UNIVERSIDAD DE GRANADA

PROGRAMA DE DOCTORADO EN BIOMEDICINA

Centro Pfizer - Universidad de Granada - Junta de Andalucía de Genómica e Investigación Oncológica (GENYO)

Departamento de Estadística e Investigación Operativa – Universidad de Granada

TESIS DOCTORAL

INFERENCIA DE PATRONES DE REGULACIÓN A PARTIR DE DATOS ÓMICOS

Raúl López Domínguez

Dirigida por

Pedro María Carmona Sáez

Marta Eugenia Alarcón Riquelme

Granada, 2023

Editor: Universidad de Granada. Tesis Doctorales
Autor: Raúl López Domínguez
ISBN: 978-84-1117-993-5
URI: https://hdl.handle.net/10481/84440

“Cuando cae la nieve y soplan blancos vientos,
el lobo solitario muere, pero la manada sobrevive”

(Eddard Stark – Juego de Tronos)

AGRADECIMIENTOS

¡Pues ya está! Parecía que nunca iba a acabar y finalmente hemos llegado al último paso. Digo hemos porque ha sido un viaje en compañía, por lo que, como es de bien nacidos ser agradecidos, me toca agradecer a todas esas personas que me han apoyado, animado, ayudado, soportado y, por qué no decirlo, emborrachado.

En primer lugar, me gustaría agradecer a mis directores de tesis por darme la oportunidad de aprender de ellos, orientarme, darme un toque de atención cuando fue necesario y ayudarme en los momentos en los que no veía el final de todo esto. Principalmente me gustaría agradecer a Pedro por tu compromiso conmigo desde que llegué a Genyo para hacer el TFM, cuando aún no tenía apenas barba, y por darme la ocasión de pertenecer todos estos años al estupendo grupo de trabajo que has formado.

Y qué decir de mis compañeros de sala, a los que tengo el gran honor de llamar amigos. No tengo palabras para agradeceros todo lo que habéis cambiado mi vida. Yo no sería el mismo sin haberos conocido a todos y cada uno de vosotros. Los viernes, las tonterías, las canciones, los detalles, los desayunos, las videollamadas… Me lo llevo todo y creo que me quedo corto al decir que somos una pequeña (cada vez más grande) familia. Me habéis enseñado a disfrutar de la vida, a reír y hasta a ser mejor persona. Desde el primer momento en el que me senté en una silla de comedor porque no había sillones hasta ahora mismo me he sentido como uno más, y lo que nos queda aún. ¡Sois muy grandes! Gracias a Jordi por su humor negro y sus paellas con canciones. Gracias a Dani por ser un director de tesis más y por darme la oportunidad de convivir contigo. Gracias a Adrián por ser tú, con tus múltiples personalidades como consecuencias del *genecodis syndrome* y que el chicleteo no decaiga. Gracias a Juanan por siempre estar ahí para ayudarnos a todos y tomarte las cosas de esa forma tan especial. Gracias a Alba por quedarte cuando eras Penny y, aunque a veces no te consideres miembro, eres unidad interestelar. Gracias a Iván por las escapadas de fin de semana y por siempre dejar un hueco en Genyo. Por último, y aunque he coincidido menos con vosotros, gracias a Jose, Samu y Marina porque habéis llenado nuestra sala de más risas y tonterías (and you know what I mean).

Pero no todo se queda en la sala de Bioinformática. Durante este trayecto he conocido a gente maravillosa que en muchas ocasiones me ha ayudado, animado, hecho reír, disfrutar… Quisiera

agradecer al gran equipo de bioinfoGRX, gracias al cual he ido obteniendo ganas de divulgar y entretener con ciencia. Y a nivel personal, me quedaría sin páginas para agradecer a la gente; Silvia, Inés, Marisa, Sergio, Jorge, Gonzalo… Gracias a todos vosotros, en representación del resto.

También quiero agradecer a una persona muy especial que apareció en mi vida hace relativamente poco. Muchas gracias Yera por animarme y aconsejarme cuando no veía la luz al final del túnel, por ayudarme con mis miedos y por dedicarte a hacerme muy feliz.

Para terminar, no me olvido de mi familia. Muchas gracias a mis padres y a mi hermano por apoyarme y preocuparse siempre por mi, sin vosotros no habría llegado aquí jamás.

TABLA DE CONTENIDO

RESUMEN

La necesidad de conocer cómo interaccionan las biomoléculas entre sí y como afectan dichas interacciones al funcionamiento global de un tejido o célula es uno de los aspectos clave que, en los últimos años, están ganando cada vez más importancia. Uno de los elementos más importantes en el contexto de las redes de regulación son los factores de transcripción (TFs). Estas proteínas son esenciales en el buen funcionamiento de la regulación de la transcripción ya que se unen a ciertas regiones del genoma, principalmente asociadas a regiones codificantes, y actúan activando o inhibiendo la transcripción de un gen determinado.

A lo largo de esta tesis doctoral, los trabajos realizados se han centrado en el análisis de datos ómicos, principalmente transcriptómica, en el contexto de las enfermedades autoinmunes y, de forma más específica, en lupus eritematoso sistémico. Esta enfermedad se caracteriza por ser una patología heterogénea en la que tiene lugar una respuesta inmune contra células y órganos del propio individuo. Sobre lupus y datos ómicos se han desarrollado múltiples estudios tanto para identificar biomarcadores, estratificar pacientes o buscar tratamientos potenciales. En el contexto de los TFs, se sabe que muchas de las mutaciones asociadas a la enfermedad se han localizado en pacientes de lupus se encuentran en regiones reguladoras, como los sitios de unión de los TFs. A pesar de este conocimiento, no se aporta información acerca de cómo afectan estas mutaciones a la actividad de dichos TFs.

Por este motivo, en esta tesis se han aplicado una serie de técnicas que permiten inferir la actividad de los TFs en base a la expresión de los genes que están regulando (también conocidos como genes diana). Para ello era necesario conocer qué genes son regulados por cada TF, cuya información se localiza en múltiples bases de datos y se obtienen a partir de una gran cantidad de artículos publicados. El conjunto de genes regulados por cada TFs (también llamado regulones) se obtuvo utilizando la base de datos desarrollada por DoRothEA que incluye interacciones de fuentes muy diversas y que se clasifican en base a la credibilidad. Para inferir las actividades de los TFs de cada muestra utilizando los niveles de expresión de sus genes diana se utilizó el método desarrollado por VIPER (*Virtual Inference of Protein-activity by Enriched Regulon).*

Sin embargo, como trabajo previo a la inferencia de los TFs se realizó una recopilación de datos ómicos públicos de tipo caso-control disponibles en el repositorio público más conocido: NCBI-GEO. Este trabajo permitió el desarrollo de ADEx, una herramienta online con datos ómicos procesados atendiendo a un mismo método estandarizado y con varios análisis de dichos datos: expresión diferencial, análisis de rutas, inferencia de redes o meta-análisis.

Posteriormente, elegimos uno de los estudios incluidos en ADEx con datos transcripcionales de lupus así como los datos de otra cohorte a la que tuvimos acceso gracias a nuestra colaboración con la doctora Michelle Petri, de la Universidad Johns Hopkins. Estos dos conjuntos de datos transcripcionales se utilizaron para inferir las actividades de los TFs en los pacientes de lupus y controles sanos. Las actividades de todos los individuos de lupus se utilizaron para estratificar pacientes y, en ambos estudios se localizaron dos grupos muy diferenciados, que además presentaban diferencias clínicas con respecto a las proporciones de algunos tipos celulares como neutrófilos y linfocitos. Finalmente, mediante análisis de actividad diferencial, localizamos una firma robusta de 14 TFs que se encontraban diferencialmente activados en lupus con respecto a controles.

Dada la importancia que se observó en cuanto a los tipos celulares, se decidió aplicar un análisis similar al anterior pero en datos transcripcionales de célula única, con el fin de analizar en profundidad los TFs en los que existen diferencias a nivel de tipo celular. Para ello, se utilizó un conjunto de datos público alojado en NCBI GEO y se aplicó un protocolo de procesamiento en línea con lo más estándar en el campo. Además, se añadió una capa adicional, la inferencia de actividad de rutas de señalización. Aunque este trabajo se encuentra incompleto, se han hallado resultados relevantes, principalmente relativos a tipos celulares en los que hay patrones de actividad diferentes.

ABREVIATURAS

ADN o DNA: Ácido desoxirribonucleico (*deoxyribonucleic acid*).

ADNc: cadena complementaria de ADN.

ARN o RNA: Ácido ribonucleico (*ribonucleic acid*).

PCR: Reacción en cadena de la polimerasa (*Polymerase Chain Reaction*).

SNP: Polimorfismo de Nucleótido Único (*Single Nucleotide Polimorfism*)

NGS: Secuenciación de última generación (*Next-Generation Sequencing*)

GWAS: Estudio de asociación de genoma completo (*Genome-wide Association Study*).

RNA-Seq: Secuenciación de ARN mediante técnicas de secuenciación masiva.

CpG: Dinucleótidos de citosina y guanina que pueden ser metilados.

scRNA-Seq: Secuenciación de ARN mediante técnicas de secuenciación masiva aplicados a células únicas.

ChIP-Seq: Secuenciación de ADN mediante técnicas de secuenciación masiva aplicadas a muestras tratadas para comprobar inmunoprecipitación.

TF/TFs: Factor de transcripción o factores de transcripción (*Transcription factor*).

SEA: Análisis de enriquecimiento simple (*Singular Enrichment Analysis*),

GSEA: Análisis de enriquecimiento con un conjunto de genes (*Gene Set Enrichment Analysis*)

MEA: Análisis de enriquecimiento modular (*Modular Enrichment Analysis*)

IFN: Interferón

ISG: Genes estimulados por interferón

NK: Células asesinas naturales (*Natural Killer*).

LES o SLE: Lupus Eritematoso Sistémico (*Sistemic Lupus Erythematosus*)

AR: Artritis reumatoide

SjS: Síndrome de Sjogren

SSc: Esclerosis sistémica

T1D: Diabetes tipo I

SLEDAI: Índice de Actividad de Enfermedad de LES (*Systemic Lupus Erythematosus Disease Activity Index*).

ADEx: Autoimmune Disease Explorer

PBMC: Célula Mononuclear de Sangre Periférica (*Peripheral Blood Mononuclear Cell*).

NLR: Ratio de neutrófilos y linfocitos *(Neutrophil-to-Lymphocyte Ratio*)

PCA: Análisis de Componentes Principales (*Principal Component Analysis*)

UMAP: *Uniform Manifold Approximation and Projection*

1 INTRODUCCIÓN

1.1 EXPANSIÓN DE LAS CIENCIAS ÓMICAS Y LA BIOINFORMÁTICA

Las ciencias ómicas son un conjunto de disciplinas científicas que se dedican al estudio global de los sistemas biológicos a nivel molecular, como el ADN, el ARN, las proteínas y otros componentes moleculares que forman parte de la célula y son esenciales para el funcionamiento de los organismos vivos. Dentro de las ciencias ómicas se encuentran la genómica, que se centra en el estudio del ADN y el genoma completo de un organismo; la transcriptómica, que se dedica al estudio del ARN; la proteómica, que se enfoca en el estudio de las proteínas; la metagenómica, una disciplina interdisciplinaria que se ocupa del estudio de los genomas completos de comunidades microbianas; y la epigenética, que se dedica al estudio de cómo las modificaciones en la información genética no codificante afectan la expresión génica. Juntas, estas disciplinas nos ayudan a entender cómo los diferentes componentes moleculares interactúan para regular los procesos biológicos y contribuyen al desarrollo de nuevas terapias y tratamientos médicos[1].

Las principales dificultades que presentan las ciencias ómicas son la cantidad masiva de datos que se generan y la complejidad de los mismos. Los datos genéticos, proteicos y metabolómicos son extremadamente complejos y voluminosos, y requieren un gran poder de procesamiento y almacenamiento para su análisis. Por este motivo el auge y expansión que han tenido las ciencias ómicas en los últimos años ha ido de la mano del desarrollo de herramientas y técnicas bioinformáticas que sirvan de apoyo para analizar estos datos. La bioinformática es la disciplina que aplica métodos informáticos, matemáticos y estadísticos para analizar y comprender los datos producidos por las ciencias ómicas y otras disciplinas de la biología.

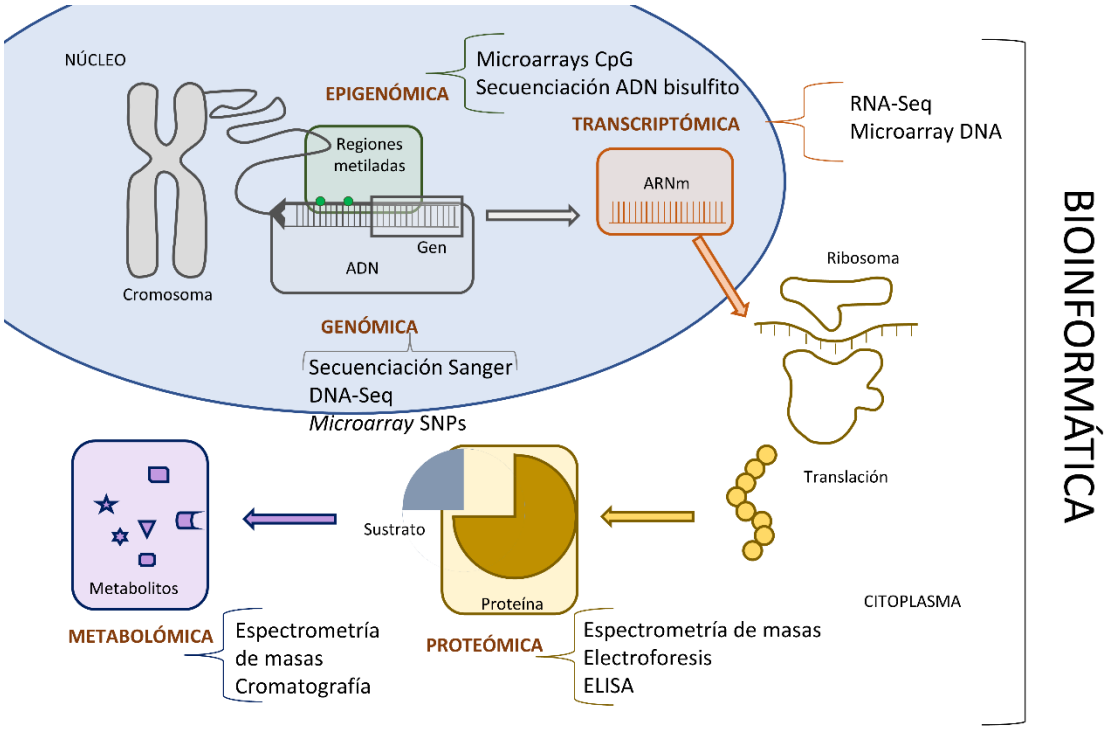

Figura 1. Esquema del dogma central de la biología molecular. En este esquema se muestran los eventos principales que ocurren en una célula eucariota, desde la molécula de ADN, con sus modificaciones epigenéticas, la transcripción, la translación o traducción y la interacción de proteínas con sustratos para generar metabolitos. En el mismo, se indican las técnicas ómicas más utilizadas para generar datos de cada una de las ciencias ómicas, todas ellas analizadas bajo el amparo de la bioinformática.

Las ómicas que se han descrito son las más importantes y relevantes en el campo biomédico y biotecnológico ya que giran en torno al dogma central de la biología molecular, que consiste en que las secuencias de ADN se transcriben a ARN, el cual se traduce a proteínas en los ribosomas. Estas proteínas intervienen en reacciones actuando como enzimas generando una serie de metabolitos necesarios para el organismo (véase Figura 1).

A continuación, se van a describir las ciencias ómicas más importantes, la aparición de la tecnología de célula única y la importancia de la bioinformática en el análisis de estos datos, así como los métodos más importantes de aplicación que tienen relevancia para esta tesis.

1.1.1 ADN, GENÓMICA Y SECUENCIACIÓN.

El ácido desoxirribonucleico, conocido como ADN, es una molécula esencial para la vida, al menos para la mayoría de los organismos conocidos y más estudiados, ya que la afirmación anterior no es cierta para todas las formas de vida. Para la totalidad de la tesis hablaremos en todo momento de mecanismos que ocurren, salvo que se indique lo contrario, en el ser humano. Descubierta en 1869 por el médico alemán Friedrich Miescher[2], esta molécula contiene toda la información genética necesaria para formar, mantener y desarrollar un ser vivo, así como para trasladarla a la descendencia a través de la herencia. La molécula está compuesta por dos cadenas de nucleótidos que forman una doble hélice, las cuales están formadas por la polimerización de monómeros, denominados nucleótidos, que se componen de una base nitrogenada, un grupo fosfato y un azúcar, la desoxirribosa. Las bases nitrogenadas del ADN son adenina, guanina, timina y citosina, las cuales quedan enfrentadas en el interior de la doble hélice mostrando complementariedad (A-T y C-G). La estructura de esta doble hélice es esencial para la estabilidad y la integridad de la información genética, así como para la replicación y la transcripción de los genes[3,4]. El ADN se encuentra en el interior de las células, tanto en el núcleo como en orgánulos como los cloroplastos o las mitocondrias.

La genómica es una disciplina científica que se encarga del estudio de la molécula de ADN y su secuencia de pares de bases en los seres vivos. Aunque puede parecer una tarea relativamente sencilla, el número de pares de bases en el ADN de la mayoría de los organismos es enorme, lo que hace difícil la obtención de la secuencia de forma manual. Por ejemplo, el genoma humano está formado por alrededor de 3 mil millones de pares de bases[5].

Los genes son las unidades funcionales del ADN y contienen la información necesaria para fabricar diferentes tipos de proteínas que son esenciales para el desarrollo y el funcionamiento del organismo. Se estima que en los humanos hay alrededor de 20.000 genes, que codifican un total de 60.000 proteínas. Sin embargo, esta cantidad de genes solo ocupa el 1-2% de las pares de bases del genoma completo[6]. Antes del desarrollo de las técnicas de secuenciación modernas, solo se solía secuenciar las regiones consideradas de importancia vital, las regiones codificantes, y se conocía como genoma al conjunto de genes de un organismo. Sin embargo, con la disponibilidad de la secuenciación de la secuencia completa de ADN es cada vez más común estudiar también las regiones no codificantes.

Es indudable que uno de los factores fundamentales relacionados con el desarrollo de una enfermedad se encuentra en la información genética de un organismo. Por este motivo, el estudio del genoma resulta esencial ya que conocer la secuencia, parcial o completa, de un individuo nos permite buscar variantes que pueden influir en el desarrollo de dicha enfermedad.

Actualmente, existen varias formas de identificar estas variantes genéticas. Una de ellas es mediante *microarrays* de ADN. Un *microarray* consiste en una placa pequeña y plana con millones de pozos pequeños en la que se colocan pequeñas moléculas de ADN de secuencias conocidas, llamadas sondas[7]. El funcionamiento de los *microarrays*, en general, se basa en la hibridación de estas sondas con las muestras que se quieren analizar. La hibridación es el proceso por el cual dos secuencias de ADN se unen por complementariedad de forma específica. Para detectar que una secuencia ha hibridado con una sonda se utilizan fluorocromos para cuantificar la intensidad lumínica. En caso particular de los microarrays planteados para identificar variantes denominadas SNPs (*Single Nucleotide Polimorphisms*), las sondas están diseñadas para hibridar con SNPs específicos y, por tanto, conocidos.

Para utilizar un microarray de ADN de SNPs, se toma una muestra de tejido o células y se extrae el ADN, luego se fragmenta y se añade al *microarray*. Las sondas que se unen a fragmentos de ADN específicos se iluminan cuando se exponen a una fuente de luz, y la intensidad de la luz que se emite se correlaciona con la cantidad de ADN unido, permitiendo identificar y analizar los SNPs presentes en la muestra. Las principales limitaciones de esta técnica son; la imposibilidad de buscar variantes que no sean SNPs, y que estas variantes deben ser conocidas, o, dicho de otro modo, no es posible buscar variantes nuevas o poco conocidas.

El resto de técnicas utilizadas para detectar variantes en el genoma se basan en la secuenciación de las cadenas de ADN, es decir, conocer la secuencia parcial o completa de un organismo o de una muestra. Existen varios tipos de secuenciación, pero nosotros vamos a destacar dos de ellos.

La técnica de secuenciación más tradicional es la conocida como secuenciación de Sanger, cuyo origen data del año 1977 y que tuvo como principal impulsor a Frederick Sanger[8]. A día de hoy, la secuenciación de Sanger se utiliza para secuenciar pequeños fragmentos de ADN o validar resultados obtenidos por secuenciación *next-generation* (NGS). El proceso de secuenciación de Sanger comienza con la amplificación del fragmento de ADN deseado mediante la técnica de la reacción en cadena de la polimerasa (PCR). Luego, se utilizan cuatro tipos diferentes de cebadores de secuencia fluorescentes, cada uno marcado con un color diferente, para iniciar la

síntesis de cadenas complementarias al fragmento de ADN amplificado. Los cebadores contienen una base diferente en su extremo 3' y cada uno se incorpora a la cadena de ADN solo cuando se encuentra su base complementaria en el fragmento de ADN. Una vez que se han sintetizado las cadenas complementarias, se utiliza una enzima conocida como terminador de cadena, que se incorpora solo en una posición específica en la cadena de ADN y detiene la síntesis de la cadena. Esto produce fragmentos de ADN de longitud variable, cada uno con un cebador de secuencia fluorescente marcado en su extremo 3'. Los fragmentos de ADN se separan por longitud mediante electroforesis en un gel de poliacrilamida, y luego se visualizan mediante un escáner de geles de fluorescencia. El escáner lee los colores de los cebadores fluorescentes y los convierte en una secuencia de bases, que se alinea para producir la secuencia completa del fragmento de ADN original. La ventaja de la secuenciación de Sanger es su alta precisión y especificidad, ya que se utilizan cebadores fluorescentes y se pueden detectar errores en la secuencia. Sin embargo, tiene la limitación de que solo se pueden secuenciar fragmentos de ADN pequeños y requiere grandes cantidades de material genético.

La técnica de secuenciación más utilizada en los últimos años es la tecnología NGS. Al contrario que las la secuenciación de Sanger, este tipo de tecnología sí que permite generar secuencias del genoma completo[9]. Esto se logra mediante la fragmentación del ADN a secuenciar, la amplificación de estos fragmentos mediante la reacción en cadena de la polimerasa (PCR) y el uso de una plataforma de secuenciación para leer las secuencias de los fragmentos amplificados.

La secuenciación NGS comienza con la fragmentación del ADN a secuenciar mediante la utilización de enzimas de restricción. Estos fragmentos son entonces ligados a adaptadores, que son moléculas de ADN sintético que contienen secuencias específicas que permiten su unión a los fragmentos de ADN. Los fragmentos de ADN ligados a los adaptadores son luego amplificados mediante PCR.

Una vez amplificados, los fragmentos de ADN se colocan en una plataforma de secuenciación, como Illumina, PacBio o Nanopore, para leer las secuencias de los fragmentos amplificados. Cada plataforma tiene un método diferente para leer las secuencias, pero en general, implica la adición de químicos que reaccionan con las bases del ADN y producen una señal fluorescente que puede ser detectada por una cámara. Las secuencias de todos los fragmentos se almacenan en ficheros de un formato especial llamado *FASTQ*, que, tras un control de calidad, pueden ser

alineadas o mapeadas contra el genoma de referencia del organismo al que pertenezca la muestra de ADN secuenciada.

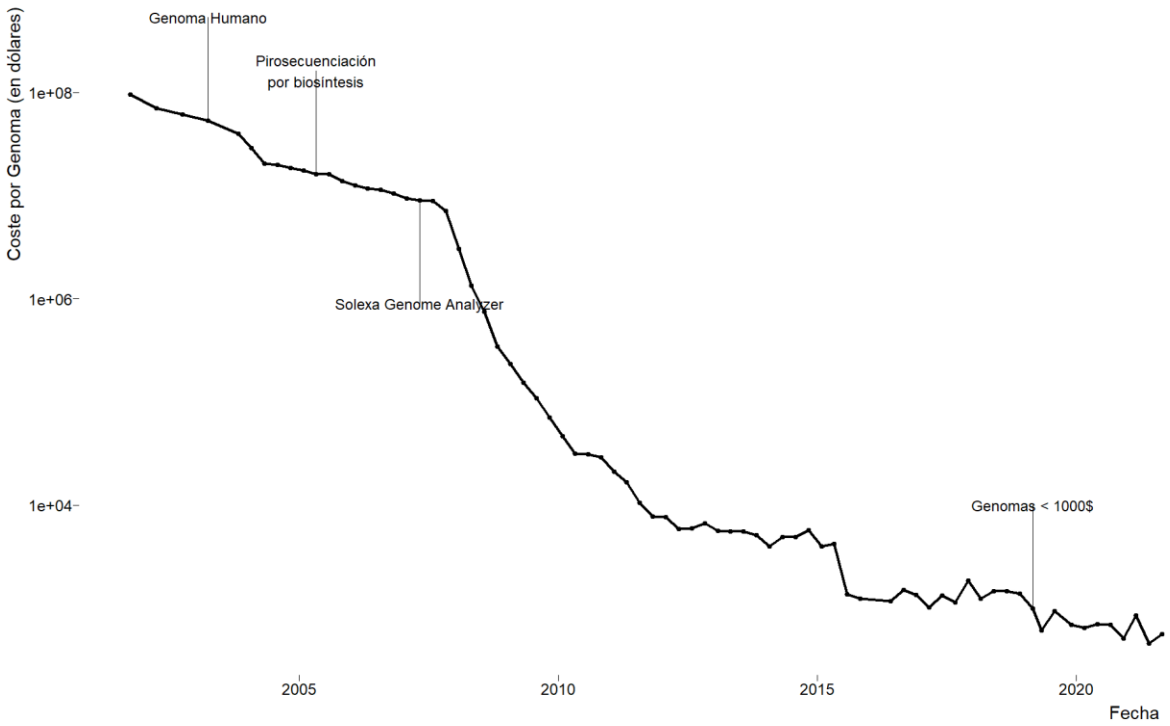

Figura 2. Abaratamiento del precio de secuenciación de un genoma completo. En esta figura se muestra cómo ha variado el precio de secuenciar un genoma desde antes de la publicación del Proyecto Genoma Humano entre 2003 y 2004 hasta la actualidad. El eje de abscisas muestra la serie temporal, en años, y el eje de ordenadas indica el coste de secuenciación del genoma en dólares. Además, se han añadido varios hitos a esta línea temporal, como el proyecto Genoma Humano, las primeras secuenciaciones de tipo *next-generation* o los primeros genomas secuenciados por menos de 1000$. Datos obtenidos de https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Los principales hitos de la secuenciación en el siglo XXI se inician con la publicación del Proyecto Genoma Humano entre los años 2003 y 2004, cuyo objetivo fue la secuenciación completa del genoma del ser humano. Este gran proyecto supuso una inversión de más de trece años de trabajo y casi 3 mil millones de dólares[10,11]. La aparición de las técnicas NGS se pueden establecer entre los años 2005 y 2007, si bien se habían desarrollado y utilizado años antes, en este espacio de tiempo la tecnología revolucionó gracias, entre otros motivos, al lanzamiento del primer secuenciador NGS por la empresa Illumina, llamado Genome Analyzer. A partir de 2008 en adelante las técnicas de secuenciación *next-generation* se expandieron de forma global, permitiendo conocer el genoma de muchas especies, tanto vegetales como animales. A partir de este momento, se estableció el reto de producir la secuencia completa del ser humano con un precio de secuenciación inferior a 1000$[12], lo cual se cumplió en el año 2014. Actualmente

hay empresas que ofrecen este servicio por menos de 500$, como Ultima Genomics. En la Figura 2 se muestra cómo, a partir de la publicación del Proyecto Genoma Humano y, sobre todo tras la aparición del secuenciador Genome Analyzer el precio comenzó a abaratarse abruptamente.

La identificación de variantes, tanto nuevas como ya conocidas ha desembocado en el auge de bases de datos que almacenan estas variantes. De entre estos proyectos destacan *International HapMap Project*[13] para identificar variantes comunes en diferentes poblaciones o el *1000 Genomes Project*, en el que se secuenciaron 2500 genomas de 26 poblaciones para aumentar el catálogo de variaciones humanas. Entre los proyectos más actuales que han surgido con el fin de incorporar más variantes, cada vez menos frecuentes y raras destacan *UK10K*[14], *100.000 Genomes Project*[15] y la *Precision Medicine Initiative*[16]. En España también ha surgido un proyecto similar, denominado Collaborative Spanish Variability Server (CSVS)[17]. La aparición de estos ambiciosos proyectos no es más que el resultado del abaratamiento y mejora de la eficacia de las técnicas de secuenciación, principalmente impulsadas, como se ha mencionado, por las innovaciones en la secuenciación NGS.

Aunque el Proyecto Genoma Humano se publicó entre los años 2003 y 2004, actualmente hay un ambicioso proyecto internacional en marcha, llamado *Telomere-to-Telomere* (T2T), cuyo objetivo es alcanzar la primera secuencia de ADN completa y de calidad del ser humano[5]. Aunque se prevé que el fin del proyecto tenga lugar en 2026, ya en 2022 se han completado muchas regiones faltantes o incompletas utilizando las técnicas de secuenciación NGS de cadena larga, entre las que destacan Oxford Nanopore y PacBio.

El análisis de asociación del genoma completo (GWAS, por sus siglas en inglés) es una herramienta poderosa en la genómica que permite identificar variantes genéticas comunes que se asocian con rasgos complejos[18]. En otras palabras, GWAS es una técnica estadística que se utiliza para explorar el genoma humano en busca de variaciones genéticas asociadas con enfermedades, características fenotípicas y rasgos complejos en grandes grupos de individuos. El análisis de GWAS ha revolucionado nuestra comprensión de la genética de enfermedades complejas, y ha llevado a la identificación de cientos de variantes genéticas asociadas con enfermedades tales como la diabetes, la enfermedad de Alzheimer, y el cáncer. En consecuencia, el análisis de GWAS se ha convertido en una herramienta esencial para el descubrimiento de nuevos objetivos terapéuticos y para el desarrollo de nuevas estrategias de tratamiento para enfermedades genéticas complejas.

1.1.2 ARN y Datos Transcripcionales, Desde los Microarrays a la Tecnología RNA-Seq

Análogamente al ADN, la molécula de ARN también está formada por una cadena de nucleótidos. No obstante, existen algunas diferencias, principalmente relativas a su composición; la timina se sustituye por uracilo y la desoxirribosa por ribosa, su estructura; ya que se trata de una simple hebra, y función; ya que mientras el ADN almacena la información genética, el ARN actúa, en la mayoría de los casos, como mensajero de esta información.

Las secuencias de ARN resultantes de la transcripción del ADN se denominan transcritos. En otras palabras, los transcritos son copias de los genes en formato de ARN, que posteriormente pueden ser traducidos a proteínas o desempeñar otras funciones importantes en el organismo. El transcriptoma de un organismo se compone de un grupo muy heterogéneo de macromoléculas de ARN. Existen varios tipos de ARN, pero los más importantes son ARN mensajero (ARNm), ARN ribosómico (ARNr), ARN de transferencia (ARNt), ARN no codificante (ARNnc), ARN de interferencia (ARNi) y ARN pequeño no codificante (ARNsnc). Cada tipo de ARN cumple una serie de funciones. El ARNm es el más conocido de todos y se encarga de transmitir la información desde el ADN hasta los ribosomas, por lo tanto, es un intermediario entre el ADN y las proteínas. Se sintetiza en un proceso llamado transcripción, que consiste en transcribir una de las hebras de ADN a una molécula de ARN. En células eucariotas, este proceso es llevado a cabo por una enzima llamada ARN polimerasa, y se realiza en el núcleo de la célula. La ARN polimerasa, se encarga de localizar el sitio de inicio de la transcripción, determinado por una secuencia específica llamada promotor, desencadenando la apertura de la doble hélice. Mediante complementariedad, la enzima se encarga de generar un molde de ARN a partir de una de las dos hebras de ADN. Este proceso finaliza cuando detecta otra secuencia concreta, llamada terminador y libera la cadena de ARN al núcleo. Tras finalizar este proceso, el ARNm inmaduro debe modificarse antes de liberarse al citoplasma, añadiendo una caperuza en el extremo 5’ y una cola poli-A en el extremo 3’. Además, tienen lugar un proceso llamado *splicing*, mediante el cual se eliminan los trozos de ARN que pertenecen a las regiones intrónicas del gen.

Las funciones en las que intervienen el resto de moléculas de ARN son: ARNr; que forma parte de los ribosomas, donde se realiza la síntesis de proteínas; ARNt; transporta los aminoácidos a los ribosomas; ARNnc; desempeñan diferentes roles en la regulación de la expresión génica,

los más conocidos dentro de este grupo son los microARN y los ARN no codificantes largos (ARNlnc); ARNi interviene en el sistema de defensa contra virus y en la regulación de la expresión génica; ARNsnc; también desempeñan un papel en la regulación de la expresión.

La transcriptómica surge de la necesidad de conocer cómo se expresan los diferentes transcritos en un tejido o tipo de célula concreto o en una condición. De este modo, podemos establecer una conexión entre los genes expresados en una muestra, la cantidad de cada uno de estos genes y las funciones biológica so rutas metabólicas en las que intervienen las proteínas que son producto de estos transcritos. Para detectarlos y cuantificarlos, la transcriptómica ha desarrollado principalmente dos métodos: mediante *microarrays* de ADN o por secuenciación de ARN (conocido comúnmente como RNA-Seq)[19,20].

Los *microarrays* de ADN son las primeras técnicas utilizadas para cuantificar las moléculas de ARN. A pesar de que estas técnicas se han visto desplazadas por la secuenciación, aún se siguen utilizando. Estos *microarrays* son superficies planas, generalmente de vidrio, que contienen oligómeros, conocidos como sondas, que son secuencias de ADN de interés, por ejemplo, la secuencia de un gen. Una vez las secuencias de ARN se han aislado y replicado mediante la reacción en cadena de la polimerasa (PCR), se les añade un grupo fluoróforo y posteriormente se añaden a los *microarrays*. Las secuencias de ARN complementarias a las secuencias de ADN de las sondas se van a unir por complementariedad por hibridación, tal y como se ha detallado con los *microarrays* de SNPs. Dicha hibridación es detectada mediante fluorescencia y la abundancia (y presencia) de los transcritos es determinada en base a la intensidad lumínica que emita cada sonda (Figura 3a). Los datos generados por *microarrays* se almacenan en matrices de expresión o ficheros específicos que contienen la intensidad de señal de cada sonda de cada una de las muestras que se ha utilizado. Aunque hay varias plataformas de *microarrays* disponibles, las más conocidas pertenecen a las empresas de Affymetrix, Illumina, Agilent, Roche-NimbleGen o Thermo Fisher Scientific que aplican diferentes técnicas basadas en los mismos principios.

La secuenciación de ARN, o RNA-Seq es actualmente una de las técnicas más utilizadas. Tiene el mismo funcionamiento que se ha descrito en la secuenciación de ADN en el apartado anterior. Solo varía en unos pocos aspectos, ya que, una vez se extrae el ARN se genera una cadena de ADN complementaria (ADNc) mediante transcripción inversa. El resto del proceso es muy similar, incluyendo la fragmentación, amplificación y secuenciación. Las secuencias obtenidas se alinean con un transcriptoma de referencia y se cuantifican las secuencias que pertenecen a

cada transcrito. En este caso, la abundancia de un transcrito viene dada por la cantidad de fragmentos que se asocian al mismo (Figura 3b). El resultado de todo este proceso se suele analizar utilizando ficheros en formato *FASTQ* que contienen las lecturas de todas las secuencias que se han secuenciado. Estas lecturas también se suelen denominar *reads*. Existen múltiples herramientas desarrolladas para llevar a cabo el alineamiento o mapeado y la cuantificación[21].

Las secuencias que se generan en un análisis de RNA-Seq suelen mapearse o alinearse a un genoma o transcriptoma de referencia. El alineamiento consiste en asignar las secuencias de lectura a una secuencias de referencia Algunos de los alineadores más conocidos son STAR[22], Bowtie2[23] o Burrows-Wheeler Aligner (BWA)[24]. Tras alinear las secuencias, es necesario cuantificar el número de *reads* que se asocian a cada una de las secuencias de referencia. Del mismo modo que con los alineadores, hay una serie de algoritmos diferentes para cuantificar los transcritos, entre ellos: RSEM[25], eXpress[26], Sailfish[27] o kallisto[28].

Tras obtener los datos de expresión a partir de cualquiera de las técnicas descritas, se pueden aplicar varios tipos de análisis según el objetivo deseado. Lo más establecido es realizar un análisis de expresión diferencial, en el cual se comparan muestras de dos condiciones para buscar identificadores que permitan diferenciarlas. Sin embargo, hay otros análisis dirigidos a la búsqueda de patrones moleculares comunes de un conjunto de muestras con el fin de establecer grupos, aplicando técnicas de clústering o agrupamiento, o buscar transcritos *de novo*, es decir, secuencias de ARN que no se encuentran reflejadas en el genoma de referencia.

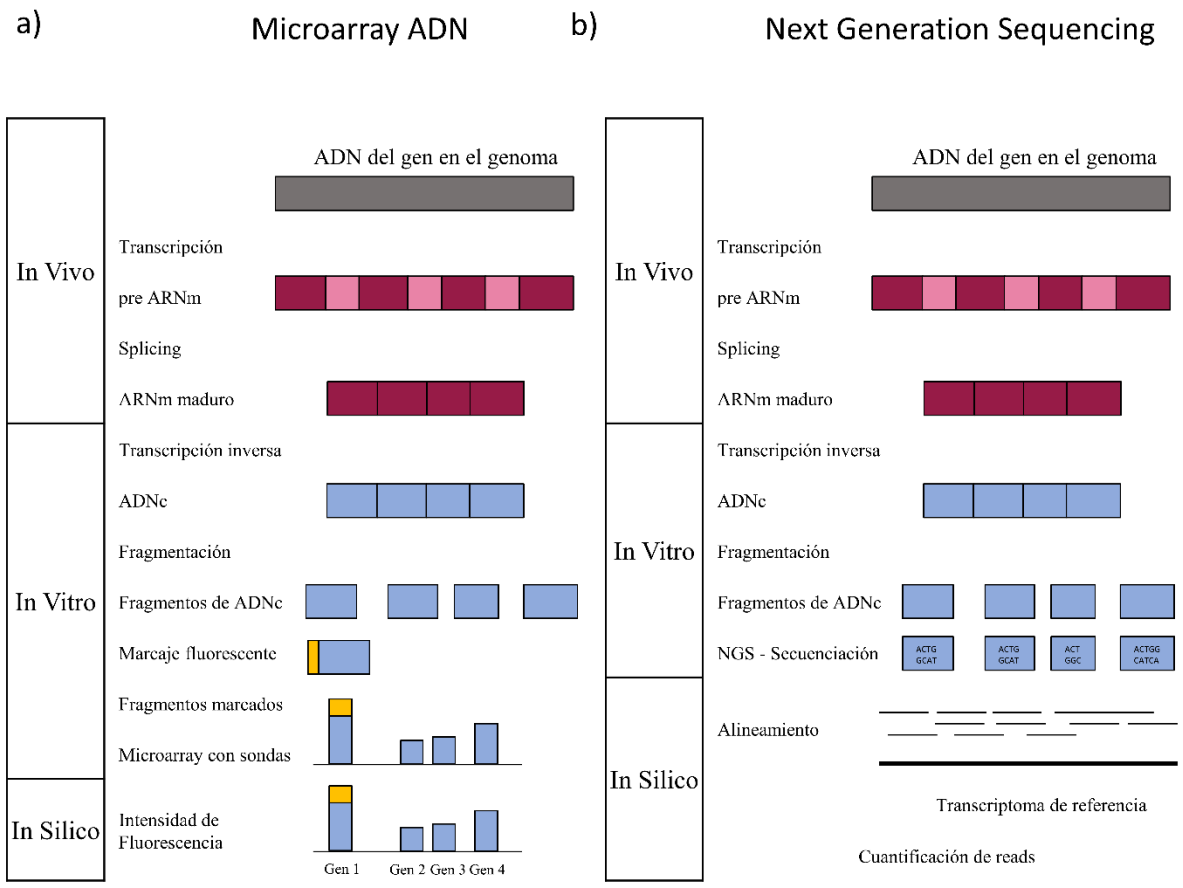

Figura 3. Métodos de obtención de datos transcriptómicos. a) Resumen de la metodología de microarrays de ADN. El ARN maduro se transcribe a ADN mediante transcripción inversa, este ADNc se fragmenta y se marca mediante fluorescencia. En el *microarray* se hibrida con una sonda determinada y, al estar marcado, se utiliza la intensidad lumínica que emiten los fragmentos marcados para determinar la cantidad de transcritos que hibridan con dicha sonda. b) Resumen de la metodología de RNA-Seq. El ARN maduro se fragmenta y se transforma en ADNc mediante transcripción inversa. Tras secuenciar estos fragmentos se alinean con un genoma de referencia y después se pueden cuantificar los diferentes transcritos. Imagen creada a partir de la idea propuesta en el artículo publicado por Lowe y colaboradores[19].

1.1.3 Detección de Fenómenos Epigenéticos.

Los cambios epigenéticos son aquellas modificaciones en la cadena de ADN que no alteran la secuencia de nucleótidos, sino que alteran la composición y estructura de esta con el fin de regular la expresión de los genes. Existen numerosos cambios epigenéticos que intervienen en la regulación génica como el remodelaje de la cromatina, la expresión de ARN no codificante, modificaciones en las histonas o la metilación de ADN[29]. El estudio global de todas ellas conforma el campo de investigación de la epigenómica, una ciencia ómica que suele considerarse como una de las ramas de la genómica.

De todas ellas las modificaciones epigenéticas, una de las más interesantes es la metilación de nucleótidos de ADN, que consiste en la adición de un grupo metilo (CH3) en una citosina siempre que esta se encuentre en un dinucleótido de citosina seguida de guanina, llamado CpG. Este fenómeno, que es reversible, es llevado a cabo por enzimas denominadas ADN metiltransferasas y generalmente los dinucleótidos CpG no se encuentran aislados, sino que se organizan en regiones ricas en sitios CpG, llamadas islas CpG, que suelen localizarse en los promotores de los genes[30]. Las regiones promotoras tienen una importancia vital, ya que son las regiones en las que se inicia la maquinaria de la transcripción.

Aunque los mecanismos por los cuales la metilación de ADN influye en la regulación de la expresión de genes aún no se conoce al completo, uno de los mecanismos supone que, debido a que la metilación de las islas CpG provoca cambios en la conformación de la cadena de ADN lo cual interviene dificultando la accesibilidad de los factores de transcripción al ADN, teniendo un efecto inhibitorio en la expresión de los genes[31]. Aunque hay varias técnicas para conocer si un nucleótido se encuentra metilado, las más utilizadas se basan en el tratamiento de bisulfito, bien utilizando microarrays o mediante secuenciación.

Existen varias plataformas de *microarrays* para detectar la metilación del ADN, pero las más usadas son las desarrolladas por Illumina: Infinium HumanMethylation450 e Infinium HumanMethylation EPIC. Las dos difieren en el número de sondas que contienen, alrededor de 450 mil la primera y cerca de 900 mil la segunda. Estas sondas están formadas por las secuencias de los sitios CpG conocidos por lo que se puede saber si hibrida una secuencia metilada o una no metilada. El proceso comprende los siguientes pasos: extraer y fragmentar el ADN, se aplica un tratamiento con bisulfito a los fragmentos de ADN, este tratamiento va a convertir todos los nucleótidos de citosina a uracilo, salvo los nucleótidos metilados, ya que el

grupo metilo protege al nucleótido de la acción del bisulfito. Posteriormente se amplifica el ADN mediante PCR y se añaden etiquetas fluorescentes a los sitios CpG. Finalmente, se añaden al *microarray* para que las sondas hibriden, de modo que se puede medir la cantidad de metilación en base a la intensidad de fluorescencia de cada secuencia[32].

