Máster en Bioinformática

Desarrollo de modelos de aprendizaje automático para la predicción de variables pronósticas y de respuesta farmacológica en pacientes de Nefritis Lúpica a partir de perfiles moleculares, orientados hacia la medicina personalizada

Autor: Elisabeth Medina Estévez

Tutor: Daniel Toro Domínguez

Curso 2023-24

Gracias a Daniel, mi tutor, por su compromiso con el proyecto, estando siempre disponible y con buena actitud para resolver todas las dudas, comprender mis tiempos y ofrecer una guía científica totalmente indispensable.

A Laura, mi compañera de datos, por estar ahí para compartir frustraciones y éxitos, motivándonos y mejorando juntas. Al final, aquí estamos y lo hemos conseguido.

A mi familia, por apoyarme y animarme a pesar de no entender del todo porqué me complico la vida de esta manera. A las amigas que siguen ahí, aunque lleve un año medio desaparecida. Y especialmente a Javi, por estar a mi lado siempre, cuidarme y comprenderme; por tener una paciencia infinita a veces, y por ser el mejor soporte emocional y soporte IT que podría tener.

RESUMEN

La nefritis lúpica (LN) es una de las manifestaciones clínicas con peor pronóstico dentro del Lupus. La heterogeneidad clínica y molecular propias de la enfermedad, junto con la incapacidad de predecir brotes que puedan derivar en fallo orgánico, son las principales causas del fracaso de los ensayos clínicos tradicionales. Un enfoque más personalizado basado en biomarcadores genéticos podría ayudar en la toma de decisiones médicas ante la predicción de grupos de pacientes no respondedores.

Objetivos:

El objetivo de este estudio es construir y evaluar la aplicabilidad clínica de modelos de aprendizaje automático basados en genes diferencialmente expresados (DEGs) para la predicción de la respuesta al tratamiento con Micofenolato de Mofetilo (MMF) en una cohorte longitudinal de pacientes de LN.

Material y métodos:

Se planteó un estudio a dos niveles: global para todas las muestras, en base a DEGs constitutivos de la respuesta al fármaco a lo largo del tiempo; y con muestras cercanas al periodo de biopsia, para la detección temprana. La respuesta a MMF se midió por uPCR y mPERR. Los modelos predictivos para cada periodo y variable respuesta se crearon con la herramienta *pathMED*, para análisis en base a datos de expresión. La selección de características de los modelos se realizó aplicando filtrado por correlación y análisis de expresión diferencial; y en los casos requeridos, análisis de componentes principales (PCA) y eliminación recursiva de características.

Resultados:

Todos los modelos obtuvieron buenos resultados, siendo los más optimizados el modelo global para uPCR con selección de DEGs por PCA, y los modelos de respuesta temprana para ambas variables (MCC > 0.8, *balance accuracy* > 0.8, AUC-ROC y AUC-PR en torno a 0.90).

Conclusiones:

El desempeño de los perfiles genéticos como predictores de pronóstico y respuesta al tratamiento en pacientes con LN es prometedor, especialmente en el caso de periodo cercano al diagnóstico o inicio del tratamiento. Sin embargo, es necesaria más investigación para resolver los problemas de reproducibilidad inherentes a los datos de expresión génica, que permitan su validación independiente para el escalado de los modelos a la práctica clínica.

Palabras clave:

Nefritis lúpica, micofenolato de mofetilo, uPCR, mPERR, expresión génica, aprendizaje automático

ABSTRACT

Lupus nephritis (LN) is one of the clinical manifestations with the worst prognosis in Lupus. The clinical and molecular heterogeneity of the disease, together with the inability to predict flares that may lead to organ failure, are the main reasons for the failure of traditional clinical trials. A more personalised approach based on genetic biomarkers could help in medical decision-making by predicting non-responsive patient groups.

Objectives:

The aim of this study is to build and evaluate the clinical applicability of machine learning models based on differentially expressed genes (DEGs) for the prediction of response to mycophenolate mofetil (MMF) treatment in a longitudinal cohort of LN patients.

Methods:

A two-level study was proposed: global for all samples, based on DEGs constitutive of drug response over time; and with samples close to the biopsy period, for early detection. MMF response was measured by uPCR and mPERR. Predictive models for each time period and response variable were created with the *pathMED* tool for analysis based on expression data. Model feature selection was performed by applying correlation filtering and differential expression analysis; and where required, principal component analysis (PCA) and recursive feature elimination.

Results:

All models performed well, the most optimised being the global model for uPCR with selection of DEGs by PCA, and the early response models for both variables (MCC > 0.8, *balance accuracy* > 0.8, AUC-ROC and AUC-PR around 0.90).

Conclusions:

The performance of genetic profiles as predictors of prognosis and treatment response in patients with LN is promising, especially in the case of the period close to diagnosis or treatment initiation. However, more research is needed to address the reproducibility issues inherent in gene expression data to allow independent validation for scaling up the models to clinical practice.

Keywords:

Lupus nephritis, mycophenolate mofetil, uPCR, mPERR, gene expression, machine learning

ÍNDICE

1. Introducción

El Lupus Eritematoso Sistémico (SLE, por sus siglas en inglés) es una enfermedad autoinmune que presenta una gran heterogeneidad tanto a nivel molecular como clínico, con patrones impredecibles de brotes y remisiones que dificultan la evaluación pronóstica de la enfermedad y la selección de tratamientos efectivos a largo plazo [1].

La heterogeneidad clínica implica la afectación de múltiples sistemas y órganos corporales, con sintomatología que incluye desde erupciones cutáneas, úlceras orales, enfermedades inflamatorias (pericarditis, pleuresía) y citopenias, hasta patologías renales y neurológicas graves [2]. Una de las manifestaciones clínicas más graves del SLE es la Nefritis Lúpica (LN, por sus siglas en inglés) que, en el peor de los casos, puede derivar en insuficiencia renal terminal.

Las terapias actuales de referencia para el tratamiento de LN incluyen inmunosupresores como el Micofenolato de Mofetilo (MMF) o la Azatriopina (AZA), Belimubab (anticuerpos monoclonales) e inhibidores de calcineurina, junto con dosis altas iniciales de fármacos estándar, como la Prednisona (glucocorticoide, GC) y las cloroquinas (CQ). Sin embargo, debido a la variabilidad intrínseca de la enfermedad y entre los propios pacientes, la mayoría de ensayos clínicos no han mostrado eficacia a largo plazo, con no más de un 40% de remisión completa al año de tratamiento [3].

Por tanto, y paralelamente a la necesidad de estudio de los mecanismos moleculares subyacentes de la enfermedad, la capacidad de anticipar la falta de respuesta de un fármaco en pacientes de LN permitiría al clínico la elección temprana de una terapia alternativa, personalizada, con mayor probabilidad de éxito.

En este sentido existen diferentes índices para medir la actividad de la enfermedad, relacionada con los periodos de exacerbación y en base a las manifestaciones clínicas observadas, que incluyen los propios síntomas, resultados de pruebas de laboratorio, afectación de órganos, informes de los pacientes y evaluaciones médicas [4].

Uno de los más utilizados en la práctica clínica y en ensayos clínicos es el Índice de Actividad de la Enfermedad de SLE (*SLE Desease Activity Index*, SLEDAI), el cual pondera 24 descriptores, según importancia del órgano afectado, de los cuales 16 son

manifestaciones clínicas, que el médico evalúa como ausente o presente en los últimos diez días; y 8 se basan en resultados de pruebas de laboratorio [5]. Otros sistemas de puntuación de la actividad global de la enfermedad son la Evaluación Global del Médico (*Physician Global Assesment*, PGA), una escala analógica visual de 0-3; o el Índice de Actividad de Lupus (*Lupus Activity Index*, LAI).

La utilidad de estos índices es la de servir como herramienta para el diagnóstico precoz y guía para la decisión terapéutica desde un enfoque "tratamiento a objetivo" o *treat to target* (T2T), mediante la monitorización continua de la eficacia del tratamiento [4]. Sin embargo, aunque los descriptores de SLEDAI, como la piuria, hematuria, etc., son en una alta proporción determinantes de LN; se ha observado que pacientes con puntuaciones similares de SLEDAI, pero con combinaciones diferentes de manifestaciones clínicas, pueden presentar diferente pronóstico y respuesta a fármacos [6]. Esto es porque los índices de actividad no reflejan la heterogeneidad molecular en las rutas de activación de la enfermedad y en los mecanismos biológicos desregulados, que pueden ser diferentes entre pacientes y a lo largo del curso de la enfermedad.

Por otro lado, SLEDAI y otros índices son indicadores generales de actividad de la enfermedad, y son poco sensibles a la evolución de fenotipos concretos dentro de la misma [5], por lo que se necesitan indicadores más específicos para medir de una forma objetiva el funcionamiento orgánico; por ejemplo, en la monitorización de la respuesta a un tratamiento. Este es el caso de los indicadores de respuesta renal en LN, como la relación proteína/creatinina en análisis de orina (*urine Protein Creatinine Ratio*, uPCR) y la Tasa de Filtración Glomerular estimada, (*estimated Glomerular Filtration Rate*, eGFR).

El enfoque T2T con terapias basadas en ómicas no está tan extendido en enfermedades autoinmunes como lo está en el cáncer, por ejemplo. En lupus, estudios de estratificación basados en datos longitudinales de expresión génica han mostrado correlación con la actividad y diferenciación clínica, funcional y a nivel celular [6].

Por otra parte, el desarrollo de biomarcadores basados en genes mediante técnicas de aprendizaje automático (*Machine Learning*, ML) supone un primer avance hacia un tratamiento de la enfermedad más personalizado. Una de las razones del fracaso de los ensayos clínicos de SLE es debido al tratamiento de los pacientes como grupos

homogéneos. Se han probado con éxito modelos ML basados en la caracterización de firmas diferencialmente expresadas tras la respuesta a fármacos y teniendo en cuenta diferencias en el tiempo en cohortes longitudinales [1].

La detección de biomarcadores de la respuesta a fármacos constituye un problema de clasificación binaria, que considera dos clases posibles para la agrupación de pacientes, respondedores y no respondedores. Dependiendo del enfoque de aprendizaje utilizado, existen diferentes técnicas ML para la predicción de clases.

Por un lado, se encuentran los algoritmos de aprendizaje supervisado, que utilizan clases predefinidas sobre los datos de entrenamiento. El algoritmo generaliza la información de las características sobre el conjunto de entrenamiento para construir un modelo que predice correctamente los resultados sobre el conjunto conocido; después aplica el modelo entrenado en conjuntos de datos no vistos para evaluar las predicciones con nuevas características. Algunos de los algoritmos de aprendizaje supervisado son los modelos lineales, como el Análisis Discriminante Lineal (*Linear Discrimination Analysis*, LDA) o el Modelo Lineal Generalizado (*Generalized Linear Model*, GLM); los Árboles de Decisión (*Decision Trees*, DT); el Bosque Aleatorio (*Random Forest*, RF), las Máquinas de Soporte Vectorial (*Support Vector Machine*, SVM) y las Redes Neuronales Artificiales (*Artificial Neural Networks*, NNET). [7]

Los modelos lineales se basan en la regresión lineal y estiman la probabilidad de que una observación, dado un determinado valor de los predictores, pertenezca a una clase u otra de una variable cualitativa. Los DT parten de un nodo de decisión inicial (raíz) sobre la clasificación de las instancias y van descomponiendo el dataset en distintas ramas y hojas (clases) en función de condiciones “if-else”, que se basan en grupos lógicos de máxima diferencia entre sí. Variantes más sofisticadas de DT son RF y los Árboles con potenciación extrema de gradiente (*Extreme Gradient Boosted Trees*, xgbTree). RF aplica DT en múltiples clasificadores (*ensemble*), estableciendo un número de árboles inicial para la toma de decisiones (*ntrees*) y un número de variables seleccionadas como candidatas en cada rama de decisión (hiperparámetro *mtry*). Por su parte, los potenciadores de gradiente, como *xgbTree*, llevan a cabo un remuestreo del dataset que pondera los resultados del conjunto de decisiones. Los algoritmos SVM se basan en

modelos *Kernel*, lineal o radial, cuyo objetivo es encontrar un hiperplano de mayor dimensión en el que clasificar las distancias medidas en el espacio conocido inicial de la red de datos de entrenamiento. Esto es, un método de aumento de la dimensionalidad para mejorar la separación y ajuste de los datos a las clases definidas, útil con relaciones no lineales entre las características. Las NNET tratan de imitar el funcionamiento neuronal para el reconocimiento de patrones en los datos. Constan de una capa de entrada, una o más capas ocultas y capas de salida, donde las unidades funcionales/neuronas de una capa están conectadas a todas las neuronas de la capa anterior y posterior. El proceso de entrenamiento se realiza mediante retropropagación entre capas ocultas, ajustando los pesos de las características en las conexiones entre neuronas [7-9].

Otro enfoque de los algoritmos de ML se basa en el aprendizaje no supervisado, en el que no se predefinen clases en los datos de entrenamiento, y que suele tener un propósito de estratificación (*clustering*) y reducción de la dimensionalidad, más que de predicción; habitualmente debido a falta de homogeneidad en los datos. Los métodos de *clustering* permiten la extracción de características y exploración de los posibles grupos o clases mediante la identificación de relaciones subyacentes en conjuntos de datos no etiquetados, agrupándolos por similitudes. Ejemplos de algoritmos ML de aprendizaje no supervisado son K-Vecinos Más Cercanos (*K Nearest Neigbourghs*, KNN) y el Análisis de Componentes Principales (*Principal Component Analysis*, PCA). KNN establece la probabilidad de que un elemento pertenezca a una clase según su cercanía en el espacio a otros elementos similares, que puedan conformar un grupo homogéneo basado en la media de grupo para la distinción entre clases. PCA aplicado a expresión diferencial, es una técnica de reducción de la dimensionalidad, por la cual se sustituyen las observaciones (genes o muestras) originales por n combinaciones lineales (componentes principales), que representan una proporción razonable de la variación total de la expresión génica [7-9].

La elección de un enfoque u otro depende del número de características, tamaño de la muestra y distribución de los datos; si bien en el ámbito de la salud es más común el uso de modelos de aprendizaje supervisado por su mayor precisión.

En el presente proyecto se lleva a cabo la construcción y evaluación de modelos ML basados en firmas moleculares de una cohorte longitudinal de pacientes de LN con datos de respuesta y no respuesta a MMF, obtenidos a partir del estudio de Toro-Domínguez et al. (2024) [2]. Para la generación de los modelos se emplea la herramienta bioinformática *pathMED*, diseñada específicamente para el uso de datos de expresión génica como predictores, y basada en lenguaje R [10].

La posibilidad de detección temprana de grupos de pacientes no respondedores, puede ser una herramienta valiosa para la toma de decisiones de primera línea de tratamiento; que puedan suponer frenar la progresión a daño renal crónico en estadios leves de la enfermedad, con la consiguiente mejora de la supervivencia y la calidad de vida de los pacientes.

2. Hipótesis y Objetivos

En el contexto de medicina personalizada, ¿puede haber vínculos relevantes entre características genéticas y manifestaciones clínicas asociadas a LN que permitan generar modelos predictivos del curso de la enfermedad y de la respuesta farmacológica, para la ayuda en la toma de decisiones en la práctica clínica?

Hipótesis:

Las manifestaciones clínicas asociadas al pronóstico de la enfermedad renal y la respuesta al tratamiento con MMF pueden predecirse mediante modelos de aprendizaje automático basados en el perfil molecular de cada paciente de LN.

Objetivo primario:

Construir y evaluar modelos de aprendizaje automático capaces de predecir variables clínicas de interés y respuesta farmacológica a partir de perfiles genéticos de pacientes de LN tratados con MMF.

