



**Universidad  
Europea**

## **Máster en Bioinformática**

**Desarrollo de un nuevo score transcriptómico para medir  
la actividad de la enfermedad en pacientes con Lupus  
Eritematoso Sistémico mediante análisis bioinformático  
en R**

**Autor: Ana Viñuela de Vicente**

**Tutor: Daniel Toro Domínguez**

**Curso 2024-25**

## AGRADECIMIENTOS

Este trabajo se lo agradezco a mi familia, que siempre confió en mí y me ayudó en cada paso. Su apoyo y paciencia han sido clave para llegar hasta aquí.

# ÍNDICE

|                                                                 |           |
|-----------------------------------------------------------------|-----------|
| <b>1. Introducción</b>                                          | <b>8</b>  |
| <b>2. Hipótesis</b>                                             | <b>9</b>  |
| <b>3. Objetivos</b>                                             | <b>9</b>  |
| <b>4. Metodología</b>                                           | <b>10</b> |
| <b>4.1. Diseño y estudio de la población</b>                    | <b>11</b> |
| <b>4.2. Población</b>                                           | <b>11</b> |
| <b>4.3. Número de pacientes estudiado</b>                       | <b>12</b> |
| <b>4.4. Ámbito y tiempo</b>                                     | <b>13</b> |
| <b>4.5. Criterios de inclusion</b>                              | <b>13</b> |
| <b>4.6. Criterios de exclusion</b>                              | <b>14</b> |
| <b>4.7. Cálculo del tamaño muestral</b>                         | <b>14</b> |
| <b>4.8. Variables</b>                                           | <b>14</b> |
| <b>4.9. Recogida y preprocesamiento de los datos</b>            | <b>15</b> |
| <b>4.10. Análisis estadístico</b>                               | <b>15</b> |
| <b>4.11. Herramientas bioinformáticas y entornos de trabajo</b> | <b>16</b> |
| <b>5. Resultados</b>                                            | <b>16</b> |
| <b>6. Discusión</b>                                             | <b>19</b> |
| <b>7. Conclusiones</b>                                          | <b>22</b> |
| <b>8. Bibliografía</b>                                          | <b>22</b> |
| <b>9. Anexos</b>                                                | <b>24</b> |

## RESUMEN

El lupus eritematoso sistémico (LES) es una enfermedad autoinmune compleja y heterogénea, cuya actividad se evalúa habitualmente mediante el índice clínico SLEDAI. Sin embargo, este enfoque se centra en manifestaciones clínicas y parámetros de laboratorio, dejando fuera alteraciones moleculares que pueden preceder a los brotes. En este trabajo proponemos un score transcriptómico complementario a SLEDAI que capta con mayor sensibilidad estados de activación inmunitaria.

### Objetivos

El objetivo general fue diseñar un nuevo score bioinformático basado en datos transcriptómicos que refleje la actividad del LES. Los objetivos específicos fueron: (1) identificar firmas génicas asociadas a la actividad clínica, (2) construir y evaluar scores moleculares a nivel de firmas génicas y rutas funcionales (*singscore/ssGSEA/GSVA*), (3) explorar la presencia de actividad subclínica en pacientes con SLEDAI bajo, y (4) integrar un panel flexible de genes no-IFN con un subpanel de genes IFN para generar un score híbrido con mayor capacidad de generalización.

### Material y métodos

Se analizaron dos cohortes públicas: una de pacientes pediátricos (Banchereau et al., GSE65391) empleada como conjunto de entrenamiento, y otra adulta (Walsh et al., SPARE study) utilizada para validación. Tras la limpieza y normalización de los datos de expresión génica, se aplicaron métodos de selección de genes basados en correlación con SLEDAI y variabilidad intra-grupo. Se calcularon scores moleculares con *singscore*, *ssGSEA* y *GSVA* con el paquete PathMED. Para explorar subgrupos dentro de pacientes con alta actividad clínica, se utilizaron algoritmos de clustering no supervisado (*k-means*). Finalmente, se construyó un score híbrido integrando genes de interferón tipo I (IFN-I) y un panel flexible no-IFN, optimizando el parámetro de ponderación  $\alpha$  en función del rendimiento en ambas cohortes.

## Resultados

En la cohorte pediátrica (entrenamiento), la firma dominada por IFN-I mostró una asociación robusta con SLEDAI y discriminó SLE frente a controles sanos (NHV). Al pasar a rutas funcionales (ssGSEA/GSVA) observamos mejor generalización en la cohorte adulta (validación) y, además, actividad molecular residual en pacientes clínicamente inactivos (SLEDAI  $<2$  y también  $\leq 4$ ) cuando se comparan inactivos vs NHV; estas mismas rutas aparecen más intensificadas en SLE con alta actividad, evidenciando un gradiente NHV  $\rightarrow$  inactivo/bajo  $\rightarrow$  alto. Con un panel flexible no-IFN identificamos subgrupos moleculares dentro del SLE alto. Finalmente, el score híbrido IFN+Flexible mantuvo la correlación con SLEDAI en pediátricos, mejoró la generalización en adultos y discriminó SLE vs NHV; la separación alto vs bajo fue más limitada.

## Conclusiones

Los resultados confirman que es posible diseñar un marcador transcriptómico que complementa al SLEDAI, captando tanto mecanismos clásicos como la heterogeneidad molecular del LES. El singscore individual destaca por su robustez, mientras que el análisis funcional aporta interpretabilidad. El hallazgo de actividad residual en pacientes clínicamente inactivos subraya la utilidad de incorporar datos transcriptómicos en el seguimiento del LES. El score híbrido desarrollado constituye un candidato prometedor para estudios futuros, orientados a la validación longitudinal y a la integración en la práctica clínica.

## **1. Introducción**

El Lupus Eritematoso Sistémico (LES) es una enfermedad autoinmune crónica caracterizada por una gran heterogeneidad clínica y molecular, que afecta de forma impredecible a múltiples órganos y sistemas. Esta complejidad dificulta su seguimiento y estratificación precisa, lo que convierte la monitorización de la actividad de la enfermedad en un reto clínico considerable. Aunque existen escalas clínicas consolidadas como el índice SLEDAI (Systemic Lupus Erythematosus Disease Activity Index), ampliamente utilizado en la práctica médica, su capacidad para reflejar actividad subclínica es limitada. El SLEDAI se basa exclusivamente en parámetros clínicos y bioquímicos observables, lo que puede dejar sin detectar alteraciones moleculares tempranas o persistentes, especialmente en pacientes aparentemente estables.

En este contexto, surge la necesidad de herramientas complementarias que integren información transcriptómica, permitiendo una evaluación más sensible, precisa y personalizada del estado inmunológico del paciente. La actividad del LES no se limita a los síntomas visibles, sino que también se manifiesta en desregulaciones profundas a nivel de expresión génica en las células del sistema inmunitario. Estas alteraciones pueden preceder a los brotes clínicos o persistir en estados de aparente remisión, lo que subraya la importancia de incorporar la dimensión molecular en la evaluación de la enfermedad.

Este trabajo se enmarca dentro de ese esfuerzo y propone el desarrollo de un nuevo score transcriptómico, basado en la expresión génica individual y funcional, utilizando métodos bioinformáticos reproducibles en R. Se exploran diferentes estrategias, como el uso de firmas génicas con singscore y el enriquecimiento funcional mediante ssGSEA y GSVA con el paquete PathMED, evaluando su correlación con SLEDAI, su capacidad de generalización en cohortes independientes, y su potencial utilidad clínica. Un score de estas características no solo permitiría detectar actividad residual o subclínica, sino que también podría optimizar las decisiones terapéuticas y facilitar una medicina más proactiva y personalizada en pacientes con LES.

