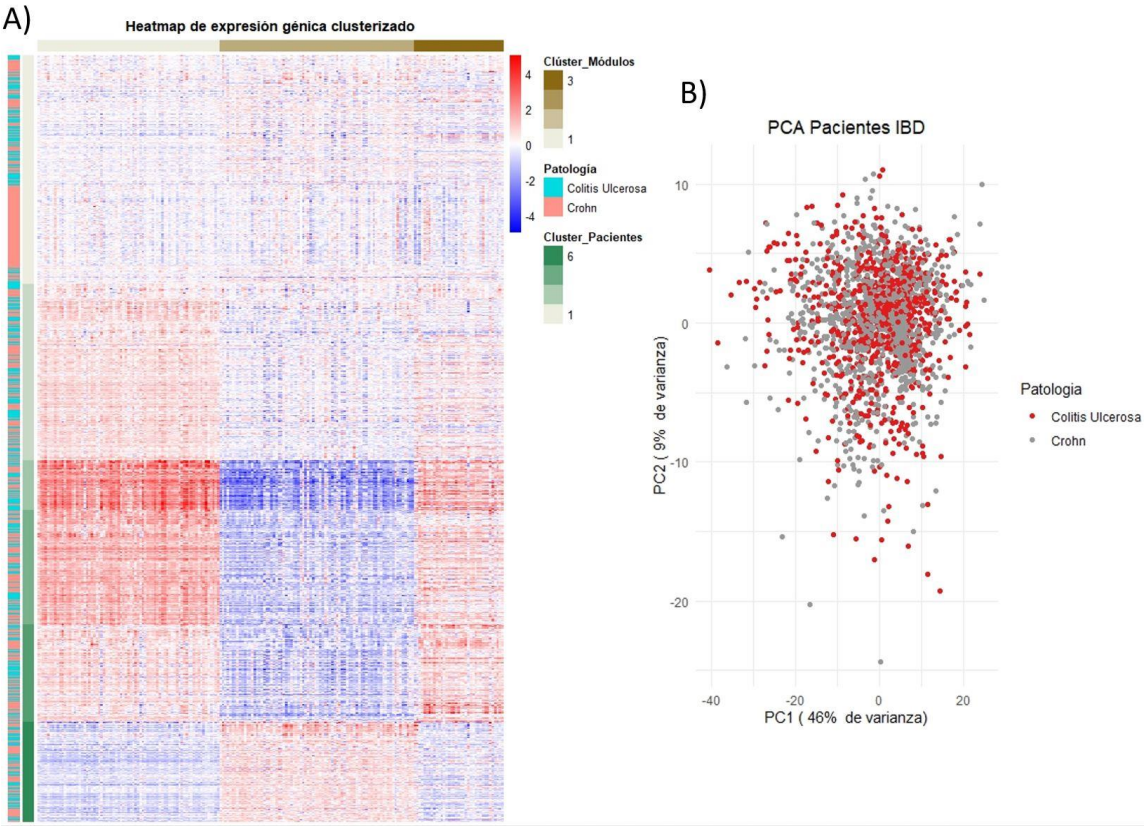

Figura 7. A) Heatmap de M-score con la agrupación de módulos en clústeres y de subgrupos de pacientes. En el lateral izquierdo está indicada la distribución de pacientes con CD y UC en cada subgrupo de pacientes. B) Análisis de Componentes Principales que muestra la agrupación de las muestras en función de la patología que sufre el paciente.

La severidad de la enfermedad es otra variable que fue sometida a análisis para observar patrones de clasificación. La gravedad de la enfermedad experimentada por cada paciente fue categorizada de acuerdo con los resultados obtenidos por exámenes estandarizados tras hacer la colonoscopia. Para la enfermedad de Crohn se empleó la puntuación endoscópica simple (SES-CD). Dicha puntuación divide el estado de la enfermedad en cuatro categorías: inactiva (0-2), leve (3-6), moderada (7-15) y grave (≥16). En el caso de los pacientes con colitis ulcerosa, se empleó la medida endoscópica Mayo para clasificar la enfermedad en categorías que incluyen inactiva (0), leve (1), moderada (2) y grave (3). Las endoscopias se hicieron el mismo día en el que se tomaron las muestras de sangre periférica que se emplean en el estudio.

Tal como se comenta en el apartado de Discusión, una de las principales desventajas del análisis de variables es que los datos clínicos que aparecen en los repositorios públicos suelen ser escasos y pobres. Debido a ello, no todas las muestras empleadas en este estudio poseen información sobre el grado de enfermedad, por lo que este análisis se

realizó con un porcentaje de ellas. A pesar de ello, los resultados fueron interesantes para analizar, como se puede ver en la Tabla 3 y en la Figura 8.

Tabla 3

Clasificación de severidad de la enfermedad por subgrupo de pacientes

Subgrupo	Inactiva	Leve	Moderada	Grave	Muestras totales
1	95 (44%)	75 (35%)	39 (18%)	8 (3%)	217
2	66 (31%)	72 (34%)	44 (21%)	28 (13%)	210
3	5 (23%)	5 (23%)	6 (27%)	6 (27%)	22
4	17 (17%)	28 (27%)	28 (27%)	28 (27%)	101
5	42 (30%)	55 (40%)	32 (23%)	9 (7%)	138
6	86 (57%)	39 (26%)	21 (14%)	5 (3%)	151
Muestras totales	311	274	170	84	839*

Nota. Las muestras con datos de gravedad de la enfermedad son 839(*), que ocupan un 42% del total de muestras del estudio

La clasificación de las muestras dentro de cada subgrupo de pacientes muestra que hay un mayor porcentaje de muestras inactivas o leves en el primer subgrupo, con una expresión similar al de los controles sanos, mientras que en los grupos subgrupos tres y cuatro, hay una notoria mayor presencia de pacientes graves, siendo el 27% muestras de pacientes con curso grave de IBD (54% entre moderadas y graves). Estos datos concuerdan con los niveles de desregulación mostrado por la estratificación de pacientes, por lo tanto, una mayor alteración a nivel molecular se refleja en una mayor gravedad clínica de la patología.