Objetivos secundarios trabajado en el proyecto:

1. Explorar la distribución y asociación entre variables clínicas y/o datos de expresión y diferenciar características en base a la respuesta positiva o negativa al fármaco.

2. Estudiar los métodos de selección de características en aprendizaje automático que sean adecuados a datos de expresión.

3. Enfrentar los retos durante la puesta a punto de modelos predictivos basados en genes y con datos de estudios longitudinales, dependientes del tiempo, para su aplicabilidad interestudio.

3. Materiales y Metodología

El procesamiento y análisis de los datos se realiza en lenguaje de programación R, con el software RStudio, versión 4.3.2.

a) Origen de los datos y Diseño del estudio

Se lleva a cabo un análisis retrospectivo de una cohorte inicial de 301 pacientes de LN, obtenida a partir de datos longitudinales procedentes de un estudio clínico observacional, en colaboración con la Universidad Johns Hopkins (EEUU) [2]. Todos los datos del estudio se encuentran pseudo-anonimizados y publicados en el repositorio público de datos genómicos Gene Expression Omnibus (GEO) del *National Center for Biotechnology Information* (NCBI), con código de acceso: **GSE224705**.

El conjunto completo se estructura en datos de expresión génica obtenidos y varios sets de datos clínicos de los pacientes de LN, que incluyen información demográfica y de dosis y respuesta a MMF, AZA y otros fármacos de tratamiento estándar, como la hidroxicloroquina en tratamiento único y junto a glucocorticoides.

Los datos originales fueron recopilados de la historia clínica de los pacientes y mediante el seguimiento de la enfermedad en visitas ambulatorias por un periodo de más de dos años; siguiendo el protocolo del estudio SPARE (*Study of biological Pathways, disease Activity and Response markers in patients with systemic lupus Erythematosus*), de la Universidad Johns Hopkins [3]. El protocolo del estudio fue aprobado por la *Johns Hopkins University School of Medicine Institutional Review Board*. 301 pacientes de SLE adultos, de entre 18 y 75 años, y que cumplían con los criterios revisados de clasificación

de la enfermedad del *American College of Rheumatology*, fueron incluidos en la *Hopkins Lupus Cohort*, previo consentimiento informado por escrito.

Sobre la cohorte inicial se lleva a cabo el análisis de variables clínicas y datos de expresión asociados a la respuesta al tratamiento con MMF, la selección de características significativas; así como el entrenamiento y optimización de modelos de aprendizaje automático para la predicción de respuesta a este tratamiento, a escala global de la población de estudio y para periodo temprano de diagnóstico/terapia (ver sección h). Al tratarse de datos recopilados en visita médica rutinaria, los tiempos de toma de muestras no se encuentran definidos y no coinciden entre los pacientes de la cohorte, como ocurre en un ensayo clínico. Por tanto, se establece la fecha de biopsia renal (medida por muestra), como delimitador de tiempo para la evaluación y predicción de respuesta en periodo temprano; ya que la biopsia se usa para confirmar el diagnóstico de LN y supone el punto de inicio de la terapia.

La respuesta al tratamiento con MMF se encuentra establecida por dos variables: la Relación Proteína/Creatinina en orina (uPCR) como *gold standard*, y la Respuesta Renal de Eficacia Primaria modificada (*modified Primary Efficacy Renal Response*, mPERR) como resultado secundario.

- La respuesta positiva por uPCR se define, de acuerdo con las guías EULAR, como la reducción y el mantenimiento del cociente proteína/creatinina medido en análisis de orina (parámetro uPCR) por debajo de 0.5 g/g a los 12 meses desde el inicio del tratamiento; o como una disminución del 25% y del 50% desde el inicio a los 3 y 6 meses, respectivamente [11].
- Para mPERR, se considera respuesta positiva al tratamiento con MMF el no empeoramiento en los criterios estimados de la Tasa de Filtración Glomerular (*estimated Glomerular Filtration Rate*, eGFR) (≥60 mL/min/1.73 m^2 o ≤20% por debajo del valor basal); y uPCR ≤0,7 (o proteína en orina de 24 horas ≤0,7 g/24 h, si faltan datos de uPCR; o puntuación semicuantitativa de la tira reactiva de proteína en orina ≤1 (≤30 mg/dL), si faltan tanto los datos de uPCR como los de proteínas en orina de 24 horas) [12].

Adicionalmente, se dispone de una cohorte separada para la validación de los modelos predictivos generados. La cohorte de prueba consta de 81 pacientes de LN y está constituida, a su vez, por dos cohortes: 43 pacientes provienen de la cohorte inicial ya mencionada, y 38 pertenecen a un estudio clínico de la *University Health Network* de Canadá [13]. Sería necesario seleccionar los 38 pacientes independientes para la validación. El estudio canadiense dispone de la aprobación del *Research Ethics Board of the University Health Network* para el reclutamiento de pacientes y la revisión de biopsias renales; así como de la firma del consentimiento informado por parte de todos los participantes. Los datos se encuentran disponibles con código de acceso GEO: **GSE99967**.

b) Tamaño muestral y Población de estudio.

El conjunto de datos completo de la cohorte inicial se obtiene mediante descarga del repositorio GitHub vinculado a la publicación de Toro-Domínguez et al. (2024) [2] [14]. El dato primario se compone de:

- Un objeto '*.RData*' que contiene la tabla de expresión génica por muestra y algunas tablas clínicas con variables de interés, pre-filtradas por tratamiento con MMF.
- Una carpeta comprimida con diferentes archivos de datos clínicos por paciente y muestra, correspondientes a la cohorte de origen con 301 pacientes. Incluye archivos Excel (.xls) con datos crudos recogidos directamente por el personal clínico, y archivos separados por comas (.csv), derivados de los .xls para su tratamiento en R.

En la Tabla 1 se describe un resumen del contenido de los archivos de datos:

Origen / Formato	Tipo de dato	Número de sets	Promedio de observaciones	Promedio de variables
.RData	Clínica por muestra	2	166	30
	Clínica por paciente	1	39	18
	Información renal	1	64	14

Origen / Formato	Tipo de dato	Número de sets	Promedio de observaciones	Promedio de variables
	Expresión génica	1	54715	747
.xls /.csv	ID visita/muestra - paciente	5	747	7
	Clínica por paciente	5	301	383
	Clínica por muestra/visita	4	714	104
	Perfiles celulares	5	714	15
.csv	Tratamiento por paciente	1	183	4
.xls	Diccionarios	4	N/A	N/A

Tabla 1. Resumen de los datos primarios de partida. Se indica el formato de archivo, tipo de dato recogido, número de tablas y un promedio de observaciones (filas) y variables (columnas) por cada tipo de tabla, al encontrarse datos duplicados entre las distintas tablas clínicas.

La población de estudio viene definida por el set de datos clínicos por paciente del objeto '*.RData*', el cual contiene los 39 pacientes que tienen asociadas las variables a predecir; la respuesta dicotómica (sí/no) a tratamiento con MMF por uPCR y mPERR, según los criterios definidos en la sección a, ambas a nivel global y tras un año de seguimiento. De estos 39 pacientes se descartan 2 con valores no válidos en ambas variables de respuesta global, que serán las estudiadas para los modelos predictivos. Con estos criterios de selección, de los 301 pacientes de la cohorte longitudinal y con datos de expresión génica para 747 muestras (que incluyen 20 controles de sanos), se identifican un total de 37 pacientes con datos de respuesta a MMF y 106 muestras asociadas (Material Suplementario, Archivo "*00. EDA_inicio.Rmd*"). Todos los pacientes seleccionados disponen de datos históricos de biopsia renal con hallazgos anormales asociados a pronóstico de LN.

c) Variables clínicas

Los datos clínicos por paciente se identifican de forma unívoca por el identificativo de cohorte o paciente (*cohortID* o *patientID*) y son, en su mayoría, variables numéricas binarias procedentes de historia clínica. Entre estas variables se encuentra un histórico de sintomatología asociada a la enfermedad (1=presencia en algún momento / 0=ausencia del síntoma); junto con la fecha (mes/año) del síntoma. También la

presencia o no de varios tipos de anticuerpos indicadores de pronóstico [2], como los anticuerpos anti-ADN bicatenario (anti-dsDNA) y antinucleares anti-LA (1=positivo/ 0=negativo); niveles del sistema del complemento (1=bajo o alto respecto al límite / 0=normal), como los componentes C3 y C4; e histórico de tratamientos (1=sí / 0=no), incluyendo la prednisona, esteroides tópicos, aspirina, citotóxicos (MMF), y la dosis más alta usada (variable continua). También dispone de datos demográficos, algunos, características modificadoras de la expresión génica como son la raza, el sexo, la edad o el peso.

Los datos clínicos por muestra son los tomados en las distintas visitas del estudio y están asociados de forma unívoca al ID de la visita (*visitID*) y/o al ID de la muestra (*sampleID*); también incluyen la correspondencia al ID del paciente, de forma que permite la conexión en todo momento entre paciente, visita y variable. En estos sets se recogen variables dicotómicas de presencia/ausencia (1/0) de síntomas para el cálculo de la actividad (*SLEDAI*, *LAI*); y distintos tratamientos (1=Sí / 0=No). Además, se recogen variables continuas medidas en analítica de laboratorio. Para evaluar la actividad de la enfermedad, se midieron en cada visita el Índice de Actividad de la Enfermedad de SLE (SLEDAI), la *Physician Global Assessment* (PGA) y el *Lupus Activity Index* o LAI (*Phys. Estimate*).

d) Limpieza y preparación de datos clínicos

Se realiza un análisis exhaustivo de las tablas clínicas crudas y del objeto '*RData*' inicial para, primero, reducir los datos a la población de estudio (37 pacientes / 106 muestras); y después estudiar las duplicidades y valores faltantes con objeto de unificar tablas y simplificar el número de variables sin perder información relevante. Para ello se emplean, fundamentalmente, los paquetes de R '*tidyverse*' y '*dyplr*', para manejo de *data.frames*, junto con funciones de `*R(base)*'. Se realizan comprobaciones de correspondencia fila a fila, columna a columna; se reordenan datos para coincidencia; se eliminan variables vacías (con datos NA) y variables de fechas no relevantes para el estudio estadístico, sí se mantienen las relativas a fechas de biopsia. Por último, se comprueban y eliminan aquellas variables que no cambian en la población de estudio;

es decir, que solo muestran un valor y, por tanto, no van a reflejar diferencias entre respuesta y no respuesta a tratamiento (Material Suplementario, Archivo "*01. Clinical_prep.Rmd*" y Archivo "*02. Environment_prep.Rmd*").

e) Análisis estadístico

Se realiza un análisis bivariado de las variables clínicas presentes en los conjuntos de datos por paciente y muestra previamente procesados frente a ambas respuestas a tratamiento, uPCR y mPERR; con objeto de obtener la distribución de valores y analizar las asociaciones más significativas frente a la respuesta a tratamiento en cada caso (Material Suplementario, Archivo "*03. Covars.Rmd*" y Archivo "*utils_clin.R*").

Se preparan dos grandes conjuntos de datos:

- *clin_by_patient*, con las variables clínicas recogidas por paciente
- *clin_by_sample*, con las variables clínicas recogidas por muestra (también asociadas a pacientes)

Con objeto de automatizar la evaluación del mayor número de variables posible se separan las variables numéricas binarias, categóricas no binarias y continuas en cada set de datos. Se elabora un archivo de funciones, *utils_clin.R*, con funciones para aplicar el Test de Fischer para variables numéricas binarias; Chi Cuadrado para variables categóricas no binarias; y T de Student o Test de Wilcoxon/U de Mann-Whitney para variables continuas en función de obtener distribuciones normales o no normales, respectivamente, en cada grupo o nivel de respuesta (respondedor y no respondedor para cada variable) [15]. Para la comprobación de la normalidad se utiliza el test de Shapiro-Wilk ($n<50$) o el test de Kolmogorov-Smirnov ($n>50$), dependiendo del número de muestras por nivel de respuesta para cada variable continua.

Una vez realizados los test, se extraen las variables significativas en base a p-valor < 0.05 y se representan gráficamente frente a cada respuesta. Para ello, se crean funciones en *utils_clin.R* para la generación de gráficos de barras agrupadas, en el caso de variables numéricas binarias y categóricas; y gráficos de cajas (*boxplot*) para variables continuas. En principio, se mantienen las variables de dosis de tratamiento como continuas;

aunque tras el análisis de variables dicotómicas (tratamiento sí/no), si el tratamiento en cuestión obtiene significancia, se pueden etiquetar los niveles para estudiar como categóricas. Ejemplo: La dosis de aspirina, variable "asa", con tres valores únicos, se puede tratar como categórica de tres niveles: no, *low* (dosis baja) y *high dose* (dosis alta).

f) Filtrado, normalización y anotación de la matriz de expresión

La matriz de expresión fue generada a partir de archivos *fastq*, utilizando sondas (*probes*) de *microarrays* de *Affymetrix* de *perfect match*. Código de acceso GEO de la plataforma: **GPL13158**, vinculado a GSE224705. Los *probes* son secuencias cortas de ADN diseñadas para registrar la expresión de genes de interés en una muestra mediante la hibridación a sus ARNm. Los *Probes Set* son los conjuntos de sondas que se dirigen a un mismo gen, los cuales disponen de un ID específico en la plataforma de *Affymetrix* [16].

La matriz inicial parte de 54715 *probes* en las filas y 747 muestras en las columnas. Primero se realiza una selección de las muestras por la población de estudio en base a los criterios descritos en la sección a, y conservando las 20 muestras de sanos para un análisis de expresión génica diferencial (*Differential Expression Gene* analysis, análisis DEG) entre casos y controles (Material Suplementario, Archivo "02. Environment_prep.Rmd"). El procesado posterior se realiza sobre la matriz de 126 muestras hasta su reducción a la población de estudio para la evaluación de la respuesta al tratamiento. La matriz de expresión se encuentra previamente normalizada con el método RMA (*Robust Multi-Array Average*), que realiza una normalización por cuartiles a través de todas las muestras. Los valores de expresión en *microarrays* se miden en intensidad de señal de fluorescencia, normalizada a escala logarítmica (se comprueba en R, rango de valores de 0 a 15) [16].

De forma adicional a RMA, se utiliza la función 'nearZeroVar()' de la librería *caret*, que devuelve aquellos genes que presenten valores de expresión con varianza cero o casi cero. Estos genes bien pueden no haber sido medidos correctamente, o bien no tienen expresión o no difieren en el conjunto de muestras. Es necesario eliminarlos ya que, en

el primer caso, pueden generar resultados irreales en el análisis; y en el segundo caso, genes con expresión no variable no resultan relevantes en la distinción entre los grupos de respuesta y no respuesta a fármaco.

Posteriormente, es necesario anotar cada *probe* a *GeneSymbol*, es decir, traducir las sondas a una nomenclatura estandarizada por genes. *GeneSymbol* es el identificativo génico reconocido por la herramienta *pathMED*, usada posteriormente para la generación de los modelos predictivos. Se creea la matriz de anotación con los *GeneSymbol* asociados a los *Probes Set ID* de *Affymetrix*, empleando el paquete de anotación *BiomaRt* [17-18]. En la correspondencia *probe*-gen pueden darse varias casuísticas: que un *probe* anote a un gen, entonces el valor de expresión del *probe* pasa al gen; que un *probe* anote a varios genes, dichos genes toman el valor de expresión del *probe*; que varios *probes* anoten un solo gen, en cuyo caso se imputa al gen el valor de la mediana de expresión de los *probes*, con objeto de contabilizar un único gen con un único valor de expresión media. Tras la anotación se realiza una limpieza de la matriz para eliminar *probes* no anotados y se fusionan los datos de expresión de aquellos *probes* que se dirigen (anotan) al mismo gen. Para ello se utiliza el valor de la mediana de expresión, calculado con la función 'AnnotateGenes()' procedente del repositorio asociado al artículo de Toro-Domínguez et al. (2022) [19] [1] (Material Suplementario, Archivo "*04. Annotation.Rmd*" y Archivo "*utils_annot.R"*).

g) Selección de características

El objetivo principal es generar modelos predictivos para la respuesta al MMF en base a los valores de expresión genética. Por tanto, se requiere establecer un flujo de trabajo para reducir el número de características a incluir en el modelo, dado el elevado número de genes de partida que lo harían computacional y clínicamente inviable. Además, la mayoría de estos genes no acaban siendo relevantes para la diferenciación entre respuesta y no respuesta, se trata de pasar de *big data* a *smart data*.