## 2. Hipótesis

La actividad del Lupus Eritematoso Sistémico (LES) no se limita únicamente a los síntomas clínicos visibles, sino que también se refleja en alteraciones profundas de la expresión génica en las células del sistema inmunitario. Aunque el índice clínico SLEDAI constituye una herramienta de referencia en la práctica médica, pacientes con valores similares pueden mostrar perfiles moleculares y respuestas terapéuticas divergentes, lo que pone de manifiesto que este índice no capta completamente la heterogeneidad biológica de la enfermedad.

Partimos de la hipótesis de que un score transcriptómico, construido a partir de datos de expresión génica y diseñado para integrar tanto la señal característica del interferón tipo I como otros módulos moleculares complementarios, permitirá reflejar con mayor fidelidad el estado inmunológico del paciente. Este marcador bioinformático podría ofrecer una visión más sensible y personalizada de la actividad del LES, detectar actividad subclínica que escapa a las métricas convencionales y, en última instancia, contribuir a optimizar el seguimiento clínico y la toma de decisiones terapéuticas.

## 3. Objetivos

El objetivo principal de la investigación es diseñar, desarrollar y validar un score bioinformático híbrido basado en datos transcriptómicos que permita cuantificar de forma más precisa la actividad de la enfermedad en pacientes con LES, complementando la información clínica proporcionada por el índice SLEDAI.

Para cumplir este objetivo final, hay varios subapartados que hay que cumplir:

- 1) **Selección y procesamiento de datos:** Investigar bases de datos transcriptómicas públicas que incluyan pacientes con LES y controles sanos, realizar el preprocesamiento y normalización de las matrices de expresión, y aplicar estrategias que permitan la comparación de resultados derivados entre cohortes pediátricas y adultas.

- 2) **Construcción de un panel flexible de genes:** Detectar genes asociados con la actividad clínica (SLEDAI) y con alta variabilidad en pacientes con LES activo.
- 3) **Evaluación de scores transcriptómicos:** Aplicar y comparar diferentes métodos de scoring para estimar la actividad transcriptómica en cada muestra. Analizar su grado de correlación con SLEDAI y su capacidad para discriminar entre pacientes con LES y controles sanos.
- 4) **Análisis del sesgo por interferón tipo I:** Examinar el peso de la firma de interferón tipo I en los paneles generados y evaluar estrategias para limitar su efecto dominante. Este paso es necesario porque la señal de interferón tipo I, aunque biológicamente relevante, puede dominar la variabilidad transcriptómica del LES e impedir la identificación de otros mecanismos activos.
- 5) **Desarrollo de un score híbrido:** Diseñar un score combinado que integre un subpanel IFN reducido y el panel flexible no-IFN, definiendo una ponderación óptima entre ambos componentes.
- 6) **Validación externa y estratificación molecular:** Validar el score híbrido en una cohorte independiente de pacientes adultos, evaluando su correlación con SLEDAI, su capacidad de discriminar entre pacientes y controles, y su utilidad para identificar subgrupos moleculares dentro de los pacientes con alta actividad clínica.

#### **4. Metodología**

Este subestudio se enmarca dentro de un proyecto más amplio centrado en el desarrollo de un marcador molecular que sirva como alternativa al índice clínico SLEDAI (Systemic Lupus Erythematosus Disease Activity Index). El objetivo principal es identificar una medida objetiva, reproducible y basada en expresión génica que permita evaluar la actividad del lupus eritematoso sistémico (LES), ya sea como complemento o incluso como posible sustituto del score clínico tradicional. Para

lograrlo, se han comparado distintas estrategias bioinformáticas de forma independiente, evaluando su capacidad para correlacionarse con SLEDAI y discriminar entre pacientes con LES y controles sanos.

- **Diseño y estudio de la población**

Este estudio se basa en dos cohortes públicas previamente descritas en el artículo de Walsh et al. (2021, *Frontiers in Immunology*) [PMCID: PMC8326877], seleccionadas para permitir la exploración longitudinal de la actividad de la enfermedad (LES) mediante expresión génica.

- **Cohorte de entrenamiento (clin.dat1):** procede del estudio de Banchereau et al., accesible públicamente en el repositorio GEO (ID: **GSE65391**). En dicho estudio se incluyeron 80 pacientes pediátricos con múltiples visitas longitudinales, datos transcriptómicos obtenidos mediante microarrays, y variables clínicas como el índice de actividad SLEDAI.
- **Cohorte de validación (clin.dat2):** corresponde a pacientes adultos del estudio SPARE (*Study of biological Pathways, disease Activity and Response markers in patients with systemic lupus erythematosus*), incluidos en el *Johns Hopkins Lupus Cohort*. En el artículo de Walsh et al., se utilizaron datos de 65 pacientes adultos con seguimiento longitudinal. Aunque los datos de expresión génica están disponibles públicamente, no se encuentran completamente accesibles a través de GEO, sino que fueron proporcionados por los autores.

Ambas cohortes incluyen muestras de pacientes con LES (SLE) y de controles sanos (NHV, *Normal Healthy Volunteers*), con sus respectivas matrices de expresión génica.

- **Población**

La muestra analizada está compuesta por pacientes con diagnóstico confirmado de lupus eritematoso sistémico (LES), junto con controles sanos (NHV) que participaron como grupo comparador en los estudios originales. Todos los datos provienen de investigaciones publicadas por centros internacionales, y han sido recogidos siguiendo los criterios definidos por los autores de cada estudio.

A continuación, se resumen los criterios de selección aplicados por los autores de cada estudio:

- **Cohorte pediátrica (Banchereau et al., GSE65391 – cohorte de entrenamiento):**
  - Pacientes con diagnóstico confirmado de LES pediátrico.
  - Inclusión de múltiples visitas longitudinales por paciente.
  - Evaluación clínica mediante SLEDAI y otras variables inmunológicas.
  - Disponibilidad de datos transcriptómicos completos (microarrays Affymetrix).
  - Para este trabajo, se seleccionaron únicamente aquellos pacientes con  $\geq 3$  visitas y variabilidad en SLEDAI a lo largo del tiempo.
- **Cohorte adulta (Walsh et al., SPARE study – cohorte de validación):**
  - Pacientes adultos (18–75 años) con LES confirmado según los criterios revisados del American College of Rheumatology (1997).
  - Seguimiento longitudinal durante al menos un año, con visitas periódicas y evaluación constante de la actividad mediante SLEDAI.
  - Recolección de información clínica, medicación, análisis de laboratorio y datos transcriptómicos.
  - Para este trabajo, se seleccionaron pacientes con  $\geq 3$  visitas y variabilidad clínica suficiente, excluyendo casos con poca o nula fluctuación de la enfermedad.

Los **controles sanos** fueron reclutados bajo protocolos específicos de cada estudio, y no presentaban enfermedades autoinmunes ni inflamatorias. Sus datos de expresión

génica se emplean como referencia para comparar perfiles moleculares con los pacientes con LES.

- **Número de pacientes estudiado**

Para este trabajo, se aplicaron **filtros adicionales** a las cohortes descritas previamente, con el objetivo de garantizar la calidad y la utilidad de los datos analizados. En concreto, se incluyeron únicamente aquellos pacientes que:

- Tuvieran al menos **tres visitas longitudinales**
- Presentaran **variación temporal en la actividad clínica** (SLEDAI)
- Contaran con **información clínica y transcriptómica completa**

Tras este filtrado, los conjuntos de datos finales utilizados fueron los siguientes:

- **clin.dat1** (cohorte pediátrica, entrenamiento): 111 pacientes con un total de 857 muestras longitudinales, y 206 variables clínicas por muestra. Matriz de expresión: SLE.dat1 y NHV.dat1, con 12.445 genes.
- **clin.dat2** (cohorte adulta, validación): 84 pacientes con un total de 358 muestras longitudinales, y 94 variables clínicas por muestra. Matriz de expresión: SLE.dat2 y NHV.dat2, con 22.048 genes.