Por último, en la Tabla 4 se puede apreciar cómo se distribuye la variable de actividad de la enfermedad en los subgrupos de pacientes. Esta medida se clasifica en activa o inactiva según la puntuación obtenida en función de los síntomas que presente el paciente. Para pacientes de CD se utilizó el índice de Harvey-Bradshaw (HBI), y para pacientes con CU se empleó el índice clínico simple de actividad de la colitis (SCCAI). La enfermedad clínicamente inactiva se definió como un HBI < 5 o un SCCAI < 5 y la enfermedad activa como un HBI >7 o un SSCAI >= 5. En este caso ocurre lo mismo que con la severidad de la enfermedad, no todas las muestras poseen datos sobre su

actividad, por lo que el análisis se realizó sobre un tamaño muestral menor. En el clúster con un patrón de expresión menos desregulado respecto a los controles sanos, solo un 8% de los pacientes tenían enfermedad activa. En contraste, en el grupo con una mayor desregulación, los pacientes con la enfermedad activa representaban un 37%.

Tabla 4

Clasificación de estado de actividad de la enfermedad por subgrupo de pacientes

Subgrupo	Inactiva	Activa	Muestras totales
1	179 (92%)	16 (8%)	195
2	148 (83%)	31 (17%)	179
3	12 (63%)	7 (37%)	19
4	64 (74%)	22 (26%)	86
5	109 (83%)	21 (17%)	130
6	113 (84%)	21 (16%)	134
Muestras totales	625	118	743*

Nota. Las muestras con datos de gravedad de la enfermedad son 743(*), que ocupan un 37% del total de muestras del estudio.

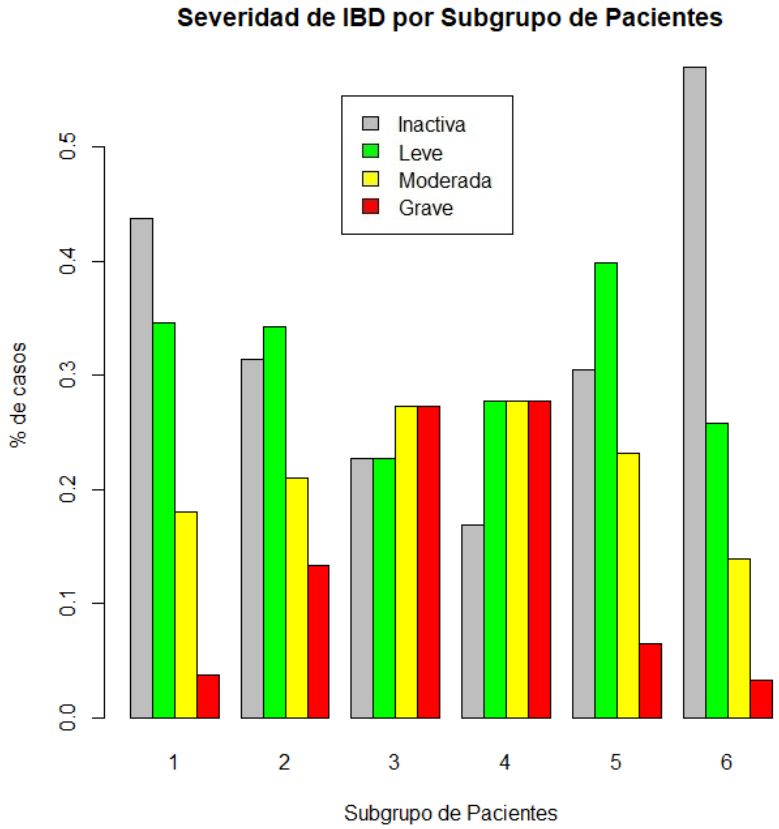

Figura 8. Barplot del porcentaje de muestras según la severidad de la enfermedad en cada subgrupo de pacientes.

En la Figura 8 está representada gráficamente los resultados expuestos en la Tabla 4. Se puede ver de manera más evidente el aumento de casos con daños intestinales de mayor gravedad en los subgrupos tres y cuatro. También es sorprendente el elevado caso de enfermedad inactiva y leve en el subgrupo seis, donde el clúster de linfocitos está sobreexpresado a expensas de una desregulación negativa de los clústeres de inflamación y coagulación (siendo el único subgrupo en el que esto ocurre).

Una vez definidos los subgrupos de pacientes según las variables de expresión se procede a realizar un análisis de asociación mediante χ^2 de independencia entre los seis subconjuntos y las diferentes variables clínicas, con el objetivo de certificar la efectividad del agrupamiento y poder clasificar de manera óptima a los pacientes. Para ello, se crean diferentes tablas de contingencia entre los subgrupos y las diversas variables, como el tipo de IBD, la actividad de la enfermedad en base a sus síntomas y a los resultados de la colonoscopia. Los resultados se pueden ver en la Tabla 5.

Tabla 5

Análisis de asociación por χ^2 de independencia.

Variable	χ^2	Grados de libertad	P-valor
Tipo de IBD	32.4	5	<0.0001
Actividad IBD	21.2	5	<0.001
Severidad IBD	105.9	15	<0.0001

La Tabla 5 muestra como hay una elevada significancia en la correlación entre los subgrupos de paciente y las variables clínicas, por lo que su distribución no es casual. La alta proporción de pacientes con CD y la elevada tasa de pacientes con la enfermedad inactiva en cuanto a sintomatología y en cuanto a lesiones intestinales en el primer subgrupo se debe a la baja desregulación transcripcional que presentan. Lo inverso ocurre con los clústeres donde presentan una mayor desregulación, en particular el tercero, donde hay una mayor proporción de pacientes con UC y con actividad en la enfermedad.

Desarrollo de predictores

Modelo de identificación de pacientes con IBD

El primer objetivo planteado fue intentar crear un modelo que a través de muestras de sangre pueda predecir si una persona padece IBD, sin necesidad de colonoscopia. De esta manera, lograría acelerar el sistema de salud, al priorizar la colonoscopia de aquellos pacientes que sean clasificados como enfermos, ya que requieren una mayor urgencia en el tratamiento. Las mejores métricas se obtuvieron con el modelo *Random Forest* (Figura 9), llegando a obtener valores muy consistentes, con una sensibilidad superior al 97%, una precisión balanceada del 87% y el área bajo la curva ROC del 93% (Tabla 6). Como medida para comprobar la habilidad de clasificación binaria del modelo se empleó el Coeficiente de Correlación de Matthews (MCC), obteniéndose un índice de 0.79, mostrando la alta efectividad de predicción en todos los casos.