La selección de características para los modelos se plantea en tres pasos, tras el filtrado por anotación de la matriz de expresión:

1) Filtrado por genes correlacionados (Material Suplementario, Archivo "04. Annotation.Rmd")

2) Filtrado por genes significativos en análisis DEG (Material Suplementario, Archivo "05. DEG.Rmd")

3) Eliminación recursiva de características dentro de *pathMED* (Material Suplementario, Archivo "06. PathMED.Rmd")

El filtrado por genes correlacionados se realiza con la función 'findCorrelation()' del paquete *caret*, la cual requiere el cálculo previo de una matriz de correlación con los genes en las columnas, como características, y las muestras en las filas. De esta forma, 'findCorrelation()' marca las variables/genes correlacionados según un criterio límite de correlación por pares establecido (*cutoff*); genes que posteriormente se eliminan de la matriz de expresión para optimizar los modelos ML. La matriz de correlación se obtiene con la función 'cor()' aplicada a la traspuesta de la matriz de expresión anotada.

A continuación, se extraen los genes expresados diferencialmente (DEGs) más significativos para la respuesta a tratamiento con MMF; evaluados a nivel global para la población de muestras de estudio, y para el subconjunto de muestras tomadas en periodo temprano (inferior a 6 meses tras la fecha de biopsia).

Se utiliza el paquete *limma* [20] para la obtención de los genes significativos de cada subconjunto y para cada variable respuesta en base a p-valor ajustado. Para ello, se establecen las muestras no respondedoras a tratamiento como referencia (grupo 1 del factor respuesta para cada variable, uPCR y mPERR), de forma que *limma* aplica el test estadístico comparando el grupo respondedor (2) frente al no respondedor (1).

Se genera la matriz de diseño a partir de la fórmula para comparar ambos grupos de respuesta, y corrigiendo por las covariables significativas para expresión definidas en el análisis estadístico de variables clínicas. Independientemente de su significancia estadística, se deben tener en cuenta aquellas variables que son *per se* modificadoras de la expresión génica; estas son la dosis de fármaco, la edad, el sexo y la raza [2]. Se ajustan los datos de expresión a modelo lineal, aplicando matriz de contraste entre grupos (respondedores/no respondedores) y ajustando p-valores mediante corrección

'eBayes()', que ordena los genes por evidencia de expresión diferencial. A continuación, con la función 'topTable()' se extrae la clasificación de genes obtenida en 'eBayes()', aplicando *Benjamini y Hochberg* (BH) como método de ajuste (más apropiado para *microarrays* [20]) y ordenando por p-valor ajustado. BH controla la tasa esperada de falsos descubrimientos (*False Discovery Rate*, FDR) por debajo del valor especificado.

Finalmente, para obtener el grado de significancia de los genes se utiliza la función 'decideTests()'; que obtiene una matriz numérica con elementos -1, 0 o 1, dependiendo de si cada estadístico t calculado para cada coeficiente del modelo lineal se clasifica como significativamente negativo (genes reprimidos), no significativo o significativamente positivo (genes sobre-expresados), respectivamente. Se aplica por defecto el mismo método de ajuste que 'topTable()', "BH", de forma que se identifican las mismas sondas como diferencialmente expresadas. La lista de genes significativos se extrae a partir del nivel de significancia estadística p-valor ajustado < 0.05 y dejando el corte logFC >= 0, para no perder significancia biológica con los genes seleccionados, ya que hay genes cuyos *logFoldChange* (logFC) pueden no ser demasiado elevados, pero sí ser consistentes entre grupos, más aún cuando se estudian fenotipos complejos, o grupos para los cuales no se esperan grandes diferencias.

Una vez obtenidas las listas de genes significativos para cada conjunto de población y variable respuesta a tratamiento, se generan gráficos para visualización de resultados:

- *Volcano Plot* con 'ggplot()' [21], para la representación de los genes según su expresión diferencial y significancia estadística.
- Mapas de calor con *complexHeatmaps* [22], que identifican grupos y subgrupos de genes y muestras con alta similitud entre sí en valores de expresión.
- Análisis de Componentes Principales (PCA) para genes y muestras. Con la función 'prcomp()' se calculan las componentes principales de la matriz de expresión; y mediante 'fviz_eig()' y 'fviz_pca()' del paquete *factoextra* [23] se identifican los perfiles genéticos de máxima varianza, en el caso de agrupar por genes; y el agrupamiento según respuesta a tratamiento, en el caso de muestras. Los perfiles genéticos de máxima varianza son otra opción de filtrado para la selección de características.

h) Aprendizaje automático

Se crean modelos ML para estudiar la capacidad de predicción de la respuesta a MMF de los DEGs seleccionados en muestras a dos periodos de tiempo:

- Periodo temprano (*early*): conjunto de muestras tomadas hasta 6 meses después de la biopsia. Es el caso ideal, se trata de predecir la respuesta al fármaco en fases iniciales del tratamiento o, incluso, antes del diagnóstico (periodos negativos, anteriores a biopsia).
- Periodo global (*global*), o usando todas las muestras a todos los tiempos. En este caso se estudian los genes que están diferencialmente expresados de forma constitutiva en el tiempo entre respondedores y no respondedores; de forma que, los modelos ML con estos DEGs, permitan predecir si un paciente es respondedor o no a cualquier tiempo, para su posible uso en la práctica clínica rutinaria.

Una vez obtenidas las listas de genes significativos (DEGs) para cada variable respuesta y por conjunto de estudio (*global* y *early*), se preparan las matrices de expresión y metadatos necesarios para generar los modelos predictivos utilizando la función 'getML()' de la herramienta *pathMED* [10]. Entre las funcionalidades de la herramienta, 'getML()' construye modelos de aprendizaje automático (ML) a partir de datos numéricos de entrada; en este caso, modelos de predicción de respuesta a fármacos y progresión de la enfermedad a partir de datos de expresión génica. La metodología de *pathMED* fue desarrollada y puesta en uso inicialmente con datos de pacientes de SLE [1].

Preparación de los datos

Para la construcción de los modelos de aprendizaje automático se requiere como principales argumentos una matriz numérica con valores de expresión o scores de características por muestra, *expData*, y una matriz de información asociada a las muestras, *metadata*. Para *expData* se emplean las matrices de expresión filtradas por los DEGs para cada variable respuesta, uPCR y mPERR; y como *metadata* cada respuesta asociada a los códigos de muestra. Se añade la columna 'patient' con el ID del paciente,

de forma que el modelo pueda distinguir por esta variable para las particiones de entrenamiento (train) y test; ya que, al tratarse de un estudio longitudinal, algunas muestras pertenecen al mismo paciente y aquellos pacientes incluidos en el entrenamiento no deben ser utilizados en el test para prevenir el sobreajuste.

Construcción de los modelos ML en base a datos de expresión génica

La función 'getML()' ajusta el mejor modelo de predicción de la variable respuesta ('var2predict') mediante la evaluación con varios algoritmos de aprendizaje automático y utilizando validación cruzada anidada (*nested k-fold cross-validation*) (Material Suplementario, Archivo "06. PathMED.Rmd").

Los objetos de entrada son la matriz de expresión, *expData*, y el marco de metadatos de las muestras, *metadata*, anteriormente preparados. En el argumento *paired* se indica la variable pareada; es decir, aquella que la función debe distinguir para la asignación de grupos de entrenamiento y test. En el presente estudio se indica la columna 'patient' del objeto *metadata*; de forma que muestras de un mismo paciente se asignen siempre al mismo grupo (train o test).

En el parámetro *models*, con la función 'methodsML()' se genera la lista de especificaciones del modelo. Los algoritmos se obtienen internamente mediante la función 'caretModelSpec()' del paquete 'caretEnsemble' para ajuste múltiple de modelos caret. Se selecciona que aplique todos los algoritmos disponibles (*algorithms* = "all") para comparación, indicando el tipo de variable a predecir en *outcomeClass*. La variable a predecir puede ser de tipo "character" para problemas de clasificación, o "numeric" para problemas de regresión. Dado que la respuesta a fármaco es una variable de clasificación binaria, la función permite aplicar hasta 11 algoritmos ML: 'glm', 'lda', 'rf', 'knn', 'nnet', 'svmLinear','svmRadial', 'nb'(naive bayes), y variantes de DT con potenciación de gradiente ('xgbTree', 'ada', 'gamboost'). También debe indicarse la clase positiva en el argumento *positiveClass,* en este caso, "Resp".

Las particiones de datos se realizan mediante validación cruzada (*Cross-Validation*, CV) anidada; es decir, aplicando CV dentro de cada subdivisión de Cross-validation. Se indica el número de pliegues (K) para el CV externo (*Koutter*) en el que se divide la matriz de

expresión, y que está influenciado por el número de muestras evaluado; y el número de CV interno (*Kinner*) para el ajuste de hiperparámetros (*tuning process*). La CV interna se repite el número de veces indicado *reapeatsCV*. Se aplican 5 *Koutter*, 4 *Kinner* y 5 repeticiones a todos los modelos, excepto los *Koutter* para *early*, que se reducen a 4 por el bajo número de muestras.

Por último, se incluye el tercer filtro de selección de características (apartado g), indicando métodos de reducción de características/genes en el argumento *filterFeatures*. Los algoritmos disponibles en la función son la eliminación recursiva de características (*Recursive Features Elimination*, "rfe") y la selección por filtrado (*Selection by Filtering*, "sbf") [24-25]. Aplicado a DEGs:

- "rfe" selecciona los genes en base a una estimación de su importancia en la diferenciación de la variable a predecir; el estimador se entrena en el conjunto inicial de genes calculando el peso o coeficiente de cada uno en la predicción de la respuesta, y en un proceso iterativo va eliminando los genes menos importantes hasta que la eliminación no aporte mejora estadística del modelo, normalmente en base a p-valor.
- "sbf" realiza también un filtrado en base a significancia, pero la evalúa fuera del modelo ML. Posteriormente incorpora aquellas características que presentan un p-valor inferior a un determinado límite.

Se utiliza preferentemente la eliminación recursiva ("rfe"), estableciendo en 2 el número mínimo de genes a probar/conservar del total en *filterSizes*, y recalculando la significancia cada vez que se elimina una característica (*rerank* = TRUE).

Los modelos generados son evaluados en base a los estadísticos de rendimiento Matthews Correlation Coefficient (MCC), *Balanced Accuracy* (balacc) y el área bajo la curva de exactitud (AUC-ROC) y precisión (AUC-PR), obtenidas a partir de la matriz de probabilidades de clase.

Validación independiente

Se seleccionan los modelos con mejor desempeño y se validan con la cohorte de prueba (GEO nº: GSE99967) (Material Suplementario, Archivo "07. VAL.rmd").

Se toma el conjunto de sets de expresión-clínica con las 81 muestras y se verifica si existe coincidencia con la cohorte de inicio. Las muestras independientes conforman el conjunto de prueba *global*; de ellas se extrae el conjunto *early* de validación a partir del tiempo de biopsia (en días), disponible en el set clínico. Tras verificar que la matriz de expresión se encuentra normalizada y anotada a *GeneSymbol*, se generan los conjuntos de expresión filtrando por los genes significativos para cada modelo.

En la validación se utiliza la función 'predict()', para la obtención de las predicciones. Se evalúan como métricas de rendimiento las curvas AUC-ROC y AUC-PR, los estadísticos MCC, *accuracy*, *kappa* y la matriz de confusión.

4. Resultados

4.1. Estadística Descriptiva

Se obtienen las muestras clasificadas por respuesta/no respuesta tanto para uPCR como para mPERR; en frecuencias absolutas mediante la función' table()' de R (Tabla 2), y gráficamente en frecuencias relativas (Figura 1):

Muestras	uPCR	mPERR
Respondedoras (y)	76	78
No Respondedoras (n)	30	28

Tabla 2. Muestras según respuesta/no respuesta a MMF en frecuencias absolutas.

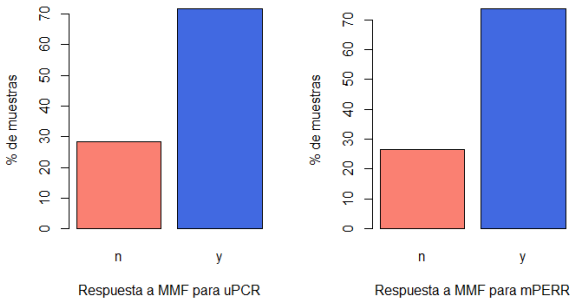


Figura 1. Muestras según respuesta/no respuesta a MMF en frecuencias relativas (%).

Las frecuencias entre muestras respondedoras y no respondedoras son prácticamente coincidentes para uPCR y mPERR, en torno al 73% de respuesta positiva. Se comprueba que, de las 106 muestras, 12 presentan respuestas no coincidentes entre ambas variables; esto es debido a que mPERR incluye a uPCR, pero además tiene en cuenta eGFR.

De forma análoga, se obtiene la clasificación para pacientes (Tabla 3 y Figura 2):

Pacientes	**uPCR**	**mPERR**
Respondedores (y)	25	24
No Respondedores (n)	12	13

Tabla 3. Pacientes según respuesta / no respuesta a MMF en frecuencias absolutas.

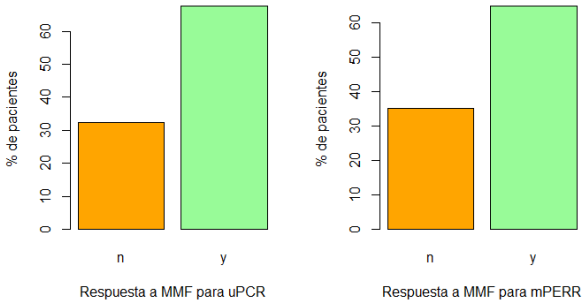


Figura 2. Pacientes según respuesta/no respuesta a MMF en frecuencias relativas (%)

De nuevo, las frecuencias entre muestras respondedoras y no respondedoras son prácticamente coincidentes para uPCR y mPERR, en torno al 66% de respuesta positiva. En este caso, de los 37 pacientes, 5 presentan respuestas no coincidentes entre ambas variables.