Este conjunto representa una versión depurada de los datos originalmente incluidos en el artículo de Walsh et al., adaptada específicamente para el desarrollo y validación de scores moleculares transcriptómicos aplicables al seguimiento clínico del LES.

- **Ámbito y tiempo**

Dado que se trata de un análisis de datos provenientes de distintos estudios internacionales, el ámbito es global. El análisis bioinformático se llevó a cabo entre los

meses de mayo y julio de 2025, utilizando muestras originalmente publicadas entre los años 2017 y 2022.

- **Criterios de de selección (inclusión/exclusión)**

Para este estudio se seleccionaron:

- Pacientes con diagnóstico confirmado de lupus eritematoso sistémico (LES).
- Sujetos sanos (NHV: Normal Healthy Volunteers) que participaron como controles en los estudios originales.
- Muestras con datos de expresión génica de calidad adecuada y con información clínica disponible.
- Al menos tres visitas por paciente y variación en el valor de SLEDAI en el tiempo.

Se excluyeron todas las muestras o pacientes que no cumplieran cualquiera de los criterios anteriores; en particular:

- registros con datos clínicos críticos ausentes o incompletos;
- pacientes con < 3 visitas o sin variación apreciable del SLEDAI en el tiempo.

- **Variables**

- **Variable dependiente:** el índice clínico de actividad de la enfermedad SLEDAI.
- **Variables independientes:** niveles de expresión génica (a nivel de genes individuales y de rutas funcionales), y grupo clínico de pertenencia (paciente con LES vs control sano), empleado en los análisis de clasificación. En algunos casos también se registraron variables demográficas como sexo y edad, aunque no fueron utilizadas para construir los scores moleculares.

- **Recogida y preprocesamiento de los datos**

Los datos transcriptómicos se obtuvieron en forma de matrices de expresión microarray previamente normalizadas, descargadas desde la plataforma GEO. Estas matrices se transformaron mediante logaritmicación, y se integró la información clínica a partir de los ficheros de anotación disponibles. Todo el procesamiento y la limpieza de datos se llevó a cabo en R, utilizando principalmente herramientas del paquete `tidyverse`, lo que permitió asegurar un formato homogéneo y fiable para el análisis.

- **Análisis estadístico**

Se evaluó la relación entre los distintos scores moleculares y el índice clínico SLEDAI mediante correlaciones de Spearman, dado que este enfoque no asume linealidad y resulta más apropiado para datos clínicos. Adicionalmente, se utilizaron pruebas no paramétricas (Wilcoxon) para comparar grupos clínicos (alto vs bajo SLEDAI, pacientes vs controles sanos).

También se aplicaron algoritmos de clustering no supervisado (k-means) para explorar subgrupos moleculares dentro de los pacientes con alta actividad clínica (SLEDAI elevado).

Finalmente, se diseñó un **score híbrido** que integra un subpanel reducido de genes de interferón tipo I (IFN-I) y un panel flexible no-IFN. Se calculó como una combinación ponderada de ambos componentes, ajustando el parámetro de ponderación  $\alpha$  para maximizar la correlación con SLEDAI en adultos sin comprometer el rendimiento en la cohorte pediátrica.

- **Herramientas bioinformáticas y entornos de trabajo**

El análisis se llevó a cabo íntegramente en el entorno de programación R (versión 4.x), utilizando RStudio como interfaz de desarrollo. Para abordar los distintos objetivos del

estudio, se emplearon diversas herramientas y librerías bioinformáticas especializadas, entre las que destacan:

- **limma**: paquete ampliamente validado para el análisis de expresión diferencial.
- **tidyverse**, **dplyr** y **ggplot2**: herramientas clave para el procesamiento, manipulación y visualización de datos.
- **Biobase** y **GSEABase**: esenciales para el manejo de objetos del tipo `ExpressionSet` y la gestión de rutas funcionales como las de Gene Ontology (GO).
- **Pheatmap**: para gráficos heatmaps.
- **matrixStats**: para calcular varianzas y ranks de genes.
- **Caret**: para la validación cruzada.
- **Msigdb**: para descargar colecciones de genes (GO:BP) en enriquecimiento funcional.
- **pathMED**: Los scores (*singscore*, **GSVA** y **ssGSEA**) se calcularon con el paquete `pathMED` (`getScores()`).

## 5. Resultados

### Primera aproximación: genes individuales dominados por interferón

El primer enfoque consistió en aplicar el método *singscore* (calculado con *pathMED*) para construir un marcador individual basado en la expresión de genes seleccionados por su correlación con la actividad clínica (SLEDAI). A partir de la cohorte pediátrica, se identificó una firma génica fuertemente enriquecida en genes estimulados por interferón tipo I, entre los que destacan IFI44L, ISG15, MX1, IFIT3 u OAS1. Esta firma mostró una correlación elevada con el índice SLEDAI ( $\rho \approx 0.483$ ,  $p < 1e-12$ ), lo que la consolidó como una primera aproximación sólida.

Además, el singscore permitió diferenciar de forma clara a pacientes con LES frente a controles sanos (Figura 1), lo que sugiere que captura procesos moleculares robustos asociados a la enfermedad. La validación cruzada reforzó estos resultados, demostrando la estabilidad del score y su resistencia a la variabilidad en la selección de muestras.

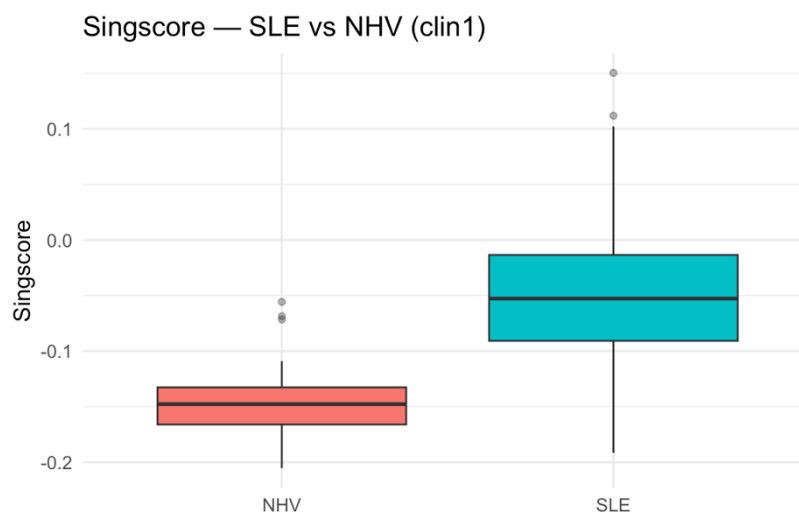


Figura 1: Boxplot de singscore SLE vs NHV (clin1)

Sin embargo, se evidenció una limitación importante: la señal estaba dominada casi en exclusiva por genes relacionados con el interferón, lo que, aunque refleja un mecanismo patológico bien descrito en LES, reduce la capacidad del marcador para captar la heterogeneidad molecular de la enfermedad. Este hallazgo motivó la búsqueda de aproximaciones más amplias.

#### Complemento funcional: rutas biológicas y actividad residual

Con el objetivo de superar la dependencia de genes individuales, se aplicaron enfoques funcionales basados en el enriquecimiento de rutas biológicas (*ssGSEA* y *GSVA*) calculados con *pathMED*. Las rutas correspondieron a Gene Ontology—Biological Process (GO:BP), obtenidas de la base pública MSigDB (colección C5, *Homo sapiens*) mediante el paquete *msigdbR*. Los datos de expresión proceden de *Longitudinal*

*stratification of gene expression reveals three SLE groups of disease activity progression* (Banchereau et al., 2016).