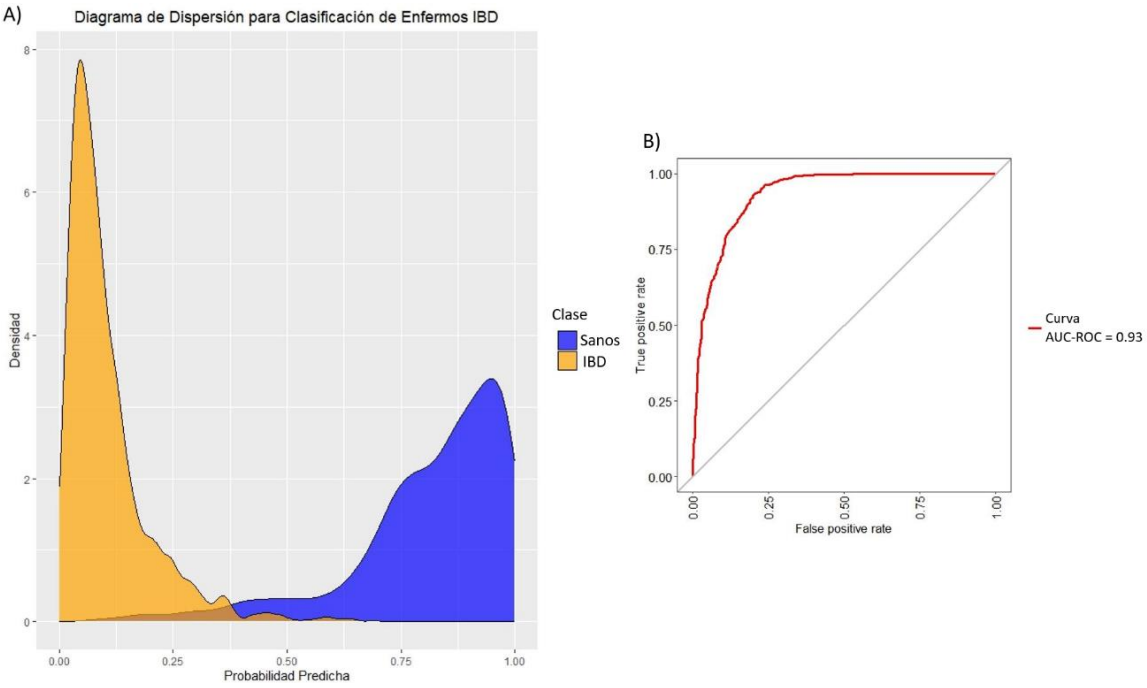

***Figura 9**. Resultados del Modelo Predictivo para detectar pacientes de IBD. A) Diagrama de dispersión donde se puede apreciar la densidad de detección del modelo para ambos grupos. B) Curva AUC-ROC del modelo.*

Tabla 6

Métricas de modelo predictivo para identificar personas con IBD.

Métrica	Valor	Métrica	Valor
Exactitud	0.92	Especificidad (Sano)	0.76
P-Valor	< 2.2e-16	Precisión Balanceada	0.87
Kappa	0.78	MCC	0.79
95% CI	(0.89, 0.93)	AUC-ROC	0.93
Sensibilidad (Enfermo)	0.98	F1-Score	0.94

Sorprendentemente, de las cinco variables más importantes para explicar el modelo, cuatro de ellas pertenecen al clúster de linfocitos, incluidas las tres primeras. Es un hallazgo revelador que ninguna ruta de inflamación sea más influyente a la hora de contribuir a las predicciones del modelo. Los prometedores resultados de este clasificador indican que el cálculo de M-scores es de gran utilidad para reducir la dimensionalidad y generar un buen modelo que diferencia entre pacientes enfermos y personas sanas con tan solo 192 predictores de expresión génica.

Modelo predictivo CD vs UC

Otra variable que sería interesante predecir sería si, una vez detectado que la persona padece algún tipo de IBD, identificar si se trata de enfermedad de Crohn o de colitis ulcerosa. Para ello se empleó el paquete "***pathMED***" para calcular qué modelo tiene una mayor efectividad. El modelo con mejor rendimiento fue el *Support Vector Machine Linear*, con unas métricas descritas en la Tabla 7.

Los resultados de este modelo demuestran que, aunque es evidente que hay diferencias significativas entre ambas enfermedades, los patrones transcripcionales basados en módulos de expresión inmune son muy similares y difíciles de clasificar con robustez. Es posible que con un mayor número de datos se pudiese obtener un mejor rendimiento en base a los resultados obtenidos, aunque esto demuestra que ambas enfermedades comparten procesos moleculares. También es destacable que, de nuevo, la variable más importante para explicar el modelo es un módulo de la actividad mitótica en linfocitos CD4, no en una función perteneciente al clúster de inflamación.

Tabla 7

Métricas de modelo predictivo para la enfermedad subyacente al IBD.

Métrica	Valor	Métrica	Valor
Exactitud	0.65	Especificidad (UC)	0.44
P-Valor	< 0.005	Precisión Balanceada	0.62
Kappa	0.25	MCC	0.26
95% CI	(0.61, 0.69)	AUC-ROC	0.73
Sensibilidad (CD)	0.8	F1-Score	0.73

Modelo predictivo estado de enfermedad: Activo o Inactivo

Otra variable con interés para predecir con los datos existentes sería la presencia de actividad de la enfermedad en el tejido intestinal. Para ello se empleó la variable de Severidad descrita anteriormente y se intentó detectar la presencia o ausencia de heridas dividiendo los datos en dos grupos: los pacientes que tenían inactivada la enfermedad en el tejido, y los que presentaban algún tipo de lesión. Los resultados del modelo óptimo fueron también de *Random Forest*, pero sin la misma efectividad que el anterior, obteniendo valores muy bajos (55% de precisión balanceada y MCC de 0.12).

Se probó de nuevo con la misma variable, pero agrupando a las muestras catalogadas como activa y como leve de bajo riesgo, y por otro lado las muestras con severidad moderada y grave, como pacientes de alto riesgo. En este caso, las métricas del modelo son algo más efectivas, mostrando una precisión balanceada de 71% y un MCC de 0.46 (Tabla 8). Una traba para la efectividad de este modelo es el desbalanceo de clases, existiendo más del doble de muestras de bajo riesgo. Sin embargo, su precisión en la detección son prometedoras cuando se pueda emplear con un mayor tamaño muestral.

Tabla 8

Métricas de modelo predictivo para muestras de alto y de bajo riesgo.