Por otro lado, la secuenciación de genoma completo con tratamiento de bisulfito aúna las características de la tecnología NGS para la secuenciación y la idea de utilizar el bisulfito para diferenciar CpG metiladas y no metiladas[33]. El procedimiento no difiere en demasía del utilizado en los *microarrays*, pero en vez de obtener los resultados midiendo la intensidad de fluorescencia de unas sondas concretas, se lleva a cabo la secuenciación y posterior alineamiento de las secuencias. Las citosinas que queden en la secuencia se deberán a ADN metilado, mientras que las citosinas que no aparezcan serán correspondientes a sitios no metilados.

Otro de los fenómenos epigenómicos más estudiados son las interacciones entre proteínas y el ADN y el efecto de las mismas en la regulación de la expresión génica. Entre las técnicas más utilizadas para detectar este efecto, las basadas en la inmunoprecipitación son las más utilizadas. Las técnicas de inmunoprecipitación son una familia de técnicas utilizadas para purificar una proteína de interés y su ADN asociado a partir de una mezcla compleja de proteínas y ADN. Estas técnicas se basan en la especificidad de la unión entre un anticuerpo específico y la proteína de interés, lo que permite la purificación de la proteína junto con su ADN asociado.

Una vez que la proteína y su ADN asociado se han purificado mediante inmunoprecipitación, se pueden utilizar diferentes técnicas para analizar la secuencia de ADN asociada. Por ejemplo, se puede utilizar la secuenciación NGS para analizar los sitios de unión proteína-ADN a lo largo del genoma, técnica conocida como ChIP-Seq[34]. Dicha técnica consiste en fijar las proteínas de una célula con formaldehido para asegurar que se unan a su sitio de unión en el ADN, e identificar la proteína de interés mediante fluorescencia. Posteriormente se rompe la cromatina, liberando el ADN en fragmentos y se añaden anticuerpos, que se unirán a la proteína diana. De esta forma se obtienen todos los fragmentos a los que se han unido los anticuerpos. Finalmente se eliminan los anticuerpos, se secuencian los fragmentos y se reconocen los sitios de unión mediante perfiles de picos de fragmentos[35].

1.1.4 TECNOLOGÍAS DE CÉLULA ÚNICA

La tendencia actual de la biotecnología y la biomedicina es el uso de técnicas orientadas al estudio a nivel ómico de células individuales (denominadas *single-cell*). Cada vez son más las publicaciones que incluyen este conjunto de técnicas ya que permite conocer el estado transcripcional, genómico o epigenómico de células individuales. Este tipo de técnicas permite, entre otras cosas, analizar la heterogeneidad celular de un tejido en una condición específica.

Lo que diferencia a las técnicas de *single-cell* de las técnicas aplicadas a tejido o conjunto de células es la capacidad de aislar individualmente a cada una de las células de una muestra. La aplicación más extendida de métodos los de *singe-cell* es el estudio de la heterogeneidad celular de un tejido a partir del análisis del ARN mediante secuenciación NGS, conocido como *single-cell* RNA-Seq (scRNA-Seq). En la actualidad, hay varias plataformas que permiten realizar este proceso, difiriendo en la estrategia de aislamiento de las células[36]:

- *Fluorescence-activated cell sorting* (FACS). Se caracteriza en la separación y análisis de las células individuales mediante el uso de anticuerpos fluorescentes y un citómetro de flujo. Las células se marcan con anticuerpos específicos, después se les añade un marcador fluorescente, por lo que pueden separarse en grupos según su fluorescencia en un citómetro. Las plataformas que utilizan este tipo de estrategias son: Smart-seq, Smart-seq2, CEL-seq, MATQ-seq o MARS-seq.
- *Micro-fluidic*. Consisten en el uso de dispositivos de flujo controlando a escala microscópica para manipular y analizar células individuales. Algunos ejemplos de técnicas microfluídicas son: flujo lateral (desviar y aislar células en un mcrocanal) o *sorting* por flujo (utiliza un flujo continuo para separar las células en base a sus propiedades físicas, como tamaño o densidad). Las plataformas que utilizan estas técnicas son Fluidigm C1 y Seq-Well.
- *Microdroplets*. Los *microdroplets* son un tipo especial de técnicas microfluídicas. Se trara de una técnica de flujo continuo de líquido que se divide en pequeñas gotas mediante un interfaz líquido-líquido o líquido-aire. Por tanto, la células o pequeño grupo de células son encapsuladas por estas gotas, lo que sirve para asilarlas. Muchas plataformas utilizan este tipo de técnicas, entre ellas: Drop-seq, 10x Genomics, InDrop-seq o DNBelab C4.

Aunque las técnicas de scRNA-Seq aún presentan muchas limitaciones, desde hace unos años hay múltiples repositorios públicos que se encargan de alojar datos y resultados de este tipo de análisis, como ExpressionAtlas o GTEX[37,38].

1.1.5 La Bioinformática y la Biología Computacional

Hasta ahora se han descrito algunas de las ciencias ómicas más relevantes para el desarrollo de esta tesis, además del caso excepcional de las tecnologías de células única. El mensaje principal que se ha querido transmitir es que todas estas metodologías generan una cantidad masiva de datos muy complejos que necesitan aproximaciones sofisticadas y eficientes para su análisis. Debido a esto, fue necesario impulsar una nueva disciplina que se encargase de evaluar y procesar toda esta información. De este modo surgió la bioinformática Debido a la enorme cantidad y complejidad de los datos generados por las ciencias ómicas, fue necesario impulsar varias disciplinas para su análisis, la bioinformática y la biología computacional, que, si bien presentan pequeñas diferencias, ambas se centran en la combinación de la informática, la estadística y las matemáticas para aplicar métodos de análisis sofisticados a estos datos de origen biológico.

La aparición de estas disciplinas ha supuesto un cambio en el paradigma del método científico clásico aplicado a la biomedicina o la biotecnología. El método científico clásico consiste en hacer hipótesis a partir de observaciones y refutar dichas hipótesis mediante la generación de datos orientados a esa hipótesis y la comprobación experimental. El actual paradigma post-ómico o post-bioinformático no parte de hipótesis específicas, sino que, a partir de un diseño experimental se analizan todos los datos, se interpretan los resultados y se proponen hipótesis a partir de dichos resultados (véase Figura 4). Un ejemplo de este cambio se puede apreciar en los estudios dirigidos a buscar fármacos que se puedan utilizar en alguna patología. Antes, la búsqueda de fármacos se basaba en enfoques empíricos, donde los compuestos eran descubiertos a través de ensayo y error, basándose en la observación de sus efectos en animales o en pacientes. Los científicos dependían principalmente de la síntesis de compuestos químicos y la realización de ensayos biológicos para encontrar posibles candidatos.

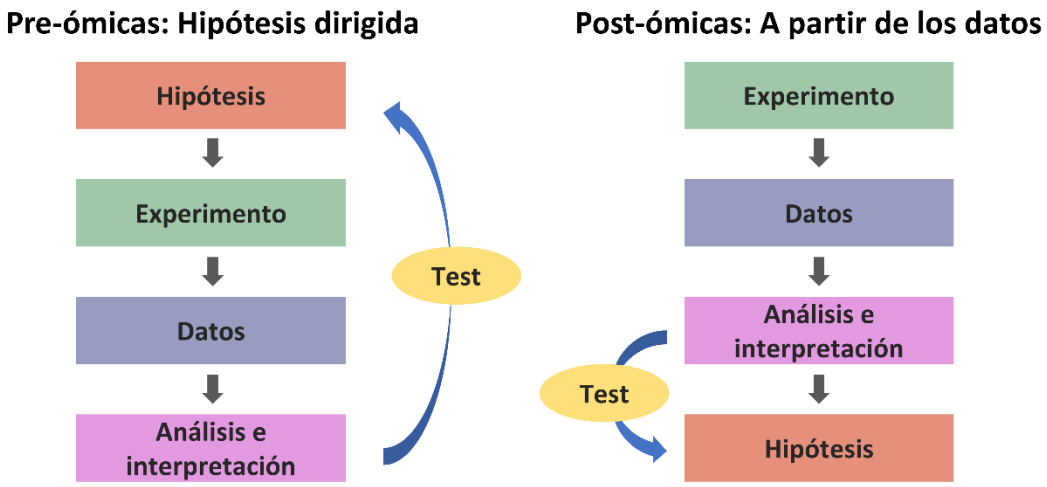

Figura 4. Cambio de paradigma en el método científico a partir de la aparición de las ciencias ómicas y la evolución de la bioinformática.

En cambio, hoy en día, la búsqueda de fármacos se ha vuelto sistemática, gracias a las tecnologías de ciencias ómicas. Ahora, los científicos pueden utilizar herramientas de análisis de datos a gran escala para descubrir biomarcadores y comprender las redes de interacción molecular involucradas en las enfermedades. Esto permite una identificación más precisa de dianas terapéuticas y una mayor eficiencia en la identificación de candidatos a fármacos. Además, se utilizan técnicas de modelado molecular y cribado virtual para identificar y optimizar moléculas prometedoras de manera más eficiente.

Los campos de aplicación de la bioinformática son cada vez más sofisticados y específicos, sin embargo, algunos de los métodos más utilizados y estandarizados son; la búsqueda de biomarcadores a partir de comparación de condiciones, métodos de agrupamiento (también llamado *clustering*), los análisis de enriquecimiento funcional y la integración de datos ómicos. A continuación, se detallarán las técnicas más conocidas de cada uno de estos campos de aplicación de la bioinformática.

1.1.5.1 COMPARACIÓN DE CONDICIONES

Gran cantidad de estudios y experimentos se han centrado en buscar biomarcadores a partir de comparaciones entre muestras afectadas y controles sanos. Esta metodología está muy extendida en todos los tipos de ómicas. De hecho, ya hemos hablado un poco de esto en el apartado de genómica y transcriptómica, pues cuando buscamos variantes genómicas o genes diferencialmente expresados realmente estamos buscando diferencias entre grupos de muestras.

Puesto que la tesis se centra en analizar, principalmente, datos transcripcionales, vamos a describir diferencias técnicas que se han aplicado para obtener un listado de genes diferencialmente expresados. La más primitiva de ellas, y la más simple es seleccionar un umbral de valores de expresión como la razón entre las dos condiciones (expresión condición A / expresión condición B). Sin embargo, como es lógico, este método es totalmente arbitrario y carece de base estadística alguna. Las técnicas estadísticas más estandarizadas para comprobar diferencias entre dos grupos son el test de las t de Student o la prueba de la U de Mann-Whitney, que se diferencian entre sí en que, mientras la primera se trata de una prueba paramétrica, la segunda es un test no paramétrico. Estos métodos sí que aportan una base estadística que permite complementar la diferencia o razón de los valores de expresión, sin embargo, en el caso de t de Student, presenta algunas limitaciones, como la asunción de que los niveles de expresión tienen varianzas iguales en las dos clases y siguen una distribución normal, por lo que pequeños cambios en la varianza de un gen pueden provocar fluctuaciones considerables en el valor de t.

Aunque se han desarrollado algunos métodos más para hacer expresión diferencial, los más utilizados, tal y como se ha comentado en la sección de transcriptómica son limma y DESeq2. El paquete limma (*Linear Models for Microarray Data*) es una herramienta de análisis que se desarrolló para llevar a cabo análisis de expresión diferenial en datos transcriptómicos de *microarray*[39]. El paquete limma ajusta un modelo lineal para cada gen y utiliza una técnica de estimación de máxima verosimilitud para estimar los coeficientes de los efectos de interés. Los efectso de interés pueden ser tratamientos o grupos de muestras. Posteriormente utiliza la técnica de regularización, llamada *empirical Bayes*, para mejorar la precisión de las estimaciones de los coeficientes y reducir el error de tipo I. Tras ajustar el modelo lineal se realiza una prueba estadística para evaluar la significancia de la diferencia en la expresión de un gen entre dos grupos de muestras, calculando el pvalor a utilizando la distribución t o F, según sea apropiado para el número de grupos y el diseño experimental. Para analizar datos de

RNA-Seq, limma presenta ciertos cambios con respecto a los datos de *microarrays*, debido a la diferencia que existe en las técnicas de medición de expresión génica de ambas aproximaciones, que ya se han descrito previamente. El análisis de RNA-Seq utilizando el paquete limma consta de varias etapas. En primer lugar, se requiere hacer una normalización de los datos para ajustar la expresión de los genes por tamaño de la muestra mediante la normalización por cuantiles. Posteriormente se estima la dispersión de los datos de cada gen con el fin de mejorar la estimación de la varianza y la sensibilidad a la hora de detectar genes diferencialmente expresados. Finalmente, se ajusta un modelo linear similar al descrito para los datos de *microarrays*. También se suele aplicar la misma metodología de limma para buscar genes diferencialmente expresados en *microarrays* con datos de metilación procedentes de metilación.

Por el contrario, el paquete DESeq2 solo es aplicable a datos de RNA-Seq. Este método utiliza un modelo estadístico negativo binomial generalizado para modelar la relación entre el número de lecturas de ARN y la expresión génica, y para estimar los parámetros de dispersión de los datos[40]. Luego, se utiliza un método de ajuste empírico de la dispersión para ajustar la varianza en función del nivel de expresión y de las diferencias entre las muestras. Una vez que se han ajustado los datos, se utiliza un análisis de contraste para identificar los genes que están diferencialmente expresados entre las condiciones experimentales. DESeq2 también proporciona herramientas para normalizar los datos, filtrar los genes con poca expresión, ajustar la varianza y corregir múltiples pruebas.

Independientemente del método seleccionado, aquellos que tienen una base estadística se obtendrá un p-valor para cada gen. Un p-valor pequeño indica que la probabilidad de que existan diferencias para un determinado gen entre dos conjuntos de muestras se deba al azar es pequeña, por lo que se habla de genes diferencialmente expresados. Además, en muchas ocasiones es recomendable tener en cuenta el *fold change*, que indica la diferencia de magnitud entre los dos grupos de muestras. Por ejemplo, si el grupo experimental tiene un nivel medio de expresión génica 2 veces mayor que el grupo de control, se dice que el *fold change* es 2. Un *fold change* de 2 implica un aumento de 2 veces en la expresión génica entre los dos grupos.

1.1.5.2 MÉTODOS DE CLÚSTERING

Una de las metodologías más usadas en el análisis de matrices de expresión génica son los algoritmos de agrupamiento o *clustering*. Estos algoritmos se utilizan para dividir los elementos de una matriz en grupos homogéneos que se distingan del resto de grupos. En el campo de aplicación de los datos de expresión génica (así como de metilación), el objetivo de utilizar estos algoritmos es encontrar conjuntos de genes (o CpGs) o de muestras que presenten patrones de expresión muy similares. De este modo, somos capaces de localizar genes que se expresen de forma similar o muestras que se agrupen en grupos uniformes. Los métodos de agrupamiento más utilizados en el análisis de datos de expresión génica son:

- Algoritmos de agrupamiento jerárquico, que consiste en ordenar los elementos de una población en base a un árbol de distancias como reflejo de la similitud entre los elementos y los grupos. De entre los algoritmos de agrupamiento jerárquico más utilizados destacan los algoritmos aglomerativos y los algoritmos divisivos. Los primeros comienzan con un conjunto de datos sin etiquetas, que se agrupan en base a las distancias que presentan todos con todos, de forma que los dos elementos más similares forman un grupo. Posteriormente se calculan la matriz de distancias, teniendo en cuenta que los dos elementos del grupo anterior forman un nodo, y se vuelven a unir los dos elementos más similares. Este proceso se repite hasta que se unen los dos últimos grupos. Por otro lado, los algoritmos divisivos comienzan con un único clúster que va dividiendo mediante iteraciones en grupos más pequeños hasta llegar a los elementos únicos. En resumen, la principal diferencia entre ambos enfoques radica en el sentido de la agrupación: los algoritmos aglomerativos combinan grupos similares, mientras que los algoritmos divisivos dividen grupos heterogéneos. Entre los algoritmos aglomerativos destacan el algoritmo de enlace único (*single linkage*), el algoritmo de enlace completo (*complete linkage*) y el algoritmo de enlace promedio (*average linkage*), mientras que de los métodos divisivos podemos destacar el algoritmo de Ward. La representación gráfica de este tipo de algoritmos es un dendograma o estructura de árbol donde se muestra la relación que existe entre los diferentes clústeres (Figura 5a)

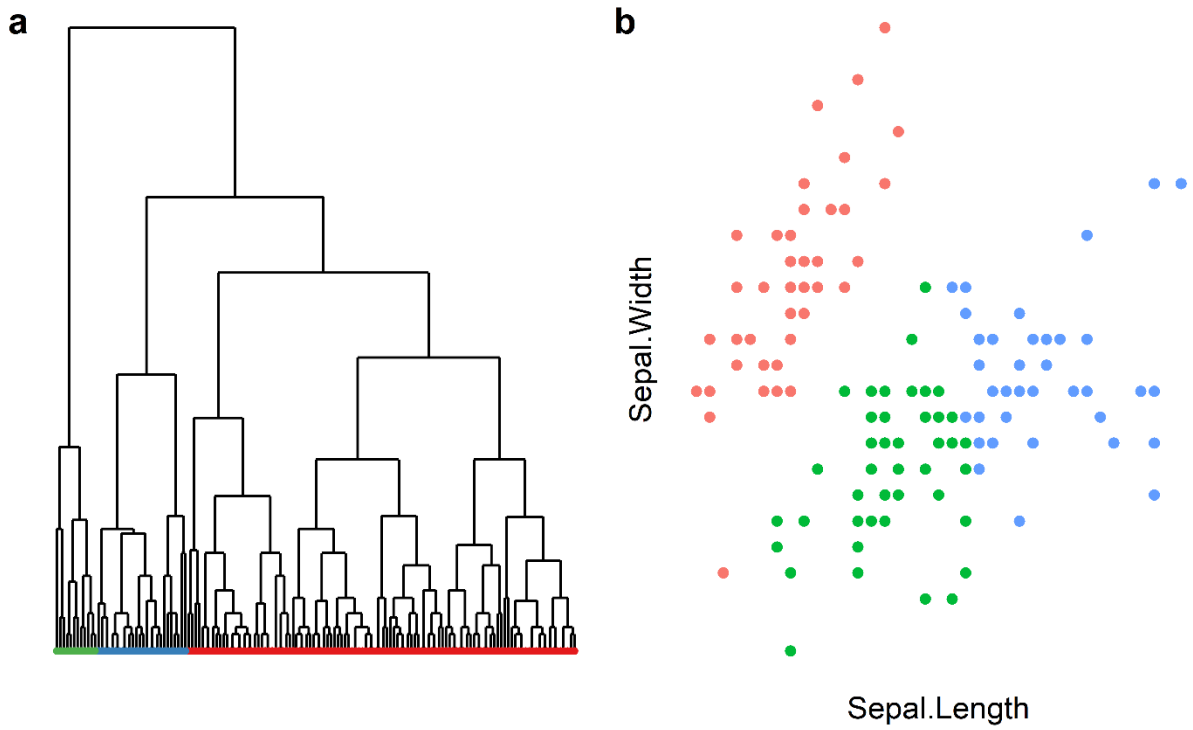

Figura 5. Representación gráfica de los principales algoritmos de agrupamiento o clustering aplicados a datos de expresión génica. a) Dendograma obtenido a partir de un clustering jerárquico. b) Representación espacial de los datos y los centroides obtenidos al aplicar un kmeans.

- Algoritmos de agrupamiento particionales, que se caracteriza porque el número de grupos o clústeres se deben definir previamente. El algoritmo más conocido de este grupo es el llamado K-medias o *kmeans*, que consiste; seleccionar el número K de clústeres, asignar aleatoriamente K puntos como los centroides iniciales de los clústeres, asignar cada punto de datos al clúster cuyo centroide esté más cerca, calcular la media de los puntos en cada clúster y actualizar la posición de su centroide. Posteriormente, se repiten los pasos anteriores hasta que los centroides no cambien o la variación se mínima, momento en el que los clústeres finales quedan definidos por la posición de los centroides. La representación gráfica de estos algoritmos suele hacerse como grupos de puntos en un espacio de dimensiones reducidas (generalmente 2D) o en un diagrama de dispersión (Figura 5b)

1.1.5.3 Integración de Datos Ómicos

Aunque hasta ahora se han explicado las ciencias ómicas como disciplinas independientes entre sí, cuando se integran dos o más ómicas puede dar un conocimiento mucho más conciso de las interacciones y rutas moleculares[41]. A este tipo de estudios se les conoce como estudios multiómicos.

Uno de los ejemplos de la integración de ómicas son las técnicas para determinar eQTL (*Expression Quantitative Trait Loci*), que son biomarcadores genómicos asociados con la expresión de genes. En resumen, estas técnicas consisten en identificar genes diferencialmente expresados y variantes genómicas entre dos conjuntos de datos[42].

Otro ejemplo de estudio de datos multiómicos es la integración de datos transcripcionales y epigenómicos, en este caso a través de la metilación. Como se ha comentado anteriormente, la metilación en la región promotora de un gen tiene el efecto de dificultar la unión de los factores de transcripción que regulan la expresión de un gen[43]. Por tanto, es posible estudiar de forma conjunta los dos efectos, la metilación de regiones específicas y la expresión de los genes, para tener un conocimiento del sistema más completo.

Existen varios métodos de integración de datos ómicos, bien utilizando el mismo tipo de ómica o a partir de estudios multiómicos. Entre los métodos de integración de datos ómicos de la misma ómica destacan las técnicas de meta-análisis. El conjunto de técnicas de meta-análisis se utilizan para integrar y resumir los resultados de múltiples estudios en una sola estimación, por lo que permiten aumentar la precisión y la potencia que si solo se utilizara un estudio. Hay múltiples técnicas de meta-análisis de datos ómicos, como las técnicas de p-valor, de tamaño de efectos o combinación de rangos para datos transcriptómicos[44], aunque también pueden aplicarse a datos epigenómicos y genómicos.

De entre los métodos de integración de datos multiómicos, destacaremos la técnica SNF (*Similarity Network Fusion*) que se basa en evaluar las distancias que existen entre las muestras en cada tipo de fuente ómica[45]. Por ejemplo, si tenemos datos transcriptómicos y epigenómicos, podemos calcular la distancia entre las muestras a partir de las matrices de expresión y metilación, de forma individual. SNF combina estas matrices de distancias para construir una red integrada que representa la similitud global entre los pacientes.

Otros métodos de integración son iCluster[46], que utiliza modelos matemáticos para agrupar los datos similares en clústeres y para analizar los patrones de asociación entre los diferentes tipos de datos, o MEFISTO, un método reciente que se utiliza para analizar patrones de datos temporales y espaciales en datos multiómicos[47].

1.1.5.4 Análisis de Enriquecimiento Funcional

En los apartados anteriores hemos comentado diferentes técnicas para procesar y generar resultados a partir de datos ómicos. Sin embargo, uno de los aspectos más importantes a tener en cuenta a la hora de enfrentarse a estos resultados es la interpretación biológica de los mismos, ya que puede ser un desafío debido a la cantidad masiva de datos. Por ejemplo, después de realizar un análisis de expresión diferencial entre dos condiciones, encontrar sitios metilados asociados a una enfermedad o identificar variantes genéticas relacionadas con una enfermedad, la lista de resultados (ya sea de genes, CpGs, variantes, proteínas, etc.) puede ser abrumadora. Por este motivo es común recurrir a técnicas de enriquecimiento y anotación funcional para indicar las funciones biológicas asociadas con los elementos de la lista de resultados.

Para evitar ser muy repetitivos, cuando hablemos de listas de genes nos referimos, por supuesto a una lista de genes que se han obtenido a partir de un análisis de expresión diferencial, pero también engloba a listas de CpGs, listas de variantes, listas de proteínas o cualquier otra lista de elementos de los cuales haya información en las bases de datos que generalmente se utilizan para hacer análisis de enriquecimiento funcional. Estas bases de datos, son, junto a las listas de genes y los métodos de enriquecimiento, los tres elementos necesarios para realizar este tipo de análisis. Ya hemos comentado algunas formas de obtener estas listas de genes, por ejemplo, a partir de un análisis de expresión diferencial, por lo que lo siguiente que vamos a detallar son las bases de datos que se utilizan con más asiduidad para obtener la información biológica, así como los métodos más usados.

1.1.5.4.1 Bases de Datos de Anotaciones Funcionales

Las bases de datos de anotaciones funcionales se utilizan para conocer las funciones en las que intervienen un gen o una proteína concreta, o en nuestro caso, una lista de genes. Podemos considerar como bases de datos de anotaciones funcionales todas aquellas bases de datos en las que un gen se asocia a una función biológica, a una enfermedad, a un fármaco, a otro gen, a un elemento regulador, a un compartimento celular o a un tipo celular concreto, entre otras muchas posibilidades. El uso de una u otra base de datos va a depender en gran medida de lo que se quiera obtener; por ejemplo, si queremos saber los genes asociados a una ruta metabólica concreta no vamos a usar la misma base de datos que si queremos conocer que genes se expresan en un tejido concreto. Debido a que el número de bases de datos de anotaciones funcionales es inmenso, nos vamos a centrar en describir las más utilizadas, que también son las más curadas y las que tienen un nivel de evidencia mucho más robusto.

Aunque las bases de datos pueden clasificarse de infinidad de formas, nosotros hemos decidido clasificarlas en base a la información que proporcionan, de forma que tenemos una serie de bases de datos que apuntan a rutas metabólicas generales o concretas, otras que asocian genes a enfermedades o a fármacos y el caso específico de *Gene Ontology*, por la cual empezaremos, ya que es la más utilizada por la comunidad científica.

1.1.5.4.1.1 GENE ONTOLOGY (GO)

Gene Ontology (GO) surgió de la colaboración del consorcio The Gene Ontology Consortium cuyo objetivo era clasificar el conocimiento biológico en una ontología[48]. Una ontología es la representación estructurada del conocimiento de un dominio, desde conceptos generales a aquellos muy específicos. GO describe el conocimiento en el dominio biológico agrupándolo en tres áreas. Estas áreas permiten clasificar y disponer de la información abarcando los tres aspectos fundamentales de la biología: procesos biológicos (GO BP), funciones moleculares (GO MF) y componentes celulares (GO CC).

Uno de los aspectos que hace de GO una de las bases de datos más consistentes es como se estructura la información, ya que se trata de una estructura jerárquica que utiliza como pilares las tres grandes áreas (GO BP, GO MF y GO CC). Esto implica que todos los términos que se incluyen en GO están relacionados entre sí mediante este modelo jerárquico de grafos (Figura 6a). De este modo, cada grafo ‘padre’ contiene a su vez uno o varios grafos ‘hijos’ que son más

específicos. La información se transmite de 'hijos' a 'padre' de forma que todos los genes implicados en una función de un 'hijo' también estarán implicados en la función de un 'padre'. Volviendo al ejemplo de la enzima ATPasa de transporte de potasio y sodio, el proceso biológico 'metal ion transport' tiene varios 'hijos', entre ellos 'sodium ion transport' y 'potasium ion transport' y a su vez tiene un 'padre'; 'monoatomic cation transport'. Al ser una estructura jerárquica, los genes implicados en 'sodium ion transport' y 'potasium ion transport' también participan en 'metal ion transport' y en 'monoatomic cation transport'.

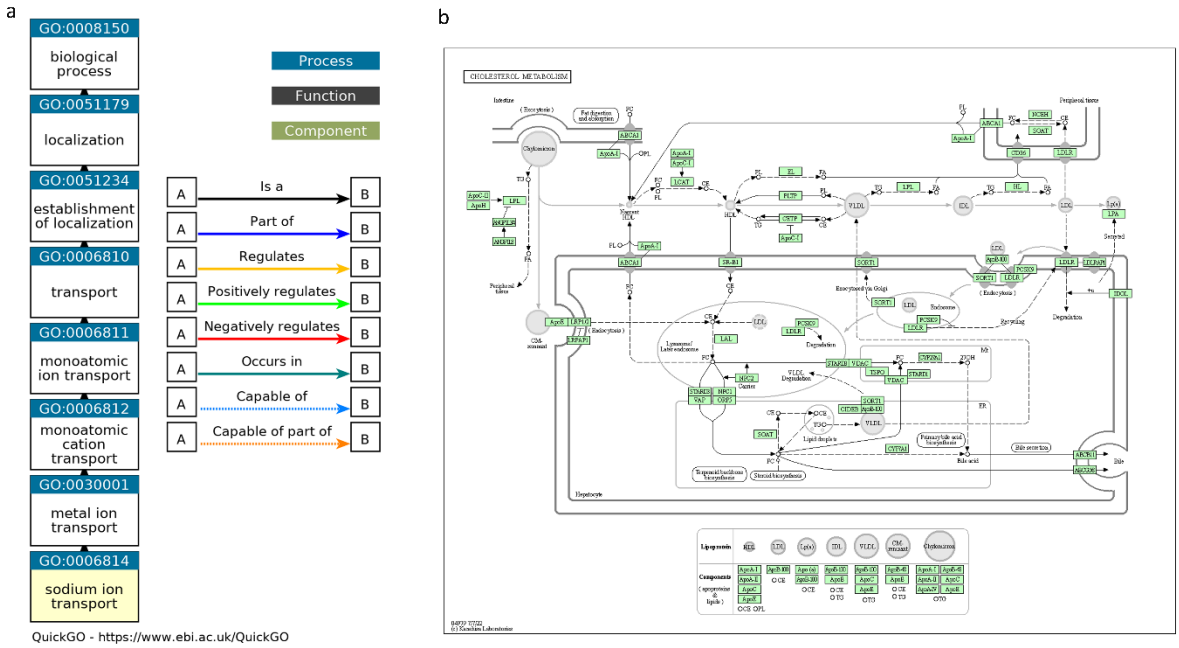

Figura 6. Estructura de información en las bases de datos de GO y KEGG. A) Estructura jerárquica de los términos de GO. El término "sodium ion transport" es un nodo hijo de "metal ion transport", que a su vez tiene una serie de grafos jerárquicos hasta llegar a "biological process". También se advierte que cada término ontológico presenta un identificador único. B) Mapa de elementos de la ruta "Cholesterol metabolism" de KEGG.

Las funciones moleculares describen acciones o actividades específicas que tienen lugar a nivel molecular, como actividad enzimática o la unión de proteínas. Esta categoría se enfoca en describir como una proteína o un producto génico contribuye a la biología molecular. Los procesos biológicos describen las series de eventos que ocurren en un organismo, desde la regulación de la expresión hasta la respuesta a un estímulo externo. Se encarga de relacionar las interacciones de las funciones moleculares para llevar a cabo procesos biológicos complejos. Los componentes celulares reflejan las estructuras y compartimentos celulares donde tienen lugar estos procesos biológicos. Para detallar estas tres áreas o categorías de GO, propondremos un pequeño ejemplo. La subunidad Alpha-1 de la enzima ATPasa de transporte de potasio y

sodio (ATP1A1) interviene en la función molecular “ATP binding”, que forma parte del proceso biológico “metal ion transport”, que sucede en el compartimento celular “plasmatic membrane”. Por tanto, esta proteína se encuentra asociada a estos tres términos.

1.1.5.4.1.2 BASES DE DATOS DE RUTAS METABÓLICAS

Obviando la base de datos de GO, las anotaciones funcionales que más se utilizan son aquellas que describen rutas metabólicas. Aunque pueda parecer que GO también describe rutas, esto no es así. En el ejemplo que describimos, la anotación ‘metal ion transport’ se refiere a un conjunto de mecanismos que se encargan de este proceso, pero no entran en detalle de cómo tiene lugar el mismo. Por ello, podemos diferenciar la base de datos de GO de aquellas en los que si se detallan los genes que participan en una ruta.

En este sentido, una lista de genes que se encuentren enriquecidos en una ruta biológica indicará que los resultados están asociados a esta ruta o rutas. De este modo, si, por ejemplo, obtenemos como resultado significativo a partir de una lista de genes asociados a una enfermedad el término “Cholesterol metabolism”, se puede inferir que existe una anormalidad en esta ruta que interviene en la enfermedad. De las bases de datos con información acerca de rutas destacan KEGG, Reactome, WikiPathways o BioPlanet.

La primera de ellas, KEGG, está diseñada para proporcionar información a los investigadores con el fin de entender cómo se relacionan los genes y las vías metabólicas en diferentes organismos[49]. Se trata de un entramado de información en la que se incluyen desde genomas hasta interacciones entre proteínas, además de vías metabólicas e información sobre enzimas y compuestos químicos. Uno de los aspectos más representativos de KEGG para realizar un análisis funcional es la información que contiene sobre los genes que intervienen en cada ruta metabólica y además permite conocer en profundidad estas rutas ya que dispone de visualizaciones en forma de mapa metabólico (Figura 6b). A pesar de que aquí hablamos de KEGG como herramienta para hacer enriquecimientos funcionales, no hay que olvidar que contiene mucha más información, incluyendo secuencia de nucleótidos y aminoácidos de genes y proteínas, respectivamente o información sobre la composición de algunos fármacos y la interacción con proteínas y vías metabólicas.

Otra de las bases de datos de rutas metabólicas más usadas es Reactome, que proporciona una versión extendida del clásico mapa de rutas metabólicas, de tal forma que ofrecen detalles

moleculares con respecto a señales de transducción, transporte, replicación de ADN u otros procesos celulares[50].

Siguiendo el formato y la idea de la famosa web de WikiPedia, en la cual cualquiera puede añadir o modificar información sobre cualquier tema, surge WikiPathways[51], una base de datos desarrollada por y para científicos. Sin embargo, no funciona de forma tan trivial como Wikipedia, sino que la información pasa por una curación manual a manos de diferentes comunidades científicas. En WikiPathways se puede encontrar desde información acerca de aspectos generales hasta de rutas muy específicas de enfermedades concretas.

Por último, como un esfuerzo con el fin de englobar éstas y otras bases de datos de rutas metabólicas, se desarrolló la base de datos de BioPlanet[52], cuya contribución es la recolección, integración y curación de la información procedente de varias bases de datos, incluyendo WikiPathways, KEGG y Reactome.

Un aspecto interesante de estas bases de datos basadas en rutas es que todas ellas proporcionan, además de la relación de genes y términos, un recurso gráfico en el cual se puede observar cada ruta de forma interactiva.

1.1.5.4.1.3 BASES DE DATOS DE PERTURBACIONES

Hay un grupo de anotaciones funcionales que contienen información sobre perturbaciones. Hemos agrupado con el concepto de perturbaciones tanto a perturbaciones biológicas (enfermedades o fenotipos anormales) como perturbaciones bioquímicas (principalmente mediante el uso de fármacos).

Las bases de datos de enfermedades o fenotipos que más se utilizan son el catálogo de Online Mendelian Inheritance in Man (OMIM)[53], que relaciona a los genes con enfermedades genéticas humanas, o Human Phenotype Ontology (HPO)[54] que proporciona la asociación de genes con fenotipos humanos anormales. También, en el campo de los fármacos, destacan Comparative Toxicogenomics Database (CTD), que contiene información curada de interacciones de compuestos químicos y genes[55] y PharmGKB, que contiene, entre otras cosas, información de los genes diana de cada fármaco incorporado en la base de datos[56]. Con estas bases de datos se pueden asociar resultados ómicos a enfermedades o compuestos químicos que afectan al conjunto de genes.

Aunque se trata de una base de datos que aporta mucha más información, el proyecto de LINCS Transcriptomics (también llamado L1000, alojado en la web de Connectivity Map) también aporta información acerca de los genes diana de los diferentes fármacos que incluye[57].

1.1.5.4.2 Métodos de Enriquecimiento Funcional

Los análisis de enriquecimiento funcional o de anotación funcional son útiles a la hora de interpretar una serie de resultados obtenidos a partir de datos ómicos. A nivel general, estos métodos se utilizan cuando se han obtenido un número elevado de genes o proteínas. Para realizar un análisis funcional se requiere, además del listado de genes y la base de datos sobre la que testar el conjunto de genes, un método estadístico que proporcione un nivel de significancia para cada uno de los términos que se pretenden testar. Los tres métodos más conocidos son: SEA (Singular Enrichment Analysis), GSEA (Gene Set Enrichment Analysis) y MEA (Modular Enrichment Analysis)[58]. En la Figura 7 se puede apreciar un esquema de cómo llevar a cabo las diferentes técnicas.

Dentro de las técnicas SEA se aplican métodos estadísticos que utilizan el tamaño de muestra como parámetro. Algunos de estos métodos son muy conocidos y muchos de ellos siguen distribuciones clásicas como la binomial, la chi-cuadrado o la hipergeométrica. Estos métodos se basan en comprobar el solapamiento entre la lista de genes que se quiere testar y la lista de genes asociados a un término concreto. Hay multitud de herramientas online y librerías desarrolladas en R que permiten hacer este tipo de análisis utilizando como entrada una lista de genes o proteínas. Entre estas herramientas, las más utilizadas son DAVID Functional Annotation Tool[59], EnrichR[60] o GOSeq[61]. A pesar de que su uso está muy extendido, presentan una serie de limitaciones ya que se asume que hay independencia entre los genes, sin tener en cuenta que existen genes que se coexpresan, se asume que las rutas son independientes entre sí y no se solapan y es un análisis subjetivo, ya que la lista de genes de entrada la introduce el propio usuario.

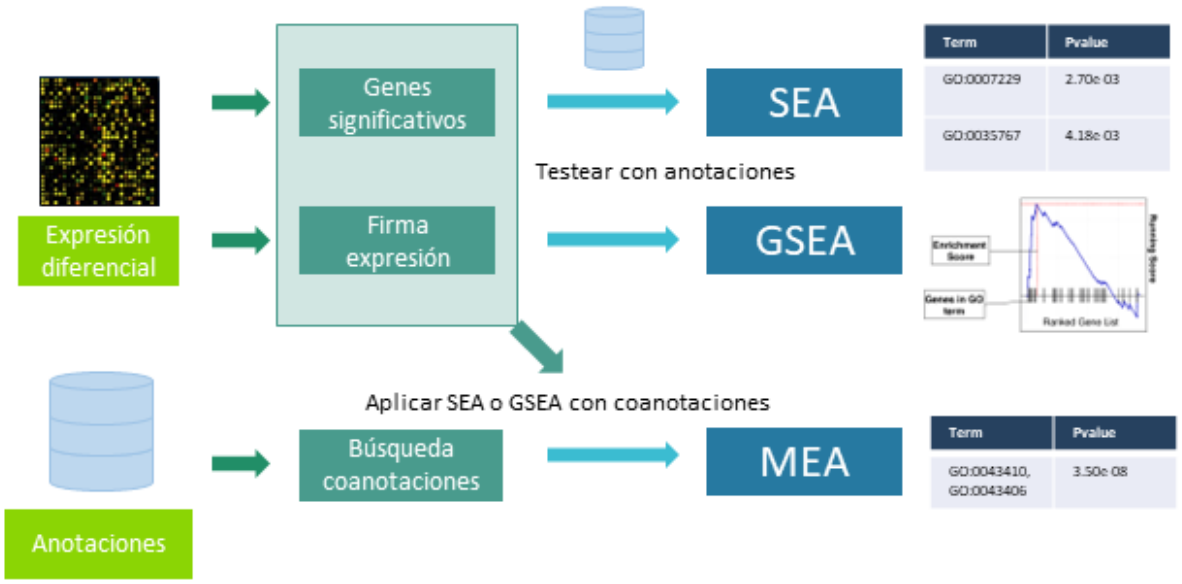

Figura 7. Esquema general de los tres tipos principales de análisis de enriquecimiento.

En el año 2005 se publicó un método para hacer enriquecimiento funcional basado en ordenar todos los genes mediante un valor estadístico a partir de la comparación entre dos grupos, como puede ser el *fold change* o el valor t que se obtienen a partir de aplicar una t de Student. Posteriormente se calcula un valor de enriquecimiento (*Enrichment Score* o ES) de cada término en base a que posiciones de la lista ordenada ocupan los genes asociados a ese término. Este método se conoce como GSEA y da nombre a diferentes variantes metodológicas desarrolladas a partir de la publicación original[62]. GSEA presenta la ventaja de que no depende de una lista de genes de entrada subjetiva, ya que se utilizan todos los genes disponibles.

Los métodos MEA consisten en analizar los términos anotados de forma modular en vez de individualmente. Es decir, se encargan de identificar relaciones entre términos y evaluar la significancia de la relación. Existen varias herramientas y softwares que emplean esta metodología, como topGO[63], Ontologizer[64] o GeneCodis[65]. Estas técnicas permiten, además de analizar los términos de forma global en lugar de individual, reducir la redundancia de resultados en anotaciones muy heterogéneas. Lo que diferencia las herramientas que aplican métodos MEA entre ellas es la estrategia que usan para identificar las relaciones entre términos,

generalmente mediante técnicas que buscan coocurrencias. La gran limitación de este tipo de técnicas es que solo son útiles cuando la anotación/es que se utilizan presentan cierto solapamiento entre ellas, es decir, que hay pocos términos o genes “huérfanos” o poco relacionados con el resto.

1.2 Redes de Regulación y Análisis de Actividad Transcripcional

Hasta ahora hemos visto la base de algunas de las ómicas más importantes, como es la genómica, la transcriptómica y la epigenómica, además de exponer la importancia de la bioinformática en el análisis de las mismas y describir algunos de los métodos más utilizados, principalmente en transcriptómica. Sin embargo, cuando hablamos de una enfermedad, generalmente no se trata de una alteración en un gen concreto, sino que la gran mayoría de ellas están causadas, entre otros factores, por un complejo entramado de genes que interactúan entre si además de con otros elementos, lo que conlleva la anomalía de algunos procesos fisiológicos que pueden conducir al desarrollo de una patología. Por ello, en los últimos años ha surgido un nuevo campo de investigación en el ámbito de las ciencias ómicas centrado en conocer la red de interacciones de genes, proteínas y otros elementos que actúan en las enfermedades[66]. A este entramado de interacciones entre elementos se le llama interactoma y, en el caso del ser humano se puede agrupar en varios conjuntos: interacciones entre proteínas (PPIs), rutas metabólicas, regulación post-translacional, redes de ARN y redes de regulación génica.

Las PPIs son el conjunto de interacciones físicas que tienen lugar entre dos o más proteínas. En este concepto se engloba a todas aquellas interacciones que implican una causalidad y que participan en un evento determinado[67]. Un ejemplo de interacciones entre proteínas pueden ser la formación de estructuras proteicas formados por subunidades. Tal es la relevancia de estas interacciones que se han desarrollado gran cantidad de bases de datos donde se alojan las PPIs que se conocen. Algunas de las más conocidas son BioGrid o IntAct[68,69].

Una red metabólica es el conjunto de rutas metabólicas que ocurren en una célula u organismo y, por lo tanto, describe las interacciones que tienen las proteínas con los diferentes metabolitos, así como el producto de esta interacción. El conocimiento de las rutas metabólicas es de gran utilidad a la hora de descifrar que ocurre en un organismo o una célula. Algunas de las bases de datos que contienen información acerca de los elementos que participan en las rutas metabólicas ya las hemos descrito en el apartado anterior, ya que se trata de bases de datos como KEGG o Reactome.

La regulación post-translacional trata de las modificaciones que sufren las proteínas. Las modificaciones post-translacionales más importantes son la fosforilación, la metilación, la acetilación o la glicosilación[70]. Estas consisten en la adición de un grupo fosfato, metilo, acetilo o glúcido, respectivamente, formando una modificación covalente que modifica la estructura y función de la proteína.