En la cohorte pediátrica (clin1) se evaluaron 4.228 rutas; la mediana de  $|\rho|$  (Spearman) entre los scores de ruta y SLEDAI fue 0,078 (IQR: 0,038–0,129), y 3.262 rutas mostraron  $FDR < 0,05$  frente a NHV. A nivel funcional, emergieron rutas de síntesis proteica/biogénesis ribosomal, procesamiento de ARN y targeting cotraduccional a membrana/ER, que mostraron asociación con la actividad clínica. Con estas rutas se construyó un score funcional agregado (promedio de las 8 rutas con mayor  $|\rho|$  y  $FDR < 0,05$ ), que presentó una correlación estable con SLEDAI en pediátricos y adultos ( $\rho \approx 0,30$ – $0,40$ ), con buena generalización inter-cohorte.

Un hallazgo particularmente relevante surgió al analizar pacientes clínicamente inactivos (SLEDAI  $< 2$ ). El análisis funcional reveló que existe una actividad molecular persistente en estos pacientes. Se identificaron numerosas rutas biológicas alteradas en comparación con los controles sanos, muchas de ellas relacionadas con procesos de síntesis proteica, biogénesis ribosomal y procesamiento de ARN.

Lo más relevante fue comprobar que estas mismas rutas también estaban activas en los pacientes con alta actividad clínica, pero con una intensidad considerablemente mayor. Es decir, los mecanismos moleculares que se observan en la fase activa del lupus ya están presentes, aunque de forma más leve, en los pacientes inactivos.

Esta observación refuerza la idea de que el estado “inactivo” según las métricas clínicas no implica una completa normalización inmunológica, sino que puede coexistir con una actividad subclínica residual a nivel transcriptómico. Desde una perspectiva clínica, esto podría explicar por qué algunos pacientes aparentemente estables acaban experimentando brotes posteriores.

El boxplot del score funcional muestra claramente esta transición, con un gradiente progresivo de señal entre controles, inactivos y pacientes con lupus activo (Fig. 2a). De forma complementaria, el diagrama comparativo de rutas (Fig. 2b) evidencia que la mayoría de los procesos que ya están alterados en los inactivos se intensifican en el

grupo de alta actividad, lo que sugiere una continuidad molecular entre ambos estados.

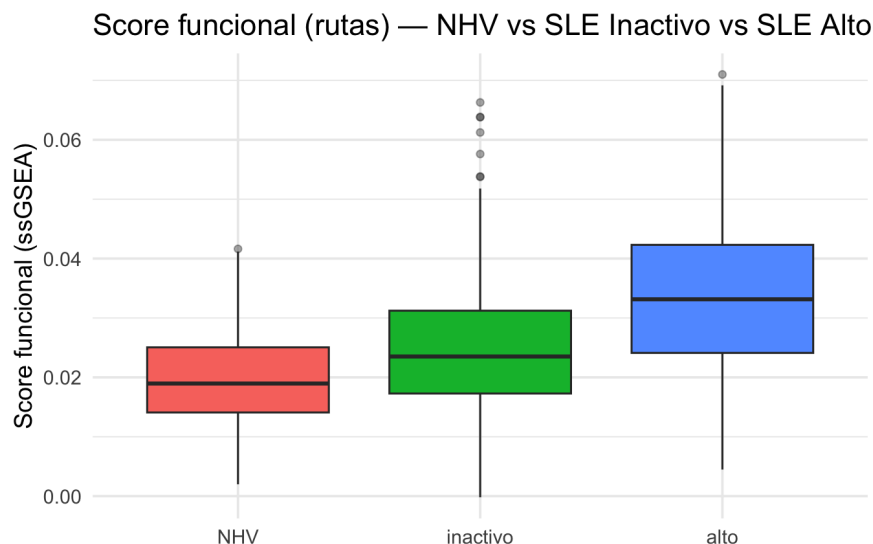


Figura 2a: Boxplot del score funcional entre NHV, SLE inactivo y SLE alto.

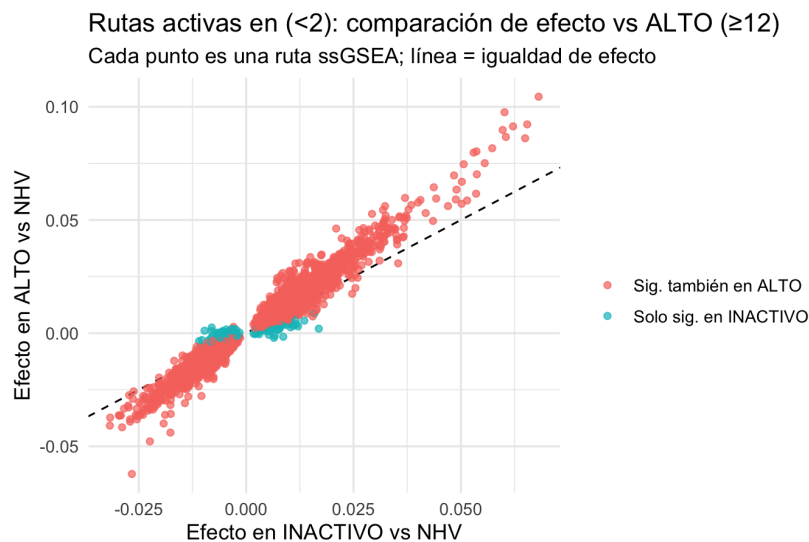


Figura 2b. Comparación magnitud de las rutas activas (inactivos vs activos).

Finalmente, el análisis se amplió al grupo de pacientes con actividad baja ( $\text{SLEDAI} \leq 4$ ), y se observó un patrón muy similar. Los resultados mostraron una expresión diferencial significativa respecto a los controles sanos, aunque con menor intensidad que en el

grupo de alta actividad. Esto confirma que la actividad biológica no desaparece tras la mejoría clínica, sino que disminuye gradualmente a medida que el paciente se estabiliza.

En conjunto, tanto el grupo inactivo como el de baja actividad comparten la activación de rutas relacionadas con la traducción, el metabolismo del ARN y la producción de proteínas, las cuales se intensifican a medida que la enfermedad progresa.

Estos hallazgos sugieren que el seguimiento molecular mediante scores funcionales podría anticipar la reactivación clínica, permitiendo una evaluación más sensible del estado inmunológico real de los pacientes.

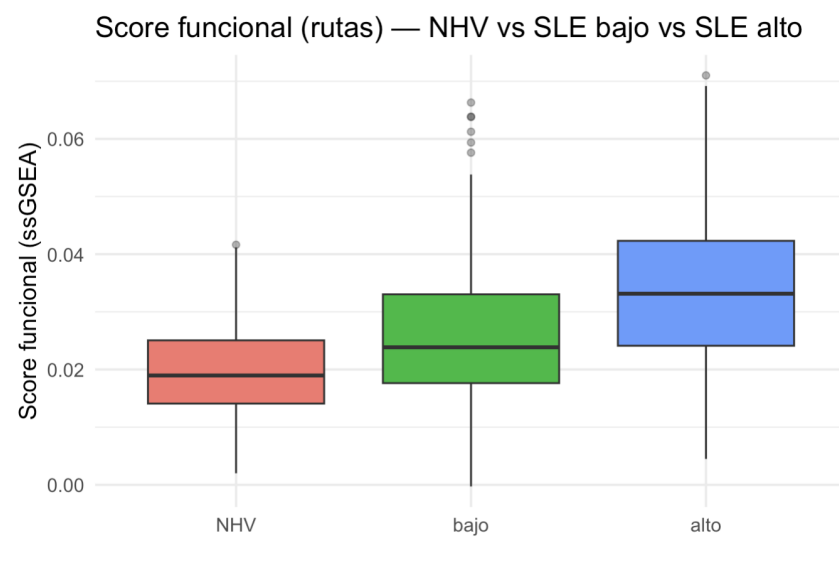


Figura 3: Boxplot del score funcional entre NHV, SLE bajo y SLE alto.