Métrica	Valor	Métrica	Valor
Exactitud	0.78	Especificidad (Bajo Riesgo)	0.9
P-Valor	< 0.005	Precisión Balanceada	0.71
Kappa	0.454	MCC	0.46
95% CI	0.53	AUC-ROC	0.71
Sensibilidad (Alto Riesgo)	0.53	F1-Score	0.6

Con estos resultados se destaca de nuevo la heterogeneidad en la expresión de estas enfermedades, incluso en episodios de remisión y de actividad, aunque es claro que existen patrones de expresión diferenciados entre pacientes con la enfermedad activa y pacientes con la enfermedad inactiva. De hecho, en este caso las variables con mayor importancia para el modelo pertenecen al clúster de inflamación. Queda demostrado que es posible predecir los daños intestinales mediate la expresión génica de las rutas inflamatorias en un análisis de sangre, y será adecuado volver a probarlo con una población mayor de estudio con el objetivo de desarrollar un modelo aún más preciso.

Modelo Respondedor a tratamiento con ICVS.

Se intentó usar el cálculo de M-scores en una base de datos obtenida en GEO (GSE21231) para intentar predecir la respuesta a pacientes con IBD tras el tratamiento con corticoesteroides intravenosos (Kabakchiev et al., 2010). Desgraciadamente, el tamaño muestral del estudio es muy pequeño, con 40 muestras, siendo 20 de cada clase, muy bajo para que ningún modelo pudiese conseguir métricas fiables (precisión balanceada de 0.56 y MCC de 0.13).

5. Discusión

Se han definido tres clústeres de co-expresión genética con una alta correlación entre los seis subgrupos de pacientes de IBD, cada uno enriquecido principalmente en una

función inmune: inflamación (Clúster 1), diferenciación y activación de linfocitos (Clúster 2), y activación de plaquetas y coagulación (Clúster 3).

De los seis subgrupos de pacientes descritos, se ha observado que cuatro de ellos tienen un patrón de expresión similar, aunque con diferentes grados de intensidad de M-score. En estos pacientes, se han visto sobreexpresados el Clúster 1 y el Clúster 3, con un descenso en la expresión del Clúster 2. Además, se ha asociado significativamente la presencia de enfermedad activa y de mayor daño intestinal con la pertenencia a estos subgrupos, especialmente a los subgrupos tres y cuatro que son los que presentan una mayor desregulación.

Estudios previos señalan que existe asociación de monocitosis en sangre en pacientes con Crohn y colitis ulcerosa. Se ha observado que padecer monocitosis con IBD está correlacionado con un aumento de biomarcadores sanguíneos de inflamación, una peor calidad de vida y prognosis de la enfermedad. También se ha reportado en varios metaanálisis de gran tamaño muestral, que los pacientes con IBD tienen un ratio más elevado de neutrófilos-linfocitos y de monocitos-linfocitos que los controles sanos, incluso se ven diferencias significativas entre pacientes con enfermedad activa frente a inactiva, lo que es un reflejo de la inflamación en curso y del estado inmune del paciente (Anderson et al., 2022; Fu et al., 2021; Gao et al., 2023). Estas observaciones han quedado contrastadas con los resultados obtenidos en este estudio, donde se ve una correlación inversa entre el clúster de inflamación con respuesta inmune innata y el clúster de linfocitos, donde predomina la respuesta inmune adaptativa.

El clúster de coagulación siempre se co-expresa positivamente con el clúster de inflamación. Estudios previos demuestran como las plaquetas están involucradas en la perpetuación de la inflamación crónica en IBD. Se ha descrito que la activación de la trombocitosis ocurre en la fase activa de la enfermedad, regulando la coagulación e inflamando el tejido mucoso con la secreción de numerosos compuestos activos. Esto se debe a que, durante el proceso de activación, las plaquetas desarrollan receptores de citoquinas y provoca que participen en la cascada de inflamación, principalmente aumentando la permeabilidad vascular. También se ha detectado un aumento de los niveles de proteínas participantes en cascadas de coagulación en pacientes con IBD,

además de un aumento en la activación plaquetaria y en micropartículas circulantes derivadas de plaquetas (PDMPs) en pacientes activos, las cuales participan en la respuesta inflamatoria. Asimismo, un estudio también reporto un significativo elevada ratio de plaquetas-linfocitos en pacientes con IBD activa. (Gao et al., 2023; Matowicka-Karna, 2016).

Numerosos estudios han evidenciado que los pacientes de IBD tienen hasta tres veces más riesgo que la población general de padecer un tromboembolismo venoso (formación de un coágulo de sangre en una vena). Y los estudios apuntan que es una característica única de IBD, ya que no aparece en otras inflamaciones crónicas, y además está asociada significativamente a episodios activos de la enfermedad y al tratamiento de esta con corticoesteroides (Cheng & Faye, 2020; Matowicka-Karna, 2016). Por ello, el método empleado podría ser de elevada utilidad para priorizar a aquellos pacientes que estén en mayor riesgo de tromboembolismo debido a un aumento de la actividad transcripcional de este clúster.

A la vista de los resultados, es aparentemente evidente que los subgrupos 2, 3, 4 y 5 (especialmente el 3 y 4) son agrupaciones que representan un estado activo de la enfermedad, y un mayor riesgo, siendo posible su identificación a través de una muestra de sangre rutinaria. También pueden representar a pacientes no respondedores bajo un tratamiento, ya que las muestras que no responden a un tratamiento pueden presentar un alto grado de desregulación respecto a los controles sanos (Pavlidis et al., 2019). Sin embargo, se ha identificado un subtipo de paciente que a nivel molecular presenta un patrón completamente opuesto a los subgrupos mencionados anteriormente. El subgrupo 6 destaca por una baja expresión del Clúster 1 y Clúster 3, mientras que el Clúster 2 se encuentra claramente sobreexpresado. Este subgrupo se caracteriza por tener la mayor proporción de casos de inactividad en la severidad endoscópica y una menor ratio de casos moderados-graves. Además, su nivel de actividad es similar al del primer subgrupo que posee una menor desregulación.

No se han descrito casos de IBD que muestren un descenso en los niveles de inflamación y coagulación, y alta actividad linfocitaria, por lo que la identificación de este subgrupo es sorprendente, ya que no es exclusivo de un dataset o de una variable específica. Una

teoría que podría explicar el bajo nivel de actividad y gravedad de estas muestras es que se traten de muestras de pacientes respondedores a un tratamiento. Se ha expuesto en otros estudios que los pacientes respondedores a terapias anti-TNF, las cuales tienen el objetivo de disminuir la inflamación, muestran un ratio NLR y PLR significativamente menor que cuando empezaron el tratamiento. También hay estudios que han observado episodios de neutropenia en el 34% de pacientes bajo tratamiento anti-TNF. El aumento de la actividad de linfocitos puede deberse a un mecanismo de respuesta compensatoria del sistema inmunológico a la regulación negativa de las rutas de inflamación y coagulación. Sin embargo, otro estudio que cuantificó la actividad molecular de pacientes respondedores con tratamientos anti-TNF en muestras de colonoscopia mostró que, aunque los niveles de actividad de linfocitos y monocitos disminuyesen en los pacientes respondedores, no eran menores que en los controles sanos (Bertani et al., 2020; Levin et al., 2016; Pavlidis et al., 2019).