Aunque cuando se habla de ARN en la mayoría de los casos nos referimos a transcritos que se asocian a la expresión de genes, no podemos obviar los distintos tipos de ARN que se han identificado. Muchos de estos tipos de ARN participan en el proceso de regulación de la expresión génica[71]. Los ARN reguladores de la expresión más conocidos y estudiados son los microARN y los ARNlnc, ambos incluidos dentro del grupo de ARNi. Los microARN son moléculas de pequeño tamaño que intervienen la degradación del ARN mensajero o la inhibición de la translación[72]. Los ARNlnc son secuencias largas, de unos 200 nucleótidos que participan a nivel de regulación de expresión, epigenética y translación[73].

Los factores de transcripción (a los que llamaremos TFs, a partir de las siglas en inglés de *Transcription Factors*) son proteínas esenciales que controlan la transcripción de los genes a través de interacciones con las secuencias específicas del ADN. Estos TFs se unen a secuencias específicas de unión de proteínas o elementos de respuesta en el ADN, también llamados motivos de unión (por la palabra *motif*, en inglés), que se localizan principalmente en zonas promotoras de los genes y pueden intervenir la transcripción de diferentes genes[74]. Las interacciones entre los TFs y sus genes diana representan las redes de regulación de la expresión génica a partir del reconocimiento de estas secuencias de unión. Una vez que se une un factor de transcripción a un elemento de respuesta, puede activar o inactivar la transcripción del gen al que está unido[75]. El hecho de unirse a secuencias específicas hace que los TFs no actúen sobre todos los genes, por lo que para conocer cómo actúan los TFs en la regulación de la transcripción es necesario conocer las relaciones entre TFs y genes diana. La colección de genes diana regulados por un mismo TF se llama regulón[76].

Para conocer que TFs intervienen en la expresión de cada gen, se han desarrollado varias técnicas, algunas de ellas basadas en evidencias experimentales y otras a partir de estimación computacional. La curación manual de interacciones de TFs y genes conlleva la búsqueda de publicaciones en los que se ha determinado esta relación de forma experimental. Hay varias bases de datos que contienen información acerca de las interacciones entre TFs y genes diana obtenidas a partir de publicaciones científicas, entre las cuales están TRRUST y ORegAnno[77,78].

TRRUST, recopila las interacciones entre TFs y genes utilizando técnicas de minería de texto con una posterior curación manual. ORegAnno contiene información sobre los TFs y sus interacciones con genes en varios organismos, incluyendo su mecanismo de acción, interacciones con otros TFs así como su estructura y función.

Otra alternativa para conocer los TFs que regulan cada gen es medir la unión de los TFs a las regiones de ADN mediante técnicas como ChIP-Seq, que ya hemos descrito anteriormente. Hay varias bases de datos y herramientas que utilizan datos de ChIP-Seq con el fin de proporcionar información acerca de los sitios de unión de diferentes proteínas al ADN, como las histonas o los TFs. Algunas de ellas son ENCODE[79] o ChEA3 (*ChIP-X Enrichment Analysis*)[80]. Los experimentos de ChIP-Seq permiten conocer los TFs u otras proteínas que se unen a regiones específicas del ADN. Sin embargo, como cada experimento de ChIP-Seq se realiza para evaluar una proteína concreta, existe cierto sesgo debido a la gran heterogeneidad de los estudios. Por ejemplo, es de esperar que los estudios que se centren en cualquier proteína relacionada con el cáncer sean muy abundantes, mientras que aquellos genes no tan estudiados presentan mucha menos información. Este fenómeno también tiene lugar por la variedad de tejidos, células o condiciones.

Se suele asumir que cada TF tiene preferencia por ciertas secuencias de unión específicas. Bajo esta premisa, se han desarrollado herramientas capaces de predecir los sitios de unión preferenciales de cada TF mediante modelos matemáticos, con el fin de construir una base de datos de interacciones entre TFs y genes. Algunas de estas herramientas son JASPAR[81], que contiene datos de una gran variedad de organismos, tanto plantas como animales, u HOCOMOCO que es específico de humano y ratón[82].

Finalmente, la última técnica que se suele utilizar para conocer los TFs que actúan sobre cada gen es mediante ingeniería inversa a partir de la expresión de los genes y la de los TFs. Los TFs son proteínas y, por tanto, su biosíntesis no difiere de la del resto de proteínas. Es decir, siguen el proceso de transcripción y traducción. Esta técnica asume que debe existir una correlación entre los niveles de expresión de un TF y sus genes diana, midiendo la expresión mediante técnicas transcripcionales. De entre las herramientas que utilizan esta técnica destaca ARACNE (*Algorithm for the Reconstruction of Accurate Cellular Networks*)[83]. ARACNE es un método para inferir redes de regulación génica a partir de datos de expresión génica. Brevemente, el proceso de búsqueda de TFs en ARACNE se realiza en dos pasos. En primer lugar, se calcula una matriz de correlación entre los niveles de expresión de todos los genes del conjunto de

datos. Posteriormente, se utiliza esta matriz de correlación para calcular una matriz de dependencia, que indica la fuerza y dirección de la relación entre los genes. En segundo lugar, se aplica un test de independencia estadística, basado en la información mutua, entre cada par de genes. El resultado de este test es una matriz de dependencia condicional, que indica la dependencia entre cada par de genes, dada la expresión de todos los demás genes. A continuación, se utiliza esta matriz de dependencia condicional para inferir las relaciones de regulación entre los TFs y los genes regulados. En particular, se utiliza un algoritmo llamado *Context Likelihood of Relatedness* (CLR) para identificar los pares de TF y gen que tienen una dependencia condicional significativa, indicando una relación de regulación.

1.2.1 MÉTODOS DE INFERENCIA DE ACTIVIDAD TRANSCRIPCIONAL

Los métodos existentes para medir la abundancia proteica basados en tecnologías de *microarrays*[84] o espectrometría de masas[85] siguen siendo laboriosos, costosos, y cubren sólo una pequeña fracción del panorama proteómico o requieren grandes cantidades de tejido. Además, estos métodos proporcionan sólo una medida indirecta de la actividad proteica, ya que ésta está determinada por una compleja cascada de eventos, incluyendo la síntesis, degradación, modificación post-translacional, formación de complejos y localización subcelular de las proteínas. La evaluación de la actividad de los TFs, por tanto, presentan las mismas inconveniencias por lo que hasta ahora la mejor opción consiste en inferir dicha actividad en base a los niveles de expresión de sus genes diana, ya que los datos de expresión génica se utilizan asiduamente en la investigación de muchas patologías.

Hay varias herramientas que evalúan la actividad de las proteínas en base a la expresión de los genes que estas proteínas están regulando, como sería el caso de los TFs. Algunas de estas herramientas son: VIPER (*Virtual Inference of Protein activity by Enriched Regulon*)[86], NCA (*Network Component Analysis*)[87], AUCell[88] o más recientemente NetAct[89]. En el desarrollo de esta tesis hemos utilizado la metodología desarrollada en VIPER a través de un paquete de Bioconductor con el mismo nombre, por lo que vamos a describir en que consiste.

VIPER aplica un enriquecimiento de regulones a partir de datos de expresión, es decir, trata de inferir la actividad de una serie de proteínas a partir de la expresión de los genes sobre los que actúa, cómo es el caso de los TFs. Hay dos formas alternativas de aplicar este método, la primera

de ellas utilizando los resultados de expresión diferencial entre las condiciones y la otra a partir de la matriz de expresión. Debido a que queríamos aplicar técnicas similares a un análisis de expresión diferencial, decidimos optar por generar una matriz de actividades de los TFs a partir de la matriz de expresión, mediante el método denominado "de muestra única". Este método compara los niveles de expresión de cada muestra con los niveles de expresión a lo largo de todas las muestras. El método de enriquecimiento de los regulones se hace utilizando implementaciones del algoritmo aREA (*analytic rank-based enrichment analysis*), un método estadístico basado en rangos que calcula un valor de enriquecimiento usado como símil de la actividad de la proteína. Brevemente, este método aplica un método similar a GSEA, el cual utiliza los genes ordenados por el nivel de expresión para calcular un valor de enriquecimiento, utilizando los genes diana de cada TF como conjunto de genes sobre el que calcular dicho valor. De este modo, cuando un TF está activando a varios genes que, en una muestra, están especialmente expresados o muy expresados, la actividad de este TF será alta, ya que se infiere que si la expresión de los genes diana es alta implica que este TF está muy activado. Como no todos los TFs actúan activando genes, sino que los hay que inhiben genes, el software tiene en cuenta el modo de regulación (MoR), de forma que los genes que están inhibiendo TFs con expresión alta también están muy activados.

1.2.2 Bases de Datos de TFs-Genes Diana

Hemos descrito previamente cuatro estrategias diferentes para determinar la relación entre los TFs y los genes diana. Hace unos años se desarrolló una base de datos, DoRothEA (*Discriminant Regulon Expression Analysis*) que incluye interacciones de TFs y genes diana de estas cuatro estrategias[76]. Esta base de datos no solo contiene información relativa a cada una de ellas, sino que clasifica las estrategias utilizadas para identificar cada interacción en base a diferentes niveles de evidencia (Figura 8a). Estos niveles de evidencia, ordenados de mayor a menor nivel de confianza, son:

- Interacciones obtenidas a partir de la curación manual de artículos (incluyendo bases de datos como TRRUST y ORegAnno, entre otras).

- Experimentos de ChIP-Seq procedentes de ReMap[90] (que incluye los datos incorporados en ENCODE y otros repositorios públicos).

- Prediccion de secuencias de unión a TFs, también llamados TFBS (del inglés, *Transcription Factor Binding Site*) basados en secuencias de los promotores (usando bases de datos como Hocomoco o JASPAR).

- Inferencia a partir de datos de expresión de GTEx utilizando ARACNE.

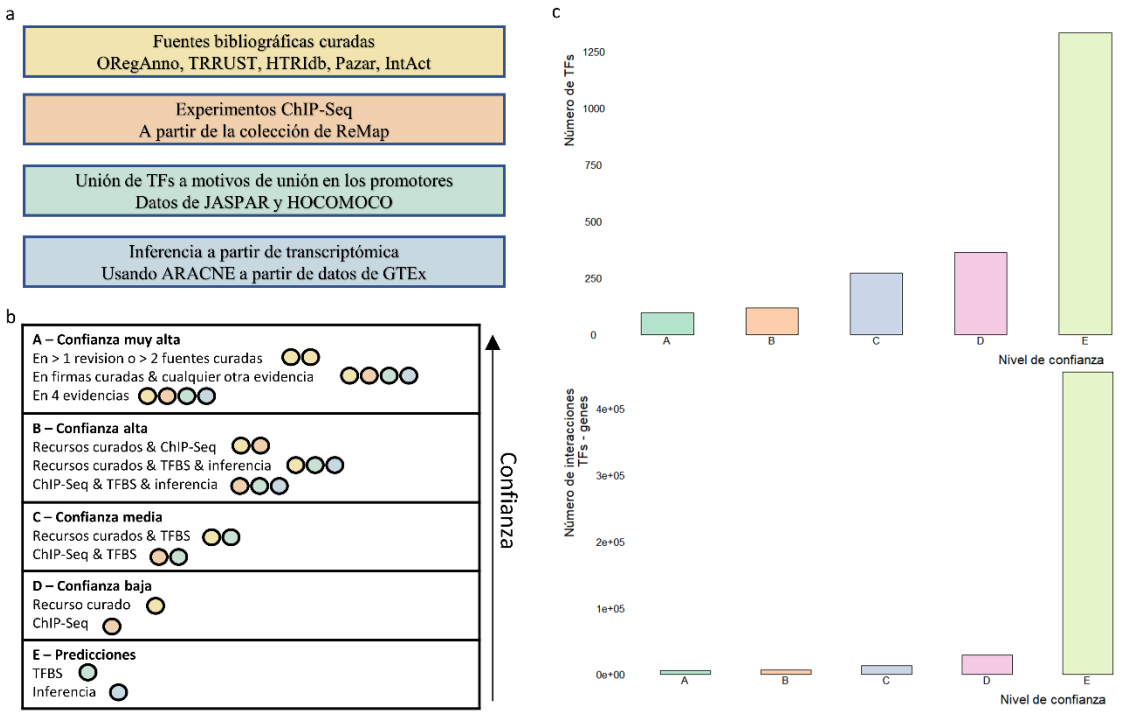

Figura 8. Descripción de la información disponible en DoRothEA. a) Los distintos niveles de evidencia caracterizados en DoRothEA, de arriba abajo clasificados en cuanto a nivel de confianza siguiendo la descripción de estas fuentes de datos disponibles en el artículo de DoRothEA[76]. b) Clasificación de los interactomas disponibles en DoRothEA, clasificados de arriba abajo en base al nivel de confianza de los mismos. Dentro de cada celda se describen los criterios necesarios para clasificar cada regulón en un nivel. Cada círculo indica una fuente de información distinta, siguiendo la gama de colores utilizada en la Figura 8b. c) Número de TFs (arriba) e interacciones de TFs y genes (abajo) según los diferentes niveles de confianza. Esta figura se ha construido con la información disponible en el paquete de Bioconductor de DoRothEA, con los datos de humano.

Además, DoRothEA utiliza estos niveles de evidencia para clasificar las interacciones de TFs y genes diana en 5 interactomas distintos estableciendo una serie de criterios para cada uno de ellos. Estos 5 interactomas engloban todas las interacciones conocidas e incluidas en DoRothEA, desde aquellas que presentan un nivel de confianza muy alto (basado principalmente en evidencia experimental), hasta las que son predicciones computacionales, pasando por interacciones de diferente nivel de confianza. Los criterios de clasificación de los regulones se pueden visualizar en la Figura 8b. Como es de esperar, los interactomas con las interacciones de mayor calidad son también los más pequeños, ya que las interacciones

clasificadas en los niveles superiores también se incluyen en las inferiores (Figura 8c). Esta base de datos se ha extendido en los últimos años, añadiendo tanto datos de ratón como datos de *single-cell*[91,92].

1.3 Repositorios de Datos Públicos

El auge de las ómicas, junto con su abaratamiento, se ha visto reflejado en la acumulación de una gran cantidad de datos. Pero este aumento de información plantea un problema: ¿dónde guardarlos y cómo acceder a ellos? Afortunadamente, desde finales del siglo XX han surgido plataformas públicas en línea que permiten almacenar y compartir estos datos con otros investigadores. De esta manera, cualquier persona puede tener acceso a información generada en otras partes del mundo. Estas plataformas son conocidas como repositorios públicos de datos ómicos y pueden incluir secuencias de ADN, proteínas, datos de expresión u otros tipos de información relevante.

Hay muchas razones para utilizar los repositorios públicos de datos ómicos. Además de permitir el almacenamiento y compartición de datos, estos repositorios también pueden ser utilizados por otros investigadores para reutilizar la información y reducir el esfuerzo de generación de nuevos datos similares[93]. También se pueden aplicar técnicas de meta-análisis, gracias a la disponibilidad de una gran cantidad de información de diferentes proyectos, para encontrar patrones robustos entre estudios[94–96]. Además, los repositorios públicos pueden ser utilizados para desarrollar y mejorar técnicas, utilizando los datos de uno o varios estudios como base (lo que se conoce como benchmarking)[76,86,97,98].

Aunque existen numerosos repositorios de datos públicos, como INSDC (compuesto por GenBank, ENA y DDBJ), ENCODE o SRA, vamos a detallar los que han tenido especial relevancia en el desarrollo de esta tesis: NCBI GEO, ENA y Expression Atlas.

1.3.1 Gene Expression Omnibus (GEO)

NCBI GEO (*Gene Expression Omnibus*) es un repositorio de datos ómicos publicado por el *National Center for Biotechnology Information* (NCBI), que forma parte de los *National Institutes of Health* (NIH) de los Estados Unidos. NCBI GEO se lanzó en el año 2000 con el fin de almacenar y dar accesibilidad a la comunidad científica a múltiples datos de expresión, principalmente obtenidos a partir de *microarrays* de expresión[99]. Sin embargo, la aparición de las técnicas de secuenciación NGS hicieron que NCBI GEO incorporara algunos matices en cuanto a la información almacenada, ya que, si bien es cierto que se siguen incorporando

estudios generados por *microarrays* de expresión, se ha incrementado la incorporación de estudios de fuentes ómicas muy distintas. La mayoría de los datos alojados en NCBI GEO pertenecen a estudios transcripcionales, bien sea de *microarrays* de expresión o de RNA-Seq, pero también almacena datos de metilación (tanto *microarrays* como secuenciación con bisulfito), de ChIP-Seq, y datos de variantes genómicas (tanto a partir de secuenciación como de *microarrays* de SNPs de ADN), de múltiples organismos. El hecho de que en muchas revistas científicas sea de obligatoriedad el acceso público a los datos ómicos, ha dado lugar al crecimiento exponencial de la cantidad de información alojada en NCBI GEO.

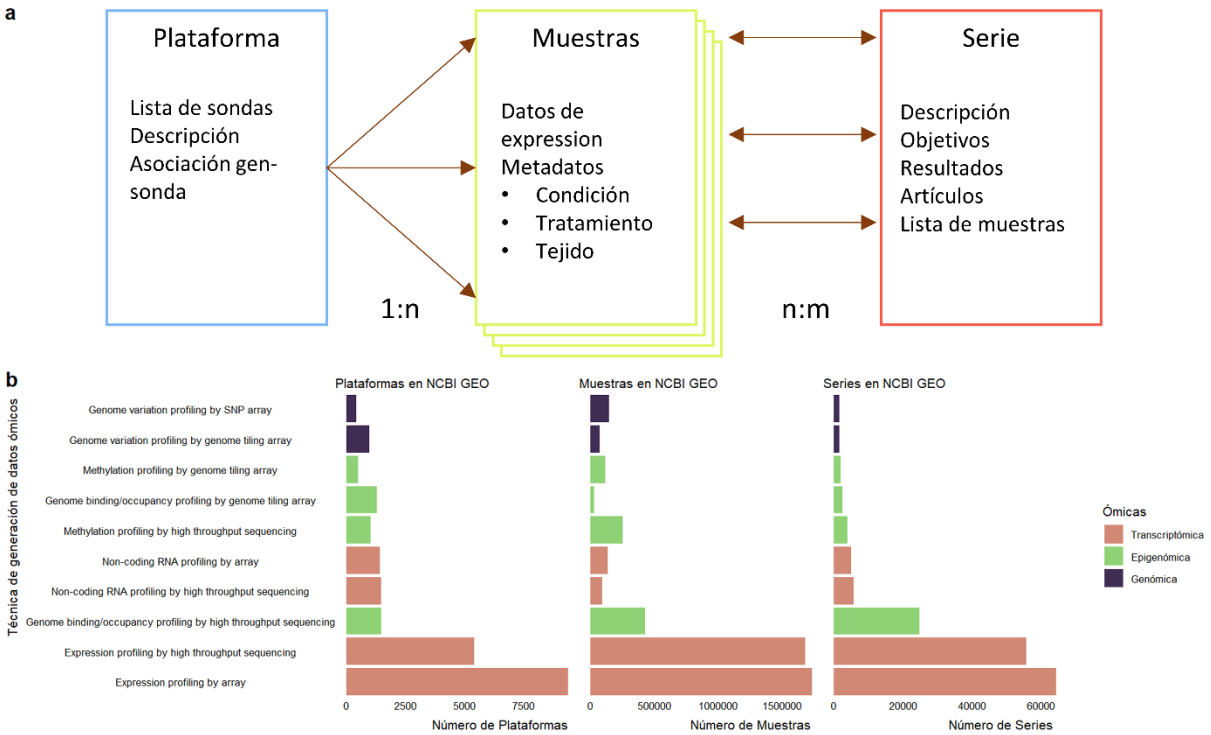

Figura 9. Estructura y composición de NCBI GEO. a) Esquema en el que se indica la relación entre las distintas entidades de NCBI GEO; la relación entre Plataformas y Muestras es de tipo 1 a muchas (1:n) mientras que en las Muestras y Series la asociación es muchas a muchas (n:m). b) Volumen de Plataformas, Muestras y Series en NCBI GEO, clasificadas por el método de obtención de estos datos, en el eje de ordenadas. El color indica el tipo de ciencia ómica al que pertenece cada método de generación de datos; por ejemplo, *Expression profiling by array* (*microarrays*) y *Expression profiling by high throughtput sequencing* (RNA-Seq) son datos transcriptómicos y son los predominantes en NCBI GEO.

Para trabajar con NCBI GEO es muy importante conocer la estructura de los datos, ya que es algo compleja. NCBI GEO distribuye la información en tres elementos principales: Plataformas, Series y Muestras. Cada uno de estos elementos hacen referencia a entidades complejas y diferentes entre sí, pero todas están relacionadas entre ellas (véase Figura 9a). Las Plataformas reflejan la tecnología que se ha utilizado para generar los datos de un estudio, ya

sean *microarrays* de ADN, secuenciadores de RNA-Seq o *microarray*s de metilación, entre otras. Cada Plataforma tiene un código identificativo único que sigue el patrón de GPLXXXX, donde XXXX es un código numérico. Actualmente (febrero 2023) hay casi 25,000 plataformas recogidas en este repositorio. Algunas de las plataformas más utilizas en NCBI GEO son Illumina HiSeq 2500 (GPL16791), Illumina NextSeq 500 (GPL18573) o Illumina NovaSeq 6000 (GPL24247) con las que se han generado alrededor de 300,000 muestras con cada una de ellas. Para volver a resaltar la relevancia de las técnicas de secuenciación por encima de los *microarrays* de ADN, de las 10 plataformas más usadas solo una, [HG-U133_Plus_2] *Affymetrix Human Genome U133 Plus 2.0 Array* (GPL570), es una plataforma de microarrays, el resto se trata de técnicas de secuenciación.

El segundo elemento que encontramos en la estructura de NCBI GEO son las Series, también llamados *datasets*. Cada Serie refleja un estudio concreto, incluyendo el conjunto de muestras que se ha utilizado en el mismo, así como la plataforma o plataformas que se han utilizado para generar los datos. Cuando buscamos un conjunto de datos en NCBI GEO generalmente lo hacemos buscando las Series, ya que incorporan detalles del estudio que se ha realizado, incluyendo tipo de datos, origen, objetivo o tratamiento de las muestras, entre otras. En la actualidad hay casi 194,000 Series alojadas en NCBI GEO. A partir de una Serie podemos acceder a la información de todas las muestras asociadas a la misma. De igual modo que las plataformas, cada Serie se caracteriza por un código único, salvo que el patrón en este caso es GSEXXXX.

Finalmente, la unidad básica sobre la que se estructura NCBI GEO son las muestras, ya que son las que contienen tanto la información obtenida a partir de las técnicas ómicas como las características particulares de la muestra. Generalmente al conjunto de características de una muestra se le denomina metadatos, y varían mucho entre los diferentes estudios, pero deben indicar información acerca del procesamiento de la muestra y características esenciales (como si es caso o control, si pertenece a un fenotipo concreto o el tejido o conjunto celular del que se ha obtenido) y pueden complementar esta información con características clínicas de la muestra (edad, género o raza). Cada muestra se ha generado con una plataforma única, pero puede pertenecer a más de una Serie. Se identifican mediante un código único que sigue el patrón GSMXXXX y actualmente en NCBI GEO hay más de 5,500,000 muestras. Hemos resumido el volumen de datos, separando por ciencias ómicas y diferentes técnicas de generación de datos en la Figura 9b.

Aunque trabajar con NCBI GEO requiere conocer esta estructura, existen herramientas que facilitan su uso. La más importante es el propio buscador de NCBI GEO, que permite hacer búsquedas generales o avanzadas, aplicando varios filtros como seleccionar los organismos, el tipo de dato y una serie de palabras clave que queramos incluir. Por ejemplo, si quisiéramos buscar las Series que tienen datos de RNA-Seq de ratón que estén relacionadas con el cáncer de pulmón, la consulta sería la siguiente: '("lung neoplasms"[MeSH Terms] OR lung cancer[All Fields]) AND "Homo sapiens"[porgn] AND "Expression profiling by high throughput sequencing"[Filter]'.

Otra herramienta muy utilizada para trabajar con NCBI GEO es la librería de R/Bioconductor GEOquery[100]. Esta librería hace de puente entre nuestra sesión de R y la web de NCBI GEO, permitiendo acceder a la información relativa a Plataformas, Series o Muestras simplemente escribiendo el código del cual se desea tener dicha información. En el caso de los *microarrays* de ADN permite descargar la matriz de expresión con los valores de expresión de todas las muestras de una Serie concreta.

1.3.2 EUROPEAN NUCLEOTIDE ARCHIVE

El *European Nucleotide Archive* (ENA) es un repositorio público de secuenciación de datos transcriptómicos y genómicos de una amplia gama de organismos, incluyendo humano y tanto animales como plantas modelo. Se trata de una iniciativa de *European Molecular Biology Laboratory* (EMBL), lanzada en 1999 y es mantenida por el *European Bioinformatics Institute* (EBI)[101].

Este repositorio se compone de una gran cantidad de datos de secuenciación, incluyendo genomas y transcriptomas completos, así como varios estudios y proyectos que contienen datos crudos de secuenciación de RNA-Seq y genómicos. Cómo es de esperar, la estructura de los datos de ENA es muy compleja y altamente organizada, ya que cada registro se identifica con un código único. Cuando queremos acceder a cualquier estudio debemos tener en cuenta que el patrón de dicho código nos va a indicar el tipo de dato al que hace referencia. De este modo, distinguimos entre Proyectos, Muestras, Experimentos y Carreras.

Los Proyectos hacen referencia a un conjunto de muestras de un estudio concreto, serían el símil de las Series en NCBI GEO, ya que contienen la misma información detallada de dicho estudio. Tienen un código que sigue el patrón PRJNAXXXX. Sin embargo, el resto de los elementos se utilizan para identificar una particularidad de una muestra concreta. Mientras el identificador de Muestra que detalla algunos metadatos de la muestra, cuyo código es SAMNXXXX, el identificador de Experimento se utiliza para describir la tecnología utilizada para generar la muestra (código SRXXXXX) y el identificador de carrera es en donde se incluyen los datos de secuenciación generados (código SRRXXXX). De este modo, por ejemplo, el Proyecto PRJNA871765 tiene un total de 30 muestras, donde cada una de ellas tiene un código único de Muestra, de Experimento y de Carrera (por ejemplo, SAMN30428146, SRX17159478, SRR21147739).

Del mismo modo que con NCBI GEO, la herramienta principal para buscar estudios y datos ómicos es la propia interfaz web de ENA, aunque dispone de una API que permite hacer búsquedas programáticamente. Es importante saber que los datos de secuenciación de las Series de NCBI GEO raramente se encuentran en NCBI GEO, sino que se añaden a ENA, por lo cual es necesario conocer el código ENA del proyecto y las muestras de NCBI GEO para descargar estos datos.

1.3.3 EXPRESSION ATLAS

Expression Atlas es una base de datos de acceso público que proporciona información sobre la expresión de genes en diferentes tejidos y condiciones biológicas en una variedad de organismos, desde bacterias hasta humanos[102]. La base de datos es mantenida por el Instituto Europeo de Bioinformática (EBI) y contiene información de más de 30.000 experimentos de expresión de genes.

La principal función de Expression Atlas es proporcionar una visión general de la expresión de genes en diferentes tejidos y condiciones biológicas. Los datos de expresión de genes se recopilan a partir de una variedad de fuentes, como *microarrays* y RNA-Seq, y se normalizan y analizan utilizando técnicas estadísticas avanzadas y estandarizadas. La información resultante se muestra en forma de mapas de calor, gráficos y tablas que muestran la expresión de genes en diferentes condiciones y tejidos.

Los usuarios de Expression Atlas pueden buscar información sobre la expresión de genes en diferentes organismos, diferentes tejidos y tipos celulares, además de condiciones biológicas. Además, a partir del año 2020 se incluyeron datos de expresión de tipos celulares obtenidos con experimentos de scRNA-Seq.

1.4 Conceptos Básicos de Inmunidad y Enfermedades Autoinmunes

La inmunología es la ciencia que estudia el sistema inmune. El sistema inmune se encarga de proteger al individuo frente a infecciones y agentes externos potencialmente dañinos para el organismo, además de células anormales que se generen en el interior del organismo, como las células cancerosas. Por tanto, es capaz de contrarrestar tantos elementos externos a cuerpo (virus, bacterias u hongos) como fallos internos (enfermedades o células cancerosas). El sistema inmune está compuesto por un complejo y heterogéneo conjunto de elementos localizados en órganos, tejidos y células muy variados que se suele dividir en dos sistemas de acuerdo a la especificidad y el modo de actuar de los mismos: el sistema inmune innato y el sistema inmune adaptativo.

La inmunidad innata (también llamada natural o inespecífica) constituye la primera línea de defensa a la entrada de agentes extraños en el cuerpo del individuo. Se caracteriza porque es muy poco específica, ya que en la mayoría de los casos no es capaz de identificar al patógeno, aunque contiene una serie de receptores que reconocen regiones muy conservadas de la mayoría de los patógenos[103]. Se trata de una respuesta rápida que carece de memoria, elementos diferenciadores con la inmunidad adaptativa, de la cual se hablará posteriormente. Los componentes de la inmunidad innata son dos: barreras naturales y las células de defensa.

Las barreras naturales, también llamadas de superficie son una serie de barreras físicas y respuestas químicas que se encargan de que las sustancias u organismos externos no penetren en nuestro cuerpo. Están formadas por un conjunto de órganos, tejidos y diferentes fluidos producidos por nuestro organismo. Entre estas barreras de superficie destacan la piel, que actúa como barrera física, las membranas mucosas, localizadas en orificios de entrada como nariz y boca y se encargan de secretar enzimas para degradar agentes externos, las vías respiratorias, con la secreción y posterior expulsión de moco o el tracto gastrointestinal, con varios mecanismos de acción como la secreción de bilis o ácido gástrico entre otros.

Los principales componentes de la inmunidad innata son las células inmunes de defensa innatas. Entre ellas destacan las células fagocíticas (neutrófilos y monocitos/macrófagos) y los linfocitos citolíticos naturales (también conocidos como células NK, del inglés *natural killer*). Tanto los neutrófilos como los monocitos, que se convierten en macrófagos cuando llegan al tejido dañado, actúan fagocitando a los agentes invasores o las células dañadas o muertas y producen

una serie de sustancias químicas que sirven como mensaje de reclutamiento para otras células del sistema inmune[104].

Por otro lado, las células NK se encargan también de destruir bacterias y células, pero no mediante fagocitosis sino provocando la citólisis de su membrana plasmática. La respuesta inmune innata siempre comienza con un proceso protector denominado inflamación, cuya función principal es aislar la célula dañada o el agente externo para que no se propague por el resto del cuerpo. Durante el proceso de inflamación se generan una serie de sustancias que participan en el reclutamiento de todas las células inmunes necesarias[105].

Cuando la inmunidad innata no es suficiente para contrarrestar la entrada de agentes extraños, se activa el sistema inmune adaptativo o inmunidad adquirida. Este tipo de inmunidad es capaz de reconocer un gran número de sustancias externas y reaccionar frente a ellas, por lo que es muy específica[106]. Además, tiene la capacidad de "memorizar" la respuesta que se da al mismo microbio cuando se expone de forma repetida al mismo. Consiste principalmente en la capacidad que tienen una serie de células inmunes, los linfocitos, de reconocer los antígenos que presentan los diferentes microbios, sustancias tóxicas o células anormales que se encuentran en el organismo. Por tanto, los actores principales de la inmunidad adaptativa son los linfocitos, los anticuerpos, que son moléculas capaces de identificar de forma específica a los antígenos y las células dendríticas.

Los linfocitos son un grupo de leucocitos que se encargan de llevar a cabo la respuesta inmune adaptativa. Se diferencian principalmente en dos tipos, linfocitos B (o células B) y linfocitos T (o células T). Cada tipo de linfocito da un tipo de respuesta diferente, mientras las células B se encargan de la respuesta humoral mediante la producción de anticuerpos, los linfocitos T participan en la respuesta celular liberando toxinas frente a antígenos específicos.

Los linfocitos B, que se producen en la médula ósea y maduran en el bazo, se encargan de producir anticuerpos específicos para cada antígeno. Cada linfocito B está programado para producir un solo tipo de anticuerpo, pero están preparados para, en caso de que un linfocito B detecte el antígeno específico de su anticuerpo, llevara a cabo un fenómeno llamado expansión clonal, mediante el cual este linfocito se divide y diferencia hasta dar un clon de células plasmáticas, que fabrican y producen grandes cantidades del anticuerpo en cuestión. Además, este mismo linfocito B original genera otra línea de descendientes que desemboca en un tipo de células B llamadas de memoria, que son linfocitos que pueden vivir durante muchos años y que

sirven como reservorio de anticuerpos. De este modo, si el organismo volviese a enfrentarse al mismo antígeno, no es necesario generar al azar un anticuerpo específico, pues ya existe en estas células. Por tanto, la respuesta humoral consiste en la identificación del antígeno del elemento anormal por medio de las células B, que lo reconocen y fabrican gran cantidad de anticuerpos específicos mediante los cuales marcan a este elemento anómalo para que pueda ser destruido por otros elementos del sistema inmune.

Por otro lado, los linfocitos T, producidos en la médula ósea y madurados en el timo, llevan a cabo la inmunidad celular, que tiene lugar en el interior de las células. Este fenómeno tiene lugar cuando las células T atacan tejidos infectados directamente. En este caso, los linfocitos T reconocen al antígeno siempre que haya sido procesado previamente. En primer lugar, las células dendríticas y otras células presentadoras de antígenos profesionales se encargan de degradar el antígeno a péptidos que son reconocidos por los receptores de los linfocitos T. Posteriormente el linfocito T se divide y se especializa en varios subtipos de células T: células T citotóxicas, que se encargan de destruir las células infectadas mediante la inyección de enzimas catalizadoras y células T colaboradoras (o helper), que se encargan de activar a los linfocitos B, los linfocitos T citotóxicos y macrófagos. De forma similar a los linfocitos B, los linfocitos T también sufren un proceso de expansión clonal mediante el cual se generan más linfocitos citotóxicos y colaboradores además de linfocitos T de memoria que se almacenan como reservorio.

Las células presentadoras de antígenos, entre las que destacan las células dendríticas son las encargadas de procesar y presentar los antígenos para que el sistema inmune sea capaz de reconocerlos e iniciar una respuesta inmune adaptativa.

Un agente de gran relevancia y que se encarga de comunicar la respuesta inmune innata y adaptativa son las citoquinas. Las citoquinas son moléculas secretadas por las distintas células inmunes cuyas funciones principales son la activación de las funciones efectoras de los fagocitos y linfocitos, la migración de células inmunitarias hacia los tejidos donde son necesarios e intervienen en la diferenciación de las células inmunitarias. Hay diferentes tipos de citoquinas: interleucinas (IL), interferones (IFN), factor de necrosis tumoral (FNT). De forma general, las interleucinas se encargan de regular la diferenciación de los distintos subtipos celulares, los interferones actúan en la respuesta inmune innata frente a virus principalmente, promoviendo la actividad antiviral, por lo que activan las células NK, mientras los factores de necrosis tumoral tienen actividad proinflamatoria por lo que intervienen en la inflamación.

Existen múltiples patologías relacionadas con deficiencias genéticas que provocan perturbaciones en el sistema inmune, como la inmunodeficiencia, o la incapacidad de hacer frente a patógenos debido a la pérdida de capacidad de llevar a cabo una respuesta inmune eficiente, la hipersensibilidad o respuesta exagerada del sistema inmune frente a un patógeno y la autoinmunidad, que es provocada cuando el sistema inmune tiene respuesta contra componentes del propio organismo[107]. Cuando una respuesta autoinmune se vuelve crónica pasa a denominarse enfermedad autoinmune.

Este tipo de patologías pueden afectar a un órgano concreto, como la diabetes tipo I (T1D) o manifestarse de forma sistémica en varias partes del organismo, en las que nos centraremos en el siguiente párrafo. Las enfermedades autoinmunes suelen ser consideradas patologías raras, pero completan un conjunto de más de 100 afecciones que podrían afectar hasta al 5% de la población, principalmente predominantes en mujeres con una ratio de 2 a 1 con respecto a hombres[108]. De hecho, el conjunto de enfermedades autoinmunes son la sexta o séptima causa más frecuente de muerte entre mujeres[109].

Las enfermedades autoinmunes sistémicas son un conjunto de trastornos que pueden afectar a varios órganos y tejidos del cuerpo, como los riñones, el corazón, la piel o las articulaciones. Entre las enfermedades autoinmunes sistémicas más importantes se encuentran el lupus eritematoso sistémico (LES o SLE), la artritis reumatoide (AR), la esclerosis sistémica (SSc), el síndrome de Sjogren (SjS), la enfermedad mixta del tejido conectivo (MCTD) y el síndrome antifosfolípido (PAPS). Aunque todas ellas están caracterizadas clínicamente, este diagnóstico no es simple ya que a menudo comparten síntomas. Por ejemplo, algunos pacientes con LES también presentan deformidades en las articulaciones, una manifestación típica de la AR[110]. De manera similar, los pacientes con MCTD pueden experimentar síntomas que se observan en LES, AR o SSc[111,112]. Asimismo, algunos pacientes con LES o AR también pueden desarrollar SjS, mientras que hay personas con SjS que no tienen síntomas de AR ni LES[113].

En este sentido, un esfuerzo europeo impulsado por el proyecto PRECISESADS ha desembocado en una clasificación de pacientes de múltiples enfermedades autoinmunes sistémicas a partir de la integración de datos ómicos, específicamente transcriptómica y epigenómica, esta última a partir de datos de metilación, junto a la información clínica de los pacientes incluidos en este estudio. Este trabajo, publicado recientemente ha desarrollado una clasificación mediante técnicas de agrupamiento a partir de la integración de datos de más de 950 pacientes de 7 enfermedades autoinmunes y más de 260 controles sanos. En la clasificación

se identificaron cuatro grupos, que, además de presentar diferencias moleculares, mostraban diferencias clínicas[114].

Puesto que la tesis se centra en la aplicación de métodos de inferencia en LES, vamos a poner el foco en la descripción de lo que se sabe y lo que se ha hecho en cuanto a la aplicación de ciencias ómicas en LES.

1.5 DATOS ÓMICOS EN LUPUS ERITEMATOSO SISTÉMICO

El lupus eritematoso sistémico (LES) es una enfermedad autoinmune sistémica con una gran prevalencia en mujeres con respecto a hombres, con una ratio de 9 a 1[115] y que actúa de forma diferente en pacientes adultos y pediátricos, siendo más grave en estos últimos[116]. Los mecanismos moleculares que han sido más ampliamente asociados a la patología son los niveles elevados de interferones de tipo I circulante[117] así como el exceso de anticuerpos antinucleares (ANAs)[118]. Además, esta patología se caracteriza por períodos fluctuantes de la actividad de la enfermedad, con períodos de remisión, actividad leve o inexistente y periodos de alta actividad llamados brotes. Los motivos o causas por las que tienen lugar estas fluctuaciones son aún un gran misterio, y la progresión de la enfermedad es diferente entre diferentes pacientes.

Actualmente existen varios criterios de clasificación de pacientes de LES, entre los que destacan el criterio del *American College of Rheumatology* (ACR). Existen también varios criterios de medición del grado de daño tisular como el *Systemic Lupus International Collaborating Clinics* (SLICC). Ambos índices se basan en criterios clínicos e inmunológicos[119]. También hay varios índices que se utilizan para caracterizar la actividad de la enfermedad, principalmente mediante formularios y criterios clínicos que conllevan una serie de puntuaciones que indiquen si el paciente se encuentra en un período de alta o baja actividad de la enfermedad. De todos ellos, el más importante es el grupo de índices conocidos como SLEDAI (*Systemic Lupus Erythematosus Disease Activity Index*), con varias versiones: SELENA-SLEDAI, SLEDAI-2000, MEX-SLEDAI o SLEDAI-2K[120].

Existen muy pocos fármacos aprobados para ser utilizados en LES, desde 1955 sólo estaban aprobados por la *Food and Drug Administration* (FDA): la aspirina, los corticosteroides, principalmente prednisona, y la hidroxicloroquina. En 2011 se aprobó un nuevo medicamento, belimumab[121], que actúa uniéndose a la proteína BAFF, una molécula importante en la diferenciación de los linfocitos B a células productoras de anticuerpos o células plasmáticas. Mediante esta unión se inhibe la supervivencia de las células B, incluyendo aquellas que son autorreactivas, y limita la diferenciación de linfocitos B a células productoras de anticuerpos.. Muy recientemente, en agosto de 2021, se ha aprobado el uso Anifrolumab-fnia en pacientes de LES. Este fármaco actúa inhibiendo los receptores de IFNα e IFNβ (IFNAR), por lo que modera los niveles de IFN de tipo I, que es una de las rutas moleculares más importantes en esta enfermedad[122]. El principal motivo de la escasez de terapias usadas en LES es la

heterogeneidad de los mecanismos moleculares desregulados entre los pacientes[123], que causa que la eficiencia de los fármacos sea muy variable entre pacientes.

Debido a la dificultad que implica tanto el diagnóstico, el tratamiento y conocer las causas que provocan la aparición de las enfermedades autoinmunes, desde hace varios años se ha promovido el estudio de estas utilizando diferentes ómicas. La aplicación de las ómicas en estudio de LES ha determinado que los factores que influyen en la aparición de la enfermedad son genéticos, epigenéticos y medioambientales[124].

1.5.1 TRANSCRIPTÓMICA EN LES

La ciencia ómica que, hasta ahora, se ha utilizado más para describir esta enfermedad es la transcriptómica. Desde 2003, año en el que dos grupos de investigación independientes describieron la presencia de una firma transcripcional comparando muestras de células mononucleares de sangre periférica (PBMCs, del inglés, *Peripheral Blood Mononuclear Cells*) de pacientes de LES y controles sanos, se conoce que existe gran cantidad de genes estimulados por el IFN tipo I que se encuentran sobreexpresados en los pacientes de LES[125,126]. Este conjunto de genes, denominado genes estimulados por interferón (ISGs, del inglés, *Interferon Stimulated Genes*) se ha convertido desde entonces en uno de los aspectos claves del estudio de esta patología. Esta misma firma se replicó en un estudio que realizó un análisis de expresión diferencial entre diferentes tipos celulares concretos (células B, células T y células mieloides) entre pacientes de LES y controles sanos[127]. Más recientemente, en un estudio de scRNA-Seq se demostró que la firma de ISGs deriva de un pequeño número de subpoblaciones celulares incluidas entre las células inmunes más importantes, cuya expansión en LES activo provoca este aumento de expresión[128]. Como se ha comentado previamente, la reciente aprobación de Anifrolumab-fnia como mecanismo de reducción de niveles de IFN tipo I, es uno de los puntos culminantes del estudio de la sobreexpresión de la firma de IFN a lo largo de todos estos años[122].

La transcriptómica también se ha utilizado para tratar de asociar la expresión de los genes con la actividad de la enfermedad, sin embargo, existe cierta controversia con respecto a esto, ya que hay publicaciones en las que se describe la existencia de esta asociación y otras en las que no se encuentra asociación[129,130]. Además, varios estudios han utilizado técnicas de *clustering*, usando datos transcriptómicos, con el fin de determinar grupos de pacientes de LES más homogéneos desde el punto de vista molecular y validados entre diferentes estudios[131,132]. También se han llevado a cabo estudios de expresión en ARN no codificante, principalmente

en microARNs, localizando una serie de biomarcadores que cumplen un papel importante en la enfermedad[133,134], algunos de ellos, como el miR-146a, asociados a la regulación de la expresión de genes implicados en las rutas relacionados con IFN[135].

1.5.2 GENÉTICA DEL LES

Mediante el estudio de datos genómicos, se han localizado más de 100 variantes asociadas a los pacientes de LES entre los que destacan las localizadas en los genes IRF5, STAT4, ITGAM y en la región del Complejo Mayor de Histocompatibilidad conocidas como Antígeno Humano Leucocitario o *Human Leukocyte Antigen* (HLA)[136]. Además, algunas de estas variantes se encuentran en los sitios de unión de los TFs que regulan estos genes[137].