### Hacia un marcador flexible y adaptable

Para capturar la heterogeneidad observada en los pacientes con SLEDAI alto, se diseñó un panel transcriptómico flexible. A partir de un conjunto inicial de 200 genes candidatos, seleccionados en función de su correlación con SLEDAI y su variabilidad en

pacientes de alta actividad, se aplicaron procedimientos de estabilidad (bootstrap) que condujeron a un panel refinado de 80 genes.

Este marcador flexible mostró una correlación significativa con la actividad clínica en la cohorte pediátrica ( $p \approx 0.34$ ,  $p < 1e-4$ ) y, lo más interesante, permitió identificar subgrupos moleculares dentro de los pacientes con actividad elevada. En concreto, se detectaron dos subpoblaciones (A1 y A2) con perfiles transcriptómicos divergentes, lo que evidencia que un mismo nivel de SLEDAI puede corresponder a mecanismos biológicos distintos. Este hallazgo apoya la necesidad de herramientas moleculares capaces de estratificar pacientes más allá de lo que permite la evaluación clínica. La heterogeneidad se visualiza de forma clara en el heatmap de la firma flexible (Figura 4), donde se aprecian patrones diferenciados de activación génica en los pacientes con SLEDAI alto.

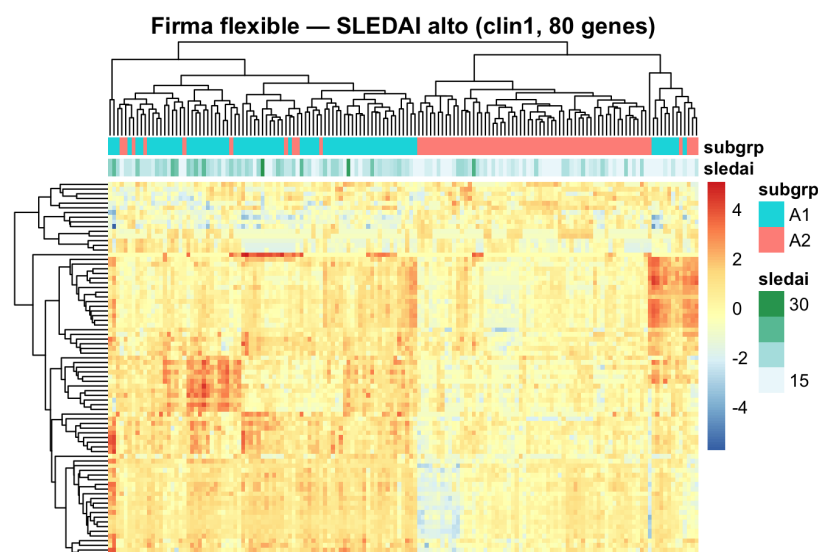


Figura 4: Heatmap de la firma flexible en pacientes con SLEDAI alto (clin1).

La validación en la cohorte adulta reforzó la utilidad de este enfoque. El score flexible alcanzó una correlación de  $p \approx 0.42$  con SLEDAI y distinguió de forma significativa a pacientes con LES frente a controles sanos ( $p \approx 5e-04$ ). Además, la identificación de subgrupos en adultos confirmó que la heterogeneidad molecular no es exclusiva de la población pediátrica, sino que también está presente en el LES adulto.

### Integración final: un score híbrido IFN + flexible

Tras analizar por separado el *singscore* dominado por interferón tipo I (IFN-I) y el panel flexible no-IFN, surgió la necesidad de integrar ambos enfoques en un modelo combinado que maximizase las ventajas de cada uno y minimizase sus limitaciones. El razonamiento detrás de esta integración fue doble. Por un lado, los genes de interferón aportaban una señal extremadamente robusta y reproducible, presente de manera transversal en cohortes y estudios previos, pero limitada en su capacidad para explicar toda la diversidad molecular del lupus. Por otro lado, el panel flexible de genes no-IFN, aunque más heterogéneo, ofrecía información complementaria sobre rutas y mecanismos adicionales relevantes en la patogenia del LES.

Para ello, se seleccionó un subpanel reducido de 7 genes IFN-I (IFI27, IFI44, IFI44L, ISG15, IFITM1, USP18 y RTP4), representativos del eje interferón, y un panel flexible de 52 genes no-IFN definido tras procesos de selección, bootstrap y filtrado por estabilidad. La combinación de ambos se operacionalizó mediante un score híbrido definido como:

$$\text{Score híbrido} = \alpha \cdot \text{Score IFN} + (1 - \alpha) \cdot \text{Score flexible}, \quad \alpha \in (0, 1)$$

Aquí, *Score IFN* cuantifica por muestra la activación del eje de interferón, mientras que *Score FLEX* resume la señal no-IFN del panel flexible; ambos se reescalan previamente a [0,1] para garantizar comparabilidad. El parámetro  $\alpha$  regula la contribución relativa: valores altos privilegian IFN y valores bajos enfatizan el componente flexible. A título ilustrativo, si tras el reescalado una muestra A presenta *Score\_IFN* = 0,80 y *Score\_FLEX* = 0,40, y otra muestra B *Score\_IFN* = 0,20 y *Score\_FLEX* = 0,70, al fijar  $\alpha = 0,25$  (25% IFN, 75% FLEX) se obtiene Híbrido(A) = 0,50 y Híbrido(B) = 0,58, lo que evidencia cómo el índice equilibra la robustez del eje IFN con la diversidad informativa no-IFN. El valor de  $\alpha$  se ajustó de manera sistemática explorando el intervalo [0,1] y eligiendo el que maximizó la correlación con SLEDAI en adultos (cohorte de validación, *clin2*),

manteniendo al mismo tiempo un rendimiento mínimo en pediátricos (entrenamiento, *clin1*).

El análisis reveló que el valor óptimo era  $\alpha = 0.25$ , lo que corresponde a asignar un 25% de peso a la señal IFN y un 75% al panel flexible. Con esta configuración, el score híbrido alcanzó una correlación de  $\rho \approx 0.32$  en pediátricos y de  $\rho \approx 0.41$  en adultos, superando los resultados de cualquiera de los enfoques individuales en términos de equilibrio y generalización. En la cohorte adulta, el score híbrido mostró una correlación robusta con SLEDAI (Figura 5) y permitió diferenciar de manera significativa a pacientes con LES frente a controles sanos (Figura 6).

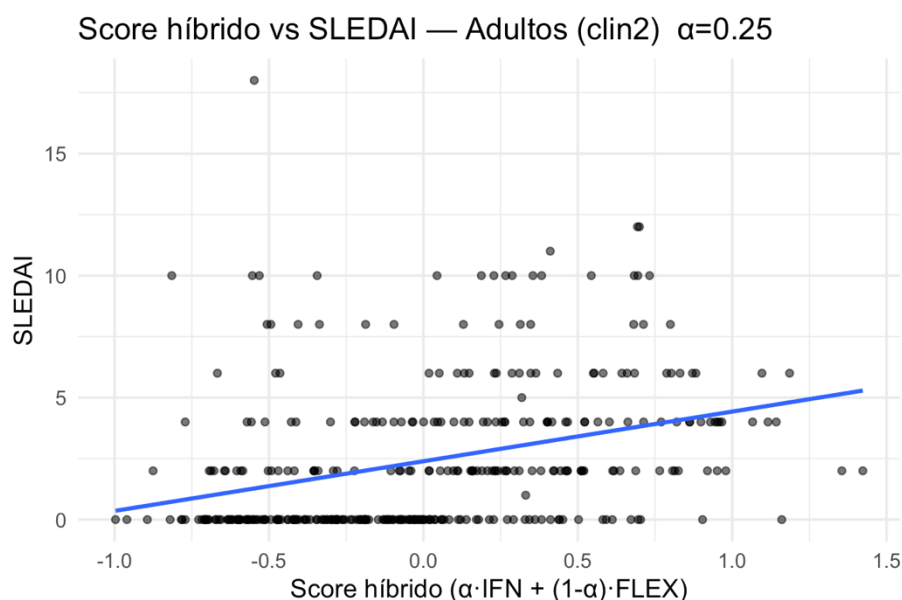


Figura 5: Scatterplot del score híbrido vs SLEDAI en adultos (clin2).