Sería necesario realizar un nuevo estudio teniendo el control de las variables como el tratamiento bajo el que está el paciente, o el nivel de neutrófilos y linfocitos en sangre, para poder realizar una aproximación más exacta. Lo que es seguro, es que existe una desregulación en las funciones moleculares de linfocitos muy alta en pacientes con IBD, tanto que es el principal factor que ha empleado el modelo clasificador para diferenciar entre personas sanas y pacientes con IBD, y entre enfermos por CD o UC. Además, este subgrupo de pacientes tiene una sintomatología más leve en el momento de la extracción y no son susceptibles de mayor riesgo que los otros subgrupos con mayor desregulación. Por lo tanto, el aumento de actividad en inflamación y la coagulación permanecen como los principales indicadores de riesgo para la enfermedad.

El primer subgrupo de pacientes es la fracción de muestras que tiene la tasa de desregulación más baja, asemejándose más al grupo de los controles sanos. Es el grupo que posee un mayor porcentaje de pacientes con enfermedad inactiva, con un 92% respecto a los pacientes con IBD activo. Los datos dan a entender que este grupo de pacientes está compuesto por enfermos que no están en un episodio activo de la enfermedad, ya bien porque se encuentren en una etapa de inactividad o porque están respondiendo positivamente a un tratamiento (Pavlidis et al., 2019)

Por otro lado, se ha conseguido desarrollar un modelo predictivo altamente efectivo para poder diagnosticar a personas con sospecha de IBD. Este modelo podría ayudar a priorizar a las personas que requieren una examinación mediante colonoscopia para identificar el tipo de patología y valorar los daños del tejido intestinal. Una detección precoz y un tratamiento para la inflamación puede mejorar la prognosis de la enfermedad, la cual tiene una incidencia de más de 3600 casos al año en España, aunque se estima que su incidencia real puede estar alrededor de 10000 casos anuales (Chaparro et al., 2021). Existen otros estudios recientes de realizar un modelo predictivo de diagnóstico de IBD sin colonoscopia. El grupo de Ungaro *et al*., realizaron un modelo predictivo con también grandes resultados a través de proteínas marcadoras de analíticas sanguíneas, de heces y de orina. En cambio, uno de los puntos fuertes de este modelo es que se puede realizar a través de la secuenciación de una muestra de sangre periférica sin la necesidad de aislar ninguna línea celular ni de emplear varios laboratorios clínicos de detección, lo que disminuye el tiempo y los recursos (Ungaro et al., 2021).

Los perfiles de expresión entre enfermos de CD y de UC son muy similares, y se ha visto que es difícil lograr un modelo que clasifique correctamente entre ambas enfermedades. Otros estudios también demuestran las complicaciones que existen para diferenciar el tipo de IBD a través de una muestra de sangre, resaltando la heterogeneidad existente dentro del abanico transcripcional de las IBD (Netz et al., 2018). Sin embargo, el modelo que se ha desarrollado ha indicado que las variables con mayor importancia para diferenciar entre ambas enfermedades pertenecen a funciones moleculares de linfocitos. Este hecho abre la puerta a nuevos estudios que investiguen el papel de los linfocitos en sangre periférica en la diferenciación de ambas enfermedades.

Por último, se ha conseguido desarrollar un modelo predictivo con un rendimiento prometedor para detectar la gravedad de las lesiones intestinales para pacientes con IBD. Este modelo puede ayudar a identificar a aquellos pacientes que sufren mayor daño intestinal y a tratarles con la prioridad necesaria para evitar una peor prognosis. Se identificó que los módulos más importantes para diferenciar entre personas de bajo y

alto riesgo de lesión pertenecían al clúster de inflamación. La teoría de que la inflamación es un buen indicador en sangre periférica está sustentada con el estudio de Planell *et al.*, que detalló que los pacientes con respuesta a tratamiento anti-TNF habían mostrado un descenso significativo en la expresión en sangre periférica de CD177, una glicoproteína de membrana en neutrófilos (Planell et al., 2017).

6. Conclusiones

Las Enfermedades Inflamatorias Intestinales son enfermedades muy complejas, que comparten grandes similitudes en cuanto a expresión génica y que presentan un mosaico de manifestaciones clínicas. Identificarlas de manera rápida es de vital importancia para mejorar el pronóstico y la calidad de vida de quienes la sufren. En este estudio, a partir de una muestra de sangre se ha podido identificar tres firmas transcripcionales de la enfermedad y seis subgrupos de pacientes que presentan diferentes patrones de expresión, relacionándose con la actividad y la gravedad de la enfermedad.

Además, se ha logrado desarrollar un modelo predictivo mediante *machine learning* para el diagnóstico de enfermos con IBD con una alta precisión que podría acelerar el tratamiento a estas personas. El análisis del perfil de expresión inmunológica a partir de muestras de sangre periférica podría desempeñar un papel fundamental en la clasificación de su gravedad y, en última instancia, en la optimización de las estrategias de tratamiento. Es vital realizar un mayor número de validaciones antes de considerarse su posible aplicación en la práctica clínica.

7. Bibliografía

Alatab, S., Sepanlou, S. G., Ikuta, K., Vahedi, H., Bisignano, C., Safiri, S., Sadeghi, A., Nixon, M. R., Abdoli, A., Abolhassani, H., Alipour, V., Almadi, M. A. H., Almasi-Hashiani, A., Anushiravani, A., Arabloo, J., Atique, S., Awasthi, A., Badawi, A., Baig, A. A. A., … Naghavi, M. (2020). The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of

Disease Study 2017. *The Lancet Gastroenterology and Hepatology*, *5*(1), 17–30. https://doi.org/10.1016/S2468-1253(19)30333-4.

Alboukadel Kassambara, & Fabian Mundt. (2016). *Package "factoextra" Type Package Title Extract and Visualize the Results of Multivariate Data Analyses*. https://github.com/kassambara/factoextra/issues.