Las alteraciones epigenéticas también se han estudiado en profundidad en pacientes de LES. Debido a la escasa susceptibilidad que confiere la herencia, se cree que los cambios epigenéticos pueden ser clave para conocer el origen de la enfermedad. Por ejemplo, existe un bajo porcentaje de concordancia entre gemelos monocigóticos[138], lo que apunta a posibles efectos no genéticos que pueden estar involucrados en el desarrollo de la enfermedad. En este aspecto, se llevó a cabo un análisis donde hallaron hasta 49 regiones hipometiladas al comparar gemelos monocigóticos con LES y sanos[139]. También se han encontrado regiones hipometiladas en muchos genes relacionados con IFN tipo I como MX1, IFI44L, PARP9, DTX3L, IFIT1, IFI44, RSAD2, PLSCR1 y IRF7 que registraron menor metilación en pacientes de LES con respecto a controles[140].

2 OBJETIVOS

En esta tesis se han aplicado diferentes métodos de inferencia de actividad para determinar distintas redes de regulación génica a partir de datos transcripcionales en el contexto de una enfermedad autoinmune como es el lupus eritematoso sistémico.

- Desarrollo de una base de datos de datos ómicos, transcripcionales y epigenómicos, de varias enfermedades autoinmunes a partir de la disponibilidad de los mismos en repositorios públicos como NCBI GEO y ENA.
- Análisis de la actividad transcripcional global en pacientes de LES para i) establecer TFs con actividad diferencial en pacientes de LES e individuos sanos y ii) análisis de estratificación de pacientes en base a las actividades transcripcionales.
- Análisis de actividad transcripcional y redes de señalización en datos de expresión de célula única con el fin de determinar patrones de actividad transcripcional y regulación específicos de tipos celulares en pacientes de LES.

3 Recopilación e Integración de Datos Ómicos de Patologías Autoinmunes Disponibles en Repositorios Públicos. Desarrollo de ADEx.

Las enfermedades autoinmunes afectan a una proporción significativa de la población mundial, a pesar de que algunas de ellas se consideran enfermedades raras. Entre ellas, se encuentran el lupus eritematoso sistémico y la artritis reumatoide, que son bastante comunes. Sin embargo, la falta de conocimiento sobre cómo se desarrollan estas enfermedades y cómo identificarlas con precisión es un desafío para la comunidad científica, ya que muchas de ellas presentan síntomas similares y son muy heterogéneas.

Aunque en los últimos años el avance de las ciencias ómicas ha permitido desentrañar algunos de los aspectos moleculares principales relacionados con estas enfermedades, tanto la búsqueda de biomarcadores como las terapias empleadas para tratarlas no son consistentes. Esto se debe, en parte, a la falta de información sobre las características moleculares específicas de cada enfermedad autoinmune y a la complejidad de los sistemas biológicos involucrados.

Por otro lado, la creación de repositorios de datos públicos que almacenen y compartan información ómica generada por la comunidad científica se está convirtiendo en una práctica cada vez más común. Aunque existen muchos esfuerzos colaborativos para desarrollar este tipo de repositorios, tanto a nivel global como NCBI GEO o ENA, como a nivel más específico, como The Cancer Genome Atlas (TCGA), que almacena datos de pacientes con cáncer de todo el mundo, no hay ninguna base de datos ómica que contenga estudios y trabajos relacionados con las enfermedades autoinmunes.

Con el objetivo de abordar esta laguna de conocimiento, nos propusimos desarrollar una base de datos que contenga información sobre estudios ómicos disponibles en los repositorios públicos de NCBI GEO y ENA que se enfoquen en enfermedades autoinmunes. A través de esta base de datos, esperamos ayudar a la comunidad científica a identificar y comprender mejor las características moleculares específicas de estas enfermedades, lo que puede contribuir a la identificación de biomarcadores y a la mejora de las terapias empleadas para tratarlas.

3.1 Recopilación de Datos Públicos en NCBI GEO

Cómo se ha descrito en la introducción, los repositorios públicos de datos ómicos resultan una eficaz herramienta para analizar los datos generados por otros investigadores, permitiendo abarcar un amplio abanico de funcionalidades. Con esta idea en mente, nuestro primer objetivo fue construir una base de datos ómicos de enfermedades autoinmunes, estandarizando los protocolos de procesamiento y aplicando una curación manual de los metadatos. Todo este esfuerzo finalizó con la publicación de la herramienta web ADEx[141], disponible en este enlace: https://adex.genyo.es/. En este apartado vamos a describir los aspectos claves de la construcción de la base de datos, incluyendo los criterios de selección de muestras y estudios de NCBI GEO, la descarga de los datos y la curación manual de los metadatos.

3.1.1 Criterios de Selección de Series y Muestras en NCBI GEO

Los datos que se han incluido en la base de datos proceden de datos de transcriptómica y epigenómica de humano, todos ellos localizados en NCBI GEO, así como los metadatos de origen biológico de las muestras.

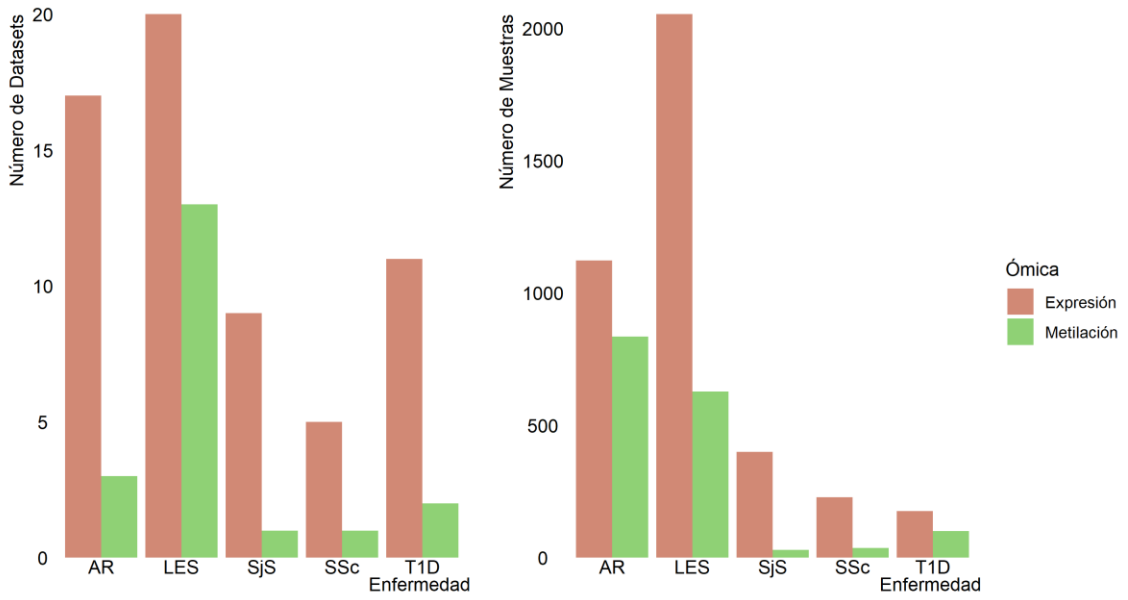

Figura 10. Datos disponibles en ADEx.

Primero realizamos una búsqueda manual de conjuntos de datos NCBI GEO de origen transcriptómico (tanto *microarray* como RNA-Seq) y/o epigenómico (*microarray*) de varias enfermedades autoinmunes (LES, AR, SSc, SjS y TD1). Se realizó una búsqueda global en NCBI GEO, utilizando el motor de búsqueda alojado en la web, para incluir estudios en humano de estas 5 enfermedades que incluyesen datos ómicos de las tecnologías mencionadas. De todos los estudios que se obtuvieron, hicimos un curado manual de los mismos en base a los siguientes criterios:

- El estudio contiene muestras de algunas de las enfermedades autoinmunes propuestas, así como controles sanos.
- Las muestras utilizadas deben ser obtenidas a partir de tejido, sin ningún tratamiento en laboratorio, más allá del necesario para procesarlas y generar los datos. Es decir, no queríamos incluir muestras sobre las que se había llevado a cabo un tratamiento *in vitro*. Esto no excluye a las muestras de pacientes que si están tratados.
- Los datos sin procesar, o crudos, deben estar disponibles.
- Las muestras de casos y controles deben pertenecer al mismo tejido.
- Debe haber al menos 10 muestras.
- Los datos procedentes de *microarrays* de ADN deben ser generados por plataformas de Affymetrix o Illumina.
- Las series que contienen muestras de diferente tejido, plataforma o enfermedad se incluirán como estudios diferentes. Por ejemplo, si un estudio tiene datos de LES y AR, se crearán dos estudios, uno para cada enfermedad, mientras que los controles sanos se duplicarán para incluir todos en ambos estudios.
- En el caso de un estudio longitudinal, en el que se generaron diferentes muestras para un mismo paciente a lo largo del tiempo, decidimos elegir sólo la primera visita.

Tras aplicar todos estos criterios, la base de datos consta un total de 5609 muestras de 82 series de estudios de casos y controles sobre la expresión y metilación de LES, AR, SSc, SjS y T1D, que se importaron a la aplicación web ADEx. La distribución de estas enfermedades en términos de número de muestra y número de serie se puede ver en la Figura 10.

En la pestaña *Overview* de la aplicación ADEx, los usuarios podrán acceder a información general de los datos contenidos en la base de datos a través de tablas o gráficos circulares. En las tablas tenemos acceso a la información de cada estudio, desde el número de casos y controles, el enlace de NCBI GEO que conduce al estudio, el tipo de datos o la plataforma con

la que se generaron. Además, también se muestran algunas variables cualitativas, como el tipo de tejido y célula, o algunas variables clínicas, como el sexo, la edad o la raza, si están disponibles. El gráfico muestra la información disponible en ADEx a nivel cuantitativo, pudiendo elegir la información a mostrar entre las diferentes variables cualitativas. Además, tanto las tablas como los gráficos se pueden filtrar para seleccionar grupos de estudio agrupados por enfermedad o seleccionar un estudio específico (véase Figura 11).

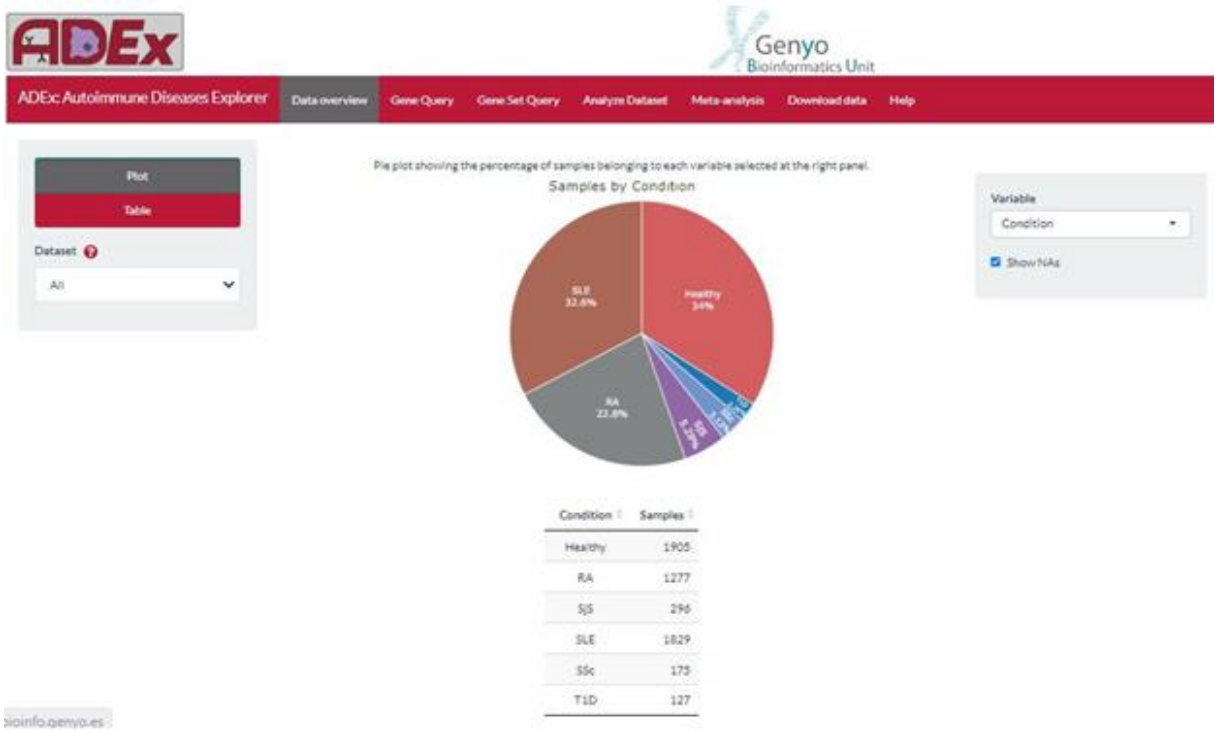

Figura 11. Imagen de la sección *Overview* de ADEx.

3.1.2 DESCARGA Y PROCESAMIENTO DE DATOS ÓMICOS Y METADATOS

Los estudios que cumplieron con los criterios anteriores se descargaron de NCBI GEO, tanto datos ómicos como metadatos. Con el fin de obtener los metadatos de los estudios seleccionados se utilizó la librería de R GEOquery. Esta librería hace de puente entre la información disponible en NCBI GEO y nuestra sesión de R. A diferencia de los metadatos, los datos ómicos sin procesar se descargaron de formas muy diferentes. A continuación, se describirá los pasos seguidos para descargar cada tipo de dato ómico.

Los datos crudos de expresión y metilación de *microarrays* se descargaron desde la web de NCBI GEO, descargando los ficheros CEL de las plataformas de Affymetrix y ficheros de texto plano, en formato de matriz, generados en las plataformas de Illumina. Para los datos de metilación, nos encontramos con la posibilidad de que los datos estuvieran en dos formatos distinto, ya sea ficheros de texto plano o fichero IDAT, por lo que decidimos usar ambos.

La descarga de los ficheros generados por la tecnología de RNA-Seq no fue tan sencilla, ya que en la web de NCBI GEO solo se alojan las matrices ya transformadas, no los datos crudos, los ficheros *FASTQ*. Por ello, se desarrolló un pequeño script mediante el cual podíamos acceder a los códigos de identificación de ENA a partir del código identificativo de cada estudio de NCBI GEO. De este modo se descargaron los ficheros *FASTQ* de los estudios de RNA-Seq. Dicho script está disponible en el siguiente enlace de GitHub. (https://github.com/ralodo93/useful_scripts/blob/main/getGEOFastq.R).

3.2 Protocolos de Procesamiento de los Datos

Como mencionamos, nuestro objetivo era crear una base de datos ómica de varias enfermedades autoinmunitarias. Por lo tanto, decidimos seguir un protocolo estándar para todos los datos. Al tener diferentes tipos de datos (microarray, RNA-Seq, metilación), cada tipo debía ser procesado de manera individual. Por lo tanto, aplicamos un método específico a cada uno de ellos para lograr el procesamiento óptimo (véase Figura 12).

Comenzamos procesando los datos (ficheros CEL) de *microarrays* de Affymetrix, cargándolos y normalizándolos con el paquete affy en R[143]. Se eliminaron aquellas sondas que presentaban una intensidad de señal menor de 100 en al menos el 10% de las muestras y normalizamos los datos transcriptómicos mediante el método RMA (*Robust Multichip Average*)[144].

Los ficheros crudos pertenecientes a los conjuntos de datos generados por las plataformas de microarrays de expresión de Illumina se leyeron como ficheros planos en R. Eliminamos aquellas sondas que tuviesen un p-valor de detección menor de 0.05 en al menos el 10% de las muestras. Posteriormente se realizó una limpieza de ruido y una normalización por cuantiles usando la función neqc del paquete de R limma[39].

Como hemos descrito anteriormente, hay algunas sondas de estos microarrays que están asociadas a varios genes, así como genes que contienen secuencias que hibridan en más de una sonda. Para poder transformas las matrices de expresión de sondas a matrices de expresión de genes, para aquellos genes que presentaban más de una sonda se calculó la mediana de la expresión de todas las sondas.

Para los datos de RNA-Seq, se llevó a cabo un protocolo estándar que incluía: alineamiento de las secuencias de los ficheros FASTQ al transcriptoma de referencia hg38 utilizando el alineador STAR (versión 2.4)[22]. La expresión cruda de los genes, manifestada en conteos (se utilizará el término *counts*) se obtuvo con RSEM (versión 1.2.31) con los parámetros por defecto[25]. Para incorporar los *counts* crudos en R, leímos los ficheros que proporciona RSEM y, para el procesamiento, utilizamos, en primer lugar, la librería NOISeq[145] para eliminar genes con baja expresión (usando los parámetros: expresión media < 0.5 *counts* por millón (CPM) y un coeficiente de variación mayor de 100), y, posteriormente el método TMM (Trimed Mean of M values) para normalizar los datos[146]. De igual modo que con los datos de microarrays debíamos asociar las sondas a los diferentes genes, en el caso de RNA-Seq se deben asociar los códigos de los transcritos de Ensembl con los nombres de los genes, lo cual se realizó con el paquete biomaRt[147].

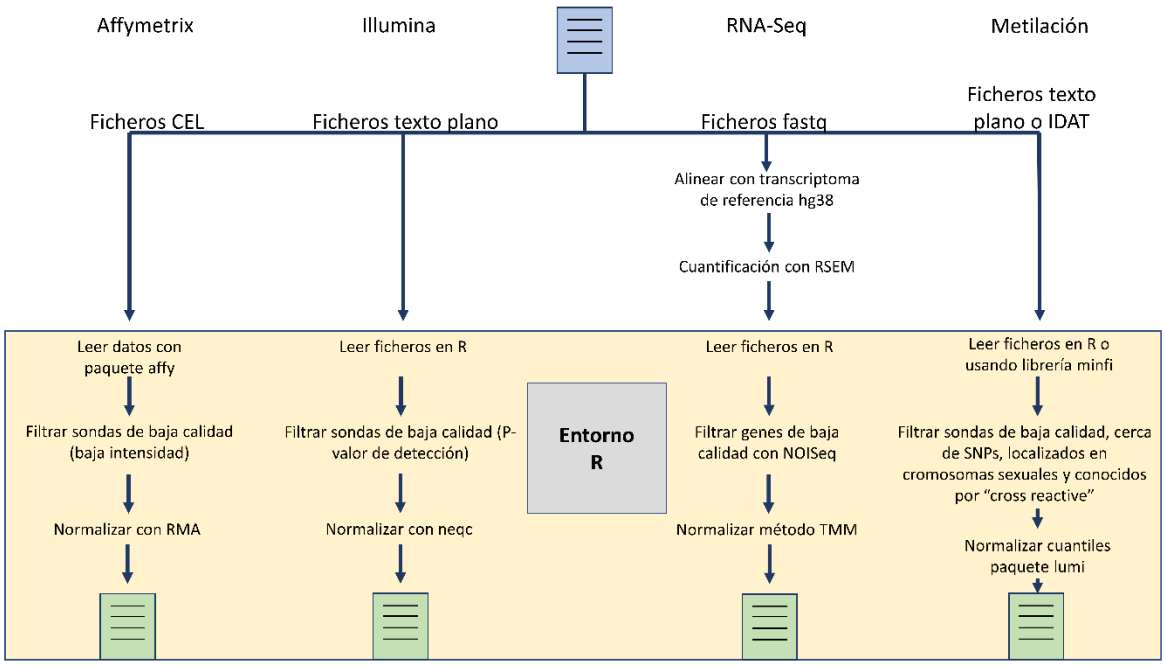

Figura 12. Diagrama con los pasos realizados para procesar los datos de diferentes fuentes. Se incluyen los pasos desde que se descargan los datos crudos hasta la obtención de datos normalizados y preparados para ser sometidos a análisis posteriores. Cada columna indica la fuente de procedencia de los datos; Affymetrix e Illumina hacen referencia a microarrays de expresión de plataformas de ambas entidades, RNA-Seq consiste en todos los datos

transcripcionales obtenidos utilizando secuenciación NGS y Metilación es el conjunto de estudios que contienen datos epigenómicos en arrays, bien sea 450K o EPIC.

Hemos comentado que los datos de metilación estaban disponibles en formato IDAT o en tablas de texto plano. Para cargarlos en R, la información de los ficheros IDAT se obtuvo con el paquete minfi[148], mientras que los datos disponibles en tablas de texto, se leyeron con el paquete de R vroom, que permite leer ficheros grandes de forma muy eficiente mediante el uso de múltiples hilos. El procesamiento de los datos una vez se han leído es común. Primero se eliminaron las sondas con un p-valor de detección mayor de 0.05 en más del 10% de las muestras, además de las que se sitúan cerca de los SNPs, las localizadas en cromosomas sexuales y las que se conocen que son "*cross reactive*", es decir, que se unen a secuencias de ADN por similitud estructural[149]. Tras eliminar estas sondas, las señales de metilación se normalizaron por cuantiles usando el paquete lumi[150] y, para los datos cuyo origen era la plataforma 450K, se aplicó una normalización BMIQ (*Beta-Mixture Quantile*)[151] con el paquete wateRmelon para corregir los dos tipos de sondas de la plataforma[152].

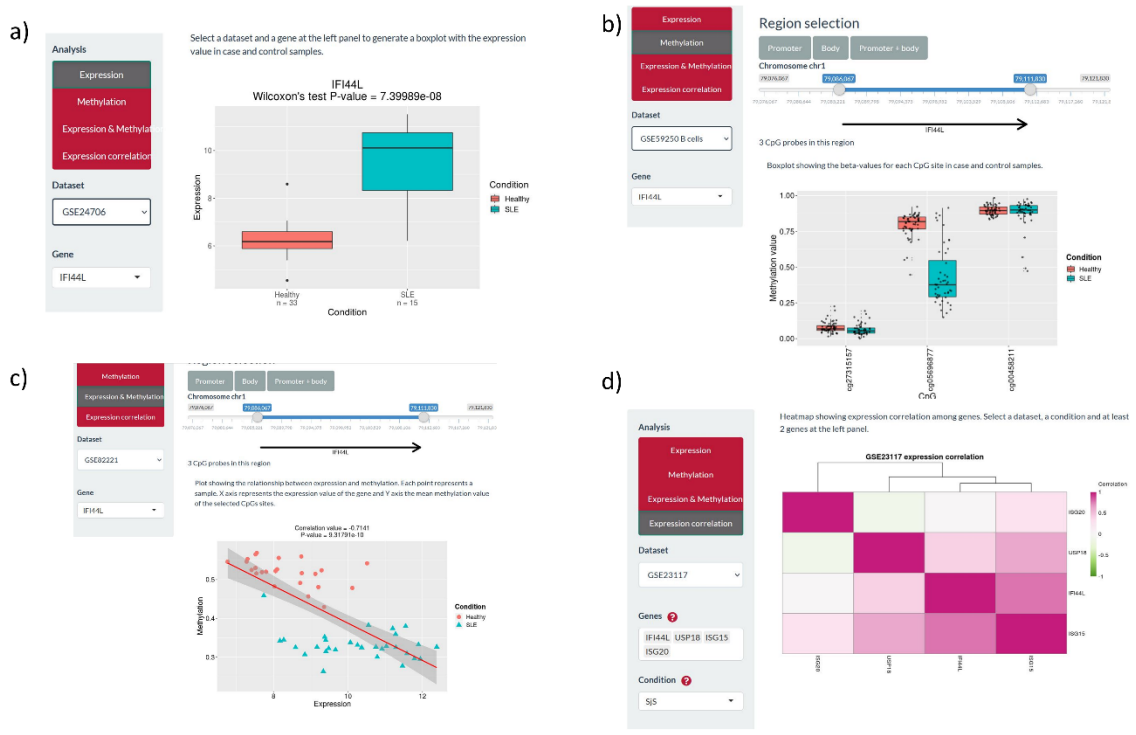

Figura 13. Funcionalidades de la sección *Gene Query* de ADEx. a) Diferencias de expresión en un gen concreto. b) Diferencias de metilación en las sondas contenidas en un gen concreto. c) En el caso de estudios con datos de metilación y expresión, podemos ver la correlación entre ambas ómicas para un gen concreto. d) Valores de correlación entre un conjunto de genes.

En la aplicación ADEx, se pueden descargar todas estas matrices de expresión y metilación, así como los metadatos curados manualmente. Además, en la sección *Gene Query*, es posible ver los datos procesados a nivel de gen (o sonda de metilación) (Figura 13a y b). Podemos elegir un gen específico de diferentes estudios para ver la diferencia entre los casos. De manera similar, las diferencias en la metilación se pueden ver en los estudios de metilación. Hubo dos estudios (GSE82221 y GSE117931) que tenían datos de expresión y metilación para la misma muestra. Para estos dos casos específicos, podemos observar la correlación entre el valor de expresión del gen y el valor de metilación de las sondas asociadas al gen anterior, lo que puede identificar las sondas ubicadas en la región desencadenante, en el cuerpo de ese gen. o en ambos (Figura 13c). La última característica de esta sección permite monitorear la correlación entre la lista de genes indicados por el usuario (Figura 13d).

3.3 DESARROLLO DE MÓDULOS DE ANÁLISIS EN ADEX

Tras seleccionar, emparejar, filtrar, normalizar y procesar los datos disponibles en la serie en NCBI-GEO, aplicamos una serie de técnicas para obtener resultados a partir de los datos homogeneizados de múltiples estudios. Las técnicas aplicadas son: análisis de expresión diferencial, análisis de enriquecimiento de rutas, análisis de redes de señalización, inferencia de redes causales y metaanálisis. Todos los resultados de esta sección se almacenaron en las pestañas *Gene Set Query*, *Analyze* Dataset y *Meta-Análisis* de ADEx. La Figura 14 detallan los pasos que se han realizado para generar todos estos resultados según lo que se puede encontrar en cada una de las secciones de ADEx de las que se va a hablar.

3.3.1 ANÁLISIS DE EXPRESIÓN DIFERENCIAL

Todos los estudios de expresión se analizaron de forma independiente y, según el tipo de datos, aplicamos diferentes métodos. El objetivo de este análisis fue investigar los genes expresados diferencialmente entre muestras pertenecientes a cada enfermedad y controles sanos para cada estudio. Los datos obtenidos del procesamiento de *microarrays* de expresión, tanto de Illumina como de Affymetrix, se analizaron de acuerdo con un *workflow* estándar con el paquete limma. La funcionalidad principal de este paquete es ajustar los datos a un modelo lineal seguido de un t-test utilizando el método empírico de Bayes para obtener los resultados de expresión diferencial (función lmFit para ajustar el modelo y eBayes para aplicar el t-test). El flujo de trabajo devuelve una tabla con información valiosa para identificar qué genes se expresan diferencialmente y en qué medida. Las variables obtenidas de este análisis fueron: p-valor, p-valor ajustado por la técnica de Benjamini-Hochberg (BH) y el valor de *fold change*.

Para los datos procedentes de RNA-Seq se utilizó el paquete DESeq2[40]. Esta librería está especializada en realizar análisis de expresión diferencial con datos de RNA-Seq, modelando los datos mediante una distribución negativa binomial, ajustando los parámetros del modelo mediante el método de máxima verosimilitud. Los resultados de expresión diferencial entre las dos condiciones de cada estudio se resumen en una tabla con las mismas variables descritas a la hora de realizar los análisis de expresión diferencial con limma.

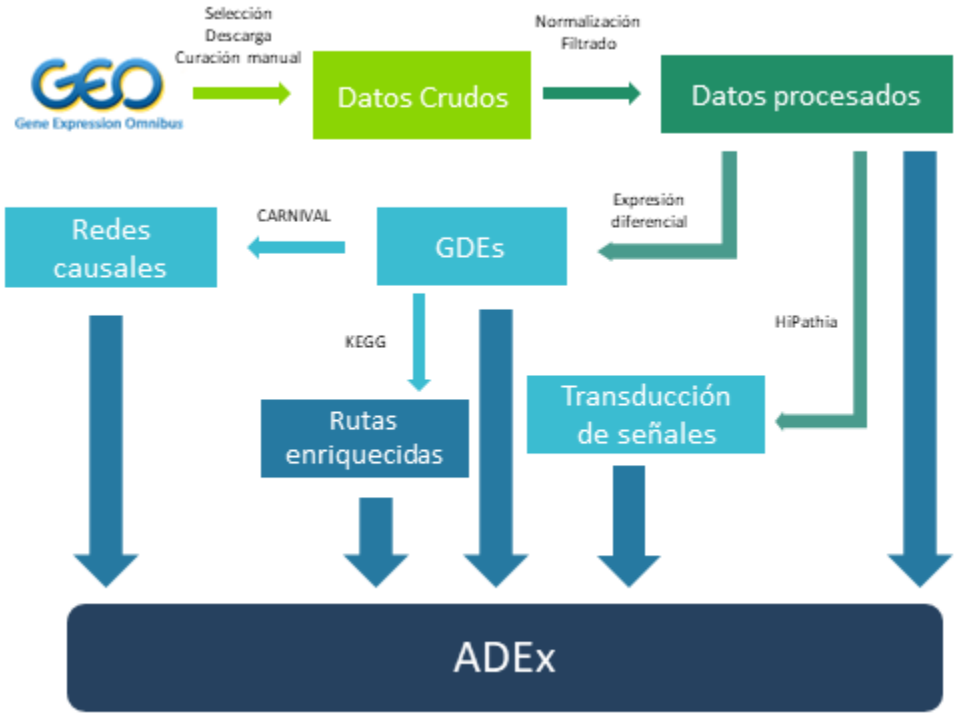

Figura 14. Análisis incluidos en ADEx. Los datos descargados de NCBI GEO son procesados y posteriormente se aplican varios tipos de análisis: Expresión diferencial, con la que se obtienen los genes diferencialmente expresados (GDEs), que son utilizados tanto para realizar un análisis de enriquecimiento como para identificar redes causales. Los datos procesados también se utilizan para aplicar métodos de transducción de señales y para ser incluidos para poder descargarse directamente desde ADEx.

En la pestaña *Gene Set Query* los usuarios podrán seleccionar tantos estudios como se desee, así como introducir una lista de genes con el fin de explorar el *fold change* entre casos y controles a lo largo de todos los estudios. Se pueden seleccionar todos los estudios de una misma enfermedad o seleccionar estudios concretos (Figura 15a). También se pueden introducir listas de genes personalizadas, aunque hay algunas listas de genes asociadas a procesos inmunes disponibles de forma automática. Estas listas de genes se han elaborado a partir de los módulos de coexpresión generados por Chaussabel y colaboradores[153]. Estos módulos o conjuntos de genes son listas de genes que están coexpresados en miles de muestras. Cada módulo se asocia a una ruta o tipo celular, muchos de ellos relacionados al sistema inmune.

En la sección *Analyze Dataset* encontramos los resultados de la expresión diferencial, que supone la búsqueda de diferencias de expresión entre dos condiciones, en nuestro caso entre casos y controles de cada estudio (Figura 15b). Los resultados se muestran en forma de tablas o heatmap, indicando por defecto el top 50 de genes significativos ordenados por p-valor ajustado mediante BH.

Figura 15. Resultados de expresión diferencial en ADEx. a) En la sección *Gene Set Query* es posible observar el log fold change de varios estudios y un conjunto de genes concreto (introducidos por el usuario o seleccionando conjuntos de genes predefinidos). b) En la sección *Analyze Dataset* podemos acceder a los resultados de expresión diferencial de cada estudio.

3.3.2 Análisis de Enriquecimiento de Rutas

Como se describió en la introducción, los análisis de enriquecimiento funcional brindan una información muy valiosa en términos de interpretabilidad de los resultados. Por lo tanto, se llevó a cabo un análisis de anotación funcional utilizando los resultados de expresión diferencial en cada uno de los estudios. Para ello, se utilizó el paquete de R llamado KEGGprofile (versión 1.24.0), que realiza un análisis de enriquecimiento basado en SEA y utiliza la prueba exacta de Fisher. Esta librería permite llevar a cabo este tipo de análisis utilizando la base de datos KEGG. Además, cuenta con una serie de funciones para manipular la visualización de los mapas de KEGG.

Para utilizar este método, se seleccionaron los genes a partir de la expresión diferencial, identificando aquellos cuyo p-valor ajustado por BH fuera inferior a 0.05. Además, con el objetivo de visualizar los resultados de cada estudio, se modificó la base de datos de KEGG,

KEGG.db, para manipular las figuras resultantes y poder colorear los genes dependiendo del valor de *fold change* entre las muestras casos y controles.

Todos los resultados de los estudios han sido calculados previamente, por lo que no es necesario esperar para acceder a ellos. Estos pueden visualizarse tanto en una tabla con los estadísticos obtenidos como en una figura modificada para mostrar la variación de los genes participantes en cada ruta. Todo esto se encuentra disponible en la sección *Analyze Dataset* de ADEx (Figura 16a).

3.3.3 Análisis de Redes de Señalización

En el ámbito biológico, las redes de señalización son fundamentales para la regulación de diversos procesos, incluyendo el crecimiento celular. La comprensión de estas redes de señalización permitirá conocer los procesos previos que tienen lugar en una célula para llevar a cabo un determinado proceso biológico. Con el objetivo de analizar estas redes, se utilizó la herramienta HiPathia, disponible a través de su paquete de R[154].

Figura 16. Métodos de análisis sobre datos ómicos en ADEx. a) Análisis de rutas en la sección *Analyze dataset*. b) Análisis de transducción de señales en la sección *Analyze dataset*. c) Análisis de inferencia causal en la sección *Analyze dataset*. d) Meta-análisis en la sección *Meta-analysis*.

HiPathia es un modelo mecanístico que ofrece una comprensión profunda de cómo la transducción de señales se relaciona con los cambios en la expresión de los genes en una red de señalización. Además, proporciona la posibilidad de predecir los efectos de intervenir en un gen específico a través de técnicas como el silenciamiento o la administración de fármacos.

Para aplicar HiPathia, se utilizaron datos de expresión normalizados para calcular la transducción de señales de diferentes rutas de señalización establecidas en HiPathia, todas ellas provenientes de KEGG, y comparar la transducción de señales entre casos y controles de cada estudio. De esta manera, se pueden identificar los circuitos de señalización asociados a enfermedades autoinmunitarias. Los resultados se encuentran disponibles en formato de tabla y en un informe interactivo generado por HiPathia, y pueden ser accedidos en la sección *Analyze Dataset* (Figura 16b).

3.3.4 INFERENCIA DE REDES CAUSALES

Las redes de causalidad son interpretaciones gráficas del funcionamiento de un sistema biológico a partir de nodos y relaciones direccionales. Son útiles para conocer cómo los cambios de un componente afectan a todo el sistema, por ejemplo, ver cómo la mutación de un gen o la aplicación de un fármaco que actúa sobre una proteína afecta a los genes y proteínas con los que tiene relación.

Para realizar este análisis se utilizó el *workflow* propuesto por los creadores del paquete de R CARNIVAL[155]. disponible en este enlace: https://github.com/saezlab/transcriptutorial. CARNIVAL es un método desarrollado para identificar la regulación de las rutas de señalización a partir de conjuntos de interacciones almacenadas en una red de conocimiento previo. Esta red está construida a partir de rutas de regulación y dianas potenciales de perturbación. Este *workflow* parte de un análisis de expresión diferencial con el fin de calcular las actividades de los TFs con DoRothEA y las actividades de rutas con PROGENy[156]. Posteriormente, utiliza la red de conocimiento previo, disponible en la herramienta Omnipath[157], para identificar los genes alterados que pueden afectar a la ruta de señalización que afecta a la actividad de los TFs. El usuario puede acceder a los resultados de inferencia de redes causales en la pestaña *Analyze Dataset* de la aplicación (Figura 16c).

3.3.5 METAANÁLISIS

Los métodos de metaanálisis en datos ómicos, principalmente en datos transcriptómicos, permiten identificar una firma transcripcional a partir de un conjunto de estudios de este tipo de datos. Estos métodos se utilizan principalmente en tres escenarios: incrementar el poder estadístico, ya que los genes diferencialmente expresados deben incluirse en el conjunto de estudios o, al menos en su mayoría, eliminando resultados sesgados por un solo estudio, buscar patrones comunes, al comparar estudios de varias enfermedades con datos de controles y buscar patrones inversos de dos enfermedades, utilizando los datos de controles de una de ellas como si fueran casos. Los métodos de metaanálisis son muy variados y utilizar uno u otro depende de la estructura y tipología de los datos[44].

En la pestaña *Meta-analysis* es posible realizar un metaanálisis en tiempo real, seleccionando los estudios que se quieran incluir, ya sea manualmente o seleccionándolos por bloques de enfermedades. El resultado consiste en una tabla de genes diferencialmente expresados y un *heatmap* donde se muestra el *fold change* de cada estudio para cada uno de los genes (Figura 16d).

El metaanálisis se realiza con la librería RankProd[158] para aplicar el algoritmo de *rankproduct*, que es un método estadístico que calcula el producto de las posiciones de los genes ranqueadas en todas las muestras de acuerdo con un estadístico. En nuestro, para evitar que el proceso se haga muy largo y costoso computacionalmente, se utilizó el orden del *fold change* calculado previamente.

4 INFERENCIA DE ACTIVIDAD DE FACTORES DE TRANSCRIPCIÓN EN PACIENTES DE LUPUS

Aunque existen métodos bioinformáticos para procesar y analizar datos ómicos, enfocados principalmente en la transcriptómica, el objetivo principal de esta tesis es inferir la actividad de los factores de transcripción a partir de los datos de expresión de pacientes con LES. La inferencia de la actividad de estos factores de transcripción puede proporcionar información valiosa sobre cómo afectan los genes y cómo se diferencian de los controles, así como dentro de la propia enfermedad. A través de estudios de asociación llevados a cabo mediante GWAS, se han identificado varios genes asociados al LES[136], incluyendo aquellos relacionados con la región HLA[159], y se ha demostrado que muchos de estos loci están enriquecidos en sitios de unión de factores de transcripción y regiones reguladoras de genes[137,160]. A pesar de todos estos estudios, no se aporta información sobre cómo los factores de transcripción actúan sobre los genes. Por lo tanto, en este trabajo se ha inferido la actividad de los factores de transcripción en dos cohortes de datos transcripcionales de LES utilizando la tecnología desarrollada por VIPER[86] y la base de datos de factores de transcripción y genes diana de DoRothEA[76], con el objetivo de identificar un conjunto de factores de transcripción cuya firma sea robusta a lo largo de todas las muestras.

Las enfermedades autoinmunes suelen presentar un alto grado de heterogeneidad entre los pacientes. En el caso de los afectados de LES esta heterogeneidad viene dada no solo por las diferentes manifestaciones que pueden tener cada uno de ellos, sino porque a lo largo del tiempo la enfermedad entra en fases de remisión y pronunciamiento. De hecho, en un artículo reciente en el que se aplicó un método de agrupamiento con datos ómicos de pacientes de varias enfermedades autoinmunes, se estableció que existían cuatro clústeres mixto de enfermedades autoinmunes[114]. Previamente, utilizando datos transcripcionales de LES se aplicaron técnicas de clustering de los datos a nivel longitudinal en el cual se concluyó que existían tres clústeres de pacientes, independientes de tratamientos, raza o cualquier otra fuente de sesgo. Estos tres clústeres presentaban una clara diferencia a nivel clínico ya que, mientras en el primero las muestras presentaban un elevado porcentaje de neutrófilos, el segundo presentaba un elevado porcentaje de linfocitos y el tercero que constituía un grupo más heterogéneo[132]. Por este

motivo, nosotros decidimos aplicar un método de agrupamiento utilizando la actividad inferida de los TFs de las muestras de LES.

En resumen, lo que se hemos realizado en este trabajo es aplicar técnicas de inferencia de actividad de TFs a partir de transcriptómica sobre dos cohortes independientes y diferenciadas de LES (una con pacientes adultos y otra con pacientes pediátricos), buscar grupos de pacientes utilizando técnicas de agrupamiento sobre los datos de actividad y buscar biomarcadores robustos comparando las actividad de los TFs entre pacientes y controles, además de indagar en las funciones biológicas que subyacen de los genes que están siendo regulados por estos TFs.

4.1 Recolección y Procesamiento de los Datos

La disponibilidad de datos ómicos públicos en NCBI GEO nos ha permitido, en parte, abordar este trabajo. Las dos cohortes de LES que se van a utilizar ya se han utilizado previamente en nuestro grupo[132], una de ellas procede de NCBI GEO y está parcialmente incluida en ADEx y la otra se obtuvo a partir de una colaboración privada.

La primera de ellas, disponible en NCBI GEO, con código de acceso GSE65391, incluye datos transcriptómicos longitudinales de pacientes pediátricos con LES, obtenidos a lo largo del tiempo a través de varias visitas. En cada visita se recopilaron datos de transcriptómica de células mononucleares de sangre periférica (PBMCs) mediante un *microarray* de Illumina en la plataforma Illumina HumanHT-12 V4.0 *expression beadchip* (GPL10558). Además, se incluyen datos transcriptómicos de controles sanos. La colecta de datos y los individuos fueron obtenidos siguiendo los protocolos aprobados por el Institutional Review Boards de la Universidad de Texas Southwestern Medical Center (092010-067) y el Baylor University Medical Center (011-200)[131].

La segunda cohorte se tomó gracias a la colaboración con los doctores Daniel Goldman y Michelle Petri, del Departamento de Medicina, División de Reumatología de la Universidad Jonh Hopkins, en Baltimore, Estados Unidos. Los datos de expresión de esta cohorte corresponden a individuos de LES, en este caso adultos, a partir del proyecto SPARE (*Study of biological Pathways, disease Activity and Response markers in patients with systemic lupus Erythematosus*) aprobado por Johns Hopkins University School of Medicine Institutional Review Board[161]. Esta cohorte, al igual que la anterior, contiene datos transcriptómicos de PBMCs, generados en la plataforma Affymetrix GeneChip HT HG-U133+ PM plate (GPL13158), de individuos a los que se les ha realizado un seguimiento, por lo que también tenemos datos de varias visitas, además de controles sanos.

Ambas poblaciones conforman un conjunto de datos de 158 pacientes de LES y 46 controles sanos en el caso del grupo de datos de pediátrico y 301 individuos con LES y 20 muestras sanas para el estudio en adultos. Es de remarcar que, al ser un estudio longitudinal, cada paciente puede tener más de una visita, por lo que, en cuanto a la cantidad de muestras, el número total de muestras de LES es de 924 en la cohorte de pediátricos y de 714 en la cohorte de adulto.

Antes de comenzar con el análisis decidimos seleccionar una visita de cada paciente, la que presentase el valor de actividad SLEDAI más alto, siempre que fuera mayor de 5, siguiendo los criterios por los cuales se determina la gravedad del paciente, ya que se estima que los pacientes que presentan un valor de SLEDAI mayor de 5 presentan un 50% más de probabilidad de ser sometidos a una terapia[162]. Con esto buscamos seleccionar solo aquellos pacientes que estuvieran graves o con enfermedad activa y, además, incluir los controles sanos. Esta estrategia se ha seguido apoyada en los estudios que utilizan los llamados fenotipos extremos, que suelen utilizarse para reducir ruido de fondo o la identificaciones de genes realmente importantes[163]. En cada una de las visitas, los investigadores recolectaron, además de la expresión de genes, una serie de variables clínicas de interés, como el SLEDAI, el porcentaje de neutrófilos, linfocitos y monocitos de la muestra, los niveles de componente 3 y 4 (C3 y C4), la cantidad de células blancas (WBC-*White Blood Cell*) o la ratio de sedimentación de eritrocitos (ESR-*Erythrocyte Sedimentation Ratio*), además de variables categóricas como si el individuo sufría o no diferentes manifestaciones clínicas, como la Piuria o la Proteinuria. La Tabla 1 nos sirve para caracterizar las muestras seleccionadas en cuento a la información clínica disponible.