4.2. Estadística Analítica: variables clínicas

Tras el procesado de tablas clínicas, las 24 tablas de la Figura 1 quedan reducidas a 7 de interés para el análisis estadístico, 4 identificadas por visita/muestra, y 3 por paciente. Los conjuntos de datos finales unificados por paciente y muestra presentan las siguientes dimensiones:

- *clin_by_patient*: 37 pacientes, 170 variables
- *clin_by_sample:* 106 muestras, 82 variables

<u>Variables dicotómicas</u>

En la Tabla 4 se muestran las variables categóricas binarias que obtienen significancia estadística en el test de Fisher, p-valor < 0.05, para cada variable respuesta, por paciente y por muestra:

	Respuesta	Variable	p.value	odds.ratio
Paciente	uPCR / mPERR	Cloroquina (alguna vez)	0.0283	0.0000
			0.0368	0.0000
		Pleuresía	0.0323	0.1659
			0.0075	0.1319
		Prednisona (actual)	0.0128	0.1199
			0.0382	0.1891
	uPCR	Alopecia	0.0173	0.1410
		Esteroides tópicos (alguna vez)	0.0486	0.2009
		Trombosis venosa	0.0184	Inf
	mPERR	Síndrome nefrórico	0.0165	0.1374
		Pericarditis	0.0282	0.1743
Muestra	uPCR / mPERR	Proteinuria (SLEDAI renal)	8.42e-06	0.0751
			0.0021	0.1812
		Bajo complemento (SLEDAI inmunología)	2.24e-04	**16.6232**
			0.0005	**14.8701**
		Aumento en unión ADN (SLEDAI inmunología)	1.17e-03	**6.7990**
			0.0023	**6.0223**
	uPCR	Hematuria (SLEDAI renal)	0.0351	0.2353
		Alopecia (SLEDAI piel)	0.0262	0.3071
	mPERR	Inhibidor ACE	0.0318	0.2846
		Antihipertensivos	0.0436	0.2572

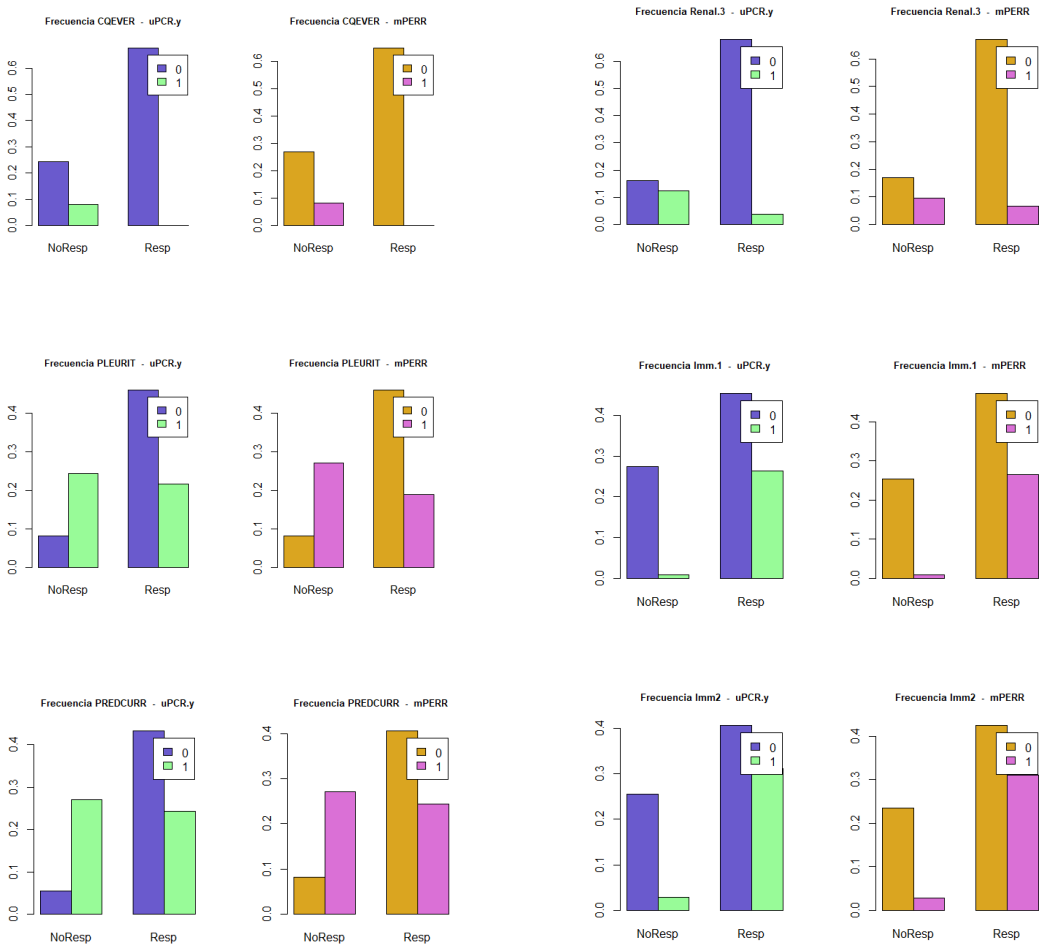
Tabla 4. Variables dicotómicas significativas por test de Fisher para la respuesta a MMF.

Por paciente, se obtienen como significativos varios aspectos sintomatológicos del SLE: como la alopecia y la trombosis venosa para uPCR, y el dolor torácico pleurítico, tanto para uPCR como para mPERR. Además, aparecen tres descriptores relacionados con terapias asociadas: el uso de esteroides tópicos para uPCR; y para ambas respuestas el uso de coloroquina (CQ) y el tratamiento con Prednisona, este último referido a tratamiento en el momento de la toma de datos, por lo que es de especial relevancia.

Para muestras, se obtienen como especialmente significativas para ambas variables los factores inmunológicos de la actividad SLEDAI, tanto para aumento de la unión de ADN, como por bajos niveles de complemento. Otras variables significativas para uPCR y relacionadas con la actividad (SLEDAI) son hematuria y proteinuria, esta última también para mPERR); así como la alopecia, ya obtenida significancia en el set de pacientes. Con mPERR, además, se obtienen como significativas los tratamientos con inhibidor de ACE o fármacos para la hipertensión (anti HTN).

El odds.ratio es una medida de la influencia de cada variable en la respuesta, razón de probabilidades de ocurrencia de cada grupo (respondedores y no respondedores). OR > 1 implica un aumento de la probabilidad de una determinada respuesta; sin embargo, el valor "Inf" es resultado de la división por cero, al carecer uno de los grupos de representación en la variable. Por tanto, este valor no es indicativo de significancia, ya que está relacionado con la falta de datos para el análisis.

En la Figura 3 se muestran gráficos de barras apiladas de las variables dicotómicas significativas para uPCR y mPERR frente a cada respuesta, a modo de comparativa:



a) Por paciente

b) Por muestra

Figura 3. Gráficos de barras apiladas con las frecuencias relativas de las variables dicotómicas significativas por test de Fisher para cada grupo de respuesta en uPCR y mPERR. a) Variables por paciente. b) Variables por muestra. Leyenda: 0 = ausencia; 1 = alguna vez presente; CQEVER = cloroquina; Pleurit = Pleuresía; PREDCURR = Prednisona (actual); Renal3 = Proteinuria, Imm1 = Bajo complemento; Imm2 = aumento en unión de ADN.

Las tres variables por paciente (tratamiento con CQ, prednisona actual, dolor pleurítico) y la actividad relativa a proteinuria presentan un comportamiento similar; con respuesta diferenciada cuando no están presentes (mayor proporción de respuesta positiva); mientras que cuando hay presencia de estos factores, no se observan grandes diferencias en los grupos de respuesta. Si bien, la presencia de proteinuria es muy elevada en la población de estudio (son pacientes con progresión a LN); y en caso de tratamiento con CQ (1) no se han registrado casos de respondedores, lo que explica el valor OR=0. Los factores inmunológicos de SLEDAI, por el contrario, muestran un aumento significativo en la respuesta a tratamiento cuando están presentes.

Variables categóricas/multivalor

Se obtiene la raza como única variable significativa para *uPCR* y *mPERR* en muestras, obteniendo p-valor < 0.05 en el test Chi-cuadrado.

En la Tabla 5 y la Figura 4, se muestran, respectivamente, los resultados estadísticos y los gráficos de barras apiladas con las frecuencias relativas de la raza para ambas respuestas a tratamiento, en el caso de pacientes.

	Respuesta	p.value	chi.squared
Raza **Paciente**	uPCR	0.0370	7.5972
	mPERR	NoSig	NoSig
Raza **Muestra**	uPCR	0.0005	29.7078
	mPERR	0.0225	9.8261

Tabla 5. Tabla de estadísticos del test Chi-Cuadrado para la variable multinivel raza.

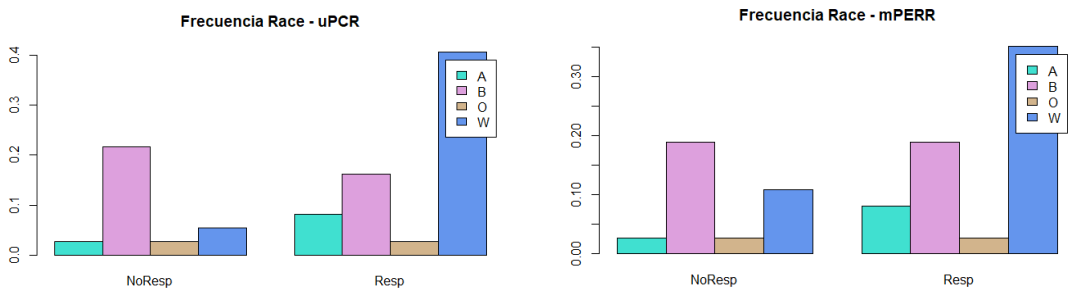


Figura 4. Gráficos de barras apiladas con las frecuencias relativas de la raza en pacientes para cada grupo de respuesta, uPCR y mPERR. Leyenda: A = asiática; B = negra; O = otra, W = blanca

Se observa que la cohorte está sesgada para la raza, ya que la mayor parte de muestras no respondedoras, especialmente en uPCR, pertenecen a raza negra; hecho ya constatado en el estudio original de Toro-Domínguez, et al [2].

Variables continuas

El análisis por paciente solo obtiene como significativas (p-value <0.05) para uPCR: la dosis más alta de MMF (CellCept) y los meses de duración del tratamiento con alta dosis de Prednisona. No se obtiene significancia para mPERR.

La mayor parte de variables continuas son medidas durante las visitas y se encuentran recogidas en el set de muestras. En la Tabla 6 se resumen aquellas que obtienen significancia tanto para uPCR como para mPERR.

Variable	**Respuesta**	**Normalidad (Grupos No resp / Resp)**	**Test**	**p.value**
PGA	uPCR	No	WX / U-MW	5.52e-09
	mPERR	No	WX / U-MW	1.20e-07
Dosis Aspirina	uPCR	No / NA	WX / U-MW	0.0268
	mPERR	No	WX / U-MW	0.0020
Dosis 1 MMF	uPCR	No / Sí	WX / U-MW	6.72e-06
	mPERR	No	WX / U-MW	0.0046
C3	uPCR	Sí	T-Test	8.23e-05
	mPERR	Sí / NA	WX / U-MW	2.96e-04
C4	uPCR	Sí	T-Test	1.92e-05
	mPERR	Sí / NA	WX / U-MW	0.0014
Anticuerpos anti-DNA	uPCR	No	WX / U-MW	0.0028
	mPERR	No / NA	WX / U-MW	0.0042
IgA.ACL (anti-cardiolipina	uPCR	No	WX / U-MW	0.0150
	mPERR	No / NA	WX / U-MW	0.0244
Phys Estimate o LAI (0-3)	uPCR	No	WX / U-MW	5.52e-09
	mPERR	No	WX / U-MW	1.20e-07
LAI renal	uPCR	No	WX / U-MW	1.78e-11
	mPERR	No	WX / U-MW	2.64e-08
UrPr (Proteína en orina)	uPCR	No / NA	WX / U-MW	1.29e-09
	mPERR	No / NA	WX / U-MW	6.98e-04

Tabla 6. Variables continuas por muestra y significativas para la respuesta a MMF por uPCR y mPERR. WX / U-MW: Test de Wilcoxon o U de Mann Whitney

No se incluyen en la Tabla 6 variables relacionadas directamente con la medida de la respuesta a fármaco, como son la propia uPCR y la relación Proteína Urina / Creatinina

(UrPrCrRatio), ambas significativas para las dos respuestas; el nivel de Creatinina (UrCrRndm) para uPCR, y eGFR para mPERR.

Excepto en el caso de los complementos C3 y C4, el resto de variables han mostrado una distribución no normal; por tanto, se selecciona la mediana como valor central para la comparación de los valores medios de las variables por grupo de respuesta. La mediana es también más representativa para variables continuas discretas, con pocos valores únicos, como las dosis de aspirina y MMF. En la Tabla 7 se muestran los resultados:

	PGA	Dosis Asp.	Dosis MMF	C3	C4	DNA	IgA. ACL	LAI	LAI Renal	UrPr
Resp uPCR	0.5	0	1500	94.5	17	0	3.0	0.5	0.0	0.0
NoResp uPCR	1.8	81	3000	127.0	25	0	4.5	1.8	1.8	2.0
Resp mPERR	0.5	0	1500	96.0	17	0	3.0	0.5	0.0	0.5
NoResp mPERR	1.8	81	3000	122.0	25	0	5.0	1.8	1.8	2.0

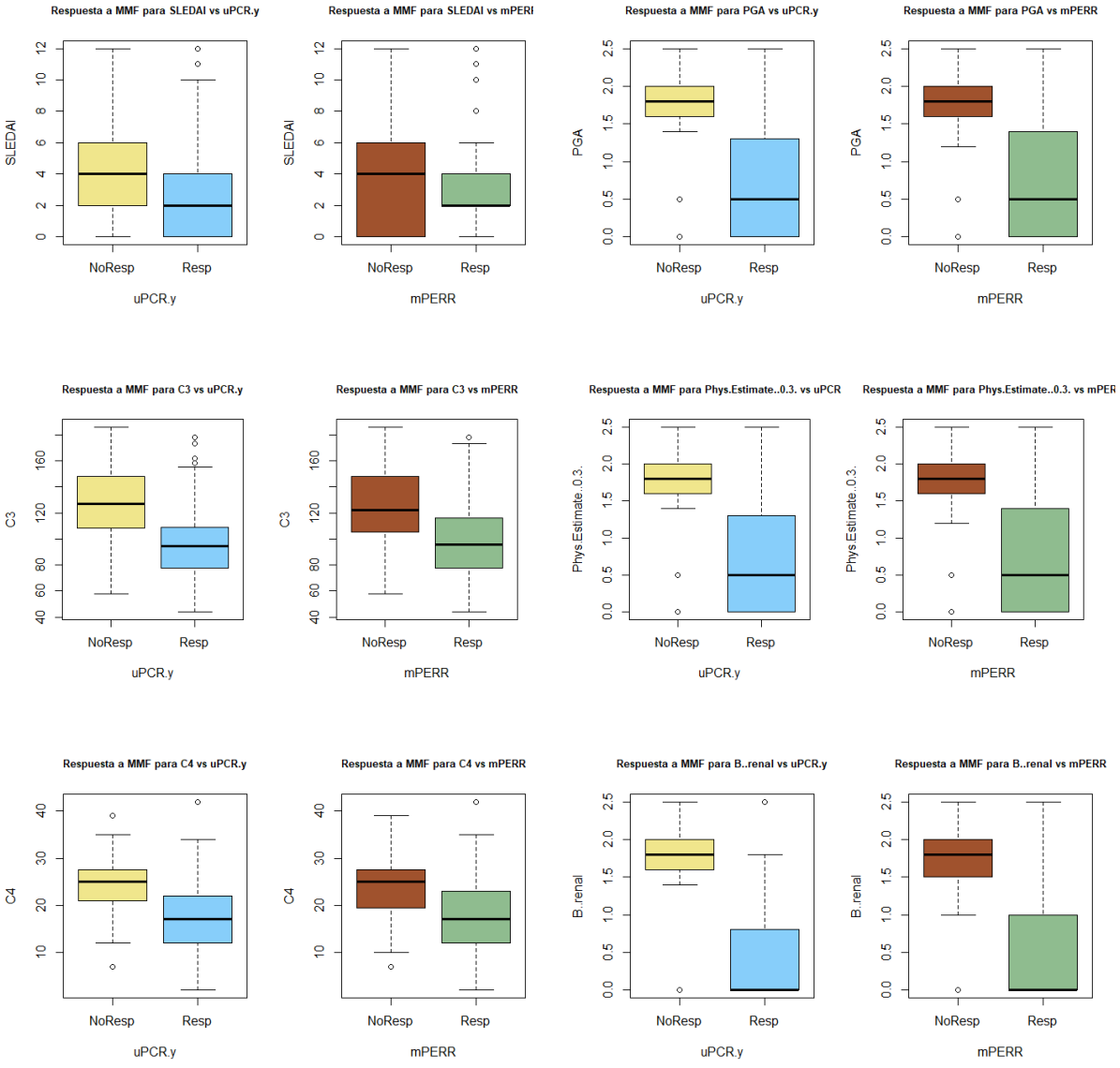
Tabla 7. Estadísticos centrales para las variables continuas significativas por grupo de respuesta a MMF, para uPCR y mPERR.