Alghoul, Z., Yang, C., & Merlin, D. (2022). The Current Status of Molecular Biomarkers for Inflammatory Bowel Disease. In *Biomedicines* (Vol. 10, Issue 7). MDPI. https://doi.org/10.3390/biomedicines10071492.

Anderson, A., Cherfane, C., Click, B., Ramos-Rivers, C., Koutroubakis, I. E., Hashash, J. G., Babichenko, D., Tang, G., Dunn, M., Barrie, A., Proksell, S., Dueker, J., Johnston, E., Schwartz, M., & Binion, D. G. (2022). Monocytosis Is a Biomarker of Severity in Inflammatory Bowel Disease: Analysis of a 6-Year Prospective Natural History Registry. *Inflammatory Bowel Diseases*, *28*(1), 70–78. https://doi.org/10.1093/ibd/izab031.

Argmann, C., Hou, R., Ungaro, R. C., Irizar, H., Al-Taie, Z., Huang, R., Kosoy, R., Venkat, S., Song, W. M., Di'narzo, A. F., Losic, B., Hao, K., Peters, L., Comella, P. H., Wei, G., Atreja, A., Mahajan, M., Iuga, A., Desai, P. T., … Suárez-Fariñas, M. (2023). Biopsy and blood-based molecular biomarker of inflammation in IBD. *Gut*, *72*(7), 1271–1287. https://doi.org/10.1136/gutjnl-2021-326451.

Bertani, L., Rossari, F., Barberio, B., Demarzo, M. G., Tapete, G., Albano, E., Svizzero, G. B., Ceccarelli, L., Mumolo, M. G., Brombin, C., Bortoli, N. De, Bellini, M., Marchi, S., Bodini, G., Savarino, E., & Costa, F. (2020). Novel prognostic biomarkers of mucosal healing in ulcerative colitis patients treated with anti-Tnf: Neutrophil-To-lymphocyte ratio and platelet-To-lymphocyte ratio. *Inflammatory Bowel Diseases*, *26*(10), 1579–1587. https://doi.org/10.1093/ibd/izaa062.

Biasci, D., Lee, J. C., Noor, N. M., Pombal, D. R., Hou, M., Lewis, N., Ahmad, T., Hart, A., Parkes, M., Mckinney, E. F., Lyons, P. A., & Smith, K. G. C. (2019). A blood-based prognostic biomarker in IBD. *Gut*, *68*(8), 1386–1395. https://doi.org/10.1136/gutjnl-2019-318343.

Burczynski, M. E., Peterson, R. L., Twine, N. C., Zuberek, K. A., Brodeur, B. J., Casciotti, L., Maganti, V., Reddy, P. S., Strahs, A., Immermann, F., Spinelli, W., Schwertschlag, U., Slager, A. M., Cotreau, M. M., & Dorner, A. J. (2006). Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *Journal of Molecular Diagnostics*, *8*(1), 51–61. https://doi.org/10.2353/jmoldx.2006.050079.

Chaparro, M., Garre, A., Núñez Ortiz, A., Diz-Lois Palomares, M. T., Rodríguez, C., Riestra, S., Vela, M., Benítez, J. M., Fernández Salgado, E., Sánchez Rodríguez, E., Hernández, V., Ferreiro-Iglesias, R., Díaz, Á. P., Barrio, J., Huguet, J. M., Sicilia, B., Martín-Arranz, M. D., Calvet, X., Ginard, D., … Gisbert, J. P. (2021). Incidence, clinical characteristics and management of inflammatory bowel disease in spain: Large-scale epidemiological study. *Journal of Clinical Medicine*, *10*(13). https://doi.org/10.3390/jcm10132885.

Charrad, M., Ghazzali, N., Laval, U., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set Véronique Boiteau. In *JSS Journal of Statistical* (Vol. 61, Issue 6). http://www.jstatsoft.org/.

Chaussabel, D., Quinn, C., Shen, J., Patel, P., Glaser, C., Baldwin, N., Stichweh, D., Blankenship, D., Li, L., Munagala, I., Bennett, L., Allantaz, F., Mejias, A., Ardura, M., Kaizer, E., Monnet, L., Allman, W., Randall, H., Johnson, D., … Pascual, V. (2008). A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus. *Immunity*, *29*(1), 150–164. https://doi.org/10.1016/j.immuni.2008.05.012.

Cheng, K., & Faye, A. S. (2020). Venous thromboembolism in inflammatory bowel disease. In *World Journal of Gastroenterology* (Vol. 26, Issue 12, pp. 1231–1241). Baishideng Publishing Group Co. https://doi.org/10.3748/WJG.V26.I12.1231.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-019-6413-7.

D'haeseleer, P. (2005). How does gene expression clustering work? In *NATURE BIOTECHNOLOGY* (Vol. 23). http://www.nature.com/naturebiotechnology.

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nature Protocols*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Fu, W., Fu, H., Ye, W., Han, Y., Liu, X., Zhu, S., Li, H., Tang, R., & Wang, Q. (2021). Peripheral blood neutrophil-to-lymphocyte ratio in inflammatory bowel disease and disease activity: A meta-analysis. *International Immunopharmacology*, *101*, 108235. https://doi.org/10.1016/j.intimp.2021.108235.

Gao, L., Zhan, Y., Hu, X., & Liao, S. (2023). Platelet–lymphocyte ratio and lymphocyte–monocyte ratio in inflammatory bowel disease and disease activity: A systematic review and meta-analysis. *Scottish Medical Journal*, *68*(3), 101–109. https://doi.org/10.1177/00369330231188962.

Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz, & Bill Venables. (2016). *Package "gplots" Title Various R Programming Tools for Plotting Data*.

Guan, Q. (2019). A Comprehensive Review and Update on the Pathogenesis of Inflammatory Bowel Disease. *Journal of Immunology Research*, *2019*. https://doi.org/10.1155/2019/7247238.

Iuchi H (2017). _mccr: The Matthews Correlation Coefficient_. R package version 0.4.4, <https://CRAN.R-project.org/package=mccr>.

Juzenas, S., Hubenthal, M., Lindqvist, C. M., Kruse, R., Steiert, T. A., Degenhardt, F., Schulte, D., Nikolaus, S., Zeissig, S., Bergemalm, D., Almer, S., Hjortswang, H., Bresso, F., Struning, N., Kupcinskas, J., Keller, A., Lieb, W., Rosenstiel, P., Schreiber, S., … Franke, A. (2022). Detailed Transcriptional Landscape of Peripheral Blood Points to Increased Neutrophil Activation in Treatment-Naïve Inflammatory Bowel Disease. *Journal of Crohn's and Colitis*, *16*(7), 1097–1109. https://doi.org/10.1093/ecco-jcc/jjac003.