Tabla 1. Caracterización clínica de las cohortes de datos de pacientes de LES. En la misma se muestra la cantidad de individuos que cumplen una condición (en el caso de género, proteinuria, o piuria) o el valor medio con la desviación estándar del mismo.

	Adulto	Pediátrico
Género	67 mujeres y 2 hombres	102 mujeres y 14 hombres
SLEDAI	8.493 ± 2.5	13.371 ± 6.6
% Neutrófilos	67.289 ± 15.4	63.963 ± 14.8
% Linfocitos	22.641 ± 12.9	24.808 ± 12.4
% Monocitos	7.68 ± 4.0	7.415 ± 3.8
C3 (mg/dL)	99.667 ± 39.0	78.645 ± 37.5
C4 (mg/dL)	17.87 ± 10.2	12.495 ± 9.5
WBC (K/cu mm)	6.396 ± 3.3	6.507 ± 2.9
ESR (mm/h)	41.765 ± 31.2	50.038 ± 37.0
Proteinuria	20 con proteinuria	69 con proteinuria
Piuria	11 con piuria	41 con piuria

El procesamiento de los datos ómicos se realizó de forma similar a la colección de datos procedentes de enfermedades autoinmunes disponibles en ADEx. Tras la selección de las muestras a incluir en el estudio de datos pediátricos, descargamos la matriz normalizada de expresión y los metadatos utilizando la librería GEOquery. Revisando los metadatos, se descubrió que algunos individuos sanos tenían dos muestras, las cuales se identificaron como replicas técnicas. Se calculó la expresión media de las sondas de estas muestras. Por otro lado,

para los datos de adulto, los ficheros CEL fueron procesados siguiendo el protocolo desarrollado en la sección anterior, tal y como se indica en la Figura 12.

El proceso de transformación de matrices de expresión de sondas a matrices de expresión de genes se realizó utilizando información de plataformas disponibles en NCBI GEO y la librería biomaRt, calculando la mediana de la expresión de aquellos genes que tenían más de una sonda asociada. Además, se filtraron los genes presentes en ambas matrices de expresión de la cohorte pediátrica y la de adultos.

4.2 Inferencia de Actividad de Factores de Transcripción

Como hemos descrito, los TFs son proteínas que se unen a regiones específicas del ADN, generalmente cerca de la región del promotor de algunos genes, con el fin de controlar la transcripción de los mismos. Estos TFs no son específicos de un gen o grupo de genes, sino que se unen por afinidad a estas regiones, llamadas sitios de unión de factores de transcripción mediante motivos de unión. Existen varias bases de datos que describen las relaciones entre TFs y los genes que están regulando, siendo DoRothEA una de las más completas, ya que incluye información de varias bases de datos y fuentes de información como ChIP-Seq, TFBSs o inferencia a partir de expresión (Figura 8a). Además, estas interacciones se han clasificado en base a varios niveles de evidencia o credibilidad, de A a E, siendo A el nivel más fiable, construido a partir de curaciones manuales y experimentales, y E el nivel menos fiable, basado en predicciones (Figura 8b).

En este caso, decidimos seleccionar los regulones con la mayor credibilidad, es decir, el nivel A, con el fin de utilizar información robusta y evitar falsos positivos. El conjunto de regulones de este nivel de evidencia consistía en 168 TFs y 2602 genes diana. Acerca de estos regulones, hay una cosa que nos gustaría comentar, ya que fueron usados en el año 2020 para realizar el trabajo que se publicó. En ese momento, el artículo de DoRothEA y las expansiones posteriores no estaban publicadas en una revista indexada, sino que estaban en un preprint. De hecho, en ese momento los datos estaban almacenados en el GitHub del grupo de Julio Sáez en ficheros independientes, mientras que actualmente, además de estar disponibles a través de una librería publicada en bioconductor, existe un solo fichero global con toda la información. Los ficheros que utilizamos nosotros ya no están disponibles.

Una vez seleccionados los regulones del nivel de confianza A, utilizamos el software de VIPER para inferir las actividades de los TFs de cada muestra. La información de los MoRs de cada interacción entre TFs y genes diana se encuentra en DoRothEA. VIPER también tiene en cuenta el grado de verosimilitud de cada interacción, sin embargo, esta parte no se aplica ya que en DoRothEA todas las interacciones tienen una ratio de verosimilitud de 1.

La inferencia de las actividades de los TFs se hace, por tanto, a partir de las matrices de expresión obtenidas a partir de las dos cohortes de pacientes de LES, tanto los casos como los controles, junto con los regulones obtenidos de DoRothEA, con nivel de evidencia A, aplicando la función de VIPER para generar matrices de actividad de TFs. Cada cohorte se analiza de

forma independiente, por lo que el resultado es la obtención de dos matrices de actividad de TFs. Los pasos que se siguen para inferir las actividades de los TFs están resumidos en la Figura 17.

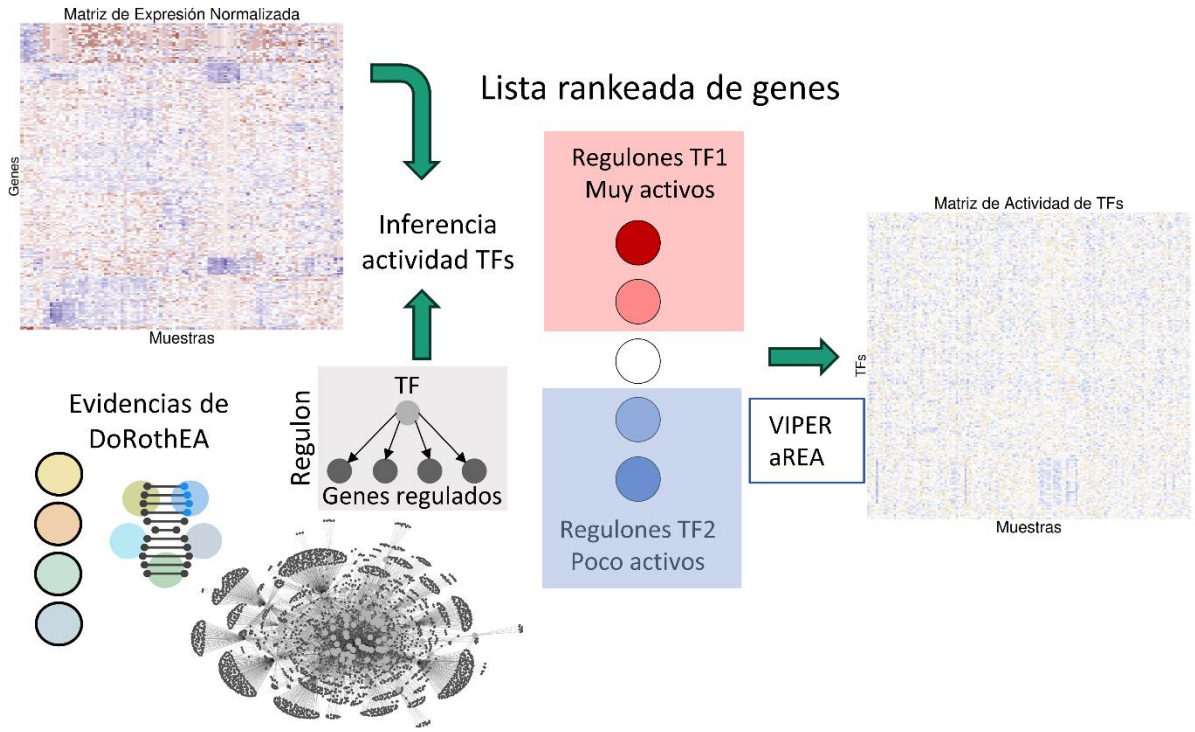

Figura 17. Esquema del funcionamiento de VIPER y aREA con los datos de expresión normalizados y utilizando la base de datos de DoRothEA.

4.3 Identificación de Subgrupos de Pacientes de Lupus en Base a la Actividad Transcripcional

Hemos remarcado en varias ocasiones la heterogeneidad que existe en algunas enfermedades autoinmunes y, específicamente, en LES. Por este motivo, decidimos aplicar un método de agrupamiento utilizando la actividad inferida de los TFs de las muestras de LES. Cada una de las cohortes fue analizada independientemente, optando por un clustering jerárquico aglomerativo mediante el método promedio (*average linkage*) utilizando la distancia euclídea. Para determinar el número óptimo de clústeres que presentaban nuestras cohortes se utilizó el índice de Calinski y Harabasz, una medida de la calidad de los clústeres que compara todas las combinaciones de clústeres de los datos y determina el número óptimo de clústeres. Este índice se encuentra integrado en el paquete NbClust[164]. Al aplicar el clustering no supervisado, o jerárquico, obtenemos un número óptimo de clústeres de 2 para ambas cohortes.

El clúster 1 de la cohorte de adultos está compuesta por 47 muestras (alrededor del 68% de las muestras de la cohorte), mientras el clúster 2 contiene 22 muestras (Figura 18a). Adicionalmente, se observa que mientras el clúster 1 estaba enriquecido en muestras con elevados niveles de neutrófilos (en porcentaje), el clúster 2 mostraba porcentajes altos de linfocitos.

En la cohorte de datos pediátrico estos dos clústeres se componían por 62 y 54 muestras respectivamente (Figura 18b), apreciándose además el mismo patrón localizado en la cohorte anterior, y es que el clúster 1 y el clúster 2 presentaban diferencias en cuanto a los porcentajes de linfocitos y neutrófilos.

Cuando observamos la distribución espacial de los datos al generar un gráfico de análisis de componentes principales (PCA), las muestras pertenecientes al clúster 2 de ambas cohortes (con altos porcentajes de linfocitos) se acercan de forma considerable a las muestras sanas en la cohorte de adultos (Figura 18c) y en la cohorte de pediátricos (Figura 18d).

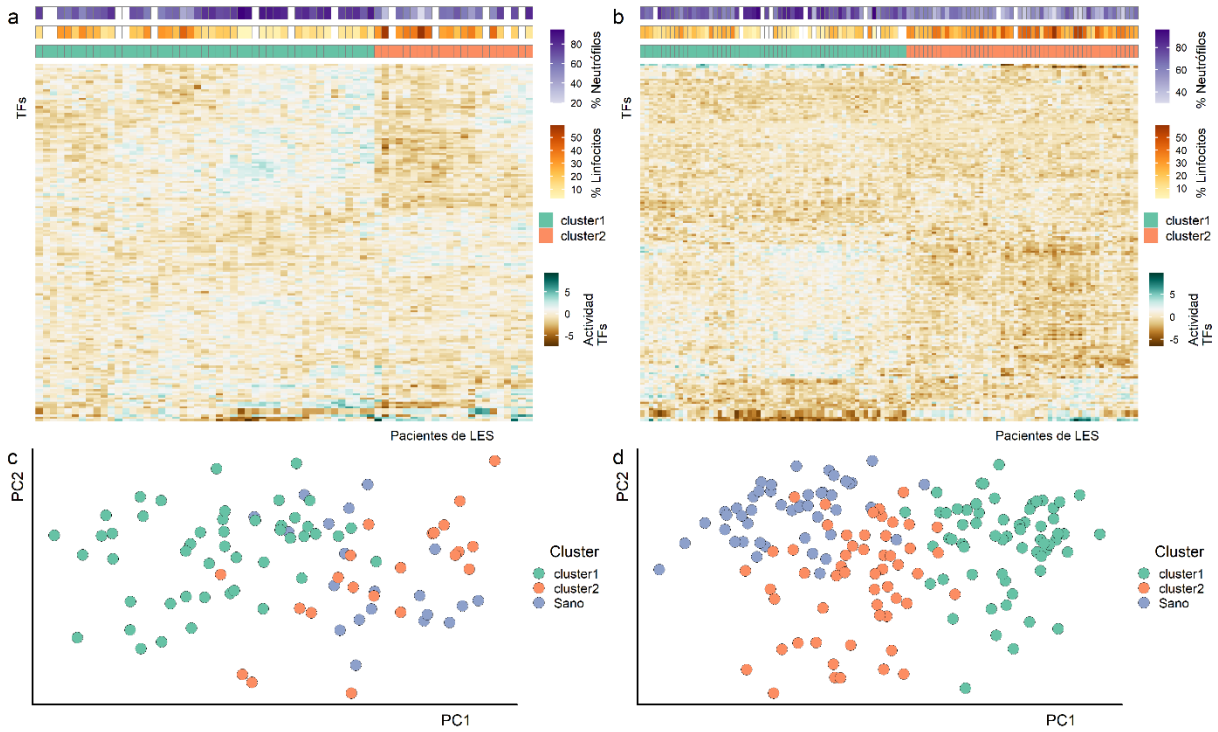

Figura 18. Clústeres obtenidos utilizando la actividad inferida de los TFs en ambas cohortes. a) Actividad de los TFs de las muestras de LES de la cohorte de adulto. b) Actividad de los TFs de las muestras de LES de la cohorte de pediátrico. c) PCA con los datos transcripcionales de la cohorte de adultos, incluyendo muestras de LES (asociadas a los clústeres obtenidos) y sanas. d) PCA con los datos transcripcionales de la cohorte de adultos, incluyendo muestras de LES (asociadas a los clústeres obtenidos) y sanas.

Ya que la distribución de los clústeres reflejaba una diferencia evidente en cuento a los porcentajes de linfocitos y neutrófilos, buscamos en la información clínica si tales diferencias resultaban estadísticamente significativas. Además de comparar los porcentajes de linfocitos y neutrófilos entre clústeres, se compararon el resto de variables cuantitativas (porcentaje de neutrófilos, porcentaje de linfocitos, porcentaje de monocitos, SLEDAI, ESR, WBC, C3 y C4). Se utilizó el test no paramétrico para dos muestras independientes de la U de Mann-Whitney, también llamada prueba de rangos de Wilcoxon, para comparar los valores de estas variables.

Cómo era de esperar, el principal hallazgo al analizar las variables clínicas de ambos clústeres de las dos cohortes consistió en la comparación del porcentaje de los tipos celulares neutrófilos y linfocitos presentando valores de significancia muy bajos (ver Tabla 2). Por este motivo, cambiamos la nomenclatura de los clústeres, de forma que, tanto en la cohorte de adultos como la de pediátricos, el clúster 1 pasó a llamarse clúster de neutrófilos (Neu Clust) y el clúster 2 será clúster de linfocitos (Lym Clust), para indicar el tipo celular enriquecido. Las

comparaciones de estas variables clínicas y las que se han mencionado anteriormente se muestran en las Figura 19a y b.

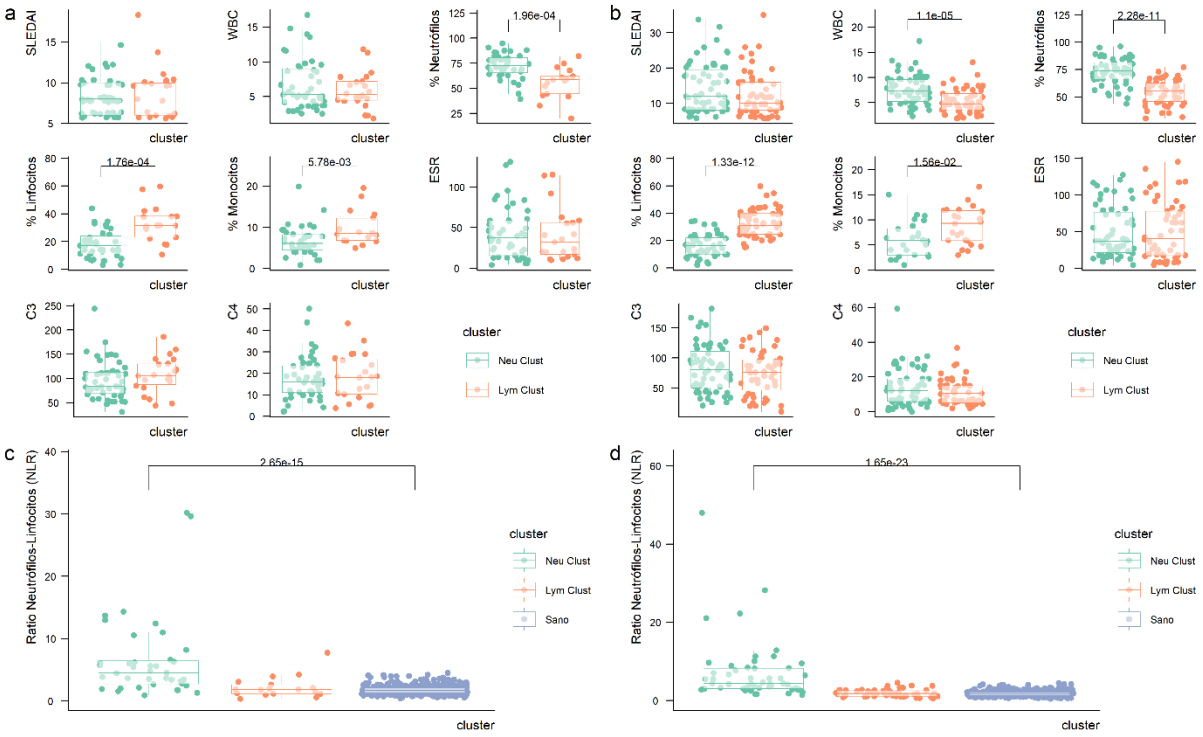

Figura 19. Comparación de variables clínicas entre clústeres. a) Gráficos de cajas en las que se observan los valores de las diferentes variables clínicas evaluadas en la cohorte de adultos. En caso de pvalor inferior a 0.05 se muestra en la gráfica. b) Gráficos de cajas en las que se observan los valores de las diferentes variables clínicas evaluadas en la cohorte de pediátricos. En caso de pvalor inferior a 0.05 se muestra en la gráfica. c) Comparación entre los valores de NLR de cada clúster de la cohorte de adultos con el estudio de nivel de NLR de personas sanas. En caso de observar un pvalor < 0.05 se muestra en la gráfica. d) Comparación entre los valores de NLR de cada clúster de la cohorte de pediátricos con el estudio de nivel de NLR de personas sanas. En caso de observar un pvalor < 0.05 se muestra en la gráfica.

Para comprobar que este resultado no se debía a un artificio, utilizamos la aplicación web de cibersort[165], que es una herramienta que permite estimar la proporción de diferentes tipos celulares a partir de datos de expresión. Para ello, incluimos las matrices de expresión de ambas cohortes y seleccionamos la base de datos que contiene cibersort, en la que se encuentran datos de expresión de tipos celulares inmunes. Utilizamos las proporciones estimadas por cibersort para aplicar comparaciones entre los clústeres de neutrófilos y los clústeres de linfocitos en las dos cohortes. Encontramos que se mantenían las diferencias entre los clústeres en las proporciones estimadas de neutrófilos. Con respecto a los linfocitos, como cibersort los divide en varios grupos de tipos celulares (células B naive, células B de memoria, células plasmáticas,

células T CD8 etc) no existe un patrón diferencial claro, aunque sí que hay tipos celulares que presentan diferencias.

Tabla 2. Significancia obtenida al comparar varias variables cuantitativas entre los dos clústeres de cada cohorte. El valor que se indica en las celdas corresponde al p-valor obtenido aplicando la prueba U de Mann-Whitney entre clústeres.

	Adulto	Pediátrico
SLEDAI	0.44471318	0.18821993
WBC	0.57539286	1.10E-05
Neu	0.00019577	2.28E-11
Lym	0.00017599	1.33E-12
Mon	0.00578086	0.01564996
ESR	0.81623418	0.84481617
C3	0.19780703	0.29140163
C4	0.54467828	0.50085811

Como se ha mencionado, al representar los datos de actividad de los TFs de los pacientes de LES y los controles sanos mediante un análisis de componentes principales (PCA), vimos que las muestras pertenecientes a los clústeres de linfocitos de las dos cohortes se asimilaban más a las muestras sanas que al otro clúster de muestras de LES. Con respecto a los porcentajes populares, se ha utilizado ampliamente en varias enfermedades, entre las que se incluye LES, el ratio de neutrófilos y linfocitos (NLR) como indicador de riesgo de las mismas[166,167]. Por ello, decidimos calcular esta variable, dividiendo la cantidad de neutrófilos entre la cantidad de linfocitos. Nuestro objetivo era comprobar si las evidencias que apreciamos en el PCA son consistentes y el NLR de los clústeres de linfocitos es similar a los rangos normales. Para determinar la NLR de los individuos sanos, ya que esta información no estaba disponible, buscamos una base de datos que nos facilitase esta información, ya que queríamos averiguar si existían diferencias entre los valores de NLR de los diferentes clústeres obtenidos con respecto a los valores de un conjunto de datos de controles sanos. Esta base de datos consta de 413 individuos sanos a los que se les había calculado el valor de NLR para determinar cuál es el rango de NLR de valores normales[168].

Se llevó a cabo una comparación entre los valores de NLR de cada clúster de nuestras cohortes con los obtenidos de la base de datos procedente de la cohorte de individuos sanos aplicando la prueba de la U de Mann-Whitney. Vimos que existía una gran diferencia entre los valores de NLR de los controles sanos y las muestras de clústeres de neutrófilos, mientras que las muestras de los clústeres de linfocitos presentaron niveles similares a los sanos (ver Figura 19c y d).

4.4 Análisis de TFs con Actividad Diferencial entre Pacientes de Lupus y Controles Sanos

A pesar de haber demostrado la existencia de una enorme heterogeneidad molecular en los pacientes de lupus, queríamos identificar patrones robustos o maestros a lo largo de los pacientes seleccionados. Para ello se realizó un análisis de actividad diferencial, de forma homóloga a los análisis de expresión diferencial. Las comparaciones entre casos y controles se llevaron a cabo siguiendo el protocolo para datos de transcriptómica de *microarrays*, utilizando el paquete limma para generar una tabla de resultados de TFs con actividad diferente entre condiciones.

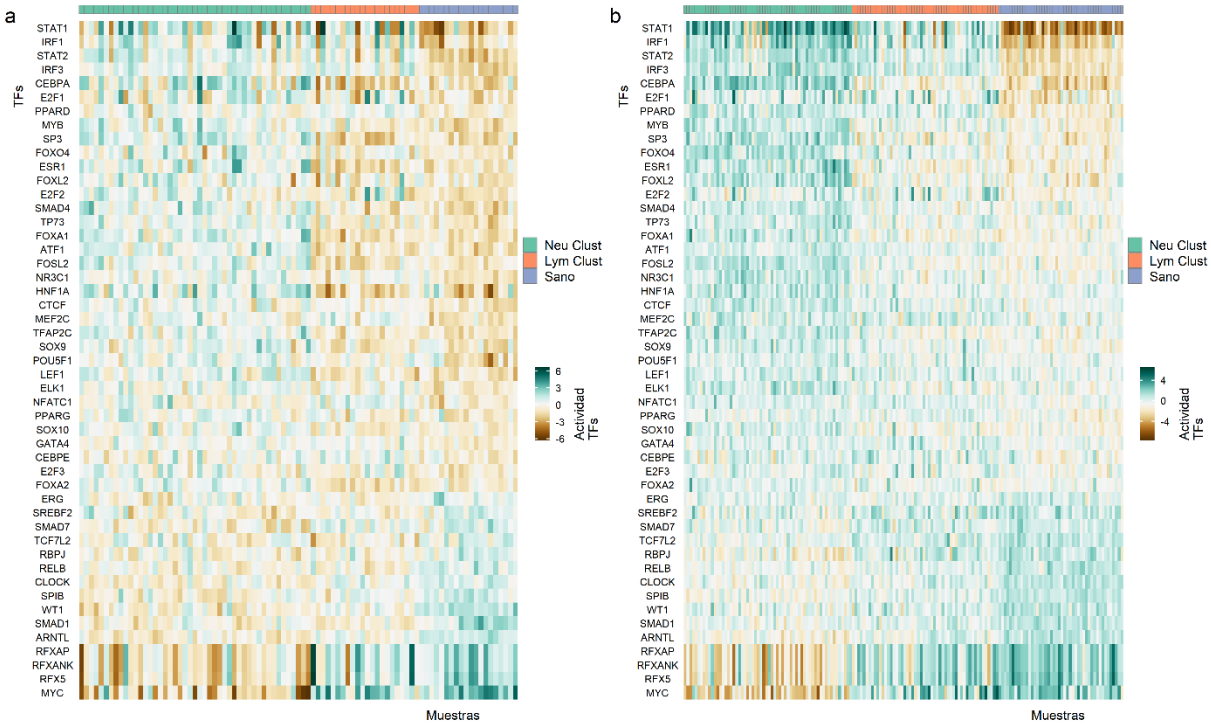

Figura 20. TFs obtenidos a partir del análisis de actividad diferencial entre muestras de LES y controles de ambas cohortes. En estas imágenes mostramos los TFs que son significativos y muestran la misma dirección, es decir, tienen actividad diferencial positiva o negativa en ambos estudios. a) Cohorte de adulto y b) cohorte pediátrica.

Al comparar los datos de LES con los controles, tal y como era de esperar según lo descrito en el apartado anterior, los resultados se veían claramente sesgados por la presencia de dos clústeres tan diferenciados. Encontramos un total de 49 TFs que presentaban actividad diferencial (BH < 0.05) entre casos y controles en ambas cohortes. Sin embargo, como se observa en la Figura 20, la heterogeneidad intragrupo que se ha establecido previamente estaba

añadiendo ruido a los resultados obtenidos. Por ejemplo, los patrones de actividad de los TFs MYC, RFX5, RFXAP o RFXANK eran muy diferentes entre los dos clústeres de LES.

Queríamos conocer el motivo de la existencia de sesgo en los resultados centrándonos en estos TFs, utilizando los genes diana de ellos en la aplicación online de ExpressionAtlas, que contiene una base de datos con la información de que genes se expresan en cada tipo celular inmune, entre otros. La expresión de muchos de los genes diana de MYC, RFX5, RFAX y RFANK es extremadamente variada entre los tipos celulares, de forma que los genes de HLA, regulados por los TFs de la familia RFX presentan una expresión mucho mayor en los linfocitos B y células dendríticas. Por otro lado, algunos genes diana de MYC, como BBC3, CXCL2, ICAM1, PTEN, EGR3, HBA2 o IMPA2 mostraban una expresión muy alta en neutrófilos. Estos dos grupos de TFs reflejan las diferencias en las proporciones de tipos celulares de ambos clústeres.

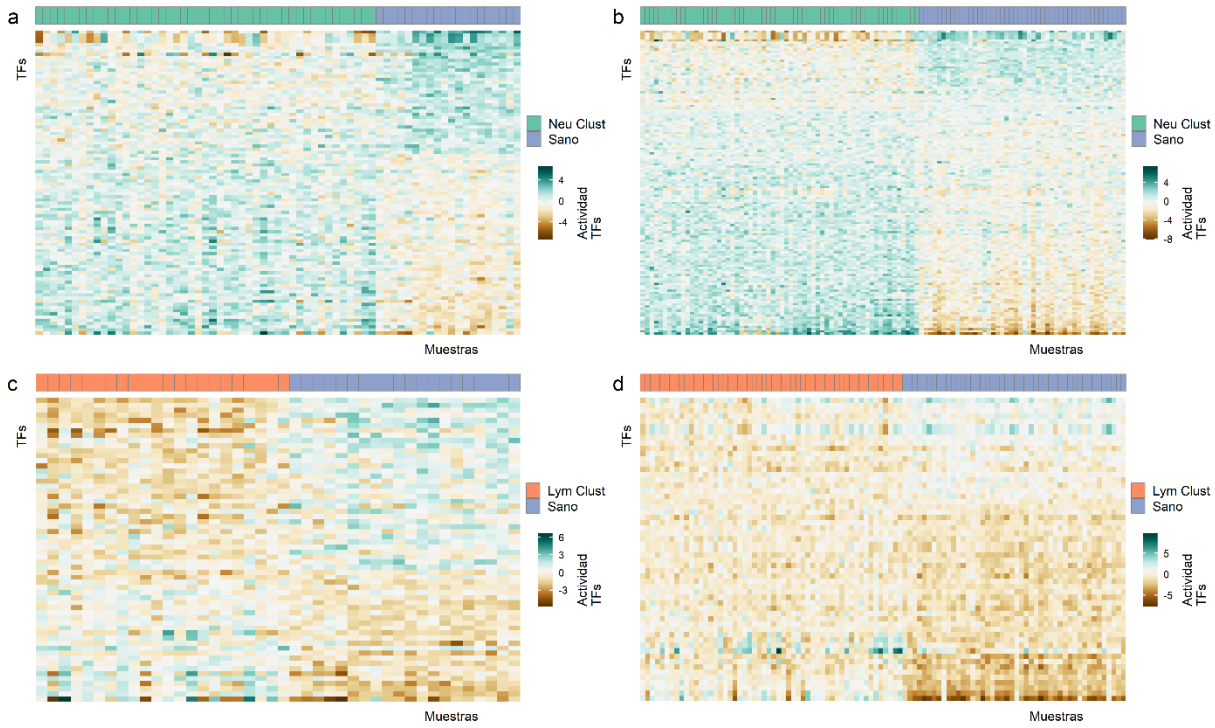

Figura 21. Actividad diferencial de los TFs al comparar cada uno de los clústeres con las muestras sanas de cada cohorte. Las figuras a) y b) se han obtenido a partir del estudio en datos de adulto y las c) y d) pertenecen al análisis en el estudio de pediátrico.

Con el fin de prevenir el sesgo en nuestro análisis, decidimos comparar cada uno de los clústeres de forma independiente con el conjunto de controles, con la intención de lograr una firma que estuviese presente en las dos cohortes y de forma simultánea en ambos clústeres de cada una de ellas. A fin de conseguir esto, optamos por quedarnos con aquellos TFs que resultaron

significativos entre cada clúster con respecto a controles ($p < 0.05$) y que además presentasen la misma dirección de actividad a nivel de clúster y de cohorte.

Este análisis reveló un conjunto de 96 y 60 TFs con actividad diferencialmente significativa en los grupos Neu Clust y Lym Clust, respectivamente, en la cohorte de adultos. Por otro lado, en el estudio que incluía datos pediátricos se obtuvieron un total de 135 y 57 TFs significativos en los clústeres Neu Clust y Lym Clust. De todos estos resultados, encontramos un sobrelapamiento de 69 TFs específicos de los clústeres de neutrófilos de ambos estudios y 21 TFs en los clústeres de linfocitos de adultos y pediátricos. Los resultados obtenidos al realizar estas comparaciones están en la Figura 21, en la que se observa que, tanto a nivel cuantitativo como visualmente, las diferencias entre las muestras de los clústeres de neutrófilos y los controles son mucho más remarcadas que las que se observan en los clústeres de linfocitos, lo cual no es extraño ya que previamente hemos demostrado que estas muestras son, tanto a nivel molecular como en relación a los datos clínicos, más similares.

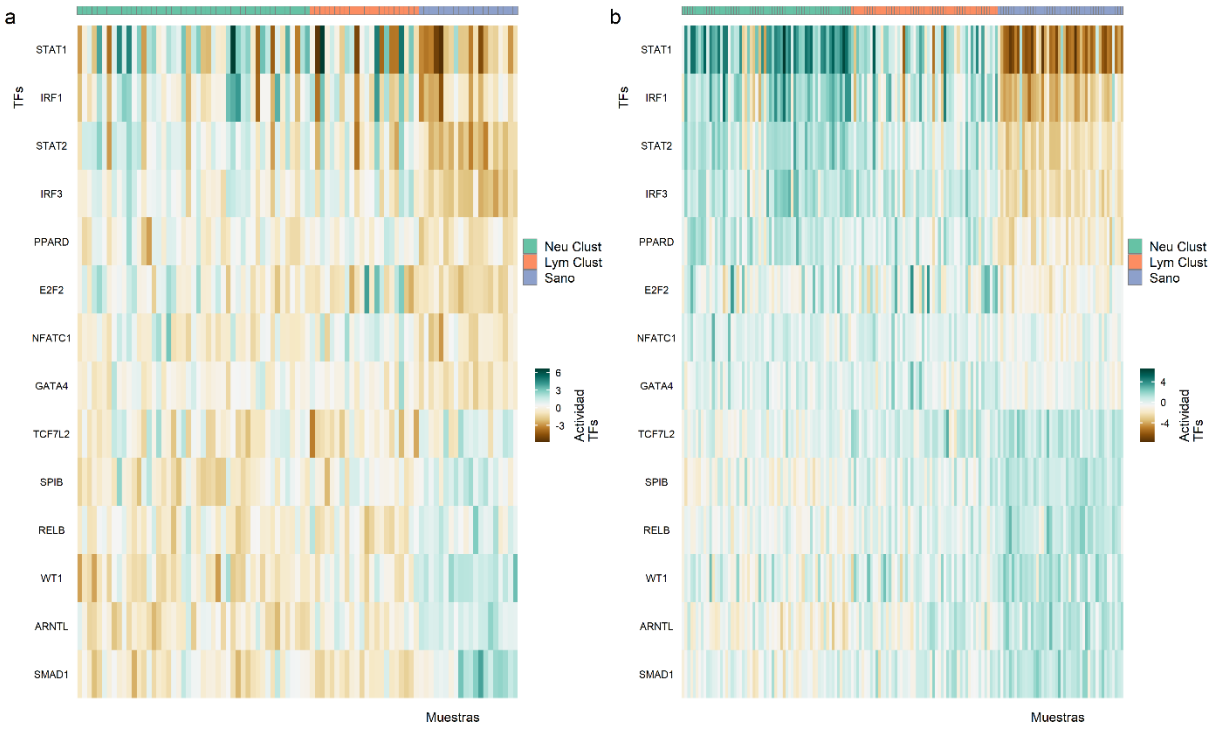

Figura 22. Firma robusta de TFs a lo largo de todas las muestras de LES en ambas cohortes. a) adulto y b) pediátrico.

Finalmente, decidimos establecer un conjunto de TFs que estuviese presente en todas las muestras de SLE, independientemente de clústeres y de cohortes. Para ello, buscamos los genes

que resultaron significativos de todas las comparaciones anteriores, siempre que mantuvieran la misma direccionalidad de diferencia de actividad entre casos y controles. Esto proceso nos devolvió una lista de 14 TFs cuya actividad era diferencialmente alta o baja a lo largo de todas las muestras de LES, independientemente de subgrupos y cohortes. Específicamente, SMAD1, ARNTL, WT1, RELB, SPIB y TCF7L2 mostró una actividad inferior en muestras de LES mientras que GATA4, NFATC1, E2F2, PPARD, IRF3, STAT2, IRF1, STAT1 se identificaron como TFs con actividad mayor, siempre con respecto a los controles (ver Figura 22).

Además, teníamos especial interés en determinar cuáles eran las funciones en las que estaban inmersas estos TFs. Sin embargo, carecía de lógica aplicar un análisis de enriquecimiento funcional sobre la lista de TFs, ya que, en la mayoría de las bases de datos de anotaciones obtendríamos resultados ligados a la regulación de la transcripción o de la expresión, lo cual no es un error ya que realmente actúan en estos procesos, pero no nos deja ver más allá. Por este motivo, decidimos centrarnos en los genes diana de los TFs significativos (aquellos con un valor de BH < 0.05). Aplicamos un análisis de GSEA, usando la librería *fgsea*[169] para evaluar que genes estaban implicados en la existencia de diferencias entre TFs entre condiciones. Esto se obtiene utilizando el *leading-edge* que proporciona GSEA, que indica los genes que tienen mayor protagonismo en la firma que se indica. Para ello, utilizamos la lista de genes diana de cada TF como conjunto de genes.

Para conseguir una firma robusta de los genes obtenidos del *leading edge*, optamos por seleccionar los genes diana que resultaron significativos de comparar cada clúster contra los controles, tanto en adultos como en pediátricos. Este análisis generó un listado de 44 genes que además de ser genes diana de uno o más de los 14 TFs identificados, eran genes diferencialmente expresados independientemente de subgrupos o edad (ver Figura 23). La figura muestra que muchos de los genes diferencialmente expresados están regulados por el TF STAT1, y son, principalmente ISGs. Para conocer el resto de rutas o procesos en los que están implicados el conjunto de genes, utilizamos la herramienta de enrichR[60].

Este análisis nos condujo a una interpretación que ya esperábamos y que hemos descrito anteriormente, y es que la mayoría de los genes sobre expresados de forma más o menos extendida entre todos los pacientes, se encuentran vinculados a rutas relacionadas con el interferón o infecciones víricas. Por otro lado, los genes infra-expresados se relacionan con procesos como el fotoperíodo y la actividad circadiana. La implicación de esta firma, pese a

que se ha estudiado previamente en LES y otras enfermedades autoinmunes[170,171], no está muy extendida y se desconoce la relación directa con la patología.

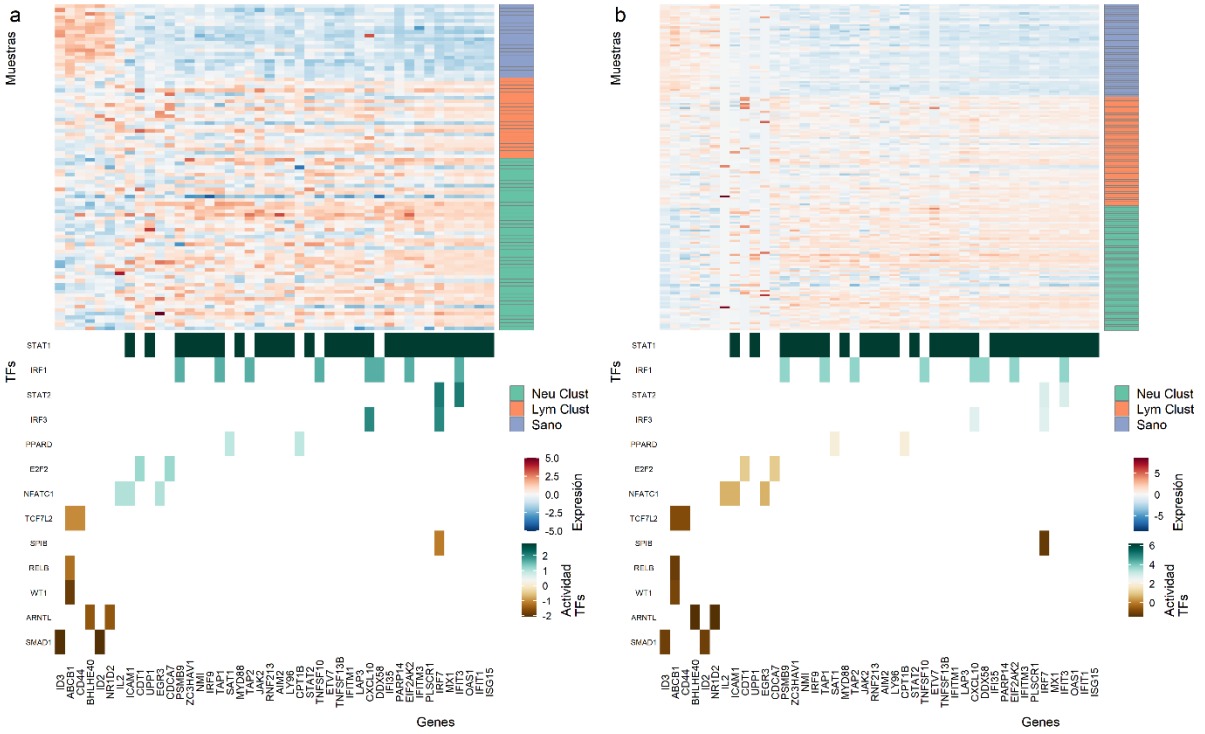

Figura 23. Genes obtenidos a partir de GSEA que son significativos de forma robusta entre casos y controles. Se aprecia también la relación entre la firma de TFs robusta y los genes diana. a) Adulto y b) pediátrico.

Finalmente, queríamos buscar una serie de fármacos cuyos target o dianas fueran los TFs robustos a lo largo de los diferentes pacientes que localizamos. Para ello recurrimos a la base de datos de CLUE/Connectivity Map[57] para asociar los fármacos y los TFs. Esta base de datos se construyó con datos transcriptómicos para un amplio abanico de compuestos químicos y experimentos de perturbaciones genéticas y permite identificar que fármacos se encuentran más ligados a una firma transcripcional, así como aquellos que potencialmente podrían revertir esta firma. En nuestro caso, tan solo utilizamos la información de los emparejamientos entre fármacos y sus genes diana. Los resultados se localizan en la Tabla 3.

Tabla 3. Información acerca de los fármacos que actúan sobre alguno de los TFs significativos.

Drug	Mechanisms of Action (MoA)	TF Target	Indication	Phase
Bezafibrate	PPAR receptor agonist	PPARD	Cholesterol	Launched
DG-172	PPAR receptor inverse agonist	PPARD		Preclinical
Elafibranor	PPAR receptor agonist	PPARD		Phase 3
FH-535	PPAR receptor antagonist, WNT signaling inhibitor	PPARD		Preclinical
GSK-0660	PPAR receptor antagonist	PPARD		Preclinical
GSK3787	PPAR receptor antagonist	PPARD		Preclinical
GW-0742	PPAR receptor agonist	PPARD		Preclinical
GW-501516	PPAR receptor agonist	PPARD		Phase 2
Icosapent	Platelet aggregation inhibitor	PPARD	Hypertriglyceridemia	Launched
L-165041	PPAR receptor agonist	PPARD		Preclinical
Sulindac	Cyclooxygenase inhibitor	PPARD	Osteoarthritis, rheumatoid arthritis, ankylosing spondylitis	Launched
Tretinoin	Retinoid receptor agonist, retinoid receptor ligand	PPARD	Leukemia	Launched
INCA-6	Calcineurin inhibitor	NFATC1		Preclinical
piceatannol	SYK inhibitor	IRF3		Preclinical
CKD-712	NFkB pathway inhibitor	STAT1		Phase 1

4.5 Discusión y Evaluación de los Resultados de la Inferencia de TFs en Cohortes de Lupus

En este trabajo, hemos inferido las actividades de los factores de transcripción utilizando los niveles de expresión génica de pacientes de LES con estado activo de la enfermedad en dos cohortes diferentes. En primer lugar, el análisis de agrupamiento no supervisado de la matriz de actividad reveló una estructura de dos clústeres en ambos conjuntos de datos, caracterizada principalmente por diferencias en las proporciones de neutrófilos y linfocitos. Nuestra clasificación molecular y caracterización clínica son coherentes con trabajos previos que estratificaron el LES o enfermedades autoinmunitarias sistémicas en las que diferentes proporciones de neutrófilos y linfocitos se asociaron con diferentes grupos[114,131,132]. En este contexto, existe evidencia reciente sobre el potencial papel de las proporciones de neutrófilos y linfocitos como posibles marcadores para estratificar a los pacientes de LES en grupos clínicamente separados[172,173]. De hecho, el potencial de NLR como biomarcador barato y efectivo de la actividad o respuesta al tratamiento en patologías autoinmunitarias está siendo analizado por diferentes grupos[174,175] y también se ha descrito que las alteraciones en el equilibrio de neutrófilos, monocitos y linfocitos explican la resistencia a los tratamientos en pacientes con artritis reumatoide, aunque los mecanismos moleculares no están totalmente caracterizados[176].

Después de comparar las actividades de los factores de transcripción de muestras de LES y muestras saludables, definimos 49 factores de transcripción con actividad diferencial significativa. Un análisis detallado reveló que algunos podrían estar sesgados por la heterogeneidad de la población celular en las muestras de LES, lo que es notable para los factores de transcripción MYC, RFX5, RFXAP y RFXANK. RFX5, RFXANK y RFXAP actúan como potenciadores de la expresión génica de los genes de clase II de histocompatibilidad mayor (MHC II)[177]. Estos se expresan en células presentadoras de antígenos profesionales (APC) como monocitos, células B y células dendríticas[178]. Esto podría explicar la baja actividad de estos factores de transcripción en el clúster de neutrófilos. Por otro lado, MYC regula un gran conjunto de genes, pero destaca la presencia de genes ribosomales como RPS15, RPL19, RPS19, RPS6, RPL3, RPL22, RPL6, RPL32, RPL27A, RPL23 y RPS16. La expresión de estos genes en diferentes tipos de células sanguíneas es muy heterogénea[179].