Se observa correspondencia entre los índices de actividad PGA y LAI, y entre grupos respondedores y no respondedores para ambas variables, uPCR y mPERR. En el caso de anti-dsADN, los valores centrales no discriminan la respuesta.

Se obtiene significancia estadística en un gran número de variables continuas: Relacionadas con índices de actividad: SLEDAI (solo uPCR), PGA, Phys.Estimate 0-3 (Índice de Actividad de Lupus, LAI); y a destacar el factor Renal para la puntuación LAI. Relacionados con dosis y periodo de tratamiento: dosis de aspirina, dosis de MMF (dose1) y la duración del tratamiento con MMF (solo uPCR). Procedentes de analítica de laboratorio, medidas del sistema inmunológico; como los niveles C3 y C4 del complemento. También procedentes de analítica y medidas del estado renal: Proteína en orina o UrPr (proteinuria), además de UrPrCrRatio y UrCrRndm ya mencionadas. Los resultados de analítica confirman la significancia obtenida en factores inmunológicos y renales de la actividad medida en variables dicotómicas.

Otras variables significativas han sido las medidas de tensión arterial para uPCR y el colesterol y el número de glóbulos blancos (WBC) para ambas respuestas; medidas con alta variabilidad en el tiempo y de naturaleza heterogénea, por lo que no se atribuye una relación directa con la enfermedad.

En la Figura 5 se muestran los gráficos de cajas para la visualización de la distribución de las variables continuas significativas en cada grupo de respuesta a MMF por uPCR y mPERR. Se incluyen como relevantes las relativas a medida de la actividad y asociadas, los niveles del complemento C3 y C4 (SLEDAI) y estado renal (LAI).



a) SLEDAI y niveles C3 y C4

b) PGA, LAI y factor renal para LAI

Figura 5. *Boxplots* con la distribución de valores de las variables continuas significativas para cada grupo de respuesta en uPCR y mPERR. a) Variables por actividad SLEDAI b) Variables por actividad PGA y LAI.

Se obtiene correlación en la respuesta a fármaco para los niveles de complemento C3 y C4 y el índice SLEDAI, por una parte; con mayor tasa de no respuesta uPCR y mPERR a niveles altos de las variables (mayor actividad, alto complemento). Por otro lado, la variación con la respuesta del índice PGA y el índice LAI son prácticamente coincidentes (como indicaba la Tabla 7); al igual que SLEDAI, presentan menos respuesta a mayor actividad, existiendo una mayor diferencia en las medianas de los grupos (valor medio de actividad por grupo de respuesta). Sin embargo, los grupos respondedores presentan una dispersión mayor, que abarca la totalidad del rango del índice. El estado renal asociado a LAI se encuentra correlacionado, con mayor tasa de no respuesta a mayor avance de la enfermedad renal.

Covariables para análisis DEG

De las variables significativas para expresión génica indicadas en la sección 3g, en este análisis se ha obtenido significancia en la dosis de MMF y la raza, aunque debido al sesgo observado en no respondedores para la raza, no se incluye esta covariable ya que el modelo no podría distinguir entre los cambios de expresión debidos a raza y los cambios debidos a no respuesta. Tampoco se incluye la edad, ya que no se dispone de información sobre la edad de los pacientes en el momento de la visita y toma de muestra, sino de la edad de detección de los primeros síntomas (Age.Onset) y la edad de diagnóstico de la enfermedad (Age.Dx). Además, en el análisis estadístico no se observaron diferencias entre grupos de respuesta para estas variables.

Por tanto, se tienen en cuenta tanto el sexo como la dosis de fármaco, que se incluye como dosis máxima calculada a partir de `dose1` y `dose2` para muestras, "*MMFdose*". A estas covariables se añade la dosis de prednisona, correspondiente al co-tratamiento base y mayoritario en la cohorte.

Por otro lado, teniendo en cuenta las variables categóricas significativas obtenidas; componentes inmunológicos, renal y alopecia, todas relacionadas con el índice SLEDAI, se puede utilizar este índice como covariable y medida general de la actividad. SLEDAI se construye a partir de la sumatoria de ciertas variables clínicas del paciente, algunas

de las cuales han mostrado significancia respecto a la respuesta a MMF. Para su inclusión como covariable, se ha recalculado la "*actividad*" agrupando los valores del índice SLEDAI en tres categorías: < 3 = 1, 3-7 = 2, >7 = 3 [26].

En la Figura 6 se muestran los gráficos de cajas para las covariables continuas, incluyendo las nuevas variables calculadas para *actividad* y *MMFdose*, y el gráfico de barras para la covariable *sex*, indicada por paciente.

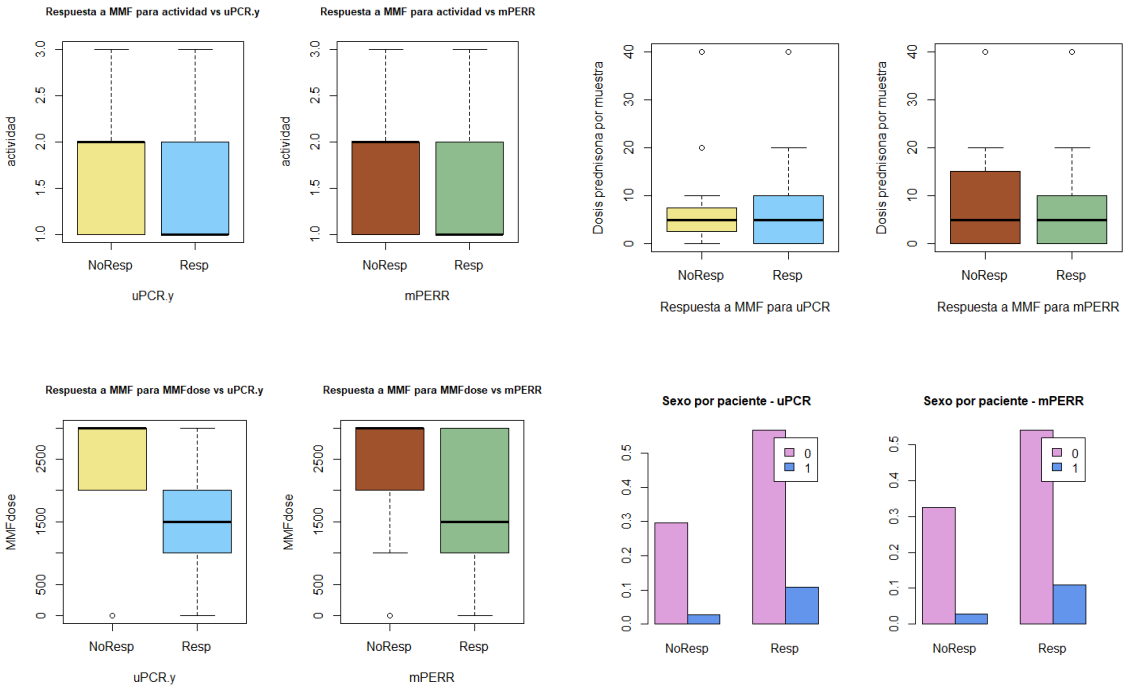


Figura 6. Gráficos de cajas con la distribución de valores de las covariables continuas para cada grupo de respuesta en uPCR y mPERR; y gráfico de barras agrupadas para la covariable *sexo* (0 = mujer; 1 = hombre).

La distribución de datos de *actividad* es homogénea para ambos grupos de respuesta, encontrándose las medianas más diferenciadas tanto para uPCR como mPERR; de modo que, índices de actividad más bajos correlacionan con respuesta positiva a MMF, y viceversa; lo cual está en consonancia con la variación observada en las variables categóricas asociadas a SLEDAI. *MMFdose* presenta una distribución diferenciada entre los grupos, con mayor número de no respondedores a dosis altas. De forma lógica, ante la no respuesta del paciente, se trata de subir la dosis; si la respuesta es buena, se intenta disminuir. En cuanto al sexo, la mayor proporción de mujeres en uno y otro grupo se debe a una mayor representación en la cohorte, por la propia prevalencia de la enfermedad (ratio 9:1) [4], y no al desbalanceo de clases.

4.3. Anotación de genes y Filtrado por correlación

Se parte de la matriz de expresión filtrada por las muestras de estudio (106) y las muestras de sanos (20), con los 54715 *Probes Set ID* en las filas y valores de expresión medidos en escala log2 y pre-normalizados por RMA.

No se encuentran genes con varianza cero tras aplicar la función 'nearZeroVar()' a la traspuesta de la matriz de expresión (por defecto evalúa los índices de las variables).

Anotación de *Probes Set ID* a *GeneSymbol*

Se realiza la búsqueda en las bases de datos de *Biomart* para establecer los atributos de la función: '*affy_ht_hg_u133_plus_pm*' (tipo de *Probe Set ID* utilizado, consultado en la plataforma de *Affymetrix* GPL13158) como valores objetivo, y '*external_gene_name'* (*GeneSymbol*) a anotar.

Se obtiene una matriz de anotación con 48891 observaciones, que contiene 43325 *probes affy* y 22566 *GeneSymbol* distintos, lo cual implica que algunos de estos genes anotan a varios *Probes Set ID*. El total de *probes* únicos se reduce a 41077 y los genes quedan en 22565 tras la eliminación de filas con datos faltantes (lo que supone eliminar *probes* que no anotan a ningún gen), y el filtrado por los *affy* presentes en la matriz de expresión.

Por último, con la función 'AnnotateGenes()' se realiza la fusión de datos de expresión de los *probes* dirigidos a un solo gen, imputando el valor de la mediana de expresión a dicho gen. La función permite realizar paralelamente todas las operaciones de anotación con *Biomart*, eliminación de datos faltantes, etc.; si bien, se parte de la matriz filtrada por *affys* anotados no únicos (48860, sin NA), de forma que solo se aplica el cálculo de las medianas de expresión para reducir al número de genes únicos.

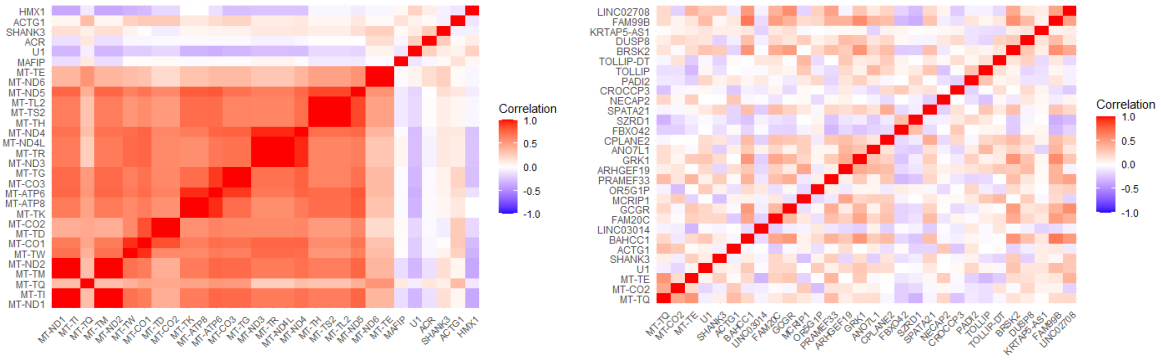
Finalmente, se obtiene la matriz de expresión génica de las 126 muestras (columnas) para los 22565 genes únicos, identificados por *GeneSymbol* en las filas (*rownames*).

Filtrado de genes por correlación

De los 22565 genes anotados en la matriz de expresión, se obtienen un total de 3113 genes correlacionados, en base al *cutoff* de 0.75 (correlación máxima permitida)

establecido en la función 'findCorrelation()'. Los genes correlacionados son eliminados de la matriz de expresión, quedando un total de 19452 genes.

En la Figura 7 se visualizan el resultado del filtrado por correlación con gráficos 'ggplot()':



a) Gráfico antes del filtrado

b) Gráfico después del filtrado

Figura 7. Gráficos de correlación por pares de genes para el subconjunto de los 30 primeros genes en a) Matriz de expresión sin filtrar por correlación; b) Matriz de expresión filtrada por correlación

4.4. Selección de características por análisis DEG

Se seleccionan las 106 muestras de la población de estudio en la matriz de expresión para la evaluar DEGs asociados a la respuesta global a MMF. Y esta matriz, a su vez, es filtrada por muestras tomadas hasta 6 meses después de la biopsia; que servirán para obtener genes significativos en periodo temprano de diagnóstico. Se obtienen 22 muestras con este criterio (*early*).

El análisis para obtener los DEGs se aplica para cada conjunto de muestras y cada variable respuesta. En la Tabla 8 se muestra la comparativa de la distribución y proporción de muestras por grupo de respuesta:

	uPCR			**mPERR**		
	Resp	NoResp	**%Resp**	Resp	NoResp	**%Resp**
global	76	30	72	78	28	74
early	16	6	73	14	8	64

Tabla 8. Distribución y proporción de muestras en los grupos de respuesta a MMF por uPCR y mPERR.

La proporción de respuesta positiva se mantiene en ambos conjuntos de muestras para uPCR, y resulta algo inferior en *early* para mPERR.

En ambos casos se aplica el flujo de análisis DEG para cada respuesta empleando el paquete *limma*, con ajuste "BH" y corrigiendo por las covariables seleccionadas en el apartado 4.2. Como referencia se establece el grupo no respondedor (NoResp).

En la Tabla 9 se muestran los 10 genes mejor clasificados según p-valor ajustado < 0.05, obtenidos con 'topTable()' del modelo de análisis DEG ajustado para cada conjunto de muestras y variable respuesta. En el caso de uPCR para *early,* solo se obtienen dos genes significativos.