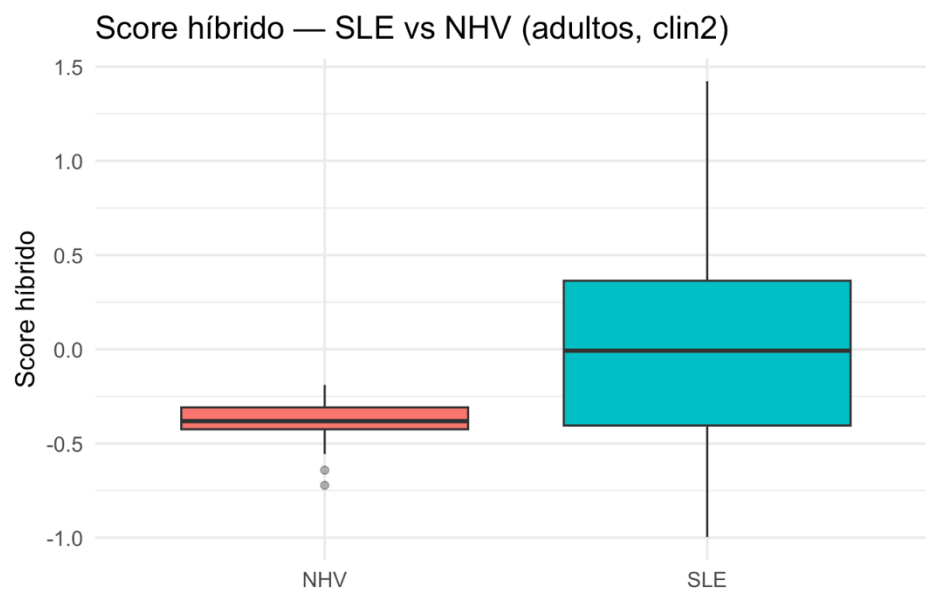


Figura 6: Boxplot del score híbrido SLE vs NHV (clin2).

Desde un punto de vista clínico, este resultado es relevante porque confirma que la información del interferón es indispensable, pero que debe complementarse con otras señales moleculares para reflejar la heterogeneidad real de la enfermedad. Además, el score híbrido permitió diferenciar de manera significativa a pacientes con LES frente a controles sanos en la cohorte adulta ( $p < 1e-06$ , test de Wilcoxon), lo que respalda su utilidad como marcador diagnóstico molecular.

Sin embargo, una limitación observada fue su menor capacidad para separar pacientes con SLEDAI alto y bajo en la cohorte adulta ( $p \approx 0.28$ ). Este hallazgo pone de manifiesto la complejidad de la enfermedad: aunque el lupus activo comparte una firma común dominada por interferón y rutas inflamatorias adicionales, la expresión diferencial entre “alto” y “bajo” no es uniforme y probablemente dependa de subtipos clínicos y moleculares que requieren un análisis más fino.

Aun con esta limitación, el desarrollo del score híbrido supone un avance conceptual importante. No se trata únicamente de identificar un conjunto de genes altamente correlacionados, sino de integrar distintas dimensiones moleculares de la enfermedad en un único índice. Esta aproximación refleja de manera más fiel el carácter multifactorial y heterogéneo del LES, y abre la puerta a futuros refinamientos, como la

ponderación dinámica de los paneles en función del contexto clínico o la inclusión de otras capas de datos (proteómica, metabolómica o variables clínicas).

En resumen, el score híbrido confirma que:

1. La señal de interferón, aunque esencial, es insuficiente por sí sola para capturar toda la complejidad del lupus.
2. El panel flexible aporta diversidad y detecta procesos no cubiertos por IFN, mejorando la sensibilidad del modelo.
3. La combinación ponderada de ambos logra un equilibrio entre robustez y capacidad de generalización, especialmente en cohortes adultas.

Este enfoque híbrido constituye el cierre natural del trabajo: un marcador **más sensible, integrador y adaptable**, que refleja tanto la activación clásica por IFN como la heterogeneidad molecular residual que define la fisiopatología del LES.

### Extensión a múltiples firmas (n)

Dado que el lupus es molecularmente heterogéneo, extendimos el híbrido a un marco general con n firmas. En lugar de combinar solo IFN y el panel flexible, incorporamos módulos biológicamente relevantes (p. ej., traducción/ribosoma, inflamación/neutrófilos) y dejamos que un optimizador asigne pesos no negativos que suman 1 a cada firma. Formalmente:

$$\text{Score híbrido} = \sum_{k=1}^K w_k \text{Score}_k, \quad w_k \geq 0, \quad \sum_{k=1}^K w_k = 1.$$

En esta expresión, Score\_k representa el valor del k-ésimo módulo o firma para cada muestra (tras reescalado previo a [0,1] para homogeneizar escalas), w\_k es su peso no negativo, y K el número total de firmas. Intuitivamente, w\_k cuantifica la contribución relativa de la firma k: a mayor asociación con la actividad clínica, mayor peso. En nuestro caso, los pesos se aprenden en la cohorte pediátrica (entrenamiento) en

función de la magnitud de su asociación con SLEDAI (p. ej.,  $|\rho \text{ Spearman}|$ ) y se evalúan en la cohorte adulta (validación), manteniendo un rendimiento mínimo en entrenamiento para evitar combinaciones inestables.

A modo ilustrativo, si para una muestra concreta (ya reescalada) se obtienen  $\text{IFN} = 0,70$ ,  $\text{FLEX} = 0,50$  y  $\text{NEUT} = 0,20$ , y en entrenamiento las asociaciones absolutas con SLEDAI son 0,40, 0,30 y 0,10, respectivamente, la normalización conduce a  $w_{\text{IFN}} = 0,50$ ,  $w_{\text{FLEX}} = 0,375$  y  $w_{\text{NEUT}} = 0,125$ ; el índice combinado sería entonces  $0,50 \cdot 0,70 + 0,375 \cdot 0,50 + 0,125 \cdot 0,20 = 0,5625$ . Este ejemplo ilustra cómo el modelo prioriza cuantitativamente las firmas más explicativas, incorpora diversidad biológica a través del panel flexible y atenúa señales débiles, manteniendo la suma de pesos igual a 1 y, por tanto, una lectura estable y directamente interpretable del marcador.

En la práctica, el modelo con  $n$  firmas repartió el peso entre el panel flexible y los módulos de traducción/ribosoma y neutrófilos, dejando un peso menor a IFN. Es coherente con lo observado en los análisis funcionales: además del eje interferón, aparecen procesos de síntesis proteica y actividad mieloide. Aun así, el rendimiento global no mejoró frente a la combinación de solo dos firmas (IFN+FLEX). Es decir, añadir más módulos no aportó señal independiente suficiente y, en conjunto, tendió a diluir la robustez del modelo más parsimonioso.