Debido a la heterogeneidad observada en las actividades de los TFs en pacientes con SLE, decidimos comparar las actividades de los TFs de cada grupo con muestras de individuos sanos con el fin de obtener patrones de regulación coherentes e imparciales en todos los pacientes con SLE. A partir de estos análisis, identificamos 14 TFs que se activan o reprimen consistentemente en SLE.

Los que presentaron mayor actividad en SLE fueron STAT1, STAT2, IRF1, IRF3, NFATC1, PPARD, E2F2 y GATA4. Este conjunto de TFs incluye factores de transcripción que son conocidos en el contexto de la patogénesis de SLE, como STAT1, STAT2, IRF1 e IRF3, que son activadores de genes de interferón[180,181]. El análisis de la asociación de fármacos con los TFs reveló que STAT1 es el objetivo de CKD-712, un inhibidor de la vía de NF-κB, un mediador proinflamatorio[182]. Por otro lado, un inhibidor de SYK actúa sobre IRF3. Este mecanismo ha demostrado ser efectivo en la artritis reumatoide y en ratones MRL/lpr propensos al lupus[183]. De hecho, la sobreexpresión de Syk en células T sanas conduce a un fenotipo de células T similar al de SLE, lo que sugiere que la inhibición de Syk tiene el efecto opuesto. Syk se ha propuesto como un objetivo terapéutico[184].

NFATC1 está sobreexpresado en ratones MRL/lpr propensos al lupus, activando la vía calcio/NF-AT[185]. Además, este TF es el objetivo de un inhibidor preclínico de calcineurina[186], el mecanismo inhibitorio a través del cual la ciclosporina y el tacrolimus ejercen sus efectos cuando se utilizan en pacientes con lupus eritematoso sistémico (LES)[187].

Los ratones deficientes en *Ppard* desarrollan autoinmunidad tipo lupus con aumento en la producción de autoanticuerpos y una eliminación anormal de células apoptóticas[188]. Existen algunos fármacos cuyo objetivo es PPARD. Uno de los fármacos más destacados que apuntan a PPARD, la tretinoína, está relacionado con el mecanismo de acción (MoA) de agonista del receptor del ácido retinoico. Se ha informado de la mejora de los síntomas inflamatorios del LES con el tratamiento con ácido retinoico en modelos murinos y enfermedades humanas[189]. La mejoría del LES se podría lograr a través del tratamiento con ácido retinoico mediante tres mecanismos[190]. Uno de ellos es mediante la reversión de la disbiosis microbiana[191]; en segundo lugar, mediante la inhibición de la actividad de Pin-1, que activa la señal TLR-7/TLR-9/IRAK-1/IRF-7 que contribuye al fenotipo del LES[192], y, en tercer lugar, mediante el restablecimiento de los niveles de vitamina A en pacientes con LES, lo que mejora el equilibrio de las células T auxiliares 17 (Th17) y las células T reguladoras (Treg)[189]. Además, PPARD es una diana de

sulindac, que es un inhibidor de la ciclooxigenasa lanzado en ensayos clínicos de artritis reumatoide o espondilitis anquilosante.

El resto de los TFs significativos tuvieron una actividad más baja en el LES que en los controles: SMAD1, ARNTL, WT1, RELB, SPIB y TCF7L2. SMAD1, junto con otros genes involucrados en la vía de señalización BMP/Smad, se reprime a través de la vía de señalización NF-κB es un factor de transcripción esencial en el proceso inflamatorio[193]. Por otro lado, RELB es una subunidad de NF-κB y participa en el desarrollo de células dendríticas[195]. En el modelo murino de lupus, las células dendríticas modificadas por *Relb* disminuyeron la expresión de interferón-γ[196]. ARNTL es un TF que forma un componente principal del reloj circadiano. Este sistema regula la expresión génica de genes involucrados en varios procesos biológicos según los ritmos circadianos. Como hemos descrito anteriormente, existen estudios que correlacionan la desregulación del reloj circadiano y la patogénesis del LES[170,171].

La nefritis lúpica es una de las manifestaciones más serias del LES, caracterizada por inflamaciones renales y la pérdida de podocitos[197]. WT1 es un marcador bien conocido de podocitos[198] y en modelo murino su expresión está disminuida[199].

SPIB pertenece a la familia de TFs ETS y promueve el desarrollo de células dendríticas plasmacitoides (pDC), los principales productores de interferón tipo I y está involucrado en el desarrollo de células B del centro germinal[200]. Sin embargo, se ha demostrado que SPIB está infra-expresado en las células B de pacientes con LES. SPIB, así como E2F2, GATA4 y TCF7L2, se han asociado con otras enfermedades autoinmunitarias a través de la expresión génica diferencial y polimorfismos genéticos, respectivamente[201–203].

Aunque muchos de los TFs asociados con pacientes de LES han sido previamente descritos, no por su actividad como TFs sino por su implicación como proteínas en diferentes procesos, en este trabajo describimos la estratificación de pacientes de LES en dos subgrupos basados en perfiles globales de actividad de TFs, los cuales se caracterizan por diferencias en las proporciones de neutrófilos y linfocitos. Además, identificamos 14 TFs significativos y robustos en pacientes de LES. Estos resultados revelan mecanismos de regulación en cuanto a la heterogeneidad del LES, los cuales podrían ser posibles blancos terapéuticos. Los grupos observados aquí son consistentes con hallazgos previos[204] y pueden relacionar la heterogeneidad molecular con manifestaciones clínicas o respuesta a terapias, proporcionando oportunidades para nuevos desarrollos terapéuticos o un mejor diagnóstico de la enfermedad.

5 ACTIVIDAD TRANSCRIPCIONAL EN DATOS DE CÉLULA ÚNICA

Las técnicas de célula única han supuesto un auténtico *boom* en los últimos años, ya que nos permiten conocer el estado de una célula concreta, al contrario que los métodos tradicionales en los que solo podíamos conocer el estado de un conjunto de células, tejidos o muestras serológicas. En este contexto, localizamos un nuevo estudio de single-cell RNA-Seq (es decir, datos transcripcionales de secuenciación a nivel de single-cell, también conocido como scRNA-Seq) en pacientes de LES, que incluían controles sanos[128]. El resultado más relevante de este análisis fue la identificación de un pequeño grupo de células dentro de cada grupo celular primario que presentaban una elevada expresión de los genes ISGs.

Al analizar los TFs inferidos a partir de datos transcripcionales de *microarrays* en muestras de PBMCs sobre pacientes de LES, observamos la estratificación en grupos que se diferenciaban, además de en la firma de TFs, en los porcentajes celulares, principalmente neutrófilos y linfocitos. Gracias a las técnicas de célula única, podemos aplicar la misma idea, pero en lugar de en sangre completa, mediante un análisis mucho más granular, centrándonos en los distintos tipos celulares. Para ello, hemos utilizado los datos disponibles del trabajo mencionado anteriormente[128], cuyos datos se encuentran en NCBI GEO (GSE135779). Este conjunto de datos consiste en datos de scRNA-Seq de 56 individuos, 40 pacientes de LES y 16 controles sanos. Además, estas muestras estaban divididas en grupos de edad, quedando estructurado finalmente como 33 pacientes de LES y 11 controles sanos de edad pediátrica y 7 pacientes de LES con 5 controles sanos de edad adulta. El conjunto de datos contiene aproximadamente 300.000 PBMCs únicas. La descarga de los datos se hizo desde la web de NCBI GEO, aunque los metadatos se obtuvieron con el paquete GEOquery.

5.1 SELECCIÓN DE MUESTRAS Y CONTROL DE CALIDAD

Se utilizaron solo las células procedentes de pacientes y controles disponibles en el grupo de edad de pediátrico, ya que eran la gran mayoría y no necesitábamos validar los resultados con los datos de adulto ya que esto era parte del trabajo publicado por los autores originales de los datos. El primer paso fue leer los ficheros que se han descargado de NCBI GEO, cuyas matrices de expresión se encuentran en formato MTX. Estos ficheros contienen la información de los *counts* de cada gen en cada célula en forma de una matriz dispersa, un tipo de matriz especial para tablas con muchos ceros, como es el caso de los datos de *single-cell*. Para leer los ficheros se utilizó la función ***Read10X*** del paquete de R Seurat. Ya que la mayoría de las funciones utilizadas para procesar los datos de *single-cell*, tanto en el control de calidad como en los análisis posteriores se llevaron a cabo con el paquete Seurat[205], salvo que se especifique otra cosa, las funciones que se describan pertenecen a esta librería.

La función ***Read10X*** se encargar de leer los ficheros que contienen la expresión de todas las células de un individuo convirtiendo esta información en una matriz dispersa, ya que es el método más eficiente de almacenar matrices gigantes con gran cantidad de ceros, como es el caso. Posteriormente se creó el objeto de Seurat con la función *CreateSeuratObject* y finalmente se añadió la información clínica o los metadatos de cada paciente. Para el paso siguiente se calcularon los porcentajes de expresión de genes mitocondriales y ribosómicos usando la función *PercentageFeatureSet*. Hasta ahora todo se ha realizado a nivel de paciente individual, por lo que debíamos reunir tanto las matrices de expresión como los metadatos en un objeto único que contuviera la información de todos los individuos. En este punto tenemos un total de 281.148 células.

Tras unir todos los datos de los pacientes se realizó un control de calidad de las muestras y los genes. Para filtrar las células, se tuvieron en cuenta los siguientes criterios: 1) células con menos de 300 genes expresados, 2) células con un % de genes mitocondriales superior al 20% y 3) células con un % de genes ribosomales inferior al 5% fueron eliminadas. En la Figura 24a se observa las células que fueron eliminadas por no cumplir estos criterios. Por motivos de elevada expresión en todas las células, se decidió eliminar también el gen MALAT1.

En este momento se llevó a cabo un análisis de duplicados (*doublets*). Los *doublets* no son más que células captadas de forma errónea en el proceso de selección de células que tiene lugar antes de la secuenciación. Se trata por tanto de "células" formadas por trozos de células muy

diferentes, o células “pegadas” unas a otras que no se han podido separar correctamente, y deben ser eliminadas para evitar que aporten ruido a nuestros resultados. Para ello se utilizó la librería scDblFinder, que evalúa todas las células y determina cuales tienen una alta probabilidad de ser *doublets*, por lo que también fueron eliminadas. Tras aplicar todos los pasos de control de calidad reducimos la cantidad de células de ~280.000 a 235.424. Estas células se distribuyen en todas las muestras de forma que tenemos más células de los individuos sanos que de los pacientes de LES (ver Figura 24b), sin embargo, como tenemos muchas más células de pacientes, el conjunto de células es mucho mayor en los individuos con LES (ver Figura 24c). El código completo de este apartado está localizado en el siguiente repositorio de GitHub: https://github.com/ralodo93/scRNA-Seq_SLE/blob/main/01_qc.R.

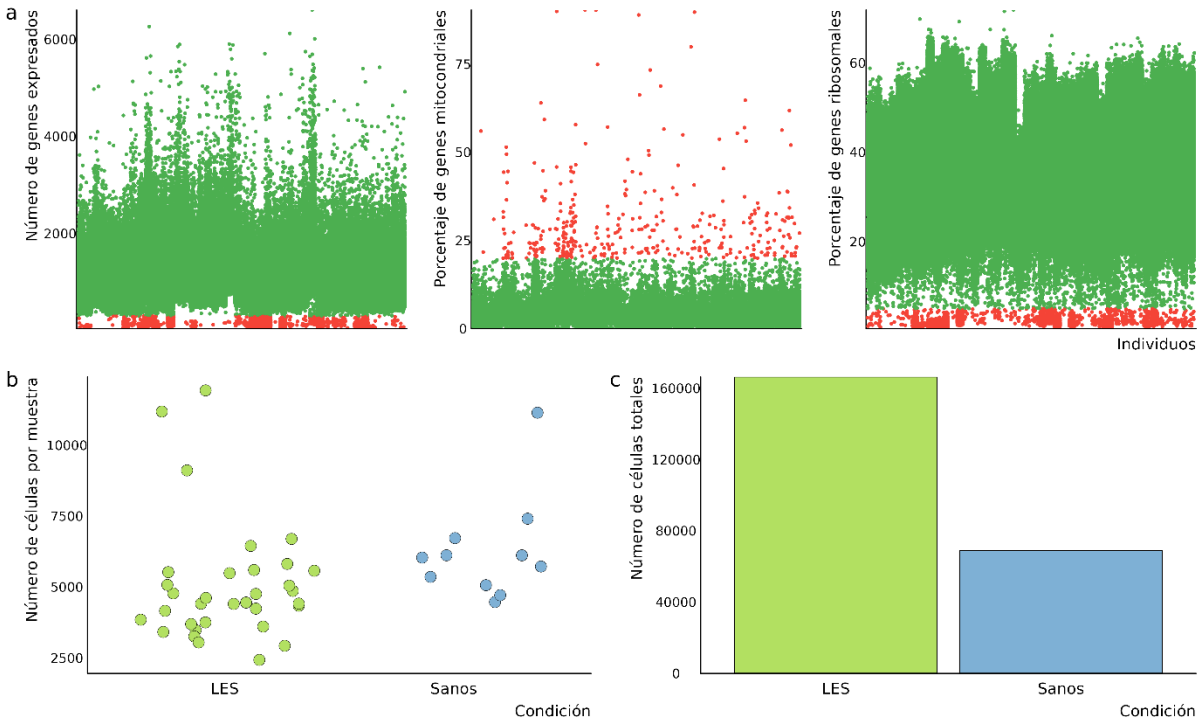

Figura 24. Vista general de los datos de scRNA-Seq. a) Valores que toma cada célula en cuanto a las variables que se han utilizado para eliminar aquellas de baja calidad. Cada uno de los paneles hace referencia a una de ellas, cuyos valores se localizan en el eje de ordenadas. En el eje de abscisas se disponen los individuos. Las células de color rojo son las que no han pasado o superan el umbral que se ha establecido de cada variable. b) Número de células por muestra, representado en un gráfico de puntos en los que se ha separado las muestras que pertenecen a individuos sanos y las que son de pacientes de LES. c) Similar al anterior, pero en este caso se representa el conjunto total de células de cada condición.

5.2 REDUCCIÓN DE DIMENSIONALIDAD

Uno de los pasos clave en los análisis de scRNA-Seq es la reducción de dimensionalidad. Se trata de un conjunto de técnicas que se emplean para analizar y visualizar grandes conjuntos de datos de expresión génica a nivel celular. Estas técnicas permiten reducir la dimensionalidad y complejidad de los datos. Los usos para los que se emplean estas técnicas son: identificar grupos de células de perfil de expresión similar (clustering o agrupamiento), identificar grupos de células que representan distintos fenotipos e identificar cambios en la expresión debido a respuesta a tratamientos o perturbaciones. Las técnicas más usadas incluyen el análisis de componentes principales (PCA), la t-*Distributed Stochastic Neighbor Embedding* (t-SNE) y *Uniform Manifold Approximation and Projection* (UMAP). Los pasos que se han seguido para realizar este proceso se resumen en la Figura 25a.

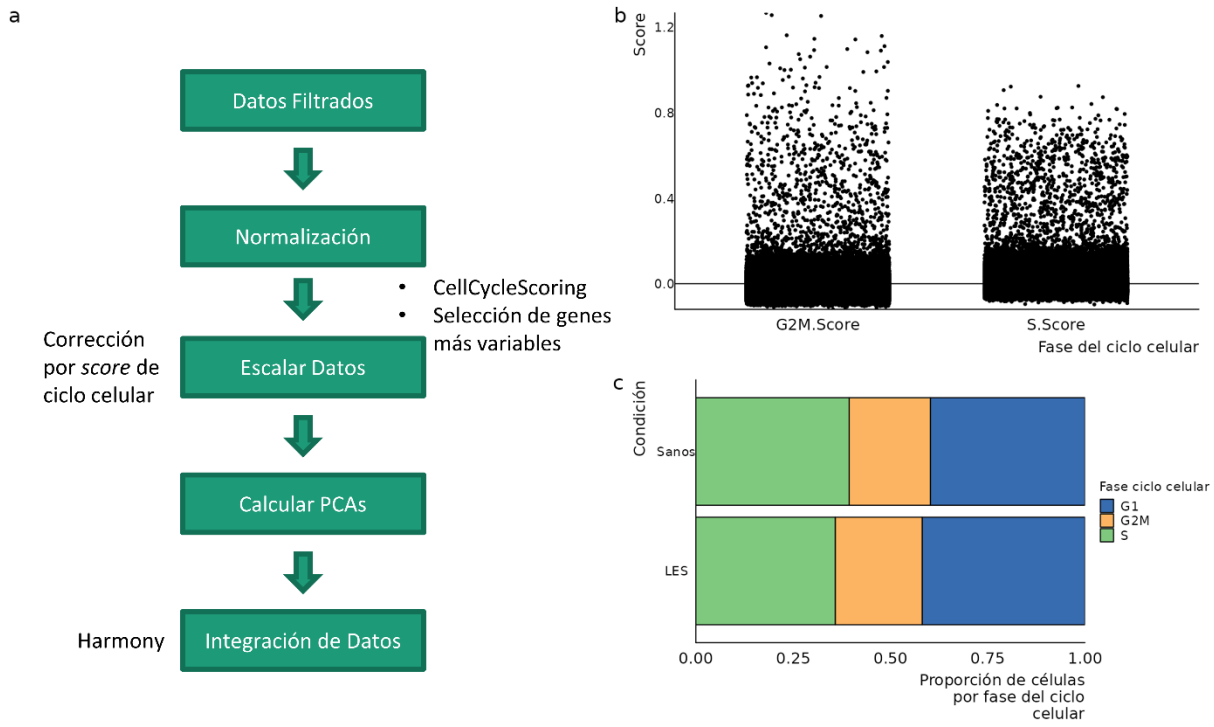

Figura 25. Reducción de dimensiones en datos de scRNA-Seq. a) Pasos llevados a cabo en el análisis de scRNA-Seq a partir de los datos filtrados del control de calidad hasta la reducción de dimensionalidad. Incluye los pasos de normalización, cálculo de las puntuaciones para ciclo celular, la corrección de las mismas junto al escalado de los datos, el cálculo de las PCAs y la integración de todos los datos con Harmony. b) Distribución de las puntuaciones obtenidas de *CellCycleScoring*, en donde se observa que muy pocas células se separan de 0. c) *CellCycleScoring* también etiqueta las células según la fase del ciclo celular en la que se encuentren. Vemos la distribución de estas etiquetas por condición.

Antes de iniciar la reducción de dimensionalidad, se llevó a cabo la normalización de los *counts*, utilizando la función *NormalizeData*, con los parámetros por defecto; *LogNormalize* como normalization.method (que lleva a cabo una transformación logarítmica natural de los *counts*) y 10000 como scale.factor (que es un factor que se utiliza para dividir los *counts* antes de normalizar).

Tras esto, calculamos una serie de variables que determinan el ciclo celular. Esto se realizó con la función *CellCycleScoring*, que determina en qué fase del ciclo celular se encuentra cada célula Calculamos una variable basada en la diferencia entre el *score* de la fase S y el *score* de la fase G2M. Debido a que muy pocas células tenían *scores* lejanos de 0 para las variables calculadas por *CellCycleScoring* (ver Figura 25b) y que la distribución de fases es similar (ver Figura 25c), decidimos no eliminar ninguna célula, aunque posteriormente tendremos en cuenta esta heterogeneidad.

El siguiente paso consiste en escalar los datos de los genes expresados de forma más variable para buscar clústeres basados en su expresión. Necesitamos seleccionar aquellos genes que se van a usar para hacer la reducción de dimensionalidad y el clustering posterior, que fueron seleccionados con la función *FindVariableFeatures* con los parámetros por defecto, es decir, buscando los 2000 genes con más variabilidad de todas las muestras. Seguidamente se realizó un escalado de los datos, con la función *ScaleData*, incluyendo en el parámetro var.regress la variable calculada en el paso anterior del ciclo celular, con el fin de corregir la heterogeneidad.

Una vez los datos están escalados, es necesario realizar la integración de las células. Esto supone que, como los datos proceden de individuos y *batch* diferentes los datos deben transformarse para que puedan ser comparables entre sí. Aunque el método estándar de hacer esto es utilizando la función *IntegrateData*, debido a problemas computacionales se utilizó la integración de la librería Harmony en el paquete de Seurat[206], para llevar a cabo la integración de los datos, a través de la función *RunHarmony*. Harmony es un algoritmo de corrección de lotes (comúnmente conocido como *batch correction*) que permite ajustar y eliminar los efectos de dichos lotes en las matrices de expresión de células individuales. Estos efectos de lotes son diferencias artificiales debido a las técnicas de procesamiento. Antes de ejecutar esta función se calculan los componentes principales (PCAs) sobre los datos (función *RunPCA*).

Para establecer el número de dimensiones de Harmony que se desean utilizar, primero se calcula el porcentaje de variación y los porcentajes de variación acumulados de cada dimensión. Para

identificar las dimensiones que queremos utilizar calculamos dos parámetros. Por un lado, se calcula que dimensión presenta un porcentaje acumulado mayor del 90% y una variación asociada inferior al 5%. Por otro lado, se determina la diferencia entre la variación de una dimensión con la subsiguiente. De estos dos parámetros, el número de dimensiones será el más bajo de ellos. Esta metodología se ha implementado siguiendo uno de los materiales de aprendizaje del equipo de Bioinformática de la Universidad de Harvard, disponible en: https://hbctraining.github.io/scRNA-seq/lessons/elbow_plot_metric.html.

Con las dimensiones que mejor explican nuestros datos, ejecutamos la función *RunUMAP* para estimar las dimensiones UMAP de las células y finalmente se buscan los vecinos más próximos con la función *FindNeighbors*, que nos servirá para buscar clústeres en los pasos sucesivos. Como para la sección anterior, el código completo se encuentra en el siguiente enlace: https://github.com/ralodo93/scRNA-Seq_SLE/blob/main/02_dim_red.R.

5.3 CLUSTERING E IDENTIFICACIÓN DE TIPOS CELULARES

Hasta ahora solo hemos trabajado con células sin saber qué tipo de célula es cada una, salvando las que fueron eliminadas por ser *doublets*. Para conocer que tipo celular es cada una de las células es necesario realizar un clustering de todas ellas, con el fin de determinar una serie de grupos de células similares en base a su perfil transcripcional e identificar qué tipo celular es mayoritario en cada grupo.

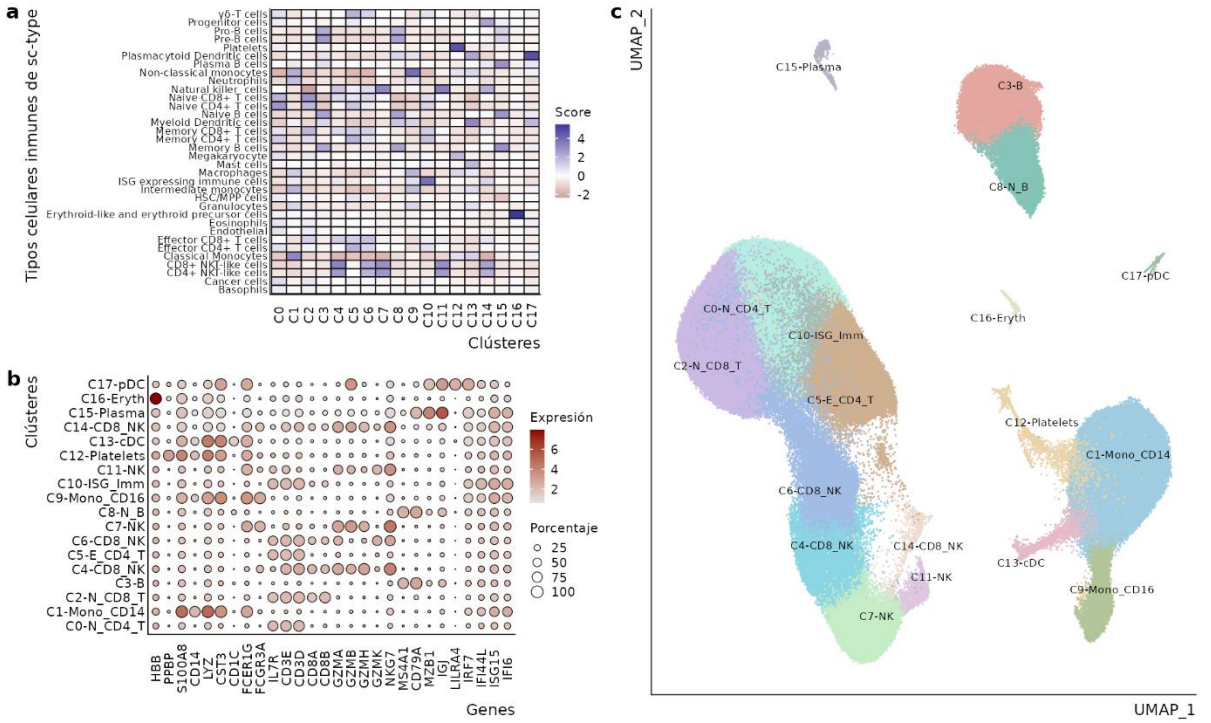

Figura 26. Descripción de los clústeres obtenidos en el análisis de scRNA-Seq con datos de LES y sanos. a) Puntuación obtenida a partir del método desarrollado por sc-type con el fin de determinar el tipo celular que se asocia mejor a cada uno de los clústeres. En el eje de ordenadas están todos los tipos celulares del sistema inmune incluidas en sc-type, en el eje de abscisas están los 18 clústeres que se han localizado y el color indica la puntuación otorgada por sc-type para cada combinación de clúster y tipo celular. Las puntuaciones altas se identifican con el color azul. b) Expresión de marcadores celulares canónicos para apoyar la estimación obtenida en sc-type. En el eje de ordenadas se distribuyen los clústeres y en el eje de abscisas una serie de genes canónicos de varios tipos celulares. El color indica la expresión media de las células de cada clúster por gen y el tamaño el porcentaje de células de un clúster que expresan dicho gen. c) Representación UMAP de la distribución espacial de los clústeres ya etiquetados según su tipo celular dominante.

Para llevar a cabo este paso se utiliza la función *FindClusters* a partir del objeto de Seurat obtenido en la sección anterior e introduciendo un valor de resolución que va a determinar cómo queremos sean los clústeres. La función por defecto aplica el algoritmo desarrollado por Louvain[207]. Un valor de resolución bajo nos va a devolver muy pocos clústeres y, generalmente,

muy grandes, mientras que un valor alto nos devuelve muchos clústeres, la gran mayoría muy pequeños. Se trata de un valor que oscila entre 0 y 1, por lo que cómo queríamos tener una visión global que nos permitiera identificar los tipos celulares principales, decidimos seleccionar un valor de 0.5, que además es el valor por defecto. Siguiendo todo el proceso que se ha descrito se obtienen un total de 18 clústeres (nombrados de 0 a 17). El código para asignar cada célula a un clúster concreto se encuentra alojado en el siguiente enlace: https://github.com/ralodo93/scRNA-Seq_SLE/blob/main/03_clustering.R.

Una vez obtenidos los clústeres, queríamos saber qué tipo celular era el predominante en cada cluster. Para ello se utilizó un software para R llamado sc-type[208], que contiene una muy completa colección de marcadores de tipos celulares para datos de scRNA-Seq. Además de dicha base de datos, en este trabajo han desarrollado un método que consiste en comparar la expresión génica de una célula con un conjunto de genes previamente conocidos calculando un valor de puntuación de cada conjunto de genes. Los conjuntos de genes (tipos de células) que presenten una mayor puntuación podrán ser seleccionados como tipo celular mayoritario en el clúster que se está testeando. Como nosotros tenemos datos de PBMCs, seleccionamos los marcados de células inmunes y aplicamos este método para identificar el tipo celular mayoritario en cada clúster. Según describen en la aplicación, la predominancia de un tipo celular se establece a través de la puntuación más alta. Siguiendo esta premisa, podemos clasificar los18 clústeres en varios tipos celulares de acuerdo con la puntuación obtenida al aplicar el método desarrollado por sc-type (ver Figura 26a).

Los clústeres se han identificado del siguiente modo: identificamos varios clústeres de células T, de células B, NK y células T citotóxicas similares a NK. Además, hemos hallado un clúster de monocitos clásicos y otro de no clásicos, un clúster de células dendríticas convencionales y otro de células dendríticas plasmacitoides, y clústeres individuales para células plasmáticas, eritrocitos y plaquetas. Para hacernos una idea de que clúster está fenotípicamente predominado por cada tipo celular es necesario recurrir a la Tabla 4.

Además de utilizar el método de sc-type, utilizamos una serie de genes canónicos, obtenidos de la publicación de Nehar-Belaid [128], para hacer una curación manual de la identificación de clústeres realizada con sc-type. Gracias a esta curación manual, vemos que el clúster 10 (C10), que se ha definido como clúster relacionado con células que presentan alta expresión de genes estimulados por interferón, realmente puede incluirse dentro del grupo de células T de tipo CD4 (ver Figura 26b). Una vez se han definido tanto los clústeres como el grupo celular principal en

cada uno de ellos, podemos generar la visualización más famosa de los estudios de scRNA-Seq, una representación espacial de cómo se disponen los clústeres, utilizando las métricas de UMAP que hemos comentado antes. Este gráfico se puede observar en la Figura 26c.

Tabla 4. Caracterización de los clústeres obtenidos a partir de los datos de scRNA-Seq. Se muestra de cada clúster la identificación del tipo celular dominante según los scores obtenidos de sc-type y el número de células que componen cada clúster.

Clúster	Identificación	Número de Células
C0	CD4 T Naïve	53667
C1	Mono CD14	42320
C2	CD8 T Naive	28692
C3	B	21322
C4	CD8 NK	20799
C5	CD4 T Efector	17784
C6	CD8 NK	11882
C7	NK	11872
C8	B Naive	7547
C9	Mono CD16	5990
C10	ISG Imm	3208
C11	NK	2266
C12	Plaquetas	2234
C13	cDC	2131
C14	CD8 NK	1076
C15	Plasma	1056
C16	Eritrocitos	894
C17	pDC	684

Queríamos caracterizar los clústeres en base al porcentaje de células de cada uno de ellos que pertenecen a cada individuo y a cada condición con el fin de comprobar que no hubiera ningún clúster que estuviera formado por células procedentes de uno o pocos individuos. En la Figura 27 tenemos una visión global de la distribución de todas las células en los diferentes clústeres. Específicamente, la Figura 27a muestra el porcentaje de células de cada individuo que se asocian a cada clúster. Adicionalmente, hemos generado la Figura 27b en la que vemos los mismos porcentajes de individuos en cada clúster, pero diferenciando entre casos y controles. Aquí se aprecia que hay algunos clústeres que muestran porcentajes de células bastante diferentes entre casos y controles. Hay algunos clústeres que están expandidos en células de LES, como los clústeres C1, C10, C15 y C16. Por otro lado, hay una serie de clústeres enriquecidos en células de individuos sanos, como los clústeres C6, C7, C11 y C17. Estos

resultados concuerdan con los obtenidos en el trabajo publicado por Nehar-Belaid. Finalmente, es posible comprobar la distribución global de células de LES y sanos en cada clúster en la Figura 27c.

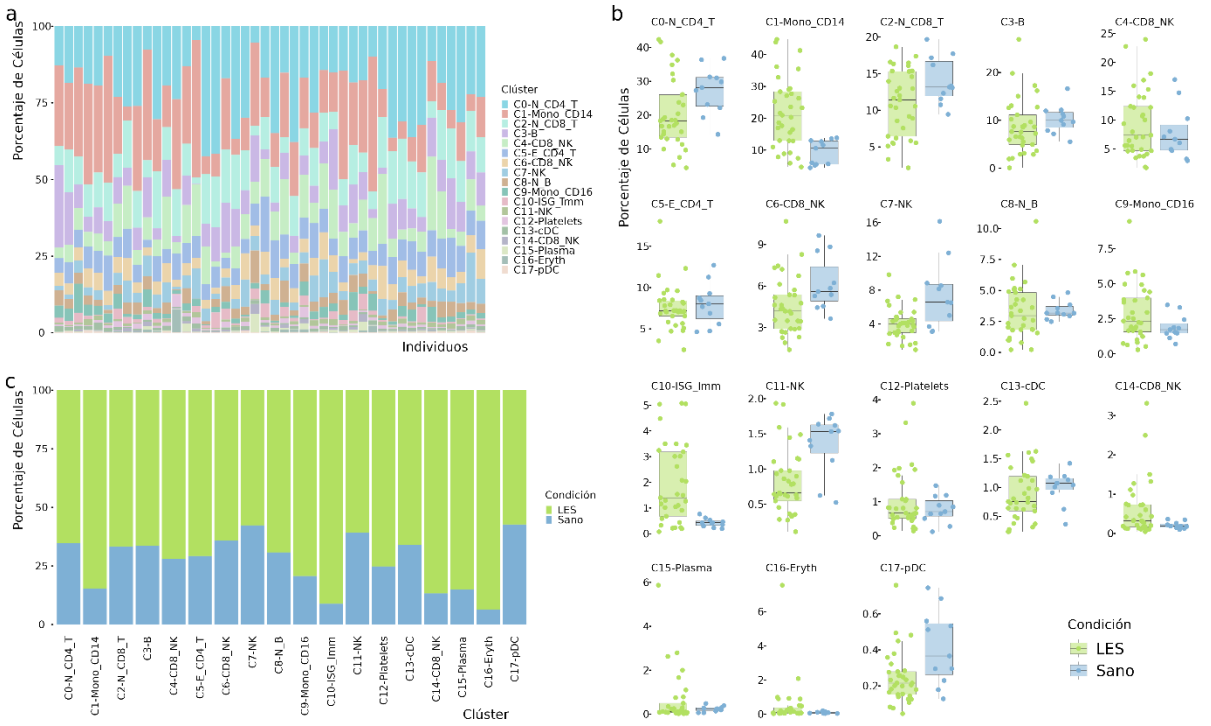

Figura 27. Caracterización de los clústeres obtenidos en los datos de scRNA-Seq. a) Distribución de los clústeres en cada individuo. b) Comparación en cuanto al porcentaje de células de cada individuo asociadas a cada clúster. Cada panel corresponde a un clúster diferente y se separan las muestras de LES y las sanas. c) Composición de cada clúster en términos de cantidad (en porcentaje) de células de ambas condiciones.

El código mediante el cual asignamos cada clúster a un tipo celular al cual se asemeja molecularmente está disponible en: https://github.com/ralodo93/scRNA-Seq_SLE/blob/main/04_identification.R.

5.4 Expresión Diferencial y Análisis Funcional

Los clústeres que se han identificado se han utilizado para realizar análisis de expresión diferencial entre casos y controles, así como inferencia de actividad de factores de transcripción y de rutas. Para ello se ha aplicado el mismo flujo de trabajo con cada clúster, es decir: aplicar un análisis de expresión diferencial de genes comparando la expresión de células de casos y controles y análisis de enriquecimiento con los TFs localizados en DoRothEA y con las rutas de PROGENy.

Sin embargo, antes de avanzar, se decidió realizar estos análisis mediante la técnica de *pseudobulk*. Esta técnica consiste en colapsar todos los conteos de cada individuo con el fin de crear una matriz de expresión que sea reflejo de la matriz de conteos típica en datos de RNA-Seq tradicional. Para ello, lo que se realizó es calcular dichas matrices para cada individuo y clúster. Por ejemplo, la matriz perteneciente al primer clúster constará un valor por cada individuo y gen, como suma total de la expresión de todas las células de dicho individuo que pertenecen a este clúster. El uso de esta técnica se considera una ventaja a la hora de realizar análisis de expresión diferencial ya que reducen considerablemente el tamaño muestral, de forma que los métodos estadísticos utilizados puedan trabajar sin problemas con estos tamaños[209]. Para llevar a cabo la transformación de datos de célula única a *pseudobulk* se han seguido los pasos utilizados por los desarrolladores de DoRothEA y PROGENy en el manual de uso de la librería decoupleR de Python (https://decoupler-py.readthedocs.io/en/latest/notebooks/pseudobulk.html).

5.4.1 Análisis de Expresión Diferencial de todos los Clústeres

Para realizar el análisis de expresión diferencial tomamos los *counts* totales de cada muestra en cada clúster a partir de la transformación de datos de célula única a *pseudobulk*. A partir de aquí, los análisis de expresión diferencial se llevan a cabo de forma independiente en todos los clústeres, comparando las muestras de LES y sanos como si se tratase de un análisis de expresión diferencial convencional con datos de RNA-Seq. Para ello usamos la librería DESeq2 y la transformación de *counts* a TMM, de forma similar a como se llevó a cabo en ADEx. Antes de realizar estos análisis nos damos cuenta que las células del clúster C16-Eryth pertenecen en su mayoría a un solo individuo y que el número de células de sanos es muy bajo. Por este motivo vamos a eliminar este clúster tanto de los análisis de expresión diferencial como del resto de análisis posteriores. Cómo los datos de scRNA-Seq se realizaron en varios lotes (conocido comúnmente como *batch*), el análisis de expresión diferencial se llevó a cabo corrigiendo por esta variable.

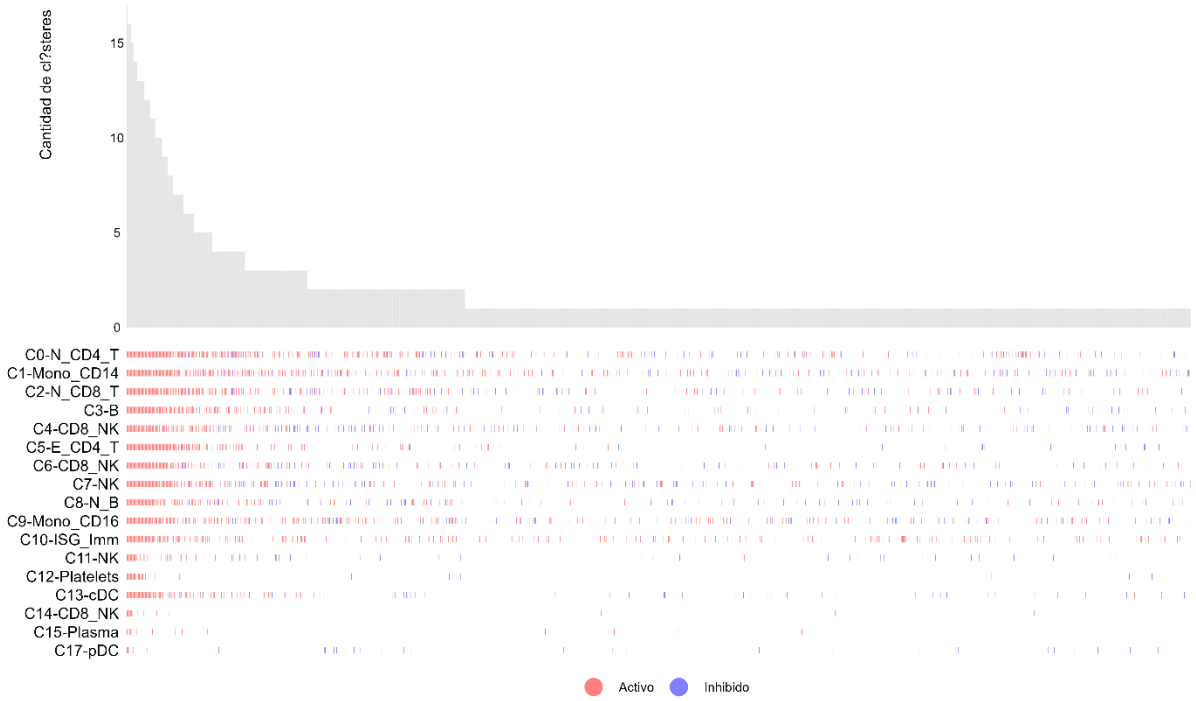

Figura 28. Gráfico de intersección (o *upset plot*) en el cual se representa el número de genes diferencialmente expresados (BH < 0.05, *fold change* absoluto > 0.5) que se comparten en cada combinación de clústeres. Los colores indican si dicho gen está sobre-expresado (activo; *fold change* > 0.5; color rojo) o infra-expresado (inhibido; *fold change* < -0.05; color azul).

Los genes diferencialmente expresados (BH < 0.05 y valor absoluto de *fold change* > 0.5) a lo largo de todos los clústeres forman un conjunto heterogéneo, lo cual se aprecia en la Figura 28, en la cual se observa la distribución de los genes diferencialmente expresados que comparten los clústeres entre sí. A pesar de dicha heterogeneidad, algunos genes son significativos en gran cantidad de clústeres. En la Tabla 5 se muestran los genes diferencialmente expresados que se localizan en más de 12 clústeres. La mayoría de estos genes pertenecen al grupo de genes que se asocian a la firma de interferón como: EPSTI1, IFI44, IFI6, ISG15, IFI27 o IFI44L, estando representados, algunos de ellos, en casi todos los clústeres. Esta firma de interferón, como hemos visto a lo largo de la tesis, no resulta en ningún resultado novedoso.

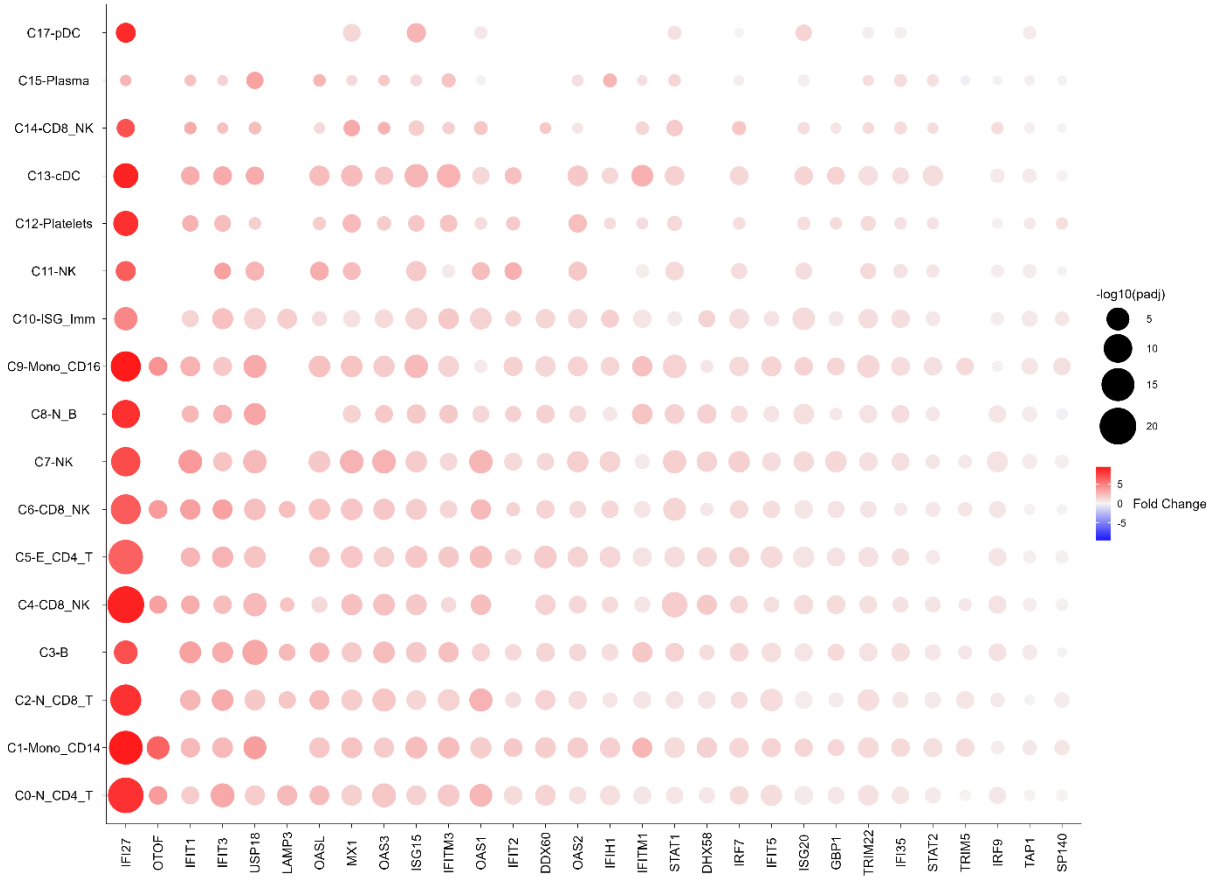

Figura 29. Resultados de expresión diferencial de los genes que pertenecen a la firma de interferón. Cada panel representa a uno de los clústeres obtenidos en el análisis de scRNA-Seq. Cada punto representa a un gen. El eje de abscisas refleja el *fold change* y el eje de ordenadas el -log(p-valor) obtenidos a partir del análisis de expresión diferencial. Las líneas de puntos paralelas de color rojo y azul indican la posición de los p-valores 0.05 y 0.01 respectivamente.