Resp	Muestras	Gen	Avg.Expr	logFC	adj.P.Val
uPCR	*global*	TUBB2A	7.1369	-1.6548	4.36e-07
		FGF13	2.5978	0.7258	4.62e-04
		CD180	6.7298	-0.6342	4.95e-04
		LINC02915	4.5552	0.6583	1.22e-03
		S100P	10.3538	1.3845	1.22e-03
		LILRA5	6.1716	0.7736	1.22e-03
		HSPA1B	7.5475	0.5146	1.27e-03
		MIR4648	9.8923	-0.2898	1.27e-03
		PSPHP1	6.1526	-2.0062	1.27e-03
		SRBD1	6.5839	0.5168	1.40e-03
	early	MCTP1	7.0043	1.4324	0.0189
		FGF13	2.9993	1.3834	0.0189
mPERR	*global*	MAP2K6	6.4296	0.7438	0.0030
		RNF182	4.1028	-2.0426	0.0030
		SLFN5	7.1226	0.8054	0.0030
		EEF2	10.3207	-0.3324	0.0030
		TUBB2A	7.1369	-1.1739	0.0082
		PNPT1	5.8560	0.9405	0.0117
		BICD1	3.8758	-0.2762	0.0124
		ZDHHC13	5.7381	0.3958	0.0156
		DMPK	5.2175	-0.4916	0.0156
		ETV7	5.9657	0.7759	0.0172
	early	BTNL3	7.0664	2.8117	0.0022
		RNF182	5.0535	-4.0760	0.0038
		PLD6	8.5075	-0.9271	0.0038
		SLFN5	7.0703	1.3750	0.0065
		MCTP1	7.0043	1.2347	0.0085
		UBL7	6.6399	-0.7442	0.0326
		ACAP1	7.3251	-0.7178	0.0399
		MAGED4B	4.4728	-0.7227	0.0438
		TUBB2A	7.0903	-1.8148	0.0457
		SIGIRR	7.9615	-0.8924	0.0471

Tabla 9. Top10 DEGs por uPCR y mPERR extraídos con topTable() para cada conjunto de muestras. Para *early* uPCR se muestran solo los 2 significativos. Se muestran los estadísticos p.valor ajustado, logFC y la expresión promedio del gen (avg). Leyenda: Verde = genes comunes para todos, excepto *early uPCR*; Morado = genes comunes para *early*; Azul = genes comunes para *uPCR*; Naranja = genes comunes para *mPERR.*

Se encuentran varios genes comunes entre grupos en el top10: TUBB2A es significativo para todas excepto uPCR *early* y MCTP1 para *early*; FGF13 es significativo para uPCR; y RNF182 y SLFN5 son significativos para mPERR.

Para extraer la lista de genes significativos calculando el tipo de expresión diferencial se emplea la función 'decideTest()'. Alternativamente, se crea una nueva columna, "expression", para guardar el tipo de DEG en cada modelo ajustado. En ambos casos se utiliza como nivel de significancia estadística los parámetros p-valor ajustado < 0.05 y se deja por defecto el corte en logFC >=0.

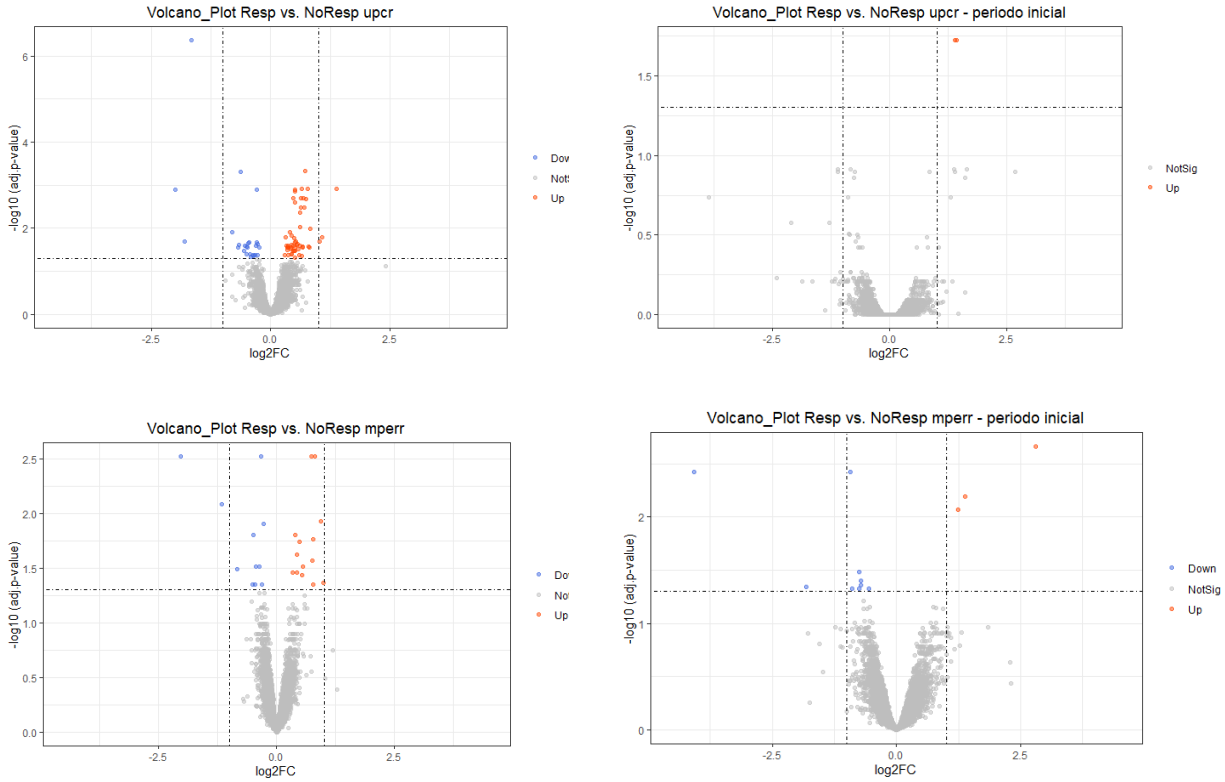
En la Tabla 10 se muestran los DEGs para cada conjunto de muestras y grupo de respuesta, indicando el signo de la expresión.

	uPCR			**mPERR**		
	Up	Down	**Total**	Up	Down	**Total**
global	52	25	77	14	11	25
early	2	0	2	3	9	12

Tabla 10. DEGs por tipo de expresión para cada conjunto de muestras y variable respuesta.

77 genes para *uPCR global*, 25 genes para *mPERR global*, 2 genes para *uPCR early* y 12 genes para *mPERR early* son los candidatos al proceso de selección de características de los modelos predictivos de pronóstico y respuesta a tratamiento con MMF.

En la Figura 8 se representan los sets completos de genes según su signo de expresión o FC comparando respondedores versus no respondedores, y significancia estadística para cada conjunto de muestras y variable respuesta, mediante Volcano_plots [21]



a) Respuesta *global*

b) Respuesta *early*

Figura 8. Volcano_plots para cada conjunto de muestras y variable respuesta.
Leyenda: rojo = sobreexpresados; azul = reprimidos; gris = no significativos.

En los ejes de los gráficos *Volcano* se delimitan los criterios logFC (x) y adj.p.value (y) para establecer la significancia del gen y el signo de la expresión. Se observa que una gran proporción de genes significativos en *global* se encuentran en torno a logFC = 0, por lo que un aumento en este parámetro reduciría ampliamente el número seleccionado. Esto se debe a que los cambios en las expresiones entre respuesta y no respuesta no son excesivamente elevados para la mayoría de los genes.

A continuación, mediante mapas de calor o *heatmaps* y en combinación con técnicas de *clustering*, se representa la matriz de expresión ordenada por similitud entre perfiles de expresión de grupos de muestras y genes. Se genera el objeto de clase 'HeatmapAnnotation()' del paquete *complexHeatmap*, en el que se indica la asignación de colores a los grupos a contrastar (respuesta a MMF) que se aplicará en el *heatmap*. Se realiza un escalado previo por filas de la matriz de expresión, mediante la función 'scale()' de R(base); y en la función 'Heatmap()' se indica que agrupe columnas y filas en base a la respuesta (muestras y genes, respectivamente).

En las Figuras 9 y 10 se muestran los *heatmaps* obtenidos para cada respuesta en los conjuntos de muestras *global* y *early*, respectivamente:

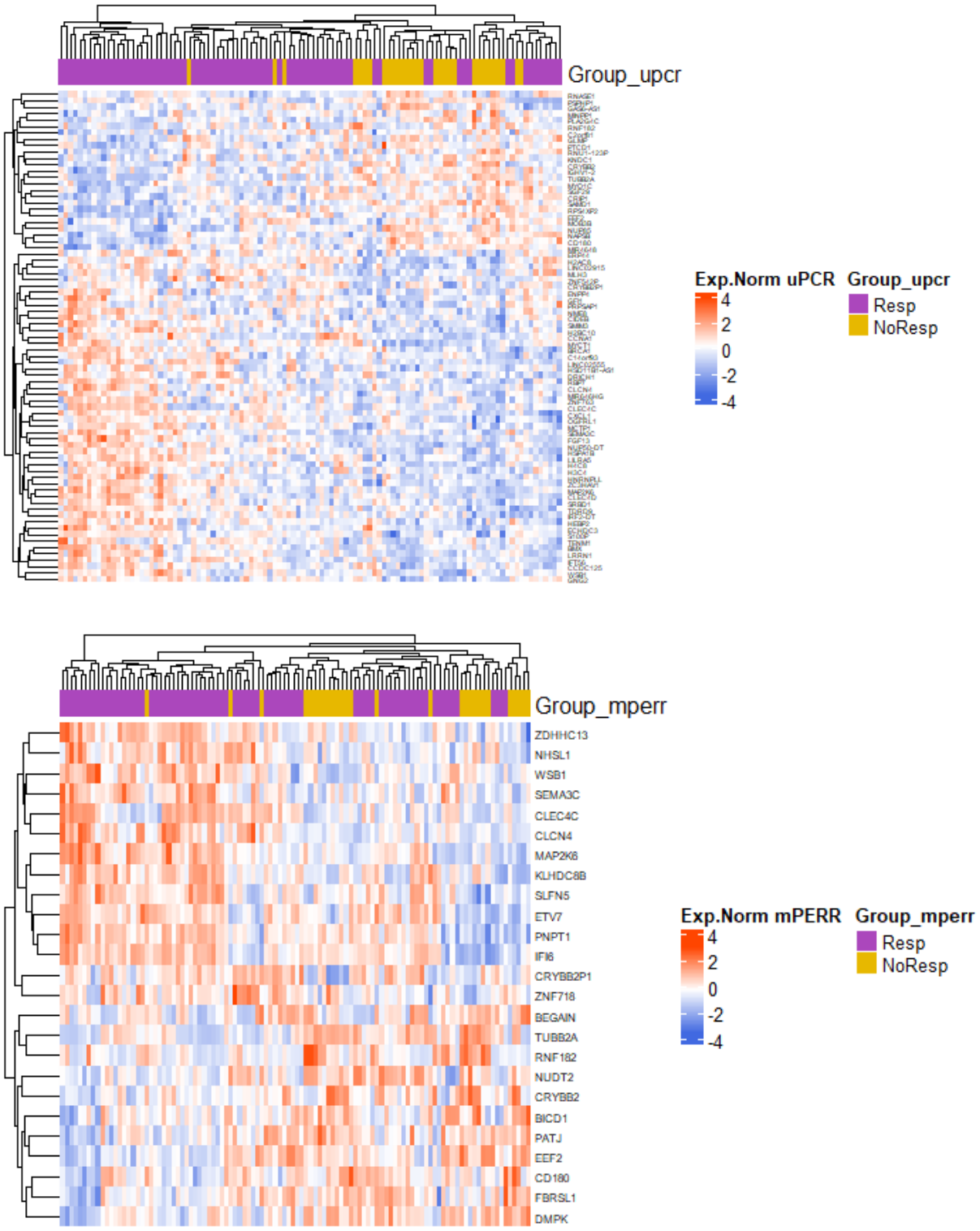


Figura 9. *Heatmaps* para conjunto *global* de muestras por grupo de respuesta a MMF. Arriba, expresión normalizada para los 77 genes significativos para uPCR. Abajo, expresión normalizada para los 25 genes significativos en mPERR.