Kabakchiev, B., Turner, D., Hyams, J., Mack, D., Leleiko, N., Crandall, W., Markowitz, J., Otley, A. R., Xu, W., Hu, P., Griffiths, A. M., & Silverberg, M. S. (2010). Gene expression changes

associated with resistance to intravenous corticosteroid therapy in children with severe ulcerative colitis. *PLoS ONE*, *5*(9), 1–8. https://doi.org/10.1371/journal.pone.0013085

Kaplan, G. G. (2015). The global burden of IBD: From 2015 to 2025. In *Nature Reviews Gastroenterology and Hepatology* (Vol. 12, Issue 12, pp. 720–727). Nature Publishing Group. https://doi.org/10.1038/nrgastro.2015.150

Kolde R (2019). *pheatmap: Pretty Heatmaps*. R package version 1.0.12, https://CRAN.R-project.org/package=pheatmap.

Kraszewski, S., Szczurek, W., Szymczak, J., Reguła, M., & Neubauer, K. (2021). Machine learning prediction model for inflammatory bowel disease based on laboratory markers. Working model in a discovery cohort study. *Journal of Clinical Medicine*, *10*(20). https://doi.org/10.3390/jcm10204745.

Kuhn, M. (2008). *Journal of Statistical Software Building Predictive Models in R Using the caret Package*. http://www.jstatsoft.org/.

Lê, S., Josse, J., Rennes, A., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. In *JSS Journal of Statistical Software* (Vol. 25). http://www.jstatsoft.org/.

Levin, A. D., Wildenberg, M. E., & van den Brink, G. R. (2016). Mechanism of action of anti-TNF therapy in inflammatory bowel disease. In *Journal of Crohn's and Colitis* (Vol. 10, Issue 8, pp. 989–997). Oxford University Press. https://doi.org/10.1093/ecco-jcc/jjw053.

Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D. S., Johnson, S. E., Milton, A., Rajam, G., Kasturi, S., Carlone, G. M., Quinn, C., Chaussabel, D., Palucka, A. K., Mulligan, M. J., Ahmed, R., Stephens, D. S., Nakaya, H. I., & Pulendran, B. (2014). Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature Immunology*, *15*(2), 195–204. https://doi.org/10.1038/ni.2789.

Matowicka-Karna, J. (2016). Markers of inflammation, activation of blood platelets and coagulation disorders in inflammatory bowel diseases. *Postępy Higieny i Medycyny Doświadczalnej*, *70*, 305–312. https://doi.org/10.5604/17322693.1199305.

Medina, E. (2013). Actualización Introducción Enfermedad inflamatoria intestinal (I): clasificación, etiología y clínica. In *An Pediatr Contin* (Vol. 11, Issue 2).

Mo, A., Marigorta, U. M., Arafat, D., Chan, L. H. K., Ponder, L., Jang, S. R., Prince, J., Kugathasan, S., Prahalad, S., & Gibson, G. (2018). Disease-specific regulation of gene expression in a comparative analysis of juvenile idiopathic arthritis and inflammatory bowel disease. *Genome Medicine*, *10*(1). https://doi.org/10.1186/s13073-018-0558-x.

Mulder, D. J., Noble, A. J., Justinich, C. J., & Duffin, J. M. (2014). A tale of two diseases: The history of inflammatory bowel disease. In *Journal of Crohn's and Colitis* (Vol. 8, Issue 5, pp. 341–348). Elsevier. https://doi.org/10.1016/j.crohns.2013.09.009.

Netea, M. G., Balkwill, F., Chonchol, M., Cominelli, F., Donath, M. Y., Giamarellos-Bourboulis, E. J., Golenbock, D., Gresnigt, M. S., Heneka, M. T., Hoffman, H. M., Hotchkiss, R., Joosten, L. A. B., Kastner, D. L., Korte, M., Latz, E., Libby, P., Mandrup-Poulsen, T., Mantovani, A., Mills, K. H. G., … Dinarello, C. A. (2017). A guiding map for inflammation. In *Nature*

Immunology (Vol. 18, Issue 8, pp. 826–831). Nature Publishing Group. https://doi.org/10.1038/ni.3790.

Netz, U., Carter, J., Eichenberger, M. R., Feagins, K., Galbraith, N. J., Dryden, G. W., Pan, J., Rai, S. N., & Galandiuk, S. (2018). Plasma microRNA Profile Differentiates Crohn's Colitis from Ulcerative Colitis. *Inflammatory Bowel Diseases*, *24*(1), 159–165. https://doi.org/10.1093/ibd/izx009.

Neuwirth E (2022). _RColorBrewer: ColorBrewer Palettes_. R package version 1.1-3, <https://CRAN.R-project.org/package=RColorBrewer>.

Nowak, J. K., Adams, A. T., Kalla, R., Lindstrøm, J. C., Vatn, S., Bergemalm, D., Keita, Å. V., Gomollón, F., Jahnsen, J., Vatn, M. H., Ricanek, P., Ostrowski, J., Walkowiak, J., Halfvarson, J., Satsangi, J., Andersson, E., Arnott, I. D., Bayes, M., Bonfiglio, F., … You, P. (2022). Characterisation of the Circulating Transcriptomic Landscape in Inflammatory Bowel Disease Provides Evidence for Dysregulation of Multiple Transcription Factors Including NFE2, SPI1, CEBPB, and IRF2. *Journal of Crohn's and Colitis*, *16*(8), 1255–1268. https://doi.org/10.1093/ecco-jcc/jjac033.

Nowak, J. K., Szymańska, C. J., Glapa-Nowak, A., Duclaux-Loras, R., Dybska, E., Ostrowski, J., Walkowiak, J., & Adams, A. T. (2022). Unexpected Actors in Inflammatory Bowel Disease Revealed by Machine Learning from Whole-Blood Transcriptomic Data. *Genes*, *13*(9). https://doi.org/10.3390/genes13091570.

Ordás, I., Eckmann, L., Talamini, M., Baumgart, D. C., & Sandborn, W. J. (2012). Ulcerative colitis. *The Lancet*, *380*(9853), 1606–1619. https://doi.org/10.1016/S0140-6736(12)60150-0.