Tabla 5. Genes diferencialmente expresados (BH < 0.05 y *fold change* absoluto > 0.5) en más de 12 clústeres.

Gen	Número de clústeres en los que es significativo
EPSTI1	17
IFI44L	17
ISG15	17
LY6E	17
MX1	17
IFI27	16
IFI44	16
IFI6	16
PLSCR1	16
TYMP	16
EIF2AK2	15
IRF7	15
S100A8	15
XAF1	15
CMPK2	14
LGALS3BP	14
OAS2	14
RSAD2	14
STAT1	14
TRIM22	14
USP18	14
BST2	13
HERC5	13
IFI35	13
IFIT1	13
IFIT3	13
IFITM1	13
IFITM3	13
LAP3	13
MX2	13
PARP9	13
S100A9	13

Además, los genes asociados a interferón presentan una firma sobreexpresada en todos los clústeres. Para comprobar esto hemos obtenido la firma de genes de interferón a partir del trabajo de Banchereau y colaboradores[131]. En la Figura 29 podemos observar cómo se distribuyen estos genes en el análisis de expresión diferencial en todos los clústeres.

5.4.2 RUTA JAK-STAT EXTENDIDA EN LES

Conocer las rutas que están perturbadas en un grupo de muestras o una enfermedad concreta puede ser útil a la hora de conocer sobre que vías metabólicas o redes de señalización habría que investigar para conocer más en profundidad dicha enfermedad. En este aspecto, hace unos años se publicó la base de datos de PROGENy, de la cual ya hemos hablado previamente y que consiste en un conjunto de 14 rutas con información de las perturbaciones que están afectando, de modo que cada ruta tiene un abanico de genes a los que regula. Por ejemplo, el gen CBX6 tiene una respuesta negativa a la vía EGFR, lo que quiere decir que cuando hay una señalización de EGFR, este gen está infra expresado. Por otro lado, siguiendo con algún ejemplo, el gen RFC2 tiene una un peso positivo para EGFR, de modo que cuando tiene lugar la señalización EGFR, el gen tiende a sobreexpresarse.

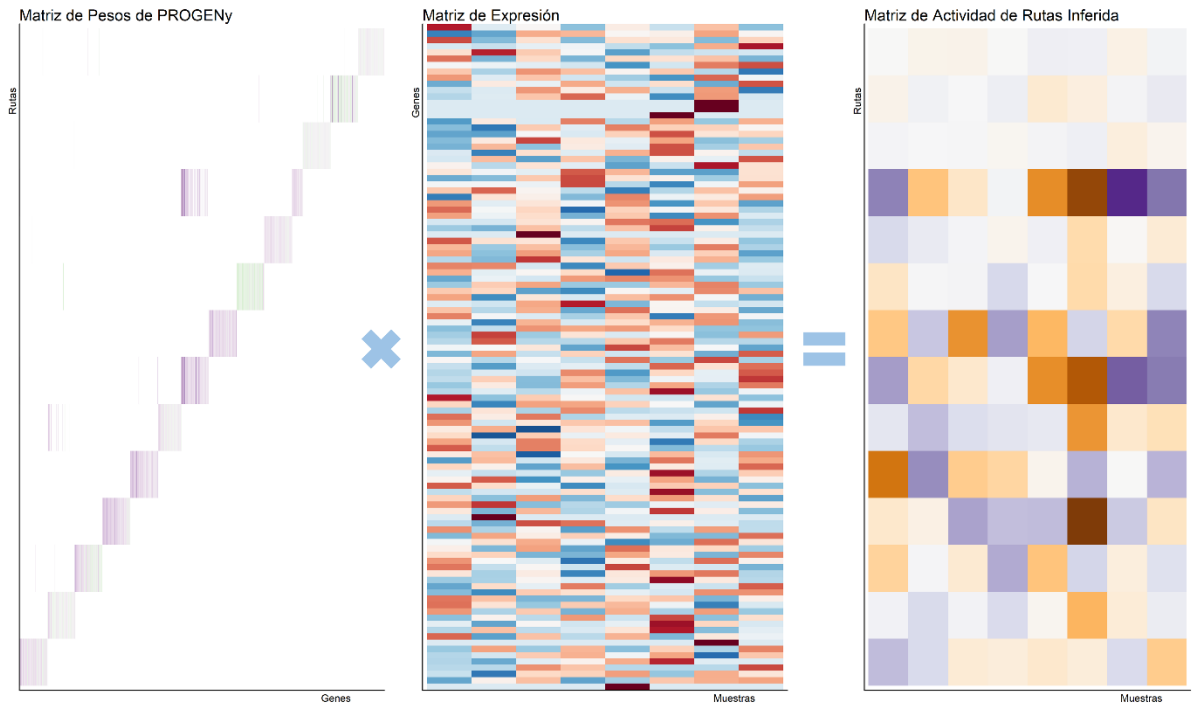

Figura 30. Esquema de la metodología empleada en PROGENy para inferir la actividad de las rutas de cada muestra en utilizando una matriz de pesos y la matriz de expresión de dichas muestras.

La base de datos de PROGENy contiene los genes que más influyen en cada ruta de modo que a cada gen se le asigna un peso, que como hemos descrito puede ser positivo o negativo. En nuestro caso hemos seleccionado la información de los 100 genes más relevantes de cada ruta. El método de inferencia de las actividades consiste en hacer una multiplicación de matrices,

utilizando la matriz de expresión normalizada y la matriz de pesos de cada ruta tal y como se indica en la Figura 30. El resultado de este proceso es una matriz de actividades de cada ruta y muestra.

Utilizando la matriz normalizada obtenida a partir de la transformación a *pseudobulk* que se ha descrito en el apartado anterior, realizamos un análisis de actividad diferencial entre casos y controles dentro de cada clúster mediante el protocolo estándar de la librería limma, con el fin de determinar las vías que presentan diferencias significativas entre las muestras de LES y sanos a nivel celular. Al seleccionar aquellas rutas que presentan un p-valor < 0.05, observamos que la ruta JAK-STAT se encuentra desregulada en la gran mayoría de los clústeres analizados, en todos ellos presentando una diferencia de actividad positiva entre casos y controles (Figura 31). Este resultado era el esperado pues, como hemos comentado previamente, la ruta JAK-STAT está muy relacionada a la firma de genes de interferón.

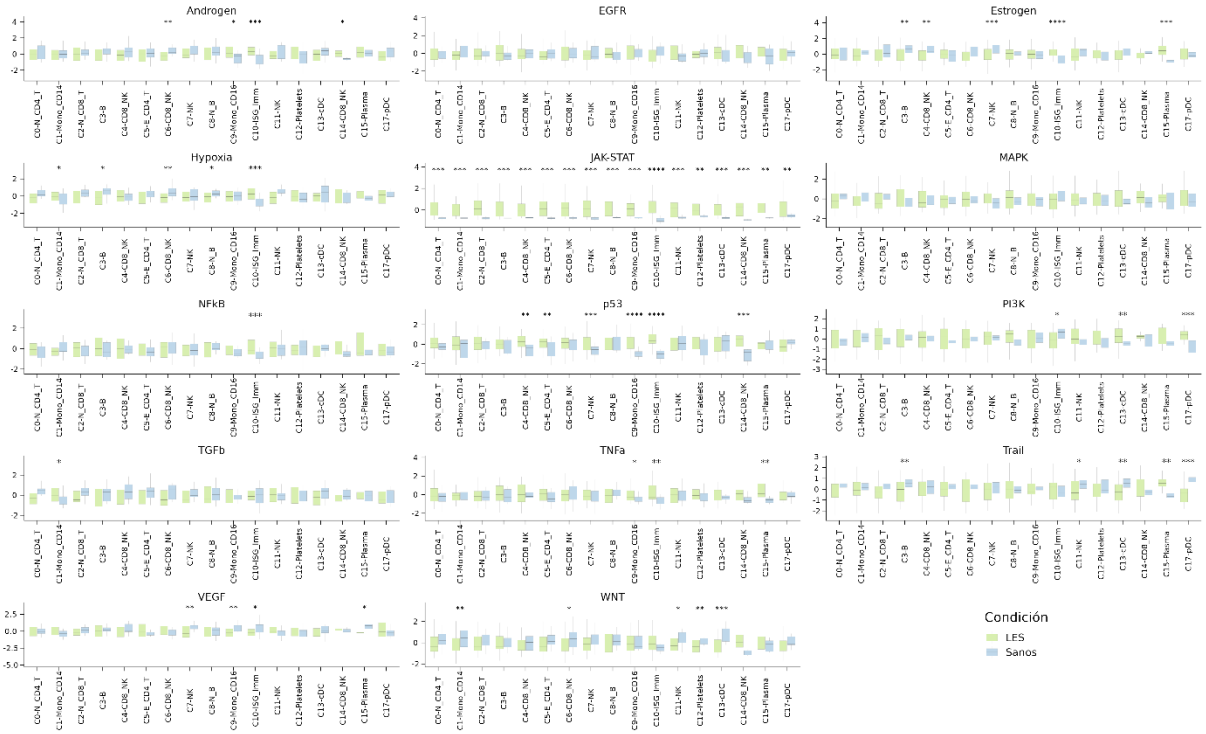

Figura 31. Resultados del análisis de actividad diferencial de rutas de PROGENy. En cada panel se representan las actividades de cada muestra, diferenciando entre casos y controles, en una determinada ruta de PROGENy. Los asteriscos indican la significancia siguiendo esta asignación: p-valor < 0.1 (*), p-valor < 0.05 (**), p-valor < 0.01 (***), p-valor < 0.001 (****).

Además de dicho patrón, observamos que en el clúster C10-ISG_Imm presenta diferencias significativas en varias rutas: p53, que es una conocida ruta relacionada con el cáncer,

encargada de procesos como la regulación del ciclo celular, la apoptosis, la supresión de tumores o la reparación de ADN, TNFa que interviene en los procesos inmunes o la regresión tumoral o NFkB que regula la respuesta inmune y la producción de citoquinas.

También observamos que los dos clústeres de células dendríticas presentan diferencias significativas en la ruta PI3K (con actividad mayor en LES) y la ruta Trail (con actividad menor en LES). Aunque ambas rutas se han asociado a LES, la primera de ellas como participante de en la apoptosis celular[210] y la segunda por su implicación en procesos proinflamatorios[211], hasta donde llega nuestro conocimiento no existe ninguna vinculación de estas rutas en estos tipos celulares específicos.

5.4.3 Inferencia de Actividades de TFs en Clústeres Celulares

La inferencia de las actividades de los TFs sigue una metodología similar a la que se realizó en el trabajo anterior (Figura 17). De igual forma que en el paso anterior y, siguiendo con las premisas aportadas por los autores de DoRothEA, los pasos a seguir en este análisis son: 1) seleccionar los regulones de DoRothEA que se van a incluir en el análisis, 2) utilizar la matriz de expresión normalizada para determinar la actividad inferida de los TFs y 3) considerar aquellos TFs que son relevantes en cada clúster. Al contrario que en el trabajo anterior, en el que solo seleccionamos los regulones del nivel A, queríamos abarcar mayor cantidad de información sin que eso supusiera una pérdida importante de credibilidad. Por un lado, queríamos mantener la seguridad de que los regulones fueran fiables y tuvieran evidencias publicadas en la bibliografía, pero queríamos extender un poco más la base de datos, ya que nuestra experiencia a la hora de analizar las dos cohortes de LES que se han descrito en el apartado anterior nos demostró que la profundidad de la base de datos era escasa. Por ello, decidimos utilizar los regulones de nivel C o superior (ver Figura 8a y b). De este modo, tenemos una base de datos balanceada entre calidad (los regulones A, B y C tienen al menos una evidencia experimental obtenida a partir de un ChIP-Seq y también están descritos en la literatura) y cantidad (alcanzando los 271 TFs, más de 5000 genes unidos por más de 13.000 interacciones). Esto está en consonancia con las recomendaciones de los autores de la librería de DoRothEA. Para inferir las actividades de los TFs se ha utilizado la función “run_viper” de la librería dorothea, que a su vez utiliza la función “viper” de la librería viper, mediante la matriz normalizada obtenida tras el proceso de transformación a *pseudobulk*.

Para comparar las actividades de los TFs utilizamos el protocolo estándar de limma, del mismo modo que con las actividades de PROGENy. De igual forma que con la expresión diferencial, existe gran heterogeneidad en los TFs significativos a lo largo de todos los clústeres (Figura 32). A pesar de dicha heterogeneidad, hay una serie de TFs que presentan diferencias significativas gran cantidad de los clústeres, cómo es el caso de IRF1, IRF2, IRF9, STAT1 y STAT2, que, como hemos indicado anteriormente regulan genes estimulados por interferón (Tabla 6).

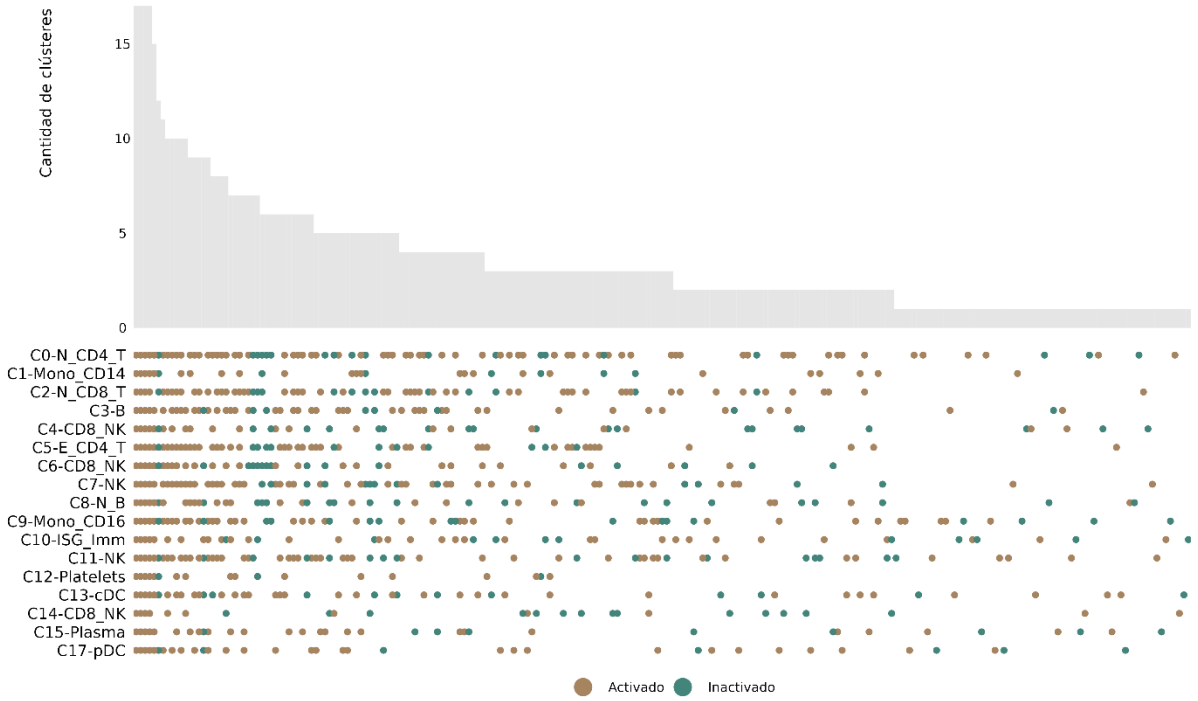

Figura 32. Gráfico de intersección en el cual se representa el número de TFs con actividad diferencial que se comparten en cada combinación de clústeres. Las barras indican el número de TFs y los puntos hacen referencia a los clústeres que se están combinando para determinar dicho número.

Tabla 6. Factores de transcripción significativos en más de 12 clústeres.

Factor de Transcripción	Número de clústeres en los que es significativo
STAT1	17
STAT2	17
IRF9	17
IRF1	17
IRF2	15

A la hora de representar los TFs que son significativos en los diferentes clústeres, observamos que, tal y como esperábamos, los TFs que están más implicados en la firma de regulón y que regulan genes inmersos en la red de señalización JAK-STAT, presentan actividades mucho más altas en las muestras de LES con respecto a las muestras sanas en todos los tipos celulares. Estos TFs, de los cuales ya hemos hablado, son: STAT2, STAT1, IRF9, IRF2 e IRF1. A pesar de que dicha firma es homogénea en todos los tipos celulares, es más que probable que, con los resultados obtenidos por Nehar-Behair y colaboradores[128], dicha firma de regulación se de solo en pequeños grupos celulares.

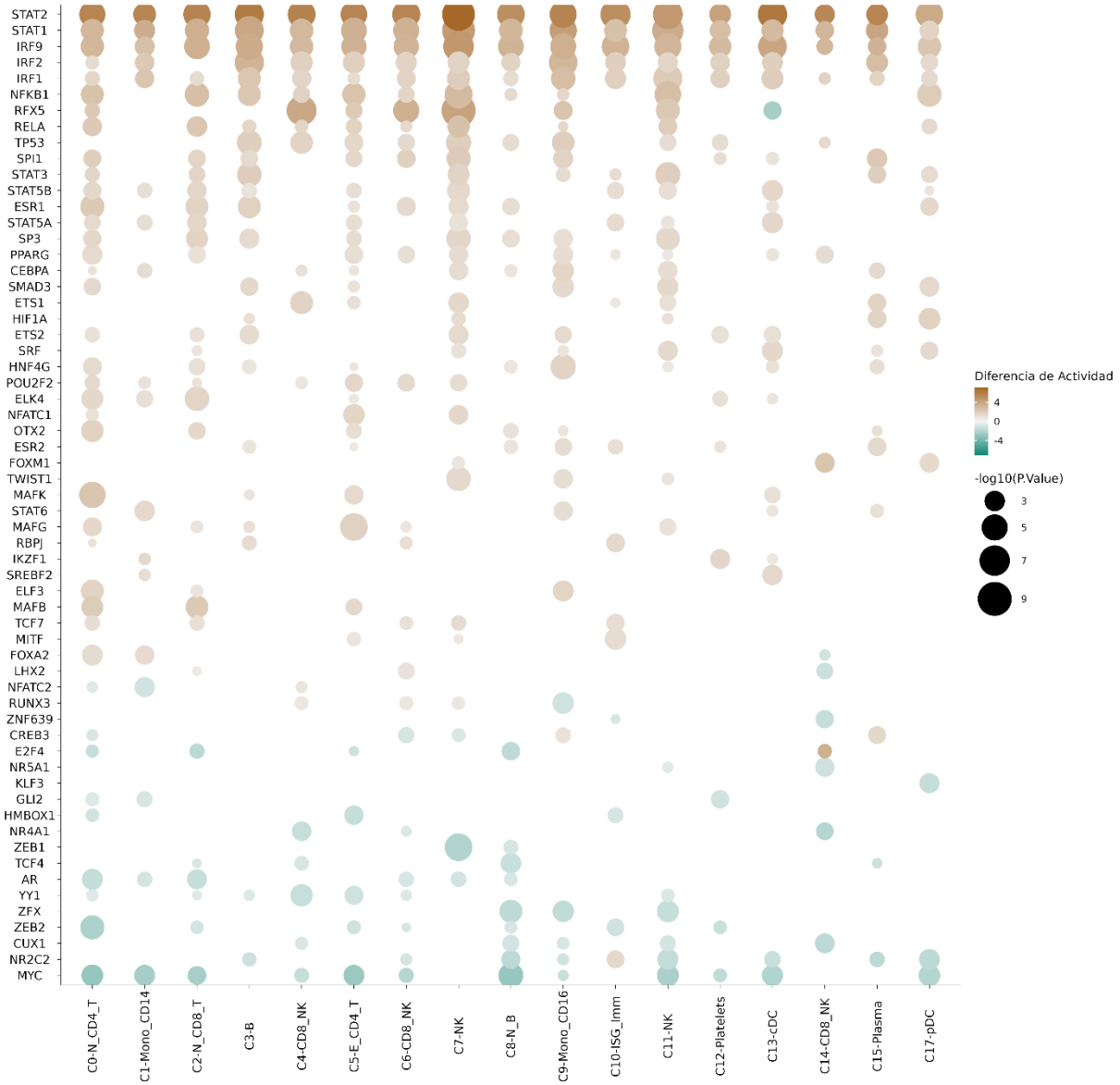

Figura 33. TFs diferencialmente activados entre LES y sanos. En el eje de abscisas están todos los clústeres analizados y en el eje de ordenadas los TFs que son significativos (y dentro del top 10) en al menos uno de los clústeres. El color indica la diferencia de actividad entre muestras de LES y sanos, marrón indica actividad mayor en LES y verde actividad menor en LES. El tamaño es la representación de la significancia a través del -log10(p-valor), de forma que a mayor tamaño mayor significancia (y menor p-valor).

Además de este patrón común, vemos algunos TFs que pueden resultar interesantes. Por ejemplo, RFX5 es significativo en varios clústeres, principalmente en aquellos relacionados con las células NK (C7-NK, C4-CD8_NK, C6-CD8-NK y C11-NK) presentando actividad mayor en muestras de LES que en sanos. Este TF se caracteriza por activar varios de los genes del complejo mayor de histocompatibilidad tipo II (MHC II) y que, por lo tanto, se encuentran desregulados en este tipo celular. Otro de los TFs que nos resultaron interesantes es MYC. MYC es un TF que regula a muchísimos genes. De hecho, en la base de DoRothEA que se ha utilizado hay un total de 386 genes regulados por dicho TF. Esto hace que sea imposible encontrar un elemento común de regulación, ya que podríamos decir que está regulando genes implicados en una gran cantidad de funciones biológicas. Sin embargo, vemos que las actividades de este TFs son diferencialmente significativas entre LES y sanos, con una actividad menor en las muestras de LES para varios de los clústeres analizados. Resulta curioso que presente actividad diferencial en todos los clústeres de linfocitos T salvo el C10-ISG_Imm, en uno de los clústeres de linfocitos B (C8-N_B) o en uno de los clústeres de células NK (C11-NK), presentando significancia también en ambos clústeres de células dendríticas y algunas células CD8 citotóxicas (C4-CD8_NK y C6-CD8_NK). Como hemos comentado, resulta muy complejo identificar las funciones biológicas que se ven alteradas por la desregulación de los genes que son controlados por MYC por lo que llevamos a cabo una serie de análisis de enriquecimiento con los genes diana de este TF utilizando el método GSEA implementado en el paquete fgsea. Para ello utilizamos el *fold change* de los genes diana de MYC para ordenar la lista de genes y testeamos dicha firma con la base de datos de REACTOME. Para determinar los genes relevantes de cada ruta, seleccionamos solo aquellos que forman parte del *leading edge* de cada una.

Como se aprecia en la Figura 34 este TF está regulando multitud de genes que codifican proteínas que forman parte de los ribosomas (genes RPL y RPS) que se encuentran inhibidos en las muestras de LES de los clústeres en los que la actividad de MYC es significativa entre casos y controles. Estos genes, como es lógico, se encuentran muy vinculados a funciones celulares como la translación, el procesamiento de ARNr o *Nonsense-mediated decay* que es un proceso de control de calidad del ARNm para impedir la traslación en contextos anormales.

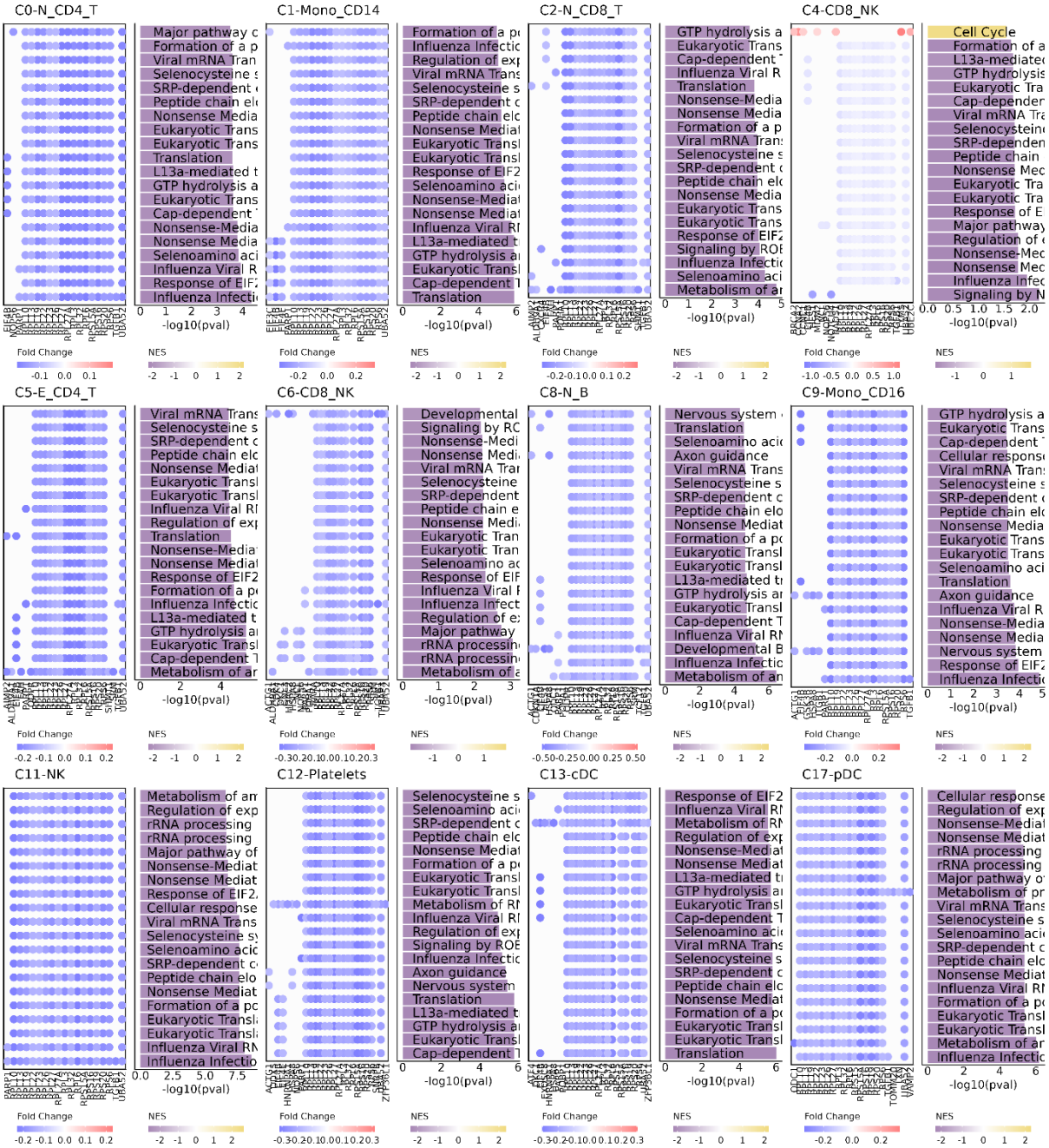

Figura 34. Análisis de enriquecimiento de los genes diana del TF MYC en cada clúster. Cada panel corresponde a un clúster y está formado por dos figuras. La figura de la derecha es el resultado del análisis de enriquecimiento funcional con los genes diana de MYC usando fgsea. La longitud de las barras indica la significancia (-log10(p-valor)) y el color muestra el valor de enriquecimeinto obtenido en el GSEA, morado negativo y amarillo positivo. La mayoría de las rutas muestran un color morado indicando que esa ruta está inactiva en LES con respecto a controles. En la subfigura de la izquierda se muestran el fold change de los genes (eje de abscisas) implicados en cada una de las rutas (eje de ordenadas).

Aunque no es la primera vez que se describe esta firma de genes ribosomales diferencialmente infraexpresados[96,212], no existe ninguna evidencia de porqué esta diferencia se da tan solo en algunos clústeres celulares y no en todos o al menos en los mismos.

5.5 Inferencia de Actividades en Tipos Celulares de Lupus

En este trabajo queríamos refutar la relevancia de los distintos tipos celulares a partir de la inferencia de actividades de rutas o vías y TFs utilizando las técnicas desarrolladas por PROGENy y VIPER, respectivamente. Para ello usamos las bases de datos de perturbaciones de rutas reclutadas por los propios desarrolladores de PROGENy así como la base de datos de interacciones de TF y genes diana de DoRothEA.

Para ello hemos analizado un conjunto de datos de RNA-Seq en célula única que se publicó en el año 2020 y cuyos datos están disponibles en NCBI GEO. La primera parte de este trabajo consta de la selección y depuración de las infinitas posibilidades a nivel de procedimientos para analizar este tipo de datos. Los métodos más conocidos se aplican en Python, a través de la librería scanpy[213] o en R, con las librerías Seurat y SingleCellExperiment[205,214]. Para este análisis se ha seleccionado la librería de Seurat, aunque el artículo original utiliza el módulo de scanpy y el código que utilizan está disponible, no fuimos capaces de replicar dicho código ya que encontramos problemas de compatibilidad con las librerías.

Recurrimos a varios tutoriales, guías y flujos de trabajo para analizar este tipo de datos con el fin de determinar los pasos que queríamos dar en el preprocesamiento de los datos que se resumen en tres pasos clave: control de calidad, reducción de dimensionalidad y clustering e identificación de tipos celulares. En resumen, la primera de ellas consiste en eliminar aquellas células de baja calidad o posibles dobletes, la segunda, que es la más costosa computacionalmente, lleva a cabo varios pasos con el fin de determinar un conjunto de genes que se utilizan para agrupar las células y que incluyen la normalización, escalar y corregir teniendo en cuenta la fase del ciclo celular de las muestras e integrar los datos para que sean comparables entre sí. Finalmente aplicamos una técnica de clustering para determinar el número de clústeres en los que se dividen las células e identificamos los tipos celulares predominantes de cada grupo tanto computacionalmente (con sc-type) como manualmente (con la expresión de genes canónicos).

Tras obtener los grupos celulares con los que vamos a trabajar, realizamos algunos análisis exploratorios de los mismos, que incluyen la visualización espacial de los clústeres mediante UMAPs (Figura 26c) así como la caracterización de los mismos (Figura 27). En este aspecto, en la Figura 27b se observa cómo se distribuyen los tipos celulares en las muestras de LES y controles sanos. Vemos algunos clústeres en los que hay mayor proporción de células de una

condición o de la otra, pero todos estos resultados concuerdan con lo que se conoce en la bibliografía, así como la descripción de los clústeres realizada en el artículo original. En este aspecto vemos que algunos de los clústeres enriquecidos en células de tipo NK (C6, C7 y C11) se encuentran expandidos en células sanas con respecto a células de LES, lo cual ya está descrito en la bibliografía[215]. Además, también se advierten diferencias en cuento a las proporciones de células dendríticas plasmacitoides (C17) presentes en mayor cantidad en muestras de sanos, lo cual va en consonancia con la literatura[216].

Por otro lado, hay otros clústeres que presentan proporciones mayores de células en los individuos de LES, como es el caso del C1 (Monocitos CD14), C9 (Monocitos CD16) o C10 (células inmunes que expresan interferón). El caso del C10 no resulta sospechoso dado el historial de esta enfermedad con la firma de interferón y la proporción desbalanceada de monocitos también se ha analizado previamente[217].

Una vez se han definido los clústeres, tanto a nivel de tipo celular predominante como a nivel de caracterización, se realizó un análisis funcional de cada clúster independiente que incluía tres apartados; expresión diferencial, inferencia de rutas de PROGENy e inferencia de TFs. Previo a iniciar este análisis, decidimos utilizar la técnica de *pseudobulk*, basándonos en recomendaciones de la literatura según las cuales esta técnica es más precisa a la hora de comparar entre condiciones dentro de un mismo clúster, que es justo nuestro objetivo. El problema de comparar los datos a nivel de célula es que, en la mayoría de los casos, la gran cantidad de células disponibles en cada clúster hace que los test estadísticos que se aplican pierdan fiabilidad debido al gran tamaño muestral. Para convertir los datos de expresión de célula única a *pseudobulk*, aplicamos la función get_pseudobulk del módulo de Python decoupler. De este modo, convertimos los datos de *single-cell* a una matriz de conteos con la suma de los *counts* de cada célula de un individuo que pertenece a cada clúster.

Con la matriz de *pseudobulk*, procedemos a realizar un análisis de expresión diferencial estándar, similar al que se llevó a cabo en la aplicación de ADEx, utilizando DESeq2 y una normalización por el método de TMM. Los resultados de expresión diferencial reportaron, sin ninguna sorpresa, que en todos los clústeres celulares se expresaba de forma anormal la firma de genes de interferón (ver Figura 29 y Tabla 5). Sin embargo, todos estos procedimientos se llevaron a cabo para desembocar en la inferencia de actividades de rutas y TFs a partir de los datos transcripcionales mediante las matrices de *pseudobulk* normalizadas.

Para el caso de PROGENy, seleccionamos los 100 genes más influyentes de cada ruta y calculamos la actividad de cada ruta por muestra en cada clúster. De igual modo, seleccionamos los regulones de DoRothEA de los niveles A, B y C para inferir la actividad de los TFs por muestra. Por tanto, en cada clúster tenemos dos matrices de actividad, de rutas y TFs, con muestras sanas y muestras de LES. La comparación de dichas actividades entre casos y controles se realizó con limma, tanto la actividad de las rutas como la actividad de los TFs.

Con respecto a las rutas, vemos una extensión total de la ruta JAK-STAT en las muestras de LES de todos los tipos celulares, siendo la ruta con mayor significancia y la que presenta mayores diferencias entre casos y controles. Esto no resulta para nada extraño, pues, como ya se ha definido, esta ruta incluye muchos de los genes asociados a la firma de interferón. Además de este patrón, vemos una serie de rutas que muestran significancia en tipos celulares o clústeres concretos, como es el caso de la ruta p53 que es significativa en hasta 6 clústeres, mostrando en todos ellos actividades mayores en muestras de LES. La relevancia del gen TP53 en LES no está muy extendida, si bien es cierto que se ha propuesto como un supresor de la autoinmunidad[218] pero en LES se ha localizado un autoanticuerpo que inhibe la función supresora de TP53[219]. Destacamos también la sobreactivación e inactivación, respectivamente de las rutas PI3K y Trail en los dos tipos de células dendríticas que se han analizado, las convencionales y las plasmacitoides. Aunque ambas rutas se han asociado a LES, la primera de ellas como participante de en la apoptosis celular[210] y la segunda por su implicación en procesos proinflamatorios[211], hasta donde llega nuestro conocimiento no existe ninguna vinculación de estas rutas en estos tipos celulares específicos. En el caso específico de la sobreactivación de la ruta PI3K en las células pDC, si que se conoce que dicha ruta es vital para la activación de la producción de interferón por estas células[220] y se ha propuesto como potencial terapia para inhibir la actividad patogénica de interferón en enfermedades autoinmunes.

Finalmente procedimos a llevar a cabo el análisis de actividad diferencial de los TFs entre casos y controles. No nos sorprendió que nuevamente vimos que la firma de TFs que regulan los genes de interferón (IRF1, IRF2, IRF9, STAT1 y STAT2, principalmente) presentaban actividades mucho más altas en las muestras de lupus en todos los tipos celulares. Además, localizamos algunos TFs que, si bien no presentan una firma robusta a lo largo de todos los clústeres, sí que tienen diferencias de actividad específicas entre LES y controles sanos.

RFX5 es significativo en varios clústeres, principalmente en aquellos relacionados con las células NK (C7-NK, C4-CD8_NK, C6-CD8-NK y C11-NK) presentando actividad mayor en

muestras de LES que en sanos. Este TF se caracteriza por activar varios de los genes del complejo mayor de histocompatibilidad tipo II (MHC II) y que, por lo tanto, se encuentran desregulados en este tipo celular. Esto se relaciona con subtipos de células NK que expresan esta firma génica y que son denominados células HLA-DR^{+} NK y MHCII^{+} NK. Aunque el mecanismo no es del todo claro y aún se desconoce si estos subtipos celulares son protectores o causantes de la enfermedad, parecen tener un papel aumentando la capacidad citotóxica de dichas células como induciendo un tipo de enfermedad similar a LES en modelos de ratón[221,222].

Otro de los TFs que nos resultaron interesantes es MYC. MYC es un TF que regula a muchísimos genes (hasta 386 genes según la base de datos de DoRothEA que hemos usado. Tal cantidad de genes hace muy difícil dilucidar una función o conjunto de funciones comunes en las que participen los genes regulados por MYC. Vemos que las actividades de este TFs son diferencialmente significativas entre LES y sanos, con una actividad menor en las muestras de LES para varios de los clústeres analizados. Como se ha comentado, resulta significativos para algunos de los clústeres, pero no sigue un patrón común entre tipos celulares. Por ejemplo, es significativo en uno de los clústeres de células B (C8-N_B) pero no en el otro (C3-B), fenómeno que ocurre también con los linfocitos T, las células T citotóxicas o las NK. Para comprobar las rutas o procesos biológicos que se verían alterados por la diferencia de actividad de este TF, utilizamos sus genes diana y la información que tenemos del análisis de expresión diferencial, aplicando un GSEA con dichos genes. Como se aprecia en la Figura 34 este TF está regulando multitud de genes que codifican proteínas que forman parte de los ribosomas (genes RPL y RPS) que se encuentran inhibidos en las muestras de LES de los clústeres en los que la actividad de MYC es significativa entre casos y controles. Estos genes, como es lógico, se encuentran muy vinculados a funciones celulares como la translación, el procesamiento de ARNr o *Nonsense-mediated decay* que es un proceso de control de calidad del ARNm para impedir la traslación en contextos anormales. Aunque esta firma se ha determinado previamente en LES aún existen lagunas acerca de por qué ocurre o si es causa o consecuencia de la enfermedad.

El objetivo de este trabajo se ha cumplido de forma parcial pues en el futuro próximo nos gustaría hacer un análisis de tipo subclustering sobre algunos de los tipos celulares de interés para comprobar si las actividades de rutas y TFs se dan a lo largo de todos los subtipos celulares o si por el contrario solo aparecen en un grupo reducido de ellas, de forma similar a como el artículo principal localizó la firma de interferón en pequeñas poblaciones celulares dentro de las mayoritarias.

6 CONCLUSIONES

A continuación, se describen las conclusiones más relevantes de esta tesis doctoral:

1. El desarrollo de la herramienta online de ADEx se ha establecido como un recurso útil para el análisis de datos omicos en el campo de enfermedades autoinmunes, ya que pone a disposición de la comunidad científica datos procesados uniformemente sin necesidad de tener habilidades bioinformáticas.
2. La inferencia de las actividades de los factores de transcripción a partir de los niveles de expresión de sus genes diana, por medio de VIPER y el uso de la base de datos de DoRothEA, nos permitió profundizar en el conocimiento de las redes de regulación relevantes en lupus.
3. Específicamente, las actividades de los TFs nos permitieron identificar dos subtipos de pacientes de lupus que han sido validados en dos cohortes independientes. Estos dos grupos difieren principalmente en las proporciones de linfocitos y neutrófilos.
4. Gracias al análisis diferencial de pacientes de lupus y controles sanos, identificamos 14 TFs que presentan diferencias significativas entre ambas condiciones, destacando principalmente aquellos relacionados con la regulación de la firma de interferón.
5. El estudio de expresión en célula única, nos permitió identificar 18 clústeres de células, los cuales se asignaron a los tipos celulares predominantes mediante técnicas automáticas (sc-type) y curación manual.
6. En el análisis de célula única se identificó la ruta JAK-STAT como diferencialmente activada en todos los tipos celulares de lupus, principalmente debido a la expresión diferencial de los genes de interferón así como TFs que regulan los genes de interferón (IRF1, IRF2, IRF9, STAT1 y STAT2) como aquellos que presentaban diferencias más significativas entre lupus y controles. Estos hallazgos concuerdan con lo conocido en el campo, pero también encontramos algunos patrones específicos en ciertos tipos celulares, remarcando de forma especial la desregulación del factor de transcripción MYC, el cual tiene actividad negativa en lupus que por el análisis de enriquecimiento funcional se observó que se debe a la infraexpresión de un conjunto de genes ribosomales.

7 ÍNDICE DE FIGURAS

8 ÍNDICE DE TABLAS

9 REFERENCIAS

1. Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat Med*. 2020;26(1):29-38. doi:10.1038/s41591-019-0727-5

2. Dahm R. Friedrich Miescher and the discovery of DNA. *Dev Biol*. 2005;278(2):274-288. doi:10.1016/j.ydbio.2004.11.028

3. Franklin RE, Gosling RG. Molecular Configuration in Sodium Thymonucleate. *Nature*. 1953;171(4356):740-741. doi:10.1038/171740a0

4. Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*. 1953;171(4356):737-738. doi:10.1038/171737a0

5. Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. *Science*. 2022;376(6588):44-53. doi:10.1126/science.abj6987

6. Venter JC, Smith HO, Adams MD. The Sequence of the Human Genome. *Clin Chem*. 2015;61(9):1207-1208. doi:10.1373/clinchem.2014.237016

7. Bumgarner R. Overview of DNA Microarrays: Types, Applications, and Their Future. *Curr Protoc Mol Biol*. 2013;101(1):22.1.1-22.1.11. doi:10.1002/0471142727.mb2201s101

8. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74(12):5463-5467.