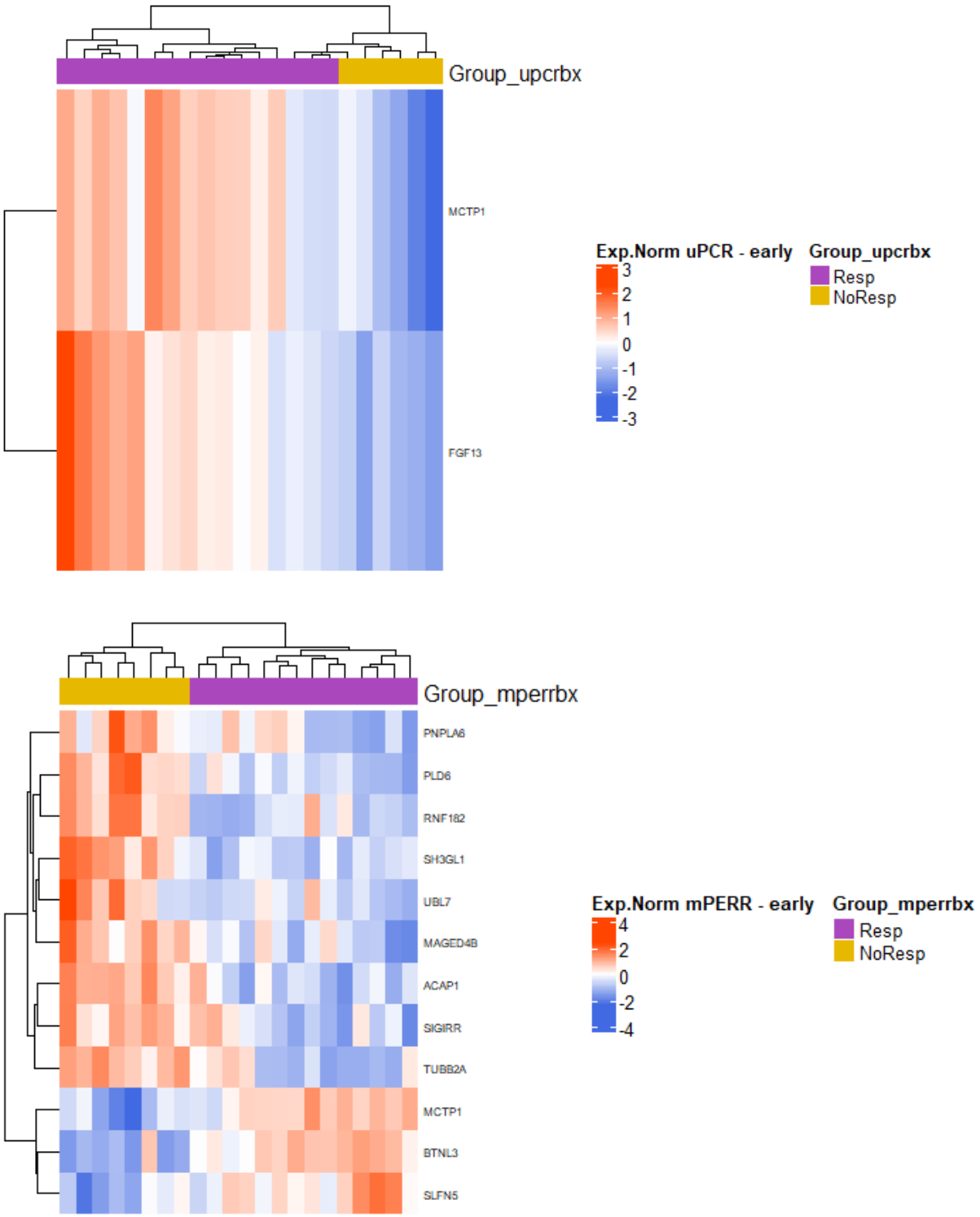


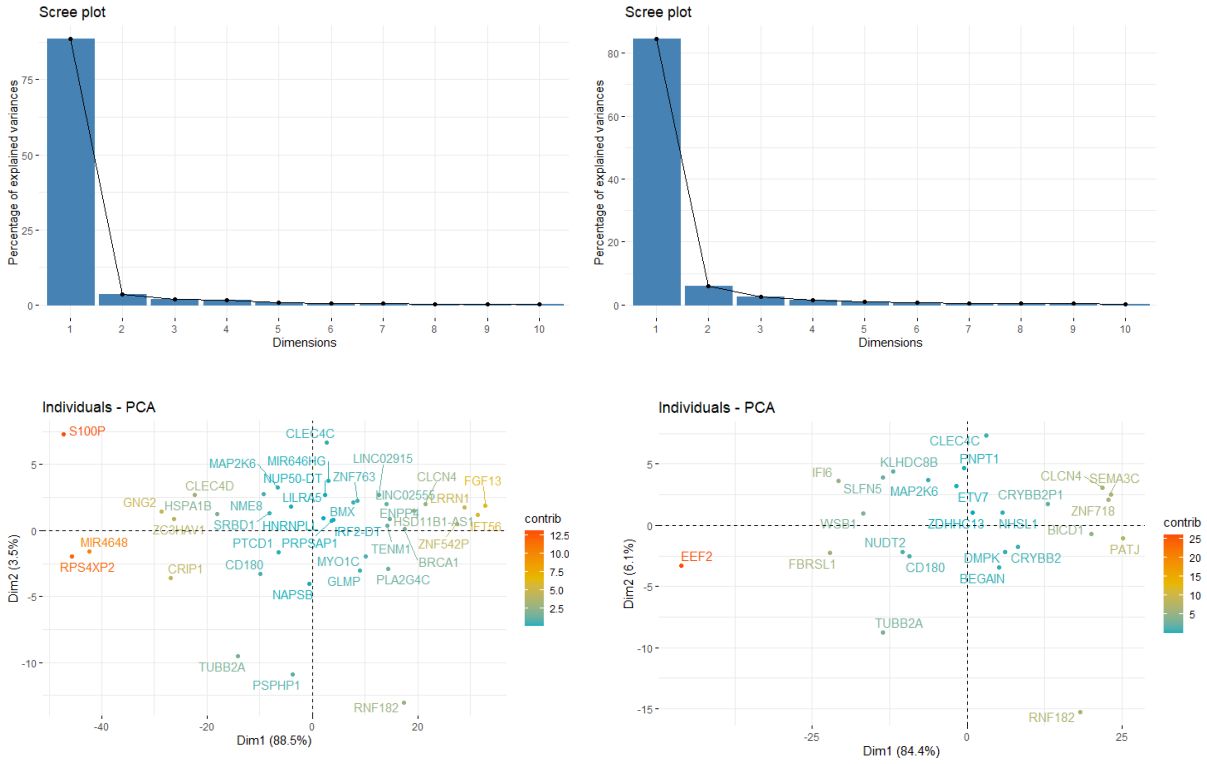
Figura 10. *Heatmaps* para conjunto *early* de muestras por grupo de respuesta a MMF. Arriba, expresión normalizada para los 2 genes significativos en uPCR. Abajo, *expresión normalizada* para los 12 genes significativos en mPERR.

Se observan patrones de expresión entre grupos de genes, diferenciados para respuesta y no respuesta a MMF, especialmente en los *heatmaps* de uPCR. Esto sugiere la posible utilidad de los genes seleccionados para la construcción de modelos predictivos de respuesta al fármaco, basados en dichos patrones moleculares.
Los dendogramas identifican grupos y subgrupos de genes y muestras con alta similitud entre sí en valores de expresión. En el conjunto *early*, con un menor número de muestras, se observa una agrupación completa por nivel de respuesta, que no llega a darse en el conjunto *global*. Esto se debe a que el conjunto *global* incluye muestras muy heterogéneas, con seguimiento de hasta 10 años. Acotando los periodos mejora la similitud de las muestras y sugiere una mayor capacidad predictiva de los modelos *early*.

Finalmente, se realiza un PCA para estudiar la relevancia de grupos de genes y posibilitar la reducción de la dimensionalidad, especialmente en el conjunto *global*; así como para visualizar la diferenciación de grupos de muestras por respuesta al fármaco.

PCA por genes

En la Figura 11 se muestran los perfiles genéticos de máxima varianza para *global*, obtenidos de la agrupación de genes significativos en las componentes principales para cada variable respuesta. Para uPCR se representa el subconjunto con los primeros 40 genes (más significativos, ordenados por p-valor ajustado).



a) uPCR *global* – PCA Genes

b) mPERR *global* – PCA Genes

Figura 11. Gráficos 'fviz_eig()' (barras) y 'fviz_pca()' (coordenadas) para PCA agrupado por genes. Los gráficos de barras representan cuánta variabilidad de expresión está explicada por cada una de las componentes principales (PC) calculadas. Los gráficos de coordenadas muestran la contribución a la variabilidad de expresión / PC de cada gen, en una escala de color.

Para ambas variables respuesta, más del 80% de la varianza en los datos está explicada por la primera componente. En los ejes del gráfico de coordenadas se muestra el porcentaje de variabilidad que representa cada componente, 88.5% PC1 uPCR y 84.4% PC1 mPERR. Los colores más anaranjados representan los genes con mayor variabilidad en expresión; en el caso de uPCR: S100P, RPS4XP2 y MIR4648. Para mPERR destaca EEF2.

PCA por muestras

En la Figura 12 se muestran los gráficos de coordenadas 'fviz_pca()' con el agrupamiento de muestras en función de la respuesta a MMF para cada variable respuesta y conjunto de estudio.

a) *global* PCA Muestras– uPCR y mPERR

b) *early* PCA Muestras– uPCR y mPERR

Figura 12. Gráficos de coordenadas ‘‘fviz_pca()’ para PCA agrupado por muestras en función de la respuesta a MMF por uPCR y mPERR.

La proporción de variabilidad está más repartida en las muestras, teniendo en cuenta todos los genes: PC1 representa menos del 30% de variabilidad para uPCR *global* y el 37% *para* mPERR. Las proporciones aumentan en *early*, al 86% y 72%, respectivamente.

Se observa cierto solapamiento entre grupos de respuesta para el conjunto *global*, tanto para uPCR como para mPERR. Dado que los gráficos por genes (Figura 11) muestran que PC1 explica más del 80% de la variabilidad en la expresión génica para ambas respuestas; se extraen los 10 genes principales para esta componente, de forma que puedan probarse como características en la generación de modelos ML, reduciendo la dimensionalidad de los conjuntos *global*.

Por otro lado, la diferenciación de muestras entre grupos de respuesta mejora para *early* con respecto a *global*, especialmente para mPERR. El gráfico muestra una mayor homogeneidad en los grupos *early* y como el conjunto de genes definido permite la diferenciación clara entre muestras de cada grupo de respuesta a fármacos.

4.5. Modelos de predicción de respuesta a tratamiento con MMF

En la Tabla 11 se resumen las características principales de los modelos entrenados: población y variable de estudio, filtrado de características (si aplica), número de características finales, modelo final obtenido, hiperparámetros y métricas de interés.

Resp	Muestras	Features Selec	Final Genes	Final Model	Hyperparam	MCC	BalAcc
uPCR	*global*	Rfe (77, 2)	71	k-Nearest Neighbors (knn)	K = 11	0.7785	0.7956
		Sbf (77)	76	SVMRadial	sigma = 0.0056 C = 1.4525	0.8408	0.8323
		PCA (10)	10	Linear Discriminant Analysis (lda)	NA	0.8225	0.8343
	early	NA (2)	2	SVMLineal	C = 0.2051	0.8849	0.8125
mPERR	*global*	Rfe (25, 2)	25	SVMLineal	C = 113.759	0.5738	0.7843
		PCA (10)	10	Boosted Classification Trees (ada)	iter = 157.8, maxdepth = 6.2 nu = 0.1914	0.5304	0.6927
	early	Rfe (12, 2)	12	Random Forest (rf)	mtry = 2.25	1.0000	0.8750

Tabla 11. Resumen de parámetros de interés de los modelos entrenados para la predicción de respuesta a MMF. Features Selec = método de reducción de características aplicado (número de genes de partida); MCC = Matthews Correlation Coefficient; BalAcc = Balance Accuracy.

Se ha aplicado "rfe" como métodos de reducción de características para todos los conjuntos, excepto para *early* uPCR que solo dispone de 2 genes. Solo se ha conseguido reducir el número de genes en el caso de uPCR *global*, de 77 a 71; lo cual indica que la mayoría de los DEGs seleccionados están influyendo para la predicción de respuesta a fármacos por el modelo propuesto. Se prueba el método de "sbf" para uPCR *global*, pero en esta ocasión solo elimina un gen. Ambas variables del conjunto *global* se evalúan con la selección de genes por PCA.

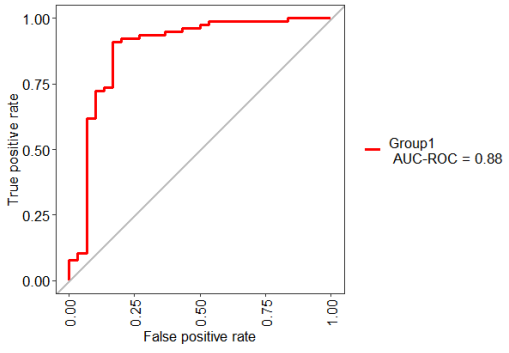
Entre las métricas obtenidas se encuentra el estadístico Matthews Correlation Coefficient (MCC), que está entre -1 y 1; siendo mejor el modelo cuanto más alto o próximo a 1. Se considera el mejor optimizador para clasificación binaria, ya que tiene en cuenta toda la matriz de confusión. También se obtiene el Balance Accuracy (balAcc), media entre la sensibilidad y especificidad del modelo. BalAcc es útil cuando existen

clases desbalanceadas, como es el caso del presente estudio, con una ratio próxima a 70/30 de respuesta / no respuesta para todos los grupos.

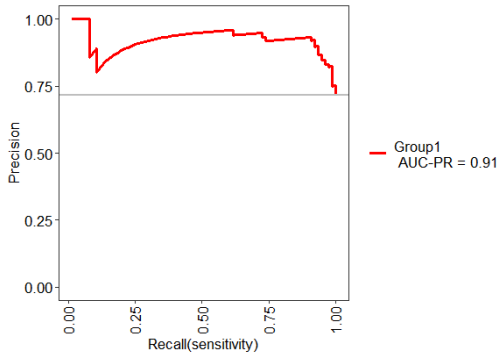
Se consideran bien optimizados todos los modelos con MCC y balacc a partir de 0.7; esto indica que todos los modelos han obtenido buenos resultados, excepto para *global* mPERR, el cual obtiene un rendimiento moderado. Si bien, para un menor número de genes a considerar, los 10 genes de *global* uPCR filtrada por PCA y los 2 genes de *early* uPCR presentan muy buena optimización, con MCC > 0.8 y BalAcc > 0.8. Aunque las métricas de los modelos *global* uPCR por "rfe" y "sbf" son adecuadas, el elevado número de genes que emplea no se considera óptimo para la aplicabilidad clínica. En cuanto a *global* mPERR, los 25 genes no obtienen buenas métricas y parecen ser los que peor ajustan las predicciones.

El resultado del modelo también almacena un objeto con la matriz de probabilidades por clase, útil para la evaluación del rendimiento de los modelos entrenados mediante representación gráfica del área bajo la curva para la exactitud (AUC-ROC) y Precision-Recall (AUC-PR).

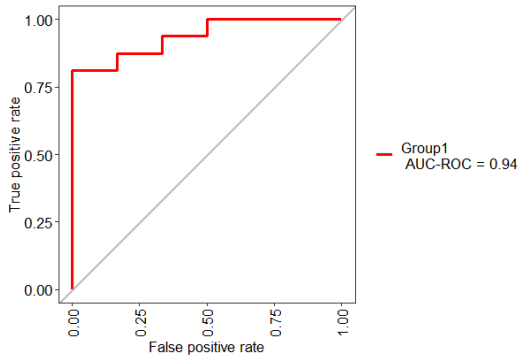
En la Figura 13 se muestra la comparativa de las curvas AUC-ROC y AUC-PR para los tres modelos seleccionados.



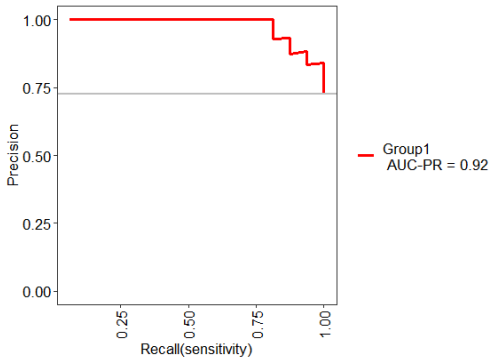
1.a) AUC-ROC – PCA global uPCR



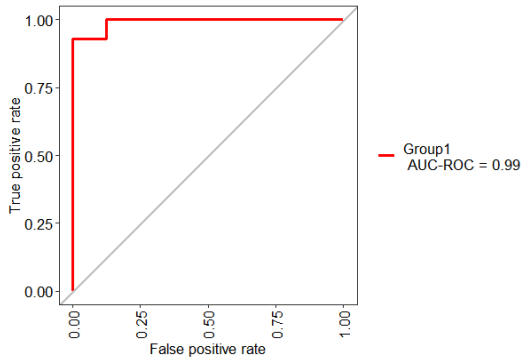
1.b) AUC-PR – PCA global uPCR



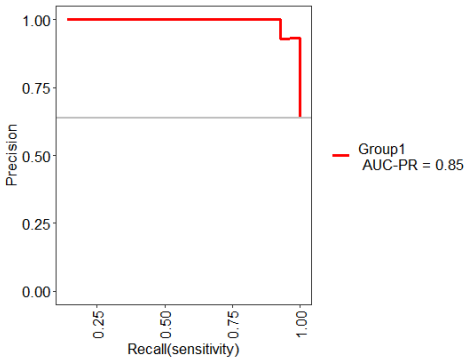
2.a) AUC-ROC – early uPCR



2.b) AUC-PR – early uPCR



3.a) AUC-ROC – early mPERR



3.b) AUC-PR – early mPERR

Figura 12. Gráficos de rendimiento del modelo a) AUC-ROC y b) AUC-PR para los tres modelos que han obtenido mejores resultados: 1) PCA *global* UPCR, 2) *early* uPCR, 3) *early* mPERR.

AUC-ROC muestra la sensitividad (tasa TP) en el eje 'y' frente a la especificidad (tasa FP) en el eje 'x', siendo mejor el modelo a medida que la curva se aproxima a la esquina superior izquierda (altos TP y bajos FP). AUC-PR presenta la precisión (aciertos de entre todos los positivos) frente a la sensitividad o recall (tasa TP). Por tanto, en este caso, el modelo es mejor cuanto más se aproxima la curva a la esquina superior derecha, lo que supone altos valores de ambas métricas. La curva AUC-PR es útil en el caso de clases desbalanceadas, para medir la capacidad de predicción de la clase minoritaria.

El valor de AUC-ROC(%) indica la capacidad predictiva entre respondedores (clase positiva) y no respondedores, por encima del 88% en el por caso. Mientras que, el valor AUC-PR por encima del 85% indica buena predicción en la asignación de NoResp (clase minoritaria).

4.6. Validación de los modelos seleccionados

Se preparan los datos para la obtención de las predicciones para el modelo *global* con los 10 DEGs obtenidos por PCA para uPCR y los dos modelos *early*, para uPCR y mPERR.

Finalmente, no es posible completar la validación tal como fue planteada, aplicando directamente los DEGs obtenidos en la cohorte inicial a los nuevos datos debido a que los modelos basados en expresión son altamente dependientes de las magnitudes, lo cual varía mucho entre estudios. Además, son sensibles a valores ausentes en los genes, algo muy común al analizar *microarrays* de diferentes plataformas.