Desde el punto de vista práctico, esto sugiere dos ideas:

1. **Parquedad antes que complejidad.** Cuando IFN y el panel flexible ya capturan la mayor parte de la variación clínicamente relevante, sumar módulos puede no traducirse en mejoría estable fuera de la cohorte de entrenamiento.
2. **Selección y regularización.** La extensión a  $n$  firmas es útil como marco y para interpretar pesos (qué procesos intentan entrar en el índice), pero conviene acompañarla de criterios de selección más estrictos (p. ej., mantener solo módulos con señal mínima en entrenamiento, o estrategias con penalización para forzar pocos pesos distintos de cero).

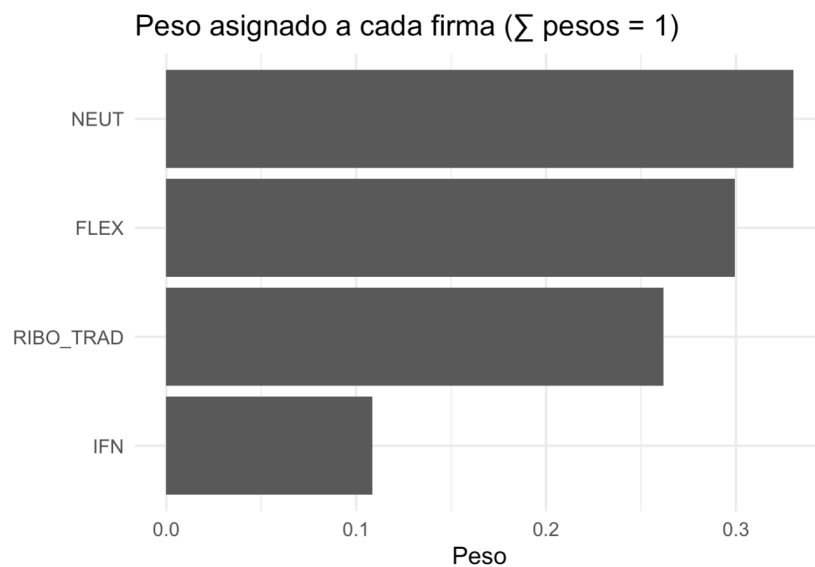


Figura 7: Gráfico de barras con los pesos aprendidos por firma (interpretabilidad)

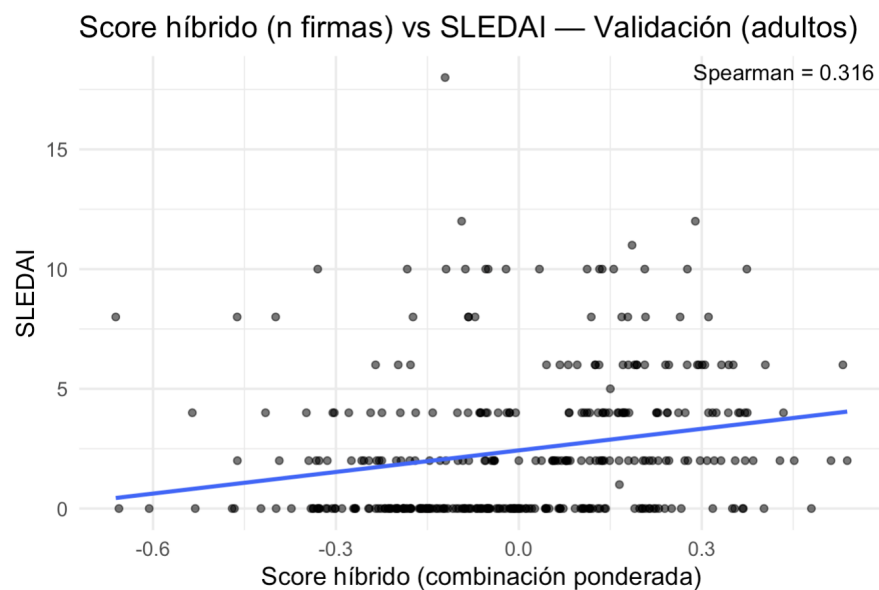


Figura 8: Scatterplot del score híbrido (n firmas) vs SLEDAI en adultos.

En conclusión, El marco con n firmas aporta flexibilidad e interpretabilidad y permite comprobar, de forma data-driven, qué procesos “quieren” entrar en el marcador. En nuestros datos, la combinación binaria IFN+FLEX sigue siendo la opción más equilibrada y generalizable, mientras que el modelo con n firmas ayuda a explicar la biología subyacente (peso de traducción y neutrófilos) pero no mejora la predicción clínica.

## 6. Discusión

El presente trabajo tuvo como objetivo explorar nuevas aproximaciones transcriptómicas que pudieran complementar o mejorar la capacidad del índice clínico SLEDAI para evaluar la actividad del lupus eritematoso sistémico (LES). Los resultados obtenidos permiten extraer varias conclusiones relevantes que contribuyen a comprender tanto los alcances como las limitaciones de las estrategias moleculares propuestas.

En primer lugar, la aplicación de *singscore* a genes individuales demostró ser una aproximación sólida y reproducible. La firma génica identificada mostró una correlación elevada con la actividad clínica en la cohorte pediátrica y permitió separar claramente pacientes de controles sanos. Estos resultados reproducen hallazgos previos de la literatura, en los que la activación del eje de interferón tipo I aparece como un sello molecular del LES. La ventaja principal de este enfoque radica en su simplicidad e interpretabilidad, ya que se basa en un conjunto relativamente reducido de genes y no requiere normalizaciones complejas entre grupos. No obstante, el hecho de que la señal esté casi monopolizada por genes inducidos por interferón representa una limitación, puesto que reduce la capacidad de este marcador para captar la heterogeneidad molecular del LES.

Con el fin de ampliar la visión proporcionada por los genes individuales, se incorporaron aproximaciones funcionales basadas en enriquecimiento de rutas biológicas mediante *ssGSEA* y *GSVA* calculadas con PathMED. Estos métodos permitieron detectar procesos celulares coordinados cuya desregulación se asocia a la actividad clínica, incluyendo rutas vinculadas con la síntesis proteica y biogénesis ribosomal, procesamiento de ARN y direccionamiento cotraduccional a membrana/retículo endoplásmico. Aunque la correlación alcanzada con SLEDAI fue más moderada que en el caso del *singscore*, los resultados se mantuvieron consistentes en la cohorte de validación adulta, lo que respalda su capacidad de generalización.

Además, este enfoque aportó un valor añadido: la posibilidad de interpretar biológicamente los mecanismos subyacentes, algo de gran relevancia en una enfermedad multifactorial como el LES.

Un hallazgo de especial interés fue la detección de actividad molecular en pacientes clínicamente inactivos (SLEDAI <2). Considerados estables desde un punto de vista clínico, estos pacientes mostraron alteraciones transcriptómicas vinculadas a procesos de traducción y biogénesis ribosomal. Aunque estos procesos no implican por sí mismos un empeoramiento clínico inmediato, su presencia consistente sugiere actividad biológica persistente no captada por SLEDAI. La identificación de esta “actividad residual” puede tener implicaciones relevantes para la monitorización terapéutica y la predicción de brotes. Este patrón se replicó en el grupo de baja actividad (SLEDAI ≤4) con menor magnitud, reforzando la idea de una actividad subclínica que disminuye progresivamente a medida que el paciente se estabiliza.