Ostrowski, J., Goryca, K., Lazowska, I., Rogowska, A., Paziewska, A., Dabrowska, M., Ambrozkiewicz, F., Karczmarski, J., Balabas, A., Kluska, A., Piatkowska, M., Zeber-Lubecka, N., Kulecka, M., Habior, A., Mikula, M., Walewska-Zielecka, B., Krawczyk, M., Cichoz-Lach, H., Milkiewicz, P., … Starzynska, T. (2019). Common functional alterations identified in blood transcriptome of autoimmune cholestatic liver and inflammatory bowel diseases. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-43699-1.

Palmer, N. P., Silvester, J. A., Lee, J. J., Beam, A. L., Fried, I., Valtchinov, V. I., Rahimov, F., Kong, S. W., Ghodoussipour, S., Hood, H. C., Bousvaros, A., Grand, R. J., Kunkel, L. M., & Kohane, I. S. (2019). Concordance between gene expression in peripheral whole blood and colonic tissue in children with inflammatory bowel disease. *PLoS ONE*, *14*(10). https://doi.org/10.1371/journal.pone.0222952.

Park, S. K., Kim, Y. B., Kim, S., Lee, C. W., Choi, C. H., Kang, S. B., Kim, T. O., Bang, K. B., Chun, J., Cha, J. M., Im, J. P., Kim, M. S., Ahn, K. S., Kim, S. Y., & Park, D. Il. (2022). Development of a Machine Learning Model to Predict Non-Durable Response to Anti-TNF Therapy in Crohn's Disease Using Transcriptome Imputed from Genotypes. *Journal of Personalized Medicine*, *12*(6). https://doi.org/10.3390/jpm12060947.

Pavlidis, S., Monast, C., Loza, M. J., Branigan, P., Chung, K. F., Adcock, I. M., Guo, Y., Rowe, A., & Baribaud, F. (2019). I_MDS: An inflammatory bowel disease molecular activity score to classify patients with differing disease-driving pathways and therapeutic response to anti-

TNF treatment. *PLoS Computational Biology*, *15*(4). https://doi.org/10.1371/journal.pcbi.1006951.

Pelé, J., Bécu, J. M., Abdi, H., & Chabbert, M. (2012). Bios2mds: An R package for comparing orthologous protein families by metric multidimensional scaling. *BMC Bioinformatics*, *13*(1). https://doi.org/10.1186/1471-2105-13-133.

Peters, L. A., Perrigoue, J., Mortha, A., Iuga, A., Song, W. M., Neiman, E. M., Llewellyn, S. R., Di Narzo, A., Kidd, B. A., Telesco, S. E., Zhao, Y., Stojmirovic, A., Sendecki, J., Shameer, K., Miotto, R., Losic, B., Shah, H., Lee, E., Wang, M., … Schadt, E. E. (2017). A functional genomics predictive network model identifies regulators of inflammatory bowel disease. *Nature Genetics*, *49*(10), 1437–1449. https://doi.org/10.1038/ng.3947.

Planell, N., Masamunt, M. C., Leal, R. F., Rodríguez, L., Esteller, M., Lozano, J. J., Ramírez, A., Ayrizono, M. de L. S., Coy, C. S. R., Alfaro, I., Ordás, I., Visvanathan, S., Ricart, E., Guardiola, J., Panés, J., & Salas, A. (2017). Usefulness of transcriptional blood biomarkers as a non-invasive surrogate marker of mucosal healing and endoscopic response in ulcerative colitis. *Journal of Crohn's and Colitis*, *11*(11), 1335–1346. https://doi.org/10.1093/ecco-jcc/jjx091.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Roda, G., Chien Ng, S., Kotze, P. G., Argollo, M., Panaccione, R., Spinelli, A., Kaser, A., Peyrin-Biroulet, L., & Danese, S. (2020). Crohn's disease. *Nature Reviews Disease Primers*, *6*(1). https://doi.org/10.1038/s41572-020-0156-2.

Rousseeuw, P. J. (1987). Silhouet tes: a graphic al aid to the interpre tation and validati on of cluster analysis. In *Journal of Computational and Applied Mathematics* (Vol. 20).

Saez, A., Herrero-Fernandez, B., Gomez-Bris, R., Sánchez-Martinez, H., & Gonzalez-Granado, J. M. (2023). Pathophysiology of Inflammatory Bowel Disease: Innate Immune System. In *International Journal of Molecular Sciences* (Vol. 24, Issue 2). MDPI. https://doi.org/10.3390/ijms24021526.

Schett, G., & Neurath, M. F. (2018). Resolution of chronic inflammatory disease: universal and tissue-specific concepts. In *Nature Communications* (Vol. 9, Issue 1). Nature Publishing Group. https://doi.org/10.1038/s41467-018-05800-6.

Sean, D., & Meltzer, P. S. (2007). GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, *23*(14), 1846–1847. https://doi.org/10.1093/bioinformatics/btm254.

Tarazona, S., Furió-Tarí, P., Turrà, D., Di Pietro, A., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, *43*(21). https://doi.org/10.1093/nar/gkv711.

Toro-Domínguez, D., Martorell-Marugán, J., Martinez-Bueno, M., López-Domínguez, R., Carnero-Montoro, E., Barturen, G., Goldman, D., Petri, M., Carmona-Sáez, P., & Alarcón-Riquelme, M. E. (2022). Scoring personalized molecular portraits identify Systemic Lupus Erythematosus subtypes and predict individualized drug responses, symptomatology and

disease progression. *Briefings in Bioinformatics*, *23*(5). https://doi.org/10.1093/bib/bbac332.

Ungaro, R. C., Hu, L., Ji, J., Nayar, S., Kugathasan, S., Denson, L. A., Hyams, J., Dubinsky, M. C., Sands, B. E., & Cho, J. H. (2021). Machine learning identifies novel blood protein predictors of penetrating and stricturing complications in newly diagnosed paediatric Crohn's disease. *Alimentary Pharmacology and Therapeutics*, *53*(2), 281–290. https://doi.org/10.1111/apt.16136.

Ventham, N. T., Kennedy, N. A., Adams, A. T., Kalla, R., Heath, S., O'Leary, K. R., Drummond, H., Wilson, D. C., Gut, I. G., Nimmo, E. R., Satsangi, J., Lauc, G., Campbell, H., McGovern, D. P. B., Annese, V., Zoldoš, V., Permberton, I. K., Wuhrer, M., Kolarich, D., … Vatn, M. H. (2016). Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms13507.

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.

Yang, K., Wang, H., Dai, G., Hu, S., Zhang, Y., & Xu, J. (2011). Determining the repeat number of cross-validation. *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI).*