9. Dijk EL van, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends Genet*. 2014;30(9):418-426. doi:10.1016/j.tig.2014.07.001

10. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;431(7011):931-945. doi:10.1038/nature03001

11. Lewin HA, Robinson GE, Kress WJ, et al. Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci*. 2018;115(17):4325-4333. doi:10.1073/pnas.1720115115

12. Check Hayden E. Technology: The $1,000 genome. *Nature*. 2014;507(7492):294-295. doi:10.1038/507294a

13. Gibbs RA, Belmont JW, Hardenbol P, et al. The International HapMap Project. *Nature*. 2003;426(6968):789-796. doi:10.1038/nature02168

14. Walter K, Min JL, Huang J, et al. The UK10K project identifies rare variants in health and disease. *Nature*. 2015;526(7571):82-90. doi:10.1038/nature14962

15. Marx V. The DNA of a nation. *Nature*. 2015;524(7566):503-505. doi:10.1038/524503a

16. Collins FS, Varmus H. A New Initiative on Precision Medicine. *N Engl J Med*. 2015;372(9):793-795. doi:10.1056/NEJMp1500523

17. Peña-Chilet M, Roldán G, Perez-Florido J, et al. CSVS, a crowdsourcing database of the Spanish population genetic variability. *Nucleic Acids Res*. 2021;49(D1):D1130-D1137. doi:10.1093/nar/gkaa794

18. Pearson TA, Manolio TA. How to Interpret a Genome-wide Association Study. *JAMA*. 2008;299(11):1335-1344. doi:10.1001/jama.299.11.1335

19. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. *PLOS Comput Biol*. 2017;13(5):e1005457. doi:10.1371/journal.pcbi.1005457

20. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57-63. doi:10.1038/nrg2484

21. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 2016;17(1):13. doi:10.1186/s13059-016-0881-8

22. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357-359. doi:10.1038/nmeth.1923

24. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324

25. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323. doi:10.1186/1471-2105-12-323

26. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods*. 2013;10(1):71-73. doi:10.1038/nmeth.2251

27. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol*. 2014;32(5):462-464. doi:10.1038/nbt.2862

28. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525-527. doi:10.1038/nbt.3519

29. Ibeagha-Awemu EM, Zhao X. Epigenetic marks: regulators of livestock phenotypes and conceivable sources of missing variation in livestock improvement programs. *Front Genet*. 2015;6. Accessed January 4, 2023. https://www.frontiersin.org/articles/10.3389/fgene.2015.00302

30. Orlando DA, Guenther MG, Frampton GM, Young RA. CpG island structure and trithorax/polycomb chromatin domains in human cells. *Genomics*. 2012;100(5):320-326. doi:10.1016/j.ygeno.2012.07.006

31. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*. 2008;9(6):465-476. doi:10.1038/nrg2341

32. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98(4):288-295. doi:10.1016/j.ygeno.2011.07.007

33. Li Y, Tollefsbol TO. DNA Methylation Detection: Bisulfite Genomic Sequencing Analysis. In: Tollefsbol TO, ed. *Epigenetics Protocols*. Methods in Molecular Biology. Humana Press; 2011:11-21. doi:10.1007/978-1-61779-316-5_2

34. Mundade R, Ozer HG, Wei H, Prabhu L, Lu T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle*. 2014;13(18):2847-2852. doi:10.4161/15384101.2014.949201

35. Nakato R, Sakata T. Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods*. 2021;187:44-53. doi:10.1016/j.ymeth.2020.03.005

36. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin Transl Med*. 2022;12(3):e694. doi:10.1002/ctm2.694

37. Papatheodorou I, Moreno P, Manning J, et al. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res*. 2020;48(D1):D77-D83. doi:10.1093/nar/gkz947

38. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318-1330. doi:10.1126/science.aaz1776

39. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47-e47. doi:10.1093/nar/gkv007

40. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

41. Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. *Nat Rev Genet*. 2014;15(1):34-48. doi:10.1038/nrg3575

42. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philos Trans R Soc B Biol Sci*. 2013;368(1620):20120362. doi:10.1098/rstb.2012.0362

43. Li Z, Heng J, Yan J, et al. Integrated analysis of gene expression and methylation profiles of 48 candidate genes in breast cancer patients. *Breast Cancer Res Treat*. 2016;160(2):371-383. doi:10.1007/s10549-016-4004-8

44. Toro-Domínguez D, Villatoro-García JA, Martorell-Marugán J, Román-Montoya Y, Alarcón-Riquelme ME, Carmona-Sáez P. A survey of gene expression meta-analysis: methods and applications. *Brief Bioinform*. 2021;22(2):1694-1705. doi:10.1093/bib/bbaa019

45. Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. 2014;11(3):333-337. doi:10.1038/nmeth.2810

46. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. 2009;25(22):2906-2912. doi:10.1093/bioinformatics/btp543

47. Velten B, Braunger JM, Argelaguet R, et al. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat Methods*. 2022;19(2):179-186. doi:10.1038/s41592-021-01343-9

48. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25-29. doi:10.1038/75556

49. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017;45(D1):D353-D361. doi:10.1093/nar/gkw1092

50. Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2020;48(D1):D498-D503. doi:10.1093/nar/gkz1031

51. Martens M, Ammar A, Riutta A, et al. WikiPathways: connecting communities. *Nucleic Acids Res*. 2021;49(D1):D613-D621. doi:10.1093/nar/gkaa1024

52. Huang R, Grishagin I, Wang Y, et al. The NCATS BioPlanet – An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics. *Front Pharmacol*. 2019;10:445. doi:10.3389/fphar.2019.00445

53. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015;43(D1):D789-D798. doi:10.1093/nar/gku1205

54. Köhler S, Carmody L, Vasilevsky N, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res*. 2019;47(D1):D1018-D1027. doi:10.1093/nar/gky1105

55. Davis AP, Grondin CJ, Johnson RJ, et al. Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res*. 2021;49(D1):D1138-D1143. doi:10.1093/nar/gkaa891

56. Thorn CF, Klein TE, Altman RB. PharmGKB: the Pharmacogenomics Knowledge Base. *Methods Mol Biol Clifton NJ*. 2013;1015:311-320. doi:10.1007/978-1-62703-435-7_20

57. Stathias V, Turner J, Koleti A, et al. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Res*. 2020;48(D1):D431-D439. doi:10.1093/nar/gkz1023

58. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37(1):1-13. doi:10.1093/nar/gkn923

59. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211

60. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44(W1):W90-W97. doi:10.1093/nar/gkw377

61. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*. 2010;11(2):R14. doi:10.1186/gb-2010-11-2-r14

62. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102

63. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*. 2006;22(13):1600-1607. doi:10.1093/bioinformatics/btl140

64. Bauer S, Grossmann S, Vingron M, Robinson PN. Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*. 2008;24(14):1650-1651. doi:10.1093/bioinformatics/btn250

65. Garcia-Moreno A, López-Domínguez R, Villatoro-García JA, et al. Functional Enrichment Analysis of Regulatory Elements. *Biomedicines*. 2022;10(3):590. doi:10.3390/biomedicines10030590

66. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet*. 2011;12(1):56-68. doi:10.1038/nrg2918

67. Rivas JDL, Fontanillo C. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLOS Comput Biol*. 2010;6(6):e1000807. doi:10.1371/journal.pcbi.1000807

68. Orchard S, Ammari M, Aranda B, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*. 2014;42(D1):D358-D363. doi:10.1093/nar/gkt1115

69. Oughtred R, Rust J, Chang C, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci*. 2021;30(1):187-200. doi:10.1002/pro.3978

70. Ramazi S, Allahverdi A, Zahiri J. Evaluation of post-translational modifications in histone proteins: A review on histone modification defects in developmental and neurological disorders. *J Biosci*. 2020;45:135.

71. Morris KV, Mattick JS. The rise of regulatory RNA. *Nat Rev Genet*. 2014;15(6):423-437. doi:10.1038/nrg3722

72. Bushati N, Cohen SM. microRNA Functions. *Annu Rev Cell Dev Biol*. 2007;23(1):175-205. doi:10.1146/annurev.cellbio.23.090506.123406

73. Statello L, Guo CJ, Chen LL, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol*. 2021;22(2):96-118. doi:10.1038/s41580-020-00315-9

74. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*. 2012;13(9):613-626. doi:10.1038/nrg3207

75. Lambert SA, Jolma A, Campitelli LF, et al. The Human Transcription Factors. *Cell*. 2018;172(4):650-665. doi:10.1016/j.cell.2018.01.029

76. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res*. 2019;29(8):1363-1375. doi:10.1101/gr.240663.118

77. Han H, Cho JW, Lee S, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res*. 2018;46(D1):D380-D386. doi:10.1093/nar/gkx1013

78. Lesurf R, Cotto KC, Wang G, et al. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res*. 2016;44(D1):D126-D132. doi:10.1093/nar/gkv1203

79. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res*. 2018;46(D1):D794-D801. doi:10.1093/nar/gkx1081

80. Keenan AB, Torre D, Lachmann A, et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res*. 2019;47(W1):W212-W224. doi:10.1093/nar/gkz446

81. Fornes O, Castro-Mondragon JA, Khan A, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2020;48(D1):D87-D92. doi:10.1093/nar/gkz1001

82. Kulakovskiy IV, Vorontsov IE, Yevshin IS, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res*. 2018;46(D1):D252-D259. doi:10.1093/nar/gkx1106

83. Lachmann A, Giorgi FM, Lopez G, Califano A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*. 2016;32(14):2233-2235. doi:10.1093/bioinformatics/btw216

84. Wolf-Yadlin A, Sevecka M, MacBeath G. Dissecting protein function and signaling using protein microarrays. *Curr Opin Chem Biol*. 2009;13(4):398-405. doi:10.1016/j.cbpa.2009.06.027

85. Božović A, Kulasingam V. Quantitative mass spectrometry-based assay development and validation: From small molecules to proteins. *Clin Biochem*. 2013;46(6):444-455. doi:10.1016/j.clinbiochem.2012.09.024

86. Alvarez MJ, Shen Y, Giorgi FM, et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet*. 2016;48(8):838-847. doi:10.1038/ng.3593

87. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. Network component analysis: Reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci*. 2003;100(26):15522-15527. doi:10.1073/pnas.2136632100

88. Aibar S, González-Blas CB, Moerman T, et al. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods*. 2017;14(11):1083-1086. doi:10.1038/nmeth.4463

89. Su K, Katebi A, Kohar V, et al. NetAct: a computational platform to construct core transcription factor regulatory networks using gene activity. *Genome Biol*. 2022;23(1):270. doi:10.1186/s13059-022-02835-3

90. Chèneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res*. 2018;46(D1):D267-D275. doi:10.1093/nar/gkx1092

91. Holland CH, Szalai B, Saez-Rodriguez J. Transfer of regulatory knowledge from human to mouse for functional genomics analysis. *Biochim Biophys Acta BBA - Gene Regul Mech*. 2020;1863(6):194431. doi:10.1016/j.bbagrm.2019.194431

92. Holland CH, Tanevski J, Perales-Patón J, et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol*. 2020;21(1):36. doi:10.1186/s13059-020-1949-z

93. Perez-Riverol Y, Zorin A, Dass G, et al. Quantifying the impact of public omics data. *Nat Commun*. 2019;10(1):3512. doi:10.1038/s41467-019-11461-w

94. Kröger W, Mapiye D, Entfellner JBD, Tiffin N. A meta-analysis of public microarray data identifies gene regulatory pathways deregulated in peripheral blood mononuclear cells from individuals with Systemic Lupus Erythematosus compared to those without. *BMC Med Genomics*. 2016;9(1):66. doi:10.1186/s12920-016-0227-0

95. Piras IS, Manchia M, Huentelman MJ, et al. Peripheral Biomarkers in Schizophrenia: A Meta-Analysis of Microarray Gene Expression Datasets. *Int J Neuropsychopharmacol*. 2019;22(3):186-193. doi:10.1093/ijnp/pyy103

96. Toro-Domínguez D, Carmona-Sáez P, Alarcón-Riquelme ME. Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis. *Arthritis Res Ther*. 2014;16(6):489. doi:10.1186/s13075-014-0489-x

97. Maksimovic J, Oshlack A, Phipson B. Gene set enrichment analysis for genome-wide DNA methylation data. *Genome Biol*. 2021;22(1):173. doi:10.1186/s13059-021-02388-x

98. Paczkowska M, Barenboim J, Sintupisut N, et al. Integrative pathway enrichment analysis of multivariate omics data. *Nat Commun*. 2020;11(1):735. doi:10.1038/s41467-019-13983-9

99. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. 2013;41(D1):D991-D995. doi:10.1093/nar/gks1193

100. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*. 2007;23(14):1846-1847. doi:10.1093/bioinformatics/btm254

101. Harrison PW, Ahamed A, Aslam R, et al. The European Nucleotide Archive in 2020. *Nucleic Acids Res*. 2021;49(D1):D82-D85. doi:10.1093/nar/gkaa1028

102. Papatheodorou I, Fonseca NA, Keays M, et al. Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res*. 2018;46(D1):D246-D251. doi:10.1093/nar/gkx1158

103. Turvey SE, Broide DH. Innate immunity. *J Allergy Clin Immunol*. 2010;125(2):S24-S32. doi:10.1016/j.jaci.2009.07.016

104. Kantari C, Pederzoli-Ribeil M, Witko-Sarsat V. The Role of Neutrophils and Monocytes in Innate Immunity. *Trends Innate Immun*. 2008;15:118-146. doi:10.1159/000136335

105. Luster AD, Alon R, von Andrian UH. Immune cell migration in inflammation: present and future therapeutic targets. *Nat Immunol*. 2005;6(12):1182-1190. doi:10.1038/ni1275

106. Bonilla FA, Oettgen HC. Adaptive immunity. *J Allergy Clin Immunol*. 2010;125(2):S33-S40. doi:10.1016/j.jaci.2009.09.017

107. Wang L, Wang FS, Gershwin ME. Human autoimmune diseases: a comprehensive update. *J Intern Med*. 2015;278(4):369-395. doi:10.1111/joim.12395

108. Angum F, Khan T, Kaler J, Siddiqui L, Hussain A. The Prevalence of Autoimmune Disorders in Women: A Narrative Review. *Cureus*. 2020;12(5). doi:10.7759/cureus.8094

109. Thomas SL, Griffiths C, Smeeth L, Rooney C, Hall AJ. Burden of Mortality Associated With Autoimmune Diseases Among Females in the United Kingdom. *Am J Public Health*. 2010;100(11):2279-2287. doi:10.2105/AJPH.2009.180273

110. Antonini L, Le Mauff B, Marcelli C, Aouba A, de Boysson H. Rhupus: a systematic literature review. *Autoimmun Rev*. 2020;19(9):102612. doi:10.1016/j.autrev.2020.102612

111. Alarcón-Segovia D, Cardiel MH. Comparison between 3 diagnostic criteria for mixed connective tissue disease. Study of 593 patients. *J Rheumatol*. 1989;16(3):328-334.

112. Sharp GC, Irvin WS, Tan EM, Gould RG, Holman HR. Mixed connective tissue disease-an apparently distinct rheumatic disease syndrome associated with a specific antibody to an extractable nuclear antigen (ENA). *Am J Med*. 1972;52(2):148-159. doi:10.1016/0002-9343(72)90064-2

113. Skopouli FN, Drosos AA, Papaioannou T, Moutsopoulos HM. Preliminary diagnostic criteria for Sjögren's syndrome. *Scand J Rheumatol Suppl*. 1986;61:22-25.

114. Barturen G, Babaei S, Català-Moll F, et al. Integrative Analysis Reveals a Molecular Stratification of Systemic Autoimmune Diseases. *Arthritis Rheumatol*. 2021;73(6):1073-1085. doi:10.1002/art.41610

115. Urowitz MB, Gladman DD, Ibañez D, et al. Evolution of disease burden over five years in a multicenter inception systemic lupus erythematosus cohort. *Arthritis Care Res*. 2012;64(1):132-137. doi:10.1002/acr.20648

116. Tarr T, Dérfalvi B, Győri N, et al. Similarities and differences between pediatric and adult patients with systemic lupus erythematosus. *Lupus*. 2015;24(8):796-803. doi:10.1177/0961203314563817

117. Bengtsson AA, Sturfelt G, Truedsson L, et al. Activation of type I interferon system in systemic lupus erythematosus correlates with disease activity but not with antiretroviral antibodies. *Lupus*. 2000;9(9):664-671. doi:10.1191/096120300674499064

118. Pisetsky DS, Lipsky PE. New insights into the role of antinuclear antibodies in systemic lupus erythematosus. *Nat Rev Rheumatol*. 2020;16(10):565-579. doi:10.1038/s41584-020-0480-7

119. Fava A, Petri M. Systemic lupus erythematosus: Diagnosis and clinical management. *J Autoimmun*. 2019;96:1-13. doi:10.1016/j.jaut.2018.11.001

120. Mikdashi J, Nived O. Measuring disease activity in adults with systemic lupus erythematosus: the challenges of administrative burden and responsiveness to patient concerns in clinical research. *Arthritis Res Ther*. 2015;17(1):183. doi:10.1186/s13075-015-0702-6

121. Dolgin E. Lupus in crisis: as failures pile up, clinicians call for new tools. *Nat Biotechnol*. 2019;37(1):7-8. doi:10.1038/nbt0119-7

122. Furie R, Khamashta M, Merrill JT, et al. Anifrolumab, an Anti–Interferon-α Receptor Monoclonal Antibody, in Moderate-to-Severe Systemic Lupus Erythematosus. *Arthritis Rheumatol*. 2017;69(2):376-386. doi:10.1002/art.39962

123. Allen ME, Rus V, Szeto GL. Leveraging Heterogeneity in Systemic Lupus Erythematosus for New Therapies. *Trends Mol Med*. 2021;27(2):152-171. doi:10.1016/j.molmed.2020.09.009

124. Teruel M, Chamberlain C, Alarcón-Riquelme ME. Omics studies: their use in diagnosis and reclassification of SLE and other systemic autoimmune diseases. *Rheumatology*. 2017;56(suppl_1):i78-i87. doi:10.1093/rheumatology/kew339

125. Baechler EC, Batliwalla FM, Karypis G, et al. Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc Natl Acad Sci*. 2003;100(5):2610-2615. doi:10.1073/pnas.0337679100

126. Bennett L, Palucka AK, Arce E, et al. Interferon and Granulopoiesis Signatures in Systemic Lupus Erythematosus Blood. *J Exp Med*. 2003;197(6):711-723. doi:10.1084/jem.20021553

127. Becker AM, Dao KH, Han BK, et al. SLE Peripheral Blood B Cell, T Cell and Myeloid Cell Transcriptomes Display Unique Profiles and Each Subset Contributes to the Interferon Signature. *PLOS ONE*. 2013;8(6):e67003. doi:10.1371/journal.pone.0067003

128. Nehar-Belaid D, Hong S, Marches R, et al. Mapping systemic lupus erythematosus heterogeneity at the single-cell level. *Nat Immunol*. 2020;21(9):1094-1106. doi:10.1038/s41590-020-0743-0

129. Landolt-Marticorena C, Bonventi G, Lubovich A, et al. Lack of association between the interferon-α signature and longitudinal changes in disease activity in systemic lupus erythematosus. *Ann Rheum Dis*. 2009;68(9):1440-1446. doi:10.1136/ard.2008.093146

130. Petri M, Fu W, Ranger A, et al. Association between changes in gene signatures expression and disease activity among patients with systemic lupus erythematosus. *BMC Med Genomics*. 2019;12(1):4. doi:10.1186/s12920-018-0468-1

131. Banchereau R, Hong S, Cantarel B, et al. Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell*. 2016;165(3):551-565. doi:10.1016/j.cell.2016.03.008

132. Toro-Domínguez D, Martorell-Marugán J, Goldman D, Petri M, Carmona-Sáez P, Alarcón-Riquelme ME. Stratification of Systemic Lupus Erythematosus Patients Into Three Groups of Disease Activity Progression According to Longitudinal Gene Expression. *Arthritis Rheumatol*. 2018;70(12):2025-2035. doi:10.1002/art.40653

133. Ceribelli A, Satoh M, Chan EK. MicroRNAs and autoimmunity. *Curr Opin Immunol*. 2012;24(6):686-691. doi:10.1016/j.coi.2012.07.011

134. Te JL, Dozmorov IM, Guthridge JM, et al. Identification of Unique MicroRNA Signature Associated with Lupus Nephritis. *PLOS ONE*. 2010;5(5):e10344. doi:10.1371/journal.pone.0010344

135. Hou J, Wang P, Lin L, et al. MicroRNA-146a Feedback Inhibits RIG-I-Dependent Type I IFN Production in Macrophages by Targeting TRAF6, IRAK1, and IRAK2. *J Immunol*. 2009;183(3):2150-2158. doi:10.4049/jimmunol.0900707

136. Teruel M, Alarcón-Riquelme ME. The genetic basis of systemic lupus erythematosus: What are the risk factors and what have we learned. *J Autoimmun*. 2016;74:161-175. doi:10.1016/j.jaut.2016.08.001

137. Harley JB, Chen X, Pujato M, et al. Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. *Nat Genet*. 2018;50(5):699. doi:10.1038/s41588-018-0102-3

138. Gan L, O'Hanlon TP, Gordon AS, Rider LG, Miller FW, Burbelo PD. Twins discordant for myositis and systemic lupus erythematosus show markedly enriched autoantibodies in the affected twin supporting environmental influences in pathogenesis. *BMC Musculoskelet Disord*. 2014;15(1):67. doi:10.1186/1471-2474-15-67

139. Javierre BM, Fernandez AF, Richter J, et al. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res*. 2010;20(2):170-179. doi:10.1101/gr.100289.109

140. Joseph S, George NI, Green-Knox B, et al. Epigenome-wide association study of peripheral blood mononuclear cells in systemic lupus erythematosus: Identifying DNA methylation signatures associated with interferon-related genes based on ethnicity and SLEDAI. *J Autoimmun*. 2019;96:147-157. doi:10.1016/j.jaut.2018.09.007

141. Martorell-Marugán J, López-Domínguez R, García-Moreno A, et al. A comprehensive database for integrated analysis of omics data in autoimmune diseases. *BMC Bioinformatics*. 2021;22(1):343. doi:10.1186/s12859-021-04268-4

142. Ziemann M, Eren Y, El-Osta A. Gene name errors are widespread in the scientific literature. *Genome Biol*. 2016;17(1):177. doi:10.1186/s13059-016-1044-7

143. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004;20(3):307-315. doi:10.1093/bioinformatics/btg405

144. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249-264. doi:10.1093/biostatistics/4.2.249

145. Tarazona S, Furió-Tarí P, Turrà D, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res*. 2015;43(21):e140. doi:10.1093/nar/gkv711

146. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):R25. doi:10.1186/gb-2010-11-3-r25

147. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*. 2009;4(8):1184-1191. doi:10.1038/nprot.2009.97

148. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10):1363-1369. doi:10.1093/bioinformatics/btu049

149. Chen Y an, Lemire M, Choufani S, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*. 2013;8(2):203-209. doi:10.4161/epi.23470

150. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinforma Oxf Engl*. 2008;24(13):1547-1548. doi:10.1093/bioinformatics/btn224

151. Teschendorff AE, Marabita F, Lechner M, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinforma Oxf Engl*. 2013;29(2):189-196. doi:10.1093/bioinformatics/bts680

152. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*. 2013;14:293. doi:10.1186/1471-2164-14-293

153. Chaussabel D, Quinn C, Shen J, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity*. 2008;29(1):150-164. doi:10.1016/j.immuni.2008.05.012

154. Hidalgo MR, Cubuk C, Amadoz A, Salavert F, Carbonell-Caballero J, Dopazo J. High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. *Oncotarget*. 2016;8(3):5160-5178. doi:10.18632/oncotarget.14107

155. Liu A, Trairatphisan P, Gjerga E, Didangelos A, Barratt J, Saez-Rodriguez J. From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *Npj Syst Biol Appl*. 2019;5(1):40. doi:10.1038/s41540-019-0118-z

156. Schubert M, Klinger B, Klünemann M, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat Commun*. 2018;9(1):20. doi:10.1038/s41467-017-02391-6

157. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods*. 2016;13(12):966-967. doi:10.1038/nmeth.4077

158. Del Carratore F, Jankevics A, Eisinga R, Heskes T, Hong F, Breitling R. RankProd 2.0: a refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets. *Bioinforma Oxf Engl*. 2017;33(17):2774-2775. doi:10.1093/bioinformatics/btx292

159. Gambino CM, Di Bona D, Aiello A, et al. HLA-C1 ligands are associated with increased susceptibility to systemic lupus erythematosus. *Hum Immunol*. 2018;79(3):172-177. doi:10.1016/j.humimm.2018.01.005

160. Dozmorov MG, Wren JD, Alarcón-Riquelme ME. Epigenomic elements enriched in the promoters of autoimmunity susceptibility genes. *Epigenetics*. 2014;9(2):276-285. doi:10.4161/epi.27021

161. Zollars E, Courtney SM, Wolf BJ, et al. Clinical Application of a Modular Genomics Technique in Systemic Lupus Erythematosus: Progress towards Precision Medicine. *Int J Genomics*. 2016;2016:7862962. doi:10.1155/2016/7862962

162. Romero-Diaz J, Isenberg D, Ramsey-Goldman R. Measures of adult systemic lupus erythematosus: Updated Version of British Isles Lupus Assessment Group (BILAG 2004), European Consensus Lupus Activity Measurements (ECLAM), Systemic Lupus Activity Measure, Revised (SLAM-R), Systemic Lupus Activity Questionnaire for Population Studies (SLAQ), Systemic Lupus Erythematosus Disease Activity Index 2000 (SLEDAI-2K), and Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index (SDI). *Arthritis Care Res*. 2011;63(S11):S37-S46. doi:10.1002/acr.20572

163. Pérez-Gracia JL, Gúrpide A, Ruiz-Ilundain MG, et al. Selection of extreme phenotypes: the role of clinical observation in translational research. *Clin Transl Oncol*. 2010;12(3):174-180. doi:10.1007/s12094-010-0487-7

164. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *J Stat Softw*. 2014;61(1):1-36. doi:10.18637/jss.v061.i06

165. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12(5):453-457. doi:10.1038/nmeth.3337

166. Qin B, Ma N, Tang Q, et al. Neutrophil to lymphocyte ratio (NLR) and platelet to lymphocyte ratio (PLR) were useful markers in assessment of inflammatory response and disease activity in SLE patients. *Mod Rheumatol*. 2016;26(3):372-376. doi:10.3109/14397595.2015.1091136

167. Han BK, Wysham KD, Cain KC, Tyden H, Bengtsson AA, Lood C. Neutrophil and lymphocyte counts are associated with different immunopathological mechanisms in systemic lupus erythematosus. *Lupus Sci Med*. 2020;7(1):e000382. doi:10.1136/lupus-2020-000382

168. Forget P, Khalifa C, Defour JP, Latinne D, Van Pel MC, De Kock M. What is the normal value of the neutrophil-to-lymphocyte ratio? *BMC Res Notes*. 2017;10(1):12. doi:10.1186/s13104-016-2335-5

169. Sergushichev AA. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*. Published online June 20, 2016:060012. doi:10.1101/060012

170. Cao Q, Zhao X, Bai J, et al. Circadian clock cryptochrome proteins regulate autoimmunity. *Proc Natl Acad Sci*. 2017;114(47):12548-12553. doi:10.1073/pnas.1619119114

171. Lin GJ, Huang SH, Chen SJ, Wang CH, Chang DM, Sytwu HK. Modulation by Melatonin of the Pathogenesis of Inflammatory Autoimmune Diseases. *Int J Mol Sci*. 2013;14(6):11742-11766. doi:10.3390/ijms140611742

172. Alarcón-Riquelme ME. New Attempts to Define and Clarify Lupus. *Curr Rheumatol Rep*. 2019;21(4):11. doi:10.1007/s11926-019-0810-4

173. Smith CK, Kaplan MJ. The role of neutrophils in the pathogenesis of systemic lupus erythematosus. *Curr Opin Rheumatol*. 2015;27(5):448. doi:10.1097/BOR.0000000000000197

174. Boulos D, Proudman SM, Metcalf RG, McWilliams L, Hall C, Wicks IP. The neutrophil-lymphocyte ratio in early rheumatoid arthritis and its ability to predict subsequent failure of triple therapy. *Semin Arthritis Rheum*. 2019;49(3):373-376. doi:10.1016/j.semarthrit.2019.05.008

175. Li L, Xia Y, Chen C, Cheng P, Peng C. Neutrophil-lymphocyte ratio in systemic lupus erythematosus disease: a retrospective study. *Int J Clin Exp Med*. 2015;8(7):11026-11031.

176. Tasaki S, Suzuki K, Kassai Y, et al. Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. *Nat Commun*. 2018;9(1):2755. doi:10.1038/s41467-018-05044-4

177. Handunnetthi L, Ramagopalan SV, Ebers GC, Knight JC. Regulation of major histocompatibility complex class II gene expression, genetic variation and disease. *Genes Immun*. 2010;11(2):99-112. doi:10.1038/gene.2009.83

178. Kambayashi T, Laufer TM. Atypical MHC class II-expressing antigen-presenting cells: can anything replace a dendritic cell? *Nat Rev Immunol*. 2014;14(11):719-730. doi:10.1038/nri3754

179. Guimaraes JC, Zavolan M. Patterns of ribosomal protein expression specify normal and malignant human cells. *Genome Biol*. 2016;17(1):236. doi:10.1186/s13059-016-1104-z

180. Goropevšek A, Holcar M, Avčin T. The Role of STAT Signaling Pathways in the Pathogenesis of Systemic Lupus Erythematosus. *Clin Rev Allergy Immunol*. 2017;52(2):164-181. doi:10.1007/s12016-016-8550-y

181. Zhang Z, Shi L, Song L, Ephrem E, Petri M, Sullivan KE. Interferon Regulatory Factor 1 Marks Activated Genes and Can Induce Target Gene Expression in Systemic Lupus Erythematosus. *Arthritis Rheumatol*. 2015;67(3):785-796. doi:10.1002/art.38964

182. Sun SC, Chang JH, Jin J. Regulation of nuclear factor-κB in autoimmunity. *Trends Immunol*. 2013;34(6):282-289. doi:10.1016/j.it.2013.01.004

183. Deng GM, Kyttaris VC, Tsokos GC. Targeting Syk in Autoimmune Rheumatic Diseases. *Front Immunol*. 2016;7. doi:10.3389/fimmu.2016.00078

184. Grammatikos AP, Ghosh D, Devlin A, Kyttaris VC, Tsokos GC. Spleen Tyrosine Kinase (Syk) Regulates Systemic Lupus Erythematosus (SLE) T Cell Signaling. *PLOS ONE*. 2013;8(8):e74550. doi:10.1371/journal.pone.0074550

185. Kyttaris VC, Zhang Z, Kampagianni O, Tsokos GC. Calcium signaling in systemic lupus erythematosus T cells: A treatment target. *Arthritis Rheum*. 2011;63(7):2058-2066. doi:10.1002/art.30353

186. Frischbutter S, Schultheis K, Pätzel M, Radbruch A, Baumgrass R. Evaluation of calcineurin/NFAT inhibitor selectivity in primary human Th cells using bar-coding and phospho-flow cytometry. *Cytometry A*. 2012;81A(11):1005-1011. doi:10.1002/cyto.a.22204

187. Mok CC. Calcineurin inhibitors in systemic lupus erythematosus. *Best Pract Res Clin Rheumatol*. 2017;31(3):429-438. doi:10.1016/j.berh.2017.09.010

188. Mukundan L, Odegaard JI, Morel CR, et al. PPAR-δ senses and orchestrates clearance of apoptotic cells to promote tolerance. *Nat Med*. 2009;15(11):1266-1272. doi:10.1038/nm.2048

189. Handono K, Firdausi SN, Pratama MZ, Endharti AT, Kalim H. Vitamin A improve Th17 and Treg regulation in systemic lupus erythematosus. *Clin Rheumatol*. 2016;35(3):631-638. doi:10.1007/s10067-016-3197-x

190. Abdelhamid L, Luo XM. Retinoic Acid, Leaky Gut, and Autoimmune Diseases. *Nutrients*. 2018;10(8):1016. doi:10.3390/nu10081016

191. Zhang H, Liao X, Sparks JB, Luo XM. Dynamics of Gut Microbiota in Autoimmune Lupus. *Appl Environ Microbiol*. 2014;80(24):7551-7560. doi:10.1128/AEM.02676-14

192. Wei S, Yoshida N, Finn G, et al. Pin1-Targeted Therapy for Systemic Lupus Erythematosus. *Arthritis Rheumatol*. 2016;68(10):2503-2513. doi:10.1002/art.39741

193. Zhang Q, Lenardo MJ, Baltimore D. 30 Years of NF-κB: A Blossoming of Relevance to Human Pathobiology. *Cell*. 2017;168(1):37-57. doi:10.1016/j.cell.2016.12.012

194. Tang Y, Xie H, Chen J, et al. Activated NF-κB in Bone Marrow Mesenchymal Stem Cells from Systemic Lupus Erythematosus Patients Inhibits Osteogenic Differentiation Through Downregulating Smad Signaling. *Stem Cells Dev*. 2012;22(4):668-678. doi:10.1089/scd.2012.0226

195. Shih VFS, Davis-Turak J, Macal M, et al. Control of RelB during dendritic cell activation integrates canonical and noncanonical NF-κB pathways. *Nat Immunol*. 2012;13(12):1162-1170. doi:10.1038/ni.2446

196. Wu H, Lo Y, Chan A, Law KS, Mok MY. Rel B-modified dendritic cells possess tolerogenic phenotype and functions on lupus splenic lymphocytes in vitro. *Immunology*. 2016;149(1):48-61. doi:10.1111/imm.12628

197. Nakamura T, Ushiyama C, Suzuki S, et al. Urinary Podocytes for the Assessment of Disease Activity in Lupus Nephritis. *Am J Med Sci*. 2000;320(2):112-116. doi:10.1097/00000441-200008000-00009

198. Kanemoto K, Takahashi S, Shu Y, et al. Variable expression of podocyte-related markers in the glomeruloid bodies in Wilms tumor. *Pathol Int*. 2003;53(9):596-601. doi:10.1046/j.1440-1827.2003.01526.x

199. Liao R, Liu Q, Zheng Z, et al. Tacrolimus Protects Podocytes from Injury in Lupus Nephritis Partly by Stabilizing the Cytoskeleton and Inhibiting Podocyte Apoptosis. *PLOS ONE*. 2015;10(7):e0132724. doi:10.1371/journal.pone.0132724

200. Willis SN, Tellier J, Liao Y, et al. Environmental sensing by mature B cells is controlled by the transcription factors PU.1 and SpiB. *Nat Commun*. 2017;8(1):1-14. doi:10.1038/s41467-017-01605-1

201. Zhang R, Wang L, Pan J hong, Han J. A critical role of E2F transcription factor 2 in proinflammatory cytokines-dependent proliferation and invasiveness of fibroblast-like synoviocytes in rheumatoid Arthritis. *Sci Rep*. 2018;8(1):2623. doi:10.1038/s41598-018-20782-7

202. Jia W, Wu W, Yang D, et al. GATA4 regulates angiogenesis and persistence of inflammation in rheumatoid arthritis. *Cell Death Dis*. 2018;9(5):1-15. doi:10.1038/s41419-018-0570-5

203. Lyssenko V, Lupi R, Marchetti P, et al. Mechanisms by which common variants in the *TCF7L2* gene increase risk of type 2 diabetes. *J Clin Invest*. 2007;117(8):2155-2163. doi:10.1172/JCI30706

204. Lopez-Dominguez R, Toro-Dominguez D, Martorell-Marugan J, et al. Transcription Factor Activity Inference in Systemic Lupus Erythematosus. *Life*. 2021;11(4):299. doi:10.3390/life11040299

205. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. Published online 2021. doi:10.1016/j.cell.2021.04.048

206. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16(12):1289-1296. doi:10.1038/s41592-019-0619-0

207. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech Theory Exp*. 2008;2008(10):P10008. doi:10.1088/1742-5468/2008/10/P10008

208. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun*. 2022;13(1):1246. doi:10.1038/s41467-022-28803-w

209. Squair JW, Gautier M, Kathe C, et al. Confronting false discoveries in single-cell differential expression. *Nat Commun*. 2021;12(1):5692. doi:10.1038/s41467-021-25960-2

210. Besliu A, Banica L, Matache C, et al. Systemic Lupus Erythematosus: the involvement of PI3K/Akt/mTOR pathway in cellular cycle and in early apoptosis. *Joint Bone Spine*. 2008;75(2):246. doi:10.1016/j.jbspin.2008.01.012

211. Nguyen V, Cudrici C, Zernetkina V, et al. TRAIL, DR4 and DR5 are upregulated in kidneys from patients with lupus nephritis and exert proliferative and proinflammatory effects. *Clin Immunol*. 2009;132(1):32-42. doi:10.1016/j.clim.2009.02.011

212. Brant EJ, Rietman EA, Klement GL, Cavaglia M, Tuszynski JA. Personalized therapy design for systemic lupus erythematosus based on the analysis of protein-protein interaction networks. *PLOS ONE*. 2020;15(3):e0226883. doi:10.1371/journal.pone.0226883

213. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):15. doi:10.1186/s13059-017-1382-0

214. Amezquita RA, Lun ATL, Becht E, et al. Orchestrating single-cell analysis with Bioconductor. *Nat Methods*. 2020;17(2):137-145. doi:10.1038/s41592-019-0654-x

215. Green MRJ, Kennell ASM, Larche MJ, Seifert MH, Isenberg DA, Salaman MR. Natural killer T cells in families of patients with systemic lupus erythematosus: Their possible role in regulation of IGG production. *Arthritis Rheum*. 2007;56(1):303-310. doi:10.1002/art.22326

216. Tucci M, Quatraro C, Lombardi L, Pellegrino C, Dammacco F, Silvestris F. Glomerular accumulation of plasmacytoid dendritic cells in active lupus nephritis: Role of interleukin-18. *Arthritis Rheum*. 2008;58(1):251-262. doi:10.1002/art.23186

217. Blanco P, Palucka AK, Gill M, Pascual V, Banchereau J. Induction of Dendritic Cell Differentiation by IFN-α in Systemic Lupus Erythematosus. *Science*. 2001;294(5546):1540-1543. doi:10.1126/science.1064890

218. Kawashima H, Takatori H, Suzuki K, et al. Tumor Suppressor p53 Inhibits Systemic Autoimmune Diseases by Inducing Regulatory T Cells. *J Immunol*. 2013;191(7):3614-3623. doi:10.4049/jimmunol.1300509

219. Herkel J, Kam N, Erez N, et al. Monoclonal antibody to a DNA-binding domain of p53 mimics charge structure of DNA: anti-idiotypes to the anti-p53 antibody are anti-DNA. *Eur J Immunol*. 2004;34(12):3623-3632. doi:10.1002/eji.200425371

220. Guiducci C, Ghirelli C, Marloie-Provost MA, et al. PI3K is critical for the nuclear translocation of IRF-7 and type I IFN production by human plasmacytoid predendritic cells in response to TLR activation. *J Exp Med*. 2008;205(2):315-322. doi:10.1084/jem.20070763

221. Voynova E, Qi CF, Scott B, Bolland S. Cutting Edge: Induction of Inflammatory Disease by Adoptive Transfer of an Atypical NK Cell Subset. *J Immunol*. 2015;195(3):806-809. doi:10.4049/jimmunol.1500540

222. Cruz-González D de J, Gómez-Martin D, Layseca-Espinosa E, et al. Analysis of the regulatory function of natural killer cells from patients with systemic lupus erythematosus. *Clin Exp Immunol*. 2018;191(3):288-300. doi:10.1111/cei.13073

10 ACTIVIDAD CIENTÍFICA

Publicados:

- Garcia-Moreno, Adrian, Raul López-Domínguez, Juan Antonio Villatoro-García, Alberto Ramirez-Mena, Ernesto Aparicio-Puerta, Michael Hackenberg, Alberto Pascual-Montano, and Pedro Carmona-Saez. 2022. "Functional Enrichment Analysis of Regulatory Elements." Biomedicines 10 (3): 590. https://doi.org/10.3390/biomedicines10030590. JCR: Q2
- Lopez-Dominguez, Raul, Daniel Toro-Dominguez, Jordi Martorell-Marugan, Adrian Garcia-Moreno, Christian H. Holland, Julio Saez-Rodriguez, Daniel Goldman, Michelle A. Petri, Marta E. Alarcon-Riquelme, and Pedro Carmona-Saez. 2021. "Transcription Factor Activity Inference in Systemic Lupus Erythematosus." Life (Basel, Switzerland) 11 (4): 299. https://doi.org/10.3390/life11040299. JCR: Q2
- Martorell-Marugán, Jordi, Raúl López-Domínguez, Adrián García-Moreno, Daniel Toro-Domínguez, Juan Antonio Villatoro-García, Guillermo Barturen, Adoración Martín-Gómez, et al. 2021. "A Comprehensive Database for Integrated Analysis of Omics Data in Autoimmune Diseases." BMC Bioinformatics 22 (1): 343. https://doi.org/10.1186/s12859-021-04268-4. JCR: Q2
- Martorell-Marugán, Jordi, Juan Antonio Villatoro-García, Adrián García-Moreno, Raúl López-Domínguez, Francisco Requena, Juan Julián Merelo, Marina Lacasaña, et al. 2021. "DatAC: A Visual Analytics Platform to Explore Climate and Air Quality Indicators Associated with the COVID-19 Pandemic in Spain." Science of The Total Environment 750 (January): 141424. https://doi.org/10.1016/j.scitotenv.2020.141424. JCR: D1
- Peris-Torres, Carlos, María del Carmen Plaza-Calonge, Raúl López-Domínguez, Silvia Domínguez-García, Antonio Barrientos-Durán, Pedro Carmona-Sáez, and Juan Carlos Rodríguez-Manzaneque. 2020. "Extracellular Protease ADAMTS1 Is Required at Early Stages of Human Uveal Melanoma Development by Inducing Stemness and Endothelial-Like Features on Tumor Cells." Cancers 12 (4): 801. https://doi.org/10.3390/cancers12040801. JCR: Q1

- Ramos-Molina, Bruno, Lidia Sánchez-Alcoholado, Amanda Cabrera-Mulero, Raul Lopez-Dominguez, Pedro Carmona-Saez, Eduardo Garcia-Fuentes, Isabel Moreno-Indias, and Francisco J. Tinahones. 2019. "Gut Microbiota Composition Is Associated With the Global DNA Methylation Pattern in Obesity." Frontiers in Genetics 10. https://doi.org/10.3389/fgene.2019.00613. JCR: Q2
- Toro-Domínguez, Daniel, Raúl Lopez-Domínguez, Adrián García Moreno, Juan A. Villatoro-García, Jordi Martorell-Marugán, Daniel Goldman, Michelle Petri, et al. 2019. "Differential Treatments Based on Drug-Induced Gene Expression Signatures and Longitudinal Systemic Lupus Erythematosus Stratification." Scientific Reports 9 (1): 15502. https://doi.org/10.1038/s41598-019-51616-9. JCR: Q1
- Toro-Domínguez, Daniel, Jordi Martorell-Marugán, Raúl López-Domínguez, Adrián García-Moreno, Víctor González-Rumayor, Marta E. Alarcón-Riquelme, and Pedro Carmona-Sáez. 2019. "ImaGEO: Integrative Gene Expression Meta-Analysis from GEO Database." Bioinformatics 35 (5): 880–82. https://doi.org/10.1093/bioinformatics/bty721. JCR: D1
- Toro-Domínguez, Daniel, Jordi Martorell-Marugán, Manuel Martinez-Bueno, Raúl López-Domínguez, Elena Carnero-Montoro, Guillermo Barturen, Daniel Goldman, Michelle Petri, Pedro Carmona-Sáez, and Marta E Alarcón-Riquelme. 2022. "Scoring Personalized Molecular Portraits Identify Systemic Lupus Erythematosus Subtypes and Predict Individualized Drug Responses, Symptomatology and Disease Progression." Briefings in Bioinformatics 23 (5): bbac332. https://doi.org/10.1093/bib/bbac332. JCR: D1

En revisión:

- Olivia Castellini-Pérez, Dr Guillermo Barturen, Manuel Martínez-Bueno, Dr Andrii Iakovliev, Dr Martin Kerick , Raúl López-Domínguez , Dr Concepción Marañón , Dr Javier Martín , Dr Esteban Ballestar , Dr María Orietta Borghi , Dr Weiliang Qiu , Dr Cheng Zhu , Srinivas Shankara , Athina Spiliopolou , Emanuele de Rinaldis , Professor Marta Alarcón-Riquelme. "The Lupus Epigenome Relates to Genetics, Transcription and Serological Profiles with Dependency on Molecular Subtypes and Informs Drug Discovery." Npj Genomic Medicine. JCR: Q1
- Villatoro-García, Juan Antonio, López-Domínguez Raúl, Martorel-Marugán, Jordi, de Dios Luna, Juan, Lorente, Jose Antonio, Carmona-Saez, Pedro. "Exploring the Interplay

between Climate, Population Immunity and SARS-CoV-2 Transmission Dynamic". Science of The Total Environment. JCR: D1