En el artículo de Toro-Domínguez et al. (2022) [1] se propone un enfoque basado en el cálculo de puntuaciones o actividades de rutas desreguladas en muestras individuales, también realizado con *pathMED*, y que mejora los problemas de reproducibilidad. Queda fuera del ámbito del presente trabajo, pero se plantea como línea futura de investigación, dada la necesidad de validar los resultados de los modelos con datos independientes y para su aplicabilidad clínica.

5. Discusión

En este trabajo se ha estudiado la capacidad y utilidad clínica de los modelos ML basados en genes para la predicción de la respuesta al tratamiento con MMF en pacientes de LN, analizando retrospectivamente una cohorte longitudinal de pacientes respondedores y no respondedores al fármaco. La respuesta al fármaco fue establecida por uPCR, según el criterio EULAR/ACR [11], considerado *gold standard*; y por mPERR, según criterio modificado de BLISS-LN [12], como medida secundaria.

La dificultad en la predicción de la respuesta a fármacos radica en la propia heterogeneidad de la enfermedad y en el carácter impredecible de los periodos de brote/remisión, brotes que en poco tiempo pueden cursar con un empeoramiento significativo del daño orgánico y afectar la capacidad de respuesta al tratamiento. Una de las razones del fracaso en los ensayos clínicos de SLE es la consideración de los pacientes como grupos homogéneos [1], por lo que un enfoque basado en perfiles

moleculares puede ser la clave para el desarrollo de nuevas terapias, más precisas y personalizadas, y para la anticipación ante tratamientos previsiblemente no efectivos.

El estudio de significancia estadística de variables clínicas manifestó la relación de valores altos en los índices de actividad de la enfermedad, medidos en base a síntomas, con la no respuesta a MMF. En el caso de SLEDAI, se contabilizaron hasta 5 descriptores significativos entre grupos de respuesta; entre ellos, los más importantes, relacionados con la respuesta inmunológica para niveles bajos de complemento y aumento en la fusión de ADN (p-valor < 1.0 e^{-2}, OR > 6), mostrando un aumento de la probabilidad de respuesta positiva al fármaco cuando estos factores inmunológicos están presentes. En la serología se corroboró la asociación de bajos niveles de C3 y C4 con un mayor índice de respuesta a MMF. Y al revés, a mayor nivel de complemento, mayor actividad, menor respuesta al fármaco. Aunque los factores inmunológicos fueron significativos para ambas respuestas, SLEDAI como medida global solo obtuvo significancia para uPCR. Los resultados concuerdan con los obtenidos en el estudio original [2].

De forma lógica en pacientes con progresión a LN, los factores renales de la actividad se encontraron altamente asociados a la respuesta al fármaco. De hecho, la medida de proteínas en orina o proteinuria establece el criterio para la respuesta por uPCR y forma parte de la respuesta a mPERR. Este hecho está relacionado con la obtención de una relación similar entre grupos de respuesta en ambas variables, en torno al 70% de respuesta positiva, y donde la diferencia la marca la componente eGFR de mPERR.

La presencia de co-tratamientos y la dosis de fármacos influyeron en la respuesta y obtuvieron significancia cuando el fármaco era de uso mayoritario en la cohorte, como es el caso de los tratamientos estándar con prednisona y CQ. La influencia de la dosis del fármaco en la respuesta está condicionada por la práctica rutinaria, observándose dosis más altas de MMF y ácido acetil salicílico en pacientes no respondedores, normalmente debido a la decisión médica de aumentar la dosis ante la ineficacia con dosis más bajas [2]. En el caso de la prednisona, se obtiene relación con la respuesta a MMF, tanto para uPCR como para mPERR, en los casos en que el paciente se encuentra en tratamiento activo e independientemente de la dosis; hecho de especial relevancia y por lo que se considera como variable de interés en el ajuste por análisis DEG para la

selección de características de los modelos. Otras variables consideradas fueron la dosis máxima de MMF, la actividad SLEDAI y el sexo. No se consideró la edad, al no obtener significancia en la diferenciación de grupos de respuesta; ni la raza, por presentar un sesgo en la población no respondedora, mayoritaria de raza negra.

Se plantearon dos líneas de estudio de la predicción de la respuesta a MMF, con distinta utilidad clínica; a nivel global, con todas las muestras de la cohorte, para extraer los DEGs fundamentales entre grupos de respuesta a lo largo del tiempo; y en un subconjunto de muestras tomadas hasta 6 meses después de la biopsia, para obtener DEGs capaces de predecir la respuesta en fases iniciales del tratamiento o, incluso, antes del diagnóstico.

El objetivo era obtener modelos ML aplicables a la práctica clínica, teniendo no solo en cuenta la precisión y exactitud de las predicciones, sino también el coste económico de la obtención de los DEGs y el computacional para su evaluación. Por ello, fue necesario plantear una estrategia de selección de características para reducir la dimensionalidad de los datos transcriptómicos de partida, 54715 *probes* de *microarrays*.

Se propuso un enfoque basado en expresión diferencial de genes, con el que se obtuvieron 77 genes significativos con uPCR y 25 con mPERR, para periodo global; 2 con uPCR y 12 con mPERR, para periodo cercano a la biopsia o *early*. La significancia de los DEGs se estableció para un p-valor ajustado < 0.05 y ajuste BH para FDR. Entre los 10 genes más significativos destacaron: TUBB2A, reprimido en respondedores en el conjunto global para ambas variables y para mPERR en periodo inicial; pudiendo considerarse un posible biomarcador de respuesta a MMF constitutivo a lo largo del tiempo; y MCTP1 sobre-expresado en respondedores del periodo inicial para las dos variables, como posible biomarcador en fase temprana de diagnóstico/tratamiento.

77 y 25 genes se siguen considerando un número elevado para la aplicabilidad clínica, por lo que se realizó un nuevo filtrado de los conjuntos para periodo global en base a PCA. Se extrajo el top10 de genes de la componente principal, dado que PC1 explicaba más del 80% de la variabilidad en la expresión génica para ambas respuestas.

Los 6 conjuntos de DEGs fueron utilizados como entrada para la construcción de modelos ML con la herramienta *pathMED*. Se aplicó *rfe* para los dos modelos globales completos y para el modelo *early* mPERR, pero solo se consiguió reducir los genes de uPCR global de 77 a 71; el resto permanecieron igual, lo que indica la influencia de todo el conjunto de genes en el modelo seleccionado.

Todos los modelos obtuvieron buenos resultados (mcc > 0.7 y balacc > 0.7), excepto para mPERR global, que obtuvo un rendimiento moderado (mcc > 0.5, balacc > 0.7). Tres se consideraron los más optimizados teniendo en cuenta el número de genes empleado, todos aplicaron aprendizaje supervisado: modelo lda con los 10 genes para uPCR global filtrada por PCA, modelo svmLineal con los 2 genes de early uPCR, y modelo RandomForest con los 12 de early mPERR (MCC > 0.8 y BalAcc > 0.8).

Se verificó que, en general, los modelos *early* obtuvieron mejores estadísticos de predicción que los globales, especialmente *early* uPCR con 2 únicos genes (AUC-ROC = 0.94 y AUC-PR = 0.91). Esto está en consonancia con una mayor homogeneidad de las muestras a medida que se acota el periodo a analizar, como ya se observó en la *clusterización* completa de las muestras por grupo de respuesta en los *heatmaps* para *early*; así como en la agrupación de muestras según la respuesta a MMF en el PCA.

El buen desempeño de los modelos early es prometedor desde el punto de vista médico, ya que la posibilidad de predecir la respuesta al fármaco en fases iniciales de la enfermedad puede suponer una mejora en las decisiones terapéuticas para la mejora de los pacientes, especialmente en el caso de no respondedores, pudiendo frenar la progresión del daño orgánico.

Sin embargo, el planteamiento inicial de reducción de características en base a DEGs requería la validación de los modelos obtenidos en una cohorte independiente, para su reproducibilidad en entorno clínico. La validación independiente de estos métodos permite evaluar la influencia de factores limitantes para la comparabilidad entre estudios, como el tipo de plataforma de secuenciación utilizada, que afecta al cálculo del tamaño muestral y las intensidades obtenidas en el experimento [27]. Estos factores han dificultado hasta ahora el descubrimiento de biomarcadores genéticos en Lupus por problemas de sensibilidad, especificidad y capacidad predictiva para su escalado a la

clínica. Finalmente, no fue posible ejecutar la validación, precisamente por estos problemas de reproducibilidad de los DEGs, cuya solución quedaba fuera del ámbito de estudio de este trabajo y no hubo tiempo de abordar. Por ello, y dados los buenos resultados de los modelos obtenidos, se plantea como línea de investigación futura y necesaria para la aplicabilidad clínica, la aplicación de un nuevo enfoque que permita hacer más comparables los modelos basados en datos de expresión.

6. Conclusiones

La combinación de unos pocos biomarcadores genéticos basados en expresión diferencial permite la obtención de modelos precisos de ML para la predicción de la respuesta al fármaco MMF en pacientes de LN. Se ha observado una mayor optimización de los modelos cuando se utiliza uPCR como medida de la respuesta, especialmente si se estudian periodos acotados para la toma de muestras, mejorando su homogeneidad. Sin embargo, para poder escalar estos modelos a la práctica clínica, se necesitan superar los problemas de reproducibilidad intrínsecos de los datos de expresión, que permitan obtener biomarcadores robustos, independientes de la tecnología utilizada para su obtención y del conjunto de muestras de estudio.

7. Bibliografía

[1] 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

[2] Toro-Domínguez, D., López-Domínguez, R., Villatoro-García, J. A., Marañón, C., Goldman, D., Petri, M., Carmona-Sáez, P., & Alarcón-Riquelme, M. (2024). Immune and molecular landscape behind non-response to Mycophenolate Mofetil and Azathioprine in lupus nephritis therapy. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-3783877/v1

[3] Zollars, E., Courtney, S. M., Wolf, B. J., Allaire, N., Ranger, A., Hardiman, G., & Petri, M. (2016). Clinical Application of a Modular Genomics Technique in Systemic Lupus Erythematosus: Progress towards Precision Medicine. International Journal Of Genomics, 2016, 1-7. https://doi.org/10.1155/2016/7862962

[4] Koo, M., & Lu, M. (2023). Performance of a New Instrument for the Measurement of Systemic Lupus Erythematosus Disease Activity: The SLE-DAS. Medicina, 59(12), 2097. https://doi.org/10.3390/medicina59122097

[5] Arora, S., Isenberg, D. A., & Castrejon, I. (2020). Measures of Adult Systemic Lupus Erythematosus: Disease Activity and Damage. Arthritis Care & Research, 72(S10), 27-46. https://doi.org/10.1002/acr.24221

[6] Toro-Domínguez, D., Martorell-Marugán, J., Goldman, D., Petri, M., Carmona-Sáez, P., & Alarcón-Riquelme, M. E. (2018). Stratification of Systemic Lupus Erythematosus Patients Into Three Groups of Disease Activity Progression According to Longitudinal Gene Expression. Arthritis & Rheumatology, 70(12), 2025-2035. https://doi.org/10.1002/art.40653

[7] Habehh, H., & Gohel, S. (2021). Machine Learning in Healthcare. Current Genomics, 22(4), 291-300. https://doi.org/10.2174/1389202922666210705124359

[8] RPubs – Machine Learning con R y caret. (s. f.). https://rpubs.com/joaquin_ar/383283

[9] Gacto, M. J., Soto-Hidalgo, J. M., Alcala-Fdez, J., & Alcala, R. (2019). Experimental Study on 164 Algorithms Available in Software Tools for Solving Standard Non-Linear Regression Problems. IEEE Access, 7, 108916-108939. https://doi.org/10.1109/access.2019.2933261

[10] Jordimartorell. (s. f.). GitHub - jordimartorell/pathMED: Development version of pathMED package. GitHub. https://github.com/jordimartorell/pathMED

[11] Wang, H., Gao, Y., Ma, Y., Cai, F., Huang, X., Lan, L., Ren, P., Wang, Y., Chen, J., & Han, F. (2021). Performance of the 2019 EULAR/ACR systemic lupus erythematosus classification criteria in a cohort of patients with biopsy-confirmed lupus nephritis. Lupus Science & Medicine, 8(1), e000458. https://doi.org/10.1136/lupus-2020-000458

[12] Blenkinsopp, S. C., Fu, Q., Green, Y., Madan, A., Juliao, P., Goldman, D. W., Roth, D. A., & Petri, M. A. (2022). Renal response at 2 years post biopsy to predict long-term renal survival in lupus nephritis: a retrospective analysis of the Hopkins Lupus Cohort. Lupus Science & Medicine, 9(1), e000598. https://doi.org/10.1136/lupus-2021-000598

[13] Wither, J. E., Prokopec, S. D., Noamani, B., Chang, N., Bonilla, D., Touma, Z., Avila-Casado, C., Reich, H. N., Scholey, J., Fortin, P. R., Boutros, P. C., & Landolt-Marticorena, C. (2018). Identification of a neutrophil-related gene expression signature that is enriched in adult systemic lupus erythematosus patients with active nephritis: Clinical/pathologic associations and etiologic mechanisms. PLoS ONE, 13(5), e0196117. https://doi.org/10.1371/journal.pone.0196117

[14] Dtordom. (s. f.). GitHub - dtordom/LNtherapy: All R codes to reproduce the entire analysis of characterization the non-response to drugs for LN. GitHub. https://github.com/dtordom/LNtherapy/tree/main

[15] RCODER. (2024, 7 septiembre). Statistics with R. https://r-coder.com/r-statistics/

[16] Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research, 31(4), 15e-115. https://doi.org/10.1093/nar/gng015

[17] 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

[18] Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics, 21(16), 3439-3440. https://doi.org/10.1093/bioinformatics/bti525

[19] GENyO-BioInformatics. (s. f.). MyPROSLE/Letter/utils.R at main · GENyO-BioInformatics/MyPROSLE. GitHub. https://github.com/GENyO-BioInformatics/MyPROSLE/blob/main/Letter/utils.R

[20] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47. https://doi.org/10.1093/nar/gkv007

[21] RPubs - volcano plot. (s. f.). https://www.rpubs.com/Knight95/volcano

[22] Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32(18), 2847-2849. https://doi.org/10.1093/bioinformatics/btw313

[23] Kassambara, A., & Mundt, F. (2016). factoextra: Extract and Visualize the Results of Multivariate Data Analyses [Conjunto de datos]. https://doi.org/10.32614/cran.package.factoextra

[24] Kuhn, M. (2007). caret: Classification and Regression Training [Conjunto de datos]. https://doi.org/10.32614/cran.package.caret

[25] Kuhn, M. (2019, 27 marzo). The caret Package. https://topepo.github.io/caret/

[26] Banchereau, R., Hong, S., Cantarel, B., Baldwin, N., Baisch, J., Edens, M., Cepika, A., Acs, P., Turner, J., Anguiano, E., Vinod, P., Khan, S., Obermoser, G., Blankenship, D., Wakeland, E., Nassi, L., Gotte, A., Punaro, M., Liu, Y., . . . Pascual, V. (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

[27] Yu, H., Nagafuchi, Y., & Fujio, K. (2021). Clinical and Immunological Biomarkers for Systemic Lupus Erythematosus. Biomolecules, 11(7), 928.https://doi.org/10.3390/biom11070928

8. Anexos

Material Suplementario (enlace): TFM_eme

9. Declaración obligatoria del uso de herramientas de IA

En el presente trabajo se ha utilizado ChatGPT como herramienta soporte para la resolución de dudas en el uso de nuevas funciones y codificación en lenguaje R, siempre complementado con documentación oficial y tutoriales online. También se ha utilizado DeepL para traducción de fragmentos de texto en inglés, y Scribbr como generador de citas bibliográficas en formato APA 7. La información aportada por la IA fue siempre revisada, contrastada y adaptada a las necesidades del trabajo.