A partir de estas observaciones, se planteó el desarrollo de una firma transcriptómica flexible que permitiera capturar la heterogeneidad de los pacientes con alta actividad clínica. Esta estrategia, basada en la selección de genes con alta variabilidad y asociación consistente con SLEDAI, condujo a la construcción de un panel de 80 genes capaz de identificar subgrupos moleculares dentro de los pacientes con SLEDAI elevado. En ambas cohortes (pediátrica y adulta) fue posible observar que los pacientes con alta actividad no son homogéneos, sino que se dividen en subpoblaciones con perfiles transcriptómicos divergentes (La traducción clínica de estos subgrupos requiere validación externa y seguimiento longitudinal). Este hallazgo es especialmente relevante desde un punto de vista clínico, ya que sugiere que el mismo nivel de actividad clínica puede estar sustentado en mecanismos moleculares distintos, lo que podría tener consecuencias para la respuesta terapéutica.

El paso final consistió en integrar la robustez de la señal de interferón con la diversidad del panel flexible mediante un score híbrido. La combinación ponderada de ambos componentes (25% IFN y 75% flexible) mejoró el rendimiento global, manteniendo resultados consistentes en la pediátrica y distinguiendo de forma significativa a

pacientes de controles sanos. En términos prácticos,  $\alpha$  actúa como un control deslizante entre la robustez del IFN y la cobertura de mecanismos del panel flexible. Aunque la discriminación entre pacientes con SLEDAI alto y bajo fue limitada en la cohorte adulta, la mejor generalización y la capacidad de reflejar tanto la activación por interferón como otras rutas moleculares convierten a este marcador híbrido en un candidato prometedor para estudios futuros. Además, exploramos una extensión con múltiples firmas ( $n$ ) combinadas con pesos no negativos que suman 1. El modelo repartió peso entre el panel flexible y módulos de traducción/ribosoma y neutrófilos, pero no superó de forma estable al híbrido de dos firmas; ello sugiere que la parquedad (IFN + flexible) ofrece el mejor equilibrio entre rendimiento e interpretabilidad en nuestros datos. Por último, señalamos dos consideraciones metodológicas: (i) los pesos se ajustaron en la cohorte pediátrica, con el consiguiente riesgo de sobreajuste; y (ii) la traslación a RNA-seq requerirá armonización y recalibración de umbrales.

En conjunto, estos resultados ponen de manifiesto la necesidad de superar los enfoques centrados exclusivamente en el interferón y avanzar hacia estrategias que integren múltiples dimensiones moleculares. El LES es una enfermedad heterogénea en su presentación clínica y también en su sustrato biológico, por lo que herramientas transcriptómicas flexibles, capaces de adaptarse a esa diversidad, ofrecen una vía prometedora para el desarrollo de biomarcadores más sensibles y personalizados. No obstante, este estudio también presenta limitaciones, entre ellas el tamaño muestral reducido en la validación adulta, la dependencia de datos de microarrays y la ausencia de validación funcional en cohortes independientes adicionales.

A pesar de estas limitaciones, los hallazgos obtenidos refuerzan la hipótesis inicial de que los cambios en la expresión génica reflejan estados de actividad de la enfermedad que no siempre son detectables mediante escalas clínicas. La integración de datos transcriptómicos en el seguimiento del LES no solo puede mejorar la estratificación de los pacientes, sino también abrir nuevas vías para la predicción de brotes y la individualización del tratamiento.

## 7. Conclusiones

El presente trabajo demuestra que es posible diseñar un marcador transcriptómico para el LES que complemente al índice clínico SLEDAI, captando tanto la activación de rutas moleculares clásicas como la heterogeneidad de la enfermedad.

- i. El *singscore* basado en genes individuales, dominado por la señal de interferón tipo I, mostró la mayor correlación con SLEDAI en la cohorte pediátrica y se consolidó como un punto de partida robusto y fácilmente interpretable.
- ii. El análisis funcional de rutas mediante *ssGSEA* y *GSVA* añadió una capa interpretativa, permitiendo identificar procesos biológicos relevantes y demostrando que pacientes clínicamente inactivos pueden mantener actividad molecular residual.
- iii. La construcción de un panel transcriptómico flexible permitió identificar subgrupos moleculares dentro de pacientes con alta actividad clínica, reflejando la diversidad biológica del LES más allá de lo que capta SLEDAI.
- iv. El desarrollo de un score híbrido, que combina genes de interferón y genes no-IFN seleccionados por estabilidad, mejoró la generalización en la cohorte adulta y demostró ser un candidato prometedor como biomarcador complementario.

En conjunto, los resultados refuerzan la necesidad de integrar información transcriptómica en la evaluación del LES para lograr una caracterización más precisa y personalizada de la actividad de la enfermedad.

## 8. Bibliografía

- Banchereau, R., Hong, S., Cantarel, B., Baldwin, N., Baisch, J., Edens, M., Cepika, A. M., Acs, P., Turner, J., Anguiano, E., Vinod, P., Kahn, C., Hong, S., et al.

- (2016). Personalized immunomonitoring uncovers molecular networks that stratify lupus patients. *Cell*, 165(3), 551–565.  
<https://doi.org/10.1016/j.cell.2016.03.008>
- Bioconductor. (2025). *Bioconductor: Open source software for bioinformatics*.  
<https://www.bioconductor.org/>
  - Chiche, L., Jourde-Chiche, N., Whalen, E., Presnell, S., Gersuk, V., Dang, K., Anguiano, E., Quinn, C., Burtsey, S., Berland, Y., Chaussabel, D., & Pascual, V. (2014). Modular transcriptional repertoire analyses of adults with systemic lupus erythematosus reveal distinct type I and type II interferon signatures. *Arthritis & Rheumatology*, 66(6), 1583–1595.  
<https://doi.org/10.1002/art.38628>
  - Crow, M. K. (2014). Type I interferon in the pathogenesis of lupus. *The Journal of Immunology*, 192(12), 5459–5468.  
<https://doi.org/10.4049/jimmunol.1002795>
  - Foroutan, M., Bhuva, D. D., Lyu, R., Horan, K., Cursons, J., & Davis, M. J. (2018). Single sample scoring of molecular phenotypes. *BMC Bioinformatics*, 19, 404.  
<https://doi.org/10.1186/s12859-018-2435-4>
  - Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data.  
<https://doi.org/10.1186/1471-2105-14-7>
  - Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. <https://doi.org/10.1016/j.cels.2015.12.004>
  - The Broad Institute. (2025). *Molecular Signatures Database (MSigDB)*.  
<https://www.gsea-msigdb.org/gsea/msigdb>
  - The R Foundation. (2025). *The R Project for Statistical Computing*.  
<https://www.r-project.org/>
  - Toro-Domínguez, D., Villatoro-García, J. A., Martorell-Marugán, J., López-Domínguez, R., Luque-Tévar, M., Carmona-Sáez, P., & Teruel, M. (2018). Stratification of systemic lupus erythematosus patients into three groups of

disease activity progression according to longitudinal gene expression.

<https://pubmed.ncbi.nlm.nih.gov/29938934/>

- Mariana Postal, Jessica F Vivaldo, Ruth Fernandez-Ruiz, Jacqueline L Paredes , Simone Appenzeller, Timothy B Niewold (2021). Type I interferon pathway transcriptomic signature in patients with systemic lupus erythematosus stratifies disease activity and clinical manifestations.  
<https://pmc.ncbi.nlm.nih.gov/articles/PMC8054829/>
- Martorell-Marugán, J., Toro-Domínguez, D., López-Domínguez, R., & Ellson, I. (2025). *pathMED: Scoring Personalized Molecular Portraits* (R package version 1.0.2) [Software]. Bioconductor.  
<https://www.bioconductor.org/packages/release/bioc/html/pathMED.html>

## 9. Anexos

[https://drive.google.com/drive/folders/18DLWMdlhmpAoVNF5pqJBpu5PPyMBo\\_yx?usp=drive\\_link](https://drive.google.com/drive/folders/18DLWMdlhmpAoVNF5pqJBpu5PPyMBo_yx?usp=drive_link)